Theoretical Chemistry Thesis

Computational Studies on Molecular Recognition in Chymosin Complexes and Related Systems

Author:

Samiul M. Ansari

Supervisor:

Dr. David S. Palmer

A thesis submitted to the University of Strathclyde

in partial fulfilment of the requirements

for the degree of

Doctor of Philosophy

Palmer Research Group

Department of Pure and Applied Chemistry

University of Strathclyde

February 2018

"Did you hear about the rose that grew from a crack in the concrete?
Proving nature's laws wrong, it learned to walk without having feet.
Funny, it seems, to by keeping its dreams; it learned to breathe fresh air.
Long live the rose that grew from concrete, when no one else even cared."

-Tupac Amaru Shakur (Lesane Parish Crooks)

Acknowledgements

First of all, I would like to give a very special thanks to my supervisor, Dr David S. Palmer, for his time, encouragement, support and empathy while also making this Ph.D. a wonderful experience. The guidance, trust and advice has proved to be invaluable in completing the work carried out for this thesis.

Next, I would like to acknowledge my exceptional colleagues, in particular Benjamin Smith, Maksim Mišin, Lucia Fusani and Christopher Faulkner, who have made my Ph.D. a most enjoyable experience. During the past three years at the University of Strathclyde I have had the privilege of meeting a number of brilliant individuals from all around the world and I would like to thank each and every one of them for their time, advice and friendship.

I would also like to thank my family, in particular my parents, Selina and Mohammad Mosaddeq, and my grandmother, Saira Chowdhury. Without their love, encouragement and belief I would not have made it this far.

Finally, I thank my sweet, caring Halima, whose support during this Ph.D. is remembered and so appreciated.

Thank you.

Many thanks to the University of Strathclyde for support through its Strategic Appointment and Investment Scheme. Computations were performed through *ESPRC* funded *ARCHIE-WeST High Performance Computer* (www.archie-west.ac.uk, ESPRC grant no. EP K0005861).

Declaration of Authorship

I declare that, except where specifically indicated, this thesis, titled 'Computational Studies on Molecular Recognition in Chymosin Complexes and Related Systems' and the work presented in it are my own and I am the sole author of all parts.

This thesis is the result of original research. It has been composed by me and has not been previously submitted for examination.

Sign: .....

Date: .....

UNIVERSITY OF STRATHCLYDE

Abstract

Department of Pure and Applied Chemistry

Doctor of Philosophy

Computational Studies on Molecular Recognition in Chymosin Complexes and Related Systems

by Samiul M. Ansari

This thesis focuses on molecular recognition in chymosin complexes using various computational approaches used for studying protein-ligand systems. Three computational investigations are presented in this thesis.

The first research project, titled, '*Allosteric-Activation Mechanism Of Bovine Chymosin*', is presented in chapter 5. The study investigates the aspartic protease, bovine chymosin, which catalyses the proteolysis of κ-casein proteins in milk. The research presented in this chapter employed two computational techniques, molecular dynamics and bias exchange metadynamics simulations, to study the mechanism of allosteric-activation and to compute the free energy surface for the process. The simulations reveal that allosteric activation is initiated by interactions between the HPHPH sequence of κ-casein and a small α-helical region of chymosin (residues 112-116). A small conformational change in the α-helix causes the side chain of Phe114 to vacate a pocket that may then be occupied by the side chain of Tyr77. The free energy surface for the self-inhibited to open transition is significantly altered by the presence of the HPHPH sequence of κ-casein.

The second research project, named, '*Effect of Mutations in Bovine or Camel Chymosin on the Thermodynamics of Binding κ-Caseins*', is presented in chapter 6. Both bovine and camel chymosin catalyse the proteolysis of a milk protein, κ-casein, which helps to initiate milk coagulation. The research in this chapter reports computational alanine scanning calculations in four chymosin–κ-casein complexes, helping to elucidate the influence that individual residues have on the protein-ligand binding thermodynamics. Of the 12 sequence differences in the binding sites of bovine and camel chymosin, eight are shown to be particularly important for understanding differences in the binding thermodynamics (Asp112Glu, Lys221Val, Gln242Arg, Gln278Lys. Glu290Asp, His292Asn, Gln294Glu, and Lys295Leu. Residue in bovine chymosin written first).

The final research project of this thesis titled, '*Comparative Molecular Field Analysis using Molecular Integral Equation Theory*', is delivered in chapter 7. The study reports, and thoroughly benchmarks, a new method for 3D-QSAR that uses a classical statistical mechanics based solvent model combined with machine learning. Recently, Güssregen *et al.* used solute–solvent distribution functions calculated by the 3D Reference Interaction Site Model (3D-RISM) in a 3D-QSAR model to predict the binding affinities of serine protease inhibitors. The work carried out for this thesis extends this idea by introducing *probe atoms* into the 3D-RISM solvent model in order to capture other molecular interactions in addition to those related to hydration/dehydration. The CARMa models have been thoroughly benchmarked against other 3D-QSAR methods across six different datasets, demonstrating that CARMa is an extremely robust method, outperforming other field-based QSAR methods.

Abbreviations

AMBER	**A**ssisted **M**odel **B**uilding with **E**nergy **R**efinement
ANN	**A**rtificial **N**eural **N**etworks
BEMD	**B**ias-**E**xchange **M**eta-**D**ynamics
BTR	**B**agging **T**ree **R**egression
CADD	**C**omputer **A**ided **D**rug **D**esign
CARMa	**C**omparative **A**nalysis of 3D-**R**ism **Ma**ps
CART	**C**lassification **A**nd **R**egression **T**rees
CoMFA	**Co**mparative **M**olecular **F**ield **A**nalysis
CoMSIA	**Co**mparative **M**olecular **S**imilarity **I**ndices **A**nalysis
CV	**C**ross-**V**alidation
CVFF	**C**onsistent **V**alence **F**orce **F**ield
FEP	**F**ree **E**nergy **P**ertubation
FES	**F**ree-**E**nergy **S**urface
GA	**G**enetic **A**lgorithm
GAFF	**G**eneral **A**MBER **F**orce **F**ield
GBSA	**G**eneralised **B**orn **S**urface **A**rea
GF	**G**aussian **F**luctuations
HNC	**H**yper-**N**etted **C**hain
IET	**I**ntegral **E**quation **T**heory
IFST	**I**nhomogeneous **F**luid **S**olvation **T**heory
KH	**K**ovalenko-**H**irata
KNN	**k**-**N**earest **N**eighbours
LIE	**L**inear-**I**nteraction-**E**nergy
LOO	**L**eave-**O**ne-**O**ut

MARSplines	**M**ultivariate **A**daptive **R**egression **Splines**
MD	**M**olecular **D**ynamics
MDIIS	**M**odified **D**irect **I**nversion of the **I**terative **S**ubspace
MLR	**M**ultiple **L**inear **R**egression
MM	**M**olecular **M**echanics
MOZ	**M**olecular **O**rnstein-**Z**ernike
NB	**N**aive **B**ayes
NN	**N**eural **N**etworks
OLR	**O**rdinary **L**inear **R**egression
PBSA	**P**oisson-**B**oltzmann **S**urface **A**rea
PC	**P**ressure **C**orrection
PC+	**P**ressure **C**orrection Plus
PCA	**P**rincipal **C**omponent **A**nalysis
PCFF	**P**olymer **C**onsistent **F**orce **F**ield
PCR	**P**rincipal **C**omponent **R**egression
PLHNC	**P**artial-**L**inearised **H**yper-**N**etted **C**hain
PLS	**P**artial **L**east **S**quares
PME	**P**article **M**esh **E**wald
PR	**P**enalised **R**egression
PSE	**P**artial **S**eries **E**xpansion
QSAR	**Q**uantitative **S**tructure **A**ctivity **R**elationship
RESP	**R**estrained **E**lectro-**S**tatic **P**otential
RF	**R**andom **F**orest
RISM	**R**eference **I**nteraction **S**ite **M**odel
SAR	**S**tructure **A**ctivity **R**elationship
SFE	**S**olvation **F**ree **E**nergy
SVM	**S**upport **V**ector **M**achine
TI	**T**hermodynamic **I**ntegration
VMD	**V**isual **M**olecular **D**ynamics

Contents

List of Figures

List of Tables

Part I

Introduction

Chapter 1

Overview

1.1 Chymosin (Rennin)

Chymosin is a mammalian digestive enzyme used in the manufacturing of cheese in industry. Archaeological findings dating back to 6000 BC suggest cheese manufacture, via chymosin, is possibly one of the earliest biotechnological applications conducted by humans.[4,5] Cave paintings from the Libyan Sahara (5500 – 2000 BC) and Sumerian relief (3500 – 2800 BC) depict a clear process of curding milk into cheese. Forensic tests on ancient Egyptian pottery dating between 3000 – 2800 BC found that the earthenware was used to store cheese.[6] Historic documents from ancient Rome report that by the time the Roman Empire reached its height, cheese production via chymosin was a well-established process.[5]

Evidence suggests that the earliest production of cheese may have been discovered by accident when milk was being stored in bags made from ruminant calf stomachs which contained traces of proteolytic enzymes.[7] The first recorded attempts to isolate the active enzyme were made in 1840, by a French pharmacist named Jean-Baptist Deschamps.[8] The name given to the enzyme was *Chymosine*, derived from the ancient Greek word for juice – *khymos*[9] (which may be due to the stomach contents appearing as juice). In 1890, Lea and Dickinson suggested

the name *rennin*, derived from the word *rennet*. However, in 1970, Foltmann returned to the prior nomenclature of chymosin to avoid confusion with *renin*, an enzyme found in kidneys.[10]

Naturally found in the stomach of young mammals, chymosin is an acid stable peptidase. As a member of the aspartic protease (also called aspartic peptidase) family it is closely related to pepsin, which is found in adult mammals. The enzymes primary function in nature is to aid digestion by selectively cleaving κ-casein proteins in milk to initiate coagulation. In industry, this procedure is exploited to initiate milk-clotting in the first stages of cheese manufacture.[11]

Latest statistics value the global cheese industry at $92 billion and it is expected to grow at a steady rate of 3% over the next 5 years due to gradual increases in demand from emerging countries.[12] Globally, over 22.5 million metric tonnes of cheese is produced annually, making cheese production the largest driver of growth in the dairy industry, as shown by Figure 1.1.[13]

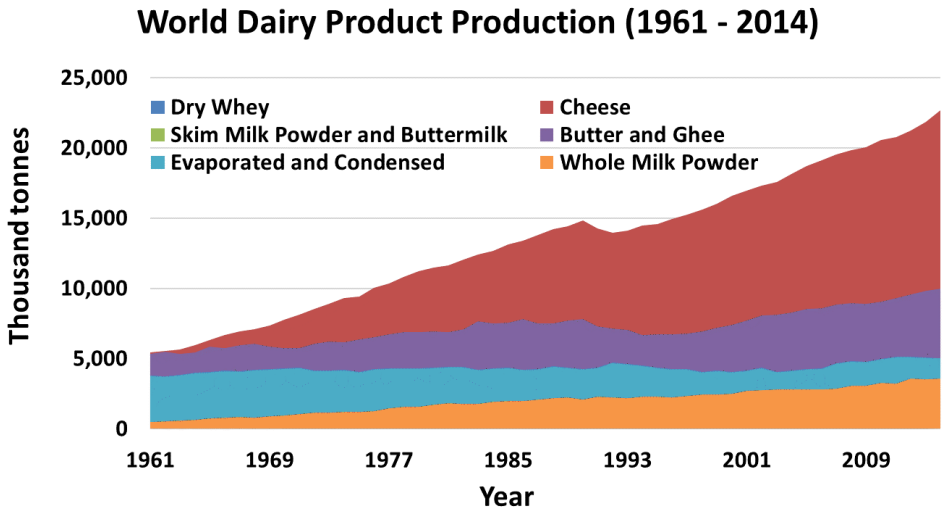


FIGURE 1.1: Cumulative world dairy production between the years of 1961 and 2014. Displays the dominance of cheese in dairy product production.

Bovine chymosin has been marketed towards the manufacturing of cheese over the past half century. However, in recent times it has been discovered that the camel

variant of the enzyme has a 70% higher clotting activity for bovine κ-casein.[11] It has also been demonstrated that camel chymosin has just 20% of the unspecific proteolytic activity that bovine chymosin shows in bovine milk.[11] By comparison, bovine chymosin performs very poorly in camel milk. Both bovine and camel chymosin are now marketed towards the food industry as enzymes that initiate milk clotting.[11]

Although the physiological effects of the two enzymes in nature are known, there are several aspects of their biological functions that are not understood at a molecular level.[2] Moreover, the reasons for the disparity in catalytic propensity between bovine and camel chymosin have not been fully explained.[14]

The crystal structures of bovine chymosin have revealed that residue Tyr77 can occupy two different positions. It can be extended over the binding pocket, referred to as the self-inhibited position, or it can be extended back into the β-hairpin flap, known as the open conformation.[2] The limited information in the literature makes it difficult to be able to deduce the dynamic procedure to convert between the two known conformations.

Experimental studies using a pentapeptide -*His-Pro*- fragment (His98-His102) of κ-casein incubated with bovine chymosin, have shown up to a 200-fold increase in the catalytic rate for hydrolysis in comparison to fragments of the native substrate of varied lengths.[15] The pentapeptide cluster in camel κ-casein is *Arg-Pro-Arg-Pro-Arg*, whereas in bovine κ-casein it is *His-Pro-His-Pro-His*. The His-Pro cluster is suggested to act as an allosteric-activator on the self-inhibited chymosin, converting it to the open conformation.[16] This theory concurs with experimental investigations which demonstrated the poor performance of bovine chymosin in camel milk. Furthermore, it is also supported by mutagenesis studies which suggest the His-Pro residues themselves are important for catalysis.[17,18]

The catalytic mechanism for the cleavage of κ-casein by chymosin is still disputed.

Although there are a number of theories on how it may occur, it is widely accepted that a nucleophilic attack on the carbonyl carbon of κ-casein via a catalytic water results in hydrolysis. However, the lack of experimental data to support this theory makes the exact details of the mechanism speculative. There are numerous other theories on the catalytic mechanism, but due to disagreement within the scientific community none are definitive.[19,20]

An improved understanding of the structures, complexes and allostery will help design mutants of the enzyme with enhanced properties, compared to the natural version. The economic impact of this is substantial as it can lead to an increase in output from industry. Designing mutant chymosin enzymes that can efficiently act on milk from mammals that are more abundant in impoverished regions of the world will help to increase cheese production. The increased yield would in turn help reduce the cost of cheese and allow for it to be more readily available across the world without added cost of transportation over long distances. Therefore, this would enable less fortunate regions of the world to assure food security, providing a humanitarian incentive to investigate these unknowns.

1.2 Aims and Objectives

This thesis explores and investigates various computational techniques used to study protein-ligand binding with an emphasis on the aspartic protease, chymosin, and approaches using the three-dimensional reference interaction site model (3D-RISM).

The first research chapter of this thesis is titled *Allosteric-Activation Mechanism Of Bovine Chymosin*. The main objective of the research conducted in this chapter is to determine the activation mechanism that takes place in chymosin to allow

for protein-ligand binding. To achieve this, molecular dynamics (MD) and bias-exchange metadynamics (BEMD) simulations were used to reveal the allosteric-activation mechanism and its associated free energy surface. The BEMD simulations have been performed by collaborators in Prof Birgit Schiøtt's group at Aarhus University. However, the analysis of the BEMD simulations presented in this thesis are the author's own.

The second research chapter titled, *Effect of Mutations in Bovine or Camel Chymosin on the Thermodynamics of Binding κ-Caseins*, investigates the protein-ligand binding thermodynamics in chymosin–κ-casein complexes. The study investigates the importance of individual amino acids in chymosin through single-point mutations to calculate the influence the residues have on the binding free energy of chymosin to κ-casein. The research aims to identify individual residues that can be mutated to effect binding thermodynamics to favour complexation. Furthermore, the study assesses the use of 3D-RISM methods for calculating binding free energy.

The final research chapter titled *Comparative Molecular Field Analysis using Molecular Integral Equation Theory*, reports the development and benchmarking of a novel 3D-QSAR technique. The purpose of this research is to develop a new 3D-QSAR method that uses a classical statistical mechanics based solvent model combined with machine learning to predict protein-ligand binding related properties. This is done by further developing a recently presented method (CARMa) by introducing probe atoms into the 3D-RISM solvent model in order to capture other molecular interactions in addition to those related to hydration/dehydration. The influence of algorithmic parameters, such as the 3D-RISM bridge-functional and grid-size, on the prediction accuracy are systematically investigated. The new method is benchmarked and results are compared to those in the literature.

1.3 Thesis Structure

The first part of this thesis is dedicated to reviewing background theory, literature results and computational protocols. In summary, chapter 2 is dedicated to the enzyme, chymosin, providing a review of its history, structure, activity, commercial production and literature investigations. Chapter 3 is a review of quantitative structure activity relationships (QSAR) with a focus on 3D-QSAR and CoMFA. The fourth chapter surveys the computational theory for molecular mechanics (MM), molecular dynamics (MD), reference interaction site model (RISM), machine learning and regression.

The second part of the thesis presents three research chapters and is concerned with new findings, discussions of the analysis about the research conducted and describe the conclusions that are made. Here, chapter 5 is concerned with the allosteric-activation mechanism of bovine chymosin. The sixth chapter presents the effects on binding thermodynamics for mutations in bovine and camel chymosin. Chapter 7 introduces an extension to CARMa and benchmarks the use of molecular integral equation theory in QSAR.

The thesis is complemented with appendices that outline additional calculations relevant to the research alongside additional tables and figures that have not been presented as part of the research chapters.

Part II

Background

Chapter 2

Chymosin

The mammalian aspartic protease, bovine chymosin, is an enzyme that aids digestion by selectively cleaving the milk protein κ-casein.[21] Chymosin is released in the fourth stomach of calves as an inactive enzyme or zymogen, referred to as pro-chymosin. Once pro-chymosin is exposed to an acidic environment inside the stomach, a 43 residue pro-peptide in the N-terminus is proteolytically cleaved to form chymosin, the active enzyme. The primary function of this enzyme is to catalytically convert the milk protein caseinogen into para-casein, which precipitates out in the stomach as a calcium salt.[22] This precipitate forms a firm curd to ensure milk remains in the stomach long enough to be exposed to other proteolytic enzymes and the gastric juice. This process ensures maximum absorption of nutrients for young mammals.[23]

2.1 History

In 1972, Christian Ditlev Ammentorp Hansen, a Danish pharmacist, was awarded the University of Copenhagen gold medal for his research on developing a chemical treatise. He developed a procedure to extract pure and functional chymosin

enzymes from calves and revolutionised the dairy industry.[24] This development directly resulted in the establishment of the first chymosin production factory in 1974.

In 1981, Burkhalter listed around 500 different varieties of cheese produced from cows milk alone,[25] and in 1993, Kalantzopoulos listed a further 500 more, produced from sheep and/or goats milk.[5] This displays, to some extent, the progression of research into the cheeses throughout modern history. However, it is still evident that much is not yet understood about the molecular mechanism involved. Important information such as the significance of the tertiary structure which promotes chemical activity and the mechanism of action which takes place in unknown. The information in the literature regarding the chemical environment in which chymosin activity is most effective is limited and the knowledge behind the self-inhibiting structure of chymosin is inconclusive.

2.2 Structural Chemistry

2.2.1 Primary Sequence Structure

Chymosin exists as a single strand polypeptide chain consisting of 323 amino acid residues with a molecular weight of around 35,000.[21] The enzyme is rich in dicarboxylic and β-hydroxy amino acids but has a low content of basic residues.[26–28]

Chymosin exists as three isozymes, chymosin A, chymosin B and chymosin C. Chymosin A has an aspartic acid residue at position 244 whereas in chymosin B a glycine residue occupies this position. This results in a higher affinity of chymosin A to κ-casein due to the additional electrostatic stabilisation of the intermediary κ-casein–chymosin A complex.[21] Chymosin C is understood to be a product of chymosin A degradation, losing residues 244 to 246.[29] This study will focus on

chymosin B since it is the enzyme preferred in industry.

2.2.2 Secondary Structure

The chymosin secondary structure mainly consists of β-sheets (48%, 29 strands, 158 residues), combined with a few small α-helices (13%, 9 helices, 44 residues).[30] Three well defined sheets are formed by the anti-parallel β-strands.[31]

2.2.3 Tertiary Structure

The three-dimensional crystal structures of a number of aspartic proteases have been solved by X-ray crystallography including chymosin, endothiapepsin[32], human renin[33], penicillopepsin[34], pepsinogen[35,36], porcine pepsin[37–39], rhizopuspepsin and a number of retroviral proteinases.[40–42]

The tertiary structure of bovine chymosin is shown in Figure 2.1. The structure is highlighted in the key regions which have been found to be of importance in the allosteric mechanism.

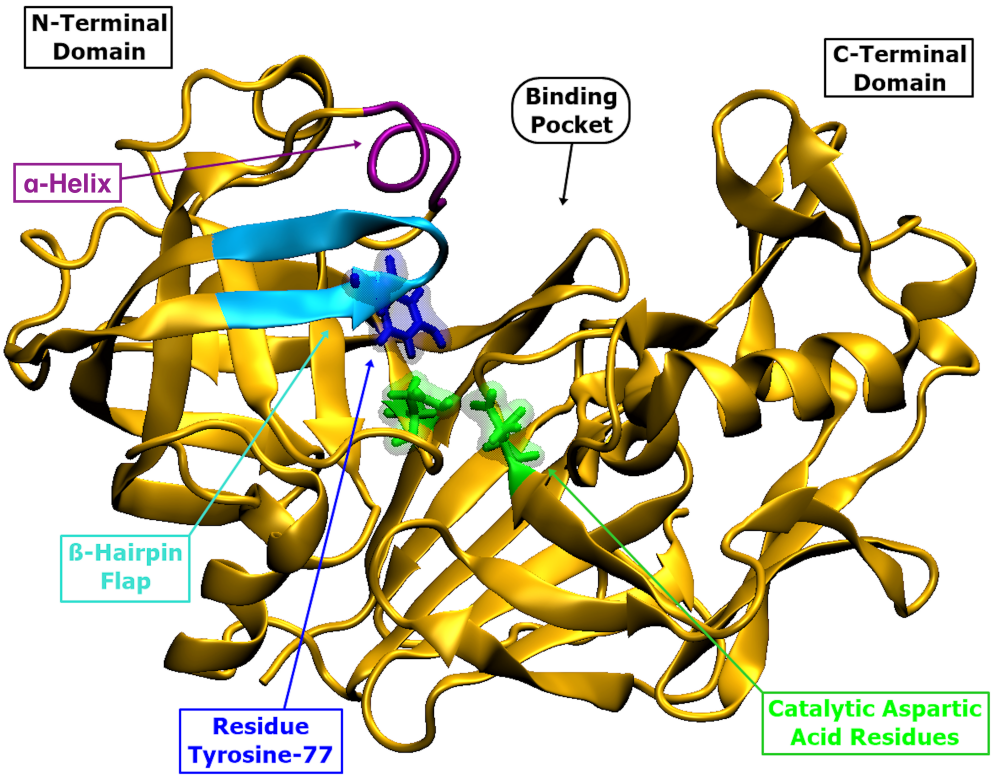


FIGURE 2.1: Depiction of bovine chymosin. 1) β-hairpin flap (light blue) and Tyr77 (dark blue). 2) α-helix (purple). 3) Catalytic residues (green).

The crystal structure of chymosin was first obtained in 1971.[43] A recombinant bovine chymosin crystal structure was solved and refined at 2.3 Å resolution in 1990 (1CMS),[44] then at 2.2 Å resolution in 1991 (4CMS).[31] There is a clear binding cleft which separates the bi-lobal folding pattern of the N- and C-terminal domains.[44] The two lobes are related by a pseudo-2-fold axis which lies between the two catalytic aspartic acid residues (Asp32 and Asp215), forming the approximate intramolecular symmetry.[31] The side chains of the aspartic acid residues are extended towards each other in an approximately planar geometry. The two catalytic aspartic acid residues are stabilised by a network of hydrogen bonds which incorporates two threonine residues, referred to as *"the fireman's grip"*.[45]

There are three disulphide bridges between residues Cys45-Cys50, Cys206-Cys210 and Cys249-Cys282. A number of ion pairs can also be found between Arg59-Asp57, Arg157-Glu308, Arg157-Ile326, Arg307-Asp11 and Arg315-Asp138.[31,44]

The structure contains a single *cis-proline* residue, Pro23, found in the N-terminus.[46] An identical *cis-proline* residue can be found in endothiapepsin, mucorpepsin and porcine pepsin.[32,47,48]

The active site of chymosin and other aspartic proteases consist of Asp, Thr and Gly residues from terminal domain and is highly conserved. There is a nine percent sequence identity between the C- and N-terminal domains of chymosin.[31] Structural comparisons of chymosin with other aspartic proteinases reveal a high degree of structural similarity. Superimpositions reveal the N-terminus has greater similarity with other aspartic proteinases than the C-terminal domain.[44] A rigid body in the C-terminus (residues 190 to 302) results in the C-terminus being more separated from the binding cleft than the N-terminal domain.[49]

2.3 Production

There are three established techniques for producing cheese in industry; rennet-curd, acid-curd and acid-heat.[50] The most used is the rennet-curd technique, where chymosin is used to cleave κ-casein, eliminating the hydrophilic region of micelles.[51]

The acid-curd method does not rely on a coagulating enzyme, instead milk is simply acidified.[50] At a pH of approximately 5.2, caseins clot and reducing the pH causes the substance to gel. This process is also referred to as acid-induced gelation, where physiochemical changes to caseins induce a gelling process.[51] The acidification of milk is understood to disintegrate calcium-phosphate complexes, causing some caseins to disassociate from micelles and fall into the micelle core.[52] A reduced net negative charge and increase in hydrophobic interactions result in aggregation of the micelles. This process is widely used in cream cheeses such as cottage cheese.[50]

The acid-heat technique is similar to the acid-curd, but involves heating the milk to 78-80°C.[52] The heating process takes place first and is then followed by acidification to a pH of 5.9 - 5.2, depending on the type of cheese desired.[50] Heating causes the flocculation of caseins and whey proteins, this in turn minimises the level of acidification required.[53] This technique is used to produce ricotta and queso-blanco, both highly desired cheeses in Europe.

2.4 Biological Aspects

2.4.1 Enzyme Stability

Chymosin is believed to be a stable enzyme between a pH range of 5.3 and 6.3, although some reports have found the structure remains relatively stable at a pH of 2.0.[54] It has also been reported the enzyme loses its activity rapidly under acidic conditions (pH 3-4), which is suggested to be caused by auto-degradation.[55] A similar loss of activity has been observed under basic conditions (pH above 9.8), which is thought to be due to conformational changes.[55] The loss of activity is reported to be larger for chymosin A when compared to chymosin B.[23] Chymosin B is favoured in industry due to its longer shelf-life, even though chymosin A has a 20% greater milk clotting activity.[56] This study solely investigates bovine chymosin B unless stated otherwise.

At temperatures of around 2°C, chymosin is more stable than at room temperature.[57] At high temperatures between 45 and 55°C, a rapid loss of activity is observed.[58] Studies have also indicated modifications on the terminal amino group of lysine residues and photo-oxidation of histidine residues may adversely affect chymosin activity.[59–61] A 30 minute incubation with 4.6M of urea at 37°C, will approximately half chymosin activity.[62] Experimental studies has shown both

cysteine residues and amino acids in the pro-region of the enzyme are crucial to refold the enzyme after denaturing.[63]

Chymosin is less stable at neutral pH compared to pro-chymosin.[23] Pro-chymosin is readily converted to chymosin at pH values below 5, whereas at pH greater than 11 a conformational change occurs, resulting in a dramatic loss of stability. Pro-chymosin can also be produced by bacteria, this variant is known as pseudo-chymosin and can be stored under acidic conditions while remaining stable. However, at pH values above 4.5 it is rapidly converted to chymosin.[64]

2.4.2 Enzyme Solubility

A number of factors affect the solubility of chymosin including the ionic strength of the solution, temperature and pH.[57] In a 1M sodium chloride (NaCl) solution, chymosin is soluble at a pH of approximately 5.5, but in the respective 2M solution it appears to be insoluble. Amorphous chymosin precipitates are more stable at 2°C than at 25°C, whereas crystalline chymosin is more soluble at 25°C. At a pH value of around 6.5, chymosin has an ionic strength of 0.005 mol/kg and is very insoluble. Solubility has been shown to increase with increases in ionic strength.[57]

2.5 Activity and Specificity

2.5.1 Catalytic Mechanisms

The earliest proposed catalytic mechanisms suggested that the catalysis was initiated by protonation of the carbonyl oxygen of the substrate by an Asp216 proton.[65] A donation of a water proton to Asp216 and a nucleophilic attack by the

generated hydroxide ion on the carbonyl carbon of substrate Asp34,[66] leading to the formation of the tetrahedral intermediate complex.[67] The intermediate complex is broken down by a protonation of the nitrogen atom by either the Asp34 catalytic carboxyl group or the bulk solvent. If broken down by Asp34 a simultaneous proton transfer may occur to the Asp216 carbonyl during the cleavage of the complex.[68]

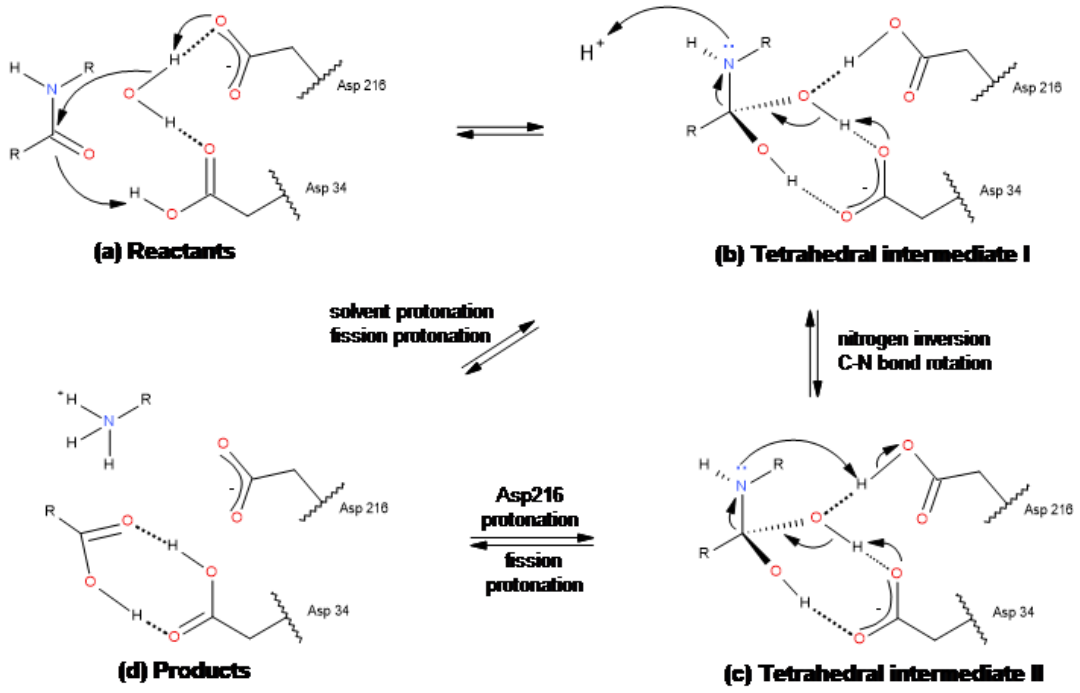


FIGURE 2.2: Catalytic mechanism proposed by Veerapandian *et al.*

The catalytic mechanistic model outlined in Figure 2.2 was proposed in 1990 by Veerapandian *et al.*[19] The carbonyl carbon of the tetrahedral carbonyl hydrate is hydrogen-bonded to the intermediate oxygen atom of Asp34-Asp216. The carboxyl oxygen of Asp34 is hydrogen-bonded to the hydroxyl oxygen of the hydrate. The scissile bond of the hydrate is protonated by Asp34 and consequently attacked by a water molecule which is polarized by Asp216 into a nucleophilic state. It is suggested that the rigid body of the enzyme-substrate complex may distort the amide bond which facilitates the nucleophilic water's attack on the hydrate. This accounts for the extensive hydrogen bonding of intermediate complex 1 to stabilise Asp34.[20] The amide nitrogen in this complex would favour protonation which can

be transferred from Asp216 or the bulk solvent.

Both chymosin and pepsin have been demonstrated to catalyse peptide synthesis.[69] For chymosin, peptide synthesis and hydrolysis is optimal between pH 4-5.[70] The activity of the enzyme is dependent on the residues neighbouring the forming or hydrolysing bond.[70]

In recent reports, in light of kinetic and *ab initio* studies, Piana and Carlini have proposed a different reaction mechanism involving a low-barrier hydrogen bond between residues Asp216 and Asp34.[71] This proposed mechanism is believed to occur through proton rearrangement around a 10-membered cyclic intermediate which is believed to occur via quantum tunnelling.[72] Neither the mechanism proposed by Veerapandian *et al.* nor the mechanism proposed by Piana and Carlini explain all of the experimental findings on chymosin catalysis. Both of these mechanisms and others proposed in the literature have not been accepted as the definitive mechanism within the field.

2.5.2 Zymogen Activation

Hammarsten was the first to identify the pro-enzyme, pro-chymosin[9] He showed that the rennet was formed and stored as the inactive, pro-chymosin. The pro-enzyme was shown to be activated by stomach acid, converting it into chymosin.[73]

Structural comparisons between porcine pepsinogen and pepsin have shown the enzyme and pro-enzyme are structurally very similar.[74] The region of pepsinogen which is defined as the pro-region and the first 13 residues of pepsin are where the differences occur between the enzyme and zymogen.[75] The side chain of these residues form different conformations in the zymogen and active enzyme. The same structural similarities between the active enzyme and zymogen have been observed in most aspartic proteases.[35]

The mechanism of zymogen activation in aspartic proteinases are different from enzyme to enzyme and dependant on pH.[46] An intramolecular mechanism occurs below pH 2.5 to convert the pepsinogen zymogen to the active pepsinogen enzyme.[64] The zymogen is cleaved at Met16P-Glu17P releasing the pseudo-enzyme and the pro-segment.[76] In calf chymosin the pro-enzyme is cleaved at Phe27P-Leu28P.[77] The same position is cleaved in chicken pepsinogen and human progastriscin.[78,79] It has been suggested the removal of the total pro-peptide segment occurs via an intermolecular mechanism at pH 3-4. The cleavage site, Phe42-Gly1 has been found to be more active at pH 2 than at pH 4.5.[64]

Site-directed mutagenesis studies on pro-chymosin have indicated that mutations of residues do not significantly impact on activity. Altering the pro-chain residues 27 to 30 resulted in normal proteolytic and activation processing.[80] It was also found when the cleavage site was removed a new cleavage site was generated by the pro-enzyme, Ser37P-Val38P, and this also left the proteolytic and activation processing unaffected.[80]

Zymogen activation processes in aspartic proteases are found to be dependent on pH, temperature and salt concentration. At room temperature and a pH of 5 pro-chymosin activation takes from 2 to 3 days.[81] On the other hand zymogen activation is completed in 5 to 10 minutes at room temperature, pH 2 and ionic strength of 0.1 mol/kg.[82]

2.6 Chymosin Flap Binding Specificity

Structural similarity between chymosin and other aspartic proteinases are well known and thoroughly reported in the literature. However, the greatest difference among these proteinases occur in the surface loop regions.[46] The most significant

difference is the position of the β-hairpin flap in chymosin, residues 73-85 (shown in light blue in Figure 2.1), which has been implicated in substrate binding specificity.[32]

The β-flap in both chymosin and pepsin appears to be able to form two different conformations, suggesting they may exist in two alternative structural forms.[83] The active form ensures the binding pocket is available for substrate binding, whereas in the inactive form the binding pocket is self-inhibited by the orientation of the Tyr77 residue.[16,84] It is widely accepted that the self-inhibited form of chymosin is converted into its active open form via an allosteric-activation procedure by the His98-His102 fragment of κ-casein, commonly referred to as the *"histidine-proline"* cluster.[16] This conversion is evident through experimental measurements of catalytic specificity of chymosin to κ-casein.

2.7 κ-Casein

The chymosin substrate, known as κ-casein is a 169 residue protein that helps solubilize the caseins in milk serum (α_{s1}-, α_{s2}- and β-caseins). κ-casein promotes the aggregation of micelles, which are macro-structures made by the four types of caseins. κ-casein is found predominantly on the surface of these.[85]

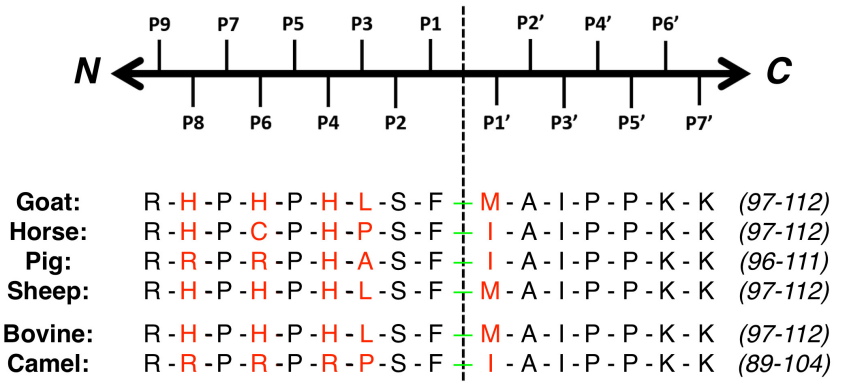


FIGURE 2.3: The aligned primary sequence of the chymosin sensitive region of κ-casein of different species. The Pn and Pn' numbering follows the Schechter and Berger nomenclature,[1] where n increases with the distance from the scissile bond. Residues that differ between some of the species are highlighted in red. The residue numbers are shown to the right in parenthesis.

The hydrophilic region of κ-casein protrudes from the micelles giving the structure stability against spontaneous aggregation.[86] Chymosin selectively cleaves bovine κ-casein at the PheP1-MetP1' bond and camel κ-casein at the PheP1-IleP1' bond, causing the hydrophilic C-terminal end of κ-casein to dissociate, thereby destabilizing the casein micelles initiating the release of insoluble casein proteins which results in milk clotting.[86] (The Px or Px' nomenclature is used to describe κ-casein residues on the two sides of the cleavage site, e.g. Ser104, Phe105, Met106 and Ala107 in bovine κ-casein are referred to as SerP2, PheP1, MetP1' and AlaP2', respectively. Regions of chymosin that interact with the P2, P1, P1' and P2' residues are denoted S2, S1, S1' and S2' pockets, respectively.[1]). The amino acid sequences of κ-caseins from different species in the region of the cleavage site are given in Figure 2.3.

The crystal structure of a chymosin-inhibitor complex (RCSB PDB ID: 1CZI) and previous molecular modelling studies, suggest that κ-casein binds to chymosin in an extended secondary structure.[15,86] This is supported by circular dichroism, solution NMR and molecular modelling studies of unbound κ-casein, all showing an extended structure in the region of the scissile bond.[87,88] Molecular modelling and mutagenesis studies, propose that the P8-P7' residues are located in the chymosin binding cleft during catalysis. Recently the arginine residue in the P9 position has

been implicated in binding because it is conserved in bovine, camel, pig, buffalo and goat chymosin.[89] Furthermore, an ArgP9His mutant is observed to be a poor substrate.[90] In κ-casein, residue SerP2 appears to be essential for the catalysis to take place.[91] The hydrophobic residues LeuP3, AlaP2' and IleP3' are crucial in giving the structure its hydrophobic qualities.[92] In camel κ-casein LeuP3 is replaced with a hydrophobic proline residue, retaining the same hydrophobic qualities as bovine chymosin.

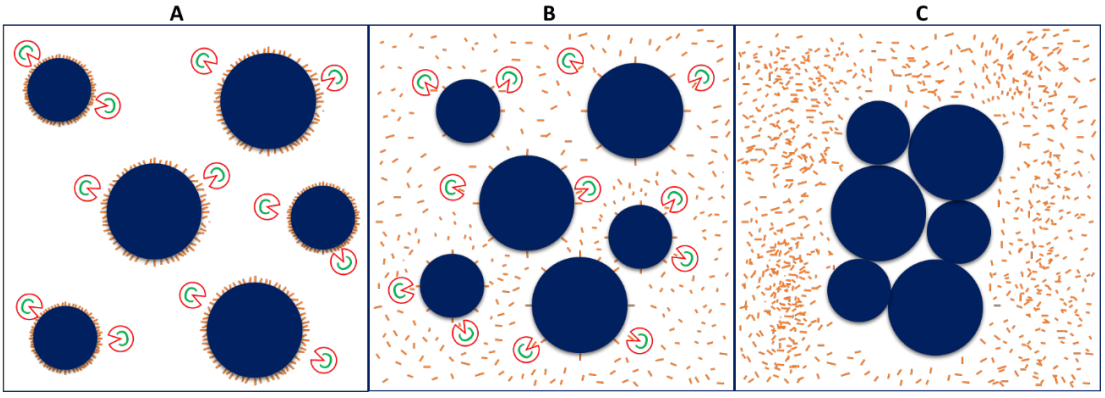


FIGURE 2.4: Schematic depicting chymosin activity on κ-casein. A) Intact casein micelles with protruding κ-casein layer preventing aggregation of micelles. B) Chymosin catalytic activity cleaves protruding κ-casein, displaying partially denatured micelles with removal of hydrophilic shell. C) Progressive coagulation of extensively denatured micelles with a hydrophobic outer layer.

At the normal pH of milk, 6.6 - 6.7, the micelles carry a net negative charge. The hydrophilic region of κ-casein protrudes from the micelles, providing the structure with added stability against spontaneous aggregation.[86] The stabilising effects of electrostatic repulsion and steric hindrance are eliminated once the coagulant, chymosin, cleaves the protruding κ-casein of the micelle. This also results in a reduction in the negative charge at the surface of the micelles.[15] Losing these chemical barriers results in the micelles coming together as shown by Figure 2.4.

Clotting tends to occur when calcium ions help adjacent micelles to aggregate through electrostatic and hydrophobic interactions. The ionic calcium trapped within micelles is released upon acidification. This is exploited in industry where

ionic calcium is sometimes added to increase the rate of aggregation. The network of aggregates that forms is known as the coagulum.[86] This coagulum entraps structures such as milk-fat, water and water soluble components including undenatured whey proteins, a mixture collectively known as serum.[86]

Upon acidification to 4.6 pH or lower, the milk proteins separate. Casein proteins are precipitated out of the coagulum and the whey/serum proteins remain soluble.[9] The ratio of casein to whey proteins vary from mammal to mammal but in bovine mammals it is approximately 80:20.[15] In 1938 two more types of proteins were discovered in milk; proteose-peptone and ionic nitrogen.[93,94]

2.8 Camel/Bovine Chymosin Relationship

Bovine and camel chymosin have high sequence similarity (94%) and identity (85%) and similar three-dimensional structures, depicted in Figure 2.5. They both comprise 323 amino acids that fold into a pseudo-symmetric bi-lobal structure forming a central binding cleft containing the catalytic residues Asp34 and Asp216. In both enzymes, the side chains of the catalytic aspartic acid residues extend towards each other in a planar geometry,[95] which is stabilised by a network of hydrogen bonds with two threonine residues, referred to as "*the fireman's grip*".[45] Similar features are found in other homologous aspartic proteases.[14,16,47,48]

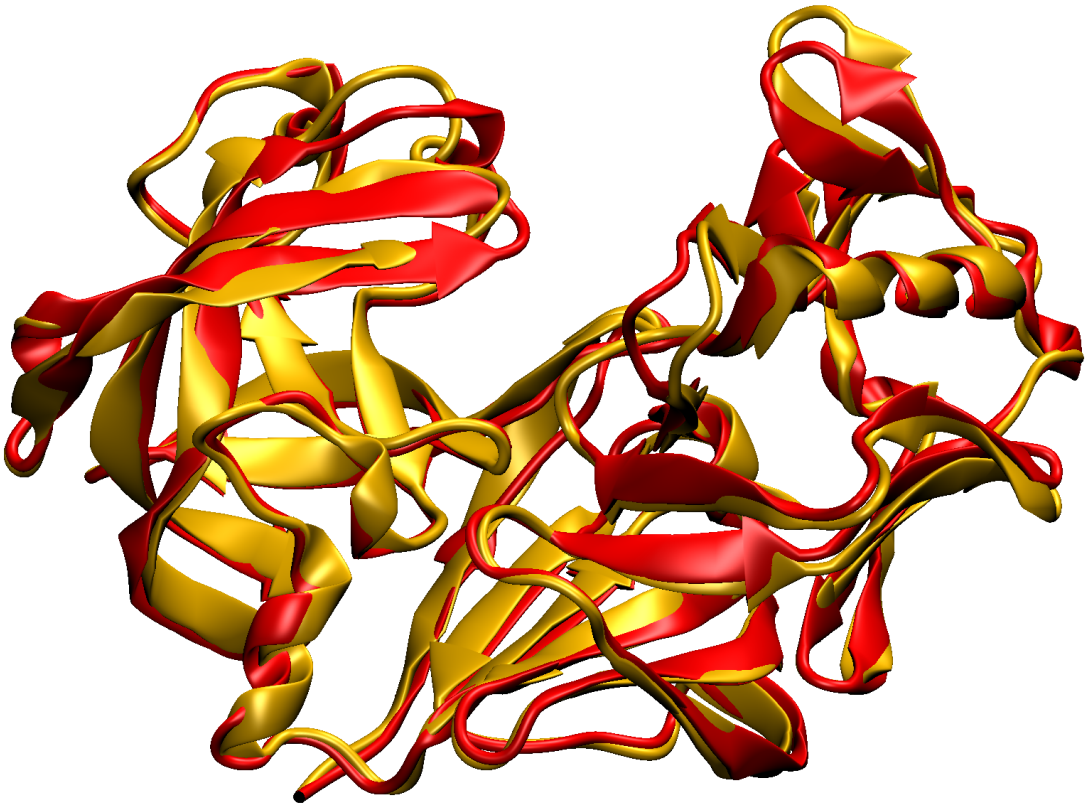


FIGURE 2.5: Depiction of structural similarities between camel (red) and bovine (gold) chymosin through structural alignment.

Within the substrate-binding cleft, there are 12 differences in the primary structure of bovine and camel chymosin. Both variants of the enzyme have a positively charged patch on the N-terminal (residues 50-60) and a negatively charged patch on the C-terminal (residues 240-260) that interact with κ-casein.[96]. The positive patch is larger in camel chymosin through the replacement of a Gln56 in bovine chymosin by His56 in camel. The negatively charged patch in camel chymosin is found to be less negative through the replacement of Asp249 and Asp251 in bovine chymosin by Asn249 and Gly251 in the camel variant. In camel chymosin, there are two additional positively charged patches that are not found in bovine chymosin.[97] The first resides in the C-terminal where a small positive patch comprises residues Arg242, Arg254 and Lys278, and the corresponding residues in bovine chymosin are hydrophilic but neutral. The second is found at the base of the binding cleft where the residues are Arg150 and Arg316 in camel chymosin,

but Gln150 and Leu316 in bovine chymosin.

2.9 Sources of Chymosin

Aspartic proteinases have been found in a number of natural sources including but not limited to viruses, plants and mammals. Generally they are divided into two groups; pepsin-like and retro-viral enzymes.[46] They have been isolated from five major sources:

1. As gastric enzymes in the stomachs of mammals i.e. pepsin A (EC 3.4.23.1), pepsin B (EC 3.4.23.2), gastricsin (EC 2.4.23.3) and chymosin (EC 2.4.23.4). These are produced in the mucosa as zymogens (an inactive precursor).[98] Gastricsins are found in all parts of the stomach, prostate gland, seminal vesicles and α-cells of pancreatic islets.[46] Chymosin is produced in the mucosa of new-born mammals during gestation, such as calf,[99] kitten,[100] lamb,[99,101] piglet[26] and seal.[102] It was established that the production of enzymes were linked directly to the age of the animals and their feeding frequency.[46]

2. Lysosomes which contain cathepsin D and cathepsin E. Cathepsin E has been found in gastric mucosa, spleen, thymus and also blood cells.[103] Cathepsin D in humans has been proposed to be involved in the degradation of endocytosed and intracellular proteins. It has also been used in the past as a prognostic tool to indicate breast tumour invasiveness.[46]

3. Rennin-producing tissues such as the kidney and sub maxillary gland.[104]

4. Plant seeds have been reported to contain aspartic proteases, i.e. barley, cucumber, lotus, rice, sorghum, squash, tomato and wheat.[105–108]

5. Micro-organisms have been reported to excrete a number of proteinases. Some fungi excrete proteinases, i.e. *Aspergillus awamori*,[109] *Aspergillus*

niger,[110] *Endothia parasitica*,[111] *Mucor pusillus*,[112] *Mucor miehei*,[113] *Penicillium janthinellum*,[114] *Rhizopus chinensis*[115] and *Trichoderma reesei*.[109] Thermopsin is secreted from a thermophilic archaebacterium named *Sulfolobus acidocaldarius*.[116] Yeast proteinases have been documented in *Candida tropicalis*,[117] *Saccharomyces cerevisiae*[118] and *Yarrowia lipolytica*.[119]

Retro-viral proteinases are usually found in a dimeric form. Each monomer carries just one of the catalytic aspartic acid residues and is approximately half the size of a eukaryotic aspartic proteinase. Retro-pepsins have been found in a number of viruses including avian myeloblastosis virus, human immunodeficiency virus (HIV), *Rous* sarcoma virus and simian immunodeficiency virus (SIV).[120,121] These types of proteinases code for the processing of RNA dimerization within a host and is therefore essential.[46]

2.10 Recombinant Calf Chymosin

A number of rennet substitutes for bovine calf chymosin have been developed in industry from various sources including adult cows, other proteolytic enzymes and fungus proteinases. The major problem with these substitutes is that they exhibit a considerably larger level of non-specific proteolytic activity.[46] They also exhibit a greater level of thermo-stability, leading to considerably lower yields as higher temperatures lead to a more complete degradation of milk proteins to form peptides.[46] Reasons for these physiochemical properties in terms of structural chains and conformations is not yet known, but it does suggest a synthetic chymosin enzyme with enhanced and desirable properties is a feasible prospect.

2.11 Molecular Modelling Studies of Chymosin

Very few molecular modelling studies have been conducted on chymosin–κ-casein complexes. In 1995, a study on the protein-ligand complex involving the HisP8-LysP6' fragment from κ-casein coupled with both bovine chymosin and porcine pepsin was investigated through computational means.[15] Short molecular dynamics simulations were carried out using a distance-dependant dielectric protocol to model the solvent. The starting position of the κ-casein fragment was deduced by a computational mutation of the pepstatin inhibitor. This was done by superimposing bovine chymosin on a rhizopuspepsin–pepstatin inhibitor complex. The authors initially suggested that a cis-peptide bond between HisP8-His99 was crucial for interactions between Asp247 and His98. However, in 1997, they revised this theory on the basis of results from molecular dynamics simulations using a longer peptide chain.[87] The study in 1997 reported favourable interactions between κ-casein and chymosin (HisP4:Glu245, HisP6:Asp297 and LysP6':Glu133 respectively). Although an implicit solvent protocol was used, no comment was made regarding the conserved or catalytic waters in this investigation. In 2002 a computational study of apo-chymosin using an explicit solvent model suggested the self-inhibited conformation of chymosin can be found in solution.[122]

In 2010, unrestrained molecular dynamics simulations were carried out on bovine chymosin complexed with a fragment of κ-casein (ArgP9-LysP7').[2] The trajectories showed that the substrate binds in an extended pose and charged residues flank the scissile bond which the authors propose stabilises the binding pose. The κ-casein fragment can be seen to bind to both terminals: residues LysP6' and LysP7' displace a conserved water molecule to bind to the N-terminal domain and the HPHPH sequence in residues P8-P4 binds to the C-terminal domain. ArgP9 of κ-casein is also proposed to be crucial for the stabilising of the binding pose.[2] However, the steric and/or electrostatic effects which cause the binding pose to stabilise remain unclear.[2]

A solvent binding and computational alanine scanning study of chymosin–κ-casein complexes in 2013 highlighted that water binding sites on the surface of bovine chymosin take part in stabilising the complex.[123] The authors demonstrated that relative binding thermodynamics of single-point mutants in bovine chymosin–bovine κ-casein complexes can be accurately calculated using molecular integral equation theory techniques. Although the water binding sites have been identified by statistical analysis of crystallographic data, as well as through simulation methods to deduce the importance and roles of the waters and their binding sites, there are a number of questions that still remain unanswered. Such as the mechanism for chymosin–κ-casein binding, the catalytic mechanism and the residues that take part in stabilising the binding pose.

Chapter 3

Quantitative Structure Activity Relationships (QSAR)

3.1 Introduction

The premise of quantitative structure-activity relationships (QSAR) is that a compound's molecular structure can be used to determine its macroscopic properties, such as binding affinity and pIC_{50}. A QSAR is derived by using experimental data to learn a statistical relationship between the physical property of interest (e.g., pIC_{50}) and molecular descriptors calculable from a simple computational representation of the molecule. The QSAR must accurately model the training data and generalize to correctly predict activities for molecules outside the representative training set.[124]

Since the concept of QSAR was first introduced by Free, Wilson, Hansch, and Fujita in 1964, a wide range of methodologies have been developed using various classes of descriptors.[125,126] For the prediction of physiochemical properties, 1D and 2D descriptors that can be calculated quickly without knowledge of molecular conformation are often considered to be satisfactory (e.g. counts of functional groups, graph indices, etc).[127,128] However, for modelling protein-ligand systems,

where ligand conformation influences the strength of binding interactions, 3D (or 4D) descriptors are usually preferred.[124,129–132]

QSAR has traditionally been applied to virtual or newly synthesised compounds to predict/classify biological activities and investigate their biology, chemistry and toxicology.[133–136] QSAR models are also commonly used in computer aided drug design (CADD) to design new chemical entities and this approach is being employed increasingly by the pharmaceutical industry to find high quality leads in the early stages of drug discovery.[136,137] The systematic application of QSAR in CADD is used to reduce costs by ensuring only the most promising hit compounds are pursued, thus reducing the number of time consuming and costly experiments.

3.2 Procedure

In general, QSAR investigations involve a multi-step systematic process (Figure 3.1). This includes selection and preparation of the dataset, selection/generation of descriptors, statistical/mathematical model derivation, training of the derived models, validation of the models using the training set and/or an internal test set and finally, testing the predictive accuracy of the model using a testing dataset.

During the dataset preparation step, it is crucial to assess the quality of the data, avoiding developing unreliable QSAR models. The data should preferably be obtained from the same bioassay protocol to avoid any inconsistencies and inter-laboratory variabilities. Furthermore, a large enough number of compounds should be included in the dataset to provide statistically valid QSAR models and the bio-activities covered by the dataset compounds should be a good representation of the bio-activity range of the compound family.[136]

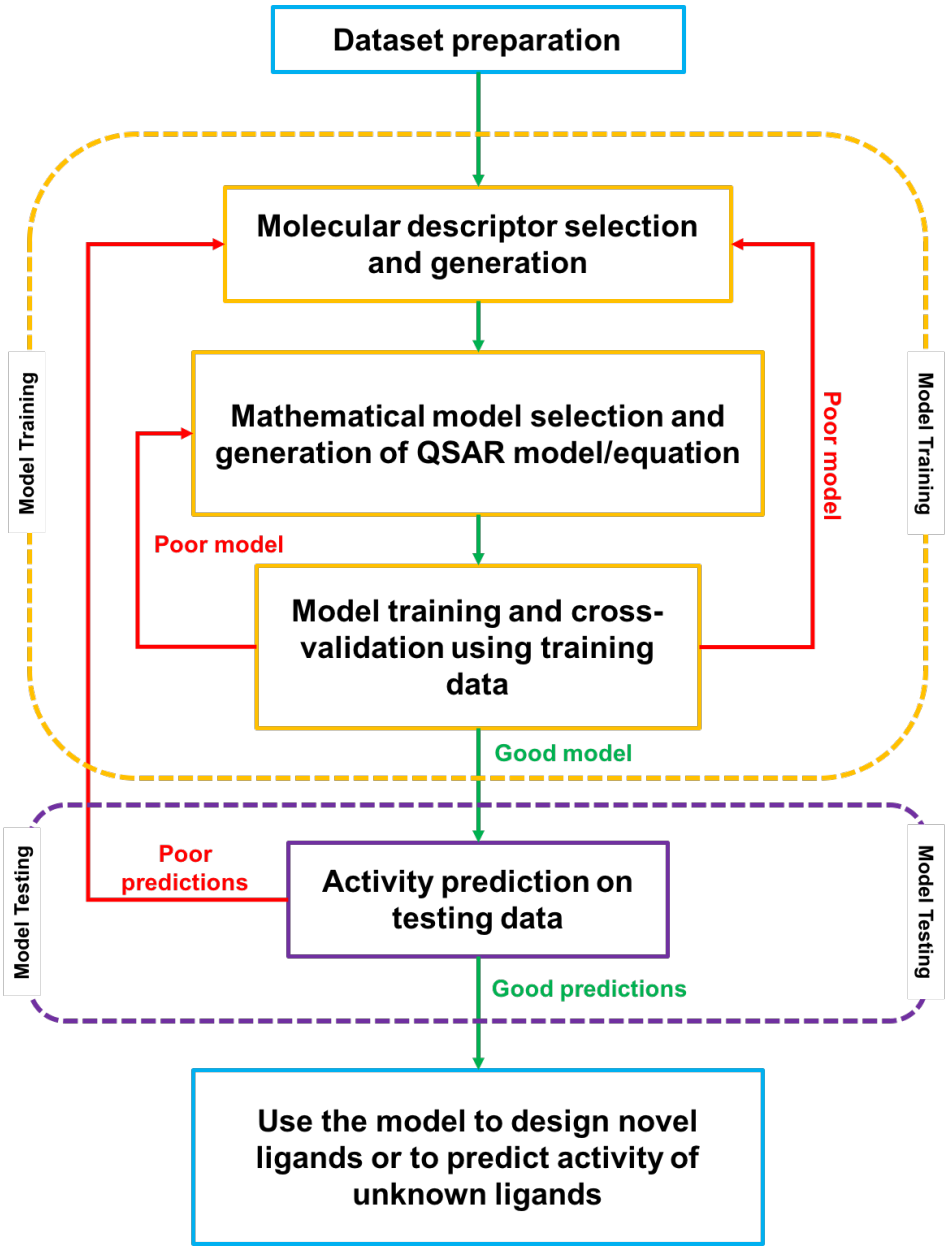


FIGURE 3.1: A scheme depicting the steps involved in QSAR model development, including the systematic training and testing processes.

The second step of QSAR involves the selection and/or generation of descriptors for each compound in the dataset. There is a vast array of possible descriptors that can be selected in this step but only a few of these are likely to be significantly correlated with the activity. Therefore, selecting the most appropriate descriptors is crucial in QSAR to build a robust model, usually done retrospectively after preliminary tests. The descriptors should appropriately capture the most useful information about the structural variation between compounds in the dataset.

Once the molecular descriptors have been defined and generated for all compounds in the dataset, the most suitable statistical or mathematical model to identify the relationship between the descriptors and biological activity must be selected. Here, regression methods can be used which require continuous variables, or classification methods, which require discrete variables (i.e. active/inactive, soluble/insoluble). For instance, linear approaches such as partial least squares (PLS) or multiple linear regression (MLR) can be used as mapping or correlating functions. Non-linear approaches such as neural networks (NN), support vector machine (SVM) and random forest (RF) can also be used. Other methods including evolutionary algorithms like the genetic algorithm (GA) have also been employed to select input variables.

The next step involves training and validating the QSAR model. QSAR models are trained on the training set which contains a subset of randomly selected compounds from the dataset, (usually a majority) leaving a minority set which can be used for testing. Often during the model training, validation is also performed. Validation methods such as leave-one-out cross-validation (LOO-CV) are used to measure the statistical stability of a QSAR model. The training process is usually repeated until a satisfactory model is achieved. Finally, the trained QSAR model is tested by using the dataset compounds in the test subset to predict their activity values, assessing the models predictive accuracy.

A critical step for QSAR studies is the splitting of the dataset into training and testing subsets, typically done during the dataset preparation. The training set should preferably cover the range of bio-activity that is included in the test set. It is also preferable that the training subset includes a good distribution of compounds which includes all atom types and molecular fragments included in the testing subset.

3.3 3D-QSAR

3D-QSAR methods were developed to provide improved prediction accuracies in comparison to 2D methods and, as such, are computationally more complex and demanding. Usually, 3D-QSAR methods are split into two families: *alignment-independent* and *alignment-dependant*.[138] The difference between the two families is straight forward, *alignment-independent* methods do not require the compounds of the data set to be structurally aligned, whereas, the *alignment-dependant* methods do require the alignment of the compounds. This thesis focuses on *alignment-dependant* 3D-QSAR methods. Both methods require the bioactive conformations of the compounds and measured bioactivity (experimentally or computationally derived). The conformations of compounds in QSAR methods are considered one of the major drawbacks. Experimentally, bioactive conformations are difficult to interpret and computational conformations are difficult to validate.

3.3.1 Comparative Molecular Field Analysis (CoMFA)

One of the most widely used 3D-QSAR methods is the comparative molecular field analysis (CoMFA), which was proposed by Cramer *et al.* in 1988.[132] CoMFA establishes a uniform grid encompassing a series of pre-aligned molecules. Electrostatic and steric potential energies are then calculated between a positively charged carbon atom probe, located at each vertex of the grid, and each of the molecules embedded within.[132] The resulting electrostatic and steric fields are used as input for partial-least-squares (PLS) regression models. Since its first publication, CoMFA has been cited in over 4000 published articles and used in numerous drug discovery programs.[139,140] Several extensions to the CoMFA methodology have been proposed, of which the highest profile is comparative molecular similarity indices analysis (CoMSIA).[141,142]

CoMFA requires all compounds in the dataset to be aligned and this can adversely affect the model/predictions if it is not done correctly. The quality of the alignment is subjective and it is both time-consuming and difficult to reproduce as the method is slightly different from software to software as well as from version to version, however, a good alignment is fundamentally required for CoMFA.[143,144] Nevertheless, numerous CoMFA models have been developed for several drug design and molecular modelling studies since its release.[135,145–149]

3.3.2 Comparative Analysis of 3D-RISM Maps (CARMa)

Although CoMFA is widely used, it relies on a relatively simple representation of molecular interactions, which does not explicitly account for solvation/de-solvation effects that can dramatically influence protein-ligand binding. Since CoMFA was first proposed, advances in theory, algorithms and computer power mean that there are now many fast and accurate methods to model molecular solvation effects. Integral equation theory approaches are of particular interest for QSAR modelling because they allow solute-solvent distributions and solvation thermodynamics to be computed at a fraction of the cost of explicit solvent numerical simulations and with no sampling error. The most widely used of these methods are the 1D and 3D Reference Interaction Site Models proposed by Chandler et al. and Beglov and Roux, respectively.[150–153] Accurate predictions of hydration free energy and Caco-2 permeability have previously been reported using 3D-QSAR models based on 1D RISM molecular descriptors.[154]. Recently, Güssregen et al. proposed the CARMa methodology, which uses solute–solvent distribution functions calculated by 3D-RISM to replace the electrostatic or steric fields in CoMFA. This approach was shown to give accurate predictions of binding affinities for a series of serine protease inhibitors, but tests on other systems have not yet been published.[155]

Chapter 4

Theory

4.1 Molecular Mechanics (MM)

The use of a molecular mechanics force field allows for the potential energy of a chemical system to be calculated as a function of its configurational and/or conformational degrees of freedom.[156–159] This thesis focuses on AMBER force fields as they are used in the research but other force fields are mentioned where appropriate. A common form for a molecular mechanics force field is given in Equation (4.1).

$$\begin{aligned} U = &\sum_{Stretch} U_{AB} + \sum_{Torsion} U_{ABCD} + \sum_{Bend} U_{ABC} + \sum_{Out-of-Plane} U \\ &(+ \sum_{Cross-Terms} U) + \sum_{Van-der-Waals} U_{AB} + \sum_{Electrostatic} U_{AB} \end{aligned} \tag{4.1}$$

Figure 4.1 provides a physical depiction of the terms in a force field as described by Equation (4.1). The sum of these terms provide the potential energy of a system. Each term in the equation describes the energy of the atoms in different positions

for a single conformation.

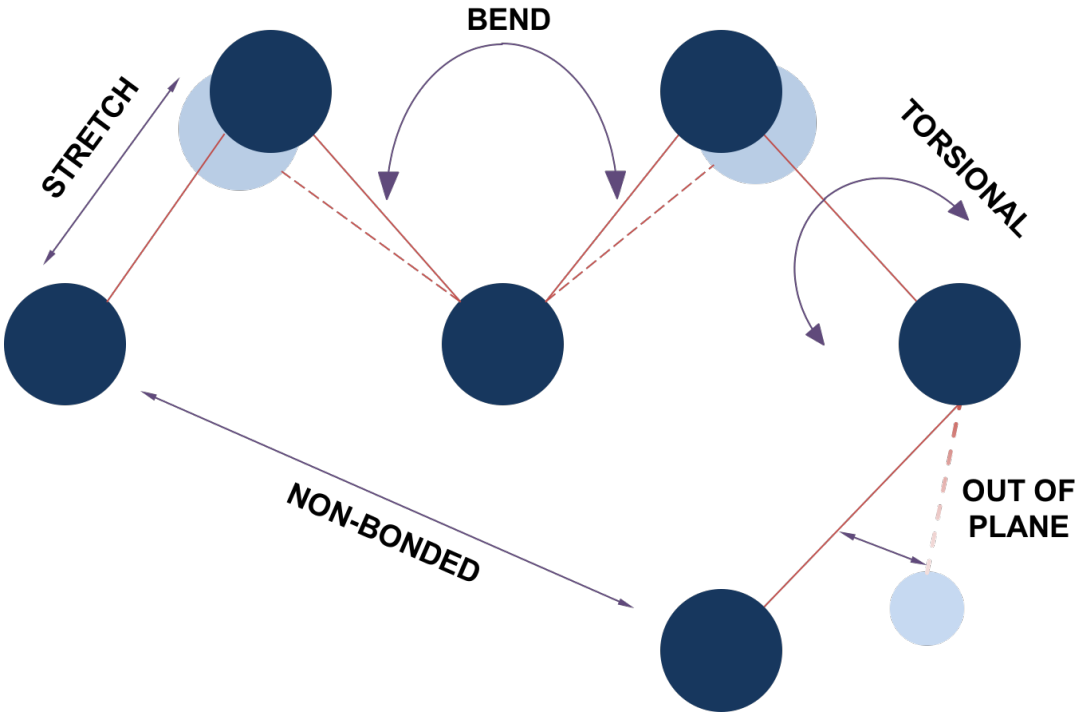


FIGURE 4.1: Depiction of a typical force field bonding and non-bonding terms.

4.1.1 Bonding Terms

4.1.1.1 Bond Stretching

The bond stretching function describes the energy needed to stretch a bond between two atoms as shown in Figure 4.2.[160]

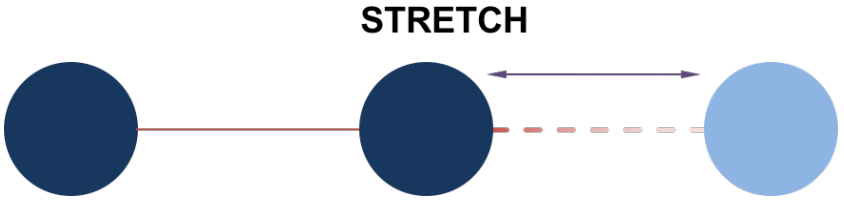


FIGURE 4.2: Depiction of a bond stretching motion between two atoms.

In the AMBER force field this is expressed by Hooke's Law:[161]

$$U_{AB} = k_{AB}(R_{AB} - R_{eAB})^2 \tag{4.2}$$

where the force constant is represented by k_{AB}. A stretched bond is described by R_{AB} and R_{eAB} represents a bond length at equilibrium.[162]

Hooke's Law provides a relatively accurate estimate for two atoms that are close to their optimum bond lengths. However, when two atoms move apart, Hooke's law assumes a harmonic correlation where the rate of energy change is the same as when two atoms are moving closer together. In reality this is not entirely true as it is an an-harmonic motion; the rate of energy change is greater when two atoms are moving closer together than when two atoms are moving further apart, described in Figure 4.3.

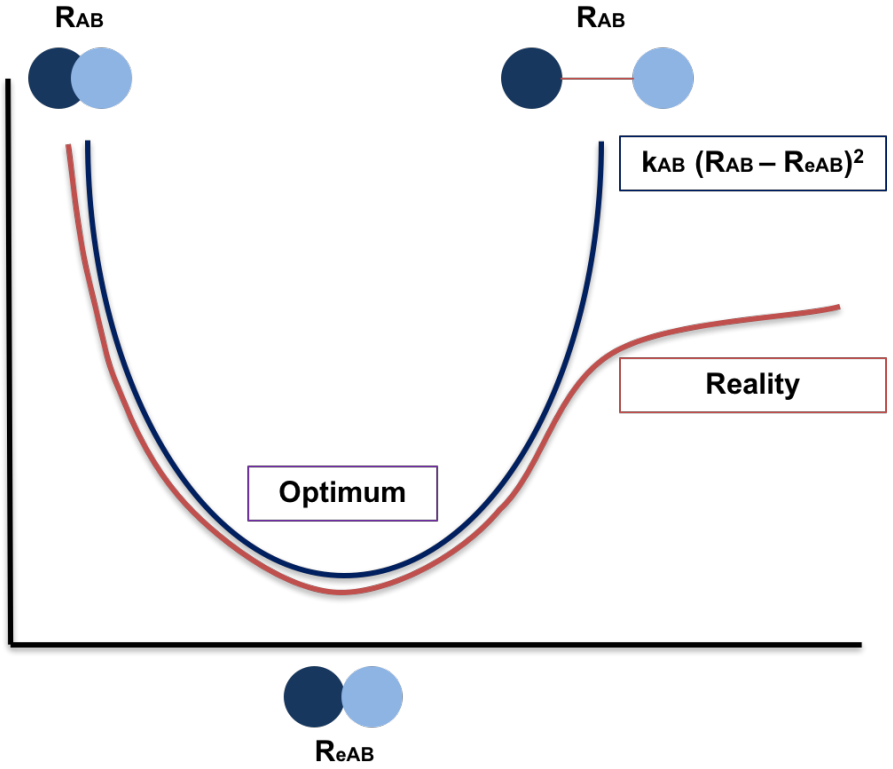


FIGURE 4.3: Graphic describing difference between the harmonic correlation assumed by Hooke's law and the an-harmonic potential that occurs in reality.

Other expressions for bond stretching potentials have been well documented, including the Morse potential[163] which is used in the CVFF[164] (consistent valence force field), as well as quartic polynomials used in PCFF[165] (polymer consistent force field). The Morse potential descriptor for bond stretching is a more qualitative measure compared to the harmonic and quartic polynomial counterparts. It provides a more accurate result for bond lengths that are close to equilibrium. For modelling high energy systems, a harmonic function is sometimes preferred because it prevents bonded atoms migrating to irrational positions, such unrealistic stretches that are possible with the Morse potential. However, the harmonic potential in Hooke's law provides comparably accurate data for a general force field like AMBER.

4.1.1.2 Bond Bending

The bond bending term of a force field represents the energy required for a bond to distort or bend in relation to its equilibrium state, as shown in Figure 4.4.[166]

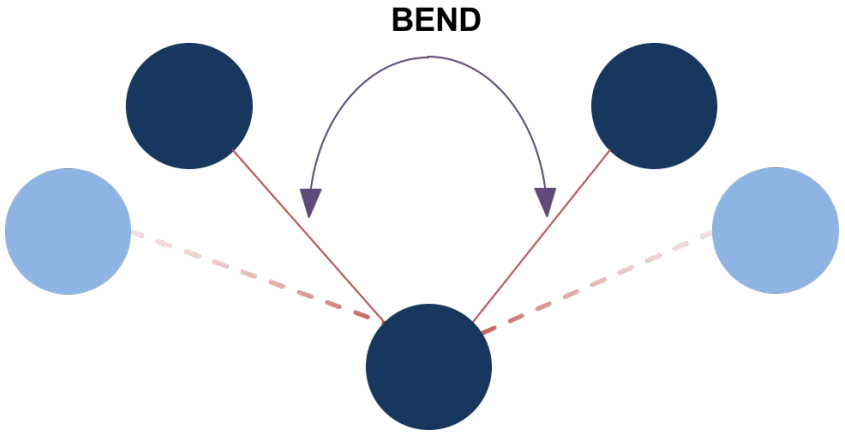


FIGURE 4.4: Depiction of a bond angle between three atoms distorting or bending.

In AMBER, the harmonic expression for bond bending is similar to the expression used for bond stretching.[167] Other expressions such as quadratic polynomial also describe bond bending distortions, however, for complex systems a harmonic expression is preferred due to its relative cost effectiveness and accuracy. The bond bending term in AMBER is expressed as:

$$U_{ABC} = k_{ABC}(\theta_{ABC} - \theta_{eABC})^2 \tag{4.3}$$

where k_{ABC} is the angular constant. The distorted bond angle is represented by θ_{ABC} and the equilibrium bond angle is θ_{eABC}. Therefore any deviation from the equilibrium bond angle will alter the potential energy.[167]

4.1.1.3 Torsion (Dihedral) Angles

The dihedral angle term is a description of the energy required to rotate a bond, shown in Figure 4.5.[168]

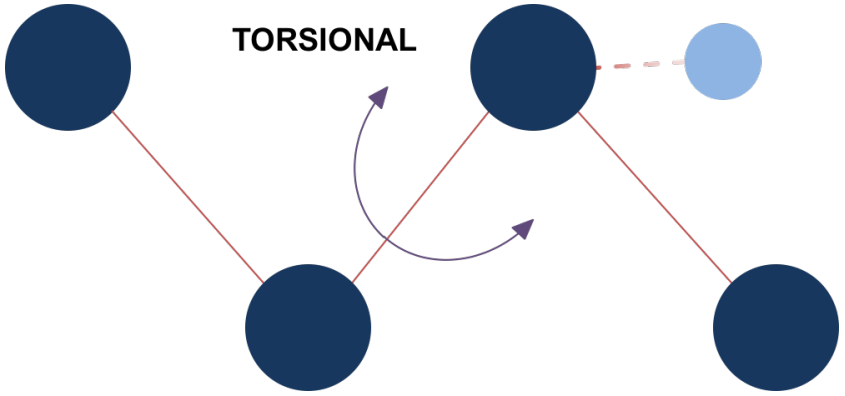


FIGURE 4.5: Depiction of the conformational freedom of a torsion angle.

The torsion angle term accounts for the amount of energy necessary for a bond to rotate. In early versions of AMBER the term was described by a harmonic expression, Equation (4.4).[169]

$$U_{ABCD} = k_{ABCD}(\chi_{ABCD} - \chi_{eABCD})^2 \tag{4.4}$$

The k_{ABCD} term in Equation (4.4) describes the dihedral constant. Comparable to the other Hooke's law expressions the rest of this expression represents the energy needed for a deviation of the torsion in relation to the equilibrium position; χ_{ABCD} is the distorted dihedral angle and χ_{eABCD} is the equilibrium dihedral angle.[170]

Dihedral angle potential terms are also expressed through a three-term Fourier expansion, a more accurate harmonic equation used in later versions of AMBER. Another form of the term is the cosine function format which depicts the periodic nature of torsions and is also used by some AMBER force fields. In AMBER03 and AMBER99-SB (version 14, 15 and 16) the torsion angle term used is a simplified

version of the cosine function, Equation (4.5).[171,172]

$$U_{ABCD} = k_{ABCD}(1 - \cos(n(\chi_{ABCD} - \chi_{eABCD}))) \tag{4.5}$$

The n term in Equation (4.5) represents periodicity, enabling the equation to more accurately represent torsional angles. The dihedral constant also accounts for amplitude which is predetermined from experimental studies.[173] Dihedral parameters are optimised on the simplest molecules and then extrapolated onto larger, more complex structures. The benefit of this, in contrast to other methods that attempt to reproduce conformational energies of large systems is that it can be applied to a wider range of systems. A lack of dependence on implicit parameters on groups of compounds is beneficial and has been found to produce accurate results.[174] A consideration must also be made for the absence of an offset term (Υ) within the expression, which would further improve accuracy without adversely affecting running time.[169]

4.1.1.4 Out-of-Plane (Inversion) Angle

The out-of-plane angle term is used to define the planar interaction of a group of atoms, shown in Figure 4.6.[175]

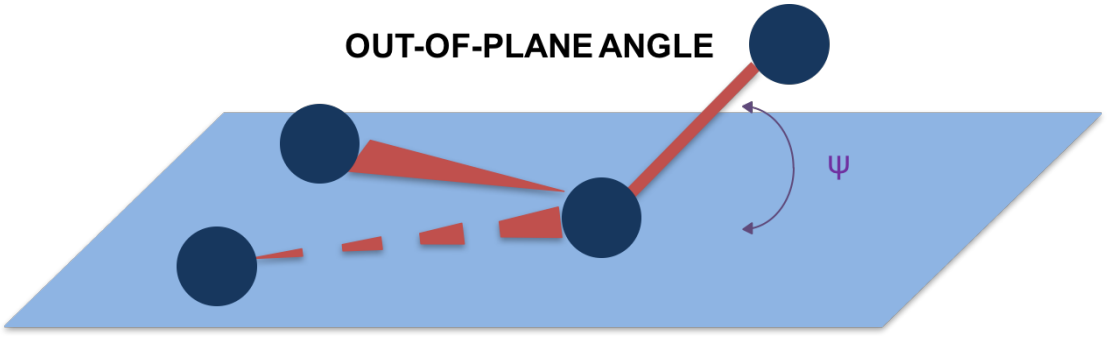


FIGURE 4.6: Depiction of conformational freedoms that go out-of-plane.

The term considers four atoms existing as a group in a single plane connected via three valence bonds from a central atom. Typically in molecular mechanics force field, Equation (4.6) is most used.[176]

$$U = \left(\frac{1}{2\sin^2\psi_e}\right) k(\cos\psi - \cos\psi_e)^2 \tag{4.6}$$

In AMBER however, the out-of-plane term is an "*umbrella function*" which is expressed in Equation (4.7).[177]

$$U = k(\psi - \psi_e)^2 \tag{4.7}$$

This empirical potential function is similar to harmonic expressions described above. The term k describes the harmonic force constant factor.[178] The term $\psi - \psi_e$ describes the energy cost for the out-of-plane angle in relation to the angle at equilibrium.[177]

4.1.2 Cross Terms

Cross-terms are sometimes included within force fields to achieve better accuracy.[179] They are designed to more accurately account for bond and/or angle distortions caused by neighbouring atoms.[180] They are implemented by reproducing experimental vibrational frequencies, hence depicting the dynamic properties of the molecule.

Cross-terms can occasionally result in unrealistic geometries so precautions must be taken. The unlikely geometries arise when the starting geometry is significantly distorted causing the optimisation to settle in a local minimum.[181] The versions

of AMBER used in this study have no built in cross-terms.

4.1.3 Non-Bonding Terms

Non-bonding interactions can be described as either intra- or intermolecular interactions. Within molecular mechanics force fields they are divided into two groups of interactions, electrostatic and van der Waals.[182] In some force fields a hydrogen bonding and non-bonding term can be present.[183] In the force fields used in this thesis, hydrogen bonds are an anticipated consequence of the electrostatic and van der Waals parameters and are not represented by a separate term in the algorithms. It has also been found that a separate hydrogen bonding term does not always improve the accuracy of a force field in relation to experimental data.[184,185]

4.1.3.1 Van der Waals

Van der Waals (VdW) interactions are described in molecular mechanics force fields by Lennard-Jones (L-J) potentials.[186] The L-J potential describes the forces of attraction and repulsion between non-ionic atoms. In AMBER the expression is defined as:

$$U_{L-J} = \varepsilon \left[\left(\frac{C}{R} \right)^{12} - 2 \left(\frac{C}{R} \right)^{6} \right] \tag{4.8}$$

where term C defines the distance between a pair of non-bonded atoms when the bonding potential is zero. This is usually set to half the distance of the inter-nuclear distance of the two atoms in molecular mechanics force fields.[187] R represents the distance which separates the two atoms; normally the distance from nucleus to nucleus. The repulsive force is described by, $\left(\frac{C}{R}\right)^{12}$, and the attraction force, $\left(\frac{C}{R}\right)^{6}$, is subtracted from it.[186] The final term ε represents the energy well

depth. It is a function that differentiates and parametrises the atom types that are interacting.

Some force fields include an additional term which is used to describe hydrogen bonding. This is typically a modified form of the 12-6 L-J potential and is usually expressed as the 12-10 L-J term, Equation (4.9).[188]

$$U_{HB} = 4\varepsilon \left[\left(\frac{C}{R_{HB}} \right)^{12} - \left(\frac{C}{R_{HB}} \right)^{10} \right] \tag{4.9}$$

The inverse of the 10^{th} power represents the attractive force rather than the 6^{th} power term that is used in the L-J potential. The remaining terms represent the same things as in the 12-6 L-J potential. The versions of AMBER used in this study, do not include any hydrogen bonding term.

4.1.3.2 Electrostatic (Coulombic)

The electrostatic interactions in many force fields are parametrised by the Coulomb potential.[189] This non-bonding interaction arises when molecules have an unequal distribution of charges. The term accounts for differences in electro-negativity within the system. For AMBER the electrostatic potential is defined as Equation (4.10):

$$U_{AB} = \frac{(Q_A Q_B)}{4\pi\varepsilon_0 R_{AB}} \tag{4.10}$$

where the term Q represents the atomic charges of the two atoms being assessed. In AMBER a predetermined (from experimental data and quantum mechanical calculations), specific charge model is used. Atomic partial charges have on occasion been assigned by empirical rules, however, they are more commonly determined

by fitting to an electrostatic potential that is calculated by electronic structure methods. The R_{AB} term accounts for the distance between the two atoms. ε_0 accounts for the electrical field in and around the free space of the atoms as well as the effect of the dielectric medium.[189]

AMBER describes them as individual atomic monopoles which can interact.[190]

4.2 AMBER Force Field

The same potential energy equations have been used in all versions of AMBER (Assisted Model Building with Energy Refinement) since 1994, including AMBER03 and AMBER99-SB (SB = improved backbone torsion potentials) (version 14, 15 and 16), shown in Equation (4.11).[174]

$$\begin{aligned} U = \sum_{Stretch} k_{AB}(R_{AB} - R_{eAB})^2 + \sum_{Torsion} k_{ABCD}(1 - \cos(n(\chi_{ABCD} - \chi_{eABCD}))) \\ + \sum_{Bend} k_{ABC}(\theta_{ABC} - \theta_{eABC})^2 + \sum_{Out-of-Plane} k(\psi - \psi_e)^2 \\ + \sum_{Van-der-Waals} \varepsilon \left[\left(\frac{C}{R}\right)^{12} - 2\left(\frac{C}{R}\right)^{6} \right] \\ + \sum_{Electrostatic} \frac{(Q_A Q_B)}{4\pi\varepsilon_0 R_{AB}} \end{aligned} \tag{4.11}$$

The differences between AMBER03 and AMBER99-SB force fields occur in two aspects.[191] The primary difference is AMBER99-SB has a new fixed-charge calculation method parametrised from the most up to date experimental data. The

second difference is an update of atom charges which are determined using an improved basis set and restrained electrostatic potential (RESP) fitting. This has shown an improved energy balance between extended and helical regions of a peptide and the protein backbone.[173] The improvements are suspected to be due to the lack of dependence of side chain parameters on backbone conformations.[172]

4.3 Molecular Dynamics (MD)

Molecular dynamics (MD) simulations are trajectories of a series of successive configurations for a system, generated by integrating Newton's laws of motion.[192] This series of configurations describes how the positions and velocities of atoms (or particles) vary with time.[193] The intra- and inter-molecular motions of biomacromolecules that are described by trajectories can be associated with complex chemical processes such as reaction potentials, zymogen activation and hormone-receptor binding just to name a few.[193]

The thermodynamics of a system identifies possible chemical states and describes the energetic relationship between them.[194] A system's kinetics describes the sequence and/or rate of change from state to state providing a more mechanical relationship between the chemical states. These intricate changes are computationally studied through MD simulations by sampling the conformational space.[194]

4.3.1 Newton's Equations

Newton's equations are solved in MD to describe the motions of atoms on a potential energy surface. Molecules are able to overcome energy barriers that are smaller

than the difference between total and potential energy.[195] This means, given there is enough energy in the system, the simulation will explore the entirety of the potential energy surface as potential energy is influenced by simulation temperature. However exploring the totality of the potential energy surface would require an unreasonable amount of time.[196]

Newton's laws of motion are defined as:

1. A body will move in a straight line at constant velocity unless a force acts upon it.
2. Force is equal to the rate of change of momentum. (Force (F) = Mass (m) **x** Acceleration (a))
3. For every action there is an equal and opposite reaction.

Molecular dynamics trajectories are obtained by solving differential equations embodied within Newton's second law, ($F = ma$), shown in Equation (4.12).

$$\frac{d^2 r_i}{dt^2} = \frac{F_{r_i}}{m_i} \tag{4.12}$$

The equation describes the movement of mass (m_i) along a single Cartesian axis (r_i) where F_{r_i} describes the force in that direction in relation to time (t). In molecular dynamics it is solved for all three Cartesian directions taking into account the position of an atom's mass relative to the other particles, affecting the strength of force exhibited. The rate and direction of an atom's motion is dictated by the force which in turn is governed by the forces atoms exert on each other in a system.[195]

$$\frac{-dE}{dr_i} = F_{r_i} \tag{4.13}$$

MD simulations calculate the force exerted on an atom by calculating the change in energy from its current position and a position a small distance away. This is known as the derivative of energy, described in Equation (4.13).[197] The energies are calculated by the MM force field used for the simulation.

Velocities of atoms are unknown at the start of any molecular dynamics simulation. Initial velocities are assigned that satisfy the total kinetic energy of the system (by obeying the Boltzmann or Gaussian distribution from the assigned temperature of the system).[198] The system is normally heated slowly during the first steps of a molecular dynamics simulation to avoid the physical and numerical instabilities that would be caused by a rapid temperature jump.[196]

4.3.2 Verlet Algorithm

The Verlet algorithm integrates Newton's equation of motion to compute new atomic positions using the positions and accelerations from the previous step.[192] Although a number of algorithms are available for the integration of equations of motion, the Verlet algorithm is widely used in MD due to its use of minimal computer memory and CPU time. The method writes two third-order Taylor expansions for atomic positions; one forward in time and one backwards. Velocity and acceleration are defined by v and a respectively in Equations (4.15) and (4.16).

$$Where: \quad v = \frac{dr_i(t)}{dt} \qquad a = \frac{d^2r_i(t)}{dt^2} \tag{4.14}$$

$$r_i\left(\overrightarrow{t+dt}\right) = r_i\left(t\right) + v\left(t\right)dt + \left(\frac{1}{2}\right)a\left(t\right)dt^2 + \left(\frac{1}{6}\right)\frac{d^3r_i\left(t\right)}{dt^3}dt^3 + O\left(dt^4\right) \quad (4.15)$$

$$r_i\left(\overleftarrow{t-dt}\right) = r_i\left(t\right) - v\left(t\right)dt + \left(\frac{1}{2}\right)a\left(t\right)dt^2 - \left(\frac{1}{6}\right)\frac{d^3r_i(t)}{dt^3}dt^3 + O\left(dt^4\right) \quad (4.16)$$

Combining the two Taylor expansion equations give the basic form of the verlet algorithm, Equation (4.17).

$$r_i\left(t+dt\right) = 2r_i\left(t\right) - r_i(t-dt) + a\left(t\right)dt^2 + O\left(dt^4\right) \quad (4.17)$$

The truncation error is to the order dt^4, even though no third derivatives are explicitly present. Acceleration can be calculated by substituting in a function of atomic position, Equation (4.18), in place of force in Newton's equation to provide Equation (4.19).

$$F(r_i) = -\frac{dU}{dr_i} \quad (4.18)$$

$$a\left(t\right) = -\frac{1}{m_i}\frac{dU}{dr_i} = \frac{d^2r_i(t)}{dt^2} \quad (4.19)$$

This form of the Verlet algorithm does not directly generate velocities which is problematic as velocities are required to calculate kinetic energy (K), which in turn is used to assess if total energy (E) is conserved, ($E = K + U$). Velocities are also required to calculate (and, using thermostats, moderate) temperature within a system. They can be calculated separately via Equation (4.20).

$$v = \frac{dr_i(t)}{dt} = \frac{r_i\,(t + dt) - r_i\,(t - dt)}{2\,(dt)} \tag{4.20}$$

Calculating kinetic energy at a point in time results in an error to the order of dt^2 rather than dt^4 as shown by Equation (4.21).

$$K\,(t) = \sum_{i=1}^{n} \frac{mv_i(t)^2}{2} \tag{4.21}$$

To overcome this difficulty variants of the Verlet algorithm have been developed which handle velocity calculation somewhat better. One such variant is known as the Velocity Verlet algorithm.

4.3.3 Velocity Verlet Algorithm

The Velocity Verlet algorithm uses the positions, accelerations and velocities of the current time step to compute the positions of the next time step.[199] This method generates a far more accurate integration. The algorithm incorporates a step which rescales the velocities to apply a correction for any minor integration errors. This ensures that the simulation is carried out at the correct temperature in a constant-temperature system.[200] An algorithm named the Gear predictor-corrector algorithm is sometimes used as an add-on to Velocity Verlet algorithm.[192] It is used to predict the next set of atomic positions and accelerations at the expense of CPU memory and time, then compares the predicted to the calculated, generating a correction so each step is refined iteratively.

The position of atoms (r_i) is calculated at every time step (dt), by Equation (4.22).

$$r_i\left(t+dt\right)=r_i\left(t\right)+v\left(t\right)dt+\left(\frac{1}{2}\right)a\left(t\right)dt^2 \tag{4.22}$$

The velocity (v) is calculated every half time step ($\frac{dt}{2}$), by Equation (4.23).

$$v\left(t+\frac{dt}{2}\right)=v\left(t\right)+\left(\frac{1}{2}\right)a\left(t\right)dt \tag{4.23}$$

The acceleration (a) is calculated every time step (dt), by Equation (4.24).

$$a\left(t+dt\right)=\frac{-1}{m_i}(dU)\left(r_i(t+dt)\right) \tag{4.24}$$

Velocity at the next step is calculated by a variation of Equation (4.23), shown by Equation (4.25).

$$v\left(t+dt\right)=v\left(t+\frac{dt}{2}\right)+\left(\frac{1}{2}\right)a\left(t+\frac{dt}{2}\right)dt$$

The Velocity Verlet algorithm is faster and more accurate compared to the basic Verlet algorithm.[161] It also requires less computer memory to run which is advantageous.

TABLE 4.1: Description of the series of calculations done by the Velocity Verlet algorithm in terms of time steps. Where r_i is the atomic coordinates, v is the velocity and a is the acceleration.

Calculations	r_i	v	r_i, v, a	v	r_i, v, a
Time Step	1		2		3

Table 4.1 shows the series of calculations that take place in the Velocity verlet algorithm. The atomic coordinates and accelerations are calculated every time

step. The velocity of atoms is calculated every half-time step. Which ensures a more accurate trajectory is obtained, satisfying the kinetic energy and in turn the total energy of the system. This also helps to ensure the system is simulated at the desired temperature.

4.3.4 Affecting Factors

4.3.4.1 Ensembles

Ensembles are simply a collection of all possible systems which have varying microscopic states but indistinguishable macroscopic or thermodynamic states. The concept was first introduced in 1878 by J. Willard Gibbs and has since been developed further for computational implementation.[201]

Simulations can be characterised by features such as volume (V), pressure (P), temperature (T), total energy (E), number of particles (N), chemical potential (μ), etc. However, these are not always independent factors.[192] For a system which requires a constant number of particles to be set, either the pressure or volume must be fixed, but both cannot be fixed simultaneously. Likewise if the temperature is fixed, total energy can't be. If a constant chemical potential is set for a system the number of particles must vary. The different ensembles are described according to the fixed parameters, shown by Table 4.2.

TABLE 4.2: Description of the constants in different ensembles and the corresponding equilibrium states. [N = number of particles; P = pressure; T = temperature; V = volume; E = total energy; μ = chemical potential].

N	P	V	T	E	μ	**Acronym**	**Name**
✓		✓	✓			NVT	Canonical
✓		✓		✓		NVE	Micro-canonical
✓	✓		✓			NPT	Isothermal-isobaric
		✓	✓		✓	VEμ	Grand canonical

MD simulations that aim to preserve energy use the NVE ensemble.[192] However, in systems where mechanisms are being investigated, the isothermal-isobaric NPT ensemble is the ensemble preferred.

4.3.4.2 Simulation Temperature and Thermostats

Computer simulations treat temperature as a statistical quantity.[192] It is typically expressed as a function of a system's atomic positions and momenta.[202] For large systems the temperature can be estimated using kinetic energy data.[161]

Thermostats are simply algorithms that are used to rescale the velocities in a system to control the temperature.[192] The Berensden thermostat is regarded as the most straight forward method, but is not the most accurate.[194] The method rescales velocities over a specified number of time steps to keep a constant temperature, but this results in small but rapid fluctuations in temperature. Langevin dynamics is a more advanced method which is more commonly used in many MD packages.[193] This method rescales velocities more often which results in less fluctuations in temperature.

4.3.4.3 Periodic Boundary Conditions

Periodic boundary conditions describe a simulation's structure using a collection of uniform subunits. The setting is primarily used to simulate and model bulk-material, crystalline systems and the movement of solvent.[194] Figure 4.7 depicts periodic boundaries, where the middle cell is the simulation cell and surrounding cells are identical copies.

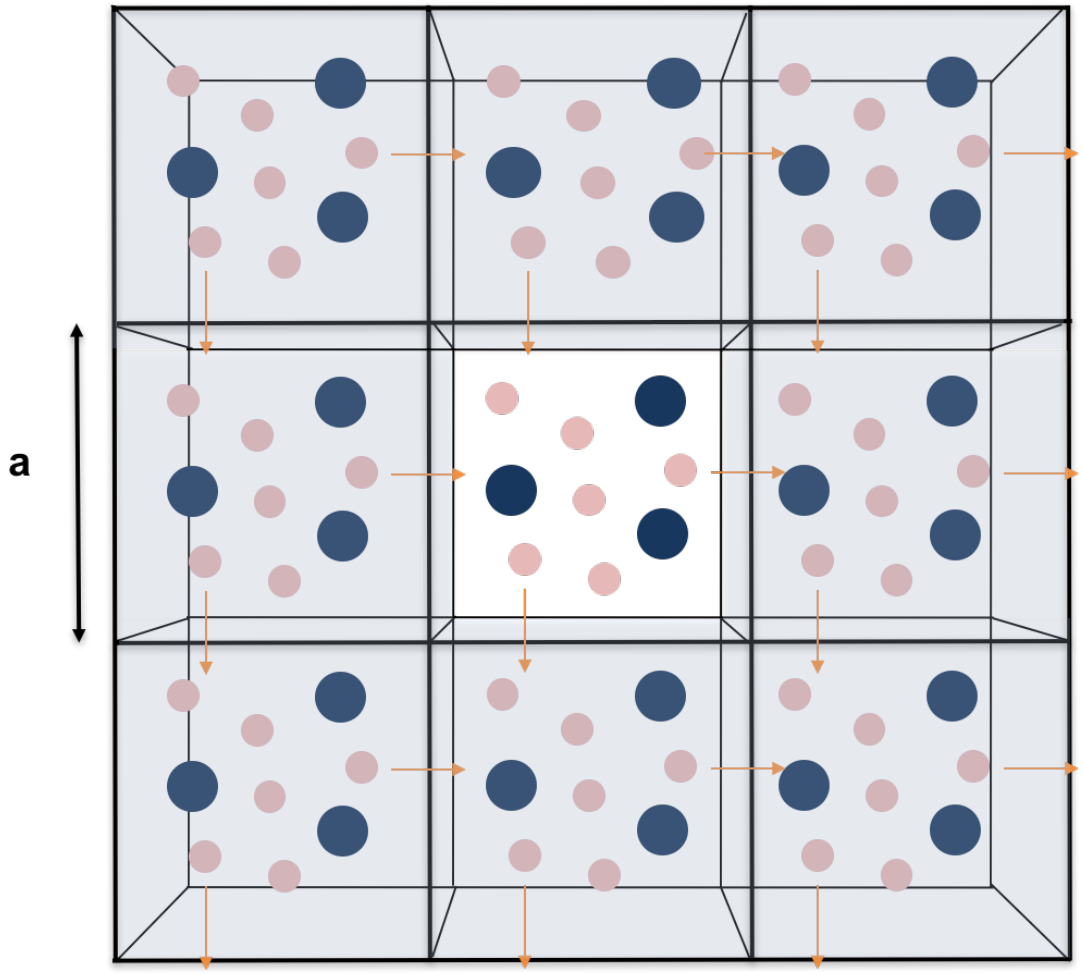


FIGURE 4.7: Schematic illustrating periodic boundary conditions.

When an atom or cell is forced to move out of the simulation cell, it theoretically enters the adjacent identical cell. This is represented in the simulation cell by the atom re-entering from the opposite face simultaneously. The simulation box is typically large enough to prevent an atom from interacting with itself. The subunit is normally set to be at least twice as big as the van der Waals and electrostatic interaction range. A cut-off can also be introduced to prevent these unwanted interactions, which is typically set as half the cell length (**a**). The use of periodic boundary conditions eliminates the occurrence of unwanted wall effects which would adversely affect results. It can also make calculations computationally less expensive as calculations are limited to the desired space.

4.3.4.4 Particle Mesh Ewald (PME)

Particle mesh Ewald (PME) is a method used to compute (electrostatic) interaction energies of periodic systems.[203] The interactions are separated into two groups; short and long range. The method replaces the direct interaction between two particles with two separate summations; a sum of short range potential in real space (simulation box) and a sum of long range potential in Fourier space (adjacent periodic boxes).[204] The summations converge quickly and can be shortened with little adversities if computational time needs reducing. The method uses the fast Fourier transform to evaluate the density field of a lattice in space but this can also be applied to periodic systems. The unit cell in periodic systems needs to be large enough to circumvent any improper interactions through a cell face but must also be small enough to be computationally inexpensive. PME is preferred in computational chemistry as it is more accurate and less expensive than using a larger cut-off.[205]

4.3.5 Cost Reductions

The most demanding steps of MD simulations are the calculation of non-bonding potentials; electrostatic and van der Waals interactions. These interactions should be calculated up to infinitum but at large distances they become infinitesimal. The range is usually limited to approximately 12 Å to reduce computational time. However, even with a cut-off these calculations are still the most demanding aspect of computational simulations.

Usually in complex systems bond stretches which include hydrogen atoms and water molecules are deemed insignificant. In computational systems these can be constrained by algorithms such as SHAKE,[206,207] which allows for the user to constrain all bond lengths including hydrogen atoms and all water molecules to be kept rigid. This greatly reduces simulation time in large systems without any

great loss in accuracy. By constraining the high-frequency vibrations involving hydrogen atoms, it also allows for a 2fs (rather than 1fs) time step to be used in the MD algorithms, further reducing computational expense.

4.4 Solvent Models

There are a number of solvent models reported in the literature that account for the behaviour of solvated condensed phases. Solvent models are employed in chemical simulations and thermodynamic calculations to study reactions and processes which take place in solution, including biological, chemical and environmental processes. Implicit models are widely reported in the literature, typically providing a reasonable description of the solvent behaviour. However, they fail to account for the local fluctuations in solvent density distribution around a solute molecule. Explicit models aim to provide a physical spatial distribution description of the solvent. However, the calculations are very demanding computations and can fail to reproduce experimental results. Hybrid methodologies aim to provide a good median between implicit and explicit methods. They incorporate aspects of implicit and explicit, minimising computational costs whilst retaining the physical spatial distribution description of the solvent. The three dimensional reference interaction site model (3D-RISM) is a hybrid solvent model that is well documented in the literature and described in the sections that follow.

4.5 Reference Interaction Site Model (RISM)

Microscopic effects of solvents near biomolecular surfaces play a critical role in mediating ligand binding. Accurate representation of these minute effects are necessary to generate highly accurate models of molecular systems. In recent times there have been increasing efforts to understand solvation effects on biomolecular complexes.

Limited capabilities and spatial resolution has hindered experimental methods from effectively analysing the behaviour of solvent molecules in protein complexes.

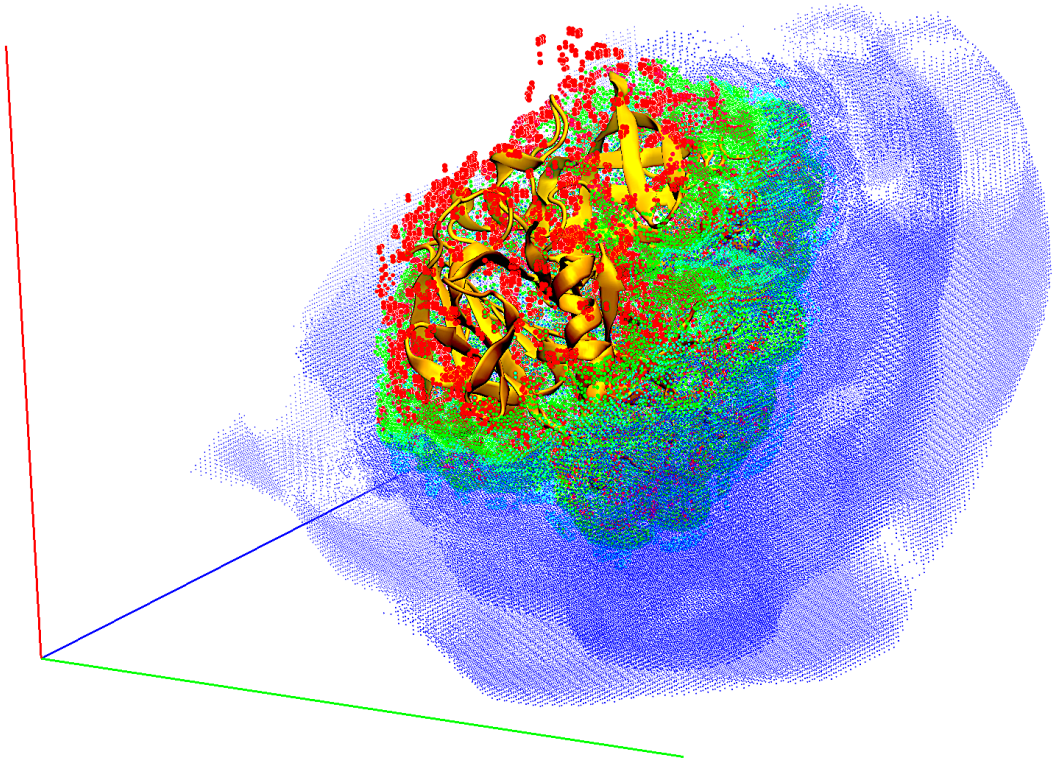


FIGURE 4.8: Depiction of 3D-RISM solvent distribution around chymosin, highlighting various solvation shells in *red*, *green* and *blue*.

The solvent sampling issue is circumvented by the three-dimensional reference interaction site model (3D-RISM) theory.[153] The method obtains a complete atomistic sampling of the solvent (including the ions) through the integral equation approach, as shown in Figure 4.8. 3D-RISM has successfully been used to locate water molecules in a number of systems of experimental proteins,[208–211] and

simulated proteins.[212,213] It has also been applied to predict protein fragment poses,[214–216] hydration free energies,[217] ion locations,[218] ion pathways[219,220] and drug poses.[216,221]

4.5.1 3D-RISM

The method is described as a solvent interaction-site interpretation of the molecular Ornstein-Zernike (MOZ) algorithm in three dimensions.[222] The solvation properties at equilibrium are obtained without the use of any dynamic simulation.

3D-RISM[153,214,223,224] is a theoretical method for modelling solution phase systems based on classical statistical mechanics. The equations relate 3D intermolecular *solvent site - solute* total correlation functions ($h_\alpha(\mathbf{r})$), and direct correlation functions ($c_\alpha(\mathbf{r})$) (index α corresponds to the solvent sites):[153,224]

$$h_\alpha(\mathbf{r}) = \sum_{\xi=1}^{N_{Solvent}} \int_{R^3} c_\xi(\mathbf{r}-\mathbf{r}')\chi_{\xi\alpha}(|\mathbf{r}'|)d\mathbf{r}' \tag{4.25}$$

where $\chi_{\xi\alpha}(r)$ is the bulk solvent susceptibility function, and $N_{Solvent}$ is the number of sites in a solvent molecule (see Figure 4.9). The solvent susceptibility function $\chi_{\xi\alpha}(r)$ describes the mutual correlations of sites ξ and α in solvent molecules in the bulk solvent. It can be obtained from the solvent intramolecular correlation function ($\omega^{Solv}_{\xi\alpha}(r)$), site-site radial total correlation functions ($h^{Solv}_{\xi\alpha}(r)$) and the solvent site number density (ρ_α): $\chi_{\xi\alpha}(r) = \omega^{Solv}_{\xi\alpha}(r) + \rho_\alpha h^{Solv}_{\xi\alpha}(r)$ (from here onwards we imply that each site is unique in the molecule, so that $\rho_\alpha = \rho$ for all α).[224]. In this work, these functions were obtained by solution of the RISM equations of the solvent.[224,225]

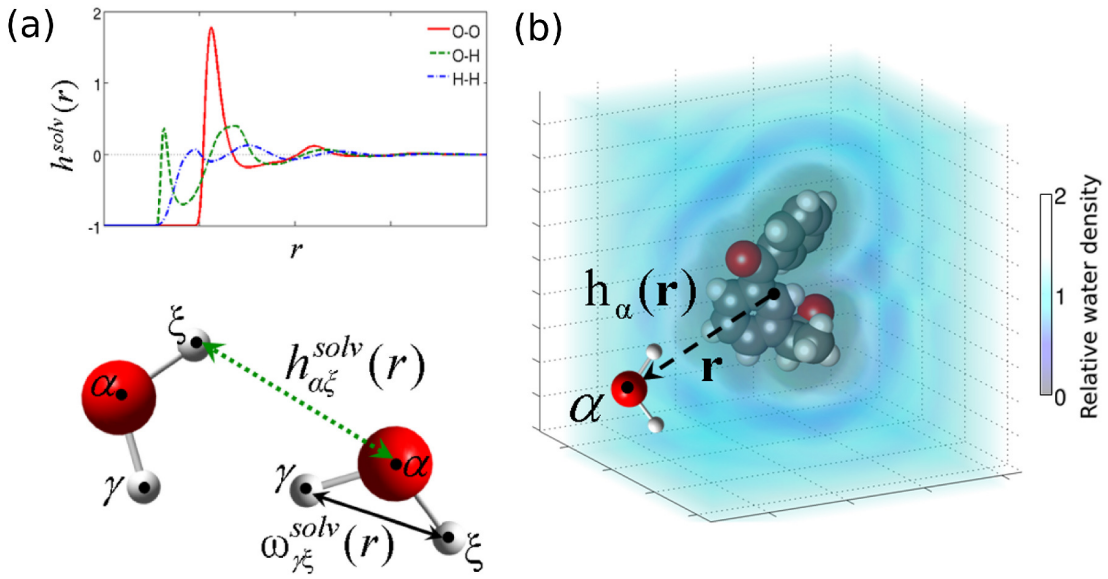


FIGURE 4.9: Correlation functions in the 3D-RISM approach. (a) Site-site intramolecular ($\omega^{solv}_{\gamma\xi}(r)$) and intermolecular ($h^{solv}_{\alpha\xi}(r)$) correlation functions between sites of solvent molecules. The graph shows the radial projections of water solvent site-site density correlation functions: oxygen-oxygen (OO, red solid), oxygen-hydrogen (OH, green dashed) and hydrogen-hydrogen (HH, blue dash-dotted); (b) Three-dimensional intermolecular solute-solvent correlation function $h_\alpha(\mathbf{r})$ around a model solute (diclofenac). This figure is based on Reference [2].[2]

In order to calculate $h_\alpha(\mathbf{r})$ and $c_\alpha(\mathbf{r})$, $N_{Solvent}$ *closure* relations are introduced:

$$\begin{aligned} h_\alpha(\mathbf{r}) &= \exp(-\beta u_\alpha(\mathbf{r}) + h_\alpha(\mathbf{r}) - c_\alpha(\mathbf{r}) + B_\alpha(\mathbf{r})) - 1 \\ \alpha &= 1, \ldots, N_{Solvent} \end{aligned} \tag{4.26}$$

where $u_\alpha(\mathbf{r})$ is the 3D interaction potential between the solute molecule and α solvent site, $B_\alpha(\mathbf{r})$ are bridge functionals, $\beta = 1/k_B T$, k_B is the Boltzmann constant, and T is the temperature.

In general, the exact bridge functionals $B_\alpha(\mathbf{r})$ in Equation 4.26 are represented as an infinite series of integrals over high order correlation functions and are therefore practically incomputable, which makes it necessary to incorporate some approximations,[224,226,227] or to estimate the form of these functionals from molecular

simulation[228]. In this thesis, a closure relationship proposed by Kovalenko and Hirata (the KH closure also known as partial series expansion order 1 (PSE-1)),[229] along side the PSE-3 closure are investigated. Both, the KH and PSE-3 closures were designed to improve convergence rates and to prevent possible divergence of the numerical solution of the RISM equations[229–231]

$$h_\alpha(\mathbf{r}) = \begin{cases} \exp(\Xi_\alpha(\mathbf{r})) - 1 & when \quad \Xi_\alpha(\mathbf{r}) \leq 0 \\ \Xi_\alpha(\mathbf{r}) & when \quad \Xi_\alpha(\mathbf{r}) > 0 \end{cases} \tag{4.27}$$

$$\text{where } \Xi_\alpha(\mathbf{r}) = -\beta u_\alpha(\mathbf{r}) + h_\alpha(\mathbf{r}) - c_\alpha(\mathbf{r}).$$

The PSE-3 closure was designed to minimise any convergence issues that arise from the KH (Equation 4.27) closure in a systematic manner by using a partial series expansion of order n (PSE-n) in the hypernetted chain (HNC) closure (Equation 4.28).[230,231]

$$h_\alpha(\mathbf{r}) = \begin{cases} \exp(\Xi_\alpha(\mathbf{r})) - 1 & when \quad \Xi_\alpha(\mathbf{r}) \leq 0 \\ \sum_{i=0}^{n} (\Xi_\alpha(\mathbf{r}))^i / i! - 1 & when \quad \Xi_\alpha(\mathbf{r}) > 0 \end{cases} \tag{4.28}$$

The PSE-n closures interpolates between the KH and the HNC closures: $n = 1$ is the KH closure; $n >1$ is the HNC closure; $n \to \infty$ will result in convergence issues.[230] PSE closure of order 3 achieves a good balance between numerical convergence and results that well approximate calculations with the HNC closure.[231] Both the KH and PSE-3 closure have extensively been applied in the study of a variety of polar and charged systems.[232–236]

The 3D interaction potential between the solute molecule and α site of solvent ($u_\alpha(\mathbf{r})$, 4.26) is estimated as a superposition of the site-site interaction potentials between solute sites and the particular solvent site, which depend only on the absolute distance between the two sites. In the research presented in this thesis, the

common form of the site-site interaction potential is used, which is represented by the long-range electrostatic interaction term and the short-range term (Lennard-Jones potential).[237]

4.5.2 Solvation Free Energy Functionals

Solvation free energy (SFE) is the reversible work required to take a compound out of the gas phase and into the solvent phase. Within the framework of the RISM theory there exist several approximate functionals that allow one to analytically obtain values of the SFE from the total $h_\alpha(\mathbf{r})$ and direct $c_\alpha(\mathbf{r})$ correlation functions.[238–240].

4.5.2.1 Partial Series Expansion-3 (PSE-3)

For the PSE-3 functional, the solute's excess chemical potential (SFE) at infinite dilution is derived from the 3D-RISM solute-solvent correlation functions as follows (Equation 4.29):

$$\Delta G_{Solv}^{PSE-3} = \Delta G_{Solv}^{HNC} - k_B T \sum_{\alpha=1}^{N_{solvent}} \rho_\alpha \int_V \left[\Theta[h_\alpha(\mathbf{r})] \frac{\Xi_\alpha(\mathbf{r})^{n+1}}{(n+1)!} \right] d\mathbf{r} \tag{4.29}$$

where ρ_α is the number density of solvent sites α, Θ is a Heaviside step function, and ΔG_{Solv}^{HNC} is the SFE calculated using the hypernetted-chain functional, which is given by:[241]

$$\Delta G_{Solv}^{HNC} = k_B T \sum_{\alpha=1}^{N_{Solvent}} \rho_\alpha \int_V \left[\frac{1}{2} h_\alpha^2(\mathbf{r}) - \frac{1}{2} h_\alpha(\mathbf{r}) c_\alpha(\mathbf{r}) - c_\alpha(\mathbf{r}) \right] \tag{4.30}$$

4.5.2.2 Gaussian Fluctuations (GF)

Developed by Chandler, Singh and Richardson, for 1D-RISM, and adopted by Kovalenko and Hirata for the 3D-RISM case[224,242], the Gaussian fluctuations (GF) free energy functional is given as:

$$\Delta G_{Solv}^{GF} = k_B T \sum_{\alpha=1}^{N_{Solvent}} \rho_\alpha \int_{R^3} \left[-\frac{1}{2} c_\alpha(\mathbf{r}) h_\alpha(\mathbf{r}) - c_\alpha(\mathbf{r}) \right] d\mathbf{r} \tag{4.31}$$

4.5.2.3 Kovalenko-Hirata (KH)

The KH free energy functional for 3D-RISM is given by:

$$\Delta G_{Solv}^{KH} = k_B T \sum_{\alpha=1}^{N_{Solvent}} \rho_\alpha \int_{R^3} \left[\frac{1}{2} h_\alpha^2(\mathbf{r}) \Theta(-h_\alpha(\mathbf{r})) - \frac{1}{2} h_\alpha(\mathbf{r}) c_\alpha(\mathbf{r}) - c_\alpha(\mathbf{r}) \right] d\mathbf{r} \tag{4.32}$$

where ρ_α is the number density of solvent sites α, and Θ is the Heaviside step function:

$$\Theta(x) = \left\{ \begin{array}{l} 1 \text{ for x>0} \\ 0 \text{ for x<0} \end{array} \right\} \tag{4.33}$$

The solute partial molar volume is estimated via *solute-solvent site* correlation functions using the standard 3D-RISM theory expression[243]:

$$V = k_B T \eta \left(1 - \rho_\alpha \sum_{\alpha=1}^{N_{Solvent}} \int_{R^3} c_\alpha(\mathbf{r}) d\mathbf{r} \right) \tag{4.34}$$

where η is the pure solvent isothermal compressibility, and ρ_α is the number density of solute sites α. The distribution functions ($g(\mathbf{r}) = h(\mathbf{r}) + 1$) calculated by 3D-RISM characterize the average density distribution of solvent molecules around solute at thermodynamic equilibrium.

4.5.3 Pressure Corrected Free Energy Functional

4.5.3.1 3D-RISM(PC)

The PC free energy functional is designed as an improvement on the standard 3D-RISM SFE functionals that over estimate the solvent pressure. To counteract this the PC functional subtracts all mechanical work required to create the cavity ($P\Delta V$) from the SFE ($\Delta G_{Solv}^{3D-RISM}$), as shown in Equation 4.35.[244,245]

$$\Delta G_{Solv}^{PC} = \Delta G_{Solv}^{3D-RISM} - P\Delta V \tag{4.35}$$

Here P represents the 3D-RISM pressure and ΔV represents the volume change of the system upon solvation. $P\Delta V$ is computed using methods described by Misin *et al.*[241,245] ΔG_{Solv}^{PC} simply refers to the pressure corrected solvation free energy, 3D-RISM(PC).

4.5.3.2 3D-RISM(PC+)

3D-RISM(PC+) The PC+ free energy functional is a further improvement on the PC functional where just the non-ideal mechanical work is subtracted from the hydration free energy. To accomplish this the ideal gas pressure, P_{id} is used to represent the ideal mechanical work, $P_{id}\Delta V$ and is added to Equation 4.35.

$$\Delta G_{Solv}^{PC+} = \Delta G_{Solv}^{3D-RISM} - P\Delta V + P_{id}\Delta V \tag{4.36}$$

Here, and in the formula for the PC functional, $\Delta G_{Solv}^{3D-RISM}$ is the SFE calculated using the PSE-3 free energy functional (Equation 4.29). Although there is no compelling explanation as to why PC+ performs better than PC, there have been numerous reports of its benefits in the literature.[241,244–247] The PC+ functional has been shown to give accurate predictions of SFE for neutral and ionised solutes, in both pure water and salt solutions at a wide-range of temperatures.[237,241,245,248,249] It has also been successfully applied to the prediction of solvation free energies in organic solvents.[246]

4.6 Calculation of ΔG_{bind}

The thermodynamic parameter that characterizes the binding of a ligand (L) by a receptor (R) is the binding free energy (ΔG_{Bind}) for the process:[250]

$$R + L \overset{\Delta G_{Bind}}{\longrightarrow} RL \tag{4.37}$$

The most common computational methods in drug design are docking and scoring.[251] These methods predict the binding mode of the drug and then go on to estimate the binding affinity. Although they are efficient methods they are not particularly accurate; they can discriminate well between binding and non-binding drugs, but fail to discriminate between drugs that differ by ΔG_{Bind} <6 kJ/mol.[252]

Methods such as thermodynamic integration (TI) and free energy perturbation (FEP) are considered thermodynamically rigorous and widely accepted in the field

for calculating relative binding free energies between two equilibrium states. However both methods require a great deal of simulation time to provide adequate sampling, making them unsuitable for large scale studies. To tackle this practical issue, many different end-point techniques have been developed to predict binding free energies at lower computational expense i.e. the linear-interaction-energy (LIE) approach,[253] and the closely related molecular mechanics generalised Born surface area (MM-GBSA), molecular mechanics Poisson-Boltzmann surface area (MM-PBSA),[254,255] and MM-3DRISM methods. Where FEP and TI spend the majority of simulation time investigating intermediate states, end-point techniques investigate just two (bound and unbound) states, significantly reducing computational cost.

Arguably, the most popular end point method is the molecular mechanics with Poisson-Boltzmann and surface area solvation (MM-PBSA) method.[255] Here, the ΔG_{Bind} is calculated from free energies of the reactants and products. The method was developed in the late 90's by Kollman *et al.* and has been cited over 2500 times since,[254] having been applied to a number of scientific studies including protein-protein interactions,[256,257] protein design,[258] conformer stability[259,260] and re-scoring.[261,262]

Since the free energy of binding of a solvated complex is very hard to calculate directly, the free energy is calculated in the gas phase first and then the SFE is calculated next, depicted in Figure 4.10.

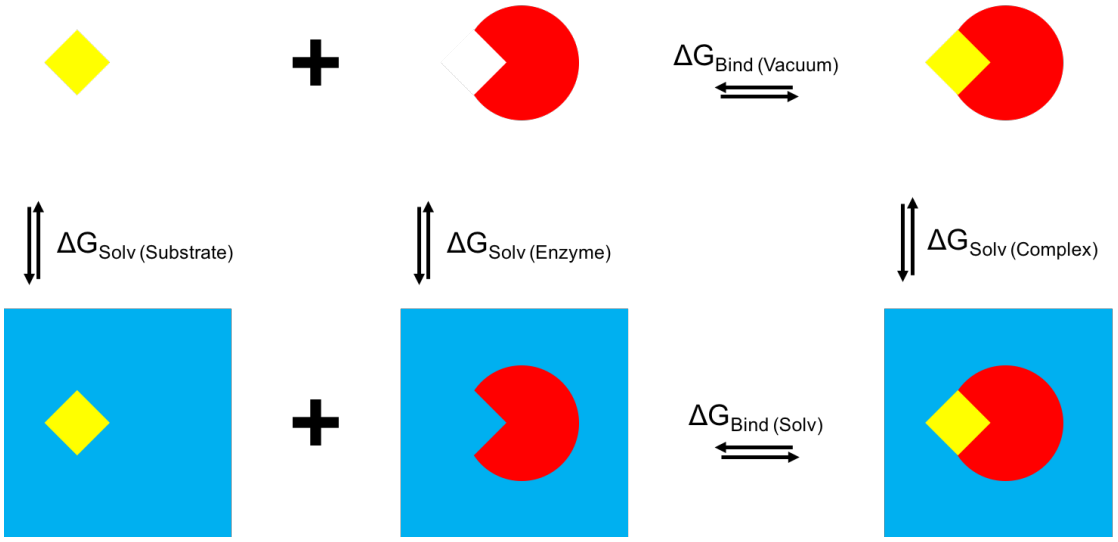


FIGURE 4.10: Representation of the calculation pathways to obtain the binding free energy of a solvated complex.

Here $\Delta G_{Bind(Solv)}$ represents free energy of binding of the complex in a solvent medium and $\Delta G_{Bind(Vacuum)}$ is the binding free energy in a vacuum. $\Delta G_{Solv(Substrate)}$, $\Delta G_{Solv(Enzyme)}$ and $\Delta G_{Solv(Complex)}$ represent the free energy of solvation.

The solvent environment strongly influences the binding free energy, modulating competitive solvent binding effects and hydrophobicity.[263] In this thesis, we employ the MM-3DRISM method, which uses a statistical mechanics based solvent model to provide a realistic model of molecular solvation effects. MM-3DRISM affords accurate estimates of binding free energy and has previously been used in modelling a wide-variety of protein-ligand complexes.[123,239]

In MM-3DRISM, the binding free energy is computed according to:

$$\Delta G_{Bind} = G_{Solvated(Complex)} - G_{Solvated(Enzyme)} - G_{Solvated(Substrate)} \tag{4.38}$$

The free energies of each species is evaluated as:

$$G_{Solvated} = \langle E_{Gas} \rangle + \langle \Delta G_{Hyd} \rangle - TS \tag{4.39}$$

$$\langle E_{Gas} \rangle = \langle E_{Internal} \rangle + \langle E_{Electrostatic} \rangle + \langle E_{vdW} \rangle \tag{4.40}$$

$$\langle E_{Internal} \rangle = \langle E_{Bond} \rangle + \langle E_{Angle} \rangle + \langle E_{Torsion} \rangle \tag{4.41}$$

E_{Gas} describes the average energy of a species in the gas phase as a sum of internal, electrostatic ($E_{Electrostatic}$) and van der Waals (E_{vdW}) energy contributions, obtained through a molecular mechanics forcefield. E_{Bond}, E_{Angle} and $E_{Torsion}$ contribute to the internal energy $E_{Internal}$ through the strain caused by deviation of bonds, angles, and torsions from their equilibrium values. ΔG_{hyd} describes the hydration free energy and is computed by the 3D-RISM calculation.

4.7 QSAR

4.7.1 Machine Learning

Machine learning is a computational approach to design and develop algorithms that can use empirical data to recognise statistical relationships and are able to automatically learn from experience with respect to a task and a performance measure. One of the major scientific applications of machine learning research is to recognise structure-activity relationships (SAR) of chemical compounds. The core part of machine learning is tasked with building predictive models. In SAR studies the models are used to predict biological activity using independent variables, also referred to as descriptors.[264–266] The idea is that the predictive models should be able to describe the data it trained on in some meaningful way (*representation*)

and perform accurately on new, unseen data (*generalisation*).

Within the machine learning umbrella, a number of advanced statistical methods exist, handling regression and classification tasks using multiple input variables to provide a single statistical output. These methods include support vector machines (SVM), naive bayes (NB), k-nearest neighbours (KNN), classification and regression trees (CART), multivariate adaptive regression splines (MARSplines), genetic algorithm (GA), random forest (RF), and others.[267] The various machine learning methods can be identified with one of two subgroups; supervised learning or unsupervised learning. Both classification and regression methods are typically supervised learning algorithms. The research presented in this thesis focuses on supervised-learning regression methods.

In this thesis two different regression methods were considered to create predictive models: partial least squares (PLS) and random forest (RF). A genetic algorithm (GA) was also tested to select input variables for the PLS model.

4.7.2 Regression

Regression models estimate the relationship between variables and a response. This includes a vast array of techniques that focus on a relationship between a dependant variable and a number of independent variables. Regression methods can be classed into three subgroups. Firstly there is linear regression which includes methods such as ordinary linear regression (OLR), partial least squares regression (PLS) and penalized regression (PR). Secondly, non-linear regression methods which include support vector machines (SVM), artificial neural networks (ANN) and multivariate adaptive regression splines (MARSplines). The final group is regression trees which are also non-linear by nature and these include bagging tree

regression (BTR), boosted tree regression and random forest (RF).

4.7.2.1 Linear Regression

Providing the dataset descriptors take the form of $\left\{x_{i_1}, x_{i_2}, ..., x_n\right\}^T_{i=1,2,...,n}$, where i is the descriptor number, the linear regression model takes on the form of Equation (4.42).

$$y = \beta_{i_1} x_{i_1} + \beta_{i_2} x_{i_2} + ... + \beta_{i_n} x_{i_n} + \epsilon \tag{4.42}$$

Here, y represents the continuous numeric response for descriptor set i. β_i is the regression coefficient associated with descriptor x_i, where x_i is the descriptor variable. ϵ accounts for the noise or random error that cannot be explained by the linear regression model. Equation (4.42) is simplified into a summarised form, Equation (4.43).

$$Y = X\beta + \epsilon \tag{4.43}$$

The main objective of the linear regression model is to estimate the regression coefficient vector (β) according to the variance-bias trade-off, where the mean squared error (MSE) is minimised. The regression coefficients possess high interpretability which means relationships between the coefficient and response as well as relationships between different regression coefficients can easily be interpreted. The performance of the predictive models can also be interpreted easily as the statistical nature of the linear method allows for the extraction of standard errors of the regression parameters.

However, as the relationship between the descriptor variable and last numeric

response is required to fall on a flat-plane (therefore be linear), a non-linear relationship between the regression coefficients and the predicted response cannot be explained by this model.

Partial-Least-Squares (PLS) Partial least squares (PLS) is a method for linear regression that has been widely used in many different fields of research, including chemistry, biology, econometrics and social science.[268] The PLS algorithm finds a linear regression model by projecting both the dependent and independent variables into a new mathematical space in which the covariance in the data structure can be explained by a small number of latent variables. As such PLS regression has some similarity to principal component regression (PCR), but the latent variables are selected for their ability to explain the variance in the dependent variable as well as in the independent variables.[269]

The main function of the PLS regression model is to determine a new set of potential components. These new potential components should then be able to explain the covariance between independent variables (X) and response (Y) by decomposing both X and Y.[270] The decomposition formula for descriptor variables is given in Equation 4.44.

$$X = TP + \epsilon \tag{4.44}$$

Here, X is the matrix score and T is the projection of X. P represents the orthogonal matrix loading, which in PCR is a simple variability loading instead and ϵ is the noise or error value.[271] A diagonal matrix of the regression weight loadings (B) then allows for the decomposition of the response Y via Equation 4.45, where C is the dependant variable weight matrix.

$$Y = TBC \tag{4.45}$$

In contrast to principle component analysis (PCA), PLS takes two steps to determine the best linear relationship; finding the linear components first then determining which components maximally correlate with the response (depicted in Figure 4.11).[272]

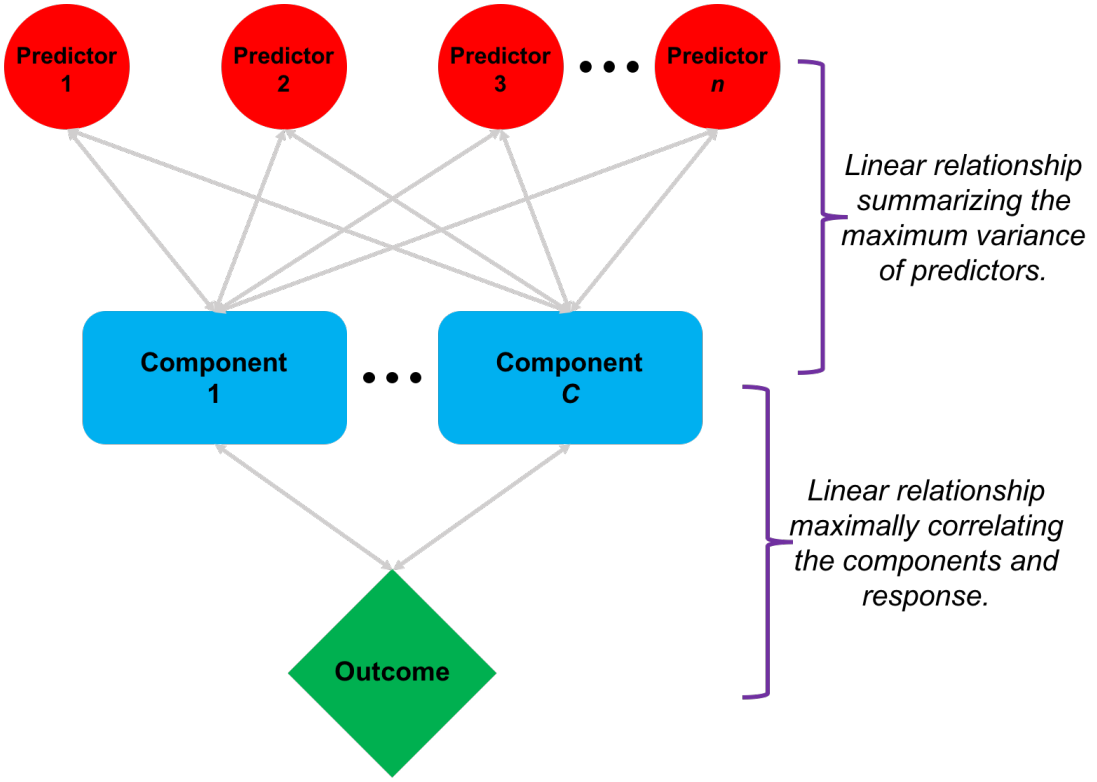


FIGURE 4.11: Structured flow of the PLS regression model.

It is worth emphasising the only tuning parameter in the PLS regression model is the number of components. This is usually determined through resampling techniques.[273]

Genetic Algorithm (GA) A genetic algorithm (GA) was used to select an optimal subset of descriptors for the PLS model. Genetic algorithms are commonly used to solve both constrained and unconstrained optimization problems

using a selection approach based on biological evolution.[274] The GA continuously modifies the population of descriptors of individual solutions at each iteration.[275] At each of these steps, the genetic algorithm selects a predetermined number of descriptors from the total population to be the initial seeds and uses them to produce the permutations for the next generation.[276,277] Over successive generations, the population "evolves" toward an optimal solution.[278]

4.7.2.2 Regression Trees

Regression tree models are a type of non-linear regression method. Typically they are used to predict continuous responses by dividing the dataset into smaller groups (i.e. trees to branches to leaves). The method allows for the descriptor variables used as input to be categorical, continuous, sparse, skewed, etc. without preprocessing requirements. The structures of these trees are easy to compute and interpret as they are intuitive. They can be allied easily to large datasets without any prior knowledge of the relationship between the predicted response and independent variables.

Random Forest (RF) Random forest is a method for classification and regression which was introduced by Breiman and Cutler.[279] The method is based upon an ensemble of decision trees, from which the prediction of a continuous variable is provided as the average of the predictions of all trees. Recent studies have suggested that random forest offers features which make it very attractive for statistical modelling studies.[280] These include relatively high accuracy of prediction, built-in variable selection, and a method for assessing the importance of each variable to the model.

In RF regression, an ensemble of regression trees is grown from separate bootstrap samples of the training data using the CART algorithm.[279] The branches

in each tree continue to be subdivided while the minimum number of observations in each leaf is greater than a predetermined value. Unlike regression trees, the branches are not pruned back. Furthermore, the descriptor selected for branch splitting at any fork in any tree is not selected from the full set of possible descriptors but from a randomly selected subset of predetermined size.

There are three possible training parameters for random forest: *ntree* - the number of trees in the forest; *mtry* - the number of different variables tried at each split; and *nodesize* - the minimum node size below which leaves are not further subdivided. In theory, as the *ntree* increases, so does the computational expense. However, due to the randomly selected descriptors being only a small part of the original descriptor set in the dataset, *ntree* can be set high and the computations can still be more efficient than other methods like bagging trees.

The bootstrap sample used during tree growth is a random selection with replacement from the molecules in the dataset. The molecules that are not used for tree growth are termed the *out-of-bag* sample. Each tree provides a prediction for its out-of-bag sample, and the average of these results for all trees provides an in situ cross-validation called the out-of-bag validation.

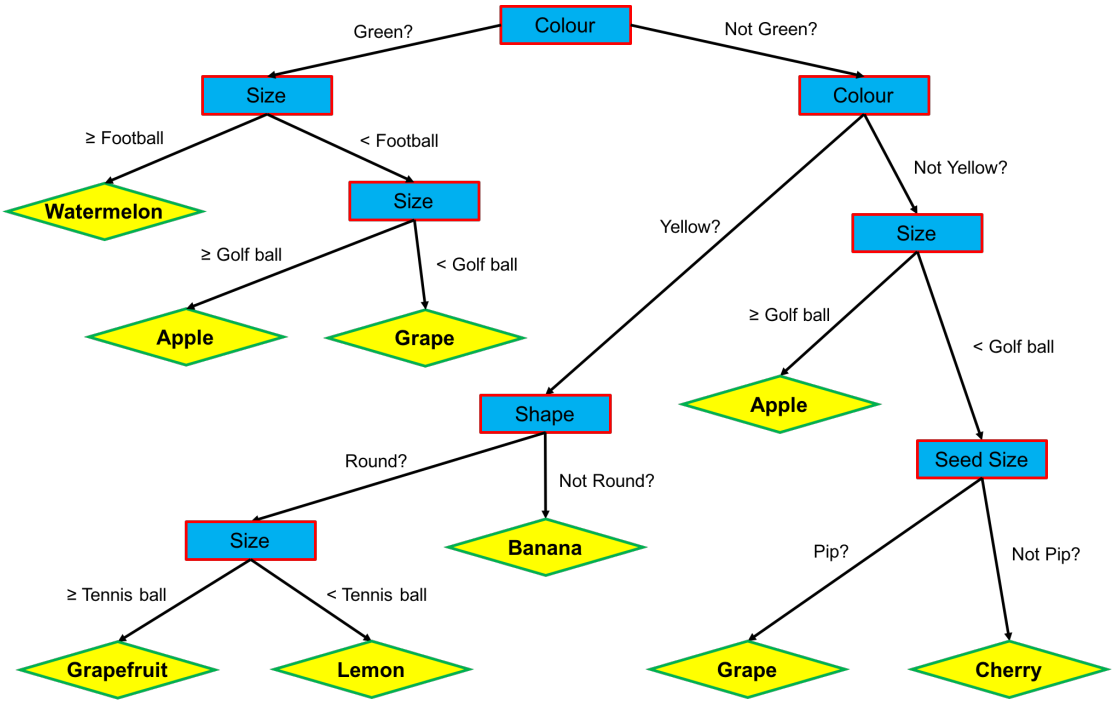


FIGURE 4.12: Example of a single RF decision tree to predict fruit type from physical data.

Figure 4.12 displays an example RF decision tree for the prediction of fruits using physical data. Here, the RF decision tree is shown to interpret missing data where there is no node that splits for red coloured fruit. Instead it interprets the group as 'not yellow' and 'not green' (shown on the right hand side of the tree) which consists of three end nodes (apple, grape and cherry). Therefore, RF is a good approach for large data sets with some missing data as it can still maintain good performance. The example in Figure 4.12 is categorical where the end nodes are categories, but the same logic would apply to numerical systems. Indeed, it is also possible to assign a numerical classification to each category (i.e. banana = 1, red apple = 2, green grape = 3 etc.) to predict a numerical value instead. One disadvantage of RF is that it is inaccurate when the predicted response is beyond the range of the observed outcomes in the training data. Although a prediction will be made it is likely to be inaccurate for those beyond this range.

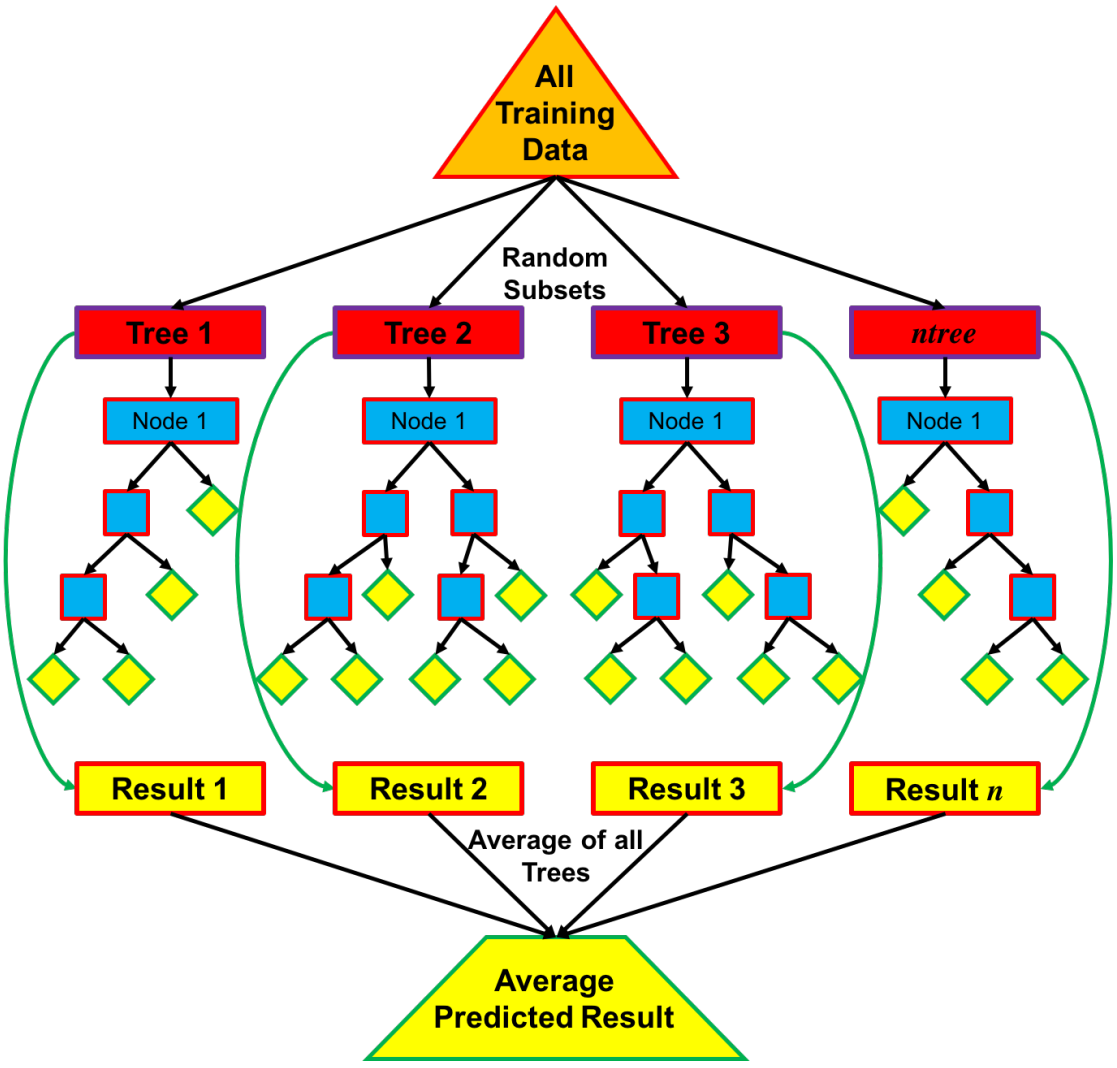


FIGURE 4.13: Schematic of an RF regression forest.

In RF numerous regression trees are built using a subset of the training data where the number of trees is defined by the user (*ntree*). A schematic of a RF regression forest is shown in Figure 4.13. Here, the training set is split into random subsets which can overlap. The model then creates the predetermined number of regression trees to generate a predictive result from each tree. In the final step, the average result from all the trees are taken to provide a single predictive result. During the training steps the model can change the random predictor subsets and trees to find the best model for the training set. When using the predictive model on a test set the trees and predictor subsets do not change.

Part III

Research

Chapter 5

Allosteric-Activation Mechanism Of Bovine Chymosin

5.1 Overview

Crystal structures of chymosin reveal that the side chain of residue Tyr77 in a β-hairpin flap region (shown in Figure 5.1) above the binding cleft in bovine chymosin can occupy two different positions.[84] The side chain can be extended over the binding pocket occupying the position where κ-casein binds (referred to as the self-inhibited conformation) or it can be extended back into the β-hairpin flap (referred to as the open conformation).[2,83] The transition from self-inhibited to open is associated with a rearrangement of the β-hairpin flap, which becomes more puckered in the open conformation. It has been widely reported that apo-chymosin exists in the self-inhibited form and that this is converted into its active open form by allosteric activation by the P8-P4 fragment of κ-casein, (His98-Pro99-His100-Pro101-His102,) the so called "*His-Pro*" cluster (shown in Figure 5.1).[16] Evidence for this allosteric activation mechanism came originally from the experiments of Visser et al.[18] and Gustchina et al.,[16] who measured the catalytic rates for proteolysis of different fragments of κ-casein. They observed a $\sim$200 fold reduction in catalytic rate for proteolysis of P2-P2' or P3-P3' fragments of κ-casein as compared

to the native substrate. However, the reduction in catalytic rate was not observed when chymosin had been pre-incubated with the P8-P4 residues of κ-casein. Taken together with crystallographic data,[83] which show that apo-chymosin occupies a self-inhibited conformation, the experiments carried out by Visser et al. suggest that the P8-P4 residues act as an allosteric-activator. Further mutagenesis studies have demonstrated that all five of the residues in the *His-Pro* cluster are important for catalysis.[17,18]

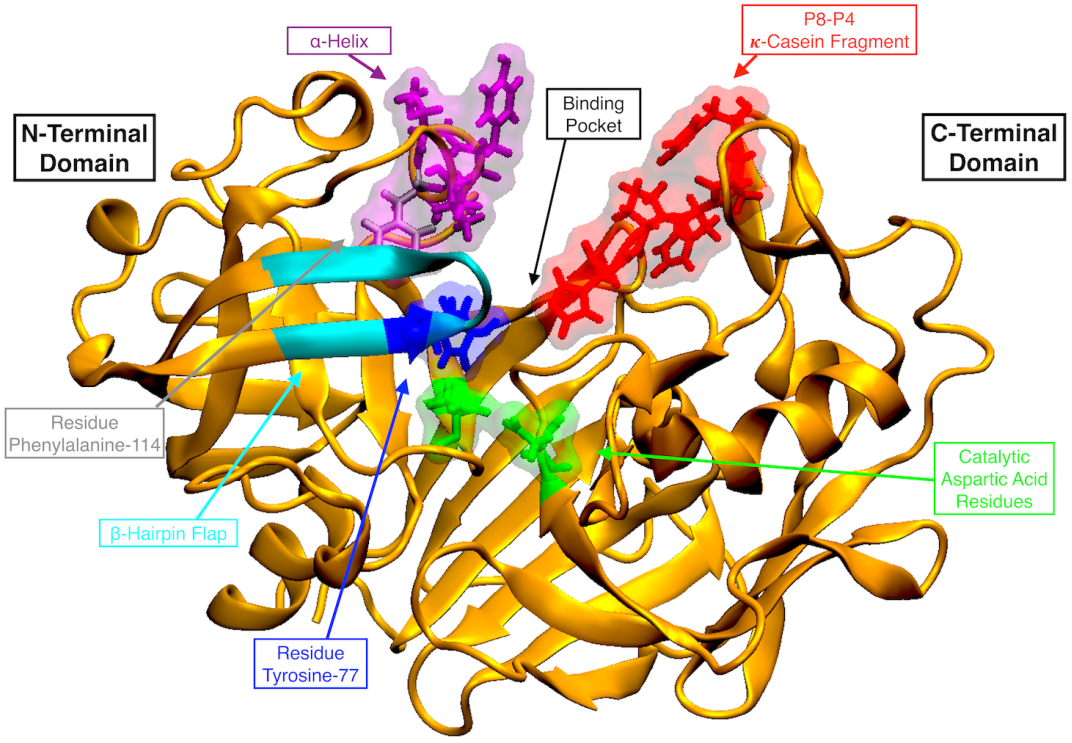


FIGURE 5.1: Depiction of bovine chymosin–κ-casein complex. 1) β-hairpin flap (light blue) and Tyr77 (dark blue). 2) α-helix (purple). 3) Catalytic residues (green). 4) His-Pro cluster (red).

The HPHPH cluster is conserved in many other mammalian κ-casein peptides including buffalo and goat, but in camel κ-casein the three histidine residues are mutated to arginines.[11,281] Although a potential allosteric-activation process has been widely discussed in the literature,[2,16,18,83,84,281] the mechanism has not been elucidated at a molecular level due to the challenges associated with studying it by experimental methods. This has hindered the development of novel enzymes

and enzymatic processes for the food industry.[281]

Here, two computational techniques are used, molecular dynamics (MD) and bias exchange metadynamics (BEMD) simulations, to reveal the allosteric activation mechanism and its associated free energy surface. (Note - the BEMD simulations were run by a collaborator, Dr Andrea Coletta at the University of Aarhus in Denmark, but the interpretation of the results, as well as all other simulations, were carried out by me.) BEMD is an enhanced sampling technique that allows the efficient exploration of complex free energy landscapes. It is well suited to studying conformational/configurational transformations in bio-macromolecules and has previously been used to study protein folding,[282,283] protein-ligand recognition[283] and allosteric transitions.[284,285]

5.2 Methods

5.2.1 Molecular Dynamics Simulations

Unrestrained MD simulations were performed for the six chemical systems described in Table 5.1, which include the open and self-inhibited conformations of apo-chymosin and four replicas of the self-inhibited chymosin - P8-P4-κ-casein complex. The four replicas differed only in whether or not capping groups were applied to the κ-casein fragment and whether the side chain of His102 in κ-casein was modelled as protonated or neutral (with a proton on the N-δ atom of the imidazole ring). Each system was simulated four times: two simulations using the AMBER ff03[286] force field and two simulations using the AMBER ff99SB-ILDN[287] force field (6 systems $\times$ 2 force fields $\times$ 2 duplicates = 24 simulations in total). Duplicate simulations were started from the same coordinates but with atoms being assigned different initial velocities in different simulations.

TABLE 5.1: Six chemical systems used as input for the molecular dynamics simulations

ID	System	Apo/Holo[a]	Capping[b]	Proton[c]
A1	Open	Apo	-	-
B1	Self-inhibited	Apo	-	-
C1	Self-inhibited	Holo	Capped	No (HID)
C2	Self-inhibited	Holo	No Cap	No (HID)
C3	Self-inhibited	Holo	Capped	Yes
C4	Self-inhibited	Holo	No Cap	Yes

[a] the holo complex is *P8-P4 κ-casein* bound to chymosin; [b] His-Pro fragment can be capped (by methyl group) at both ends or can be uncapped; [c] κ-casein fragment protonated at *P4His* position.

5.2.1.1 Input coordinates

The initial coordinates for the MD simulations were taken from previous work in the literature.[2,123,288,289] A brief summary of the steps used to prepare the input coordinates will be provided, since the details have previously been reported.[2,123,288,289] In summary, the coordinates of chymosin were taken from the crystal structure of 3CMS (where Tyr77 is resolved in both open and self-inhibited forms). Chymosin coordinates were modified to: insert missing residues (Asn291-His292-Ser293); reverse the Val111Phe mutation; introduce disulphide bonds between Cys47-Cys52, Cys207-211 and Cys250-Cys283; assign amino acid protonation states appropriately for pH 6.5; include 16 conserved water molecules identified by Prasad *et al.*[290]

Since there are no crystal structures of chymosin-κ-casein complexes, the coordinates of the chymosin sensitive regions of κ-casein (residues P9-P7') in the complex were generated by a two step process.[2] Firstly, the P2-P2' residues were docked into the 3CMS structure of apo chymosin and relaxed by MD simulations. The binding pose has been shown to be in the correct geometry to allow proteolysis of the P1-P1' (Phe105-Met106) amide bond in κ-casein via the established

reaction mechanisms.[2,291–295] Secondly, the remaining residues were grown one by one in the binding cleft using a series conformational search algorithms and MD simulations per residue. The resulting binding pose has previously been shown to be in good agreement with a crystal structure of a chymosin-inhibitor bound complex[2,86] and with previous computational studies of the same system.[15] Free energy calculations using this bound pose also agree with the results of experimental mutagenesis studies.[2,123,288].

In these previous models of the chymosin – P9-P7' κ-casein complex, the side chains of HisP8 and HisP4 in κ-casein were modelled as positively charged, while HisP4 was modelled as the neutral N-δ tautomer (since close contacts with the side chain of Lys221 disfavour the protonated form).[2] These protonation states were assigned based on predictions from PROPKA2.0 and comparisons of binding energies computed using Poisson-Boltzmann solvent models.[2,296]

In the simulations reported here, tests were conducted on the positively charged form of HisP4 as well as the N-δ tautomer because HisP4 is more solvent exposed in the complex of chymosin – P8-P4 κ-casein. However, as demonstrated later, the HisP4 protonation state was not deemed to affect the conclusions. The chymosin – P8-P4-κ-casein complex that was simulated here was obtained by deleting the P9 and P3-P7' residues and adding hydrogen atoms or capping groups to complete the valency, as necessary. The self-inhibited complexes were obtained by copying the P8-P4-κ-casein residues from the open complex into self-inhibited, apo chymosin (after alignment on the chymosin coordinates), followed by relaxation of the coordinates of the complex by constrained minimisation and MD simulations, as described below.

5.2.1.2 System Preparation

Molecular dynamics simulations were performed in NAMD.[297] Each protein or protein-ligand complex was solvated by TIP3P[298] water molecules using the

XLEAP module in AmberTools14.0.[299] Over 15000 water molecules were placed around the protein in a (rectangular cuboid) periodic box. All systems were neutralised then given an ionic strength of 0.07 mol dm^{-1} using chloride and sodium ions as required.

5.2.1.3 Simulations

The solvated complexes were relaxed by conjugate gradient energy minimisation in four steps of 5000 iterations. In steps 1 to 3, the whole protein, the protein backbone, and the α-carbon atoms, respectively, were held fixed. All constraints were removed in the fourth step. The systems were gradually heated to 300 K in the NVT ensemble over 10 ps with the α-carbons held fixed, followed by a 4 ns equilibration at 300 K with all constraints removed.

Equilibration and production simulations were performed in the isothermal-isobaric (NPT) ensemble[300] at 300 K and 1 atm. The pressure was regulated by the Nosé-Hoover Langevin piston pressure control[301] with the piston set up to a target of 1.01325 bar, a period of 200 fs, a decay of 100 fs and a temperature of 300 K.[302] The temperature of the system was maintained by means of Langevin dynamics with the dampening coefficient set to 2 ps^{-1}, but not affecting hydrogen atoms. Periodic boundary conditions were applied to the systems and electrostatic interactions were calculated by the particle mesh Ewald (PME) method.[203–205] A cut-off distance of 10 Å was set for van der Waals' interactions using a switching distance of 9 Å. The pair list was updated every 20 ns for atom pairs within 11 Å. The distances of all bonds between hydrogen atoms and hetero-atoms were constrained by the SHAKE algorithm.[206,207] The velocity Verlet algorithm was used to update the equations of motions every 2 fs, and snapshots were taken every 2 ps. For each system, a 2 ns equilibration was performed, followed by 80-100 ns of production dynamics.

5.2.1.4 Analysis

The open and self-inhibited forms of chymosin are distinguished by the N–C_α–C_β–C_γ (χ_{77}) dihedral angle in Tyr77 (Figure 5.2), which is approximately 300° in the open form and approximately 175° in the self-inhibited form (the dihedral angle is expressed on a scale from 0° to 360°, rather than the more common -180° to 180° scale, because it simplifies the resulting figures and free energy surface diagrams). In the open form, the side chain of Tyr77 tucks into a pocket under the β-hairpin flap formed by residues 74 to 82 of chymosin, while in the self-inhibited form it occludes the binding site. Measuring the Tyr77 dihedral as a function of simulation time is therefore a convenient method to identify transitions between open and self-inhibited forms.

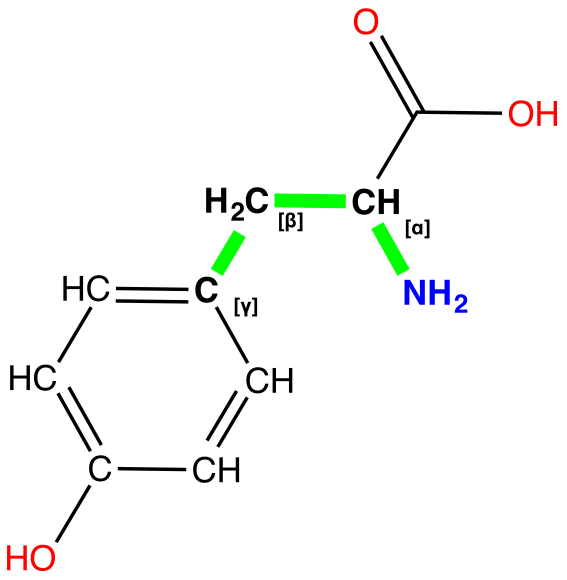


FIGURE 5.2: Tyr77 dihedral angle definition. Dihedral angle bonds in bold green.

To provide further insight into the observed allosteric activation mechanisms, the dihedral angle in residue Phe114 (C–C_α–C_β–C_γ), important residue-residue close-contacts, and hydrogen bonding networks stabilising Tyr77 were also measured as a function of simulation time. All measurements were automated using bespoke Tcl scripts in VMD.

5.2.2 Bias-Exchange Metadynamics (BEMD) Simulations

Well-tempered bias-exchange metadynamics (BEMD) simulations [303] of apo-chymosin (A1, B1 in Table 5.1) and chymosin in complex with P8-P4 κ-casein residues (C1 in Table 5.1) were performed in GROMACS-5.0.4 using the PLUMED-2.1 plug-in [304]. The free-energy surfaces (FES) were reconstructed using two collective variables (CV):

1. The dihedral angle χ of Tyr77 (χ_{77}, defined by the N–C_α–C_β–C_γ atoms).

2. The number of contacts between the side chains of Tyr77 and Phe114, measured using the PLUMED implemented CV coordination number (CN) :

$$\mathrm{CN} = \begin{cases} 1 & \text{if } r_{ij} \leq 0 \\ \sum_{ij} \frac{1-\left(\frac{r_{ij}}{r_0}\right)^6}{1-\left(\frac{r_{ij}}{r_0}\right)^{12}} & \text{if } r_{ij} > 0 \end{cases}$$

   where $r_{ij} = |r_i - r_j| - d_0$, ($r_i$ and r_j being the coordinates of Tyr77 and Phe114 atoms respectively). The values of the pair-wise switching function parameters d_0 and r_0 have been set to 4.0 and 3.0 Å, respectively.

In the case of apo-chymosin the BEMD simulations were performed using 4 replicas (one for each combination of the two structures A1 and B1 and the two CVs) while in the case of the chymosin - P8-P4 κ-casein complexes two replicas were used (one for each of the two CVs). In order to ensure the sampling of the free energy surface in the presence of the P8-P4 residues of κ-casein the inclusion of two piecewise linear/harmonic distance restraints were used: one between the terminal cap of P8 and N_δ of Asn241 and one between the terminal cap of P4 and O_γ of Ser220. The BEMD simulations were subsequently analysed using the VMD plug-in "METAGUI" [305].

In order to characterise the correlation between the Tyr77 dihedral angle and

the conformation of the protein, the mutual information entropy[306] value of χ_{77} dihedral was used along with the protein secondary structure, calculated as:

$$\mu_{res_i} = -\sum_{ss_i} \int \rho(\chi, ss_i) \log_2 \left(\frac{\rho(\chi, ss_i)}{\rho(\chi)\rho(ss_i)} \right) d\chi \tag{5.1}$$

where ss_i is the DSSP secondary structure[307] and the probability densities were estimated from the metadynamics simulation using 144 and 8 bins for χ_{77} and ss_i respectively.

This statistical measure is a generalization of the linear correlation coefficient and gives an estimate of the extra-information gained using the joint distribution function $\rho(\chi, ss_i)$ instead of the two single distributions $\rho(\chi)$ and $\rho(ss_i)$.

It can be shown that mutual information $\mu(a, b)$ between two random variables a and b can be expressed as $\mu(a, b) = (H(a) + H(b)) - H(a, b)$.

H is the information entropy:

$$H(x) = -\sum_i p(x_i) log_2(p(x_i)) \tag{5.2}$$

where $p(x_i)$ is the probability of event x_i. Information entropy is a measure of the information content in the variable x, and of the number of "bits" needed to efficiently encode a time series of that random variable. Mutual Information $\mu(a, b)$ can be interpreted as a measure of the reduced number of bits needed to encode the information content in the joint distribution (a, b) with respect to the

total amount needed to encode two single distributions for a and b separately.

Since it can be shown that $H(a,b) \leq H(a) + H(b)$ with the equality standing only in the case of a and b being independent, MI can be used as a generalised measure of correlation between a and b being not restricted to pure linear relationship between the two variables. The usual unit of measure for mutual information is the "bit".

5.3 Results and Discussion

5.3.1 Molecular Dynamics

5.3.1.1 Apo-Chymosin

None of the eight unrestrained MD simulations of *apo*-chymosin (>800 ns simulation time) exhibited a transition between the open and self-inhibited forms of the enzyme. Analysis of the Tyr77 dihedral angle reveals a clear distinction between simulations started from either the open or self-inhibited forms (Figure 5.3).

Since the transition between open and self-inhibited forms was not observed in these simulations, it suggests that there is a high-barrier for rotation around the Tyr77 dihedral angle (and the associated rearrangement of the β-hairpin flap) in the absence of the P8-P4 κ-casein penta-peptide, which agrees with the proposed allosteric activation method (further sampling of the dihedral angle is carried out using BEMD in the next section).

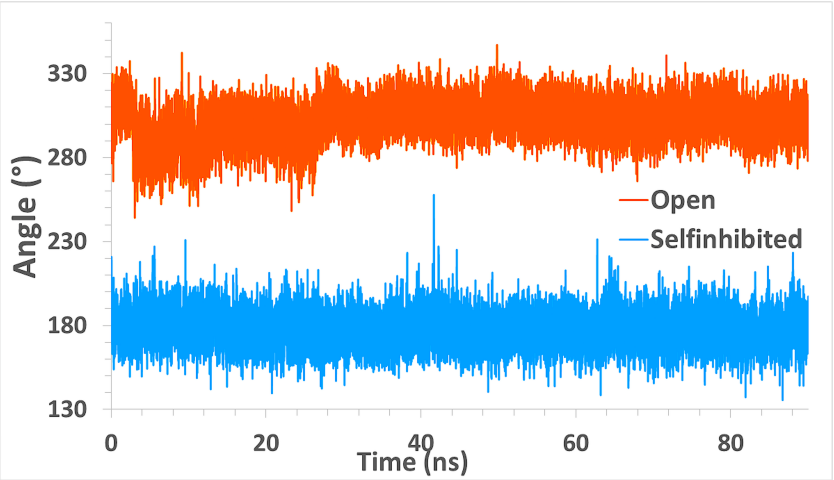


FIGURE 5.3: Tyr77 dihedral angle at open (orange, A1) and self-inhibited (blue, B1) conformations.

Tyr77 in its open conformation is found to be stabilised by a single water molecule. This stabilising water forms hydrogen bonds with Tyr77 and residues Ser37 and Asp39 of chymosin, as depicted in Figure 5.4. The water molecule has previously been shown to be conserved in crystal structures of aspartic proteases. [290]

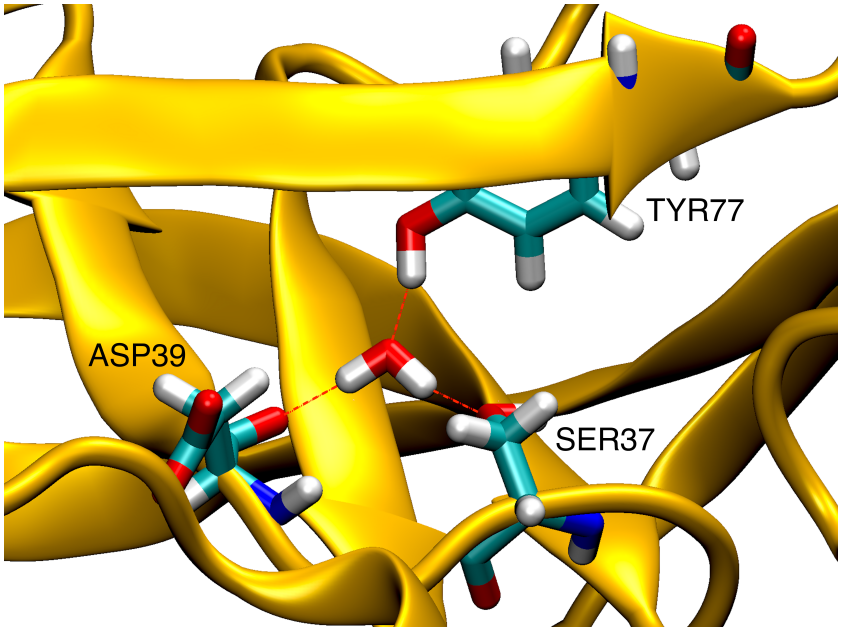


FIGURE 5.4: Open Tyr77 stabilisation

A different stabilising network is observed in the simulations in which the enzyme is in the self-inhibiting conformation. A single water molecule forms hydrogen

bonds with Tyr77, Ser14 and Gly218, both of which reside in the binding pocket Figure (5.5). The same hydrogen bonding pattern is also observed for short recurring periods in the three simulations in which a change in Tyr77 conformation takes place. In these simulations, the water molecule was regularly displaced in short succession prior to the conformational change occurring, which suggests that the κ-casein fragment affects this hydrogen bonding pattern.

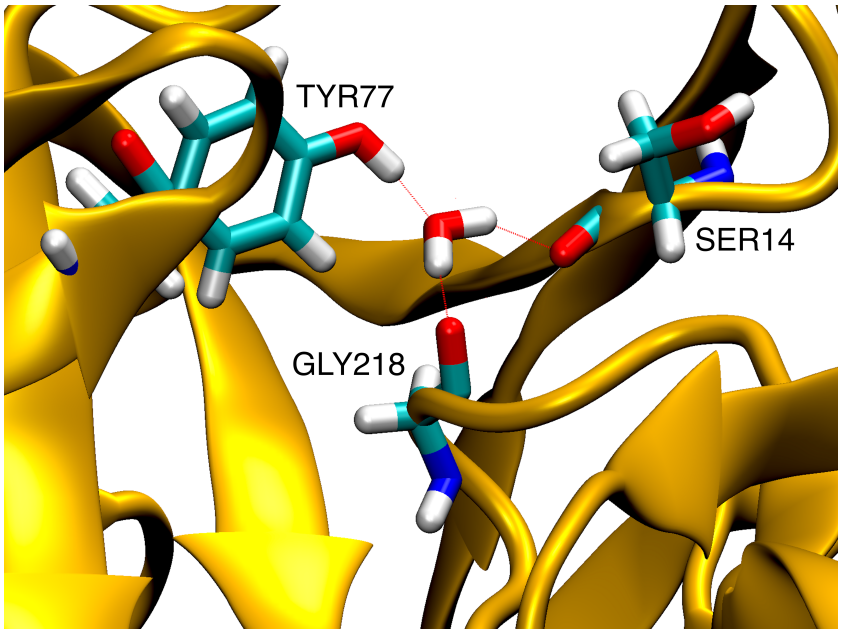


FIGURE 5.5: Self-inhibited Tyr77 stabilisation

5.3.1.2 Chymosin – P8-P4 κ-Casein

The allosteric transition from self-inhibited to open conformation in the presence of the κ-casein fragment is expected to occur on a sufficiently long timescale (seconds) making it difficult to completely sample using regular MD simulations (nanoseconds) on current computational hardware (this sampling problem is addressed later by the use of bias-exchange metadynamics simulations). Nevertheless, the allosteric transition was observed in three of the regular MD simulations that included the κ-casein fragment (The C1 simulation using the AMBER ff99SB-ILDN force field and simulations C2 and C3 using the AMBER ff03 force field).

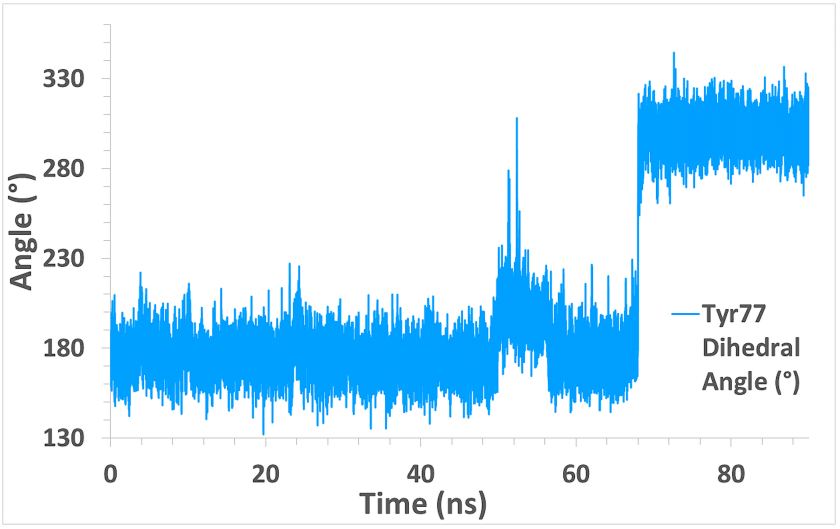


FIGURE 5.6: Tyr77 dihedral angle (blue) – AMBER-ff99SB-ILDN Self-inhibited-HPHPH complex (C1).

In the C1 simulation the dihedral angle of Tyr77 changes from self-inhibited to open at 70 ns in the 90 ns trajectory, shown in Figure 5.6. At 50 ns there is a deviation which lasts for 6 ns but this is not sustained and the dihedral angle of Tyr77 returns to self-inhibited until the change in form at 70 ns. The first conformational change at 50 ns is not sustained because residue Phe114 on the adjacent α-helix remains in the space that Tyr77 would occupy in its open conformation. The open conformation observed from 70 ns onwards is stabilised by the same hydrogen bonding pattern between Tyr77, water, Asp39 and Ser37 as observed in the simulation of apo-chymosin complex (and depicted in Figure 5.4).

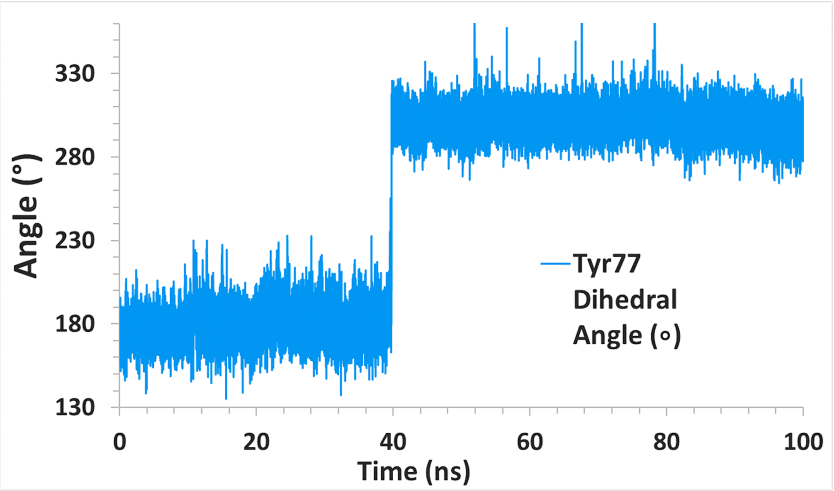


FIGURE 5.7: Tyr77 dihedral angle (blue) – AMBER-ff03 Self-inhibited-HPHPH complex with no capping group (C2).

The change in form in the C2 simulation occurs at 40 ns where Tyr77 moves from a self-inhibiting position to its open conformation Figure (5.7). This change is sustained for the remainder of the 100 ns trajectory. After the conformational change, Tyr77 is observed to make the same interactions that it does in the simulation of apo-chymosin in its open form (A1), including the hydrogen-bonding network between Tyr77, water, Asp39 and Ser37 that is depicted in Figure 5.4.

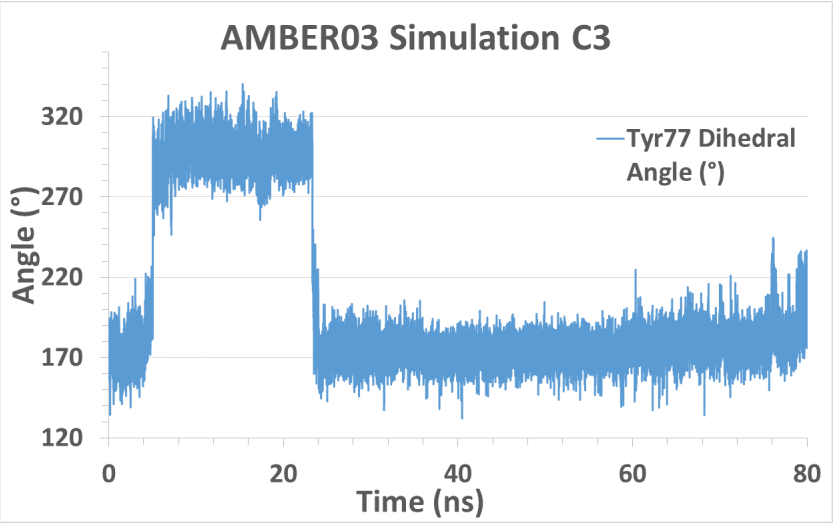


FIGURE 5.8: Tyr77 dihedral angle (blue) - AMBER-ff03 Self-inhibited-HPHPH protonated complex (C3).

The conformational change occurs early on in the 80 ns simulation of system C3 (see Figure 5.8). At 6 ns Tyr77 moves from a self-inhibiting position to open and remains in this state for 18 ns. In this pose the water molecule in the hydrogen bonding network that stabilises the open conformation is continuously displaced and replaced in short succession, which suggests that the stable open conformation has not been fully reached. There is an increase in contact between Phe114 and Trp41 (located at the base of the β-flap) when Tyr77 is in its open conformation (self-inhibited:35% $\rightarrow$ open:51%), which is the opposite of what is found in simulations C1 and C2, 37% $\rightarrow$ 17% and 66% $\rightarrow$ 32% respectively. This suggests steric interference inhibits the stabilisation of open Tyr77. The reverse transformation occurs at 24 ns where Tyr77 returns to its self-inhibited position suggesting that the open form was not stabilised.

Analysis of the MD simulations provides an initial indication of the mechanism by which the κ-casein fragment induces allosteric activation. The κ-casein fragment interacts with the α-helical region of chymosin causing a sequence of changes, all of which must occur to give allosteric activation. The key changes include: (i) disruption of the hydrogen-bonding network between Tyr77, water, Ser14 and Gly218 that would otherwise help to stabilise Tyr77 in the self-inhibited form; (ii) interaction of the P8-P4 residues of κ-casein with the short α-helix in residues 112-116 of chymosin, which causes movement of the side chain of Phe114 such that it vacates the pocket that is occupied by the side chain of Tyr77 in the open conformation, (iii) rearrangement of the β-hairpin flap to allow rotation of the Tyr77 dihedral from its self-inhibited to open conformation. The steps in this pathway were not observed in their entirety in any of the simulations that did not show allosteric activation.

5.3.2 BEMD – Bias-Exchange Metadynamics Simulations

5.3.2.1 Free Energy Surface

To further investigate the influence of the P8-P4 κ-casein residues on the Tyr77 conformation, two bias-exchange metadynamics simulations were performed, one for the apo-enzyme and one for the chymosin – P8-P4 κ-casein fragment complex. The free-energy surface (FES) as a function of χ_{77} and CN is reported in Figure 5.9-A1.

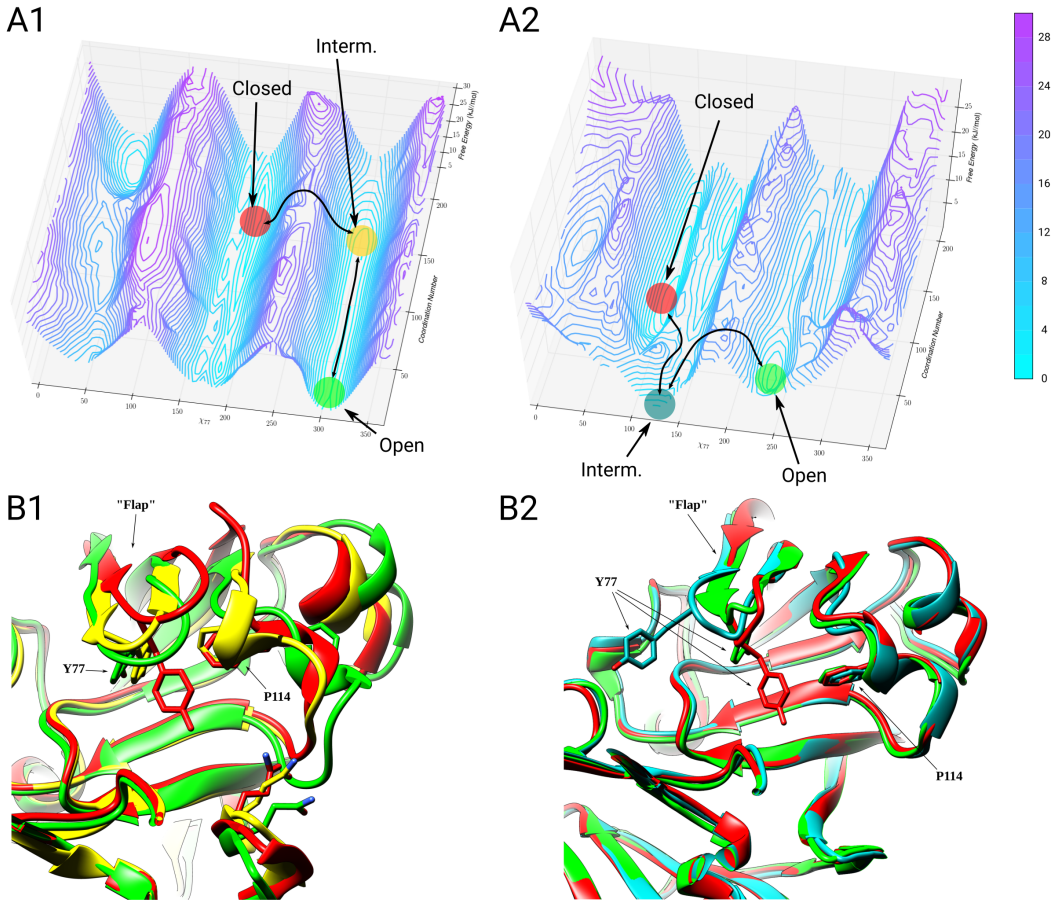


FIGURE 5.9: Top Panel: free-energy surface (FES) as a function of χ_{77} and CN obtained from the BEMD simulation of apo-chymosin (A1) and chymosin – P8-P4 κ-casein complex (A2). The open (green), self-inhibited (red) and intermediate (yellow) state of the enzyme are indicated with coloured spots. A simplified picture of the transition path from open to self-inhibited state is reported as a black line. Bottom Panel: Representative structure of the FES minima obtained from the BEMD simulation of apo-chymosin (B1) and chymosin – P8-P4 κ-casein complex (B2). The enzyme is represented as ribbon and the colouring scheme is the same used in the top panel.

The minima corresponding to the open and closed (self-inhibited) conformation of the apo-enzyme are highlighted with a green and a red spot respectively. The dihedral angles and the CN values of the stable states are reported in Table 5.2. The open state minimum is found to have a low number of contacts between Tyr77 and Phe114 and a dihedral angle of $305° \pm 5$ while the self-inhibited state minimum has a high number of contacts and a Tyr77 dihedral angle of $185° \pm 5$. The path connecting the open and the self-inhibited states is divided in two sub-steps joined by an intermediate state with a Tyr77 dihedral value of $305° \pm 5$ (equal to the open

state), and a high coordination number between Tyr77 and Phe114 (similarly to the self-inhibited state). These findings provide further evidence that the small helix spanning from residue 112 to residue 116 of chymosin plays an important role in the interconversion from the self-inhibited to the open state of the enzyme.

TABLE 5.2: χ_{77} dihedral angle and CN values of the stable states observed in the apo-chymosin BEMD simulation. The errors are the bin-widths used to calculate free energy in the VMD plugin METAGUI

State	χ_{77}	**CN**
Open	$305° \pm 5$	7 ± 5
Intermediate	$305° \pm 5$	150 ± 5
Self-inhibited	$185° \pm 5$	172 ± 5

A close-up of the structural change at the interface between the flap region and the small 112-116 helix in the three minimal states found in the apo-chymosin BEMD simulation is visible in Figure 5.9-B1 where a coordinate shift of the flap and the 2-turn helix is observed. In the first phase, the side chain of Tyr77 passes from the self-inhibited conformation (red) to the intermediate state (yellow) maintaining contact with Phe114, while the α-helix changes its conformation in concurrence with the β-flap region. The intermediate state observed here is considered unstable and is quickly transformed into the open conformation (green) in the trajectory. In the second phase the number of contacts between Phe114 and Tyr77 is reduced and the small helix returns to a conformation close to the original. This is confirmed through residue contacts analysis carried out in simulations A1 and B1 systems used in regular MD, see Appendices A. The findings are also in agreement with the regular MD simulations of holo-chymosin, where a simultaneous movement of Tyr77 and Phe114 is observed as Tyr77 moves from its self-inhibited to its open pose.

TABLE 5.3: χ_{77} dihedral angle and CN values of the stable states observed in the BEMD simulation of the chymosin – P8-P4 κ-casein complex. The errors are the bin-widths used to calculate free energy in the VMD plugin METAGUI

State	χ_{77}	**CN**
Open	$235° \pm 5$	7 ± 5
Intermediate	$125° \pm 5$	5 ± 5
Self-inhibited	$105° \pm 5$	105 ± 5

When the protein is in complex with the κ-casein fragment a dramatic change in the FES is observed (see Figure 5.9-A2 and Table 5.3). The minimum corresponding to the open state is shifted to a Tyr77 dihedral angle of $235° \pm 5$ and a low coordination number with Phe114 while the minimum corresponding to the closed state is characterised by a shifted Tyr77 dihedral angle of $105° \pm 5$ and a high coordination number. The systematic shift in the dihedral angles is possible because of a pronounced twisting of the β-hairpin flap in the BEMD simulations, which allows the side chain of Tyr77 to occupy the normal pockets in the open and self-inhibited conformations despite the change in angles. It is believed that the twisting of the β-hairpin flap is more pronounced in the BEMD simulations than the regular MD simulations because the former allows a more thorough sampling of the conformational change. Nonetheless, the difference in the Tyr77 dihedral angles remains $\sim$ 130 +/-5 degrees in the BEMD simulations (similar to that observed in the MD simulations and the 3CMS crystal structure). The importance of the β-hairpin flap in aspartic proteases has previously been highlighted in studies of mammalian (chymosin, BACE) and viral (HIV-protease) enzymes.[308,309] Interestingly, in the BEMD simulation of the complex, an intermediate state is found, but with a dihedral angle similar to the closed state ($105° \pm 5$) and a low coordination number giving a different picture to what is observed in the apo-enzyme FES. Since free energy estimates are less accurate for higher-energy regions of phase-space (which are less well sampled during simulations), some caution must be exercised in estimating barrier heights from the data in Figures 5.9-A1 and 5.9-A2. Nonetheless, in the apo-enzyme, the open and intermediate states are clearly separated by a very low energy barrier while the intermediate and closed state are

separated by a high energy barrier. By contrast, in the chymosin – P8-P4 κ-casein complex the closed and intermediate states are in the same kinetic basin while the intermediate and open states are divided by a high energy barrier.

From observations of the minimal structure states found for chymosin – P8-P4 κ-casein complex Figure (5.9-B2), it appears that in the intermediate state (cyan) the side chain of Tyr77 is pointing away from Phe114 and the active site of chymosin. This intermediate state is nominally an active conformation because the side chain of Tyr77 does not occlude the binding site. The conformation is observed only fleetingly in the regular MD simulations, however. The coordinated motions of the small helix and the β-flap regions found in the apo-enzyme transition are less obvious in the holo-enzyme transition; the small helix where Phe114 resides, conserves its structure in all three states, which is in good agreement with what is observed in the regular MD simulations.

5.3.2.2 Mutual Information

Mutual information is the measure of mutual dependence between two random variables. The regions in the enzyme having a high mutual information (MI) value (Figures 5.10-A and 5.10-C) are those in proximity to Tyr77, roughly from residue 70 to 80 (β-flap region), residues 110 to 120 (where the small helix and Phe114 are found) plus some individual residues (148,162 and 163), a small loop (residues 240-246) and a β-hairpin (residues 276-283) constituting the binding site of P8-P4 κ-casein fragment on the chymosin C-terminal domain.

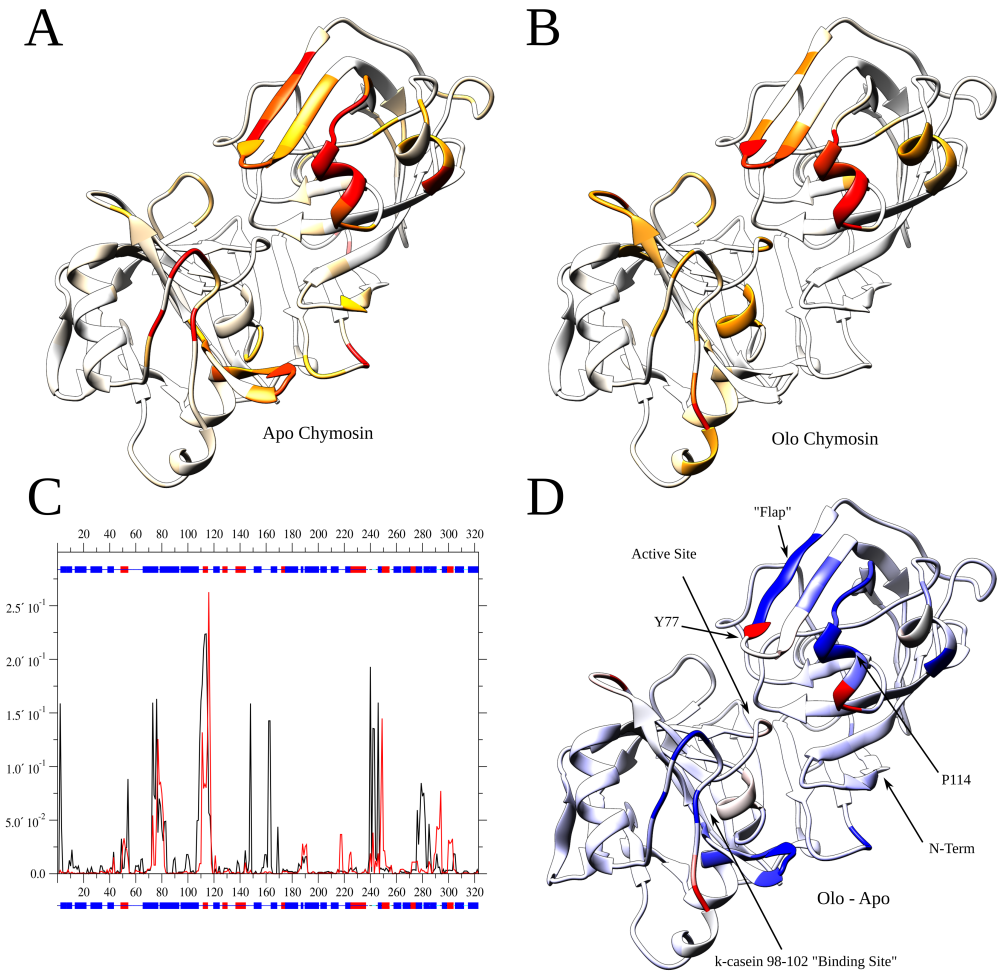


FIGURE 5.10: Mutual information (MI) entropy between χ_{77} and secondary structure of chymosin residues. Panel A: Residues on the apo-enzyme system with a MI greater than 0.25 or 0.5 of the maximum value are coloured in orange and red respectively. Panel B: MI of the holo-enzyme complex (same colouring scheme as Panel A). Panel C: Comparison of MI for apo- (black line) and holo-enzyme (red line), as a function of residue number; Secondary structure on the residues (as observed in the PDB 3CMS) is reported (red: helix, blue: beta). Panel D: Change of MI upon P8-P4 fragment binding. Residues for which a decrease of MI is observed are coloured in blue, while residues for which an increase of MI is observed are coloured in red.

A general reduction of the mutual information between Tyr77 rotation and the enzyme domain change is observed in the chymosin – P8-P4 κ-casein complex. The binding of the His-Pro cluster between the N-terminal β-hairpin and the C-terminal domain disrupts the communication network observed in the apo enzyme by modifying the conformation of the small α-helix, which, by a cascade effect, allows the β-flap to deform and explore different paths for the conversion from

the self-inhibited to the open state. The new path involves an intermediate state in the same kinetic basin as the self-inhibited state and, therefore, a higher rate of conversion between those two conformations. This intermediate state permits access to the κ-casein cleavage site within the chymosin binding.

5.4 Conclusions

The conformational change occurring in the allosteric activation of bovine chymosin has been observed by both regular MD and BEMD simulations. In agreement with previous proposals based on kinetic, mutagenesis and crystallographic experiments,[16–18,83,84] the simulations show that the HPHPH sequence from the P8-P4 residues of bovine κ-casein initiates a conformational change in the side chain of Tyr77 and the β-hairpin region of bovine chymosin. The allosteric activation mechanism occurs via the following steps: (i) the P8-P4 κ-casein fragment binds with chymosin and disrupts the hydrogen bonding network that stabilises the self-inhibiting pose of Tyr77 Figure 5.5; (ii) the P8-P4 κ-casein peptide interacts with the short α-helix in residues 112-116 of chymosin, which both allows the β-hairpin flap in residues 72 to 84 of chymosin to twist, and also causes the side chain of Phe114 to vacate the pocket that is occupied by Tyr77 in the open conformation; (iii) as Phe114 moves, Tyr77 simultaneously changes conformation from self-inhibiting to open and is stabilised by a hydrogen bonding network below the β-hairpin flap (Figure 5.4). Subtle variations in the simulation trajectories suggest that allosteric activation is possible by multiple related pathways, but these all go through the general steps described above, which were observed in their entirety in all of the relevant MD and BEMD simulations.

Chapter 6

Effect of Mutations in Bovine or Camel Chymosin on the Thermodynamics of Binding κ-Caseins

6.1 Overview

Bovine and camel chymosin have high sequence similarity (94%) and identity (85%) and similar three-dimensional structures. They both comprise 323 amino acids that fold into a pseudo-symmetric bi-lobal structure forming a central binding cleft containing the catalytic residues Asp34 and Asp216. The side chains of the catalytic aspartic acid residues extend towards each other in a planar geometry,[95] which is stabilised by a network of hydrogen bonds with two threonine residues, referred to as "*the fireman's grip*".[45] Within the substrate-binding cleft, there are 12 residue differences in the primary structure of bovine and camel chymosin.

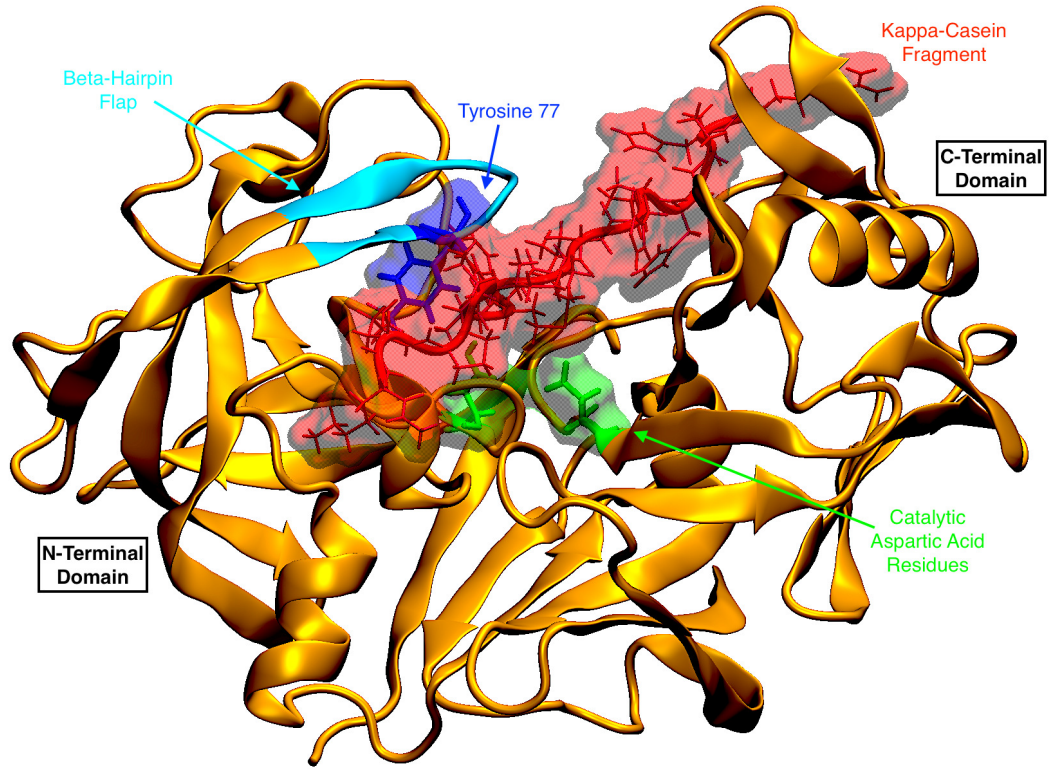


FIGURE 6.1: Depiction of bovine chymosin–bovine κ-casein complex. κ-casein fragment in red aligned across the binding cleft of chymosin. Catalytic aspartic acid residues in green located within the binding cleft.

Here, the importance of individual amino acid residues in chymosin is investigated by calculating the influence each residue has on the binding free energy of chymosin to its substrate, κ-casein. This study focuses on the 12 residues that are naturally different in the active sites of bovine and camel chymosin in order to elucidate the influence of each residue on chymosin-κ-casein binding. For each of these residues, computational alanine scanning calculations are performed in all four chymosin-κ-casein complexes (Bov/Bov, Bov/Cam, Cam/Bov, Cam/Cam, chymosin type first and κ-casein type second, Bov/Bov shown in Figure 6.1), using the molecular mechanics three-dimensional reference interaction site model (MM-3DRISM) methodology. Using MM-3DRISM permits accurate estimates of solvation and binding free energies and allows for the investigation of solvent density effects that could not be studied by implicit continuum solvent models (as in e.g. MM-PBSA).

6.2 Methods

6.2.1 MD Simulations

A total of 100 ns of unrestrained molecular dynamics simulations were run for each of the four chymosin–P9-P7'- κ-casein complexes (Bov/Bov, Bov/Cam, Cam/Bov, Cam/Cam). Input coordinates for each complex were taken from previous work in the literature.[95,289]

The molecular dynamic simulations were run with the AMBER-ff03 force field parameters developed by Duan *et al.* using NAMD[297] and the TIP3P water model.[298] Production simulations were run in the isothermal-isobaric (NPT) ensemble[300] at 1 atm. Langevin dynamics maintained the system temperature at 300K and the pressure was regulated by the Nosé-Hoover Langevin piston pressure control[301], the piston was set to a target of 1.01325 bar, period set to 200fs, and decay set to 100 fs.[302] Periodic boundary conditions were applied to each system and the electrostatic interactions were calculated using the particle mesh Ewald (PME) method.[203–205] Van der Waals interactions had a cut-off distance of 10 Å and a switching distance of 9 Å. All hydrogen to hetero-atom bond distances were constrained by the SHAKE algorithm.[206,207] The velocity Verlet algorithm was set to update the equations of motion every 2 fs, and snapshots were stored every 2 ps.

Each simulation system was equilibrated for 4 ns. A 96 ns simulation was generated with a snapshot every 400 ps providing a 240 frame trajectory for analysis. To reduce unnecessary computational expense, MM-3DRISM and normal-mode entropy calculations were carried out on every third frame of this trajectory, as per previous MM-3DRISM studies in the literature.[123,239]

6.2.2 MM-3DRISM Calulations

Binding free energy of the κ-casein fragment to chymosin was calculated using the MM-3DRISM[239,310] method at 298.15 K as implemented in AmberTools15,[172,225] using a locally modified version of the MMPBSA.py program which implemented the PC and PC+ free energy functionals.[246] The calculations were carried out on single trajectories of each complex, this has proven to be both computationally more efficient and provides results closer to experimental values through cancellation of errors.[256] All interactions were computed with the AMBER-ff03 force field. 3D-RISM calculations were performed with the assumption of an infinitely dilute solute. Solvent was modelled using a modified SPC water model (as implemented in the AmberTools package) with a water density of 55.343 mol/l. The modified SPC water model was used to avoid numerical convergence issues.[217] The buffer parameter was set to give a minimum distance of 18 Å between the solute and the edge of the solvent box. The calculations employed the MDIIS iterative scheme,[311] where 5 MDIIS vectors were used, and a MDIIS step size of 0.7. Solvent susceptibility functions required as input to the 3D-RISM calculations were calculated with the dielectrically consistent 1D-RISM. The grid spacing for 1D functions was 0.025 Å, which gave a total of 16,384 grid points. The MDIIS iterative scheme was employed, using 20 MDIIS vectors, an MDIIS step size of 0.3, and a residual tolerance of 10^{-12}. The solvent was considered to be pure water with a number density 0.0333 Å^3 and a dielectric constant of 78.497. Salt water was also considered at various concentrations in preliminary test but due to the lengthy calculation time and results being similar to the pure water results, the salt water calculations were not conducted for this study.

Entropic contributions were calculated from rotational, translational and vibrational contributions, with the latter computed by normal mode analysis.[312] The binding free energy of a single complex is calculated through the average binding free energy from a set of different conformations of the protein-ligand complex. ΔG_{Hyd} has been calculated using the advanced pressure correction (PC+) free energy functional. The PC+ functional contains no empirical parameters and has

been shown to give accurate predictions of hydration free energies for neutral and ionised solutes, in both pure water and salt solutions at a wide-range of temperatures.[237,241,245,248,249] As a comparison, results from three other free energy functionals are presented: partial series expansion-3 (PSE-3), pressure correction (PC) and Gaussian fluctuations (GF). For the calculation of relative solvation free energies, as is required in computational alanine scanning, the investigation shows that similar results are obtained using the PSE-3, PC and PC+ functionals because the differences between these functionals partially cancel out.

6.2.3 Computational Alanine Scanning

Computational alanine scanning calculations were carried out for 12 residues in the binding site that are natural mutants (different amino acids) in bovine and camel chymosin. All of these residues were within 4 Å of a residue in κ-casein for at least 70% of each of the molecular dynamics trajectories (measured using bespoke VMD Tcl scripts)(Figure 6.2).

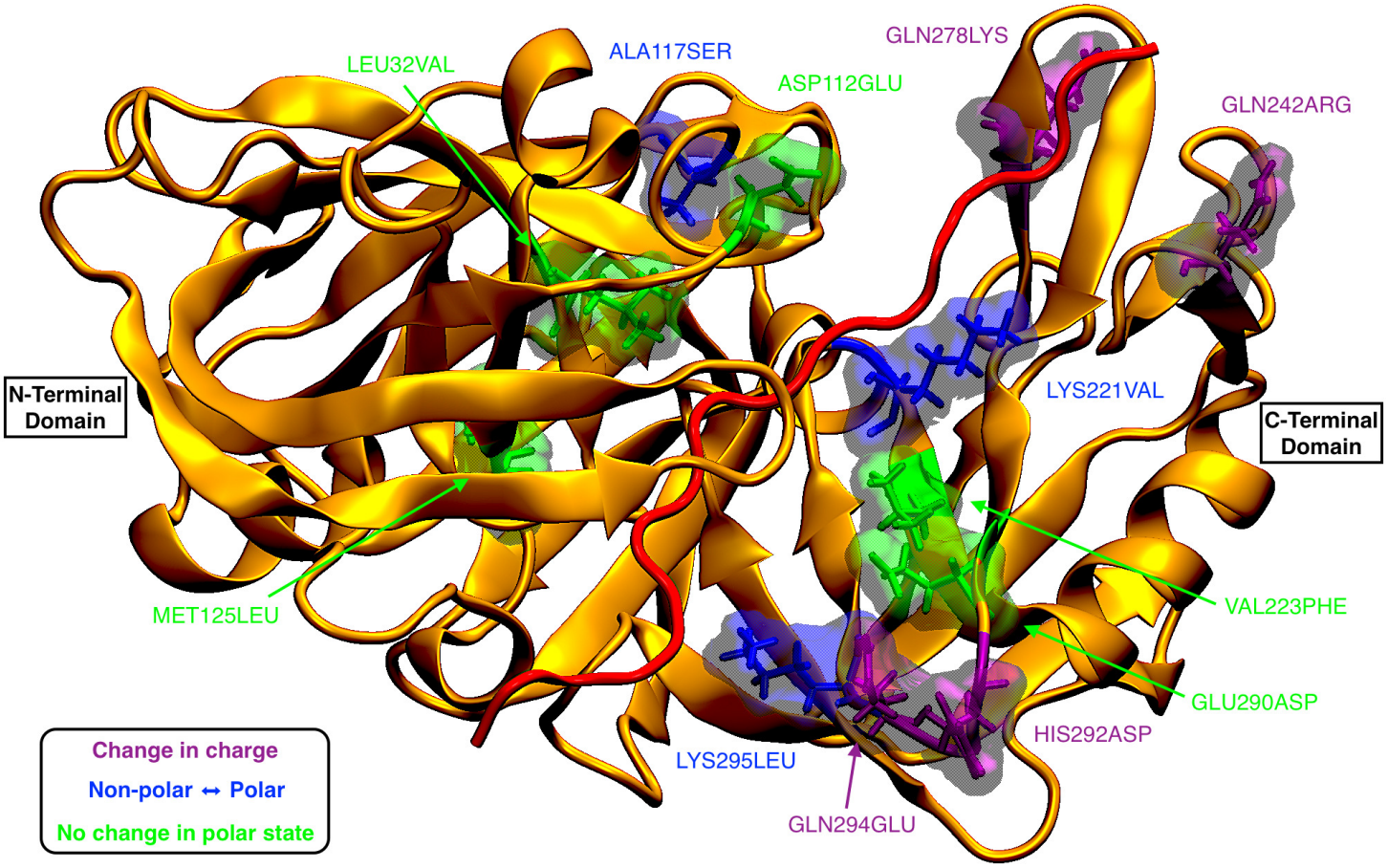


FIGURE 6.2: Depiction of residues that are naturally different between bovine and camel chymosin (*natural mutants*), shown on a bovine chymosin–bovine κ-casein complex. The labels refer to the bovine residue first and the camel counterpart second. Purple residues show that the amino acid has a different polar charge in each version of chymosin. Blue residues show the amino acid has gone from a polar to a non-polar one, or *vice-versa*. Green residues show that the polarity (and charge) remains the same in camel and bovine chymosin even though the amino acids are different.

The Massova and Kollman[313] protocol was used to carry out the alanine scanning calculations, employing the same trajectories used in the binding free energy calculations. The Massova and Kollman method assumes that the mutations do not significantly change the dynamics of the enzyme-substrate system, thereby introducing the mutation after the simulations have been performed should provide accurate results. This assumption has been shown to be valid for a wide-range of protein-ligand systems.[314] It is also supported by molecular dynamics simulations of bovine chymosin complexes (unbound, inhibitor bound and substrate bound) which show binding incurs no significant change in the conformation of the protein backbone.[83,86,95] The protocol has a number of advantages including being computationally much less demanding, and most importantly the use of the same trajectories allows for the cancellation of errors, resulting in more accurate results.[313,315] In accordance with previous studies[313,315,316] and limited by computational expense, the entropy term was neglected for all alanine scanning calculations, since using the Massova and Kollman protocol the difference in entropy between mutant and wildtype proteins is expected to be negligible.

The difference that an alanine mutation makes on binding free energy was calculated by subtracting the wild-type binding free energy from the mutant to give $\Delta\Delta G_{bind}$, ($\Delta\Delta G_{bind} = \Delta G_{mutant} - \Delta G_{wildtype}$). A negative $\Delta\Delta G_{bind}$ indicates a favourable mutation (the native system has higher binding free energy than the mutant), and positive results are unfavourable mutations (the native system has a lower binding free energy compared to the mutant). As reported in previous work[3,123] residues can be classified on a basis of magnitude of $\Delta\Delta G_{bind}$; warm ($\geq$ 1 kcal/mol) or hot-spots ($\geq$ 2 kcal/mol), representing a disproportionate contribution to the binding free energy.

6.3 Results and Discussion

6.3.1 Binding Free Energies

The binding free energies of the wildtype chymosin–κ-casein complexes were calculated using four different solvation free energy functionals in MM-3DRISM. Extensive previous benchmarking on solvation free energy data of organic molecules indicates that the PC+ functional gives more accurate results than the GF, PSE-3 or PC functionals.[237,241,245,248,249]

In this section, which discusses binding free energies, the focus is on the results obtained using the PC+ functional, while the results of the other functionals are provided in Appendix B. The calculated binding free energies must be interpreted with caution because they do not include some terms relating to the loss of conformational freedom on binding (due to the use of a single-trajectory approach to the MM-3DRISM calculations) and because they only include harmonic contributions to the vibrational entropy; neither of these problems unduly affect the computational alanine scanning results because of favourable cancellation of errors.

Table 6.1 presents enthalpic and entropic components of the free energy calculated by MM-3DRISM(PC+). The results indicate that binding is thermodynamically favourable for all four complexes, but that the native complexes exhibit more favourable binding than the cross-complexes. For all four complexes, a large favourable change in the gas phase contribution to the binding free energy (ΔG_{gas}) is opposed by an unfavourable change in the hydration free energy (ΔG_{hyd}) of similar magnitude. The entropic contributions to the binding free energy are of similar magnitude for all complexes. The binding free energies calculated by MM-3DRISM(PC+) for the Bov/Bov and Cam/Bov systems are observed to be in good general agreement with those obtained from MM-PBSA by Sorensen *et al.* (Table 6.1). The binding free energies obtained by these two methods differ by $\approx$ 2 kcal/mol in both cases, which is not negligible, but is surprisingly consistent

TABLE 6.1: Different Components of Binding Free Energy Calculated for the Various Chymosin–κ-Casein Complexed using MM-3DRISM(PC+) Methodology. MM-PBSA results taken from reports by Sørensen *et al.*[3] All Values Given are in units of kcal/mol.

	CAM/CAM		**CAM/BOV**		**BOV/CAM**		**BOV/BOV**	
Energy	**Mean**	**SE**[a]	**Mean**	**SE**	**Mean**	**SE**	**Mean**	**SE**
ΔG_{gas}[b]	-1330.1	*9.8*	-1029.6	*6.6*	-1595.4	*7.2*	-1388.0	*9.9*
ΔG_{hyd}[c]	1211.7	*9.1*	932.4	*6.1*	1494.7	*6.8*	1277.3	*9.1*
ΔG_{total}[d]	-118.4	*1.4*	-97.2	*1.2*	-100.7	*1.1*	-110.8	*1.4*
$T\Delta S$[e]	-76.5	*1.5*	-61.9	*1.3*	-68.1	*1.4*	-70.2	*1.1*
ΔG_{bind}[f]	-41.9	*2.1*	-35.3	*1.8*	-32.6	*1.8*	-40.6	*1.8*
$\Delta G_{bind}^{MM-PBSA}$ (Ref.[3])[g]	-	-	-33.4	*0.8*	-	-	-42.8	*0.7*

[a] Standard Error; [b] Total Gas Phase Free Energy; [c] Total Hydration Free Energy; [d] Total Energy; [e] Total Entropy; [f] Total Binding Free Energy. [g] Results taken from reports by Sørensen *et al.*[3]

given the size of the peptide substrates and the difficulties associated with predicting absolute binding free energies from molecular simulation. Unfortunately, it is impossible to compare either of these sets of results to experiment, since neither the binding free energies nor the hydration free energies of the protein-ligand complexes considered here have been reported.

6.3.2 Alanine Scanning

To determine the importance of individual residues for free energy of binding, alanine scanning calculations have been performed in all four complexes (Bov/Bov, Bov/Cam, Cam/Bov and Cam/Cam), from 96 ns MD simulations of each complex. The alanine scanning results that reveal a significant difference for a given amino acid position in the four complexes will be grouped into two classes, corresponding to whether it is the native residue in bovine or in camel chymosin that contributes more favourably to the binding free energy. A negative $\Delta\Delta G$ shows that mutating the natural residue to alanine will result in a stronger binding, and a positive $\Delta\Delta G$ means the alanine mutation will result in a weaker binding.

6.3.2.1 Favoured Native Camel Residues

LYS221VAL Residue 221 lies in the S4 pocket in close contact with either HisP4 (bovine) or ArgP4 (camel) of κ-casein.[3] On the basis of binding free energy calculations of the wildtype bovine complex only, it has previously been suggested that Val221 should be more favoured than Lys221 for the binding of bovine κ-casein.[3] The results presented in Figure 6.3 are in good agreement with that prediction. Figure 6.3 shows that the Lys221Ala mutation in bovine chymosin is favoured (because it reduces unfavourable polar interactions with HisP4 in bovine κ-casein), whereas by contrast the Val221Ala mutation in camel chymosin is unfavoured (because it reduces favourable non-polar interactions). For the camel κ-casein substrate (6.4), the same trend is observed, but the effects are greater because ArgP4 (camel) is larger and more basic than HisP4 (bovine). The unfavourable interaction between ArgP4 and Lys221 correlates with the observation that camel κ-casein is a poor substrate for bovine chymosin. This is supported by the experimental observation that a LysP4 mutant of bovine κ-casein is also a poor substrate for bovine chymosin.[18]

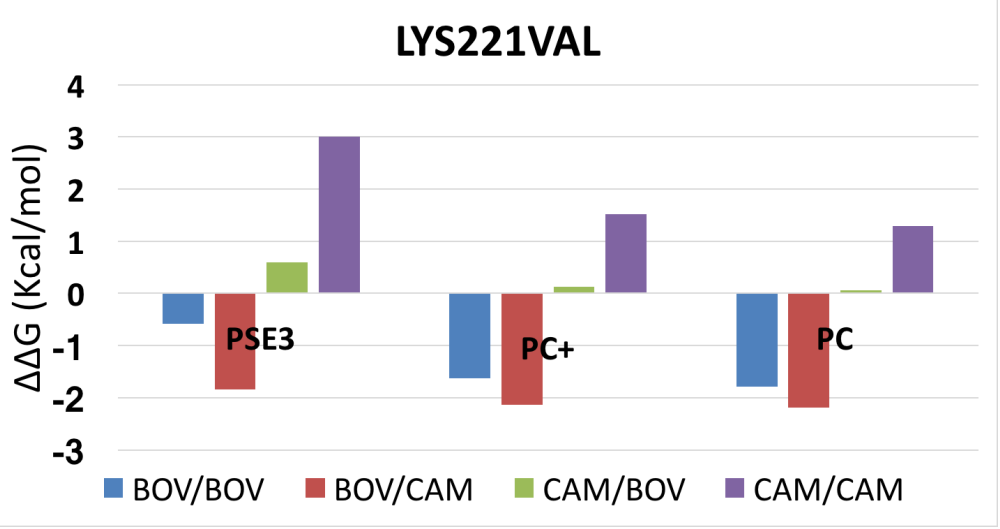


FIGURE 6.3: Comparison of alanine scanning results of residue 221 (Lys221 in bovine chymosin, Val221 in camel chymosin), on the four different chymosin–κ-casein complexes with three different MM-3DRISM calculation methods. A negative $\Delta\Delta G$ represents a favourable mutation, and positive results are unfavourable.

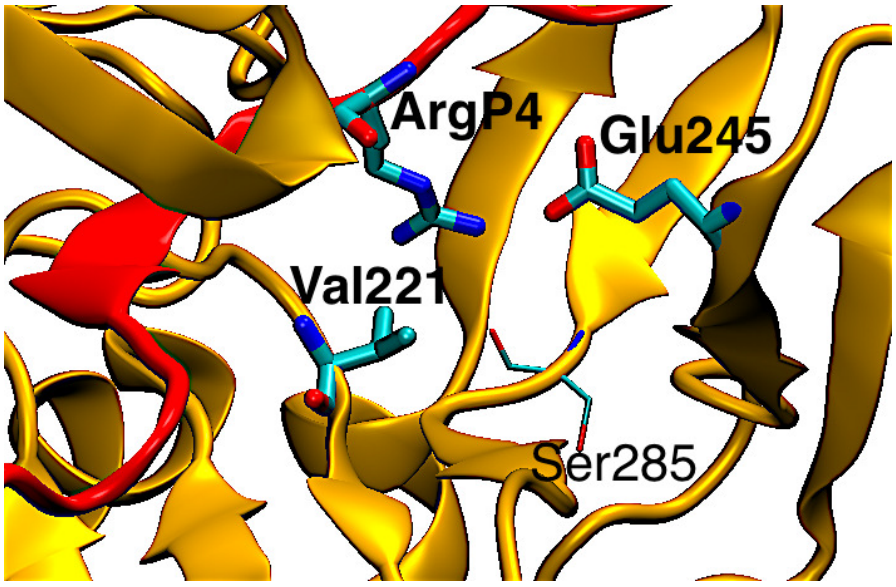


FIGURE 6.4: Snapshot of Val221 and ArgP4 (in bold for clarity) at 75.2 ns of Cam/Cam complex MD simulation. Residue ArgP4 interacts with both Val221 and Glu245.

ASP112GLU Residue 112 (Asp112 in bovine chymosin, Glu112 in camel chymosin) lies at one end of a short α-helical region of chymosin (residues 112 to 116) near the surface of the binding cleft. The importance of the Asp112Glu mutation to chymosin κ-casein binding thermodynamics is not immediately obvious from crystallographic data since the mutation lies in the N-terminal domain of chymosin, whereas the neighbouring P9-P1 residues of κ-casein bind predominantly to the C-terminal domain. In all four complexes, mutating residue 112 to alanine is shown to be thermodynamically unfavourable (Figure 6.5). Furthermore, a clear difference is observed in the values of $\Delta\Delta G_{bind}$ for alanine scanning in the four complexes. Both the importance of residue 112 and the trend in the alanine scanning results can be partly explained by a salt bridge between residue 112 of chymosin and P8 of κ-casein, which is observed to form for some part of each of the simulations. For example, in the Cam/Cam complex, a relatively stable Glu112-ArgP9 salt bridge is observed throughout the majority of the simulation (Figure 6.6). Consequently, mutating Glu112 to alanine results in an unfavourable change in binding free energy because of the loss of the salt bridge. By contrast, in the Bov/Cam system, in which Asp112 in chymosin interacts with ArgP8 in κ-casein, the salt bridge is observed less frequently during the molecular dynamics

simulation and the loss of binding free energy due to an Asp112Ala mutation is lower. This trend agrees with recent experimental and computational research that suggests that in solution Arg forms weaker salt bridges with Asp than Glu.[317] However, in the context of chymosin–κ-casein complexes, it may also be partly due to the fact that the side chain of Glu is longer than Asp and, hence, it can orientate itself better with respect to ArgP8. The alanine scanning data for bovine or camel chymosin binding to bovine κ-casein reveals a similar trend with the stronger salt bridge in the Cam/Bov complex (Glu112-HisP8) giving rise to a slightly larger value of $\Delta\Delta G_{bind}$ than that in the Bov/Bov complex (Asp112-HisP8, HisP8 was modelled as the charged histidinium ion in agreement with previous work).[2]

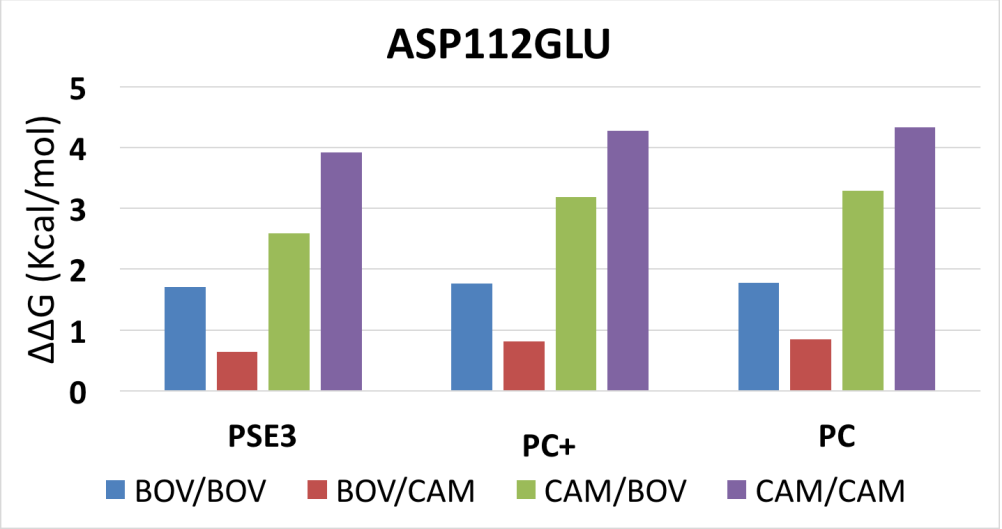


FIGURE 6.5: Comparison of alanine scanning results of residue 112 (Asp112 in bovine chymosin, Glu112 in camel chymosin).

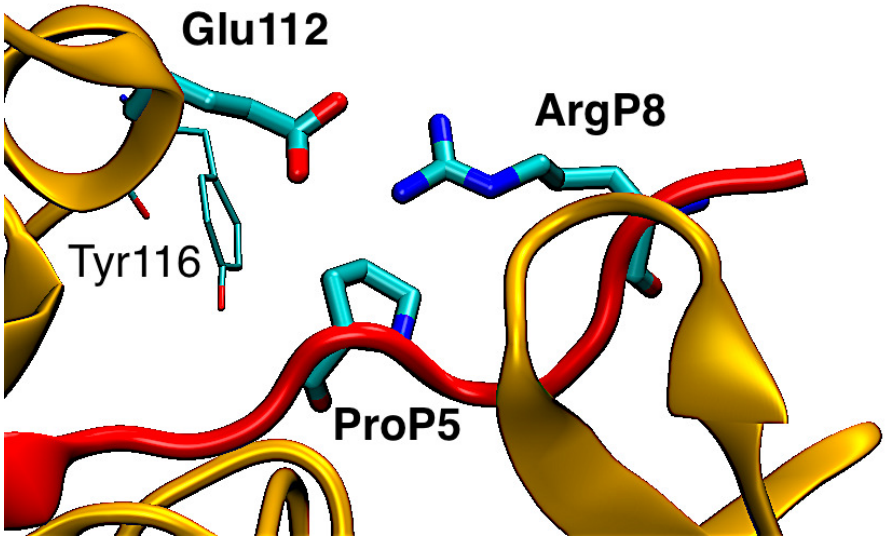


FIGURE 6.6: Snapshot of Glu112 and ArgP8 (in bold for clarity) at 18.8 ns of Cam/Cam complex MD simulation.

The α-helical region containing Asp112Glu has been implicated in allosteric activation of bovine chymosin by the P8-P4 residues of bovine κ-casein.[318] In this mechanism, the *His-Pro cluster* (HPHPH in P8-P4 of bovine κ-casein) interacts with the α-helix, which both allows the β-hairpin flap in residues 72-84 of chymosin to twist and causes the side chain of Phe114 to vacate a pocket that is occupied by Tyr77 in the open conformation. The interaction between Asp112 and the P8 residue of κ-casein is therefore a potential target for protein engineering aimed at modifying the self-inhibited to open transition of Tyr77 in the bovine complex. However, further experimental research would be required to verify how these processes occur in the complexes involving camel chymosin or camel κ-casein.

GLU290ASP, HIS292ASN, GLN294GLU and LYS295LEU Residues 290 to 295 form an unstructured loop region above the centre of the binding cleft in the C-terminal domain, opposite the β-hairpin flap in the N-terminal domain. The loop region is known to be more flexible than the surrounding residues as indicated by the crystallographic B-factors of the backbone atoms, which are ~40 A^2 in the loop compared to an average of ~21-22 A^2 in the protein.[97] Indeed,

in earlier crystallographic studies, residues 291 to 293 were considered to be too flexible to be resolved accurately.[83] The primary sequence of residues 290 to 295 is ENHSQK in bovine chymosin and DNNSEL in camel chymosin, which reveals four mutations: Glu290Asp, His292Asn, Gln294Glu and Lys295Leu. Analysis of the molecular dynamics trajectories shows that the side chains of residues Glu290Asp, Gln294Glu and Lys295Leu point towards κ-casein in all four complexes, while Ser293 and, to a lesser extent, Asn291 are solvent exposed. His292Asn lies in the most flexible region at the tip of the loop (B-factor $>$ 40 A^2 in both bovine and camel crystal structures). An ensemble of different conformations are observed throughout the molecular dynamics simulations, but on average Asn292 in camel chymosin is more solvent exposed than His292 in bovine chymosin regardless of the identity of the substrate.

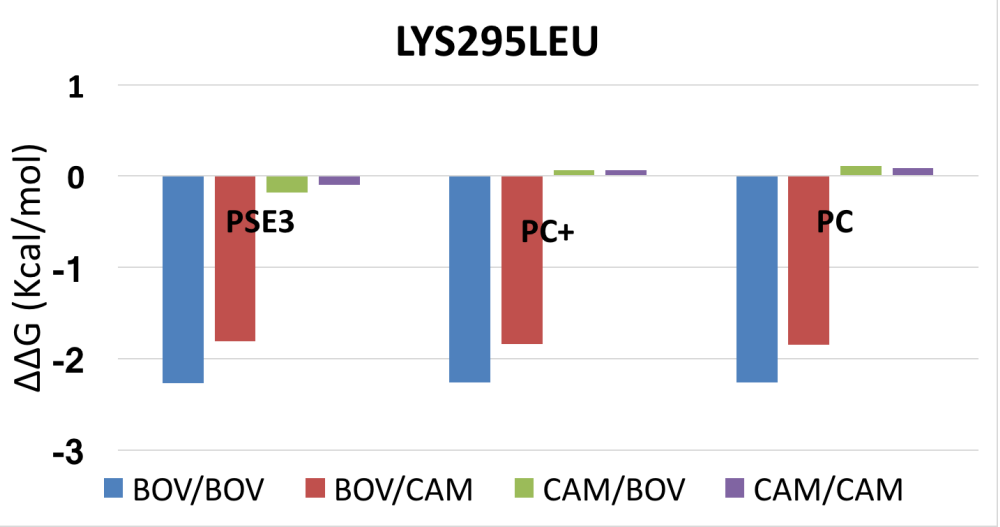


FIGURE 6.7: Comparison of alanine scanning results of residue 295 (Lys295 in bovine chymosin, Leu295 in camel chymosin).

In all four complexes, residue 290 (Glu290 in bovine chymosin, Asp290 in camel chymosin) forms intermittent hydrogen bonds with the side chain of SerP2 and non-specific interactions with IleP3'; both of these residues are conserved in bovine and camel κ-casein (as well as goat, horse, pig and sheep κ-casein. Figure 2.3). The alanine scanning calculations show a weak preference for Asp in the 290 position, but mutating either Glu290 or Asp290 to alanine is unfavourable, since it incurs

the loss of a hydrogen bond to SerP2 (Figure B.6, Appendix B). A more selective influence on the binding free energy is observed in positions 292, 294 and 295. Mutating His292 to alanine strengthens the binding by $\sim$ 2 kcal/mol in the bovine chymosin complexes, whilst mutating Asn292 to alanine in the camel chymosin complexes has essentially no effect because the residue is largely solvent exposed (Figure B.7, Appendix B). Similarly, mutating Lys295 to alanine strengthens the binding by $\sim$ 2 kcal/mol in both bovine chymosin complexes (Figure 6.7), whilst mutating Leu295 to alanine in the camel chymosin complexes has little effect because both Leu295 and Ala295 make similar weak van der Waals interactions with κ-casein (IleP3' and ProP5'). (Figure 6.8). In the 294 position, there is a weak preference for the Glu294 residue in camel chymosin, but the Gln294 residue in bovine chymosin contributes approximately the same amount to the binding free energy as an Ala residue (Figure B.8, Appendix B). The 294 residue points towards the side chains of the P1 and P3 residues in κ-casein and is partially solvated in all four complexes.

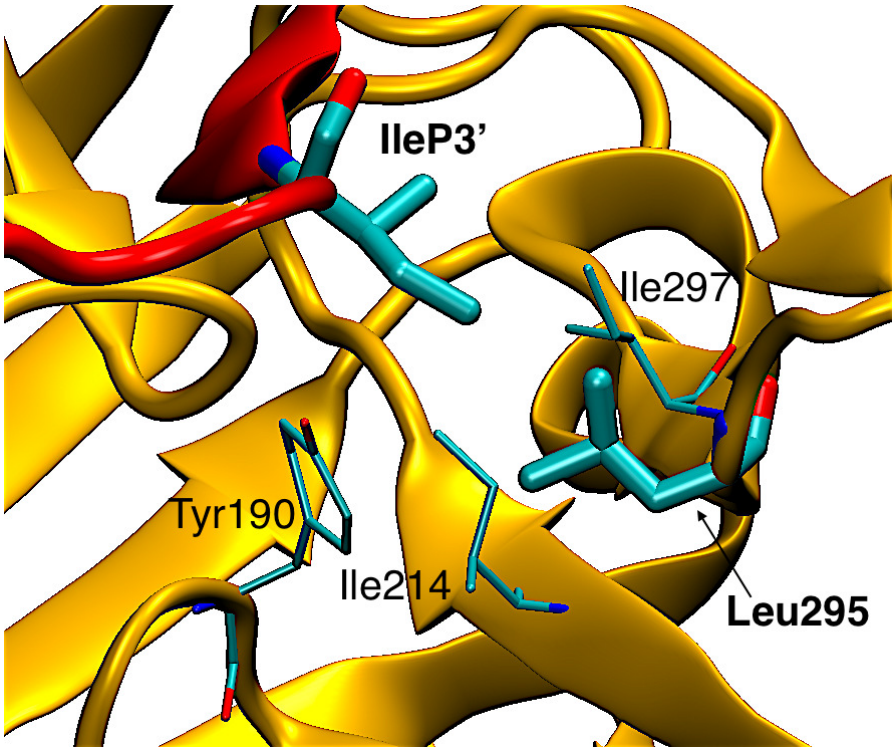


FIGURE 6.8: Snapshot of Leu295 and IleP3' at 17.2 ns of Cam/Cam complex MD simulation.

An additional consideration is that Asn291 is a known glycosylation site in both bovine and camel chymosin.[97] Here, considerations were made for the unglycosylated variants since these predominate in commercial products. However, *A. Niger* fermentation can glycosylate proteins at $N^{\delta 2}$ atoms of Asn in Asn-X-Thr/Ser sequences, of which there are two in both bovine (Asn252 and Asn291) and camel chymosin (Asn100 and Asn291). Glycosylation is favoured at Asn-X-Thr sequences (Asn100 Camel) as compared to Asn-X-Ser sequences (Asn252 and Asn291 bovine, Asn291 camel). Approximately 10% of bovine chymosin produced by *A. Niger* fermentation is glycosylated. A reduction in clotting activity is observed when camel chymosin is glycosylated at Asn291.[97]

6.3.2.2 Favoured Native Bovine Residues

GLN242ARG Residue 242 (Gln242 in bovine chymosin, Arg242 in camel chymosin) resides in a predominantly uncharged polar region on the surface of the C-terminal domain, where it interacts with the ArgP9 residue of κ-casein. Although early structural studies focused on the P8-P7' residues of κ-casein only,[15] the importance of the P9 position for binding has since been recognised because ArgP9 is conserved in bovine, camel, pig, buffalo, horse, and goat chymosin.[319] Furthermore, a variant of bovine κ-casein, in which the P9 position is occupied by a histidine, has been shown to be a poor substrate for bovine chymosin.[90]. The Gln242Arg mutation is the only sequence difference in the S9 pocket of bovine and camel chymosin.

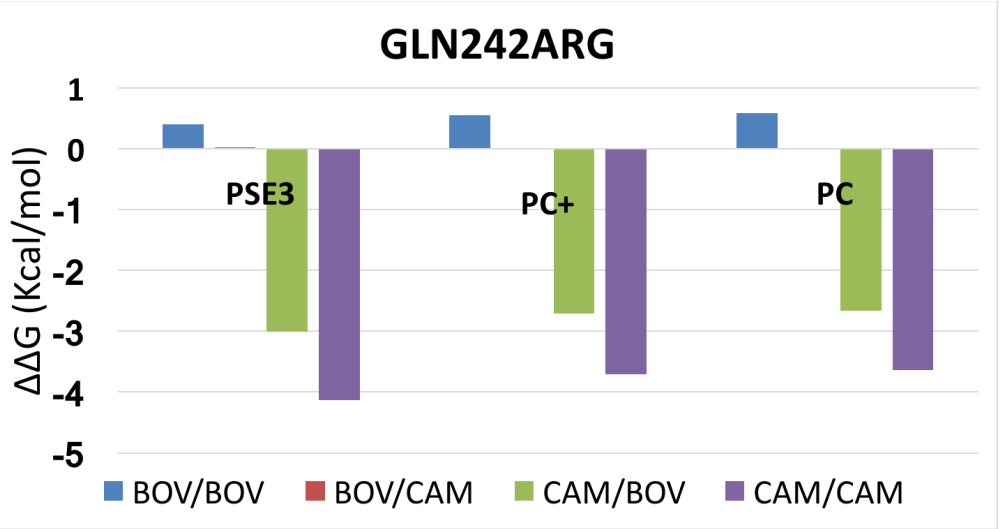


FIGURE 6.9: Comparison of alanine scanning results of residue 242 (Gln242 in bovine chymosin, Arg242 in camel chymosin)

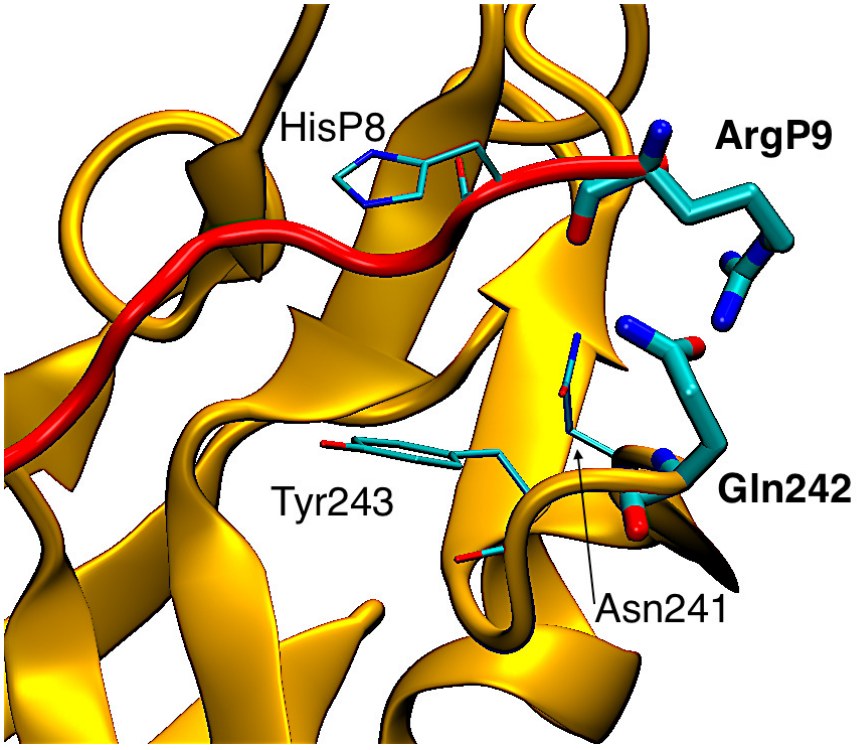


FIGURE 6.10: Snapshot of Gln242 and ArgP9 (in bold for clarity) at 15.2 ns of Bov/Bov complex MD simulation. The side chains of both residues extended towards each other.

The alanine scanning results for residue 242 (Figure 6.9) reveal a significant difference between the bovine and camel variants. For the Bov/Bov and Bov/Cam systems, Gln242 can be seen extending towards ArgP9 throughout the molecular

dynamics simulations (Figure 6.10). Consequently, mutating Gln242 to alanine results in weaker binding in these systems, as indicated by the positive $\Delta\Delta G$ obtained by alanine scanning (Figure 6.9). By contrast, in camel chymosin, where the interaction of Arg242 with ArgP9 is electrostatically and sterically unfavourable, the side chain of Arg242 is observed to extend partly out of the binding pocket (Figure 6.11). Here mutating Arg242 in camel chymosin to alanine shows a clear improvement in binding free energy with a reduction in $\Delta\Delta G$ by ~3.5 kcal/mol (Figure 6.9).

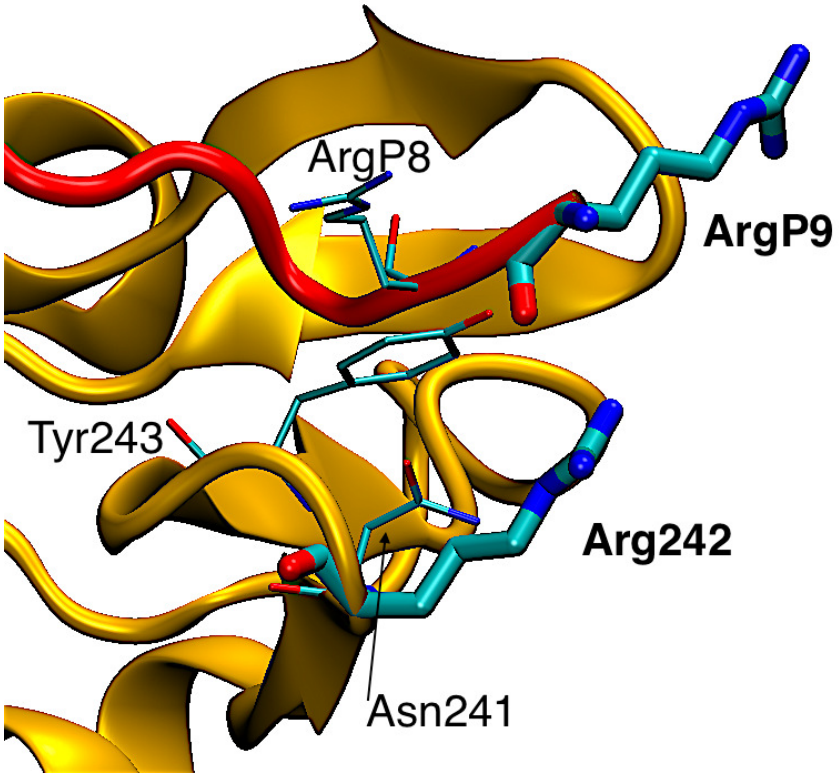


FIGURE 6.11: Snapshot of Arg242 and ArgP9 (in bold for clarity) at 53.2 ns of Cam/Cam complex MD simulation. The side chains of both residues extended away from the binding pocket.

The results suggest that the Gln residue observed in wildtype bovine enzyme is more favoured for binding than the Arg residue from its camel counterpart. The role of the residues in the S9 pocket has not previously been defined, but it may be to help orientate the neighbouring P8-P4 residues of κ-casein, which have been

implicated by both experimental and modelling studies in the allosteric activation of chymosin.[17,18,318]

GLN278LYS Residue 278 (Gln278 in bovine chymosin, Lys278 in camel chymosin) resides in a predominantly uncharged region on the surface of the N-terminal domain of chymosin, where it interacts with the P6 residue in κ-casein, which is HisP6 in bovine and the larger and more basic ArgP6 in camel. The S6 pocket is an open cleft formed by the side chains of the Ser277, Asp279 and Thr284 residues in the C-terminal domain and Asp13 and Ser14 residues in the N-terminal domain, all of which are conserved in bovine and camel chymosin. The only mutation site near the S6 pocket, Gln278Lys, is not a common target for protein engineering because crystallographic data shows that it lies at the bottom of the open cleft with the Gln or Lys side chain pointing away from the binding site in apo bovine or camel chymosin, respectively. In solution, as revealed by the molecular dynamics simulation, however, the flexibility of the side chain of residue 278 allows it to extend over the open cleft of the S6 pocket bringing it closer to the P6 residue of κ-casein. Alanine scanning results for residue 278 show that mutating the natural camel residue to alanine favours binding of bovine or camel κ-casein, as shown by the negative $\Delta\Delta G$ in Figure 6.12. Here the mutation to alanine removes an unfavourable Lys-Arg (Cam/Cam) or Lys-His(Cam/Bov) interaction. By contrast, mutating the bovine chymosin residue, Gln278 to alanine shows no significant change in binding free energy for either the Bov/Bov or Bov/Cam complex. The results suggest that Gln or Ala are favoured over Lys in the 278 position of chymosin.

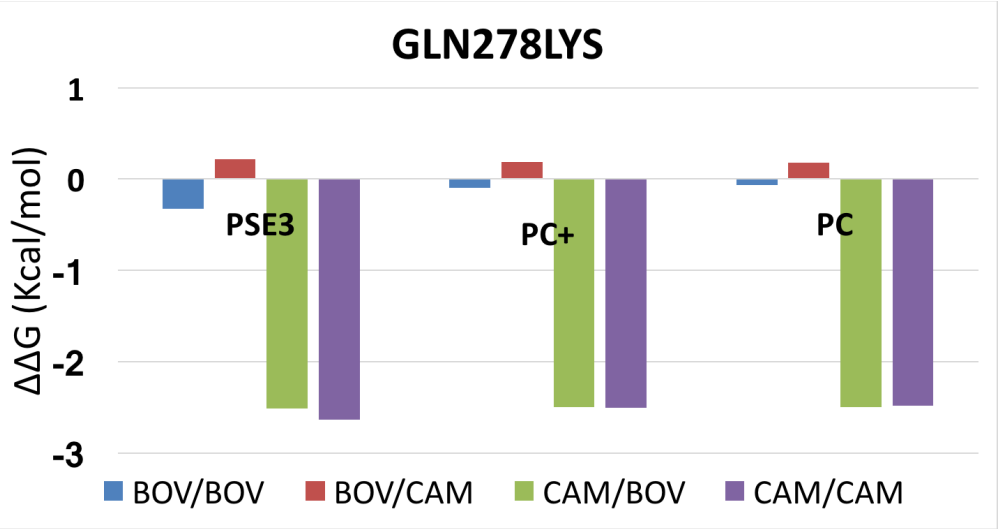


FIGURE 6.12: Comparison of alanine scanning results of residue 278 (Gln278 in bovine chymosin, Lys278 in camel chymosin), on the four different chymosin–κ-casein complexes with three different MM-3DRISM calculation methods. A negative $\Delta\Delta G$ represent a favourable mutation, and positive results are unfavourable.

6.3.2.3 Other Residues

The remaining four residues that were analysed through computational alanine scanning show no significant difference in $\Delta\Delta G_{bind}$ between bovine or camel chymosin complexes (Leu32Val, Ala117Ser, Met125Leu, and Val223Phe). The results for these residues are presented in Appendix B. It was found that changes in binding free energy were either too little to be considered a significant change, or the change was the same throughout all systems.

6.4 Changes in Solvent Density Distribution Due to Single-Point Mutations

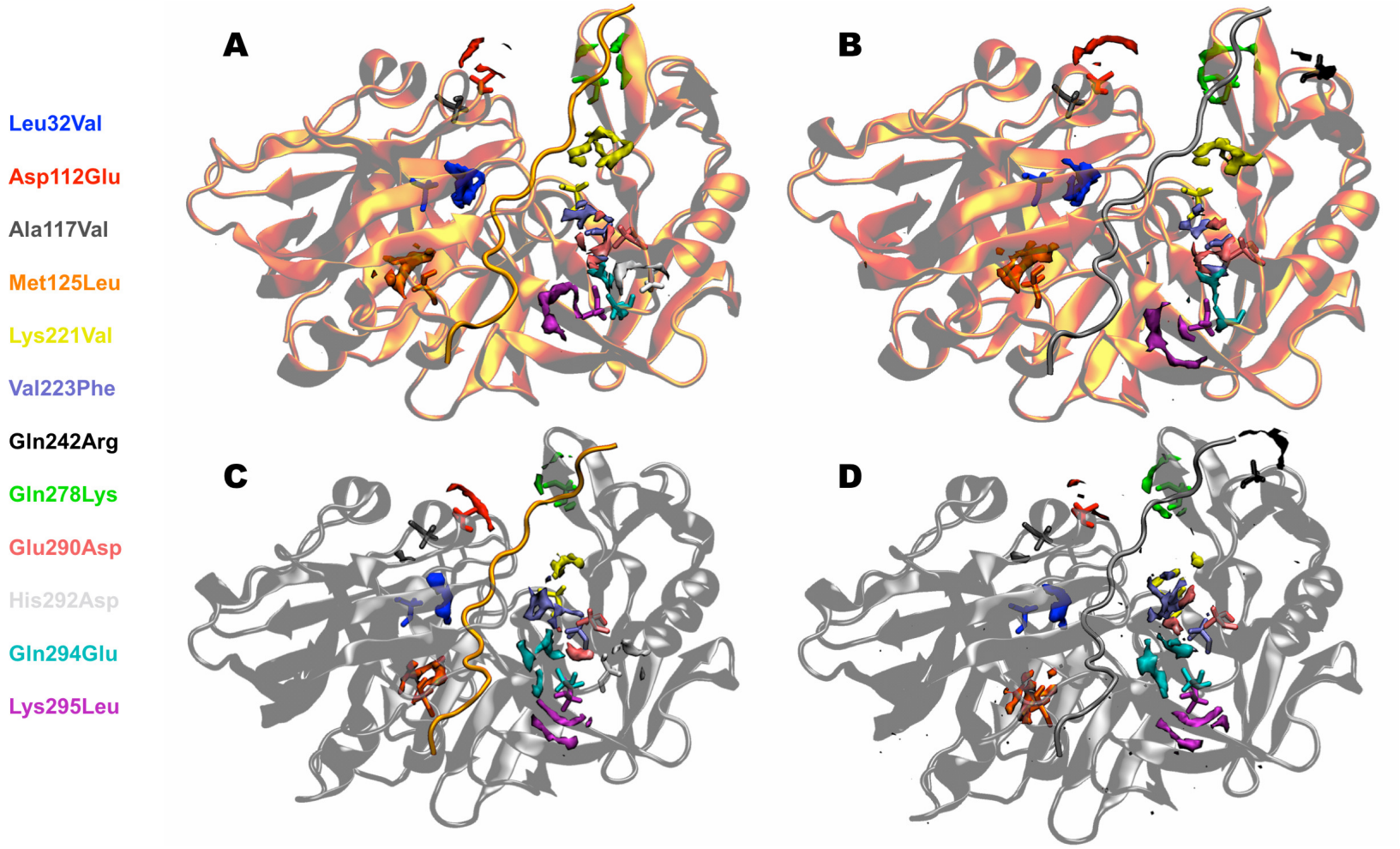


FIGURE 6.13: Illustration of the changes in local solvation density for 12 single-point mutations on the four different complexes used in this study. ***A*** *= Bovine chymosin (gold) in complex with bovine κ-casein (gold).* ***B*** *= Bovine chymosin (gold) in complex with camel κ-casein (silver).* ***C*** *= Camel chymosin (silver) in complex with bovine κ-casein (gold).* ***D*** *= Camel chymosin (silver) in complex with camel κ-casein (silver).* Each of the the coloured surfaces corresponds to an isosurface of a residue on which $g(\mathbf{r})_{\text{mutant}} - g(\mathbf{r})_{\text{wildtype}} = 3$. The colours of the isosurfaces correspond to alanine mutation of: *Leu32Val (blue), Asp112Glu (red), Ala117Val (grey), Met125Leu (orange), Lys221Val (yellow), Val223Phe (lilac), Gln242Arg (black), Gln278Lys (green), Glu290Asp (pink), His292Asp (off-white), Gln294Glu (cyan) and Lys295Leu (purple).*

As well as permitting estimates of solvation thermodynamics, the 3D-RISM calculations provide information about the local solvation of protein-ligand binding sites that can be readily visualised. The 3D-RISM solvent density functions, $g_\alpha(\mathbf{r}) = h_\alpha(\mathbf{r}) + 1$, give the spatial distribution of solvent density on a grid around the protein-ligand complex. The change in solvent density distribution that occurs due to a single-point mutation in the chymosin–κ-casein complex can be illustrated by taking the difference between the spatial density distribution functions of the mutant and wildtype complexes:

$$\Delta \mathrm{g}(\mathbf{r})\mathrm{m/w} = \mathrm{g}(\mathbf{r})_{\mathrm{mutant}} - \mathrm{g}(\mathbf{r})_{\mathrm{wildtype}} \text{ [123]} \tag{6.1}$$

Figure 6.13 shows the corresponding isosurfaces at $\Delta \mathrm{g}(\mathbf{r})\mathrm{m/w} = 3$, for each of the single point mutations introduced in the alanine scanning experiments in the four complexes. Changes in the local solvation are observed around each single-point mutation. As would be expected, the largest changes in solvation are localised within the binding site, with the most significant changes occurring due to changes in excluded volume.

6.5 Conclusions

Using molecular dynamics simulations and free energy calculations, binding in four different chymosin–κ-casein complexes (Bov/Bov, Bov/Cam, Cam/Bov, Cam/-Cam) have been investigated. By way of computational alanine scanning calculations, the influence that differences in the primary sequence of bovine and camel chymosin ("natural mutations") have on the binding thermodynamics in these complexes have been identified. Four of the natural mutations investigated here do not appear to differ in their contribution to $\Delta\Delta G_{bind}$ as both the bovine and

camel variants produce similar alanine scanning results. It is worth remembering that these residues can have specific interactions that assist in forming the optimal orientation of the complex, or facilitate the correct binding of κ-casein to chymosin. The alanine scanning results shows that there are eight important residues (112, 221, 242, 278, 290, 292, 294, and 295) that selectively influence binding thermodynamics. For Gln242Arg and Gln278Lys, the residue in bovine chymosin is more energetically favourable for binding. In the camel chymosin systems the alanine mutations are energetically favoured suggesting the polar positive residues in camel chymosin adversely influence the binding thermodynamics with κ-casein. By contrast, for mutations Asp112Glu, Lys221Val and Lys295Leu, the native camel variant is most favoured. All of these residues occupy separate and predominantly non-polar pockets along the binding cleft where the natural polar positive residues in bovine chymosin adversely influence the binding thermodynamics. Analysis of the solvent density distributions obtained by 3D-RISM illustrate that, as might be expected, mutation of binding site residues to alanine leads to localised changes in solvent density, with the largest contributions coming from excluded volume effects and polar functional groups.

It should be noted that there are a number of factors that are a part of the enzymatic process, and binding free energy is just one of them. Factors not considered in this study such as covalent bond breaking/forming and association/dissociation kinetics also affect the enzymatic process. Nonetheless, on the basis of the analysis carried out here, several residues have been identified for mutation with the aim of selectively modifying the binding free energy. Other aspects of chymosin catalysis, including the enzymatic reaction mechanism, are the subject of ongoing investigation.

Chapter 7

Comparative Molecular Field Analysis using Molecular Integral Equation Theory

7.1 Overview

The two previous chapters reported studies of protein-ligand binding in chymosin complexes that were carried out using existing computational methodologies (MD, MM-3DRISM, etc). In this chapter, a new method that has been developed for predicting protein-ligand binding affinities based on 3D-RISM and 3D-QSAR is discussed, extended and benchmarked.

One of the most widely used 3D-QSAR methods is the comparative molecular field analysis (CoMFA), which was proposed by Cramer *et al.* in 1988.[132] CoMFA establishes a uniform grid encompassing a series of pre-aligned molecules. Electrostatic and Lennard-Jones potential energies are then calculated between a positively charged carbon atom probe, located at each vertex of the grid, and each of the molecules embedded within.[132] The resulting electrostatic and "steric" fields

are used as input for partial-least-squares regression models. Since its first publication, CoMFA has been cited in over 4000 published articles and used in numerous drug discovery programs.[139,140] Several extensions to the CoMFA methodology have been proposed, of which the highest profile is comparative molecular similarity indices analysis (CoMSIA).[141,142] CoMFA considers both electrostatic and steric fields to correlate activity. While CoMSIA considers electrostatic, steric, hydrophobic and hydrogen bonding potentials.

Although CoMFA is widely used, it relies on a relatively simple representation of molecular interactions, which does not explicitly account for solvation/desolvation effects that can dramatically influence protein-ligand binding. Since CoMFA was first proposed, advances in theory, algorithms and computer power mean that there are now many fast and accurate methods to model molecular solvation effects. Some success has been achieved using numerical simulation (e.g. Monte Carlo or molecular dynamics simulations) to compute solute-solvent descriptors for QSAR models,[320] but such methods are computationally expensive and subject to sampling errors that reduce the signal-to-noise ratio in the modelling dataset. Integral equation theory approaches are of particular interest for QSAR modelling because they allow solute-solvent distributions and solvation thermodynamics to be computed at a fraction of the cost of explicit solvent numerical simulations and with no sampling error.[2,238,248] The most widely used of these methods are the 1D and 3D reference interaction site models (RISM) proposed by Chandler et al.[321] and Beglov and Roux,[151–153] respectively. Accurate predictions of hydration free energy and Caco-2 permeability have previously been reported using QSAR models based on 1D RISM molecular descriptors.[154]. Recently, Güssregen et al. proposed the CARMa methodology, which uses solute–solvent distribution functions calculated by 3DRISM to replace the electrostatic or steric fields in CoMFA.[155] This approach was shown to give accurate predictions of binding affinities for a series of serine protease inhibitors, but tests on other systems have not yet been published.[155]

The purpose of the research conducted in this chapter is two-fold. Firstly, the study proposes an extension to the CARMa methodology. CARMa uses a statistical mechanics solvent model to capture solvation effects, but does not directly model the electrostatic and steric effects probed by CoMFA. Solving the 3D-RISM equations for a solvent comprising CoMFA probes in aqueous solution addresses this issue and results in predictions that are more accurate than either CoMFA or the original CARMa model. Secondly, an extensive benchmark of both CARMa models is conducted over six different protein-ligand systems and the results are compared to previously published CoMFA and 3D-QSAR results. The influence of algorithmic parameters, such as the 3D-RISM bridge-functional and grid-size, on the prediction accuracy are systematically investigated.

7.2 Methods

7.2.1 QSAR Data Sets

Six datasets were selected to benchmark the CARMa predictions. Firstly, the 21 steroids selected by Cramer et al. were used to provide a direct comparison between CARMa and CoMFA (Figure 7.1).[132,140] Optimised and aligned structures for all 21 molecules were taken from Coates et al.[140]; these files resolve some errors in the original structures reported by Cramer *et al.*[140]

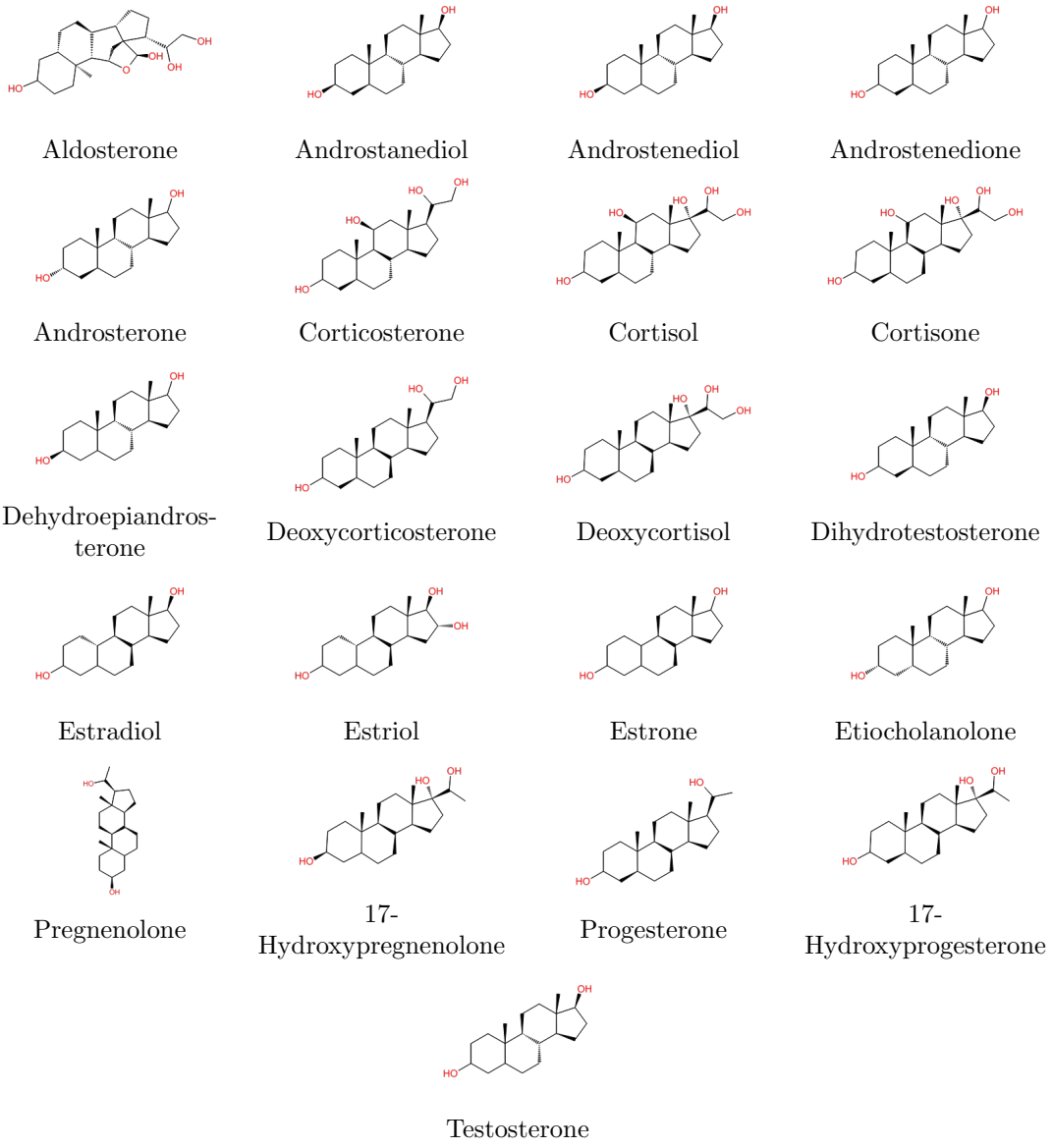


FIGURE 7.1: A depiction of steroids training set.

Secondly, five pIC_{50} data sets published by Sutherland *et al.* were used to compare CARMa to a wide-range of 3D-QSAR methods (including CoMFA). The compounds with literature references, aligned molecular structures, and grid parameters for field based QSAR are all described by Sutherland *et al.*[124] Briefly, the datasets are: **ACE dataset** – 114 angiotensin converting enzyme (ACE) inhibitors with pIC_{50} values ranging between 2.1 and 9.9;[322] **AchE dataset** – 111 acetylcholinesterase (AchE) inhibitors with pIC_{50} values ranging between 4.3 and 9.5;[323] **BZR dataset** – 163 ligands for the benzodiazepine receptor (BZR) with pIC_{50} values ranging between 5.5 and 8.9;[324] **COX2 dataset** – 322 cyclooxygenase-2

(COX2) inhibitors with pIC_{50} values ranging between 4.0 and 9.0;[325] **DHFR dataset** – 397 dihydrofolate reductase (DHFR) inhibitors with pIC_{50} values ranging between 3.3 and 9.8.[326] Sutherland *et al* used a "cherry picking" with maximum dissimilarity algorithm to assign 33% of the dataset to the test set and the remaining compounds to the training set.[327,328] To allow a direct comparison with Sutherland's results, the same aligned molecular conformations and training/test sets reported by Sutherland *et al.* are used here.

7.2.2 3D-RISM

The 3DRISM calculations were performed using AmberTools16.[299] The KH closure was used for solution of the 3D-RISM equations. The linear grid spacing in each of the three directions was 0.5 Å. The MDIIS iterative scheme was employed,[311] using 5 MDIIS vectors, MDIIS step size of 0.7, residual tolerance of 10^{-10}.

Solvent susceptibility functions required as input to 3D-RISM were calculated using dielectrically consistent 1D-RISM[329] with the KH closure. The grid size for 1D-functions was 0.025 Å, which gave a total of 16384 grid points. The MDIIS iterative scheme was used, having a total of 20 MDIIS vectors, MDIIS step size of 0.3, and residual tolerance of 10^{-12}. The solvent was considered to be pure water with a number density 0.0333 $Å^{-3}$, and a dielectric constant of 78.497. The solvent isothermal compressibility evaluated from the 1D-RISM calculation was $k_B T\eta = 1.949459$ $Å^3$.

The Lue and Blankschtein version of the SPC/E model of water (MSPC/E) was used.[330] This differs from the original SPC/E water model[331] by the addition of modified Lennard-Jones (LJ) potential parameters for the water hydrogen, which

were altered to prevent possible divergence of the algorithm.[332–335] The Lorentz-Berthelot mixing rules were used to generate the solute-water LJ potential parameters[336]. The following LJ parameters (for water hydrogen) were used to calculate the interactions between solute sites and water hydrogens: $\sigma_{H_w}^{LJ} = 1.1657$ Å and $\epsilon_{H_w}^{LJ} = 0.0155$ kcal/mol.

7.2.3 3D-RISM–QSAR

Two different classes of functions were tested as input to CARMa analyses: *solvent density distribution functions*, $g(r)$, which represent the local solvent density at grid points around the solute; *solvation free energy density functions*, which indicate the local contribution to the excess chemical potential of the solute.

7.2.3.1 Solvent Density Distribution Functions

Solving the 3D-RISM equations gives a solvent density distribution function, $g(r)$, for each interaction site (atom) in the solvent. Four different $g(r)$ functions were tested as input to CARMa: (i) *water density distribution functions*, $g_O(r)$ or $g_H(r)$, computed for pure aqueous solvent; (ii) *solvent-probe density distribution functions*, $g_{C+}(r)$ or $g_{C-}(r)$, obtained by solving the 3D-RISM equations with 0.1 M $g_{C+}(r)$ and 0.1 M $g_{C-}(r)$ probe atoms as co-solvents in aqueous solution. The $g_{C+}(r)$ and $g_{C-}(r)$ probes are positively or negatively charged sp^3 carbon atoms with Lennard-Jones parameters taken from the AmberGAFF2 forcefield.

7.2.3.2 Solvent Free Energy Density

Within the framework of the RISM theory there exist several approximate functionals that allow one to analytically obtain values of the solvation free energy

from the total $h_\alpha(\mathbf{r})$ and direct $c_\alpha(\mathbf{r})$ correlation functions.[238–240]. These can be derived analytically from the appropriate 3D-RISM closure relationship. In this chapter the KH (PSE-1) closure is used primarily unless explicitly stated that the PSE-3 closure has been used. More information about the different closure relationships can be found in Chapter 4.

7.2.4 Grids

Both the local solvent density (as given by $g_O(\mathbf{r})$) and the solvation free energy density ($w(\mathbf{r})$) are represented on discrete grids. In principle, the values of these functions at specific grid points could be used directly as input to the CARMa models. Normally 3D-RISM calculations are carried out on a large grid with a small grid spacing (0.3-0.5 Å), this would lead to many redundant variables, thus making the numerical data sets too large to be processed easily. A simple solution would be to solve the 3D RISM calculations on a small and coarse grid, but this would reduce the accuracy of the obtained density distribution functions. Instead, in this study, all 3D-RISM calculations were performed on a large and fine grid (>50 Å^3 grid with a 0.5 Å spacing). The output from 3D-RISM was then modified by reducing the size of the grid to 22 Å^3 by removing layers of each grid face as appropriate (using custom Python scripts). To provide a further filter to remove some of the unnecessary variables, two different approaches have been tested: (i) mapping the 3D-RISM results onto a coarser grid; (ii) selecting only those grid points that were within a distance, d, from the solute. The latter method increased computational expense without improving prediction accuracy and, therefore, is not discussed further. Prior to statistical modelling, all variables that had a variance of zero were removed.

7.2.5 CARMa

7.2.5.1 Statistical and Machine Learning Algorithms

To derive the predictive CARMa models, two different methods of regression were considered: partial-least-squares (PLS) and random forest (RF). A genetic algorithm (GA) was also tested to select input variables for the PLS model. The regression methods are described in detail in Chapter 4.

CARMa models were setup and trained using a combination of bespoke Python and R scripts.

7.2.5.2 Partial-Least Squares

Partial-Least Squares regression models were trained using the *pls* library[337] in the R statistical computing environment.[338] All PLS models were trained with 3 latent variables, which was selected as optimal balance between model size and prediction accuracy based on consideration of the residual error sum of squares and the percentage of variance explained.

7.2.5.3 Random Forest

Random forest models were trained with the *randomForest* library[339] in the R statistical computing environment,[338] using standard parameters: $mtry = N/3$, $nodesize = 5$, and $ntree = 500$, where N is the number of input variables and $mtry$ is rounded down to the nearest integer. There is extensive evidence in the literature that the random forest algorithm is insensitive to training parameters,[280,340] so that variation of $mtry$ between 40 and N, of $ntree$ from 250 upward, and of $nodesize$ in the region 5 to 10 has little effect on prediction accuracy. As has been done previously, these standard random forest parameters are used without

further optimization.[280,340]

7.2.6 Statistical Analysis

To compare computational predictions with experimental data, a correlation coefficient and the root mean squared deviation ($RMSD$) were evaluated:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(x^i - y^i)^2}{\sum_{i=1}^{n}(x^i - M(y^i))^2}, \tag{7.1}$$

$$RMSD(x, y) = \sqrt{\frac{1}{N}\sum_i (x^i - y^i)^2} \tag{7.2}$$

where index i runs through the set of N selected molecules, and x^i and y^i are values calculated by different computational methods, for molecule i for a given property. The total deviation can be split into two parts: bias (or mean displacement, M) and standard deviation (σ), which are calculated by the formulae:

$$bias = M(x - y) = \frac{1}{N}\sum_{i \in S}\left(x^i - y^i\right) \tag{7.3}$$

$$\sigma(x - y) = \sqrt{\frac{1}{N}\sum_{i \in S}\left(x^{(i)} - y^{(i)} - M(x - y)\right)^2} \tag{7.4}$$

The bias gives a systematic error, which can be corrected by a simple constant term. The standard deviation gives the random error that is not explained by the model. The connection between these three formulae can be seen in Equation 7.5.

$$RMSD(x,y)^2 = M(x-y)^2 + \sigma(x-y)^2 \tag{7.5}$$

Statistical analyses were carried out in the R Statistical Computing Environment.[341] Python scripts were used to manipulate raw data files.

7.2.7 Computational Expense

The CARMa calculations reported here were performed using a quad-core, 3.4GHz Intel Core i5 iMac desktop with 16GB RAM (late 2013, operating system version 10.12.2). The most time-consuming step in making a prediction with a pre-trained 3D-RISM–CARMa model is solving the 3D-RISM equations; the remaining steps require negligible computational expense. The mean time required to solve the 3D-RISM equations for an individual molecule in the steroid dataset was ∼1 min. By their nature, 3D-RISM calculations are trivially parallel (e.g. one calculation per node), but the time required for a single calculation could be significantly reduced by using advanced numerical algorithms[342,343] or by performing the simulations using parallel computation.[225]

7.3 Results

7.3.1 Steroid dataset

The steroids dataset consists of 21 compounds with corticosteroid-binding globulins (CBG) binding affinity data. Cramer et al. report a $q^2 = 0.734$ for leave-one-out cross-validation of a CoMFA model,[140] which represents a relatively accurate prediction of the CBG binding affinity data.

TABLE 7.1: Steroids leave-one-out cross-validation statistics (q^2) using CARMa with various descriptors and grid spacings.

Grid Spacing (Å)	$g_O(r)$	$g_O^{PSE3}(r)^a$	$g_H(r)$	*SFED*^{b}	$g_{C-}(r)^c$	$g_{C+}(r)^d$	*CoMFA*
			PLS				
1.0	0.84	0.85	0.84	0.68	0.84	0.84	-
1.5	0.86	0.86	0.85	0.67	0.85	0.84	-
2.0	0.84	0.84	0.83	0.69	0.85	0.84	0.73
2.5	0.81	0.81	0.85	0.74	0.83	0.83	-
3.0	0.85	0.86	0.85	0.67	0.83	0.84	-

[a] Partial Series Expansion-3 closure; [b] Solvation Free Energy Density; [c] sp^3 Carbon probe atom with -1 charge; [d] sp^3 Carbon probe atom with +1 charge.

Table 7.1 presents q^2 values for LOO-CV of CARMa models built using the PLS method and trained on six different distribution functions represented on six different grids. Several different trends are evident in Table 7.1. Firstly, the choice of bridge functional used to solve the 3D RISM equations (KH or PSE-3) does not significantly influence the results. The q^2 values for CARMa models built on $g_O^{KH}(r)$ or $g_O^{PSE3}(r)$ are nearly identical for all grid sizes. A similar conclusion was reached in literature reports that used PLS models trained on 1D-RISM descriptors to predict hydration free energy and Caco-2 permeability. Secondly, for this dataset, the PLS models trained on solvation density distributions ($g_O(r)$, $g_O^{PSE3}(r)$, $g_H(r)$, $g_{C-}(r)$ and $g_{C+}(r)$) perform better than those trained on solvation free energy density (SFED). Thirdly, there is no obvious trend between the various grid spacings. Although finer grids might be expected to lead to more

accurate models, this is not evident in the data, which suggests that some redundancy is present in the finer grids.

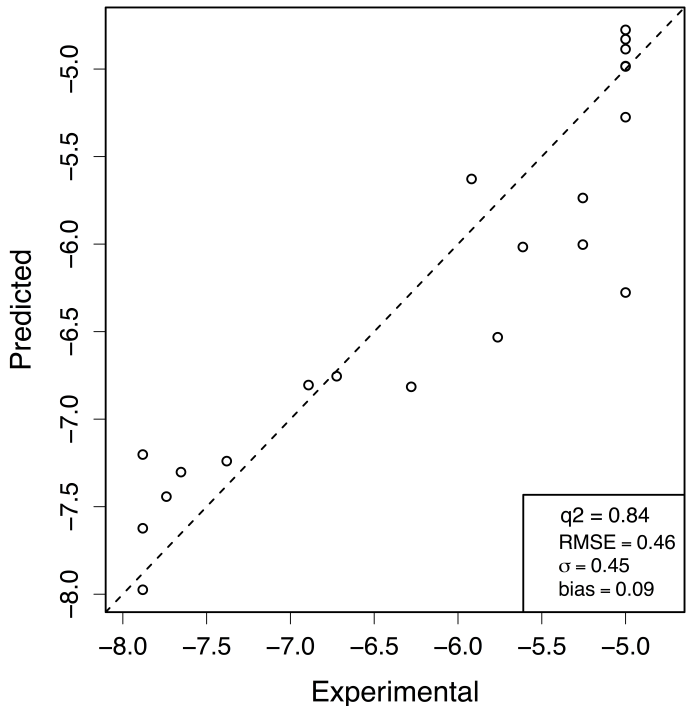
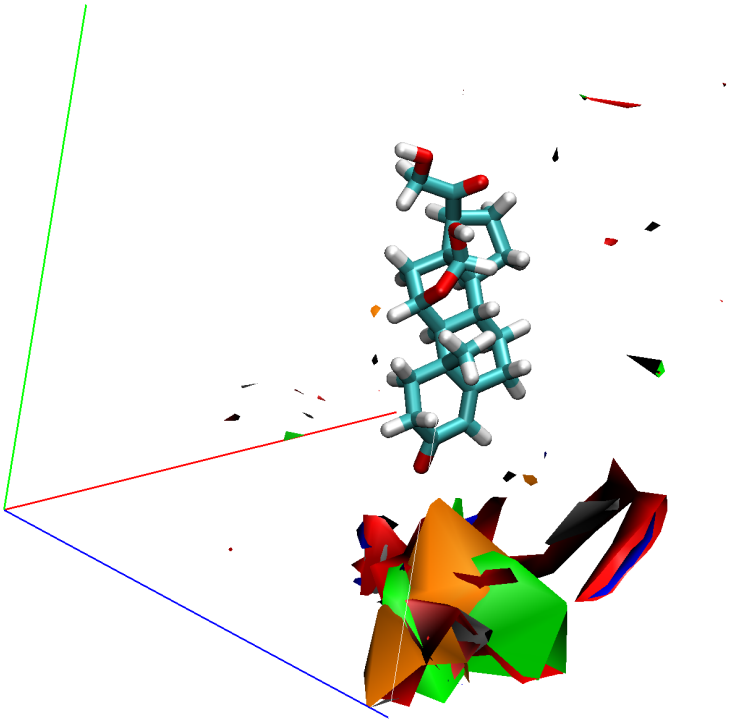
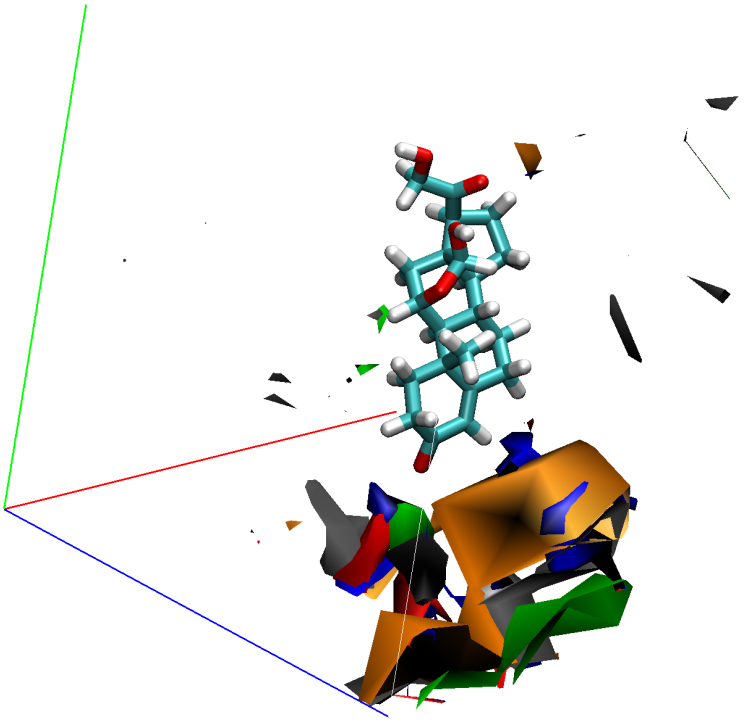


FIGURE 7.2: Correlation graph of leave-one-out cross-validation (LOO-CV) for PLS models using the $g_O(r)$ distribution data at 2.0 Å grid spacing.

Figure 7.2 shows the cross-validated predictions obtained for PLS models trained on $g_O(r)$ distribution functions represented on a 2 Å grid; the same grid spacing used in the CoMFA models. The CARMa model explains more of the variance in the experimental data than the CoMFA model, as exemplified by $q^2 = 0.84$ for CARMa compared to $q^2 = 0.73$ for CoMFA. The residual cross-validated error in the CARMa model ($RMSE = 0.46$) is predominantly due to random error ($\sigma = 0.45$) with a relatively small systematic error ($bias = 0.09$).



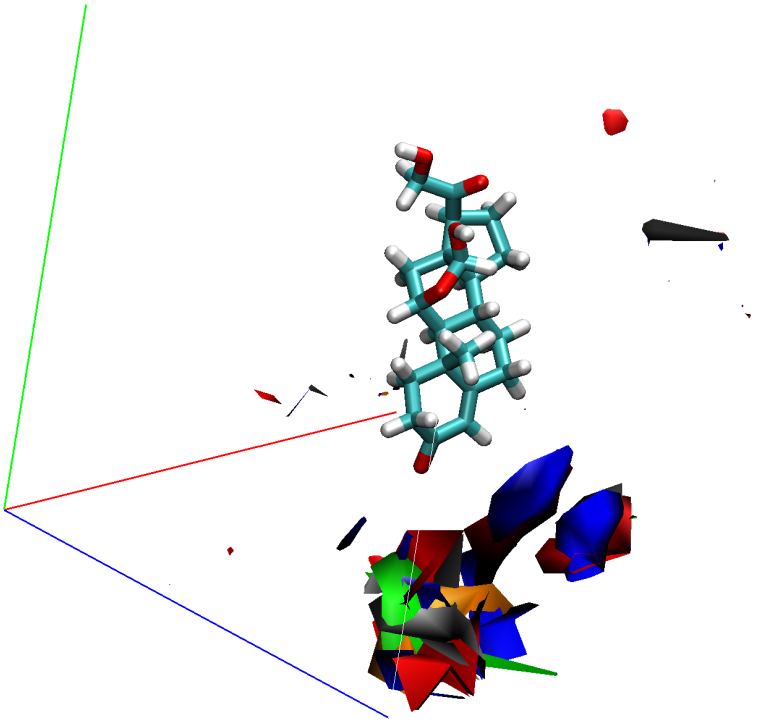
(A) $g_O(r)$ distribution importance.



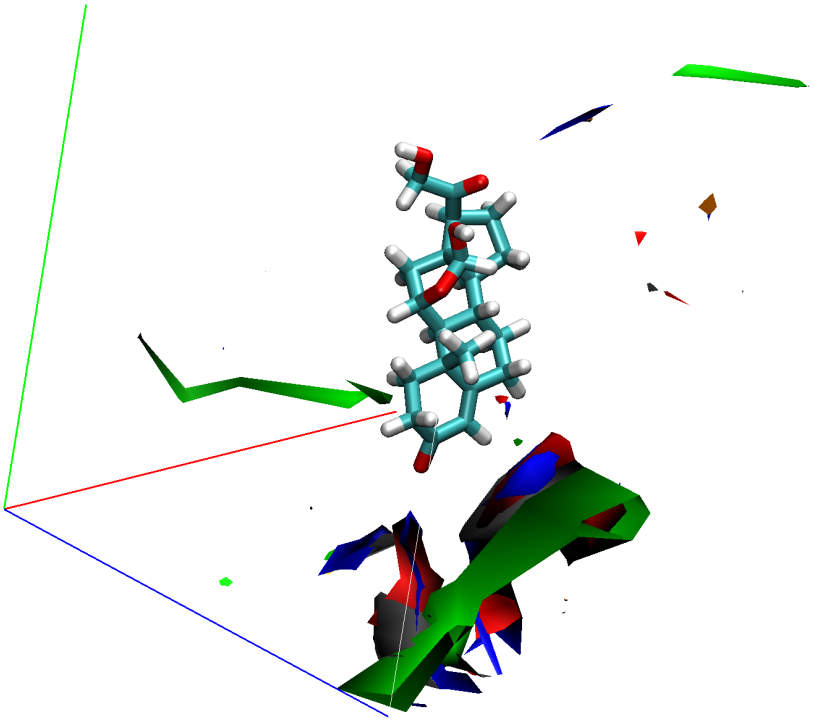
(B) $g_H(r)$ distribution importance.

FIGURE 7.3: Aldosterone is shown with PLS importance of $g_O(r)$ and $g_H(r)$ distributions at grid spacings 1.0 (*blue*), 1.5 (*red*), 2.0 (*grey*), 2.5 (*orange*) and 3.0 (*green*). The graphics show 10% of the most important regions for the PLS models, derived from the importance metric.

The total contribution that each input variable made to the PLS latent variables was used as a metric to assess its importance to the model. Figure 7.3 depicts the most important 10% of the $g_O(r)$ and $g_H(r)$ functions as assessed from the PLS models. There is little difference between the $g_O(r)$ and $g_H(r)$ descriptor models, which is perhaps not surprising given that oxygen and hydrogen atoms are covalently bonded in water. In Figures 7.3a and 7.3b, the regions highlighted are located by the terminal cyclo-hexane (ring A) of the steroids for all grid spacings. A similar trend is observed in the importance graphics for the $g_{C-}(r)$ and $g_{C+}(r)$ probe atom distributions (see Figure 7.4), but here the distributions seem to be more localised in space in comparison to those for $g_O(r)$ and $g_H(r)$ (Figure 7.3).



(A) $g_{C-}(r)$ distribution importance.



(B) $g_{C+}(r)$ distribution importance.

FIGURE 7.4: Aldosterone is shown with PLS importance of $g_{C-}(r)$ and $g_{C+}(r)$ distributions at grid spacings 1.0 (*blue*), 1.5 (*red*), 2.0 (*grey*), 2.5 (*orange*) and 3.0 (*green*). The graphics show 10% of the most important regions for the PLS models, derived from the importance metric.

7.3.2 pIC_{50} Data Sets

To further validate the methodology, CARMa models were developed to predict pIC_{50} values for five datasets collated by Sutherland et al.[124]. In each case, the training/testing datasets and aligned molecular structures selected by Sutherland et al. were used to provide a direct comparison to their CoMFA and 3D-QSAR results.

In total, 450 different CARMa models were considered (5 3D-RISM fields $\times$ 6 grid spacings $\times$ 3 regression methods $\times$ 5 datasets). All of the results are compiled in Table 7.2 (training dataset) and Table 7.3 (testing dataset). Since correlation coefficients (q^2 or R^2) and predictive errors ($RMSE$) were found to be highly correlated for these datasets, only the correlation coefficients are presented in Tables 7.2 and 7.3, but all other statistics (RMSE, σ, bias) are provided in Appendix C. The "-" entries in Tables 7.2 and 7.3 indicate that training PLS or RF models on 3D-RISM fields with a 0.5 Å grid spacing was found to be prohibitively computationally expensive. The best predictions for the external test set are summarised in Table 7.4.

TABLE 7.2: Leave-one-out cross-validation statistics (q^2) for 5 pIC_{50} data sets using CARMa with various descriptors and grid spacings. In bold are the best models and each dataset using the PLS, GA-PLS and RF models.

GS^a (Å)	$g_O(r)^b$			$g_H(r)^c$			$SFED^d$			$g_{C-}(r)^e$			$g_{C+}(r)^f$		
	PLS	GA-PLS	RF	PLS	GA-PLS	RF	PLS	GA-PLS	RF	PLS	GA-PLS	RF	PLS	GA-PLS	RF
ACE															
0.5	-	0.702	-	-	0.698	-	-	0.666	-	-	0.714	-	-	0.709	-
1.0	0.640	0.759	0.683	0.629	0.751	0.692	0.622	0.695	0.682	0.649	0.752	0.675	0.649	0.747	0.695
1.5	0.621	0.780	0.696	0.621	0.777	0.676	0.620	0.731	0.684	0.648	0.774	0.692	0.658	0.754	0.675
2.0	0.619	0.743	0.679	0.642	***0.795***	0.681	0.619	0.732	0.687	0.626	0.758	0.695	0.637	0.741	0.680
2.5	0.634	0.672	0.697	0.609	0.722	0.701	0.638	0.743	0.696	**0.670**	0.752	0.707	0.658	0.716	0.702
3.0	0.633	0.685	0.698	0.623	0.717	0.693	0.617	0.685	0.688	0.631	0.672	**0.717**	0.659	0.714	0.692
AchE															
0.5	-	0.564	-	-	0.579	-	-	0.363	-	-	0.572	-	-	0.559	-
1.0	0.490	0.639	0.399	0.497	0.629	0.402	0.276	0.459	0.411	0.493	0.634	0.411	0.487	0.629	0.431
1.5	0.482	0.639	0.427	0.492	0.643	0.426	0.261	0.491	0.426	0.483	0.641	0.383	0.488	***0.682***	0.412
2.0	0.455	0.644	0.395	0.487	0.640	0.414	0.271	0.550	0.418	0.446	0.604	0.403	0.499	0.659	0.398
2.5	**0.508**	0.565	0.412	0.504	0.633	0.392	0.269	0.487	0.436	0.463	0.584	0.395	0.430	0.601	0.392
3.0	0.444	0.623	0.407	0.493	0.590	0.417	0.240	0.446	0.424	0.449	0.627	0.426	0.502	0.628	**0.449**
BZR															
0.5	-	0.314	-	-	0.309	-	-	0.214	-	-	0.310	-	-	0.306	-
1.0	0.232	0.389	0.314	0.235	0.373	0.306	0.117	0.323	0.281	0.232	0.397	0.367	0.211	0.369	0.363
1.5	0.230	0.416	0.309	0.234	0.397	0.296	0.114	0.365	0.286	0.238	0.439	0.337	0.224	0.442	**0.378**
2.0	0.217	0.404	0.353	0.223	0.390	0.309	0.111	0.366	0.298	0.234	0.431	0.356	0.228	0.426	0.341
2.5	0.202	0.456	0.304	0.240	0.390	0.310	0.116	0.410	0.279	**0.243**	0.416	0.354	0.208	0.419	0.346
3.0	0.210	0.417	0.297	0.247	***0.458***	0.328	0.098	0.391	0.263	0.219	0.427	0.311	0.179	0.406	0.336
COX2															
0.5	-	0.460	-	-	0.451	-	-	0.357	-	-	0.462	-	-	0.457	-
1.0	0.426	0.483	0.462	0.421	0.477	0.457	0.276	0.397	0.416	0.427	0.503	0.438	0.426	0.508	0.447
1.5	0.420	0.494	0.457	0.416	0.498	0.473	0.266	0.412	0.405	0.421	0.515	0.458	0.426	***0.536***	0.449
2.0	0.421	0.476	0.466	0.424	0.467	**0.473**	0.262	0.385	0.418	0.416	0.480	0.434	0.415	0.473	0.435
2.5	0.419	0.441	0.459	0.397	0.468	0.431	0.281	0.347	0.411	**0.439**	0.460	0.428	0.412	0.462	0.443
3.0	0.424	0.464	0.463	0.412	0.465	0.454	0.320	0.351	0.427	0.427	0.437	0.440	0.422	0.436	0.458
DHFR															
0.5	-	0.619	-	-	0.622	-	-	0.523	-	-	0.633	-	-	0.627	-
1.0	0.581	0.647	0.634	0.586	0.659	0.636	0.488	0.601	0.642	0.591	0.649	0.646	0.594	0.667	**0.644**
1.5	0.574	0.628	0.636	0.585	***0.668***	0.634	0.485	0.586	0.642	0.587	0.657	0.642	0.593	0.663	0.632
2.0	0.578	0.593	0.631	0.576	0.630	0.634	0.487	0.599	0.642	0.587	0.637	0.638	0.580	0.618	0.640
2.5	0.562	0.523	0.626	0.573	0.602	0.651	0.480	0.581	0.643	0.584	0.599	0.639	0.589	0.609	0.642
3.0	0.568	0.623	0.633	0.583	0.626	0.625	0.482	0.530	0.651	0.589	0.550	0.659	**0.598**	0.632	0.624

[a] Grid spacing; [b] Oxygen distribution; [c] Hydrogen distribution; [d] Solvent free energy distribution; [e] Carbon probe distribution with -1 charge; [f] Carbon probe distribution with +1 charge.

TABLE 7.3: Test set predictive accuracy statistics (r^2) for 5 pIC_{50} data sets using CARMa with various descriptors and grid spacings. In bold are the best models and each dataset using the PLS, GA-PLS and RF models.

GS[a] (Å)	$g_O(r)$[b]			$g_H(r)$[c]			*SFED*[d]			$g_{C-}(r)$[e]			$g_{C+}(r)$[f]		
	PLS	GA-PLS	RF	PLS	GA-PLS	RF	PLS	GA-PLS	RF	PLS	GA-PLS	RF	PLS	GA-PLS	RF
ACE															
0.5	-	0.541	-	-	0.565	-	-	0.472	-	-	0.597	-	-	**0.615**	-
1.0	0.558	0.531	0.621	0.571	0.459	0.595	0.447	0.474	0.615	0.611	0.564	**0.636**	0.605	0.582	0.614
1.5	0.532	0.460	0.612	0.561	0.510	0.613	0.440	0.559	0.603	0.605	0.513	0.621	0.585	0.502	0.599
2.0	0.519	0.156	0.592	0.572	0.373	0.559	0.450	0.487	0.575	0.596	0.345	0.621	0.623	0.497	0.589
2.5	0.542	0.481	0.578	0.531	0.467	0.605	0.458	0.507	0.577	***0.638***	0.458	0.625	0.631	0.502	0.616
3.0	0.505	0.395	0.601	0.513	0.501	0.550	0.433	0.397	0.608	0.578	0.424	0.631	0.602	0.459	0.608
AchE															
0.5	-	0.670	-	-	0.673	-	-	0.438	-	-	0.676	-	-	***0.697***	-
1.0	0.626	0.629	0.454	0.632	0.637	0.476	0.404	0.474	0.506	0.660	0.587	0.405	**0.665**	0.601	0.402
1.5	0.632	0.422	0.460	0.623	0.490	0.488	0.414	0.423	0.518	0.658	0.696	0.443	0.659	0.595	0.385
2.0	0.587	0.459	0.454	0.583	0.317	0.484	0.366	0.494	**0.537**	0.634	0.491	0.445	0.644	0.500	0.359
2.5	0.603	0.393	0.493	0.648	0.481	0.471	0.373	0.364	0.495	0.608	0.345	0.456	0.601	0.364	0.359
3.0	0.606	0.429	0.468	0.637	0.365	0.472	0.383	0.335	0.526	0.621	0.208	0.431	0.654	0.269	0.397
BZR															
0.5	-	0.186	-	-	0.166	-	-	0.088	-	-	0.183	-	-	0.187	-
1.0	0.177	0.142	0.202	0.190	0.142	0.198	0.095	0.114	0.205	0.203	0.125	0.198	0.197	0.165	0.192
1.5	0.184	0.078	0.197	0.181	0.205	0.202	0.084	0.045	0.180	**0.209**	0.195	0.196	0.192	0.092	0.203
2.0	0.171	0.130	0.214	0.194	**0.208**	0.193	0.092	0.156	0.193	0.184	0.055	0.199	0.191	0.150	0.198
2.5	0.189	0.116	0.189	0.188	0.033	0.186	0.111	0.125	0.188	0.172	0.114	0.183	0.184	0.074	0.185
3.0	0.155	0.076	0.203	0.166	0.060	0.193	0.071	0.118	0.188	0.193	0.102	***0.217***	0.151	0.095	0.209
COX2															
0.5	-	0.327	-	-	0.334	-	-	0.200	-	-	**0.351**	-	-	0.336	-
1.0	0.343	0.322	0.347	0.346	0.341	0.347	0.176	0.241	0.355	0.365	0.342	0.353	0.366	0.282	0.341
1.5	0.326	0.243	0.353	0.348	0.224	0.357	0.166	0.260	0.372	0.344	0.248	0.364	0.363	0.266	0.370
2.0	0.308	0.303	0.338	0.331	0.257	0.335	0.188	0.183	0.341	0.334	0.216	0.357	0.342	0.269	0.339
2.5	0.303	0.164	0.349	0.299	0.251	0.339	0.176	0.182	0.346	0.312	0.225	0.338	0.323	0.228	0.374
3.0	0.323	0.249	0.343	0.323	0.188	0.318	0.156	0.139	0.367	0.348	0.198	0.318	***0.382***	0.172	**0.375**
DHFR															
0.5	-	0.548	-	-	0.545	-	-	0.421	-	-	**0.567**	-	-	0.548	-
1.0	0.540	0.513	0.603	0.539	0.548	0.604	0.397	0.485	0.567	0.533	0.524	0.597	0.538	0.514	0.600
1.5	0.534	0.532	0.606	0.535	0.527	0.601	0.392	0.486	0.566	0.529	0.560	0.590	0.537	0.504	0.630
2.0	0.517	0.439	0.610	0.531	0.519	0.612	0.390	0.455	0.555	0.510	0.430	0.601	0.532	0.475	0.604
2.5	0.518	0.396	0.621	0.515	0.371	0.622	0.381	0.410	0.540	0.524	0.497	0.598	0.538	0.453	0.620
3.0	0.536	0.425	***0.652***	0.530	0.463	0.612	0.375	0.351	0.528	0.532	0.454	0.589	**0.562**	0.515	0.613

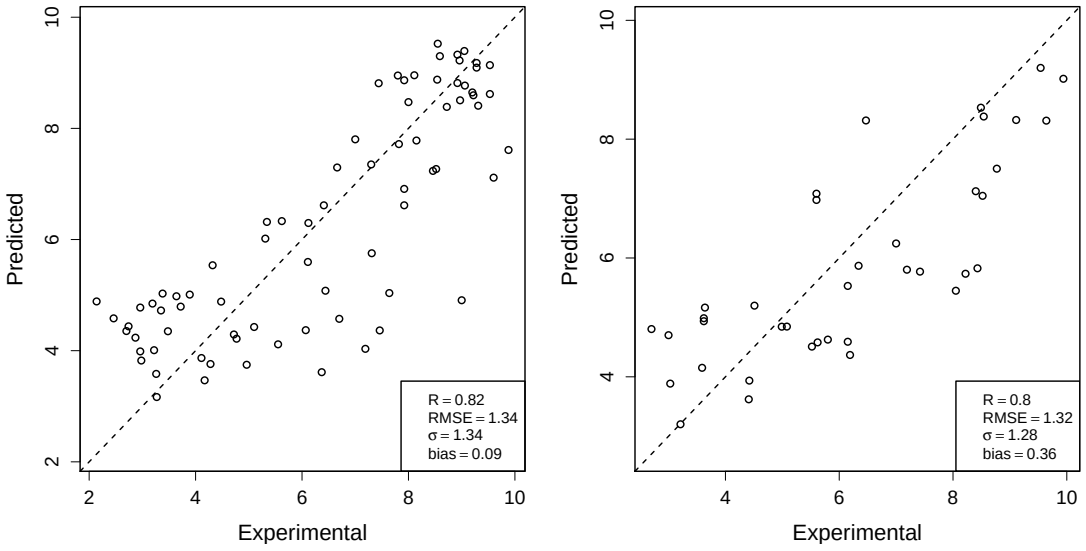
[a] Grid Spacing; [b] Oxygen Distribution; [c] Hydrogen Distribution; [d] Solvent Free Energy Distribution.

TABLE 7.4: Best test set predictive accuracy statistics (r^2) for the pIC_{50} data sets compared to CoMFA and best literature model.

	r^2	Grid Spacing (Å)	RMSE[a]	Descriptor
ACE				
CoMFA	0.490	2.0	1.520	-
CoMSIA Basic	0.520	2.0	1.460	-
PLS	0.638	2.5	1.325	$g_{C-}(r)$
GA-PLS	0.615	0.5	1.366	$g_{C+}(r)$
RF	0.636	1.0	1.304	$g_{C-}(r)$
AchE				
CoMFA	0.470	2.0	0.937	-
PLS	0.665	1.0	0.791	$g_{C+}(r)$
GA-PLS	0.697	0.5	0.761	$g_{C+}(r)$
RF	0.537	2.0	0.918	SFED
BZR				
CoMFA	0.000	2.0	0.960	-
2.5D	0.200	2.0	0.861	-
PLS	0.209	1.5	0.878	$g_{C-}(r)$
GA-PLS	0.208	2.0	0.848	$g_{H}(r)$
RF	0.217	3.0	0.863	$g_{C-}(r)$
COX2				
CoMFA	0.290	2.0	1.233	-
CoMSIA Extra	0.370	2.0	1.164	-
PLS	0.382	3.0	1.159	$g_{C+}(r)$
GA-PLS	0.351	0.5	1.211	$g_{C-}(r)$
RF	0.375	3.0	1.252	$g_{C+}(r)$
DHFR				
CoMFA	0.590	2.0	0.886	-
HQSAR	0.630	2.0	0.837	-
PLS	0.562	3.0	0.913	$g_{C+}(r)$
GA-PLS	0.567	0.5	0.913	$g_{C-}(r)$
RF	0.652	3.0	0.837	$g_{O}(r)$

[a] For literature results this has been recalculated from the standard error of prediction (s) reported by Sutherland *et al.*[124] as: RMSE = $\sqrt{((s^2)(N - 1/N)}$.

ACE Dataset. The ACE dataset comprises pIC_{50} data for 114 inhibitors of angiotensin converting enzyme separated into a training dataset of 76 and a test dataset of 38 molecules. The pIC_{50} values range between 2.1 and 9.9. Inspection of the data in Tables 7.2 and 7.3 show that the CARMa models are relatively insensitive to the choice of 3D-RISM field or grid-spacing for this dataset. The most accurate predictions were obtained using either PLS or RF regression on $g_{C-}(r)$ variables. For the external test set, the RF model has a slightly smaller error ($RMSE = 1.304$) than the PLS model ($RMSE = 1.325$), but both methods report $R^2 = 0.64$ (2 decimal places). The correlation between experimental and predicted pIC_{50} data for the PLS model is illustrated Figure 7.5. By comparison, the most accurate predictions reported by Sutherland et al. were obtained using CoMSIA ($R^2 = 0.520, RMSE = 1.46$), which was found to be slightly more accurate than CoMFA ($R^2 = 0.490, RMSE = 1.52$).



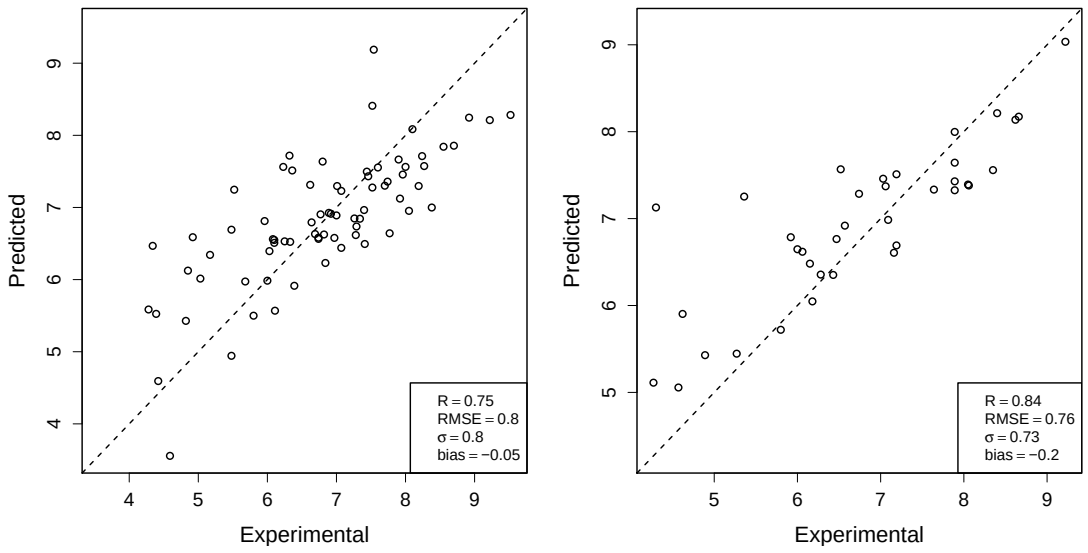
(A) q^2 (B) r^2

FIGURE 7.5: ACE correlation graphs of leave-one-out cross-validation (LOO-CV) and test set predictive accuracy for the CARMa PLS model using the $g_{C-}(r)$ probe atom distribution descriptor at 2.5 Å grid spacing.

Using a genetic algorithm (GA) to select input variables for the PLS method leads to a high q^2 for cross-validation, which is not surprising given that the GA fitness function was $RMSE$ for 3-fold cross-validation, but these models do not generalise

as well as the PLS or RF models; the best GA-PLS prediction of the test set is $R^2 = 0.615$ and $RMSE = 1.366$.

AchE Dataset. The pIC_{50} values for the 111 acetylcholinesterase inhibitors in the AchE dataset range from 4.3 to 9.5. Sutherland et al. found CoMFA to be more accurate than other QSAR methods for modelling this dataset ($R^2 = 0.47$ and $RMSE = 0.937$). Tables 7.3 and 7.4 show that an improvement in accuracy can be made by replacing CoMFA's electric/steric fields with $g_{C+}(r)$ variables giving $R^2 = 0.665$ and $RMSE = 0.791$ for PLS regression. Using a GA to select input variables for PLS further improves the accuracy for most 3D-RISM fields and grid-spacings. The best CARMa model was obtained with GA-PLS regression on $g_{C+}(r)$ variables giving $R^2 = 0.697$ and $RMSE = 0.761$ (Table 7.4). The correlation diagrams for this model are presented in Figure 7.6.

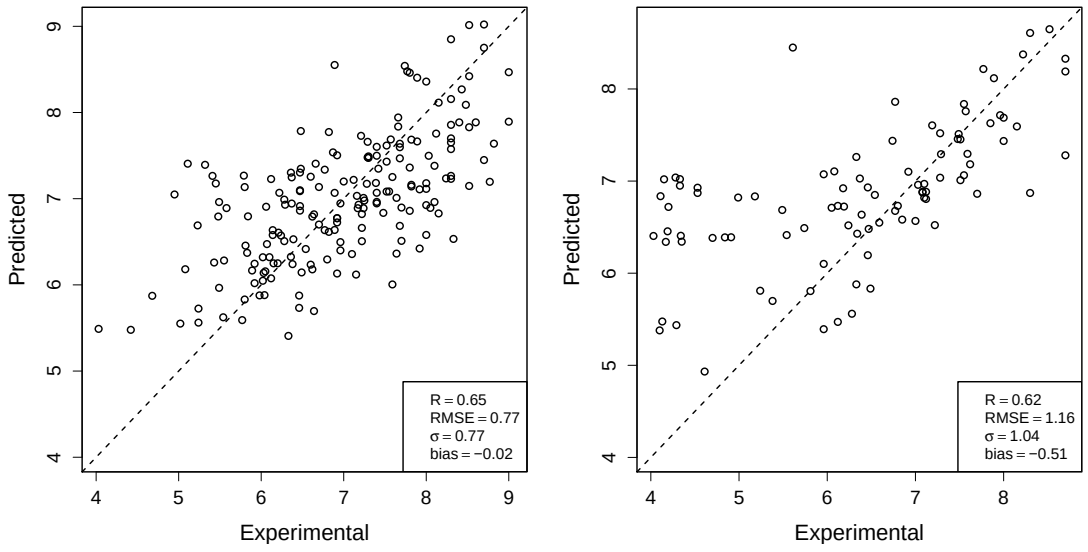


(A) q^2 (B) r^2

FIGURE 7.6: AchE correlation graphs of leave-one-out cross-validation (LOO-CV) and test set predictive accuracy for the CARMa GA-PLS model using the $g_{C+}(r)$ probe atom distribution descriptor at 0.5 Å grid spacing.

BZR and COX2 datasets. The BZR and COX2 data have proven to be almost impossible to model accurately using QSAR methods. Sutherland et al. reported

$R^2 = 0$ and $R^2 = 0.29$ for CoMFA predictions of the BZR and COX2 test sets, respectively. The best results were $R^2 = 0.200$ and $RMSE = 0.861$ for a "2.5D" QSAR model of the BZR data and $R^2 = 0.370$ and $RMSE = 1.164$ for a CoMSIA Extra model of the COX2 data; both of these models were considered to be too poor to be particularly useful. As would be expected, the CARMa method is also not able to produce very accurate models for these datasets, but in both cases it improves on the CoMFA results and matches or improves upon the other predictions. For the BZR dataset, a CARMa model using $g_{C-}(r)$ variables and RF regression gives $R^2 = 0.217$ and $RMSE = 0.863$, while for the COX2 dataset a PLS model trained on $g_{C+}(r)$ variables gives $R^2 = 0.217$ and $RMSE = 1.159$.



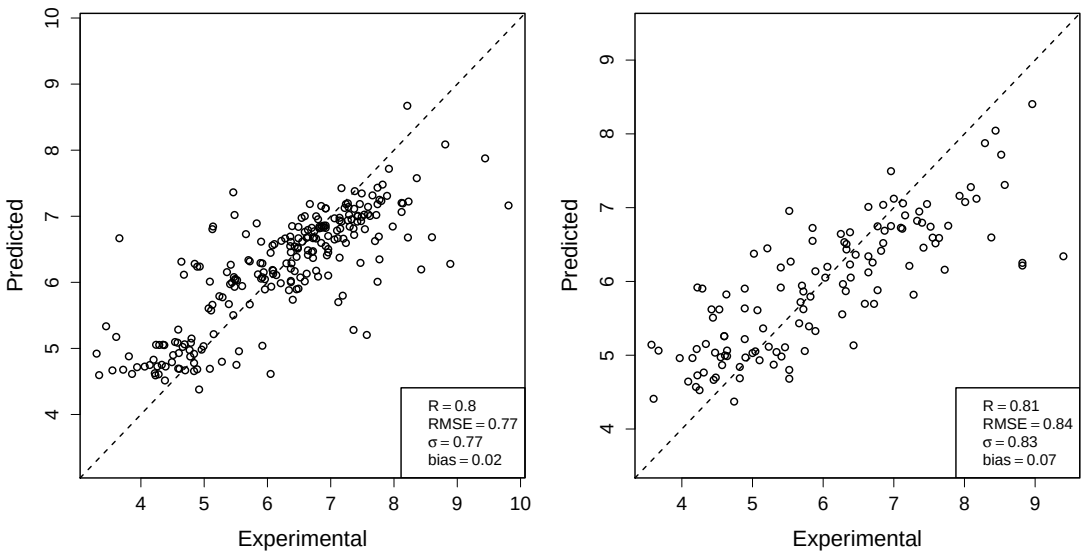
(A) q^2 (B) r^2

FIGURE 7.7: COX2 correlation graphs of leave-one-out cross-validation (LOO-CV) and test set predictive accuracy for the CARMa PLS model using the $g_{C+}(r)$ probe atom distribution descriptor at 3.0 Å grid spacing.

For the COX2 dataset, part of the reason for the poor test set prediction is that the training and test sets cover different ranges of property space. The correlation diagram for the PLS model on $g_{C+}(r)$ variables is given in Figure 7.7a. There are only three compounds with pIC_{50} values below 5 in the training set, whereas in the test set there are 19 compounds fitting this criteria. The structures of these compounds, although from the same family, do not show a stand out chemical sub

structure that can explain the poor predictions below a pIC_{50} value of 5. Figure 7.7b shows that compounds with pIC_{50} values above 5 are relatively well predicted, with the exception of one or two outliers, but the 19 compounds with pIC_{50} values below 5 have all been overestimated.

DHFR Dataset. A CoMFA model of the DHFR data has previously been reported to give a $R^2 = 0.590$ and $RMSE = 0.886$, while the HQSAR produces an improved result $R^2 = 0.630$ and $RMSE = 0.837$. The best CARMa model is found using the RF method and $g_O(r)$ variables at 3.0 Å grid-spacing, which has $R^2 = 0.652$ and $RMSE = 0.837$.



(A) q^2 (B) r^2

FIGURE 7.8: DHFR correlation graphs of leave-one-out cross-validation (LOO-CV) and test set predictive accuracy for the CARMa RF model using the $g_O(r)$ distribution descriptor at 3.0 Å grid spacing.

In Table 7.4, CARMa is shown to improve R^2 in comparison to CoMFA by 6.2% when the RF method is used with the $g_O(r)$ descriptor. In fact, the RF method produces the best result for all five descriptors tested here. The poorest results are obtained from the SFED descriptors as shown in Table 7.3. The PLS and GA-PLS methods produce results comparable to the literature when used with

$g_O(r)$ and $g_H(r)$ descriptors, but improved results when used with the probe atom descriptors. Figure 7.8a shows the correlation diagram for cross-validation of the DHFR training data using the best RF model. It is apparent that the models do not make very accurate predictions for molecules with pIC_{50} values above 8, which may partly be because they are under-represented in the dataset.

7.4 Discussion

The predictive accuracy of CARMa using various parameters and descriptors has been examined using the steroid dataset defined by Coates in 1988 and the five largest data sets reported by Sutherland *et al* in 2004.[124,140] The physiochemical properties of compounds were encoded using 3D-RISM calculations for application in field-based QSAR. The 3D-RISM calculations provided oxygen ($g_O(r)$), hydrogen ($g_H(r)$), SFED, $g_{C-}(r)$ probe atom and $g_{C+}(r)$ probe atom distribution functions to be used as fields. The models were implemented using PLS, GA-PLS and RF regression.

Two different approaches were taken to asses the predictive model accuracy. First, the cross-validation (CV) was examined: a measure of a models ability to generalise for a given group of compounds (training set). Second, the ability to extrapolate the generalisation was assessed by measuring the accuracy of predictions for test set compounds.

7.4.1 Steroids

For the steroid dataset, only the CV is assessed as the dataset is considered too small for any model to extrapolate any generalisation well. The GA-PLS is also

omitted as the CV here is representative of a fitness function rather than a measure of generalisation, therefore is not comparable to other methods or the literature.

From the CV of the PLS models, it emerges that all methods perform well and are better than CoMFA. With the exception of the SFED, all other descriptors clearly perform better than CoMFA, Table 7.1. The results suggest that there is no preferred grid spacing that is obvious which is not unexpected as the coarser grids tend to contain much of the information from the finer grids. Overall, we find the CARMa models have performed better than the best CoMFA model for the steroids dataset.

7.4.2 pIC_{50} Data Sets

For the pIC_{50} data sets both aspects of predictive model accuracy are examined. Here, the GA-PLS CV is included as the test sets are used to calculate predictive accuracy statistics, providing a validation for the GA-PLS models fitness function.

Comparing the predictive accuracy for the CARMa models as shown in Tables 7.3 and 7.4, suggests that CARMa has outperformed the best literature results in all 5 data sets. The RF models seem to generate the good results consistently and some other trends have been noted.

(1) The probe atom distribution descriptors consistently produce very good results. The best predictive model derived from the probe atom descriptors performs substantially better that the best literature model for the ACE, AchE, BZR and COX2 data sets. In the case of DHFR the best model derived from the probe atom descriptors generates the same result as the literature best (HQSAR).

(2) In comparison to CoMFA, the CARMa results are a significant improvement in all 5 data sets. With the exception of the DHFR dataset, the best CARMa

PLS and GA-PLS models are better than CoMFA. The best RF model for each dataset, outperforms CoMFA significantly.

(3) The SFED descriptor tends to perform poorly in comparison to other CARMa models with the exception of the AchE dataset, for which the RF method works best with the SFED descriptors. In most cases the SFED descriptor models are outperformed by $g_O(r)$ and $g_H(r)$ descriptor models.

(4) The $g_O(r)$ and $g_H(r)$ models tend to generate similar results. This is not unexpected as oxygen and hydrogen atoms are covalently bonded in water molecules, resulting in similar information being captured in the distribution functions.

(5) The results show that the training/test set split can create a biased generalisation when training models. This can be happen if an adequate sampling of the full range of activities is not done in the training set and results in poorer predictive models. For example, in the BZR, COX2 and DHFR data sets a small portion of the activity range is not adequately accounted for in the training sets. This has resulted in poor predictions for compounds in the test set with activities within this range for all three data sets. However, it is noted that for the activity ranges that have been adequately sampled, the prediction of compounds within this range are done well.

(6) The RF models show the most consistency across the range of grid spacings. The PLS method is consistent within the grid spacing range tested here but in some cases can be thrown off, shown by the AchE dataset using the $g_H(r)$ descriptors. The GA-PLS method is highly inconsistent across the range of spacings and can go from being a very good model to being a very poor one by changing the grid spacing. This is seen in the AchE dataset using $g_{C+}(r)$ probe descriptors, where the 0.5 Å GA-PLS model has a predictive accuracy of 0.697, and the 3.0 Å GA-PLS model gives a poor score of 0.269.

(7) For the PLS and RF models there are no real trends that determine the best grid spacing. However, for the GA-PLS model with the exception of SFED descriptors, the trend seems to be that the finer grid spacings are favoured. For example, when using the $g_O(r)$ descriptors with GA-PLS the best result is obtained using 0.5 Å spacing for all 5 data sets.

This study has compared q^2 and r^2 statistics and determined that most models have been overfitted to some extent ($q^2 > r^2$). For the PLS and RF there seems to be a reasonable correspondence between cross-validation and predictive accuracy with negligible overfitting but for the GA-PLS models overfitting seems to be problematic. The q^2 statistics of the GA-PLS models are quite high and suggests that they are the best predictive models, Table 7.2. However, a look at the r^2 statistics reveal that the GA-PLS method tends to be outperformed by the RF and PLS regularly, Tables 7.3 and 7.4. This reinforces literature reports that measuring q^2 alone is not enough to determine the predictive accuracy of a QSAR model.[344]

7.5 Conclusions

In summary, this investigation has examined the predictive accuracy of numerous CARMa models applied to 5 pIC_{50} data sets and compared the results against the best models reported in the literature. The results conclusively show that CARMa is the best method available for predictive accuracy for all the data sets used. The CARMa regression models are able to extract structure-activity relationship (SAR) trends from a training set and extrapolate them over a test set relatively accurately. The RF method produce the best predictive models in comparison to PLS and GA-PLS, and although no grid spacing could be determined as the optimal, it can be concluded that for RF models the differences are negligible.

The best descriptors are the probe atom distributions, resulting in better predictive models for PLS, GA-PLS and RF in comparison to models using $g_O(r)$, $g_H(r)$ and SFED descriptors. These finding suggest CARMa (3D-RISM) captures more information about the field that is required for predictive models in comparison to other field-based QSAR methods.

While there are discrepancies in usefulness of the predictive models from dataset to dataset, this study has demonstrated that CARMa is an extremely robust QSAR method and will be given serious consideration for applications by QSAR practitioners. As reported in the literature, a model that works sufficiently for one data set may not work very well for others and it is fair to say many models are unlikely to be good for all data sets available.[124] However, the benchmarking done as a part of this research has demonstrated that CARMa works extremely well for the 5 pIC_{50} data sets employed here.

A number of changes are likely to improve the predictive accuracy of the CARMa models. This investigation has shown the discrepancies in the training/test set split which is believed to have adversely affected the predictive models for the BZR, COX2 and DHFR data sets. Further work would include, but not be limited to; investigating different 3D-RISM parameter such as forcefields, partial charges, solvent model and bridge functionals; calculate other solvent probe descriptors and improving the structural alignment.

Chapter 8

Conclusions

This thesis has investigated and developed various computational techniques used to study protein-ligand binding with an emphasis on the aspartic protease, chymosin, and approaches using the three-dimensional reference interaction site model (3D-RISM). Three computational investigations have been presented in this thesis and the key findings from each investigation are summarised in the sections that follow.

8.1 Allosteric-Activation Mechanism Of Bovine Chymosin

The conformational change that occurs in the allosteric-activation mechanism of bovine chymosin has been observed in both regular MD and BEMD simulations. This is in agreement with literature kinetic, mutagenesis and crystallographic experiments, showing the HPHPH (P8-P4) residue sequence of bovine κ-casein initiates the conformational change in the side chain of Tyr77 and the β-hairpin region of bovine chymosin. The investigation has led to the the proposal of a new allosteric-activation mechanism that occurs via the following steps: (i) the P8-P4 κ-casein fragment binds with chymosin and disrupts the hydrogen bonding network that stabilises the self-inhibiting pose of Tyr77 Figure 5.5; (ii) the P8-P4

κ-casein peptide interacts with the short α-helix in residues 112-116 of chymosin, which both allows the β-hairpin flap in residues 72 to 84 of chymosin to twist, and also causes the side chain of Phe114 to vacate the pocket that is occupied by Tyr77 in the open conformation; (iii) as Phe114 moves, Tyr77 simultaneously changes conformation from self-inhibiting to open and is stabilised by a hydrogen bonding network below the β-hairpin flap.

Further work is warranted as the investigation has highlighted multiple related pathways for the proposed activation mechanism. The same experiment can be conducted on other chymosin–κ-casein complexes such as the bovine-camel, camel-bovine and camel-camel (chymosin variant is named first and κ-casein second). Longer chains of κ-casein can be investigated to asses the impact it would have on the allostery and if the additional residues create a bias for a particular pathway.

8.2 Effect of Mutations in Bovine or Camel Chymosin on the Thermodynamics of Binding κ-Caseins

Through the use of molecular dynamics simulations and free energy calculations, binding in four different chymosin–κ-casein complexes (Bov/Bov, Bov/Cam, Cam/Bov, Cam/Cam) has been investigated. By way of computational alanine scanning calculations, the influence that differences in the primary sequence of bovine and camel chymosin ("natural mutations") have on the binding thermodynamics in these complexes has been identified. Of the 12 sequence differences in the binding sites of bovine and camel chymosin, eight are shown to be particularly important for understanding differences in the binding thermodynamics (Asp112Glu, Lys221Val, Gln242Arg, Gln278Lys. Glu290Asp, His292Asn, Gln294Glu, and Lys295Leu).

For Gln242Arg and Gln278Lys, the natural residue in bovine chymosin is energetically more favourable for binding. In the camel chymosin systems the alanine mutations are energetically favoured suggesting the polar positive residues in camel chymosin adversely influence the binding thermodynamics with κ-casein. For residues Asp112Glu, Lys221Val and Lys295Leu, the native camel variant is most favoured. All of these residues occupy separate and predominantly non-polar pockets along the binding cleft where the natural polar positive residues in bovine chymosin adversely influence the binding thermodynamics.

The research conducted in this chapter has identified a number of residues in variants of chymosin that are particularly important for influencing protein-ligand binding thermodynamics for chymosin–κ-casein complexes. Further investigations for this study would include the effect of single-point mutations on natural variants in the substrate as well as the effect of multiple point mutations for the complexes.

8.3 Comparative Molecular Field Analysis using Molecular Integral Equation Theory

The results show that CARMa is the best method available for predictive accuracy for all the data sets used. The CARMa regression models are able to extract structure-activity relationships (SAR) trends from a training set and extrapolate them over a test set relatively accurately. The results show that the RF method produces the best predictive models in comparison to PLS and GA-PLS. Although no grid spacing could be determined as the optimal, it can be concluded that for RF models the differences between the various grid spacings are negligible. The best descriptors are the probe atom distributions, resulting in better predictive models for PLS, GA-PLS and RF in comparison to models using $g_O(r)$, $g_H(r)$ and SFED descriptors. These findings show that CARMa (3D-RISM) captures

more information about the field that is required for predictive models in comparison than other field-based QSAR methods.As QSAR methods are increasingly being applied in the early stages of drug discovery to identify high quality leads/ligands, PCARMa has demonstrated that it can outperform commonly used QSAR methods consistently and should be given serious consideration for applications by QSAR practitioners.

Further work could include investigating different 3D-RISM parameters (e.g. forcefields, partial charges, solvent models and bridge functionals), applying other solvent probe descriptors, or improving the structural alignment.

Part IV

Appendices

Appendix A

Allosteric-Activation Mechanism Of Bovine Chymosin

A.1 Tyr77 and Phe114 interactions

A.1.1 Apo-Chymosin

In the simulations of apo-chymosin, Tyr77 makes close contacts with two residues in the α-helix, Val113 and Phe114. Phe114 extends towards the β-hairpin flap of chymosin when Tyr77 is in its self-inhibited state and extends away from the flap when Tyr77 is in its open conformation. In the *A1* simulation, in which Tyr77 is in the open conformation, contact with Phe114 occurs for 55% of the trajectory (using a distance of 4 Å between any non-hydrogen atoms to define a contact in the AMBER ff99SB-ILDN simulation). This increases to 94% in simulation *B1* where Tyr77 is in the self-inhibited conformation.

A.1.2 Chymosin – P8-P4 κ-Casein

In simulation *C1* using AMBER ff99SB-ILDN, Tyr77 remains in contact with the Phe114 in both its selfinhibited (73%) and open conformations (78%). There is an increase in atom-atom contacts between Tyr77.N:Phe114.CZ (15% $\rightarrow$ 36%) and Tyr77.H:Phe114.CE2 (6% $\rightarrow$ 11%) after the self-inhibited to open conformation. Phe114 makes fewer contacts with Trp41 (which is located under the β-hairpin flap in the binding pocket) when Tyr77 is in its open conformation (37% $\rightarrow$ 17%).

In simulation *C2* using the AMBER03 force field, Tyr77 remains in contact with residue Phe114 in both the self-inhibited (67%) and open (74%) conformations. Close contacts are shown to increase for the same atoms as in the previous simulations when Tyr77 moves to its open state, Tyr77.N:Phe114.CZ (11% $\rightarrow$ 39%) and Tyr77.H:Phe114.CE2 (5% $\rightarrow$ 24%). Contact between Trp41 and Phe114 is also seen to decrease when Tyr77 is in its open state (66% $\rightarrow$ 32%).

Simulation *C3* using the AMBER03 force field contains a short 17ns period where Tyr77 changes to its open form before returning to the self-inhibiting state. Residue contacts analysis reveals Tyr77 and Phe114 remains in contact in both the selfinhibited (84%) and open (80%) conformations. Atom contacts follow the same trend as seen in the simulations described above with an increase when Tyr77 is in its open conformation for Tyr77.N:Phe114.CZ (11% $\rightarrow$ 39%) and Tyr77.H:Phe114.CE2 (5% $\rightarrow$ 24%). However analysis of the residue contact between Trp41 and Phe114 reveal that there is an increase when Tyr77 is in its open conformation (selfinhibited-35% $\rightarrow$ open-51%).

A.2 Hydrogen-Bonding

TABLE A.1: Percentage of total simulation time in which specific hydrogen bonds were observed in the Bias-Exchange Metadynamics simulations

		Apo Enzyme			Olo Enzyme		
Donor	Acceptor	Closed	Intermediate	Open	Closed	Intermediate	Open
ASN10	GLY161	63	57	94	1	2	2
ASN10	ASP158	9	17	96	66	26	53
SER164	ASN10	74	59	3	93	61	93
SER220	ASP13	48	54	93	0	0	0
ARG304	ASP13	37	33	95	68	18	50
GLU118	GLN15	97	97	73	0	4	13
GLN15	GLY218	0	0	0	74	63	70
TYR16	TYR156	64	82	92	0	0	0

Appendix B

Effect of Mutations in Bovine or Camel Chymosin on the Thermodynamics of Binding κ-Caseins

B.1 Binding free energies

Binding free energies calculated using the GF, PSE-3 and PC functionals in the scope of MM3DRISM are presented in Tables B.1, B.2 and B.3, respectively, and plotted in Figure B.1. The GF free energy functional is well known to give wildly inaccurate estimates of hydrogen free energies for anything but simple model solutes. As was expected, therefore, the binding free energies computed using the GF free energy functional were very different from those computed by the other functionals. Extensive previous benchmarking on solvation free energy data of organic molecules indicates that the PC+ functional gives more accurate results than the GF, PSE-3 or PC functionals, which is why only the PC+ binding free energy data were considered in the manuscript.

TABLE B.1: Binding free energy results using the GF functional.

Energy	GF							
	CAM-CAM		CAM-BOV		BOV-CAM		BOV-BOV	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE
ΔG_{gas}	-1330.1	9.8	-1029.6	6.6	-1595.4	7.2	-1388.0	9.9
ΔG_{hyd}	1381.6	11.0	1088.1	8.1	1663.1	8.0	1454.3	10.7
ΔG_{total}	51.5	3.7	58.4	3.8	67.7	3.1	66.3	2.7
$T\Delta S$	-76.5	1.5	-61.9	1.3	-68.1	1.4	-70.2	1.1
ΔG_{bind}	128.0	4.0	120.3	4.0	135.8	3.4	136.5	2.9

TABLE B.2: Binding free energy results using the PSE-3 functional.

Energy	PSE-3							
	CAM-CAM		CAM-BOV		BOV-CAM		BOV-BOV	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE
ΔG_{gas}	-1330.1	9.8	-1029.6	6.6	-1595.4	7.2	-1388.0	9.9
ΔG_{hyd}	1246.1	9.3	966.1	6.1	1529.4	6.9	1312.6	9.2
ΔG_{total}	-84.0	1.6	-63.6	1.5	-66.0	1.3	-75.5	1.5
$T\Delta S$	-76.5	1.5	-61.9	1.3	-68.1	1.4	-70.2	1.1
ΔG_{bind}	-7.5	2.2	-1.7	2.0	2.1	1.9	-5.3	1.9

TABLE B.3: Binding free energy results using the PC functional.

Energy	PC							
	CAM-CAM		CAM-BOV		BOV-CAM		BOV-BOV	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE
ΔG_{gas}	-1330.1	9.8	-1029.6	6.6	-1595.4	7.2	-1388.0	9.9
ΔG_{hyd}	1206.2	9.1	927.1	6.1	1489.2	6.8	1271.6	9.0
ΔG_{total}	-123.9	1.4	-102.6	1.2	-106.2	1.0	-116.4	1.4
$T\Delta S$	-76.5	1.5	-61.9	1.3	-68.1	1.4	-70.2	1.1
ΔG_{bind}	-47.4	2.0	-40.7	1.8	-38.1	1.8	-46.2	1.8

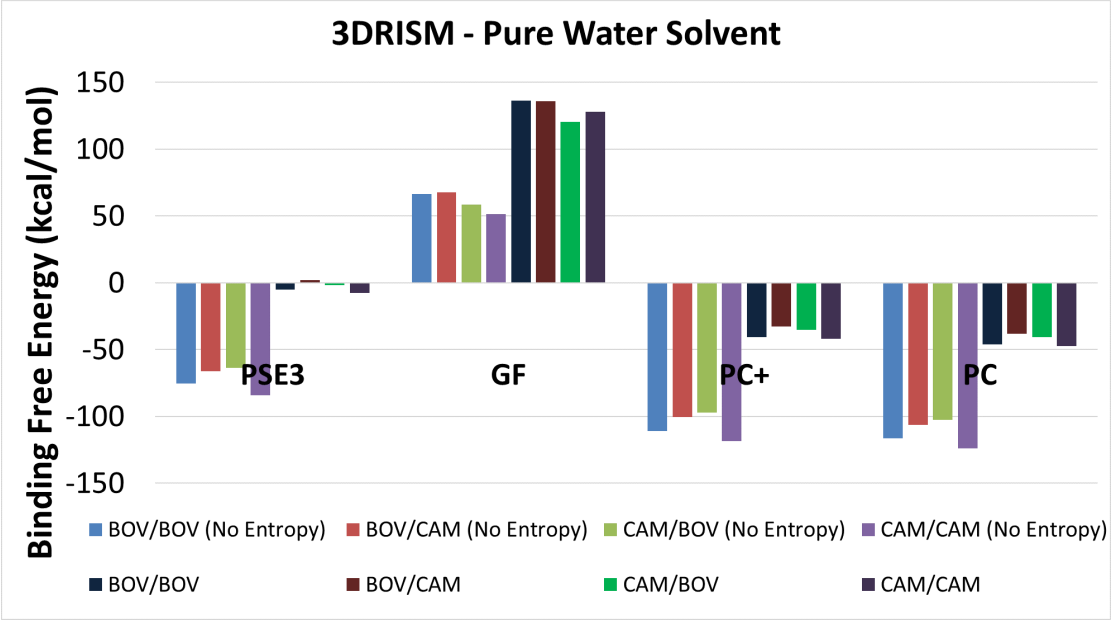


FIGURE B.1: Binding free energy results with or without entropy term, indicating the error associated with the Gaussian Fluctuation free energy functional

B.2 Additional computational alanine scanning results

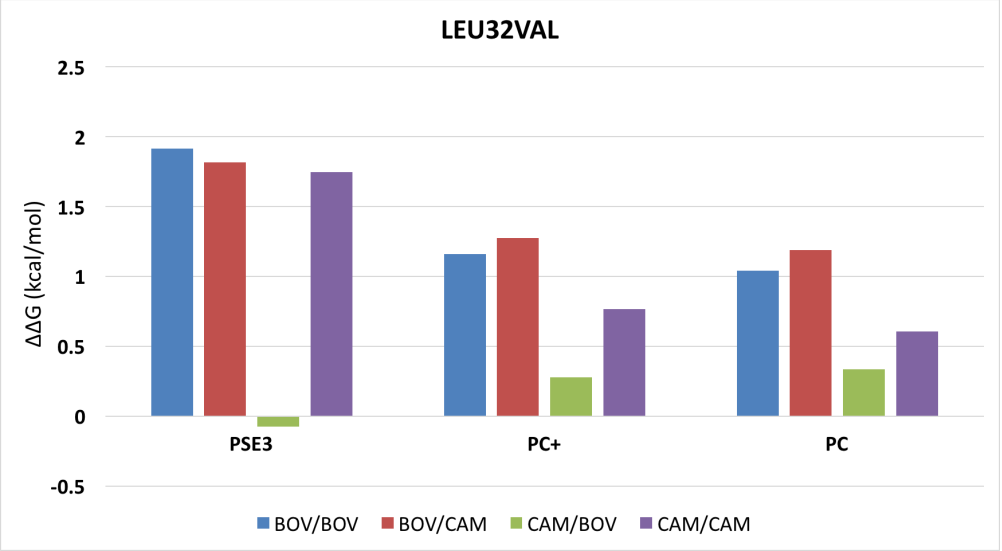


FIGURE B.2: Comparison of alanine scanning results of residue 32 (Leu32 in bovine chymosin, Val32 in camel chymosin)

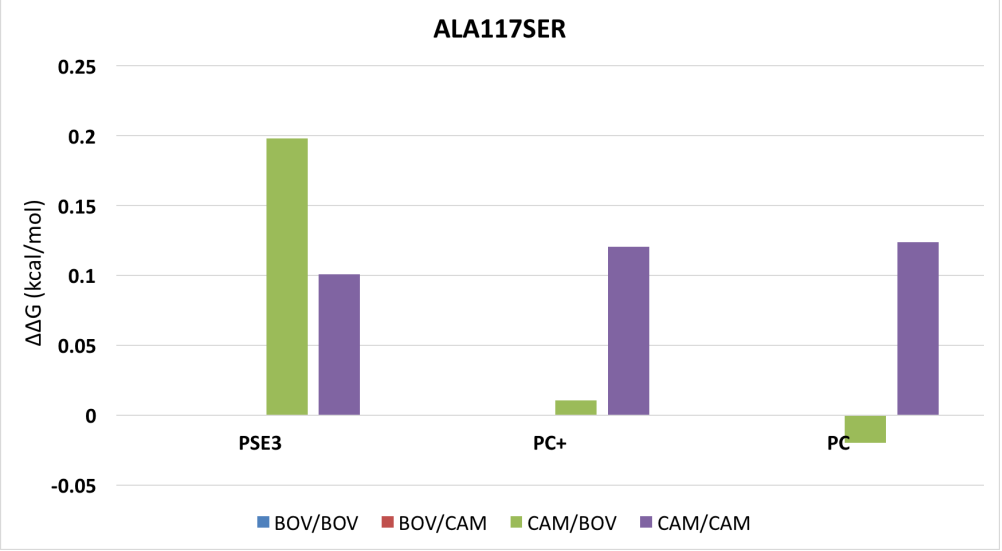


FIGURE B.3: Comparison of alanine scanning results of residue 117 (Ala117 in bovine chymosin, Ser117 in camel chymosin)

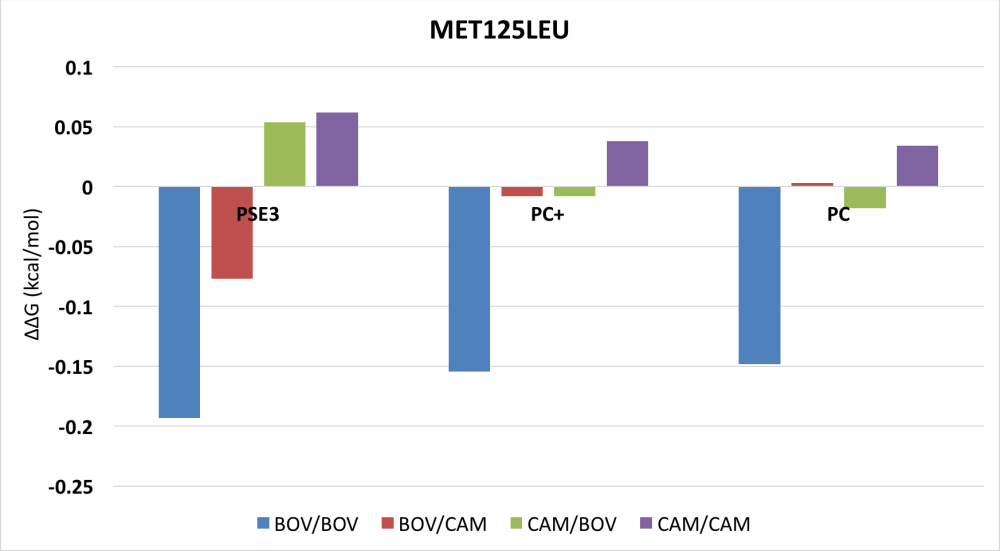


FIGURE B.4: Comparison of alanine scanning results of residue 125 (Met125 in bovine chymosin, Leu125 in camel chymosin)

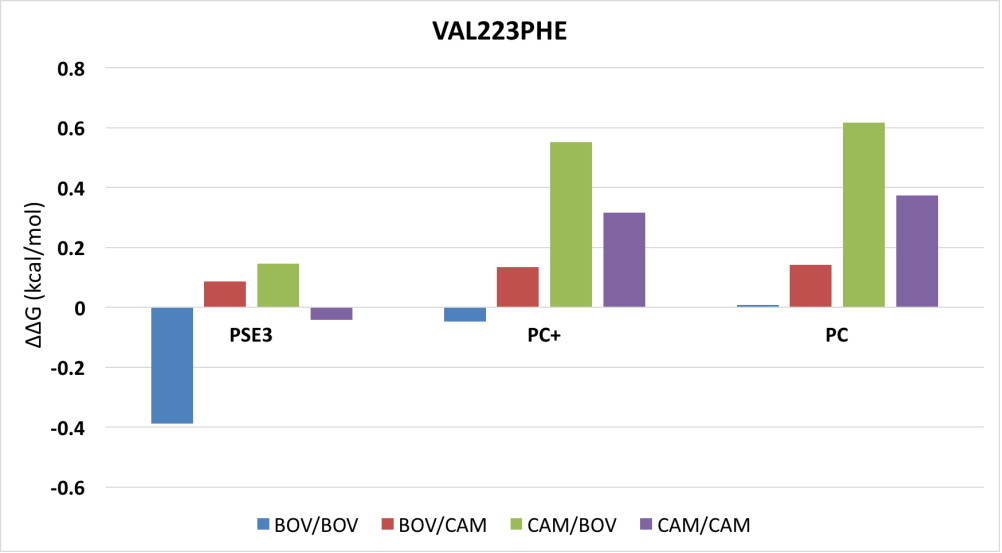


FIGURE B.5: Comparison of alanine scanning results of residue 223 (Val223 in bovine chymosin, Phe223 in camel chymosin)

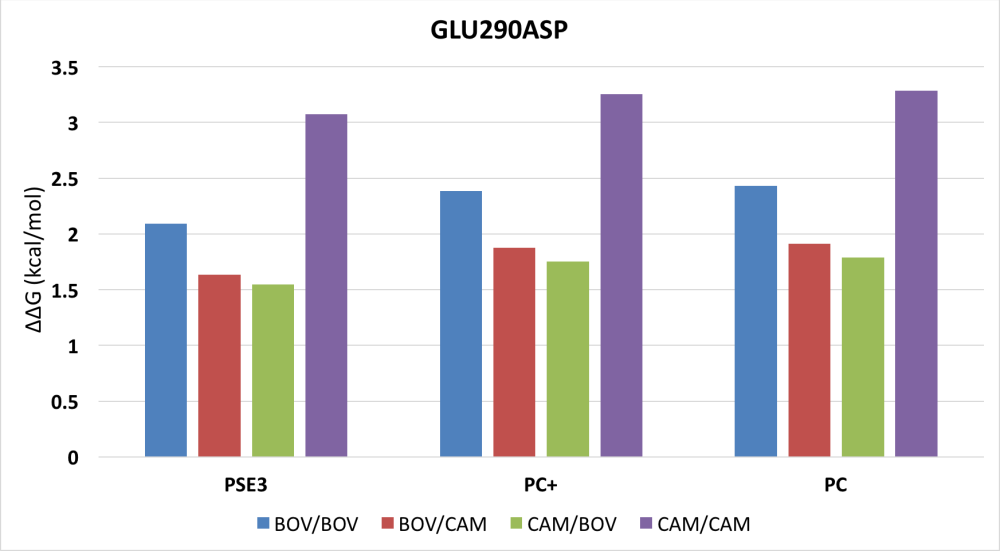


FIGURE B.6: Comparison of alanine scanning results of residue 290 (Glu290 in bovine chymosin, Asp290 in camel chymosin)

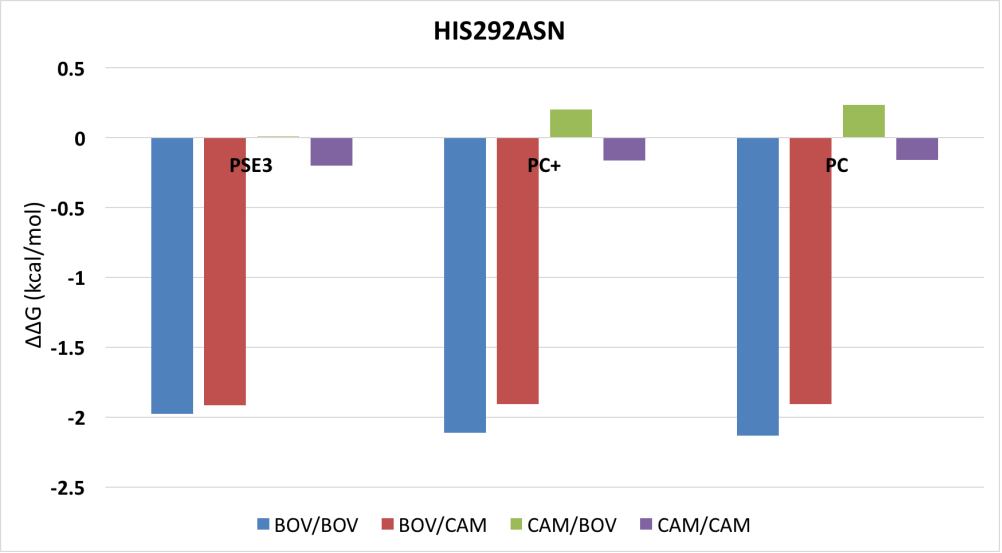


FIGURE B.7: Comparison of alanine scanning results of residue 292 (His292 in bovine chymosin, Asn292 in camel chymosin)

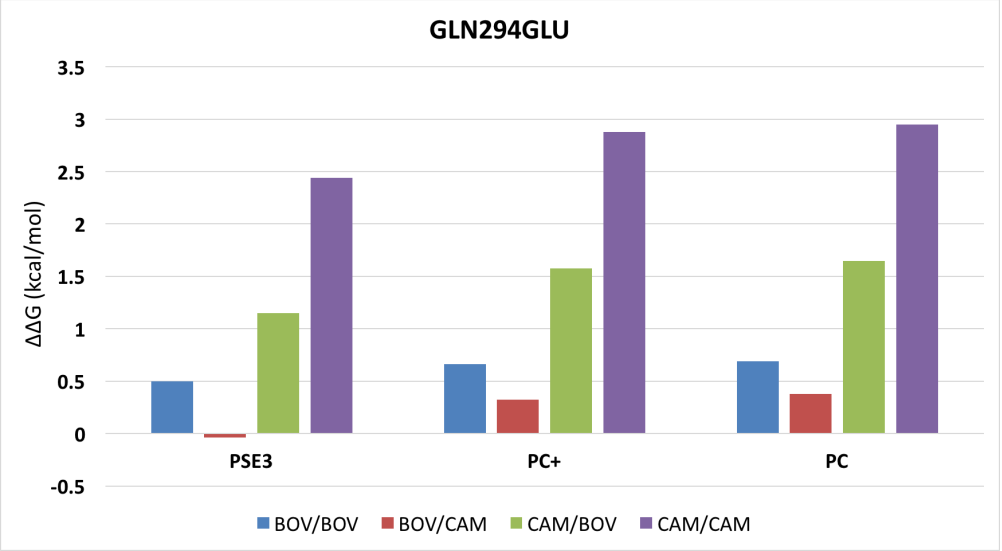


FIGURE B.8: Comparison of alanine scanning results of residue 294 (Gln294 in bovine chymosin, Glu294 in camel chymosin)

Appendix C

Comparative Molecular Field Analysis using Molecular Integral Equation Theory

C.1 Benchmarking Tables

Below are a series of tables containing the raw data for all models tested. The following key is true for all tables that follow.

[a] – All grid spacings are in Å.

[b] – Counts all descriptors that have a standard deviation of 0.

[c] – Total number of descriptors once those with a standard deviation of 0 have been removed.

[d] – Machine learning model used.

[e] – Time measurements are reported in minutes.

[f] – Root mean square error statistic (RMSE).

[g] – Standard deviation statistic (σ).

[h] – Pearson's correlation coefficient squared.

[i] – Model bias.

[j] – Cross-validation correlation statistic.

[k] – Predicted residual error sum of squares statistic (PRESS).

TABLE C.1: Benchmarking statistics for ACE dataset using $g_O(r)$ descriptors.

ACE - OXYGEN

GS[a]	σ=0[b]	TNC[c]	Method[d]	Training			Cross-Validation					Test				Time[e]
				RMSE[f]	σ[g]	r^2[h]	RMSE	σ	Bias[i]	q^2[j]	PRESS[k]	RMSE	σ	Bias	r^2	
0.5	498	84686	PLS	-	-	**-**	-	-	-	**-**	-	-	-	-	**-**	-
			GA-PLS	0.79	0.79	**0.89**	1.28	1.27	0.16	**0.70**	123.91	1.51	1.44	0.47	**0.54**	31
			RF	-	-	**-**	-	-	-	**-**	-	-	-	-	**-**	-
1.0	60	10588	PLS	0.79	0.79	**0.89**	1.40	1.39	0.19	**0.64**	149.86	1.49	1.41	0.49	**0.56**	6
			GA-PLS	0.75	0.75	**0.90**	1.15	1.14	0.14	**0.76**	100.42	1.50	1.46	0.36	**0.53**	3
			RF	0.51	0.51	**0.97**	1.14	1.31	0.12	**0.68**	131.93	1.34	1.34	-0.04	**0.62**	582
1.5	15	3360	PLS	0.79	0.79	**0.89**	1.44	1.43	0.19	**0.62**	157.59	1.54	1.45	0.52	**0.53**	1
			GA-PLS	0.85	0.85	**0.87**	1.10	1.10	0.06	**0.78**	91.65	1.63	1.56	0.47	**0.46**	1
			RF	0.51	0.51	**0.97**	1.29	1.29	0.10	**0.70**	126.44	1.35	1.35	0.02	**0.61**	100
2.0	6	1325	PLS	0.77	0.77	**0.89**	1.46	1.43	0.18	**0.62**	158.65	1.54	1.47	0.45	**0.52**	1
			GA-PLS	0.97	0.97	**0.83**	1.19	1.19	0.00	**0.74**	106.78	2.33	2.31	0.34	**0.16**	1
			RF	0.53	0.53	**0.96**	1.33	1.32	0.11	**0.68**	133.45	1.38	1.37	-0.06	**0.59**	33
2.5	3	726	PLS	0.84	0.84	**0.87**	1.42	1.41	0.16	**0.63**	152.10	1.50	1.43	0.45	**0.54**	1
			GA-PLS	1.20	1.20	**0.74**	1.34	1.34	0.01	**0.67**	136.39	1.65	1.58	0.45	**0.48**	1
			RF	0.51	0.51	**0.97**	1.29	1.28	0.09	**0.70**	125.95	1.38	1.38	-0.01	**0.58**	19
3.0	2	510	PLS	0.83	0.83	**0.87**	1.42	1.41	0.18	**0.63**	152.61	1.56	1.49	0.47	**0.51**	1
			GA-PLS	1.12	1.12	**0.77**	1.31	1.31	0.02	**0.69**	131.05	1.80	1.67	0.69	**0.40**	1
			RF	0.51	0.51	**0.97**	1.29	1.28	0.15	**0.70**	125.58	1.38	1.38	0.04	**0.60**	2

TABLE C.2: Benchmarking statistics for ACE dataset using Solvation Free Energy Density (SFED) descriptors.

ACE - SFED

GS[a]	σ=0[b]	TNC[c]	Method[d]	Training			Cross-Validation					Test				Time[e]
				RMSE[f]	σ[g]	r^2[h]	RMSE	σ	Bias[i]	q^2[j]	PRESS[k]	RMSE	σ	Bias	r^2	
0.5	3	85181	PLS	-	-	**-**	-	-	-	**-**	-	-	-	-	**-**	-
			GA-PLS	1.11	1.11	**0.78**	1.35	1.35	0.01	**0.67**	139.06	1.62	1.61	0.24	**0.47**	84
			RF	-	-	**-**	-	-	-	**-**	-	-	-	-	**-**	-
1.0	1	10647	PLS	1.14	1.14	**0.76**	1.44	1.44	0.01	**0.62**	157.32	1.68	1.65	0.28	**0.45**	31
			GA-PLS	1.10	1.10	**0.78**	1.29	1.29	0.02	**0.70**	126.72	1.62	1.59	0.35	**0.47**	10
			RF	0.52	0.52	**0.96**	1.32	1.32	0.01	**0.68**	132.32	1.32	1.31	-0.07	**0.62**	600
1.5	1	3374	PLS	1.14	1.14	**0.76**	1.44	1.44	0.01	**0.62**	158.08	1.69	1.67	0.28	**0.44**	2
			GA-PLS	1.00	1.00	**0.82**	1.21	1.21	0.04	**0.73**	112.10	1.46	1.41	0.40	**0.56**	5
			RF	0.51	0.51	**0.96**	1.32	1.32	-0.02	**0.68**	131.42	1.34	1.34	-0.01	**0.60**	94
2.0	0	1331	PLS	1.15	1.15	**0.76**	1.44	1.44	0.02	**0.62**	158.58	1.68	1.65	0.00	**0.45**	1
			GA-PLS	1.09	1.09	**0.78**	1.21	1.21	0.01	**0.73**	111.40	1.55	1.54	0.14	**0.49**	2
			RF	0.52	0.52	**0.96**	1.31	1.31	-0.01	**0.69**	130.22	1.38	1.38	-0.01	**0.58**	33
2.5	0	729	PLS	1.14	1.14	**0.76**	1.41	1.41	0.02	**0.64**	150.46	1.66	1.63	0.31	**0.46**	1
			GA-PLS	1.03	1.03	**0.81**	1.19	1.19	0.02	**0.74**	107.03	1.55	1.53	0.24	**0.51**	1
			RF	0.52	0.52	**0.96**	1.29	1.29	-0.01	**0.70**	126.41	1.38	1.38	0.04	**0.58**	19
3.0	1	511	PLS	1.16	1.16	**0.75**	1.45	1.45	0.01	**0.62**	159.55	1.70	1.68	0.25	**0.43**	1
			GA-PLS	1.21	1.21	**0.73**	1.31	1.31	0.00	**0.69**	130.90	1.79	1.71	0.55	**0.40**	1
			RF	0.52	0.52	**0.96**	1.31	1.31	0.01	**0.69**	129.93	1.33	1.33	-0.01	**0.61**	14

TABLE C.3: Benchmarking statistics for ACE dataset using $g_H(r)$ descriptors.

ACE - $g_H(r)$

GS[a]	σ=0[b]	TNC[c]	Method[d]	Training			Cross-Validation					Test				Time[e]
				RMSE[f]	σ[g]	r^2[h]	RMSE	σ	Bias[i]	q^2[j]	PRESS[k]	RMSE	σ	Bias	r^2	
0.5	57	85127	PLS	-	-	**-**	-	-	-	**-**	-	-	-	-	**-**	-
			GA-PLS	0.77	0.77	**0.89**	1.29	1.28	0.15	**0.70**	125.58	1.46	1.40	0.42	**0.57**	96
			RF	-	-	**-**	-	-	-	**-**	-	-	-	-	**-**	-
1.0	5	10643	PLS	0.82	0.82	**0.88**	1.42	1.42	0.16	**0.63**	154.22	1.46	1.39	0.46	**0.57**	23
			GA-PLS	0.76	0.76	**0.90**	1.17	1.16	0.10	**0.75**	103.70	1.61	1.57	0.38	**0.46**	5
			RF	0.50	0.50	**0.97**	1.30	1.29	0.12	**0.69**	128.33	1.37	1.37	-0.06	**0.60**	674
1.5	1	3374	PLS	0.83	0.83	**0.88**	1.44	1.43	0.17	**0.62**	157.86	1.48	1.40	0.47	**0.56**	2
			GA-PLS	0.87	0.87	**0.86**	1.11	1.11	0.03	**0.78**	92.96	1.56	1.49	0.47	**0.51**	3
			RF	0.53	0.52	**0.97**	1.33	1.33	0.14	**0.68**	134.93	1.34	1.34	0.01	**0.61**	91
2.0	2	1329	PLS	0.84	0.84	**0.87**	1.40	1.39	0.15	**0.64**	149.11	1.45	1.39	0.43	**0.57**	1
			GA-PLS	0.89	0.89	**0.86**	1.06	1.06	0.04	**0.80**	85.11	1.84	1.84	0.12	**0.37**	2
			RF	0.52	0.52	**0.96**	1.32	1.32	0.11	**0.68**	132.65	1.42	1.42	-0.04	**0.56**	22
2.5	1	728	PLS	0.87	0.87	**0.86**	1.46	1.45	0.15	**0.61**	162.57	1.55	1.45	0.53	**0.53**	1
			GA-PLS	1.09	1.09	**0.78**	1.23	1.23	0.00	**0.72**	115.59	1.64	1.62	0.25	**0.47**	2
			RF	0.52	0.52	**0.97**	1.28	1.28	0.10	**0.70**	124.39	1.36	1.36	-0.03	**0.61**	14
3.0	0	512	PLS	0.82	0.82	**0.88**	1.44	1.43	0.18	**0.62**	156.73	1.55	1.48	0.46	**0.51**	1
			GA-PLS	1.10	1.10	**0.78**	1.25	0.25	0.02	**0.72**	117.96	1.60	1.51	0.53	**0.50**	2
			RF	0.53	0.53	**0.96**	1.30	1.29	0.10	**0.69**	127.59	1.43	1.43	0.04	**0.55**	11

TABLE C.4: Benchmarking statistics for ACE dataset using $g_{C-}(r)$ descriptors.

ACE - $g_{C-}(r)$

GS[a]	σ=0[b]	TNC[c]	Method[d]	Training			Cross-Validation					Test				Time[e]
				RMSE[f]	σ[g]	r^2[h]	RMSE	σ	Bias[i]	q^2[j]	PRESS[k]	RMSE	σ	Bias	r^2	
0.5	204	84980	PLS	-	-	**-**	-	-	-	**-**	-	-	-	-	**-**	-
			GA-PLS	0.79	0.79	**0.89**	1.25	1.25	0.12	**0.71**	119.13	1.40	1.34	0.39	**0.60**	50
			RF	-	-	**-**	-	-	-	**-**	-	-	-	-	**-**	-
1.0	26	10622	PLS	0.86	0.86	**0.86**	1.39	1.38	0.11	**0.65**	146.24	1.39	1.32	0.41	**0.61**	6
			GA-PLS	0.75	0.75	**0.90**	1.17	1.16	0.07	**0.75**	103.29	1.41	1.40	0.16	**0.56**	8
			RF	0.52	0.51	**0.97**	1.33	1.33	0.09	**0.68**	135.33	1.30	1.30	-0.10	**0.64**	1212
1.5	8	3367	PLS	0.85	0.85	**0.87**	1.39	1.38	0.12	**0.65**	146.49	1.39	1.33	0.40	**0.61**	5
			GA-PLS	0.81	0.81	**0.88**	1.11	1.11	0.03	**0.77**	93.92	1.54	1.49	0.39	**0.51**	3
			RF	0.52	0.52	**0.96**	1.30	1.30	0.08	**0.69**	128.00	1.32	1.32	-0.08	**0.62**	193
2.0	3	1328	PLS	0.84	0.84	**0.87**	1.43	1.43	0.10	**0.63**	155.75	1.39	1.35	0.32	**0.60**	1
			GA-PLS	0.95	0.95	**0.84**	1.15	1.15	0.01	**0.76**	100.84	1.82	1.80	0.26	**0.35**	2
			RF	0.51	0.51	**0.96**	1.29	1.29	0.06	**0.70**	127.00	1.31	1.31	-0.13	**0.62**	69
2.5	3	726	PLS	0.84	0.84	**0.87**	1.35	1.34	0.09	**0.67**	137.46	1.33	1.28	0.36	**0.64**	1
			GA-PLS	1.00	1.00	**0.82**	1.17	1.17	0.02	**0.75**	103.26	1.66	1.65	0.15	**0.46**	1
			RF	0.51	0.51	**0.96**	1.27	1.27	0.06	**0.71**	122.10	1.30	1.30	-0.11	**0.63**	40
3.0	1	511	PLS	0.85	0.85	**0.87**	1.42	1.42	0.13	**0.63**	153.74	1.43	1.38	0.39	**0.58**	1
			GA-PLS	1.13	1.13	**0.77**	1.34	1.34	0.05	**0.67**	136.43	1.82	1.75	0.49	**0.42**	1
			RF	0.50	0.50	**0.97**	1.25	1.24	0.07	**0.72**	117.94	1.31	1.30	-0.17	**0.63**	29

TABLE C.5: Benchmarking statistics for ACE dataset using $g_{C+}(r)$ descriptors.

ACE - $g_{C+}(r)$

GS[a]	σ=0[b]	TNC[c]	Method[d]	Training			Cross-Validation					Test				Time[e]
				RMSE[f]	σ[g]	r^2[h]	RMSE	σ	Bias[i]	q^2[j]	PRESS[k]	RMSE	σ	Bias	r^2	
0.5	75	85109	PLS	-	-	**-**	-	-	-	**-**	-	-	-	-	**-**	-
			GA-PLS	0.80	0.80	**0.88**	1.26	1.26	0.13	**0.71**	120.92	1.37	1.32	0.37	**0.62**	50
			RF	-	-	**-**	-	-	-	**-**	-	-	-	-	**-**	-
1.0	8	10640	PLS	0.87	0.87	**0.86**	1.39	1.38	0.13	**0.65**	146.08	1.39	1.33	0.40	**0.61**	6
			GA-PLS	0.82	0.82	**0.88**	1.18	1.17	0.07	**0.75**	105.17	1.39	1.37	0.24	**0.58**	8
			RF	0.50	0.50	**0.97**	1.29	1.29	0.08	**0.70**	126.81	1.34	1.34	-0.05	**0.61**	1205
1.5	4	3371	PLS	0.86	0.86	**0.86**	1.37	1.36	0.13	**0.66**	142.15	1.43	1.36	0.41	**0.59**	5
			GA-PLS	0.88	0.88	**0.86**	1.16	1.16	0.04	**0.75**	102.52	1.54	1.52	0.25	**0.50**	3
			RF	0.52	0.52	**0.96**	1.34	1.33	0.09	**0.68**	135.35	1.35	1.35	-0.05	**0.60**	194
2.0	3	1328	PLS	0.89	0.89	**0.86**	1.41	1.40	0.12	**0.64**	150.97	1.37	1.31	0.40	**0.62**	1
			GA-PLS	0.94	0.94	**0.84**	1.19	1.19	0.05	**0.74**	107.70	1.56	1.54	-0.25	**0.50**	2
			RF	0.54	0.53	**0.96**	1.33	1.32	0.10	**0.68**	133.35	1.36	1.36	-0.03	**0.60**	69
2.5	2	727	PLS	0.87	0.87	**0.86**	1.37	1.36	0.11	**0.66**	142.36	1.36	1.29	0.42	**0.63**	1
			GA-PLS	1.07	1.07	**0.79**	1.25	1.25	0.03	**0.72**	118.22	1.58	1.58	0.03	**0.50**	2
			RF	0.50	0.50	**0.97**	1.28	1.27	0.09	**0.70**	124.07	1.33	1.33	-0.09	**0.62**	40
3.0	1	511	PLS	0.82	0.82	**0.88**	1.37	1.36	0.14	**0.66**	142.05	1.38	1.34	0.35	**0.60**	1
			GA-PLS	1.04	1.04	**0.80**	1.25	1.25	0.01	**0.71**	118.80	1.67	1.67	0.04	**0.46**	2
			RF	0.52	0.52	**0.96**	1.30	1.30	0.08	**0.69**	128.03	1.34	1.34	-0.03	**0.61**	29

TABLE C.6: Benchmarking statistics for AchE dataset using $g_O(r)$ descriptors.

AchE - OXYGEN

GS[a]	σ=0[b]	TNC[c]	Method[d]	Training			Cross-Validation					Test				Time[e]
				RMSE[f]	σ[g]	r^{2h}	RMSE	σ	Bias[i]	q^{2j}	PRESS[k]	RMSE	σ	Bias	r^2	
0.5	2201	82983	PLS	-	-	**-**	-	-	-	**-**	-	-	-	-	**-**	-
			GA-PLS	0.52	0.52	**0.81**	0.80	0.80	-0.03	**0.56**	47.05	0.77	0.75	-0.18	**0.67**	26
			RF	-	-	**-**	-	-	-	**-**	-	-	-	-	**-**	-
1.0	274	10374	PLS	0.57	0.57	**0.78**	0.86	0.86	-0.03	**0.49**	55.02	0.82	0.80	-0.17	**0.63**	10
			GA-PLS	0.46	0.46	**0.85**	0.73	0.72	-0.05	**0.64**	38.99	0.80	0.79	-0.15	**0.63**	3
			RF	0.35	0.35	**0.96**	0.94	0.94	-0.02	**0.40**	64.82	0.98	0.96	-0.20	**0.45**	566
1.5	84	3291	PLS	0.58	0.58	**0.77**	0.87	0.87	-0.03	**0.48**	55.88	0.81	0.79	-0.16	**0.63**	1
			GA-PLS	0.53	0.53	**0.81**	0.73	0.73	-0.03	**0.64**	38.94	1.02	0.99	-0.24	**0.42**	1
			RF	0.36	0.35	**0.95**	0.91	0.91	-0.03	**0.43**	61.87	0.99	0.96	-0.23	**0.46**	97
2.0	33	1298	PLS	0.59	0.59	**0.76**	0.89	0.89	-0.04	**0.46**	58.85	0.86	0.84	-0.20	**0.59**	1
			GA-PLS	0.57	0.57	**0.78**	0.72	0.72	-0.04	**0.64**	38.36	0.96	0.96	-0.10	**0.46**	1
			RF	0.36	0.36	**0.95**	0.94	0.94	-0.03	**0.40**	65.32	0.97	0.96	-0.16	**0.45**	33
2.5	19	710	PLS	0.59	0.59	**0.76**	0.85	0.85	-0.02	**0.51**	53.05	0.84	0.82	-0.02	**0.60**	1
			GA-PLS	0.67	0.67	**0.70**	0.80	0.80	-0.01	**0.57**	46.88	1.01	1.01	-0.09	**0.39**	1
			RF	0.36	0.36	**0.95**	0.93	0.93	-0.02	**0.41**	63.43	0.97	0.95	-0.20	**0.49**	18
3.0	9	503	PLS	0.58	0.58	**0.77**	0.90	0.90	-0.03	**0.44**	60.02	0.82	0.81	-0.11	**0.61**	1
			GA-PLS	0.64	0.64	**0.72**	0.74	0.74	0.00	**0.62**	40.70	1.29	1.28	0.08	**0.43**	1
			RF	0.36	0.36	**0.95**	0.93	0.93	-0.01	**0.41**	64.00	0.97	0.95	-0.20	**0.47**	13

TABLE C.7: Benchmarking statistics for AchE dataset using Solvation Free Energy Density (SFED) descriptors.

AchE - SFED

GS[a]	σ=0[b]	TNC[c]	Method[d]	Training			Cross-Validation					Test				Time[e]
				RMSE[f]	σ[g]	r^{2h}	RMSE	σ	Bias[i]	q^{2j}	PRESS[k]	RMSE	σ	Bias	r^2	
0.5	20	85164	PLS	-	-	**-**	-	-	-	**-**	-	-	-	-	**-**	-
			GA-PLS	0.73	0.73	**0.63**	0.96	0.96	-0.11	**0.36**	68.78	1.00	0.97	-0.25	**0.44**	83
			RF	-	-	**-**	-	-	-	**-**	-	-	-	-	**-**	-
1.0	6	10642	PLS	0.76	0.76	**0.61**	1.03	1.02	-0.11	**0.28**	78.13	1.02	1.00	-0.21	**0.40**	29
			GA-PLS	0.69	0.69	**0.67**	0.89	0.88	-0.08	**0.46**	58.34	0.96	0.94	-0.19	**0.47**	11
			RF	0.36	0.36	**0.95**	0.93	0.93	-0.06	**0.41**	63.55	0.94	0.92	-0.18	**0.51**	596
1.5	4	3371	PLS	0.75	0.75	**0.61**	1.04	1.03	-0.12	**0.26**	79.72	1.02	0.99	-0.22	**0.41**	2
			GA-PLS	0.70	0.70	**0.66**	0.86	0.86	-0.06	**0.49**	54.93	1.01	0.99	-0.19	**0.42**	5
			RF	0.35	0.35	**0.95**	0.92	0.91	-0.06	**0.43**	61.98	0.93	0.91	-0.19	**0.52**	94
2.0	2	1329	PLS	0.75	0.75	**0.62**	1.03	1.03	-0.11	**0.27**	78.60	1.06	1.04	-0.21	**0.37**	1
			GA-PLS	0.72	0.72	**0.65**	0.81	0.81	-0.04	**0.55**	48.59	0.94	0.92	-0.19	**0.49**	3
			RF	0.36	0.36	**0.95**	0.92	0.92	-0.06	**0.42**	62.83	0.92	0.90	-0.19	**0.54**	33
2.5	1	728	PLS	0.74	0.74	**0.62**	1.03	1.03	-0.12	**0.27**	78.85	1.06	1.03	-0.25	**0.37**	1
			GA-PLS	0.73	0.73	**0.63**	0.86	0.86	-0.06	**0.49**	55.30	1.13	1.12	-0.09	**0.36**	2
			RF	0.35	0.35	**0.94**	0.91	0.91	-0.06	**0.44**	60.89	0.95	0.93	-0.20	**0.50**	19
3.0	2	510	PLS	0.75	0.75	**0.62**	1.05	1.04	-0.15	**0.24**	81.95	1.05	1.03	-0.21	**0.38**	1
			GA-PLS	0.76	0.76	**0.60**	0.90	0.90	-0.06	**0.45**	59.74	1.09	1.08	-0.07	**0.34**	2
			RF	0.35	0.34	**0.95**	0.92	0.91	-0.08	**0.42**	62.14	0.91	0.90	-0.15	**0.53**	14

TABLE C.8: Benchmarking statistics for AchE dataset using $g_H(r)$ descriptors.

AchE - $g_H(r)$

GS[a]	σ=0[b]	TNC[c]	Method[d]	Training			Cross-Validation					Test				Time[e]
				RMSE[f]	σ[g]	r^2[h]	RMSE	σ	Bias[i]	q^2[j]	PRESS[k]	RMSE	σ	Bias	r^2	
0.5	744	84440	PLS	-	-	**-**	-	-	-	**-**	-	-	-	-	**-**	-
			GA-PLS	0.53	0.53	**0.81**	0.78	0.78	-0.04	**0.58**	45.47	0.79	0.76	-0.21	**0.67**	94
			RF	-	-	**-**	-	-	-	**-**	-	-	-	-	**-**	-
1.0	99	10549	PLS	0.57	0.57	**0.77**	0.86	0.86	-0.03	**0.49**	54.88	0.81	0.79	-0.18	**0.63**	22
			GA-PLS	0.52	0.52	**0.81**	0.74	0.74	-0.04	**0.63**	40.03	0.81	0.80	-0.14	**0.64**	5
			RF	0.36	0.36	**0.95**	0.93	0.93	-0.04	**0.40**	64.48	0.97	0.95	-0.20	**0.48**	662
1.5	26	3349	PLS	0.58	0.58	**0.77**	0.86	0.86	-0.03	**0.49**	54.84	0.82	0.80	-0.18	**0.62**	2
			GA-PLS	0.52	0.52	**0.81**	0.72	0.72	-0.02	**0.64**	38.48	0.95	0.94	-0.15	**0.49**	3
			RF	0.35	0.35	**0.95**	0.92	0.91	-0.04	**0.43**	61.91	0.96	0.94	-0.19	**0.49**	89
2.0	13	1318	PLS	0.58	0.58	**0.77**	0.87	0.86	-0.02	**0.49**	55.32	0.86	0.84	-0.20	**0.58**	1
			GA-PLS	0.59	0.59	**0.77**	0.72	0.72	0.01	**0.64**	38.83	1.15	1.09	-0.37	**0.32**	2
			RF	0.36	0.36	**0.95**	0.92	0.92	-0.04	**0.41**	63.22	0.96	0.94	-0.18	**0.48**	21
2.5	4	725	PLS	0.57	0.57	**0.78**	0.85	0.85	-0.04	**0.50**	53.47	0.79	0.77	-0.18	**0.65**	1
			GA-PLS	0.61	0.61	**0.75**	0.73	0.73	-0.02	**0.63**	39.57	0.96	0.94	-0.19	**0.48**	2
			RF	0.36	0.36	**0.95**	0.94	0.94	-0.03	**0.39**	65.57	0.97	0.95	-0.21	**0.47**	14
3.0	3	509	PLS	0.58	0.58	**0.77**	0.86	0.86	-0.02	**0.49**	54.70	0.80	0.78	-0.17	**0.64**	1
			GA-PLS	0.67	0.67	**0.69**	0.77	0.77	-0.02	**0.59**	44.20	1.07	1.06	-0.15	**0.37**	2
			RF	0.37	0.37	**0.95**	0.92	0.92	-0.04	**0.42**	62.87	0.97	0.95	-0.18	**0.47**	10

TABLE C.9: Benchmarking statistics for AchE dataset using $g_{C-}(r)$ descriptors.

AchE - $g_{C-}(r)$

GS[a]	σ=0[b]	TNC[c]	Method[d]	Training			Cross-Validation					Test				Time[e]
				RMSE[f]	σ[g]	r^2[h]	RMSE	σ	Bias[i]	q^2[j]	PRESS[k]	RMSE	σ	Bias	r^2	
0.5	1309	83875	PLS	-	-	**-**	-	-	-	**-**	-	-	-	-	**-**	-
			GA-PLS	0.52	0.52	**0.82**	0.79	0.79	-0.04	**0.57**	46.23	0.77	0.75	-0.16	**0.68**	93
			RF	-	-	**-**	-	-	-	**-**	-	-	-	-	**-**	-
1.0	163	10485	PLS	0.56	0.56	**0.79**	0.86	0.86	-0.05	**0.49**	54.74	0.79	0.77	-0.20	**0.66**	5
			GA-PLS	0.47	0.47	**0.85**	0.73	0.73	-0.06	**0.63**	39.41	0.84	0.83	-0.11	**0.59**	8
			RF	0.36	0.36	**0.95**	0.93	0.93	-0.03	**0.41**	63.60	1.02	1.00	-0.20	**0.41**	1184
1.5	43	3332	PLS	0.56	0.56	**0.78**	0.87	0.87	-0.05	**0.48**	55.76	0.79	0.77	-0.19	**0.66**	4
			GA-PLS	0.55	0.55	**0.80**	0.72	0.72	-0.05	**0.64**	38.76	0.72	0.72	-0.11	**0.70**	3
			RF	0.37	0.37	**0.95**	0.95	0.95	-0.03	**0.38**	66.59	0.99	0.97	-0.19	**0.44**	188
2.0	20	1311	PLS	0.59	0.59	**0.77**	0.90	0.90	-0.07	**0.45**	59.72	0.83	0.80	-0.21	**0.63**	1
			GA-PLS	0.57	0.57	**0.78**	0.76	0.76	-0.03	**0.60**	42.77	1.00	0.94	-0.35	**0.49**	2
			RF	0.36	0.36	**0.95**	0.93	0.93	-0.03	**0.40**	64.42	1.00	0.97	-0.21	**0.45**	68
2.5	10	719	PLS	0.60	0.60	**0.76**	0.88	0.88	-0.04	**0.46**	57.89	0.85	0.82	-0.22	**0.61**	1
			GA-PLS	0.61	0.61	**0.74**	0.78	0.78	0.01	**0.58**	44.91	1.14	1.06	-0.40	**0.35**	1
			RF	0.36	0.36	**0.95**	0.94	0.94	-0.03	**0.40**	65.24	0.99	0.97	-0.20	**0.46**	39
3.0	5	507	PLS	0.58	0.58	**0.77**	0.90	0.90	-0.05	**0.45**	59.41	0.81	0.80	-0.13	**0.62**	1
			GA-PLS	0.62	0.62	**0.73**	0.74	0.74	0.00	**0.63**	40.27	1.41	1.41	0.10	**0.21**	1
			RF	0.35	0.35	**0.96**	0.92	0.91	-0.04	**0.43**	61.94	1.00	0.98	-0.20	**0.43**	29

TABLE C.10: Benchmarking statistics for AchE dataset using $g_{C+}(r)$ descriptors.

AchE - $g_{C+}(r)$

GS[a]	σ=0[b]	TNC[c]	Method[d]	Training			Cross-Validation					Test				Time[e]
				RMSE[f]	σ[g]	r^2[h]	RMSE	σ	Bias[i]	q^2[j]	PRESS[k]	RMSE	σ	Bias	r^2	
0.5	808	84376	PLS	-	-	**-**	-	-	-	**-**	-	-	-	-	**-**	-
			GA-PLS	0.54	0.54	**0.80**	0.80	0.80	-0.05	**0.56**	47.53	0.76	0.73	-0.20	**0.70**	95
			RF	-	-	**-**	-	-	-	**-**	-	-	-	-	**-**	-
1.0	111	10537	PLS	0.57	0.57	**0.78**	0.87	0.86	-0.05	**0.49**	55.38	0.79	0.77	-0.20	**0.67**	5
			GA-PLS	0.50	0.50	**0.83**	0.74	0.73	-0.06	**0.63**	40.03	0.83	0.82	-0.16	**0.60**	8
			RF	0.35	0.35	**0.96**	0.91	0.91	-0.03	**0.43**	61.40	1.02	1.00	-0.20	**0.40**	1179
1.5	29	3346	PLS	0.56	0.56	**0.78**	0.86	0.86	-0.06	**0.49**	55.21	0.80	0.77	-0.21	**0.66**	4
			GA-PLS	0.47	0.47	**0.85**	0.68	0.68	-0.05	**0.68**	34.28	0.86	0.82	-0.26	**0.60**	3
			RF	0.36	0.36	**0.95**	0.93	0.93	-0.02	**0.41**	63.40	1.03	1.01	-0.18	**0.39**	190
2.0	15	1316	PLS	0.58	0.58	**0.77**	0.86	0.85	-0.06	**0.50**	54.11	0.81	0.79	-0.18	**0.64**	1
			GA-PLS	0.57	0.57	**0.78**	0.71	0.70	-0.03	**0.66**	36.77	0.96	0.95	-0.17	**0.50**	2
			RF	0.37	0.37	**0.95**	0.94	0.94	-0.03	**0.40**	64.93	1.05	1.03	-0.19	**0.36**	68
2.5	4	725	PLS	0.61	0.61	**0.75**	0.91	0.91	-0.06	**0.43**	61.50	0.84	0.83	-0.16	**0.60**	1
			GA-PLS	0.63	0.63	**0.73**	0.76	0.76	-0.02	**0.60**	43.10	1.03	1.03	-0.08	**0.36**	2
			RF	0.37	0.37	**0.95**	0.94	0.94	-0.02	**0.39**	65.61	1.05	1.04	-0.15	**0.36**	39
3.0	3	509	PLS	0.58	0.58	**0.77**	0.85	0.85	-0.04	**0.50**	53.78	0.82	0.79	-0.23	**0.65**	1
			GA-PLS	0.61	0.61	**0.74**	0.74	0.74	0.02	**0.63**	40.10	1.33	1.33	-0.08	**0.27**	2
			RF	0.35	0.35	**0.96**	0.90	0.90	-0.02	**0.45**	59.42	1.02	1.01	-0.19	**0.40**	29

TABLE C.11: Benchmarking statistics for BZR dataset using $g_O(r)$ descriptors.

BZR - OXYGEN

GS[a]	σ=0[b]	TNC[c]	Method[d]	Training			Cross-Validation					Test				Time[e]
				RMSE[f]	σ^g	r^{2h}	RMSE	σ	Bias[i]	q^{2j}	PRESS[k]	RMSE	σ	Bias	r^2	
0.5	2116	83068	PLS	-	-	**-**	-	-	-	**-**	-	-	-	-	**-**	-
			GA-PLS	0.42	0.42	**0.59**	0.55	0.55	-0.01	**0.31**	29.34	0.89	0.80	-0.38	**0.19**	31
			RF	-	-	**-**	-	-	-	**-**	-	-	-	-	**-**	-
1.0	271	10377	PLS	0.42	0.42	**0.59**	0.58	0.58	-0.01	**0.23**	32.85	0.90	0.81	-0.39	**0.18**	5
			GA-PLS	0.41	0.41	**0.61**	0.52	0.52	0.00	**0.39**	26.12	0.93	0.83	-0.42	**0.14**	3
			RF	0.21	0.21	**0.94**	0.55	0.55	-0.01	**0.31**	29.34	0.85	0.80	-0.31	**0.20**	649
1.5	77	3298	PLS	0.43	0.43	**0.59**	0.58	0.58	-0.01	**0.23**	32.91	0.89	0.80	-0.39	**0.18**	1
			GA-PLS	0.43	0.43	**0.58**	0.51	0.51	0.00	**0.42**	24.96	0.97	0.89	-0.37	**0.08**	2
			RF	0.21	0.21	**0.94**	0.54	0.54	0.00	**0.31**	29.55	0.86	0.80	-0.33	**0.20**	146
2.0	35	1296	PLS	0.43	0.43	**0.58**	0.59	0.58	-0.02	**0.22**	33.49	0.90	0.81	-0.39	**0.17**	1
			GA-PLS	0.45	0.45	**0.55**	0.51	0.51	0.00	**0.40**	25.46	0.97	0.85	-0.46	**0.13**	1
			RF	0.20	0.20	**0.94**	0.53	0.53	0.00	**0.35**	27.65	0.84	0.79	-0.29	**0.21**	59
2.5	15	714	PLS	0.43	0.43	**0.58**	0.59	0.59	-0.02	**0.20**	34.10	0.89	0.80	-0.38	**0.19**	1
			GA-PLS	0.44	0.44	**0.56**	0.49	0.49	-0.01	**0.46**	23.26	0.96	0.87	-0.41	**0.12**	1
			RF	0.21	0.21	**0.94**	0.55	0.55	-0.01	**0.30**	29.74	0.86	0.80	-0.32	**0.19**	33
3.0	10	502	PLS	0.44	0.44	**0.55**	0.59	0.59	-0.01	**0.21**	33.76	0.91	0.82	-0.41	**0.16**	1
			GA-PLS	0.47	0.47	**0.49**	0.50	0.50	0.00	**0.42**	24.94	0.98	0.88	-0.41	**0.08**	1
			RF	0.22	0.22	**0.94**	0.55	0.55	0.00	**0.30**	30.05	0.87	0.80	-0.34	**0.20**	25

TABLE C.12: Benchmarking statistics for BZR dataset using Solvation Free Energy Density (SFED) descriptors.

BZR - SFED

GS[a]	σ=0[b]	TNC[c]	Method[d]	Training			Cross-Validation					Test				Time[e]
				RMSE[f]	σ^g	r^{2h}	RMSE	σ	Bias[i]	q^{2j}	PRESS[k]	RMSE	σ	Bias	r^2	
0.5	18	85166	PLS	-	-	**-**	-	-	-	**-**	-	-	-	-	**-**	-
			GA-PLS	0.49	0.49	**0.44**	0.59	0.59	-0.03	**0.21**	33.61	0.94	0.86	-0.38	**0.09**	88
			RF	-	-	**-**	-	-	-	**-**	-	-	-	-	**-**	-
1.0	1	10647	PLS	0.51	0.51	**0.41**	0.62	0.62	-0.03	**0.12**	37.74	0.93	0.85	-0.39	**0.10**	26
			GA-PLS	0.47	0.47	**0.49**	0.54	0.54	-0.02	**0.32**	28.94	0.93	0.84	-0.39	**0.11**	13
			RF	0.21	0.21	**0.94**	0.56	0.56	-0.02	**0.28**	30.74	0.87	0.80	-0.34	**0.21**	683
1.5	1	3374	PLS	0.51	0.51	**0.41**	0.62	0.62	-0.03	**0.11**	37.88	0.94	0.85	-0.39	**0.08**	2
			GA-PLS	0.47	0.47	**0.49**	0.53	0.53	-0.01	**0.37**	27.13	0.98	0.91	-0.37	**0.05**	6
			RF	0.21	0.21	**0.94**	0.56	0.56	-0.02	**0.29**	30.53	0.88	0.81	-0.35	**0.18**	146
2.0	0	1331	PLS	0.50	0.50	**0.42**	0.62	0.62	-0.03	**0.11**	37.98	0.94	0.85	-0.39	**0.09**	1
			GA-PLS	0.49	0.49	**0.46**	0.53	0.53	-0.01	**0.37**	27.09	0.88	0.82	-0.32	**0.16**	4
			RF	0.22	0.22	**0.94**	0.55	0.55	-0.01	**0.30**	30.00	0.88	0.80	-0.36	**0.19**	58
2.5	0	729	PLS	0.50	0.50	**0.42**	0.62	0.62	-0.03	**0.12**	37.77	0.92	0.84	-0.37	**0.11**	1
			GA-PLS	0.47	0.47	**0.48**	0.51	0.51	-0.01	**0.41**	25.22	0.91	0.84	-0.36	**0.13**	3
			RF	0.21	0.21	**0.94**	0.56	0.56	-0.02	**0.28**	30.83	0.89	0.80	-0.38	**0.19**	34
3.0	1	511	PLS	0.52	0.52	**0.39**	0.63	0.63	-0.03	**0.10**	38.54	0.95	0.86	-0.39	**0.07**	1
			GA-PLS	0.48	0.48	**0.47**	0.52	0.52	0.00	**0.39**	26.02	0.91	0.85	-0.34	**0.12**	3
			RF	0.22	0.22	**0.94**	0.57	0.57	-0.01	**0.26**	31.52	0.89	0.80	-0.37	**0.19**	26

TABLE C.13: Benchmarking statistics for BZR dataset using $g_H(r)$ descriptors.

BZR - $g_H(r)$

GS[a]	σ=0[b]	TNC[c]	Method[d]	Training			Cross-Validation					Test				Time[e]
				RMSE[f]	σ[g]	r^2[h]	RMSE	σ	Bias[i]	q^2[j]	PRESS[k]	RMSE	σ	Bias	r^2	
0.5	792	84392	PLS	-	-	**-**	-	-	-	**-**	-	-	-	-	**-**	-
			GA-PLS	0.41	0.41	**0.61**	0.55	0.55	-0.01	**0.31**	29.56	0.91	0.82	-0.40	**0.17**	107
			RF	-	-	**-**	-	-	-	**-**	-	-	-	-	**-**	-
1.0	98	10550	PLS	0.43	0.43	**0.58**	0.58	0.58	-0.01	**0.24**	32.71	0.89	0.80	-0.39	**0.19**	23
			GA-PLS	0.42	0.42	**0.59**	0.52	0.52	0.00	**0.37**	26.82	0.92	0.84	-0.38	**0.14**	5
			RF	0.21	0.21	**0.94**	0.55	0.55	-0.01	**0.31**	29.66	0.86	0.80	-0.32	**0.20**	767
1.5	28	3347	PLS	0.43	0.43	**0.58**	0.58	0.58	-0.01	**0.23**	32.76	0.89	0.81	-0.39	**0.18**	2
			GA-PLS	0.43	0.43	**0.57**	0.51	0.51	0.00	**0.40**	25.78	0.86	0.80	-0.32	**0.21**	4
			RF	0.21	0.21	**0.94**	0.55	0.55	-0.01	**0.30**	30.08	0.86	0.80	-0.31	**0.20**	143
2.0	12	1319	PLS	0.44	0.44	**0.56**	0.58	0.58	-0.01	**0.22**	33.23	0.88	0.80	-0.38	**0.19**	1
			GA-PLS	0.44	0.44	**0.56**	0.52	0.52	-0.01	**0.39**	26.08	0.85	0.80	-0.28	**0.21**	3
			RF	0.21	0.21	**0.94**	0.55	0.55	-0.01	**0.31**	29.55	0.85	0.80	-0.30	**0.19**	32
2.5	8	721	PLS	0.43	0.43	**0.57**	0.58	0.58	-0.01	**0.24**	32.49	0.89	0.80	-0.38	**0.19**	1
			GA-PLS	0.47	0.47	**0.50**	0.52	0.52	0.00	**0.39**	26.08	1.05	0.97	-0.42	**0.03**	2
			RF	0.21	0.21	**0.94**	0.55	0.55	0.00	**0.31**	29.51	0.87	0.80	-0.32	**0.19**	21
3.0	2	510	PLS	0.43	0.43	**0.58**	0.57	0.57	-0.01	**0.25**	32.19	0.90	0.81	-0.39	**0.17**	1
			GA-PLS	0.45	0.45	**0.54**	0.49	0.49	-0.01	**0.46**	23.18	1.02	0.93	-0.41	**0.06**	2
			RF	0.21	0.21	**0.95**	0.54	0.54	0.00	**0.33**	28.75	0.86	0.80	-0.30	**0.19**	17

TABLE C.14: Benchmarking statistics for BZR dataset using $g_{C-}(r)$ descriptors.

BZR - $g_{C-}(r)$

GS[a]	σ=0[b]	TNC[c]	Method[d]	Training			Cross-Validation					Test				Time[e]
				RMSE[f]	σ[g]	r^2[h]	RMSE	σ	Bias[i]	q^2[j]	PRESS[k]	RMSE	σ	Bias	r^2	
0.5	1299	83885	PLS	-	-	**-**	-	-	-	**-**	-	-	-	-	**-**	-
			GA-PLS	0.42	0.42	**0.60**	0.55	0.55	-0.02	**0.31**	29.51	0.89	0.81	-0.37	**0.18**	106
			RF	-	-	**-**	-	-	-	**-**	-	-	-	-	**-**	-
1.0	161	10487	PLS	0.43	0.43	**0.58**	0.58	0.58	-0.02	**0.23**	32.84	0.88	0.79	-0.38	**0.20**	8
			GA-PLS	0.40	0.40	**0.63**	0.51	0.51	-0.01	**0.40**	25.76	0.94	0.86	-0.38	**0.13**	10
			RF	0.20	0.20	**0.95**	0.53	0.53	0.00	**0.37**	27.07	0.86	0.80	-0.32	**0.20**	523
1.5	47	3328	PLS	0.43	0.43	**0.58**	0.58	0.58	-0.02	**0.24**	32.59	0.88	0.79	-0.38	**0.21**	5
			GA-PLS	0.42	0.42	**0.60**	0.49	0.49	0.00	**0.44**	23.96	0.90	0.81	-0.40	**0.20**	3
			RF	0.21	0.21	**0.95**	0.54	0.54	0.00	**0.34**	28.32	0.86	0.80	-0.31	**0.20**	286
2.0	19	1312	PLS	0.43	0.43	**0.57**	0.58	0.58	-0.02	**0.23**	32.75	0.89	0.80	-0.39	**0.18**	1
			GA-PLS	0.43	0.43	**0.58**	0.50	0.50	-0.01	**0.43**	24.33	0.99	0.91	-0.39	**0.06**	2
			RF	0.20	0.20	**0.95**	0.53	0.53	-0.01	**0.36**	27.51	0.86	0.80	-0.32	**0.20**	117
2.5	10	719	PLS	0.43	0.43	**0.57**	0.58	0.57	-0.02	**0.24**	32.37	0.90	0.81	-0.38	**0.17**	1
			GA-PLS	0.46	0.46	**0.51**	0.51	0.51	0.00	**0.42**	24.95	0.96	0.85	-0.45	**0.11**	2
			RF	0.21	0.21	**0.95**	0.53	0.53	-0.01	**0.35**	27.59	0.87	0.81	-0.32	**0.18**	70
3.0	7	505	PLS	0.45	0.45	**0.53**	0.58	0.58	-0.01	**0.22**	33.41	0.90	0.80	-0.41	**0.19**	1
			GA-PLS	0.46	0.46	**0.52**	0.50	0.50	0.00	**0.43**	24.49	0.97	0.87	-0.43	**0.10**	1
			RF	0.22	0.22	**0.94**	0.55	0.55	-0.01	**0.31**	29.44	0.86	0.79	-0.34	**0.22**	51

TABLE C.15: Benchmarking statistics for BZR dataset using $g_{C+}(r)$ descriptors.

BZR - $g_{C+}(r)$

GS[a]	σ=0[b]	TNC[c]	Method[d]	Training			Cross-Validation					Test				Time[e]
				RMSE[f]	σ[g]	r^2[h]	RMSE	σ	Bias[i]	q^2[j]	PRESS[k]	RMSE	σ	Bias	r^2	
0.5	874	84310	PLS	-	-	**-**	-	-	-	**-**	-	-	-	-	**-**	-
			GA-PLS	0.43	0.43	**0.58**	0.55	0.55	-0.01	**0.31**	29.67	0.89	0.80	-0.37	**0.19**	107
			RF	-	-	**-**	-	-	-	**-**	-	-	-	-	**-**	-
1.0	113	10535	PLS	0.43	0.43	**0.58**	0.58	0.58	-0.02	**0.22**	33.29	0.89	0.80	-0.38	**0.20**	8
			GA-PLS	0.41	0.41	**0.61**	0.53	0.52	-0.01	**0.37**	26.96	0.90	0.83	-0.35	**0.17**	10
			RF	0.20	0.20	**0.95**	0.53	0.53	-0.01	**0.36**	27.21	0.86	0.80	-0.32	**0.19**	526
1.5	32	3343	PLS	0.43	0.43	**0.57**	0.58	0.58	-0.02	**0.22**	33.17	0.89	0.80	-0.38	**0.19**	5
			GA-PLS	0.41	0.41	**0.61**	0.49	0.49	-0.01	**0.44**	23.85	0.97	0.88	-0.39	**0.09**	4
			RF	0.20	0.20	**0.95**	0.52	0.52	-0.01	**0.38**	26.61	0.86	0.80	-0.32	**0.20**	289
2.0	13	1318	PLS	0.43	0.43	**0.58**	0.58	0.58	-0.01	**0.23**	33.01	0.89	0.80	-0.39	**0.19**	1
			GA-PLS	0.44	0.44	**0.57**	0.50	0.50	0.00	**0.43**	24.54	0.94	0.84	-0.43	**0.15**	2
			RF	0.20	0.20	**0.94**	0.54	0.54	0.00	**0.34**	28.18	0.87	0.80	-0.33	**0.20**	117
2.5	8	721	PLS	0.44	0.44	**0.56**	0.59	0.59	-0.02	**0.21**	33.84	0.90	0.80	-0.40	**0.18**	1
			GA-PLS	0.47	0.47	**0.49**	0.50	0.50	0.00	**0.42**	24.83	1.00	..93	-0.36	**0.07**	2
			RF	0.20	0.20	**0.95**	0.53	0.53	0.00	**0.35**	27.97	0.87	0.81	-0.34	**0.19**	69
3.0	3	509	PLS	0.44	0.44	**0.56**	0.60	0.60	-0.02	**0.18**	35.10	0.91	0.83	-0.39	**0.15**	1
			GA-PLS	0.46	0.46	**0.51**	0.51	0.51	0.00	**0.41**	25.40	0.93	0.87	-0.33	**0.10**	2
			RF	0.21	0.21	**0.95**	0.54	0.54	0.00	**0.34**	28.40	0.86	0.80	-0.31	**0.21**	51

TABLE C.16: Benchmarking statistics for COX2 dataset using $g_O(r)$ descriptors.

COX2 - OXYGEN

GS[a]	σ=0[b]	TNC[c]	Method[d]	Training			Cross-Validation					Test				Time[e]
				RMSE[f]	σ^g	r^{2h}	RMSE	σ	Bias[i]	q^{2j}	PRESS[k]	RMSE	σ	Bias	r^2	
0.5	3001	82183	PLS	-	-	**-**	-	-	-	**-**	-	-	-	-	**-**	-
			GA-PLS	0.63	0.63	**0.62**	0.75	0.75	-0.01	**0.46**	104.99	1.24	1.09	-0.60	**0.33**	51
			RF	-	-	**-**	-	-	-	**-**	-	-	-	-	**-**	-
1.0	374	10274	PLS	0.64	0.64	**0.61**	0.77	0.77	-0.01	**0.43**	111.56	1.22	1.07	-0.59	**0.34**	6
			GA-PLS	0.63	0.63	**0.61**	0.73	0.73	-0.01	**0.48**	100.61	1.23	1.09	-0.58	**0.32**	6
			RF	0.28	0.28	**0.96**	0.75	0.75	0.01	**0.46**	104.69	1.26	1.08	-0.64	**0.35**	1057
1.5	108	3267	PLS	0.64	0.64	**0.61**	0.77	0.77	0.00	**0.42**	112.75	1.23	1.09	-0.59	**0.33**	2
			GA-PLS	0.64	0.64	**0.61**	0.72	0.72	-0.01	**0.49**	98.32	1.33	1.16	-0.65	**0.24**	3
			RF	0.29	0.29	**0.96**	0.75	0.75	0.00	**0.46**	105.65	1.25	1.08	-0.63	**0.35**	371
2.0	49	1282	PLS	0.65	0.65	**0.60**	0.77	0.77	-0.01	**0.42**	112.56	1.25	1.10	-0.59	**0.31**	1
			GA-PLS	0.67	0.67	**0.57**	0.74	0.74	0.00	**0.48**	101.95	1.27	1.10	-0.63	**0.30**	2
			RF	0.29	0.29	**0.95**	0.74	0.74	0.00	**0.47**	103.78	1.27	1.09	-0.65	**0.34**	178
2.5	23	706	PLS	0.65	0.65	**0.60**	0.78	0.78	0.00	**0.42**	112.93	1.24	1.10	-0.57	**0.30**	1
			GA-PLS	0.71	0.71	**0.52**	0.76	0.76	0.00	**0.44**	108.77	1.44	1.24	-0.73	**0.16**	2
			RF	0.29	0.29	**0.95**	0.75	0.75	0.01	**0.46**	105.27	1.26	1.08	-0.64	**0.35**	116
3.0	11	501	PLS	0.64	0.64	**0.60**	0.77	0.77	-0.01	**0.42**	112.03	1.24	1.09	-0.59	**0.32**	1
			GA-PLS	0.68	0.68	**0.55**	0.75	0.75	0.00	**0.46**	104.27	1.33	1.16	-0.65	**0.25**	1
			RF	0.29	0.29	**0.95**	0.75	0.75	0.00	**0.46**	104.31	1.26	1.09	-0.63	**0.34**	90

TABLE C.17: Benchmarking statistics for COX2 dataset using Solvation Free Energy Density (SFED) descriptors.

COX2 - SFED

GS[a]	σ=0[b]	TNC[c]	Method[d]	Training			Cross-Validation					Test				Time[e]
				RMSE[f]	σ^g	r^{2h}	RMSE	σ	Bias[i]	q^{2j}	PRESS[k]	RMSE	σ	Bias	r^2	
0.5	64	85120	PLS	-	-	**-**	-	-	-	**-**	-	-	-	-	**-**	-
			GA-PLS	0.74	0.74	**0.47**	0.82	0.82	-0.03	**0.36**	124.98	1.38	1.21	-0.67	**0.20**	109
			RF	-	-	**-**	-	-	-	**-**	-	-	-	-	**-**	-
1.0	6	10642	PLS	0.76	0.76	**0.44**	0.87	0.87	-0.03	**0.28**	140.85	1.40	1.22	-0.67	**0.18**	27
			GA-PLS	0.73	0.73	**0.48**	0.79	0.79	-0.02	**0.40**	117.24	1.37	1.16	-0.72	**0.24**	20
			RF	0.30	0.30	**0.95**	0.78	0.78	0.01	**0.42**	113.46	1.24	1.08	-0.61	**0.36**	1112
1.5	3	3372	PLS	0.76	0.76	**0.44**	0.87	0.87	-0.03	**0.27**	142.70	1.40	1.23	-0.67	**0.17**	3
			GA-PLS	0.72	0.72	**0.50**	0.78	0.78	-0.01	**0.41**	114.29	1.32	1.15	-0.66	**0.26**	11
			RF	0.30	0.30	**0.95**	0.79	0.79	0.00	**0.41**	115.72	1.22	1.06	-0.60	**0.37**	386
2.0	0	1331	PLS	0.77	0.77	**0.42**	0.87	0.87	-0.03	**0.26**	143.43	1.38	1.21	-0.67	**0.19**	1
			GA-PLS	0.76	0.76	**0.45**	0.80	0.80	-0.01	**0.39**	119.59	1.44	1.21	-0.78	**0.18**	7
			RF	0.30	0.30	**0.95**	0.78	0.78	0.01	**0.42**	113.06	1.24	1.08	-0.60	**0.34**	183
2.5	0	729	PLS	0.77	0.77	**0.43**	0.86	0.86	-0.03	**0.28**	139.87	1.39	1.22	-0.66	**0.18**	1
			GA-PLS	0.78	0.78	**0.41**	0.82	0.82	-0.01	**0.35**	126.88	1.41	1.21	-0.72	**0.18**	5
			RF	0.31	0.31	**0.94**	0.78	0.78	0.00	**0.41**	114.51	1.24	1.08	-0.61	**0.35**	115
3.0	2	510	PLS	0.75	0.75	**0.46**	0.84	0.84	-0.03	**0.32**	132.26	1.43	1.25	-0.68	**0.16**	1
			GA-PLS	0.76	0.76	**0.44**	0.82	0.82	-0.01	**0.35**	126.13	1.46	1.24	-0.76	**0.14**	5
			RF	0.30	0.30	**0.95**	0.77	0.77	0.01	**0.43**	111.44	1.22	1.07	-0.59	**0.37**	88

TABLE C.18: Benchmarking statistics for COX2 dataset using $g_H(r)$ descriptors.

COX2 - $g_H(r)$

GS[a]	σ=0[b]	TNC[c]	Method[d]	Training			Cross-Validation					Test				Time[e]
				RMSE[f]	σ[g]	r^{2}[h]	RMSE	σ	Bias[i]	q^{2}[j]	PRESS[k]	RMSE	σ	Bias	r^{2}	
0.5	1209	83975	PLS	-	-	**-**	-	-	-	**-**	-	-	-	-	**-**	-
			GA-PLS	0.64	0.64	**0.61**	0.75	0.75	-0.01	**0.45**	106.66	1.23	1.08	-0.59	**0.33**	140
			RF	-	-	**-**	-	-	-	**-**	-	-	-	-	**-**	-
1.0	157	10491	PLS	0.65	0.65	**0.59**	0.77	0.77	-0.01	**0.42**	112.48	1.22	1.07	-0.58	**0.35**	24
			GA-PLS	0.65	0.65	**0.60**	0.74	0.74	-0.01	**0.48**	101.64	1.22	1.07	-0.59	**0.34**	8
			RF	0.29	0.29	**0.95**	0.75	0.75	0.00	**0.46**	105.61	1.26	1.08	-0.65	**0.35**	1284
1.5	40	3335	PLS	0.65	0.65	**0.59**	0.78	0.78	-0.02	**0.42**	113.48	1.21	1.07	-0.58	**0.35**	3
			GA-PLS	0.65	0.65	**0.60**	0.72	0.72	-0.01	**0.50**	97.65	1.32	1.18	-0.60	**0.22**	5
			RF	0.28	0.28	**0.96**	0.74	0.74	0.00	**0.47**	102.43	1.25	1.07	-0.63	**0.36**	410
2.0	18	1313	PLS	0.65	0.65	**0.59**	0.77	0.77	0.00	**0.42**	111.96	1.23	1.08	-0.58	**0.33**	1
			GA-PLS	0.67	0.67	**0.56**	0.74	0.74	0.00	**0.47**	103.67	1.28	1.15	-0.56	**0.26**	4
			RF	0.28	0.28	**0.96**	0.74	0.74	0.00	**0.47**	102.37	1.27	1.09	-0.65	**0.34**	178
2.5	10	719	PLS	0.66	0.66	**0.58**	0.79	0.79	0.00	**0.40**	117.16	1.25	1.11	-0.58	**0.30**	1
			GA-PLS	0.69	0.69	**0.55**	0.74	0.74	-0.01	**0.47**	103.39	1.29	1.16	-0.56	**0.25**	4
			RF	0.29	0.29	**0.95**	0.77	0.77	0.01	**0.43**	110.58	1.27	1.09	-0.65	**0.34**	98
3.0	5	507	PLS	0.66	0.66	**0.58**	0.78	0.78	-0.02	**0.41**	114.33	1.23	1.09	-0.58	**0.32**	1
			GA-PLS	0.70	0.70	**0.53**	0.74	0.74	0.00	**0.47**	104.10	1.41	1.22	-0.72	**0.19**	4
			RF	0.30	0.30	**0.95**	0.75	0.75	0.01	**0.45**	106.12	1.27	1.10	-0.64	**0.32**	63

TABLE C.19: Benchmarking statistics for COX2 dataset using $g_{C-}(r)$ descriptors.

COX2 - $g_{C-}(r)$

GS[a]	σ=0[b]	TNC[c]	Method[d]	Training			Cross-Validation					Test				Time[e]
				RMSE[f]	σ[g]	r^{2}[h]	RMSE	σ	Bias[i]	q^{2}[j]	PRESS[k]	RMSE	σ	Bias	r^{2}	
0.5	1963	83221	PLS	-	-	**-**	-	-	-	**-**	-	-	-	-	**-**	-
			GA-PLS	0.63	0.63	**0.62**	0.75	0.75	-0.01	**0.46**	104.59	1.21	1.07	-0.58	**0.35**	144
			RF	-	-	**-**	-	-	-	**-**	-	-	-	-	**-**	-
1.0	245	10403	PLS	0.64	0.64	**0.61**	0.77	0.77	-0.01	**0.43**	111.41	1.21	1.06	-0.57	**0.36**	6
			GA-PLS	0.62	0.62	**0.63**	0.72	0.72	-0.01	**0.50**	96.62	1.22	1.07	-0.58	**0.34**	21
			RF	0.29	0.29	**0.95**	0.76	0.76	0.01	**0.44**	109.26	1.26	1.08	-0.65	**0.35**	986
1.5	73	3302	PLS	0.64	0.64	**0.61**	0.77	0.77	0.00	**0.42**	112.53	1.22	1.07	-0.58	**0.34**	5
			GA-PLS	0.64	0.64	**0.60**	0.71	0.71	-0.01	**0.52**	94.30	1.33	1.16	-0.65	**0.25**	6
			RF	0.29	0.29	**0.96**	0.75	0.75	0.01	**0.46**	105.41	1.25	1.07	-0.63	**0.36**	755
2.0	31	1300	PLS	0.65	0.65	**0.60**	0.78	0.78	-0.01	**0.42**	113.57	1.21	1.08	-0.54	**0.33**	2
			GA-PLS	0.68	0.68	**0.55**	0.73	0.73	0.00	**0.48**	101.19	1.36	1.18	-0.68	**0.22**	3
			RF	0.30	0.30	**0.95**	0.77	0.77	0.01	**0.43**	109.98	1.26	1.08	-0.65	**0.36**	357
2.5	14	715	PLS	0.64	0.64	**0.61**	0.76	0.76	-0.01	**0.44**	109.09	1.23	1.10	-0.56	**0.31**	1
			GA-PLS	0.70	0.70	**0.53**	0.75	0.75	0.00	**0.46**	105.03	1.37	1.18	-0.69	**0.23**	2
			RF	0.30	0.30	**0.95**	0.77	0.77	0.01	**0.43**	111.16	1.26	1.09	-0.64	**0.34**	226
3.0	10	502	PLS	0.64	0.64	**0.60**	0.77	0.77	-0.01	**0.43**	111.37	1.22	1.07	-0.58	**0.35**	1
			GA-PLS	0.71	0.71	**0.51**	0.76	0.76	0.00	**0.44**	109.37	1.35	1.20	-0.62	**0.20**	2
			RF	0.29	0.29	**0.95**	0.76	0.76	0.01	**0.44**	108.86	1.27	1.10	-0.64	**0.32**	173

TABLE C.20: Benchmarking statistics for COX2 dataset using $g_{C+}(r)$ descriptors.

COX2 - $g_{C+}(r)$

GS[a]	σ=0[b]	TNC[c]	Method[d]	Training			Cross-Validation					Test				Time[e]
				RMSE[f]	σ[g]	r^2[h]	RMSE	σ	Bias[i]	q^2[j]	PRESS[k]	RMSE	σ	Bias	r^2	
0.5	1325	83859	PLS	-	-	**-**	-	-	-	**-**	-	-	-	-	**-**	-
			GA-PLS	0.63	0.63	**0.61**	0.75	0.75	-0.01	**0.46**	105.53	1.22	1.08	-0.56	**0.34**	145
			RF	-	-	**-**	-	-	-	**-**	-	-	-	-	**-**	-
1.0	165	10483	PLS	0.64	0.64	**0.61**	0.77	0.77	-0.01	**0.43**	111.65	1.19	1.05	-0.56	**0.37**	6
			GA-PLS	0.62	0.62	**0.63**	0.71	0.71	-0.01	**0.51**	95.60	1.26	1.12	-0.58	**0.28**	16
			RF	0.29	0.29	**0.95**	0.76	0.76	0.01	**0.45**	107.48	1.25	1.09	-0.62	**0.34**	1004
1.5	43	3332	PLS	0.63	0.63	**0.61**	0.77	0.77	-0.01	**0.43**	111.59	1.19	1.06	-0.55	**0.36**	5
			GA-PLS	0.62	0.62	**0.63**	0.69	0.69	-0.01	**0.54**	90.14	1.27	1.14	-0.56	**0.27**	8
			RF	0.29	0.29	**0.95**	0.76	0.76	0.01	**0.45**	107.05	1.25	1.07	-0.65	**0.37**	758
2.0	20	1311	PLS	0.64	0.64	**0.61**	0.78	0.78	0.00	**0.42**	111.82	1.20	1.07	-0.53	**0.34**	2
			GA-PLS	0.67	0.67	**0.57**	0.74	0.74	-0.01	**0.47**	102.50	1.27	1.14	-0.56	**0.27**	3
			RF	0.29	0.29	**0.95**	0.76	0.76	0.02	**0.44**	109.82	1.25	1.09	-0.61	**0.34**	357
2.5	11	718	PLS	0.65	0.65	**0.59**	0.78	0.78	-0.01	**0.41**	114.23	1.21	1.09	-0.54	**0.32**	1
			GA-PLS	0.68	0.68	**0.55**	0.75	0.75	0.00	**0.46**	104.51	1.36	1.17	-0.69	**0.23**	3
			RF	0.30	0.30	**0.95**	0.76	0.76	0.01	**0.44**	108.27	1.25	1.07	-0.65	**0.37**	226
3.0	5	507	PLS	0.64	0.64	**0.60**	0.77	0.77	-0.02	**0.42**	112.31	1.16	1.04	-0.51	**0.38**	1
			GA-PLS	0.71	0.71	**0.52**	0.76	0.76	-0.01	**0.44**	109.70	1.34	1.22	-0.55	**0.17**	3
			RF	0.29	0.29	**0.96**	0.75	0.75	0.01	**0.46**	105.44	1.25	1.07	-0.65	**0.38**	173

TABLE C.21: Benchmarking statistics for DHFR dataset using $g_O(r)$ descriptors.

DHFR - OXYGEN

GS[a]	σ=0[b]	TNC[c]	Method[d]	Training			Cross-Validation					Test				Time[e]
				RMSE[f]	σ[g]	r^{2}[h]	RMSE	σ	Bias[i]	q^{2}[j]	PRESS[k]	RMSE	σ	Bias	r^{2}	
0.5	1558	83626	PLS	-	-	**-**	-	-	-	**-**	-	-	-	-	**-**	-
			GA-PLS	0.67	0.67	**0.72**	0.78	0.78	-0.01	**0.62**	144.63	0.93	0.92	0.09	**0.55**	90
			RF	-	-	**-**	-	-	-	**-**	-	-	-	-	**-**	-
1.0	192	10456	PLS	0.71	0.71	**0.69**	0.82	0.82	-0.01	**0.58**	158.91	0.94	0.93	0.10	**0.54**	16
			GA-PLS	0.64	0.64	**0.75**	0.75	0.75	-0.02	**0.65**	134.18	0.98	0.98	0.10	**0.51**	19
			RF	0.30	0.30	**0.96**	0.77	0.77	0.01	**0.63**	139.02	0.87	0.87	0.05	**0.60**	1392
1.5	60	3315	PLS	0.72	0.72	**0.68**	0.83	0.83	-0.01	**0.57**	161.57	0.94	0.94	0.10	**0.53**	4
			GA-PLS	0.67	0.67	**0.72**	0.77	0.77	0.00	**0.63**	141.22	0.94	0.94	-0.03	**0.53**	10
			RF	0.30	0.30	**0.96**	0.76	0.76	0.01	**0.64**	138.12	0.88	0.88	0.05	**0.61**	538
2.0	25	1306	PLS	0.72	0.72	**0.68**	0.82	0.82	-0.01	**0.58**	160.17	0.97	0.96	0.11	**0.52**	1
			GA-PLS	0.73	0.73	**0.67**	0.81	0.81	0.01	**0.59**	154.60	1.07	1.07	0.03	**0.44**	6
			RF	0.30	0.30	**0.96**	0.77	0.77	0.02	**0.63**	140.20	0.87	0.87	0.06	**0.61**	270
2.5	14	715	PLS	0.73	0.73	**0.67**	0.84	0.84	-0.01	**0.56**	166.20	0.95	0.95	0.09	**0.52**	1
			GA-PLS	0.82	0.82	**0.58**	0.87	0.87	0.01	**0.52**	181.21	1.08	1.07	-0.10	**0.40**	5
			RF	0.30	0.30	**0.96**	0.77	0.77	0.02	**0.63**	142.02	0.87	0.86	0.07	**0.62**	176
3.0	6	506	PLS	0.72	0.72	**0.68**	0.83	0.83	-0.01	**0.57**	163.86	0.94	0.94	0.11	**0.54**	1
			GA-PLS	0.73	0.73	**0.67**	0.78	0.78	-0.01	**0.62**	143.26	1.12	1.11	0.11	**0.43**	5
			RF	0.30	0.30	**0.96**	0.77	0.77	0.02	**0.63**	139.36	0.84	0.83	0.07	**0.65**	138

TABLE C.22: Benchmarking statistics for DHFR dataset using Solvation Free Energy Density (SFED) descriptors.

DHFR - SFED

GS[a]	σ=0[b]	TNC[c]	Method[d]	Training			Cross-Validation					Test				Time[e]
				RMSE[f]	σ[g]	r^{2}[h]	RMSE	σ	Bias[i]	q^{2}[j]	PRESS[k]	RMSE	σ	Bias	r^{2}	
0.5	30	85154	PLS	-	-	**-**	-	-	-	**-**	-	-	-	-	**-**	-
			GA-PLS	0.81	0.81	**0.59**	0.87	0.87	-0.02	**0.52**	181.12	1.06	1.06	-0.07	**0.42**	128
			RF	-	-	**-**	-	-	-	**-**	-	-	-	-	**-**	-
1.0	8	10640	PLS	0.84	0.84	**0.56**	0.91	0.91	-0.02	**0.49**	194.43	1.09	1.08	-0.10	**0.40**	27
			GA-PLS	0.74	0.74	**0.66**	0.80	0.80	0.01	**0.60**	151.65	1.02	1.02	-0.04	**0.49**	27
			RF	0.29	0.29	**0.96**	0.76	0.76	-0.02	**0.64**	136.03	0.90	0.90	-0.03	**0.57**	1438
1.5	4	3371	PLS	0.84	0.84	**0.56**	0.91	0.91	-0.02	**0.49**	195.41	1.09	1.09	-0.10	**0.39**	3
			GA-PLS	0.75	0.75	**0.65**	0.81	0.81	0.00	**0.59**	157.00	1.02	1.02	-0.04	**0.49**	13
			RF	0.29	0.29	**0.96**	0.76	0.76	-0.02	**0.64**	136.10	0.90	0.90	-0.04	**0.57**	553
2.0	3	1328	PLS	0.84	0.84	**0.56**	0.91	0.91	-0.02	**0.49**	194.89	1.10	1.09	-0.10	**0.39**	1
			GA-PLS	0.76	0.76	**0.64**	0.80	0.80	-0.01	**0.60**	152.12	1.11	1.11	0.05	**0.46**	8
			RF	0.29	0.29	**0.96**	0.76	0.76	-0.02	**0.64**	135.78	0.91	0.91	-0.01	**0.56**	276
2.5	2	727	PLS	0.85	0.85	**0.55**	0.91	0.91	-0.02	**0.48**	197.58	1.11	1.10	-0.09	**0.38**	1
			GA-PLS	0.78	0.78	**0.62**	0.82	0.82	-0.01	**0.58**	159.27	1.11	1.11	-0.05	**0.41**	7
			RF	0.30	0.30	**0.96**	0.76	0.76	-0.02	**0.64**	135.56	0.93	0.93	0.00	**0.54**	181
3.0	2	510	PLS	0.85	0.85	**0.55**	0.91	0.91	-0.02	**0.48**	196.63	1.11	1.10	-0.13	**0.38**	1
			GA-PLS	0.83	0.83	**0.57**	0.87	0.87	-0.01	**0.53**	178.39	1.20	1.20	-0.10	**0.35**	6
			RF	0.30	0.30	**0.96**	0.75	0.75	-0.03	**0.65**	132.34	0.94	0.94	-0.06	**0.53**	139

TABLE C.23: Benchmarking statistics for DHFR dataset using $g_H(r)$ descriptors.

DHFR - $g_H(r)$

GS[a]	σ=0[b]	TNC[c]	Method[d]	Training			Cross-Validation					Test				Time[e]
				RMSE[f]	σ[g]	r^2[h]	RMSE	σ	Bias[i]	q^2[j]	PRESS[k]	RMSE	σ	Bias	r^2	
0.5	486	84698	PLS	-	-	**-**	-	-	-	**-**	-	-	-	-	**-**	-
			GA-PLS	0.67	0.67	**0.72**	0.78	0.78	-0.01	**0.62**	143.59	0.94	0.93	0.13	**0.55**	164
			RF	-	-	**-**	-	-	-	**-**	-	-	-	-	**-**	-
1.0	60	10588	PLS	0.71	0.71	**0.69**	0.82	0.82	-0.01	**0.59**	157.33	0.94	0.94	0.11	**0.54**	25
			GA-PLS	0.64	0.64	**0.75**	0.74	0.74	-0.01	**0.66**	129.48	0.96	0.94	0.18	**0.55**	10
			RF	0.29	0.29	**0.96**	0.76	0.76	0.02	**0.64**	138.03	0.87	0.87	0.08	**0.60**	1662
1.5	18	3357	PLS	0.71	0.71	**0.69**	0.82	0.82	-0.01	**0.59**	157.68	0.95	0.94	0.11	**0.54**	3
			GA-PLS	0.64	0.64	**0.74**	0.73	0.73	-0.01	**0.67**	125.95	0.97	0.96	0.16	**0.53**	6
			RF	0.30	0.30	**0.96**	0.77	0.77	0.02	**0.63**	138.78	0.88	0.87	0.09	**0.60**	616
2.0	7	1324	PLS	0.72	0.72	**0.67**	0.82	0.82	-0.01	**0.58**	161.02	0.95	0.94	0.10	**0.53**	2
			GA-PLS	0.71	0.71	**0.68**	0.77	0.77	0.00	**0.63**	140.34	0.96	0.96	0.06	**0.52**	4
			RF	0.30	0.30	**0.96**	0.77	0.77	0.02	**0.63**	138.97	0.87	0.87	0.08	**0.61**	289
2.5	2	727	PLS	0.73	0.73	**0.67**	0.83	0.83	-0.01	**0.57**	162.26	0.96	0.96	0.09	**0.52**	1
			GA-PLS	0.74	0.74	**0.66**	0.80	0.80	0.00	**0.60**	151.05	1.16	1.16	0.15	**0.37**	4
			RF	0.29	0.29	**0.97**	0.75	0.75	0.01	**0.65**	132.54	0.86	0.86	0.05	**0.62**	177
3.0	2	510	PLS	0.71	0.71	**0.68**	0.82	0.82	-0.01	**0.58**	158.51	0.95	0.94	0.10	**0.53**	1
			GA-PLS	0.73	0.73	**0.67**	0.78	0.78	0.00	**0.63**	142.17	1.02	1.02	0.02	**0.46**	4
			RF	0.31	0.31	**0.96**	0.78	0.77	0.02	**0.63**	142.19	0.87	0.87	0.08	**0.61**	127

TABLE C.24: Benchmarking statistics for DHFR dataset using $g_{C-}(r)$ descriptors.

DHFR - $g_{C-}(r)$

GS[a]	σ=0[b]	TNC[c]	Method[d]	Training			Cross-Validation					Test				Time[e]
				RMSE[f]	σ[g]	r^2[h]	RMSE	σ	Bias[i]	q^2[j]	PRESS[k]	RMSE	σ	Bias	r^2	
0.5	860	84324	PLS	-	-	**-**	-	-	-	**-**	-	-	-	-	**-**	-
			GA-PLS	0.64	0.64	**0.75**	0.77	0.77	-0.02	**0.63**	139.44	0.91	0.91	0.10	**0.57**	169
			RF	-	-	**-**	-	-	-	**-**	-	-	-	-	**-**	-
1.0	108	10540	PLS	0.70	0.70	**0.70**	0.81	0.81	-0.01	**0.59**	155.39	0.96	0.95	0.12	**0.53**	6
			GA-PLS	0.65	0.65	**0.74**	0.75	0.75	-0.01	**0.65**	133.19	0.99	0.98	0.12	**0.52**	28
			RF	0.29	0.29	**0.96**	0.75	0.75	0.01	**0.65**	134.59	0.88	0.88	0.05	**0.60**	1451
1.5	31	3344	PLS	0.70	0.70	**0.69**	0.81	0.81	-0.01	**0.59**	156.76	0.96	0.95	0.12	**0.53**	6
			GA-PLS	0.64	0.64	**0.74**	0.74	0.74	0.00	**0.66**	130.40	0.93	0.92	0.09	**0.56**	7
			RF	0.29	0.29	**0.96**	0.76	0.76	0.01	**0.64**	135.96	0.89	0.88	0.05	**0.59**	1096
2.0	15	1316	PLS	0.71	0.71	**0.68**	0.81	0.81	-0.01	**0.59**	156.83	0.99	0.98	0.13	**0.51**	2
			GA-PLS	0.70	0.70	**0.70**	0.76	0.76	0.00	**0.64**	137.83	1.07	1.07	0.04	**0.43**	4
			RF	0.30	0.30	**0.96**	0.76	0.76	0.01	**0.64**	137.28	0.88	0.88	0.05	**0.60**	547
2.5	6	723	PLS	0.70	0.70	**0.69**	0.82	0.82	-0.02	**0.58**	158.13	0.97	0.96	0.14	**0.52**	1
			GA-PLS	0.76	0.76	**0.64**	0.80	0.80	0.01	**0.60**	152.37	0.97	0.97	-0.01	**0.50**	3
			RF	0.29	0.29	**0.96**	0.76	0.76	0.00	**0.64**	136.92	0.88	0.88	0.05	**0.60**	353
3.0	4	508	PLS	0.70	0.70	**0.70**	0.81	0.81	-0.01	**0.59**	156.02	0.97	0.96	0.16	**0.53**	1
			GA-PLS	0.79	0.79	**0.61**	0.85	0.85	0.00	**0.55**	170.93	1.03	1.03	-0.02	**0.45**	3
			RF	0.29	0.29	**0.96**	0.74	0.74	0.01	**0.66**	129.58	0.89	0.88	0.06	**0.59**	274

TABLE C.25: Benchmarking statistics for DHFR dataset using $g_{C+}(r)$ descriptors.

DHFR - $g_{C+}(r)$

GS[a]	σ=0[b]	TNC[c]	Method[d]	Training			Cross-Validation					Test				Time[e]
				RMSE[f]	σ[g]	r^2[h]	RMSE	σ	Bias[i]	q^2[j]	PRESS[k]	RMSE	σ	Bias	r^2	
0.5	547	84637	PLS	-	-	**-**	-	-	-	**-**	-	-	-	-	**-**	-
			GA-PLS	0.67	0.67	**0.72**	0.77	0.77	-0.01	**0.63**	141.67	0.94	0.93	0.11	**0.55**	170
			RF	-	-	**-**	-	-	-	**-**	-	-	-	-	**-**	-
1.0	67	10581	PLS	0.70	0.70	**0.70**	0.81	0.81	-0.01	**0.59**	154.13	0.95	0.94	0.12	**0.54**	6
			GA-PLS	0.63	0.63	**0.76**	0.73	0.73	-0.01	**0.67**	126.38	0.99	0.98	0.11	**0.51**	21
			RF	0.30	0.30	**0.96**	0.76	0.76	0.01	**0.64**	135.26	0.88	0.88	0.05	**0.60**	1448
1.5	20	3355	PLS	0.70	0.70	**0.70**	0.81	0.81	-0.01	**0.59**	154.38	0.95	0.94	0.12	**0.54**	6
			GA-PLS	0.66	0.66	**0.73**	0.74	0.74	-0.01	**0.66**	128.07	0.98	0.98	0.03	**0.50**	12
			RF	0.29	0.29	**0.97**	0.77	0.77	0.02	**0.63**	139.79	0.86	0.86	0.06	**0.63**	1106
2.0	8	1323	PLS	0.72	0.72	**0.68**	0.82	0.82	-0.01	**0.58**	159.41	0.95	0.94	0.13	**0.53**	2
			GA-PLS	0.71	0.71	**0.69**	0.78	0.78	0.01	**0.62**	144.96	1.04	1.04	0.09	**0.48**	4
			RF	0.29	0.29	**0.96**	0.76	0.76	0.02	**0.64**	136.66	0.88	0.87	0.07	**0.60**	556
2.5	2	727	PLS	0.71	0.71	**0.68**	0.81	0.81	-0.01	**0.59**	155.89	0.95	0.94	0.14	**0.54**	1
			GA-PLS	0.73	0.73	**0.66**	0.79	0.79	0.00	**0.61**	148.33	1.07	1.06	0.14	**0.45**	4
			RF	0.30	0.30	**0.96**	0.76	0.76	0.01	**0.64**	135.93	0.86	0.86	0.06	**0.62**	355
3.0	3	509	PLS	0.69	0.69	**0.70**	0.80	0.80	-0.01	**0.60**	152.78	0.91	0.92	0.10	**0.56**	1
			GA-PLS	0.72	0.72	**0.68**	0.77	0.77	0.00	**0.63**	139.87	0.96	0.96	0.06	**0.52**	3
			RF	0.30	0.30	**0.96**	0.78	0.78	0.01	**0.62**	142.73	0.88	0.88	0.05	**0.61**	277

TABLE C.26: Benchmarking statistics for Steroids dataset using $g_O(r)$ descriptors.

STEROIDS - OXYGEN

GS[a]	σ=0[b]	TNC[c]	Method[d]	Training			Cross-Validation					Time[e]
				RMSE[f]	σ[g]	r^2[h]	RMSE	σ	Bias[i]	q^2[j]	PRESS[k]	
1.0	365	10283	PLS	0.21	0.21	**0.97**	0.45	0.44	0.10	**0.84**	4.32	4
1.5	107	3268	PLS	0.21	0.21	**0.97**	0.44	0.42	0.10	**0.86**	3.98	1
2.0	46	1285	PLS	0.23	0.23	**0.96**	0.46	0.45	0.09	**0.84**	4.49	1
2.5	23	706	PLS	0.23	0.23	**0.96**	0.50	0.48	0.11	**0.81**	5.18	1
3.0	14	498	PLS	0.21	0.21	**0.97**	0.44	0.43	0.10	**0.85**	4.11	1

TABLE C.27: Benchmarking statistics for Steroids dataset using $g_O(r)$ descriptors using the PSE-3 closure.

STEROIDS - OXYGEN (PSE-3)

GS[a]	σ=0[b]	TNC[c]	Method[d]	Training			Cross-Validation					Time[e]
				RMSE[f]	σ[g]	r^2[h]	RMSE	σ	Bias[i]	q^2[j]	PRESS[k]	
1.0	366	10282	PLS	0.20	0.20	**0.97**	0.45	0.44	0.11	**0.85**	4.25	4
1.5	107	3268	PLS	0.19	0.19	**0.97**	0.44	0.42	0.11	**0.86**	3.98	1
2.0	47	1284	PLS	0.22	0.22	**0.96**	0.46	0.45	0.10	**0.84**	4.40	1
2.5	23	706	PLS	0.22	0.22	**0.96**	0.49	0.48	0.12	**0.82**	5.08	1
3.0	14	498	PLS	0.19	0.19	**0.97**	0.43	0.42	0.11	**0.86**	3.90	1

TABLE C.28: Benchmarking statistics for Steroids dataset using $g_H(r)$ descriptors.

Steroids - $g_H(r)$

GS[a]	σ=0[b]	TNC[c]	Method[d]	Training			Cross-Validation					Time[e]
				RMSE[f]	σ[g]	r^{2}[h]	RMSE	σ	Bias[i]	q^{2}[j]	PRESS[k]	
1.0	141	10507	PLS	0.21	0.21	**0.97**	0.45	0.44	0.11	**0.84**	4.29	4
1.5	44	3331	PLS	0.21	0.21	**0.97**	0.45	0.44	0.10	**0.85**	4.24	1
2.0	15	1316	PLS	0.22	0.22	**0.96**	0.47	0.46	0.09	**0.83**	4.60	1
2.5	8	721	PLS	0.21	0.21	**0.97**	0.45	0.44	0.10	**0.85**	4.21	1
3.0	5	507	PLS	0.21	0.21	**0.97**	0.44	0.43	0.10	**0.85**	4.08	1

TABLE C.29: Benchmarking statistics for Steroids dataset using Solvation Free Energy Density (SFED) descriptors.

STEROIDS - SFED

GS[a]	σ=0[b]	TNC[c]	Method[d]	Training			Cross-Validation					Time[e]
				RMSE[f]	σ[g]	r^{2}[h]	RMSE	σ	Bias[i]	q^{2}[j]	PRESS[k]	
1.0	8	10640	PLS	0.34	0.34	**0.91**	0.65	0.64	0.09	**0.68**	8.74	1
1.5	6	3369	PLS	0.36	0.36	**0.90**	0.66	0.66	0.09	**0.67**	9.19	1
2.0	2	1329	PLS	0.34	0.34	**0.91**	0.63	0.63	0.09	**0.70**	8.40	1
2.5	2	727	PLS	0.32	0.32	**0.92**	0.58	0.57	0.10	**0.74**	7.11	1
3.0	4	508	PLS	0.36	0.36	**0.90**	0.65	0.65	0.08	**0.67**	9.00	1

TABLE C.30: Benchmarking statistics for Steroids dataset using $g_{C-}(r)$ descriptors.

Steroids - $g_{C-}(r)$

GS[a]	σ=0[b]	TNC[c]	Method[d]	Training			Cross-Validation					Time[e]
				RMSE[f]	σ[g]	r^2[h]	RMSE	σ	Bias[i]	q^2[j]	PRESS[k]	
1.0	260	10388	PLS	0.21	0.21	**0.97**	0.45	0.44	0.11	**0.84**	4.29	4
1.5	77	3298	PLS	0.21	0.21	**0.97**	0.45	0.43	0.11	**0.85**	4.18	1
2.0	29	1302	PLS	0.21	0.21	**0.97**	0.45	0.43	0.12	**0.85**	4.21	1
2.5	16	713	PLS	0.22	0.22	**0.96**	0.48	0.46	0.11	**0.83**	4.76	1
3.0	10	502	PLS	0.21	0.21	**0.97**	0.47	0.46	0.10	**0.83**	4.66	1

TABLE C.31: Benchmarking statistics for Steroids dataset using $g_{C+}(r)$ descriptors.

Steroids - $g_{C+}(r)$

GS[a]	σ=0[b]	TNC[c]	Method[d]	Training			Cross-Validation					Time[e]
				RMSE[f]	σ[g]	r^2[h]	RMSE	σ	Bias[i]	q^2[j]	PRESS[k]	
1.0	155	10493	PLS	0.20	0.20	**0.97**	0.45	0.44	0.11	**0.84**	4.33	5
1.5	46	3329	PLS	0.21	0.21	**0.97**	0.46	0.45	0.11	**0.84**	4.48	1
2.0	17	1314	PLS	0.20	0.20	**0.97**	0.45	0.44	0.12	**0.85**	4.28	1
2.5	9	720	PLS	0.21	0.21	**0.97**	0.48	0.46	0.11	**0.83**	4.77	1
3.0	5	507	PLS	0.20	0.20	**0.97**	0.46	0.45	0.12	**0.84**	4.52	1

Original Publications

The following is a list of publications originating from the work presented in this thesis:

1. Ansari, S. M., Coletta, A., Kirkeby S. K., Sørensen, J., Schiøtt, B. and Palmer, D. S. (2016) Allosteric-Activation Mechanism of Bovine Chymosin Revealed by Bias-Exchange Metadynamics and Molecular Dynamics Simulations. *J. Phys. Chem. B.* 120, 10453 – 10462.

2. Ansari, S. M., Sørensen, J., Schiøtt, B. and Palmer, D. S. (submitted) On The Effect of Mutations in Bovine or Camel Chymosin on the Thermodynamics of Binding κ-Caseins. *Proteins.*

3. Ansari, S. M. and Palmer, D. S. (in preparation) Comparative Molecular Field Analysis using Molecular Integral Equation Theory.

Bibliography

[1] Schechter, I., and Berger, A. (1967) On the size of the active site in proteases. I. Papain. *Biochem. Biophys. Res. Commun. 27*, 157–162.

[2] Palmer, D. S., Frolov, A. I., Ratkova, E. L., and Fedorov, M. V. (2010) Towards a universal method to calculate hydration free energies: a 3D reference interaction site model with partial molar volume correction. *J. Phys. Cond. Matt. 22*, 492101.

[3] Sørensen, J., Palmer, D. S., and Schiøtt, B. (2013) Hot-Spot Mapping of the Interactions between Chymosin and Bovine kappa-Casein. *J. Agr. Food Chem. 61*, 7949–7959.

[4] Szecsi, P. B. (1992) The aspartic proteases. *Scandinavian J. Clin. Lab. Investig. 52*, 5–22.

[5] Law, B. A. *Microbiology and biochemistry of cheese and fermented milk*; Springer Science & Business Media, 2012.

[6] Barrett, A. J., Woessner, J. F., and Rawlings, N. D. *Handbook of proteolytic enzymes*; Elsevier, 2012; Vol. 1.

[7] Robinson, R. In *Modern Dairy Technology*; Robinson, R. K., Ed.; Springer US, 1993.

[8] Deschamps, J. B. (1840) De la pressure. *J. Pharmacol 26*, 412–420.

[9] Rawlings, N., and Salvesen, G. *Handbook of Proteolytic Enzymes*; Elsevier Science: San Diego, 2012.

[10] Foltmann, B. (1970) Prochymosin and chymosin (prorennin and rennin). *Meth. Enzy 19*, 421–436.

[11] Kappeler, S. R., van den Brink, H. J., Rahbek-Nielsen, H., Farah, Z., Puhan, Z., Hansen, E. B., and Johansen, E. (2006) Characterization of recombinant camel chymosin reveals superior properties for the coagulation of bovine and camel milk. *Biochem. Biophys. Res. Commun. 342*, 647–654.

[12] Giglierano, J. (2011) Comparative Review of IBISWorld Global Industry Reports, Euromonitor Market Research Monitor, and Mintel Global Market Navigator. *The Charleston Advisor 13*, 10–15.

[13] Dairyco, World production trends for dairy products. 2015.

[14] Ansari, S. M., Coletta, A., Kirkeby Skeby, K., Sørensen, J., Schiøtt, B., and Palmer, D. S. (2016) Allosteric-Activation Mechanism of Bovine Chymosin Revealed by Bias-Exchange Metadynamics and Molecular Dynamics Simulations. *J. Phys. Chem. B 120*, 10453–10462.

[15] Plowman, J. E., and Creamer, L. K. (1995) Restrained molecular dynamics study of the interaction between bovine kappa-casein peptide 98-111 and bovine chymosin and porcine pepsin. *J. Dairy Res. 62*, 451–467.

[16] Gustchina, E., Rumsh, L., Ginodman, L., Majer, P., and Andreeva, N. (1996) Post X-ray crystallographic studies of chymosin: the existence of two structural forms and the regulation of activity by the interaction with the histidine-proline cluster of kappa-casein. *FEBS Lett. 379*, 60–62.

[17] Visser, S., Rooijen, P. J. V., Schattenkerk, C., and Kerling, K. E. (1976) Peptide substrates for chymosin (rennin). Kinetic studies with peptides of different chain length including parts of the sequence 101-112 of bovine k-casein. *BBA 438*, 265–272.

[18] Visser, S., Rooijen, P. J. V., Schattenkerk, C., and Kerling, K. E. (1977) Peptide substrates for chymosin (rennin). Kinetic studies with bovine kappa-casein-(103-108)-hexapeptide analogues. *BBA 481*, 171–176.

[19] Veerapandian, B., Cooper, J. B., Sali, A., and Blundell, T. L. (1990) X-ray analyses of aspartic proteinases. III. Three-dimensional structure of endothiapepsin complexed with a transition-state isostere inhibitor of renin at 1.6A resolution. *J. Mol. Biol.* 1017–1029.

[20] James, M. N., Sielecki, A. R., Hayakawa, K., and Gelb, M. H. (1992) Crystallographic analysis of transition state mimics bound to penicillopepsin: difluorostatine- and difluorostatone-containing peptides. *Biochem. 31*, 3872–3886.

[21] Foltmann, B., Pedersen, V. B., Kauffman, D., and Wybrandt, G. (1979) The primary structure of calf chymosin. *J. Biol. Chem. 254*, 8447–8456.

[22] Darke, P. (1988) HIV-1 protease specificity of peptide cleavage is sufficient for processing of gag and pol polyproteins. *Biochem. Biophys. Res. Comm. 156*, 297–303.

[23] Foltmann, B. (1966) A review on prorennin and rennin. *Chem. Rev. 35*, 143–231.

[24] Hansen, E. C., and Miller, A. K. *Practical studies in fermentation: Being contributions to the life history of micro-organisms*; E. & F. N. Spon, 1896.

[25] Burkhalter, G. (1981) IDF-Catalogue of cheeses. *FIL-IDF Bulletin* 3.

[26] Foltmann, B., Lonblad, P., and Axelsen, N. H. (1978) Demonstration of chymosin (EC 3.4.23.4) in the stomach of new born pig. *Biochem. 160*, 425–427.

[27] Hidaka, M., Sasaki, K., Uozumi, T., and Beppu, T. (1986) Cloning and structural analysis of the calf prochymosin gene. *Gene 43*, 197–203.

[28] Moir, D. T., Moa, J., Duncan, M. J., Smith, R. A., and Kohno, T. In *Production of calf chymosin by the yeast Saccharomyces cerevisiae*; Underkofler, L., Ed.; 1985; Vol. 26; pp 75–85.

[29] Danley, D. E., and Geoghegan, K. F. (1988) Structure and mechanism of formation of recombinant-derived chymosin C. *J. Biol. Chem. 263*, 9785–9789.

[30] Blundell, T. L., Jenkins, J. A., Pearl, L. H., Sewell, B. T., and Pedersen, V. In *Aspartic Proteinases and Their Inhibitors*; Kostka, V., Ed.; Walter de Gruyter, 1985; Chapter The high resolution structure of endothiapepsin, pp 151–161.

[31] Newman, M., Safro, M., Frazao, C., Khan, G., Zdanov, A., Tickle, I. J., Blundell, T. L., and Andreeva, N. (1991) X-ray analyses of aspartic proteinases. IV. Structure and refinement at 2.2 A resolution of bovine chymosin. *J. Mol. Biol. 221*, 1295–1309.

[32] Blundell, T. L., Jenkins, J. A., Sewell, B. T., Pearl, L. H., Cooper, J. B., Tickle, I. J., Veerapandian, B., and Wood, S. P. (1990) X-ray analyses of aspartic proteinases: the three-dimensional structure at 2.1 A resolution of endothiapepsin. *J. Mol. Biol. 1*, 919–941.

[33] Sielecki, A. R., Hayakawa, K., Fujinaga, M., Murphy, M. E. P., Fraser, M., Muir, A. K., Carilli, C. T., Lewicki, J. A., Baxter, J. D., and James, M. N. G. (1989) Structure of recombinant human renin, a target for cardiovascular-active drugs, at 2.5 A resolution. *Science 5*, 1346–1351.

[34] James, M. N. G., and Sielecki, A. R. (1983) Structure and refinement of penicillopepsin at 1.8 A resolution. *J. Mol. Biol. 8*, 299–361.

[35] James, M. N. G., and Sielecki, A. R. (1986) Molecular structure of an aspartic proteinase zymogen, porcine pepsinogen, at 1.8 A resolution. *Nature 8*, 33–38.

[36] Hartsuck, J. A., and Remington, S. J. *Linderstrom-Lang conference proceeding*; Elsinore: Denmark, 1988; Vol. 18; Chapter Porcine pepsinogen crystallography.

[37] Andreeva, N. S., Zdanov, A. S., Gustchina, A. E., and Fedorov, A. A. (1984) Structure of ethanol-inhibited porcine pepsin at 2A resolution and binding of the methyl ester of phenylalanyl-diiodotyrosine to the enzyme. *J. Biol. Chem. 259*, 11353–11365.

[38] Abad-Zapatero, C., Rydel, T. J., and Erickson, J. (1990) Revised 2.3A structure of porcine pepsin: evidence for a flexible subdomain. *Prot.: Struc., Func. Gen. 3*, 62–81.

[39] Sielecki, A. R., Fedorov, A. A., Boodhoo, A., Andreeva, N. S., and James, M. N. G. (1990) The molecular and crystal structures of monoclinic porcine pepsin refined at 1.8A resolution. *J. Mol. Biol. 8*, 143–170.

[40] Lapatto, R., Blundell, T., Hemmings, A., Overington, J., Wilderspin, A., Wood, S., Merson, J. R., Whittle, P. J., Danley, D. E., Geoghegan, K. F., Hawrylik, S. J., Lee, S. E., Scheld, K. G., and Hobart, P. M. (1989) X-ray analysis of HIV-I proteinase at 2.7A resolution confirms structural homology among retroviral enzymes. *Nature 7*, 299–302.

[41] Miller, M., Jaskalski, M., Rao, J. K. M., Leis, J., and Wlodawer, A. (1989) Crystal structure of a retroviral protease proves relationship to aspartic protease family. *Nature 337*, 576–579.

[42] Wlodawer, A., Miller, M., Jaskalski, M., Sathyanarayana, B. K., Baldwin, E., Weber, I. T., Selk, L. M., Clawson, L., Schneider, J., and Kent, S. B. H. (1989) Conserved folding in retroviral proteases: crystal structure of a synthetic HIV-I protease. *Science 245*, 616–621.

[43] Bunn, C. W., Moews, P. C., and Baumber, M. E. (1971) The crystallography of calf rennin (chymosin). *Phil. Tranc. R. Soc. Lond.* 245–258.

[44] Gilliland, G. L., Winborne, E. L., Nachman, J., and Wlodawer, A. (1990) The three-dimensional structure of recombinant bovine chymosin at 2.3 A resolution. *Proteins 8*, 82–101.

[45] Pearl, L., and Blundell, T. (1984) The active site of aspartic proteinases. *FEBS lett. 174*, 96–101.

[46] Chitpinityol, S., and Crabbe, M. J. C. (1998) Chymosin and Aspartic Proteinases. *Food Chem. 61*, 395–418.

[47] Newman, M., Watson, F., Roychowdhury, P., Jones, H., Badasso, M., Cleasby, A., Wood, S. P., Tickle, I. J., and Blundell, T. L. (1993) X-ray analyses of aspartic proteinase: V structure and refinement at 2.0A resolution of the aspartic proteinase from Mucor pusillus. *J. Mol. Biol.* 260–283.

[48] Cooper, J. B., Khan, G., Taylor, G., Tickle, I. J., and Blundell, T. L. (1990) X-ray analyses of aspartic proteinases II: Three-dimensional structure of the hexagonal crystal form of porcine pepsin at 2.3A. *J. Mol. Biol. 3*, 199–222.

[49] Sali, A., Veerapandian, B., Cooper, J. B., Moss, D. S., Hofmann, T., and Blundell, T. L. (1992) Domain flexibility in aspartic proteinase. *Prot.: Struc., Func. Gen. 12*, 158–170.

[50] Marth, E., and Steele, J. *Applied dairy microbiology*; Marcel Dekker: New York, 1998.

[51] Andrews, A., and Varley, J. *Biochemistry of milk products*; Royal Society of Chemistry: Cambridge, 1994.

[52] Fox, P. *Advanced dairy chemistry*; Kluwer Academic/Plenum: New York, 2003.

[53] Fox, P., and McSweeney, P. *Dairy chemistry and biochemistry*; Blackie Academic and Professional: London, 1998.

[54] Foltmann, B. (1959) Studies on rennin: II on the crystallisation, stability and proteolytic activity of rennin. *Acta. Chem. Scand 13*, 1927–1935.

[55] Cheeseman, G. C. (1965) Denaturation of rennin: Effect on activity and molecular configuration. *Nature 205*, 1011–1012.

[56] Williams, M. G., Wilsher, J., Nugent, P., Mills, A., Dhanaraj, V., Fabry, M., Sedlacek, J., Uusitalo, J. M., Penttila, M. E., Pitts, J. E., and Blundell, T. L. (1997) Mutagenesis, biochemical characterization and X-ray structural analysis of point mutants of bovine chymosin. *Protein Eng. 10*, 991–997.

[57] Foltmann, B. (1959) Studies on rennin: III on the solubility of rennin. *Acta Chem. Scand. 13*, 1936–1942.

[58] Kawaguchi, Y. (1987) Production of chymosin in Escherichia coli cells and its enzymatic properties. *Agric. Biol. Chem. 51*, 1871–1877.

[59] Hill, R. D., and Laing, R. R. (1965) The action of rennin on casein: The effect of modifying functional groups on the rennin. *Biochim. Biophys. Acta 99*, 352–359.

[60] Smith, J. L. (1991) Chemical modification of amino groups in Mucor miehei aspartic proteinases, porcine pepsin, and chymosin: I Structure and function. *Agric. Biol. Chem. 55*, 2009–2016.

[61] Smith, J. L. (1991) Chemical modification of amino groups in Mucor miehei aspartic proteinases, porcine pepsin, and chymosin: II Conformation stability. *Agric. Biol. Chem. 55*, 2017–2024.

[62] Sugrue, R., Marston, F. A. O., Lowe, P. A., and Freedman, R. B. (1990) Denaturation studies on natural and recombinant bovine prochymosin (prorennin). *Biochem. 271*, 541–547.

[63] Huang, K., Zhang, Z., Liu, N., Zhang, Y., Zhang, G., and Yang, K. (1992) Functional implication of disulfide bond, Cys250 -Cys283, in bovine chymosin. *Biochem. Biophys. Res. Comm. 187*, 692–696.

[64] Barkholt, P. V., Christensen, K. A., and Foltmann, B. (1979) Investigations on the activation of bovine prochymosin. *Eur. J. Biochem 94*, 573–580.

[65] James, M. N. G., Hsu, I., and Delbaere, L. (1977) Mechanism of acid protease catalysis based on the crystal structure of penicillopepsin. *Nature 267*, 808–813.

[66] James, M. N. G., Sielecki, A., Salituro, F., Rich, D. H., and Hofmann, T. (1982) Conformational flexibility in the active sites of aspartyl proteinases revealed by a pepstatin fragment binding to penicillopepsin. *Proc. Natl. Acad. Sci. USA 79*, 6137–6141.

[67] James, M. N. G., and Sielecki, A. R. (1985) Stereochemical analysis of peptide bond hydrolysis catalyzed by the aspartic proteinase penicillopepsin. *Biochem. 24*, 3701–3713.

[68] Polgar, L. (1987) The mechanism of action of aspartic proteases involves 'push-pull' catalysis. *FEBS lett. 219*, 1–4.

[69] Fruton, S. J. (1982) Proteinase-catalysed synthesis of peptide-bonds. *Adv. Enzy. 53*, 239–306.

[70] Abdel-Malak, C. A. (1992) Calf chymosin as a catalyst of peptide synthesis. *Biochem. 288*, 941–943.

[71] Piana, S., and Carloni, P. (2000) Conformational flexibility of the catalytic Asp dyad in HIV-1 protease: An ab initio study on the free enzyme. *Proteins 39*, 26–36.

[72] Northrop, D. B. (2001) Follow the protons: a low-barrier hydrogen bond unifies the mechanisms of the aspartic proteases. *Accounts Chem. Res. 34*, 790–797.

[73] Foltmann, B., and Axelsen, N. H. In *Enzyme Regulation and Mechanism of Action*; Mildner, P., and Reis, B., Eds.; Pergamon: London, 1980; Chapter Gastric proteinases and their zymogens. Phylogenetic and developmental aspects, pp 271–280.

[74] Sielecki, A. R., Fujinaga, M., Read, R. J., and James, M. N. G. (1991) Refined structure of porcine pepsinogen at 1.8A resolution. *J. Mol. Biol. 8*, 671–692.

[75] Hartsuck, J. A., Koelsch, G., and Remington, S. J. (1992) The high-resolution crystal structure of porcine pepsinogen. *Prot.: Struc., Func. Gen. 13*, 1–25.

[76] Turk, V., Lah, T., Puizdar, V., Babnik, J., Kotnik, M., and Kregar, I. In *Aspartic Proteinases and Their Inhibitors*; Kostka, V., Ed.; Walter de Gruyter and Co.: Berlin, 1985; Chapter Cathepsins D and E: Molecular characteristics and mechanism of activation, pp 283–297.

[77] Barkholt, P. V., and Foltmann, B. (1975) Amino-acid sequence of the peptide segment liberated during activation of pro-chymosin (prorennin). *Eur. J. Biochem. 55*, 95–103.

[78] Foltmann, B. *Abstract of the Fifth International Conference on Aspartic Proteinases*; Gifu: Japan, 1993; Chapter Ontogeny and characterization of porcine prochymosin, pepsinogen B, progastricsin and pepsinogen A.

[79] Larsen, L. B., Boisen, A., and Peterson, T. (1993) Procathepsin D cannot auto-activate to cathepsin D at acid pH. *FEBS Lett. 319*, 54–58.

[80] McCamon, M. T., and Cummings, D. B. (1988) Unusual zymogen processing properties of a mutated form of pro-chymosin. *Prot.: Struc., Func. Gen. 3*, 256–261.

[81] Rand, A. G., and Ernstrom, C. A. (1964) Effect of pH and sodium chloride on activation of prorennin. *J. Dairy Sci. 47*, 1181–1187.

[82] Foltmann, B. (1962) Studies on rennin: VI the heterogeneity of prorennin and its transformation into rennin. *C. R. Travaux Lab 32*, 425–444.

[83] Strop, P., Sedlacek, J., Stys, J., Kaderabkova, Z., Blaha, I., Pavlickova, L., Pohl, J., Fabry, M., Kostka, V., and Newman, M. (1990) Engineering enzyme subsite specificity: preparation, kinetic characterization, and X-ray analysis at 2.0-A resolution of Val111Phe site-mutated calf chymosin. *Biochem. 29*, 9863–9871.

[84] Andreeva, N., Dill, J., and Gilliland, G. L. (1992) Can enzymes adopt a self-inhibited form? Results of x-ray crystallographic studies of chymosin. *Biochem. Biophys. Res. Commun. 184*, 1074–1081.

[85] Dalgleish, D. G. (1998) Casein Micelles as Colloids: Surface Structures and Stabilities. *J. Dairy Sci. 81*, 3013–3018.

[86] Groves, M. R., Dhanaraj, V., Badasso, M., Nugent, P., Pitts, J. E., Hoover, D. J., and Blundell, T. L. (1998) A 2.3 A resolution structure of chymosin complexed with a reduced bond inhibitor shows that the active site beta-hairpin flap is rearranged when compared with the native crystal structure. *Protein Eng. Des. Sel. 11*, 833–840.

[87] Plowman, J. E., Creamer, L. K., Smith, M. H., and Hill, J. P. (1997) Restrained molecular dynamics investigation of the differences in association of chymosin to k-caseins A and C. *J. Dairy Res. 64*, 299–304.

[88] Plowman, J. E., Smith, M. H., Creamer, L. K., Liddell, M. J., Coddington, J. M., Gibson, J. J., and Engelbretsen, D. R. (1994) Proton assignment and structural features of a peptide from the chymosin-sensitive region of bovine 0 false 18 pt 18 pt 0 0 false false false k-casein determined by 2D-NMR spectroscopy. *Magn. Reson. Chem. 32*, 458–464.

[89] Bairoch, A. (2005) The universal protein resource (UniProt). *Nucleic Acids Res. 33*, D154–D159.

[90] Macheboef, D., Coulon, J. B., and D'Hour, P. (1993) Effect of breed, protein genetic variants and feeding on cows' milk coagulation properties. *J. Dairy Res. 9*, 373–374.

[91] Post, A. *Fractionation of bovine casein and enrichment of functional casein peptides*; Dr. Hut.: Munchen, 2012.

[92] Doi, H., Park, B., Ibuki, F., and Kanamori, M. (1980) Heterogeneity and composition of K-casein from bovine colostrum. *Agr. Bio. Chem. 44*, 813–820.

[93] Paul, M. *Formation of bioactive peptides from dairy products*; Gene-Tech Books: New Delhi, 2007.

[94] JAMA, (1938) Formation of Milk Proteins. *J. Am Med. Ass. 111*.

[95] Palmer, D. S., Christensen, A. U., Sørensen, J., Celik, L., Qvist, K. B., and Schiøtt, B. (2010) Bovine Chymosin: A Computational Study of Recognition and Binding of Bovine -Casein. *Biochem. 49*, 2563–2573.

[96] Jensen, J., Molgaard, A., Poulsen, J., van den Brink, H., Harboe, M., Simonsen, J., Qvist, K., and Larsen, S. (2011) Structural comparison of the milk-clotting enzymes bovine and camel chymosin. *Acta Cryst. Sect. A 67*, C772–C772.

[97] Langholm Jensen, J., Mølgaard, A., Navarro Poulsen, J.-C., Harboe, M. K., Simonsen, J. B., Lorentzen, A. M., Hjernø, K., van den Brink, J. M., Qvist, K. B., and Larsen, S. (2013) Camel and bovine chymosin: the relationship between their structures and cheese-making properties. *Acta Crystallogr. Sect. D-Biol. Crystallogr. 69*, 901–913.

[98] Tang, J., Sepulveda, P., Marciniszyn, J., Chen, K. C. S., W-Y.,, Tao, N., Liu, D., and Lanier, J. P. (1973) Amino-acid sequence of porcinec pepsin. *Proc. Natl. Acad. Sci. USA 70*, 3437–3439.

[99] Baudys, M., Erdene, T. G., Kostka, V., Pavlik, M., and Foltmann, B. *Comparison between prochymosin and pepsinogen from lamb and calf*; Comp. Biochem. Physiol., 1988; pp 385–391.

[100] Jensen, T., Axelsen, N. H., and Foltmann, B. (1982) Isolation and partial characterization of prochymosin and chymosin from cat. *Biochim. Biophys. Acta 705*, 249–256.

[101] Pungercar, J., Strukelj, B., Gubensek, F., Turk, V., and Kregar, I. In *Structure and Function of Aspartic Proteinases*; Dunn, B. M., Ed.; Plenum Press: New York, 1991; Chapter Amino acid sequence of lamb pre-prochymosin and its comparison to other chymosin, pp 127–131.

[102] Shamsuzzaman, K., and Haard, N. F. (1984) Purification and characterization of a chymosin-like protease from the gastric mucosa of the harp seal (Pagphilusgroen landicus). *Can. J. Biochem. Cell Biol. 62*, 699–708.

[103] Kageyama, T. (1995) Procathepsin E and cathepsin E. *Met. Enz. 248*, 120–136.

[104] Kay, J. (1985) Aspartic proteinases and their inhibitors. *Bio-chem. Soc. Trans. 13*, 1027–1029.

[105] Doi, E., Shibata, D., Matoba, T., and Yonezawa, D. (1980) Characterization of pepstatin-sensitive acid protease in resting rice seeds. *Agric. Biol. Chem. 44*, 741–747.

[106] Morris, P. C., Miller, R. C., and Bowles, D. J. (1985) Endo-peptidase activity in dry harvest-ripe wheat and barley grains. *Pla. Sci. 39*, 121–124.

[107] Polanowski, A., Wilusz, T., Kolaczkowska, M. K., Wieczorek, M., and Wilimowska-Pelc, A. In *Aspartic Proteinases and Their Inhibitors*; Kostka, V., Ed.; Walter de Gruyter and Co.: Berlin, 1985; Chapter Purification and characterisation of aspartic proteinases from Cucumis sativus and Cucurbita maxima seeds, pp 49–52.

[108] Belozersky, M. A., Sarbakanova, S. T., and Dunaevsky, Y. E. (1989) Aspartic proteinase from wheat seeds: isolation, properties and action on gliadin. *Planta. 177*, 321–326.

[109] Pitts, J. E., Quinn, D., Uusitalo, J. M., and Penttila, M. E. (1991) Protein engineering of chymosin and expression in Trichoderma reesei. *F. Biotech. 19*, 663–666.

[110] Koaze, Y., Goi, H., Ezawa, K., Yamada, Y., and Hara, T. (1964) Fungal proteolytic enzymes. Part 1. Isolation of two kinds of acid-proteases excreted by Aspergillus niger var. macrosporus. *Agri. Biol. Chem. 28*, 216.

[111] Sardinas, J. L. (1968) Rennin enzymes of Endothia parasitica. *Appl. Microbiol. 16*, 248–253.

[112] Arima, K., Yu, J., and Lwasaki, S. (1970) Milk-clotting enzyme from Mucor pusillus var. lindt. *Met. Enz. 19*, 446–459.

[113] Sternberg, M. Z. (1971) Crystalline milk-clotting protease from Mucor miehei and some of its properties. *J. Dairy Sci. 54*.

[114] Hofmann, T., and Shaw, R. (1964) Proteolytic enzymes of Penicillium janthinellum: I Purification and properties of a trypsinogen-activation enzyme (peptidase A). *Biochim. Biophys. Acta 92*, 543–557.

[115] Fumamoto, J., Tsuru, D., and Yamamoto, T. (1967) A renin-like enzyme from Rhizopus chinensis. *Agri. Biol. Chem. 31*, 710–717.

[116] Lin, X., and Tang, J. (1990) Purification, characterization and gene cloning of thermopepsin, a thermostable acid protease from Sulfolobus acidocaldarius. *J. Biol. Chem. 265*, 1490–1495.

[117] Hube, B., Turver, C. J., Odds, F. C., Eiffert, H., Boulnois, G. J., Kochel, H., and Ruchel, R. (1991) Sequence of the Candida albicans gene encoding the secretory aspartate proteinase. *J. Med. Vet. Mycol. 29*, 129–132.

[118] MacKay, V. L., Welch, S. K., Insley, M. Y., Manney, T. R., Holly, J., Saari, G. C., and Parker, M. L. (1988) The Saccharomyces cerevisiae BAR1 gene encodes an exported protein with homology to pepsin. *Proc. Natl. Acad. Sci. USA 85*, 55–59.

[119] Yamada, T., and Ogrydziak, D. M. (1983) Extracellular acid proteases produced by Saccharomycopsis lipolytica. *J. Bact. 154*, 23–31.

[120] Toh, H., Ono, M., Saigo, K., and Miyata, T. (1985) Retroviral protease-like sequence in yeast transposon Ty1. *Nature 315*, 691.

[121] Kotler, M., Danho, W., Katz, A. A., Leis, J., and Skalka, A. M. (1989) Avian retroviral protease and cellular aspartic proteases are distinguished by activities on peptide substrates. *J. Biol. Chem. 264*, 3428–3435.

[122] Kashparov, I. V., Russ, A. V., and Andreeva, N. S. (2002) Molecular dynamics analysis of chymosin conformations in solution and in crystalline environment. *Molek. Biol. 36*, 931–938.

[123] Palmer, D., Sorensen, J., Schiott, B., and Fedorov, M. (2013) Solvent Binding Analysis and Computational Alanine Scanning of the Bovine Chymosin-Bovine -Casein Complex Using Molecular Integral Equation Theory. *J. Chem. Theory Comput. 9*, 5706–5717.

[124] Sutherland, J. J., O'Brien, L. A., and Weaver, D. F. (2004) A comparison of methods for modeling quantitative structure- activity relationships. *J. Med. Chem. 47*, 5541–5554.

[125] Free, S. M., and Wilson, J. W. (1964) A mathematical contribution to structure-activity studies. *J. Med. Chem. 7*, 395–399.

[126] Hansch, C., and Fujita, T. (1964) ρ-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc. 86*, 1616–1626.

[127] Hall, L. H., and Kier, L. B. (2007) The molecular connectivity chi indexes and kappa shape indexes in structure-property modeling. *Rev. Comput. Chem., Volume 2* 367–422.

[128] Stone, M., and Jonathan, P. (1993) Statistical thinking and technique for QSAR and related studies. Part I: General theory. *J. Chemomet. 7*, 455–475.

[129] Oprea, T. I., and Waller, C. L. (2007) Theoretical and Practical Aspects of Three-Dimensional Quantitative Structure-Activity Relationships. *Rev. Comput. Chem., Volume 11* 127–182.

[130] Hopfinger, A., Wang, S., Tokarski, J. S., Jin, B., Albuquerque, M., Madhav, P. J., and Duraiswami, C. (1997) Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. *J. Am. Chem. Soc. 119*, 10509–10524.

[131] Andrade, C. H., Pasqualoto, K. F., Ferreira, E. I., and Hopfinger, A. J. (2010) 4D-QSAR: perspectives in drug design. *Mol. 15*, 3281–3294.

[132] Cramer, R. D., Patterson, D. E., and Bunce, J. D. (1988) Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc. 110*, 5959–5967.

[133] Bradbury, S. P. (1995) Quantitative structure-activity relationships and ecological risk assessment: an overview of predictive aquatic toxicology research. *Toxicol. Lett. 79*, 229–237.

[134] Hansen, C., Telzer, B. R., and Zhang, L. (1995) Comparative QSAR in Toxicology: Examples from Teratology and Cancer Chemotherapy of Aniline Mustards. *Crit. Rev. Toxicol. 25*, 67–89.

[135] Chen, J.-Z., Han, X.-W., Liu, Q., Makriyannis, A., Wang, J., and Xie, X.-Q. (2006) 3D-QSAR Studies of Arylpyrazole Antagonists of Cannabinoid Receptor Subtypes CB1 and CB2. A Combined NMR and CoMFA Approach. *J. Med. Chem. 49*, 625–636.

[136] Perkins, R., Fang, H., Tong, W., and Welsh, W. J. (2003) Quantitative structure-activity relationship methods: Perspectives on drug discovery and toxicology. *Environ. Toxicol. Chem. 22*, 1666.

[137] Salum, L. B., and Andricopulo, A. D. (2009) Fragment-based QSAR: perspectives in drug design. *Molecular Divers. 13*, 277–285.

[138] Roy, K., Ed. *Quantitative Structure-Activity Relationships in Drug Design, Predictive Toxicology, and Risk Assessment*; IGI Global, 2015.

[139] Clark, M., Cramer, R. D., Jones, D. M., Patterson, D. E., and Simeroth, P. E. (1990) Comparative molecular field analysis (CoMFA). 2. Toward its use with 3D-structural databases. *Tetrahed. Comp. Meth. 3*, 47–59.

[140] Coats, E. A. (1998) The CoMFA steroids as a benchmark dataset for development of 3D QSAR methods. *Perspect. Drug Discov. Des. 12*, 199–214.

[141] Klebe, G., Abraham, U., and Mietzner, T. (1994) Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *J. Med. Chem. 37*, 4130–4146.

[142] Klebe, G., and Abraham, U. (1999) Comparative molecular similarity index analysis (CoMSIA) to study hydrogen-bonding properties and to score combinatorial libraries. *J. Comput. Aid. Mol. Des. 13*, 1–10.

[143] Dudek, A., Arodz, T., and Galvez, J. (2006) Computational Methods in Developing Quantitative Structure-Activity Relationships (QSAR): A Review. *Comb. Chem. High Throughput Screen 9*, 213–228.

[144] Cramer, R. D., Cruz, P., Stahl, G., Curtiss, W. C., Campbell, B., Masek, B. B., and Soltanshahi, F. (2008) Virtual Screening for R-Groups, including Predicted pIC50 Contributions, within Large Structural Databases, Using Topomer CoMFA. *J. Chem. Inf. Mod. 48*, 2180–2195.

[145] Avram, S., A.-Milac,, and Flonta, M. (2005) Computer-Aided Drug Design for Typical and Atypical Antipsychotic Drugs: A Review of Application of QSAR and Combinatorial Chemistry Methods - Tools for New Antipsychotics Design. *Curr. Comp. Aid. Drug Des. 1*, 347–364.

[146] Nunthanavanit, P., Anthony, N. G., Johnston, B. F., Mackay, S. P., and Ungwitayatorn, J. (2008) 3D-QSAR Studies on Chromone Derivatives as HIV-1 Protease Inhibitors: Application of Molecular Field Analysis. *Archiv der Pharmazie 341*, 357–364.

[147] Labrie, P., Maddaford, S. P., Fortin, S., Rakhit, S., Kotra, L. P., and Gaudreault, R. C. (2006) A Comparative Molecular Field Analysis (CoMFA) and Comparative Molecular Similarity Indices Analysis (CoMSIA) of Anthranilamide Derivatives That Are Multidrug Resistance Modulators. *J. Med. Chem. 49*, 7646–7660.

[148] Jeong, J. A., Cho, H., Jung, S. Y., Kang, H. B., Park, J. Y., Kim, J., Choo, D. J., and Lee, J. Y. (2010) 3D QSAR studies on 3,4-dihydroquinazolines as T-type calcium channel blocker by comparative molecular similarity indices analysis (CoMSIA). *Bioorg. Med. Chem. Lett. 20*, 38–41.

[149] Dayan, F. E., Singh, N., McCurdy, C. R., Godfrey, C. A., Larsen, L., Weavers, R. T., Klink, J. W. V., and Perry, N. B. (2009) β-Triketone Inhibitors of Plantp-Hydroxyphenylpyruvate Dioxygenase: Modeling and Comparative Molecular Field Analysis of Their Interactions. *J. Agric. Food Chem. 57*, 5194–5200.

[150] Chandler, D. *Introduction to Modern Statistical Mechanics*; Oxford University Press, 1988.

[151] Beglov, D., and Roux, B. (1995) Numerical-Solution Of The Hypernetted-Chain Equation For A Solute Of Arbitrary Geometry In 3 Dimensions. *J. Chem. Phys. 103*, 360–364.

[152] Beglov, D., and Roux, B. (1996) Solvation of complex molecules in a polar liquid: An integral equation theory. *J. Chem. Phys. 104*, 8678–8689.

[153] Beglov, D., and Roux, B. (1997) An integral equation to describe the solvation of polar molecules in liquid water. *J. Phys. Chem. 101*, 7821–7826.

[154] Palmer, D. S., Mišin, M., Fedorov, M. V., and Llinas, A. (2015) Fast and General Method To Predict the Physicochemical Properties of Druglike Molecules Using the Integral Equation Theory of Molecular Liquids. *Mol. Pharmaceut. 12*, 3420–3432.

[155] Gussregen, S., Matter, H., Hessler, G., Lionta, E., Heil, J., and Kast, S. M. (2017) Thermodynamic Characterization of Hydration Sites from Integral Equation-Derived Free Energy Densities: Application to Protein Binding Sites and Ligand Series. *J. Chem. Inf. Mod.*

[156] Rappe, A. K., and Casewit, C. L. *J. Stiefel Molecular Mechanics Across Chemistry*; University Science Books: California, 1997; Chapter Overview: Introduction, pp 1–4.

[157] Lubich, C. *From Quantum Mechanics to Classical Molecular Dynamics: Reduced Models and Numerical Analysis*; European Mathematical Society: Switzerland, 2008; Chapter Quantum vs. Classical Dynamics: A First Look - Classical Mechanics, p 1.

[158] Kendrick, J., Leusen, F. J. J., Neumann, M. A., and van de Streek, J. (2011) Progress in Crystal Structure Prediction. *C. E. J. 17*, 10736–10744.

[159] Laaksonen, A., and Tu, Y. In *Molecular Dynamics: From Classical to Quantum Methods*; Balbuena, P., and Serminario, J. M., Eds.; Elsevier Science: Amsterdam, 1999; Chapter Methods of Incorporating Quantum Mechanical Calculations into Molecular Dynamics Simulations: Introduction., p 1.

[160] Rappe, A. K., and Casewit, C. J. *J. Stiefel Molecular Mechanics Across Chemistry*; University Science Books: California, 1997; Chapter Overview: Bond Stretch, pp 5–6.

[161] Rapaport, D. C. *The Art of Molecular Dynamics Simulation*; Cambridge University Press, 2004; Vol. 2.

[162] Fung, Y. C., and Tong, P. *Classical and Computational Solid Mechanics*; World Scientific Printers: Singapore, 2001; Chapter Introduction: Hooke's Law, pp 2–8.

[163] Morse, P. M. (1929) Diatomic Molecules According to the Wave Mechanics. II. Vibrational Levels. *Phys. Rev. 34*, 57–64.

[164] Dauber-Osguthorpe, P., Roberts, V. A., Osguthorpe, D. J., Wolff, J., Genest, M., and Hagler, A. T. (1988) Structure and Energetics of Ligand Binding to Protiens: E Coli dihydrofolate reductase-trimethoprim, a drug-receptor system. *Prot.: Struc., Func. Gen. 4*, 31–47.

[165] Sun, H., Mumby, S. J., Maple, J. R., and Hagler, A. T. (1994) An ab initio CFF93 all-atom forcefield for Polycarbonates. *J. Am. Chem. Soc. 116*, 2978–2987.

[166] Rappe, A. K., and Casewit, C. J. *J. Stiefel Molecular Mechanics Across Chemistry*; University Science Books: California, 1997; Chapter Overview: Angle Bend, p 7.

[167] Sekercioglu, A. S., and Duysak, A. (2009) Application of molecular modeling with mass-spring. *Int. J. Phys. Sci. 4*, 500–504.

[168] Rappe, A. K., and Casewit, C. J. *J. Stiefel Molecular Mechanics Across Chemistry*; University Science Books: California, 1997; Chapter Overview: Torsion., p 8.

[169] Anslyn, E. V., and Dougherty, D. A. *J. Murdzek Modern Physical Organic Chemistry*; University Science Books: United States of America, 2006; Chapter Molecular Mechanics: Torsion, p 130.

[170] Hinchliffe, A. *Molecular Modelling for Beginners*; John Wiley and Sons Ltd: West Sussex, 2003; Chapter Molecular Mechanics: Dihedral Motions, pp 69–70.

[171] Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., Profeta, S., and Weiner, P. (1984) A New Force Field For Molecular Mechanical Simulation Of Nucleic Acids And Proteins. *J. Am. Chem. Soc. 106*, 765–784.

[172] Case, D. A. et al. AMBER. 2015.

[173] Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A., and Simmerling, C. (2006) Comparison of multiple amber force fields and development of improved protein backbone parameters. *Protein Struc. Func. Bioinf. 65*, 712–725.

[174] 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., and Kollman, P. A. (1995) A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc. 117*, 5179–5197.

[175] Comba, P., Hambley, T. W., and Martin, B. *Molecular Modeling of Inorganic Compounds*; John Wiley and Sons Ltd: West Sussex, 2009; Chapter Parameterization, Approximation and Limitations of Molecular Mechanics: Out-Of-Plane Deformations, pp 47–48.

[176] Hinchliffe, A. *Molecular Modelling for Beginners*; John Wiley and Sons Ltd: West Sussex, 2005; Chapter Molecular Mechanics: Out-of-Plane Angle Potential (Inversion), pp 70–71.

[177] Wilson, P. C., Decius, E. B., and Cross, J. C. *Molecular Vibrations*; McGraw-Hill: New York, 1955.

[178] Lee, S.-H., Palmo, K., and Krimm, S. (1999) New Out-of-Plane Angle and Bond Angle. *J. Comp. Chem. 20*, 1067–1084.

[179] Maple, J. R., Dinur, U., and Hagler, A. T. (1988) Derivatisation of Forcefields for Molecular Mechanics and Dynamics from ab initio Energy Surfaces. *Proc. Nature Ac. Sci. Uni. Sta. Am. 85*, 5350–5354.

[180] Maple, J. R., Thacher, T. S., Dinur, U., and Hagler, A. T. (1990) Biosym Forcefield Research Results in New Techniques for the Extraction of Inter- and Intramolecular Forces. *Chem. Des. Autom. News. 5*, 5–10.

[181] Mezei, M., and Beveridge, D. (1986) Free Energy Simulations. *Ann. N.Y Ac. Sci. 482*, 1–23.

[182] Ha, S. N., Giammona, A., and Field, M. (1988) A Revised Potential-Energy Surface for Molecular Mechanics Studies of Carbohydrates. *Carb. Res. 180*, 207–221.

[183] MacKerell, A. D. et al. (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B 102*, 3586–3616.

[184] Hagler, A. T., Dauber, P., and Lifson, S. (1979) Consistent forcefield studies of intermolecular forces in hydrogen bonded crystals. II. A benchmark for the objective comparison of alternative force fields. *J. Am. Chem. Soc. 101*, 5122–5130.

[185] Hagler, A. T., Dauber, P., and Lifson, S. (1979) Consistent forcefield studies of intermolecular forces in hydrogen bonded crystals. III. The C=O...H-O hydrogen bond and the analysis of the energetics and packing of carboxylic acids. *J. Am. Chem. Soc. 101*, 5131–5141.

[186] Lennard-Jones, J. E. (1924) On the Determination of Molecular Fields. *Proc. R. Soc. Lond. 106*, 463–477.

[187] Atkins, P., and de Paula, J. *Physical Chemistry for the Life Sciences*; Freeman Company, 2006; pp 469–472.

[188] Mayo, S. L., Olafson, B. D., and III., W. A. G. (1990) DREIDING: A Generic Force Field for Molecular Simulations. *J. Phys. Chem. 94*, 8897–8909.

[189] Poltev, V. *J. Leszczynski Handbook of Computational Chemistry*; Springer Science and Business Media: New York, 2012; Chapter Molecular Mechanics: Method and Application, pp 259–292.

[190] Burkert, U., and Allinger, N. L. *Molecular Mechanics ACS Monograph 177*; Am. Chem. Soc.: Washington DC, 1982.

[191] NSCCS, EPSRC UK National Service for Computational Chemistry Software. 2015.

[192] Jensen, F. *Introduction to computational chemistry*; Wiley: Chichester, 1999.

[193] Leach, A. R. *Molecular modelling*; Prentice Hall: Harlow, England, 2001.

[194] Cramer, C. J. *Essential of Computational Chemistry: Theories and Models*; John Wiley and Sons Ltd: West Sussex, 2004; Vol. 2; Chapter 2.Molecular Mechanics.

[195] Brooks, C. L., Karplus, M., and Pettitt, B. M. In *Proteins: A Theoretical Perspective of Dynamics, Structure, and Thermodynamics*; Prigogine, I., and Rice, S. A., Eds.; Advances in Chemical Physics; John Wiley & Sons, 1988; Vol. 71.

[196] Hoover, W. *Molecular dynamics*; Springer-Verlag: Berlin, 1986.

[197] Yonezawa, F. *Molecular dynamics simulations*; Springer-Verlag: Berlin, 1992.

[198] Goldstein, H. *Classical mechanics*; Addison-Wesley Press: Cambridge, 1950.

[199] Young, D. *Computational chemistry*; Wiley: New York, 2001.

[200] Har, J., and Tamma, K. *Advanced computational dynamics of particles, materials and structures*; John Wiley and Sons: West Sussex, England, 2012.

[201] Gibbs, J. (1878) On the equilibrium of heterogeneous substances. *Am. J. Sci. 16*, 441–458.

[202] Ciccotti, G., and Hoover, W. *Molecular-dynamics simulation of statistical-mechanical systems*; North-Holland: Amsterdam, 1986.

[203] Ewald, P. P. (1921) Die Berechnung optischer und elektrostatischer Gitterpotentiale. *Ann. Phys. 64*, 253–287.

[204] Darden, T., York, D., and Pedersen, L. (1993) Particle mesh Ewald: An N log (N) method for Ewald sums in large systems. *J. Chem. Phys. 98*, 10089–10092.

[205] York, D. M., Wlodawer, A., Pedersen, L. G., and Darden, T. A. (1994) Atomic-level accuracy in simulations of large protein crystals. *P. Natl. Acad. Sci. USA 91*, 8715–8718.

[206] Ryckaert, J.-P., Ciccotti, G., and Berendsen, H. J. C. (1977) Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes. *J. Comp. Phys. 23*, 327–341.

[207] Weinbach, Y., and Elber, R. (2005) Revisiting and parallelizing SHAKE. *J. Comput. Phys. 209*, 193–206.

[208] Imai, T., Hiraoka, R., Kovalenko, A., and Hirata, F. (2005) Water molecules in a protein cavity detected by a statistical-mechanical theory. *J. Am. Chem. Soc. 127*, 15334–15335.

[209] Imai, T., Hiraoka, R., and A Kovalenko, F. H. (2006) Locating missing water molecules in protein cavities by the three-dimensional reference interaction site model theory of molecular solvation. *Proteins 66*, 804–813.

[210] Yokogawa, D., Sato, H., and Sakaki, S. (2009) The position of water molecules in Bacteriorhodopsin: A three-dimensional distribution function study. *J. Mol. Liq. 147*, 112–116.

[211] Sindhikara, D. J., Yoshida, N., and Hirata, F. (2012) Placevent: An algorithm for prediction of explicit solvent atom distribution-Application to HIV-1 protease and F-ATP synthase. *J. Comput. Chem. 33*, 1536–1543.

[212] Stumpe, M. C., Blinov, N., Wishart, D., Kovalenko, A., and Pande, V. S. (2011) Calculation of Local Water Densities in Biological Systems: A Comparison of Molecular Dynamics Simulations and the 3D-RISM-KH Molecular Theory of Solvation. *J. Phys. Chem. B 115*, 319–328.

[213] Watanabe, H., Welke, K., Sindhikara, D., Hegemann, P., and Elstner, M. (2013) Towards an Understanding of Channelrhodopsin Function: Simulations Lead to Novel Insights of the Channel Mechanism. *J. Mol. Bio. 425*, 1795–1814.

[214] Imai, T., Oda, K., Kovalenko, A., Hirata, F., and Kidera, A. (2009) Ligand mapping on protein surfaces by the 3D-RISM theory: toward computational fragment-based drug design. *J. Am. Chem. Soc. 131*, 12430–12440.

[215] Imai, T., Miyashita, N., Sugita, Y., Kovalenko, A., Hirata, F., and Kidera, A. (2011) Functionality Mapping on Internal Surfaces of Multidrug Transporter AcrB Based on Molecular Theory of Solvation: Implications for Drug Efflux Pathway. *J. Phys. Chem. B 115*, 8288–8295.

[216] Nikolić, D., Blinov, N., Wishart, D., and Kovalenko, A. (2012) 3D-RISM-Dock : A New Fragment-Based Drug Design Protocol. *J. Chem. Theory Comput. 8*, 3356–3372.

[217] Palmer, D., Frolov, A., Ratkova, E., and Fedorov, M. (2010) Towards a universal method for calculating hydration free energies: a 3D reference interaction site model with partial molar volume correction. *J. Phys. 22*, 492101.

[218] Sindhikara, D., Yoshida, N., Kataoka, M., and Hirata, F. (2011) Solvent penetration in photoactive yellow protein R52Q mutant: A theoretical study. *J. Mol. Liq. 162*, 120–122.

[219] Yoshida, N., Phongphanphanee, S., Maruyama, Y., Imai, T., and Hirata, F. (2006) Selective ion-binding by protein probed with the 3D-RISM theory. *J. Am. Chem. Soc. 128*, 12042–12043.

[220] Phongphanphanee, S., Yoshida, N., and Hirata, F. (2009) The potential of mean force of water and ions in aquaporin channels investigated by the 3D-RISM method. *J. Mol. Liq. 147*, 107–111.

[221] Kiyota, Y., Yoshida, N., and Hirata, F. (2011) A New Approach for Investigating the Molecular Recognition of Protein: Toward Structure-Based Drug Design Based on the 3D-RISM Theory. *J. Chem. Theory Comput. 7*, 3803–3815.

[222] Sindhikara, D., and Hirata, F. (2013) Analysis of Biomolecular Solvation Sites by 3D-RISM Theory. *J. Phys. Chem. B 117*, 6718–6723.

[223] Du, Q. H., Beglov, D., and Roux, B. (2000) Solvation free energy of polar and nonpolar molecules in water: An extended interaction site integral equation theory in three dimensions. *J. Phys. Chem. B 104*, 796–805.

[224] Hirata, F., Ed. *Molecular theory of solvation*; Kluwer Academic Publishers, Dordrecht, Netherlands, 2003.

[225] Luchko, T., Gusarov, S., Roe, D. R., Simmerling, C., Case, D. A., Tuszynski, J., and Kovalenko, A. (2010) Three-Dimensional Molecular Theory of Solvation Coupled with Molecular Dynamics in Amber. *J. Chem. Theory Comput. 6*, 607–624.

[226] Hansen, J.-P., and McDonald, I. R. *Theory of Simple Liquids, 4th ed*; Academic Press, 2000.

[227] Duh, D. M., and Haymet, A. D. J. (1995) Integral-Equation Theory For Uncharged Liquids: The Lennard-Jones Fluid And The Bridge Function. *J. Chem. Phys. 103*, 2625–2633.

[228] Chuev, G. N., Vyalov, I., and Georgi, N. (2013) Extraction of atom-atom bridge and direct correlation functions from molecular simulations: A test for ambient water. *Chem. Phys. Lett. 561-562*, 175–178.

[229] Kovalenko, A., and Hirata, F. (1999) Potential of mean force between two molecular ions in a polar molecular solvent: A study by the three-dimensional reference interaction site model. *J. Phys. Chem. B 103*, 7942–7957.

[230] Kast, S. M., and Kloss, T. (2008) Closed-form expressions of the chemical potential for integral equation closures with certain bridge functions. *J. Chem. Phys. 129*, 236101.

[231] Ratkova, E. L., Palmer, D. S., and Fedorov, M. V. (2015) Solvation Thermodynamics of Organic Molecules by the Molecular Integral Equation Theory: Approaching Chemical Accuracy. *Chemical Reviews 115*, 6312–6356.

[232] Giambaşu, G. M., Luchko, T., Herschlag, D., York, D. M., and Case, D. A. (2014) Ion Counting from Explicit-Solvent Simulations and 3D-RISM. *Biophys. J. 106*, 883–894.

[233] Luchko, T., Joung, I. S., and Case, D. A. *RSC Biomolecular Sciences*; Royal Society of Chemistry, 2012; pp 51–86.

[234] Heil, J., Tomazic, D., Egbers, S., and Kast, S. M. (2014) Acidity in DMSO from the embedded cluster integral equation quantum solvation model. *J. Mol. Mod. 20*.

[235] Frach, R., and Kast, S. M. (2014) Solvation Effects on Chemical Shifts by Embedded Cluster Integral Equation Theory. *J. Phys. Chem. A 118*, 11620–11628.

[236] Joung, I. S., Luchko, T., and Case, D. A. (2013) Simple electrolyte solutions: Comparison of DRISM and molecular dynamics results for alkali halide solutions. *J. Chem. Phys. 138*, 044103.

[237] Palmer, D. S., Sergiievskyi, V. P., Jensen, F., and Fedorov, M. V. (2010) Accurate calculations of the hydration free energies of druglike molecules using the reference interaction site model. *J. Chem. Phys. 133*, 044104.

[238] Ratkova, E. L., Chuev, G. N., Sergiievskyi, V. P., and Fedorov, M. V. (2010) An Accurate Prediction of Hydration Free Energies by Combination of Molecular Integral Equations Theory with Structural Descriptors. *J. Phys. Chem. B 114*, 12068–12079.

[239] Genheden, S., Luchko, T., Gusarov, S., Kovalenko, A., and Ryde, U. (2010) An MM/3D-RISM Approach for Ligand Binding Affinities. *J. Phys. Chem. B 114*, 8505–8516.

[240] Ten-no, S., Jung, J., Chuman, H., and Kawashima, Y. (2010) Assessment of free energy expressions in RISM integral equation theory: theoretical prediction of partition coefficients revisited. *Mol. Phys. 108*, 327–332.

[241] Misin, M., Fedorov, M. V., and Palmer, D. S. (2015) Communication: Accurate hydration free energies at a wide range of temperatures from 3D-RISM. *J. Chem. Phys. 142*, 091105.

[242] Chandler, D., Singh, Y., and Richardson, D. M. (1984) Excess Electrons In Simple Fluids .1. General Equilibrium-Theory For Classical Hard-Sphere Solvents. *J. Chem. Phys. 81*, 1975–1982.

[243] Imai, T., Harano, Y., Kovalenko, A., and Hirata, F. (2001) Theoretical study for volume changes associated with the helix-coil transition of peptides. *Biopolymers 59*, 512–519.

[244] Sergiievskyi, V. P., Jeanmairet, G., Levesque, M., and Borgis, D. (2014) Fast computation of solvation free energies with molecular density functional theory: thermodynamic-ensemble partial molar volume corrections. *J. Phys. Chem. Lett. 5*, 1935–1942.

[245] Misin, M., Fedorov, M. V., and Palmer, D. S. (2016) Hydration free energies of molecular ions from theory and simulation. *J. Phys. Chem. B 120*, 975–983.

[246] Misin, M., Palmer, D. S., and Fedorov, M. V. (2016) Predicting solvation free energies using parameter-free solvent models. *J. Phys. Chem. B*

[247] Li, B., Matveev, A. V., and Rösch, N. (2015) Three-dimensional reference interaction site model solvent combined with a quantum mechanical treatment of the solute. *Comput. Theoret. Chem. 1070*, 143–151.

[248] Palmer, D. S., Frolov, A. I., Ratkova, E. L., and Fedorov, M. V. (2011) Toward a Universal Model To Calculate the Solvation Thermodynamics of Druglike Molecules: The Importance of New Experimental Databases. *Mol. Pharmaceut. 8*, 1423–1429.

[249] Misin, M., Vainikka, P. A., Fedorov, M. V., and Palmer, D. S. (2016) Salting-out effects by pressure-corrected 3D-RISM. *J. Chem. Phys. 145*, 194501.

[250] Jorgensen, W. L. (2004) The many roles of computation in drug discovery. *Science 303*, 1813–1818.

[251] Gohlke, H., and Klebe, G. (2002) Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angew. Chemie Intern. Ed. 41*, 2645–2676.

[252] Genheden, S., and Ryde, U. (2015) The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Exp. Op. Drug Disc. 10*, 449–461.

[253] Aqvist, J., Medina, C., and Samuelsson, J. E. (1994) A New Method For Predicting Binding-Affinity In Computer-Aided Drug Design. *Protein Eng. 7*, 385–391.

[254] Srinivasan, J., Cheatham, T. E., Cieplak, P., Kollman, P. A., and Case, D. A. (1998) Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate - DNA helices. *J. Am. Chem. Soc. 120*, 9401–9409.

[255] Kollman, P. A., Massova, I., Reyes, C., Kuhn, B., Huo, S. H., Chong, L., Lee, M., Lee, T., Duan, Y., Wang, W., Donini, O., Cieplak, P., Srinivasan, J., Case, D. A., and Cheatham, T. E. (2000) Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. *Acc. Chem. Res. 33*, 889–897.

[256] Gohlke, H., and Case, D. A. (2004) Converging free energy estimates: MM-PB(GB)SA studies on the protein-protein complex Ras-Raf. *J. Comput. Chem. 25*, 238–250.

[257] Moreira, I. S., Fernandes, P. A., and Ramos, M. J. (2006) Unravelling Hot Spots: a comprehensive computational mutagenesis study. *Theoret. Chem. Acc. 117*, 99–113.

[258] Sirin, S., Kumar, R., Martinez, C., Karmilowicz, M. J., Ghosh, P., Abramov, Y. A., Martin, V., and Sherman, W. (2014) A Computational Approach to Enzyme Design: Predicting ω-Aminotransferase Catalytic Activity Using Docking and MM-GBSA Scoring. *J. Chem. Inf. Mod. 54*, 2334–2346.

[259] Homeyer, N., and Gohlke, H. (2012) Free Energy Calculations by the Molecular Mechanics Poisson-Boltzmann Surface Area Method. *Mol. Inf. 31*, 114–122.

[260] Reblova, K., Strvelcova, Z., Kulhanek, P., Besvsveova, I., Mathews, D. H., Nostrand, K. V., Yildirim, I., Turner, D. H., and Svponer, J. (2010) An RNA Molecular Switch: Intrinsic Flexibility of 23S rRNA Helices 40 and 68 5′-UAA/5′-GAN Internal Loops Studied by Molecular Dynamics Methods. *J. Chem. Theory Comput. 6*, 910–929.

[261] Sun, H., Li, Y., Shen, M., Tian, S., Xu, L., Pan, P., Guan, Y., and Hou, T. (2014) Assessing the performance of MM/PBSA and MM/GBSA methods. 5. Improved docking performance using high solute dielectric constant MM/GBSA and MM/PBSA rescoring. *Phys. Chem. Chem. Phys. 16*, 22035–22045.

[262] Hou, T., Wang, J., Li, Y., and Wang, W. (2010) Assessing the performance of the molecular mechanics/Poisson Boltzmann surface area and molecular mechanics/generalized Born surface area methods. II. The accuracy of ranking poses generated from docking. *J. Comput. Chem. 32*, 866–877.

[263] Baron, R., Setny, P., and McCammon, J. A. (2010) Water in Cavity-Ligand Recognition. *J. Am. Chem. Soc. 132*, 12091–12097.

[264] Ghahramani, Z. *Advanced Lectures on Machine Learning*; Springer Berlin Heidelberg, 2004; pp 72–112.

[265] King, R. D., Muggleton, S., Lewis, R. A., and Sternberg, M. J. (1992) Drug design by machine learning: the use of inductive logic programming to model the structure-activity relationships of trimethoprim analogues binding to dihydrofolate reductase. *P. Natl. Acad. Sci. 89*, 11322–11326.

[266] King, R. D., Hirst, J. D., and Sternberg, M. J. E. (1993) New approaches to QSAR: Neural networks and machine learning. *Perspect. Drug Discov. Des. 1*, 279–290.

[267] Nordhausen, K. (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition by Trevor Hastie, Robert Tibshirani, Jerome Friedman. *Int. Stat. Rev. 77*, 482–482.

[268] Oladipupo, T. *New Advances in Machine Learning*; InTech, 2010; pp 19–48.

[269] Friedman, J., Hastie, T., and Tibshirani, R. (2010) Regularization paths for generalized linear models via coordinate descent. *J. Stat. Soft. 33*, 1.

[270] Abdi, H. (2010) Partial least squares regression and projection on latent structure regression (PLS Regression). *Wiley Interdis. Rev. Comput. Stat. 2*, 97–106.

[271] Jolliffe, I. T. *Principal Component Analysis*; Springer New York, 1986; pp 115–128.

[272] Kuhn, M., and Johnson, K. *Applied Predictive Modeling*; Springer New York, 2013.

[273] Kumar, S., Mohri, M., and Talwalkar, A. *Ensemble Machine Learning*; Springer US, 2012; pp 203–223.

[274] Conrads, M., Nordin, P., and Banzhaf, W. *Lecture Notes in Computer Science*; Springer Berlin Heidelberg, 1998; pp 113–129.

[275] Bies, R. R., Muldoon, M. F., Pollock, B. G., Manuck, S., Smith, G., and Sale, M. E. (2006) A Genetic Algorithm-Based, Hybrid Machine Learning Approach to Model Selection. *J. Pharmacokin. Pharmacodyn. 33*, 195–221.

[276] Schmitt, L. M., Nehaniv, C. L., and Fujii, R. H. (1998) Linear analysis of genetic algorithms. *Theoret. Comp. Sci. 200*, 101–134.

[277] Schmitt, L. M. (2001) Theory of genetic algorithms. *Theoret. Comp. Sci. 259*, 1–61.

[278] Schmitt, L. M. (2004) Theory of Genetic Algorithms II: models for genetic operators over the string-tensor representation of populations and convergence to global optima for arbitrary fitness function under scaling. *Theoret. Comp. Sci. 310*, 181–231.

[279] Breiman, L. (2001) Random forests. *Machine learning 45*, 5–32.

[280] Palmer, D. S., O'Boyle, N. M., Glen, R. C., and Mitchell, J. B. O. (2007) Random forest models to predict aqueous solubility. *J. Chem. Inf. Model. 47*, 150–158.

[281] Dunn, B. M. (2002) Structure and mechanism of the pepsin-like family of aspartic peptidases. *Chem. Rev. 102*, 4431–4458.

[282] Piana, S., and Laio, A. (2007) A bias-exchange approach to protein folding. *J. Phys. Chem. B 111*, 4553–4559.

[283] Baftizadeh, F., Cossio, P., Pietrucci, F., and Laio, A. (2012) Protein folding and ligand-enzyme binding from bias-exchange metadynamics simulations. *Curr. Phys. Chem. 2*, 79–91.

[284] Corbi-Verge, C., Marinelli, F., Zafra-Ruano, A., Ruiz-Sanz, J., Luque, I., and Faraldo-Gómez, J. D. (2013) Two-state dynamics of the SH3-SH2 tandem of Abl kinase and the allosteric role of the N-cap. *Proc. Natl. Acad. Sci. U.S.A. 110*, E3372–E3380.

[285] Marinelli, F., Kuhlmann, S. I., Grell, E., Kunte, H.-J., Ziegler, C., and Faraldo-Gómez, J. D. (2011) Evidence for an allosteric mechanism of substrate release from membrane-transporter accessory binding proteins. *Proc. Natl. Acad. Sci. U.S.A. 108*, E1285–E1292.

[286] Case, D. A. et al. AMBER 9. 2006.

[287] Lindorff-Larsen, K., Piana, S., Palmo, K., Maragakis, P., Klepeis, J. L., Dror, R. O., and Shaw, D. E. (2010) Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins 78*, 1950–1958.

[288] Sørensen, J., Palmer, D. S., and Schiøtt, B. (2013) Hot-spot mapping of the interactions between chymosin and bovine κ-casein. *J. Agr. Food Chem. 61*, 7949–7959.

[289] Sørensen, J., Palmer, D. S., Qvist, K. B., and Schiøtt, B. (2011) Initial stage of cheese production: A molecular modeling study of bovine and camel chymosin complexed with peptides from the chymosin-sensitive region of κ-casein. *J. Agr. Food Chem. 59*, 5636–5647.

[290] Prasad, B. V., and Suguna, K. (2002) Role of water molecules in the structure and function of aspartic proteinases. *Acta Crystallogr. Sect. D-Biol. Crystallogr. 58*, 250–259.

[291] Veerapandian, B., Cooper, J. B., Sali, A., Blundell, T. L., Rosati, R. L., Dominy, B. W., Damon, D. B., and Hoover, D. J. (1992) Direct observation by X-ray analysis of the tetrahedral "intermediate" of aspartic proteinases. *Protein Sci. 1*, 322–328.

[292] Piana, S., Sebastiani, D., Carloni, P., and Parrinello, M. (2001) Ab initio molecular dynamics-based assignment of the protonation state of pepstatin A/HIV-1 protease cleavage site. *J. Am. Chem. Soc. 123*, 8730–8737.

[293] Piana, S., Bucher, D., Carloni, P., and Rothlisberger, U. (2004) Reaction mechanism of HIV-1 protease by hybrid Car-Parrinello/Classical MD simulations. *J. Phys. Chem. B 108*, 11139–11149.

[294] Northrop, D. B. (2001) Follow the protons: A low-barrier hydrogen bond unifies the mechanisms of the aspartic proteases. *Acc. Chem. Res. 34*, 790–797.

[295] Coates, L., Erskine, P. T., Wood, S. P., Myles, D. A. A., and Cooper, J. B. (2001) A neutron Laue diffraction study of endothiapepsin: Implications for the Aspartic Proteinase Mechanism. *Biochem. 40*, 13149–13157.

[296] Bas, D. C., Rogers, D. M., and Jensen, J. H. (2008) Very fast prediction and rationalization of pKa values for proteinligand complexes. *Protein Struc. Func. Bioinf. 73*, 765–783.

[297] Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kale, L., and Schulten, K. (2005) Scalable molecular dynamics with NAMD. *J. Comput. Chem. 26*, 1781–1802.

[298] Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. (1983) Comparison Of Simple Potential Functions For Simulating Liquid Water. *J. Chem. Phys. 79*, 926–935.

[299] Case, D. A. et al. AMBER 2015. 2015.

[300] Hoover, W. G. (1985) Canonical Dynamics: Equilibrium Phase-Space Distributions. *Phys. Rev. A 31*, 1695–1697.

[301] Nose, S., and Klein, M. L. (1983) Constant pressure molecular dynamics for molecular systems. *Mol. Phys. 50*, 1055–1076.

[302] Martyna, G. J., Tobias, D. J., and Klein, M. L. (1994) Constant pressure molecular dynamics algorithms. *J. Chem. Phys. 101*, 4177–4189.

[303] Baftizadeh, F., Cossio, P., Pietrucci, F., and Laio, A. (2012) Protein Folding and Ligand-Enzyme Binding from Bias-Exchange Metadynamics Simulations. *Curr. Phys. Chem. 2*, 79–91.

[304] Bonomi, M., Branduardi, D., Bussi, G., Camilloni, C., Provasi, D., Raiteri, P., Donadio, D., Marinelli, F., Pietrucci, F., Broglia, R. A., and Parrinello, M. (2009) PLUMED: A portable plugin for free-energy calculations with molecular dynamics. *Comput. Phys. Commun. 180*, 1961–1972.

[305] Biarnés, X., Pietrucci, F., Marinelli, F., and Laio, A. (2012) METAGUI. A VMD interface for analyzing metadynamics and molecular dynamics simulations. *Comput. Phys. Commun. 183*, 203–211.

[306] Lange, O. F., and Grubmüller, H. (2006) Generalized correlation for biomolecular dynamics. *Proteins 62*, 1053–1061.

[307] Kabsch, W., and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers 22*, 2577–2637.

[308] Mahanti, M., Bhakat, S., Nilsson, U. J., and Söderhjelm, P. (2016) Flap dynamics in aspartic proteases: A computational perspective. *Chem. Biol. Drug Des. 88*, 159–177.

[309] Hornak, V., Okur, A., Rizzo, R. C., and Simmerling, C. (2006) HIV-1 protease flaps spontaneously open and reclose in molecular dynamics simulations. *Proc. Natl. Acad. Sci. USA 103*, 915–920.

[310] Blinov, N., Dorosh, L., Wishart, D., and Kovalenko, A. (2010) Association Thermodynamics and Conformational Stability of beta-Sheet Amyloid beta(17-42) Oligomers: Effects of E22Q (Dutch) Mutation and Charge Neutralization. *Biophys. J. 98*, 282–296.

[311] Kovalenko, A., Ten-No, S., and Hirata, F. (1999) Solution of three-dimensional reference interaction site model and hypernetted chain equations for simple point charge water by modified method of direct inversion in iterative subspace. *J. Comput. Chem. 20*, 928–936.

[312] Jensen, F., and Palmer, D. S. (2011) Harmonic Vibrational Analysis in Delocalized Internal Coordinates. *J. Chem. Theory Comput. 7*, 223–230.

[313] Massova, I., and Kollman, P. A. (1999) Computational Alanine Scanning To Probe Protein-Protein Interactions: A Novel Approach To Evaluate Binding Free Energies. *J. Am. Chem. Soc. 121*, 8133–8143.

[314] Bradshaw, R. T., Patel, B., Tate, E., Leatherbarrow, R., and Gould, I. (2011) Comparing experimental and computational alanine scanning techniques for probing a prototypical protein-protein interaction. *Protein Eng. Des. Sel. 24*, 197–207.

[315] Huo, S., Massova, I., and Kollman, P. (2002) Computational alanine scanning of the 1 : 1 human growth hormone-receptor complex. *J. Comput. Chem. 23*, 15–27.

[316] Moreira, I., Fernandes, P., and Ramos, M. (2007) Computational alanine scanning mutagenesis - An improved methodological approach. *J. Comput. Chem. 28*, 644–654.

[317] White, A. D., Keefe, A. J., Ella-Menye, J.-R., Nowinski, A. K., Shao, Q., Pfaendtner, J., and Jiang, S. (2013) Free Energy of Solvated Salt Bridges: A Simulation and Experimental Study. *J. Phys. Chem. B 117*, 7254–7259.

[318] Ansari, S., Coletta, A., Skeby, K. K., Sørensen, J., Schiøtt, B., and Palmer, D. S. (2016) Allosteric-Activation Mechanism of Bovine Chymosin Revealed by Bias-Exchange Metadynamics and Molecular Dynamics Simulations. *J. Phys. Chem. B 120*, 10453–10462.

[319] Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N., and Yeh, L.-S. L. (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res. 33*, D154–D159.

[320] Riniker, S. (2017) Molecular Dynamics Fingerprints (MDFP): Machine Learning from MD Data To Predict Free-Energy Differences. *J. Chem. Inf. Model. 57*, 726–741.

[321] Chandler, D., and Andersen, H. C. (1972) Optimized cluster expansions for classical fluids. 2. Theory of molecular liquids. *J. Chem. Phys. 57*, 1930–1937.

[322] DePriest, S. A., Mayer, D., Naylor, C. B., and Marshall, G. R. (1993) 3D-QSAR of angiotensin-converting enzyme and thermolysin inhibitors: a comparison of CoMFA models based on deduced and experimentally determined active site geometries. *J. Am. Chem. Soc. 115*, 5372–5384.

[323] Golbraikh, A. (2000) Validation of protein-based alignment in 3D quantitative structure–activity relationships with CoMFA models. *European J. Med. Chem. 35*, 123–136.

[324] Maddalena, D. J., and Johnston, G. A. R. (1995) Prediction of Receptor Properties and Binding Affinity of Ligands to Benzodiazepine/GABAA Receptors Using Artificial Neural Networks. *J. Med. Chem. 38*, 715–724.

[325] Chavatte, P., Yous, S., Marot, C., Baurin, N., and Lesieur, D. (2001) Three-Dimensional Quantitative Structure-Activity Relationships of Cyclo-oxygenase-2 (COX-2) Inhibitors: A Comparative Molecular Field Analysis. *J. Med. Chem. 44*, 3223–3230.

[326] Mattioni, B. E., and Jurs, P. C. (2003) Prediction of dihydrofolate reductase inhibition and selectivity using computational neural networks and linear discriminant analysis. *J. Mol. Graph. Mod. 21*, 391–419.

[327] Lajiness, M., Johnson, M., and Maggiora, G. (1989) Implementing drug screening programs using molecular similarity methods. *Prog. Clin. Biol. Res. 291*, 173.

[328] Hassan, M., Bielawski, J. P., Hempel, J. C., and Waldman, M. (1996) Optimization and visualization of molecular diversity of combinatorial libraries. *Molecular Divers. 2*, 64–74.

[329] Perkyns, J. S., and Pettitt, B. M. (1992) A Dielectrically Consistent Interaction Site Theory For Solvent Electrolyte Mixtures. *Chem. Phys. Lett. 190*, 626–630.

[330] Lue, L., and Blankschtein, D. (1992) Liquid-state theory of hydrocarbon water-systemsapplication to methane, ethane, and propane. *J. Phys. Chem. 96*, 8582–8594.

[331] Berendsen, H. J. C., Grigera, J. R., and Straatsma, T. P. (1987) The missing term in effective pair potentials. *J. Phys. Chem. 91*, 6269–6271.

[332] Hirata, F., and Rossky, P. J. (1981) An Extended Rism Equation For Molecular Polar Fluids. *Chem. Phys. Lett. 83*, 329–334.

[333] Lee, P. H., and Maggiora, G. M. (1993) Solvation Thermodynamics Of Polar-Molecules In Aqueous-Solution By The Xrism Method. *J. Phys. Chem. 97*, 10175–10185.

[334] Kovalenko, A., and Hirata, F. (2000) Hydration free energy of hydrophobic solutes studied by a reference interaction site model with a repulsive bridge correction and a thermodynamic perturbation method. *J. Chem. Phys. 113*, 2793–2805.

[335] Chuev, G., Fedorov, M., and Crain, J. (2007) Improved estimates for hydration free energy obtained by the reference interaction site model. *Chem. Phys. Lett. 448*, 198–202.

[336] Allen, M. P., and Tildesley, D. J., Eds. *Computer Simulation of Liquids*; Clarendon Press, Oxford, 1987.

[337] Wehrens, R., and Mevik, B.-H. (2007) The pls package: principal component and partial least squares regression in R.

[338] Team, R. C. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2014. 2014.

[339] Liaw, A., and Wiener, M. (2002) Classification and regression by randomForest. *R news 2*, 18–22.

[340] Eklund, M., Norinder, U., Boyer, S., and Carlsson, L. (2014) Choosing feature selection and learning algorithms in QSAR. *J. Chem. Inf. Mod. 54*, 837–843.

[341] R Development Core Team, R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing: Vienna, Austria, 2012.

[342] Sergiievskyi, V. P., Hackbusch, W., and Fedorov, M. V. (2011) Multigrid Solver for the Reference Interaction Site Model of Molecular Liquids Theory. *J. Comput. Chem. 32*, 1982–1992.

[343] Sergiievskyi, V., and Fedorov, M. (2012) 3DRISM Multigrid Algorithm for Fast Solvation Free Energy Calculations. *J. Chem. Theory. Comput. 8*, 2062–2070.

[344] Golbraikh, A., and Tropsha, A. (2002) Beware of Q2! *J. Mol. Graph. Mod. 20*, 269–276.