COMPUTATIONAL STUDIES OF BIOMOLECULES

Sih-Yu Chen

A Thesis Submitted for the Degree of PhD
at the
University of St Andrews

2017

Computational Studies of Biomolecules

Sih-Yu Chen

This thesis is submitted in partial fulfilment for the degree of
Doctor of Philosophy

School of Chemistry
University of St Andrews

May 2017

1. Candidate's declarations:

I, Sih-Yu Chen, hereby certify that this thesis, which is approximately 43800 words in length, has been written by me, and that it is the record of work carried out by me, or principally by myself in collaboration with others as acknowledged, and that it has not been submitted in any previous application for a higher degree.

I was admitted as a research student in September 2012; the higher study for which this is a record was carried out in the University of St Andrews between 2012 and 2016.

I, Sih-Yu Chen, received assistance in the writing of this thesis in respect of grammar and spelling, which was provided by Miss Rachael Skyner and Dr Luke Crawford.

Date Signature of candidate

2. Supervisor's declaration:

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of Doctor of Philosophy (PhD) in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree.

Date Signature of supervisor

3. Permission for publication:

No embargo on printed or electronic copy.

Date Signature of candidate

Date Signature of supervisor

ACKNOWLEDGEMENTS

My deepest gratitude goes first and foremost to my supervisors, Dr John Mitchell and Prof. Nicholas Westwood, for the tremendous opportunities I had to work with them and for their support and guidance during my studies, which was an extraordinary experience in my life.

To my family, and particularly my parents for giving me the time and financial support, and their unconditional love.

To Dr Jan Goetze for numerous discussions, emails and encouragements which were of enormous help.

To Dr Carlos Penedo-Esteiro, who I have been fortunate to work with on one of the projects of this thesis.

To Rachael, for her support and encouragement. especially proofreading of thesis, and to my many friends: Jasmine, Chengsi, James, Luke, Rosie, Leo, Ludovic, Bilal and Rajamalli and colleagues in Purdie building room 150.

To Dr. Herbert Fruchtl and Professor Michael Buehl for the computational source and Dr Tanja van Mourik for advice in the course of my research.

PUBLICATION

Chen, A. S. Y., Westwood, N. J., Brear, P., Rogers, G. W., Mavridis, L., & Mitchell, J. B. O. (2016). A Random Forest Model for Predicting Allosteric and Functional Sites on Proteins. *Molecular Informatics*. 35, 125-135.

ABSTRACT

In modern drug discovery, lead discovery is a term used to describe the overall process from hit discovery to lead optimisation, with the goal being to identify drug candidates. This can be greatly facilitated by the use of computer-aided (or in silico) techniques, which can reduce experimentation costs along the drug discovery pipeline. The range of relevant techniques include: molecular modelling to obtain structural information, molecular dynamics (which will be covered in Chapter 2), activity or property prediction by means of quantitative structure activity/property relationship models (QSAR/QSPR), where machine learning techniques are introduced (to be covered in Chapter 1) and quantum chemistry, used to explain chemical structure, properties and reactivity.

This thesis is divided into five parts. **Chapter 1** starts with an outline of the early stages of drug discovery; introducing the use of virtual screening for hit and lead identification. Such approaches may roughly be divided into structure-based (docking, by far the most often referred to) and ligand-based, leading to a set of promising compounds for further evaluation. Then, the use of machine learning techniques, the issue of which will be frequently encountered, followed by a brief review of the "no free lunch" theorem, that describes how no learning algorithm can perform optimally on all problems. This implies that validation of predictive accuracy in multiple models is required for optimal model selection. As the dimensionality of the feature space increases, the issue referred to as "the curse of dimensionality" becomes a challenge. In closing, the last sections focus on supervised classification Random Forests. Computer-based analyses are an integral part of drug discovery.

Chapter 2 begins with discussions of molecular docking; including strategies incorporating protein flexibility at global and local levels, then a specific focus on an automated docking program – AutoDock, which uses a Lamarckian genetic algorithm and empirical binding free energy function. In the second part of the chapter, a brief introduction of molecular dynamics will be given.

Chapter 3 describes how we constructed a dataset of known binding sites with co-crystallised ligands, used to extract features characterising the structural and chemical properties of the binding pocket. A machine learning algorithm was adopted to create a three-way predictive model, capable of assigning each case to one of the classes (regular, orthosteric and allosteric) for *in silico* selection of allosteric sites, and by a feature selection algorithm (Gini) to rationalize the selection of important descriptors, most influential in classifying the binding pockets.

In **Chapter 4,** we made use of structure-based virtual screening, and we focused on docking a fluorescent sensor to a non-canonical DNA quadruplex structure. The preferred binding poses, binding site, and the interactions are scored, followed by application of an ONIOM model to re-score the binding poses of some DNA-ligand complexes, focusing on only the best pose (with the lowest binding energy) from AutoDock. The use of a pre-generated conformational ensemble using MD to account for the receptors' flexibility followed by docking methods are termed “relaxed complex” schemes.

Chapter 5 concerns the BLUF domain photocycle. We will be focused on conformational preference of some critical residues in the flavin binding site after a charge redistribution has been introduced. This work provides another activation model to address controversial features of the BLUF domain.

TABLE OF CONTENTS

CHAPTER 1. Theory of Chemoinformatics Methods

Computer-aided drug design (CADD) is a relatively time saving yet cost effective procedure. This is the term generally referring to the use of computational tools and resources that help identify potential lead candidates in the drug discovery process. At the early stage of drug discovery, an ensemble of computer-based methods known as virtual (compound) screening (VS) methods, also referred to as in-silico screening, are emerging as an efficient approach to screen large compound databases. These methods aim to identify compounds with a higher probability of binding to a given target, analogously to the role of high-throughput screening (HTS) in experiments. [1] Later in the drug discovery process, different VS methods are developed as a means to narrow down the size of databases or libraries to be screened for hits or leads, such as to discard compounds which are not drug-like (i.e. do not adhere to measures such as Lipinski's Rule of Five) and those that are likely to have poor ADME (absorption, distribution, metabolism and excretion) and toxicity properties, thereby reducing the cost of *in vitro* research for synthesis or assays. In general, VS methods have been shown to help in three ways: a) to identify new active compounds as chemical starting points, b) to improve the molecules as leads and c) as a filtering method used to prioritise compounds throughout the lead optimisation process. [2]

1.1 Drug discovery

A potential drug candidate is obtained via a costly and time consuming multi-step process (Figure 1), starting from target identification and validation, where validation refers to the process of identifying a molecular target with therapeutic benefit. The primary goal of this step is to identify targets that are druggable, to which small molecules or biologic therapeutics (e.g. antibodies known to block protein-protein interactions) can bind and induce a biological response. This is where data mining comes into the picture. One relatively cost-effective way to do this is by making use of genomics in combination with bioinformatics approaches [3,4] mining available sources from publications, patent data, clinical research and gene expression data etc. This identification process also helps to uncover genes whose expression is correlated with disease progression. Using genome sequences to identify genes responsible for coding target proteins allows prioritisation of the most promising potential targets, or identification of any genetic associations, polymorphism, or mutations making a person more susceptible to the disease or to the disease progression. In addition, bioinformatics is further used in the process of validating the putative target as relevant to treatment of the disease. [5] Traditionally, the process of target validation has relied on a multi-validation of *in vitro* and animal model approaches, antisense oligonucleotides which reversibly bind to target mRNA by inhibiting the synthesis of the target protein [6], antibodies or transgenic animals to confirm a phenotypic response due to lack of a given gene. [7]

Target identification and validation have given rise to the term chemical genomics (or chemogenomics) [8], where the aim is to extend chemical genetics to the genomic scale. By analogy to classical genetics, chemical genomics uses small molecules ('chemical') to elucidate biological processes in which its effects are equivalent to mutations in classical genetics. [9] In practice, chemical genomics often focuses on a class of protein targets, allowing screenings to be performed in parallel. In "reverse" chemical genomics, libraries of small molecules are tested for their ability to specifically modulate a target of interest (from protein to phenotype), akin to the high-throughput screening (HTS) process (covered in the next paragraph), and thereby knowledge gained from informatics and data mining tools will help to identify tool molecules that can be used to validate the therapeutic hypothesis of a given drug target during the target validation process. In contrast to "forward" chemical genomics, the active compounds that produce a desired phenotypic change are identified from screening of any pool of compounds so as to discover a target (from phenotype to protein) in which no prior knowledge of the target is assumed. [10]

Once a validated target has been chosen, the next step involves compound screening using either *in vitro* or *in silico* based methods to identify "hits" in a screen of a library for compounds having the desired activity at the target of interest. *In vitro* assays are either done with a biochemical or cell-based assay based upon agreed selection criteria. A miniaturised and automated process, high throughput screening (HTS), has been employed since the 1980s [2] at the hit discovery stage, whereby large numbers of compounds are assayed against the target in a time-efficient manner. Focused or knowledge-based screening [11] is commonly adopted when the knowledge of the target structure and the binding site location from literature, from the constructed pharmacophores, molecular modelling, or from patent precedents is available to be used as the basis for a smaller subset consisting of compounds having desired chemical features. [12] There are other compound screens, such as fragment screening which uses libraries of smaller compounds or fragments, and tissue-based screens which look for desired *in vivo* effects, etc.

Often, the whole screening process can be broken down into two stages: 1) the primary screen, in which the activities of compounds are measured at a single concentration (single-dose screen) to yield "primary hits", and 2) the secondary screen, where the most active compounds are re-tested to confirm their activity. This activity can be further confirmed in follow-up screens. [13] High-throughput screening methods have a certain rate of false positive and false negative results occurring for a number of reasons: the causes of poor solubility (compound precipitation in aqueous media), low purity, incorrect compound concentration, fluctuations in environmental factors or other experimental errors. False positives can generally be recognised by the follow up confirmation tests. However, false negatives, which occur when active compounds are detected as inactive, may remain undetected. [14,15]

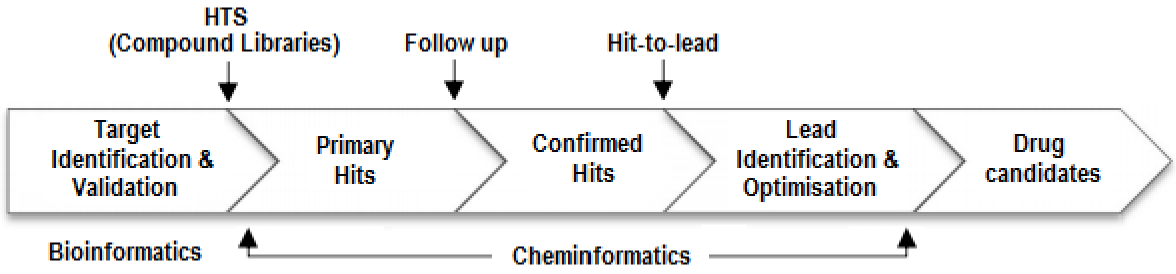


Figure 1 - Early stages of lead discovery.

The resulting hits found through virtual screening or high-throughput screening will need to be refined and prioritised. Clustering based on structural similarity is frequently applied throughout the process to ensure that compounds of diverse chemical classes are selected for consequent follow-up studies. Thereafter, compounds are analysed as groups, which could later form the basis of a lead series of compounds which share similar properties. Further studies are conducted to assess dose-response for each chemical, to identify chemicals with desired competitive behaviour. The assay is conducted for each compound at multiple concentrations and then plotted against the resulting percent inhibition, from which the concentration required for 50% inhibition can be determined, giving the IC50 value as a measure of potency for each candidate compound. At this stage, the representatives of each cluster are characterised for physicochemical properties (logP, pKa, solubility etc), ADME and pharmacokinetics (PK) using various *in vitro* assays, and checked for selectivity to other related targets using cross screening techniques to reduce off-target effects. [16] Tissue- or cell-based models are used to look at the functional response exerted by the compounds in more intact systems, however such an approach does not really replicate the true physiological environment and thus has its limitations. [17] At the end of this process, the most promising series are selected for further studies.

The compounds in a hit series appear to have different potencies due to different chemical groups attached to the core compound structure; these are used to derive structure-activity relationships (SAR) which can be used to identify substructures that may contribute to the activities. From a practical point of view, this step includes assessing whether compound series can be synthesised in parallel and thus allow a diverse set to be generated. Confirmed hits with biological activity predicted from a preconstructed SAR are termed "leads", from which researchers hope eventually to develop drug candidates. Following hit identification, this stage of the drug development process is known as "hit-to-lead", with the goal being to optimise the hits to yield compounds with improved potency and selectivity. [15] Leads are compounds that, in addition to their promising activity, should have the potential to be further developed from their relatively simple features, belong to a well-defined SAR series, possess the desired ADME properties, and have the potential to be patented. [20]

Studies of SAR are useful to establish biological activity for each hit and to determine potential structural modifications of a lead to increase its potency and selectivity. The purpose of SAR is to relate structural changes to changes in biological activity. Screening a higher number of compounds might help to identify structural features associated with certain properties. Structure-based drug design techniques and methodologies to gain structural information of the targets are able to facilitate the process of establishing SARs. Screening at this stage provides reports on the activity of the compounds and target selectivity profile which address the efficacy and off-target safety issues. Furthermore, this stage involves examining whether the compound could be active in primary assays to protein orthologues of other species. Consequently, animal models are frequently used to study the *in vivo* effects, pharmacodynamic (PD)/PK profiles and preclinical toxicity of the compounds in disease models, and in a high throughput fashion for detailed understanding of the physicochemical properties, including the solubility and permeability profile and the ADME properties. Assays are developed to examine if compounds can be introduced to the body orally and absorbed via the digestive tract, or introduced alternatively by injection directly into the bloodstream, and whose effects can be diminished as it is metabolised by enzymes and eliminated from the body. A limited number of compounds are selected for PK evaluations to study how a drug is processed by the body of animals. As a general principle, it is desirable for a candidate to be higher than 20 percent absorbed after an oral dosing, or have a half-life longer than 60 minutes, given an IV-injected dose; a concept known as “bioavailability”, which refers to the amount of a drug dose reaching to the systemic circulation (plasma). These values are typically dependent upon the targets. [17]

With the help of these assays, the initial lead compounds are selected and subjected to "lead optimisation" before being declared as a "drug candidate" in clinical trials. At this stage, further work is carried out to improve deficiencies in the lead structures while still maintaining their favourable properties. The process of further characterising the leads varies by company, but in general, the measurement and reporting of certain properties must be in accordance with regulatory bodies throughout the development process. These include the genotoxicity tests by means of *in vitro* (for example, Ames test) or *in vivo* testing that are used to reveal possible genetic damage done by the new active compounds, and others tests assess the PK/PD profiles and the PK response after repeated dosing. All nonclinical information gathered by the end of this stage will be used later, during regulatory submission for approval to test on human subjects. [17]

1.1.1 Filters for Druglikeness

As mentioned previously, *in silico* methods are used to predict properties of chemical compounds based on their structures, and the predicted results can be used to prioritise the compounds which best match the design criteria related to potency, selectivity, ADME properties, etc.; thus, to identify “leads” out of the pool of hits. Nowadays the key considerations of library design have changed to consider the druglikeness of members, rather than their size, and the diversity of the library. [18] Compounds selected for use in libraries usually adhere to "Lipinski's Rule of Five" which defines the following limiting property criteria to be satisfied as a filter for drug: a molecular weight of less than 500 daltons, an octanol-water partition coefficient (which is a measure of lipophilicity, logP) of less than 5, less than 5 hydrogen donors and 10 hydrogen acceptors. These rules were empirically derived from existing drugs that can be administered orally.

In general, the lead optimisation process is often accompanied by an increase in molecular weight and changes in logP. The lead is only used as a starting point. Because of this, Teague et al. [19] argued that the properties used to design a library of leads that will need further optimisation may be different to those for constructing a druglike library. Besides, a novel library design has to depend on the target, their routes of administration and the result of pharmacokinetics to identify proper profile. [20]

Teague et al. [19] broadly divided the leads into three types: the leadlike, the druglike and the high-affinity leads; leading to developing different optimisation scenarios. The leadlike leads consisted of small molecules with low cLogP, typically associated with low affinity are to be improved by increasing the molecular weight and lipophilicity. To convert high-affinity leads (mostly peptidic compounds with molecular weight > 350Da) into drugs with good pharmacokinetic properties, involves retaining its potency while increasing cLogP and reducing the molecular weight. The druglike leads, on the other hand perhaps are the most commonly reported type of lead emerging from HTS by filtering the combinatorial libraries; the resulting compounds tend to have low affinity. (Figure 2) They proposed that using the library of leads with molecules in the molecular weight range 100-350 and cLogP range 1-3.0 (filters defined for leadlike leads) would give results superior to the molecules in the druglike leads library, allowing additional interactions to be explored when optimising leads. In addition, smaller molecules are easier to locate at the desired binding site, and also more easily adapt to enhance selectivity, affinity, and other properties; often achieved by introducing lipophilic groups. [20,21] These preferable leadlike leads, averagely speaking, have lower molecular weight, fewer rings and rotatable bonds and are less hydrophobic and have lower polarisability. [20]

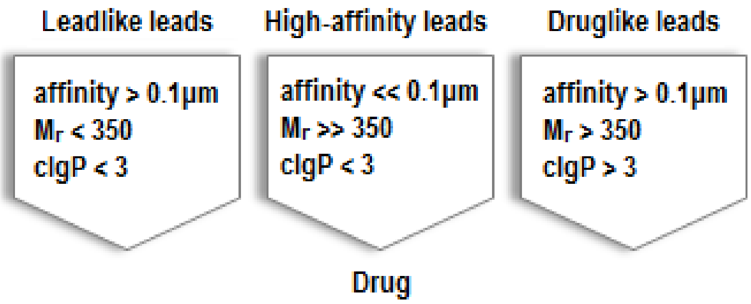


Figure 2 - Modified from Teague et al. [19] Classification of leads according to their molecular weight (Mr) and cLogP values.

1.1.2 *In silico* Virtual Screening

Virtual screening (VS) methods can generally be categorised depending on the availability of experimental data into: 1) the structure-based (SBVS) or target-based approaches, when the structure of the receptor is available or can be determined by homology modelling and 2) the ligand-based (LBVS) approaches. [1] The process of VS is often described by analogy to a funnel, where the compounds are filtered to exclude inactive molecules, or ranked based on their predicted activity as assessed using a computational algorithm. Ultimately, this process is expected to result in a more manageable set of active compounds to be tested experimentally. VS methods can be used as an alternative to HTS when no suitable assays are available (i.e. HTS assays require adequate sensitivity in identifying modulators of the enzyme activity). Compared to HTS, VS is lower in cost. Also, it is not limited to physically existing screening libraries, allowing the access to a larger chemical space. VS does not suffer from experimental deficiencies (i.e. poor solubility) that interfere with the assays and its readout. One the other hand, the success of VS depends heavily on reliable methods to predict binding modes and binding affinities. [22] Alternatively, VS methods can be complementary to HTS in identifying new hit compounds. VS can be performed prior to a HTS to enrich active compounds in a library, or after a HTS, to identify false negatives. [23] A comparison on the performance of HTS and VS (docking) against the same target has been published by Doman et al. [24]. The aim was to identify potential inhibitors of protein-tyrosine phosphatase 1B (for treatment of diabetes). In the end, they discovered two distinct sets of active compounds (hits), suggesting the complementary nature of the two methods.

Tanrikulu et al. [2] classified the applications of VS into classic VS, parallel VS, iterative VS and integrated VS on the basis of various integral strategies that exist in the literature. Parallel VS makes use of multiple VS techniques running in parallel (each of which works at its own classic VS). The resulting hits from various methods which give complementary results are combined, either directly or by a fusion method [25] to increase the true positive and reduce the false positive hits. Iterative VS applies VS sequentially to rounds of refinement processes at various stages of drug discovery. At each iteration, the *in silico* selected compounds and their similarities, identified at various threshold levels, are used as starting points. These compounds are subjected to *in vitro* evaluations where experimental results are incorporated, to improve the subsequent *in silico* model leading to the discovery of more potent hits. Integrated VS, which is found to be the most advanced application of VS, integrates a number of different validated and parameterised screening procedures into a tailor-made protocol for a specific compound or compound type. Subsequent to HTS results, compounds are re-evaluated by the VS methods and with different subsequent arrangements (so as to take advantage of their complementary nature), to reduce the false positive hits from *in vitro* screening [26] or to reduce the size of chemical space to be searched. [27]

Structure-based virtual screening (SBVS) involves using a known 3D biological receptor structure (obtained by X-ray crystallography or NMR) or 3D model (homology modelling) as a template to screen for potential binders (typically a 3D representation). Usually this is done by employing molecular docking techniques, as an estimate of how likely it is that this compound will bind to the target with high affinity [28] and in an effort to gain information on how the ligand interacts with the receptor. The structure-based approaches include molecular docking and scoring (which will be covered in

Chapter 2), molecular dynamics, pharmacophore modelling, and *de novo* ligand design methods. Since structure-based screening relies on a static structure of the target, from which ligand binding is modelled, the results of a screening are limited by (and to) fixed/rigid molecular structures. Thus, a number of strategies have been proposed to account for receptor flexibility (explored more in Chapter 2) in SBVS, to avoid biasing towards a single rigid conformation.

Ligand-based virtual screening (LBVS) emerged during the 1980s and early 1990s and uses known bioactive compounds as reference molecules (usually 2D representations, but can be 3D representations) to extract SAR. This method allows one to search for new hits sharing shape and/or pharmacophore features identified as being responsible for the activity (to identify compounds with similar bioactivity profiles). LBVS is used when 3D structural information on the target is not available. The "similar property principle" (SPP) [29], formalised by Johnson and Maggiora in 1990, provides a rationale for a structural similarity searching, which is what the ligand-based approach relies on. The principle states that similar molecules are prone to display similar biological properties. In contrast to LBVS, SBVS employs docking, and generally depends on structural complementarities between the macromolecular target and its ligand.

However, structurally similar compounds can have distinct SARs. [30] Similarity searching does not consider the stereochemistry, which can affect the ability of a molecule to bind to a target, especially where the ability of the molecule to change 'shape' is limited by strong intramolecular interaction. Consequently, ligand-based approaches were used for targeted (or focused) library design to select compounds from the initial compound collection enriched with specific properties for a target or a protein family based on the known target and literature (or patent precedents).

Approaches for ligand-based screening can be divided into similarity search and compound classification techniques. In practice, each molecule is represented using a set of descriptors encoding chemical features. Traditional virtual screening efforts focus on similarity searching using fingerprints which are binary descriptors, or various other similarity descriptors (see 'Molecular Descriptors') to provide chemical features. It is these fingerprints that are compared to perform similarity measures and will produce a similarity score (or similarity coefficient) between pairs of molecules, from which the clusters are based. Compounds from the same cluster that have a high similarity measure value are expected to interact with proteins of the same group. [31] Descriptors, calculated with the aim of predicting a given property, can be defined in terms of a subset of the chemical space (descriptor space), where each molecule is represented as a point with n-descriptor dimensions. The success of similarity searching is dependent upon the choice of this descriptor space. The best-defined are those categorised within the criteria for drug-like structures, which follow the Lipinski, the lead-like [20] and fragment-based [32] definitions. Such constraints help to reduce the chemical space, limiting the chemical space to those regions containing molecules with favourable (drug) properties; for instance, chemical space in which the active compounds reside, which is referred to as the 'biologically relevant chemical space'. (Figure 3) However, not all compounds with proved activity fall within predefined limits or other criteria (rules). [33] Compound classification techniques, on the other hand, can be divided further according to the training procedure implemented, with an ultimate goal to predict the class label for compounds using either an unsupervised (clustering) or a supervised (classification) machine learning algorithm to learn decision rules from a training set of known compounds. Then the resulting models are used to predict whether a given molecule will bind to a target on the basis of physicochemical properties. [34,35]

However, several limitations of ligand-based approaches are listed. Firstly, the rationale behind a similarity search, the similar property principle (SPP), is not always valid; in some cases, similar compounds have dissimilar properties. The SAR data can be conceptualised using an "activity landscape" model proposed by Gerald Maggiora (2006) [36], which is similar to geographical landscapes where the third dimension forms the surface of an "activity landscape" that accounts for compound biological activity, and is added to a 2D projection of chemical space. As such, it provides a graphical representation of the relationships between structural similarity and biological activity. SARs are distinguished by how molecules respond to structural modifications. In their terminology, "continuous" SARs would correspond to smooth regions or gently rolling hills of the structure activity landscape; the areas where gradual changes in chemical structure result in a small or moderate effect on biological activity, and thus which would make reliable predictions of activities (potency) for other similar compounds. [31] In an extreme case, when large changes of structure result in a very small change in activity, SARs are known as "flat", and most optimization efforts on this kind of SAR are fruitless. [37] By contrast, rugged areas represent regions of "discontinuous" SARs, where "activity cliffs" are more likely to be found. In the presence of "activity cliffs" the results appear to be inconsistent with the concept of SPP; small changes in chemical structure can result in a dramatic change in biological activity, yielding models with limited predictive power and often leading to the failure of QSAR (Quantitative Structure-Activity Relationship). [36] The presence of "activity cliffs" is a challenge to similarity-based approaches since they assume a continuous SAR. "Similarity cliffs" or scaffold hops, which in contrast to activity cliffs, occur when structures that are dissimilar but exhibit similar activities.

To put SAR in the context of target-ligand interactions, the biological activity is the result of interactions between small molecules and their biological targets, and since the ligand binding depends on chemical (i.e. hydrogen bonding, electrostatic interactions, etc.) and geometrical (shape) complementarity of ligand and receptor, which enables only certain specific interactions to occur. The "activity cliffs", here refer to "structure-based activity cliffs", in this sense, are associated with critical interactions required for the binding, regardless of ligand structure and properties, leading to the concepts of the term "activity cliff hot spots" [38,39], which are regions or atoms in the target site directly involved in interactions with the ligand, associated with the formation of activity cliffs. [40] "Continuous SAR" regions indicate permissive binding. The binding sites are able to accommodate ligand variability to some extent and such a binding would require a high degree of shape complementarity between binding site and ligand to result different potential interactions. [41] Evidence shows that different SAR types coexist at the active site. In the case of heterogeneous SARs (also termed variable activity landscapes) this coexistence comprises smooth regions intersected by (rough) activity cliffs. The SAR analysis of activity cliffs can help to drive the ligand improvement task.

Secondly, the relative positions of compounds in chemical space vary depending on the particular selection of a descriptor set defining a chemical space, and based on the chemical representation used to describe molecules, with the "activity landscape" changing accordingly. This may lead to different (chemical) neighbourhood relationships in chemical space. [33] The aim of similarity searching is to find similar compounds (neighbours) to a given query, with points close in chemical space considered to be similar. Such a lack of consensus in listing of similar (neighbour) compounds leads to inaccurate ligand similarity predictions. [31] Moreover, the use of active compounds used as the reference structures in similarity searches may result in limited diversity amongst the compounds retrieved.

Lastly, similarity searching relies heavily on the accuracy of input data in building reliable models. There are, however, significant error rates observed from the chemical data available in the literature and public databases. [31]

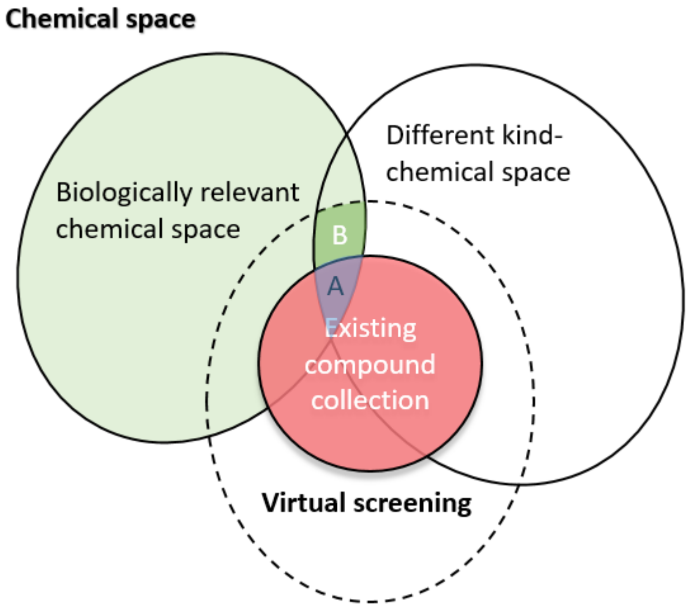

Chemical space
Different kind-chemical space
Biologically relevant chemical space
B
A
Existing compound collection
Virtual screening

Figure 3 - Modified from [33]. The existing compound collection contained only limited coverage of the biologically relevant chemical space (molecules with biological activity), which lead to discovery of drugs (A). Virtual screening offers opportunities to expand on an existing collection for a greater coverage of previously unexplored chemical space (dashed ellipse), where drugs are likely to be found (B).

1.1.3 SAR/QSAR

The ligand-based approach employs quantitative structure-activity relationships (QSAR). SAR may be designed to provide either quantitative (as in QSAR) or qualitative predictions, based on the relationships developed using continuous data or discrete data (presence or absence of a particular structural feature), respectively. It is assumed in QSAR models that a mathematical relationship (often a statistical correlation) can be found between the activity (or other relevant property) of query molecules and measurable or computable physico-chemical descriptors used to quantify the chemical structure. For this, a number of models were built to relate molecular descriptors to biological activity with the objective to firstly, predict properties for untested data, secondly, to select compounds to be prioritised for synthesis or screening, and thirdly, to extract patterns or SARs from analogues of the lead compounds. Thus, one may study the effect of structure on activity (or potency), and the resulting knowledge can be used to guide the lead optimisation process. Lastly, QSAR can be used to get insights into the characteristics of the receptor binding site. [42] A QSAR model's value depends on the quality of input data; they provide only limited precision due to experimental variation and the incompleteness of the compound set, the choice of descriptors and statistical methods.

The history of SAR originated in 1863, when Cross [43]observed that the toxicity of aliphatic alcohols is inversely related to their aqueous solubility, and just a few years later, in 1868, Crum-Brown and Fraser [44] published that the physiological action (Φ) of quaternized strychnine derivatives which would produce muscle paralysis (effect on blocking neuromuscular receptors, competitively inhibiting acetylcholine binding) is a function (f) of its chemical constituents (C), thus proposing the first quantitative relationship in pharmacology and medicinal chemistry by the Equation 1:

Equation 1 – Crum-Brown and Fraser's formula relates chemical structure to a biological response:

$$\Phi = f(\mathrm{C})$$

which is the first general form of a QSAR relationship. Due to a change in chemical constituent, $\Delta\mathrm{C}$, the effect is reflected in the biological activity, $\Delta\Phi$. Richardson (1869) [45] later observed that the hypnotic activity of aliphatic alcohols satisfies a proportional relationship with their molecular weight, then followed by Meyer [46] and Overton [47] in 1890s who correlated the toxicity of organic components to their lipophilicity (lipid-water partition coefficient). Trabe [48] and Seidell [49] were pioneers in using physicochemical properties as descriptors in their study. The QSAR concept was proposed initially by Corwin Hansch (who is honoured as the "father of QSAR") and his co-workers in the 1960s showing that a variety of biological activities could be modelled as a function of physicochemical attributes, followed by their first QSAR publication, which concerned the herbicidal effects of phenoxyacetic acids and their derivatives, in 1962. [50]

Various applications of QSAR are reviewed. Any QSAR study requires a data set with known values of activity, a set of molecular descriptors (structure-related) and a mathematical method. QSAR focuses on local structural features known to be relevant to biological activity, which is in contrast to the similarity searching based on the whole structure. [51] On the other hand, large numbers of descriptors would increase the chance of producing deceptively good models due to overfitting or chance correlations. It is found that predictions obtained from a model with a small number of simple descriptors can often outperform those from complex ones. [52] Traditionally, QSAR has been applied to a congeneric series of chemicals sharing a common scaffold, which ideally should not contribute to differences in activity but with adequate diverse substituents. In contrast, recent studies focus on data with a wider chemical space. [53] Typically, the whole dataset is partitioned into training (used to build the model) and testing for model evaluation or selection. In the early stages of QSAR method development, linear methods such as multiple linear regression, which was first introduced to QSAR by Hansch and is still commonly adopted due to its simplicity and interpretability, and partial least squares (PLS) were often applied to generate QSAR models, and over time, with increased computer power, more complex machine learning methods such as artificial neural networks, support vector machines, random forest and k-nearest neighbours have come to be used in QSAR modelling. [53]

1.1.4 Chemical space

The depiction of chemical space varies from the intuitive perspective of Lipinski, to the review of Hopkins (2004) [54] who analogises compounds in chemical space to the stars in the universe, whereby the chemical space is very large in size. Others suggest that chemical space is composed of all possible organic molecules. [33] As introduced above, compounds characterised by the same set of descriptors are mapped onto the coordinates of a multidimensional descriptor space defined as chemical space. Each molecular descriptor adds a dimension to the space. Molecules are located according to their descriptor values [51], where similarity and dissimilarity are defined based on their intermolecular distance in chemical space. The choice of molecular descriptors is decisive for a meaningful chemical space in which if the compounds happen to be similar, they would be located in contiguous regions.

1.1.5 Molecular Descriptors

Descriptors are numerical values, which can be scalars, vectors or bit strings etc., created using a defined algorithm that transforms chemical information contained in a molecule or fragments of a molecule into a number used to establish QSAR. Each descriptor encodes only a certain subset of the information contained in a molecule. They include physicochemical, geometric or topological properties associated with a molecular structure. These could either be obtained experimentally (physicochemical properties) or calculated theoretically (based on fragments or the whole molecule), or both. In other words, descriptors are mathematical representations of a molecule.

Based on the data representations of these properties, descriptors can be hierarchically ordered as: zero-dimensional (0D), the simplest, calculated from the chemical composition of the molecules by simple counts of the number of atom types or bonds. Examples are molecular weight, atomic count descriptors etc., which are not calculated from a molecular graph but usually from atomic information; one-dimensional (1D) based on the substructure (fragment); and two-dimensional (2D) descriptors, often termed as "connectivity (or topological) indices" computed from a molecular graph or matrices reflecting the connectivity between atoms; these include binary representations such as structural keys and fingerprints, feature trees, etc. 2D descriptors are the most widely used molecular representation to the field of chemoinformatics and ligand screening. In practice, the 2D molecular graph of a chemical is converted to a 1D string (the SMILES format [55]) to calculate the structural descriptors. [31] Binary descriptors are represented by a Boolean array of a set of 1 or 0 values, typically encoding the presence or absence of a specific substructure, allowing a chemical database to be screened at low computational costs by simple Boolean operations. There exist a variety of three dimensional (3D) descriptors, derived from 3D structure of molecules such as pharmacophores, considering the spatial configuration of essential features conferring specificity for a ligand binding to a specific binding site. 3D descriptors depend on the geometrical coordinates of the molecule's atoms required in a valid conformation, which varies depending on the representative physical state. [56]

To note that there is no "best" descriptor available as a general rule. The information content of the "best" descriptors should be comparable with the experimentally determined properties. High order descriptors (3D or higher) but irrelevant information with respect to the properties are usually regarded as noise on behalf of the model, which in turn produce instable or not predictive models. [57]

1.2 Machine Learning models

One of the major applications of machine learning is data mining. Due to the rapid advances in high-throughput instruments and database technologies, collections of data have become more readily available. Methods enabling new discoveries derived from the analysis of large amounts of complex data have become increasingly demanded. It is desirable for an algorithm to be able to train on data contaminated with experimental errors or missing values, and to derive empirical correlations to estimate properties of new data. One can perform these analyses using a machine learning approach which involves processing and modelling of massive amounts of experimental and computer simulation data (usually referred to as descriptors in chemoinformatics) to retrieve and discover data patterns and to establish quantitative relationships between multiple features. Nowadays, machine learning has been widely used in many areas including computer science, bio- and chemoinformatics and biostatistics.

Each instance in any database is represented by the same set of features which can be numerical (discrete and continuous), categorical and binary that is used to train a machine learning algorithm. The instances in a multivariate dataset can be pictured as points in a multidimensional space where each variable (measurement) is a dimension of the search space. In general, models can be trained with supervised and unsupervised learning algorithms. As they are named, the former contains known labels associated with each training instance. Other finer categories of machine learning algorithms include hybrid models, semi-supervised learning and reinforcement learning, which are beyond the scope of this thesis.

Clustering (or segmentation) is an unsupervised classification task that requires no previously provided class labels. The purpose of clustering is to group related entities (observations) based on the hidden relationships found from the data or as a preprocessing step to be performed prior to actual processing. This is in contrast to supervised machine learning referring to a learning process from training data and the resulting classifier is used to generalise on previous unseen instances.

In the case of supervised classification (or usually just classification), a classifier is a learning algorithm which takes a vector of feature values as inputs and returns a class label. The learning process starts with data preparation and data-preprocessing which may involve detection of outliers (noise), handling missing or imbalanced data, feature construction and transformation, collecting relevant or (if known) informative features from the dataset, and, in the case of data sets with large numbers of variables/entries, feature/instance selection prior to the learning process is required to reduce the number of features captured/data to a manageable amount, which also helps to increase performance. Apart from the data used to train the algorithm, a subset called the test set is kept independently for evaluation. In cross-validation, the training data is randomly partitioned into subsets of equal size, and each is used in turn for testing, treating as a proxy for true test data, while using the remaining data for training. The average across all repetitions is used to compensate for the bias caused by the reduction of the training set size.

The typical goal of machine learning is to generalise on new data on the basis of examples in the training set. The following section addresses each of these issues on fix experiment setting especially on construction of a random forest, a decision tree based classifier.

1.2.1 Pre-processing

The raw data has to undergo pre-processing which includes "data clearing" to fill in or remove the missing values, remove the noise (random errors), and curated data (errors in public sources), and "data reduction" to obtain a reduced representation of the original data, while at the same time eliminating irrelevant or redundant data. Instance selection involves reducing the sample size while maintaining the required quality of the estimates. This is often achieved by random sampling which randomly selects instances from the original data, or stratified sampling, to increase the sample size of minority groups. Whenever necessary, the data are transformed for better model development. In autoscaling, the variables are rescaled to have zero mean and unit standard deviation. Also, new features can be derived from the original features in the process of feature construction. Data pre-processing aims at improving data quality, to reduce the size of data and computational complexity, and improve the performance of the models. [58]

Note that, the challenge to generalise well to new samples increases drastically with increased number of features, which easily lead to "the curse of dimensionality". Typically, the performance of a classifier increases as the number of features increases; until an optimum is reached, beyond which the accuracy of the model decreases. This is due to the data points in space becoming increasingly sparse with increasing dimensionality. This will result in misleading approximations of the boundaries between classes and as a result, overfitting occurs. Most clustering algorithms rely on distance or similarity measures, whereby the data are partitioned into groups (clusters), such that the data in the same cluster are closer to each other than to the data from other clusters. However, as the dimensionality tends to infinity, distances for sparse high dimensional data follow:

Equation 2 – The relative distance between points converges to zero with increasing dimensionality (d):

$$\lim_{\mathrm{d} \to \infty} \frac{\mathrm{dist}_{\mathrm{max}} - \mathrm{dist}_{\mathrm{min}}}{\mathrm{dist}_{\mathrm{min}}} = 0$$

The relative difference in distance of the nearest and the farthest data points to the centroid tends toward zero (Equation 2), this means the distances (dissimilarity measures) between data points become relatively uniform so that distance measures to select the nearest points (as to be assigned in the same cluster) becomes meaningless in high dimensional spaces. Also, the sparseness of the data in high dimensional spaces is not uniformly distributed in which data points mostly lie near the edges of the space far away from one another, with empty space in between. [59]

To reduce the dimensionality of the data while retaining data information, feature selection is generally part of preprocessing. In other words, feature selection allows identification of relevant features. The process involves removing features that have no significant variation, or are correlated with other features. Alternatively, a method such as principal component analysis (PCA) is used to project data to a lower dimensionality, which would also help to remove noise from the model. Feature selection is a common step, appropriate to select descriptors for QSAR models, and can be done in a number of ways, such as stepwise selection, all possible subset selection, genetic methods and factor analysis.

1.2.2 Imbalanced data

Imbalanced data occurs when the number of instances of each class is not evenly distributed. Firstly, classifiers tend to optimise prediction for the largest class, while treating all others as noise so as to maximise prediction accuracy. Accordingly, there will be a bias produced. An instance in a minority class is more likely to be misclassified. Secondly, instances from the minority class may provide insufficient information to build a model. The collections of a class for learning shall be adequately sized to insure they sufficient reflect the complete chemical space. Lastly, the noise (outliers) from the majority class may mask the information contained in the minority class. One way to reduce this size-related effect is to weight the training set inversely proportionally to the size of the class, however this in turn causes a higher rate of misclassification. The other way is to even up the number of samples in each class by resampling of the training data (either by down- or up-sampling) in the pre-processing stage. Down-sampling (or under-sampling) to down-size the majority class may result in a loss of data whereas up-sampling (or over-sampling) to make exact copies of the minority class may result in overfitting. [60]

1.2.3 Decision trees

The decision tree algorithm is a non-parametric supervised learning method. Non-parametric refers to no assumptions about the space distribution and the classifier structure which can grow with the data. Classification is performed by routing from the root node of a decision tree from where the tree starts growing. At each node (except the terminal nodes), the data is split into two or more subsets according to the value of a selected feature, and the process is repeated until homogeneous groups remain or the stopping criteria have met. The prediction of each instance is made by routing it down the tree according to its attribute values tested in successive nodes until a leaf node is reached, with which a class label is associated. Decision trees are flexible yet data-driven classifiers. Small differences in training data may lead to great variations to the classification results. [61] The complexity of a tree is directly affected by the splitting criteria, stopping criteria deciding when to stop growing the tree and the pruning methods. Most existing decision tree algorithms, including ID3 [62], C4.5 [63] and CART (Classification and Regression Trees) [64] are greedy as the best attribute is searched to split the data at each node. Typically, an optimal tree has minimum generalisation error.

In general, decision trees tend to have low bias but high variance. To improve the performance, one can use an ensemble of models. It is found that the generalisation (test) error can be improved by adding classifiers. Bagging, boosting and stacking are ensemble methods. In bagging (aka bootstrap aggregation) [65], each classifier is learned from a different training set by resampling with replacement to create random variations in the training in such a way to reduce the variance. Also, all variables are considered at each split in the tree. The result is made by combining the votes from each classifier. Boosting [66] decreases the bias by using all data to train a classifier. However, each training sample carries a weight so that misclassified examples could have more focus in subsequent learning processes until all the samples are correctly learned. Stacking is similar to boosting, where the outputs of classifiers become input of another classifier as to combine the results.

1.2.4 Splitting rule

The feature resulting in the best partition of the training instances is chosen at each node according to a selection criterion. The tree is grown by recursively partitioning until a pure subset is formed or the size of subset is sufficiently small. Among numerous splitting criteria, however, none was observed to be superior from one another [67], Quinlan's C4.5 uses information gain splitting [63] and CART uses the Gini index [64].

CART only allows binary splitting whereas C4.5 and ID3 use multi-way splits.

To speed up the most time-consuming step of the training algorithm in determining the threshold of a split at a node for a numeric feature, one approach is to restrict the threshold to be based on only a subset of the instances, or to discretise the original values of features to intervals as a way to reduce computational efforts. Generally, decision trees are better for handling discrete/categorical features. The univariate decision trees which use a single attribute to test at each internal node are restricted to axis-orthogonal splits. In each partitioning, the instance space is partitioned into two hyperrectangles (sub-regions); this process is recursively repeated until every square region contains homogeneous data sets. Thus, univariate decision trees do not work well with problems that require diagonal splits with respect to the features' axes of the feature space. (Figure 4) [68]

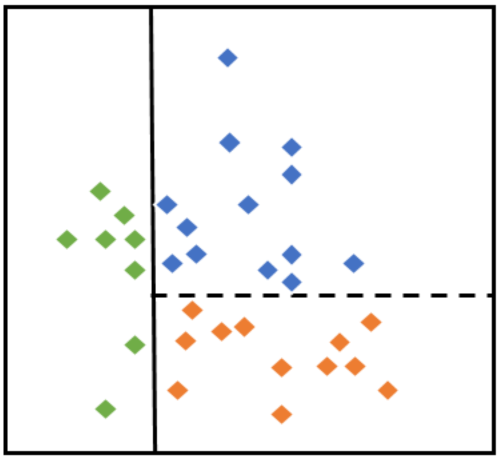


Figure 4 – Orthogonal splits divide the feature space into axis-parallel sub-regions, each with a single classification label.

1.2.5 Overfitting

A decision tree may result in overfitting of the training data due to outliers and noise (mislabelled instances) or by the lack of representative examples, resulting either from insufficient amount of training data being available or as a consequence of sampling error. Furthermore, high dimensionality of data may lead to overfitting. Generally, overfitting occurs when a model learns more of the detail and the noise in the training data rather than underlying relationship, thereby creating a tree that is excessively complex and does not generalise well to new data.

To avoid this, one can pre-prune the decision tree by some stopping criteria to stop the tree growth before it gets too complex, or alternatively employ a post-pruning method to remove useless branches. Since the objective function based on a subsample of the training data is only a proxy for our goal function; one wants to generalise beyond the training examples, thus there is no need to fully optimise it. However, even using noise free training samples can still result in overfitting because any chosen sample would most likely not be a perfect representative of the entire sample. A tree-based ensemble classifier, Random Forest, is considered relatively immune to overfitting.

Every learning algorithm has an inductive bias (or named learning bias) referring to a set of prior assumptions that a classifier holds in order to perform induction, that is to generalise beyond the training examples. In other words, inductive bias specifies a preference for which types of generalisation to use to bias toward a particular set of predictors. In decision tree models, shorter trees are preferred which naturally tend to avoid overfitting. As a result of inductive bias, some potential solutions cannot be reached. [69] This was formalised in the "no free lunch" theorem by Wolpert [70], which states for machine learning that there is no universal learning algorithm that can deal with all possible situations. Of course, at least some prior knowledge is required for induction. It is common to build multiple models and compare their performance since every model represents a certain simplification of the reality; cross validation is often used to determine the best model that suit the needs based on predictive accuracies.

1.2.6 Evaluation functions

The prediction accuracy is used to evaluate the performance of the classifiers by either the cross-validation or the out-of-bag (OOB) estimations or alternatively, leave-one-out validation estimate strategy. Predicting new data on the basis of an induction process for machine learning involves uncertainty. The classification results of a supervised machine learning (an induction process) is justified with reference (known responses to the data); thereby, the quality of the predicted value can only be guaranteed probabilistically. The theoretical guarantee is that given enough training data, there is a high probability that the learner will return a hypothesis that would either generalise well, or otherwise be unable to find a hypothesis that consistently classifies data correctly. [71]

1.3 Random Forest

Random Forest [72] is a supervised machine learning approach used for classification and regression. The predictions are made based on a stochastically built ensemble of decision trees. Each tree is grown from a particular bootstrap sample by the CART (classification and regression trees) algorithm using data drawn randomly with replacement from the entire dataset, leaving approximately a third of the data for the internal validation as an out-of-bag (OOB) sample for each tree. Random Forest improves bagged trees by way of firstly, having trees grow to their maximum to reduce the correlation between trees; secondly, each split is determined by a random subset of features thus to induce diversity of the resulting trees.

Rather than using all descriptors, the split at each node of the tree as the tree grows is determined by choosing the best division possible using any of a randomly chosen subset of *mtry* descriptors, where *mtry* is by default the square root of the number of descriptors available. Each splitting is based on a single valued attribute that best divides the training set. Random Forest is an improvement over bagging, as descriptors are not equally important. The Gini criterion [73] is used to select the split resulting in the greatest decrease in impurity. This process is iteratively continued, with a freshly chosen random sampling of descriptors at each node, until all the training data have been classified into their appropriate leaf nodes. At this point, the tree building ends and no pruning is carried out. Running Random Forest with the default setting of *mtry* speeds up the process compared to using larger values, as the number of splitting tests required at each node is smaller. The tree building process is repeated for each of the *ntree* trees in the forest, with each tree being based upon a different bootstrap sample of the instances from the dataset.

The classifier is trained with labelled samples to construct a model to predict the category of unseen data. The predictive performance of the model is evaluated through internal validation. During the training, the out-of-bag (OOB) sample, which does not participate in the tree building, is used in parallel to evaluate the prediction accuracy of the trees. Each tree being trained, that is built, on a bootstrap sample comprising one or more occurrences of approximately 2/3 of the full training set. The class prediction of each sample is made based on a majority vote of those trees (averaging for regression) for which the given instance is in the OOB sample (and therefore uses those trees for which the instance has not been part of the training set). The OOB error rate is obtained by dividing the number of misclassified data points by the total number of points. [72]

In CART, the Gini criterion (or index), introduced by Breiman in 1984 [64], is used to select the best split for each node that leads to the greatest reduction in impurity between parent and child nodes. The impurity $i(t)$ of a node t is defined as:

Equation 3 – The Gini impurity measure:

$$i(t) = \sum_{i=1}^{n} p_i(1 - p_i) = 1 - \sum_{i=1}^{n} (p_i)^2$$

where $i(t)$ is zero when the node is pure, p_i is the proportion of class i at node t, n denotes the number of classes. The reduction in impurity $\Delta i(t)$ after a split is given by:

Equation 4 – The reduction in impurity of a binary split:

$$\Delta i(t) = i(t) - P_L i(t_L) - P_R i(t_R)$$

where $i(t_L)$ and $i(t_R)$ are impurity measures, and P_L and P_R are the proportion of cases that go from parent to the left and the right child nodes, respectively. [74]

Alternatively, there are other machine learning techniques, such as support vector machine (SVM), that can be used for classification and regression tasks. SVM, proposed by Cortes and Vapnik [75], uses a kernel function to transform the data nonlinearly to a higher dimensional space. This spreads out the data, such that an optimal hyperplane can be constructed that separates the data into two classes on either side of the hyperplane with a maximum margin. The k-Nearest Neighbours (kNN) [76], where k is typically a small odd integer, calculates proximity between a query point and all of its neighbours using a distance function (e.g. Euclidean distance), and the class is assigned by voting among its k closest neighbours. In contrast, Naïve Bayes by applying Bayes' rule computes the probabilities of a query belonging to each class based on various attributes. [77]

1.3.1 Applications

Random Forest has been applied in several contexts. A Random Forest model was built to classify HIV-1 protease binding pockets to one of the nine FDA approved protease inhibitors; while to obtain the Gini importance to identify the essential features responsible for the binding with various protease inhibitors. At the end, Ko et al. [78] identified 12 top ranked descriptors quantified the geometric and electronic properties of the binding pockets which can be used to aid the design of novel HIV-1 protease inhibitors. Palmer et al. [79] used a data set of 988 organic molecules with which to train and test a Random Forest to predict aqueous solubility based on 2D descriptors and reported that Random Forest achieved better accuracy compared to the other models. Random Forest has been successfully applied to predict protein-ligand binding affinity. [80,81,82]

CHAPTER 2. Theory of Molecular Dynamics and Molecular Docking

2.1 Introduction to Molecular Docking

As mentioned in the previous chapter, molecular docking is the most used SBVS method, to ultimately identify lead compounds. Docking methods have been employed to study the interaction of small molecules with the target which helps to design a ligand with the necessary feature(s) to achieve high affinity binding. [83]

Molecular docking relies on the availability of 3D structural data of the receptors obtained experimentally, or through computational modelling. Docking inherits the intrinsic limitations of structure-based methods, including the challenges to resolve the 3D structure of membrane-bound receptors because of the difficulty in crystallising them. In addition, the crystallisation environment is not physiological, this means the determined structure may adopt a non-physiological fold [84] and may not retain in its native conformation in water or organic solvents. The choice of solvents for crystallisation is shown to affect protein conformation. [85] The structures of the unbound receptor may be of less biological relevance. This is especially true when receptor flexibility is not allowed in simulations. The binding site may have been shielded by other parts of the molecule which do not allow access of small molecules. Binding results in a structural rearrangement (induced fit) of the receptor. On the other hand, error can be found in the structures derived using homology modelling based on analogy or other simulation approaches, when the 3D receptor structure is not available. [86]

Molecular docking simulates the recognition process of ligands to receptors. Typically, a docking protocol comprises: a) a global search (or a reduced search at potential binding sites) for exploring all the possible conformations of a protein-ligand complex, referred to as sampling, for predicting the binding modes, and b) a scoring method used to evaluate the binding energy for a particular binding mode to identify the binding mode with the lowest energy-reflecting shape complementarity and electrostatic amity to the rigid target. *In silico* simulation of receptor-ligand recognition processes can be justified theoretically, depending on the level of flexibility accounted for by a docking algorithm (see the “relaxed complex scheme” section), which can be described by the “lock and key” model, where the rigid target and ligand have exactly matching binding surfaces, or by models accounting for protein flexibility, including two competing hypotheses, “induced fit” and “conformational selection” (or population shift), with both models describing the mechanisms underlying molecular recognition. [87] The induced fit, proposed by Daniel Koshland in 1958, [88] suggests binding induces a conformational change of the receptor, whereby the receptor and ligand are fitted by the binding event. Conformational selection [89] suggests the ligand binds selectively from the unbound ensemble in a "lock and key manner", and as a result of binding causes a population shift toward this bound conformational state.

2.1.1 Sampling

The earliest docking programs focused on rigid-docking where docking is performed through rigid body translations and rotations. The first docking program, DOCK, was published in 1982 [90] and employs matching methods. Conformational sampling methods considering ligand flexibility are divided up into three categories, as follows: a) systematic searching, in which ligand is divided up into rigid (core fragment) and flexible (side chains) parts. This is achieved by either docking core fragments prior to adding the side chains in an incremental fashion or by covalent linking of various molecular fragments that were docked previously into the active site region, referring to as de novo ligand design strategy. [91] e.g. DOCK, b) a stochastic search e.g. genetic algorithm AutoDock, which will be detailed later ; and, c) simulation methods [92], such as the use of molecule dynamics.

2.1.2 Relaxed complex scheme (RCS)

Using a fixed receptor structure can lead to 50-70% of binding poses being predicted incorrectly. [93] A number of ways have been developed to account for conformational rearrangements upon complex formation with the ligand. (Figure 5) For example, multiple receptor conformations (MRCs) can be used as docking targets (a.k.a ensemble docking), where docking is performed on an ensemble of pre-generated conformations provided by experimental techniques such as NMR spectroscopy or X-ray crystallography, or by computational techniques such as Monte Carlo, normal mode analysis or a MD run (methods involving docking to multiple MD conformations have been termed "relaxed complex" (RC) schemes) etc.

MRCs could be regarded as an analogy to the "conformation selection" model, but with protein flexibility modelled implicitly, as the conformation of the receptors is kept rigid during the docking process. However, there is a possibility that a ligand may bind to less frequent but important conformations with this method, and that the true bound-conformation may not be selected from the resulting conformational ensembles.

Alternatively, to MRCs, the "Soft receptor" method is designed to allow partial overlap of the ligand and receptor, by specifying a smaller van der Waals repulsive term to reduce steric penalties. The effect of this adjustment on the van der Waals potential is to simulate a larger binding site, to account for a certain degree of conformational plasticity, thereby modelling protein flexibility implicitly. Nevertheless, there are some drawbacks: soft docking is limited to small side-chain rearrangements of the target (in the order of 1Å) and has an unfortunate tendency to increase false positives.

Another example, "selective docking" aims to select a few "critical" side chains in the binding site to explicitly model their dynamics during the binding process, related to the "induced fit" model, by allowing some small rearrangement of the residues in the pocket to accommodate ligand binding. Most uses of this method will require some structural knowledge of the receptor and its function.

Finally, "on-the-fly" docking models, which change the receptor's conformations during docking, use various sampling and optimisation techniques. [87,94] One such strategy, based on the "induced fit" concept, is applied by RosettaLigand. [95] This method does not simultaneously sample both ligand and receptor flexibility, thus allows consideration of only small-scale induced-fit effects. The ligand is first docked in a rigid receptor, the side chains of the receptor are later changed by the use of rotamer libraries, followed by a minimisation of the ligand-receptor complex. Other strategies, such as energy minimisation, Monte Carlo or molecular dynamics simulations, are used to perform a post-processing refinement step after a rigid docking.

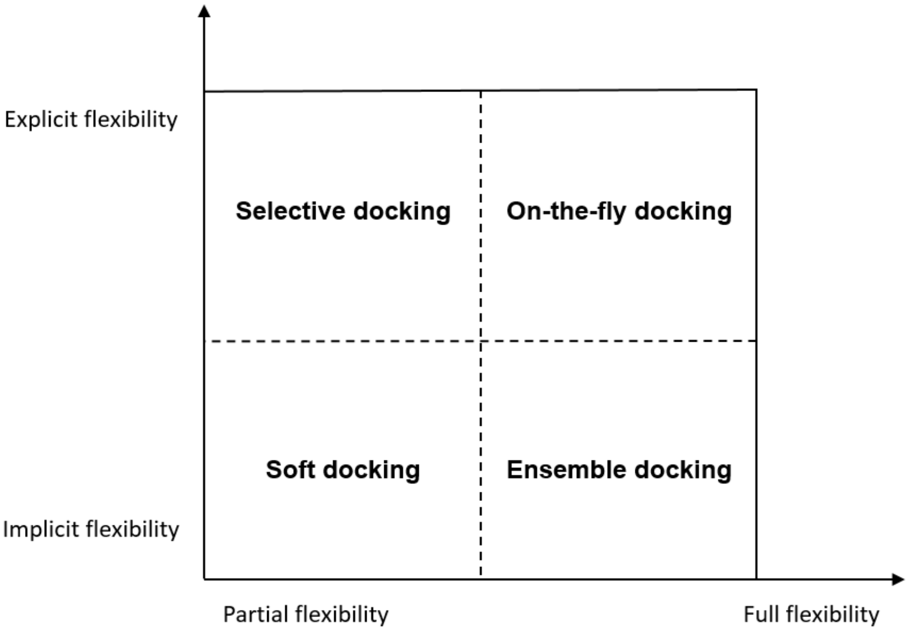


Figure 5 - Modified from [87]. Classification of approaches accounting for protein flexibility.

The “relaxed complex” method was proposed by McCammon et al. [96] inspired by two experimental works. [97,98] They observed that ligand may bind to conformations that rarely exist in the dynamics of the receptor. Therefore, MD simulation is applied first to explore conformational space to discover novel binding sites, or to find conformations that are rare. Docking to different snapshots taken at different time scales aims to overcome the intrinsic limit in docking, by accounting for protein receptor flexibility. Application of this method led to the discovery of the first clinically approved HIV drug. [99] and has been used to screen a library of ~12,500 compounds against DNA polymerase. [100] Several selection techniques have been proposed to extract representative conformations from MD simulations [101,102,103].

2.1.3 Scoring

Scoring (or potential) functions were previously categorised as either physical-based, empirical-based or knowledge-based scoring.

Physical-based (force-field) energy functions (e.g., GOLD, AutoDock, Dock), which are derived from the laws of physics, use atomic force fields; a set of functional forms and parameters to calculate the energy. Force fields, also referred to as molecular mechanics

(MM), ignore the electronic effects of nuclear motions, and electrons are treated as implicit within the MM variables.

Empirical energy functions (e.g., F-score, ChemScore, SCORE) are expressed as a sum of various energetic contributions, can be written in form of Equation 5, where ΔG is the binding free energy. Each term is scaled by a coefficient W_i derived from linear regression in order to fit known (receptor-ligand) binding affinities. Both force field-based and empirical scoring functions are functions of different energy terms. The difference is that the force field adopts an energy function derived from well-established theoretical models; whereas an empirical scoring method is built using a best-fit function obtained from regression analysis.

Equation 5 - General empirical energy functions:

$$\Delta G = \sum_i W_i \Delta G_i$$

Knowledge-based (or statistical) (e.g., PMF, DrugScore), energy functions use experimentally determined structural data to derive distance-dependent potentials for interactions between pairs of receptor and ligand atoms. Pairwise potentials are calculated by computing the frequencies of observed structural features, occurring as atom-atom pairs, and converting them into free energies using the inverse Boltzmann relationship; which states that probability of occurrence of a given state can be derived by the given energy of that state.

The lowest energy corresponds to the thermodynamically most stable complex. Docking can be regarded as an optimisation problem, including the global positioning of the ligand and a local refinement step, finding the optimal pose of small molecules in the receptor, as characterised by the position, orientation and shape of the ligand. The binding energy of every possible pose of the ligand is computed from a minimisation process to find the minimum of the binding energy function. Docking generally adopts a simpler force field, allowing a wider computational space to be explored. This allows blind docking to be performed, when no prior knowledge of the binding site is known. Recently, docking methods based on scoring have been used to prioritise those to be screened in vitro. In order to have a high binding affinity, a ligand must be electronically and sterically matched to the pocket.

2.2 AutoDock

AutoDock is an automated docking tool. In a docking simulation, a ligand starts from a random position. The flexibility of the ligand is modelled, and the translations, orientations, and conformation of the ligand are explored by a search method until putative binding sites have been found. AutoDock employs a force field energy evaluation method. AutoDock uses a united atom model, in which non-polar hydrogens are merged and the charges are assigned to the corresponding carbon, leaving only the heavy atoms and the polar hydrogen atoms.

2.2.1 Search method

Genetic algorithms (GA) [104] apply the concept of Darwinian natural evolution and Mendelian genetics, and are suitable for problems that suffer from combinatorial explosions as the ligands complexity increases with increased degrees of freedom. AutoDock uses GA to perform a global search for best docking pose. The arrangement of a ligand with respect to a protein, is described with a set of real values and these refer to a ligand’s “state variables”, analogised as a gene in a chromosome (the chromosome itself refers to solutions of a problem), while the atomic coordinates are analogised to the phenotype. Chromosomes are evaluated for the fitness based on the calculation of interaction energy and decide which to pass down to the next generation. Mutation and crossover operations which offer a larger degree of alternation, are passed from parents to their children, increasing the chance of exploring other areas of the conformational space. Based on the evaluation of the fitness, selection allows reproduction of offspring better suited to the environment.

In detail, each ligand is defined by a set of real values including three Cartesian coordinates, describing the ligand's translation, four values for quaternions defining the ligand's orientation [105], and one value for each defined conformation torsion of the ligand with respect to the receptor. The search process starts by creating a user defined number of random individuals that makes the initial population, followed by iteration over a number of generations until any of the termination criteria are met. A generation consists of mapping, fitness (energy) evaluation, selection, crossover, mutation, and elitist selection. (Figure 6)

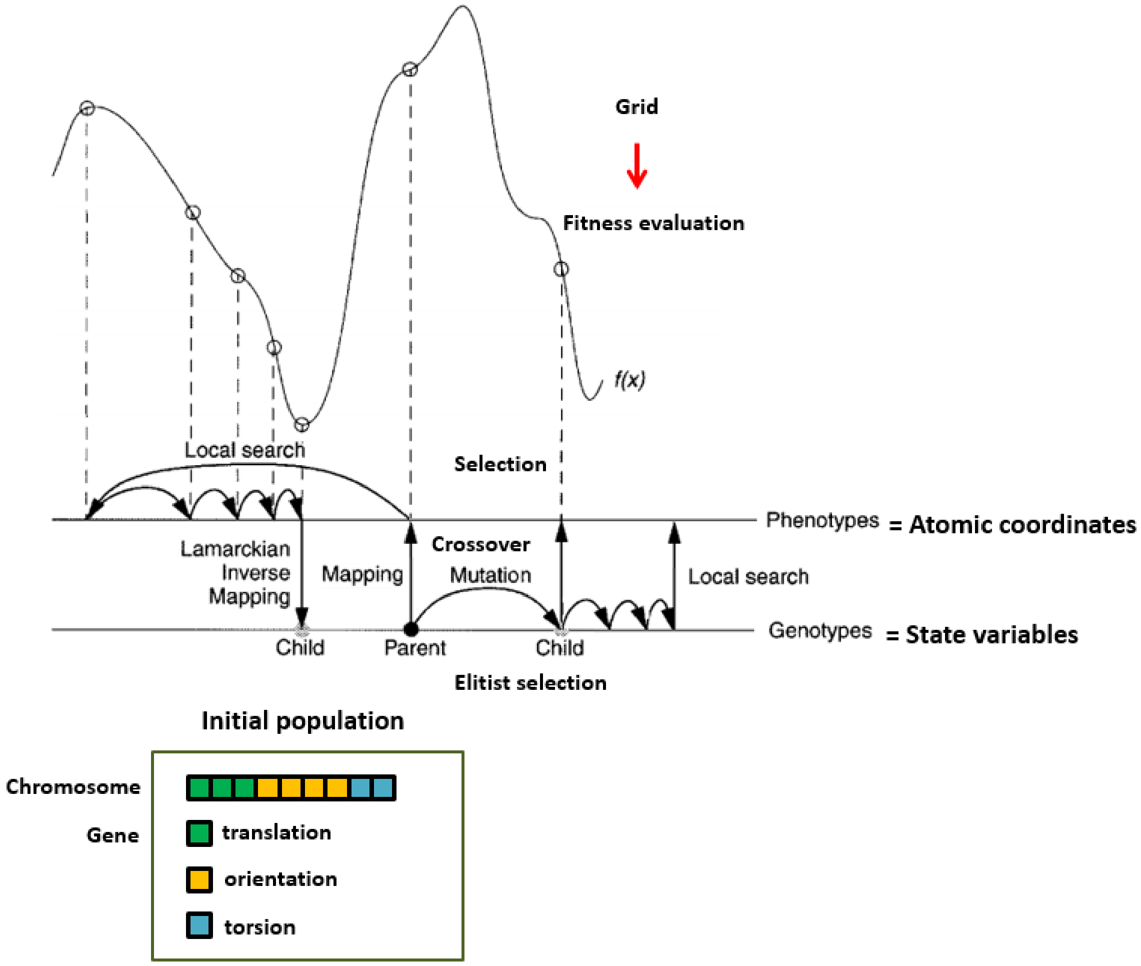


Figure 6 - Modified from [106]. The process of a genetic algorithm. Each generation consists of the following, in the order given: mapping, fitness evaluation, selection, crossover, mutation, elitist selection and local search.

Firstly, each genotype is "mapped" to corresponding phenotype (that is the ligand's atomic coordinates) to allow interaction energies to be evaluated by a grid-based approach. Then, it goes through a selection process based on fitness value, such that individuals with better than average fitness are ensured to have at least one offspring. Crossover and mutation are used to apply random perturbations to the parents. Crossover and mutation occur between random individuals at a user defined rate, and the resulting offspring replace their parents to keep the population size constant. The value replaced by a mutation is a random real number that has a Cauchy distribution for small deviates. The new generated population from the proportional selection, crossover, and mutation is ranked according to fitness.

A user defined fraction (0.06 was found to be in maximum efficiency) of populations undergo a local search, based on that of Solis and Wets method [107]. The local search is performed in genotypic space (as illustrated in Figure 6) rather than in the phenotypic space, so that the acquired traits from local adaptation can be inherited by their offspring; otherwise an inverse mapping approach is required to convert the phenotypic result of local search into its corresponding genotype. Instead of dedicating effort to an inverse mapping, AutoDock performs a local search operation in genotypic space using modified Solis and Wets method with a translational step size of 0.2, and orientation and torsional step size of 5. The step size of the local search is adaptive.

The combination of the global and adaptive local search method results in the Lamarckian genetic algorithm. The Lamarckian method was developed, so that the results from LS are heritable to offsprings. A genotypic space is used in the Lamarckian search method, in contrast to the typical phenotypic search space.

2.2.2 Energy evaluation during sampling

AutoDock uses a grid-based approach [108,109] to reduce the run time spent for evaluating candidate conformations. AutoGrid is used to generate grid maps. To do this, a protein target is placed in a 3D grid box and a probe atom systematically visits every grid point. (Figure 7) The pairwise interaction energies of the probe, and of the protein atoms positioned within a cut-off radius of 8Å at each grid point are summed and stored in the grid maps, providing a pre-calculated lookup table to speed up energy evaluation. Grid maps, including dispersion/repulsion terms and a hydrogen-bonding energy, are created for each atom type in the ligand. A separate electrostatic potential grid is also created. During the docking, interaction energies are calculated from the pre-calculated grid maps using trilinear interpolation.

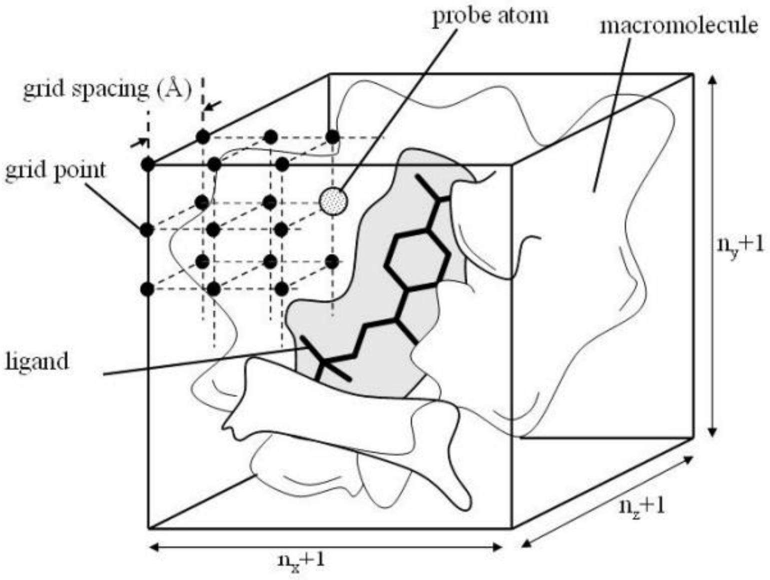


Figure 7 - Adopted from [110] illustrates the grid-based method.

The pair-wise atomic terms include evaluations of dispersion/repulsion energies, calculated using the Lennard-Jones 6-12 potential with coefficient A_{ij} and B_{ij} calculated from the well depths and equilibrium distances of homogeneous pairs using AutoDock force-field parameters (AMBER) [111]. Hydrogen bonds are described by the 12-10 potentials with well depths increased by a factor of 10. The electrostatic interaction energy is calculated using Coulomb's potential using a single positive charge probe. The resulting electrostatic interaction energy of each ligand atom is the multiplication of the trilinearly interpolated electrostatic potential with the ligand's partial charge. Intramolecular energies of the ligand are calculated at each time step using the functional forms described above, but with a factor of 4 in the dielectric constant. [112]

2.2.3 Scoring

AutoDock4 uses a semiempirical free energy force field, combining molecular mechanics force field with an empirical method, to predict binding free energies. It uses pair-wise terms to evaluate the interaction between the two molecules and an empirical method to consider the contribution of water implicitly. The free energy of binding ΔG (see Equation 6) is estimated to be equal to the sum of the intramolecular energies by consideration of the bound and unbound states of the ligand and the protein respectively; the third term gives the intermolecular energy between the bound and unbound states of the complex. It is assumed that the two molecules are sufficiently separate in an unbound state, making $(V_{unbound}^{P-L})$ zero. As the protein (P) is kept frozen, and considering there is no interaction in the unbound state, the energies are set to zero. An extended conformation of the ligand (L) in which atoms could be fully solvated, with few internal contacts, is used as unbound state. The conformational entropy lost upon binding (ΔS_{conf}) is estimated to be directly proportional to the number of rotatable bonds (N_{tor}) in the molecule. [113]

Equation 6 - The free energy of binding is estimated by:

$$\Delta \mathrm{G} = \left(V_{bound}^{L-L} - V_{unbound}^{L-L}\right) + \left(V_{bound}^{P-P} - V_{unbound}^{P-P}\right) + \left(V_{bound}^{P-L} - V_{unbound}^{P-L} + \Delta S_{conf}\right)$$

$$\Delta S_{conf} = W_{conf} N_{tor}$$

Each pair-wise evaluation (V) can be expressed as Equation 7. The weighting constants (W) are optimised to calibrate the empirical free energy based on experimental data and are applied to each term. Dispersion/repulsion interactions are calculated using the Lennard-Jones 6-12 potential, where A_{ij} and B_{ij} are taken from the AMBER force field. The hydrogen bonding is based on the 12-10 potential where a directional weight $\mathrm{E}(t)$ is used to calculate the divergence from ideal bonding geometry. A maximal well depths of 5 kcal/mol at 1.9Å for O-H and N-H and 1 kcal/mol at 2.5Å for S-H are used to derived the parameters C_{ij}, D_{ij}. Electrostatic interaction is evaluated using a screened Coulomb potential. The desolvation potential is based on the volume of the atoms surrounding a given atom, weighted by a solvation parameter and an exponential term. (weighted factor=3.5Å) [113]

Equation 7 - The pair-wise atomic terms include:

$$\mathrm{V} = W_{vdW} \sum_{i,j} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) + W_{hbond} \sum_{i,j} \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right) \mathrm{E}(t) + W_{elec} \sum_{i,j} \left(\frac{q_i q_j}{\varepsilon(r_{ij}) r_{ij}} \right) + W_{sol} \sum_{i,j} \left(S_i V_j + S_j V_i \right) e^{\frac{r_{ij}^2}{2\sigma^2}}$$

The desolvation term is based on Wesson and Eisenberg's model [114], where S_i is the atomic solvation parameter calculated from the energy needed to transfer the atom from a fully hydrated to a fully buried state. They postulate that the desolvation energy is proportional to the change of surface area exposed to the water by comparing the solvent accessible surface area of the bound and unbound states. The amount of desolation V_i is calculated by a modified method using volume-summing method from Stouton et al. [115]

One issue of the data used in calibration is that the cost of burying a hydrogen bond without forming a bond to the protein is unknown. Desolvation of polar atoms is modelled by a constant, added in the hydrogen bonding function. It is assumed that hydrogen bonding to the complex is the same as bonding to water. Polar atoms that do not form hydrogen bonds will have an unfavourable effect on binding. Polar atoms are modelled by combination of the favourable hydrogen potential and the unfavourable desolvation potential.

2.3 Introduction of molecular dynamics (MD)

2.3.1 Energy minimisation

Energy minimisation is a prerequisite for other simulation techniques (i.e. molecular dynamics) to relieve strain in the initial structure and to reduce the thermal noise allowing better comparison between structures. This is because the energy of a system, composed of kinetic and potential energies, is conserved during MD simulations. The kinetic energy will increase significantly if low potential energy structure is sampled, which can distort the structure. Therefore, it is essential for a MD to start from the minimised structure. An energy minimisation moves atoms systemically toward the atomic positions at the closest (local) minimum of the potential energy surface, resulting in a local minima stable state. The minimisation methods can be energy-based, gradient-based or a minimisation taking account of the second derivative of the potential energy function.

The steepest descent (or gradient descent) minimisation algorithms [116] use the first order derivative from the potential energy function to determine the direction for the next move to move toward a negative gradient (more negative, stable conformation). Convergence can be slow whilst in the vicinity of the local minimum, as the algorithm does not consider the previous steps.

2.3.2 Force field

Force field methods are also referred to as molecular mechanics (MM). It describes the potential energy surface by treating the electrons implicitly and expressing the energy with respect to nuclear coordinates, in which the electrons are not treated, and the quantum character of nuclear motion is neglected, and with the atoms treated with classical mechanics. The interactions are parameterised beforehand and are not changed during the calculation. The electronic energy is calculated from summation of potentials, which can be divided into non-bonded, bonded and restraints.

Equation 8 - A typical force fields:

$$\mathrm{V} = \sum_{bonds} \frac{k_i^b}{2}\left(l_i - l_i^{ref}\right)^2 + \sum_{angles} \frac{k_i^a}{2}\left(\theta_i - \theta_i^{ref}\right)^2 + \sum_{torsions} \frac{V_T}{2}[1 + \cos(n\omega - r)]$$
$$+ \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left\{4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right]\right\} + \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}}$$

The potential energy (V) is generally expressed as sum of firstly, the bonded interactions including (1) the bond stretching (a two-body term), modelled by a harmonic potential with spring constant k_i^b. The deviation of the bond length l_i from its equilibrium (reference) distance l_i^{ref}, either stretching or compressing, result in an increase of energy; (2) the angle bending (a three-body term), modelled by a harmonic potential on the angle. $\theta_i - \theta_i^{ref}$ equals to the difference between the angle θ_i to its reference bond angle θ_i^{ref} and k_i^a denotes the bending force constant; and (3) the proper and improper dihedrals (four-body terms), modelled by a periodic cosine function, where n denotes the dihedral multiplicity, ω the torsion, r the phase shift and V_T the force constant refers to the energy barrier associated with the rotation of a dihedral angle, computed based on fixed lists. The latter is used for keeping planar groups planar and to prevent a flip to their mirror images (Equation 8)

Secondly, the non-bonded interactions, which include (4) the Lennard-Jones (a two-body term) consist of a repulsion term caused by the Pauli exclusion principle and a dispersion term, where the equilibration distance σ_{ij} between atom i and j depends on pairs of atom types, can be taken from look up table of Lennard-Jones parameters, ε_{ij} is the potential well depth and r_{ij} is the distance between pairs of atoms i and j; and (5) Coulomb and modified Coulomb interactions. The expression is pair-additive, in which all pairs of atoms separated by at least three bonds are calculated intra- and inter-molecularly with the Coulombic law, where r_{ij} is the distance between two partial atomic charges q_i and q_j, and ε_0 is the dielectric constant. Calculations are computed based on a neighbour list, listing non-bonded atoms within a certain radius and are typically solved by assuming a constant dielectric environment beyond the cut off with a dielectric constant. (Equation 8)

Thirdly, constraints including position, angle distance etc. are based on fixed lists, which is not shown in the Equation 8.

2.4 Molecular dynamics (MD)

MD is used to simulate the dynamical behaviour of the system in real time, and under real conditions whilst considering solvent and ions etc. MD solves Newton's equations of motion (Equation 9) to derive atomic positions to describe how the system evolves with time. The output coordinates at regular intervals are saved to a trajectory; an ensemble of conformations which will reach to an equilibrium state. The method of MD is a deterministic method, that is the state of the system of the future can be predicted from current state.

Equation 9 - Newton's equation of motion, the forces are the negative derivatives of the potential function V:

$$m_i \frac{\partial^2 r_i}{\partial t^2} = F_i = -\frac{\partial V}{\partial r_i}, \qquad i = 1 \cdots N$$

, where m_i is the mass of the ith atom, and the V is the potential energy of the system with respect to position r_i. MD calculations are broken down into small time steps to simulate the real time potential. At each step, the atomic forces on the atoms are computed and with the current positions and velocities new positions and velocities are generated.

Velocities are randomly generated (temperature) as an initial step acting on all atoms. The system then evolves from the starting point using the velocity (v) at time t = 0 by solving Newton's laws of motion; resulting in a set of coordinates at time t = i. From the coordinates, the potential energy can be obtained. The first derivative of this energy gives the force acting on each atom, which become the new velocities for the next step. This strategy allows exploration of a greater fraction on the energy landscape. The force acting on each atom is constant during the time interval, and is implemented using a force field. The positions of the atoms in a small-time interval can be expressed by a Taylor expansion that depends on the velocities, acceleration and hyperaccelerations.

Periodic boundary conditions are applied to prevent artefacts arising from the box edges, and are used with the minimum image conversion where only the nearest image of each particle is used to compute short range non-bonded interactions. Integration methods of MD can be either the velocity Verlet (Tinker) [117] or the leap-frog integrator (Gromacs) [118] used to update position and velocities of the MD simulations.

In velocity Verlet algorithm, positions $\mathbf{r}_i(t)$ and velocities $\mathbf{v}_i(t)$ are defined at each time step and the trajectory for N particular are generate iteratively using Equation 10:

Equation 10 - The velocity Verlet algorithm:

$$\mathbf{r}_i(t+\delta t) = \mathbf{r}_i(t) + \delta t \mathbf{v}_i(t) + \frac{(\delta t)^2}{2m_i}\mathbf{F}_i(t)$$

$$\mathbf{v}_i(t+\delta t) = \mathbf{v}_i(t) + \frac{\delta t}{2m_i}[\mathbf{F}_i(t) + \mathbf{F}_i(t+\delta t)]$$

, where $\mathbf{F}_i(t)$ is the force on particle i at time t and is calculated from the potential energy function.

The drawbacks of the MD approach, include (a) when the dynamics of the particle is described by the classical mechanics, hydrogen atoms and high frequency vibrations which require quantum mechanical treatments to properly represent are not modelled. One could either apply corrections to the total internal energy or constrains on bond lengths (or bond angles). Default setting in gromacs is LINCS or P-LINCS [119] which enable constraints to be parallel processing across the nodes. (b) MD describes the time evolution of nuclear positions alone and neglects of electronic motions, limiting MD to model the dynamics of reactions, breaking and formation of chemical bonds. (c) MD uses force field or potential energy function, which is semi-empirically derived using experiments and quantum mechanics calculations, to define the interactions between atoms. As a result, force fields are fixed during the course of simulation to mainly pair additive (including non-bonded forces), with the exception of long range Coulomb forces and that the polarizabilities are not considered.

CHAPTER 3. A Random Forest Model for Predicting Allosteric and Functional Sites on Proteins

This chapter is based on my publication:

Chen, A. S. Y., Westwood, N. J., Brear, P., Rogers, G. W., Mavridis, L., & Mitchell, J. B. O. (2016). A Random Forest Model for Predicting Allosteric and Functional Sites on Proteins. *Molecular Informatics*. **35**, 125-135.

3.1 Introduction to allostery

Allostery is a universal mechanism for regulation of a protein's activity, typically an enzyme, by the binding of a ligand molecule to a cleft other than the protein's active site. In contrast to conventional use of orthostery as a simple on-off device, allosteric regulation can act as a dimmer switch, and offer greater fine modulatory control over the level of protein activity. [120] A typical enzyme has one active site, but may have multiple allosteric sites.

3.1.1 Old view

The traditional understanding of allostery focuses on those binding events that induce a conformational change affecting the activity of another site of the protein. The classical explanation of how allosteric regulation is achieved was proposed in the Monod-Wyman-Changeux (MWC) [121] and the Koshland-Nemethy-Filmer (KNF) [122] models (Figure 8), where the cooperativity between subunits of an oligomeric protein is coupled with a conformational change. According to the MWC model, cooperativity is achieved by a concerted transition between two alternative states, the protein being in either the T (tense) or R (relaxed) state. For the KNF model, a binding-induced conformational change in one subunit is propagated sequentially among other subunits. Both models imply that the conformational change at the substrate binding site results from the transmission of a signal initiated by allosteric effector binding. [123,124]

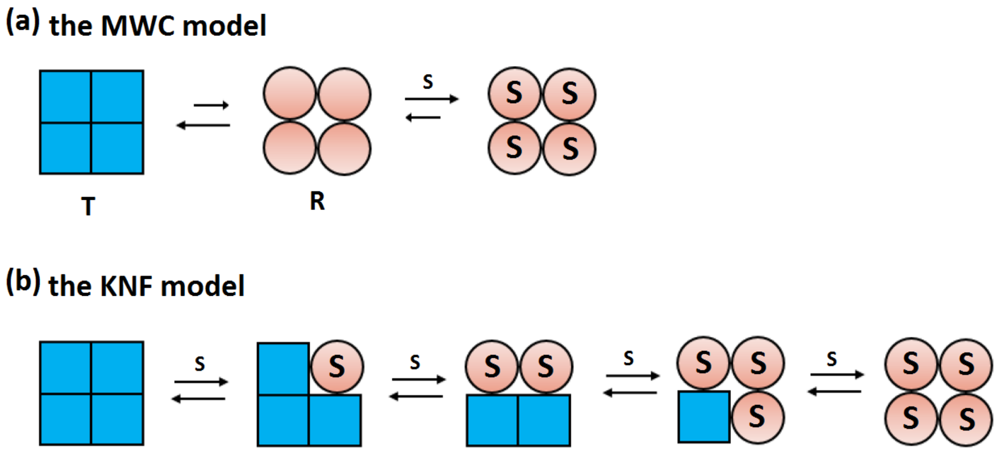


Figure 8 – Modified from [125]. Classic models of allostery (a) the concerted MWC model and (b) the sequential KNF model, where S represents substrates. Each subunit is in either a tense (T) or a relaxed (R) state, where the R state is more receptive to ligand binding.

3.1.2 New view

Conformational state redistribution is a concept that has been proposed to explain allosteric regulation. The native protein appears to exist as a conformational ensemble. [126,127] In contrast to the oversimplified classical models, Weber proposed that the binding results merely in a population shift of conformational states which were experimentally proved to have an effect on function. [128] Population redistribution enriches certain pre-existing conformations which were previously hardly seen due to low population. It is through the interconversion of the functional conformations that allosteric regulation is achieved. [129,130]

Thus, Del Sol et al. [120] think of allosteric regulation as redirecting the levels of traffic on dynamic communication pathways that already existed prior to effector binding, rather than establishing new pathways. They note that allosteric regulation can occur in the absence of significant conformational change, though some kind of communication between sites must take place.

3.2 Allostery and Drug Design

The discovery of new allosteric sites is of interest for drug design. In contrast to active site inhibitors, allosteric binding can lead to either an increase or decrease in activity of a protein. In addition, allosteric effectors do not necessarily share similar chemical properties with the natural substrate, as a site distinct from the active site is targeted. This provides an alternative route for the discovery of promising new leads for regulation of the same target. Allosteric sites on proteins are also subject to lower evolutionary pressure compared to the active site, which is beneficial when designing target-specific inhibitors. [131]

Despite the advantage of variation among homologs that an allosteric site has, this may cause difficulty in studying allosteric mechanisms, since the allosteric sites are hard to predict by traditional homology methods based on sequence similarity. [131] For protein families where a reasonably large number of sequences are available, a more effective approach to sequence-based allosteric site prediction is to assume that allosteric sites are associated with networks of co-evolving residues. [132,133] In this way, Novinec et al. [134] identified a network of co-evolving residues putatively responsible for communication between allosteric and functional sites from a multiple sequence alignment of papain-like cysteine peptidases. This prediction, along with associated experimental work, allowed them to identify a promising inhibitor candidate.

Panjkovich and Daura [135] applied normal mode analysis (NMA) to consider changes in the flexibility of a protein upon ligand binding. To achieve this, ligands were represented as dummy atoms arranged in an octahedron. For each putative binding site, the NMA-derived B-factors of the apo and the bound states were compared in order to identify any large changes in the B-factors, these indicating potential allosteric sites.

A two-way classification model was proposed to differentiate allosteric from non-allosteric sites by Huang et al. [136] They developed a support vector machine (SVM) based machine learning model, based on 90 allosteric sites selected from allosteric database (ASD) and 1360 predicted non-allosteric sites from the same set of proteins using the Fpocket algorithm. For their SVM model, sets of site descriptors were derived to characterise the topological structure and physicochemical properties of both types of sites, obtaining a total of 41 site descriptors. A somewhat related method has been adopted by van Westen et al. [137] to select allosteric modulators based on the physicochemical and structural descriptors calculated for those molecules from the ChEMBL database. [138] Several machine learning approaches have also been used with other dynamic-or NMA-based approaches to predict the location of allosteric sites.

[139,140]

Other studies relevant to the prediction of allosteric interactions focus on simplified models of protein dynamics, using approaches like NMA [135], energy exchange [141], and Monte Carlo path generation [142].

3.3 Serendipitous binder

Almost all protein crystal structures contain non-cognate bound ligand molecules, such as stabilising agents and buffers used during crystallisation [143], which was originally regarded as a crystallisation artefact and was once believed to have no effect on protein function, despite their role in maintaining protein solubility and stability for NMR experiments. Yet growing evidence of its effect on protein dynamics implies that protein function will be affected by ligand binding [144]. These molecules represent a potential starting point to design novel probes for new allosteric sites and as a tool to study changes in protein dynamics induced upon the binding of a buffer molecule. [145] In this study, buffer molecules are introduced as potential binders to identify locations of possible allosteric sites.

3.4 Aims and Objectives

In this work, we focused on identifying potential allosteric sites, while making better use of available crystal data in the PDB. We used the co-crystallised ligands to calculate descriptors from the ligand and from the structures of the sites, thus building a machine learning model. Our aim is to identify binding sites which are purely crystal contacts from potential allosteric sites. These bound ligands could be a starting point to guide experiments aimed at probing the nature of the sites.

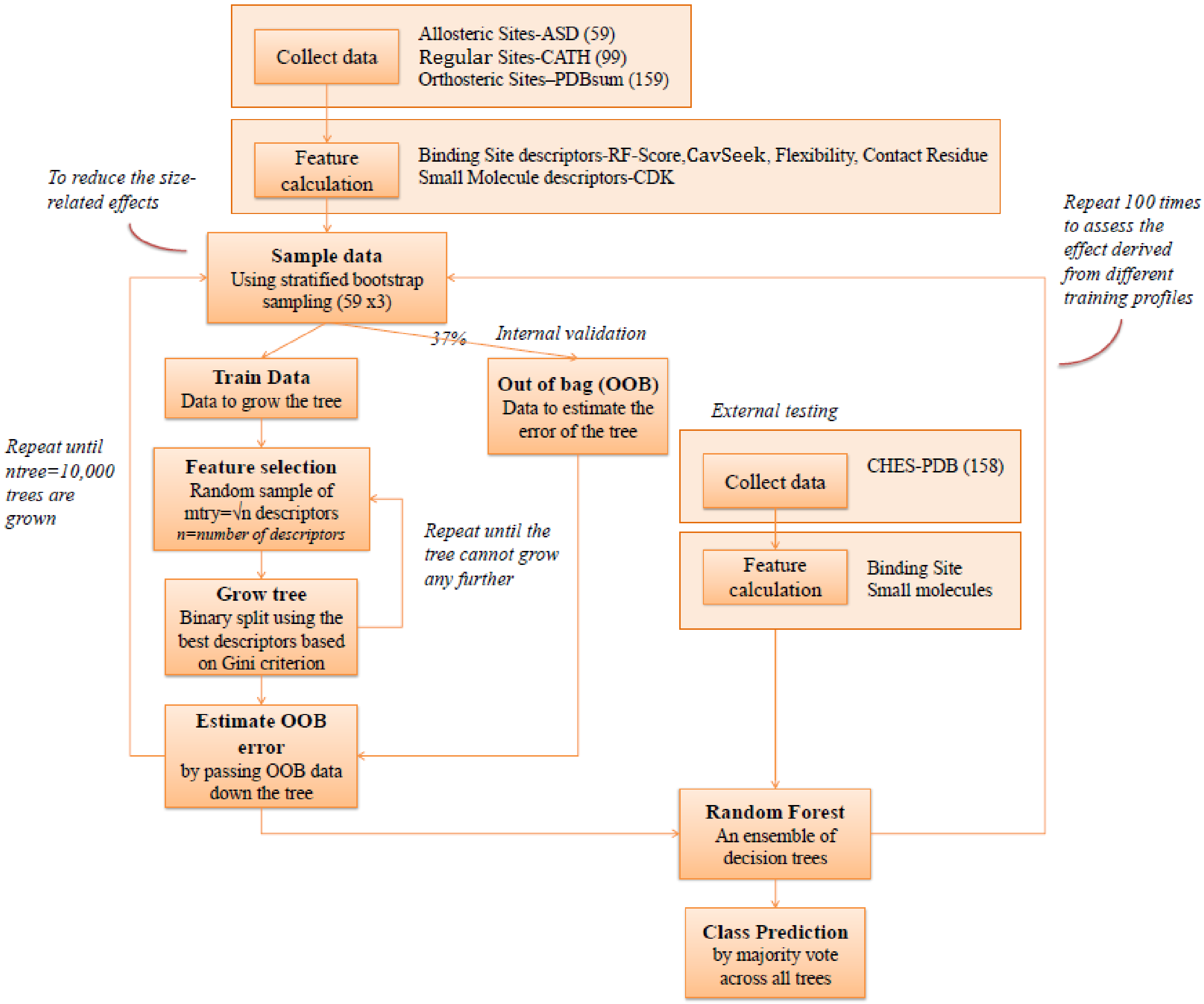
3.5 Method

To do this, we use a complementary approach, founded on a deeper analysis of the structures of potential binding sites. We assemble collections of three kinds of site based on its function: first, known orthosteric functional (active) sites of proteins in which the main cognate ligand binds; second, allosteric sites in which allosteric effectors can bind; third, a structurally representative set of other protein clefts, expected to be neither functional nor allosteric. For these three sets of sites, descriptors are proposed to identify and discriminate the binding state of individual ligands between the three different subsets. We use our existing protein-ligand scoring function RF-Score [146] and a new accessibility-like algorithm called CavSeek to compute structurally-based binding descriptors and descriptors pertaining to the composition and flexibility of the clefts. We use these as features in a ternary predictive model, employing the Random Forest machine learning algorithm. We take advantage of the out-of-bag data, [147] and separately those instances omitted from the stratified balanced samples, to conduct a fair validation, which uses only data excluded from model building. Then the model is subsequently used to predict the types of sites where CHES binds, with the objective of identifying candidate allosteric sites on proteins. The challenge was to differentiate the binding sites based on a combination of descriptors. In presenting our result, we investigated whether the results previously obtained through manual inspection corresponded to those obtained with our computational approach.

Our work is distinct from Huang et al.’s [136] in our design of a three-way predictive model containing two classes other than allosteric, and distinct from van Westen et al.’s [137] in that our work predicts allosteric sites (not molecules) using co-crystallised molecules and descriptors derived from the structure of the sites as well as from the ligands.

Workflow

Figure 9 - A flowchart of the methodology.


Collect data
Allosteric Sites-ASD (59)
Regular Sites-CATH (99)
Orthosteric Sites–PDBsum (159)
Feature calculation
Binding Site descriptors-RF-Score,CavSeek, Flexibility, Contact Residue
Small Molecule descriptors-CDK
To reduce the size-related effects
Repeat 100 times to assess the effect derived from different training profiles
Sample data
Using stratified bootstrap sampling (59 x3)
37%
Internal validation
Train Data
Data to grow the tree
Out of bag (OOB)
Data to estimate the error of the tree
External testing
Repeat until ntree=10,000 trees are grown
Feature selection
Random sample of mtry=√n descriptors
n=number of descriptors
Repeat until the tree cannot grow any further
Collect data
CHES-PDB (158)
Feature calculation
Binding Site
Small molecules
Grow tree
Binary split using the best descriptors based on Gini criterion
Estimate OOB error
by passing OOB data down the tree
Random Forest
An ensemble of decision trees
Class Prediction
by majority vote across all trees

3.6 Collection of training data

We have annotated our data according to where the ligand has bound to its protein using three classes: allosteric, regular and orthosteric sites. Each subset was included independently, and for convenience these are denoted by the capital letters A, R and T, respectively.

3.6.1 Allosteric Sites (A)

A total of 91 proteins adopted from Panjkovich and Daura's work were initially used to represent the subset of allosteric sites in the training set. [131] The data were primarily collected from the online AlloSteric Database (ASD) and from the literature, and were further filtered to be structurally non-redundant by the sequence clustering program BLASTClust. The protein with the highest resolution structure of each of the resulting 91 groups was selected to represent that group. ASD [148] provides a list of the allosteric residues in the given protein. We compared those residues, thus annotated as comprising an allosteric site, to the list of residues involved in ligand binding extracted from PDBsum. [149] From this, we can identify any ligand that is bound in the allosteric site in order to obtain descriptors which capture the binding profile of the ligand in the allosteric pocket. If there are many instances in which the same ligand adopts an equivalent binding mode, the one with the highest RF-score value is kept in the subset to represent the particular binding pattern. Thus, the list has been whittled down to 59 representative allosteric (A) protein pockets. (see Table 1 in Appendix for protein lists)

3.6.2 Orthosteric Sites (T)

A total of 195 protein-ligand complexes representing the subset of orthosteric (T) sites were retrieved from the PDBbind database (version 2007). [150] These data were originally used for the purpose of validating scoring functions in Cheng et al.'s study. [151] The data contain experimentally determined binding affinity values obtained from the literature. Cheng et al. further filtered their initial collection of data to account for the quality of structures, the quality of binding data, the components of complexes and redundancy of protein sequences, to avoid over-representing certain families. They clustered the remaining complexes according to sequence similarity and selected the complexes with the highest, the median and the lowest binding affinity to represent each of the 65 clusters, giving 195 complexes in total. In this study, we have further whittled down the number to 159 complexes which have only small molecules in the pocket.

3.6.3 Regular Sites (R)

The regular site subset was derived from a representative set of protein domain structures, each of which is given by CATH [152] as an example representing the homologous superfamily to which it belongs. From each such structure, one ligand binding site is selected according to PDBsum. [149] For enzymes, we chose sites having a ligand which is neither a cofactor nor similar to the enzyme's product or substrate. Ligands were selected to have no contact with any residues of any allosteric site given in ASD. Therefore, the sites occupied by the selected ligands are unlikely to be active sites or allosteric pockets. The regular subset is expected to have the weakest binding affinity and the lowest burial value of the three subsets. These weak interactions correspond to the regular binding events by which non-cognate ligands bind, possibly as accidents of crystallisation.

A total of 99 instances were selected for the subset of regular (R) sites. The number representing each class was designed to be proportional to the prevalence of that structural class amongst all CATH [152] superfamilies (2620 superfamilies in total). There are four top C-level classes defined in the CATH database. Table 1 shows the number of entities included from each CATH class.

Table 1 - Distribution of regular sites amongst CATH C-level classes.

Class	No.
Mainly alpha	32
Mainly beta	20
Alpha beta mixture	42
Few Secondary Structures	5

3.7 Collection of external testing data (CHES)

The PDB crystal structures containing the buffer molecule CHES (*2-(N-cyclohexylamino) ethane sulfonic acid)* were investigated. CHES is one of the many buffer molecules that commonly complex with proteins during the crystallisation process.

In total, 82 CHES containing entries had been released in the PDB up to Dec 2013. From these, our external testing CHES set of 158 CHES-protein binding sites (some proteins have multiple CHES ligands) has been identified and each site is characterised by a set of descriptors individually calculated for it. We noted 14 cases in which CHES was bound in a protein's defined pockets [145], from which only one of these 14 CHES molecules was found in an allosteric site, that of a bacterial sialidase (NanB). [153] There results were manually identified by Brear and Westwood [145], who were hoping to see if the CHES was bound at the site where other small molecules can also bind. We have further specified which one or more of multiple CHES molecules in a given structure were being referred to in their review results by using literature searching to identify the cavities as allosteric sites or otherwise.

3.8 Random Forest

We used the randomForest package in R [154,155] to build predictive models with the default setting of *mtry* to the square root of the number of descriptors and with *ntree* set to 10,000. Random Forest is widely considered relatively immune to overfitting. Each tree is grown by stochastic recursive binary partitioning, and the individual trees carry independent information because of the substantial random element in their construction.

Three further considerations apply to the use of Random Forest in this work. First, each tree is built by bootstrap sampling from the same balanced dataset, which we constructed by stratified sampling (subsampling) to include an equal number of objects (53) from each class, a total of 159 sites to avoid bias due to imbalanced class sizes. Within this stratified balanced set, the bootstrap sampling means that approximately 63% (≈ 1 – 1/e) of the data are used once or more in the building of each constructed tree, and the remaining 37% (≈ 1/e) or so are reserved for OOB validation of that tree. The bootstrap sampling from the balanced set is repeated afresh for each of the 10,000 trees.

Further, the performance of the Random Forest model is assessed firstly on the OOB data and secondly on the external test set consisting of the158 sites (46 R, 106 T and 6 A) omitted from the stratified (balanced) dataset. Those data excluded from the stratified balanced set in advance of the bootstrap sampling form an external test set which was separately used for further validation. This entire process of generating 10,000 trees was itself repeated ten times with different random seeds to avoid losing information from the majority class in training the models, see Figure 9.

Finally, Random Forest is designed to handle the inclusion of redundant and irrelevant descriptors through the process of selecting possible splits at each node from a substantial set of randomly chosen options. [147] This obviates any need for an explicit descriptor selection step, and is particularly useful when a bespoke descriptor set is used, as in the present work.

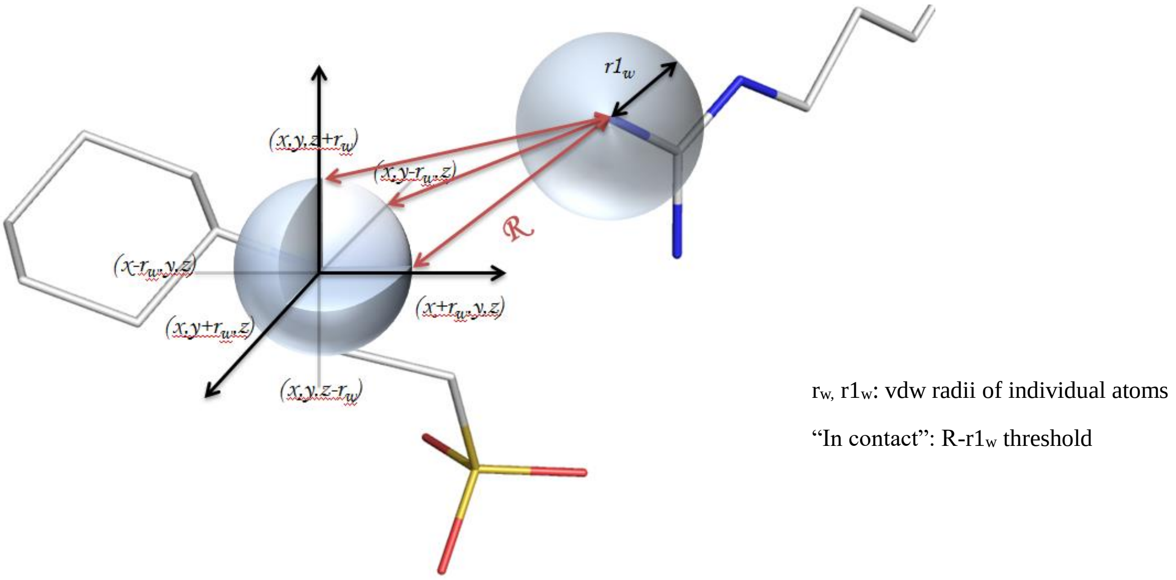
3.9 Descriptor

Every binding instance was characterised by the same set of descriptors.

Table 2 - List of descriptors and their abbreviations.

RFSCxCSK	The RF-Score (R) times average burial over nine thresholds estimated by CavSeek	Binding Site
RF.score	The unadjusted RF-Score (R)	Binding Site
NormRFScore	The normalised RF-score (R/E)	Binding Site
Function_F195	The expected RF-score (E), calculated by a fitting function E=2.222 N⅓	Binding Site
B_protein	Average B-factor of the protein	Binding Site
B_pocket	Average B-factor of the contact residues defined as protein residues <4Å to the ligand	Binding Site
noContact_resi	Number of contact residues	Binding Site

3.9.1 CavSeek



r_w, $r1_w$: vdw radii of individual atoms

“In contact”: R-$r1_w$ threshold

Figure 10 - The burial percentage is calculated as the number of distances considered to be “in contact” (less than or equal to a certain threshold) over the total number of measurements taken. The distance is measured between points of any CHES atom (on the left) to the van der Waals sphere of any protein atom.

In order to measure the burial of a ligand within the cavity of the protein binding site, we developed an accessibility-like program called CavSeek using a script written in Java. For a given protein and ligand, we calculate the percentage of possible point-to-atom contacts which are shorter than a given threshold value and hence are said to be "in contact". A number of thresholds from 0.5 to 2Å have been selected to profile optimally and identify a ligand's binding site. The aim of this program is to make it possible to discriminate computationally between surface-binding molecules and pocket-binding molecules. Ligands that are found within a protein cleft in a small binding pocket will have a higher percentage of sub-threshold contact distances. The percentage burial increases with the size of the thresholds as more points are counted. To generate descriptors from CavSeek, one can either include the result at each different threshold as a separate descriptor, or calculate an average burial as a single descriptor. For this study, we have included the burial at nine individual thresholds and also the average burial.

In detail, CavSeek first centres the protein-ligand complex at the geometric centre of the ligand. We remove all protein atoms which are more than 20Å away from this origin, since there is a very little prospect of those atoms having a significant interaction with the modestly sized ligands that we study. We then represent each ligand atom as a sphere using the following van der Waals radii (r) in Ångstron: Br (1.85), C (1.7), Cl (1.75), F (1.47), Fe (2.0), I (1.98), N (1.55), O (1.52), P (1.8) and S (1.8). [156] For an atom at (x, y, z), we define six points on the van der Waals sphere along the co-ordinate axes at (x±r, y, z), (x, y±r, z) and (x, y, z±r). (Figure 10) For each of these six points, we calculate the shortest distance to the corresponding van der Waals sphere around any protein atom. The hydrogen atoms in both ligand and protein are ignored. For a ligand with M atoms, this results in $6M$ distances, each of which is compared with the threshold. All distances less than or equal to the threshold are taken to be "in contact" at that threshold, and the percentage of the $6M$ distances that are "in contact" is recorded. This is repeated for all nine chosen threshold values. In the cases when crystal structure reveals alternative conformations of the side chain due to a partial occupancy, the first conformation listed in the PDB file was kept for this study.

3.9.2 RF-score

RF-Score [146] is our group's machine learning approach to predicting protein-ligand binding affinity, especially for docked structures. Previous knowledge-based approaches used ensembles of observed protein-ligand crystal structures to infer binding energies from atom-atom distance distributions. That approach makes the dubious assertion that

Boltzmann energetics apply, assuming a particular exponential functional form to transform distance distributions into binding energies. [157] RF-Score uses Random Forest to predict binding affinities from both structural data and the affinity data that are left unused in most knowledge-based approaches, yielding a much more accurate and flexible scoring function.

In order to make the scores of differently sized ligands comparable, and to compensate for the intrinsic size-dependency of scoring functions, we calibrate RF-Score according to the number of heavy atoms (N) of its ligand. [158,159] Figure 11 illustrates the variation of the unadjusted scores, which we empirically fitted to a small number of physically justifiable functional forms. We empirically found that the best fitting function defining the expected score (E) for a ligand of given size was

$E = 2.222\ N^{1/3}$

For each ligand, we calculate the unadjusted RF-Score (R), the expected score (E), and the normalised score (R/E).

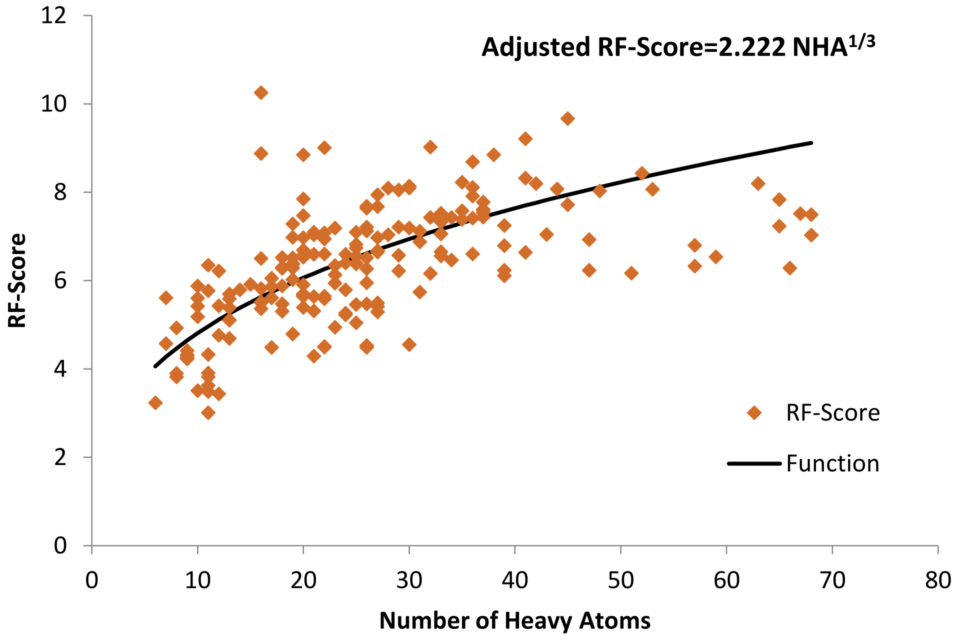


Figure 11 – Normalisation of RF-Score.

Each point represents an individual RF-Score of a different protein-ligand complex selected from the PDBbind database [150] used in this study as part of the subset of orthosteric sites within the training set. The fitted curve illustrates the function used to calibrate the scores with the ligand size.

3.9.3 Temperature Factor

To include features that describe flexibility, we have used the temperature factor (or B-factor). The B-factor, which reflects the degree of atomic displacement from their equilibrium positions in the crystals due to thermal motion, was extracted from the X-ray crystallographic structures of the protein-ligand complexes in the PDB. A higher B-factor implies that the atom has greater mobility. The average B-factor of the contact residues is divided by that of the protein to obtain values that reveal the differences in flexibility of the ligand binding region with respect to the entire protein. Firstly, to consider the bias arising from chain termini; the average B-factor of the protein with gradual omission of up to 10 residues at both ends was calculated. The results showed no significant change in the average B-factor between each omission; accordingly, proteins have been kept without terminal elimination. Secondly, the solvent and other ligands or cofactors were removed to obtain a B-factor resulting solely from the protein residues. The contact residues herein were defined as residues having at least one atom within 4Å of the centre of any atom of the ligand. B-factors of all the atoms of the contact residues and the protein are averaged and were included both as ratios and as separate descriptors in this study.

3.9.4 Contact Residues

Contact residues, which were defined as residues with an atom (or atoms) that are closer than 4Å to any atom of the ligand (as defined above), were utilized as descriptors to reflect the physicochemical composition of the ligand binding site. This includes a simple count of the total number of residues and the occurrence frequency of each of the 20 amino acids. Moreover, the contact residues are further grouped according to their side chain chemistry into charged (R, H, K, D and E), polar (S, T, N and Q), hydrophobic (A, V, I, L, M, F, Y and W), aromatic (F, Y, W and H) and special (C, G and P) categories. Each count was taken as an individual descriptor.

3.9.5 Small Molecule Descriptors

We used the Chemistry Development Kit (CDK) [160] to compute descriptors for small molecules. CDK is an open source library written in Java for structural informatics calculations. First, the chemical structures of the ligand were inputted as SMILES extracted from each ligand structure file (in SDF format). Second, we calculated 277 CDK descriptors for each compound, and removed features without discriminant power, those having either the same or an undefined value for all compounds in any of the training subsets. As a result, only the remaining 141 CDK descriptors were kept for further analysis.

3.10 Result and Discussion

3.10.1 Predictive performance (OOB)

Prediction is based on a majority vote over the set of 10,000 trees (*ntree*). One vote is made by each tree for each instance that is OOB (not used in building the tree, because it was not chosen during the bootstrap sampling) by passing the OOB data down each tree to obtain a class prediction. From the aggregated OOB predictions, classes are assigned to each OOB instance by a majority vote of the trees. The OOB error, which shows the percentage of misclassification in the dataset, was calculated based on the known and predicted class labels. Separately, we also test the Random Forest's predictivity by passing down each tree the external test set comprising those data that were omitted from the balanced set (46 R, 106 T and 6 A sites).

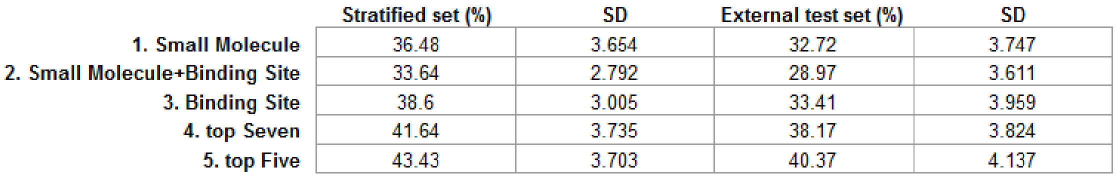
Random Forest is insensitive to values of *mtry* except close to its high and low extreme values. [161] It was empirically shown in [147] that the performance of the Random Forest remains unchanged over a wide range of *mtry* values, and the defaults work the best for the majority of cases. For all 10 repeats (each of 10,000 trees per model), the default *mtry* was used. Five models were built using various sets of descriptors, which are classified as either small molecule or binding site descriptors according to the physicochemical features captured. Some of the most significant descriptors are listed in Table 2. For each model, we computed the average OOB error to estimate the prediction error thus to assess the accuracy that is independent of particular repeats; see Table 3.

The OOB error is sensitive to the random determination of which protein-ligand complexes are kept in the training set, in general, with 3-4% deviation from the average. The first Random Forest model was trained using a total of 151 small molecule descriptors including 141 CDK descriptors and the heavy atom counts of each ligand. (see Table 2 in Appendix for the lists) The average OOB error of the Random Forest models obtained is 36.48% on the stratified balanced set, in which the pocket has been assigned a class label solely based on the small molecule descriptors of the ligand that binds to it. By the addition of 43 binding site descriptors, the second Random Forest model which includes properties of all calculated descriptors of both the bound ligand

and the site has a slightly improved error of 33.64%. Both models contained descriptors based on the structures of the small ligand molecules. These are invariant within the CHES set as the same compound CHES was used to characterise the pocket in each case. Thus, those models are not used in predicting our CHES set since these are descriptors without discriminating power for that set.

Our third model used 43 binding site descriptors that describe ligand binding in terms of predicted affinity (RF-score), a percentage scoring scale for ligand burial (CavSeek), binding site flexibility (B-factor) and binding site hydrophilicity or hydrophobicity derived from analysis of the pocket composition. The model produced an average OOB error of 38.6% on the stratified balanced set. Subsequently, it was used on the CHES set to generate predictions for the CHES-protein binding pockets.

Table 3 - Average OOB error rates for the different models.



	Stratified set (%)	SD	External test set (%)	SD
1. Small Molecule	36.48	3.654	32.72	3.747
2. Small Molecule+Binding Site	33.64	2.792	28.97	3.611
3. Binding Site	38.6	3.005	33.41	3.959
4. top Seven	41.64	3.735	38.17	3.824
5. top Five	43.43	3.703	40.37	4.137

The OOB errors are presented as the percentage of misclassified data points in the stratified balanced set and separately in the external test set (comprising data excluded from the stratified set). Standard deviations are calculated over a hundred runs using different random seeds (10,000 trees per run), using N-1 = 99 in the denominator.

3.10.2 Descriptor importance

The importance of the individual descriptors can be evaluated either with the permutation method by observing the effect on the predictivity of Random Forest models of 'noising up' each descriptor, or alternatively with the Gini index, an impurity measure. The mean decrease Gini (MDG) (calculated over all trees) is a measure of improvement to the purity when that descriptor is made available to split the trees, thus producing greater purity in the resulting nodes. The decreases in Gini impurity for each descriptor used to form splits are summed over all trees and then normalised. A higher value implies greater importance of the variable concerned. Here, we report the results of variable importance as measured by impurity reduction, see Figure 12.

The top ranked binding site descriptors obtained by averaging the Gini importance values from 10 repeats are obtained. The leading descriptors are: first the product of the RF-Score and the average score of CavSeek (RFSC×CSK), second the RF-Score values (RF.score), followed in third place (but with a significant decrease in importance) by protein flexibility (abbreviated to B_protein), fourth the residue count of the ligand binding site (noContact_resi), and fifth the normalised RF-Score (NormRFScore).

The subsequent important descriptors are: sixth the flexibility based on the contact residues (B_pocket), and seventh the expected RF-Score (Function_F195, computed by size calibration with the number of heavy atoms N to the original score as $2.222N^{1/3}$). Those two have very similar Gini importance values.

Similar importance rankings were found in all ten repeats, but they sometimes slightly differed in order. The calculation of relative importance allows a further assessment firstly of the classifiers based on the full set and secondly on classifiers based only on a few of the most important descriptors as a potential way to improve the performance, since generalisation tends to perform better in a lower dimension. To achieve this, we select the top 7 descriptors (from which to build the fourth model) and top 5 descriptors (for the fifth model) due to the breaks in the curve of the Gini importance plot, Figure 12, indicating a considerable drop of importance from the fifth to the sixth variables and similarly from the seventh to the eighth. The predictive ability of the models with reduced numbers of descriptors, as measured by the OOB error, is shown in Table 3. An increased overall OOB error is observed as the number of variables is decreased by 3.04% and 4.83% for the stratified set, relative to the model based on all binding site descriptors. Apart from the OOB error calculated, we also look for consensus of the results of computational predictions and literature findings, as discussed below.

The results also show that our largest threshold of 2Å is desirable for CavSeek to achieve optimal discrimination between binding sites, based on the relative descriptor importance of the CavSeek scores at different thresholds. The version with the 2Å threshold is listed eighth in the variable importance ranking. CavSeek is combined with RF-Score by multiplication to increase their discriminative power, hence avoiding the difficulties inherent in adding or subtracting quantities with different dimensions. The combined descriptor of RF-Score and CavSeek improved the RF-Score by itself and is listed as the most important variable averaged from ten runs. RF-Score itself is listed as the second.

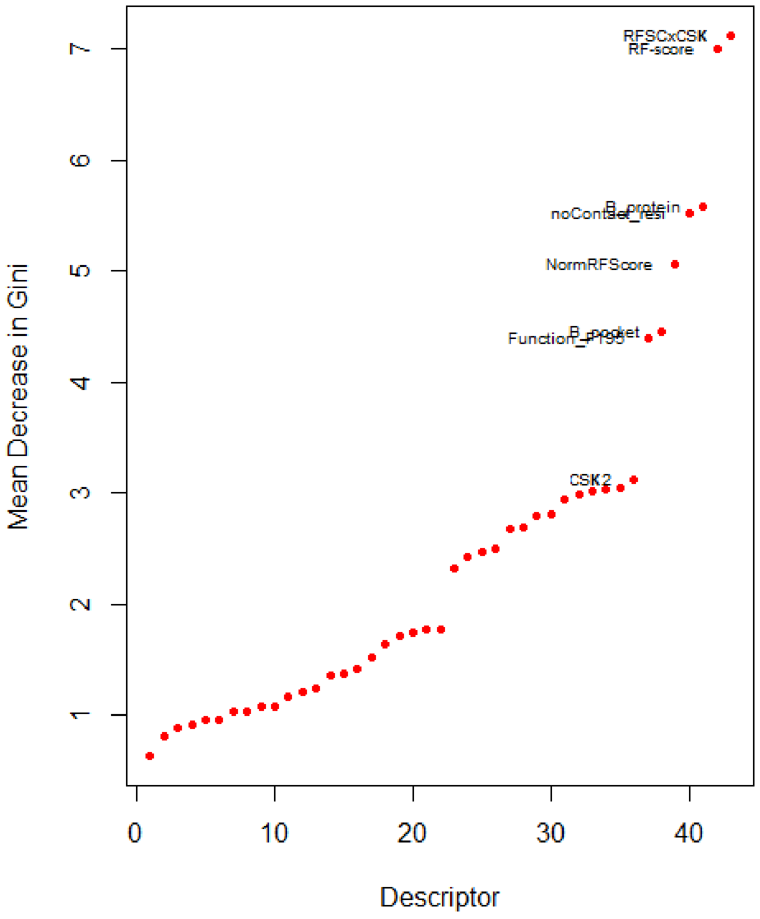


Figure 12 - The mean Gini importance values of each descriptor from the third model, averaged over ten repeats.

The plot shows variable importance on the y-axis ordered from the most to the least important. The descriptors with the highest decrease in Gini impurity make the major contributions to partitioning the data into homogeneous classes.

3.10.3 Predictions for the CHES set

Here we collate the number of times each class is predicted for each CHES binding instance and report the class with majority votes from hundred repeats. (Appendix Table 3 and 4) The numbers assigned to each class are given so as to express the approximate level of confidence with which a class has been assigned from hundred repeats.

The model trained using all binding site descriptors returns six orthosteric (T) sites, of which four (pdb codes: 2VW2, two sites in 3OQI, and 3NOQ) showed matches with the manual annotation. The remaining two were known bind to the domain interface (both in 2ICH) interacting with conserved residues which were inferred to have functional role among homologs. [162]

Among the 15 CHES binding instances predicted as allosteric (A), there is lack of literature for 4DQ0 and 3G8W. Both contain multiple CHES binding instances. Our results uncover three potential allosteric sites, which are not known orthosteric sites, supported by the literature. Four were found experimentally in sites considered [145] likely to be orthosteric (two in 3RIG, 1Q1Q and 1V30), see Table 3 in Appendix.

Since CHES is not a natural cognate ligand for any protein it binds to, it is perhaps not surprising that orthosteric sites where the CHES binds (active sites evolved to bind other ligands) have been predicted as allosteric. The ligands in the orthosteric (T) subset of the training set from which the model was built were chosen to be more specific to the corresponding protein; thus, the more buried and stronger binding ligands were expected to be the cognate ones. In the potential future use of this methodology to predict allosteric sites using serendipitous binders, the workflow would therefore be designed to filter out known orthosteric sites from the set of allosteric predictions.

In contrast, our fifth model using the top 5 descriptors resulted in more promising results. Five orthosteric sites have been predicted of which four are consistent with the previously discussed full binding site descriptor model (2VW2, two sites in 2ICH, and 3OQI). An equal number of predictions amongst the 10 repeats assigned 3OQI to the orthosteric and allosteric classes. Three out of five orthosteric predictions were indeed experimentally determined to be orthosteric (2VW2, 3OQI and 1V30), while the remaining two are found at the domain interface (2ICH).

The top 5 descriptor model identified 30 allosteric sites, of which 15 lack definitive description in the literature, six pockets correspond to manually annotated orthosteric sites (two in 3RIG, and one in each of 1Q1Q, 3OB9, 3NIB, and 4H75), and nine pockets were potential allosteric sites. The allosteric sites we have referred to are non-orthosteric clefts, based on the literature. Yet, it is not known whether those pockets are functional allosteric sites, see Table 4 in Appendix.

Unfortunately, the allosteric site obtained manually (2VW2, A1001) by Brear and Westwood [145] was not predicted correctly by either model; the cleft was identified as regular (R). We observed that CHES shares this pocket with a glycerol molecule which is lying deeper in the cavity. (Figure 13)

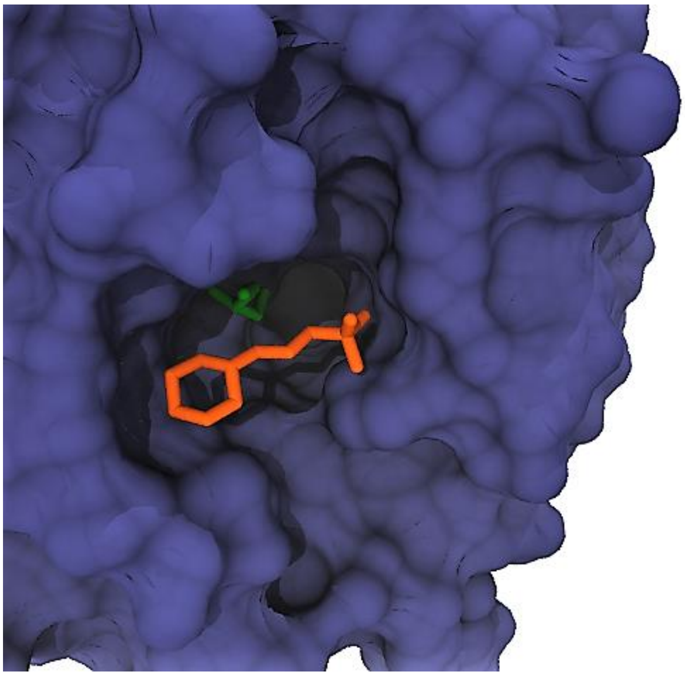


Figure 13 – View of a glycerol (coloured in green) bound to the same pocket with CHES (orange). Structure taken from PDB ID: 2VW2 (purple).

We notice that three interface cavities were assigned to classes A or T (two in 2ICH, and 4ATG), implying that there may be shared features of interface interaction common to the allosteric and orthosteric subsets. Indeed, the interface can potentially act as a binding site for an allosteric modulator. Binding of allosteric modulators at the interface between subunits of GABA receptor has been shown to have varying effects on the receptor's function. [163] Stanget et al. [164] revealed an allosteric binding site at the homodimeric interface of caspase-6 zymogen that impairs function. Descriptors to identify specifically the interaction interface can be exploited; perhaps interface cavities might be included in future work as an independent subset.

One positive note is that, in spite of high error rates (38.6% for the full binding site descriptor model and 43.43% for the top 5 descriptor model) estimated using OOB data, both models have given promising results for potential allosteric sites. Nearly half of our prediction instances are not confirmed by the literature, yet instances that can be found in the literature are annotated as either orthosteric or in a binding cavity different from the orthosteric site. In fact, our top 5 descriptor model predicts most of the defined pockets (10 out of 14) that have been identified by Brear and Westwood to be either allosteric or orthosteric.

Our method provides a fast and low computational cost way to identify potential allosteric sites on large number of crystal structures. The co-crystallised non-cognate ligands and buffers that are commonly seen on most crystal structures are used, from which we extract binding site features. It is implicitly assumed that the crystal structures downloaded from the PDB have the correct ligand binding orientations, though we note that docking and ligand reoptimisation may in future play a role. The predictions were made based only on structures with non-cognate ligand bound. Thus, an adequate description of a binding cleft might not be possible. Also, potential allosteric sites containing no co-crystallised compound are invisible to our trained algorithm. The models were not trained to predict based on specific families. Thus, the number of regular sites included for each of the four structural classes at the C-level of the CATH classification [152] is roughly proportional to its prevalence in the CATH database. However, we noted that a known allosteric site is dominant in some families [137] or perhaps may only exist in particular families, thus introducing a systematic bias. Even though these issues may have contributed to the difficulties in predicting allosteric sites, resulting in a higher-than-ideal error rate, many of our allosteric sites predictions are in agreement with literature findings. Moreover, those non-cognate ligands that co-crystallised with potential allosteric sites can be used as starting structures for the design of probes specifically created for these sites.

3.11 Conclusion

Allostery is a regulatory mechanism that affects protein function by the binding of small molecules to a site distinct from the active site. In contrast to traditional drug design by mimicking natural substrates, allosteric effectors offer therapeutic benefit for target-specific drug design. The discovery of new allosteric sites in protein cavities has emerged as a new drug design approach to identify novel pharmaceutical agents.

In this study, we have used Random Forest to build a three-way classification model for predicting allosteric pockets. We then report the results for a test set in which we consider instances of a buffer molecule, CHES, as a potential binder to allosteric sites; Brear and Westwood [145] observe 14 matches supported by the literature and structural analysis, wherein 10 of these 14 pockets were identified as either the allosteric or orthosteric sites of the protein by our top 5 descriptor (final) model. Although it is questionable whether other predicted pockets are truly functional, the implementation of a machine learning scheme allows discrimination between binding sites according to features that are captured from the protein-bound ligand conformations. This can help reduce the number of PDB files needing to be looked at when hunting for potential novel allosteric sites, prioritising those which are predicted to belong to the allosteric category. Thus, this study shows promising results from using adventitiously binding buffer molecules as agents for allosteric site discovery. However, we also note that predictions of orthosteric pockets were hardly ever made for binding sites of CHES, a non-natural ligand for any protein. CHES appeared to be associated with lower binding affinity and lower burial in protein cavities compared to the ligands of the orthosteric subset used in the model's training. However, mispredictions of orthosteric sites as allosteric will be easy to remove from a set of allosteric predictions, since the orthosteric sites are generally known for the PDB structures we are using. We found several CHES molecules bound in sites predicted to be either allosteric or orthosteric, though actually located at an interface. These can potentially be allosteric modulator binding sites.

We have evaluated the descriptor importance by the Gini importance measure. RF-Score and its combination with CavSeek appeared to have significant discriminative power in identifying the binding pockets. These descriptors reflected the binding states of ligands with respect to their strength of interaction and to their degree of burial in the cleft of the protein.

CHAPTER 4. Docking Novel Ratiometric Sensors for Guanine Quadruplex Structures

4.1 Introduction to DNA G-quadruplex

A non-canonical DNA structure, a G-quadruplex, is composed of stacks of guanine tetrads, G-quartets, stabilised by coordinated cations within the central cavity and by π-π stacking interactions between adjacent G-quartets. A G-quartet is a planar association of four Hoogsteen hydrogen-bonded guanines arranged in a cyclic fashion. These four-stranded DNA structures can be formed in guanine-rich DNA and RNA sequences, and adopt different conformations as a result of different sequences, experimental conditions such as different cations (Na^+ or K^+) or different concentrations and degree of molecular crowding. [165] For example, experiments have demonstrated that G-quadruplexes are stabilised by K^+ ions at 10-50 mM concentration. [166] Molecular crowding, chaperones and dehydrating conditions would accelerate the folding of G-quadruplexes. [167]

In 1910, Bang [168] observed a gel-like aggregate formed from concentrated guanylic acid. X-ray diffraction studies in the 1960's found this aggregate to be four guanine bases hydrogen bonded together in one plane. [169] A decade later, it was found that guanine repeats can form four-stranded nucleic acid secondary structures, named G-quadruplexes. In 1989, Williamson et al. [170] reported G-quadruplex to be stabilised by central monovalent cations.

Quadruplex structures can be of two types, depending on strand polarity: parallel quadruplexes, in which all guanine glycosidic bonds are in the anti-conformation, and anti-parallel quadruplexes, in which both syn- and anti-conformations are present. [171] Balagurumoorthy and Brahmachari [172] suggest that Na^+-rich solutions seem to promote the formation of intramolecular anti-parallel folds, whereas the presence of K^+ appeared to promote parallel fold conformation.

Repetitive G-rich sequences can also be found in promoter regions (i.e. c-Myc, H-Ras and K-Ras), 5' and 3' untranslated regions (UTRs), introns, and the 3' single-stranded end of the telomere (which is 100-200 bases long in a human). In general, G-quadruplexes could fold into either: 1) intramolecular structures formed by folding of a single strand, that could potentially regulate gene expression and chromosome stability or 2) intermolecular G-quadruplexes formed by two (dimeric) or four (tetrameric) separate strands, which are structures formed as intermediates or precursors to recombination (DNA exchange between chromatids) and/or viral integration. (Figure 14) [173]

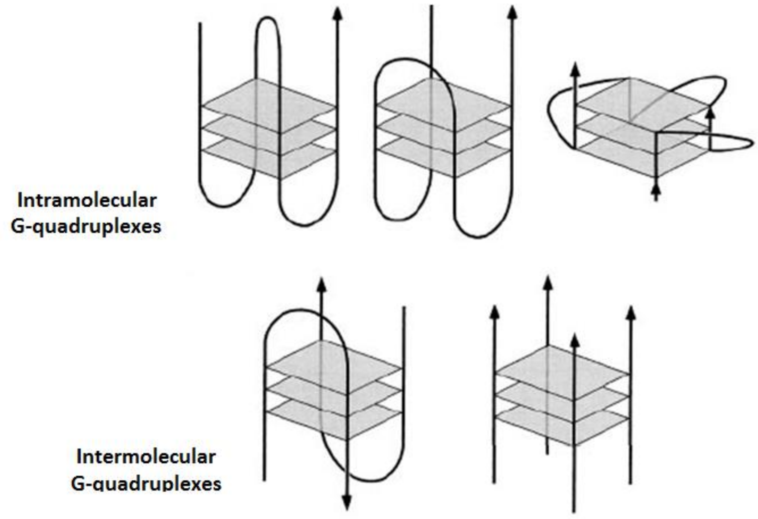


Figure 14 - Modified from [174] shows the structure of intra- and inter-molecular G-quadruplexes.

G-quadruplex specific nucleases have been identified in yeast KEM1 [175] and in human GQN1 (G-quartet nuclease 1) [176], which cut single-stranded DNA located upstream (toward 5') of a quadruplex structure, releasing intact quartets. These endonucleases function to cleave: a) intramolecular G-quadruplexes at telomeres, to allow access of telomerase for telomere maintenance, and b) intermolecular G-quadruplexes that form during chromosome pairing in meiosis. [177] G-quadruplexes function as regulatory elements in gene expression; the characterisation of G-quadruplex specific nucleases clearly supports this view, indicating that G-quadruplexes could be a viable target for therapy. Efforts have been made to identify small molecules that can bind to G-quadruplexes with high specificity, such as N-methyl mesoporphyrin (NMM) [178] and telomestatin [179]. (Figure 15)

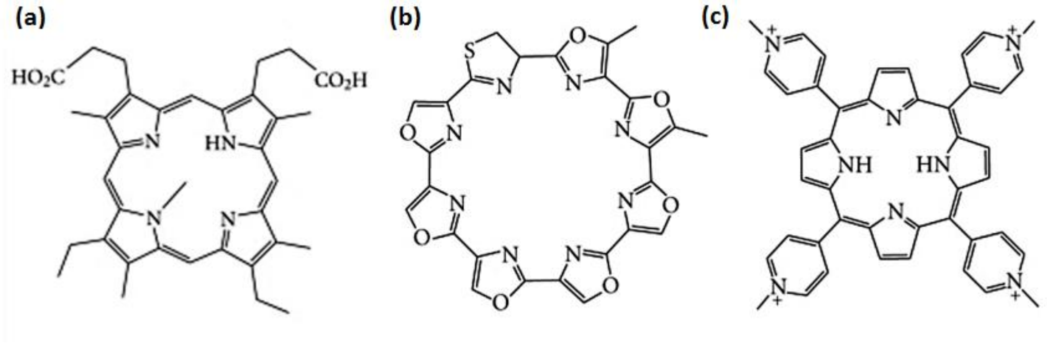


Figure 15– Chemical structures of (a) N-methyl mesoporphyrin (NMM) [180] (b) telomestatin and (c) TMPyP4 [181].

4.2 The significance of docking to G-quadruplexes

G-quadruplexes have been shown to influence carcinogenesis through transcription regulation, and inhibition of telomere elongation by telomerase (Figure 16). The Hurley Lab has uncovered a correlation between G-quadruplex stabilisation and the suppression of promoter activity. Binding of small molecules such as TMPyP4 (Figure 15), a well-known G-quadruplex stabiliser, to G-quadruplex was found to decrease the promoter activity by more than 50%. [182] Potential G-quadruplex-forming motifs $(G_{3+}N_{1-7})_{4+}$ (where G is guanosine and N is any nucleotide) have been found in 30-40% of human promoters. The formation of a G-quadruplex in the promoter region of c-MYB oncogene containing GGA repeats has been reported to be involved in both transcriptional activation and repression. Targeting G-quadruplexes has emerged as a strategy to deactivate the promoters of oncogenes, suppressing transcription by using G-quadruplex-targeting ligands. [183]

Human telomeric DNA typically consists of tandem 5’-TTAGGG-3’ repeats with a G-rich 3’ overhang (capable of folding into G-quadruplex). Telomeres protect chromosome ends from recombination, from degradation and end to end fusions, and from inappropriate repair processes (to distinguish it from double-stranded breaks). Telomeres, however, shorten with every cell cycle, and this leads to aging. The formation of G-quadruplex structures in telomeric DNA has been shown to disrupt telomeric capping and maintenance (to be sensed as DNA damage, inducing apoptosis) and to disrupt telomerase from telomere, thereby inhibiting over-elongation of telomeres in 80-85% of human tumor cells. Such DNA-containing G-quadruplex structures are no longer recognised as substrates by telomerase. Cellular events such as recombination and replication involve separation of the DNA strands, thus providing opportunities for the

G-rich strand to form the G-quadruplex structures. G-quadruplex ligands that induce and/or stabilise G-quadruplexes are promising therapeutic agents for the treatment of cancer, and are able to inhibit telomerase activity and activate a DNA-damage response, leading to apoptosis or replicative senescence. [184]

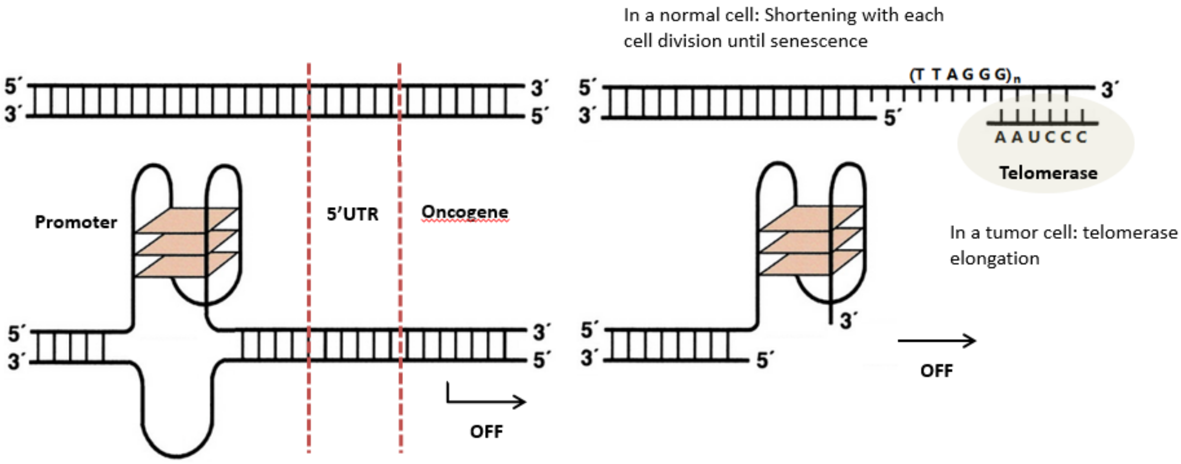

In a normal cell: Shortening with each cell division until senescence
5´
3´
3´
5´
Promoter
5'UTR
Oncogene
OFF
(T T A G G G)n
A A U C C C
Telomerase
In a tumor cell: telomerase elongation
OFF

Figure 16 - Modified from [173]. G-quadruplexes can reduce the expression of oncogenes and inhibit telomerase activity.

Unfortunately, most chemotherapy agents used bind nonspecifically to DNA. Possible anticancer drugs, however, occupy two types of binding sites for G-quadruplex ligands. (Figure 17) Firstly, co-facial end-stacking or hemi-intercalation binding mechanisms involve polyaromatic molecules (called end-stackers) with planar geometries, for binding to the ends of the G-quartets; stabilising the quadruplex via π-π stacking interactions. Many of these compounds are porphyrin derivatives. However, these molecules generally have poor drug-like properties and selectivity. Secondly, a small drug molecule may bind to grooves and/or loop regions. Binding is sensitive to subtle variations in topologies, groove widths, and loop sequences conferring selectivity. Groove binders are more selective than end-stackers, and are specific to different groove conformations, suggesting promising potential for site-specific design. [185] Examples are steroids and their derivatives, which recognise DNA particularly, and show a 15-fold higher selectivity for G-quadruplex compared to dsDNA.

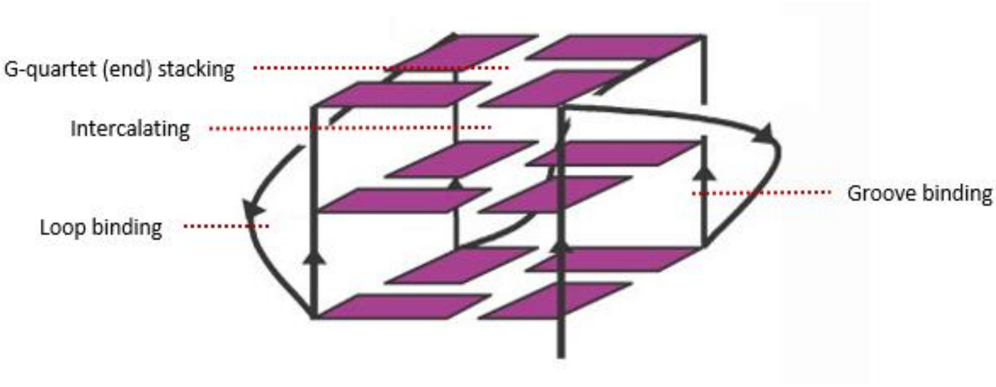


Figure 17 - Binding modes of small molecules with G-quadruplexes.

4.3 Ligand-ratiometric sensor

The ligand that was used is *2-(2′-hydroxyphenyl)-3H-imidazo[4,5-b]pyridine* (HPIP-b) which binds all kinds of G-quadruplex structures.

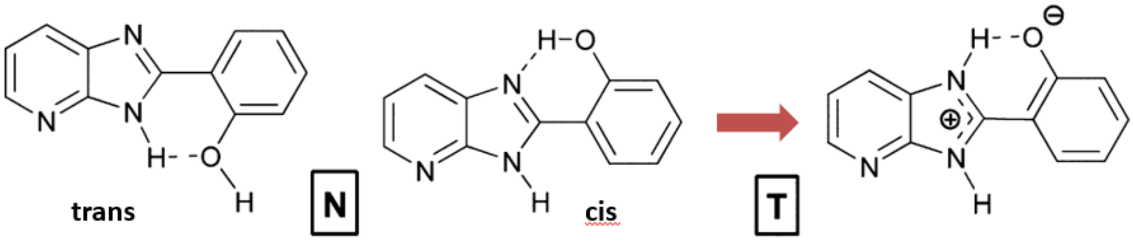


Figure 18 - Molecular structures of the normal (N) form of HPIP-b and the tautomer (T) form obtained after ESIPT.

HPIP-b is a ratiometric fluorescent sensor. In the ground state, HPIP-b exists in equilibrium between two isomeric normal (N) species: the cis-form, which is the most stable form, and the trans-form. Upon excitation (which triggers protonation or deprotonation), an intramolecular proton transfer occurs along the hydrogen bond coordinate to give the tautomer (T). (Figure 18) This process of excited state intramolecular proton transfer (ESIPT) occurs only in the cis-form, as it requires both the acidic and basic groups to be in close proximity in the same molecule. [186]

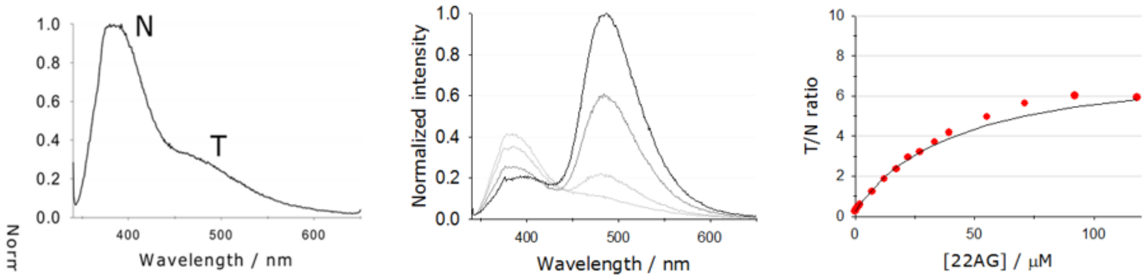


Figure 19 - Obtained from Penedo-Esteiro's group, fluorescence spectra of HPIP-b in aqueous buffer with emission of tautomer (T) and normal (N), Tautomer/normal ratio of HPIP-b at different concentration of telomeric sequence (middle and right graph) The darker the line, the higher the concentration of telomeric DNA.

HPIP-b exhibits a dual fluorescence band emission (Figure 19, Left). The normal (N) species is responsible for the emission band at 380nm, and the emission at 480nm is due to the tautomer (T). From experiment, we known that 80% of HPIP-b is in the trans-form in water, and perhaps the majority of the ligand is still in the trans-form upon binding. So, we used the trans-form to carry out this study. However, as the concentration of DNA increases, the equilibrium shifts towards a higher concentration of the bound T form conformer, so that the ligand somehow undergoes a change from the N/trans- to the N/cis-form. Thus, it becomes structurally possible to undergo an ESIPT process upon excitation, giving the tautomer, as shown by a binding isotherm graph (Figure 19, Right). The middle graph shows their relative intensity changes with the concentration of quadruplex. The tautomer becomes dominant at a higher concentration of quadruplex. Their ratio is affected by solvent polarity, and by the pH of the medium. These two isomers display different excited-state properties, making HPIP-b an excited-state proton transfer probe with high sensitivity to the environment. Due to this dual-band ratiomatic property, HPIP-b is an attractive sensor for studying binding interactions of G-quadruplexes. HPIP-b also interacts with double strand DNA, though this is thought to be a much weaker interaction.

4.4 Method

We performed a docking study to investigate the binding pattern of HPIP-b upon G-quadruplexes [AGGG(TTAGGG)$_3$]. We used molecular dynamics to firstly simulate large-scale conformational changes of G-quadruplex in the presence of explicit solvent and ions, and secondly to explore more possible binding sites, from which snapshots of the G-quadruplex structures were extracted for subsequent docking. An ensemble docking strategy was adopted to implicitly account for receptor flexibility. At the end of the simulation, we extracted five snapshots at 200ps intervals. Subsequently, each rigid receptor structure will undergo an independent docking run. AutoDock [187] was used to perform a blind docking against each of the snapshots, to obtain the most probable binding modes presented, indicated by better estimated binding energies. Due to the limitations posed by docking, there arises the need to adopt a theoretical method to account for a certain degree of protein flexibility. In this regard, we carried out a post-docking refinement step using a two-layer QM:QM ONIOM model to optimise and rescore docked complexes. The ONIOM method (our own N-layer integrated molecular orbital molecular mechanics) developed by Morokuma et al. [188], employs a subtractive (or extrapolative) QM/MM scheme, in which the total energy of the system, $E_{QM/MM}(system)$, is calculated using Equation 11. Here, $E_{MM}(system)$, is the MM energy of the whole system, $E_{QM}(QM)$, is the QM energy of the QM region and $E_{MM}(QM)$, is the MM energy of the QM region. [189]

Equation 11 – The energy of a two-layer ONIOM(QM:MM) calculation:

$$E_{QM/MM}(system) = E_{MM}(system) + E_{QM}(QM) - E_{MM}(QM)$$

In the end, we reported our docking studies from AutoDock, which revealed electrostatic effects between HPIP-b and G-quadruplexes.

4.4.1 Structure preparation

DNA

The crystal structure of the parallel 22-mer human telomeric G-quadruplex (PDB entry 1KF1), sequence d [AGGG(TTAGGG)$_3$], was obtained from the PDB. The structure consists of three G-quartets and three external TTA loops extended outward from the guanine core resulting in a propeller-like shape. (Figure 20) Adjacent G-quartets are stacked with a 30° twist and are separated by 3.13Å. [190] The loops connect the adjacent parallel chains, from the top of one strand to the bottom of the other. The second thymine of each TTA loop is located at the tip of the loop, within the adenine base swung back, intercalated between two thymine bases. These loops are thought to be involved in intermolecular interactions, for example hydrogen bonding, and stacking interactions with telomeric proteins.

This initial structure was arranged into a dimer using Chimera, making use of crystallographic symmetry information from the PDB file. This dimer was used in this study to model the structure of extended quadruplex telomeric sequences. Packing of two G-quadruplexes through G-quartets results in a stacked 5' to 5' hydrophobic surface. In contrast, a 3' surface is more hydrophilic. The dimer structure consists of a core of six G-quartets and five K^+ ions in the central channel.

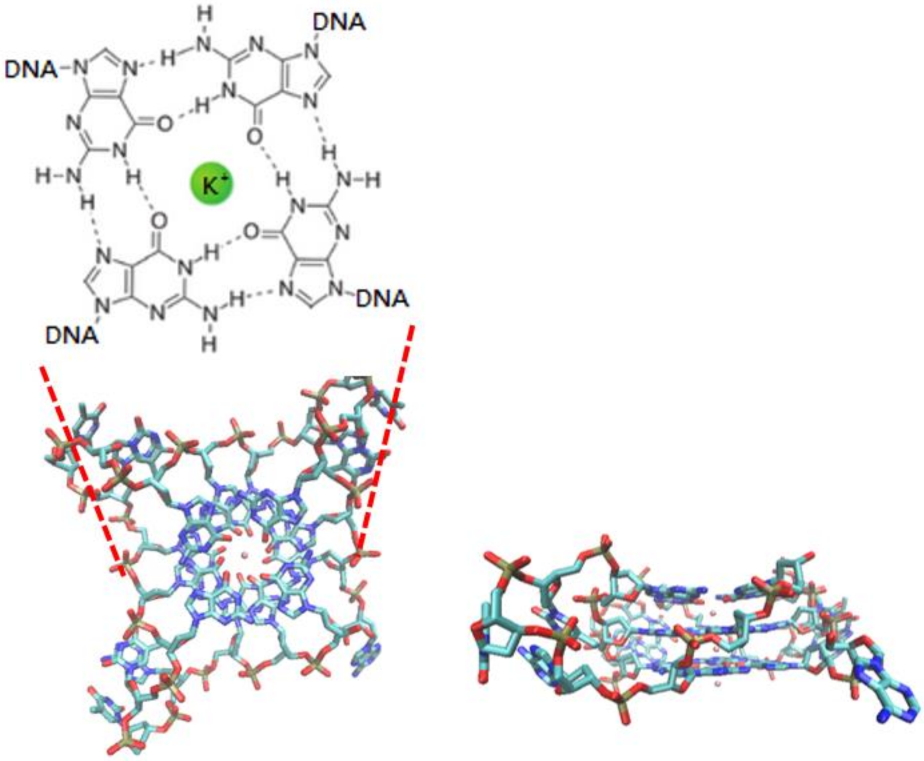


Figure 20 - Chemical structure of the G-quartets. (PBD ID: 1KF1)

Internal K^+ ions were retained between consecutive G-quartets throughout these studies. The double-stranded DNA of similar sequence 5’-GTTAGGGTTAGGG-3’ (PDB entry 1IV6) was used for comparison. (Table 4)

G-quadruplex	dsDNA
1KF1 (Human telomeric DNA)	1IV6
5’-AGGG(TTAGGG)$_3$-3’	5’-GTTAGGGTTAGGG-3’ 3’-CAATCCCAATCCC-5’
Single stranded	Double stranded with base pairing

Table 4 - The structures and sequences of the DNA used in this study.

4.4.2 Molecular Dynamics

MD simulations were carried out with GROMACS version 4.6.7 [191,192,193] using the Amber03 force field. The dimeric G-quadruplex was solvated in a 6.55Å^3 box, with 8687 TIP3P water molecules and neutralised by replacing random water molecules with 18 Mg^{2+} ions and 1 K^+ ion to bring the net charge to zero. Periodic boundary conditions were applied to minimise edge effects.

During pre-optimisation, the solvent was relaxed at 0K by a steepest descent optimisation with restrained nucleic acids, followed by an overall system optimisation. The steepest descent method, which uses first derivatives in energy minimisation, is applied in all simulations to avoid failure when the force is large, and to remove thermal noise. The method takes a step downhill by moving in the direction of the greatest negative gradient.

The hydrogen bonds are constrained with the P-LINC algorithm. The system was gradually heated from 0 to 300K with 100ps equilibration for each successive temperature step and equilibrated at 300K for 100ps, using a time step of 0.001ps. This equilibration was followed by 1000ps simulations under NPT (fixed pressure, temperature and number of atoms) conditions at 300K, during which coordinates were saved to the trajectory every 10ps, resulting in a set of 101 coordinates.

The same procedure was repeated with dsDNA in a cubic box of 6.41Å edges with 8452 TIP3P water molecules. Five structures saved after every 200ps along a 1000ps MD simulation were obtained for use in the following docking studies.

4.4.3 Molecular Docking

Dockings of HPIP-b to parallel telomeric G-quadruplexes (PDB ID: 1KF1) were carried out using AutoDock 4.2 [187], employing a Lamarckian genetic search algorithm (LGA) to generate docked poses, and a semi-empirical force-field-based scoring function to estimate the free energy of binding. A HPIP-b molecule was built and optimised with the semi-empirical PM3 method using Gaussian 09, and was in the trans-configuration during docking.

Five G-quadruplex structures were extracted from the MD trajectory for docking using AutoDock Tools (ADT) version 1.5.6 [187], a graphical user interface, used to a) merge non-polar hydrogens, by adding Gasteiger charges [194] to each constituent atom of the ligand and the receptor, and b) assign rotatable bonds prior to the docking. AutoGrid was used to generate grid maps for each atom type in the docked ligand, which store grids of interaction energy used as a lookup table, to speed up the interaction energy calculation during the conformational search (sampling stage). Default values were used for AutoGrid parameters. A grid map with 126 x 126 x 126 points, and a grid spacing of 0.375Å was used, and the maps were centred on the DNA, covering the entire DNA. The DNA was kept rigid, while the ligand was allowed to be flexible during sampling.

50 independent docking runs were performed for each G-quadruplex structure, with an initial population size of 50 individuals, a maximum number of 5 x 10^7 energy evaluations per run, and a maximum number of 27000 generations. Mutation and crossover were applied at rates of 0.02 and 0.8, respectively. All other parameters were set to default values. The local search was based on the Solis and Wets method, with a maximum of 300 iterations per search, local search rate of 0.06, step sizes of 0.2Å for translations and 5° for orientations and torsions. The resulting docked structures were clustered, using the Clusterings module in ADT, by the conformation with the lowest free binding energy (with a 2Å cut-off RMSD), to obtain clusters of similar binding modes. The low-energy conformations were chosen from the largest and lowest-energy cluster for further processing. This same process was applied to the dsDNA for comparative purposes in this work.

This docking process is semi-flexible, where the ligands are allowed to explore their conformational space, while keeping the DNA rigid. The flexibility of a ligand molecule is modelled with six external degrees of freedom (three translations along the coordinate axes and three rotations) plus internal (conformational) degrees of freedom due to rotations around bonds. Since it is impractical to fully explore the conformational space in practice, an approximation (i.e. the rigid body approximation, treating each DNA as a rigid body) is assumed to reduce the dimensionality of the space. [195] In the next experiment, ONIOM was employed to optimise and rescore the docked complexes, for considering a certain degree of protein flexibility explicitly after the docking process (post-docking refinement).

4.4.4 Thermodynamic cycle-ONIOM QM/QM

To quantify the electrostatic effects arising from the binding region of the docked ligand, a two-layer ONIOM QM/QM calculation using hybrid density functional theory (DFT) and PM6 was performed with Gaussian 09. Using the ONIOM scheme the binding region was divided into subsystems that are treated at different levels of theory. This is in the interest of making a challenging and time consuming calculation more readily tackled on currently available hardware while still providing a reasonable degree of accuracy and insight. The system was built upon the docked position of HPIP-b containing residues that are within 4Å of the ligand. This included water molecules in order to account for electrostatic solvation effects and hydrogen bonding (which are crucial in describing ligand effects). As shown in Figure 21, hydrogen atoms were used for capping the lower layer. The ligand was fully optimised using DFT at the M06/6-31G(d,p) level, whereas in its surroundings atoms heavier than hydrogen are fixed and treated by PM6.

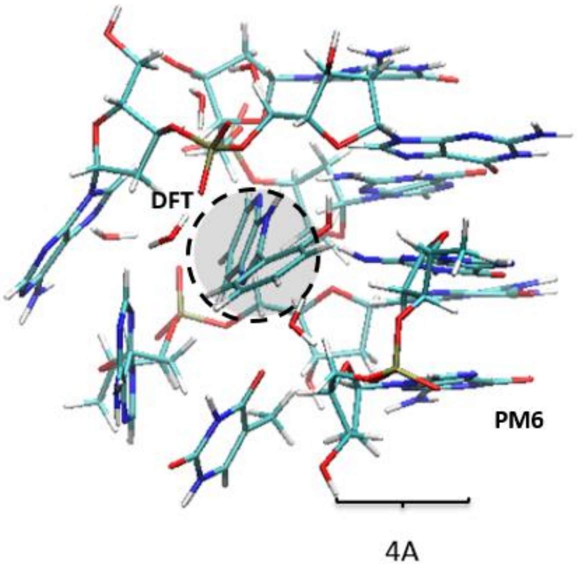


Figure 21 - ONIOM was applied upon the docked position of the ligand including 4Å around the ligand. The ligand (circled in dash) is treated at DFT (M06) level, while the remainder is treated at a lower level of theory (PM6).

We constructed a thermodynamic cycle (Figure 22), used to determine the electrostatic forces in the ligand docked region. In this cycle, the capital in the bracket indicates where the ligand has been optimised (W refers to water) while the superscript indicates where the calculation was performed; either in the gas phase, in a PCM solvent model, or in the point charge field of DNA. The ligand was optimised with DFT at M06/6-31G(d,p) level via a self-consistent reaction field (SCRF) method using the polarisable continuum model (PCM), developed by Tomasi et al. [196]. PCM is used to simulate implicit solvent effects, [Ligand, W], wherein the solute is embedded in a spherical cavity surrounded by a homogeneous dielectric continuum, which is dependent on its dielectric constant. The optimised ligand was then calculated in the gas phase at the same geometry using DFT at M06/6-31G(d,p) level, from which the energy differences are calculated to obtain the solvation energy of the ligand (left arrow of Figure 22 (a)).

The ligand was optimised in DNA (by ONIOM). The optimised ligand was calculated in the gas phase and subsequently in the point charge field of neighbouring structures updated with optimised (PM6) hydrogen position. In this portion of the model, every atom, including the whole DNA and solvent within 15Å of the optimised ligand, was considered as a point charge using Amber03 and TIP3P parameters from Gromacs. We calculated the energy difference between the ligand in the gas phase and embedded in the point charge field, to obtain a measure of electrostatic forces at the ligand docked region. (right arrow in Figure 22 (b))

The energy differences (EGeom) between optimised structure of the ligand from the water and DNA calculated in the gas phase allowed us to calculate the energy cost of changing the geometry upon binding, represented by the top arrow in the thermodynamic cycle shown in Figure 22 (c).

The energy differences of the ligand between its different optimised structures in water and its binding conformation in the DNA represented by a point charge field (PCF) is calculated to obtain the binding energy (bottom arrow, Figure 22 (d)).

$$\begin{tikzcd}
{[\text{Ligand, W}]^{\text{GAS}}} \arrow[r, "{(c)\ E^{\text{Geom}}}"] \arrow[d, "E^{\text{Solv}}", "{(a)\ \text{SCRF}}"'] & {[\text{Ligand, DNA}]^{\text{GAS}}} \arrow[d, "E^{\text{embed}}=E^{\text{electrostatic}}", "{(b)\ \text{PCF}}"'] \\
{[\text{Ligand, W}]^{\text{WATER}}} \arrow[r, "{(d)\ E^{\text{Bind}}}"] & {[\text{Ligand, DNA}]^{\text{DNA}}}
\end{tikzcd}$$

Figure 22 - Thermodynamic cycle employed for electrostatic calculations.

4.4.5 Potential energy surface

We explore a potential energy surface (PES) that gives the ground state electronic energy contained in a molecule at a given nuclear geometry. Since there is a relatively large difference in mass between electrons and nuclei, the nuclear equations of motion can be described either by quantum mechanics, or else classically via Newton's equations. The nuclei appear to stay motionless (be fixed) in their equilibrium positions in space compared to the fast-moving electrons that adjust themselves instantly to changes in molecular conformation, thereby following the Born-Oppenheimer approximation (formulated by Max Born and Julius Robert Oppenheimer in 1927), according to which electronic and nuclear motions are treated separately. The dynamics of nuclear motions on the ground state electronic surface are obtained from solving the time independent Schrödinger equation for the electrons, and the changes in the electronic energy due to nuclear displacement are independent of the kinetic energy of the nuclei, and are independent too of the nucleus mass. Hence, isotopic species result in the same potential surface, unless vibrations of the nuclei have been taken into consideration. [197]

A relaxed PES scan with respect to two dihedral angles, for rotation around the hydroxyl and the C3-C4 bond (see Figure 25 for structure), was performed by DFT at the M06/6-31G(d,p) level, each to be scanned from 0 to 360 degrees by varying the torsion angle H(11)-O(10)-C(5)-C(4) and N(8)-C(3)-C(4)-C(12) in steps of 15 degree increments of 24 steps in total to get the potential energy surface and the lowest energy configuration of HPIP-b.

4.5 Result and Discussion

4.5.1 RMSD plot

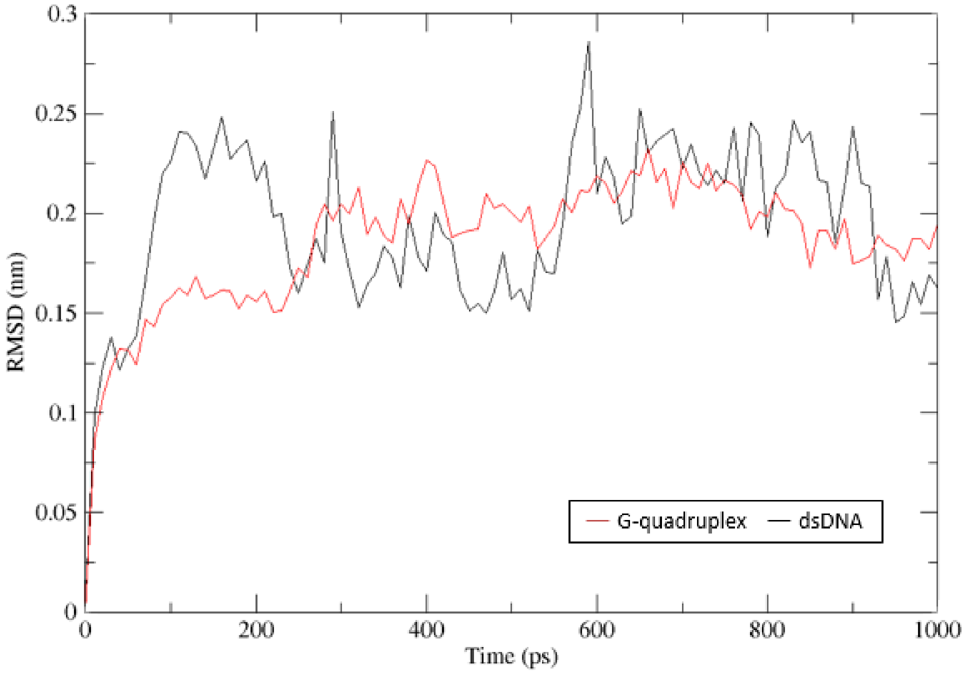


Figure 23 – Backbone RMSD from the trajectories of the simulations of dsDNA and G-quadruplex.

The RMSD (the root mean square difference) of the DNA backbone with respect to the staring structure (the first frame of the production run) were calculated over the 1ns trajectory, as can be seen a rapid increase occurs during the first 100ps and then levels off, and remains stable for the rest of the simulation time. The differences in fluctuation amplitudes show that G-quadruplex (indicated by the red line) is more rigid than dsDNA (black). (Figure 23)

4.5.2 Ligand Scan

A contour plot of the PES is shown in Figure 24. The resulting energy surface has two minima and the lowest corresponds to the cis-form (Figure 25, C), which forms an intramolecular hydrogen bonded cyclic ring with a planar geometry, and is as expected, the most stable form [198]. The trans-form (Figure 25, A), which can be stabilised by an intramolecular hydrogen bond between the OH-group and the N-H hydrogen on imidazole ring, corresponds to a local minimum (Figure 24A). It is found that trans-form is more stable than cis-form in solution due to hydrogen bonding with the solvents. [199] The energy difference between the trans- and cis-form (A and C) is about 7.51 kJ/mol, and between conformers A and B is about 22.81 kJ/mol.

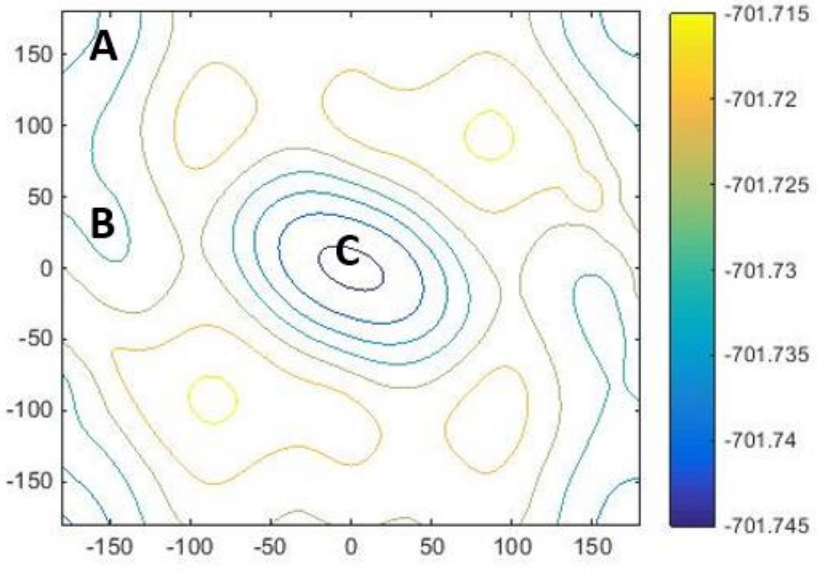


Figure 24 - Contour plot of the potential energy surface for HPIP-b resulting from the DFT scan, with energy in atomic units (a.u.). The dihedral angle H(11)-O(10)-C(5)-C(4) (y-axis) and the N(8)-C(3)-C(4)-C(12) (x-axis) in degree.

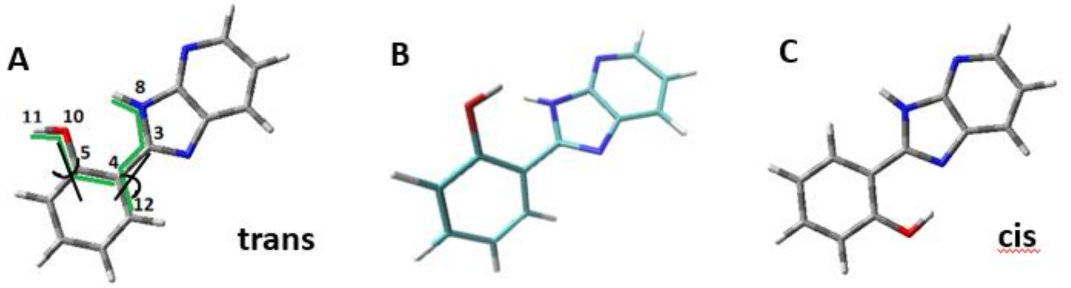


Figure 25 - The cis- and trans- conformers of HPIP-b, corresponding to the minima on the PES (Figure 24).

4.5.3 Predicted binding modes (Docking)

To obtain probable binding sites and orientations of HPIP-b on G-quadruplex structures, 50 independent docking runs to each of the five snapshot G-quadruplex structures were performed (saved every 200ps from the MD trajectory), and the resulting docked poses were clustered (independently for each snapshot) using a cut-off distance of 2Å.

As expected, no end-stacking was observed. All 5 x 50 runs with HPIP-b resulted in binding to either the loop or groove regions of the G-quadruplexes. Among all docking runs, the best docked conformation, with the lowest binding free energy of -7.5 kcal/mol, is formed through the loop, binding to the equilibrated structure obtained at 200ps from the MD trajectory (Table 5a). This is followed by the second-best conformation (-7.13 kcal/mol), which also binds to the loop region to the same snapshot structure. HPIP-b displays a higher affinity overall to this structure (at 200ps). However, these configurations were not in the most populated cluster. Instead, the fifth (second to last) cluster consisting of 20 members out of 50 was the most populated, with an average binding free energy of -6.58 kcal/mol by binding to the groove, see Figure 26(a).

As can be seen from the lowest-energy cluster of each snapshot (the charts of Figure 26), the loops form the preferred site of binding of HPIP-b to the 200ps and 1000ps snapshot structures, with the latter having an average binding free energy of -6.78 kcal/mol. However, this finding is not conclusive, as the lowest-energy cluster and the most populated cluster obtained from other structures (extracted at different time points) suggested that HPIP-b binds at the groove. (Figure 26(b), 26(c), 26(d)) In fact, only groove bindings have been reported for a snapshot at 800ps with the lowest binding free energy of -6.50 kcal/mol; and no cluster is significantly populated for a snapshot at 400ps, in which the differences in binding free energy and population between clusters are small; ranging from 2 (the least energetically favourable conformation -5.76 kcal/mol) for the smallest cluster to 8 (the first and the third cluster that the HPIP-b binds to the groove and the loop, respectively) for the largest cluster.

	Clus Rank	Run	Num in Clus	ΔGbest	ΔGaverage	Binding mode
a	1	42	5	-7.50	-7.50	Loop
b	1	43	8	-7.02	-7.00	groove
c	1	33	26	-6.32	-6.32	groove
d	1	4	24	-6.50	-6.48	groove
e	1	25	16	-6.78	-6.78	Loop

Table 5 - Shows the binding free energy of the best docked conformation (ΔGbest), corresponding to the docked complex in Figure 26, and the average energy of the lowest-energy cluster (ΔGaverage) for each G-quadruplex structure. Units are in kcal/mol.

Selection of the best conformer is typically primarily based on the lowest-energy cluster, and secondly the most populated cluster. Cluster analysis reveals that, in general, the binding energy differences between clusters are small. In cases where the lowest-energy cluster is less populated, the binding free energies of the most populated cluster are all within 2.5 kcal/mol energy difference, the estimated standard error of the AutoDock scoring function. Thus, it is hard to say which of the configurations is the more probable solution. The overall finding seems to suggest that groove and loop binding showed no significant differences in terms of binding free energy. To be specific, we defined 1) groove binding in which HPIP-b was either in the cavities bound by the neighboring phosphate backbones or within the extra cavities adjacent to the loops at the sides of the G-quadruplex units and 2) loop binding occurred when HPIP-b was buried in between the external loop and the G-quartets. HPIP-b is nearly rigid and conformationally restricted with only a single rotatable bond, via the hydroxyl group, during docking. Such docking is similar to a lock and key mechanism, as AutoDock does not account for receptor flexibility during docking. Thus, docking of HPIP-b is likely to be dependent on the target selected conformation. The resulting docked conformation of HPIP-b from the best cluster (both the most populated and the lowest-energy cluster) have unfavourable (positive) internal energy (not shown). However, the binding free energy obtained from the default value of Autodock 4.2 assume that the internal energy of the ligand in solution is the same as in the complex, thus results in a zero contribution of internal energy to the total binding energy. The results given were all calculated based on the above assumption. In general, binding a frozen ligand is more likely to overestimate the gain of free energy of binding, mostly due to the energy required to distort the ligand from the preferred unbound to the bound conformation. The consideration of the internal energy is necessary to obtain a result that is close to reality.

A comparative study has been applied on dsDNA of a similar sequence. All runs of dsDNA resulted in minor groove recognition, with a very favourable average binding energy of -6.61 kcal/mol and the most favourable docked conformation (-6.62 kcal/mol). (Table 6c) This is in comparison to docking G-quadruplexes, where a relatively larger number of clusters were found, possibly reflecting the complexity and the large surface tested of the G-quadruplex structures. The results are in agreement with experimental findings that HPIP-b displays slightly stronger interactions with the G-quadruplex than with dsDNA.

	Clus Rank	Run	Num in Clus	ΔGbest	ΔGaverage	Binding mode
a	1	32	34	-6.31	-6.30	groove
b	1	36	41	-6.02	-6.01	groove
c	1	39	46	-6.62	-6.61	groove
d	1	49	22	-6.31	-6.31	groove
e	1	35	29	-6.29	-6.14	groove

Table 6 - Shows the binding free energy of the best docked conformation (ΔGbest), corresponding to the docked complex in Figure 27, and the average energy of the lowest-energy cluster (ΔGaverage) for each dsDNA structure. Units are in kcal/mol.

In summary, we performed blind docking to obtain putative binding sites of HPIP-b on G-quadruplex and dsDNA with no prior assumption of binding sites. Autodock does not account for the receptor flexibility. We obtained a rather small difference between energetically similar cluster for HPIP-b with the binding energies ranging from -4.5 kcal/mol to -7.5 kcal/mol with results in either loop or groove binding. However, these docking studies could be the start of using a more sophisticated force field or methods for post-processing of docking results. To further ensure the stability of the binding structure of ligand, we carried out a higher level of calculation for a comparison between MM energies and QM energies. We selected the lowest energy conformation from each cluster to represent the whole cluster for further analysis.

Docking can produce both false positives and false negatives. [200] It is known that most docking programs use rigid receptor models, and that these errors could occur because DNA is a rather flexible molecule existing as an ensemble of isoforms, not modelled well with rigidity. To take receptor flexibility into account, we performed ensemble docking, in which the ligand is docked to a pre-generated ensemble of rigid structures, to implicitly account for receptor flexibility. However, this leads to two observations. Firstly, the docking performance is biased by the selected structural ensembles. A selection strategy to choose representative conformations from MD trajectories is still needed; and secondly, MD simulations normally explore only local minima and have difficulty overcoming high energy barriers. In some cases, using an ensemble of conformations did not outperform using a single structure. [201]

Figure 26 - Cluster profiles from docking of HPIP-b to G-quadruplex structures extracted from MD every 200ps, (a) at 200ps, (b) at 400ps, and so on. Each cluster is a bar, where the height of the bar indicates the number of conformations in the cluster (50 in total) and the colour indicates the binding mode: loop in blue and groove in green. Only the best docked pose with the lowest binding free energy score are shown, with unit in kcal/mol.

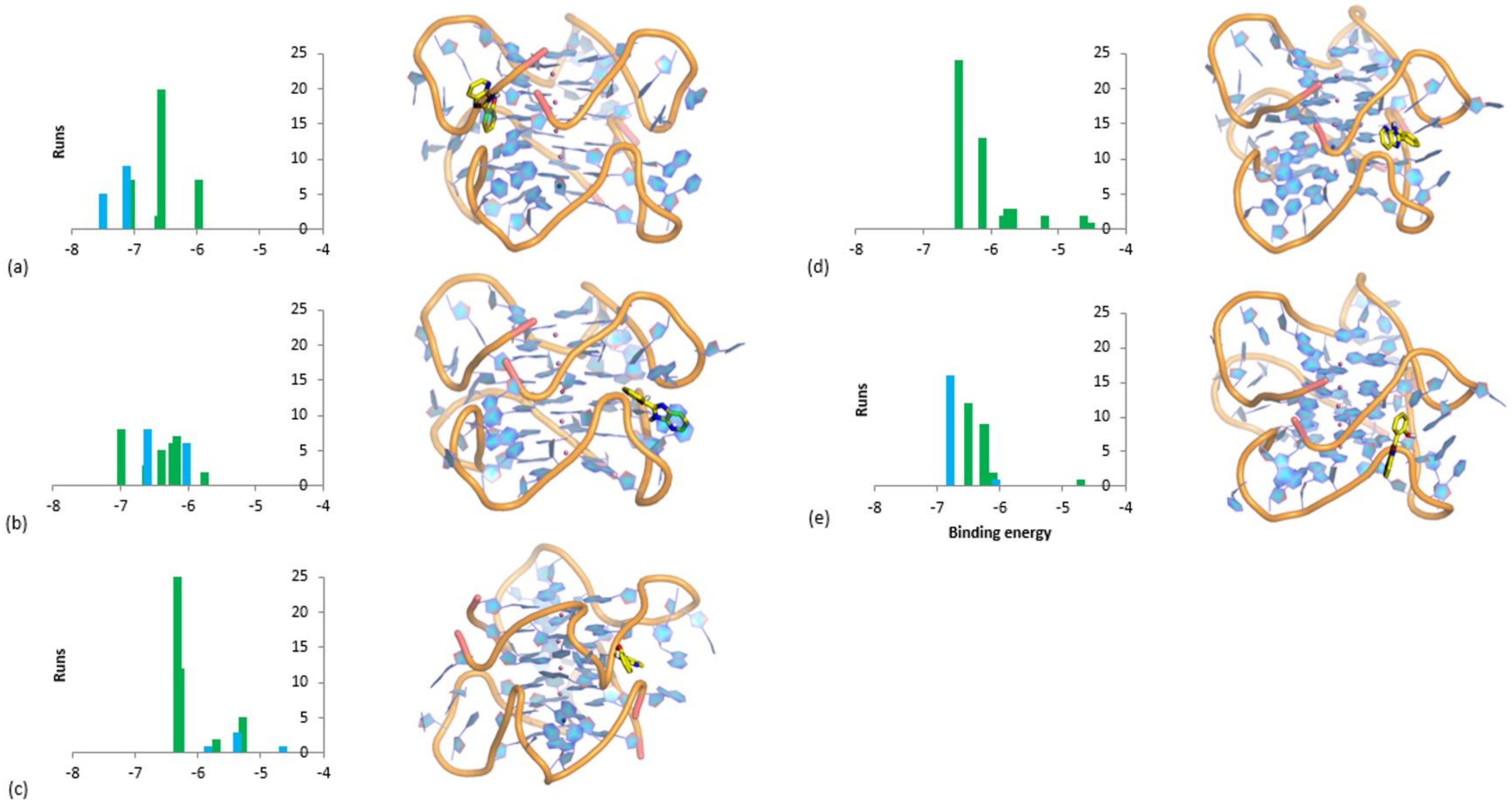
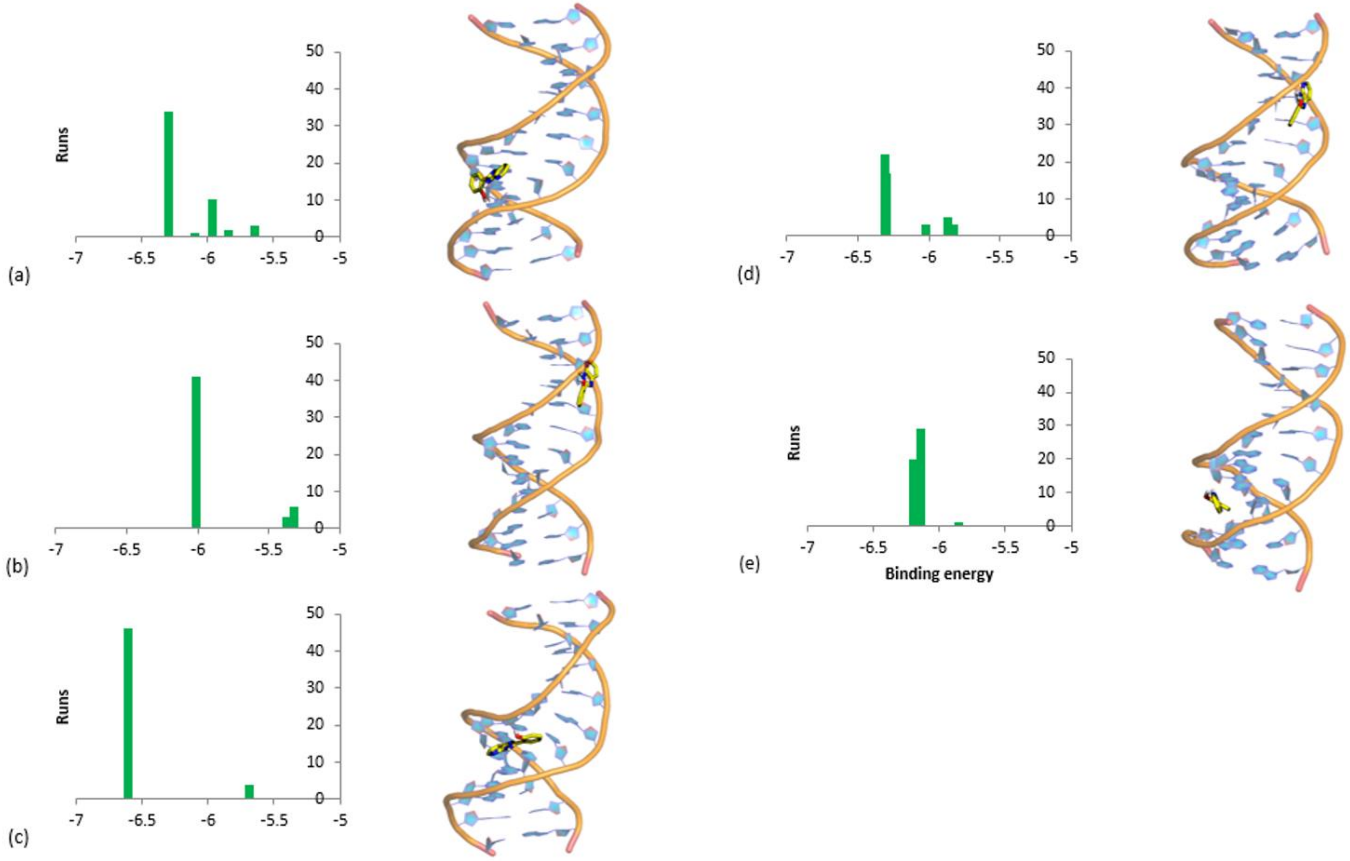


Figure 27 - Cluster profiles from docking of HPIP-b to dsDNA structures extracted from MD every 200ps, (a) at 200ps, (b) at 400ps, and so on. Each cluster is a bar, where the height of the bar indicates the number of conformations in the cluster (50 in total) and the colour indicates the binding mode: loop in blue and groove in green. Only the best docked pose with the lowest binding free energy score are shown, with unit in kcal/mol.



4.5.4 QM Binding energy

ONIOM was used to refine the docking poses, allowing limited ligand/receptor induced fit effects to be modelled, with optimised positions of the hydrogen atoms of the receptor, while holding all other atoms fixed. We selected the lowest-energy docked structure from each cluster as a sample conformation for each pocket.

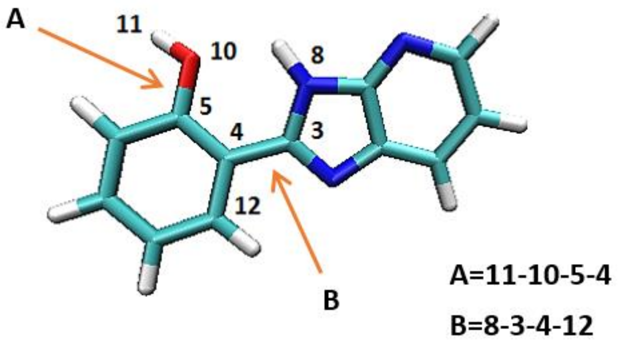


Figure 28 - Structure of HPIP-b with atomic labelling and definition of the dihedral angles A and B considered in the PES scan.

As expected, the ligand adopts an energetically more favourable conformation in solvent than in DNA (positive values for EGeom). This may arise as the ligand does not adopt its lowest energy conformation in the binding site with the energetic cost of binding from solvent offset by interactions with the binding site. Normally, this higher energy conformation can be compensated with the gain of the aforementioned interaction energy with the receptor such as forming of hydrogen bonds or van der Waals interactions. The ligand in PCM solvent model adopts a virtually planar conformation (the torsion angle of H11-O10-C5-C4 and N8-C3-C4-C12 are -179.9° and 179.9°, respectively; see Figure 28) corresponding to the local minimum (A) on the potential energy surface (PES) in Figure 24. Comparing the energy value of EGeom to the dsDNA, the bound ligand in G-quadruplex may be more restricted in terms of the conformation it can adopt. As a result, the geometric energy cost that is paid for binding with the G-quadruplex is greater than that associated with binding to dsDNA. We expect that this is due to the more complex structure of G-quadruplex. The solvation energy is -33.926 kJ/mol. The experimental measured desolvation of polar group OH is 36.4 kJ/mol, [202] and the desolvation of a polar group is enthalpically unfavourable. The nitrogen offers an additional hydrogen bonding moiety, so might impact the total solvation energy accordingly.

We inspected the docked poses and positions which give the lowest binding energy to study how these changes contribute to the electrostatic and binding energy. The analysis of the hydrogen-bonding patterns was carried out using VMD (version1.9.3) [203], with a distance cut-off of 4Å and an angle cut-off of 40°, and for producing figures (Figure 29 and 30). The obtained energy considered only the effect of electrostatics (previous work has found that electrostatic interactions are the primary interactions from DNA [204]). A positive value for Ebind corresponds to an unfavourable interaction. We obtained five most favourable binding energies which are lower than -100 kJ/mol (four for G-quadruplex, one for the dsDNA). HPIP-b interacts with G-quadruplex through both the grooves (in cases of 60_39 and 20_39) and the loops (for 60_04 and 20_11), see Table 7.

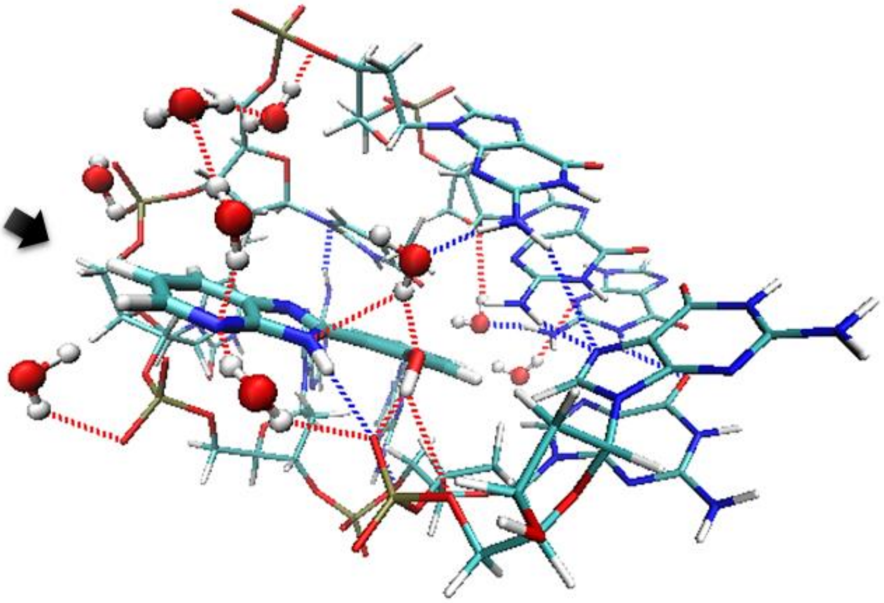


Figure 29 - The lowest energy configuration of docking to G-quadruplex. (60_04 in Table 7)

The energy cost from the conformational change of the ligand on binding to the receptor is 50.347 kJ/mol (60_04 in Table 7). The dihedrals are measured: 11-10-5-4=84.6°, 8-3-4-12=164.7° (see Figure 28 for labelled structure). The ligand does not hydrogen bond to its docked cleft, formed by a TTA propeller-type loop, but is hydrogen bonded by its OH group and N-H moiety of the heteroaromatic ring, with two vicinity oxygens of the phosphate group on the stand connecting the G-quartets. It also shows hydrogen bonding with three nearby water molecules. Here, the nitrogen lone pair (an area of negative charge) is creating the hydrogen bond to the water hydrogen (positively charged). The water is also used to hydrogen bond with other sections of the G-quadraplex. The electrostatic energy obtained is -210.029 kJ/mol, which is the largest amongst the 27 binding instances studied. However, from manual inspection, a large amount of the electrostatic interactions that result in a total favourable binding energy is more likely due to a hydrogen bond network formed by water molecules around the ligand.

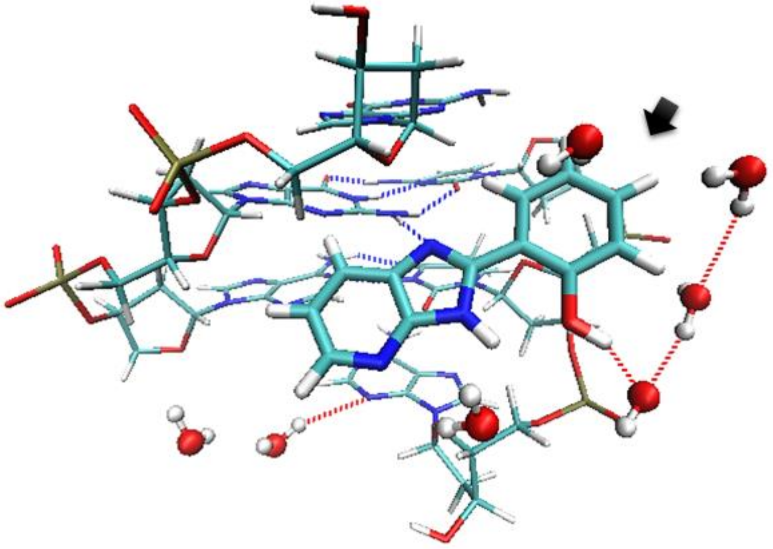


Figure 30 - The lowest energy configuration of docking to dsDNA. (ds80_27 in Table 7)

We note that a binding instance to dsDNA (ds80_27 in Table 7) achieved the lowest binding energy value (-132.027 kJ/mol) of all complexes in this test set, which is unexpected as we know from the experimental collaboration and results furnished from these studies that HPIP-b binds more strongly to the G-quadruplex. However, from the energy calculations, we see that the docked conformation is relatively energetically favourable with only a small additional conformational energy cost of 7.549 kJ/mol. This is because the ligand adopts an approximately planar conformation 11-10-5-4=161.3°, 8-3-4-12=-170.8° when complexed with the DNA. The energy cost might be due to the hydroxyl group pointing to the solvent, which was found able to form a hydrogen bond with nearby water (1.641Å). On the opposite site of the ligand, the nitrogen (N7) in imidazole could hydrogen bond to the guanine backbone. The electrostatic energy from the point charge field of the surrounding it is the largest at -173.502 kJ/mol.

Mutual polarisation of the ligand and the DNA structures are included in ONIOM, while the docking program used a force field method; such a MM approach is limited to set parameters which may not completely or accurately describe the system in question. Binding energies obtained by the docking do not show significant difference. As the protein motion is ignored, the difference in the intramolecular energy of the bound and unbound state of the receptor is zero. The intramolecular energy difference is also zero, as it is assumed the bound and unbound configuration of HPIP-b is the same, nevertheless HPIP-b is rather rigid.

DNA flexibility plays an important role during ligand binding. Our best four results from docking to G-quadruplex structures seem to focus on two snapshots of the conformation (60ps and 20ps). Water molecules contribute electrostatic effects to the binding energy. As observed in Figure 29, a hydrogen bond network is formed which leads to strong electrostatic effects, it appears that hydrogen bonding has occurred between the ligand and G-quadruplex backbone, the ligand and the water molecules, and water to G-quadruplex. A possible improvement to the methodology is the expulsion of water from the binding site and calculation of the electrostatic effects purely due to the DNA.

The loop confers conformational polymorphism, and is considered to make intermolecular interactions. The grooves in G-quadruplexes are V-shaped and do not simply comprise phosphate-sugar backbones, distinct from double-stranded DNA. A dimeric G-quadruplex structure has eight phosphate grooves, together with the extra cavities adjacent to the loops, providing extra binding surface for binding. [190] The complexity of the loop conformation may confer greater specificity.

The special structural arrangement of G-quadruplexes makes it possible to provide distinct surfaces for interaction with small molecules in which quartets provide a hydrophobic aromatic planar surface, and loops provide a hydrogen bonding surface. Also, the grooves are involved in hydrogen bonding to water molecules with the sugar (O3'), the phosphate O1P atom or guanine (N2 atom) in the quadruplex grooves.

Table 7 - The docked poses are ranked according to the binding energy (Ebind). The top panel is computed from the G-quadruplex structures, whereas the bottom is obtained from dsDNAs. The name of the docked ligand is expressed as 60_04 where this denotes that the structure was taken at 60ps of the MD trajectory, and it is the 4th out of 50 docking runs. Units are in kJ/mol.

kJ/mol	Docking		E(L,DNA)DNA	E(L,DNA)GAS	E(L,W)Water		E(L,W)GAS		PCF	EGeom	SCRP	Ebind	Boltzmann
60_04	-22.468	Loop	-701.797	-701.717	-701.749	-701.749	-701.736	-701.736	-210.029	50.347	-33.926	-125.756	100%
60_39	-23.932	Groove	-701.791	-701.725	-701.742	-701.749	-701.731	-701.736	-170.725	28.348	-33.926	-108.451	0%
20_11	-29.832	Loop	-701.790	-701.720	-701.749	-701.749	-701.736	-701.736	-184.595	43.513	-33.926	-107.156	0%
20_39	-29.539	Groove	-701.789	-701.714	-701.749	-701.749	-701.736	-701.736	-197.479	58.439	-33.926	-105.113	0%
80_28	-24.309	Groove	-701.787	-701.730	-701.749	-701.749	-701.736	-701.736	-148.946	15.615	-33.926	-99.405	0%
20_32	-24.978	Groove	-701.784	-701.725	-701.749	-701.749	-701.736	-701.736	-153.251	29.131	-33.926	-90.193	0%
40_04	-26.024	Groove	-701.775	-701.728	-701.749	-701.749	-701.736	-701.736	-123.873	21.427	-33.926	-68.520	0%
80_49	-25.606	Loop	-701.774	-701.723	-701.749	-701.749	-701.736	-701.736	-133.095	35.137	-33.926	-64.032	0%
20_42	-31.380	Loop	-701.772	-701.725	-701.749	-701.749	-701.736	-701.736	-122.536	28.295	-33.926	-60.314	0%
100_25	-28.368	Loop	-701.771	-701.730	-701.749	-701.749	-701.736	-701.736	-106.355	15.287	-33.926	-57.142	0%
60_45	-24.435	Loop	-701.767	-701.726	-701.749	-701.749	-701.736	-701.736	-107.187	25.771	-33.926	-47.490	0%
100_44	-26.066	Groove	-701.767	-701.724	-701.749	-701.749	-701.736	-701.736	-111.516	31.638	-33.926	-45.951	0%
100_23	-26.276	Groove	-701.766	-701.729	-701.749	-701.749	-701.736	-701.736	-98.263	19.145	-33.926	-45.192	0%
80_04	-27.196	Groove	-701.764	-701.732	-701.749	-701.749	-701.736	-701.736	-85.414	12.390	-33.926	-39.097	0%
40_02	-27.656	Groove	-701.759	-701.734	-701.749	-701.749	-701.736	-701.736	-66.216	5.868	-33.926	-26.422	0%
20_28	-27.656	Groove	-701.759	-701.723	-701.742	-701.749	-701.731	-701.736	-92.759	34.378	-33.926	-24.455	0%
80_23	-24.225	Groove	-701.758	-701.719	-701.742	-701.749	-701.731	-701.736	-103.292	45.006	-33.926	-24.360	0%
60_33	-26.443	Groove	-701.757	-701.719	-701.742	-701.749	-701.731	-701.736	-99.342	45.374	-33.926	-20.042	0%
100_07	-25.313	Loop	-701.751	-701.727	-701.749	-701.749	-701.736	-701.736	-62.641	23.341	-33.926	-5.373	0%
40_07	-26.108	Groove	-701.751	-701.720	-701.749	-701.749	-701.736	-701.736	-81.517	43.684	-33.926	-3.906	0%
40_19	-27.573	Loop	-701.750	-701.718	-701.742	-701.749	-701.731	-701.736	-85.994	48.857	-33.926	-3.210	0%
60_41	-26.317	Groove	-701.748	-701.726	-701.742	-701.749	-701.731	-701.736	-57.940	26.490	-33.926	2.477	0%
40_43	-29.372	Groove	-701.748	-701.726	-701.749	-701.749	-701.736	-701.736	-58.825	27.883	-33.926	2.984	0%
20_38	-27.531	Groove	-701.747	-701.734	-701.749	-701.749	-701.736	-701.736	-35.064	6.569	-33.926	5.432	0%
80_17	-23.765	Groove	-701.747	-701.726	-701.749	-701.749	-701.736	-701.736	-54.971	27.391	-33.926	6.346	0%
40_13	-25.271	Loop	-701.745	-701.727	-701.749	-701.749	-701.736	-701.736	-47.198	24.038	-33.926	10.767	0%
100_22	-27.322	Groove	-701.743	-701.681	-701.742	-701.749	-701.731	-701.736	-163.020	146.005	-33.926	16.911	0%
80_27	-26.359	Groove	-701.799	-701.733	-701.749	-701.749	-701.736	-701.736	-173.502	7.549	-33.926	-132.027	100%
20_15	-24.978	Groove	-701.784	-701.734	-701.749	-701.749	-701.736	-701.736	-131.494	5.869	-33.926	-91.698	0%
20_39	-25.522	Groove	-701.781	-701.730	-701.749	-701.749	-701.736	-701.736	-133.814	15.241	-33.926	-84.646	0%
40_15	-22.259	Groove	-701.781	-701.735	-701.749	-701.749	-701.736	-701.736	-122.198	4.403	-33.926	-83.868	0%
100_27	-25.983	Groove	-701.779	-701.730	-701.749	-701.749	-701.736	-701.736	-126.625	15.172	-33.926	-77.526	0%
100_07	-24.476	Groove	-701.777	-701.734	-701.749	-701.749	-701.736	-701.736	-115.227	7.077	-33.926	-74.223	0%
40_36	-25.188	Groove	-701.777	-701.733	-701.749	-701.749	-701.736	-701.736	-115.576	8.013	-33.926	-73.636	0%
80_17	-25.188	Groove	-701.775	-701.731	-701.749	-701.749	-701.736	-701.736	-113.563	12.555	-33.926	-67.081	0%
20_08	-26.359	Groove	-701.768	-701.734	-701.749	-701.749	-701.736	-701.736	-89.238	6.579	-33.926	-48.732	0%
20_32	-26.401	Groove	-701.761	-701.728	-701.749	-701.749	-701.736	-701.736	-88.541	22.736	-33.926	-31.879	0%
60_19	-23.849	Groove	-701.760	-701.731	-701.749	-701.749	-701.736	-701.736	-76.157	13.303	-33.926	-28.928	0%
80_49	-26.401	Groove	-701.759	-701.729	-701.742	-701.749	-701.731	-701.736	-80.268	19.665	-33.926	-26.676	0%
60_39	-27.698	Groove	-701.757	-701.730	-701.742	-701.749	-701.730	-701.736	-70.593	17.264	-33.926	-19.403	0%
100_35	-26.317	Groove	-701.750	-701.727	-701.742	-701.749	-701.731	-701.736	-61.203	25.085	-33.926	-2.192	0%
100_16	-24.727	Groove	-701.745	-701.728	-701.742	-701.749	-701.731	-701.736	-43.775	21.739	-33.926	11.891	0%
40_16	-22.468	Groove	-701.730	-701.734	-701.749	-701.749	-701.736	-701.736	10.488	5.294	-33.926	49.709	0%

4.6 Conclusion

G-quadruplexes have emerged as an attractive target for site-specific drugs for anti-cancer therapy. Guanine-rich sequences, known to form polymorphic quadruplexes, can be found in the promoter regions of oncogenes, introns and human telomeres.

A docking study to predict the binding mode of HPIP-b to G-quadruplex was performed using ensemble docking. Our docking protocol involves three steps. First, multiple conformations of the target are generated by molecular dynamic simulations, using NPT and periodic boundary conditions for the production run of 1000ps, from which five snapshot are extracted at 200ps intervals. Second, docking is performed to each snapshot structure, to implicitly account for receptor flexibility. Third, we use a two-layer ONIOM QM:QM to perform a post-docking refinement to optimise and rescore the docked complexes.

Semi-flexible docking does not simulate conformational changes of the receptor which occur upon ligand binding, unlike the real docking process, and thus limits its applicability in practice. Of the 50 x5 docking trials, HPIP-b resulted in binding to either the loop 55 (22%) or groove regions (78%) of the G-quadruplexes.

Our findings indicate that the inclusion of water molecules lead to strong electrostatic effects which are related to the formation of a hydrogen bond network between the aromatic NH and OH groups of HPIP-b to (a) G-quadruplex backbone (b) to water molecules and (c) between water molecules with other sections of the G-quadruplex, thus result in a total favourable binding energy.

CHAPTER 5. How Conformations Change: A new model for activating the blue-light sensing using flavin photoreceptor domain

5.1 Introduction to the photoreceptor system

Plants and photosynthetic organisms rely on photosensory receptors that allow them to perceive the changing environment and to adjust their metabolism or behaviour accordingly to better adapt to variations in light conditions (including the intensity, wavelength, direction and duration of ambient light) which is essential for optimal photosynthesis. Currently, six photoreceptor protein (referred to as light sensory proteins) families are known: the rhodopsins, (bacterio) phytochromes, xanthopsins, and the three blue-light photoreceptor families utilising flavin conjugates as cofactors; these are the cryptochromes (cry), phototropins (containing light-oxygen-voltage (LOV) domains) and BLUF (sensor of Blue Light Using FAD or flavin adenine dinucleotide) containing photoreceptors. Blue light which falls within the wavelength range of 455nm to 492nm is known to be an important signal for its effects upon mediating phototropism and photosynthesis, and is used to activate photolyases, enzymes involved in UV-damaged DNA repair.

The photoreceptor consists of a protein moiety and an (or several) embedded light-absorbing molecule(s) either covalently or non-covalently attached, known as chromophores involved as cofactors and whose presence is essential for photoreceptors to function. On absorption of a photon, the chromophore induces the so-called primary photoreaction (photochemical reaction); that is, the initial chemical changes result directly from irradiation with light. It begins with a photochemical change of the chromophore making a transition to an excited state [205], in which the electrons are rearranged, thereby making the various photochemical reactions possible. The photochemical reactions are specific with regard to the photoreceptor domain at hand but are consistently accompanied by a conformational change of the chromophore-protein complex that moves from the ground equilibrium (resting) state to the electronic excited metastable (signaling) state conformation for a period of time before returning back to the ground state. Accordingly, the photoreceptor serves as a signal converter whereby the external light signal perceived by the light sensitive portion of the photoreceptors is transduced from the signalling molecule into a biological signalling cascade, which is known to affect different downstream effectors involved in mediating various physiological responses [206].

The photochemical reaction leading to the formation of the signalling state is specific to each photoreceptor type. The photochemical reaction of xanthopsins, phytochromes and rhodopsins causes a cis-trans isomerization at the C=C double bond of the chromophore upon light irradiation. This excitation involves a transition of an electron from π to π* orbital, whereby the planar geometry of the double bond is distorted, and a 90° rotation about the bond becomes energetically favourable in the excited state. The subsequent relaxation of the molecule back to planarity can lead to either the cis- or trans-configuration, resulting in cis-trans isomerization of the double bond. In contrast, the photoexcited flavin-binding photoreceptors are known to undergo an intramolecular electron transfer (ET) from the close-by aromatic residues (from an adjacent tyrosine for BLUF and a conserved triad of tryptophan for cry) to the excited state (oxidised) flavin, where it acts as an electron acceptor. Electron transfer gives rise to the radical intermediate which undergoes a photochemical reaction, that produces a conformational change of the protein.

Each of the three types of flavin-binding photoreceptors gives rise to different photochemistry even though they share the same flavin chromophore; any difference in photochemistry must be attributed to the protein environment surrounding the flavin. The photochemically active flavin in cry undergoes photoreduction by which it formed a relatively long lived radical (signalling) state. [207] For LOV domains, light excitation of the flavin leads to a covalent adduct between a conserved cysteine residue in the LOV domain and the flavin at the C-4a position (Figure 31) as the signalling state and this adduct is reversed in the dark. The BLUF domains, on the other hand, undergo a photocycle involving a subtle rearrangement of the hydrogen-bond network between the flavin and nearby residue side chains, and such an alteration in hydrogen bonding is thought to be dependent on flavin-radical intermediates. [208]

Focusing on the blue light system, the essential cofactors used in the blue-light photoreceptors are flavins, specifically riboflavin (vitamin B2) derivatives (flavoprotein), most commonly seen in the form of flavin mononucleotide (FMN) or flavin adenine dinucleotide (FAD), which consists of a tricyclic isoalloxazine ring (Figure 31) substituted at the N-10 position with a ribityl phosphate (for FMN) or a ribityl adenine diphosphate (in case of FAD).

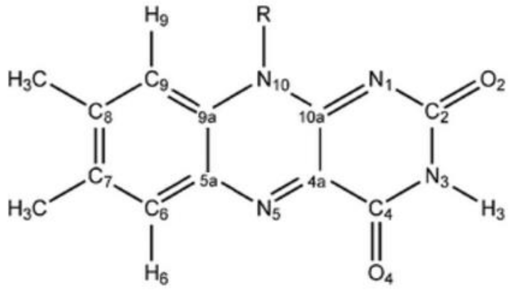


Figure 31 - Structure with atom numbering of the isoalloxazine moiety.

5.2 BLUF photocycle

BLUF domains are found primarily in prokaryotes and eukaryotic algae, but not in plants. In the dark, the BLUF domains contain flavin in the oxidised state (FAD), which acts as an electron acceptor upon irradiation. Light excitation of the flavin leads to an electron transfer from the proximal tyrosine to the excited flavin (which occurs on timescale of fs or a few ps depending on their distances and orientations relative to one another) and thereby initiates a reversible photocycle.

The formation of the anionic semiquinone (one electron reduced) radical (FAD•⁻) resulting from the single electron transfer is subsequently protonated by the neighbouring tyrosine to form the neutral semiquinone radical (FADH•) in 7-9ps, with this process generally referred to as proton-coupled electron transfer (PCET). (Figure 32) PCET is an essential part of signalling state formation. [209] The changes in the redox state of the flavin (formation of radical intermediates) will induce a conformational change that leads to reorientation of several residues (see later on "photoactivation mechanism"), affecting the hydrogen bond network in the flavin-binding pocket. The hydroxyl proton of Tyr21 is presumably attracted by the negatively charged N5 of the flavin as a result of electron transfer and thereby disrupts its hydrogen bond with Gln63, which in turn destabilises the hydrogen bonding network, enabling Gln63 reorientation to occur, leading to an altered hydrogen bonding pattern between Gln63 with its immediate environment. [210]

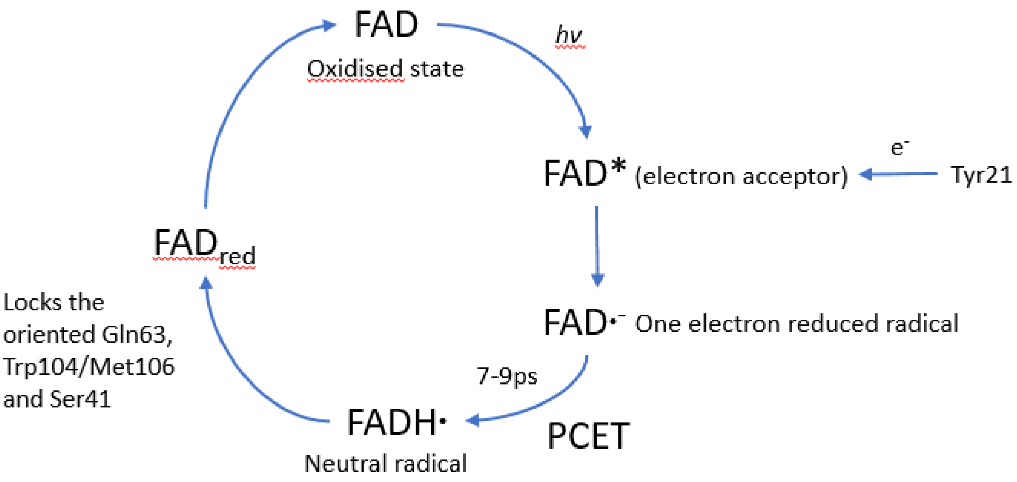


Figure 32 - Modified from [206]. The BLUF domains undergo a photocycle, in which two intermediates anionic (FAD•−) and neutral (FADH•) flavin semiquinones are formed, the latter subsequently decay to form the red-shifted signaling state (FAD_{red}).

Indeed, the side chain of the nearby Gln63 is assumed to rotate (and/or tautomerize [211]) to form a new hydrogen bond to the C4=O carbonyl of the flavin, followed by the decay of the semiquinone radical, probably occurring through a radical-pair recombination which locks the newly oriented Gln63 into place and results in the signalling state. [212,213] A semi-empirical calculation (AMPAC program using AM1 Hamiltonian) [214] shows that an increase in electron density at the N5 of the flavin due to photoexcitation does appear to be correlated with an increase in the proton affinity. [215] The fast formation of the signalling state (photoproduct) of the BLUF domain (in less than 1ns) would cause a spectral red shift of 10-15nm^{-1} in the UV-vis spectrum with respect to the initial dark state. The red shift is due to the additional hydrogen bond to flavin at O(=C4) as evidenced by the FTIR spectrum of AppA (activation of photopigment and *puc* expression A), which shows the stretching vibration of C(4)=O is downshifted by about 20cm^{-1} upon light excitation. [216] As observed, the neutral flavin radical decays into the red-shifted photoproduct within 10ns. [212] Kraft et al. described photoproduct formation as a biphasic process in which 10nm red-shifted of flavin absorption is quickly formed after light excitation (in less than 1µs), followed by a slower conformational change of the protein on a 5ms timescale. [217]

The photocycle is completed by the decay of the red-shifted signalling state back to the initial (ground) state, and known to have a relatively long half-life ranging from seconds to tens of minutes. A multi-exponential excited state decay of FAD was observed, which is attributed to structural microheterogeneity of the mutual orientation of Tyr21 and flavin in the resting state. Irradiation of flavin results in only minor changes in the structure, unlike in LOV domains, where light causes the formation of a covalent adduct. Furthermore, the flavin redox state is unchanged in both the dark-adapted and light-induced states, remaining in the fully oxidised state as evident from the UV-visible absorption spectrum. [209]

Fukushima et al. [218] reported on the photocycle of T110078 (cyanobacterial BLUF protein). They suggest that the 5-nm red shift occurring at a low temperature (below 50K) is caused by local changes limited to the chromophore and/or its immediate surroundings, and that a 10-nm red shift occurs above 50K, indicative of a further conformational change that is allowed only at higher temperatures.

The BLUF domain as a FAD-containing photosensor domain was first discovered in AppA, which is a regulatory protein negatively regulating the photosynthetic gene expression in the purple bacterium *Rhodobacter sphaeroides* in response to light and oxygen. The BLUF domain, as the name suggests, contains flavin as the light receptor molecule specialised for blue light, and was later found in many bacteria and algae. The N-terminal photosensory domain of about 100-110 residues is organised in a ferredoxin-like βαββαββ fold in which the isoalloxazine ring of flavin is bound non-covalently in a cleft between two α-helices (α1 and α2), oriented perpendicular to a five-stranded antiparallel/parallel β-sheet with a strand order of 4-1-3-2-5. (Figure 33) The ends of the domain are capped by α-helices of roughly 40-50 residues acting as linkers, connecting to the C-terminal effector domain mostly involved in cyclic nucleotide metabolism. [219] Light induced conformational changes are propagated to the effector domain where regulation takes place affecting enzymatic activity and quaternary structure (signal transduction) [220]. AppA acts as an antirepressor to the photopigment suppressor protein R (PpsR) protein, a transcription repressor of photosynthetic genes, through the activity of the cysteine-rich C-terminal catalytic domain to reduce a disulfide bond in PpsR and due to the formation of the PpsR2-AppA complex, the PpsR lost its ability to bind DNA in the anaerobic dark state. On the other hand, upon aerobic illumination, a conformational change occurred that prevented the light-activated AppA from interacting with PpsR; as a means to achieve regulation. [221] Essentially, most of the conserved residues that are located around the flavin appeared to be involved in hydrogen-bonding interactions.

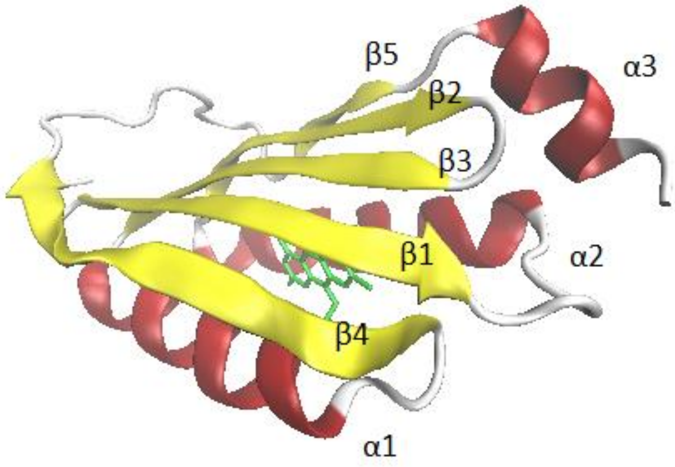


Figure 33 – The crystal structure of AppA BLUF (PDB ID 2IYG) contains a ferredoxin-like domain. The flavin (shown in green stick) is noncovalently bound between the two α-helices.

5.3 Structural Models

There remains a debate due to the contradictory findings among studies of the side-chain orientation and/or position of several key residues close to the flavin chromophore for both the resting and signalling states of BLUF, these include:

1) Gln63, which is known essential for the photocycle (which actually involves Tyr21 and Gln63, as previously discussed), and depending on its orientation, the Gln63 side chain can give rise to two sets of hydrogen bonds. So far, the available crystal structures of BLUF are not informative enough to distinguish between the side chain amidic oxygen and nitrogen atoms.

2) Trp104 (and Met106) was found to adopt two different conformations; either with its side chain (Nε-H) in close proximity to flavin located at hydrogen-bonding distance to Gln63(O), denoted as Trp-in, or on the surface of the β-sheet, denoted as Trp-out. The β-sheet was found either to act as a dimerization interface or to be shielded from solvent by the C-terminal capping helices. The positioning of Trp104 and Met106 leads to different hydrogen bonding patterns: Gln63 to Trp104 (Trp-in) or as a possible hydrogen bond donor to Met106 (Trp-out).

3) Other residues, Ser41 and Asn45 (Figure 34), whose mechanisms have remained obscure, are found conserved among the BLUF domains. Ser41 [222] appears to have two conformations, with the oxygen atom oriented towards (Sflav) or away from flavin (Sback, for backbone). Since different forms exhibit different spectral behavior, the light-induced switching of Ser41 from Sflav to Sback has a red-shifted response, thereby making a contribution to the overall (by ~10nm) red shift observed in the absorption spectra. Since the different orientations of Trp104 do not affect the absorption spectrum [222], the other contribution arises from the switching of Gln63. This movement breaks the hydrogen bond with Trp104 leading to a change in the β5 strand, and a consequent exchange of the Trp104/Met106 pair occurs [223] (possibly Trp104 moves out while Met106 moves in to fill the void [222] or changes orientation [224] upon illumination). Also, as a result of this change, a hydrogen bond is formed between Gln63 and the (C4=O) flavin; this new additional hydrogen bond to flavin is indicated by a red shift in the absorption spectra. However, experimental evidence for a direct effect of the Gln63 rotation on the conformational switch of Trp104 and Met106 is still lacking. Asn45 [224], on the other hand, forms hydrogen bonding with flavin, with an increase in the strength of the hydrogen bond upon photoexcitation promoting a red shift.

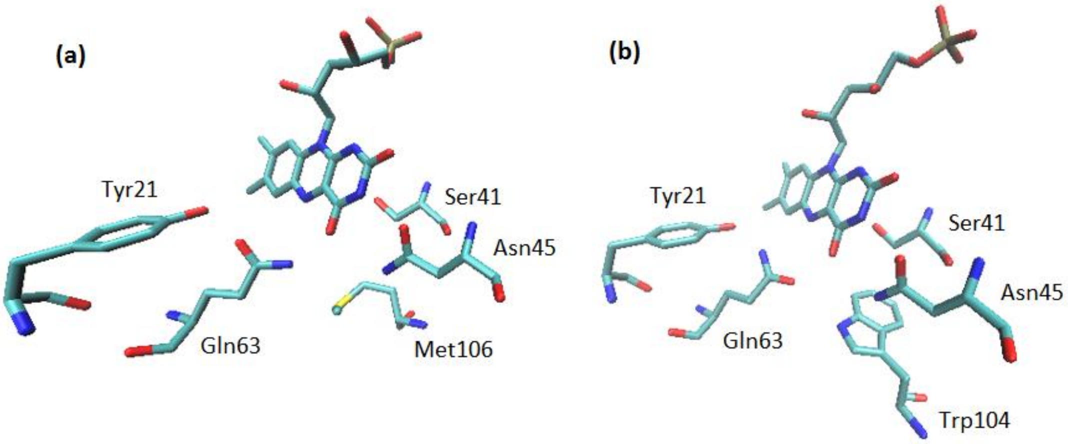


Figure 34 – The key residues close to the flavin, from PDB ID (a) 2IYG (b) 1YRX showing the position of residues Ser41 and Asn45 relative to flavin. Met106 is located near the flavin (the Trp-out conformation) instead of Trp104.

The role of relevant residues has been investigated. Site-directed mutants of AppA replacing Tyr21 with either leucine or phenylalanine were found to abolish the photocycle activity. The results from an early study assumed that Tyr21 may form a π-π stacking interaction with the isoalloxazine of flavin which would induce the photocycle. [217] However, the structural analysis reveals that the two rings are actually perpendicular to one another. Mutating Tyr21 to phenylalanine disrupts its hydrogen bond with Gln63 and the photochemical reaction. Tyr21 and Gln63 are considered as being responsible for the photocycle of AppA. Mutation of a conserved Trp104 to alanine or phenylalanine reduces the stability of the signalling state, and no structural changes of the β-sheet are observed, suggesting Trp104 has a role in transforming the light signal into changes in β-sheet structure. [223] Confusingly, by comparison with analogous mutations in SyPixD and bPAC, Trp104Phe leads to stabilisation but Trp104Ala leads to destabilisation of the structure. [225] It was observed that some variants lack of essential residues (i.e. glutamine and tyrosine) show a weaker signalling, suggesting the formation of a transient radical is sufficient to drive the signalling transduction. [226]

5.4 Photoactivation Mechanisms

Critical residues in the flavin binding pocket are in dynamic conformational exchange of both side chains and the backbone. Since NMR experiments suffer from the difficulty of assigning resides due to line broadening, multiple orientations can be proposed on the basis of NMR data. Thus, there are debates on what immediately surrounds the flavin. Two conflicting experimental structures of the BLUF domains are the PDB entries 1YRX (Trp-in) and 2IYG (Trp-out) which show considerable differences in:

(1) the backbone conformations of β5, where a kink is formed that introduces a shift of two residues in 2IYG. The carbonyl oxygen of His105 in 2IYG forms a hydrogen bond with the amide of Asn45**.** Met106 is hydrogen bonded to Gln63, whereas Trp104 is exposed to the solvent, as shown in Figure 35 (a). In contrast, in the Anderson dark-state structure (Figure 35 (b)), the amide of Asn45 is perpendicular to the indole ring of Trp104, found in close proximity to flavin. The carbonyl oxygen of His105 is hydrogen bonded to Leu54 (not shown). Met106 and Trp104 are both relevant for the photocycle, conserved in the BLUF domains, and both forming a H-bond to Gln63.

(2) the side chain of Gln63. Anderson et al. [221] claimed that the Trp-in conformation belongs to the dark-state. The Gln63 makes hydrogen bonds with the Tyr21, Trp104 and to the FMN(N5), and upon excitation this hydrogen bond breaks and Gln63 undergoes a rotation. A light-induced rotation of the Gln63 side chain was shown to disrupt its hydrogen bonding with Trp104, thus it adopts a conformation that allows hydrogen bond formation between Gln63(NH2) and FMN(O4).

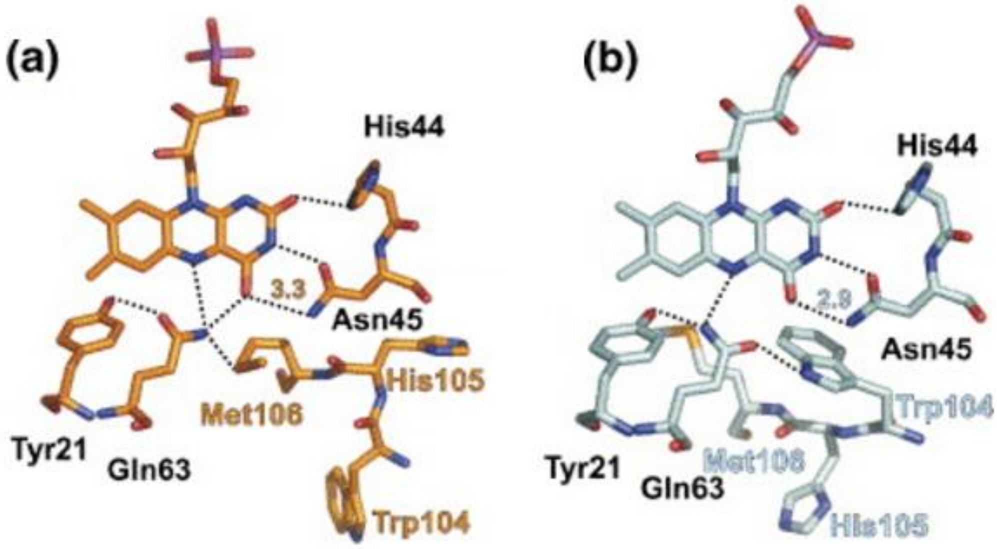


Figure 35 - Modified from [224]. The dark-state of the flavin-binding pocket (a) 2IYG (Trp-out conformation) (b)1YRX (Trp-in conformation).

Alternatively, Jung et al. [224] proposed a different orientation of Gln63, with oxygen toward the Tyr21 in the ground state such that the formation of protonated N5 leads to unfavourable interactions which promote the reorientation of Gln63.

Overall, the experimental evidence agrees: a hydrogen bond is formed between FMN(O4) and the protein matrix in the signalling state. [227] This observation could only be explained by a Gln63 rotation as the oxygen atom of the Gln63 must be pointed toward the Trp104 so that the Gln63(NH2) to FMN(O4) hydrogen bond formation could be possible, assuming the Trp-in is in the signalling state. Domratcheva et al. suggest a Gln63 rotation occurring in the signalling state. [228] Other groups say that Gln63 could be presented in a tautomeric form, thus suggesting that it could be the bond donor and acceptor at the same time. [229]

Comparisons among available studies are difficult. First of all, this could be due to variations in the BLUF sequence. For example, BlrP1 BLUF contains a threonine instead of the conserved Trp104, which is known to be of functional importance as discussed above. This significant difference from the BLUF domain of AppA makes comparison difficult. Secondly, the models are different in size, depending on the presence of capping of the helices (as linkers) at the C-terminus. Some studies work on BLUF domains in solution by using dimeric models derived from the x-ray crystal structure. Those structures variance will expect to perform differently during simulations. In this work, we cut the chains to analogous length to obtain equally sized proteins. [230]

5.5 Aims and Objectives

The aim of this project is to investigate whether the redistributed charges in the flavin binding site of the BLUF domain would result in a driving force to change the conformational preference of the protein domain, in such a way to simulate the formation of FAD•⁻ after a light-induced electron transfer from tyrosine to the flavin chromophore as the first step of the BLUF photocycle. At the end of this work we will be able to assign the Trp-in and Trp-out conformations to either of the two states 'active' or 'resting'. A better understanding of the BLUF domains would lead to applications in optogenetics, which uses light to control cellular events.

To do this, we used two conflicting structures, differing in the position of the tryptophan (Trp104 in AppA), which can be either exposed to the solvent (the 'Trp-out') or close to flavin (the 'Trp-in' conformation). Both of these structures have been attributed to the dark state (as reported by Jung et al (PDB code: 2IYG) [224] and by Anderson et al (PDB code: 1YRX) [221]). Both structures were used as the starting geometry of the protein domain to construct a "2 by 2" scheme of four calculation sets. Due to various contradictory claims on the assignment of the Trp-in and Trp-out conformations to the functional states of the BLUF domain, we make no assumptions as to the state of the structure prior to our experiments. In this work, we considered the state of the flavin binding site, described by two set of charges for each structure, the regular AmberGS charges and an altered set of charges as mimicking the electronically excited state of the flavin binding site.

A further aim of this work was to investigate the structural basis in response to the charge difference. We analyse the repositioning of the nearby residues around the Flavin: 1) we measured the distance between two atoms focusing on functionally relevant residues. 2) we studied the side chain orientation of Gln63 as it might convert to the other conformation over the course of the simulations.

5.6 Methods

As observed earlier, on photoexcitation, the N-terminal domain of AppA undergoes a photocycle that is indistinguishable from that of the full-length AppA [210]. As a result, current computational works on BLUF are mainly on the N-terminal domain. According to Rieff et al. (2011) [231], the structures of 1YRX and 2IYG are the most reliable. We adopted 1YRX Trp-in and 2IYG Trp-out conformations as starting geometries. We cut the origin structure extracted from the PDB to include equal size (including residue 13-121) for a fair comparison.

We considered excitation from the starting geometry of the protein focusing on the flavin binding site, applying individually both ground state (GS) charge taken from AmberGS and a set of relocated charges based on the AmberGS force field, from which charges for the excited Tyr21-flavin charge transfer state (ES) were calculated using time-dependent density functional theory.

Previous studies mainly focus on reproduction of UV/vis and IR spectra from MD snapshots, which are normally carried out on a single structure. This procedure is potentially statistically insufficient, as can be seen from the diversity of results obtained. This indicates that it is necessary to study the system dynamically. We applied MD to study the thermal movement of BLUF using snapshots from a preliminary MD trajectory generated by Bilal as described in the preprocessing section. Götze et al. (2012) have proposed that the initial velocity could affect the consequent hydrogen bonding behaviour of Gln63. [232] Therefore, we have repeated eight times for each of four calculation sets (16 trajectories for each calculation set, 8 for the GS and 8 for ES) to obtain the average effects.

As a result of this work, we aim to provide another activation model for the BLUF domain. The proposed driving force for activation is provided by a charge distribution. We used standard MD simulations to investigate the conformational change. Previous pure MD studies on multiple ns trajectories mainly focused on hydrogen bonding between BLUF domain and its effector domain. [233] Other MD studies which aim at resolving the hydrogen bonding patterns in the flavin binding site, however, are hard to compare due to the use of different force field, equilibrium parameters and protonated state of histidine. [232,234,235]

5.6.1 Preprocessing

All simulations were carried out with Gromacs 4.3.6 using the AmberGS force field [236] for the protein and ions, and an explicit solvent (TIP3P model). The charges of the FMN were taken from Schneider & Sühnel (1999). [237] The force field was chosen by comparing the optimised structure obtained using different force fields with the original crystal structure, and choosing that which gave the smallest RMSD.

The structures of the BLUF domains of AppA were taken from PDB code 2IYG (the Trp-out) [224] and 1YRX (the Trp-in conformation) [221]. Both were cut out into equal sized residues comprising the residues 13-121 of the chain A. The MD simulation systems consisted of either structure solvated in a box of TIP3P water then subjected to energy minimisation using the steepest descent method with convergence criterion of either a maximum number of 50000 steps or until forces reached 10 kJ mol-1 nm-1. Periodic boundary conditions were applied and electrostatic forces were evaluated using the particle Mesh Ewald algorithm with a short range cut off radius of 1nm.

After energy minimisation, the systems were gradually heated up from 0K to 300K, during which a 100ps NPT equilibration was performed at each successive temperature step, followed by a 1ns production run at 300K with a time step of 0.001ps. The H-bonds were constrained using P-LINCS. The initial structure was heated up (from initial temperature 0K) eight times with velocities randomly assigned from a Maxwell-Boltzmann distribution, resulting in eight 1ns trajectories for each structure.

5.6.2 Assignment of atomic charges for the excited state

A random snapshot was taken from each of the resulting 1ns trajectories for each structure. (eight for 1YRX and eight for 2IYG) from Bilal's work. To obtain average charge redistribution between the Tyr21 and the flavin (FMN) upon excitation, we have limited the system to tyrosine and the isoalloxazine ring in which the link atoms are replaced by hydrogens. (Figure 36) The system was calculated using TD-DFT at the CAM-B3LYP/6-31G* level with the rest of the protein atoms and the solvent 10 Ångstron radii around the protein represented as point charges. We identified the excited state where electron density moves from Tyr21 to FMN and used the excited state number (root) with key word density to set up a calculation for the excited state charge distribution. The changes in redistribution of the charge from the ground to excited state for each Tyr21 and flavin pair over eight were averaged and the amount of change was applied to the regular charges to obtain the excited Tyr21-flavin charge.

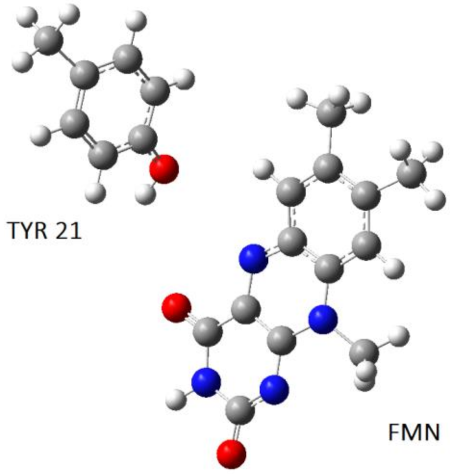


Figure 36 - Setup for TD-DFT.

5.6.3 Dynamic simulations

Snapshots were extracted every 100ps from each 1ns trajectory (resulting in 10×8 for each structure), as shown in Figure 37. We performed a 100ps run for each snapshot with regular AmberGS charges and the excited charges. We made comparisons between the two trajectories, analysing the conformational changes during the course of the simulations. All simulations were performed at 300K using an NVT ensemble.

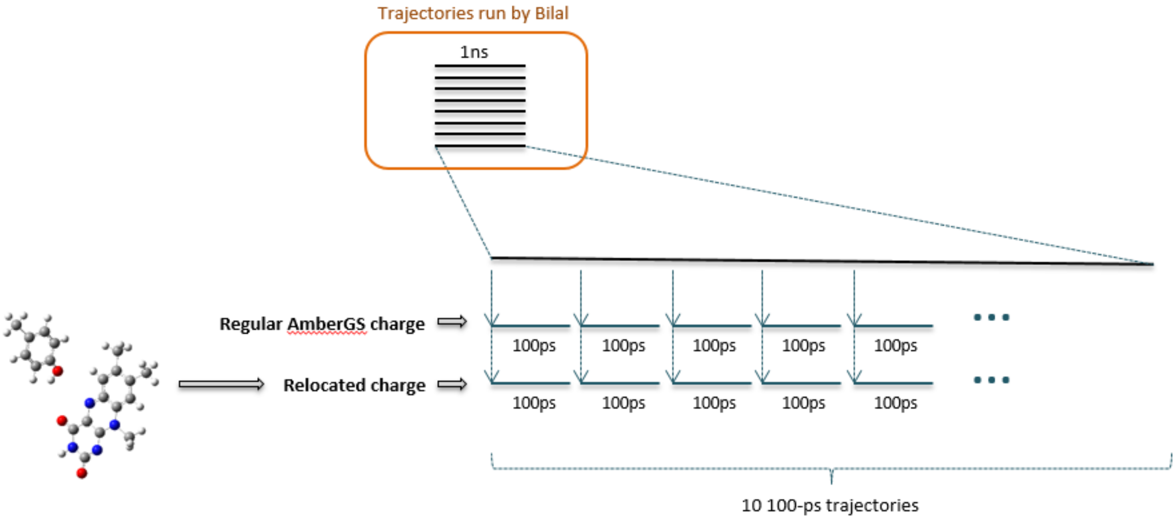


Figure 37 - Scheme of the methodology.

5.6.4 Criteria of comparison between GS & ES

5.6.4.1 Distance

To investigate the repositioning of residues at the flavin-binding site in response to the change in charges, we calculated the average distance between each of the selected two atoms restricted to functionally relevant residues around flavin over the 100ps simulation. (see Table 8) We excluded the first 20ps to allow the conformations to reach their equilibrium.

	Atom pairs
1	OH(Y21) - N5(FMN)
2	OH(Y21) - OE1(Q63)
3	OH(Y21) - NE2(Q63)
4	NE1(W104) / S(M106) - N5(FMN)
5	NE1(W104) / S(M106) - OE1(Q63)
6	NE1(W104) / S(M106) - NE2(Q63)
7	OH(S41) - N5(FMN)
8	OH(S41) - S(M106)
9	N5(FMN) - OE1(Q63)
10	N5(FMN) - NE2(Q63)

Table 8 - Lists pairs of atoms for distance measures.

5.6.4.2 Dihedral angle

Current crystal structures have insufficient resolution to distinguish between side chain amidic oxygen and nitrogen of Gln63. One could have a hydrogen bond between Tyr21(OH) to Gln63(OE1) or to Gln63(NE2) for the dark-state 2IYG and 1YRX, respectively, as shown in Figure 35. In order to study the changes in the orientation of the Gln63 side chain over the 100ps simulation, two dihedral angles were measured formed by the atoms Y21(CZ)-Q63(CG)-Q63(CD)-Q63(NE2) and Y21(CZ)-Q63(CG)-Q63(CD)-Q63(OE1).

We plotted the angle vs time graphs taking the whole 100ps simulation time and rescaled the angles and count the number of switching events. We averaged the distance between Tyr21(OH) and Gln63 before and after the switching event. Depending on the starting geometry of the respective trajectory, switches may or may not occur. We aim to make a conclusion as to which of the two structures is more disturbed by the change in Coulomb parameters.

5.7 Results & Discussion

5.7.1 Charge

Light excitation of the flavin initiates an electron transfer from the proximal tyrosine (Y21) to the flavin creating Tyr21 (+)/FMN (-) radical pair. The electron density increases at the isoalloxazine ring and, as expected, an increased electron density at N5 of the flavin is observed from the relocated (ES) charges, see Table 5 in Appendix.

5.7.2 Distance

The system behaves as expected, with a drop in Tyr21 (+) and FMN (-) distance compared to the distance between their neutral counterparts. A shorter distance can be seen from both structures which is a favourable conformation for the electron transfer. Previous work [238] observed a decrease in distance between 1) Tyr21(OH) and Gln63(O) and 2) Gln63(NH2) and FMN(N5) based on 2IYG, our results clearly confirmed these findings. In the “Trp-out” conformation, a more negative FMN(O4) (also FMN(N5)) will result in Gln63(NH2) getting closer to FMN(O4). However, we observe an increase in distance between Gln63(NH2) and FMN(N5) in 1YRX. We rationalise this is due to the Gln63 rotation, as the Gln63(NH2) is often hydrogen bonding to Tyr21(OH), thus with the Gln63(O) oriented towards the FMN. A larger negative charge at FMN would repel the Gln63(O) unless a switch occurs. (see Table 11 and 12)

5.7.3 Angle

The conformational transition (switch) of the Gln63 side chain was analysed. Table 13 is a summary of the number of switches on picosecond timescale for 1YRX. The angle vs time graphs are plotted in the Appendix. The "Trp-in" conformation (1YRX) with the Q63(NH2) oriented towards OH(Y21) should have a dihedral angle Y21(CZ)-Q63(CG)-Q63(CD)-Q63(NE2) close to zero for the ground state.

1YRX unfortunately has plenty (>50%) of geometries that effectively start in the switched conformation. (see Appendix Figure 1-8) In four (2,6,7,8) trajectories, it is completely in the switched Gln63 orientation, which with no evaluation value. Because it is no longer in its initial conformation before we apply anything, we hope to see if any switch will happen from its initial conformation. No switch is observed for the first set of trajectories. Three (3,4,5) trajectories each have one case starting in the switched orientation so that the corresponding trajectories are not evaluated. The life time of the switched conformation is prolonged in the excited state in comparison to the corresponding ground state trajectory. In future work, one may produce more trajectories on unswitched 1YRX.

Here, we reported the average distance between Tyr21(OH) and Gln63 (O) before and after the switching event (Table 9). This effect of the switch is the most significant among all of our distance measures, as the Gln63(OE) is close to hydrogen-bonding distance upon switching, after being far away before the switch.

ES		switch time (ps)	avg distance before switching (nm)	avg distance after switching (nm)	Back switching (nm)
3	1000	55	0.761494	0.452574	
4	600	20	0.521298	0.260246	
	700	85	0.559946	0.26549	
	800	5	0.445159	0.278662	
5	300	35,65	0.424301	0.266453	0.471271
	900	55	0.710382	0.266118	

Table 9 - Average distance between Tyr21(OH) and Gln63 (O) before and after a switch is occurred. 1nm=10Å.

In contrast, the position of the Gln63(NH2) group is very random, but appears to form weak interactions with FMN in all switching states. (see Appendix Table 6-8) The Tyr21(OH) and Gln63(NH2) distance is slightly larger in the switched state, which makes sense as the Gln63 (O) forms a hydrogen bond to Tyr21(OH) and Gln63 (O) and Gln63(NH2) are located at opposite ends of the end group of Gln63. (Table 10)

ES		switch time (ps)	avg distance before switching (nm)	avg distance after switching (nm)	Back switching (nm)
3	1000	55	0.613805	0.607003	
4	600	20	0.408162	0.454528	
	700	85	0.469891	0.411732	
	800	5	0.431713	0.43768	
5	300	35,65	0.356078	0.421399	0.331078
	900	55	0.568704	0.429289	

Table 10 - Average distance between Tyr21(OH) and Gln63 (NH2) before and after a switch is occurred.

Switching only happens in the 1YRX (Trp-in conformation) state. The 2IYG seems to be the signalling state because no switching event has been found after applying an excited charged state. (see Appendix Figure 9-16) We have artificially created a state of charge reversal, an artificial Tyr (-)/FMN (+) state, where we forced the system to make switching happen in the Trp-out. (Table 14)

The "Trp-out" conformation (2IYG) with the Q63(O) oriented towards OH(Y21) should have a dihedral angle Y21(CZ)-Q63(CG)-Q63(CD)-Q63(OE1) close to zero for the ground state. The Q63(NE2) graph which is made for comparison will yield a switch for the same cases. Using an artificial charge state, switching was observed in 15 out of 80 trajectories leading to OH(Y21) and Q63(NE2) bridge. In general, those switches happened very early, typically within 5ps. (Table 14 and Appendix Figure 17-24)

5.8 Conclusion

The BLUF domain is a FAD-binding blue-light-sensing protein, which regulates photosynthesis gene expression in the purple bacterium. The BLUF domain undergoes a photocycle upon illumination involving a subtle rearrangement of the hydrogen network in the flavin binding site. The photochemical reaction involves both electron and proton transfer from nearby residues. In this work, we aim to investigate whether the redistributed charges in the flavin binding site would result in a driving force to change the conformational preference of the protein domain so as to assign the functional state of the two conflicting geometries "Trp-in" and "Trp-out". Our results show that switching only happens in the 1YRX ("Trp-in" conformation) state. The switched conformation prefers the ES charge distribution, as seen from the few cases where the switches happened very early. Our study shows that 2IYG to be the signaling state as no switching event was observed from all eight trajectories, suggesting the "Trp-out" conformation appears to favour the ES charge distribution. Switching only happens in the Trp-out case using an artificial Try21 (-)/FMN (+) state.

Table 11 - Presents the distance averaged over the course of the 100ps simulations, excluding the first 20ps, for 1YRX (Trp-in conformation).

	GS	ES	Change		GS	ES	Change
OH(Y21)-N5(FMN)-100	0.5045	0.4489	-0.0557	OH(S41)-N5(FMN)-100	0.3631	0.3782	0.0150
OH(Y21)-N5(FMN)-200	0.5194	0.4453	-0.0741	OH(S41)-N5(FMN)-200	0.3585	0.3695	0.0110
OH(Y21)-N5(FMN)-300	0.5174	0.4361	-0.0813	OH(S41)-N5(FMN)-300	0.3529	0.3674	0.0145
OH(Y21)-N5(FMN)-400	0.5472	0.4288	-0.1185	OH(S41)-N5(FMN)-400	0.3551	0.3696	0.0145
OH(Y21)-N5(FMN)-500	0.5404	0.4510	-0.0894	OH(S41)-N5(FMN)-500	0.3558	0.3660	0.0102
OH(Y21)-N5(FMN)-600	0.5095	0.4570	-0.0525	OH(S41)-N5(FMN)-600	0.3540	0.3705	0.0165
OH(Y21)-N5(FMN)-700	0.5117	0.4537	-0.0581	OH(S41)-N5(FMN)-700	0.3568	0.3634	0.0066
OH(Y21)-N5(FMN)-800	0.5169	0.4253	-0.0916	OH(S41)-N5(FMN)-800	0.3593	0.3688	0.0095
OH(Y21)-N5(FMN)-900	0.5287	0.4460	-0.0827	OH(S41)-N5(FMN)-900	0.3560	0.3655	0.0095
OH(Y21)-N5(FMN)-1000	0.4976	0.4317	-0.0659	OH(S41)-N5(FMN)-1000	0.3568	0.3731	0.0163
OH(Y21)-OE1(Q63)-100	0.4024	0.3323	-0.0700	OE1(Q63)-N5(FMN)-100	0.3873	0.3995	0.0121
OH(Y21)-OE1(Q63)-200	0.4762	0.4430	-0.0332	OE1(Q63)-N5(FMN)-200	0.4239	0.4598	0.0360
OH(Y21)-OE1(Q63)-300	0.5008	0.4116	-0.0892	OE1(Q63)-N5(FMN)-300	0.4367	0.4361	-0.0006
OH(Y21)-OE1(Q63)-400	0.5279	0.4203	-0.1076	OE1(Q63)-N5(FMN)-400	0.4273	0.4427	0.0154
OH(Y21)-OE1(Q63)-500	0.5070	0.4768	-0.0302	OE1(Q63)-N5(FMN)-500	0.4249	0.4539	0.0290
OH(Y21)-OE1(Q63)-600	0.4826	0.4042	-0.0785	OE1(Q63)-N5(FMN)-600	0.4251	0.4239	-0.0012
OH(Y21)-OE1(Q63)-700	0.4591	0.4498	-0.0093	OE1(Q63)-N5(FMN)-700	0.4269	0.4466	0.0197
OH(Y21)-OE1(Q63)-800	0.4946	0.3790	-0.1156	OE1(Q63)-N5(FMN)-800	0.4254	0.4367	0.0113
OH(Y21)-OE1(Q63)-900	0.5058	0.4431	-0.0628	OE1(Q63)-N5(FMN)-900	0.4260	0.4307	0.0047
OH(Y21)-OE1(Q63)-1000	0.4767	0.4203	-0.0564	OE1(Q63)-N5(FMN)-1000	0.4306	0.4652	0.0346
OH(Y21)-NE2(Q63)-100	0.4667	0.4571	-0.0096	NE2(Q63)-N5(FMN)-100	0.3445	0.3567	0.0121
OH(Y21)-NE2(Q63)-200	0.5049	0.4602	-0.0447	NE2(Q63)-N5(FMN)-200	0.3540	0.3779	0.0240
OH(Y21)-NE2(Q63)-300	0.5177	0.4597	-0.0580	NE2(Q63)-N5(FMN)-300	0.3658	0.3793	0.0135
OH(Y21)-NE2(Q63)-400	0.5346	0.4418	-0.0928	NE2(Q63)-N5(FMN)-400	0.3417	0.3552	0.0135
OH(Y21)-NE2(Q63)-500	0.5186	0.4920	-0.0267	NE2(Q63)-N5(FMN)-500	0.3426	0.3798	0.0372
OH(Y21)-NE2(Q63)-600	0.4949	0.4732	-0.0218	NE2(Q63)-N5(FMN)-600	0.3554	0.3617	0.0062
OH(Y21)-NE2(Q63)-700	0.4768	0.4813	0.0044	NE2(Q63)-N5(FMN)-700	0.3538	0.3625	0.0087
OH(Y21)-NE2(Q63)-800	0.4988	0.4446	-0.0542	NE2(Q63)-N5(FMN)-800	0.3573	0.3773	0.0200
OH(Y21)-NE2(Q63)-900	0.5094	0.4702	-0.0393	NE2(Q63)-N5(FMN)-900	0.3542	0.3636	0.0093
OH(Y21)-NE2(Q63)-1000	0.4880	0.4749	-0.0131	NE2(Q63)-N5(FMN)-1000	0.3650	0.3970	0.0320
NE1(W104)-N5(FMN)-100	0.5703	0.5587	-0.0116				
NE1(W104)-N5(FMN)-200	0.6042	0.6185	0.0143				
NE1(W104)-N5(FMN)-300	0.6258	0.6001	-0.0257				
NE1(W104)-N5(FMN)-400	0.6245	0.6255	0.0010				
NE1(W104)-N5(FMN)-500	0.6016	0.6234	0.0218				
NE1(W104)-N5(FMN)-600	0.6060	0.6018	-0.0043				
NE1(W104)-N5(FMN)-700	0.6206	0.6224	0.0018				
NE1(W104)-N5(FMN)-800	0.6134	0.6223	0.0089				
NE1(W104)-N5(FMN)-900	0.6292	0.6311	0.0018				
NE1(W104)-N5(FMN)-1000	0.6188	0.6436	0.0248				

NE1(W104)-OE1(Q63)-100	0.4826	0.5107	0.0281
NE1(W104)-OE1(Q63)-200	0.4633	0.4431	-0.0203
NE1(W104)-OE1(Q63)-300	0.4509	0.4460	-0.0049
NE1(W104)-OE1(Q63)-400	0.4592	0.4419	-0.0173
NE1(W104)-OE1(Q63)-500	0.4542	0.4291	-0.0250
NE1(W104)-OE1(Q63)-600	0.4374	0.4737	0.0363
NE1(W104)-OE1(Q63)-700	0.4626	0.4531	-0.0095
NE1(W104)-OE1(Q63)-800	0.4437	0.4664	0.0228
NE1(W104)-OE1(Q63)-900	0.4501	0.4696	0.0194
NE1(W104)-OE1(Q63)-1000	0.4390	0.4429	0.0039
NE1(W104)-NE2(Q63)-100	0.4351	0.3921	-0.0430
NE1(W104)-NE2(Q63)-200	0.4457	0.4181	-0.0276
NE1(W104)-NE2(Q63)-300	0.4442	0.4136	-0.0306
NE1(W104)-NE2(Q63)-400	0.4539	0.4258	-0.0281
NE1(W104)-NE2(Q63)-500	0.4430	0.4169	-0.0261
NE1(W104)-NE2(Q63)-600	0.4346	0.4169	-0.0178
NE1(W104)-NE2(Q63)-700	0.4549	0.4238	-0.0311
NE1(W104)-NE2(Q63)-800	0.4419	0.4226	-0.0193
NE1(W104)-NE2(Q63)-900	0.4489	0.4400	-0.0089
NE1(W104)-NE2(Q63)-1000	0.4382	0.4171	-0.0211

Table 12 - Presents the distance averaged over the course of the 100ps simulations, excluding the first 20ps, for 2IYG (Trp-out conformation).

	GS	ES	Change		GS	ES	Change
OH(Y21)-N5(FMN)-100	0.4499	0.4293	-0.0205	OH(S41)-S(M106)-100	0.5551	0.5568	0.0016
OH(Y21)-N5(FMN)-200	0.4554	0.4218	-0.0336	OH(S41)-S(M106)-200	0.5735	0.5753	0.0018
OH(Y21)-N5(FMN)-300	0.4623	0.4299	-0.0324	OH(S41)-S(M106)-300	0.5704	0.5771	0.0066
OH(Y21)-N5(FMN)-400	0.4571	0.4325	-0.0246	OH(S41)-S(M106)-400	0.5710	0.5659	-0.0051
OH(Y21)-N5(FMN)-500	0.4573	0.4298	-0.0275	OH(S41)-S(M106)-500	0.5670	0.5603	-0.0067
OH(Y21)-N5(FMN)-600	0.4543	0.4281	-0.0262	OH(S41)-S(M106)-600	0.5640	0.5741	0.0100
OH(Y21)-N5(FMN)-700	0.4504	0.4326	-0.0178	OH(S41)-S(M106)-700	0.5644	0.5573	-0.0071
OH(Y21)-N5(FMN)-800	0.4571	0.4384	-0.0187	OH(S41)-S(M106)-800	0.5644	0.5802	0.0158
OH(Y21)-N5(FMN)-900	0.4605	0.4402	-0.0203	OH(S41)-S(M106)-900	0.5698	0.5642	-0.0056
OH(Y21)-N5(FMN)-1000	0.4614	0.4310	-0.0304	OH(S41)-S(M106)-1000	0.5683	0.5662	-0.0021
OH(Y21)-OE1(Q63)-100	0.2726	0.2629	-0.0097	OH(S41)-N5(FMN)-100	0.3556	0.3639	0.0084
OH(Y21)-OE1(Q63)-200	0.2717	0.2621	-0.0097	OH(S41)-N5(FMN)-200	0.3592	0.3753	0.0162
OH(Y21)-OE1(Q63)-300	0.2715	0.2620	-0.0095	OH(S41)-N5(FMN)-300	0.3574	0.3755	0.0181
OH(Y21)-OE1(Q63)-400	0.2722	0.2615	-0.0107	OH(S41)-N5(FMN)-400	0.3537	0.3771	0.0235
OH(Y21)-OE1(Q63)-500	0.2723	0.2619	-0.0104	OH(S41)-N5(FMN)-500	0.3552	0.3794	0.0242
OH(Y21)-OE1(Q63)-600	0.2729	0.2623	-0.0106	OH(S41)-N5(FMN)-600	0.3516	0.3773	0.0257
OH(Y21)-OE1(Q63)-700	0.2730	0.2622	-0.0108	OH(S41)-N5(FMN)-700	0.3596	0.3849	0.0253
OH(Y21)-OE1(Q63)-800	0.2729	0.2619	-0.0111	OH(S41)-N5(FMN)-800	0.3508	0.3722	0.0214
OH(Y21)-OE1(Q63)-900	0.2726	0.2620	-0.0107	OH(S41)-N5(FMN)-900	0.3635	0.3832	0.0197
OH(Y21)-OE1(Q63)-1000	0.2725	0.2619	-0.0107	OH(S41)-N5(FMN)-1000	0.3536	0.3666	0.0130
OH(Y21)-NE2(Q63)-100	0.4333	0.4314	-0.0019	OE1(Q63)-N5(FMN)-100	0.3533	0.3530	-0.0003
OH(Y21)-NE2(Q63)-200	0.4371	0.4273	-0.0098	OE1(Q63)-N5(FMN)-200	0.3522	0.3537	0.0014
OH(Y21)-NE2(Q63)-300	0.4348	0.4319	-0.0029	OE1(Q63)-N5(FMN)-300	0.3575	0.3542	-0.0033
OH(Y21)-NE2(Q63)-400	0.4343	0.4317	-0.0026	OE1(Q63)-N5(FMN)-400	0.3574	0.3582	0.0007
OH(Y21)-NE2(Q63)-500	0.4361	0.4297	-0.0064	OE1(Q63)-N5(FMN)-500	0.3558	0.3557	-0.0001
OH(Y21)-NE2(Q63)-600	0.4345	0.4267	-0.0078	OE1(Q63)-N5(FMN)-600	0.3561	0.3566	0.0005
OH(Y21)-NE2(Q63)-700	0.4283	0.4333	0.0050	OE1(Q63)-N5(FMN)-700	0.3587	0.3542	-0.0046
OH(Y21)-NE2(Q63)-800	0.4348	0.4318	-0.0030	OE1(Q63)-N5(FMN)-800	0.3592	0.3583	-0.0010
OH(Y21)-NE2(Q63)-900	0.4374	0.4351	-0.0024	OE1(Q63)-N5(FMN)-900	0.3599	0.3574	-0.0025
OH(Y21)-NE2(Q63)-1000	0.4371	0.4303	-0.0068	OE1(Q63)-N5(FMN)-1000	0.3607	0.3578	-0.0029
S(M106)-N5(FMN)-100	0.5712	0.5733	0.0021	NE2(Q63)-N5(FMN)-100	0.3170	0.3014	-0.0156
S(M106)-N5(FMN)-200	0.5762	0.5784	0.0022	NE2(Q63)-N5(FMN)-200	0.3144	0.3027	-0.0118
S(M106)-N5(FMN)-300	0.5707	0.5736	0.0029	NE2(Q63)-N5(FMN)-300	0.3147	0.3020	-0.0127
S(M106)-N5(FMN)-400	0.5750	0.5749	-0.0001	NE2(Q63)-N5(FMN)-400	0.3168	0.3077	-0.0091
S(M106)-N5(FMN)-500	0.5786	0.5750	-0.0036	NE2(Q63)-N5(FMN)-500	0.3161	0.3033	-0.0128
S(M106)-N5(FMN)-600	0.5722	0.5723	0.0001	NE2(Q63)-N5(FMN)-600	0.3167	0.3026	-0.0141
S(M106)-N5(FMN)-700	0.5711	0.5743	0.0032	NE2(Q63)-N5(FMN)-700	0.3161	0.3032	-0.0129
S(M106)-N5(FMN)-800	0.5666	0.5692	0.0026	NE2(Q63)-N5(FMN)-800	0.3211	0.3035	-0.0177
S(M106)-N5(FMN)-900	0.5698	0.5700	0.0002	NE2(Q63)-N5(FMN)-900	0.3175	0.3028	-0.0147
S(M106)-N5(FMN)-1000	0.5708	0.5719	0.0011	NE2(Q63)-N5(FMN)-1000	0.3174	0.3031	-0.0143

S(M106)-OE1(Q63)-100	0.5470	0.5430	-0.0040
S(M106)-OE1(Q63)-200	0.5440	0.5412	-0.0029
S(M106)-OE1(Q63)-300	0.5432	0.5355	-0.0077
S(M106)-OE1(Q63)-400	0.5434	0.5434	-0.0001
S(M106)-OE1(Q63)-500	0.5459	0.5369	-0.0090
S(M106)-OE1(Q63)-600	0.5423	0.5378	-0.0045
S(M106)-OE1(Q63)-700	0.5397	0.5423	0.0026
S(M106)-OE1(Q63)-800	0.5400	0.5367	-0.0033
S(M106)-OE1(Q63)-900	0.5390	0.5333	-0.0057
S(M106)-OE1(Q63)-1000	0.5415	0.5343	-0.0071
S(M106)-NE2(Q63)-100	0.3544	0.3507	-0.0037
S(M106)-NE2(Q63)-200	0.3489	0.3529	0.0040
S(M106)-NE2(Q63)-300	0.3481	0.3465	-0.0016
S(M106)-NE2(Q63)-400	0.3479	0.3537	0.0058
S(M106)-NE2(Q63)-500	0.3512	0.3475	-0.0037
S(M106)-NE2(Q63)-600	0.3476	0.3475	-0.0001
S(M106)-NE2(Q63)-700	0.3453	0.3507	0.0055
S(M106)-NE2(Q63)-800	0.3472	0.3454	-0.0018
S(M106)-NE2(Q63)-900	0.3439	0.3433	-0.0006
S(M106)-NE2(Q63)-1000	0.3469	0.3450	-0.0019

Table 13 - Number of switches (on ps timescale) observed from the eight trajectories of 1YRX.

ps	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3 (1)	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
4 (1)	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	3	3	3	3
5 (1)	0	0	0	0	0	0	1	1	1	1	2	2	2	2	2	2	2	2	2	2
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Sum	1	1	1	2	2	2	3	3	3	3	5	5	5	5	5	5	6	6	6	6

Table 14 - Number of switches (on ps timescale) observed from the eight trajectories of 2IYG by applying an artificial Tyr (-)/FMN (+) state.

ps	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
2	3	4	4	4	4	4	5	5	5	6	6	6	6	6	6	6	6	6	6	6
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	1	1	1	1	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3
7	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
8	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2
Sum	9	10	10	10	11	11	12	12	12	13	13	13	14	14	14	14	14	14	15	15

Appendix

Table 1 – List of 59 representative allosteric pockets.

	PDB	Ligand	
1	1AO0	ADP	468A
2	1CSM	TRP	501B
3	1CZA	ADP	922N
4	1EGY	9AP	801A
5	1ESM	COA	403D
6	1EYJ	AMP	341A
7	1H5U	CHI	920A
8	1I1Q	TRP	1001A
9	1IE9	VDX	500A
10	1JL0	PUT	2020B
11	1JQN	ASP	884A
12	1KFL	PHE	5354E
13	1LTH	FBP	320R
14	1M8P	PPS	576C
15	1NTK	AY1	383C
16	1NXG	NAD	3000A
17	1PJ3	FUM	700A
18	1PSD	SER	451A
19	1RD4	L08	3328D
20	1S9J	BBM	1001A
21	1SHJ	NXN	401B
22	1T36	U	1093E
23	1V4S	MRK	501A
24	1W25	C2E	505B
25	1W96	S1A	1567A
26	1X88	NAT	802B
27	1XJF	DTP	1001B
28	1XXA	ARG	1F
29	1YXD	LYS	1300B
30	1ZDQ	MSM	1509C
31	2AL5	4MP	801A
32	2CLK	G3H	1268A
33	2D5Z	L35	1200C
34	2FZC	CTP	901B
35	2G50	ALA	6106F
36	2I80	G1L	400A
37	2JC9	ADN	1497A
38	2JHR	PBQ	1780A
39	2JJX	ATP	1246C
40	2ONB	7PA	319A
41	2P2N	ASN	8004D
42	2P9H	IPT	998A
43	2PIV	T3	933A
44	2Q5Q	KPV	5004A
45	2V8Q	AMP	1327E
46	2VD4	P21	1454A
47	3BEO	UD1	372B
48	3D2P	ARG	438B
49	3DC2	SER	600A

50	3F6G	ILE	1A
51	3FZY	IHP	8000A
52	3GCP	SB2	361A
53	3H30	RFZ	337A
54	3HRF	P47	1374A
55	3I0S	RT7	601A
56	3IAD	15X	901D
57	3JVR	AGX	901A
58	3KCC	CMP	302B
59	3R1R	ATP	762B

Table 2 – List of 151 small molecule descriptors. Detailed can be found in the CDK documentation.

1	AtomCount:C
2	AtomCount:O
3	AtomCount:N
4	AtomCount:P
5	AtomCount:Cl
6	AtomCount:F
7	AtomCount:S
8	AtomCount:Br
9	AtomCount:I
10	Total_HA
11	ZagrebIndex:Zagreb
12	XLogP:XLogP
13	WienerNumbers:WPOL
14	WienerNumbers:WPATH
15	WeightedPath:WTPT-5
16	WeightedPath:WTPT-4
17	WeightedPath:WTPT-3
18	WeightedPath:WTPT-2
19	WeightedPath:WTPT-1
20	Weight:MW
21	VAdjMa:VAdjMat
22	TPSA:TopoPSA
23	RuleOfFive:LipinskiFailures
24	RotatableBondsCount:nRotB
25	PetitjeanShapeIndex:geomShape
26	PetitjeanNumber:PetitjeanNumber
27	MolWt:MWt
28	MDE:MDEN-33
29	MDE:MDEN-23
30	MDE:MDEN-22
31	MDE:MDEN-13
32	MDE:MDEN-12
33	MDE:MDEN-11
34	MDE:MDEO-22
35	MDE:MDEO-12
36	MDE:MDEO-11
37	MDE:MDEC-44
38	MDE:MDEC-34
39	MDE:MDEC-33
40	MDE:MDEC-24
41	MDE:MDEC-23
42	MDE:MDEC-22
43	MDE:MDEC-14
44	MDE:MDEC-13
45	MDE:MDEC-12
46	MDE:MDEC-11
47	MannholdLogP:MLogP
48	LongestAliphaticChain:nAtomLAC
49	LargestPiSystem:nAtomP
50	LargestChain:nAtomLC
51	KierHallSmarts:khs.sssSnH
52	KierHallSmarts:khs.sSnH3

53	KierHallSmarts:khs.sBr
54	KierHallSmarts:khs.ddssSe
55	KierHallSmarts:khs.dssSe
56	KierHallSmarts:khs.aaSe
57	KierHallSmarts:khs.dSe
58	KierHallSmarts:khs.sSeH
59	KierHallSmarts:khs.sssssAs
60	KierHallSmarts:khs.sssdAs
61	KierHallSmarts:khs.sssAs
62	KierHallSmarts:khs.ssAsH
63	KierHallSmarts:khs.sAsH2
64	KierHallSmarts:khs.sssGeH
65	KierHallSmarts:khs.ssGeH2
66	KierHallSmarts:khs.sGeH3
67	KierHallSmarts:khs.ddssS
68	KierHallSmarts:khs.dssS
69	KierHallSmarts:khs.aaS
70	KierHallSmarts:khs.ssS
71	KierHallSmarts:khs.dS
72	KierHallSmarts:khs.sSH
73	KierHallSmarts:khs.sssssP
74	KierHallSmarts:khs.dsssP
75	KierHallSmarts:khs.sssP
76	KierHallSmarts:khs.ssPH
77	KierHallSmarts:khs.sOH
78	KierHallSmarts:khs.aasN
79	KierHallSmarts:khs.sssN
80	KierHallSmarts:khs.sssNH
81	KierHallSmarts:khs.tCH
82	KappaShapeIndices:Kier3
83	KappaShapeIndices:Kier2
84	KappaShapeIndices:Kier1
85	HybridizationRatio:HybRatio
86	HBondDonorCount:nHBDon
87	HBondAcceptorCount:nHBAcc
88	FragmentComplexity:fragC
89	FMF:FMF
90	EccentricConnectivityIndex:ECCEN
91	ChiPath:VP-7
92	ChiPath:VP-6
93	ChiPath:VP-5
94	ChiPath:VP-4
95	ChiPath:VP-3
96	ChiPath:VP-2
97	ChiPath:VP-1
98	ChiPath:VP-0
99	ChiPath:SP-7
100	ChiPath:SP-6
101	ChiPath:SP-5
102	ChiPath:SP-4
103	ChiPath:SP-3
104	ChiPath:SP-2
105	ChiPath:SP-1
106	ChiPath:SP-0
107	ChiPathCluster:VPC-6
108	ChiPathCluster:VPC-5

109	ChiPathCluster:VPC-4
110	ChiPathCluster:SPC-6
111	ChiPathCluster:SPC-5
112	ChiPathCluster:SPC-4
113	ChiCluster:VC-6
114	ChiCluster:VC-5
115	ChiCluster:VC-4
116	ChiCluster:VC-3
117	ChiCluster:SC-6
118	ChiCluster:SC-5
119	ChiCluster:SC-4
120	ChiCluster:SC-3
121	ChiChain:VCH-5
122	ChiChain:VCH-4
123	ChiChain:VCH-3
124	ChiChain:SCH-5
125	ChiChain:SCH-4
126	ChiChain:SCH-3
127	CarbonTypes:C4SP3
128	CarbonTypes:C3SP3
129	CarbonTypes:C2SP3
130	CarbonTypes:C1SP3
131	CarbonTypes:C3SP2
132	CarbonTypes:C2SP2
133	CarbonTypes:C1SP2
134	CarbonTypes:C2SP1
135	CarbonTypes:C1SP1
136	BPol:bpol
137	BondCount:nB
138	BCUT:BCUTp-1h
139	BCUT:BCUTp-1l
140	BCUT:BCUTc-1h
141	BCUT:BCUTc-1l
142	BCUT:BCUTw-1h
143	BCUT:BCUTw-1l
144	BasicGroupCount:nBase
145	AtomCount:nAtom
146	AromaticBondsCount:nAromBond
147	APol:apol
148	AminoAcidCount:nW
149	AminoAcidCount:nM
150	AminoAcidCount:nP
151	AcidicGroupCount:nAcid

Table 3 - Results of Model 1 (43 descriptors) for the CHES set.

A	R	T	Prediction	PDB	Ligand	Literature
1	5	94	T	2VW2	A1000	active site
19	5	76	T	3OQI	A258	catalytic pocket
25	0	75	T	2ICH	B2	interdomain interface
10	16	74	T	3NOQ	A501	active site
29	0	71	T	2ICH	A1	interdomain interface
41	1	58	T	3OQI	B258	catalytic pocket
90	9	1	A	3RIG	A2001	active site-acyl pocket
84	16	0	A	1YI1	A696	≈20Å from the active site
68	32	0	A	3RIG	B2002	active site-acyl pocket
67	33	0	A	1Z8N	A696	≈20Å from the active site
64	35	1	A	1Q1Q	A354	active site-competes with substrate
58	5	37	A	1V30	A2854	probably active site
51	49	0	A	1YI0	A696	≈20Å from the active site
95	5	0	A	4DQ0	C202	to be published
93	7	0	A	4DQ0	B201	to be published
66	34	0	A	4DQ0	C201	to be published
74	26	0	A	3G8W	A168	to be published
68	32	0	A	3G8W	A167	to be published
62	32	6	A	3G8W	C167	to be published
57	24	19	A	3G8W	A169	to be published
43	39	18	A	3G8W	C170	to be published

Table 4 - Results of Model 5 (five descriptors) for the CHES set.

A	R	T	Prediction	PDB	Ligand	Literature
15	0	85	T	2ICH	A1	interdomain interface
16	0	84	T	2ICH	B2	interdomain interface
1	20	79	T	2VW2	A1000	active site
23	27	50	T	1V30	A2854	probably active site
41	18	41	T/A	3OQI	B258	catalytic pocket
100	0	0	A	1Q1Q	A354	active site-competes with substrate
95	5	0	A	1YI1	A696	≈20Å from the active site
92	7	1	A	3RIG	A2001	active site-acyl pocket
91	9	0	A	1Z8N	A696	≈20Å from the active site
88	12	0	A	3RIG	B2002	active site-acyl pocket
86	14	0	A	4EV1	A302	hydrophobic pocket
84	16	0	A	3GXZ	A301	primary oligosaccharide-binding sites
79	21	0	A	3OB9	A540	peptide (methyllysine) binding pocket
78	22	0	A	3NIB	A276	active site
72	28	0	A	4H75	A304	aromatic cage for methyllysine binding
71	29	0	A	1YI0	A696	≈20Å from the active site
67	33	0	A	1YBH	A696	≈20Å from the active site
61	39	0	A	1YHY	A696	≈20Å from the active site
61	39	0	A	1YHZ	A696	≈20Å from the active site
58	42	0	A	4ATG	A401	interface of symmetry-related molecules
95	5	0	A	3G8W	A168	to be published
90	9	1	A	3G8W	A167	to be published
88	12	0	A	3G8W	C167	to be published
87	12	1	A	3G8W	C170	to be published
72	28	0	A	3G8W	C168	to be published
62	38	0	A	3G8W	A169	to be published
62	38	0	A	3G8W	C171	to be published
52	48	0	A	3G8W	C169	to be published
95	5	0	A	3IXS	C1	no comment
92	8	0	A	3CGG	A195	to be published
91	9	0	A	1L5B	A301	no comment
91	9	0	A	3KSP	A129	to be published
83	17	0	A	3R97	A315	to be published
83	17	0	A	4DQ0	B201	to be published
76	24	0	A	4DQ0	C202	to be published

Table 5 - Lists the atomic charges used to perform MD simulations. The ground or dark-state (GS) charges were taken from AmberGS force field and a set of relocated charges (ES), which was obtained from a TD-DFT CAM-B3LYP/6-31G* calculation.

	1YRX		**2IYG**	
	GS	**ES**	**GS**	**ES**
N	-0.4157	-0.4157	-0.4157	-0.4157
H	0.2719	0.2719	0.2719	0.2719
CA	-0.0014	0.0665	-0.0014	0.0677
HA	0.0876	0.0876	0.0876	0.0876
CB	-0.0152	-0.0991	-0.0152	-0.1062
HB1	0.0295	0.0860	0.0295	0.0812
HB2	0.0295	0.0832	0.0295	0.0826
CG	-0.0011	0.1497	-0.0011	0.1683
CD1	-0.1906	-0.1516	-0.1906	-0.1628
HD1	0.1699	0.2014	0.1699	0.2056
CE1	-0.2341	-0.1593	-0.2341	-0.1790
HE1	0.1656	0.2206	0.1656	0.2161
CZ	0.3226	0.4721	0.3226	0.4592
OH	-0.5579	-0.3814	-0.5579	-0.3602
HH	0.3992	0.4213	0.3992	0.4168
CE2	-0.2341	-0.1823	-0.2341	-0.1378
HE2	0.1656	0.2156	0.1656	0.2143
CD2	-0.1906	-0.1394	-0.1906	-0.1864
HD2	0.1699	0.1992	0.1699	0.2070
C	0.5973	0.5973	0.5973	0.5973
O	-0.5679	-0.5679	-0.5679	-0.5679
SUM	0.0000	0.9757	0.0000	0.9597
C10	0.0540	0.0530	0.0540	0.0372
N1	-0.3250	-0.3756	-0.3250	-0.3717
C2	0.5490	0.5652	0.5490	0.5780
O2	-0.4940	-0.5703	-0.4940	-0.5677
N3	-0.3840	-0.4192	-0.3840	-0.4344
HN	0.3050	0.2889	0.3050	0.2957
C4	0.3730	0.3849	0.3730	0.3772
O4	-0.5350	-0.6298	-0.5350	-0.6285
C4A	0.5180	0.3623	0.5180	0.3890
N5	**-0.5500**	**-0.6804**	**-0.5500**	**-0.6963**
C5A	0.3630	0.4407	0.3630	0.4383
C6	-0.2660	-0.4065	-0.2660	-0.3792
C7	0.0480	0.0657	0.0480	0.0587
H	0.1820	0.1574	0.1820	0.1549
C7M	-0.2640	-0.2360	-0.2640	-0.2446

HM7	0.0780	0.0498	0.0780	0.0563
HM7	0.0780	0.0442	0.0780	0.0491
HM7	0.0780	0.0458	0.0780	0.0398
C8	0.1860	0.0741	0.1860	0.0753
C9	-0.3040	-0.3033	-0.3040	-0.2862
H	0.1770	0.1671	0.1770	0.1669
C8M	-0.2710	-0.2230	-0.2710	-0.2467
HM8	0.0860	0.0563	0.0860	0.0573
HM8	0.0860	0.0545	0.0860	0.0615
HM8	0.0860	0.0507	0.0860	0.0620
C9A	-0.0070	-0.0956	-0.0070	-0.1308
N10	-0.0990	-0.0886	-0.0990	-0.0509
C1*	-0.0060	0.0190	-0.0060	-0.0248
H11	0.1040	0.0765	0.1040	0.0960
H12	0.1040	0.0875	0.1040	0.0875
C2*	0.2140	0.1732	0.2140	0.1855
O2*	-0.6790	-0.6790	-0.6790	-0.6790
H2'	0.0400	0.0400	0.0400	0.0400
HO2	0.4450	0.4450	0.4450	0.4450
C3*	0.1910	0.1910	0.1910	0.1910
O3*	-0.7890	-0.7890	-0.7890	-0.7890
H3'	0.0350	0.0350	0.0350	0.0350
HO3	0.5030	0.5030	0.5030	0.5030
C4*	0.0100	0.0100	0.0100	0.0100
O4*	-0.7270	-0.7270	-0.7270	-0.7270
H4'	0.1000	0.1000	0.1000	0.1000
HO4	0.5180	0.5180	0.5180	0.5180
C5*	0.3210	0.3210	0.3210	0.3210
H5'	0.0000	0.0000	0.0000	0.0000
H5'	0.0000	0.0000	0.0000	0.0000
O5*	-0.6750	-0.6750	-0.6750	-0.6750
P	1.5650	1.5650	1.5650	1.5650
O1P	-1.0090	-1.0090	-1.0090	-1.0090
O2P	-1.0090	-1.0090	-1.0090	-1.0090
O3P	-1.0090	-1.0090	-1.0090	-1.0090
SUM	-2.0050	-2.9807	-2.0050	-2.9647

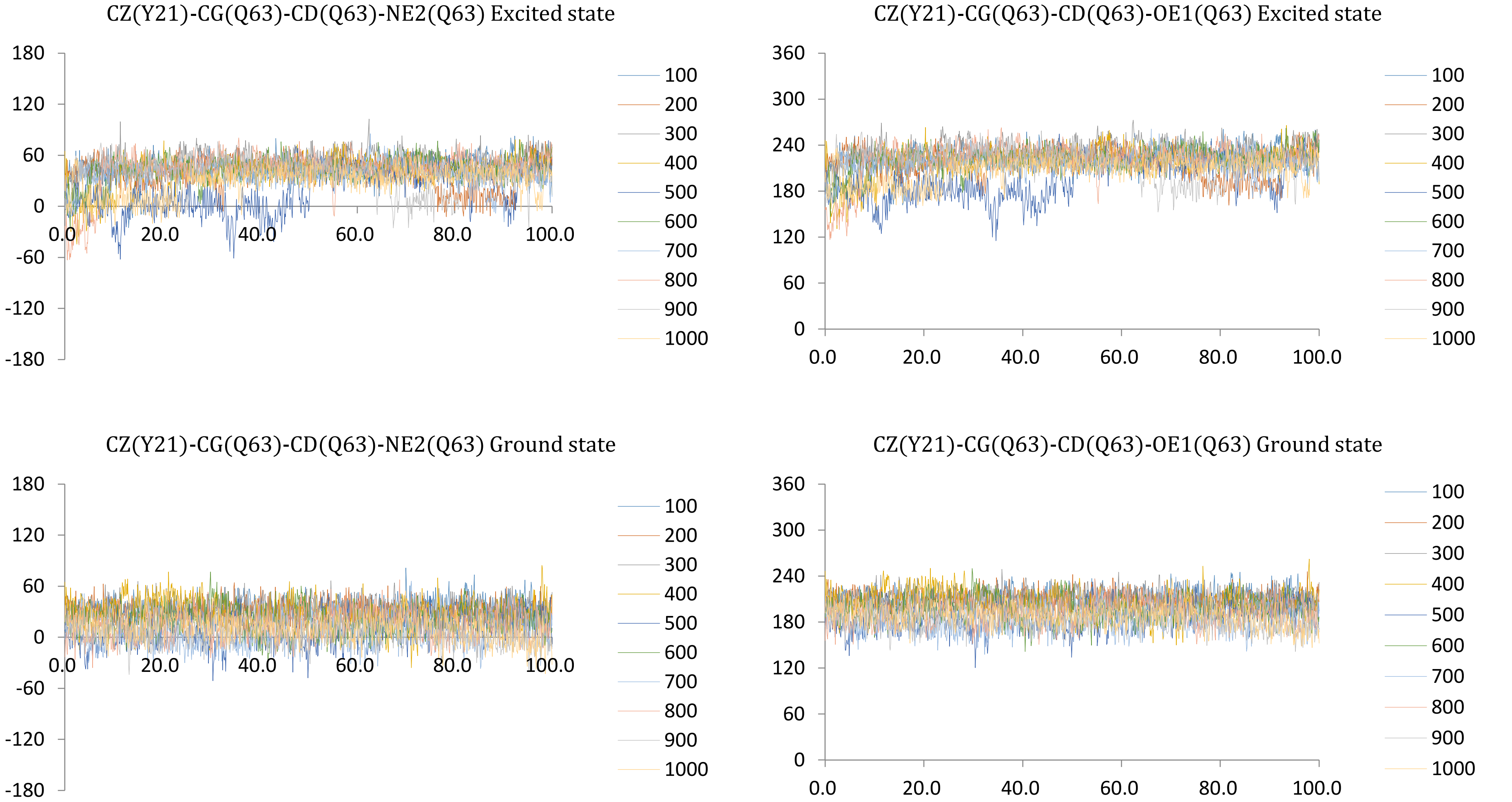


Figure 1 - Time evolution of the dihedral angle from the first trajectory, which characterise the orientation of the Gln63 side chain in the 1YRX (Trp-in structure) in the ground state and in the excited state. The X-axis shows the time in ps and Y-axis shows the degree.

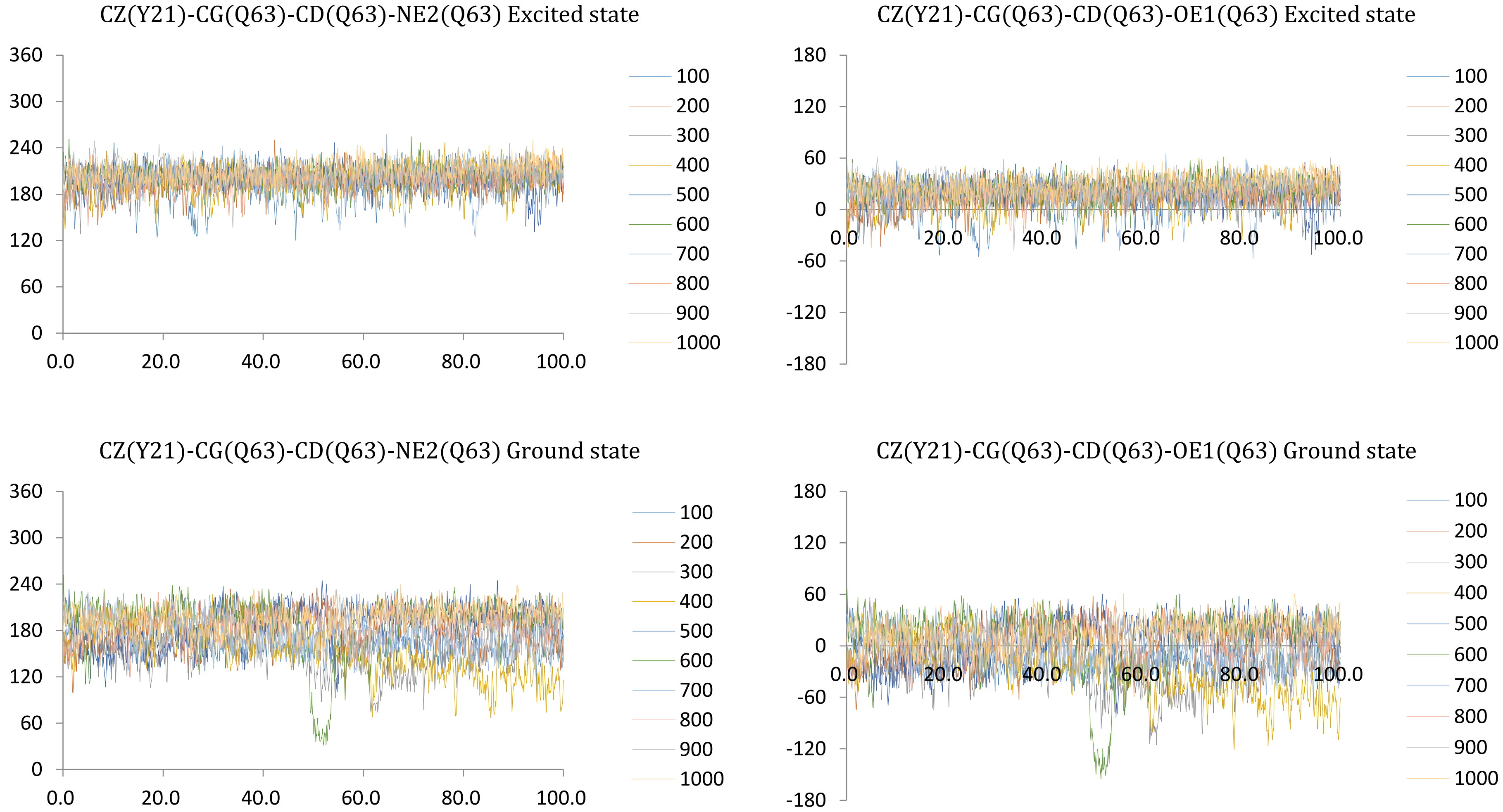
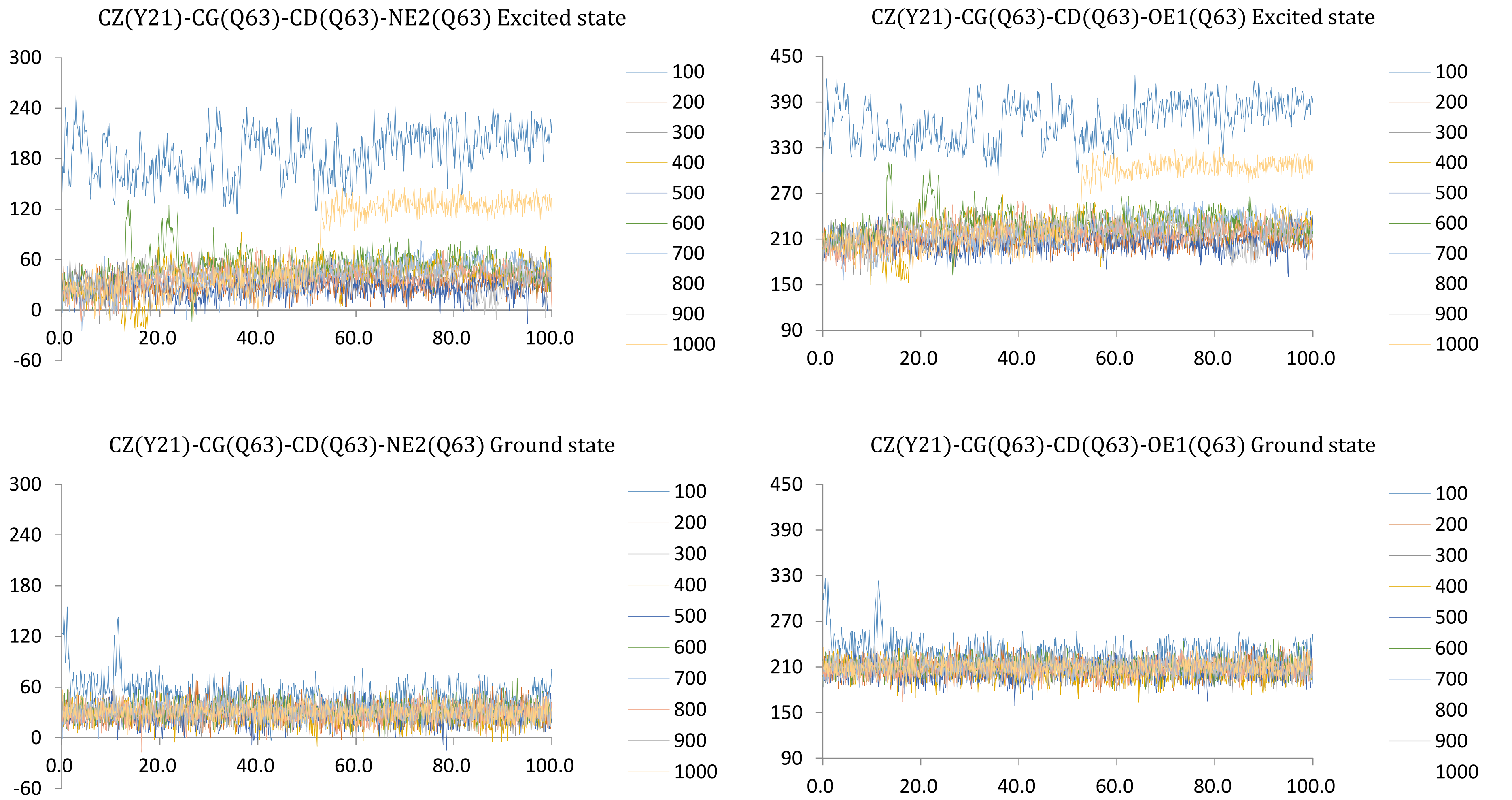


Figure 2 - Time evolution of the dihedral angle from the second trajectory, which characterise the orientation of the Gln63 side chain in the 1YRX (Trp-in structure) in the ground state and in the excited state. The X-axis shows the time in ps and Y-axis shows the degree.



CZ(Y21)-CG(Q63)-CD(Q63)-NE2(Q63) Excited state

CZ(Y21)-CG(Q63)-CD(Q63)-OE1(Q63) Excited state

CZ(Y21)-CG(Q63)-CD(Q63)-NE2(Q63) Ground state

CZ(Y21)-CG(Q63)-CD(Q63)-OE1(Q63) Ground state

Figure 3 - Time evolution of the dihedral angle from the third trajectory, which characterise the orientation of the Gln63 side chain in the 1YRX (Trp-in structure) in the ground state and in the excited state. The X-axis shows the time in ps and Y-axis shows the degree.

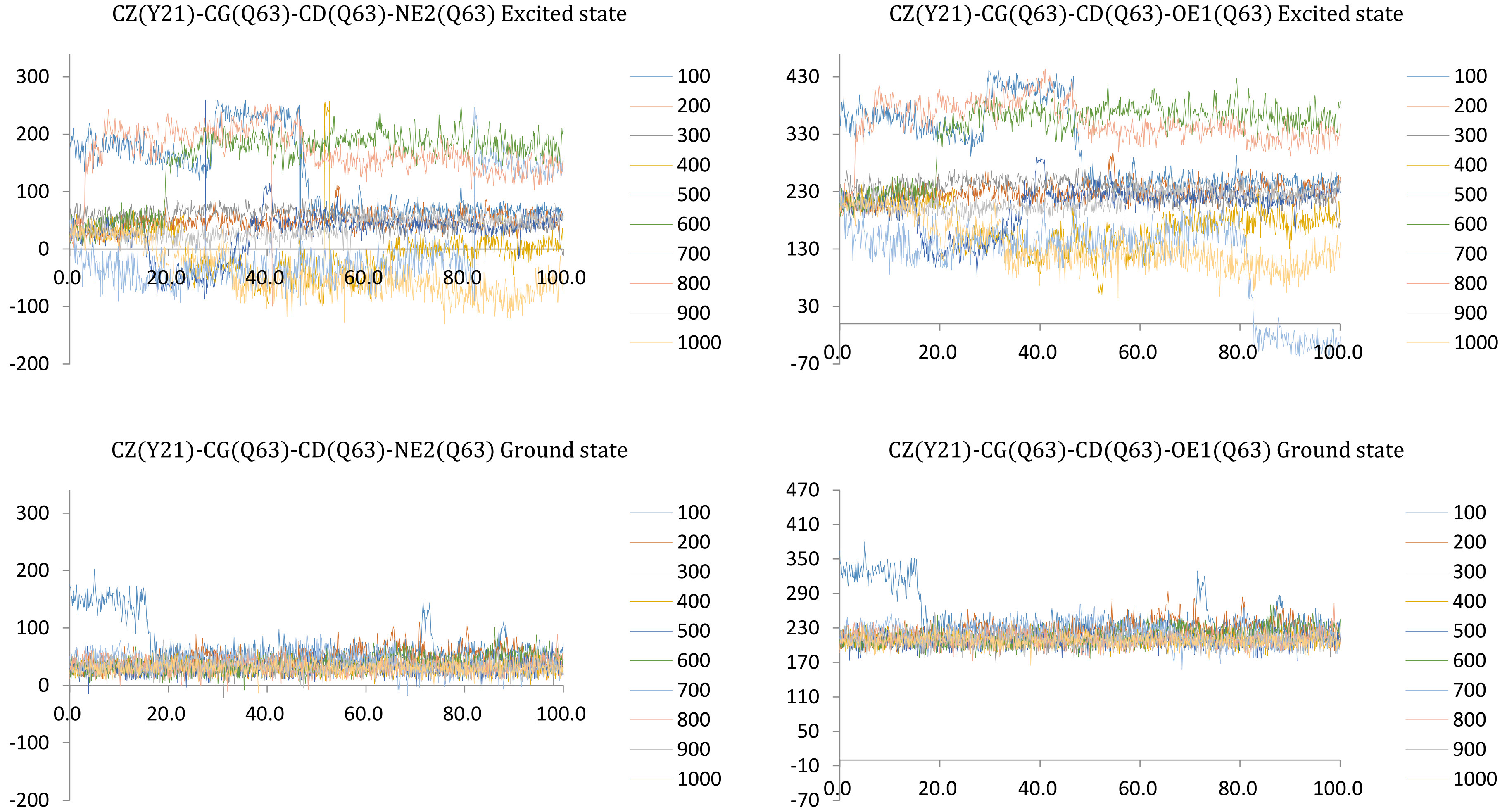


Figure 4 - Time evolution of the dihedral angle from the fourth trajectory, which characterise the orientation of the Gln63 side chain in the 1YRX (Trp-in structure) in the ground state and in the excited state. The X-axis shows the time in ps and Y-axis shows the degree.

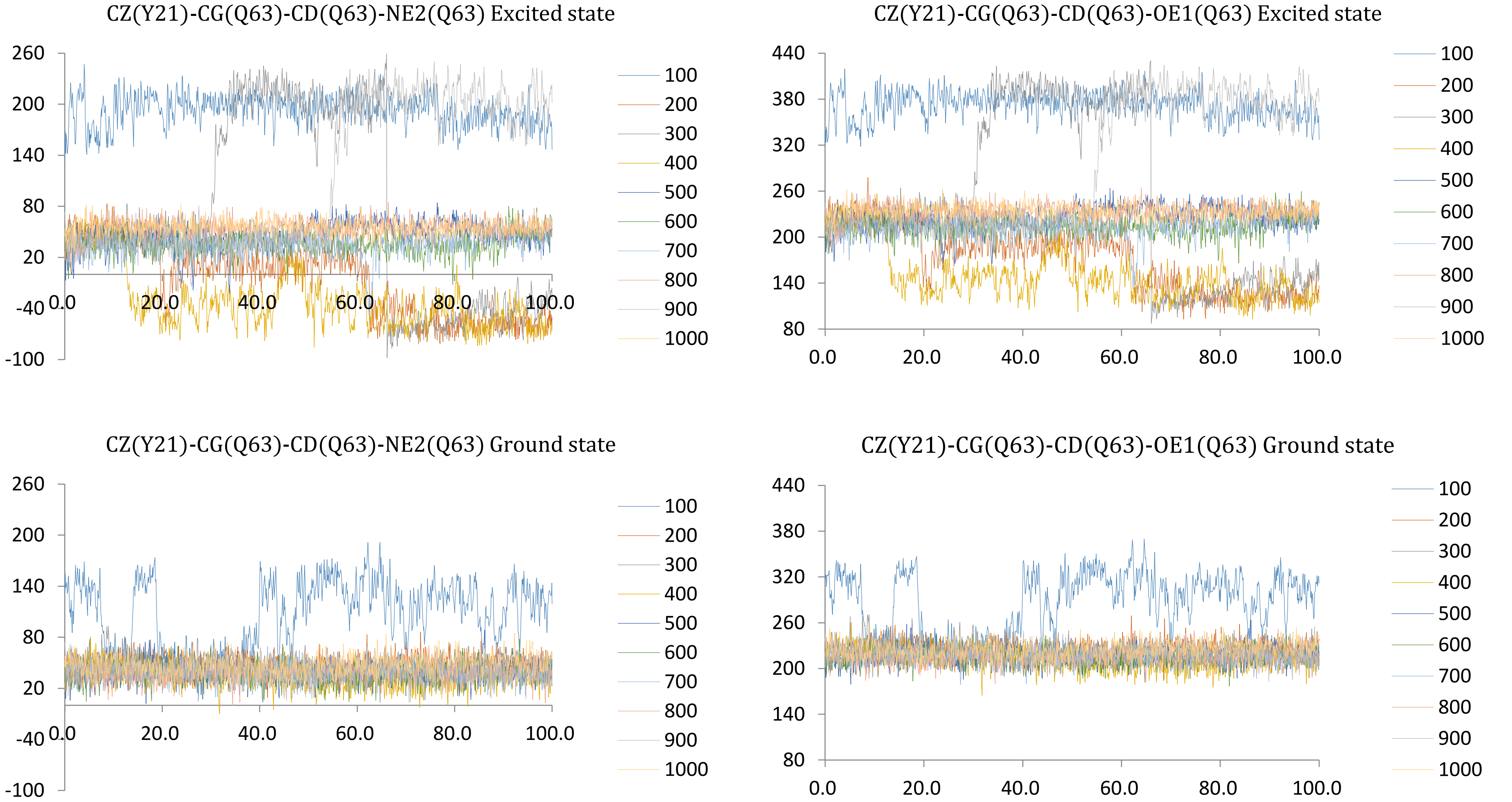


Figure 5 - Time evolution of the dihedral angle from the fifth trajectory, which characterise the orientation of the Gln63 side chain in the 1YRX (Trp-in structure) in the ground state and in the excited state. The X-axis shows the time in ps and Y-axis shows the degree.

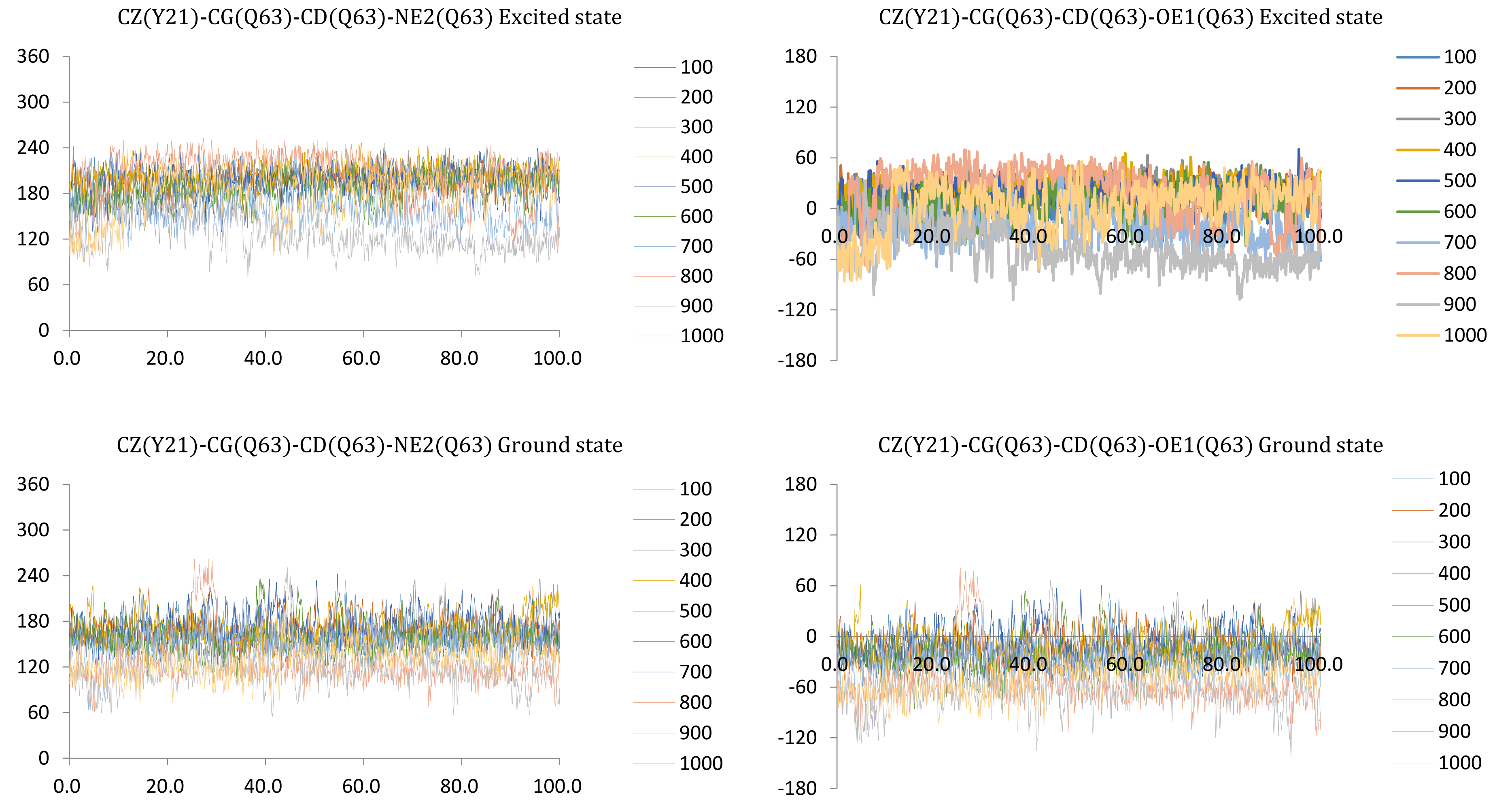
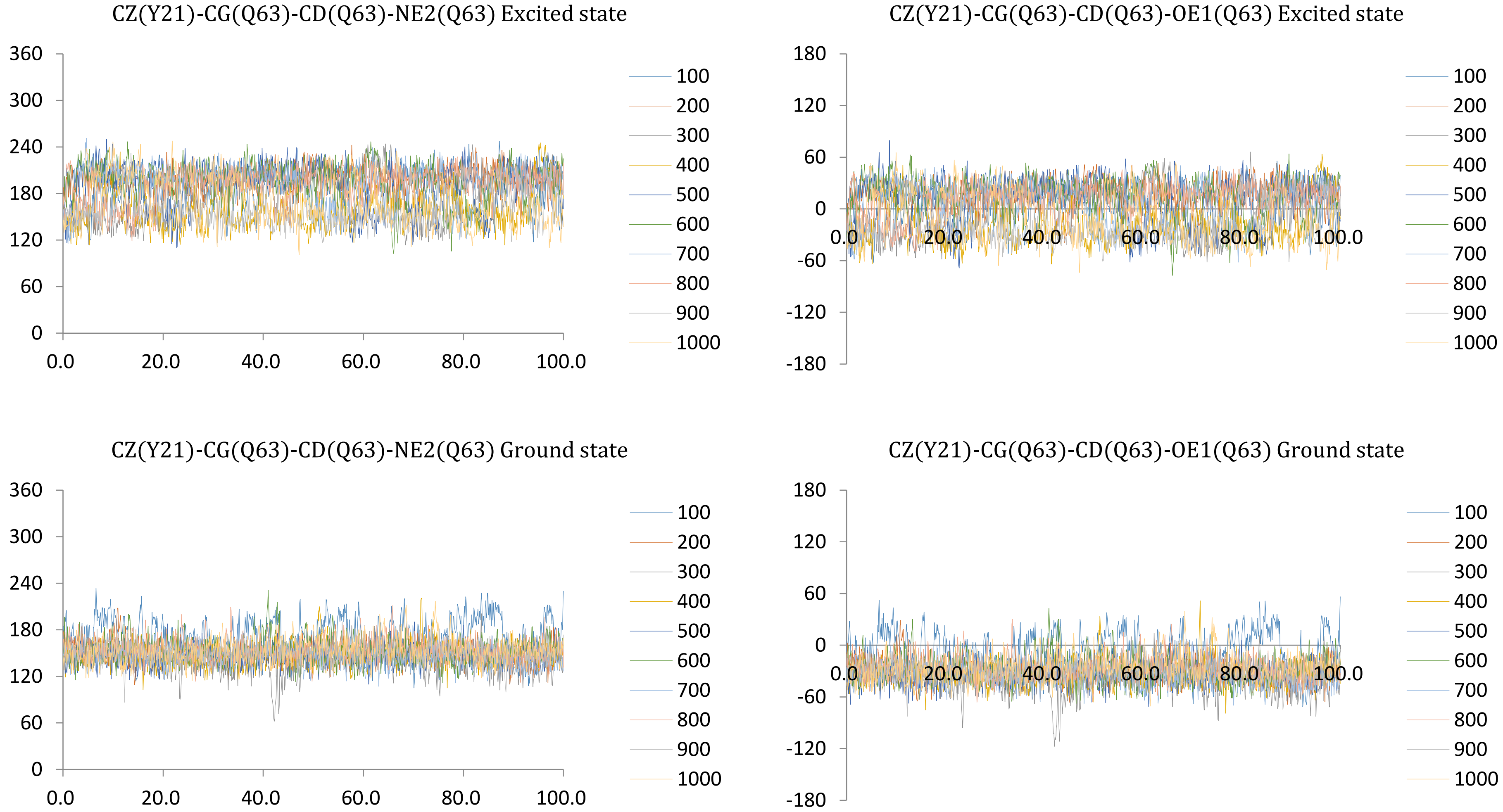


Figure 6 - Time evolution of the dihedral angle from the sixth trajectory, which characterise the orientation of the Gln63 side chain in the 1YRX (Trp-in structure) in the ground state and in the excited state. The X-axis shows the time in ps and Y-axis shows the degree.

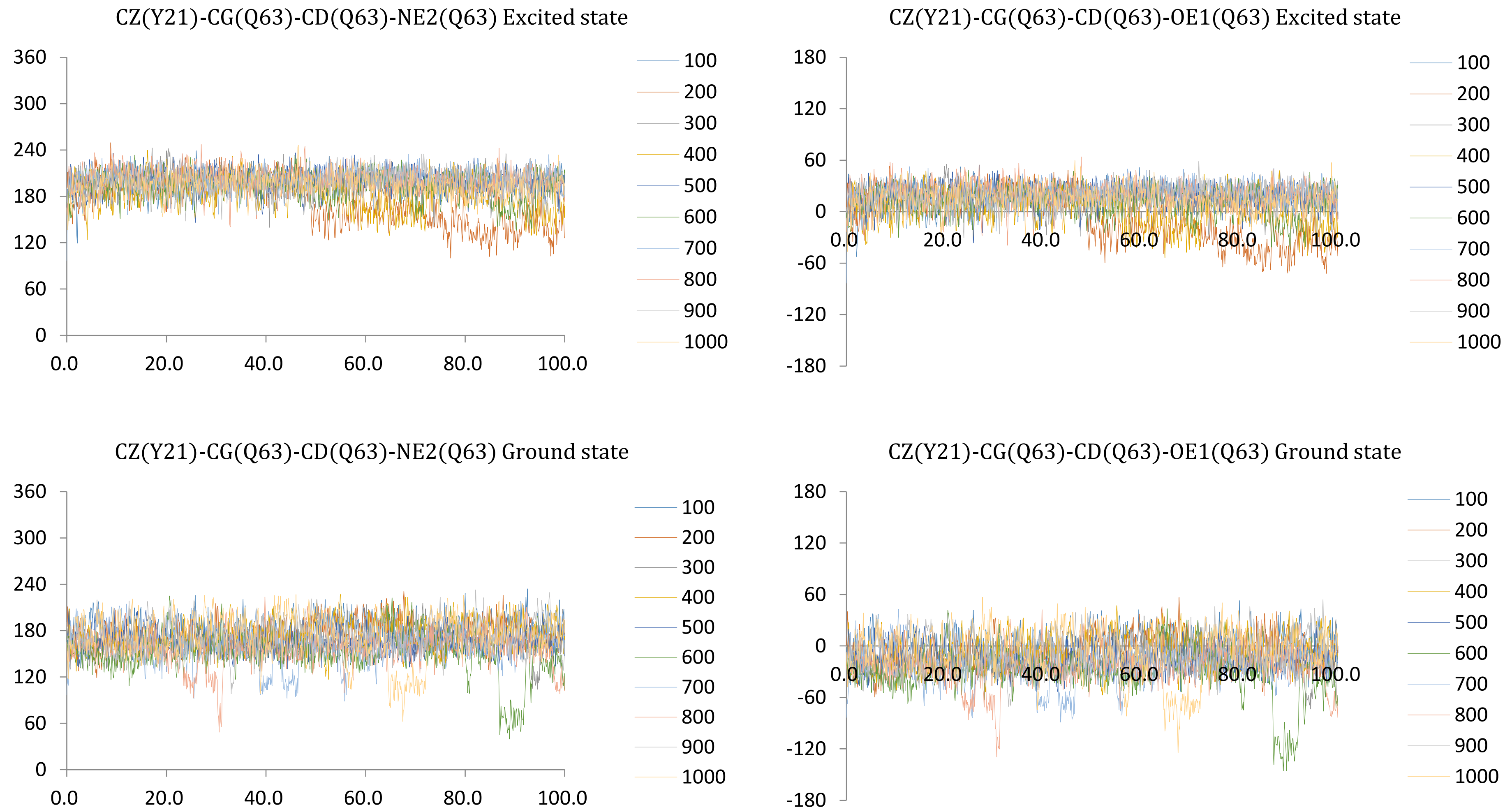

CZ(Y21)-CG(Q63)-CD(Q63)-NE2(Q63) Excited state

CZ(Y21)-CG(Q63)-CD(Q63)-OE1(Q63) Excited state

CZ(Y21)-CG(Q63)-CD(Q63)-NE2(Q63) Ground state

CZ(Y21)-CG(Q63)-CD(Q63)-OE1(Q63) Ground state

Figure 7 - Time evolution of the dihedral angle from the seventh trajectory, which characterise the orientation of the Gln63 side chain in the 1YRX (Trp-in structure) in the ground state and in the excited state. The X-axis shows the time in ps and Y-axis shows the degree.



CZ(Y21)-CG(Q63)-CD(Q63)-NE2(Q63) Excited state

CZ(Y21)-CG(Q63)-CD(Q63)-OE1(Q63) Excited state

CZ(Y21)-CG(Q63)-CD(Q63)-NE2(Q63) Ground state

CZ(Y21)-CG(Q63)-CD(Q63)-OE1(Q63) Ground state

Figure 8 - Time evolution of the dihedral angle from the eighth trajectory, which characterise the orientation of the Gln63 side chain in the 1YRX (Trp-in structure) in the ground state and in the excited state. The X-axis shows the time in ps and Y-axis shows the degree.

ES		switch time (ps)	avg distance before switching (nm)	avg distance after switching (nm)	Back switching (nm)
3	1000	55	0.499774	0.669862	
4	600	20	0.370991	0.303814	
	700	85	0.306844	0.29752	
	800	5	0.323879	0.302298	
5	300	35,65	0.383846	0.327932	0.355653
	900	55	0.389436	0.337309	

Table 6 - Average distance between Q63(NE2) and FMN(N5) before and after a switch is occurred.

ES		switch time (ps)	avg distance before switching (nm)	avg distance after switching (nm)	Back switching (nm)
3	1000	55	0.303034	0.481254	
4	600	20	0.286649	0.313888	
	700	85	0.328016	0.332979	
	800	5	0.297015	0.318334	
5	300	35,65	0.294771	0.304491	0.315038
	900	55	0.29051	0.303333	

Table 7 - Average distance between Q63(NE2) and FMN(O4) before and after a switch is occurred.

ES		switch time (ps)	avg distance before switching (nm)	avg distance after switching (nm)	Back switching (nm)
3	1000	55	0.613805	0.607003	
4	600	20	0.408162	0.454528	
	700	85	0.469891	0.411732	
	800	5	0.431713	0.43768	
5	300	35,65	0.356078	0.421399	0.331078
	900	55	0.568704	0.429289	

Table 8 - Average distance between Q63(NE2) and Y21(OH) before and after a switch is occurred.

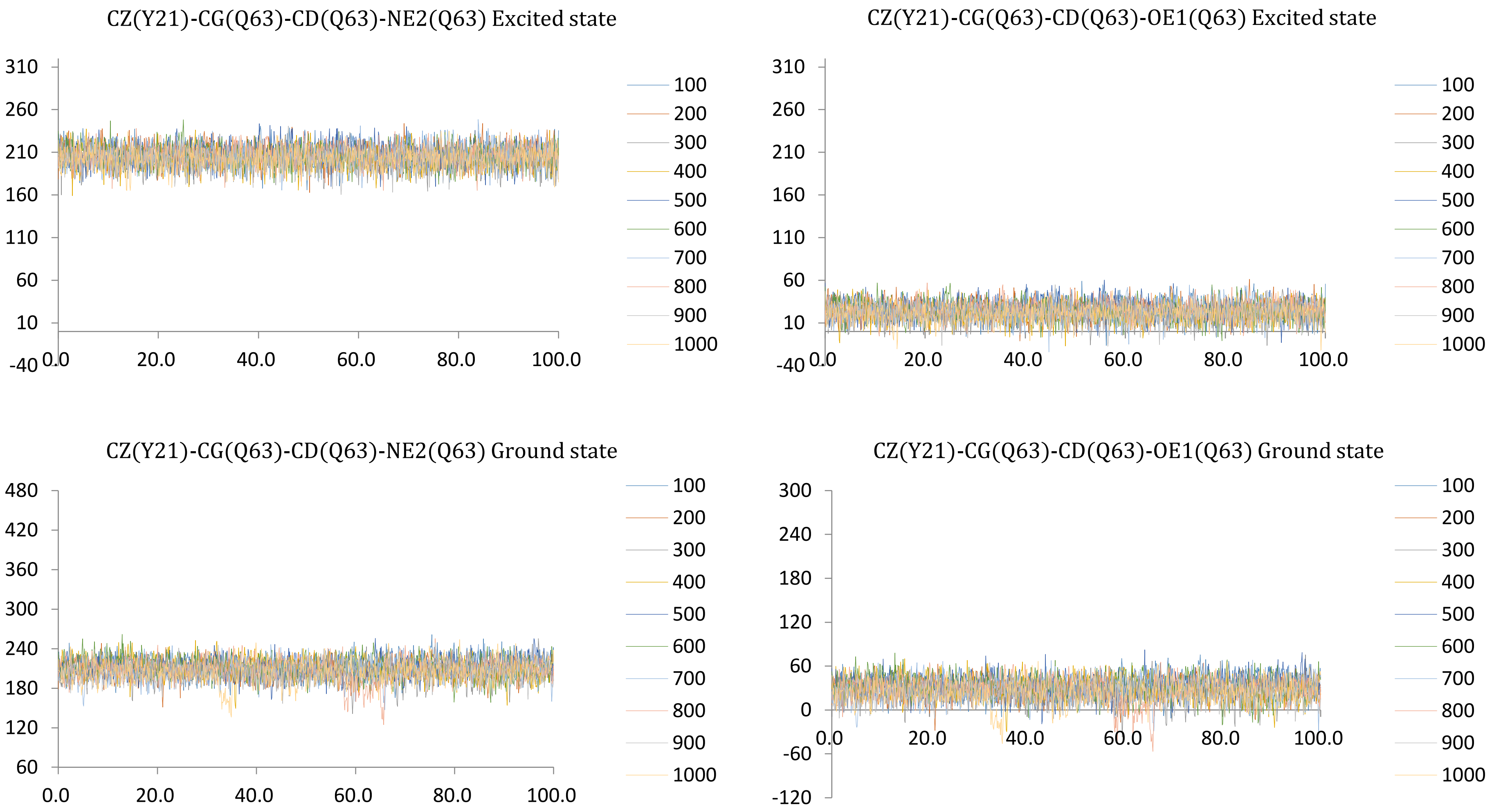


Figure 9 - Time evolution of the dihedral angle from the first trajectory, which characterise the orientation of the Gln63 side chain in the 2IYG (Trp-out structure) using the AmberGS charges (ground state) and a set of relocated charges (mimicking the excited state). The X-axis shows the time in ps and Y-axis shows the degree.

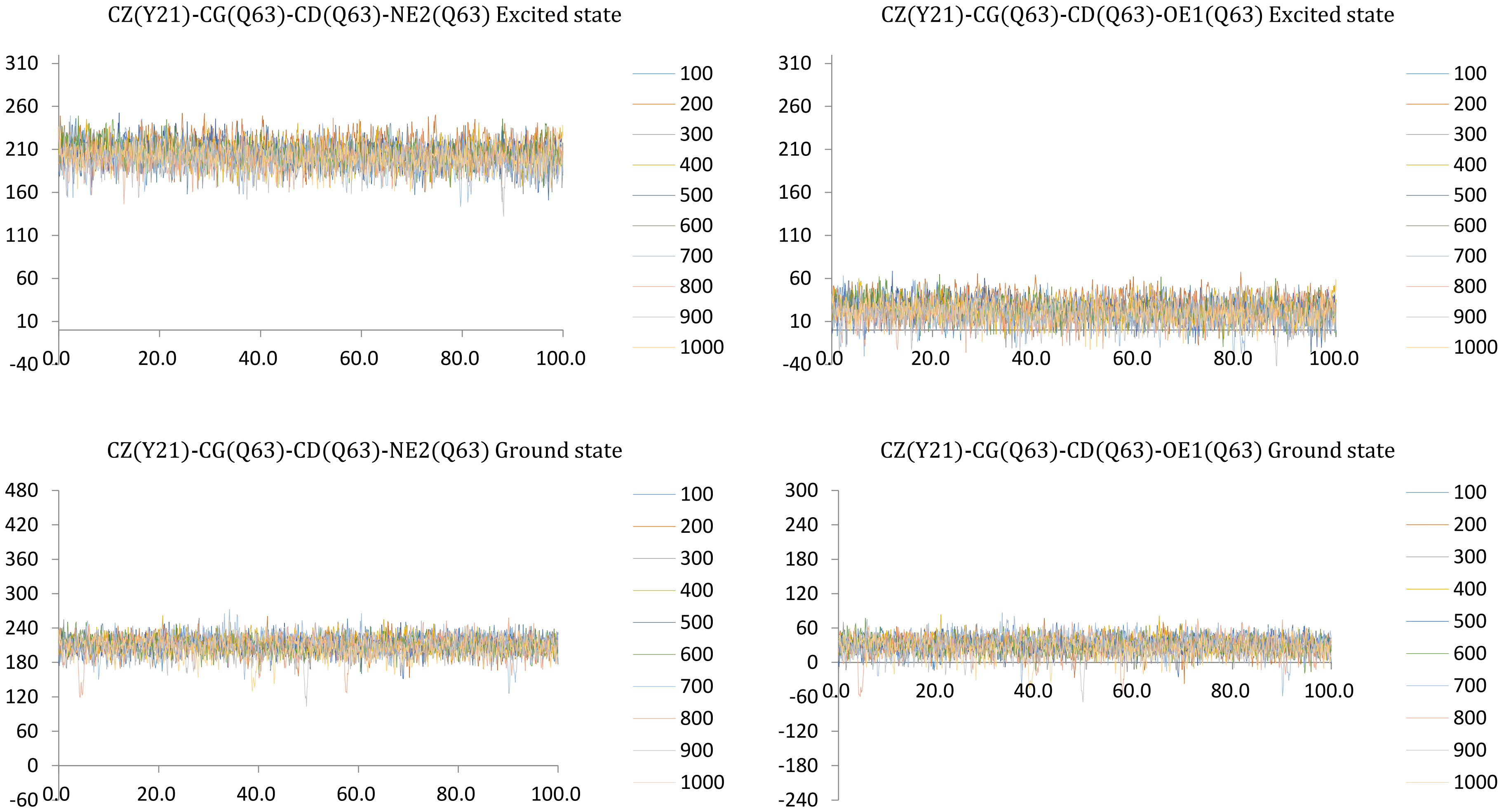


Figure 10 - Time evolution of the dihedral angle from the second trajectory, which characterise the orientation of the Gln63 side chain in the 2IYG (Trp-out structure) using the AmberGS charges (ground state) and a set of relocated charges (mimicking the excited state). The X-axis shows the time in ps and Y-axis shows the degree.

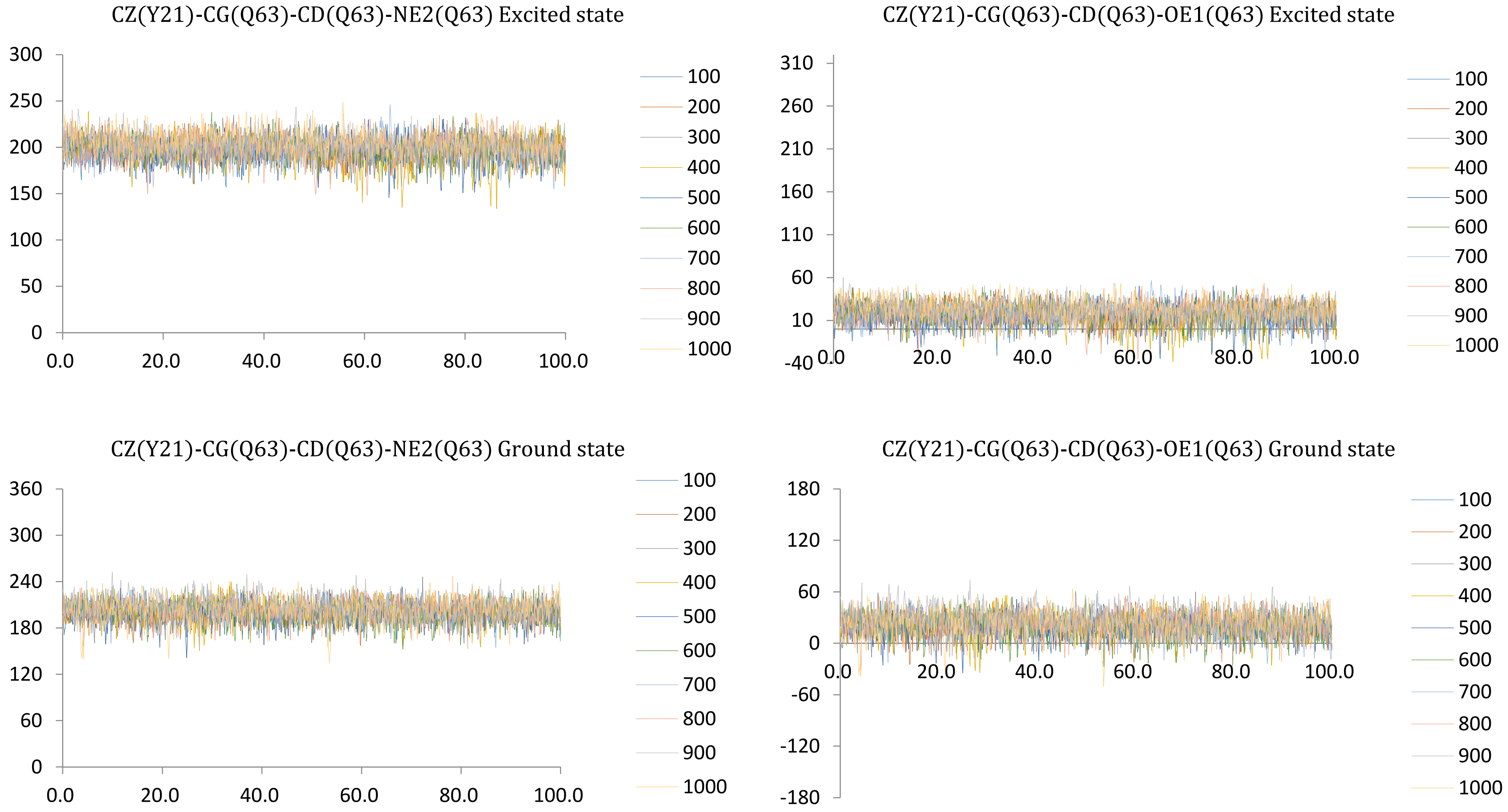


Figure 11 - Time evolution of the dihedral angle from the third trajectory, which characterise the orientation of the Gln63 side chain in the 2IYG (Trp-out structure) using the AmberGS charges (ground state) and a set of relocated charges (mimicking the excited state). The X-axis shows the time in ps and Y-axis shows the degree.

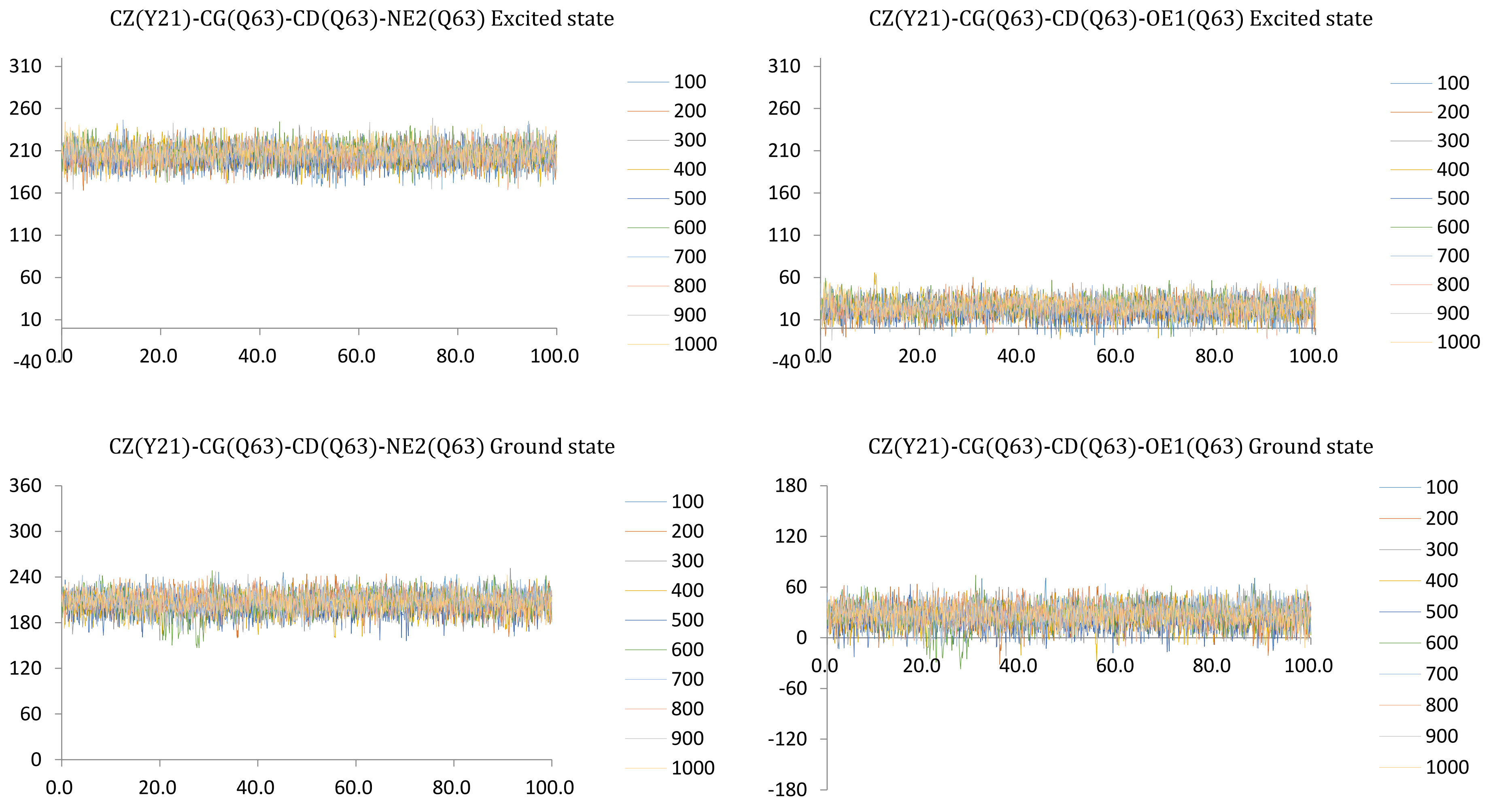
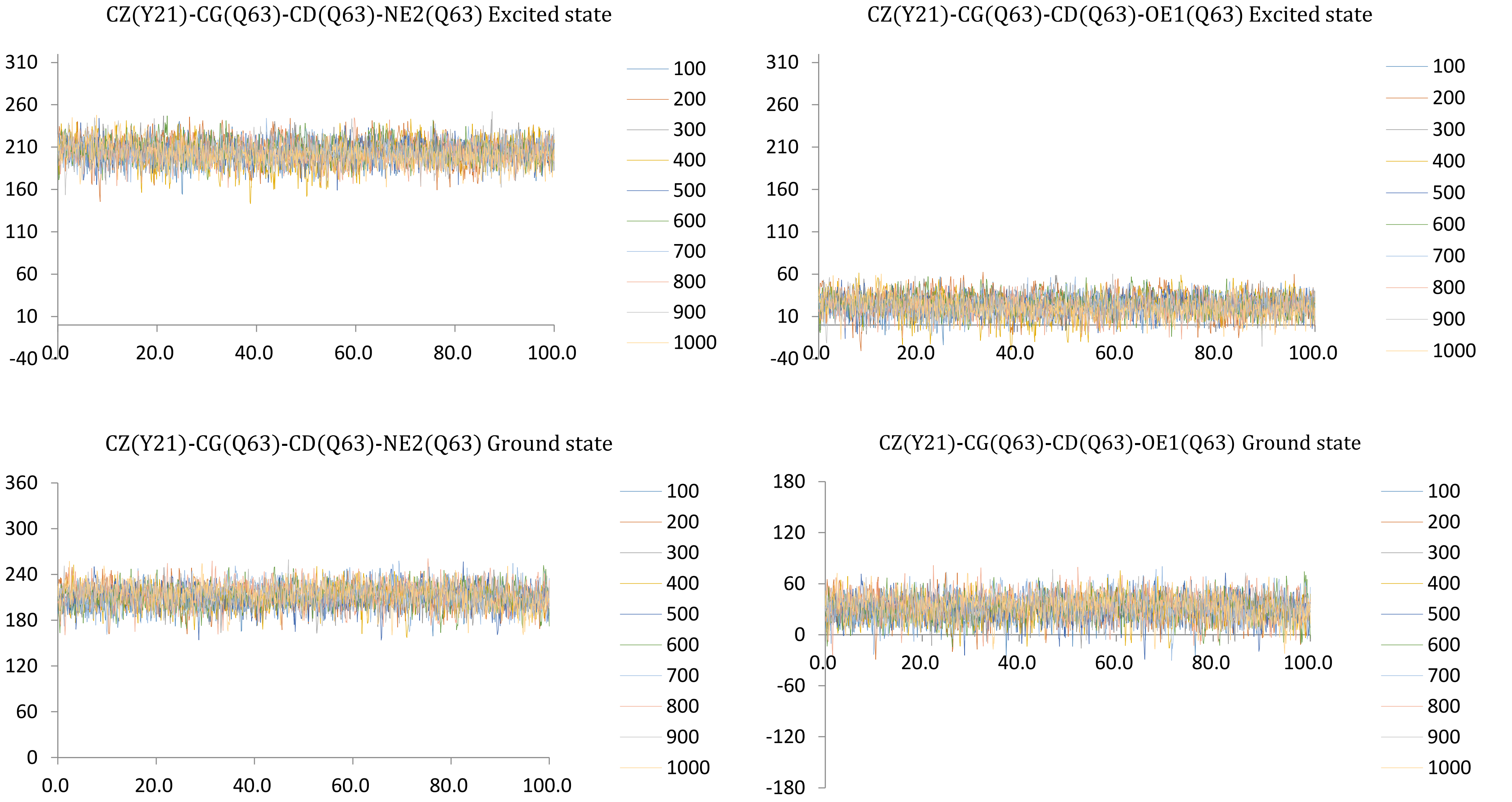


Figure 12 - Time evolution of the dihedral angle from the fourth trajectory, which characterise the orientation of the Gln63 side chain in the 2IYG (Trp-out structure) using the AmberGS charges (ground state) and a set of relocated charges (mimicking the excited state). The X-axis shows the time in ps and Y-axis shows the degree.

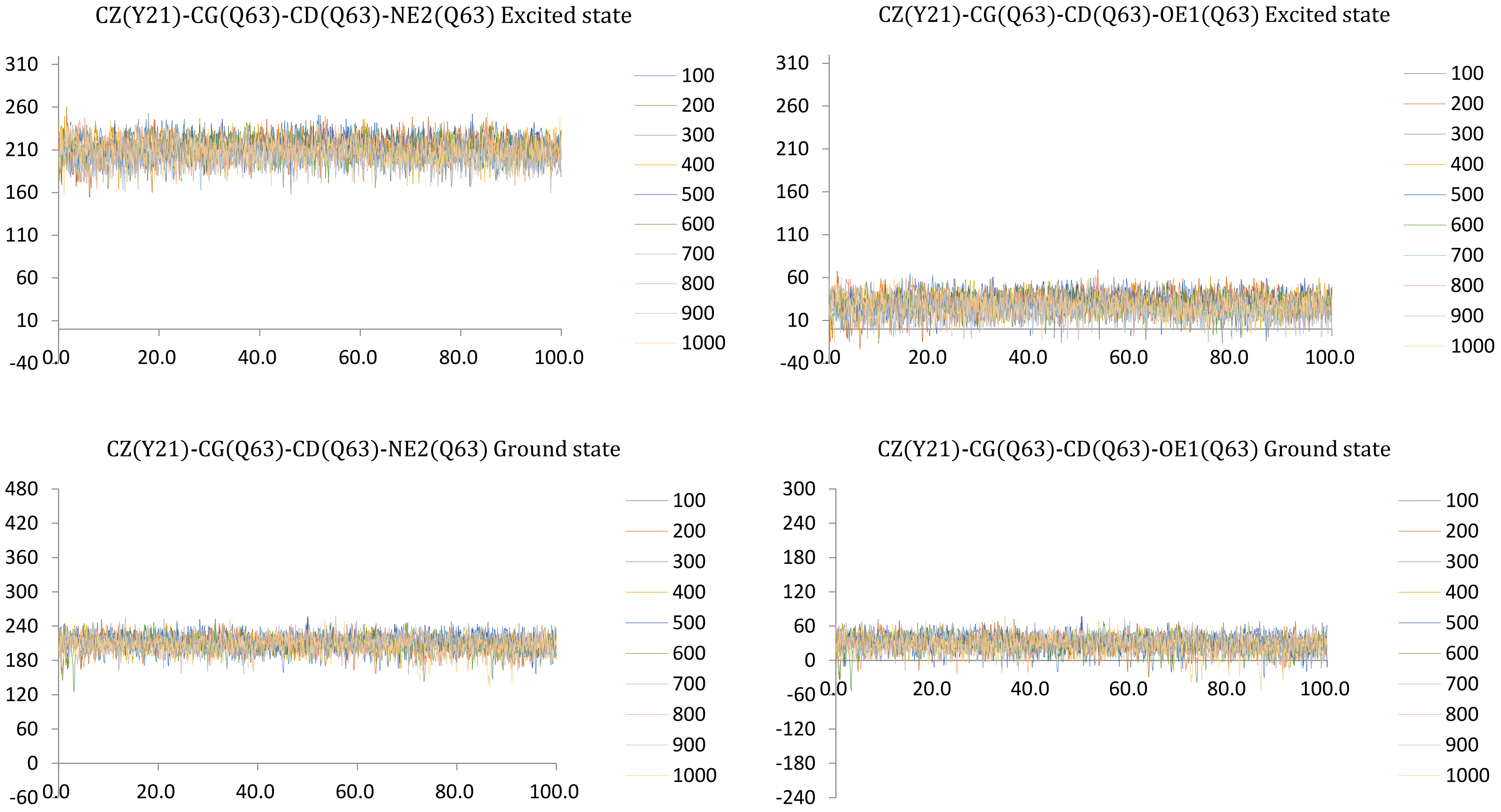

CZ(Y21)-CG(Q63)-CD(Q63)-NE2(Q63) Excited state

CZ(Y21)-CG(Q63)-CD(Q63)-OE1(Q63) Excited state

CZ(Y21)-CG(Q63)-CD(Q63)-NE2(Q63) Ground state

CZ(Y21)-CG(Q63)-CD(Q63)-OE1(Q63) Ground state

Figure 13 - Time evolution of the dihedral angle from the fifth trajectory, which characterise the orientation of the Gln63 side chain in the 2IYG (Trp-out structure) using the AmberGS charges (ground state) and a set of relocated charges (mimicking the excited state). The X-axis shows the time in ps and Y-axis shows the degree.


CZ(Y21)-CG(Q63)-CD(Q63)-NE2(Q63) Excited state

CZ(Y21)-CG(Q63)-CD(Q63)-OE1(Q63) Excited state

CZ(Y21)-CG(Q63)-CD(Q63)-NE2(Q63) Ground state

CZ(Y21)-CG(Q63)-CD(Q63)-OE1(Q63) Ground state

Figure 14 - Time evolution of the dihedral angle from the sixth trajectory, which characterise the orientation of the Gln63 side chain in the 2IYG (Trp-out structure) using the AmberGS charges (ground state) and a set of relocated charges (mimicking the excited state). The X-axis shows the time in ps and Y-axis shows the degree.

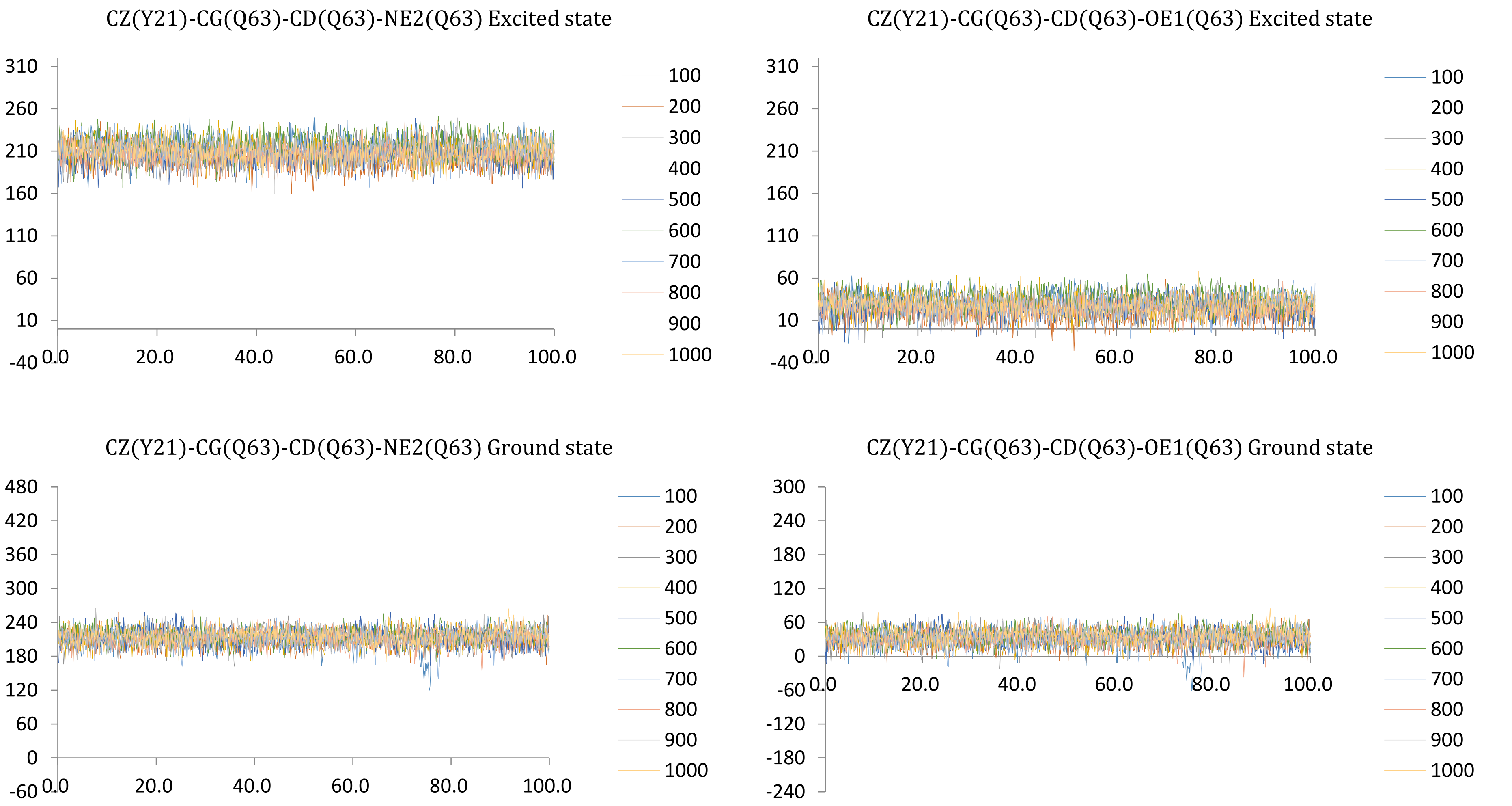
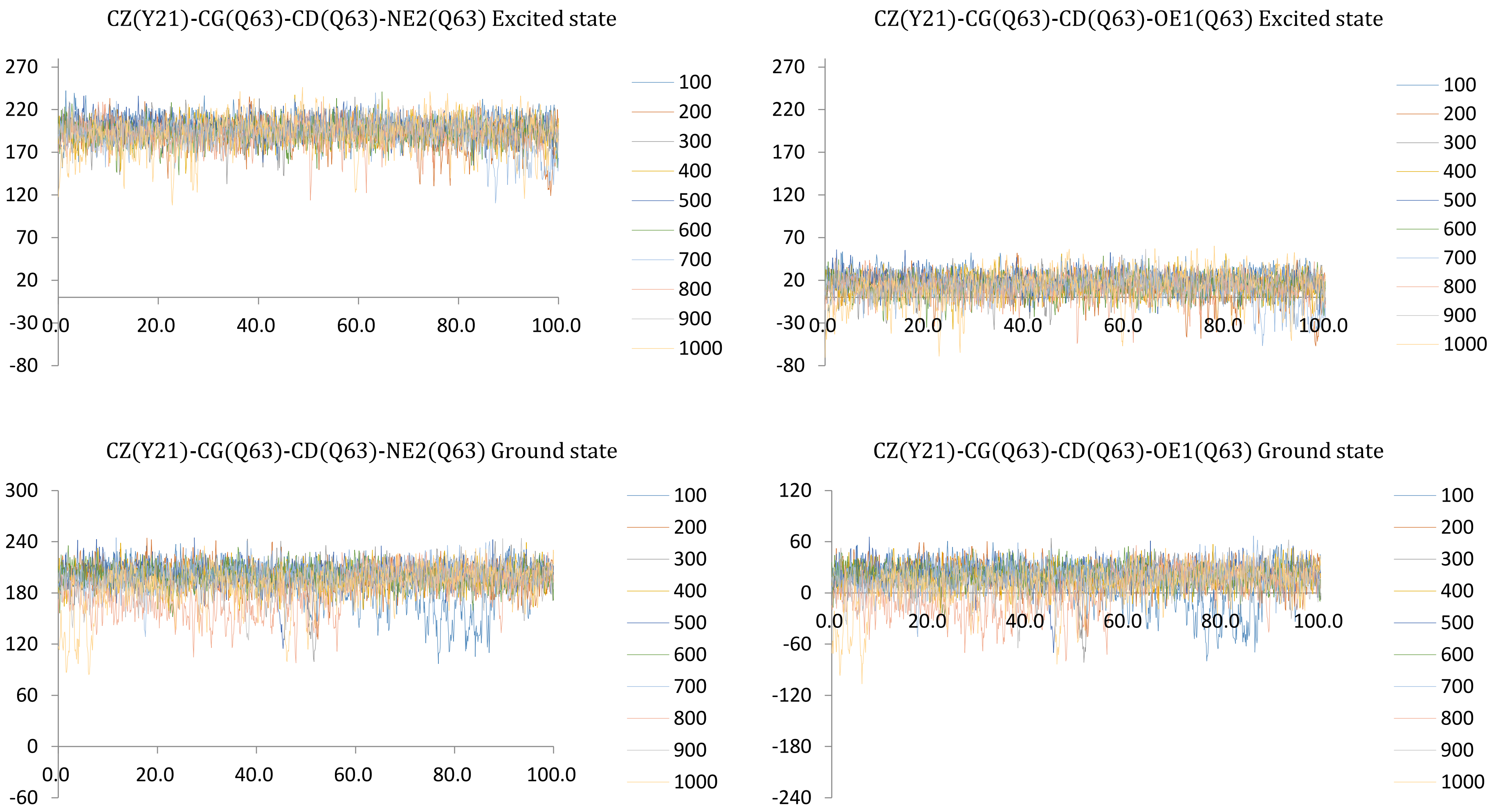


Figure 15 - Time evolution of the dihedral angle from the seventh trajectory, which characterise the orientation of the Gln63 side chain in the 2IYG (Trp-out structure) using the AmberGS charges (ground state) and a set of relocated charges (mimicking the excited state). The X-axis shows the time in ps and Y-axis shows the degree.

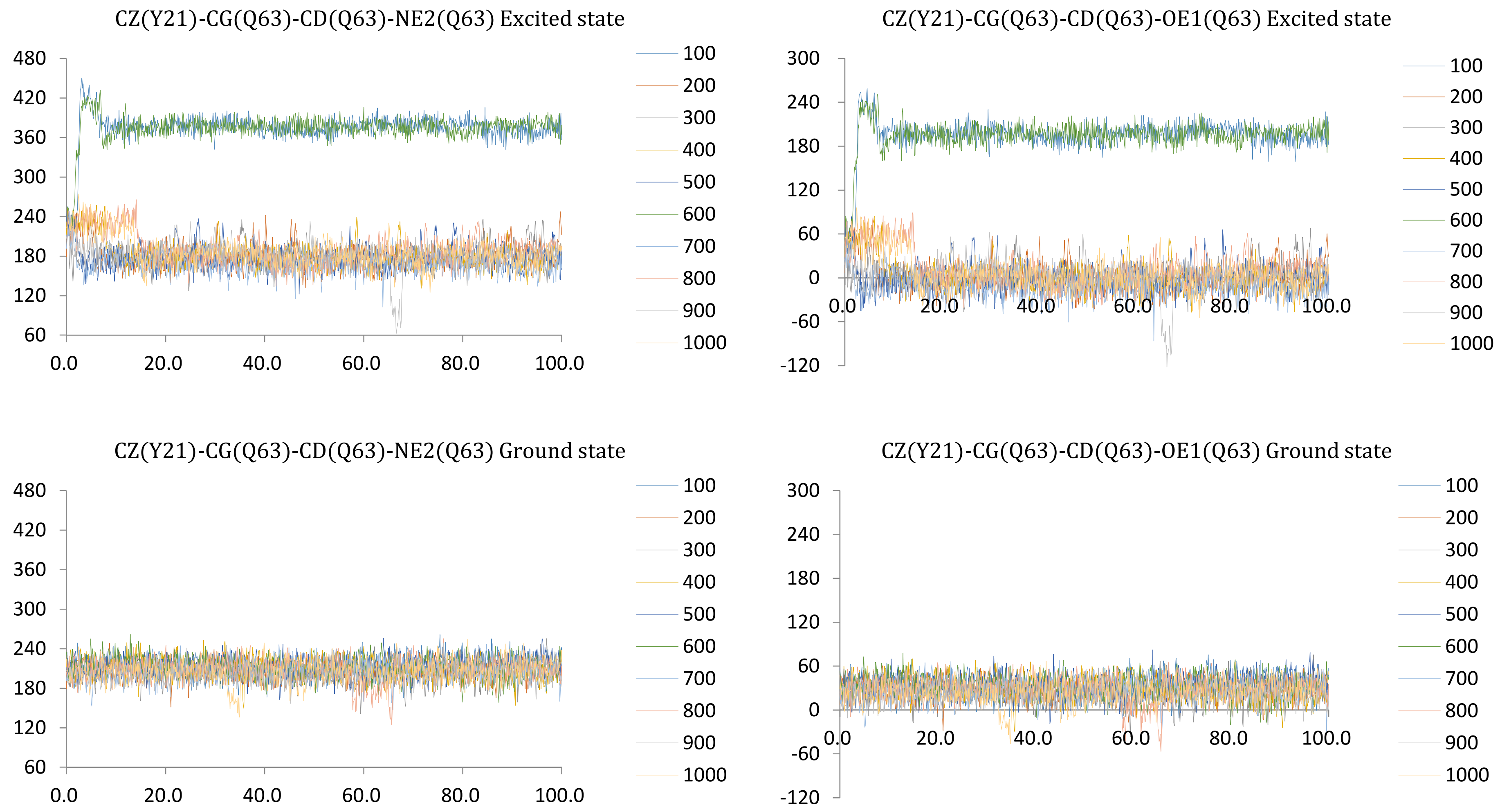

CZ(Y21)-CG(Q63)-CD(Q63)-NE2(Q63) Excited state

CZ(Y21)-CG(Q63)-CD(Q63)-OE1(Q63) Excited state

CZ(Y21)-CG(Q63)-CD(Q63)-NE2(Q63) Ground state

CZ(Y21)-CG(Q63)-CD(Q63)-OE1(Q63) Ground state

Figure 16 - Time evolution of the dihedral angle from the eighth trajectory, which characterise the orientation of the Gln63 side chain in the 2IYG (Trp-out structure) using the AmberGS charges (ground state) and a set of relocated charges (mimicking the excited state). The X-axis shows the time in ps and Y-axis shows the degree.


CZ(Y21)-CG(Q63)-CD(Q63)-NE2(Q63) Excited state

CZ(Y21)-CG(Q63)-CD(Q63)-OE1(Q63) Excited state

CZ(Y21)-CG(Q63)-CD(Q63)-NE2(Q63) Ground state

CZ(Y21)-CG(Q63)-CD(Q63)-OE1(Q63) Ground state

Figure 17 - Time evolution of the dihedral angle from the first trajectory, which characterise the orientation of the Gln63 side chain in the 2IYG (Trp-out structure) using the AmberGS charges (ground state) and a set of artificial charges. The X-axis shows the time in ps and Y-axis shows the degree.

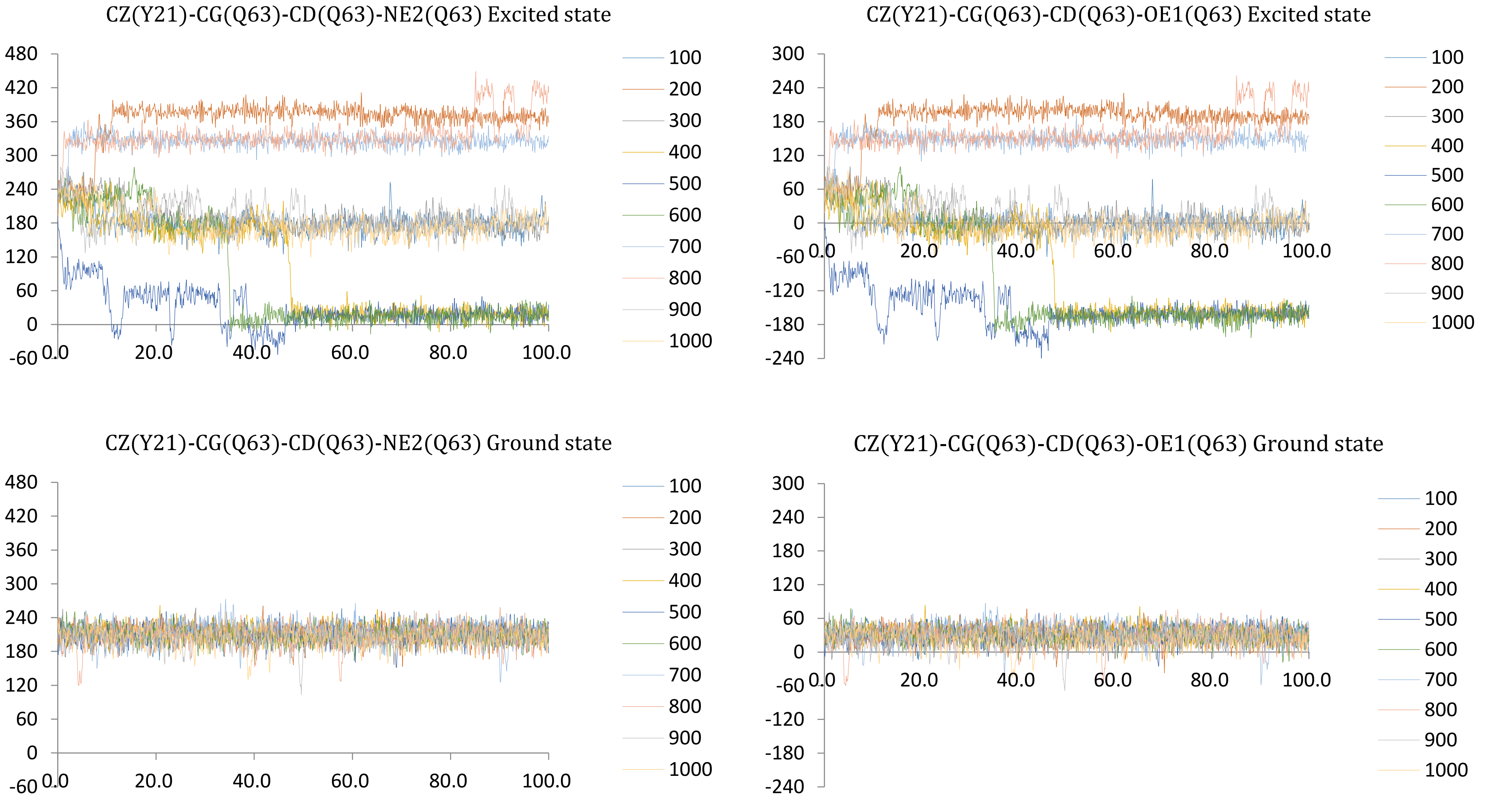
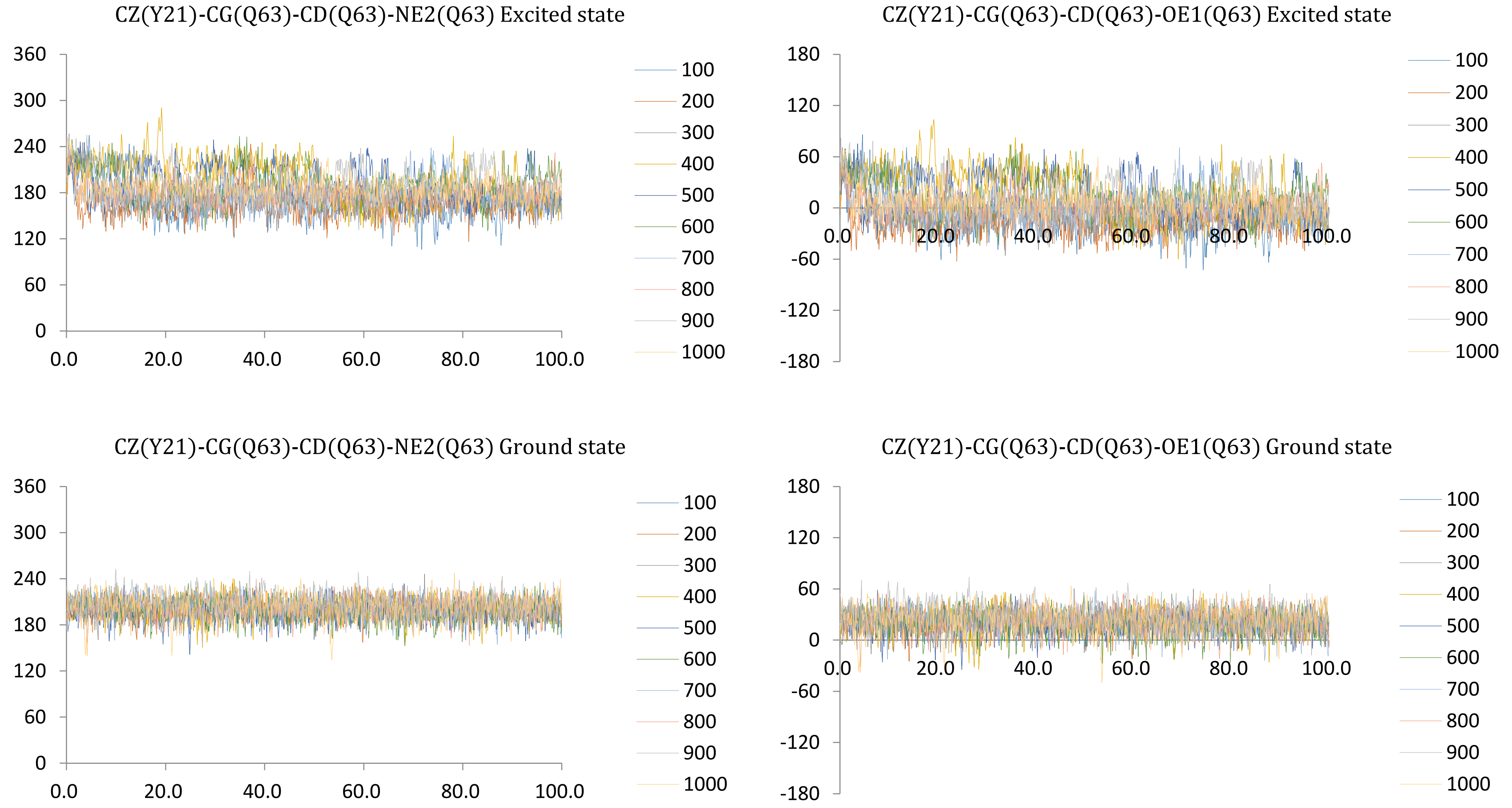


Figure 18 - Time evolution of the dihedral angle from the second trajectory, which characterise the orientation of the Gln63 side chain in the 2IYG (Trp-out structure) using the AmberGS charges (ground state) and a set of artificial charges. The X-axis shows the time in ps and Y-axis shows the degree.



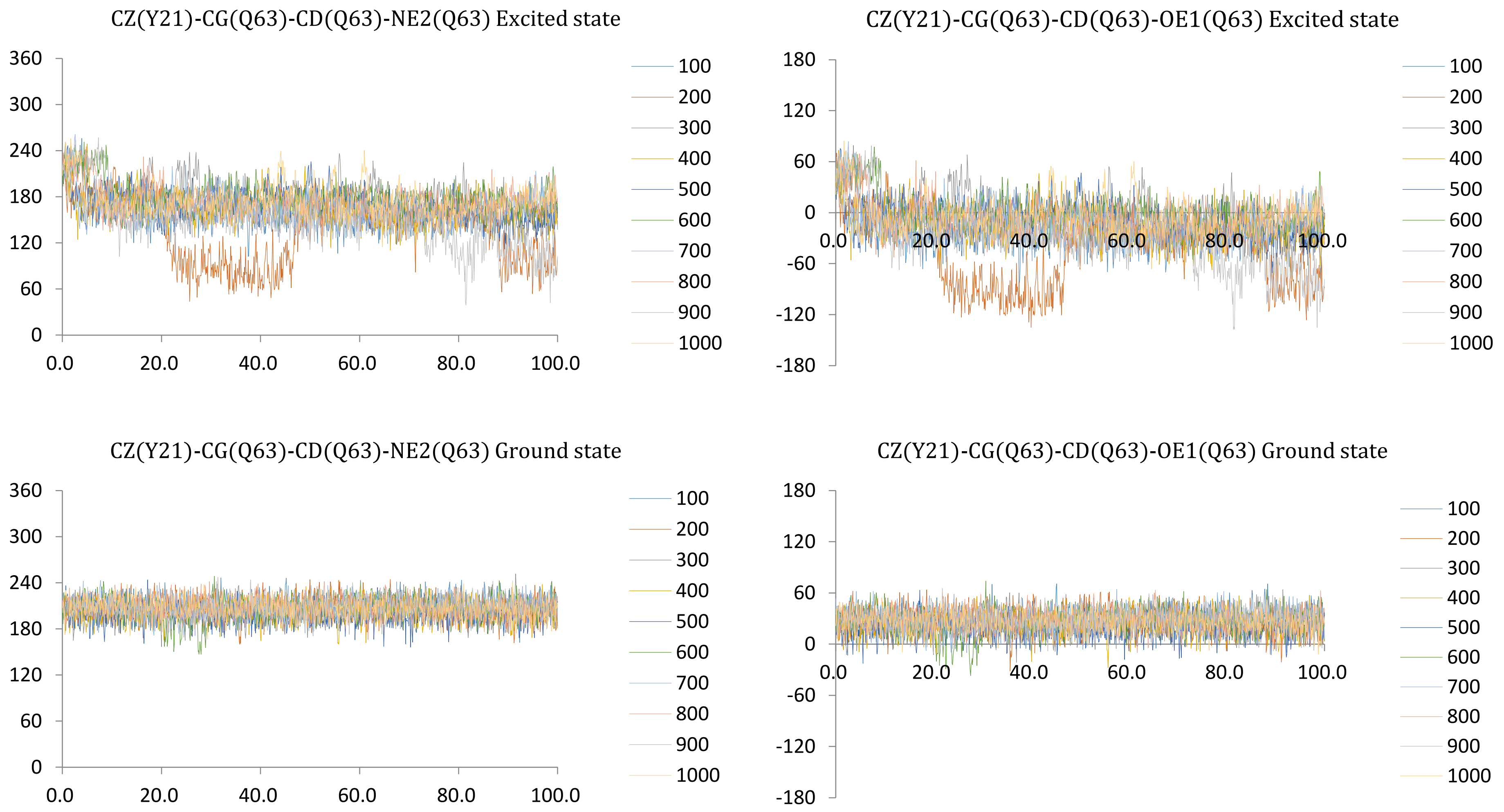
CZ(Y21)-CG(Q63)-CD(Q63)-NE2(Q63) Excited state

CZ(Y21)-CG(Q63)-CD(Q63)-OE1(Q63) Excited state

CZ(Y21)-CG(Q63)-CD(Q63)-NE2(Q63) Ground state

CZ(Y21)-CG(Q63)-CD(Q63)-OE1(Q63) Ground state

Figure 19 - Time evolution of the dihedral angle from the third trajectory, which characterise the orientation of the Gln63 side chain in the 2IYG (Trp-out structure) using the AmberGS charges (ground state) and a set of artificial charges. The X-axis shows the time in ps and Y-axis shows the degree.

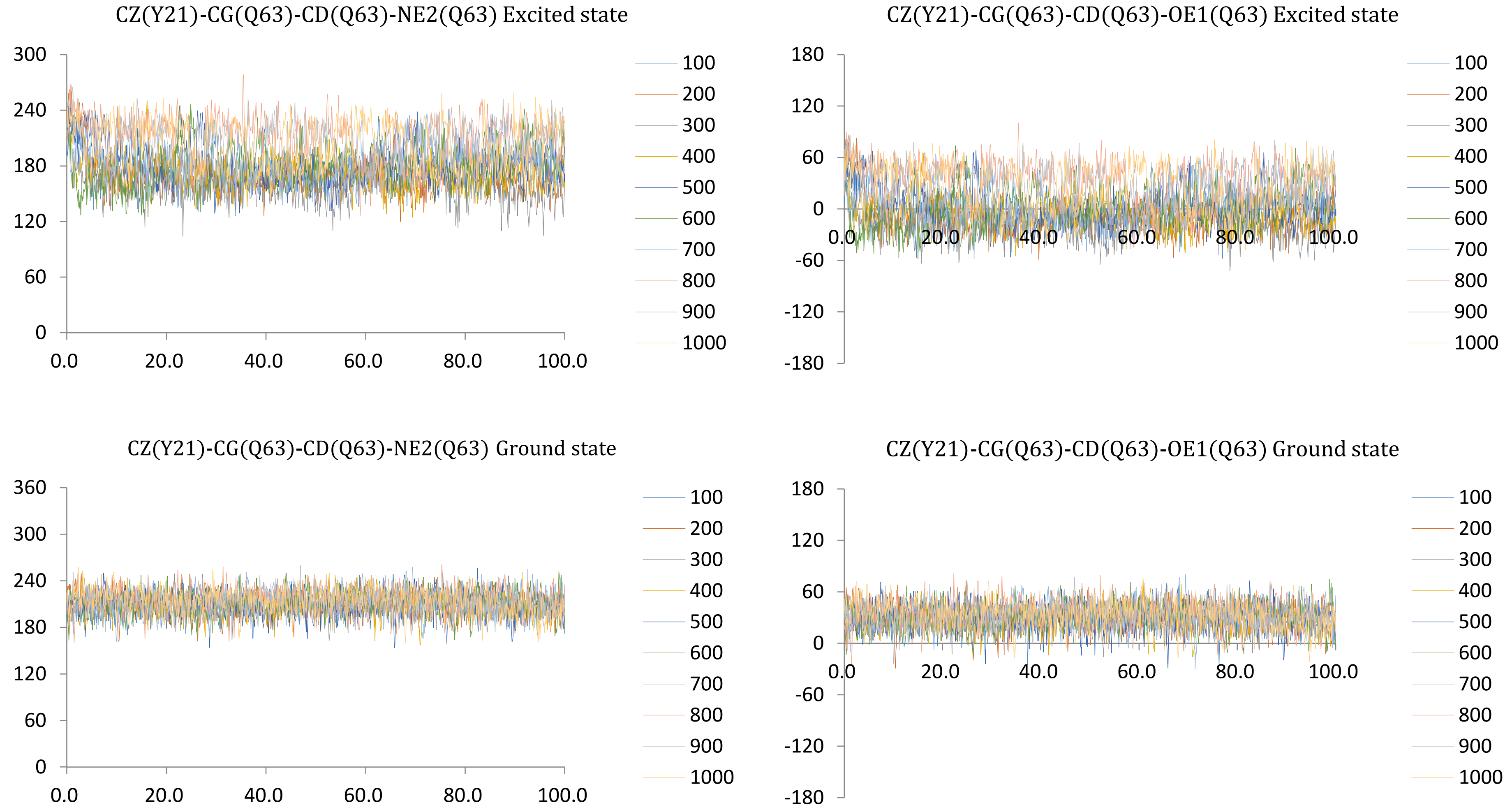

CZ(Y21)-CG(Q63)-CD(Q63)-NE2(Q63) Excited state

CZ(Y21)-CG(Q63)-CD(Q63)-OE1(Q63) Excited state

CZ(Y21)-CG(Q63)-CD(Q63)-NE2(Q63) Ground state

CZ(Y21)-CG(Q63)-CD(Q63)-OE1(Q63) Ground state

Figure 20 - Time evolution of the dihedral angle from the fourth trajectory, which characterise the orientation of the Gln63 side chain in the 2IYG (Trp-out structure) using the AmberGS charges (ground state) and a set of artificial charges. The X-axis shows the time in ps and Y-axis shows the degree.


CZ(Y21)-CG(Q63)-CD(Q63)-NE2(Q63) Excited state

CZ(Y21)-CG(Q63)-CD(Q63)-OE1(Q63) Excited state

CZ(Y21)-CG(Q63)-CD(Q63)-NE2(Q63) Ground state

CZ(Y21)-CG(Q63)-CD(Q63)-OE1(Q63) Ground state

Figure 21 - Time evolution of the dihedral angle from the fifth trajectory, which characterise the orientation of the Gln63 side chain in the 2IYG (Trp-out structure) using the AmberGS charges (ground state) and a set of artificial charges. The X-axis shows the time in ps and Y-axis shows the degree.

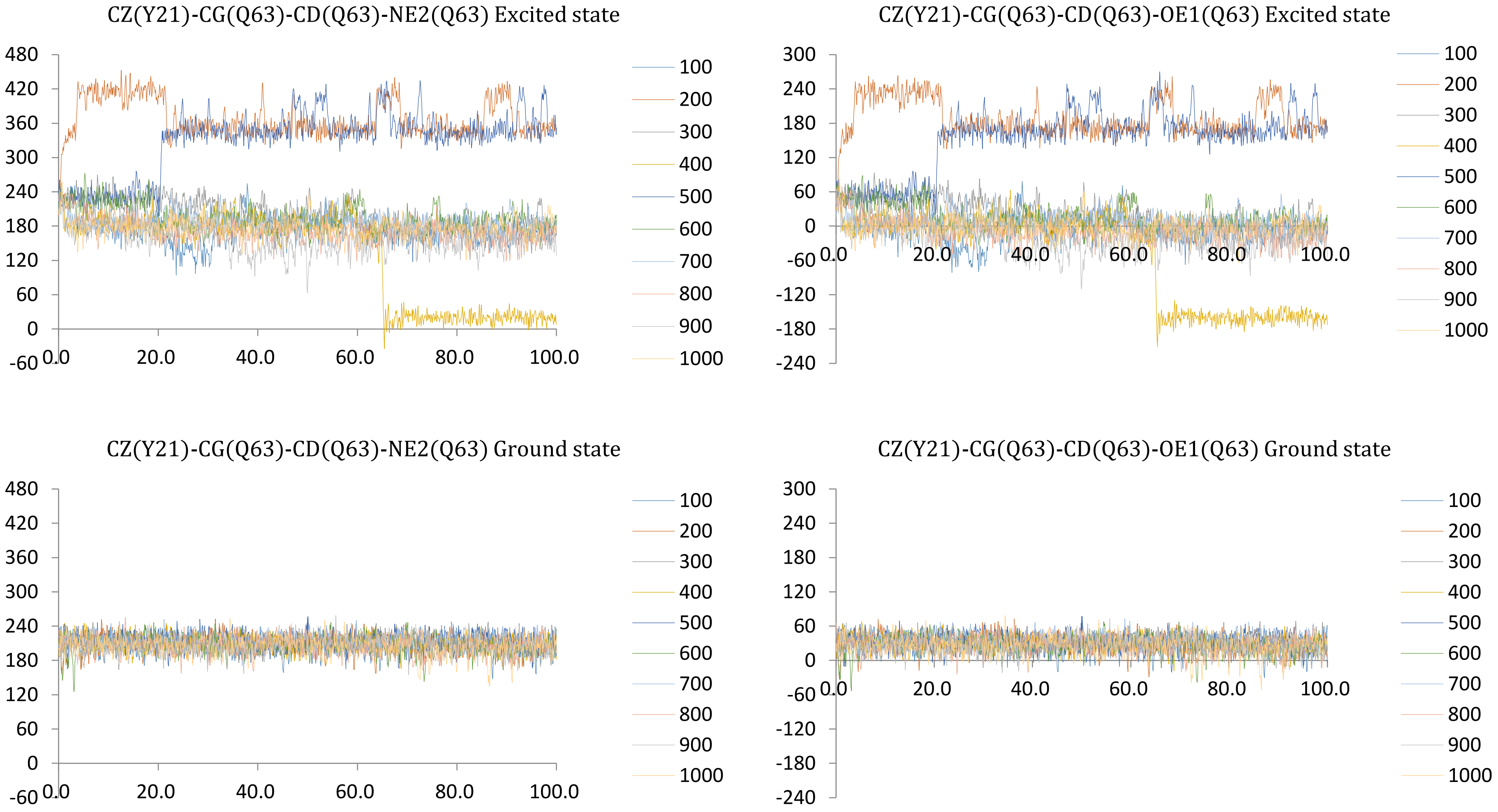
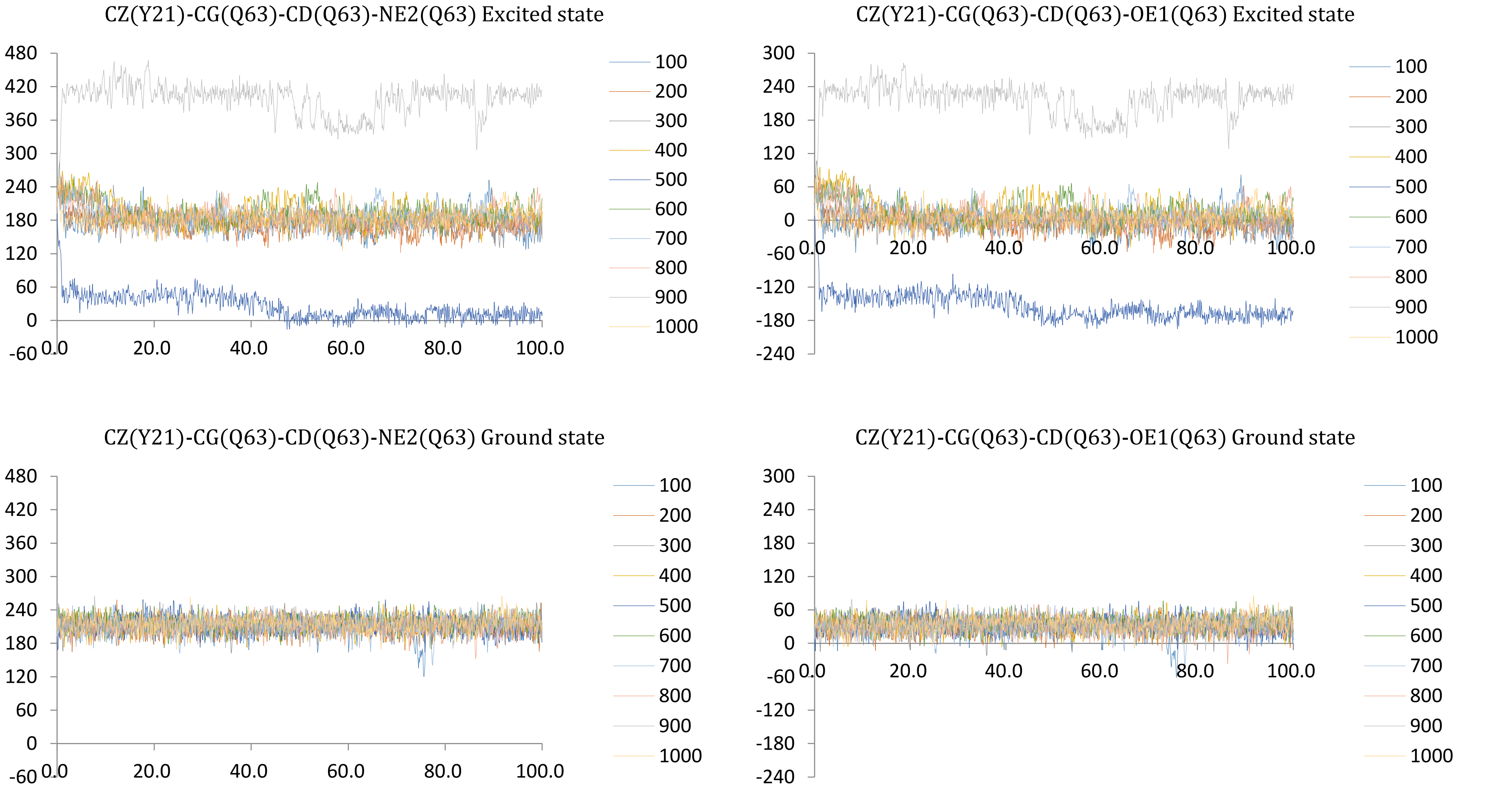


Figure 22 - Time evolution of the dihedral angle from the sixth trajectory, which characterise the orientation of the Gln63 side chain in the 2IYG (Trp-out structure) using the AmberGS charges (ground state) and a set of artificial charges. The X-axis shows the time in ps and Y-axis shows the degree.


CZ(Y21)-CG(Q63)-CD(Q63)-NE2(Q63) Excited state

CZ(Y21)-CG(Q63)-CD(Q63)-OE1(Q63) Excited state

CZ(Y21)-CG(Q63)-CD(Q63)-NE2(Q63) Ground state

CZ(Y21)-CG(Q63)-CD(Q63)-OE1(Q63) Ground state

Figure 23 - Time evolution of the dihedral angle from the seventh trajectory, which characterise the orientation of the Gln63 side chain in the 2IYG (Trp-out structure) using the AmberGS charges (ground state) and a set of artificial charges. The X-axis shows the time in ps and Y-axis shows the degree.

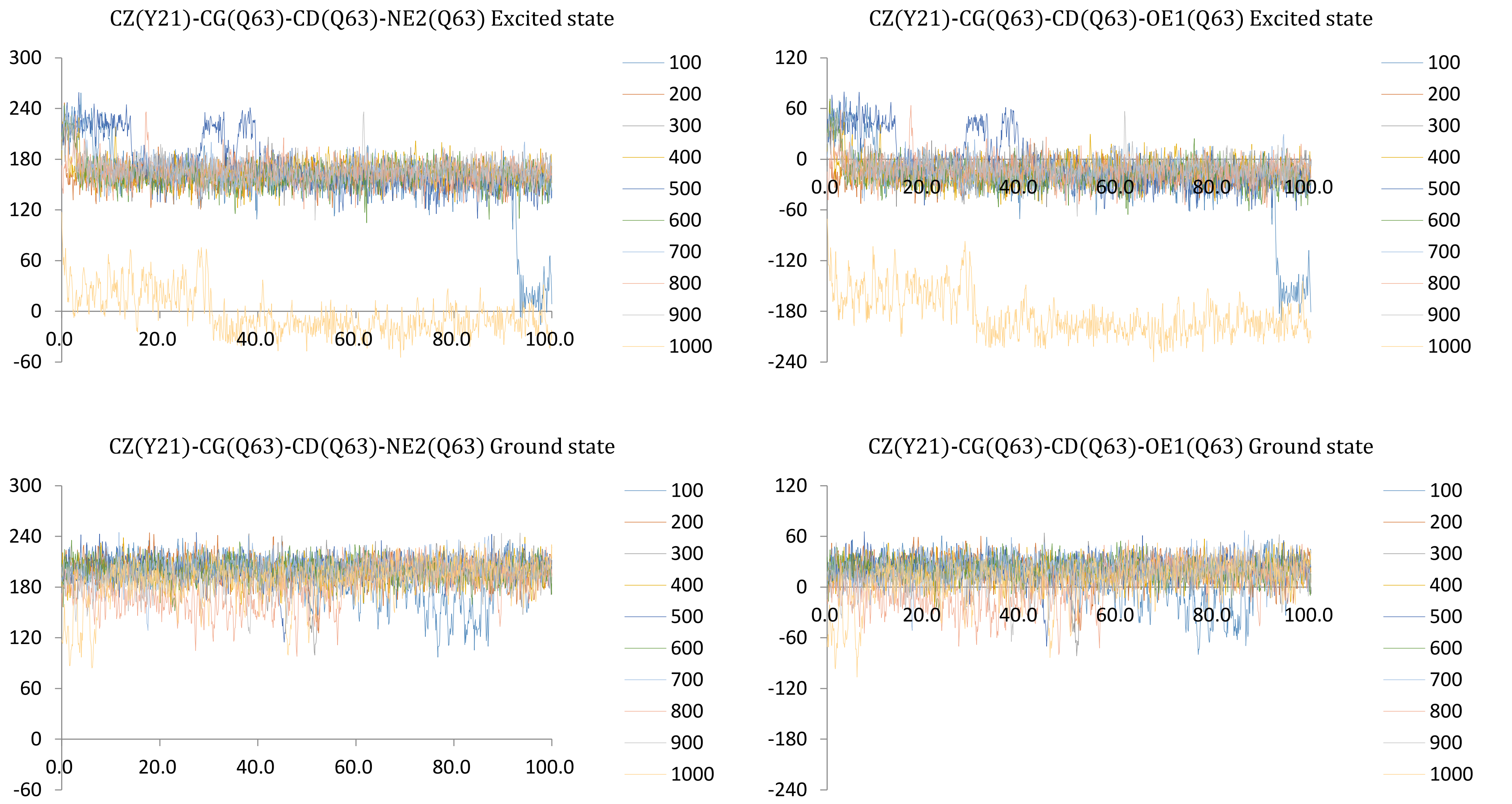


Figure 24 - Time evolution of the dihedral angle from the eighth trajectory, which characterise the orientation of the Gln63 side chain in the 2IYG (Trp-out structure) using the AmberGS charges (ground state) and a set of artificial charges. The X-axis shows the time in ps and Y-axis shows the degree.

BIBLIOGRAPHY

1 Stumpfe, D., & Bajorath, J. (2013). Critical assessment of virtual screening for hit identification. *Chemoinformatics for Drug Discovery*, 113-130.

2 Tanrikulu, Y., Krüger, B., & Proschak, E. (2013). The holistic integration of virtual screening in drug discovery. *Drug Discovery Today*, *18*(7), 358-364.

3 Katara, P. (2013). Role of bioinformatics and pharmacogenomics in drug discovery and development process. *Network Modeling Analysis in Health Informatics and Bioinformatics*, *2*(4), 225-230.

4 Blundell, T. L., Sibanda, B. L., Montalvão, R. W., Brewerton, S., Chelliah, V., Worth, C. L., Harmer, N. J., Davies, O. & Burke, D. (2006). Structural biology and bioinformatics in drug design: opportunities and challenges for target identification and lead discovery. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *361*(1467), 413-423.

5 Searls, D. B. (2000). Using bioinformatics in gene and drug discovery. *Drug Discovery Today*, *5*(4), 135-143.

6 Honore, P., Kage, K., Mikusa, J., Watt, A. T., Johnston, J. F., Wyatt, J. R., Faltynek, C.R., Jarvis, M.F. & Lynch, K. (2002). Analgesic profile of intrathecal P2X 3 antisense oligonucleotide treatment in chronic inflammatory and neuropathic pain states in rats. *Pain*, *99*(1), 11-19.

7 Chessell, I. P., Hatcher, J. P., Bountra, C., Michel, A. D., Hughes, J. P., Green, P., Egerton, J., Murfin, M., Richardson, J., Peck, W.L. & Grahames, C. B. (2005). Disruption of the P2X 7 purinoceptor gene abolishes chronic inflammatory and neuropathic pain. *Pain*, *114*(3), 386-396.

8 Zheng, X. S., & Chan, T. F. (2002). Chemical genomics: a systematic approach in biological research and drug discovery. *Current issues in molecular biology*, *4*, 33-44.

9 MacBeath, G. (2001). Chemical genomics: what will it take and who gets to play?. *Genome biology*, *2*(6), 1.

10 Salemme, F. R. (2003). Chemical genomics as an emerging paradigm for postgenomic drug discovery. *Pharmacogenomics*, *4*(3), 257-267.

11 Boppana, K., Dubey, P. K., Jagarlapudi, S. A., Vadivelan, S., & Rambabu, G. (2009). Knowledge based identification of MAO-B selective inhibitors using pharmacophore and structure based virtual screening models. *European journal of medicinal chemistry*, *44*(9), 3584-3590.

12 Lionta, E., Spyrou, G., K Vassilatis, D., & Cournia, Z. (2014). Structure-based virtual screening for drug discovery: principles, applications and recent advances. *Current topics in medicinal chemistry*, *14*(16), 1923-1938.

13 Malo, N., Hanley, J. A., Cerquozzi, S., Pelletier, J., & Nadon, R. (2006). Statistical practice in high-throughput screening data analysis. *Nature biotechnology*, *24*(2), 167-175.

14 Sergienko, E. A., & Heynen-Genel, S. (2013). Experimental Approaches to Rapid Identification, Profiling, and Characterization of Specific Biological Effects of DOS Compounds. *Diversity-Oriented Synthesis: Basics and Applications in Organic Synthesis, Drug Discovery, and Chemical Biology*, 401-429.

15 Keserű, G. M., & Makara, G. M. (2006). Hit discovery and hit-to-lead approaches. *Drug discovery today*, *11*(15), 741-748.

16 Iftekhar, M. & Jameel, S. (2015). *Computational Drug Discovery: Drug Discovery Process & Methods*. Biocuration Labs.

17 Hughes, J. P., Rees, S., Kalindjian, S. B., & Philpott, K. L. (2011). Principles of early drug discovery. *British journal of pharmacology*, *162*(6), 1239-1249.

18 Fecik, R. A., Frank, K. E., Gentry, E. J., Menon, S. R., Mitscher, L. A., & Telikepalli, H. (1998). The search for orally active medications through combinatorial chemistry. *Medicinal research reviews*, *18*(3), 149-185.

19 Teague, S. J., Davis, A. M., Leeson, P. D., & Oprea, T. (1999). The design of leadlike combinatorial libraries. *Angewandte Chemie International Edition*, *38*(24), 3743-3748.

20 Oprea, T. I., Davis, A. M., Teague, S. J., & Leeson, P. D. (2001). Is there a difference between leads and drugs? A historical perspective. *Journal of chemical information and computer sciences*, *41*(5), 1308-1315.

21 Teague, S. J., Davis, A. M., Leeson, P. D., & Oprea, T. (1999). The design of leadlike combinatorial libraries. *Angewandte Chemie International Edition*, *38*(24), 3743-3748.

22 Grüneberg, S., Stubbs, M. T., & Klebe, G. (2002). Successful virtual screening for novel inhibitors of human carbonic anhydrase: strategy and experimental confirmation. *Journal of medicinal chemistry*, *45*(17), 3588-3602.

23 Polgár, T., & M Keseru, G. (2011). Integration of virtual and high throughput screening in lead discovery settings. *Combinatorial chemistry & high throughput screening*, *14*(10), 889-897.

24 Doman, T. N., McGovern, S. L., Witherbee, B. J., Kasten, T. P., Kurumbail, R., Stallings, W. C., Connolly, D. T. & Shoichet, B. K. (2002). Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. *Journal of medicinal chemistry*, *45*(11), 2213-2221.

25 Holliday, J. D., Kanoulas, E., Malim, N., & Willett, P. (2011). Multiple search methods for similarity-based virtual screening: analysis of search overlap and precision. *Journal of cheminformatics*, *3*(1), 1.

26 Jenkins, J. L., Kao, R. Y., & Shapiro, R. (2003). Virtual screening to enrich hit lists from high-throughput screening: A case study on small-molecule inhibitors of angiogenin. *Proteins: Structure, Function, and Bioinformatics*, *50*(1), 81-93.

27 Betzi, S., Restouin, A., Opi, S., Arold, S. T., Parrot, I., Guerlesquin, F., Morelli, X. & Collette, Y. (2007). Protein–protein interaction inhibition (2P2I) combining high throughput and virtual screening: application to the HIV-1 Nef protein. *Proceedings of the National Academy of Sciences*, *104*(49), 19256-19261.

28 Danishuddin, M., & Khan, A. U. (2015). Structure based virtual screening to discover putative drug candidates: necessary considerations and successful case studies. *Methods*, *71*, 135-145.

29 Johnson, M. A., & Maggiora, G. M. (1990). *Concepts and applications of molecular similarity*. Wiley.

30 Martin, Y. C., Kofron, J. L., & Traphagen, L. M. (2002). Do structurally similar molecules have similar biological activity?. *Journal of medicinal chemistry*, *45*(19), 4350-4358.

31 Barratt, M. J., & Frail, D. E. (2012). *Drug repositioning: Bringing new life to shelved assets and existing drugs*. John Wiley & Sons.

32 Chen, Y., & Shoichet, B. K. (2009). Molecular docking and ligand specificity in fragment-based inhibitor discovery. *Nature chemical biology*, *5*(5), 358-364.

33 Boehm, M. (2011). Virtual Screening of Chemical Space: From Generic Compound Collections to Tailored Screening Libraries. *Virtual Screening: Principles, Challenges, and Practical Guidelines*, 1-33.

34 Leach, A. R., & Gillet, V. J. (2007). *An introduction to chemoinformatics*. Springer Science & Business Media.

35 Mannhold, R., & Folkers, G. (2006). *Chemogenomics in Drug Discovery: A Medicinal Chemistry Perspective* (Vol. 22). H. Kubinyi, & G. Müller (Eds.). John Wiley & Sons.

36 Maggiora, G. M. (2006). On outliers and activity cliffs why QSAR often disappoints. *Journal of chemical information and modeling*, *46*(4), 1535-1535.

37 Rankovic, Z., & Morphy, R. (Eds.). (2010). *Lead generation approaches in drug discovery*. John Wiley & Sons.

38 Méndez-Lucio, O., Kooistra, A. J., Graaf, C. D., Bender, A., & Medina-Franco, J. L. (2015). Analyzing Multitarget Activity Landscapes Using Protein–Ligand Interaction Fingerprints: Interaction Cliffs. *Journal of chemical information and modeling*, *55*(2), 251-262.

39 Furtmann, N., Hu, Y., Gütschow, M., & Bajorath, J. (2015). Identification of Interaction Hot Spots in Structures of Drug Targets on the Basis of Three-Dimensional Activity Cliff Information. *Chemical biology & drug design*, *86*(6), 1458-1465.

40 Medina-Franco, J. (Ed.). (2016). *Epi-Informatics: discovery and development of small molecule epigenetic drugs and probes*. Academic Press.

41 Varnek, A., & Tropsha, A. (2008). *Chemoinformatics approaches to virtual screening*. Royal Society

of Chemistry.

42 Roy, K., Kar, S., & Das, R. N. (2015). *Understanding the basics of QSAR for applications in pharmaceutical sciences and risk assessment*. Academic press.

43 Cross, A. F. A. (1863). *Action de l'alcool amylique sur l'organisme* (Doctoral dissertation). Faculty of Medicine, University of Strasbourg.

44 Brown, A. C., & Fraser, T. R. (1868). On the connection between chemical constitution and physiological action; with special reference to the physiological action of the salts of the ammonium bases derived from strychnia, brucia, thebaia, codeia, morphia, and nicotia. *Journal of anatomy and physiology*, *2*(2), 224-242.

45 Richardson, B. W. (1869). On Bichloride of Methylene. *British medical journal*, *2*(462), 487.

46 Meyer, H. (1899). Zur theorie der alkoholnarkose. *Naunyn-Schmiedeberg's Archives of Pharmacology*, *42*(2), 109-118.

47 Overton, E. (1901). *Studien uber die Narkose: zugleich ein Beitrag zur allgemeinen Pharmakologie*. Gustav Fischer.

48 Trabe, J. (1904). Theorie der osmose und narkose. *Pflügers Archiv European Journal of Physiology*, *105*(11), 541-558.

49 Seidell, A. (1912). A new bromine method for the determination of thymol, salicylates, and similar compounds. *Am Chem J*, *47*, 508-526.

50 Hansch, C., Maloney, P. P., Fujita, T., & Muir, R. M. (1962). Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. *Nature*, 194(4824),178-180

51 Varnek, A., & Tropsha, A. (2008). *Chemoinformatics approaches to virtual screening*. Royal Society of Chemistry.

52 Cherkasov, A., Muratov, EN., Fourches, D., Varnek, A., Baskin, II., Cronin, M., Dearden, J., Gramatica, P., Martin, YC., Todeschini, R., Consonni, V., Kuz'Min, VE., Cramer, R., Benigni, R., Yang, C., Rathman, J., Terfloth, L., Gasteiger, J., Richard, A., & Tropsha, A. (2014). QSAR modeling: where have you been? Where are you going to?. *Journal of medicinal chemistry*, *57*(12), 4977-5010.

53 Mitchell, J. B. (2014). Machine learning methods in chemoinformatics. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, *4*(5), 468-481.

54 Lipinski, C., & Hopkins, A. (2004). Navigating chemical space for biology and medicine. *Nature*, *432*(7019), 855-861.

55 Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, *28*(1), 31-

36.

56 Hinselmann, G., Jahn, A., Fechner, N., & Zell, A. (2009, April). Chronic Rat Toxicity Prediction of Chemical Compounds Using Kernel Machines. In *European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics* (pp. 25-36). Springer Berlin Heidelberg.

57 Tauler, R., Walczak, B., & Brown, S. D. (2009). *Comprehensive chemometrics: chemical and biochemical data analysis*. Elsevier.

58 Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, *26*(3), 159-190.

59 Kouiroukidis, N., & Evangelidis, G. (2011, September). The effects of dimensionality curse in high dimensional knn search. In *Informatics (PCI), 2011 15th Panhellenic Conference on* (pp. 41-45). IEEE.

60 Ali, A., Shamsuddin, S. M., & Ralescu, A. L. (2015). Classification with class imbalance problem: A Review. *Int. J. Advance Soft Compu. Appl*, *7*(3).

61 Bastos, J. (2008). Credit scoring with boosted decision trees. *Munich Personal RePEc Archive - MPRA*, 8156, 2-4

62 Quinlan, J. R. (1979). *Discovering rules by induction from large collections of examples* (pp. 168-201). Expert systems in the micro electronic age. Edinburgh University Press.

63 Quinlan, J. R. (1993). C4. 5: Programming for machine learning. *Morgan Kauffmann*, 38.

64 Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

65 Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123-140.

66 Freund, Y., & Schapire, R. E. (1996, July). Experiments with a new boosting algorithm. In *Icml* (Vol. 96, pp. 148-156).

67 Berzal, F., Cubero, J. C., Cuenca, F., & Martín-Bautista, M. J. (2003). On the quest for easy-to-understand splitting rules. *Data & Knowledge Engineering*, *44*(1), 31-48.

68 Maglogiannis, I. G. (2007). *Emerging artificial intelligence applications in computer engineering: real word AI systems with applications in eHealth, HCI, information retrieval and pervasive technologies* (Vol. 160). Ios Press.

69 Coppin, B. (2004). *Artificial intelligence illuminated*. Jones & Bartlett Learning.

70 Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms. *Neural computation*, *8*(7), 1341-1390.

71 Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the*

ACM, *55*(10), 78-87.

72 Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

73 Raileanu, L. E., & Stoffel, K. (2004). Theoretical comparison between the gini index and information gain criteria. *Annals of Mathematics and Artificial Intelligence*, *41*(1), 77-93.

74 Venkatesan, P., & Yamuna, N. R. (2013). Treatment response classification in randomized clinical trials: a decision tree approach. *Indian Journal of Science and Technology*, *6*(1), 3912-3917.

75 Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273-297.

76 Beyer, K., Goldstein, J., Ramakrishnan, R., & Shaft, U. (1999). When is "nearest neighbor" meaningful?. In *International conference on database theory*, 217-235. Springer Berlin Heidelberg.

77 Lewis, D. D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval. In *European conference on machine learning*, 4-15. Springer Berlin Heidelberg.

78 Ko, G. M., Reddy, A. S., Kumar, S., & Garg, R. Data Mining Analysis of HIV-1 Protease Crystal Structures.

79 Palmer, D. S., O'Boyle, N. M., Glen, R. C., & Mitchell, J. B. (2007). Random forest models to predict aqueous solubility. *Journal of chemical information and modeling*, *47*(1), 150-158.

80 Ballester, P. J., & Mitchell, J. B. (2010). A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. *Bioinformatics*, *26*(9), 1169-1175.

81 Springer, C., Adalsteinsson, H., Young, M. M., Kegelmeyer, P. W., & Roe, D. C. (2005). PostDOCK: a structural, empirical approach to scoring protein ligand complexes. *Journal of medicinal chemistry*, *48*(22), 6821-6831.

82 Zilian, D., & Sotriffer, C. A. (2013). SFCscore RF: a random forest-based scoring function for improved affinity prediction of protein–ligand complexes. *Journal of chemical information and modeling*, *53*(8), 1923-1933.

83 Kalyaanamoorthy, S., & Chen, Y. P. P. (2011). Structure-based drug design to augment hit discovery. *Drug discovery today*, *16*(17), 831-839.

84 Kalman, M., & Ben-Tal, N. (2010). Quality assessment of protein model-structures using evolutionary conservation. *Bioinformatics*, *26*(10), 1299-1307.

85 Engle, J. M., Lakshminarayanan, P. S., Carroll, C. N., Zakharov, L. N., Haley, M. M., & Johnson, D. W. (2011). Molecular Self-Assembly: Solvent Guests Tune the Conformation of a Series of 2, 6-Bis (2-anilinoethynyl) pyridine-Based Ureas. *Crystal growth & design*, *11*(11), 5144-5152.

86 Zhang, Y., & Skolnick, J. (2004). Automated structure prediction of weakly homologous proteins on a genomic scale. *Proceedings of the National Academy of Sciences of the United States of*

America, *101*(20), 7594-7599.

87 Antunes, D. A., Devaurs, D., & Kavraki, L. E. (2015). Understanding the challenges of protein flexibility in drug design. *Expert opinion on drug discovery*, *10*(12), 1301-1313.

88 Koshland, D. E. (1958). Application of a theory of enzyme specificity to protein synthesis. *Proceedings of the National Academy of Sciences*, *44*(2), 98-104.

89 Burgen, A. S. (1981, November). Conformational changes and drug action. In *Federation proceedings* (Vol. 40, No. 13, p. 2723).

90 Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R., & Ferrin, T. E. (1982). A geometric approach to macromolecule-ligand interactions. *Journal of molecular biology*, *161*(2), 269-288.

91 DesJarlais, R. L., Sheridan, R. P., Dixon, J. S., Kuntz, I. D., & Venkataraghavan, R. (1986). Docking flexible ligands to macromolecular receptors by molecular shape. *Journal of medicinal chemistry*, *29*(11), 2149-2153.

92 Kitchen, D. B., Decornez, H., Furr, J. R., & Bajorath, J. (2004). Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature reviews Drug discovery*, *3*(11), 935-949.

93 Totrov, M., & Abagyan, R. (2008). Flexible ligand docking to multiple receptor conformations: a practical alternative. *Current opinion in structural biology*, *18*(2), 178-184.

94 Teodoro, M. L., & Kavraki, L. E. (2003). Conformational flexibility models for the receptor in structure based drug design. *Current pharmaceutical design*, *9*(20), 1635-1648.

95 Meiler, J., & Baker, D. (2006). ROSETTALIGAND: Protein–small molecule docking with full side-chain flexibility. *Proteins: Structure, Function, and Bioinformatics*, *65*(3), 538-548.

96 Lin, J. H., Perryman, A. L., Schames, J. R., & McCammon, J. A. (2002). Computational drug design accommodating receptor flexibility: the relaxed complex scheme. *Journal of the American Chemical Society*, *124*(20), 5632-5633.

97 Shuker, S. B., Hajduk, P. J., Meadows, R. P., & Fesik, S. W. (1996). Discovering high-affinity ligands for proteins: SAR by NMR. *Science*, *274*(5292), 1531.

98 Erlanson, D. A., Braisted, A. C., Raphael, D. R., Randal, M., Stroud, R. M., Gordon, E. M., & Wells, J. A. (2000). Site-directed ligand discovery. *Proceedings of the National Academy of Sciences*, *97*(17), 9367-9372.

99 Barakat, K., & Tuszynski, J. (2011). Relaxed complex scheme suggests novel inhibitors for the lyase activity of DNA polymerase beta. *Journal of Molecular Graphics and Modelling*, *29*(5), 702-716.

100 Barakat, K., & Tuszynski, J. (2011). Relaxed complex scheme suggests novel inhibitors for the lyase activity of DNA polymerase beta. *Journal of Molecular Graphics and Modelling*, *29*(5), 702-716.

101 Lin, J. H., Perryman, A. L., Schames, J. R., & McCammon, J. A. (2002). Computational drug design accommodating receptor flexibility: the relaxed complex scheme. *Journal of the American Chemical Society*, *124*(20), 5632-5633.

102 Amaro, R. E., Baron, R., & McCammon, J. A. (2008). An improved relaxed complex scheme for receptor flexibility in computer-aided drug design. *Journal of computer-aided molecular design*, *22*(9), 693-705.

103 Machado, K. S., Winck, A. T., Ruiz, D. D., & de Souza, O. N. (2010). Mining flexible-receptor docking experiments to select promising protein receptor snapshots. *BMC genomics*, *11*(5), 1.

104 Holland, J. H. (1975). *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. U Michigan Press.

105 Maillot, P. G., & Glassner, A. S. (1990). Graphics gems. *Academic Press, London*, *498*.

106 Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R., Hart, W. E., Belew, R. K., & Olson, A. J. (1998). Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *Journal of computational chemistry*, *19*(14), 1639-1662.

107 Solis, F. J., & Wets, R. J. B. (1981). Minimization by random search techniques. *Mathematics of operations research*, *6*(1), 19-30.

108 Pattabiraman, N., Levitt, M., Ferrin, T. E., & Langridge, R. (1985). Computer graphics in real-time docking with energy calculation and minimization. *Journal of computational chemistry*, *6*(5), 432-436.

109 Goodford, P. J. (1985). A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *Journal of medicinal chemistry*, *28*(7), 849-857.

110 Morris, G. M., Goodsell, D. S., Huey, R., Hart, W. E., Halliday, S., Belew, R., & Olson, A. J. (2001). Automated Docking of Flexible Ligands to Receptors, User's Guide.

111 Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., Profeta, S. & Weiner, P. (1984). A new force field for molecular mechanical simulation of nucleic acids and proteins. *Journal of the American Chemical Society*, *106*(3), 765-784.

112 Morris, G. M., Goodsell, D. S., Huey, R., & Olson, A. J. (1996). Distributed automated docking of flexible ligands to proteins: parallel applications of AutoDock 2.4. *Journal of computer-aided molecular design*, *10*(4), 293-304.

113 Huey, R., Morris, G. M., Olson, A. J., & Goodsell, D. S. (2007). A semiempirical free energy force field with charge-based desolvation. *Journal of computational chemistry*, *28*(6), 1145-1152.

114 Wesson, L., & Eisenberg, D. (1992). Atomic solvation parameters applied to molecular dynamics of proteins in solution. *Protein Science*, *1*(2), 227-235.

115 Stouten, P. F., Frömmel, C., Nakamura, H., & Sander, C. (1993). An effective solvation term based on atomic occupancies for use in protein simulations. *Molecular Simulation*, *10*(2-6), 97-120.

116 Courant, R. (1943). Variational methods for the solution of problems of equilibrium and vibrations. *Bull. Amer. Math. Soc*, *49*(1), 1-23.

117 Swope, W. C., Andersen, H. C., Berens, P. H., & Wilson, K. R. (1982). A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *The Journal of Chemical Physics*, *76*(1), 637-649.

118 Hockney, R. W., Goel, S. P., & Eastwood, J. W. (1974). Quiet high-resolution computer models of a plasma. *Journal of Computational Physics*, *14*(2), 148-158.

119 Hess, B. (2008). P-LINCS: A parallel linear constraint solver for molecular simulation. *Journal of Chemical Theory and Computation*, *4*(1), 116-122.

120 del Sol, A., Tsai, C. J., Ma, B., & Nussinov, R. (2009). The origin of allosteric functional modulation: multiple pre-existing pathways. *Structure*, *17*(8), 1042-1050.

121 Monod, J., Wyman, J., & Changeux, J. P. (1965). On the nature of allosteric transitions: a plausible model. *Journal of molecular biology*, *12*(1), 88-118.

122 Koshland Jr, D. E., Nemethy, G., & Filmer, D. (1966). Comparison of experimental binding data and theoretical models in proteins containing subunits*. *Biochemistry*, *5*(1), 365-385.

123 Tsai, C. J., Del Sol, A., & Nussinov, R. (2009). Protein allostery, signal transmission and dynamics: a classification scheme of allosteric mechanisms. *Molecular Biosystems*, *5*(3), 207-216.

124 Laskowski, R. A., Gerick, F., & Thornton, J. M. (2009). The structural basis of allosteric regulation in proteins. *FEBS letters*, *583*(11), 1692-1698.

125 Lu, S., Li, S., & Zhang, J. (2014). Harnessing allostery: a novel approach to drug discovery. *Medicinal research reviews*, *34*(6), 1242-1285.

126 Boehr, D. D., McElheny, D., Dyson, H. J., & Wright, P. E. (2006). The dynamic energy landscape of dihydrofolate reductase catalysis. *science*, *313*(5793), 1638-1642.

127 Malmendal, A., Evenäs, J., Forsén, S., & Akke, M. (1999). Structural dynamics in the C-terminal domain of calmodulin at low calcium levels. *Journal of molecular biology*, *293*(4), 883-899.

128 Volkman, B. F., Lipson, D., Wemmer, D. E., & Kern, D. (2001). Two-state allosteric behavior in a single-domain signaling protein. *Science*, *291*(5512), 2429-2433.

129 Kumar, S., Ma, B., Tsai, C. J., Sinha, N., & Nussinov, R. (2000). Folding and binding cascades: dynamic landscapes and population shifts. *Protein Science*, *9*(1), 10-19.

130 Tsai, C. J., Ma, B., & Nussinov, R. (1999). Folding and binding cascades: shifts in energy landscapes. *Proceedings of the National Academy of Sciences*, *96*(18), 9970-9972.

131 Panjkovich, A., & Daura, X. (2012). Exploiting protein flexibility to predict the location of allosteric sites. *BMC bioinformatics*, *13*(1), 273.

132 Lockless, S. W., & Ranganathan, R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, *286*(5438), 295-299.

133 Lu, S., Huang, W., & Zhang, J. (2014). Recent computational advances in the identification of allosteric sites in proteins. *Drug discovery today*, *19*(10), 1595-1600.

134 Novinec, M., Korenč, M., Caflisch, A., Ranganathan, R., Lenarčič, B., & Baici, A. (2014). A novel allosteric mechanism in the cysteine peptidase cathepsin K discovered by computational methods. *Nature communications*, *5*, 3287.

135 Panjkovich, A., & Daura, X. (2014). PARS: a web server for the prediction of protein allosteric and regulatory sites. *Bioinformatics*, *30*(9), 1314-1315.

136 Huang, W., Lu, S., Huang, Z., Liu, X., Mou, L., Luo, Y., Zhao Y, Liu Y, Chen Z, Hou T, & Zhang, J. (2013). Allosite: a method for predicting allosteric sites. *Bioinformatics*, *29*(18), 2357-2359.

137 van Westen, G. J., Gaulton, A., & Overington, J. P. (2014). Chemical, target, and bioactive properties of allosteric modulation. *PLoS Comput Biol*, *10*(4), e1003559.

138 Bento, A. P., Gaulton, A., Hersey, A., Bellis, L. J., Chambers, J., Davies, M., Krüger, F.A., Light, Y., Mak, L., McGlinchey, S. & Nowotka, M. (2014). The ChEMBL bioactivity database: an update. *Nucleic acids research*, *42*(D1), D1083-D1090.

139 Greener, J. G., & Sternberg, M. J. (2015). AlloPred: prediction of allosteric pockets on proteins using normal mode perturbation analysis. *BMC bioinformatics*, *16*(1), 335.

140 Demerdash, O. N., Daily, M. D., & Mitchell, J. C. (2009). Structure-based predictive models for allosteric hot spots. *PLoS Comput Biol*, *5*(10), e1000531.

141 Erman, B. (2011). Relationships between ligand binding sites, protein architecture and correlated paths of energy and conformational fluctuations. *Physical biology*, *8*(5), 056003.

142 Kaya, C., Armutlulu, A., Ekesan, S., & Haliloglu, T. (2013). MCPath: Monte Carlo path generation approach to predict likely allosteric pathways and functional residues. *Nucleic acids research*, *41*(W1), W249-W255.

143 Hardy, J. A., & Wells, J. A. (2004). Searching for new allosteric sites in enzymes. *Current opinion in structural biology*, *14*(6), 706-715.

144 Long, D., & Yang, D. (2009). Buffer interference with protein dynamics: a case study on human

liver fatty acid binding protein. *Biophysical journal*, *96*(4), 1482-1488.

145 Paul Brear, PhD Thesis, University of St Andrews 2012

146 Ballester, P. J., & Mitchell, J. B. (2010). A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. *Bioinformatics*, *26*(9), 1169-1175.

147 Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., & Feuston, B. P. (2003). Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of chemical information and computer sciences*, *43*(6), 1947-1958.

148 Huang, Z., Zhu, L., Cao, Y., Wu, G., Liu, X., Chen, Y., Wang, Q., Shi, T., Zhao, Y., Wang, Y. & Li, W. (2011). ASD: a comprehensive database of allosteric proteins and modulators. *Nucleic acids research*, *39*(suppl 1), D663-D669.

149 Laskowski, R. A. (2001). PDBsum: summaries and analyses of PDB structures. *Nucleic acids research*, *29*(1), 221-222.

150 Wang, R., Fang, X., Lu, Y., Yang, C. Y., & Wang, S. (2005). The PDBbind database: methodologies and updates. *Journal of medicinal chemistry*, *48*(12), 4111-4119.

151 Cheng, T., Li, X., Li, Y., Liu, Z., & Wang, R. (2009). Comparative assessment of scoring functions on a diverse test set. *Journal of chemical information and modeling*, *49*(4), 1079-1093.

152 Cuff, A. L., Sillitoe, I., Lewis, T., Redfern, O. C., Garratt, R., Thornton, J., & Orengo, C. A. (2009). The CATH classification revisited—architectures reviewed and new ways to characterize structural divergence in superfamilies. *Nucleic acids research*, *37*(suppl 1), D310-D314.

153 Xu, G., Potter, J. A., Russell, R. J., Oggioni, M. R., Andrew, P. W., & Taylor, G. L. (2008). Crystal structure of the NanB sialidase from Streptococcus pneumoniae. *Journal of molecular biology*, *384*(2), 436-449.

154 Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, *2*(3), 18-22.

155 Team, R. C. (2013). R: A language and environment for statistical computing.

156 Bondi, A. (1964). van der Waals volumes and radii. *The Journal of physical chemistry*, *68*(3), 441-451.

157 Kirtay, C. K., Mitchell, J. B., & Lumley, J. A. (2005). Knowledge based potentials: The reverse Boltzmann methodology, virtual screening and molecular weight dependence. *QSAR & Combinatorial Science*, *24*(4), 527-536.

158 Cheng, T., Li, X., Li, Y., Liu, Z., & Wang, R. (2009). Comparative assessment of scoring functions on a diverse test set. *Journal of chemical information and modeling*, *49*(4), 1079-1093.

159 Kuntz, I. D., Chen, K., Sharp, K. A., & Kollman, P. A. (1999). The maximal affinity of

ligands. *Proceedings of the National Academy of Sciences*, *96*(18), 9997-10002.
160 Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E., & Willighagen, E. (2003). The Chemistry Development Kit (CDK): An open-source Java library for chemo-and bioinformatics. *Journal of chemical information and computer sciences*, *43*(2), 493-500.

161 Svetnik, V., Liaw, A., Tong, C., & Wang, T. (2004, June). Application of Breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules. In *International Workshop on Multiple Classifier Systems* (pp. 334-343). Springer Berlin Heidelberg.

162 Chiu, H. J., Bakolitsa, C., Skerra, A., Lomize, A., Carlton, D., Miller, M. D., Krishna, S.S., Abdubek, P., Astakhova, T., Axelrod, H.L. & Clayton, T. (2010). Structure of the first representative of Pfam family PF09410 (DUF2006) reveals a structural signature of the calycin superfamily that suggests a role in lipid metabolism. *Acta Crystallographica Section F: Structural Biology and Crystallization Communications*, *66*(10), 1153-1159.

163 Sancar, F., & Czajkowski, C. (2011). Allosteric modulators induce distinct movements at the GABA-binding site interface of the GABA-A receptor. *Neuropharmacology*, *60*(2), 520-528.

164 Stanger, K., Steffek, M., Zhou, L., Pozniak, C. D., Quan, C., Franke, Y., Tom, J., Tam, C., Krylova, I., Elliott, J.M. & Lewcock, J. W. (2012). Allosteric peptides bind a caspase zymogen and mediate caspase tetramerization. *Nature chemical biology*, *8*(7), 655-660.

165 Wang, Y., & Patel, D. J. (1993). Solution structure of the human telomeric repeat d [AG3(T2AG3)3] G-tetraplex. *Structure*, *1*(4), 263-282.

166 Hud, N. V., Smith, F. W., Anet, F. A., & Feigon, J. (1996). The selectivity for K^+ versus Na^+ in DNA quadruplexes is dominated by relative free energies of hydration: a thermodynamic analysis by 1H NMR. *Biochemistry*, *35*(48), 15383-15390.

167 Armas, P., Nasif, S., & Calcaterra, N. B. (2008). Cellular nucleic acid binding protein binds G-rich single-stranded nucleic acids and may function as a nucleic acid chaperone. *Journal of cellular biochemistry*, *103*(3), 1013-1036.

168 Bang, I. (1910). Untersuchungen über die Guanylsäure. *Biochemische Zeitschrift*, *26*, 293-311.

169 Gellert, M., Lipsett, M. N., & Davies, D. R. (1962). Helix formation by guanylic acid. *Proceedings of the National Academy of Sciences*, *48*(12), 2013-2018.

170 Williamson, J. R., Raghuraman, M. K., & Cech, T. R. (1989). Monovalent cation-induced structure of telomeric DNA: the G-quartet model. *Cell*, *59*(5), 871-880.

171 Burge, S., Parkinson, G. N., Hazel, P., Todd, A. K., & Neidle, S. (2006). Quadruplex DNA: sequence, topology and structure. *Nucleic acids research*, *34*(19), 5402-5415.

172 Balagurumoorthy, P., & Brahmachari, S. K. (1994). Structure and stability of human telomeric sequence. *Journal of Biological Chemistry*, *269*(34), 21858-21869.

173 Luedtke, N. W. (2009). Targeting G-quadruplex DNA with small molecules.*CHIMIA International*

Journal for Chemistry, *63*(3), 134-139.

174 Kim, M. Y., Gleason-Guzman, M., Izbicka, E., Nishioka, D., & Hurley, L. H. (2003). The different biological effects of telomestatin and TMPyP4 can be attributed to their selectivity for interaction with intramolecular or intermolecular G-quadruplex structures. *Cancer research*, *63*(12), 3247-3256.

175 Liu, Z., & Gilbert, W. (1994). The yeast KEM1 gene encodes a nuclease specific for G4 tetraplex DNA: implication of in vivo functions for this novel DNA structure. *Cell*, *77*(7), 1083-1092.

176 Sun, H., Yabuki, A., & Maizels, N. (2001). A human nuclease specific for G4 DNA. *Proceedings of the National Academy of Sciences*, *98*(22), 12444-12449.

177 Oganesian, L., & Bryan, T. M. (2007). Physiological relevance of telomeric G-quadruplex formation: a potential drug target. *Bioessays*, *29*(2), 155-165.

178 Ren, J., & Chaires, J. B. (1999). Sequence and structural selectivity of nucleic acid binding ligands. *Biochemistry*, *38*(49), 16067-16075.

179 Reed, J., Gunaratnam, M., Beltran, M., Reszka, A. P., Vilar, R., & Neidle, S. (2008). TRAP–LIG, a modified telomere repeat amplification protocol assay to quantitate telomerase inhibition by small molecules. *Analytical biochemistry*, *380*(1), 99-105.

180 Nicoludis, J. M., Barrett, S. P., Mergny, J. L., & Yatsunyk, L. A. (2012). Interaction of human telomeric DNA with N-methyl mesoporphyrin IX. *Nucleic acids research*, *40*(12), 5432-5447.

181 Monchaud, D., Granzhan, A., Saettel, N., Guédin, A., Mergny, J. L., & Teulade-Fichou, M. P. (2010). "One Ring to Bind Them All" - Part I: The Efficiency of the Macrocyclic Scaffold for G-Quadruplex DNA Recognition. *Journal of nucleic acids*, *2010*, 1-19

182 Siddiqui-Jain, A., Grand, C. L., Bearss, D. J., & Hurley, L. H. (2002). Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. *Proceedings of the National Academy of Sciences*, *99*(18), 11593-11598.

183 Palumbo, S. L., Memmott, R. M., Uribe, D. J., Krotova-Khan, Y., Hurley, L. H., & Ebbinghaus, S. W. (2008). A novel G-quadruplex-forming GGA repeat region in the c-myb promoter is a critical regulator of promoter activity. *Nucleic acids research*, *36*(6), 1755-1769.

184 Gomez, D., O'Donohue, M. F., Wenner, T., Douarre, C., Macadré, J., Koebel, P., Giraud-Panis, M.J., Kaplan, H., Kolkes, A., Shin-ya, K. & Riou, J. F. (2006). The G-quadruplex ligand telomestatin inhibits POT1 binding to telomeric sequences in vitro and induces GFP-POT1 dissociation from telomeres in human cells. *Cancer Research*, *66*(14), 6908-6912.

185 Patel, D. J., Phan, A. T., & Kuryavyi, V. (2007). Human telomere, oncogenic promoter and 5′-UTR G-quadruplexes: diverse higher order DNA and RNA targets for cancer therapeutics. *Nucleic acids research*, *35*(22), 7429-7455.

186 Brenlla, A., Veiga, M., Pérez Lustres, J. L., Ríos Rodríguez, M. C., Rodríguez-Prieto, F., & Mosquera,

M. (2013). Photoinduced Proton and Charge Transfer in 2-(2′-Hydroxyphenyl) imidazo [4, 5-b] pyridine. *The Journal of Physical Chemistry B*, *117*(3), 884-896.

187 Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., & Olson, A. J. (2009). AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of computational chemistry*, *30*(16), 2785-2791.

188 Maseras, F., & Morokuma, K. (1995). IMOMM: A new integrated ab initio+ molecular mechanics geometry optimization scheme of equilibrium structures and transition states. *Journal of Computational Chemistry*, *16*(9), 1170-1179.

189 Groenhof, G. (2013). Introduction to QM/MM simulations. *Biomolecular simulations: methods and protocols*, 43-66.

190 Parkinson, G. N., Lee, M. P., & Neidle, S. (2002). Crystal structure of parallel quadruplexes from human telomeric DNA. *Nature*, *417*(6891), 876-880.

191 Berendsen, H. J., van der Spoel, D., & van Drunen, R. (1995). GROMACS: a message-passing parallel molecular dynamics implementation. *Computer Physics Communications*, *91*(1), 43-56.

192 Lindahl, E., Hess, B., & Van Der Spoel, D. (2001). GROMACS 3.0: a package for molecular simulation and trajectory analysis. *Molecular modeling annual*, *7*(8), 306-317.

193 Hess, B., Kutzner, C., Van Der Spoel, D., & Lindahl, E. (2008). GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of chemical theory and computation*, *4*(3), 435-447.

194 Gasteiger, J., & Marsili, M. (1980). Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. *Tetrahedron*, *36*(22), 3219-3228.

195 Hamzeh-Mivehroud, M., Sokouti, B., & Dastmalchi, S. (2016). Molecular Docking at a Glance. In *Methods and Algorithms for Molecular Docking-Based Drug Design and Discovery* (pp. 1-38). IGI Global.

196 Miertuš, S., Scrocco, E., & Tomasi, J. (1981). Electrostatic interaction of a solute with a continuum. A direct utilizaion of AB initio molecular potentials for the prevision of solvent effects. *Chemical Physics*, *55*(1), 117-129.

197 Sim, G. A., & Sutton, L. E. (1974). Molecular Structure by Diffraction Methods. Vol. 1. *Structural Science*.

198 Chipem, F. A., & Krishnamoorthy, G. (2009). Comparative theoretical study of rotamerism and excited state intramolecular proton transfer of 2-(2′-hydroxyphenyl) benzimidazole, 2-(2′-hydroxyphenyl) imidazo [4, 5-b] pyridine, 2-(2′-hydroxyphenyl) imidazo [4, 5-c] pyridine and 8-(2′-hydroxyphenyl) purine. *The Journal of Physical Chemistry A*, *113*(44), 12063-12070.

199 Chipem, F. A., & Krishnamoorthy, G. (2013). Temperature effect on dual fluorescence of 2-(2′-

Hydroxyphenyl) benzimidazole and its nitrogen substituted analogues. *The Journal of Physical Chemistry B*, *117*(45), 14079-14088.

200 Des Jarlais, R. L., Cummings, M. D., & Gibbs, A. C. (2007). Virtual docking: how are we doing and how can we improve. *Front Drug Des Discov Struct-Based Drug Des 21st Century*, *3*, 81-103.

201 Nichols, S. E., Baron, R., Ivetac, A., & McCammon, J. A. (2011). Predictive power of molecular dynamics receptor structures in virtual screening. *Journal of chemical information and modeling*, *51*(6), 1439-1446.

202 Cabani, S., Gianni, P., Mollica, V., & Lepori, L. (1981). Group contributions to the thermodynamic properties of non-ionic organic solutes in dilute aqueous solution. *Journal of Solution Chemistry*, *10*(8), 563-595.

203 Humphrey, W., Dalke, A., & Schulten, K. (1996). VMD: visual molecular dynamics. *Journal of molecular graphics*, *14*(1), 33-38.

204 Medhi, C., Mitchell, J. B.O., Price, S. L., & Tabor, A. B. (1999). Electrostatic factors in DNA intercalation. *Biopolymers*, *52*(2), 84-93.

205 Bouly, J. P., Schleicher, E., Dionisio-Sese, M., Vandenbussche, F., Van Der Straeten, D., Bakrim, N., Meier, S., Batschauer, A., Galland, P., Bittl, R. & Ahmad, M. (2007). Cryptochrome blue light photoreceptors are activated through interconversion of flavin redox states. *Journal of Biological Chemistry*, *282*(13), 9383-9391.

206 Kennis, J. T., & Mathes, T. (2013). Molecular eyes: proteins that transform light into biological information. *Interface focus*, *3*(5), 20130005.

207 Chaves, I., Pokorny, R., Byrdin, M., Hoang, N., Ritz, T., Brettel, K., Essen, L.O., van der Horst, G.T., Batschauer, A. and Ahmad, M. (2011) The cryptochromes: blue light photoreceptors in plants and animals. *Annual review of plant biology*, *62*, 335-364.

208 Fudim, R., Mehlhorn, J., Berthold, T., Weber, S., Schleicher, E., Kennis, J., & Mathes, T. (2015). Photoinduced formation of flavin radicals in BLUF domains lacking the central glutamine. *FEBS journal*, *282*(16), 3161-3174.

209 Hasegawa, K., Masuda, S., & Ono, T. A. (2005). Spectroscopic analysis of the dark relaxation process of a photocycle in a sensor of blue light using FAD (BLUF) protein Slr1694 of the cyanobacterium Synechocystis sp. PCC6803. *Plant and cell physiology*, *46*(1), 136-146.

210 Dragnea, V., Waegele, M., Balascuta, S., Bauer, C., & Dragnea, B. (2005). Time-resolved spectroscopic studies of the AppA blue-light receptor BLUF domain from Rhodobacter sphaeroides. *Biochemistry*, *44*(49), 15978-15985.

211 Domratcheva, T., Hartmann, E., Schlichting, I., & Kottke, T. (2016). Evidence for Tautomerisation of Glutamine in BLUF Blue Light Receptors by Vibrational Spectroscopy and Computational Chemistry. *Scientific reports*, *6*.

212 Gauden, M., van Stokkum, I. H., Key, J. M., Lührs, D. C., Van Grondelle, R., Hegemann, P., & Kennis, J. T. (2006). Hydrogen-bond switching through a radical pair mechanism in a flavin-binding photoreceptor. *Proceedings of the National Academy of Sciences*, *103*(29), 10895-10900.

213 Möglich, A., Yang, X., Ayers, R. A., & Moffat, K. (2010). Structure and function of plant photoreceptors. *Annual review of plant biology*, *61*, 21-47.

214 Heelis, P. F., Parsons, B. J. & Yano, Y. (1997). Spectral and redox properties of benzodipteridine. A pulse radiolysis, laser flash photolysis and semi-empirical molecular orbital study. *Journal of the Chemical Society, Perkin Transactions 2*(4), 795-798.

215 Arnaut, L. G., & Formosinho, S. J. (1993). Excited-state proton transfer reactions I. Fundamentals and intermolecular reactions. *Journal of Photochemistry and Photobiology A: Chemistry*, *75*(1), 1-20.

216 Masuda, S., Hasegawa, K., Ishii, A., & Ono, T. A. (2004). Light-induced structural changes in a putative blue-light receptor with a novel FAD binding fold sensor of blue-light using FAD (BLUF); Slr1694 of Synechocystis sp. PCC6803. *Biochemistry*, *43*(18), 5304-5313.

217 Kraft, B. J., Masuda, S., Kikuchi, J., Dragnea, V., Tollin, G., Zaleski, J. M., & Bauer, C. E. (2003). Spectroscopic and mutational analysis of the blue-light photoreceptor AppA: a novel photocycle involving flavin stacking with an aromatic amino acid. *Biochemistry*, *42*(22), 6726-6734.

218 Fukushima, Y., Okajima, K., Shibata, Y., Ikeuchi, M., & Itoh, S. (2005). Primary intermediate in the photocycle of a blue-light sensory BLUF FAD-protein, Tll0078, of Thermosynechococcus elongatus BP-1. *Biochemistry*, *44*(13), 5149-5158.

219 Gomelsky, M., & Klug, G. (2002). BLUF: a novel FAD-binding domain involved in sensory transduction in microorganisms. *Trends in biochemical sciences*, *27*(10), 497-500.

220 Barends, T. R., Hartmann, E., Griese, J. J., Beitlich, T., Kirienko, N. V., Ryjenkov, D. A., Reinstein, J., Shoeman, R.L., Gomelsky, M. & Schlichting, I. (2009). Structure and mechanism of a bacterial light-regulated cyclic nucleotide phosphodiesterase. *Nature*, *459*(7249), 1015-1018.

221 Anderson, S., Dragnea, V., Masuda, S., Ybe, J., Moffat, K., & Bauer, C. (2005). Structure of a novel photoreceptor, the BLUF domain of AppA from Rhodobacter sphaeroides. *Biochemistry*, *44*(22), 7998-8005.

222 Götze, J., & Saalfrank, P. (2009). Serine in BLUF domains displays spectral importance in computational models. *Journal of Photochemistry and Photobiology B: Biology*, *94*(2), 87-95.

223 Masuda, S., Tomida, Y., Ohta, H., & Takamiya, K. I. (2007). The critical role of a hydrogen bond between Gln63 and Trp104 in the blue-light sensing BLUF domain that controls AppA activity. *Journal of molecular biology*, *368*(5), 1223-1230.

224 Jung, A., Reinstein, J., Domratcheva, T., Shoeman, R. L., & Schlichting, I. (2006). Crystal structures of the AppA BLUF domain photoreceptor provide insights into blue light-mediated signal transduction. *Journal of molecular biology*, *362*(4), 717-732.

225 Bonetti, C., Stierl, M., Mathes, T., Van Stokkum, I. H., Mullen, K. M., Cohen-Stuart, T. A., Van Grondelle, R., Hegemann, P. & Kennis, J. T. (2009). The role of key amino acids in the photoactivation pathway of the Synechocystis Slr1694 BLUF domain. *Biochemistry*, *48*(48), 11458-11469.

226 Fujisawa, T., Takeuchi, S., Masuda, S., & Tahara, T. (2014). Signaling-state formation mechanism of a BLUF protein PapB from the purple bacterium Rhodopseudomonas palustris studied by femtosecond time-resolved absorption spectroscopy. *The Journal of Physical Chemistry B*, *118*(51), 14761-14773.

227 Masuda, S., Hasegawa, K., & Ono, T. A. (2005). Light-induced structural changes of apoprotein and chromophore in the sensor of blue light using FAD (BLUF) domain of AppA for a signaling state. *Biochemistry*, *44*(4), 1215-1224.

228 Domratcheva, T., Grigorenko, B. L., Schlichting, I., & Nemukhin, A. V. (2008). Molecular models predict light-induced glutamine tautomerization in BLUF photoreceptors. *Biophysical journal*, *94*(10), 3872-3879.

229 Stelling, A. L., Ronayne, K. L., Nappa, J., Tonge, P. J., & Meech, S. R. (2007). Ultrafast structural dynamics in BLUF domains: transient infrared spectroscopy of AppA and its mutants. *Journal of the American Chemical Society*, *129*(50), 15556-15564.

230 Mathes, T., & Götze, J. P. (2015). A proposal for a dipole-generated BLUF domain mechanism. *Frontiers in molecular biosciences*, *2*.

231 Rieff, B., Bauer, S., Mathias, G., & Tavan, P. (2011). DFT/MM description of flavin IR spectra in BLUF domains. *The Journal of Physical Chemistry B*, *115*(38), 11239-11253.

232 Götze, J. P., Greco, C., Mitrić, R., Bonačić-Koutecký, V., & Saalfrank, P. (2012). BLUF hydrogen network dynamics and UV/Vis spectra: a combined molecular dynamics and quantum chemical study. *Journal of computational chemistry*, *33*(28), 2233-2242.

233 Khrenova, M., Domratcheva, T., Grigorenko, B., & Nemukhin, A. (2011). Coupling between the BLUF and EAL domains in the blue light-regulated phosphodiesterase BlrP1. *Journal of molecular modeling*, *17*(7), 1579-1586.

234 Obanayama, K., Kobayashi, H., Fukushima, K., & Sakurai, M. (2008). Structures of the chromophore binding sites in BLUF domains as studied by molecular dynamics and quantum chemical calculations. *Photochemistry and photobiology*, *84*(4), 1003-1010.

235 Meier, K., Thiel, W., & van Gunsteren, W. F. (2012). On the effect of a variation of the force field, spatial boundary condition and size of the QM region in QM/MM MD simulations. *Journal of computational chemistry*, *33*(4), 363-378.

236 García, A. E., & Sanbonmatsu, K. Y. (2002). α-Helical stabilization by side chain shielding of backbone hydrogen bonds. *Proceedings of the National Academy of Sciences*, *99*(5), 2782-2787.

237 Schneider, C., & Sühnel, J. (1999). A molecular dynamics simulation of the flavin mononucleotide–RNA aptamer complex. *Biopolymers*, *50*(3), 287-302.

238 Udvarhelyi, A., & Domratcheva, T. (2011). Photoreaction in BLUF Receptors: Proton-coupled Electron Transfer in the Flavin-Gln-Tyr System. *Photochemistry and photobiology*, *87*(3), 554-563.