# IN-SILICO GUIDED PREDICTION OF ALLOSTERIC SITES ON PROTEINS: APPLICATION TO CYCLIN-DEPENDENT KINASE 2

**JOE G GREENER**

A dissertation submitted for the degree of Doctor of Philosophy of Imperial College London

*Supervisor*
Prof Michael JE Sternberg

*Co-supervisors*
Dr David J Mann
Prof Alan Armstrong

Structural Bioinformatics Group
Department of Life Sciences
Imperial College London
South Kensington
London SW7 2AZ

# Copyright declaration

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work.

# Declaration of originality

I hereby declare that, except where otherwise acknowledged and appropriately referenced, the material presented in this thesis is my own work. In particular, all figures stated as being based on figures from Greener and Sternberg 2015 or Greener et al. 2017 were originally produced by me.

**Abstract**

Allostery is the functional change at one site on a protein caused by a change at a distant site. Despite being discovered more than 50 years ago allostery has remained a mystery, and there is no unified scheme to understand and predict it. In order for the the benefits of allostery to be taken advantage of, both for basic understanding of proteins and to develop new classes of drugs, computational methods to predict allosteric sites on proteins need to be developed and validated. This thesis introduces two computational methods to predict allosteric sites on proteins and describes experiments to validate a predicted allosteric site.

The concepts of allostery, allosteric prediction, the protein structural ensemble and protein kinases are explained. AlloPred, which uses perturbation of normal modes alongside pocket descriptors in a machine learning approach to predict allosteric pockets, is introduced. AlloPred shows comparable and complementary performance to two existing methods. The AlloPred web server allows visualisation and analysis of predictions.

ExProSE (Exploration of Protein Structural Ensembles), a distance geometry-based method that generates an ensemble of protein structures from two input structures, is described. ExProSE is able to access conformational changes inaccessible to classical molecular dynamics. By adding additional constraints to the method, the effect of allosteric modulators can be predicted. ExProSE is shown to be effective at allosteric site prediction in a systematic comparison of methods.

A predicted allosteric pocket on cyclin-dependent kinase 2, a medically-important protein kinase involved in cell cycle regulation, is explored experimentally. Selected compounds from virtual screening are tested in two assays, though no conclusive results are obtained. The development and adoption of methods such as those presented here is essential or the long-preached potential of allostery will remain elusive.

# Acknowledgements

This project was funded by Biotechnology and Biological Sciences Research Council grant BB/J014575/1.

I would like to thank the many people without whom this work would not have happened. My supervisor Mike Sternberg always offered support and allowed me to pursue my own ideas. The whole of the Sternberg group provided discussions and a friendly environment - particularly Ioannis for making the work in Chapter 3 so much better, Suhail for keeping the computers working, Alessia for the kind words, Lawrence and Stefans for the pub chats, and Alex and Matt for the lunches over the years. My co-supervisor David Mann and his group welcomed me in and taught me how to do experiments - Greg in particular was a great teacher, sorry for all the things I broke. I had an enjoyable 3 months at BenevolentAI on my industry placement and want to thank everyone there for the opportunity. Thanks also to my progress review panel, Alfonso De Simone and Matthew Fuchter, my other co-supervisor Alan Armstrong, the Stumpf group, and the open source contributors whose software I used.

I would also like to thank my family and friends, who have been so supportive over the years. My parents and siblings always believed in me more than I believed in myself. I have known my school friends for 14 or more years now and they still manage to surprise me. My uni friends have provided such friendship and good times. Various sports clubs have offered a very direct form of therapy. And to Joanna, for making the last year better with no $p$-value required.

# Contents


**1 Introduction** **12**

1.1 Overview of this thesis . . . . . 13
1.2 Allostery in proteins . . . . . 14
1.2.1 Computational methods for allosteric prediction . . . . . 15
1.2.2 Experimental methods for allosteric prediction . . . . . 22
1.2.3 Cryptic allosteric sites . . . . . 23
1.2.4 Design of allosteric sites . . . . . 24
1.2.5 Discussion of allostery . . . . . 24
1.2.6 Challenges in allosteric prediction . . . . . 26
1.3 Protein structural ensembles . . . . . 28
1.4 Protein kinases . . . . . 31
1.5 Cyclin-dependent kinase 2 . . . . . 34
1.6 Aims . . . . . 37

**2 AlloPred** **38**

2.1 Materials and Methods . . . . . 39
2.1.1 Data selection . . . . . 39
2.1.2 Pocket prediction . . . . . 39
2.1.3 Normal mode analysis . . . . . 40
2.1.4 Machine learning . . . . . 43
2.1.5 Web server . . . . . 46
2.2 Results . . . . . 47
2.2.1 Validation . . . . . 47
2.2.2 Web server . . . . . 48

2.3 Discussion . . . . . 52

**3 ExProSE** **55**

3.1 Materials and Methods . . . . . 56
3.1.1 Distance constraint generation . . . . . 56
3.1.2 Protein structure generation . . . . . 60
3.1.3 Ensemble analysis . . . . . 63
3.1.4 Modulator constraint generation . . . . . 64
3.1.5 Datasets . . . . . 65
3.1.6 Method comparison . . . . . 66
3.2 Results . . . . . 69
3.2.1 Ensemble generation . . . . . 69
3.2.2 Ensemble perturbation for CDK2 . . . . . 83
3.2.3 Allosteric site prediction . . . . . 85
3.2.4 Dynamic allostery in CAP . . . . . 88
3.2.5 Activation by phosphorylation . . . . . 91
3.3 Discussion . . . . . 94

**4 CDK2** **98**

4.1 Materials and Methods . . . . . 99
4.1.1 Bioinformatics resources . . . . . 99
4.1.2 Virtual screening . . . . . 99
4.1.3 Reagents and compounds . . . . . 99
4.1.4 Purification of cyclin A2 . . . . . 100
4.1.5 TR-FRET assay . . . . . 100
4.1.6 Binding assay . . . . . 101
4.2 Results . . . . . 102
4.2.1 Virtual screening . . . . . 103
4.2.2 Compound selection . . . . . 106
4.2.3 Experimental aims . . . . . 107
4.2.4 Purification of cyclin A2 . . . . . 107
4.2.5 TR-FRET assay . . . . . 110
4.2.6 Binding assay . . . . . 112
4.3 Discussion . . . . . 115

**5 Conclusion** **118**

**Appendices** **120**

AlloPred documentation . . . . . 121

ExProSE documentation . . . . . 123

BioJulia Bio.Structure module . . . . . 128

**References** **133**

# List of Figures


1.1 The current conception of allostery as a property of the conformational ensemble . . . 16
1.2 Steps in normal mode analysis . . . 19
1.3 Timescales of protein motions . . . 29
1.4 Binding sites of known type IV allosteric inhibitors of protein kinases . . . 33
1.5 Structure of CDK2 . . . 35

2.1 The change in flexibility of a protein on modulator binding at an allosteric site used as the basis for AlloPred . . . 41
2.2 Method used by AlloPred to calculate how much of an effect a change in the normal modes has at the active site . . . 42
2.3 Venn diagram comparing the performance of AlloPred on the test set to existing methods . . . 49
2.4 Flowchart showing the stages involved in running a job submitted to the AlloPred web server . . . 50
2.5 Screenshots of the AlloPred web server results page . . . 51

3.1 Calculation of the upper and lower distance constraints between two atoms given two input protein structures . . . 59
3.2 Overview of the ExProSE computational method to generate and perturb ensembles of protein structures . . . 70
3.3 Convergence of the ExProSE SPE error score for 5 separate runs on CDK2 . . . 71

3.4 SPE error score plotted against PROCHECK overall G-factor for an ExProSE ensemble . . . 72
3.5 Structures and principal components analysis of T4-lysozyme ensembles . . . 78
3.6 Ensemble generation for T4-lysozyme with different parameters . 80
3.7 T4-lysozyme ensembles generated using MD and targeted MD . . 82
3.8 Closest models from each ensemble to T4-lysozyme crystal structures . . . 84
3.9 CDK2 pockets and projections of perturbed ExProSE ensembles . 86
3.10 Mean square fluctuations across perturbed CAP ensembles compared to Apo-CAP . . . 92
3.11 Ensembles of NtrC generated with ExProSE . . . 93

4.1 Conformational variability and virtual screening of CDK2 . . . . 104
4.2 Chemical structures of selected compounds to screen experimentally against a potential allosteric site on CDK2 . . . . 109
4.3 SDS-PAGE results for purification of cyclin A2 . . . . 111
4.4 TR-FRET assay principles and results . . . . 113
4.5 Results of binding assay with selected compounds . . . . 114

5.1 Hierarchy of types in the BioJulia Bio.Structure module . . . . 130
5.2 Example functionality of the Bio.Structure module in the Jupyter Notebook . . . . 131
5.3 Benchmarks on common tasks for open source packages to read and manipulate PDB files in various programming languages . . 132

# List of Tables

1.1 Computational allosteric prediction methods currently available to run locally or as a web server . . . . . 17

3.1 Interaction types between atom pairs in ExProSE . . . . . 58
3.2 Ability of ensemble generation methods to sample conformational space . . . . . 74
3.3 Ability of ExProSE ensembles to sample conformational space . . 75
3.4 Improvement in stereochemical quality of generated structures on energy minimisation . . . . . 81
3.5 Performance of allosteric site prediction methods on a dataset of 58 known allosteric proteins - summary . . . . . 88
3.6 Performance of allosteric site prediction methods on a dataset of 58 known allosteric proteins - detail . . . . . 89

4.1 Selected compounds to screen experimentally against a potential allosteric site on CDK2 . . . . . 108

5.1 Comparison of open source packages to read and manipulate PDB files in various programming languages . . . . . 129

# Abbreviations

| | |
|---|---|
| **ANS** | 8-anilino-1-naphthalene sulfonate |
| **ASD** | AlloSteric Database (`http://mdl.shsmu.edu.cn/ASD`) |
| **CAP** | Catabolite activator protein |
| **CDK2** | Cyclin-dependent kinase 2 |
| **CSA** | Catalytic Site Atlas (`http://www.ebi.ac.uk/thornton-srv/databases/CSA`) |
| **ENM** | Elastic network model |
| **ExProSE** | Exploration of Protein Structural Ensembles |
| **FRET** | Förster resonance energy transfer |
| **GPCR** | G protein-coupled receptor |
| **IDP** | Intrinsically-disordered protein |
| **KNF** | Koshland-Némethy-Filmer (model of allostery) |
| **MD** | Molecular dynamics |
| **MWC** | Monod-Wyman-Changeux (model of allostery) |
| **NMA** | Normal mode analysis |
| **NMR** | Nuclear magnetic resonance |
| **NtrC** | Nitrogen regulatory protein C |
| **PC** | Principal component |
| **PCA** | Principal components analysis |
| **PDB** | Protein Data Bank (`https://www.rcsb.org`) |
| **PRS** | Perturbation response scanning |
| **SPE** | Stochastic Proximity Embedding |
| **SPR** | Surface plasmon resonance |
| **SVM** | Support vector machine |
| **TR-FRET** | Time-resolved Förster resonance energy transfer |
| **TrpAB** | Tryptophan synthase |

# Chapter 1

# Introduction

This chapter introduces the thesis along with the concepts of allostery in proteins, structure-based allosteric prediction, normal mode analysis (NMA), the protein structural ensemble, protein kinases and cyclin-dependent kinase 2 (CDK2). Further details on concepts and methods will be described in later chapters as required.

## 1.1 Overview of this thesis

The aim of this thesis is to contribute to the field of allosteric site prediction by developing computational methods and testing them experimentally. A summary of the concept of allostery in proteins and the recent emergence of approaches to predict allostery will be given in Chapter 1. The protein structural ensemble and methods to generate ensembles will be introduced. Protein kinases are an important family of proteins that are amenable to allosteric regulation. Protein kinases in general, and CDK2 in particular, will be described.

Chapters 2 and 3 describe two independent computational methods to predict allosteric sites on proteins, AlloPred and ExProSE. AlloPred uses NMA and machine learning to rank pockets in terms of their allosteric character. ExProSE uses distance geometry to generate structural ensembles that can be perturbed to explore dynamics and allostery. The approach and validation of each method are discussed in the respective chapters.

Chapter 4 describes the further testing of a potential allosteric site on CDK2 predicted as allosteric. Virtual screening led to experimental assays being carried out on selected compounds. Some conclusions are drawn in Chapter 5. Finally, there are the appendices. The technical documentation for the AlloPred and ExProSE source codes is attached along with the description of a module to read, write and manipulate protein structure files in the Julia programming language.

## 1.2 Allostery in proteins

Allostery in its broadest sense is the functional change at one site on a protein caused by a change at a distant site. The perturbation at the allosteric site can be non-covalent binding of a molecule (e.g. small molecule, ions, RNA, DNA), covalent binding (e.g. phosphorylation) or light absorption [1]. Changes in structure or dynamics lead to effects such as a reduction or increase in catalytic activity, changes in disordered regions or changes in oligomerisation state. All proteins are potentially allosteric [2]. This intrinsic and widespread property of proteins is important in processes such as cellular signalling and disease, yet most allosteric mechanisms remain an enigma and the foundational principles are not fully understood [3].

Since the first discovery of allosteric systems more than 50 years ago there have been various models put forward to describe the phenomenon. The dominant proposals for many years were the Monod-Wyman-Changeux (MWC) model, which posited that pre-existing states are subject to an equilibrium shift on modulator binding [4], and the Koshland-Némethy-Filmer (KNF) model, which advanced the idea that there was an induced fit of a binding site on interaction with a modulator [5]. The structural view of allostery [6], which aimed to elucidate the allosteric mechanism by finding structural changes on effector binding, began to fill the gaps left by the phenomenological MWC and KNF descriptions. The discovery that entropic contributions to allostery can be significant predicted the phenomenon of allostery without conformational change [7], where the allosteric effect is communicated by a change in protein dynamics rather than protein structure [3].

More recently these views on allostery have been revisited and reconciled in approaches that focus on the ensemble of conformational states that proteins exist in [3, 8, 9, 10]. Figure 1.1 outlines the current understanding of allostery. A perturbation at any site in the structure leads to a shift in the occupancy of states by the population, so allostery is a property of the conformational ensemble. The effect at the allosteric site is linked to the active site by small conformational changes that transmit the allosteric effect in a wave-like man-

ner along pathways of amino acids in the protein [11]. These pathways may be conserved by evolution [12]. It is also important to consider the effect of allostery on cellular networks and reaction pathways [1], with allosteric effects propagating via protein-protein interactions.

Allosteric drugs have hardly been explored and are a major avenue of research for the pharmaceutical industry [13]. They hold many potential benefits over orthosteric (non-allosteric) drugs: they do not bind to active sites that are often conserved in protein families, and are hence highly specific; they can activate as well as inhibit a protein; they can have a ceiling to their effect; and they can be used effectively in combination with orthosteric drugs [13]. The discovery of positive and negative allosteric modulators of G protein-coupled receptors (GPCRs) has highlighted these advantages [14, 15]. GPCR modulators include the positive regulator cinacalcet at the $Ca^{2+}$-sensing receptor and the negative modulator maraviroc of the chemokine CCR5 [1]. Allosteric modulators have been elucidated for targets as diverse as the GABA receptor, hepatitis C virus polymerase and RNA. Numerous other allosteric modulators are in various stages of human clinical trails. However, discovery of allosteric drugs presents challenges beyond those encountered in orthosteric drug discovery - see later.

In order to understand and utilise allostery it is necessary to be able to predict allosteric sites, allosteric modulators and residues involved in propagating the allosteric signal. Advances from the last few years in the structure-based prediction of protein allostery are outlined here, largely focusing on computational approaches. Review papers have covered similar topics [16, 17, 18, 19]. The emerging fields of cryptic allosteric site discovery and allosteric site design are described and some of the issues surrounding allostery are discussed.

### 1.2.1 Computational methods for allosteric prediction

The last few years have seen the emergence of the first general methods that predict allostery based on protein structure. Table 1.1 summarises these methods, many of which are available as web servers.

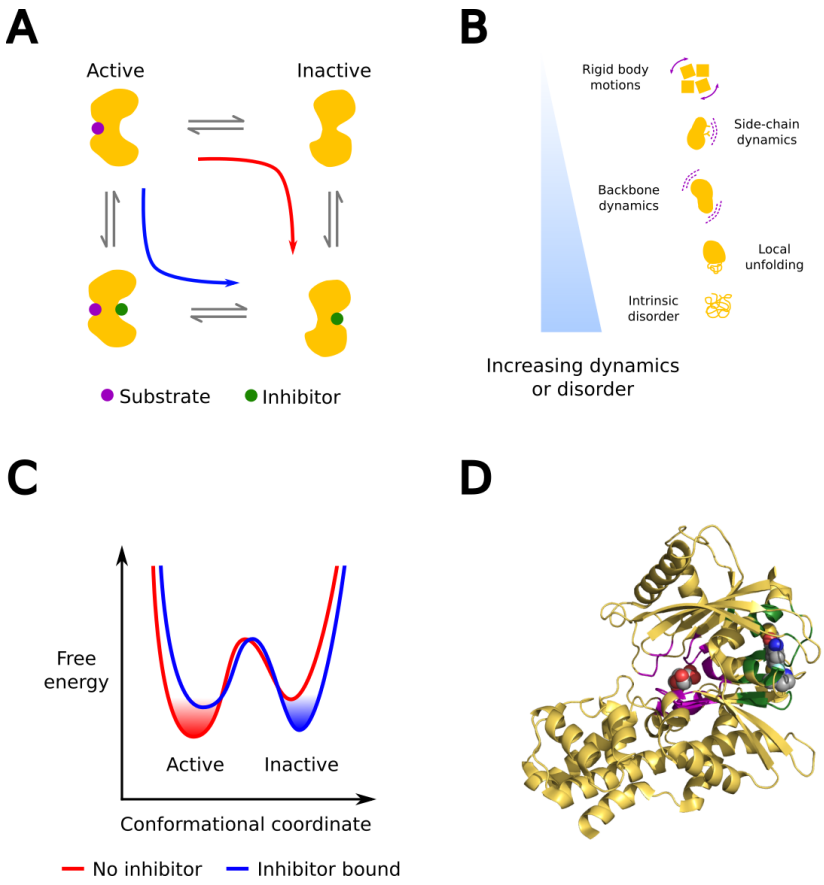



**Figure 1.1** The current conception of allostery. (A) A two-state model of allostery where a protein has an active and an inactive conformation. In the presence of the allosteric inhibitor the inactive state is favoured either by the inhibitor binding to the protein when it is in the inactive state (red arrow - conformational selection) or by the inhibitor binding to the active state and causing inactivation (blue arrow - induced fit). (B) The variety of motions that can lead to allostery. Larger motions or more disorder are shown further down the vertical axis. Figure based on Figure 2 from Motlagh et al. 2014 [3]. (C) A simplified representation of the change in the energy landscape on binding of an allosteric inhibitor. The shaded regions show the main occupied conformation in each case. On inhibitor binding the relative energies of the active and inactive states are altered. For example, disruption of a hydrogen bond could destabilise the active state and stabilise the inactive state. (D) Glucokinase, a well-studied example of allostery [20], is shown as a yellow cartoon. The glucose substrate and the allosteric modulator are shown as spheres coloured by element. The active site and allosteric site are coloured purple and green respectively.

| Name | Reference(s) | Output(s) | Web server available | Source code available online |
|---|---|---|---|---|
| AlloPred | [21] | Predicted allosteric pockets | `http://www.sbg.bio.ic.ac.uk/allopred/home` | Yes, MIT licence |
| AlloSigMA | [22] | Allosteric free energies | `http://allosigma.bii.a-star.edu.sg/home` | No |
| AlloSitePro | [23, 24] | Predicted allosteric pockets | `http://mdl.shsmu.edu.cn/AST` | No |
| AllosMod | [25] | Modelled energy landscapes | `http://modbase.compbio.ucsf.edu/allosmod` | No |
| ENM method | [26] | Residues coupled to normal modes | `http://enm.pitt.edu` | Partly as ProDy, MIT licence |
| ExProSE | [27] | Ensemble of protein structures, predicted allosteric pockets | No | Yes, MIT licence |
| MCPath | [28] | Allosteric communication pathways | `http://safir.prc.boun.edu.tr/clbet_server` | No |
| PARS | [29, 30] | Predicted allosteric pockets | `http://bioinf.uab.cat/pars` | No |
| SPACER | [31, 32] | Predicted allosteric residues, exploration of allosteric communication | `http://allostery.bii.a-star.edu.sg` | No |
| STRESS | [33] | Predicted surface-critical and interior-critical residues | No | Yes |

**Table 1.1** Computational allosteric prediction methods currently available to run locally or as a web server, ordered alphabetically. The methods presented in this thesis (AlloPred and ExProSE) are included. In addition there are various pocket prediction methods that aim to predict binding pockets on proteins, but not specifically allosteric pockets [34, 35, 36], that are not included here.

**Normal mode analysis**

In NMA the structural fluctuations of a protein around an equilibrium conformation are decomposed into harmonic orthogonal modes [37]. Each mode has all parts of the system moving sinusoidally, in phase and with the same frequency. All observed configurations of the system can be generated from a linear combination of its normal modes. The process of NMA is shown in Figure 1.2. The normal modes are found by diagonalising the Hessian matrix - the matrix of second derivatives of the potential energy with respect to the mass-weighted atomic coordinates. NMA is effective at describing protein dynamics, despite ignoring the complex nature of the protein energy landscape [38]. Even considering the $C^{\alpha}$ atoms alone as a network of balls and springs can be sufficient, meaning that NMA is a computationally efficient way of exploring protein flexibility compared to molecular dynamics (MD). This simplification is called the elastic network model (ENM) and the mathematical formulation is given in Section 2.1. The long-range nature of allosteric communication is often well-described by low-frequency modes that involve the motion of many atoms. However, allostery does involve local effects so higher-frequency modes should also be taken into account [39].

The binding leverage approach [32] predicts how ligand binding couples to the intrinsic motions of a protein. Sites with high binding leverage are predicted to be allosteric. Binding leverage was developed into the web server SPACER [31], and into the general predictor STRESS [33] by a different group. The PARS method [30, 29] calculates normal modes in the presence and absence of a simulated allosteric modulator. If the motions are significantly different the site is predicted as allosteric. The DynOmics ENM server [26] finds hinge residues that control the two slowest normal modes of a protein, and hence are able to influence its dynamics. Several studies have used NMA to model allosteric regulation in specific sets of proteins [40, 41, 42, 43]. NMA is suitable for high-throughput, automated approaches as it can be computationally inexpensive. However whilst NMA-based methods might be expected to reveal perturbations to vibrations, the assumption of harmonic fluctuations around an energetically-minimum structure means that other contributing motions to

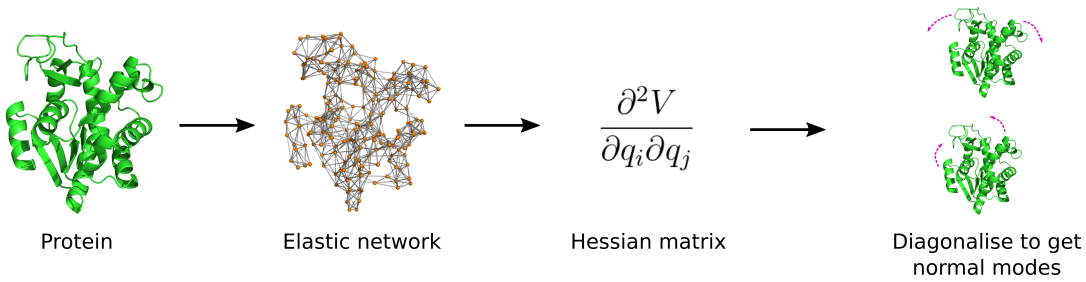



**Figure 1.2** Steps in NMA using the ENM. A protein structure is converted into an elastic network where balls are $C^{\alpha}$ atoms and springs are between $C^{\alpha}$ atoms within a certain distance. A Hessian matrix of second derivatives of the potential energy $V$ in terms of the mass-weighted atomic coordinates $q_i$ is derived from the potential energy function of this elastic network. Diagonalising the Hessian matrix yields eigenvectors and eigenvalues that correspond to the displacements for each $C^{\alpha}$ atom in a mode and the square of the mode frequency respectively. See [37] for more.

allostery such as local unfolding and rigid body movements [3] are not taken into account.

### Machine learning

A few methods have used machine learning to predict allostery. AlloSite [24] uses a support vector machine (SVM) and features from Fpocket [35] to re-rank pockets in terms of their allosteric character. However the results are often found to be similar to the Fpocket ranking, showing the difficulty of distinguishing pockets that have specific allosteric character from those that are generally suitable for ligand binding. This method has been updated as AlloSitePro to include NMA perturbation [23]. A Random Forest approach [44] uses descriptors for binding sites and associated ligands to assign protein cavities as allosteric, regular or orthosteric.

### Molecular dynamics

MD remains the standard computational tool for structural analysis when structures are available. A study on the nitrogen regulatory protein C (NtrC) com-

bined MD simulations and nuclear magnetic resonance (NMR) data to explore the free energy landscape and investigate at atomic resolution the transition from active to inactive state [45]. Perturbation response scanning (PRS), in which the response of the structure to random perturbations at specific positions is examined, is a popular tool for allosteric prediction. For example, allosteric hotspot residues were predicted using PRS for the chaperone Hsp70 [46]. Weinkam et al. constructed energy landscapes and explored them with MD [25]. They were able to study the allosteric mechanisms involved in three proteins. The method is available as the AllosMod web server.

**Evolutionary methods**

Classic work has shown that allosteric communication can be mediated by networks of residues conserved by evolution [47, 12]. One study developed previous work on protein sectors, groups of co-evolving residues physically contiguous in structure, to link sector-connected surface sites to allosteric sites [48]. A recent approach found that surface and interior critical residues tend to be conserved [33]. The recent discovery that most directly co-evolving residues distant in 3D structure are close in related structures or assemblies [49] brings into question the concept of allosteric and active sites that directly co-evolve. Other studies have found structural conservation in allosteric pockets but a lack of sequence conservation [50, 51], positing that allostery makes use of pre-existing features. As more structural and conservation information is acquired it will be important to discover to what degree allostery in proteins is a result of selection on specific pathways, and to what degree novel allostery can be discovered on proteins in the absence of previous evolutionary pressure.

**Other methods**

A recent study [52] constructs an all-atom graph and calculates for each bond the bond propensity, the strength of coupling to the active site through the graph. The method is used to reproduce observed results for three proteins

in detail and is also able to predict allosteric sites in a dataset of 20 allosteric proteins. Coarse-grained simulation approaches such as a two-state Gō model have also been utilised. In one approach an ensemble of structures is generated and the response of the ensemble distribution to an effector is used to predict allostery [53].

**Methods not specific to allostery**

The identification of binding sites on the protein surface is a problem that has long pre-dated the search for pockets that are specifically allosteric. These methods are however useful in the structure-based prediction of allostery - the identification of a high-affinity binding site distant from a known active site could present an opportunity for allosteric regulation, for example. The FTMap family of web servers [54] predicts ligand-binding hotspots using small organic molecules as probes on the protein surface. By using mixed-solvent MD this principle has been extended to the prediction of allosteric sites in particular, with success on some test cases [55].

Common pocket prediction methods such as LIGSITE$^{csc}$ [34] and Fpocket [35] are able to find pockets on a protein large enough to bind small molecules, and these often correspond to allosteric sites [24]. These methods will be used throughout this thesis, and their ability to predict allosteric pockets (either alone or when re-ranked by allosteric prediction methods) will be determined. Fpocket uses the concept of alpha spheres (spheres touching four atoms at an equal distance), related to Voronoi tessellation (the separation of space by lines equidistant to two points), to find clusters of spheres of a certain size that correspond to pockets. LIGSITE$^{csc}$ adds the concept of conservation of surface residues to the original LIGSITE algorithm [56], which finds cavities using simple operations on a cubic grid. Recently, deep learning has been used to predict binding sites [57].

**Allosteric pathway prediction**

Allosteric signals can be propagated by multiple communication pathways [11]. Understanding these pathways is necessary in order to predict sites that are able to communicate with the active site [58]. A machine learning approach to predict residues involved in allosteric communication uses a variety of structural and network features and is able to predict these hotspots with reasonable accuracy [59]. A different approach, McPath, uses a Monte Carlo algorithm to define likely allosteric pathways by examining inter-residue interactions in a residue network [28]. A study that added an allosteric domain to a protein analysed residue contact maps to find loops mechanically-coupled to the active site [60]. An investigation on the PDZ domain using MD found that allosteric changes are non-linear and occur in a non-local fashion, and are similar in many ways to protein folding [61]. This study highlights that the idea of a single set of connected residues transmitting the allosteric signal is not adequate to explain allosteric communication. In multimeric proteins it has been suggested that 'global communication networks' of quaternary and tertiary motions transmit allosteric signals, and that considering these motions separately is inadequate [62]. Successful frameworks for predicting allosteric communication should take this into account.

### 1.2.2 Experimental methods for allosteric prediction

Experimental studies such as crystallography, NMR and site-directed mutagenesis remain the best tools for exploring allostery in a particular protein. A synthetic azetidine derivative that kills Mycobacterium tuberculosis through allosteric inhibition of tryptophan synthase (TrpAB), a previously untargeted enzyme, was found by a high-throughput screen [63]. The inhibition is not easily overcome by changes in metabolic environment due to the modulator binding at the TrpAB $\alpha$-$\beta$-subunit interface and affecting multiple steps in the overall reaction of the enzyme. A study on the proteasome [64] crystallised the complex in the presence and absence of an allosteric modulator. Having the active and inactive structures allowed the authors to propose a detailed mechanism of

inactivation, which has implications for future allosteric proteasome inhibitors. A study on flavovirus protease [65] used a virtual screen to select 29 potential allosteric compounds that were tested experimentally. One showed an ability to inhibit the conformational change and also inhibit flavovirus growth. Allosteric pathways in ERG proteins were proposed using fluctuation correlation data and validated by mutating residues in the pathways [66]. However, there are limits to the use of mutational studies to validate allosteric mechanisms. It has been found that mutational data can give evidence for a deliberately poorly-conceived allosteric mechanism [67]. In the future it is to be hoped that experimental screens specifically for allosteric sites [68, 69, 70, 71, 72] become more widespread, opening the path to conventional large-scale screens for allosteric drugs.

### 1.2.3 Cryptic allosteric sites

The discovery of cryptic binding pockets - pockets that are only available in some conformations of the protein and may not have an associated experimental structure - has the potential to vastly increase the number of druggable sites on proteins [73] and is directly relevant to allosteric prediction. A recent study [74] showed using NMR data that ligands of the LpxC enzyme access a cryptic site that is invisible to crystallography. One study used Markov state modelling and MD to predict multiple hidden allosteric sites on $\beta$-lactamase and tested these using thiol labelling experiments [75], later finding modulators for the sites [76].

The general approach CryptoSite uses machine learning to predict cryptic pockets on proteins using sequence and structural features [36]. However, two problems affect the use of cryptic allosteric pockets over allosteric sites where the pocket is present in most or all conformations. Firstly, the shape of the pocket is not known so rational drug design is difficult. Secondly, there is potentially an energetic cost associated with the protein adopting the conformation required for the cryptic pocket [77]. However, the discovery of ligands with inhibition constants in the low picomolar range in the above study [74] show that these

sites are druggable. Further computational and experimental studies are required to explore this promising area.

### 1.2.4 Design of allosteric sites

The rational design of allosteric sites is a problem closely related to structure-based prediction of allostery. Introducing allosteric sites into existing proteins, or creating fusion proteins to add activity switches, has many potential applications including in biotechnology [78]. A recent study added a PDZ domain into the Cas9 protein at a site that did not disrupt enzyme action [79]. The protein showed modulator-dependent activity in cells, establishing a system for Cas9 activation.

Another study created fusion proteins that use conformational entropy to respond to temperature or pH as a switch [80]. Taylor et al. engineered *E. coli* LacI to respond to one of four new inducer molecules using computational design and mutagenesis [81]. Dagliyan et al. designed a protein with a unique topology, uniRapR, whose conformation is controlled by the binding of a small molecule [82]. The switching and control ability of uniRapR was confirmed in silico, in vitro, and in vivo. uniRapR was used as an artificial regulatory domain to control activity of kinases as a proof of concept. The same group built on this and inserted the light-sensitive LOV2 domain into 3 proteins at non-conserved, surface-exposed loops identified computationally using residue contact analysis as being allosterically coupled to active sites [60].

### 1.2.5 Discussion of allostery

It is challenging to compare different methods for allosteric prediction. The different inputs and, more commonly, outputs make systematic comparisons difficult. One of the Critical Assessment of Genome Interpretation challenges in 2015-16 focused on predicting the influence of mutations on the allosteric regulation of human liver pyruvate kinase [83]. However the uptake was limited to four groups and the predictive ability was marginally better than random.

In the long run a dedicated community-wide initiative similar to the Critical Assessment of Structure Prediction [84] would be beneficial to the field of allosteric prediction.

One factor holding allosteric prediction back is the lack of a varied and robust set of benchmarks to test methods against. ASBench [85] is a curated set of allosteric proteins, and has been used for example to benchmark the method presented in Chapter 2. It is a subset of the AlloSteric Database (ASD, `http://mdl.shsmu.edu.cn/ASD`) [86], which was set up in 2011. ASD v3.0 contains over 1,400 proteins and 70,000 modulators and also includes allosteric mechanisms, allosteric networks of proteins and 'allosteromes' of the allostery involved in protein kinases and GPCRs. The allosteric proteins in the ASD are those that have experimental evidence for allostery. The increasing number of entries in this database shows that a large number of proteins have allosteric character, and implies that many proteins have allosteric character yet to be discovered. Improvements in such resources are necessary to prevent the developers of new methods having to assemble their own datasets [30, 32, 52] and to allow systematic comparisons between methods.

An issue that requires more study in the field of allosteric prediction is the exact relationship between an allosteric modulator and whether it acts as an activator or inhibitor. It has been shown that under different conditions the same allosteric modulator can have opposite effects [87]. Another viewpoint is the anchor/driver model of allostery, with the concept of a pushing or pulling driver determining which way the ligand acts [88]. An approach to study this would be a quantitative structure activity relation-like study where a variety of modulators and conditions are explored on the same protein. This would give evidence as to whether small structural differences causing a pushing or pulling effect are enough to reliably switch activator/inhibitor action.

The mechanism of dynamic allostery, where the allosteric effect is transmitted through changes in dynamics and the average structure does not necessarily change, also requires further investigation. While experimental studies [89, 90, 63] have found evidence for dynamic allostery, Nussinov and Tsai [91] warn that an apparent lack of conformational change can be an artefact of

various factors such as crystal packing, crystallisation conditions, disorder to order transitions, incremental activation, synergy between allosteric sites and changes in oligomeric state. A recent MD study proposes that allostery in the well-studied PDZ domain is driven by changes in electrostatic effects rather than solely changes in dynamics [92, 93]. The role of water in allostery also needs to be further explored as evidence has been found that re-arrangement of water molecules is a possible mechanism of allostery [94, 52]. Studies that combine experimental and computational approaches [64, 95] are well-suited to exploring these issues.

### 1.2.6 Challenges in allosteric prediction

There are a number of challenges faced in the structure-based prediction of protein allostery:

1. As shown in Figure 1B, an allosteric effect can arise from a variety of different mechanisms. A general predictor would have to account for these in a unified manner. This is particularly challenging when disorder is involved, as approaches based on a defined structure are less applicable. Some approaches to studying disorder and allostery have been proposed [96, 97].

2. The conformational changes that cause allostery are often large enough to occur on timescales of microseconds or milliseconds. This makes them too computationally expensive to study using MD without the use of accelerated or targeted MD. NMA is more computationally feasible but the assumption of a harmonic motion around an energy minimum does not correspond well to two distinct states with differing conformations.

3. The properties of active site pockets and small molecules that target the active site have been well-studied, for example Lipinski's rule of five [98]. Allosteric pockets and modulators may have generally different properties that we are not yet fully aware of, so we do not know exactly what to look for [99, 100].

4. The effect of an allosteric modulator is difficult to predict and can range from activation to inhibition, partial or complete. This is in comparison to orthosteric drug discovery, where drug action is presumed to be by competitive inhibition at the active site.

5. The effort of researchers and the protein structural data available is biased towards certain types of protein, such as those relevant in disease. For example, the allostery of GPCRs has been studied in detail [14, 15]. There is a lack of protein structural data for important types of proteins such as membrane proteins and proteins with significant disorder, but these proteins have considerable potential to be allosteric [3]. There may be different mechanisms or approaches to prediction that are relevant to less-studied protein families. The development of experimental methods such as cryo-electron microscopy should go some way to resolve this discrepancy [95].

## 1.3 Protein structural ensembles

Proteins move on a variety of timescales, encompassing motions from the vibration of a single bond to the collective movement of whole domains [101, 102] - see Figure 1.3. X-ray crystallography provides a static view of the structure of proteins. However, when only static structures are available the dynamic processes crucial to protein function [103] are hard to elucidate. Experimental techniques to explore the dynamics of proteins, such as NMR [104], are sophisticated and time-consuming. Significant effort has gone into determining the protein structural ensemble when constraints such as those from NMR are available but a crystal structure is lacking, for instance due to disorder in the protein [105]. However, proteins with an ordered structure that can be crystallised still have significant dynamics and in this case it makes sense to use the crystallographic data to explore the dynamics computationally.

MD is a widespread computational method for predicting protein motions and generating ensembles of protein structures. It is effective at modelling motions up to the timescale of nanoseconds. However, the computational cost of modelling proteins on the scale of microseconds or milliseconds means MD is not suitable for larger-scale transitions. Advanced MD methods such as targeted or accelerated MD can overcome this sampling problem [106, 107], but these methods are not yet routinely applicable due to the parameterisation and analysis required for each protein. The force fields used in MD simulations are also only approximations of the actual interatomic interactions and this can lead to errors in predicted properties, particularly for highly dynamic systems [108].

Various non-MD methods have been used to generate ensembles of protein structures from a crystal input structure, and hence explore protein dynamics. CONCOORD [109, 110] is a distance geometry method to generate structures from an input structure and consists of a two-step process. First, the different types of chemical interactions in the input structure, e.g. H-bonding and hydrophobic interactions, are converted to distance constraints with a given tolerance. Next, an iterative minimisation procedure is performed to move a set of randomly-placed coordinates such that most distance constraints are

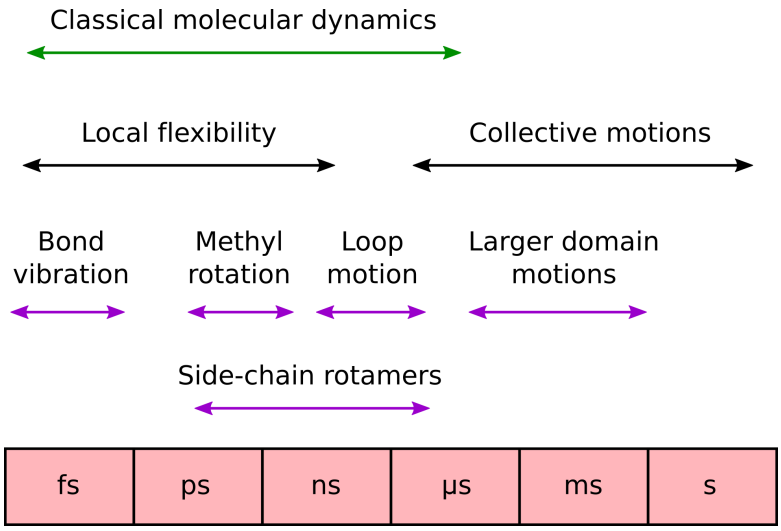



**Figure 1.3** Dynamic processes in proteins and their associated timescales. Also shown is the range of timescales that can be studied with classical MD. Figure based on Figure 1 from Henzler-Wildman and Kern 2007 [101].

satisfied. This generates a protein structure in a similar manner to the way a structure is produced from NMR constraints. The process is repeated to obtain an ensemble of structures. tCONCOORD extends CONCOORD and gives better sampling of proteins with large conformational changes by predicting H-bonds in the structure that are liable to break [111].

NMA is also well-suited to studying flexibility and conformational change in proteins [112], as discussed in Section 1.2. Conformations of proteins can be generated using NMA, usually by modelling the protein along relevant vibrations. The NMSim web server [113, 114] finds flexible and rigid protein regions using the graph theoretical approach FIRST [115], then generates conformations along low-frequency normal modes. This gives sampling similar to MD but is more computationally efficient.

Ensembles of protein structures have uses in flexible ligand docking [116], generating poses for protein-protein docking [117], predicting structures on trajectories between two crystal structures [118], and predicting flexible regions in proteins [114]. They can also be used to explore intrinsically-disordered proteins (IDPs) [119, 120]. As has been discussed in Section 1.2, allostery is a prop-

erty of the conformational ensemble. An ensemble generation method that can be modified to take allostery into account could be used to explore allostery in a protein - see Figure 1.1C. This is the main motivation for developing an ensemble generation method in this work.

## 1.4 Protein kinases

Protein kinases modify other proteins by chemically adding phosphate groups to them (phosphorylation). This leads to a functional change due to an alteration in enzyme activity, cellular location and/or protein-protein interactions. Protein kinases regulate almost all aspects of cellular physiology, from proliferation and generation of biomass to gene expression and protein production [121]. The human genome contains around 500 protein kinase genes ($\sim$2% of all human genes) and up to 30% of human proteins are modified by kinase activity. In addition to medical benefits, regulation of kinase activity in mammalian cells is important in industrial production of biomolecules of high value, for example to prevent apoptosis and maximise yield.

There is a conserved structure across the protein kinase family, particularly at the catalytic core [122]. The ATP-binding site lies between the N-terminal region (mainly $\beta$-sheet) and the larger C-terminal region (mainly $\alpha$-helix). The N-terminal region contains a glycine-rich sequence of residues close to a lysine, which has been shown to be involved in ATP binding. A conserved aspartic acid residue is found in the catalytic region and is important for catalysis. In order to modulate kinase activity we need to develop specific regulators of protein kinase function, a process that is complicated by the conserved catalytic architecture [123, 124]. One way to achieve kinase regulation with enhanced selectivity is through the isolation of allosteric regulators, as their target sites are likely to be less structurally conserved across the protein kinase family. Kinases are known to have large conformational plasticity [125, 126] and this supports the prospect of allosteric modulation.

A few notable examples have highlighted this potential [127, 128]. Serine/threonine-protein kinase Chk1 has been the target of high-throughput screening efforts [129], leading to the discovery of an inhibitor that binds 13 Å from the active site. This inhibitor binds largely to the protein surface, with part sliding into a narrow hydrophobic cleft, indicating that unexpected sites on proteins may reveal allosteric properties. Despite locating the allosteric binding site, the mechanism for inhibition is not currently known [130]. Study of the tyrosine-protein

kinase Abl1 has revealed an inhibitor that binds far from the ATP-binding site [131, 132]. Binding of the modulator leads to changes in structural dynamics at the ATP-binding site, preventing the binding of ATP and leading to inhibition.

Other targets for which allosteric modulators have been discovered include the MAP kinases [133] and CDK2 [134] - see Section 1.5. Modulators such as these that bind protein kinases at sites separate from the ATP-binding site are known as type IV inhibitors. The binding sites of the above type IV inhibitors, along with others, are shown on the conserved protein kinase structure in Figure 1.4. Type I inhibitors are directly-competitive with ATP as they target the active conformation. Type II inhibitors bind to the DFG-out conformation and occupy the ATP-binding site and the surrounding hydrophobic region. Type III inhibitors bind the hydrophobic cleft adjacent to the ATP-binding site but do not bind the ATP-binding site itself.

A novel class of inhibitors, the type V bisubstrate and bivalent inhibitors, has emerged recently [135]. For example, a bivalent inhibitor was designed for a tyrosine kinase that binds to the ATP-binding site and to the regulatory domain SH3 simultaneously via a linker [136]. Such binding is able to have both high selectivity and high affinity. The protein kinase 'allosterome' at the ASD contains 51 allosteric protein kinases, 12 of which have a crystal structure with an allosteric modulator. The recent discovery of such sites and the conserved architecture of the eukaryotic protein kinases suggest there are many allosteric sites, particularly for type IV and type V inhibitors, yet to be discovered.

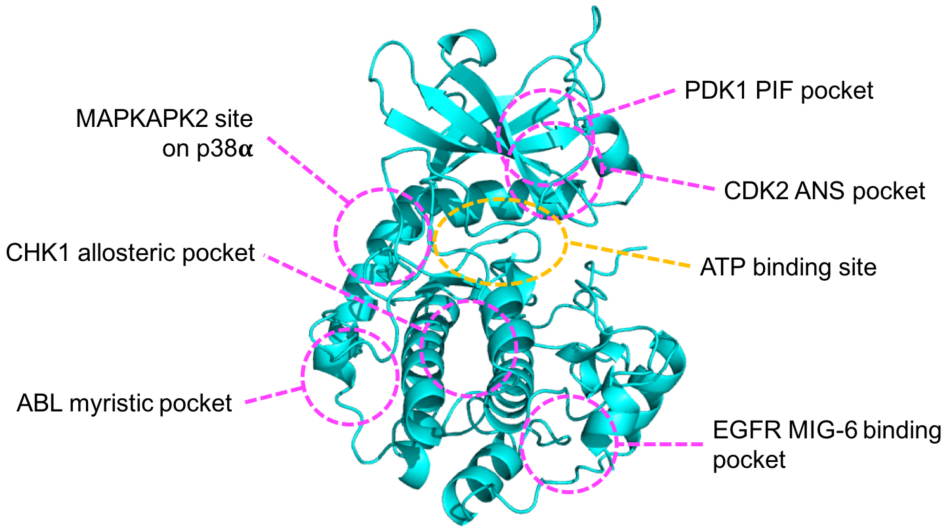



**Figure 1.4** Binding sites of known type IV allosteric inhibitors are shown in purple on the cAMP-dependent protein kinase (PKA) structure (PDB ID 1ATP). The ATP-binding site is shown in yellow. Figure based on Figure 3 from Lamba and Ghosh 2012 [135].

## 1.5 Cyclin-dependent kinase 2

CDK2 is a protein kinase important in regulating cell cycle progression [137]. Its deregulation has been linked to a number of diseases. CDK2 associates with, and is regulated by, the cyclin proteins. The G1 to S phase checkpoint of the cell cycle and progression through the S phase are largely controlled by the CDK2-cyclin E1 and CDK2-cyclin A2 complexes respectively. The CDK2-cyclin A2 complex is therefore a major target of drug discovery efforts to arrest or recover control of the cell cycle in dividing cells [134]. However, the validity of CDKs as cancer targets is not without its issues: knocking out both CDK2 and CDK4 does not stop the proliferation of adult cells in mice, for example [138]. In this case CDK1 is up-regulated to compensate for the absence of CDK2 [139].

One analysis indicated that CDK2 was the ninth most common protein in the Protein Data Bank (PDB) [140], with over 350 structures deposited since the first structure was crystallised in 1993 [141]. See Figure 1.5 for an overview of the structure of CDK2. While the crystallisation of CDK2 is relatively routine, one of the difficulties in studying the CDK2-cyclin A2 interaction is the difficulty of purifying soluble cyclin A2 in the absence of the CDK2 binding partner [142]. The mechanism of activation of CDK2 by cyclin A2 has been elucidated [143, 144, 145]. Changes in the αC-helix on cyclin A2 binding realign active site residues, and the T-loop moves which reveals the active site and makes activation phosphorylation easier. The αC-helix is an essential part of the CDK2-cyclin interface and its displacement is implicit in allosteric modulation of various kinases [146, 147]. Computational studies have indicated a large degree of flexibility in CDK2 [148, 149].

Only two inhibitors have been approved by the US Food and Drug Administration targeting the CDK family of proteins - ribociclib and palbociclib, CDK4/6 inhibitors used in the treatment of breast cancer [150]. No CDK2 inhibitors are currently approved for clinical use. This is at least in part due to the high conservation of the ATP-binding site among protein kinases, making the discovery of a specific CDK2 inhibitor that targets this site challenging - see Section 1.4. Several compounds are in clinical trials, but most of them target mul-

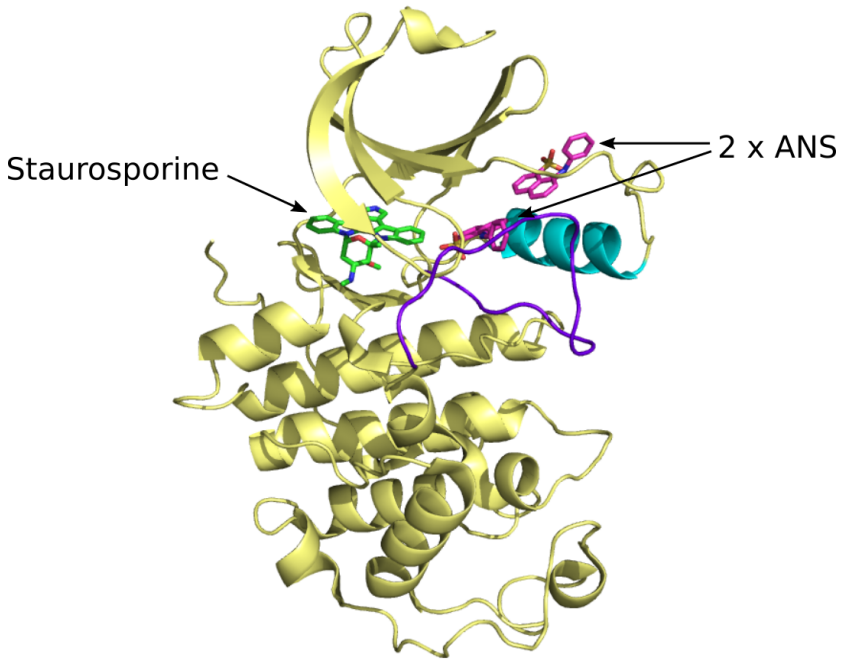



**Figure 1.5** The structure of CDK2 crystallised with the ATP-binding site inhibitor staurosporine and two molecules of the fluorophore 8-anilino-1-naphthalene sulfonate (ANS) at the allosteric site (PDB ID 4EZ7). CDK2 is shown as yellow cartoon with the αC-helix region coloured cyan and the T-loop coloured purple. The ATP-binding site inhibitor staurosporine and the two ANS molecules are shown as sticks and coloured green and magenta respectively. See also Figure 4.1. Figure based on Figure 1 from Pisani et al. 2016 [148].

tiple CDKs.

It has been found that the fluorophore ANS binds at a site removed from the ATP-binding site on CDK2 [134]. A screening study against this site revealed compounds that bind with affinity down to the micromolar range [151], and a fluorescence-based assay has the potential to facilitate high-throughput screening [68]. Binding of these compounds does not affect ATP binding and appears to inhibit the action of CDK2, indicating a likely allosteric site. Another study used virtual screening to find compounds that disrupt the interaction between CDK2 and cyclin A3 [152] and verified them experimentally with antiproliferation and binding studies. Small peptides are also able to break the CDK2-cyclin E1 interface [153]. A large scale screen of compounds against CDK2 found a family of compounds that hinder T-loop dynamics and hence could be used as allosteric modulators [70]. It is hoped that one of these approaches will lead to a viable drug.

## 1.6 Aims

The primary aim of this work was to develop computational approaches to predict allosteric sites on proteins. Two methods were developed in order to achieve this. AlloPred uses NMA and machine learning to rank pockets in a protein based on their allosteric character - see Chapter 2. ExProSE uses distance geometry to generate ensembles of protein structures from two input structures - see Chapter 3. Ensembles from ExProSE can be perturbed to predict the effect of allosteric modulators, as well as having other uses due to the size of the conformational space covered. These two methods are shown to be competitive and complementary to existing methods and represent the main contribution to knowledge of this work.

Although computational methods can be validated with existing data, ultimately the measure of their success is whether new predictions can be confirmed experimentally. The secondary aim of this work was to experimentally test the computational prediction of an allosteric site. A prediction made by ExProSE of a potential allosteric site on CDK2 was explored with bioinformatics methods and potential modulators selected using a virtual screen - see Chapter 4. Purchased compounds were tested using two experimental assays.

During the process of developing and testing the above approaches the first systematic comparison of allosteric prediction methods was carried out. The difficulty of distinguishing allosteric pockets from cavities with generally good binding properties is established and quantified by assessing the ability of generic pocket predictors to predict allosteric pockets. The variety of allosteric mechanisms [3] and the problems this poses for generic prediction methods is a common theme of the work. The strength and limitations of NMA and MD methods for studying large conformational changes and allosteric mechanisms is discussed. Ultimately, we give evidence for the current view that allostery is a property of the protein structural ensemble, and only by taking this into account can we move towards a unified scheme for understanding and predicting allostery.

# Chapter 2

# AlloPred

This chapter describes the development and validation of AlloPred, a computational approach based on the perturbation of normal modes to predict allosteric sites on proteins. AlloPred is available as a web server (`http://www.sbg.bio.ic.ac.uk/allopred/home`) that so far has had over 400 submissions from around the world. The work is published in Greener and Sternberg 2015 [21], on which this chapter is based.

## 2.1 Materials and Methods

### 2.1.1 Data selection

ASBench [85], a benchmarking set for allosteric discovery, was used as a source of known allosteric proteins. ASBench is a curated subset of the ASD [86]. The 'Core-Diversity set' contains 147 structurally-diverse allosteric sites on 127 proteins from a variety of protein classes such as transferases, hydrolases and transcription factors. The PDB files, allosteric site data and active site data were obtained for each protein from ASBench. UniProt [154] and the Catalytic Site Atlas (CSA, `http://www.ebi.ac.uk/thornton-srv/databases/CSA`) [155] were used to find active site data when it was not available from ASBench. In each PDB file, only the chain(s) containing the active and allosteric sites, and any chains linking them, were considered. This was in order to keep the size of the proteins manageable, as using entire protein assemblies would lead to a large number of pockets. It also allowed comparison with existing methods, which use similar criteria. In practise the use of larger assemblies was tried during development and did not have a large effect on the results. 7 proteins were removed from the set as the PDB file did not contain the active site, i.e. the PDB file represented the allosteric section of a larger protein. 1 protein was removed as Fpocket did not run successfully. This left 119 proteins in the dataset. The dataset was randomly split into a training set of 79 proteins and a test set of 40 proteins.

### 2.1.2 Pocket prediction

Potential binding pockets on the proteins were predicted using the open-source Fpocket v2.0 algorithm, which has been shown to be effective in comparison to other methods [35]. The default parameters used in the Fpocket calculation produced pockets that were large enough to place most (average 86%) allosteric binding residues in pockets but not so large that identifying a pocket as having allosteric effect was of little use. Sometimes multiple allosteric pockets on

the same protein represented different and physically-separated allosteric sites, and sometimes adjacent calculated pockets covered a single allosteric binding site. The pockets also covered much of the protein surface, which allowed the method to detect allosteric sites that could be found anywhere on the surface. On average 41% of residues in each protein appeared in a pocket.

Fpocket output 2,201 pockets for the 119 proteins (average 18.5 per protein), of which 389 (18% of pockets, average 3.3 per protein) contained at least one residue identified as binding to an allosteric modulator and were hence labelled as allosteric pockets. Although being defined as an allosteric pocket in this manner does not necessarily mean that binding to that pocket causes the allosteric effect, the average number of allosteric binding residues in an allosteric pocket was 4.3, indicating the utility of locating such pockets. All but 5 proteins in the dataset had at least one allosteric binding residue placed in a pocket. We treated pockets without known allosteric binding residues as negative examples during machine learning. It should be noted that these pockets may not correspond directly to the actual pockets on the protein, or may have latent allosteric character yet to be discovered.

### 2.1.3 Normal mode analysis

In NMA the Hessian matrix - the matrix of second derivatives of the potential energy $V$ with respect to the mass-weighted atomic coordinates - is diagonalised to yield the normal modes [37]. The potential energy $V$ was described according to the ENM [156] as a set of harmonic springs of strength $k$ between every pair of $C^{\alpha}$ atoms no further than distance $R_c$ apart:

$$V = \sum_{\substack{r_{ij}^0 < R_c \\ i<j}} k(r_{ij} - r_{ij}^0)^2$$

where $r_{ij}^0$ is the Euclidean distance between atoms $i$ and $j$ in the PDB file. We used values of 1 kcal mol$^{-1}$ Å$^{-2}$ and 15 Å for $k$ and $R_c$ respectively.

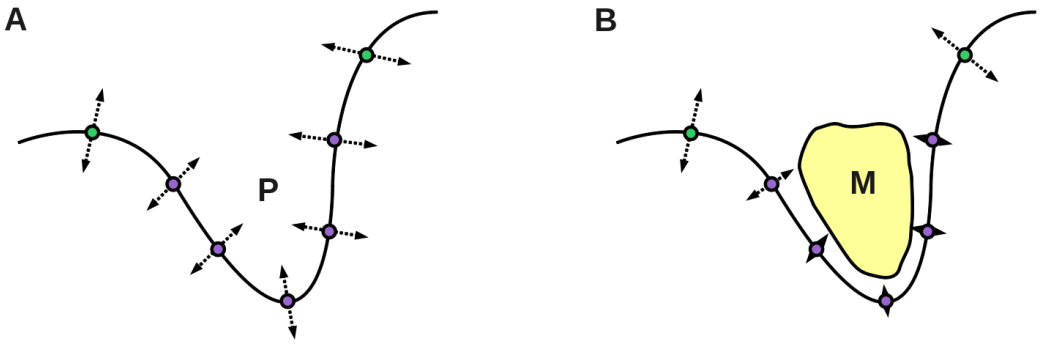



**Figure 2.1** The change in flexibility of a protein on modulator binding at an allosteric site. The black line indicates the surface of the protein and circles show residues: purple circles represent residues that are part of pocket P and green circles represent other residues. Dashed arrows represent the magnitude of the fluctuations of a residue about equilibrium. (A) shows the protein in the absence of a modulator. All residues can vibrate. (B) shows the effect of modulator M binding in pocket P. The residues in the pocket have restricted motion and are less able to vibrate around their equilibrium positions. Our method sought to approximate the effect of ligand binding by artificially restricting the flexibility of residues in a pocket using a higher spring constant in the elastic network. Figure based on Figure 1 from Greener and Sternberg 2015 [21].

The reduction in flexibility of an allosteric pocket on modulator binding is shown in Figure 2.1. To model this, the unperturbed normal modes were first calculated for the protein. The calculation was then repeated, each time perturbing one of the pockets in the protein. If either atom $i$ or $j$ was in the pocket to be perturbed then a higher value of 1.5 kcal mol$^{-1}$ Å$^{-2}$ for $k$ (1.5 times the previous value) was used instead. This higher value was chosen after values from 1.2 to 2.5 kcal mol$^{-1}$ Å$^{-2}$ were examined. Active site residues were not counted as being in any pocket for this alteration of $k$, in order to avoid direct perturbation of the site at which the effect was measured. This approach assumes nothing about the shape of the modulator other than that it affects the flexibility of the whole pocket to which it binds.

Once the perturbed NMA had been carried out, the degree of change caused by the perturbation needed to be measured. Since changes at the active site

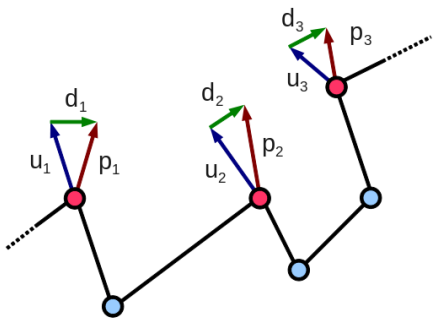



**Figure 2.2** The method to calculate how much of an effect a change in the normal modes has at the active site. The black lines represent the amino acid backbone of the protein. Red circles represent amino acids identified as being active site residues; light blue circles represent other amino acids. The dark blue arrows $\mathbf{u_j}$ represent the motion of $C^{\alpha}$ atoms in a particular normal mode before perturbation. The brown arrows $\mathbf{p_j}$ represent the motion after perturbation at a particular pocket. The green arrows $\mathbf{d_j}$ represent the vector difference between the motions, $\mathbf{d_j} = \mathbf{p_j} - \mathbf{u_j}$. The differences are averaged across the active site, then a weighted average is taken across a number of normal modes to get a single numerical measure of the effect of the perturbation.

will likely determine how strong an effect a modulator has, the effect of the perturbation on the active site should be considered. Within each individual normal mode the effect of the perturbation was measured by averaging across all identified active site residues the magnitude of the difference between the perturbed and the unperturbed displacements from equilibrium. This is shown visually in Figure 2.2 and is given by:

$$v_i = \frac{1}{N_a} \sum_{j=1}^{N_a} |\mathbf{p_j} - \mathbf{u_j}| = \frac{1}{N_a} \sum_{j=1}^{N_a} |\mathbf{d_j}|$$

where $v_i$ is the effect of the perturbation in normal mode $i$, $\mathbf{p_j}$ is the displacement of residue $j$ in the perturbed normal mode, $\mathbf{u_j}$ is the displacement of residue $j$ in the unperturbed normal mode, $\mathbf{d_j}$ is the vector difference in the harmonic motion of residue $j$ before and after perturbation, and $N_a$ is the number of active site residues.

The effects of the perturbation within each normal mode then needed to be averaged across the modes in order to get a single numeric measure for the strength of the effect arising from perturbation at one pocket. The effect within each of the normal modes was weighted by the frequency such that the lowest-frequency mode of the chosen modes had the greatest influence on the results. The equation to determine the effect of a perturbation $C_m$ is:

$$C_m = \sum_{i=1}^{m} \frac{v_i}{\omega_i}$$

where $v_i$ is defined above, $\omega_i$ is the frequency of mode $i$ and is hence equal to the square root of the eigenvalue $E_i$, and $m$ is the number of normal modes chosen for the calculation. The justification for this method was that lower-frequency modes within the range selected are likely to be more important in allosteric communication because they consist of the long-range motions of many atoms [41].

It might be expected that larger pockets will have a higher $C_m$ value simply by virtue of having more residues perturbed. In order to account for this a second measure, $E_m$, was defined as:

$$E_m = \frac{C_m}{N_p}$$

where $N_p$ is the number of residues in the pocket and $C_m$ was defined previously. $E_m$ is a measure of the amount of change caused at the active site per residue in the perturbed pocket. A Python script utilising the ProDy package [157] was used to perform NMA on the proteins.

### 2.1.4 Machine learning

Values of $C_m$ and $E_m$ with $m$ equal to 20, 50, 100, 200 and all modes were chosen as features in a SVM. SVM is a non-probabilistic binary linear classifier which maps training examples in a multi-dimensional space and finds categories such that the gap between them is as large as possible [158]. New examples can

be classified based on which side of the gap they fall. The features from the Fpocket output used in the SVM were:

- Rank
- Score
- Druggability score
- Number of alpha spheres
- Total SASA
- Polar SASA
- Apolar SASA
- Volume
- Mean local hydrophobic density
- Mean alpha sphere radius
- Mean alpha sphere solvent accessibility
- Apolar alpha sphere proportion
- Hydrophobicity score
- Volume score
- Polarity score
- Charge score
- Proportion of polar atoms
- Alpha sphere density
- Centre of mass - alpha sphere max distance
- Flexibility

See the Fpocket documentation for more details on each of these measures. Distance to the active site, number of residues in the pocket and number of

pockets in the protein were also used as features. The distance to the active site for each pocket was calculated as the distance between the geometric centre of the active site residues and the geometric centre of the residues in the pocket. Each feature (apart from number of pockets) was utilised in two different ways: the feature value normalised across all proteins (termed raw); and the ranking of the feature value within the values for that protein, where the ranks were scaled between 0 and 1 (termed ranked).

The 65 features were ranked in Weka explorer [159] using the ChiSquared attribute evaluator and the Ranker search method. This evaluates the worth of a feature by computing the value of the chi-squared statistic with respect to the class. The top 7 features only were retained, as features below this added little value. The retained features, in descending order of descriptive power, were:

- Number of alpha spheres (raw)
- $E_{200}$ (ranked)
- Score (raw)
- $E_{all}$ (ranked)
- Distance to active site (raw)
- Pocket size (raw)
- Fpocket rank (raw)

This indicates the utility of our NMA features. The SVM-Light package [160] was used to run the SVM. The Gaussian kernel was selected, containing internal parameters $C$ and $\gamma$. The cost factor by which training errors on positive examples outweigh errors on negative examples was set as the ratio of negative to positive examples in the training set (6.19). A leave-one-out parameterisation procedure was carried out over a grid of parameters with $C$ equal to 0.01, 0.1, 1 or 10 and $\gamma$ equal to $10^{-3}$, $10^{-4}$ or $10^{-5}$. The procedure consisted of training the SVM on pockets from 78 of the 79 proteins in the training set and testing on pockets from the one left out. The process was repeated for each protein in the set. Performance was similar across the parameter range, with the parameters

$C = 1$ and $\gamma = 10^{-4}$ being selected for the final SVM. Due to the low number of allosteric pockets on each protein, only the top prediction was chosen as being allosteric.

### 2.1.5 Web server

The web server was implemented using the Django extension to Python and a SQLite database. JSmol, a JavaScript implementation of the Jmol package, was used for molecular visualisation. Bootstrap was used for page styling. The standalone version of the code runs faster and it is recommended that users who intend to use the method extensively or in batch download the code for local use.

## 2.2 Results

The AlloPred computational procedure, which uses NMA and other features to predict the allosteric pockets on a protein, was developed. AlloPred models how the dynamics of a protein would be altered in the presence of a modulator at a specific pocket. As described in Section 2.1 if the binding of a modulator causes a change in dynamics at the active site, it can be predicted to have an allosteric effect. Pockets on the protein were first predicted using the Fpocket [35] algorithm, which locates pockets using Voronoi tessellation and alpha spheres. The normal modes of the protein were then calculated using the ENM, except the spring constant of any atom pair including a residue in a chosen pocket was set to be a higher value. The effect of this perturbation was measured at the active site. Intuitively, the interpretation of this change is that the binding of a small molecule in a pocket is expected to quench the vibrations of the surrounding residues. These results were combined with output from Fpocket in a SVM to predict allosteric pockets on proteins.

### 2.2.1 Validation

The AlloPred SVM was trained on a set of 79 known allosteric proteins - see Section 2.1 for the selection criteria. The top 7 retained features included 2 normal mode features, indicating the predictive power of normal mode perturbation. AlloPred was tested on a test set of 40 known allosteric proteins. For each protein AlloPred ranked the pockets and the top ranked pocket was examined. For 23 of 40 proteins AlloPred ranked top a pocket containing an allosteric binding residue (termed an allosteric pocket), when 18% of pockets were allosteric pockets. For 28 of 40 proteins an allosteric pocket was ranked first or second. It is possible that some of the predicted pockets have allosteric character yet to be discovered. The results were compared to two existing methods for allosteric site prediction. The AlloSite server uses the Fpocket algorithm and a machine learning approach [24], whereas the PARS server combines changes in protein flexibility and a structural conservation score [29]. The correct predic-

tions made by each method, and the overlap between the methods, are shown in Figure 2.3. AlloSite ranked an allosteric pocket top in 21 of 40 cases and is suitable for direct comparison to AlloPred as both methods rank pockets from Fpocket. PARS, however, makes predictions of single points; a point was considered allosteric for our purposes if it was within 10 Å of an allosteric modulator atom in the protein-modulator crystal structure. It is important to note the different criteria for a correct prediction when considering the results. PARS ranked an allosteric pocket top in 10 of 40 cases. Figure 2.3 shows that AlloPred compares well to other methods and makes 4 correct predictions that neither of the other methods do. This suggests that users of other allosteric prediction methods would benefit from the additional use of AlloPred.

In order to reduce the effects of bias during the split of the dataset into training and test sets, the dataset of 119 proteins was additionally split randomly 20 times into training and tests sets of 79 and 40 proteins respectively. The SVM was then trained on the training set, using the previous parameters, and tested on the test set. The average number of correct predictions across the 20 runs was 23.6 out of 40. This shows that the above results used for comparison to other methods are indicative of the performance of the method.

### 2.2.2 Web server

The publicly-available AlloPred web server (`http://www.sbg.bio.ic.ac.uk/allopred/home`) allows users to analyse the prediction results via an intuitive interface. A flowchart outlining the process of running a job is shown in Figure 2.4. Users can either input a PDB ID and chain(s) or upload a PDB file. The active site residues of the protein must be given but there is an option to retrieve this data, if it is available, from the CSA [155]. The CSA has catalytic site data for around 70% of enzymatic proteins in the PDB. After submission a progress page shows the log file of the job and gives an estimated completion time. When the results are available a link appears to the results page. An example results page is shown in Figure 2.5. All pockets are displayed in a table with their AlloPred rankings and Fpocket output. The table can be sorted and

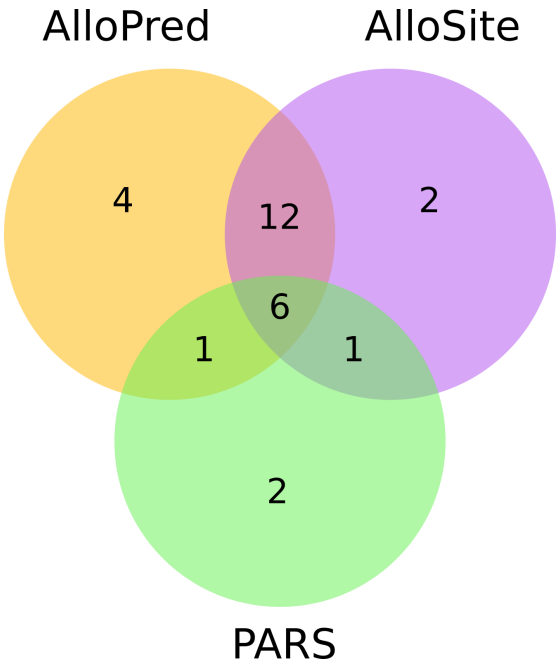



**Figure 2.3** Results comparison by method. Venn diagram showing the number of top predictions for each protein by each method that were correct, from the test set of 40 proteins. For AlloPred and AlloSite a correct prediction was prediction of a pocket containing at least one allosteric binding residue. For PARS a correct prediction was prediction of a site within 10 Å of at least one atom of the allosteric modulator in the protein-modulator crystal. Figure based on Figure 3 from Greener and Sternberg 2015 [21].

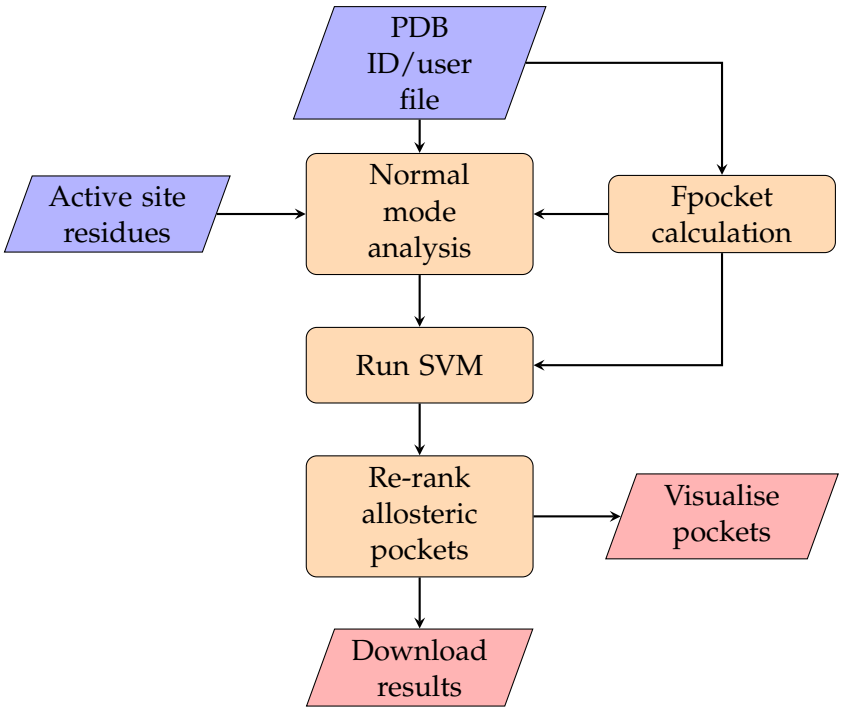



**Figure 2.4** Flowchart showing the stages involved in running a job submitted to the AlloPred web server. Trapeziums represent inputs or outputs available to the user via the web front end (inputs purple, outputs red). Rounded rectangles represent stages in the calculation pipeline that occur via the web back end. Figure based on Figure 2 from Greener and Sternberg 2015 [21].

filtered by any one or more of the 29 AlloPred and Fpocket features. The page also allows users to visualise each pocket on the protein in a JSmol window that lets the user explore the protein and its predicted allosteric sites. Features include highlighting the active site residues, selecting one of three visualisation options and a JSmol terminal to insert custom commands. The results, including full details of each pocket, can be downloaded for further analysis as a tab-delimited text file. The exploratory features on the website and results file containing all the pocket information set AlloPred apart from similar servers by allowing more extensive analysis of the results. The calculation time is fast, with a 400 residue protein ($\sim$15 predicted pockets) analysed within 5 minutes. The web server also includes a tutorial and a link to the source code to run AlloPred offline.

A

Select columns | Reset JSmol | 1 to 8 of 8 rows | 10

| View | AlloPred ranking | Fpocket ranking | NMA effect (200 modes) | NMA effect per residue (200 modes) |
|---|---|---|---|---|
| | e.g. "< 5" | | | |
| ☑ | 0 | 2 | $2.33 \times 10^{-2}$ | $1.45 \times 10^{-3}$ |
| ☑ | 1 | 6 | $1.87 \times 10^{-2}$ | $1.44 \times 10^{-3}$ |
| ☐ | 2 | 1 | $2.41 \times 10^{-2}$ | $1.61 \times 10^{-3}$ |
| ☐ | 3 | 0 | $2.18 \times 10^{-2}$ | $1.68 \times 10^{-3}$ |
| ☐ | 4 | 5 | $2.1 \times 10^{-2}$ | $1.62 \times 10^{-3}$ |
| ☐ | 5 | 4 | $2.12 \times 10^{-2}$ | $1.77 \times 10^{-3}$ |
| ☑ | 6 | 3 | $2.03 \times 10^{-2}$ | $2.03 \times 10^{-3}$ |
| ☐ | 7 | 7 | $2.03 \times 10^{-2}$ | $2.53 \times 10^{-3}$ |

B

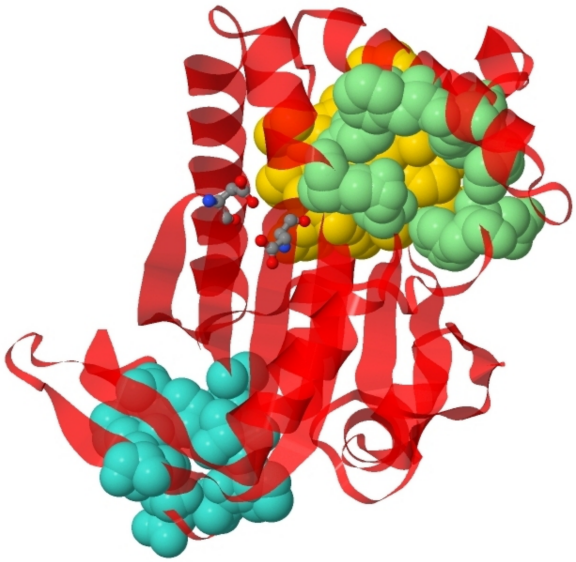


**Figure 2.5** Screenshots of the AlloPred results page for the receptor-type adenylate cyclase with PDB ID 1FX2. (A) The results table with default columns selected. Three pockets have been chosen for visualisation. (B) The JSmol window shown when the boxes are selected as in (A). The ribbon visualisation option is used and the residues identified as being part of the active site are shown as balls and sticks. Three pockets are shown in green, yellow and blue. AlloPred correctly predicts the green pocket as being allosteric. Figure based on Figure 4 from Greener and Sternberg 2015 [21].

## 2.3 Discussion

Over the last few years a renewed interest in allostery, perhaps due to the potential benefits of allosteric drugs, has led to the development of a number of computational approaches to understanding allostery - see Section 1.2. Some of these are directly associated with predicting allosteric sites on proteins from structure alone and are available as web servers that can take any PDB file as input.

The AlloSite server is similar to the method presented here in that it uses the Fpocket algorithm and attempts to elucidate allosteric pockets [24]. Whereas AlloSite solely uses the Fpocket output, our method uses an approach that combines flexibility with the Fpocket output. A combination of methods may give better predictions than either method individually, as indicated by the unique predictions made by both methods during testing. In fact the AlloSite predictions were found in every case to correspond to the pocket ranked top by Fpocket. This highlights the difficulty in separating the prediction of allosteric sites from the prediction of binding sites in general. This will be explored further in Chapter 3. It seems that Fpocket is able to predict good binding sites and AlloPred uses the extra flexibility information to bias this prediction towards sites with allosteric character. The complete ranking of pockets provided by AlloPred may also be useful, as pockets ranked second were often found to be allosteric in the test set.

An approach that combines flexibility analysis using normal modes and structural conservation scores [30] is also similar to the method presented here and was recently turned into a web server, PARS [29]. Although direct comparison is difficult due to the differences in site calculation, definition of allosteric sites and datasets used, the method presented here again may be used well in combination as shown by Figure 2.3.

The lack of input about the shape of the ligand and the large coverage of the protein in terms of pockets (average 18.5 pockets per protein) used by our method mean that it may be able to predict novel or unusual sites that methods which explicitly model the modulator might not. This is important, for

example when searching for allosteric sites on proteins believed to be non-allosteric. Whilst the use of evolutionary information is clearly beneficial in allosteric prediction [48] the use of structure alone is appealing for a general prediction method as it facilitates prediction of sites not currently conserved by evolution. This is useful due to the large variety of allosteric modulators [100] and mechanisms [3], suggesting potential novel modulators for proteins with known allosteric pathways.

Other promising approaches [59, 161, 28] investigate the allosteric pathway and are not directly comparable with this method, which is only concerned with how the pathways transmit the effects of perturbations to the normal modes and does not directly reveal any information about the pathways themselves. Again, a combination of our method with these approaches may be useful, as pockets predicted using our or other methods can be further investigated to reveal information about the underlying allosteric communication.

The main limitation of our method is related to the diversity found in allosteric systems. Rigid-body motions of oligomers, side-chain dynamics, backbone motions and local unfolding are all mechanisms of allostery, with allosteric effects even present in IDPs [3]. A method based around the changes in dynamics on ligand binding is likely to miss many allosteric effects, and this can go some way to explaining the predictions of our method that were incorrect. In particular, classic examples of allostery such as haemoglobin that involve oligomeric re-organisation to affect ligand cooperativity are not suitable for use with this method. The method also relies upon knowledge of the active site location. However, the results shown here and in other studies are encouraging and indicate a future where we can pick modulating sites on proteins with reasonable confidence. Our method, for example, successfully predicts allosteric sites on proteins with a variety of sizes and functions. The development of such methods that can be used by non-specialists without extensive parameter selection, and where the results are presented in a clear manner allowing exploration, will help bridge the gap between computational and experimental allosteric site prediction.

AlloPred could be further developed by introducing different rules for per-

turbation, such as altering the way normal modes are weighted or investigating different groupings of normal modes. Structural prediction methods [162] could also be utilised to allow prediction of allosteric sites from sequence via a predicted structure, or prediction from the structure of a homologous protein. Specific residues within allosteric pockets that have the largest impact on the allosteric effect could also be discovered by changing the amount of perturbation at each residue individually in a pocket. This could allow prediction of recently proposed 'anchor' and 'driver' atoms that contribute to the allosteric effect [88], aiding the rational design of drugs that bind to the predicted pockets.

The AlloPred procedure is ultimately a computationally-light means of ranking pockets on a protein in terms of their predicted allosteric character. It is suitable for use by non-bioinformaticians via the web server, and amenable to high throughput via the source code. Such methods will hopefully be developed, validated and utilised more widely as allosteric prediction becomes more powerful.

# Chapter 3

# ExProSE

This chapter describes ExProSE (Exploration of Protein Structural Ensembles), a computational method based on distance geometry that generates an ensemble of protein structures from two input structures. The structures span conformational space and the ensemble generation procedure can be perturbed to predict allosteric sites. The first systematic comparison of methods to predict allosteric sites is also described. The work is published in Greener et al. 2017 [27], on which this chapter is based.

## 3.1 Materials and Methods

ExProSE is based on the CONCOORD distance geometry method [109] and consists of a two-step process of (i) converting interactions in one or more crystal structures into distance constraints and (ii) generating new structures from these constraints. ExProSE has important differences to CONCOORD that make it suitable for modelling conformational transitions and ensemble perturbations. These are primarily the use of two input structures instead of one, a different procedure for achieving convergence, the ability to predict the effect of a modulator and an auto-parameterisation procedure. The effect of a potential modulator can be explored by adding new distance constraints arising due to the modulator and generating new structures with these additional constraints. The original ensemble and the ensemble generated with the extra constraints can then be compared to see how the modulator affects the protein.

ExProSE is implemented in Julia [163], a language that combines readable syntax similar to Python or MATLAB with performance approaching statically-compiled languages like C. Work was initially begun in a combination of Python and C++ before switching to Julia. Use of Julia allows good computational performance at the limiting steps, but also allows compact and easy-to-use code that others can modify. The code, documentation, details of the datasets and instructions for reproducing the data are freely-available under the MIT license as a Julia package at `https://github.com/jgreener64/ProteinEnsembles.jl`. The code is written in a modular way with associated unit tests and an automated building and testing procedure.

### 3.1.1 Distance constraint generation

The first step is to obtain a set of distance constraints from a protein structure. Files for selected structures were obtained from the PDB. Contrary to similar studies [30, 24] the smallest biological assembly of the protein is used, rather than only the chain containing the allosteric modulator. For each protein structure, interatomic distances are calculated from the PDB coordinates.

Hetero atom records, including the allosteric modulators, are removed. Any existing hydrogens are removed and polar hydrogens are added using an in-house script. For disordered atoms the location with the highest occupancy in the PDB is selected. Secondary structure assignments, required to obtain additional distance constraints, are obtained using the DSSP software [164]. As two structures for the same protein are utilised to generate distance constraints, only atoms common to both structures are used. Every atom pair is examined and assigned an interaction type. The criteria for each interaction are the same as in CONCOORD [109] and are shown in Table 3.1.

Each atom pair is assigned the first interaction for which it fulfils the criterion. If an atom pair is not assigned any of the first 14 specific interactions, it is assigned the generic 'All other pairs' interaction type. Lower and upper distance constraints $l_{ij}$ and $u_{ij}$ are generated for each atom pair $ij$ based on the interatomic distance $d_{ij}$, the constraint tolerance for the interaction $t_{ij}$ and a tolerance weighting factor $W_B$ that is between 0.0 and 1.0:

$$l_{ij} = d_{ij} - W_B t_{ij}, \quad u_{ij} = d_{ij} + W_B t_{ij}$$

The justification for and selection of $W_B$ is described below. For example two atoms 1.54 Å apart and in a covalent bond with $W_B$ equal to 0.5 would have a lower distance constraint of 1.53 Å and an upper distance constraint of 1.55 Å, as the constraint tolerance multiplied by $W_B$ is 0.01 Å. This process yields a set of distance constraints for each crystal structure of a protein.

The distance constraints generated from the two structures for the same protein are combined to get a set of combined constraints. This is shown visually in Figure 3.1. The constraints are combined in such a way that the new constraints for a given atom pair cover the distance of both the individual constraints for that pair. For example if two atoms have a lower and upper distance constraint of 6.0 Å and 7.0 Å in structure one, and 6.5 Å and 7.5 Å in structure two, then the new constraints will be 6.0 Å and 7.5 Å.

It is undesirable to retain all the 'All other pairs' interactions (type 15 in Table 3.1) as they vastly outnumber the specific interactions (types 1-14). Specific

| Number | Interaction name | Constraint tolerance / Å | Definition |
|---|---|---|---|
| 1 | Covalent bond | 0.02 | Pairs that are covalently bonded |
| 2 | Bond angle | 0.05 | Pairs where both atoms are covalently bonded to the same atom |
| 3 | Ring | 0.1 | Pairs that are part of ring systems |
| 4 | Double bond 1-4 | 0.1 | 1-4 dihedral angle restricted pairs in side chain double bonds (found in ASN, GLN and ARG) |
| 5 | Omega 1-4 | 0.1 | 1-4 pairs constrained by the rigid $\omega$ dihedral angle |
| 6 | Tight phi/psi 1-4 | 0.2 | 1-4 pairs constrained by the $\varphi/\psi$ dihedral angle where one residue is a proline or both residues are in the same helix/strand |
| 7 | Loose phi/psi 1-4 | 0.4 | 1-4 pairs constrained by the $\varphi/\psi$ dihedral angle where one residue is a glycine or both residues are in a loop region |
| 8 | Other phi/psi 1-4 | 0.3 | 1-4 pairs constrained by the $\varphi/\psi$ dihedral angle that do not fall into the above two categories |
| 9 | Other 1-4 | 0.4 | Other 1-4 dihedral angle restricted pairs that do not fall into the above categories |
| 10 | Secondary structure | 0.5 | Pairs of backbone atoms that are in the same helix/strand and are not more than 4 residues apart |
| 11 | Salt bridge | 0.75 | Pairs from oppositely-charged groups in close proximity (less than 4 Å apart) |
| 12 | Hydrogen bond | 0.5 | Pairs that are part of a hydrogen bond; donor-acceptor distance is no more than 3.5 Å, hydrogen-acceptor distance is no more than 2.5 Å and the donor-hydrogen-acceptor angle is at least 90 degrees |
| 13 | Tight hydrophobic | 0.5 | Pairs where the interatomic distance is less than the sum of the van der Waals radii of the atoms plus 0.5 Å; only C and H atoms are counted |
| 14 | Loose hydrophobic | 1 | Pairs where the interatomic distance is less than the sum of the van der Waals radii of the atoms plus 1.0 Å; only C and H atoms are counted |
| 15 | All other pairs | 5 | Pairs that do not fall into any of the above categories |

**Table 3.1** Interaction types between atom pairs. These are the same as in CONCOORD [109]. The constraint tolerance values are used to generate lower and upper distance constraints between atoms.

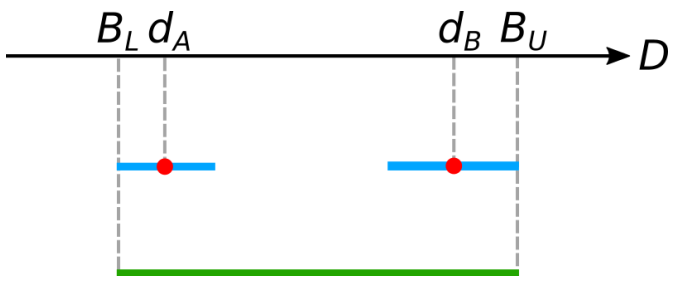



**Figure 3.1** The method to calculate the upper and lower distance constraints between two atoms given two input protein structures. First, the distance between the two atoms in each of the two input structures is calculated ($d_A$ and $d_B$). An upper and lower distance constraint is then calculated for each of these distances by multiplying the constraint tolerance for the interaction in that structure by the tolerance weighting factor (these constraints are shown as blue bars). The upper and lower limits of these constraints are combined to get the new distance constraints $B_U$ and $B_L$ (green bar).

interactions scale with the atom number $N_A$ whereas other pairs scale as $N_A^2$. Hence only a fraction of the other pairs are retained as distance constraints. The probability of retaining an other pair is chosen so that the final number of other pairs is roughly $20N_A$, the value used by studies utilising CONCOORD [110].

The parameter $W_B$ determines how much the structures are allowed to vary from the input structures. $W_B$ is chosen for each protein in the apo/holo and allosteric datasets by a process of auto-parameterisation. $W_B$ equal to 0.0 usually results in a narrow range of structures that are midway between the two input structures. By contrast, $W_B$ equal to 1.0 usually results in structures that cover a wide conformational space beyond the input structures and may be of poor stereochemical quality. A measure for the conformational spread of the ensemble was developed. This measure $F$ is the fraction of structures $S$ in the ensemble for which $TM(S,A) > TM(B,A)$ and $TM(S,B) > TM(A,B)$ where $TM(X,Y)$ is the TM-score between model $X$ and reference $Y$, and $A$ and $B$ are the two input crystal structures. The TM-score is a measure of similarity between two protein structures, with a higher score indicating more similar structures. $F$ therefore gives the proportion of structures that are closer to both input structures than the input structures are to each other. $F$ equal to 0.9 indicates an ensemble that effectively covers the conformational space of the input

structures. Ensembles of 50 structures are generated with $W_B$ starting at 1.0 and decreasing in steps of 0.1. When the ensemble generated has an $F$ value of at least 0.9, that $W_B$ is chosen. For the specific examples T4-lysozyme and CDK2, $W_B$ is equal to 0.2 and 0.3 respectively. It should be noted that the above auto-parameterisation procedure to select $W_B$ can be implemented automatically by the ExProSE software and requires no input by the user. For catabolite activator protein (CAP) only one input structure is used so $W_B$ is selected manually as 0.4. This value allows flexibility in the ensemble whilst giving good quality structures.

### 3.1.2 Protein structure generation

Once the distance constraints have been generated, an iterative process is used to generate structures that satisfy the constraints. CONCOORD uses a procedure where an atom pair violating a constraint is moved together or apart such that the new distance between the atoms falls randomly within the allowed region. A new procedure was used here that was found to provide better convergence when utilising combined distance constraints from two structures. Stochastic Proximity Embedding (SPE) [165] was selected, as it has been shown to converge effectively and scales well with system size. This procedure provides better convergence than the CONCOORD procedure of moving atoms to a random distance within the distance constraints. The pseudocode for the SPE algorithm, rephrased from an existing review [165], is shown in Algorithm 1. The distance constraints do not include favourability for a particular chirality, so coordinates produced from SPE are examined and structures with the incorrect chirality are reversed by mirroring all coordinates in the $xy$ plane.

Once a set of coordinates has been generated, an SPE error score can be calculated that measures how well the distance constraints are satisfied [165]. This score is calculated as shown in Algorithm 2. Structures with a high error score tend to have more violations of allowed stereochemistry, which is to be expected as there are more violations of allowed constraints. In order to account for this, more structures are generated than required and those with the high-

**Algorithm 1** SPE algorithm

Define lower and upper distance constraints $l_{ij}$ and $u_{ij}$ for atom pairs $i$ and $j$
Define an initial learning rate $\lambda_d = 1.0$
Randomise atomic coordinates $x_i$ within a cube of 100 Å
**for** $C$ cycles **do**
  **for** $S$ steps **do**
    Randomly select a pair of atoms $i$ and $j$ for which a constraint exists
    Compute the distance $d_{ij} = \|x_i - x_j\|$
    **if** $d_{ij} < l_{ij}$ or $d_{ij} > u_{ij}$ **then**
      Update the coordinates $x_i$ and $x_j$ by

$$x_i = x_i + \frac{\lambda_d}{2}\frac{t_{ij} - d_{ij}}{d_{ij}}(x_i - x_j)$$

$$x_j = x_j + \frac{\lambda_d}{2}\frac{t_{ij} - d_{ij}}{d_{ij}}(x_j - x_i)$$

      where $t_{ij}$ is the nearest constraint to $d_{ij}$
    **end if**
  **end for**
  Decrease the learning rate $\lambda_d$ by $1/C$
**end for**

est scores are discarded. The ratio is set to be 1.5. So if the final ensemble had 200 structures, initially 300 are generated and the 100 with the highest error score are discarded. This was found during development to generally produce ensembles of structures with acceptable stereochemical quality.

**Algorithm 2** Scoring algorithm

Set score $s = 0$
**for** each atom pair $i$, $j$ with a distance constraint **do**
  **if** $d_{ij} < l_{ij}$ or $d_{ij} > u_{ij}$ **then**
    Increase $s$ by

$$\frac{(d_{ij} - t_{ij})^2}{\max(u_{ij} - l_{ij}, 0.001)}$$

    where $t_{ij}$ is the nearest constraint to $d_{ij}$
  **end if**
**end for**

The number of iterations per atom, the product of the number of cycles $C$ and the number of steps $S$ from Algorithm 1, is taken as 60,000. This was chosen as the SPE error score did not generally decrease for iterations beyond this - see Figure 3.3. The ratio of $S$ to $C$ is taken as 50:1 as in practice any value of $S > C$ will give similar results [165]. The reduction in learning rate over the course of the minimisation makes this process similar to simulated annealing. Initially large movements through the conformational space allow the correct region to be found. The movements are dampened over time to allow the system to converge to a solution. This procedure is carried out separately multiple times to obtain an ensemble. The structures are different from an MD ensemble. They are not a trajectory where each structure is a snapshot in time; instead, each structure is produced independently from the others. The structures are also not a statistical ensemble as each structure does not have an associated probability [105]. Instead, it can be imagined that the conformational space spanned by all structures in the ensemble generated by ExProSE is an approximation of the conformational space spanned by all structures in the actual ensemble.

### 3.1.3 Ensemble analysis

Ensembles of structures produced were iteratively aligned following the procedure described in the methodology of a previous study [166]. This aligns the ensemble without the use of a reference structure as follows:

1. Each structure in the ensemble is pairwise superimposed onto a randomly selected reference structure using the Kabsch algorithm [167].
2. An average set of coordinates is calculated for the superimposed set.
3. Each structure is pairwise superimposed onto the average coordinates using the Kabsch algorithm.
4. Steps 2-3 are repeated until the average model generated in two successive iterations changes by less than a threshold RMSD of 0.001 Å.

The average structure of the ensemble is taken as the centroid of the coordinates across the ensemble following this superimposition.

Principal components analysis (PCA) is carried out on the generated ensemble. The coordinates across the ensemble are compared to the average coordinates and a set of orthogonal motions are found that describe the variation in the ensemble. The covariance matrix $C_{ij}$ is a matrix where $i$ and $j$ represent the indices of the $3N_C$ atomic coordinates of the $N_C$ $C^{\alpha}$ atoms. $C_{ij}$ is calculated as

$$C_{ij} = \langle (x_i - \langle x_i \rangle) \cdot (x_j - \langle x_j \rangle) \rangle$$

where the averages in angle brackets are over the ensemble and $x$ represents the atomic coordinates. $C$ is then diagonalised to yield the principal components (PCs). The mean square fluctuation of each $C^{\alpha}$ atom in the structure is also calculated, giving a measure of how much the location of each atom varied across the ensemble.

PROCHECK checks the stereochemical quality of protein structures [168]. The PROCHECK overall G-factor is a log-odds score based on the observed distributions of various stereochemical parameters in reference proteins. A lower overall G-factor represents a low-probability conformation, and indicates a less

stereochemically-valid structure. Ideally, scores should be above -0.5, and values below -1.0 may need investigation [169]. PROCHECK was run for each structure generated in order to assess whether the method was able to produce stereochemically-valid protein structures.

### 3.1.4 Modulator constraint generation

In order to predict how a modulator binding to the protein affects the distribution of structures in conformational space, additional distance constraints representing the modulator need to be generated. Addition of the new distance constraints leads to ensembles that may differ significantly from the unperturbed ensemble. Potential binding sites are predicted using LIGSITE$^{cs}$ [34], which is a development of the original LIGSITE algorithm [56]. Additional constraints are generated based on pocket points predicted by LIGSITE$^{cs}$. In order to keep the number of additional points the same for pockets of different sizes, 120 points are chosen randomly. If fewer than 120 points are predicted by LIGSITE$^{cs}$, points are re-sampled. Using 120 points was found for CDK2 and CAP to add enough constraints to potentially alter the distribution of the ensemble and see an effect, but not so many that invalid structures are produced. Changing this parameter changes the strength of the perturbations but does not generally change the ranking of pockets by RMSD (see below). For CAP a different procedure was used as the location of the bound cAMP molecules is known from the crystal structure. In this case 120 fake points are added at 1.2 Å gaps in a ball around the location of the C1' atom in cAMP, while the cAMP molecules are themselves omitted from the simulation. A similar procedure to CAP is carried out for NtrC, with the oxygen to which the phosphate binds (OD1 on ASP54) taken as the centre. Selected points have distance constraints of tolerance 0.1 Å with all protein atoms within 7 Å.

In the allosteric prediction procedure ensembles are generated with additional constraints (termed 'perturbation') at selected pockets in turn, then compared to the original 'unperturbed' ensemble. Each pocket greater than a size cutoff of 13 Å$^3$ is selected, up to a maximum of 8 pockets per protein. Below this

size a small-molecule modulator is unlikely to have enough space to bind. 8 pockets gives a reasonable sampling of the surface of a protein and generally includes all sizeable pockets. The $C^{\alpha}$ RMSD between the average structure in the unperturbed ensemble and the average structure in the perturbed ensemble is used to compare ensembles. This RMSD is used to rank the perturbed pockets in terms of their predicted allosteric nature (largest to smallest RMSD). This RMSD measure may seem simplistic but gave similar or superior performance to statistical measures, such as the Mann-Whitney $U$ test on the ensemble fluctuations between the unperturbed and perturbed ensembles, during development. A pocket is considered allosteric for validation purposes if the pocket centre is within 6 Å of at least one atom of the modulator defined as the allosteric modulator in the ASD. This is similar to previous studies [30].

### 3.1.5 Datasets

#### Apo/holo dataset

Out of the 25 proteins used in a prior study [170], the 12 with apo/holo all-atom RMSD greater than 2 Å are selected in order to focus on larger conformational changes.

#### Allosteric dataset

In order to assess the performance of ExProSE as an allosteric site predictor, a dataset of known allosteric proteins was assembled. All proteins in the ASD with apo and holo structures available in the PDB were retrieved (150 proteins). Those with ASD RMSD between the apo and holo structures less than 0.25 Å were removed (14 removed) as the apo and holo structures were too similar. The proteins with more than two chains in the smallest biological assembly were removed (43 removed) as the method presented here, and the methods it is compared to, are more suited to monomers and dimers than large oligomeric assemblies. The proteins were clustered by sequence identity at a threshold of

30% and the proteins with the largest ASD RMSD were kept (17 removed). This was to ensure a diverse dataset. TM-score measures the structural similarity of proteins on a scale from 0 to 1. Some apo-holo pairs had TM-score less than 0.5 and these were removed (5 removed) as the proteins were too different from each other. Proteins with more than 1000 residues in the smallest biological assembly were also removed (9 removed). Finally, proteins with missing regions or other problems were removed (4 removed). 58 proteins with apo and holo structures remained in the dataset. Details of this dataset are made available with the source code.

### 3.1.6 Method comparison

Existing methods for ensemble generation and allosteric site prediction were compared against the method described here.

#### Ensemble generation

tCONCOORD [111] is run with default parameters. NMSim is run via the NMSim web server [113] with the default parameters for large scale motions. This produces 5 trajectories of 500 structures. Every tenth structure is taken from each trajectory to yield representative ensembles of 250 structures. A 'targeted' simulation on the NMSim web server using both apo and holo structures was also carried out but did not produce better results. Alternative parameters for tCONCOORD and NMSim are used to generate the results in Figure 3.6 and these are described in the figure.

#### Molecular dynamics

All MD runs are carried out using the GROMACS package [171]. Energy minimisation to improve the stereochemistry of T4-lysozyme structures is conducted using a steepest descent energy minimisation of 5000 steps in a vacuum and the OPLS-AA force field. MD runs of T4-lysozyme are conducted

using periodic boundary conditions, SPC water, charge-neutralising counter ions, the OPLS-AA force field and a 2 fs timestep. An initial energy minimisation is followed by a constant temperature and volume equilibration for 100 ps, then a constant pressure and temperature equilibration for 100 ps. MD is run for 50 ns. PLUMED [172] with GROMACS is used to carry out targeted MD. $C^{\alpha}$ RMSD to the target structure is used as a collective variable with a $\kappa$ value starting at 0 kJ mol$^{-1}$ Å$^{-2}$ and increasing linearly to 1000 kJ mol$^{-1}$ Å$^{-2}$ over 10 ps, and remaining at this value for the rest of the run.

**Allosteric site prediction**

LIGSITE$^{cs}$ [34] and Fpocket [35] are run with default parameters. The procedure for determining if an Fpocket pocket is allosteric is as follows: the average of the locations of the vertices in the pocket is taken as the pocket centre, and the pocket is considered allosteric if this centre was within 6 Å of at least one atom of the modulator defined as the allosteric modulator in the ASD. This is consistent with the criterion for determining LIGSITE$^{cs}$ allosteric pockets defined previously. PARS results are obtained by using the PARS web server [29]. PARS uses LIGSITE$^{cs}$, so the same criterion as LIGSITE$^{cs}$ is used to determine allosteric pockets. AlloPred is run using the offline version and default parameters. The active site residues are retrieved from the CSA [155], or from literature inspection when not available in the CSA. AlloPred uses Fpocket, so the same criterion as Fpocket is used to determine allosteric pockets. STRESS [33] is run offline using the source code. Since the output of STRESS is pocket residues, a pocket is called as allosteric if there is at least one modulator atom within 3 Å of any atom in the given residues of the pocket. This represents the modulator being close to part of the predicted pocket. This value of 3 Å is less than the value of 6 Å used previously as there are many residues which the modulator can be close to, rather than a single pocket centre.

**Computation time**

ExProSE generates 250 structures in ∼20 minutes for T4-lysozyme on a 3.1 GHz Intel Core i7 processor. For tCONCOORD the time is ∼10 minutes. NMSim is run via the NMSim web server and takes ∼5 hours. MD and targeted MD use considerably more resources, with a 50 ns run taking ∼60 hours on 16 cores (2.3 GHz Intel Xeon CPU E5-2698) or ∼20 days on the single processor above.

The ExProSE computation time is significantly improved over initial attempts due to optimisation of the Julia code. Comparison of the performance-critical step (moving the atoms to satisfy the distance constraints) shows it is significantly faster that the same algorithm implemented in Python. The stochastic nature of the simulation means it is not easily amenable to vectorisation in NumPy. The performance-critical step in Julia showed speed around 1.5 times slower than an implementation in C.

## 3.2 Results

The motivation for developing ExProSE was to explore allostery in the framework of the protein structural ensemble - see Figure 1.1. A method was needed that could use existing structural information to generate ensembles that spanned the large conformational changes relevant in allostery. We decided to develop the popular existing method CONCOORD and its extension tCONCOORD. CONCOORD converts a protein structure to a series of distance constraints based on chemical interactions (e.g. hydrogen bonds) then uses an iterative procedure to move randomly-placed atoms to satisfy the distance constraints. We alter this by using two structures as input, meaning the structures generated can span the conformational space spanned by the input structures. There are also other important differences to CONCOORD - see Section 3.1 - and the algorithm is implemented from scratch. An overview of the method is shown in Figure 3.2.

ExProSE is able to (i) generate ensembles of protein structures from two input structures and (ii) predict allosteric pockets on proteins. First, it is shown using a dataset of structural pairs that two widely-used methods for generating ensembles cannot span large conformational changes, whereas ExProSE can. The ability of ExProSE to produce native-like ensembles, i.e. ensembles close to other crystal structures, is exemplified with T4-lysozyme. ExProSE ensembles can be perturbed to reveal the location of allosteric sites, as demonstrated on CDK2. The performance of ExProSE in predicting allosteric sites is assessed on a dataset of 58 known allosteric proteins and compared to existing methods. Finally, a well-studied example of dynamic allostery and an example of activation by phosphorylation are examined to show that ExProSE can explore changes in dynamics in proteins.

### 3.2.1 Ensemble generation

Before describing ExProSE results it is instructive to examine the convergence of the structural generation simulations. The SPE error score is a measure of

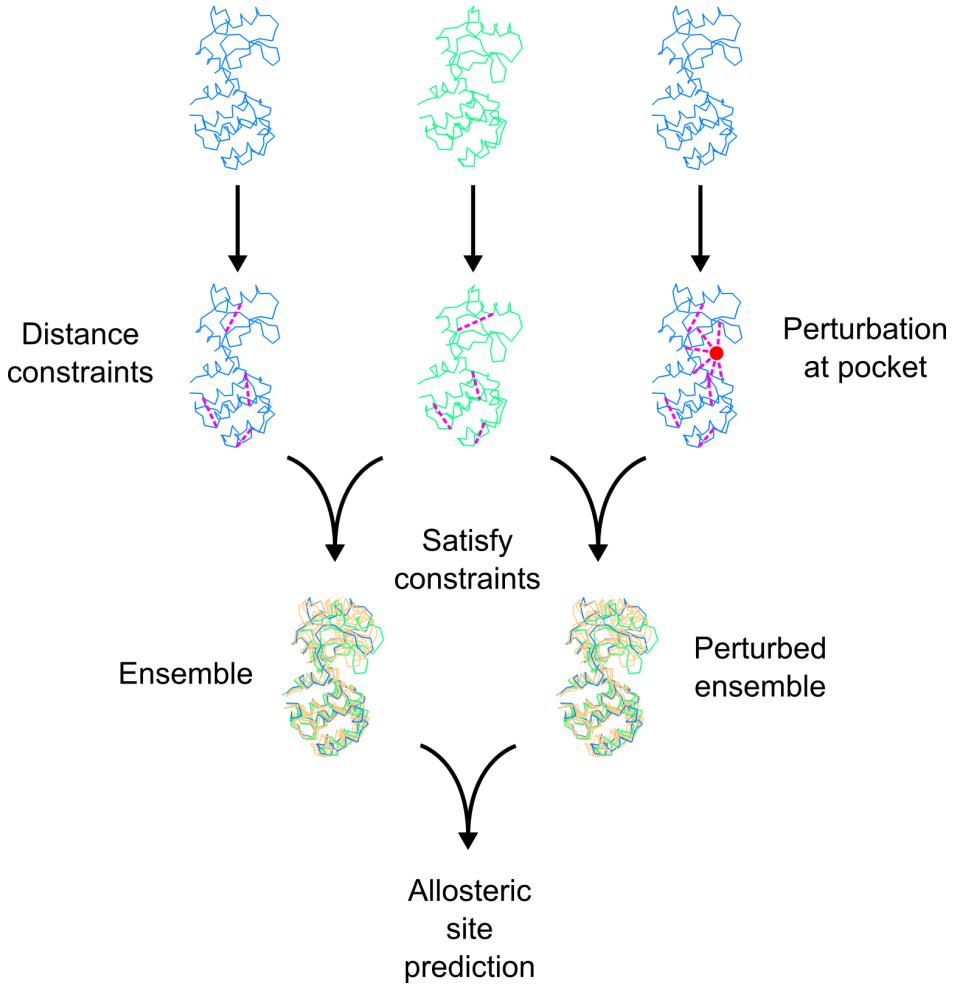



**Figure 3.2** Overview of the ExProSE computational method. Two structures of a protein (shown in blue and green) are converted to distance constraints. These constraints are combined and satisfied by an iterative procedure to generate an ensemble of structures. By adding extra constraints to one structure, for example representing perturbation at a pocket, a perturbed ensemble can be generated. This perturbed ensemble can be compared to the unperturbed ensemble to predict allosteric sites. Figure based on Graphical Abstract from Greener et al. 2017 [27].

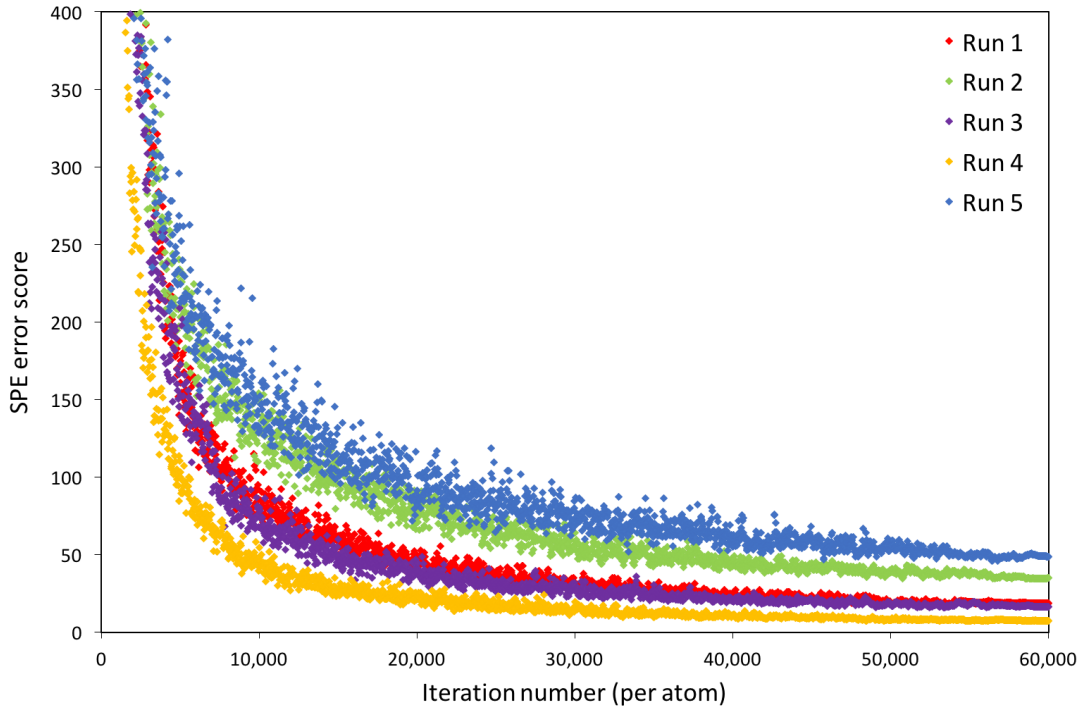



**Figure 3.3** The SPE error score, a measure of how many distance constraints are violated, shown over the course of 5 separate ExProSE runs on CDK2. Initially the random atomic coordinates give a high score. As the iterative procedure occurs atoms that violate distance constraints are moved, reducing the SPE error score.

how many distance constraints are violated - see Section 3.1. Figure 3.3 shows the SPE error score over the course of the iterative constraint correcting process for 5 independent runs of CDK2 structure generation. It can be seen that 60,000 iterations per atom is enough for the structure to improve as much as it can. Further iterations beyond this do not have a meaningful effect on the SPE error score. Similar convergence behaviour was seen with proteins across a variety of sizes. This value of 60,000 iterations per atom was used for all ExProSE runs.

It is also instructive to show how the SPE error score is related to the stereochemical quality of generated structures. Figure 3.4 shows the final SPE error score for each structure generated in an ensemble of CDK2 structures plotted against the PROCHECK overall G-factor. It can be seen that most structures have an overall G-factor between -0.4 and -0.6, and that there is an inverse cor-

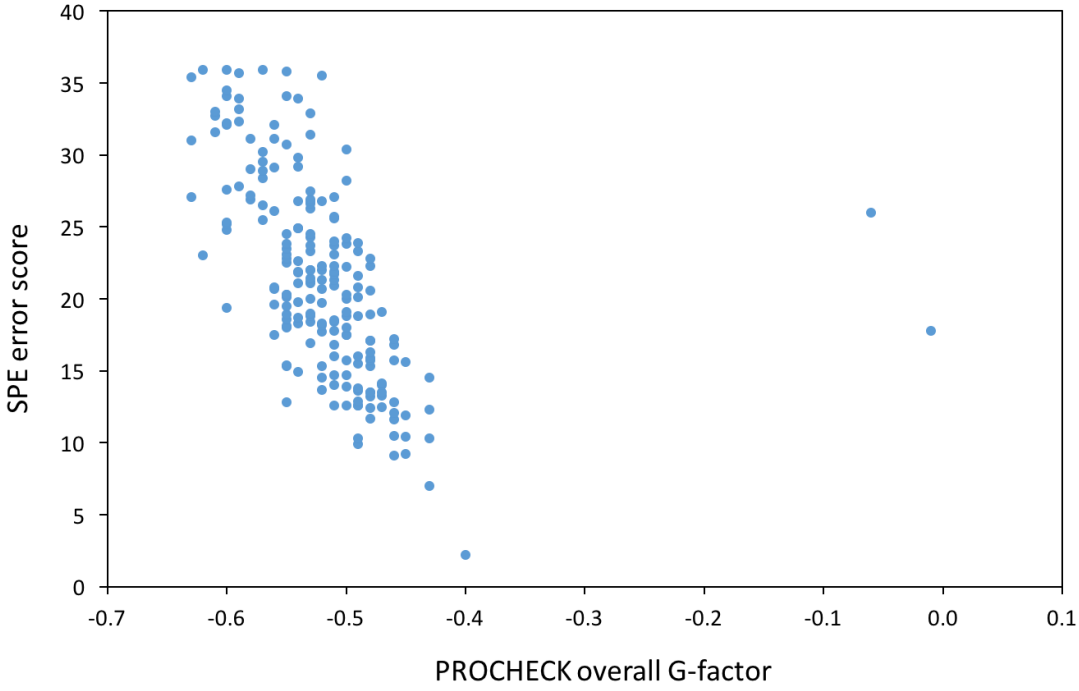



**Figure 3.4** The SPE error score of each structure in an ensemble of CDK2 structures generated using ExProSE is plotted against the PROCHECK overall G-factor of the structure. Note that an ensemble of 300 structures was generated and only the 200 structures with the lowest SPE error score were retained - see Section 3.1.

relation between overall G-factor and SPE error score. This is to be expected as structures which have more violations of the distance constraints are more likely to have poor stereochemical quality.

### Apo/holo dataset

In order to examine the ability of existing non-MD methods to generate ensembles that span conformational changes, a dataset of apo (no modulator) and holo (modulator bound) structures was used [170]. The proteins have RMSD between apo and holo structures ranging from 2 Å to 19 Å, and represent a variety of domain, subdomain and subunit motions. tCONCOORD [111] and

NMSim [113] both seek to model conformational changes such as those in the dataset. Default parameters were used to produce 250 structures for each protein from tCONCOORD and NMSim. The lowest RMSD of the structures in an ensemble to a particular crystal structure was taken as a measure of how close the ensemble came to exploring the conformational space of that crystal structure. This can be seen in Table 3.2.

When the apo structure is used as input, structures similar to the apo structure are generated by both methods. The median lowest RMSD to the apo crystal is 1.44 Å for tCONCOORD and 0.71 Å for NMSim. However, structures similar to the holo crystal are not sampled. The median lowest RMSD to the holo crystal is 4.15 Å for tCONCOORD and 4.68 Å for NMSim. In a similar manner, when the holo structure is used as input to tCONCOORD and NMSim the ensembles sample the holo structure but not the apo structure.

ExProSE, as expected because it uses both the apo and holo crystals as input, is able to generate structures close to both crystals - see Table 3.3. For 11 out of the 12 proteins ExProSE can generate a structure closer to the holo crystal than the other methods, where the other methods use the apo structure as input. For the opposite case, comparing to the apo crystal, ExProSE also generates a closer structure for 11 out of 12 proteins. Hence ExProSE is useful for generating ensembles when two or more structures are available.

The median PROCHECK overall G-factor across all generated structures is -0.99 for ExProSE, indicating that PROCHECK produces structures that are generally acceptable. The values for NMSim and tCONCOORD are -0.32 and -1.83 respectively, indicating that NMSim produces good quality structures and tCONCOORD produces structures with poor stereochemical quality. The stereochemistry of generated structures can be improved by energy minimisation (see below).

**T4-lysozyme**

Here, we demonstrate that ExProSE can generate structures close to crystals not used as input. Lysozymes damage bacterial cell walls by catalysing the

| | | | | | tCONCOORD from apo | | tCONCOORD from holo | | NMSim from apo | | NMSim from holo | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Lowest RMSD from 250 structures to apo/holo crystal / Å | | | | | | | |
| **Protein name** | **Apo PDB** | **Holo PDB** | **RMSD/ Å** | ***N*** | **Apo** | **Holo** | **Apo** | **Holo** | **Apo** | **Holo** | **Apo** | **Holo** |
| OxyR transcription factor | 1I6A | 1I69 | 2.44 | 206 | 1.18 | 2.69 | 2.66 | 1.12 | 1.04 | 2.61 | 2.51 | 0.72 |
| Ferric binding protein | 1D9V | 1MRP | 2.68 | 309 | 1.22 | 1.81 | 1.88 | 1.41 | 0.62 | 2.07 | 2.31 | 0.71 |
| Aspartate receptor | 1LIH | 2LIG | 2.77 | 157 | 1.16 | 2.73 | 2.94 | 1.48 | 0.94 | 2.45 | 2.65 | 0.80 |
| HIV-1 rev. transcriptase | 2HMI | 3HVT | 3.81 | 555 | 2.49 | 4.11 | 4.66 | 3.44 | 0.64 | 3.28 | 3.14 | 0.78 |
| Maltose binding protein | 1OMP | 3MBP | 3.88 | 370 | 0.97 | 2.62 | 2.66 | 0.89 | 0.71 | 2.35 | 2.39 | 0.57 |
| Small G protein Arf6 | 1E0S | 2J5X | 4.44 | 164 | 0.99 | 4.18 | 4.23 | 0.96 | 0.66 | 4.00 | 4.23 | 0.86 |
| Immunoglobulin | 1MCP | 4FAB | 5.95 | 214 | 1.65 | 3.60 | 3.80 | 1.51 | 0.62 | 5.35 | 3.63 | 0.79 |
| Myosin | 1VOM | 2AKA | 6.23 | 730 | 2.60 | 5.11 | 5.63 | 2.38 | 0.73 | 5.53 | 5.77 | 0.63 |
| Adenylate kinase | 4AKE | 1AKE | 7.19 | 214 | 1.70 | 4.88 | 6.00 | 1.18 | 0.58 | 6.16 | 6.09 | 0.74 |
| Serpin | 1PSI | 7API | 8.96 | 372 | 1.20 | 8.71 | 8.93 | 1.51 | 0.71 | 8.22 | 8.97 | 0.97 |
| GroEL | 1AON | 1OEL | 12.6 | 524 | 3.01 | 9.72 | 9.61 | 2.45 | 0.87 | 10.8 | 10.1 | 0.48 |
| Topoisomerase II | 1BGW | 1BJT | 18.8 | 664 | 3.36 | 17.5 | 17.0 | 3.34 | 0.81 | 18.0 | 17.3 | 0.65 |
| | | **Median across all proteins** | | | 1.44 | 4.15 | 4.45 | 1.50 | 0.71 | 4.68 | 3.93 | 0.73 |

**Table 3.2** Comparison of ensemble generation methods. The columns Apo PDB and Holo PDB refer to the PDB IDs of the apo and holo structures used. RMSD is the all-atom RMSD in Å between the apo and holo structures. The rows are ordered by increasing RMSD. *N* is the number of residues in common between the apo and holo chains used. The values on the right are the lowest RMSD in Å of the structures in an ensemble produced using the method and input indicated, to the crystal structure indicated. A low value indicates that the ensemble sampled a structure close to the crystal structure. The median of the lowest RMSDs for each method/input combination is also given.

| | | | | | Lowest RMSD from 250 ExProSE structures to apo/holo crystal / Å | |
|---|---|---|---|---|---|---|
| **Protein name** | **Apo PDB** | **Holo PDB** | **RMSD / Å** | ***N*** | **Apo** | **Holo** |
| OxyR transcription factor | 1I6A | 1I69 | 2.44 | 206 | 1.02 | 1.16 |
| Ferric binding protein | 1D9V | 1MRP | 2.68 | 309 | 0.90 | 1.08 |
| Aspartate receptor | 1LIH | 2LIG | 2.77 | 157 | 1.25 | 0.88 |
| HIV-1 rev. transcriptase | 2HMI | 3HVT | 3.81 | 555 | 1.84 | 1.45 |
| Maltose binding protein | 1OMP | 3MBP | 3.88 | 370 | 0.85 | 1.50 |
| Small G protein Arf6 | 1E0S | 2J5X | 4.44 | 164 | 1.70 | 1.88 |
| Immunoglobulin | 1MCP | 4FAB | 5.95 | 214 | 3.90 | 5.33 |
| Myosin | 1VOM | 2AKA | 6.23 | 730 | 2.38 | 1.89 |
| Adenylate kinase | 4AKE | 1AKE | 7.19 | 214 | 3.15 | 1.98 |
| Serpin | 1PSI | 7API | 8.96 | 372 | 1.08 | 1.01 |
| GroEL | 1AON | 1OEL | 12.6 | 524 | 3.13 | 3.70 |
| Topoisomerase II | 1BGW | 1BJT | 18.8 | 664 | 3.54 | 5.10 |
| | | | **Median across all proteins** | | 1.77 | 1.69 |

**Table 3.3** Ability of ExProSE ensembles to reach apo and holo structures. The columns Apo PDB, Holo PDB, RMSD and *N* are the same as in Table 3.2. The values on the right are the lowest RMSD in Å of the structures in an ExProSE ensemble to the crystal structure indicated. A low value indicates that the ensemble sampled a structure close to the crystal structure. The median of the lowest RMSDs is also given.

hydrolysis of peptidoglycans. Bacteriophage T4-lysozyme is a suitable protein for analysing conformational variability as there are many crystal structures available and MD simulations of the protein have shown that simulations up to 200 ns do not reliably reach both the open and closed conformations [111]. The pairwise RMSDs of the crystals range from 0.64 Å to 4.25 Å.

An ensemble was generated using ExProSE from the open (PDB ID 169L, chain E) and closed (PDB ID 2LZM) conformations. Four random structures from this ensemble are shown compared to the open and closed crystal structures in Figure 3.5A. PCA can be carried out on an ensemble of structures to find the orthogonal motions that describe the variation in the ensemble. Figure 3.5B shows the projections of the generated ensemble and the 38 crystal structures used in a prior study [173] onto the first and second PCs, which account for 70% and 12% of the motion respectively. The dominant first eigenvector corresponds to opening and closing of the structure. It can be seen that the method is able to sample conformations corresponding to experimentally-observed structures, as the ensembles largely overlap.

Ensembles produced by tCONCOORD starting from the open and closed structures separately are shown in Figure 3.5C. As demonstrated previously on other proteins, the ensemble generated from the open structure cannot reach all the way to the closed structure, and vice versa. The tCONCOORD ensembles also sample structures not found in the ensemble of crystal structures, particularly when using the open conformation as input. This tendency of tCONCOORD to produce ensembles with too much structural variability was also noted by the authors [174].

Ensembles produced by NMSim starting from the open and closed structures separately are shown in Figure 3.5D. In this case, the ensemble generated from the open and closed structures can largely span the conformational space. Similar to tCONCOORD, regions not explored by the crystals are sampled by NMSim. For example, there is one model in the ensemble generated from the open structure that has an RMSD of 7.38 Å to the nearest crystal structure.

Alternative parameters were also used for tCONCOORD and NMSim to see

how the ensembles varied - see Figure 3.6. For tCONCOORD decreasing the upper bound for long range constraints, and/or turning off close pairs as constraints, had little effect on the distribution of the ensembles. For NMSim using the parameters for small scale motions led to ensembles that could not span the conformational space. In each case the default parameters gave similar or better coverage of the conformational space of the crystals by visual inspection, and were hence used for the analysis below.

T4-lysozyme was also studied with MD. 50 ns MD runs starting from the closed conformation were not able to reach the open conformation and vice versa - see Figure 3.7A and 3.7B. Targeted MD runs starting from the closed conformation and targeting the open conformation (and vice versa) were also carried out. In targeted MD the atoms are guided to a target structure with the use of a steering force that seeks to minimise the RMSD of the structure to the target structure. These ensembles can be seen in Figure 3.7C and 3.7D, and are generally able to cross conformational space over the course of around 20 ns. However beyond this time they show unpredictable behaviour and can deviate from the experimental structures. Retaining only the structures up to 20 ns, as in Figure 3.7C and 3.7D, gives ensembles that largely overlap with the experimental structures.

By combining tCONCOORD, NMSim and targeted MD ensembles generated using the open and closed structures as input, a fair comparison to ExProSE can be made. A generated ensemble should ideally contain models close to all the crystal structures. The degree to which this occurs for ExProSE ensembles, and combined ensembles for tCONCOORD, NMSim and targeted MD up to 20 ns, is shown in Figure 3.8A. It can be seen that ExProSE is able to generate structures close to all crystals, with all crystals having an RMSD of 1.7 Å or less to a generated structure. For 26 out of 38 crystals ExProSE generates a model closer to the crystal than NMSim. It generates a closer model than tCONCOORD in all cases. For 15 out of 38 crystals ExProSE generates a model closer to the crystal than targeted MD. However, this is the case for 14 out of the 27 structures that have an RMSD of more than 1.0 Å to both the open and closed reference structures. Of these 27, ExProSE performs better for all of the

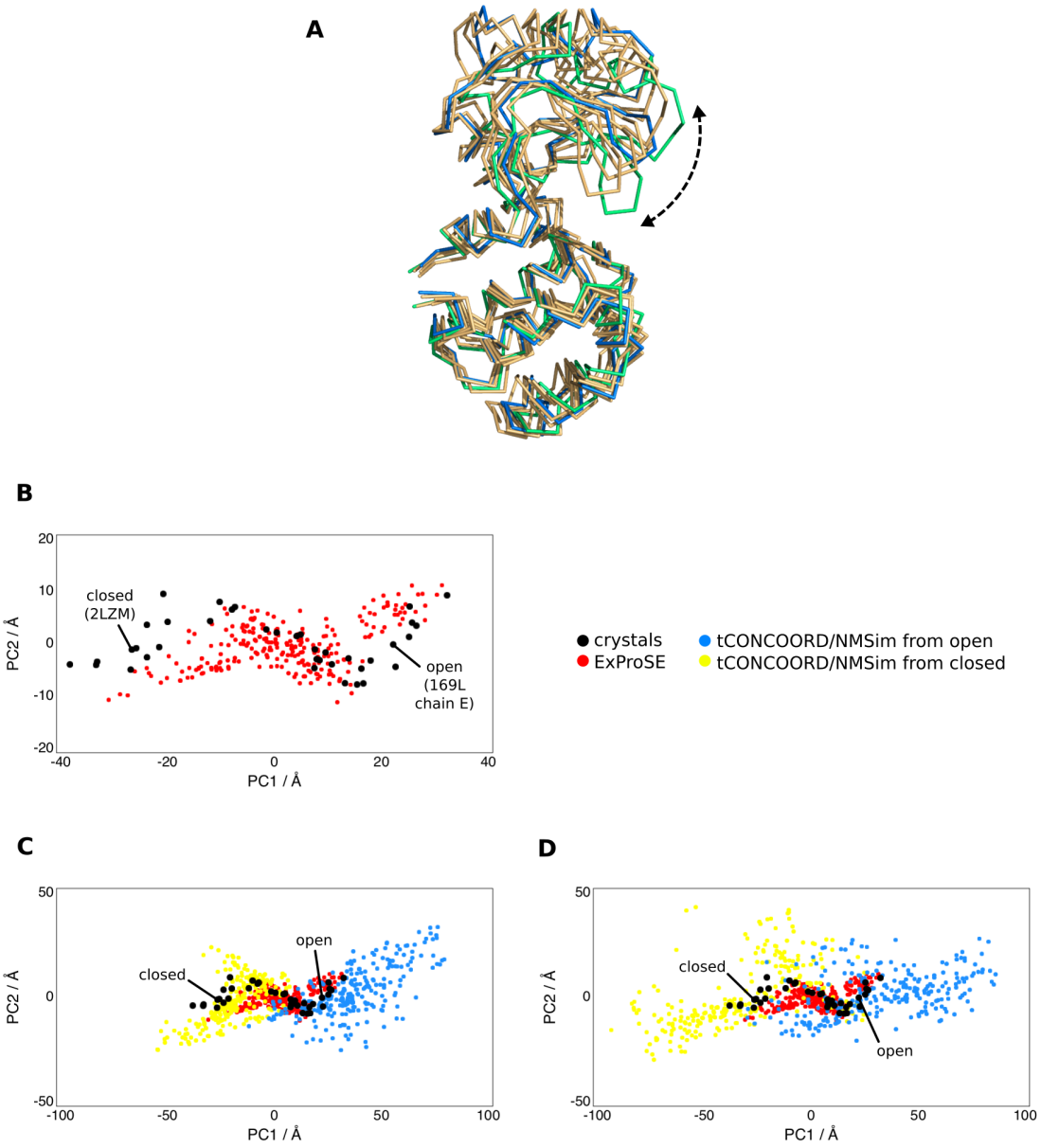



**Figure 3.5** Caption on following page.

**Figure 3.5** T4-lysozyme ensembles. (A) Four structures generated from ExProSE using the open (PDB ID 169L, chain E) and closed (PDB ID 2LZM) conformations as input are shown in orange. The crystal structures of the open and closed conformations are shown in blue and green respectively for reference. The arrow shows the opening motion caused by the breaking of a hydrogen bond between ARG137 and GLU22. (B) Projections of the 38 crystal structures used in a prior study [173] onto the first (x-axis) and second (y-axis) PCs of the PCA of the crystal structures, which account for 70% and 12% of the motion respectively (black dots). Projections from the ensembles generated with ExProSE are also shown (red dots). (C) Projections of two tCONCOORD ensembles on the same plot as (B). An ensemble using the open structure as input (blue dots) and an ensemble using the closed structure as input (yellow dots) are shown. (D) Projections of two NMSim ensembles with parameters for large scale motions on the same plot as (B). An ensemble using the open structure as input (blue dots) and an ensemble using the closed structure as input (yellow dots) are shown. Figure based on Figure 1 from Greener et al. 2017 [27].

4 structures that have an RMSD of more than 1.5 Å. Hence ExProSE is able to generate better models than the other methods for crystals which are far from either input structure, as seen on the right side of Figure 3.8A. The PROCHECK overall G-factor of the closest models for each method is shown in Figure 3.8B. ExProSE is able to produce models of acceptable quality close to all the crystals, even for those further from the input structures.

In order to determine whether the stereochemical quality of generated structures could be improved, energy minimisation was carried out on all structures. For all methods, energy minimisation improved median PROCHECK overall G-factors. Across the ensembles the median values increased from the range [-2.23, -0.45] to the range [-0.31, -0.17] - see Table 3.4. This shows that stereochemical problems in generated structures can in general be improved by energy minimisation, important if using generated structures for docking studies.

By using two input structures rather than one, ExProSE is able to produce models of acceptable quality close to other crystal structures. It can explore conformational space better than methods that use a single structure as input.

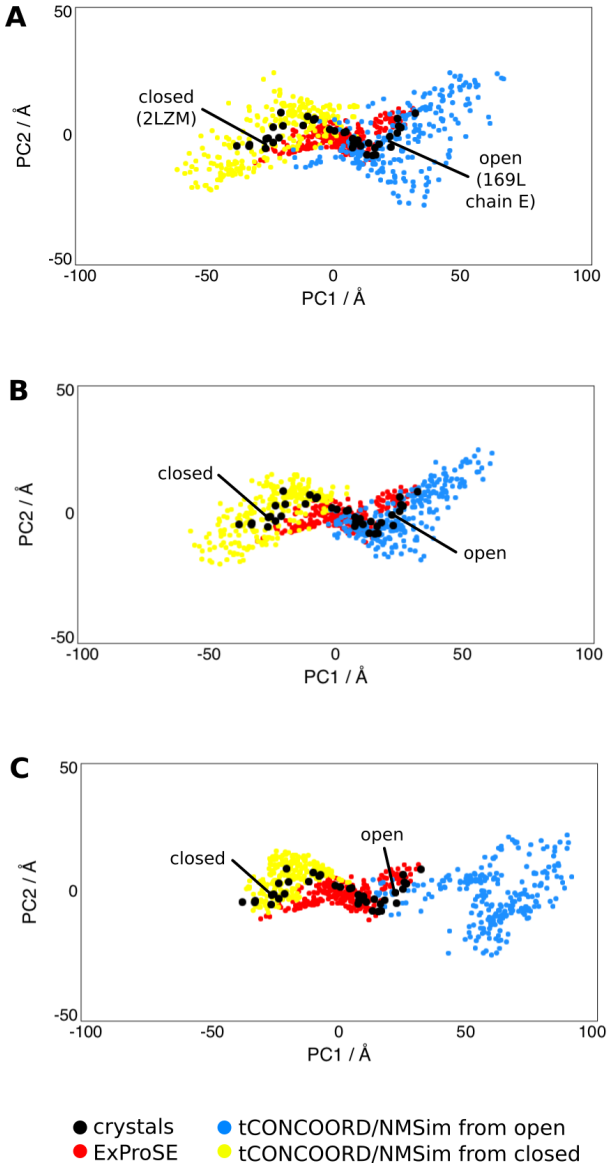



**Figure 3.6** Caption on following page.

**Figure 3.6** Ensemble generation for T4-lysozyme with different parameters. Projections of tCONCOORD/NMSim ensembles from the open (blue dots) and closed (yellow dots) structures onto the PCA of the crystal structures are shown. Similar to Figure 3.5, in each graph the projections of the crystals (black dots) and projections from the ensembles generated with ExProSE (red dots) are also shown. (A) tCONCOORD ensembles with the upper bound for long range constraints set to 1.3 Å (default 2.0 Å). (B) tCONCOORD ensembles with the upper bound for long range constraints set to 1.3 Å and close pairs not used as constraints. (C) NMSim ensembles using the default parameters for small scale motions. Figure based on Figure S1 from Greener et al. 2017 [27].

| **Method** | **Structure(s) used** | **Median overall G-factor before energy minimisation** | **Median overall G-factor after energy minimisation** |
|---|---|---|---|
| ExProSE | open and closed | -0.58 | -0.26 |
| tCONCOORD | open | -2.23 | -0.31 |
| tCONCOORD | closed | -2.09 | -0.27 |
| NMSim | open | -0.50 | -0.31 |
| NMSim | closed | -0.45 | -0.29 |
| targeted MD | starting open, targeting closed | -0.56 | -0.17 |
| targeted MD | starting closed, targeting open | -0.57 | -0.20 |

**Table 3.4** Improvement in stereochemical quality on energy minimisation. The structures in each ensemble were analysed with PROCHECK and the median overall G-factor across the ensemble was noted. The median of the overall G-factor of each structure after energy minimisation was also recorded.

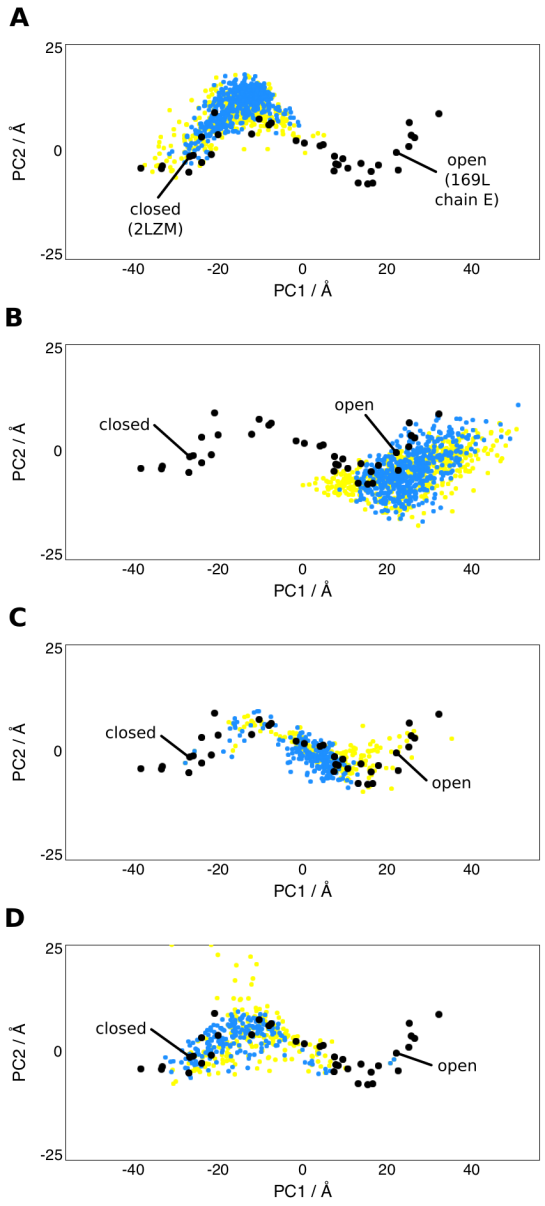



**Figure 3.7** Caption on following page.

**Figure 3.7** MD T4-lysozyme ensembles. Projections of two repeats of a particular MD run onto the PCA of the crystal structures are shown (blue and yellow dots), with snapshots taken every 100 ps. Similar to Figure 3.5, in each graph the projections of the crystals are also shown (black dots). (A) 50 ns MD runs starting from the closed structure (PDB ID 2LZM). (B) 50 ns MD runs starting from the open structure (PDB ID 169L). (C) 20 ns targeted MD runs starting from the closed structure and targeting the open structure. (D) 20 ns targeted MD runs starting from the open structure and targeting the closed structure. Figure based on Figure 2 from Greener et al. 2017 [27].

### 3.2.2 Ensemble perturbation for CDK2

Here, we demonstrate that ExProSE ensembles can be perturbed to reveal modulating sites. CDK2 is a protein kinase important in cell cycle progression - see Section 1.5. An ExProSE ensemble was generated using the apo native structure (PDB ID 1HCL) and the holo structure bound to two ANS molecules in an allosteric site (PDB ID 3PXF). The ANS-bound structure is inactive, as ANS binding causes a conformational shift in the αC-helix that prevents cyclin binding [134]. A further screening study has found potential modulators for the ANS binding site [151].

Figure 3.9A shows the pockets predicted by LIGSITE$^{cs}$ [34] on CDK2 bound to two ANS molecules. The ensemble perturbation procedure was carried out at each of the 8 pocket centres as described in Section 3.1. In brief, additional constraints are added representing a modulator bound in the selected pocket. Projections of the structures of the unperturbed ensemble and the structures of the ensemble with perturbation at the pocket centre are shown in Figure 3.9B, one graph per pocket centre. The third PC was chosen for visualisation instead of the second as it represents the inactivating motion of the αC-helix, whereas the second PC represents a rotation in the region of the protein considered to be functionally less important, the C-lobe.

Site 1 in Figure 3.9A and 3.9B is the known allosteric pocket - see Section 1.5. Simulating a modulator there shifts the ensemble towards the inactive state, agreeing with previous experimental data [134]. Site 2 is the ATP-binding site and there is no change in the ensemble by simulating a modulator there. This is

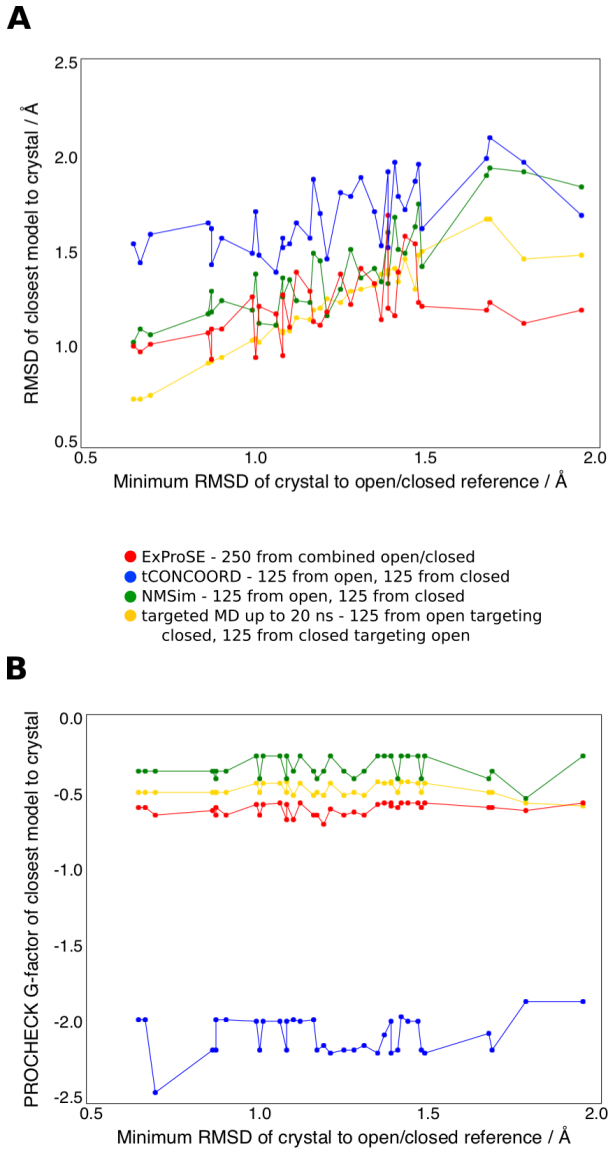



**Figure 3.8** Closest models from each ensemble to T4-lysozyme crystal structures. (A) The RMSD of the closest model from each generated ensemble to the crystal structures. The crystal structures are sorted by the lower of the two RMSD values to the open and closed crystals used as input. The crystals used as inputs are omitted from the graph. (B) PROCHECK overall G-factors of the closest model from each generated ensemble to the crystal structures. The crystal structures are sorted as in (A). Figure based on Figure 3 from Greener et al. 2017 [27].

encouraging as ATP binding does not cause structural changes that lead to cyclin dissociation. Site 3 is found in a pocket near the activation segment. A shift in the ensemble towards the inactive state is seen on perturbation at this site. In fact this site is close to a potential allosteric site suggested in another computational study [175], and is part of the region associated with cyclin binding. This indicates that the site could potentially be an allosteric site, though further effort would be required to determine whether it is druggable. See Chapter 4 for further studies on this site. Simulating modulators at sites 4 to 8 does not shift the ensemble, suggesting that binding at these sites is unable to cause an allosteric effect. No allosteric modulators have been reported experimentally for these sites.

### 3.2.3 Allosteric site prediction

Systematic methods to predict allosteric sites on proteins are necessary to utilise the potential advantages of allosteric drugs. A diverse dataset of 58 apo/holo pairs representing the unbound protein and the protein bound to a known allosteric modulator was assembled from the ASD [86]. This dataset showed a large range in protein size (153 to 955 residues) and included a variety of proteins including transcriptional regulators, transporters and protein kinases.

LIGSITE$^{cs}$ was used to predict pockets on the holo crystal structures and ExProSE was used to generate a perturbed ensemble for each pocket centre as described in Section 3.1. These perturbed ensembles were used to rank the pockets in terms of predicted allosteric effect. In this study a correct prediction for a protein indicates that an allosteric pocket was ranked first or second. This criterion was chosen as a measure of success because typically the top few pockets predicted by a method would be examined and studied further.

The ability of ExProSE to predict allosteric pockets on the dataset is compared to existing allosteric prediction methods, which are run with the holo crystal structures as input. This was found to give better results for the existing methods than using the apo crystals. PARS [29] uses NMA with and without a predicted modulator to predict changes in flexibility. STRESS [33] is an

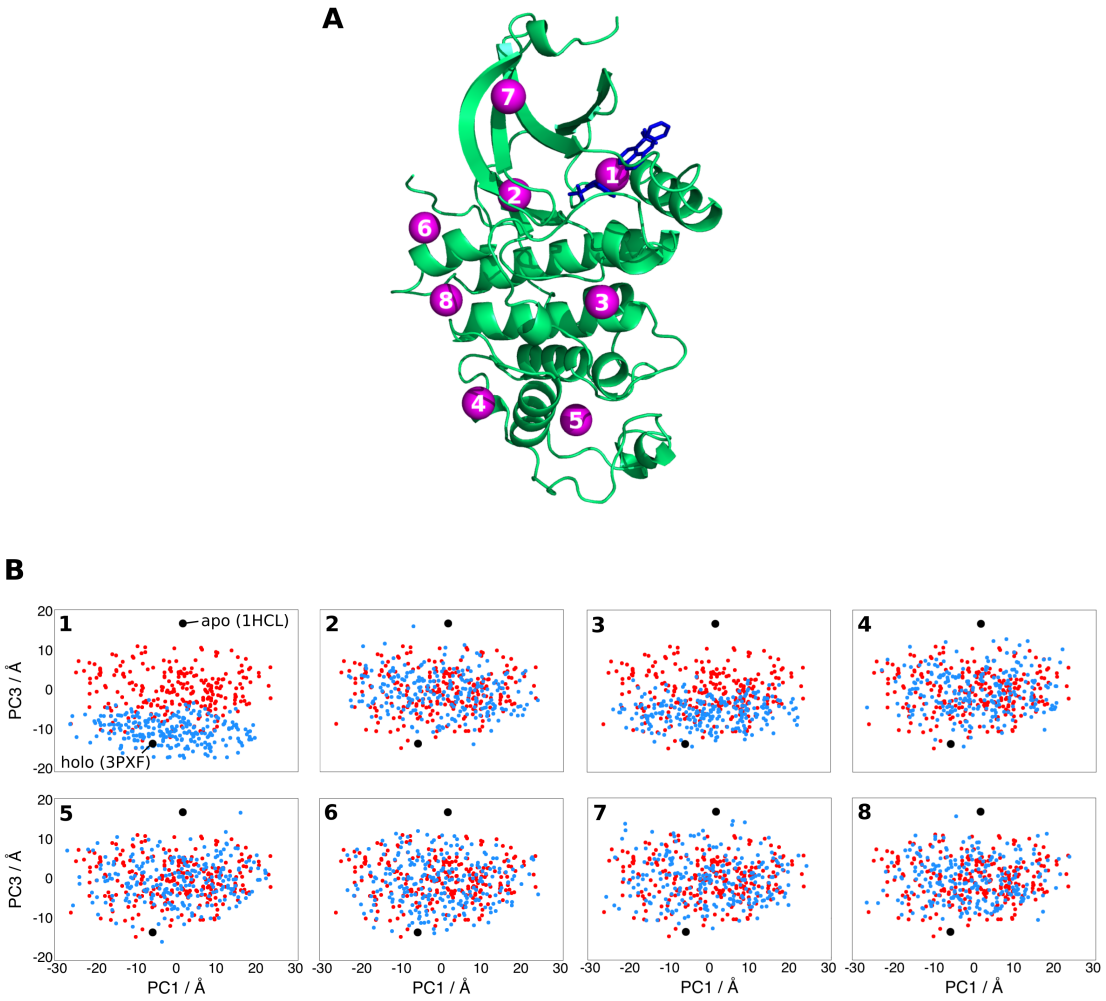



**Figure 3.9** CDK2 pockets and projections of ensembles. (A) CDK2 in its holo conformation bound to two ANS molecules in the allosteric site (PDB ID 3PXF). CDK2 is shown as a green cartoon with the two bound ANS shown as blue sticks. Pocket centres predicted by LIGSITE$^{cs}$ are shown as purple spheres. The pockets are numbered by descending volume. Pocket 1 represents the ANS allosteric pocket. Pocket 2 represents the ATP-binding pocket. (B) Structures generated using ExProSE, with input structures the apo and holo structures (PDB ID 1HCL and 3PXF respectively), are shown as red dots. The axes are projections onto the first (x-axis) and third (y-axis) PCs of the ExProSE ensemble, which account for 35% and 8% of the motion respectively. The blue dots represent the structures in the ensemble with perturbation at pocket centres 1-8 from (A). Figure based on Figure 4 from Greener et al. 2017 [27].

implementation of the earlier binding leverage algorithm [32], which models how perturbations due to binding couple to the motions of the protein as expressed by low-frequency normal modes. AlloPred [21] uses perturbation of normal modes and pocket features in a machine learning approach to predict allosteric pockets - see Chapter 2. The methods compared differ from those in Section 2.2. This is because STRESS was not released during the work in Chapter 2 so was not used there, and AlloSite was found to correspond exactly to Fpocket predictions so was not used here (see Section 2.3). It should be noted that different criteria are used to define an allosteric pocket for each method, due to the nature of their output - see Section 3.1. For 27 of 58 proteins ExProSE ranked an allosteric pocket first or second, performing better than the other three methods. This is shown in Table 3.5. Only 7 proteins have an allosteric pocket ranked first or second by all four methods. In 3 cases ExProSE makes a correct prediction for a protein while none of the other methods did.

The performance of the allosteric prediction methods is also compared to the pocket prediction methods LIGSITE$^{cs}$ and Fpocket [35] in Table 3.5. LIGSITE$^{cs}$ and Fpocket are effective at finding allosteric sites, both ranking an allosteric pocket first or second for 31 out of 58 proteins, even though they are not designed specifically for allosteric site prediction. This is not too surprising as the holo structures were used as input, so the modulator had a reasonable chance of being in one of the two largest pockets. However ExProSE is still valuable as it finds smaller, less obvious allosteric pockets. This could be due to the extra structural information used as input. For example, ExProSE in 6 cases finds sites not ranked in the top 2 by LIGSITE$^{cs}$ and in 8 cases finds sites not ranked in the top 2 by Fpocket. ExProSE shows the best complementarity to the pocket prediction methods along with STRESS, which makes fewer correct predictions. ExProSE also gives information on how the ensemble may be affected by the modulators, as demonstrated for CDK2 in Figure 3.9, allowing inspection of the predicted structural and dynamic changes arising from perturbation.

The performance on each protein by each method is shown in Table 3.6. This is to our knowledge the first systematic comparison of multiple allosteric predic-

| Method | Correct in top 2 (out of 58) | Unique from LIGSITE$^{cs}$ | Unique from Fpocket |
|---|---|---|---|
| ExProSE | 27 | 6/27 | 8/27 |
| PARS | 25 | 3/25 | 7/25 |
| STRESS | 18 | 6/18 | 8/18 |
| AlloPred | 26 | 5/26 | 1/26 |
| LIGSITE$^{cs}$ | 31 | - | 8/31 |
| Fpocket | 31 | 8/31 | - |

**Table 3.5** Performance of allosteric site prediction methods on a dataset of 58 known allosteric proteins. Correct in top 2 is the number of proteins for which the method successfully ranked an allosteric pocket first or second. The definition for an allosteric pocket is given in Section 3.1. The number of correct predictions by each method that are unique from the correct predictions of LIGSITE$^{cs}$ and Fpocket is also shown. STRESS could not run on 4 proteins as they were too small.

tion methods. 49 of 58 proteins had an allosteric pocket ranked first or second by at least one of the 6 methods compared. This complementarity indicates the potential for a meta approach that combines predictions from multiple methods.

### 3.2.4 Dynamic allostery in CAP

Catabolite activator protein (CAP) is a transcriptional activator that exists as a homodimer. Each subunit has a ligand-binding domain at the N-terminus and a DNA-binding domain at the C-terminus. Two cAMP molecules bind CAP with negative cooperativity and increase the affinity of the protein for DNA. The negative cooperativity of cAMP binding is a well-studied example of dynamic, or entropic, allostery [89]. The binding of one cAMP does not significantly change the structure of the other cAMP-binding site, but changes in the dynamics at the other site make binding entropically unfavourable [89, 176].

ExProSE was used to explore the dynamic allostery in CAP. A single structure was used as input (PDB ID 1G6N) and four ensembles were generated with perturbations representing no cAMP bound (Apo-CAP), cAMP bound to chain A, cAMP bound to chain B, and cAMP bound to both chains A and B. Note this

| Protein name | Apo PDB | Holo PDB | Apo chains | Holo chains | *N* | ExProSE | PARS | STRESS | AlloPred | LIGSITEcs | Fpocket |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Pyruvate kinase | 1A3X | 1A3W | AB | AB | 955 | | | | | | |
| Antithrombin-III | 1ANT | 3KCG | I | I | 399 | | | | | | |
| HIV-1 integrase | 1BIZ | 4CHO | AB | AB | 276 | | | | | | |
| Chorismate mutase | 1CSM | 2CSM | AB | 2 x A | 490 | | | | | | |
| Plasminogen activator inhibitor 1 | 1DB2 | 4AQH | A | A | 377 | | | | | | |
| HTH-type transcriptional repressor purR | 1DBQ | 1JH9 | AB | 2 x A | 550 | | | | | | |
| Ribose-phosphate pyrophosphokinase | 1DKR | 1DKU | AB | AB | 588 | | | | | | |
| Fatty acid metabolism regulator protein | 1E2X | 1H9G | 2 x A | 2 x A | 444 | | | | | | |
| Androgen receptor | 1E3G | 4K7A | A | A | 242 | | | | | | |
| Herpesvirus protease | 1FL1 | 4P3H | A | A | 153 | | | | | | |
| Glutamate receptor 2 | 1FTO | 3ILT | A | B | 257 | | | | | | |
| Annexin A5 | 1HVG | 1HAK | A | A | 313 | | | | | | |
| Neurolysin, mitochondrial | 1I1I | 4FXY | P | P | 664 | | | | | | |
| Cell division control protein 4 | 1NEX | 3MKS | AB | CD | 572 | | | | | | |
| Phospho-2-dehydro-3-deoxyheptonate aldolase | 1OFP | 1OFR | AB | GH | 628 | | | | | | |
| Organophosphorus hydrolase | 1PTA | 1QW7 | 2 x A | AB | 636 | | | | | | |
| Ribonucleotide reductase | 1RLR | 3UUS | A | A | 727 | | | | | | |
| Cytochrome P450 3A4 | 1W0E | 1W0F | A | A | 452 | | | | | | |
| Acetyl-CoA carboxylase | 1W93 | 1W96 | A | A | 549 | | | | | | |
| Hypothetical biotin–[acetyl-CoA-carboxylase] ligase | 1WQ7 | 2DVE | AB | AB | 456 | | | | | | |
| Putative uncharacterized protein PH0207 | 1X0M | 3ATH | A | A | 403 | | | | | | |
| Integrin α-L | 1ZON | 1RD4 | A | A | 181 | | | | | | |
| Pyruvate dehydrogenase kinase isoform 2 | 2BTZ | 2BU2 | 2 x A | 2 x A | 708 | | | | | | |
| Farnesyl pyrophosphate synthase | 2F7M | 3N45 | 2 x F | 2 x F | 682 | | | | | | |
| Fructose-1,6-bisphosphatase | 2FBP | 1Q9D | AB | AB | 630 | | | | | | |
| Protein arginine N-methyltransferase 3 | 2FYT | 3SMQ | A | A | 299 | | | | | | |
| Glycogen phosphorylase | 2GPN | 1PYG | A | A | 787 | | | | | | |
| Glutamate racemase | 2JFX | 4B1F | AB | AB | 498 | | | | | | |
| Myosin-2 heavy chain | 2JJ9 | 2JHR | A | A | 692 | | | | | | |
| Ubiquitin-conjugating enzyme E2 R1 | 2OB4 | 3RZ3 | A | A | 153 | | | | | | |
| Cytosolic purine 5′-nucleotidase | 2XCX | 2JC9 | 2 x A | 2 x A | 916 | | | | | | |
| cAMP receptor protein | 3D0S | 3I54 | AB | AB | 422 | | | | | | |
| Endothelial PAS domain-containing protein 1 | 3F1P | 3H82 | AB | AB | 222 | | | | | | |
| Acetylcholinesterase | 3GEL | 2J3Q | A | A | 527 | | | | | | |
| NAD-dependent deacetylase sirtuin-3, mitochondrial | 3GLU | 4C7B | AB | AB | 261 | | | | | | |
| FimX | 3HV9 | 3HV8 | A | A | 242 | | | | | | |
| Glucokinase | 3IDH | 4ISE | A | A | 419 | | | | | | |
| Glutamate receptor ionotropic, NMDA 2B | 3JPW | 3QEL | A | B | 349 | | | | | | |
| Global nitrogen regulator | 3LA7 | 3LA3 | AB | AB | 382 | | | | | | |
| Genome polyprotein | 3MWV | 4JTZ | A | A | 559 | | | | | | |
| β-lactamase SHV-1 | 3N4I | 1VM1 | A | A | 265 | | | | | | |
| DNA double-strand break repair Rad50 ATPase | 3QG5 | 3THO | A | A | 349 | | | | | | |
| N-acetylglutamate kinase / N-acetylglutamate synthase | 3S7Y | 4KZT | AX | AX | 862 | | | | | | |
| Leucine transporter | 3TU0 | 2QEI | A | A | 509 | | | | | | |
| 6-phosphofructokinase isozyme 2 | 3UMP | 3CQD | AB | AB | 612 | | | | | | |
| Kinesin-like protein KIF11 | 4A28 | 4BXN | A | A | 330 | | | | | | |
| Eukaryotic translation initiation factor 4E | 4BEA | 4TQC | A | A | 174 | | | | | | |
| Penicillin binding protein 2 prime | 4BL2 | 3ZG0 | A | A | 636 | | | | | | |
| CAMP-dependent protein kinase | 4DFY | 4DFX | A | E | 311 | | | | | | |
| Casein kinase II | 4DGL | 3H30 | C | A | 333 | | | | | | |
| Mitogen-activated protein kinase 14 | 4E5B | 3NNX | A | A | 321 | | | | | | |
| PelD | 4ETX | 4ETZ | A | A | 285 | | | | | | |
| Caspase 7 | 4FDL | 4FEA | AB | AB | 365 | | | | | | |
| Glucose-1-phosphate thymidylyltransferase | 4HO0 | 4HO9 | A | A | 285 | | | | | | |
| GTPase Kras | 4LPK | 4LUC | A | A | 156 | | | | | | |
| CRP transcriptional dual regulator | 4N9H | 4N9I | AB | AB | 402 | | | | | | |
| 2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase | 4NAI | 2YC3 | 2 x A | 2 x A | 428 | | | | | | |
| Adenylate cyclase type 10 | 4OYW | 4USW | A | A | 458 | | | | | | |
| **Found in top 2 predicted - out of 58 (54 for STRESS)** | | | | | | 27 | 25 | 18 | 26 | 31 | 31 |

**Table 3.6** Caption on following page.

**Table 3.6** Performance of allosteric site prediction methods on a dataset of 58 known allosteric proteins. Apo PDB and Holo PDB refer to the PDB IDs of the apo and holo structures used. Apo chains and Holo chains are the chains utilised from the apo and holo structures. 2 x A means chain A is duplicated as part of a biological assembly. *N* is the number of residues in common between the apo and holo chains used. A green square indicates that the method ranked an allosteric pocket first or second for that protein. The definition for an allosteric pocket is given in Section 3.1. A red square indicates that the method failed to rank an allosteric pocket first or second. STRESS could not run on 4 proteins as they were too small - this is indicated by a yellow square.

is the only case in this study where a single structure, rather than two, was used as input. The mean square fluctuation across each ensemble was calculated for each residue and gives a measure of the conformational flexibility of the residue across the ensemble. By comparing the mean square fluctuation of the ensembles with one or two cAMP bound to the ensemble of Apo-CAP we can see how the binding of cAMP affects the conformational flexibility of the protein. Figure 3.10 shows this visually.

On binding cAMP to chain A, the surrounding regions of chain A become more rigid. This is to be expected on ligand binding. However, significant regions of chain B have the same flexibility (grey regions in Figure 3.10) or are more flexible (red regions) on ligand binding to chain A. The corresponding effect happens on a single cAMP binding to chain B. However on cAMP binding to both chains, both binding sites become significantly rigid and nearly all regions of the protein are more constrained than in Apo-CAP. The ratio of mean square fluctuations as seen in Figure 3.10 follow the order parameter data and amide exchange rates, which are a measure of flexibility in the protein, from a previous study [89]. The explanation for the negative cooperativity given in the existing study is that the binding of the second cAMP significantly quenches motions in the protein - this has an associated entropic cost that leads to negative cooperativity between the cAMP sites. The data from ExProSE support this conclusion.

The structural changes on cAMP binding were also measured using ExProSE. The average structures across the ensembles of Apo-CAP, and CAP with cAMP

bound to chain A, were compared. The RMSD of chain A (B) between the averages of the ensembles was 0.16 Å (0.08 Å). This indicates minor structural rearrangement in chain A due to ligand binding, but almost no change in chain B. This agrees with chemical shift mapping in the existing study [89]. These results indicate that ExProSE is able to reproduce dynamic allostery in a model system. The same process could be used to explore positive or negative cooperativity in other proteins. The results also demonstrate that ExProSE can give useful information with only a single structure used as input.

### 3.2.5 Activation by phosphorylation

The nitrogen regulatory protein C (NtrC) is a signaling protein. Phosphorylation of the receiver domain causes a large structural change [177] that leads to oligomerisation and promotion of transcription. ExProSE was used to explore this system with the distance constraints coming from the inactive and active states (PDB ID 1DC7 and 1DC8 respectively), and extra constraints at the phosphorylation site in the active state - see Section 3.1. Figure 3.11 shows the mean square fluctuation across the ensemble in the unperturbed ensemble (A, shown on the inactive state) and the ensemble with perturbation at the phosphorylation site (B, shown on the active state). Phosphorylation leads to considerably less variation across the ensemble, and the regions which are effected are the regions which show structural differences between the active and inactive state.

A comprehensive NMR study has been carried out on NtrC [178] and ExProSE produced results consistent with those findings. In that study the exchange term $R_{ex}$ showed that inactive NtrC had considerable motions on the microsecond to millisecond timescale - particularly around the $\alpha$3 and $\alpha$4 helices, the surrounding loops and the $\beta$5 sheet. These motions are seen in the ensemble produced by ExProSE (see Figure 3.11). On phosphorylation NMR indicates that most of these dynamics disappear, with only some motion in the $\beta$5 sheet and loop regions. This is largely recreated by ExProSE, with the remaining flexible region close to that in the NMR data. This supports the conformational selection hypothesis outlined in [178]. It also indicates that conforma-

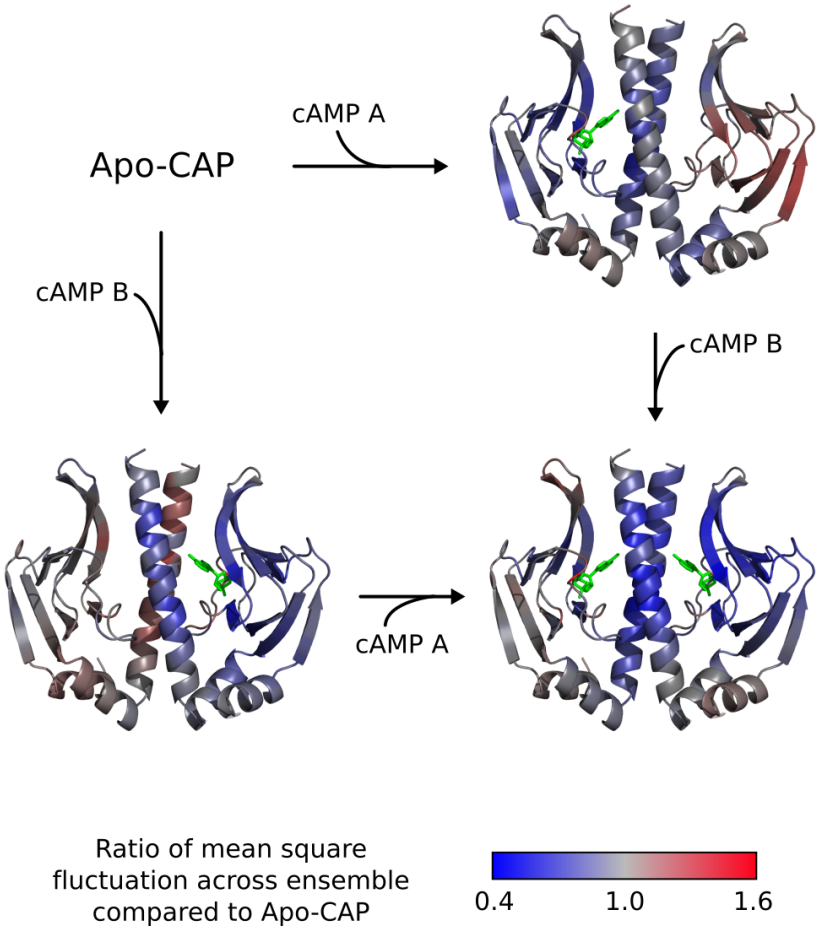



**Figure 3.10** Mean square fluctuations across CAP ensembles compared to Apo-CAP. The four ensembles are generated separately. Apo-CAP has no cAMP. The other ensembles have additional constraints (see Section 3.1) representing cAMP bound to chain A, cAMP bound to chain B, or cAMP bound to both chains A and B. The bound cAMP molecules are shown for reference as green sticks. Red regions indicate residues with more flexibility compared to Apo-CAP, and blue regions indicate residues with less flexibility compared to Apo-CAP. Figure based on Figure 5 from Greener et al. 2017 [27].

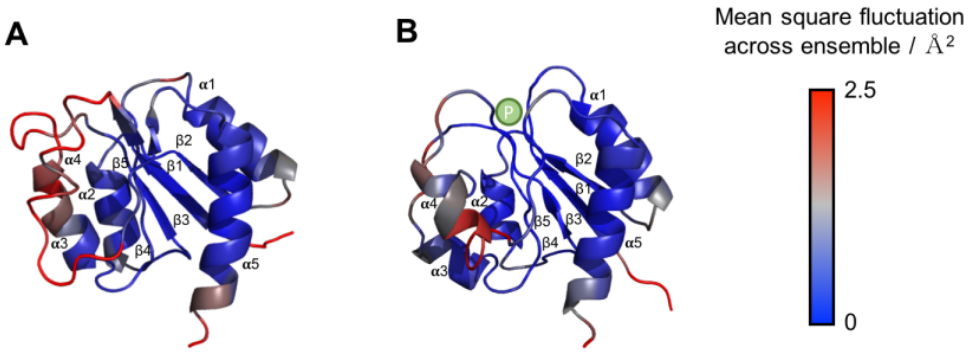



**Figure 3.11** Ensembles of NtrC. (A) The mean square fluctuation in the unperturbed ensemble is shown on the inactive structure (PDB ID 1DC7). (B) The mean square fluctuation in the ensemble with additional constraints representing phosphorylation is shown in the active structure (PDB ID 1DC8). Labels are taken from [178].

tional variability as displayed in ExProSE is linked to the micro/millisecond timescale dynamics of the system. This indicates why ExProSE is effective as an allosteric prediction tool as the motions relevant to allostery are often on these timescales, which are usually inaccessible to MD studies. These results also show that ExProSE can be used to study covalent modifications (here, phosphorylation) as well as the small molecule non-covalent interactions explored previously.

## 3.3 Discussion

The allosteric prediction methods PARS, STRESS and AlloPred all use NMA to predict allosteric sites. NMA is computationally-inexpensive and hence suitable for high-throughput, automated approaches. However the assumption of harmonic fluctuations around an energetically-minimum structure often makes prediction of conformational changes difficult, particularly for transitions with a low degree of collectivity [179]. In addition, the choice of which normal modes to use is non-trivial. Larger conformational changes are associated with low-frequency normal modes but higher-frequency modes are also required to take into account local effects.

The focus of NMA on changes in dynamics is also important - whilst NMA-based methods might be expected to reveal perturbations to vibrations in proteins there are a variety of other motions that contribute to allostery, such as local unfolding and rigid body movements [3]. By contrast, ExProSE generates native-like protein structures that can span large conformational changes. The structure generation process is then perturbed to predict allosteric sites. This has the potential to discover effects not revealed by NMA-based methods, whilst retaining the low computational cost and ease of use. It also provides an ensemble of structures under the influence of the predicted modulator that can be used, for example, in flexible ligand docking. Energy minimisation provides a way to improve the stereochemistry of generated structures for use in such approaches.

ExProSE requires two structures for each protein, whereas other methods only require one. It also requires the structures to be different from each other in order to generate structures that span the conformational space. This makes the method unsuitable for use on proteins where only one structure, or highly similar structures, are available. However, many medically-important proteins have multiple structures available, including the examples used in this study. In these cases, it makes sense to use the additional structural information. There are some methods which can predict holo structures from an apo input structure or vice versa [180, 181, 161] and these could be used to generate

apo/holo pairs. The pE-DB database of structural ensembles of IDPs could also provide a source of structural data [182]. The method also was successful at reproducing the allostery in CAP using only one structure as input. For proteins with multiple different conformational states, more than two structures could be used as input to ExProSE to explore further regions of conformational space - the constraint combination procedure can be applied to an arbitrary number of structures.

For many ensemble generation methods, such as MD and tCONCOORD, the choice of parameters has a large effect on the structures produced. The parameter in ExProSE with the largest effect is $W_B$ (see Section 3.1), which affects the conformational spread of the ensemble. Without any user input, the auto-parameterisation step of ExProSE selects a value that gives an ensemble well-spread over the conformational space between the two input structures. Once $W_B$ has been selected automatically, an ensemble that spans the correct space is generally produced without any further choice of parameters. This makes the method suitable for high-throughput structure generation across multiple proteins as the user does not need to make any parameter choices themselves. The auto-parameterisation procedure can be adjusted to obtain the desired level of structural flexibility using the parameter $F$, which is intuitive in terms of the spread of structures over conformational space. This provides a way to generate an ensemble with more flexibility if the input structures are similar, as mentioned above.

In this study, LIGSITE$^{cs}$ was used to predict pockets for ExProSE. However, it is worth noting that any pocket prediction method that outputs pocket points is compatible with ExProSE without modification. One of the challenges in allosteric site prediction is discovery of cryptic or transient pockets - pockets that are only present in some structures of the ensemble. These pockets are an appealing target as they have thus far been unavailable to pocket prediction methods on crystal structures - see Section 1.2. There are currently no general methods that use transient pockets for allosteric site prediction [73], though recent studies have used Markov state models on MD simulations to predict cryptic allosteric sites on multiple proteins [183, 75]. These studies concluded

that cryptic allosteric sites are more ubiquitous than previously thought. For example, transient sites were found on TEM-1 β-lactamase, which has been studied extensively without observing these sites [75]. ExProSE has the potential to identify transient pockets and predict their ability as allosteric sites. For example, an ensemble could be generated from apo/holo pairs and clustered into a few representative structures. LIGSITE$^{cs}$ could then be run on these structures to find potential transient pockets, and the perturbation procedure carried out to assess their allosteric character. In initial tests of this approach, the known transient allosteric site on TEM-1 β-lactamase opened up in some representative structures from an ExProSE ensemble. Methods for predicting pockets on an ensemble of protein structures, such as MDpocket [184] or the method in Polyphony [175], could also be run on ensembles generated using ExProSE.

The results from ExProSE have implications for the conformational selection/induced fit debate [73]. The principle of ExProSE is based on the conformational selection paradigm - an existing ensemble is perturbed by an effector to change the populations, as shown in Figure 1.1. This suggests that cases for which ExProSE is successful adopt a conformational selection mechanism at least in part. The conformational selection present in NtrC was reproduced by the method, for example. In general this focus on conformational selection can be seen as a positive because small molecules binding with weaker interactions has been associated with a preference for conformational selection over induced fit [185], and it is these type of interactions that are important for allosteric modulator discovery. It was found in some cases, however, that the additional constraints due to a predicted modulator caused the system to sample structures not present in the unperturbed ensemble. This can be viewed as an induced fit effect, with the modulator binding allowing the ensemble to access structures which are not available in the absence of the modulator. Study of perturbed ensembles using ExProSE could therefore give evidence for conformational selection or induced fit mechanisms in specific proteins.

ExProSE has the potential to be further improved. Other numerical methods for satisfying the distance constraints could be explored, as there are many

available [186]. The interactions used to generate distance constraints could be further refined; this was carried out for CONCOORD and gave slightly improved results [110]. Given the ease of automation of the whole pipeline, ExProSE has the potential to be turned into a web server. Users could generate an ensemble of protein structures from one or two input structures, or predict allosteric sites from two input structures. The ability of the method to generate structures for protein-protein docking or for flexible ligand docking could be explicitly investigated. For example, structures generated from unbound protein interacting partners could be generated and docked to see if better models for complexes are produced. This work is currently ongoing in the lab.

ExProSE builds on existing methods by using more structural information as input. It is able to generate ensembles of protein structures that span relevant conformational changes in proteins. This makes it an effective alternative to similar methods and to MD, which it is often not feasible to run on timescales long enough to explore large motions of interest without specialist approaches. The perturbation procedure can be applied systematically to predict allosteric sites. In a comparison of multiple allosteric site predictors, ExProSE showed performance similar to and complementary with existing methods. Experimental results in the well-studied CAP were also reproduced by ExProSE. The ability to generate ensembles of protein structures and investigate the response of an ensemble to perturbations should prove useful for both the exploration of individual proteins and the systematic study of the whole PDB. Such methods are required to make sense of the increasing volume of structural data, and to understand the crucial importance of dynamics to protein function.

# Chapter 4

# CDK2

This chapter describes computational and experimental work to test a potential allosteric site on cyclin-dependent kinase 2 (CDK2), a protein important in cell cycle regulation. This allosteric site was predicted by ExProSE in Section 3.2. A virtual screen of small molecules was carried out against the pocket. Selected molecules were purchased and tested using two experimental assays to see if they could inhibit the interaction between CDK2 and cyclin A2.

## 4.1 Materials and Methods

### 4.1.1 Bioinformatics resources

FTMap [54], the Fragment Hotspot Map [187], ResiCon [188] and the DynOmics server [26] were run with default settings using the ANS-bound CDK2 structure (PDB ID 3PXF) as input.

### 4.1.2 Virtual screening

AutoDock Vina [189] and DOCK [190] were run in line with the respective documentation. ChemMine [191] clustering of structures used binning clustering at a similarity cutoff of 0.5. The highest ranked by docking scores was retained in each case.

The ExProSE structures selected with an open pocket have the maximum sum of two distances across the opening of the pocket: the distance between atom OH on residue TYR180 and atom NH1 on ARG150, and the distance between atom OG on residue SER188 and atom O on THR165. These structures are energy minimised as described in Section 3.1 to reduce any violations of stereochemistry.

### 4.1.3 Reagents and compounds

The buffers and media used for experimental work were as follows:

- Auto-induction media: 10 g/L tryptone, 5 g/L yeast extract, 5052, NPS, 1 M $MgSO_4$, Ampicillin.
- Lysogeny broth: 10 g/L tryptone, 5 g/L yeast extract, 10 g/L NaCl.
- Lysis buffer: 50 mM HEPES pH 7.0, 150 mM NaCl, 10 mM $MgCl_2$, 2 mM DTT, 1 mM EGTA, Triton.
- Sodium phosphate pH 8 buffer: 100 mM sodium phosphate pH 8.0.

- Tris pH 8 buffer: 100 mM Tris-HCl pH 8.0, 150 mM NaCl, 10 mM $MgCl_2$.
- TBS-T buffer: 50 mM Tris-HCl pH 7.5, 150 mM NaCl, 0.1% Tween 20.
- Transfer buffer: 20% methanol, 190 mM glycine, 25 mM Tris.

### 4.1.4 Purification of cyclin A2

The plasmid for cyclin A2 with either a His tag or GST tag (source - David J Mann unpublished) was transformed into *E. coli* Tuner cells. Various conditions for growth were tried in auto-induction media and lysogeny broth, similar to the approach in [192]. The final conditions used for His-tagged cyclin A2 were growth of cultures at 37°C until optical density in auto-induction media, then incubation at 18°C for 40 hours. Cells were harvested by centrifugation (10 min at 5,000 rpm). Harvested cells were resuspended in lysis buffer with 0.5 mg $ml^{-1}$ lysozyme. After sonication and centrifugation (30 min at 16,000 rpm) the supernatant was incubated on Ni Sepharose Fast Flow beads in lysis buffer for 1 hour. After washing the beads, cyclin A2 was eluted with 500 mM imidazole in lysis buffer, dialysed into lysis buffer and stored at -20°C in 50% glycerol.

### 4.1.5 TR-FRET assay

To label CDK2 with Cy5 dye, CDK2 (source - Gregory Craven unpublished, 10 mg/ml) in sodium phosphate pH 8 buffer was incubated overnight with Cy5 NHS ester dye in anhydrous DMF. Excess dye was removed by gel filtration in sodium phosphate pH 8 buffer with the fractions containing CDK2 identified by fluorescent SDS-PAGE.

For the time-resolved Förster resonance energy transfer (TR-FRET) CDK2 titration each well contained 20 μL volume of 1 nM cyclin A2-His, 1 nM Eu-anti-His antibody and CDK2 at various concentrations in 50% glycerol/50% sodium phosphate pH 8 buffer. A control with no CDK2 and 10 CDK2 concentrations were used - 1 μM, 300 nM, 100 nM, 30 nM, 10 nM, 3 nM, 1 nM, 300 pM, 100

pM and 50 pM. The plate was incubated for 1 hour and centrifuged at 800 rpm before being read. 20 recordings were made by the plate reader and the values averaged.

### 4.1.6 Binding assay

His-tagged cyclin A2 (75 nM) was incubated on Ni Sepharose Fast Flow beads in Tris pH 8 buffer with 12% DMSO. CDK2 (source - Gregory Craven unpublished) was added (8 nM) along with compounds in 5 concentrations: 0 M control, 8 μM, 40 μM, 200 μM and 1 mM. After incubation for 1 hour the beads were washed with Tris pH 8 buffer and heated in gel-loading dye at 100°C for 10 minutes to release bound protein. The proteins in the supernatant were separated by SDS-PAGE and transferred to a nitrocellulose membrane by electroblotting in transfer buffer for 1 hour at 100 V. The membrane was incubated for 15 minutes in TBS-T buffer with 1.5% skimmed milk powder. The primary antibody, CDK2 rabbit polyclonal antibody was added and the membrane incubated overnight at 4°C. After washing with TBS-T buffer the secondary antibody was added and incubation for 1 hour carried out. The membrane was washed with TBS-T buffer and imaged with chemiluminescence by adding ECL dye and using a 1 s exposure.

## 4.2 Results

Having predicted a new allosteric site on CDK2 using ExProSE, we wished to explore this pocket further computationally and test experimentally whether it was in fact an allosteric site. Section 3.2 and Figure 3.9 describe how ExProSE predicts the third predicted pocket, henceforth referred to as the pocket of interest, as being allosteric. See Figure 1.5 for the structural elements of CDK2. These pockets are also shown in Figure 4.1A. The pocket of interest is close to the region of cyclin binding, as shown in Figure 4.1B with the non-crystallised portion of cyclin A2 modelled with Phyre2 [162].

This pocket is not open in the apo CDK2 structure (PDB ID 1HCL) or in the cyclin A2-bound structure (PDB ID 1FIN). In the cyclin A2-bound structure there is in fact a protrusion instead of a pocket opening, suggesting that a small molecule would not be able to bind the pocket at the same time as cyclin A2 is bound to CDK2. The pocket is shown in Figure 4.1C in various structures where it is open. The residues that make up the pocket are shown in Figure 4.1D. The pocket is open but small in the structure with two ANS molecules bound in the known allosteric site (PDB ID 3PXF), with a size of 45 $Å^3$ predicted by LIGSITE$^{cs}$. It is a similar shape in the structure with two ANS molecules and the ATP-binding site inhibitor staurosporine (PDB ID 4EZ7). This structure also has the crystallisation artefact ethanediol crystallised at the pocket of interest, implying that binding there is possible.

A structure with ethanediol bound at the ANS site and an ATP-binding site inhibitor (PDB ID 4EZ3) appears to have the pocket slightly open. This indicates that occupation of the ANS site and opening of the pocket of interest are linked, suggesting that if the pocket can be bound then the inactivating motions of the αC-helix will prevent cyclin binding and inactivate the protein. The idea is that by stabilising the pocket of interest the inactive state is favoured and activation by cyclin A2 cannot occur. The hope is that the pocket of interest can accommodate a small molecule ligand, and potentially that it can open further to present more binding interactions.

One risk of targeting this site is that the pocket is small, meaning a limited num-

ber of interactions can be made with a ligand. The pocket is also only available in some states, meaning that there may be an energetic cost to stabilising it [77]. Off site binding may also be a problem - the ATP-binding site and ANS allosteric site present large pockets on CDK2 known to bind small molecules, and a molecule may bind there rather than at the pocket of interest.

CDK2 was examined with available bioinformatics resources. FTMap [54] and the Fragment Hotspot Map [187] sample the protein surface with molecular probes to find fragment binding hotspots. FTMap [54] does not predict the pocket of interest as a binding site. The Fragment Hotspot Map [187] does not reveal the pocket of interest as a fragment binding site in the default map. However adjusting the score cutoff does show there is some ability for fragment to bind to the pocket.

ResiCon calculates dynamic regions in proteins by considering contacts between residues [188]. ResiCon was run on CDK2. The pocket of interest seems to be on the edge of dynamic domains across a variety of given numbers of domain outputs, including the default result of two domains. This indicates that the site could have the ability to cause structural rearrangement in distal parts of the protein by acting as a hinge that responds to ligand binding. This is further evidenced by examining CDK2 using the DynOmics server [26]. Residue that are close to the pocket of interest, such as ASP127 in the pocket, are predicted to be hinge residues that control the two slowest normal modes.

PDBFlex finds flexible regions in proteins based on all PDB depositions of the protein [193]. For CDK2 it indicates that the only real areas of variation in the structure is around the αC-helix and the T-loop, which is near the pocket of interest.

### 4.2.1 Virtual screening

A virtual screening approach was carried out to predict compounds that bind to the pocket of interest. ZINC is a free database of commercially-available compounds for virtual screening [194]. The ZINC12 LeadsNow subset con-

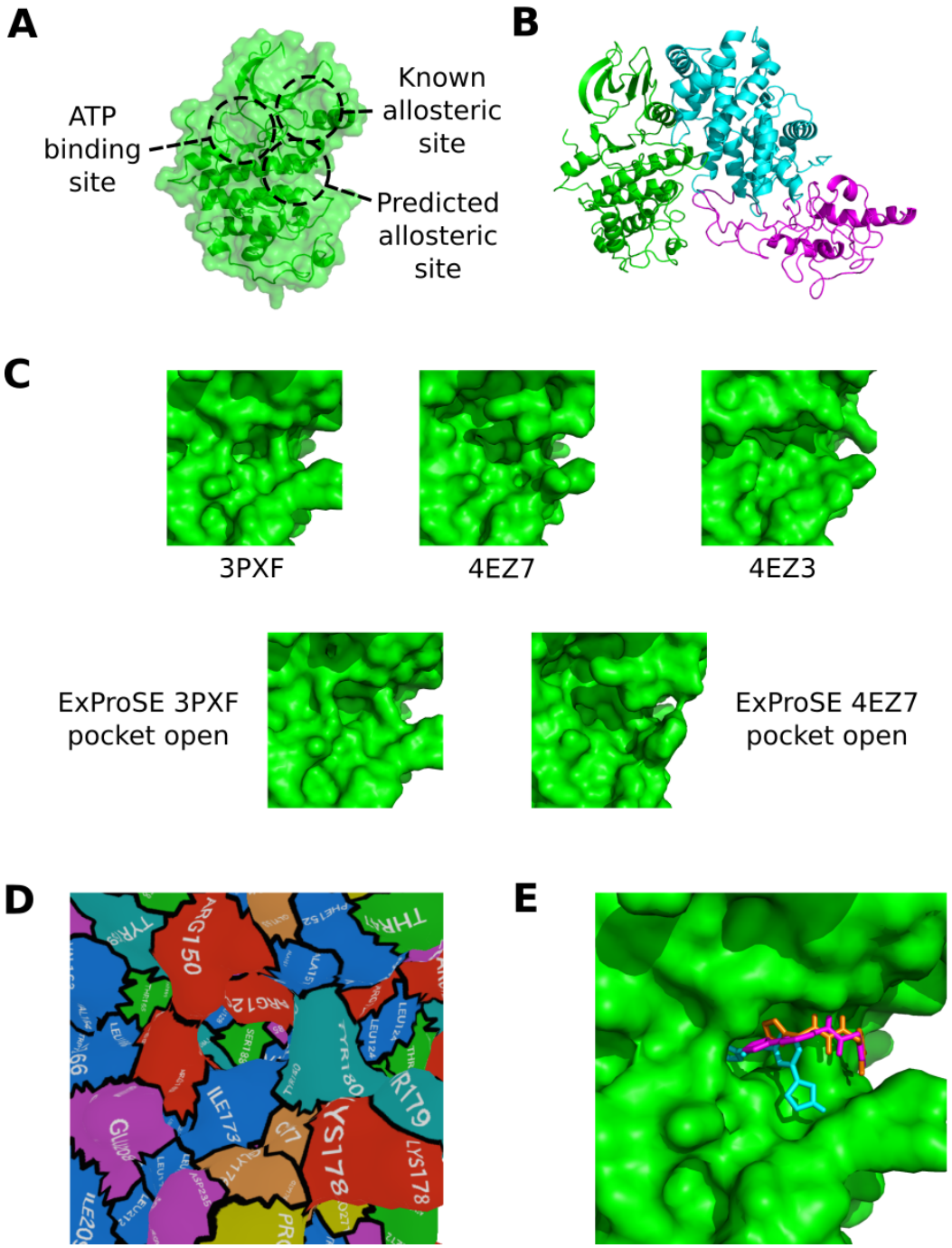



**Figure 4.1** Caption on following page.

**Figure 4.1** Structure, conformational variability and virtual screening of CDK2. See also Figure 1.5. (A) The structure of CDK2 (PDB ID 1HCL) with the ATP-binding site, the known allosteric site (ANS site) and the potential allosteric site (the pocket of interest) marked. (B) The CDK2-cyclin A2 complex required for CDK2 activation. The crystallised complex (PDB ID 1FIN) is shown with CDK2 in green and the crystallised portion of cyclin A2 in cyan. The whole cyclin A2 sequence was modelled using Phyre2 [162] and aligned to the above complex and the ab initio modelled region (with no structural templates) is shown in magenta. Part of this region potentially interacts with CDK2. (C) The pocket of interest on CDK2 shown in various structures. This is pocket 3 from Figure 3.9. The structures shown are three crystal structures with the PDB IDs given, and two structures with the pocket open generated using ExProSE from the structures indicated. (D) Residues displayed around the pocket of interest using SurfStamp (`https://yamule.github.io/SurfStamp-public`). The orientation is the same as in C and the ANS-bound structure (PDB ID 3PXF) is used. (E) Example docking poses of screened compounds using AutoDock Vina on the structure with PDB ID 3PXF. Shown are the lowest energy poses for compound A (magenta), B (cyan) and C (orange) from Table 4.1.

tains over 3 million lead-like molecules in stock at chosen suppliers. Lead-like molecules have a molecular weight between 250 and 350, an octanol-water partition coefficient not greater than 3.5 and no more than 7 rotatable bonds. These criteria give smaller and less lipophilic molecules than those that would conventionally end up being drugs, i.e. those that fit Lipinski's rule of five [98]. This is because at the hit stage the priority is finding effective scaffolds with high affinity per atom (ligand efficiency). Elaboration of the structure with further functional groups generally occurs later and would usually result in the addition of chemical groups, raising the molecular weight.

The LeadsNow subset is clustered as described at `http://zinc.docking.org` to get a representative subset of ∼250,000 compounds. Each molecule has a Tanimoto coefficient, a similarity score that compares chemical fingerprints, of no more than 0.9 with any other molecule. These compounds are docked using AutoDock Vina [189] with standard settings to the pocket of interest in the ANS-bound structure (PDB ID 3PXF). Examples of docking poses can be seen in Figure 4.1E. The top 2,000 structures by affinity score from this screen were retained for further docking studies. This was due to the computational cost of

virtual screening - the score from AutoDock Vina on the ANS-bound structure was used to eliminate most compounds and the remaining compounds were taken forward for further studies. The 2,000 retained molecules were docked using AutoDock Vina and DOCK [190] onto four structures:

1. The ANS-bound structure, PDB ID 3PXF.
2. The ANS-bound structure with the ATP-binding site inhibitor staurosporine, PDB ID 4EZ7.
3. A structure selected from an ensemble of structures generated from (1) using ExProSE. The structure selected is the one with the pocket of interest most open.
4. The same as (3) but the ensemble is generated using ExProSE from (2).

The structure of the pocket of interest in each of these structures is shown in Figure 4.1C. Docking to multiple structures means the compounds are scored in multiple conformations of the binding site, which is an approximation of ensemble docking. This is important in general to take into account the flexibility of binding pockets, and especially important for this pocket as it is a flexible pocket not present in all structures. Using two different docking algorithms, AutoDock Vina and DOCK, provides two different scores for each structure and goes some way to reducing the inaccuracies of virtual screening. Compounds were in general better scored for structures (3) and (4), those with the pocket of interest open. This is to be expected as a larger pocket presents more opportunities for interactions with the ligand.

### 4.2.2 Compound selection

Compounds were ranked by the average of the 8 scores from the above docking (4 structures, 2 docking methods). In addition to the lead-like properties described previously, compounds were further filtered to remove compounds that would potentially give erroneous assay results. Compounds violating various criteria set out by the SwissADME server for assessing medicinal chem-

istry friendliness [195], such as properties of pan-assay interference compounds [196], were removed. Compounds marked as not having benign chemical functionality in ZINC12 were also removed. Although compounds were clustered by ZINC12 to remove similar compounds, some compounds were relatively similar to each other on visual inspection. Compounds were hence removed by similarity in ChemMine [191]. The top 20 ranked compounds remaining that were available from Enamine (`http://www.enamine.net`) were purchased and taken forward for experimental testing. Docking scores and compound IDs are shown in Table 4.1. The structures of the compounds are shown in Figure 4.2. A search of these structures in the PDB indicates that none are currently ligands in the PDB. None appear in the ChEMBL database of bioactive drug-like small molecules [197].

### 4.2.3 Experimental aims

The main aim of the experimental work was to screen the purchased compounds using the TR-FRET assay and a binding assay to see if the compounds could inhibit the CDK2-cyclin A2 interaction. This required cyclin A2 to be purified.

### 4.2.4 Purification of cyclin A2

For optimal TR-FRET signal (see below) it is be beneficial to have a single surface-exposed cysteine labelled with Cy5 dye. Cyclin A2 has two surface-exposed cysteine residues. Hence, initial purification was attempted for cyclin A2 residues 169-432 with GST tag and the C327A mutation. The GST tag allowed selective separation of the protein during purification. The His tag could not be used as this was going to be present on CDK2 to bind the donor fluorophore. This purification resulted in low quantities of soluble protein ($< 1$ mg from 4 L media) and purity was low, as shown in Figure 4.3A. Different purification strategies including varying the media, incubation temperature and incubation times were attempted but did not improve yield.

| Compound ID | ZINC12 ID | Enamine ID | Molecular weight (g/mol) | AutoDock Vina best energy / kcal mol$^{-1}$ | | | | DOCK best grid score | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | A | B | C | D | A | B | C | D |
| A | ZINC06731189 | Z28083007 | 279.3 | -7.1 | -8.1 | -7.6 | -7.4 | -37.3 | -40.0 | -40.8 | -35.1 |
| B | ZINC29799246 | Z2241108787 | 281.3 | -7.4 | -7.9 | -7.9 | -7.7 | -31.2 | -38.9 | -39.7 | -32.4 |
| C | ZINC58182552 | Z953947716 | 273.3 | -7.2 | -7.4 | -7.8 | -7.4 | -32.8 | -39.3 | -41.5 | -36.0 |
| D | ZINC03275010 | Z56813876 | 280.3 | -7.3 | -7.3 | -7.3 | -7.7 | -36.5 | -39.2 | -40.6 | -36.3 |
| E | ZINC25129280 | Z125831222 | 281.3 | -7.3 | -7.8 | -7.1 | -7.5 | -32.8 | -41.2 | -39.9 | -38.8 |
| F | ZINC97022380 | Z1537396696 | 280.3 | -7.7 | -7.6 | -7.4 | -7.8 | -35.0 | -37.0 | -38.2 | -32.6 |
| G | ZINC84057181 | Z1367181624 | 280.3 | -7.2 | -7.8 | -7.9 | -7.4 | -30.9 | -37.8 | -35.5 | -37.1 |
| H | ZINC30691564 | Z383528790 | 281.3 | -7.2 | -7.2 | -7.8 | -7.8 | -35.3 | -35.3 | -41.6 | -33.8 |
| I | ZINC36390489 | Z381531134 | 279.3 | -7.6 | -7.3 | -7.5 | -7.5 | -31.2 | -39.3 | -37.5 | -34.7 |
| J | ZINC89878745 | Z1159552572 | 279.3 | -7.0 | -8.0 | -8.3 | -7.2 | -37.3 | -38.3 | -38.9 | -32.5 |
| K | ZINC12812500 | Z220404550 | 279.3 | -7.0 | -7.9 | -7.5 | -7.7 | -31.8 | -38.1 | -39.0 | -34.6 |
| L | ZINC75147268 | Z1262428103 | 281.3 | -7.3 | -7.8 | -7.8 | -7.5 | -28.8 | -36.0 | -37.6 | -34.4 |
| M | ZINC71914433 | Z1232176487 | 274.3 | -7.0 | -7.4 | -7.4 | -7.5 | -36.2 | -37.6 | -40.2 | -35.7 |
| N | ZINC72288573 | Z1229931451 | 268.3 | -7.7 | -7.8 | -7.9 | -7.4 | -29.4 | -36.2 | -33.3 | -34.1 |
| O | ZINC69453509 | Z1030096350 | 287.4 | -7.2 | -7.6 | -7.1 | -7.7 | -33.9 | -40.5 | -33.1 | -35.8 |
| P | ZINC79097391 | Z1408168262 | 281.3 | -7.0 | -7.6 | -7.6 | -7.3 | -33.7 | -38.5 | -37.5 | -33.8 |
| Q | ZINC16497227 | Z66926805 | 280.3 | -6.9 | -7.4 | -7.8 | -7.9 | -29.2 | -38.8 | -40.0 | -33.5 |
| R | ZINC23143433 | Z352550190 | 271.3 | -7.5 | -7.3 | -6.9 | -7.1 | -32.5 | -39.6 | -38.4 | -36.2 |
| S | ZINC89858262 | Z1162482939 | 273.3 | -7.2 | -7.3 | -8.3 | -7.4 | -23.9 | -35.1 | -39.6 | -36.2 |
| T | ZINC69369685 | Z1097406485 | 272.3 | -7.2 | -7.1 | -7.0 | -7.1 | -37.0 | -40.6 | -43.1 | -35.6 |
| | | **Mean values** | 278.3 | -7.3 | -7.6 | -7.6 | -7.5 | -32.8 | -38.4 | -38.8 | -35.0 |

**Table 4.1** Selected compounds to screen experimentally against a potential allosteric site on CDK2. ZINC12 ID is the ID in the ZINC12 database (`http://zinc.docking.org`). Enamine ID is the ID at Enamine Ltd (`http://www.enamine.net`). The AutoDock Vina best energy and DOCK best grid score are shown for each ligand docked to four structures: (A) PDB ID 3PXF, (B) PDB ID 4EZ7, (C) ExProSE pocket open structure from 3PXF, (D) ExProSE pocket open structure from 4EZ7. See the main text for more information on these structures. The mean values across the 20 compounds are also shown.

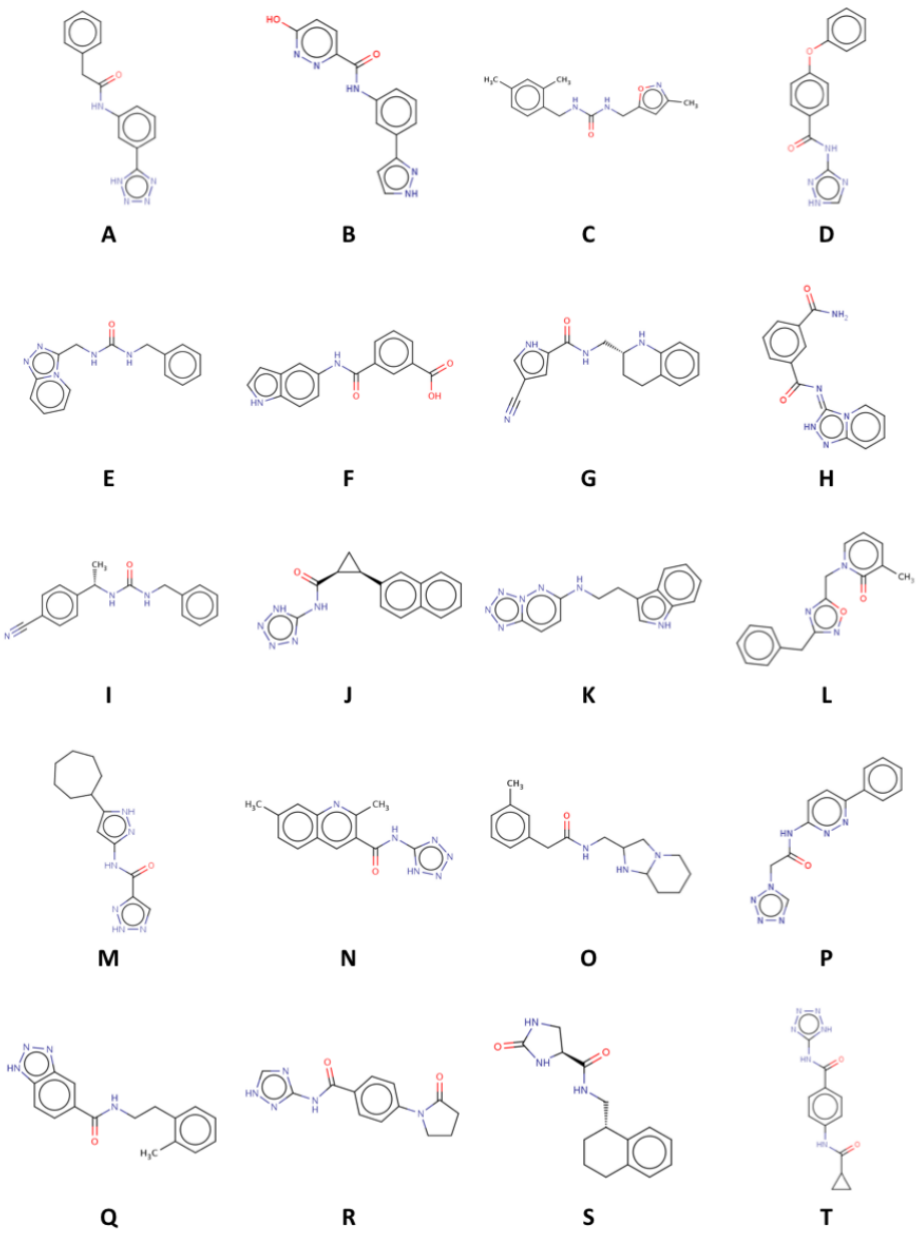



**Figure 4.2** Chemical structures of selected compounds to screen experimentally against a potential allosteric site on CDK2. The compounds are labelled as in Table 4.1. Images generated using `http://cdb.ics.uci.edu/cgibin/Smi2DepictWeb.py`.

The difficulty of purifying this cyclin A2 mutant led us to try purification of cyclin A2 residues 169-432 with GST tag and no mutation. This purified more readily than the mutant, as shown in Figure 4.3B. However the yield was still low (1 mg from 8 L media). This could be because the cysteine mutation destabilises the protein, increasing its tendency to aggregate. Concentration of the protein using spin filtration led to loss of the protein. This, along with the low yield, suggests that cyclin A2-GST is unstable and has a tendency to aggregate and become insoluble [198].

Due to the difficulty of purifying cyclin A2-GST, a change was made to the experimental strategy. The donor and acceptor fluorophores for the TR-FRET assay were switched so that the Eu anti-His antibody was targeted at cyclin A2 and CDK2 was conjugated with Cy5 dye - see later. This required purification of cyclin A2 residues 169-432 with His tag. This purification proved considerably more successful than cyclin A2-GST, with 10 mg produced from 4 L media - see Figure 4.3C. This indicates that the large GST tag may disrupt the stability or folding of cyclin A2, but the shorter His tag does not have the same effect. Purified cyclin A2-His was taken forward for the TR-FRET assay.

### 4.2.5 TR-FRET assay

This popular assay for drug discovery research is the combination of time-resolved fluorometry with Förster resonance energy transfer (FRET) [199]. Figure 4.4A and 4.4B outline the principles of a TR-FRET assay. FRET involves two fluorophores, a donor and an acceptor. A time delay between excitation and detection means measurement occurs after the timescale of background fluorescence. The long emission time of the donor fluorophore means a signal is obtained after the time delay.

CDK2 prepared previously in the lab was labelled with Cy5 dye. The TR-FRET assay was tested using a titration of CDK2 with constant cyclin A2 and donor fluorophore. An increase in signal with increasing CDK2 concentration would be expected. Figure 4.4C shows the result of this titration. Whilst there is a higher signal at higher CDK2 concentrations, this trend is also present for the

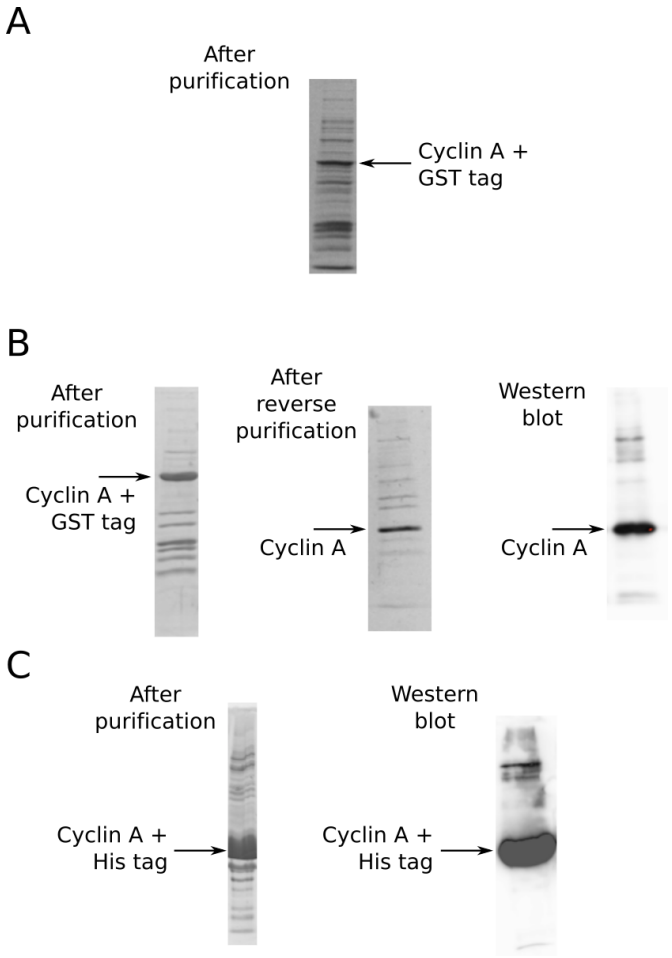



**Figure 4.3** SDS-PAGE results for purification of cyclin A2. (A) Purification of cyclin A2 residues 169-432 with GST tag and the C327A mutation. (B) Purification of cyclin A2 residues 169-432 with GST tag. (C) Purification of cyclin A2 residues 169-432 with His tag.

control of cyclin A2-GST. Cyclin A2-GST lacks the His tag required to bind to the donor fluorophore so should not lead to a TR-FRET signal. This indicates that the increased signal with CDK2 concentration is likely due to excess, unbound Cy5 dye in the CDK2 solution. Time constraints meant that a further gel filtration could not be carried out to remove this unbound dye. The high background signal means the screen would not be effective at finding compounds that inhibit the cyclin A2-CDK2 interaction, so the compounds were not put through the TR-FRET screen.

### 4.2.6 Binding assay

A binding assay was carried out to test whether the compounds could inhibit the CDK2-cyclin A2 interaction. Cyclin A2-His and CDK2 were incubated with beads that selectively bind the His tag. After washing the beads to remove unbound protein only cyclin A2 and proteins bound to it should remain. In the absence of a binding inhibitor a signal for CDK2 would be expected in immunoblotting as CDK2 binds to cyclin A2. This signal would be expected to disappear in the presence of a modulator that prevented the interaction. This is only a semi-quantitative assay at best and TR-FRET would be much more informative. However, it was carried out due to time constraints and the difficulties with obtaining results from TR-FRET.

The binding assay results can be seen in Figure 4.5 for the 12 compounds it was carried out on. A problem with the assay was binding of CDK2 to the beads despite the lack of a His tag on CDK2. This gave a substantial background signal making a signal from the compounds hard to distinguish from the background. However, some compounds do appear to remove the CDK2 signal at high concentrations. Compound D at 200 μM and 1 mM, and compound J at 1 mM, cause the CDK2 signal to decrease. At these concentrations Compounds D and J had solubility issues in the assay. As the expected background is high the solubility issues are likely the cause of the observed signal rather than genuine binding inhibition.

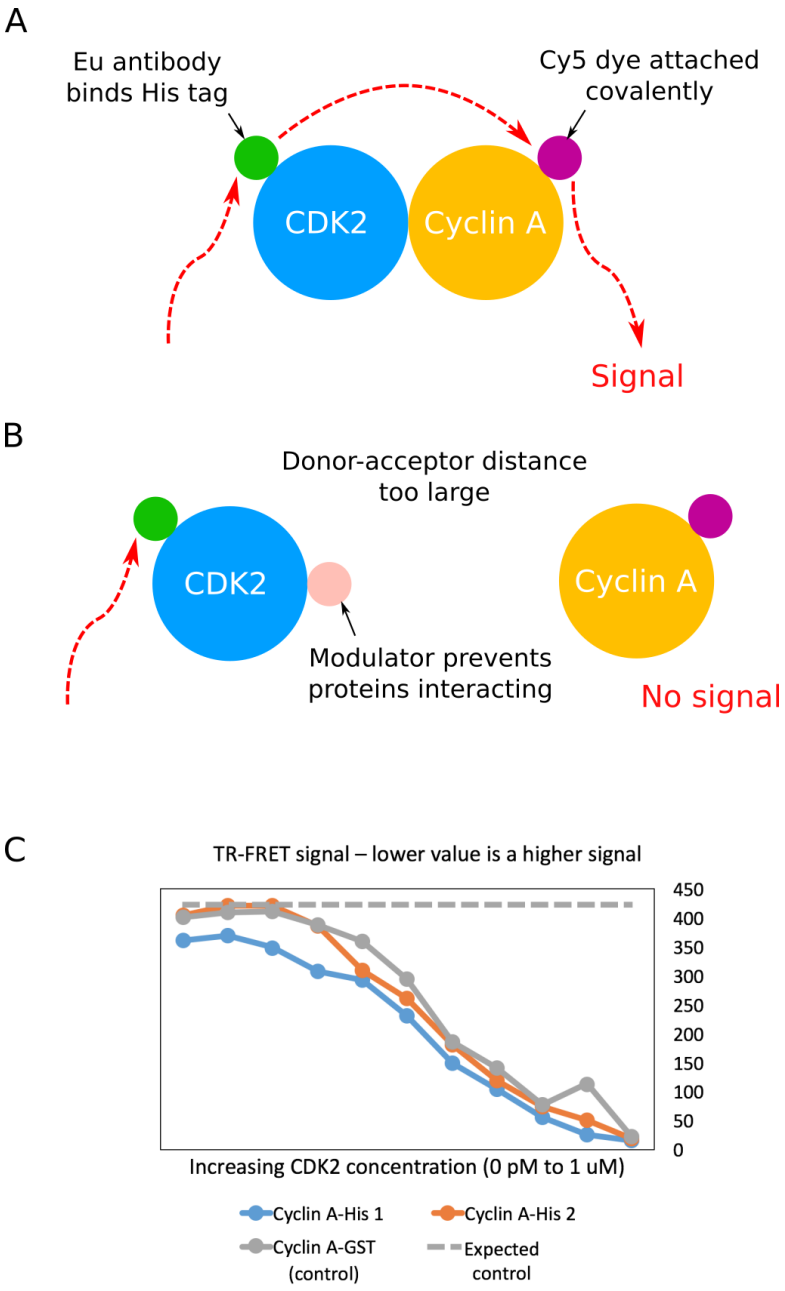



**Figure 4.4** Caption on following page.

**Figure 4.4** (A) The principle of the TR-FRET assay used. One protein in the binding pair is covalently-bound to Cy5 dye (the acceptor fluorophore). The other protein is targeted with an antibody containing the lanthanide Eu (the donor fluorophore). Light is shone at the excitation wavelength of the donor fluorophore. This emits at the excitation wavelength of the acceptor fluorophore. After a delay to allow background emission to recede, emission from the acceptor fluorophore is measured. (B) If a modulator prevents the proteins interacting, emission from the donor to the acceptor fluorophore is not possible and there is no signal. After difficulty purifying cyclin A2-GST, the strategy was switched so cyclin A2-His binds the Eu antibody and CDK2 is labelled with the dye. (C) Results of a CDK2 titration TR-FRET assay. The TR-FRET signal is shown for two repeats using cyclin A2-His, and for the control cyclin A2-GST which did not show the expected behaviour of a flat signal.

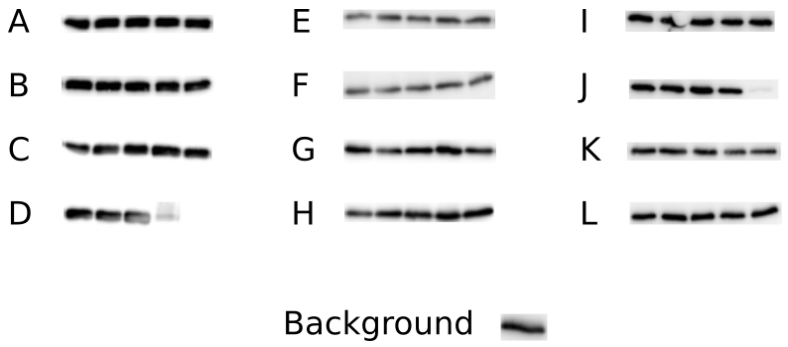



**Figure 4.5** Binding assay results for the 12 compounds tested. The five bands represent increasing compound concentrations from left to right: 0 M control, 8 μM, 40 μM, 200 μM and 1 mM. The strength of the band represents the presence of CDK2. The background signal, in the absence of cyclin A2, is also shown. This shows non-specific binding of CDK2 to the beads and indicates that a signal in the assay would be hard to distinguish from the background.

## 4.3 Discussion

The difficulty of purifying cyclin A2 is likely due to the propensity of the protein to aggregate. This is probably due to large flexibility, a lack of stability in the absence of the binding partner and/or a partially folded state during expression [142]. This is exemplified by the fact that the crystal structure of the CDK2-cyclin A2 complex has only been found for part of the cyclin A2 structure [143], as shown in Figure 4.1B. In addition the full length of human cyclin A2 has not been crystallised in the absence of CDK2, with only the fragment corresponding to residues 171-432 in bovine cyclin A having a structure [200]. This is likely due to reasons related to the difficulties in purification. The presence of the GST tag made the yield from cyclin A2 purification considerably lower than for the His tag. This could be due to the large GST segment causing problems with folding leading to aggregation in the fusion protein.

Though time constraints meant that no more experiments could be carried out, there are a number of other tests that could be done to probe the allostery at the pocket of interest. Primarily this would involve further work on the TR-FRET and binding assays to reduce background signal and allow the compounds to be screened. Other tests that could be carried out include:

- Thermal shift assay: circular dichroism can be used to measure the unfolding of a protein by observing changes in plane-polarised light. If a compound bound to CDK2 it may change the temperature at which the protein unfolds. This assay is comparatively quick and easy but provides limited information.
- Anti-proliferation assay: adding a compound to cancer cells should stop cell proliferation if the compound completely inhibits CDK2-cyclin A2 interaction. However this assay does not indicate that the compound is specific to the CDK2-cyclin A2 interaction, as any effect that prevents proliferation will appear in the assay.
- Mutagenesis: a mutation in the putative binding site of a compound will likely disrupt binding. If inhibition occurs for the wild-type CDK2 but not

for a mutant with a mutation, this gives evidence that the mutation site is the binding site or part of a crucial allosteric communication pathway. This assumes that the mutation has no other effect on the structure such as destabilisation. In addition, the effectiveness of mutagenesis studies to explore allostery has been questioned [67].

- Cyclin E1: another cyclin that binds to CDK2 is cyclin E1. Carrying out the above experiments with cyclin E1 rather than cyclin A2 would give an idea of the specificity of allosteric modulation of the interaction of CDK2 with cyclins. Use of the different structures of cyclin A2 and cyclin E1 in complex with CDK2 may also help elucidate the allosteric mechanism [201].
- Bovine cyclin A: bovine cyclin A purifies more readily than the human protein [198] and this would provide an alternative source to the human protein.
- Kinase assay: cyclin A2 is required for the activation, and hence kinase activity, of CDK2 [143]. Assays that measure the kinase activity in the presence of the compounds indicate whether they are able to prevent CDK2 activation by cyclin A2.
- Surface plasmon resonance (SPR): one binding partner is immobilised on a surface and the other partner is flowed across it. The change in mass on the surface due to complex formation leads to a change in the refractive index of the solvent, which can be measured in real time. SPR is a commonly-used tool to probe binding and would be suitable for the case of CDK2.
- X-ray crystallography: hit compounds could be incubated with CDK2 and an attempt could be made at obtaining a crystal. If successful this would indicate where the compound binds on CDK2 and show conformational changes due to binding. It would also facilitate further computational studies on the new structure.

The difficulty of validating a predicted allosteric site experimentally acts as a deterrent to many groups. However, the success of computational prediction

methods is ultimately determined by whether they can make predictions that are validated experimentally. In addition, iterative cycles of computational prediction and experimental validation can be used to develop and improve the computational approach, as has been seen recently for protein design applications [202]. The work in this chapter represents the first steps that would be taken on the path to experimental validation of a predicted site. As more groups proceed from computational prediction to experimental validation [65] it is hoped that the path will become more established and systematic.

# Chapter 5

# Conclusion

This thesis has explored the concept of allostery in proteins and developed methods for allosteric site prediction. The background of allostery, a property of the protein structural ensemble, was introduced. Two computational approaches were described. AlloPred uses normal modes and machine learning to predict allosteric pockets on proteins. ExProSE uses two structures of a protein to generate ensembles that span conformational space and can be perturbed to predict and explore allostery. Allosteric prediction methods and virtual screening were used to predict modulators for a potential allosteric site on a protein kinase important for cell cycle control, CDK2. These modulators were tested experimentally but the results were inconclusive. Discussion, implications and further work for each of these approaches is described in the relevant chapters.

Themes of this work have included the benefits and drawbacks of NMA, MD and distance geometry methods; the separation of finding good binding sites in general with potential allosteric sites; the importance of experimental validation of computational predictions; the difficulty of comparing allosteric prediction methods; and the necessity of taking account of the protein structural ensemble when studying allostery. Ultimately, the variety of allosteric mechanisms and frameworks for studying allostery makes prediction challenging and indicates that more experimental and computational work is required in

this important area.

For many years papers have pointed to the immense potential of allostery for both understanding and drugging proteins. Yet they regularly contain the qualification that a unified framework of allostery remains 'elusive', and approved allosteric drugs remain rare more than 50 years after the first descriptions of allostery. In order to unlock the true potential of allostery, predictive methods need to be as established and robust as those in other areas of bioinformatics. When allosteric prediction is as effective as prediction of secondary structure or disordered regions, the power of allostery will be truly revealed. However the recent emergence of methods such as those presented here means the future of allosteric prediction looks bright. In an analogous way to allostery itself, it is hoped that the effects of exploring allostery will propagate to all areas of structural biology.

# Appendices

The appendices include documentation for the AlloPred source code, documentation for the ExProSE source code and a description of the Bio.Structure module contributed to the BioJulia organisation.

# AlloPred documentation

This section contains documentation for the AlloPred source code. This documentation can be found along with the source code at `https://github.com/jgreener64/allopred`. The current released version of the source code is v1.0.0.

## Requirements

- Python 2.7 with the NumPy and ProDy packages installed.
- Fpocket v2.0, which can be downloaded from `http://fpocket.sourceforge.net`. Follow the installation instructions to compile the executables.
- SVM-light, which can be downloaded from `http://svmlight.joachims.org`. Follow the installation instructions to compile the executables.

## Usage

Follow these steps to set up AlloPred - the shell commands are for bash:

1. Download the files and extract them as usual, or clone the repository.
2. The environmental variables `$ALLOPRED_DIR` and `$SVM_LIGHT_DIR` need to be set as the filepaths to the AlloPred directory and the SVM-light directory respectively:

   ```
   export ALLOPRED_DIR=/path/to/allopred/
   export SVM_LIGHT_DIR=/path/to/svm_light/
   ```

   Consider adding these lines to your profile so you don't have to run them every session.

Follow these step to run AlloPred:

1. Obtain a PDB format file (`in_file.pdb`), e.g. from the Protein Data Bank.
2. Create a one-line file (`act_res.txt`) containing the active site residues of the protein. The format is `10:A,11:B` for residue 10 on chain A and

residue 11 on chain B. These can be found using resources such as the Catalytic Site Atlas. An example PDB file and active residue file can be found in the example directory of AlloPred.

3. Run Fpocket v2.0 on the PDB file:

```
fpocket -f in_file.pdb
```

   This assumes `fpocket` is on the path. This produces the directory `in_file_out`. AlloPred is optimised on the default Fpocket parameters but you can change these in accordance with the Fpocket documentation if you wish.

4. The following command, from the directory containing `in_file.pdb` and `in_file_out`, runs the AlloPred pipeline:

```
python $ALLOPRED_DIR/run_allopred.py in_file act_res.txt
```

   The arguments are the input file prefix and the path to the active site residue file. Running the `run_allopred.py` script with fewer than 2 arguments returns these instructions for the command.

5. The output files are:

   - `in_file.out`: the AlloPred output file containing the input parameters and the values for each pocket in order of AlloPred ranking.
   - `in_file.svm`: the SVM input file in the SVM-light format.

## Other files

- `dataset` contains information on the training and testing sets.
- `example` contains the inputs and outputs of an example run using the PDB entry with ID 1FX2.
- `svm_model.txt` is the optimised SVM built on the whole training set.

# ExProSE documentation

This section contains documentation for the ExProSE source code. This documentation can be found along with the source code at `https://github.com/jgreener64/ProteinEnsembles.jl`. The current released version of the source code is v0.1.1.

## Summary

Install using `Pkg.add("ProteinEnsembles")` from within Julia v0.5 or v0.6. Run using

```
exprose --i1 input_1.pdb --d1 input_1.dssp \
    --i2 input_2.pdb --d2 input_2.dssp \
    -n 50 -o exprose_out
```

where `exprose` is in the `bin` directory.

## Installation

Julia v0.5 or v0.6 is required and can be downloaded from `http://julialang.org/downloads`. Install ProteinEnsembles.jl by running `Pkg.add("ProteinEnsembles")` from the Julia REPL. This will also automatically install a few other required Julia packages. If you want, the tests can be run using `Pkg.test("ProteinEnsembles")`. If you wish to use the auto-parameterisation procedure (see below) you must also have TM-score installed.

## Requirements

To use ProteinEnsembles.jl you will need the following:

- PDB files of the protein of interest. Two is best, but one may be used (see the paper). They must have polar hydrogens only added; this can be done using tools such as Chimera or pdbtools. The chain labelling and

residue numbering must be consistent between the files as this is used to find common atoms. Alternative atom locations are discarded. PDB files must also be a single model and not have any inserted residues. HETATM records are discarded by default.

- DSSP files corresponding to the PDB files above. These can be obtained using dssp.

## Usage

These instructions are tailored towards Mac/Unix. However they could be modified to work on Windows.

Although organised as a Julia package, ProteinEnsembles.jl is primarily designed for use from the command line. The `exprose` script in the `bin` directory implements this. For example, to see the command line options run

```
~/.julia/v0.6/ProteinEnsembles/bin/exprose -h
```

For easy access to the `exprose` command you might like to add the following line to your profile:

```
export PATH=$PATH:~/.julia/v0.6/ProteinEnsembles/bin
```

Then, if all input files are in your current directory, run the program as follows:

```
# Generate an ensemble of 50 structures with an output directory exprose_out
exprose --i1 input_1.pdb --d1 input_1.dssp --i2 input_2.pdb \
    --d2 input_2.dssp -n 50 -o exprose_out

# Use a tolerance weighting of 0.5
exprose --i1 input_1.pdb --d1 input_1.dssp --i2 input_2.pdb \
    --d2 input_2.dssp -n 50 -o exprose_out -w 0.5

# Generate an ensemble from a single structure with a tolerance weighting of 1.0
exprose --i1 input_1.pdb --d1 input_1.dssp -n 50 -o exprose_out -w 1.0
```

The method may also be run from within Julia. The below Julia script does the same thing as the first example above:

```
using ProteinEnsembles
runpipeline(
```

```
    i1="input_1.pdb",
    d1="input_1.dssp",
    i2="input_2.pdb",
    d2="input_2.dssp",
    n_strucs=50,
    out_dir="exprose_out"
)
```

Or, to split it up a little into the constituent functions:

```
using ProteinEnsembles
constraints_com, constraints_one, constraints_two = interactions(
    "input_1.pdb",
    "input_1.dssp",
    "input_2.pdb",
    "input_2.dssp"
)
ensemble_com = generateensemble(constraints_com, 50)
runanalysis("exprose_out", ensemble_com, constraints_one, constraints_two)
```

### Selecting parameters

The auto-parameterisation procedure can select a more suitable tolerance weighting value (see the paper). TM-score must be installed to do this. For example:

```
exprose-param --i1 input_1.pdb --d1 input_1.dssp --i2 input_2.pdb \
    --d2 input_2.dssp -o exprose_param -t TMscore
```

runs the auto-parameterisation procedure with the `-t` option specifying the command to run TM-score. The last line of the output gives a suggested tolerance weighting. This value is also written out to `suggested.tsv`. Use this value in a normal `exprose` run as above.

### Allosteric site prediction

To predict allosteric sites you should run LIGSITE$^{cs}$ on the second input structure (the one you give as `--i2`). You then need to run the `cluster-ligsite` script in `bin` to assign the points to pockets:

```
cluster-ligsite pocket_r.pdb pocket_all.pdb pocket_points.pdb
```

where `pocket_r.pdb` and `pocket_all.pdb` are in the LIGSITE$^{cs}$ output. Then carry out an `exprose` run with the `pocket_points.pdb` file (`-l`) and the number of pockets (e.g. top 4) to perturb at (`-m`) as parameters:

```
exprose --i1 input_1.pdb --d1 input_1.dssp --i2 input_2.pdb \
    --d2 input_2.dssp -n 50 -o exprose_out -l pocket_points.pdb -m 4
```

A tolerance weighting from an auto-parameterisation run can also be used here. View the `predictions.tsv` output file to get the order of allosteric pocket predictions. Note that other pocket prediction software can be used provided you can get the output into the same format as `pocket_points.pdb`, i.e. pocket cavity points with the pocket number in the residue number column.

### Output

The output directory contains the following:

- `input_1.pdb` and `input_2.pdb`: atoms used from the input structures are written back out and superimposed.
- `pdbs`: generated structures in PDB format. Superimposed to `input_1.pdb` and `input_2.pdb`.
- `pcs`: projections onto the principal components (PCs) from the principal component analysis of the generated structures. Contains files for generated (`pcs.tsv`) and input structures (`pcs_input_1.tsv` and `pcs_input_2.tsv`) - line n corresponds to structure n and column c corresponds to PC c. Has graphs of these for the first few PCs (`pc_x_y.png`). Also includes a list of PCs ordered by decreasing distance between the input structures (`pcs_input_dist.tsv`) and the percentage variation explained by each PC (`evals_spread.tsv`).
- `pymol`: PyMol scripts to view PCs on `input_1.pdb`, e.g. run `pymol input_1.pdb pymol/view_pc_1.pml`.
- `rmsds_input_1.tsv` and `rmsds_input_2.tsv`: RMSDs of generated structures to the input structures. Line n corresponds to structure n.

- `rmsfs.tsv` and `rmsfs.png`: RMSFs of each residue over the ensemble of generated structures, and a plot of this. Line n corresponds to residue index n.
- `spe_scores.tsv`: SPE error scores of generated structures (see paper). Line n corresponds to structure n.

For allosteric site prediction there will be `pdbs_mod_n` and `mod_n` containing similar information for each perturbed ensemble, as well as the ratio of RMSF values to the unperturbed ensemble (`rmsfs_ratio.tsv`). There will also be the order of allosteric predictions (`predictions.tsv`) and the size of the perturbation on modulating each site (`perturbations.tsv`), which is the RMSD between the centroid structure of the perturbed and unperturbed ensembles.

The default plot colours are blue for generated structures, red for input structure 1, green for input structure 2 and orange for perturbed ensemble structures.

# BioJulia Bio.Structure module

This section contains information on the Bio.Structure module contributed to the BioJulia organisation.

The Julia language [163] is a new programming language that has grown quickly in the field of scientific computing. It has syntax similar to MATLAB but execution speeds approaching statically-compiled languages like C, making it suitable for scientific research applications where both usability and execution speed are important.

Open source software packages to parse PDB files and manipulate protein structures exist in many programming languages. There are a lack of such packages in the Julia language so a new module, Bio.Structure, was contributed to the BioJulia project (`http://biojulia.net`). Features of the package include fast PDB parsing, easy access to structural elements, iteration over elements, selector functions, downloading of PDB files, writing PDB files and spatial functions such as distances and Ramachandran angles. The type hierarchy is based on Biopython [203] and is shown in Figure 5.1. Examples of basic use cases for Bio.Structure are shown in Figure 5.2.

As part of the development of Bio.Structure information was collected on existing packages with similar functionality in various programming languages. These comparisons are shown in Table 5.1. The results of each package running various benchmarks are shown in Figure 5.3. Bio.Structure is able to read in PDB file 1CRN in 1.4 ms (after just-in-time compilation) on average. This is faster than any other package tested.

| | **BioJulia** | **MIToS** | **Biopython** | **ProDy** | **MDAnalysis** | **Bio3D** | **Rpdb** | **BioPerl** | **BioRuby** | **Victor** | **ESBTL** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Parse 1CRN / ms | 1.4 | 2.4 | 9.1 | 2.2 | 6.4 | 31 | 19 | 63 | 25 | 10 | 6.8 |
| Parse 3JYV / s | 0.49 | 0.74 | 1.0 | 0.28 | 0.80 | 14 | 2.2 | 3.8 | 0.98 | 7.7 | 0.95 |
| Parse 1HTQ / s | 27 | 28 | 25 | 1.7 | 3.0 | 60 | 34 | 71 | 18 | 17 | - |
| Count / ms | 0.91 | 0.16 | 0.48 | 8.9 | 5.7 | 0.53 | 0.46 | 0.79 | 0.19 | - | - |
| Distance / ms | 0.11 | 0.011 | 0.39 | 5.6 | 3.3 | 1.4 | 1.9 | 0.85 | 0.51 | - | - |
| Ramachandran / ms | 13 | - | 130 | 180 | 3500 | - | - | - | - | - | - |
| Language | Julia | Julia | Python | Python | Python | R | R | Perl | Ruby | C++ | C++ |
| Parses header | × | × | ✓ | ✓ | × | ✓ | ✓ | × | ✓ | ✓ | × |
| Hierarchichal parsing | ✓ | × | ✓ | ✓ | ✓ | × | × | ✓ | ✓ | ✓ | ✓ |
| Supports disorder | ✓ | × | ✓ | × | × | × | × | × | × | × | ✓ |
| Writes PDB files | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | × | ✓ | ✓ |
| Superimposition | × | ✓ | ✓ | ✓ | ✓ | ✓ | × | × | × | × | × |
| PCA | × | × | × | ✓ | ✓ | ✓ | × | × | × | × | × |
| Software license | MIT | MIT | Biopython | MIT | GPLv2 | GPLv2 | GPL | GPL/ Artistic | Ruby | GPLv3 | GPLv3 |

**Table 5.1** Comparison of open source packages to read and manipulate PDB files in various programming languages. See Figure 5.3 for descriptions and a visual representation of the benchmarks.

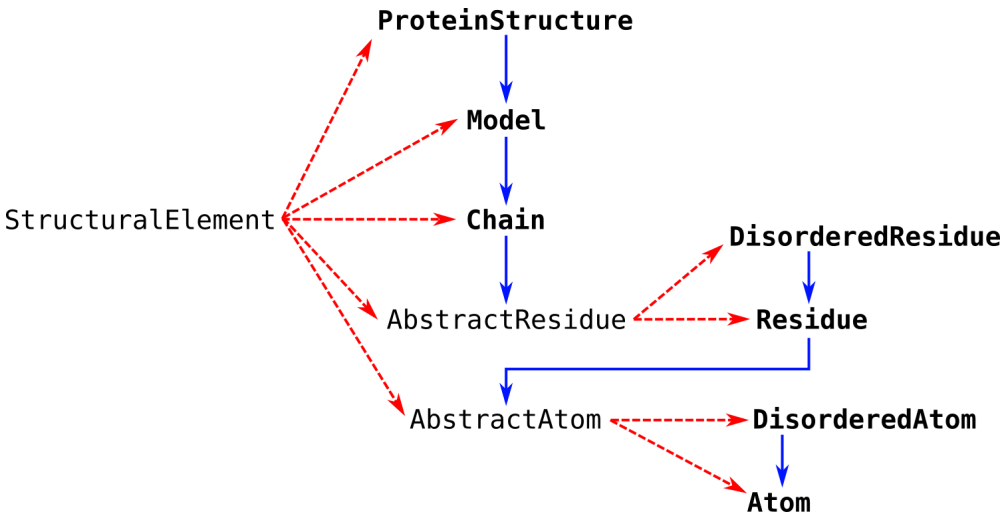



**Figure 5.1** Hierarchy of types in the BioJulia Bio.Structure module. Types, analogous to classes in other languages, are shown in text. `StructuralElement`, `AbstractResidue` and `AbstractAtom` are abstract types and may not themselves be instantiated. Concrete (i.e. not abstract) types are shown in bold text. Subtypes, analogous to subclasses, are indicated by red dotted lines. A blue line indicates that an instance of the type contains a list of the indicated type. For example, a `Chain` contains multiple `AbstractResidues`.

```
In [9]: using Bio.Structure;

In [10]: struc = read("1AKE.pdb", PDB)

Out[10]: Bio.Structure.ProteinStructure
         Name                        -  1AKE.pdb
         Number of models            -  1
         Chain(s)                    -  AB
         Number of residues          -  428
         Number of point mutations   -  0
         Number of other molecules   -  2
         Number of water molecules   -  378
         Number of atoms             -  3312
         Number of hydrogens         -  0
         Number of disordered atoms  -  5

In [11]: struc['A'][100]

Out[11]: Bio.Structure.Residue
         Residue ID                  -  100:A
         Residue name                -  GLY
         Number of atoms             -  4
         Number of hydrogens         -  0
         Number of disordered atoms  -  0

In [12]: calphas = collectatoms(struc['A'], calphaselector);

In [13]: seq = AminoAcidSequence(struc['A'], standardselector)

Out[13]: 214aa Amino Acid Sequence:
         MRIILLGAPGAGKGTQAQFIMEKYGIPQISTGDMLRAAV…APLIGYYSKEAEAGNTKYAKVDGTKPVAEVRAD
         LEKILG

In [14]: for mod in struc
             for ch in mod
                 for res in ch
                     for at in res
                         # Do something
                     end
                 end
             end
         end
```

**Figure 5.2** Example functionality of the Bio.Structure module. The Jupyter Notebook [204] and Julia v0.5.2 are used. Cases shown are importing the module, reading in a PDB file, accessing residues by index, extracting $C^{\alpha}$ atoms, extracting the amino acid sequence and iterating over sub-elements.

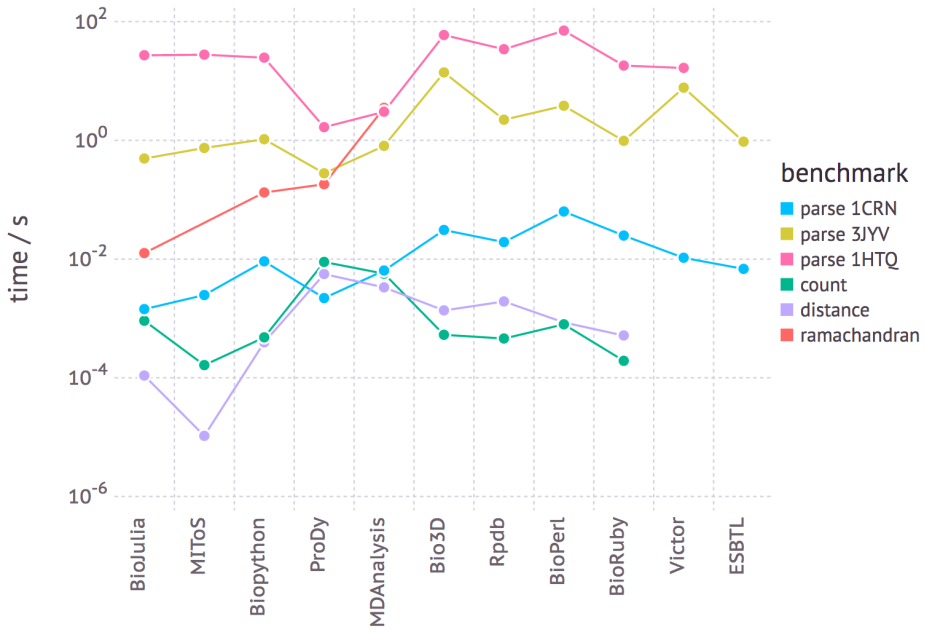



**Figure 5.3** Benchmarks on common tasks for open source packages to read and manipulate PDB files in various programming languages. Benchmarks were carried out on a 3.1 GHz Intel Core i7 processor with 16 GB 1867 MHz DDR3 RAM. The operating system was Mac OS X Yosemite 10.10.5. Time is the elapsed time. The mean over a number of runs is taken for each benchmark. The three PDB files parsed are 1CRN (327 atoms), 3JYV (57,327 atoms) and 1HTQ (10 models of 97,872 atoms). These are taken from the benchmarking in [205]. 'Count' is a count of the number of alanine residues in adenylate kinase (PDB ID 1AKE). 'Distance' is a calculation of the distance between residues 50 and 60 of chain A in adenylate kinase. 'Ramachandran' is a calculation of the Ramachandran $\varphi/\psi$ angles in adenylate kinase.

# References


[1] R Nussinov and C J Tsai. Allostery in Disease and in Drug Discovery. *Cell*, 153:293–305, 2013.

[2] K Gunasekaran, B Ma, and R Nussinov. Is Allostery an Intrinsic Property of All Dynamic Proteins? *Proteins*, 57:433–443, 2004.

[3] H N Motlagh, J O Wrabl, J Li, and V J Hilser. The ensemble nature of allostery. *Nature*, 508:331–339, 2014.

[4] J Monod, J Wyman, and J P Changeux. On the nature of allosteric transitions: A plausible model. *J Mol Biol*, 12:88–118, 1965.

[5] D E Koshland, G Némethy, and D Filmer. Comparison of experimental binding data and theoretical models in proteins containing subunits. *Biochemistry*, 5(1):365–385, 1966.

[6] M F Perutz. Stereochemistry of cooperative effects in haemoglobin. *Nature*, 228(5273):726–739, 1970.

[7] A Cooper and D T Dryden. Allostery without conformational change: A plausible model. *Eur Biophys J*, 11(2):103–109, 1984.

[8] V J Hilser, J O Wrabl, and H N Motlagh. Structural and energetic basis of allostery. *Annu Rev Biophys*, 41:585–609, 2012.

[9] Q Cui and M Karplus. Allostery and cooperativity revisited. *Protein Sci*, 17(8):1295–1307, 2008.

[10] C J Tsai and R Nussinov. A unified view of 'how allostery works'. *PLoS Comput Biol*, 10(2):e1003394, 2014.

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

[12] G M Suel, S W Lockless, M A Wall, and R Ranganathan. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat Struct Biol*, 10(1):59–69, 2003.

[13] C J Wenthur, P R Gentry, T P Mathews, and C W Lindsley. Drugs for Allosteric Sites on Receptors. *Annu Rev Pharmacol*, 54:165–184, 2014.

[14] D Wootten, A Christopoulos, and P M Sexton. Emerging paradigms in GPCR allostery: implications for drug discovery. *Nat Rev Drug Discov*, 12(8):630–644, 2013.

[15] P J Conn, A Christopoulos, and C W Lindsley. Allosteric modulators of GPCRs: a novel approach for the treatment of CNS disorders. *Nat Rev Drug Discov*, 8(1):41–54, 2009.

[16] O Schueler-Furman and S J Wodak. Computational approaches to investigating allostery. *Curr Opin Struct Biol*, 41:159–171, 2016.

[17] J R Wagner, C T Lee, J D Durrant, R D Malmstrom, V A Feher, and R E Amaro. Emerging Computational Methods for the Rational Discovery of Allosteric Drugs. *Chem Rev*, 116(11):6370–6390, 2016.

[18] E Guarnera and I N Berezovsky. Allosteric sites: remote control in regulation of protein activity. *Curr Opin Struct Biol*, 37:1–8, 2016.

[19] S Lu, W Huang, and J Zhang. Recent computational advances in the identification of allosteric sites in proteins. *Drug Discov Today*, 19(10):1595–1600, 2014.

[20] K Kamata, M Mitsuya, T Nishimura, J Eiki, and Y Nagata. Structural basis for allosteric regulation of the monomeric allosteric enzyme human glucokinase. *Structure*, 12(3):429–438, 2004.

[21] J G Greener and M J E Sternberg. AlloPred: prediction of allosteric pockets on proteins using normal mode perturbation analysis. *BMC Bioinformatics*, 16(335):1–7, 2015.

[22] E Guarnera, Z W Tan, Z Zheng, and I N Berezovsky. AlloSigMA: allosteric signaling and mutation analysis server. *Bioinformatics*, 33(24):3996–3998, 2017.

[23] K Song, X Liu, W Huang, S Lu, Q Shen, L Zhang, and J Zhang. Improved Method for the Identification and Validation of Allosteric Sites. *J Chem Inf Model*, 2017.

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

[25] P Weinkam, J Pons, and A Sali. Structure-based model of allostery predicts coupling between distant sites. *Proc Natl Acad Sci USA*, 109(13):4875–4880, 2012.

[26] E V Wasmuth and C D Lima. The Rrp6 C-terminal domain binds RNA and activates the nuclear RNA exosome. *Nucleic Acids Res*, 45(2):846–860, 2017.

[27] J G Greener, I Filippis, and M J E Sternberg. Predicting Protein Dynamics and Allostery Using Multi-Protein Atomic Distance Constraints. *Structure*, 25(3):546–558, 2017.

[28] C Kaya, A Armutlulu, S Ekesan, and T Haliloglu. MCPath: Monte Carlo path generation approach to predict likely allosteric pathways and functional residues. *Nucleic Acids Res*, 41:W249–W255, 2013.

[29] A Panjkovich and X Daura. PARS: a web server for the prediction of Protein Allosteric and Regulatory Sites. *Bioinformatics*, 30(9):1314–1315, 2014.

[30] A Panjkovich and X Daura. Exploiting protein flexibility to predict the location of allosteric sites. *BMC Bioinformatics*, 13(273):1–12, 2012.

[31] A Goncearenco, S Mitternacht, T Yong, B Eisenhaber, F Eisenhaber, and I N Berezovsky. SPACER: server for predicting allosteric communication and effects of regulation. *Nucleic Acids Res*, 41:W266–W272, 2013.

[32] S Mitternacht and I N Berezovsky. Binding Leverage as a Molecular Basis for Allosteric Regulation. *PLoS Comput Biol*, 7(9):e1002148, 2011.

[33] D Clarke, A Sethi, S Li, S Kumar, R W Chang, J Chen, and M Gerstein. Identifying Allosteric Hotspots with Dynamics: Application to Inter- and Intra-species Conservation. *Structure*, 24(5):826–837, 2016.

[34] B Huang and M Schroeder. LIGSITE$^{csc}$: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct Biol*, 6(19), 2006.

[35] V Le Guilloux, P Schmidtke, and P Tuffery. Fpocket: An open source platform for ligand pocket detection. *BMC Bioinformatics*, 10(168), 2009.

[36] P Cimermancic, P Weinkam, T J Rettenmaier, L Bichmann, D A Keedy, R A Woldeyes, D Schneidman-Duhovny, O N Demerdash, J C Mitchell, J A Wells, J S Fraser, and A Sali. CryptoSite: Expanding the Druggable Proteome by Characterization and Prediction of Cryptic Binding Sites. *J Mol Biol*, 428(4):709–719, 2016.

[37] S Hayward and B L de Groot. Normal Modes and Essential Dynamics. *Methods in Molecular Biology*, 443:89–106, 2008.

[38] I Bahar and A J Rader. Coarse-grained normal mode analysis in structural biology. *Curr Opin Struc Biol*, 15:586–592, 2005.

[39] G Collier and V Ortiz. Emerging computational approaches for the study of protein allostery. *Arch Biochem Biophys*, 538:6–15, 2013.

[40] I A Balabin, W Yang, and D N Beratan. Coarse-grained modeling of allosteric regulation in protein receptors. *Proc Natl Acad Sci USA*, 106(34):14253–14258, 2009.

[41] T L Rodgers, P D Townsend, D Burnell, M L Jones, S A Richards, T C B McLeish, E Pohl, M R Wilson, and M J Cann. Modulation of Global Low-Frequency Motions Underlies Allosteric Regulation: Demonstration in CRP/FNR Family Transcription Factors. *PLoS Biol*, 11(9):e1001651, 2013.

[42] W Zheng, B R Brooks, and D Thirumalai. Allosteric Transitions in the Chaperonin GroEL are Captured by a Dominant Normal Mode that is Most Robust to Sequence Variations. *Biophysical J*, 93(7):2289–2299, 2007.

[43] J G Su, L S Qi, C H Li, Y Y Zhu, H J Du, Y X Hou, R Hao, and J H Wang. Prediction of allosteric sites on protein surface with an elastic-network-model-based thermodynamic method. *Phys Rev E*, 90:022719, 2014.

[44] A S Y Chen, N J Westwood, P Brear, G W Rogers, L Mavridis, and J B O Mitchell. A Random Forest Model for Predicting Allosteric and Functional Sites on Proteins. *Mol Inf*, 35(3-4):125–135, 2016.

[45] F Pontiggia, D V Pachov, M W Clarkson, J Villali, M F Hagan, V S Pande, and D Kern. Free energy landscape of activation in a signalling protein at atomic resolution. *Nat Commun*, 6:7284, 2015.

[46] D Penkler, O Sensoy, C Atilgan, and O Tastan Bishop. Perturbation-Response Scanning Reveals Key Residues for Allosteric Control in Hsp70. *J Chem Inf Model*, 57(6):1359–1374, 2017.

[47] S W Lockless and R Ranganathan. Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families. *Science*, 286:295–299, 1999.

[48] K A Reynolds, R N McLaughlin, and R Ranganathan. Hot spots for allosteric regulation on protein surfaces. *Cell*, 147(7):1564–1575, 2011.

[49] I Anishchenko, S Ovchinnikov, H Kamisetty, and D Baker. Origins of coevolution between residues distant in protein 3D structures. *Proc Natl Acad Sci USA*, 114(34):9122–9127, 2017.

[50] A Panjkovich and X Daura. Assessing the structural conservation of protein pockets to study functional and allosteric sites: implications for drug discovery. *BMC Struct Biol*, 10(9):1–14, 2010.

[51] A Panjkovich. Structure and evolution of protein allosteric sites. *PhD thesis, Universitat Autònoma de Barcelona*, 2013.

[52] B R Amor, M T Schaub, S N Yaliraki, and M Barahona. Prediction of allosteric sites and mediating interactions through bond-to-bond propensities. *Nat Commun*, 7:12477, 2016.

[53] Y Qi, Q Wang, B Tang, and L Lai. Identifying Allosteric Binding Sites in Proteins with a Two-State Gō Model for Novel Allosteric Effector Discovery. *J Chem Theory Comput*, 8:2962–2971, 2012.

[54] D Kozakov, L E Grove, D R Hall, T Bohnuud, S E Mottarella, L Luo, B Xia, D Beglov, and S Vajda. The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins. *Nat Protoc*, 10(5):733–755, 2015.

[55] P Ghanakota and H A Carlson. Moving Beyond Active-Site Detection: MixMD Applied to Allosteric Systems. *J Phys Chem B*, 120(33):8685–8695, 2016.

[56] M Hendlich, F Rippmann, and G Barnickel. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *J Mol Graph Model*, 15:359–363, 1997.

[57] J Jiménez, S Doerr, G Martínez-Rosell, A S Rose, and G De Fabritiis. DeepSite: protein-binding site predictor using 3D-convolutional neural networks. *Bioinformatics*, 33(19):3036–3042, 2017.

[58] N V Dokholyan. Controlling Allosteric Networks in Proteins. *Chem Rev*, 116(11):6463–6487, 2016.

[59] O N A Demerdash, M D Daily, and J C Mitchell. Structure-Based Predictive Models for Allosteric Hot Spots. *PLoS Comput Biol*, 5(10):e1000531, 2009.

[60] O Dagliyan, M Tarnawski, P H Chu, D Shirvanyants, I Schlichting, N V Dokholyan, and K M Hahn. Engineering extrinsic disorder to control protein activity in living cells. *Science*, 354(6318):1441–1444, 2016.

[61] S Buchenberg, F Sittel, and G Stock. Time-resolved observation of protein allosteric communication. *Proc Natl Acad Sci USA*, 114(33):E6804–E6811, 2017.

[62] M D Daily and J J Gray. Allosteric communication occurs via networks of tertiary and quaternary motions in proteins. *PLoS Comput Biol*, 5(2):e1000293, 2009.

[63] S Wellington, P P Nag, K Michalska, S E Johnston, R P Jedrzejczak, V K Kaushik, A E Clatworthy, N Siddiqi, P McCarren, B Bajrami, N I Maltseva, S Combs, S L Fisher, A Joachimiak, S L Schreiber, and D T Hung. A small-molecule allosteric inhibitor of Mycobacterium tuberculosis tryptophan synthase. *Nat Chem Biol*, 13(9):943–950, 2017.

[64] D Haselbach, J Schrader, F Lambrecht, F Henneberg, A Chari, and H Stark. Long-range allosteric regulation of the human 26S proteasome by 20S proteasome-targeting cancer drugs. *Nat Commun*, 8:15578, 2017.

[65] M Brecher, Z Li, B Liu, J Zhang, C A Koetzner, A Alifarag, S A Jones, Q Lin, L D Kramer, and H Li. A conformational switch high-throughput screening assay and allosteric inhibition of the flavivirus NS2B-NS3 protease. *PLoS Pathog*, 13(5):e1006411, 2017.

[66] W Ye, T Qian, H Liu, R Luo, and H F Chen. Allosteric Autoinhibition Pathway in Transcription Factor ERG: Dynamics Network and Mutant Experimental Evaluations. *J Chem Inf Model*, 57(5):1153–1165, 2017.

[67] Q Tang, A Y Alontaga, T Holyoak, and A W Fenton. Exploring the limits of the usefulness of mutagenesis in studies of allosteric mechanisms. *Hum Mutat*, 38(9):1144–1154, 2017.

[68] M P Martin, R Alam, S Betzi, D J Ingles, J Y Zhu, and E Schonbrunn. A novel approach to the discovery of small-molecule ligands of CDK2. *Chembiochem*, 13(14):2128–2136, 2012.

[69] S S Jayakar, G Ang, D C Chiara, and A K Hamouda. Photoaffinity Labeling of Pentameric Ligand-Gated Ion Channels: A Proteomic Approach to Identify Allosteric Modulator Binding Sites. *Methods Mol Biol*, 1598:157–197, 2017.

[70] M Pellerano, S Tcherniuk, C Perals, T N Ngoc Van, E Garcin, F Mahuteau-Betzer, M P Teulade-Fichou, and M C Morris. Targeting Conformational Activation of CDK2 Kinase. *Biotechnol J*, 12(8):1600531, 2017.

[71] J P Pisco, C de Chiara, K J Pacholarz, A Garza-Garcia, R W Ogrodowicz, P A Walker, P E Barran, S J Smerdon, and L P S de Carvalho. Uncoupling conformational states from activity in an allosteric enzyme. *Nat Commun*, 8(203):1–10, 2017.

[72] S Raman, N Taylor, N Genuth, S Fields, and G M Church. Engineering allostery. *Trends Genet*, 30(12):521–528, 2014.

[73] D D Boehr, R Nussinov, and P E Wright. The role of dynamic conformational ensembles in biomolecular recognition. *Nat Chem Biol*, 5(11):789–796, 2009.

[74] C J Lee, X Liang, Q Wu, J Najeeb, J Zhao, R Gopalaswamy, M Titecat, F Sebbane, N Lemaitre, E J Toone, and P Zhou. Drug design from the cryptic inhibitor envelope. *Nat Commun*, 7:10638, 2016.

[75] G R Bowman, E R Bolin, K M Hart, B C Maguire, and S Marqusee. Discovery of multiple hidden allosteric sites by combining Markov state models and experiments. *Proc Natl Acad Sci USA*, 112(9):2734–2739, 2015.

[76] K M Hart, K E Moeder, C M W Ho, M I Zimmerman, T E Frederick, and G R Bowman. Designing small molecules to target cryptic pockets yields both positive and negative allosteric modulators. *PLoS One*, 12(6):e0178678, 2017.

[77] V Oleinikovas, G Saladino, B P Cossins, and F L Gervasio. Understanding Cryptic Pocket Formation in Protein Targets by Enhanced Sampling Simulations. *J Am Chem Soc*, 138(43):14257–14263, 2016.

[78] O V Makhlynets, E A Raymond, and I V Korendovych. Design of allosterically regulated protein catalysts. *Biochemistry*, 54(7):1444–1456, 2015.

[79] B L Oakes, D C Nadler, A Flamholz, C Fellmann, B T Staahl, J A Doudna, and D F Savage. Profiling of engineering hotspots identifies an allosteric CRISPR-Cas9 switch. *Nat Biotechnol*, 34(6):646–651, 2016.

[80] J H Choi, A H Laurent, V J Hilser, and M Ostermeier. Design of protein switches based on an ensemble model of allostery. *Nat Commun*, 6(6968):1–9, 2015.

[81] N D Taylor, A S Garruss, R Moretti, S Chan, M A Arbing, D Cascio, J K Rogers, F J Isaacs, S Kosuri, D Baker, S Fields, G M Church, and S Raman. Engineering an allosteric transcription factor to respond to new ligands. *Nat Methods*, 13(2):177–183, 2016.

[82] O Dagliyan, D Shirvanyants, A V Karginov, F Ding, L Fee, S N Chandrasekaran, C M Freisinger, G A Smolen, A Huttenlocher, K M Hahn, and N V Dokholyan. Rational design of a ligand-controlled protein conformational switch. *Proc Natl Acad Sci USA*, 110(17):6800–6804, 2013.

[83] Q Xu, Q Tang, P Katsonis, O Lichtarge, D Jones, S Bovo, G Babbi, P L Martelli, R Casadio, G R Lee, C Seok, A W Fenton, and R L Dunbrack. Benchmarking predictions of allostery in liver pyruvate kinase in CAGI4. *Hum Mutat*, 38(9):1123–1131, 2017.

[84] J Moult, K Fidelis, A Kryshtafovych, T Schwede, and A Tramontano. Critical assessment of methods of protein structure prediction: Progress and new directions in round XI. *Proteins*, 84 Suppl 1:4–14, 2016.

[85] W Huang, G Wang, Q Shen, X Liu, S Lu, L Geng, Z Huang, and J Zhang. ASBench: benchmarking sets for allosteric discovery. *Bioinformatics*, 31(15):2598–2600, 2015.

[86] Q Shen, G Wang, S Li, X Liu, S Lu, Z Chen, K Song, J Yan, L Geng, Z Huang, W Huang, G Chen, and J Zhang. ASD v3.0: unraveling allosteric regulation with structural mechanisms and biological networks. *Nucleic Acids Res*, 44:D527–D535, 2016.

[87] H N Motlagh and V J Hilser. Agonism/antagonism switching in allosteric ensembles. *Proc Natl Acad Sci USA*, 109(11):4134–4139, 2012.

[88] R Nussinov and C J Tsai. Unraveling structural mechanisms of allosteric drug action. *Trends Pharmacol Sci*, 35(5):256–264, 2014.

[89] N Popovych, S Sun, R H Ebright, and C G Kalodimos. Dynamically driven protein allostery. *Nat Struct Mol Biol*, 13(9):831–838, 2006.

[90] D A Capdevila, J J Braymer, K A Edmonds, H Wu, and D P Giedroc. Entropy redistribution controls allostery in a metalloregulatory protein. *Proc Natl Acad Sci USA*, 114(17):4424–4429, 2017.

[91] R Nussinov and C J Tsai. Allostery without a conformational change? Revisiting the paradigm. *Curr Opin Struc Biol*, 30:17–24, 2015.

[92] A Kumawat and S Chakrabarty. Hidden electrostatic basis of dynamic allostery in a PDZ domain. *Proc Natl Acad Sci USA*, 114(29):E5825–E5834, 2017.

[93] J Liu and R Nussinov. Energetic redistribution in allostery to execute protein function. *Proc Natl Acad Sci USA*, 114(29):7480–7482, 2017.

[94] B Buchli, S A Waldauer, R Walser, M L Donten, R Pfister, N Blochliger, S Steiner, A Caflisch, O Zerbe, and P Hamm. Kinetic response of a photoperturbed allosteric protein. *Proc Natl Acad Sci USA*, 110(29):11725–11730, 2013.

[95] G Ozorowski, J Pallesen, N de Val, D Lyumkis, C A Cottrell, J L Torres, J Copps, R L Stanfield, A Cupo, P Pugach, J P Moore, I A Wilson, and A B Ward. Open and closed structures reveal allostery and pliability in the HIV-1 envelope spike. *Nature*, 547:360–363, 2017.

[96] S Singh and G R Bowman. Quantifying Allosteric Communication via Both Concerted Structural Changes and Conformational Disorder with CARDS. *J Chem Theory Comput*, 13(4):1509–1517, 2017.

[97] J Wang, G Custer, D Beckett, and S Matysiak. Long Distance Modulation of Disorder-to-Order Transitions in Protein Allostery. *Biochemistry*, 56(34):4478–4488, 2017.

[98] C A Lipinski, F Lombardo, B W Dominy, and P J Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev*, 46(1-3):3–26, 2001.

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

[100] Q Wang, M Zheng, Z Huang, X Liu, H Zhou, Y Chen, T Shi, and J Zhang. Toward understanding the molecular basis for chemical allosteric modulator design. *J Mol Graph Model*, 38:324–333, 2012.

[101] K Henzler-Wildman and D Kern. Dynamic personalities of proteins. *Nature*, 450(7172):964–972, 2007.

[102] G Wei, W Xi, R Nussinov, and B Ma. Protein Ensembles: How Does Nature Harness Thermodynamic Fluctuations for Life? The Diverse Functional Roles of Conformational Ensembles in the Cell. *Chem Rev*, 116(11):6516–6551, 2016.

[103] K A Henzler-Wildman, V Thai, M Lei, M Ott, M Wolf-Watz, T Fenn, E Pozharski, M A Wilson, G A Petsko, M Karplus, C G Hubner, and D Kern. Intrinsic motions along an enzymatic reaction trajectory. *Nature*, 450(7171):838–844, 2007.

[104] P Sormanni, D Piovesan, G T Heller, M Bonomi, P Kukic, C Camilloni, M Fuxreiter, Z Dosztanyi, R V Pappu, M M Babu, S Longhi, P Tompa, A K Dunker, V N Uversky, S C Tosatto, and M Vendruscolo. Simultaneous quantification of protein order and disorder. *Nat Chem Biol*, 13(4):339–342, 2017.

[105] M Bonomi, G T Heller, C Camilloni, and M Vendruscolo. Principles of protein structural ensemble determination. *Curr Opin Struct Biol*, 42:106–116, 2017.

[106] T Maximova, R Moffatt, B Ma, R Nussinov, and A Shehu. Principles and Overview of Sampling Methods for Modeling Macromolecular Structure and Dynamics. *PLoS Comput Biol*, 12(4):e1004619, 2016.

[107] B K Ho and D A Agard. Probing the flexibility of large conformational changes in protein structures through local perturbations. *PLoS Comput Biol*, 5(4):e1000343, 2009.

[108] F Palazzesi, M K Prakash, M Bonomi, and A Barducci. Accuracy of current all-atom force-fields in modeling protein disordered states. *J Chem Theory Comput*, 11(1):2–7, 2015.

[109] B L de Groot, D M van Aalten, R M Scheek, A Amadei, G Vriend, and H J Berendsen. Prediction of protein conformational freedom from distance constraints. *Proteins*, 29(2):240–251, 1997.

[110] B L de Groot, G Vriend, and H J Berendsen. Conformational changes in the chaperonin GroEL: new insights into the allosteric mechanism. *J Mol Biol*, 286:1241–1249, 1999.

[111] D Seeliger, J Haas, and B L de Groot. Geometry-based sampling of conformational transitions in proteins. *Structure*, 15(11):1482–1492, 2007.

[112] S E Dobbins, V I Lesk, and M J E Sternberg. Insights into protein flexibility: The relationship between normal modes and conformational change upon protein-protein docking. *Proc Natl Acad Sci USA*, 105(30):10390–10395, 2008.

[113] D M Kruger, A Ahmed, and H Gohlke. NMSim web server: integrated approach for normal mode-based geometric simulations of biologically relevant conformational transitions in proteins. *Nucleic Acids Res*, 40:W310–W316, 2012.

[114] A Ahmed, F Rippmann, G Barnickel, and H Gohlke. A normal mode-based geometric simulation approach for exploring biologically relevant conformational transitions in proteins. *J Chem Inf Model*, 51(7):1604–1622, 2011.

[115] D J Jacobs, A J Rader, L A Kuhn, and M F Thorpe. Protein flexibility predictions using graph theory. *Proteins*, 44(2):150–165, 2001.

[116] M Totrov and R Abagyan. Flexible ligand docking to multiple receptor conformations: a practical alternative. *Curr Opin Struct Biol*, 18:178–184, 2008.

[117] D Mustard and D W Ritchie. Docking Essential Dynamics Eigenstructures. *Proteins*, 60(2):269–274, 2005.

[118] D R Weiss and M Levitt. Can morphing methods predict intermediate structures? *J Mol Biol*, 385(2):665–674, 2009.

[119] P Tompa. Intrinsically disordered proteins: a 10-year recap. *Trends Biochem Sci*, 37(12):509–516, 2012.

[120] R B Best. Computational and theoretical advances in studies of intrinsically disordered proteins. *Curr Opin Struc Biol*, 42:147–154, 2017.

[121] G Manning, D B Whyte, R Martinez, T Hunter, and S Sudarsanam. The protein kinase complement of the human genome. *Science*, 298:1912–1934, 2002.

[122] J A Endicott, M E Noble, and L N Johnson. The structural basis for control of eukaryotic protein kinases. *Annu Rev Biochem*, 81:587–613, 2012.

[123] S Muller, A Chaikuad, N S Gray, and S Knapp. The ins and outs of selective kinase inhibitor development. *Nat Chem Biol*, 11:818–821, 2015.

[124] R A Norman, D Toader, and A D Ferguson. Structural approaches to obtain kinase selectivity. *Trends Pharmacol Sci*, 33(5):273–278, 2012.

[125] A P Kornev and S S Taylor. Dynamics-Driven Allostery in Protein Kinases. *Trends Biochem Sci*, 40(11):628–647, 2015.

[126] M Huse and J Kuriyan. The conformational plasticity of protein kinases. *Cell*, 109:275–282, 2002.

[127] F De Smet, A Christopoulos, and P Carmeliet. Allosteric targeting of receptor tyrosine kinases. *Nat Biotechnol*, 32(11):1113–1120, 2014.

[128] L K Gavrin and E Saiah. Approaches to discover non-ATP site kinase inhibitors. *Med Chem Commun*, 4:41–51, 2013.

[129] A Converso, T Hartingh, R M Garbaccio, E Tasber, K Rickert, M E Fraley, Y Yan, C Kreatsoulas, S Stirdivant, B Drakas, E S Walsh, K Hamilton, C A Buser, X Mao, M T Abrams, S C Beck, W Tao, R Lobell, L Sepp-Lorenzino, J Zugay-Murphy, V Sardana, S K Munshi, S M Jezequel-Sur, P D Zuck, and G D Hartman. Development of thioquinazolinones, allosteric Chk1 kinase inhibitors. *Bioorg Med Chem Lett*, 19(4):1240–1244, 2009.

[130] D Vanderpool, T O Johnson, C Ping, S Bergqvist, G Alton, S Phonephaly, E Rui, C Luo, Y L Deng, S Grant, T Quenzer, S Margosiak, J Register, E Brown, and J Ermolieff. Characterization of the CHK1 allosteric inhibitor binding site. *Biochemistry*, 48(41):9823–9830, 2009.

[131] J Zhang, F J Adrian, W Jahnke, S W Cowan-Jacob, A G Li, R E Iacob, T Sim, J Powers, C Dierks, F Sun, G R Guo, Q Ding, B Okram, Y Choi, A Wojciechowski, X Deng, G Liu, G Fendrich, A Strauss, N Vajpai, S Grzesiek, T Tuntland, Y Liu, B Bursulaya, M Azam, P W Manley, J R Engen, G Q Daley, M Warmuth, and N S Gray. Targeting Bcr-Abl by combining allosteric with ATP-binding-site inhibitors. *Nature*, 463(7280):501–506, 2010.

[132] J Yang, N Campobasso, M P Biju, K Fisher, X Q Pan, J Cottom, S Galbraith, T Ho, H Zhang, X Hong, P Ward, G Hofmann, B Siegfried, F Zappacosta, Y Washio, P Cao, J Qu, S Bertrand, D Y Wang, M S Head, H Li, S Moores, Z Lai, K Johanson, G Burton, C Erickson-Miller, G Simpson, P Tummino, R A Copeland, and A Oliff. Discovery and characterization of a cell-permeable, small-molecule c-Abl kinase activator that binds to the myristoyl binding site. *Chem Biol*, 18(2):177–186, 2011.

[133] K M Comess, C Sun, C Abad-Zapatero, E R Goedken, R J Gum, D W Borhani, M Argiriadi, D R Groebe, Y Jia, J E Clampit, D L Haasch, H T Smith, S Wang, D Song, M L Coen, T E Cloutier, H Tang, X Cheng, C Quinn, B Liu, Z Xin, G Liu, E H Fry, V Stoll, T I Ng, D Banach, D Marcotte, D J Burns, D J Calderwood, and P J Hajduk. Discovery and characterization of non-ATP site inhibitors of the mitogen activated protein (MAP) kinases. *ACS Chem Biol*, 6(3):234–244, 2011.

[134] S Betzi, R Alam, M Martin, D J Lubbers, H Han, S R Jakkaraj, G I Georg, and E Schonbrunn. Discovery of a potential allosteric ligand binding site in CDK2. *ACS Chem Biol*, 6:492–501, 2011.

[135] V Lamba and I Ghosh. New directions in targeting protein kinases: focusing upon true allosteric and bivalent inhibitors. *Curr Pharm Des*, 18:2936–2945, 2012.

[136] Z B Hill, B G Perera, and D J Maly. A chemical genetic method for generating bivalent inhibitors of protein kinases. *J Am Chem Soc*, 131:6686–6688, 2009.

[137] M Peyressatre, C Prével, M Pellerano, and M C Morris. Targeting Cyclin-Dependent Kinases in Human Cancers: From Small Molecules to Peptide Inhibitors. *Cancers*, 7:179–237, 2015.

[138] C Barriere, D Santamaria, A Cerqueira, J Galan, A Martin, S Ortega, M Malumbres, P Dubus, and M Barbacid. Mice thrive without Cdk4 and Cdk2. *Mol Oncol*, 1(1):72–83, 2007.

[139] M K Diril, C K Ratnacaram, V C Padmakumar, T Du, M Wasser, V Coppola, L Tessarollo, and P Kaldis. Cyclin-dependent kinase 1 (Cdk1) is essential for cell division and suppression of DNA re-replication but not for liver regeneration. *Proc Natl Acad Sci USA*, 109(10):3826–3831, 2012.

[140] H M Berman, B C Narayanan, L Di Costanzo, S Dutta, S Ghosh, B P Hudson, C L Lawson, E Peisach, A Prlić, P W Rose, C Shao, H Yang, J Young, and C Zardecki. Trendspotting in the Protein Data Bank. *FEBS Lett*, 587(8):1036–1045, 2013.

[141] H L De Bondt, J Rosenblatt, J Jancarik, H D Jones, D O Morgan, and S H Kim. Crystal structure of cyclin-dependent kinase 2. *Nature*, 363(6430):595–602, 1993.

[142] A I Grigoroudis, C McInnes, P N Premnath, and G Kontopidis. Efficient soluble expression of active recombinant human cyclin A2 mediated by E. coli molecular chaperones. *Protein Expr Purif*, 113:8–16, 2015.

[143] P D Jeffrey, A A Russo, K Polyak, E Gibbs, J Hurwitz, J Massague, and N P Pavletich. Mechanism of CDK activation revealed by the structure of a cyclinA-CDK2 complex. *Nature*, 376(6538):313–320, 1995.

[144] A A Russo, P D Jeffrey, and N P Pavletich. Structural basis of cyclin-dependent kinase activation by phosphorylation. *Nat Struct Biol.*, 3(8):696–700, 1996.

[145] M C Morris, C Gondeau, J A Tainer, and G Divita. Kinetic mechanism of activation of the Cdk2/cyclin A complex. Key role of the C-lobe of the Cdk. *J Biol Chem*, 277(26):23847–23853, 2002.

[146] L Palmieri and G Rastelli. αC helix displacement as a general approach for allosteric modulation of protein kinases. *Drug Discov Today*, 18(7-8):407–414, 2013.

[147] E S Child, T Hendrychová, K McCague, A Futreal, M Otyepka, and D J Mann. A cancer-derived mutation in the PSTAIRE helix of cyclin-dependent kinase 2 alters the stability of cyclin binding. *Biochim Biophys Acta*, 1803(7):858–864, 2010.

[148] P Pisani, F Caporuscio, L Carlino, and G Rastelli. Molecular Dynamics Simulations and Classical Multidimensional Scaling Unveil New Metastable States in the Conformational Landscape of CDK2. *PLoS One*, 11(4):e0154066, 2016.

[149] J Gu and P E Bourne. Identifying allosteric fluctuation transitions between different protein conformational states as applied to Cyclin Dependent Kinase 2. *BMC Bioinformatics*, 8:45, 2007.

[150] G N Hortobagyi, S M Stemmer, H A Burris, Y S Yap, G S Sonke, S Paluch-Shimon, M Campone, K L Blackwell, F Andre, E P Winer, W Janni, S Verma, P Conte, C L Arteaga, D A Cameron, K Petrakova, L L Hart, C Villanueva, A Chan, E Jakobsen, A Nusch, O Burdaeva, E M Grischke, E Alba, E Wist, N Marschner, A M Favret, D Yardley, T Bachelot, L M Tseng, S Blau, F Xuan, F Souami, M Miller, C Germa, S Hirawat, and J O'Shaughnessy. Ribociclib as First-Line Therapy for HR-Positive, Advanced Breast Cancer. *N Engl J Med*, 375(18):1738–1748, 2016.

[151] G Rastelli, A Anighoro, M Chripkova, L Carrassa, and M Broggini. Structure-based discovery of the first allosteric inhibitors of cyclin-dependent kinase 2. *Cell Cycle*, 13(14):2296–2305, 2014.

[152] Y Hu, S Li, F Liu, L Geng, X Shu, and J Zhang. Discovery of novel nonpeptide allosteric inhibitors interrupting the interaction of CDK2/cyclin A3 by virtual screening and bioassays. *Bioorg Med Chem Lett*, 25(19):4069–4073, 2015.

[153] H Chen, Y Zhao, H Li, D Zhang, Y Huang, Q Shen, R Van Duyne, F Kashanchi, C Zeng, and S Liu. Break CDK2/Cyclin E1 interface allosterically with small peptides. *PLoS One*, 9(10):e109154, 2014.

[154] The UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res*, 43:D204–D212, 2015.

[155] N Furnham, G L Holliday, T A P de Beer, J O B Jacobsen, W R Pearson, and J M Thornton. The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes. *Nucleic Acids Res*, 42:D485–D489, 2014.

[156] M M Tirion. Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. *Phys Rev Lett*, 77(9):1905–1908, 1996.

[157] A Bakan, L M Meireles, and I Bahar. ProDy: Protein Dynamics Inferred from Theory and Experiments. *Bioinformatics*, 27(11):1575–1577, 2011.

[158] C Cortes and V Vapnik. Support-Vector Networks. *Machine Learning*, 20:273–297, 1995.

[159] E Frank, M Hall, L Trigg, G Holmes, and I H Witten. Data mining in bioinformatics using Weka. *Bioinformatics*, 20(15):2479–2481, 2004.

[160] T Joachims. Making Large-Scale SVM Learning Practical. *Advances in Kernel Methods - Support Vector Learning*, 1998.

[161] B A Kidd, D Baker, and W E Thomas. Computation of Conformational Coupling in Allosteric Proteins. *PLoS Comput Biol*, 5(8):e1000484, 2009.

[162] L A Kelley, S Mezulis, C M Yates, M N Wass, and M J Sternberg. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc*, 10(6):845–858, 2015.

[163] J Bezanson, A Edelman, S Karpinski, and V B Shah. Julia: A Fresh Approach to Numerical Computing. *SIAM Review*, 59(1):65–98, 2017.

[164] W G Touw, C Baakman, J Black, T A te Beek, E Krieger, R P Joosten, and G Vriend. A series of PDB-related databanks for everyday needs. *Nucleic Acids Res*, 43:D364–D368, 2015.

[165] D K Agrafiotis, D Bandyopadhyay, and E Yang. Stochastic proximity embedding: a simple, fast and scalable algorithm for solving the distance geometry problem. *Distance Geometry: Theory, Methods and Applications (Springer)*, 2013.

[166] A Bakan and I Bahar. The intrinsic dynamics of enzymes plays a dominant role in determining the structural changes induced upon inhibitor binding. *Proc Natl Acad Sci USA*, 106(34):14349–14354, 2009.

[167] Kabsch W. A solution for the best rotation to relate two sets of vectors. *Acta Cryst*, 32(922), 1976.

[168] R A Laskowski, M W MacArthur, D S Moss, and J M Thornton. PROCHECK: a program to check the stereochemical quality of protein structures. *J Appl Cryst*, 26:283–291, 1993.

[169] E X Esposito, D Tobi, and J D Madura. Comparative Protein Modeling. *Rev Comp Ch*, 22(2):57–168, 2006.

[170] C Atilgan, Z N Gerek, S B Ozkan, and A R Atilgan. Manipulation of conformational change in proteins by single-residue perturbations. *Biophysical J*, 99(3):933–943, 2010.

[171] M J Abraham, T Murtola, R Schulz, S Páll, J C Smith, B Hess, and E Lindahl. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1-2:19–25, 2015.

[172] G A Tribello, M Bonomi, D Branduardi, C Camilloni, and G Bussi. PLUMED 2: New feathers for an old bird. *Comput Phys Commun*, 185:604–613, 2014.

[173] B L de Groot, S Hayward, D M van Aalten, A Amadei, and H J Berendsen. Domain motions in bacteriophage T4 lysozyme: a comparison between molecular dynamics and crystallographic data. *Proteins*, 31:116–127, 1998.

[174] D Seeliger and B L de Groot. tCONCOORD-GUI: Visually Supported Conformational Sampling of Bioactive Molecules. *J Comput Chem*, 30:1160–1166, 2009.

[175] W R Pitt, R W Montalvão, and T L Blundell. Polyphony: superposition independent methods for ensemble-based drug discovery. *BMC Bioinformatics*, 15(324):1–18, 2014.

[176] M Louet, C Seifert, U Hensen, and F Grater. Dynamic Allostery of the Catabolite Activator Protein Revealed by Interatomic Forces. *PLoS Comput Biol*, 11(8):e1004358, 2015.

[177] D Kern, B F Volkman, P Luginbuhl, M J Nohaile, S Kustu, and D E Wemmer. Structure of a transiently phosphorylated switch in bacterial signal transduction. *Nature*, 402:894–898, 1999.

[178] B F Volkman, D Lipson, D E Wemmer, and D Kern. Two-state allosteric behavior in a single-domain signaling protein. *Science*, 291:2429–2433, 2001.

[179] L Yang, G Song, and R L Jernigan. How well can we understand large-scale protein motions using normal modes of elastic network models? *Biophysical J*, 93(3):920–929, 2007.

[180] D Seeliger and B L de Groot. Conformational transitions upon ligand binding: holo-structure prediction from apo conformations. *PLoS Comput Biol*, 6(1):e1000634, 2010.

[181] L E Grove, D R Hall, D Beglov, S Vajda, and D Kozakov. FTFlex: accounting for binding site flexibility to improve fragment-based identification of druggable hot spots. *Bioinformatics*, 29(9):1218–1219, 2013.

[182] M Varadi, S Kosol, P Lebrun, E Valentini, M Blackledge, A K Dunker, I C Felli, J D Forman-Kay, R W Kriwacki, R Pierattelli, J Sussman, D I Svergun, V N Uversky, M Vendruscolo, D Wishart, P E Wright, and P Tompa. pE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins. *Nucleic Acids Res*, 42:D326–D335, 2014.

[183] G R Bowman and P L Geissler. Equilibrium fluctuations of a single folded protein reveal a multitude of potential cryptic allosteric sites. *Proc Natl Acad Sci USA*, 109(29):11681–11686, 2012.

[184] P Schmidtke, A Bidon-Chanal, F J Luque, and X Barril. MDpocket: open-source cavity detection and characterization on molecular dynamics trajectories. *Bioinformatics*, 27(23):3276–3285, 2011.

[185] K Okazaki and S Takada. Dynamic energy landscape view of coupled binding and protein conformational change: induced-fit versus population-shift mechanisms. *Proc Natl Acad Sci USA*, 105(32):11182–11187, 2008.

[186] P I W de Bakker, N Furnham, T L Blundell, and M A DePristo. Conformer generation under restraints. *Curr Opin Struct Biol*, 16:160–165, 2006.

[187] C J Radoux, T S Olsson, W R Pitt, C R Groom, and T L Blundell. Identifying Interactions that Determine Fragment Binding at Protein Hotspots. *J Med Chem*, 59(9):4314–4325, 2016.

[188] M Dziubiński, P Daniluk, and B Lesyng. ResiCon: a method for the identification of dynamic domains, hinges and interfacial regions in proteins. *Bioinformatics*, 32(1):25–34, 2016.

[189] O Trott and A J Olson. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem*, 31(2):455–461, 2010.

[190] W J Allen, T E Balius, S Mukherjee, S R Brozell, D T Moustakas, P T Lang, D A Case, I D Kuntz, and R C Rizzo. DOCK 6: Impact of new features and current docking performance. *J Comput Chem*, 36(15):1132–1156, 2015.

[191] T W Backman, Y Cao, and T Girke. ChemMine tools: an online service for analyzing and clustering small molecules. *Nucleic Acids Res*, 39:W486–W491, 2011.

[192] X Wang, M Fu, J Ren, and X Qu. Evaluation of different culture conditions for high-level soluble expression of human cyclin A2 with pET vector in BL21 (DE3) and spectroscopic characterization of its inclusion body structure. *Protein Expr Purif*, 56(1):27–34, 2007.

[193] T Hrabe, Z Li, M Sedova, P Rotkiewicz, L Jaroszewski, and A Godzik. PDBFlex: exploring flexibility in protein structures. *Nucleic Acids Res*, 44:D423–D428, 2016.

[194] T Sterling and J J Irwin. ZINC 15 - Ligand Discovery for Everyone. *J Chem Inf Model*, 55(11):2324–2337, 2015.

[195] A Daina, O Michielin, and V Zoete. SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Sci Rep*, 7:42717, 2017.

[196] J Baell and M A Walters. Chemical con artists foil drug discovery. *Nature*, 513(7519):481–483, 2014.

[197] A Gaulton, A Hersey, M Nowotka, A P Bento, J Chambers, D Mendez, P Mutowo, F Atkinson, L J Bellis, E Cibrian-Uhalte, M Davies, N Dedman, A Karlsson, M P Magarinos, J P Overington, G Papadatos, I Smit, and A R Leach. The ChEMBL database in 2017. *Nucleic Acids Res*, 45:D945–D954, 2017.

[198] N R Brown, S Korolchuk, M P Martin, W A Stanley, R Moukhametzianov, M E Noble, and J A Endicott. CDK1 structures reveal conserved and unique features of the essential cell cycle CDK. *Nat Commun*, 6:6769, 2015.

[199] J Comley. TR-FRET based assays - getting better with age. *Drug Discov World*, pages 22–37, 2006.

[200] N R Brown, M E Noble, J A Endicott, E F Garman, S Wakatsuki, E Mitchell, B Rasmussen, T Hunt, and L N Johnson. The crystal structure of cyclin A. *Structure*, 3(11):1235–1247, 1995.

[201] R Honda, E D Lowe, E Dubinina, V Skamnaki, A Cook, N R Brown, and L N Johnson. The structure of cyclin E1/CDK2: implications for CDK2 activation and CDK2-independent roles. *EMBO J*, 24(3):452–463, 2005.

[202] G J Rocklin, T M Chidyausiku, I Goreshnik, A Ford, S Houliston, A Lemak, L Carter, R Ravichandran, V K Mulligan, A Chevalier, C H Arrowsmith, and D Baker. Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science*, 357(6347):168–175, 2017.

[203] P J Cock, T Antao, J T Chang, B A Chapman, C J Cox, A Dalke, I Friedberg, T Hamelryck, F Kauff, B Wilczynski, and M J de Hoon. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009.

[204] T Kluyver, B Ragan-Kelley, F Pérez, B Granger, M Bussonnier, J Frederic, K Kelley, J Hamrick, J Grout, S Corlay, P Ivanov, D Avila, S Abdalla, C Willing, and Jupyter Development Team. Jupyter Notebooks - a publishing format for reproducible computational workflows. *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, 2016.

[205] M J Gajda. hPDB - Haskell library for processing atomic biomolecular structures in protein data bank format. *BMC Res Notes*, 6:483, 2013.