In silico identification and assessment of novel allosteric protein binding sites to expand the "druggable" human proteome

This thesis is submitted to the School of Natural and Environmental Sciences, Newcastle University for the degree of Doctor of Philosophy

Francesc Sabanés Zariquiey

November 2019

Abstract

Throughout the last years there has been a considerable number of drugs that were discovered thanks to computer aided drug design (CADD) techniques. Using the 3D information, such as protein structures obtained by X-ray crystallography or nuclear magnetic resonance (NMR), it is possible to identify the binding sites and to design molecules that may specifically target these sites. This approach saves a lot of time and money, as the lead search is more accurate: less compounds need to be synthesised and tested. Although a great number of proteins have been successfully targeted with this structure-based approach, there are a lot of disease-linked proteins that have been considered “undruggable” by conventional structure-based techniques. This is mainly due to failure in detection of potential binding sites, which precludes the structure-guided design of suitable ligands.

There is the presumption that the “druggable” human proteome may be larger than previously expected. Protein structures may present multiple binding sites (allosteric and/or cryptic) that cannot be targeted by the means of conventional CADD techniques. In the past years, several novel methods have been developed to identify and/or unveil these binding hotspots. Amongst them cosolvent Molecular Dynamics (MD) simulations are increasingly popular techniques developed for prediction and characterisation of allosteric and cryptic binding sites, which can be rendered “druggable” by small molecule ligands. Despite their conceptual simplicity and effectiveness, the analysis of cosolvent MD trajectories relies on pocket volume data, which requires a high level of manual investigation and may introduce a bias. The present study focused on the development of the novel cosolvent analysis toolkit (denoted as CAT), as an open-source, freely accessible analytical tool, suitable for automated analysis of cosolvent MD trajectories. CAT is compatible with popular molecular graphics software packages such as UCSF Chimera and VMD. Using a novel hybrid empirical force field scoring function, CAT accurately ranked the dynamic interactions between the macromolecular target and cosolvent molecular probes.

Alongside the development of CAT, this work investigated the signal transducer activator of transcription 3 (STAT3) as the case study. STAT3 is among the most investigated oncogenic transcription factors, as it is highly associated with cancer initiation, progression, metastasis, chemoresistance, and immune evasion. Constitutive activation of STAT3 by mutations occurs frequently in tumour cells, and directly contributes to many malignant phenotypes. The evidence from both preclinical and clinical studies have demonstrated that STAT3 plays a critical role in several malignancies associated with poor prognosis such as glioblastoma and triple-negative breast cancer (TNBC), and STAT3 inhibitors have shown efficacy in inhibiting cancer growth and metastasis. Unfortunately, detailed structural biology studies on STAT3 as well as target-based drug discovery efforts have been hampered by difficulties in the expression and purification of the full length STAT3 and a lack of ligand-bound crystal structures. Considering these, computational methods offer an attractive strategy for the assessment of "druggability" of STAT3 dimers and allow investigations of reported activating and inhibiting STAT3 mutants at the atomistic level of detail. This work studied effects exerted by reported STAT3 mutations on the protein structure, dynamics, DNA binding and dimerisation, thus linking structure, dynamics, energetics, and the biological function. By employing a combination of equilibrium molecular dynamics (MD) and umbrella sampling (US) simulations to a series of human STAT3 dimers, which comprised wild-type protein and four mutations; the work presented herein explains the modulation of STAT3 activity by these mutations. The binding sites were mapped by the combination of MD simulations, molecular docking, and CAT analysis, and the binding mode of a clinical candidate napabucasin/BBI-608 at STAT3, which resembles the effect of D570K mutation, has been characterised.

Collectively the results of this study demonstrate the robustness of the newly developed CAT methodology and its applicability in computational studies aiming at identification of protein "hotspots" in a wide range of protein targets, including the challenging ones. This work contributes to understanding the activation/inhibition mechanism of STAT3, and it explains the molecular mechanism of STAT3 inhibition by BBI-608. Alongside the characterisation of the BBI-608 binding mode, a novel binding site amenable to bind small molecule

ligands has been discovered in this work, which may pave the way to design novel STAT3 inhibitors and to suggest new strategies for pharmacological intervention to combat cancers associated with poor prognosis. It is expected that the results presented in this dissertation will contribute to an increase of the size of the potentially “druggable” human proteome.

Declaration

The research described within this document was performed between September 2016 and October 2019 in the Computational Medicinal Chemistry Laboratories, Bedson Building, School of Natural and Environmental Sciences, Newcastle University, Newcastle-upon-Tyne, UK, NE1 7RU. This research was conducted in collaboration with scientists at the IQS School of Engineering, Grup de Química Farmacèutica, Barcelona, Spain, 08017.

All research described within this thesis is original in content and does not incorporate any material or ideas previously published or presented by other authors except where due reference is given.

No part of this thesis has been previously submitted for a degree, diploma or any other qualification at any other university.

List of publications resulting from this research

Sabanés, F; de Souza, JV; Estrada-Tejedor, R; Bronowska AK. **If you cannot win them join them: Understanding new ways to target STAT3 by Small Molecules**, *ACS Omega, 2019*

Sabanés, F; de Souza, JV; Bronowska AK., **Cosolvent Analysis Toolkit (CAT): a robust hotspot identification platform for cosolvent simulations of proteins to expand the druggable proteome**, *Scientific Reports – Just accepted*, 2019

Siguis on siguis:
allà dalt,
a la riera d'Arbúcies,
o dintre nostre.
Això va per tu

Acknowledgments

First of all, I would like to thank Dr. Agnieszka Bronowska, whom without her help I would not be writing these words. Professional and (most importantly) personally she has been the best supervisor I could have asked for.

Additionally, I would like to thank everyone in the Computational Medicinal Chemistry group for their help and friendship during these past years, specially Joao for all his advices, lunches, trips and projects that we shared together. I would like to thank also all current and past members of the group: Sylvia, Matt, Danlin, Ayaz, Rhys, Ruidi, Mete, Piotr, Weikang and Lanyu.

També vull tenir un petit record pel Dr. Roger Estrada, que sense ell no hagués descovert aquest camp que tant m'apassiona.

Evidentment no puc em oblidar de la meva familia. Als meus germans Arantxa, Xavi i Richi que probablament confiaven més en mi que jo mateix. Als seus fills: Germán, Aina, Martina i Luca, que t'arrencaven un somriure innocent quan més ho necessitaves. Però sobretot a la meva mare, amb qui sobretot en aquests últims dos anys ens hem recolzat mutuament encara que ens separessin centenars de kilòmetres. Perdre la persona que més t'entén és molt dur, sobretot en aquesta etapa, però entre tots han sapigut omplir aquest buit amb escreix.

A Lucía, por ser la persona que más ha tenido que soportarme y ayudado a ser mejor persona y compañero. Tanto en los buenos como en los malos momentos, siempre estaba ahí.

And obviously I should thank Zero, my cat. My buddy during all those writing sessions.

Abbreviations

A	A/Ala	Alanine
	ACE	Angiotensin-converting enzyme
	ADME	Absorption, distribution, metabolism and excretion
	AF-2	Activation function 2
	AMPA	α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid
	AR	Androgen receptor
	Å	Angstrom
B	BCL6	B-cell lymphoma 6
	BF-3	Binding function 3
C	C/Cys	Cysteine
	CADD	Computer Aided Drug Design
	CAT	Cosolvent Analysis Toolkit
	CC	Coiled-Coil

	CCR5	C-C chemokine receptor type 5
	CDK2	Cyclin dependent kinase 2
	CRD	Cysteine-rich domain
	CryoEM	Electron cryo-microscopy
	CV	Collective variable
D	D/Asp	Aspartic acid
	Da	Dalton
	DBD	DNA-binding domain
	DHT	Dihydrotestoterone
	DNA	Deoxyribonucleic acid
E	E/Glu	Glutamic acid
	EGFR	Epidermal growth factor receptor
	Epo	Erythropoietin
F	F/Phe	Phenylalanine

	FBDD	Fragment-based drug design
	FDA	Food and drug administration
	FGFR	Fibroblast growth factor receptor
G	G/Gly	Glycine
	GBP	Pound sterling
	GB/VI	Generalized-Born volume integral
	GBVI/WSA	Generalized-Born volume integral/weighted surface area
	GH	Growth hormone
	GPCR	G-protein coupled receptor
	GPU	Graphics processing unit
	GTP	Guanosine triphosphate
H	H/His	Histidine
	HIV	Human immunodeficency virus
	HTS	High throughput screening

I	I/Ile	Isoleucine
	IC_{50}	half maximal inhibitory concentration
	IL-6	Interleukin-6
	IMPS	Invalid metabolic panaceas
J	JAK	Janus kinase
K	K/Lys	Lysine
	kcal	kilocalories
	K_d	Dissociation constant
L	L/Leu	Leucine
	L16	Loop 16
	LBDD	Ligand-based drug design
	LBP	Ligand binding pocket
	LGA	Lamarckian genetic algorithm
	LJ	Lennard-Jones

	logP	Partition coefficient
	m/m	mass/mass
M	M/Met	Methionine
	MCSS	Multiple Copy Simultaneous Search
	MD	Molecular dynamics
	MEK	Mitogen-activated protein kinase kinase
	MicroED	Microcrystal electron diffraction
	MM	Molecular modelling
	MM-PBSA	Molecular mechanics - Poisson-Boltzmann surface area
	MoA	Mode of action
	MSCS	Multiple Solvent Crystal Structures
	MW	Molecular weight
N	N/Asn	Asparagine
	nm	nanometer

	NMA	Normal mode analysis
	NMR	Nuclear magnetic resonance
	NPT	Isothermal-isobaric ensemble
	ns	nanosecond
	NSAID	Non-steroideal antiinflammatory drug
	NSCLC	Non-small cell lung cancer
	NVT	Canonical ensemble
P	P/Pro	Proline
	PAINS	Panassay interference compounds
	PC	Principal component
	PCA	Principal component analysis
	PDB	Protein data bank
	PK	Pharmacokinetics
	PMF	Potential of mean force
	Prl	Prolactin

	ps	picosecond
	PTP1B	protein-tyrosine phosphatase 1B
Q	Q/Gln	Glutamine
	QM	Quantum mechanics
R	R/Arg	Arginine
	RBD	Ras binding domain
	RMSD	Root-mean square deviation
	RMSF	Root-mean square fluctuation
	rMUP	Recombinant major urinary protein
	rRaHBP2	Recombinant histamine binding protein
	RTK	Receptor tyrosine kinase
S	S/Ser	Serine
	SAR	Structure-activity relationship
	SASA	Solvent accesible surface area

	SBDD	Structure-based drug design
	SFK	Src family kinase
	SH2	Src homology 2
	SILCS	Site identification by ligand competitive saturation
	SQM	Semi-empirical quantum mechanics
	STAT	Signal transducer and activator of transcription
	SWISH	Sampling water interfaces through scaled hamiltonians
T	T/Thr	Threonine
	TCPTP	T-cell protein tyrosine phosphatase
	TGFa	Transforming growth factor alpha
	TIP3P	Transferable intermolecular potential with 3 points
	TNBC	Triple negative breast cancer
U	US	Umbrella sampling
V	v/v	volume/volume

	V/Val	Valine
	vdw	van der Waals
	VEGFR	Vascular endothelial growth factor receptor
	VS	Virtual screening
W	W/Trp	Tryptophan
	WHAM	Weighted histogram analysis method
	WT	Wild type
Y	Y/Tyr	Tyrosine

Table of Contents

List of figures

List of tables

Chapter 1 In the search of allosteric binding sites

1.1 Allostery

Many protein drug targets exert their activity, when a substrate binds to their distinctive active site, promoting their biological function. Classic orthosteric ligand may be compared to a competitive inhibitor: a small molecule that competes with the endogenous substrate for its occupancy at its cognate binding site, thus blocking the protein's activity (Figure 1). Limitations in applications of orthosteric ligands in clinics include a decreased efficacy due to chronic administration, limited or poor selectivity, and chemoresistance, occurring when crucial binding site residues mutate, changing the landscape of the protein and affecting the ligands binding to the protein.[1–3] These limitations may be tackled by targeting a protein of interest via allosteric regulation. A protein can be modulated by small molecules that bind at other regions of the protein (allosteric sites), either alone or in the presence of the orthosteric ligand, to stabilise either an active or inactive conformation of the system. Regarding the concept of allostery, there are two main schools of thought. The first and most classical one relies on the MWC model[4], which assumes that proteins exist in different interconvertible states in the absence of any regulator. The change between the different states is regulated by a thermodynamic equilibrium. Furthermore, ligands can bind to the receptor in either conformation, which can be altered by its affinity with the ligand. Therefore, the binding of a ligand in a state might regulate protein activity as it induces a conformational change. A more recent and updated view on allostery is the Nussinov model, which insists to put the concept of allostery in the framework of cells.[5–7] Allosteric effects are propagated through their mechanism pathway, and as a consequence, they are likely to further affect multiprotein complexes, which are shared by several pathways. These observations increase considerably the number of possible combinations that an allosteric effect can trigger. The binding of an allosteric ligand unveils a unique conformation of the protein that in essence is a new receptor that has a propensity for unique pharmacology. Allosteric ligands possess a series of advantages that overcome some of the biggest challenges in orthosteric drug design. Since allosteric ligands bind at pockets different from the protein , they

can afford higher levels of selectivity. This feature is crucial for ligands that attempt to target specific receptors that belong to a large family of proteins, such as protein kinases or G-protein coupled receptors (GPCR).[8] There has been a huge effort, albeit with poor results, to produce a series of compounds specific to these proteins, mainly due to the highly conserved orthosteric sites, and/or due to unfavourable physicochemical and drug metabolism/pharmacokinetic properties (ADME) of synthethic orthosteric ligands. Furthermore, it has been proven that many direct-acting agonists are toxic or lead to target desensitisation, internalisation, or downregulation when they are activated for prolonged periods. Allosteric ligands can reach unprecedented levels of selectivity as they target less conserved, thus more unique, binding sites at their cognate receptors.[1–3,9] Furthermore, as allosteric ligands bind to a different binding site, there is no need to design a candidate that competes with the substrate, meaning that a less potent ligand than the orthosteric substrate can show efficacy (Figure 1).

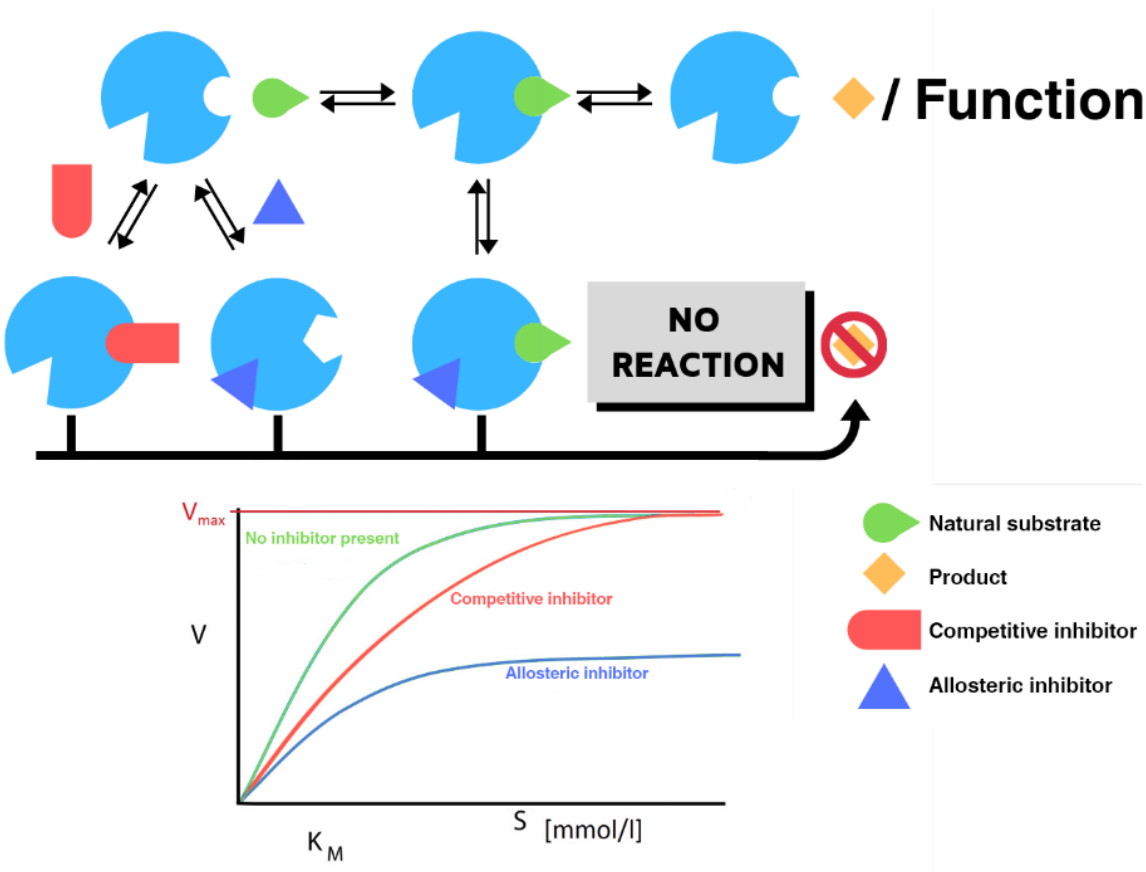


Figure 1 Graphical definition of allostery versus competitive orthosteric inhibition. Small molecule inhibitors can be divided in two groups: competitive orthosteric inhibitors and allosteric inhibitors.

A competitive inhibitor (red) binds to the proteins binding site competing with the natural substrate (green) An allosteric inhibitor (blue) binds to a distinct site on the proteins surface to prevent substrate binding (non-competitive inhibition)

Despite the numerous advantages that allostery can offer over orthosteric modulation, it is not a panacea for drug discovery, and there are many pharmacological and chemical issues to consider when developing allosteric ligands.

From the pharmacological point of view, allosteric ligands could lead to adverse effects, as they might trigger homo- and heterodimer formations for multimeric proteins, with unnecessary and/or unknown physiological responses. Furthermore, allosteric modulation could induce the formation of different oligomeric species, leading to a loss of the orthosteric function of the protein[10,11].

Success stories regarding allosteric sites targeting include Maraviroc, an allosteric modulator of CCR5 chemokine receptor that helps tackling the HIV infection[12]. Benzodiazepines have been highly successful therapeutics that allosterically regulate ion channels[13,14] as well as the AMPA receptors[15–17]. Trametinib is another allosteric inhibitor for kinases MEK1 and MEK2 that was approved by the Food and drug administration (FDA) in 2013.[18,19]

These are numerous reasons for growing interest in the search of allosteric ligands. The discovery of a new allosteric site could bring back to life the interest in targets that previously were considered “undrugabble”. Discovery of new binding sites means a whole new series of compounds for these targets to be discovered.

1.2 Computer aided drug design

The main principle of drug design is the assumption that drug activity is exerted through the binding to the pocket of a macromolecular target. Chemical and geometric/shape complementarity between the ligand and the binding site is essential for a successful binding event. The interaction between a ligand and its target is usually driven by non-covalent interactions, such as hydrogen-bonds and aromatic π-π stacking. The receptors binding site can have a hydrophobic or

hydrophilic character, depending on the residues that form them. For example, a group of non-polar amino acids like alanine, leucine and/or valine will form a hydrophobic pocket.[20,21] An example of this feature would be the comparison between the recombinant mouse major urinary protein (rMUP) and the recombinant histamine binding protein (rRaHBP2), both proteins with similar folds but different binding signatures.[22] Both systems form part of the lipocalin family and present a similar entropy of binding, but while rMUP binds small hydrophobic ligands, rRaHBP2 is a “hydrophilic” binder with high affinity for histamine and related amines. Nevertheless, binding sites tend to be hydrophobic.However, even if a compound binds well to its target, i.e. with high affinity, it does not necessarily mean that it is a good drug candidate. The drug must be transported from the site of administration (oral, intravenous, etc.) to their target. For intracellular targets, this process would involve passing through cell membranes, either via pressure (diffusion) or active transport. Once inside the cell, the inhibitor must reach its target and later be metabolised and excreted. Therefore, properties like solubility or the partition coefficient (logP) are fundamental in the small molecule drug development.[20,21] Furthermore an additional set of features is recommended to be met in order to consider a molecule a drug. Known as Lipinski’s rule of five, an orally available drug should have no more than one violation to the following criteria: no more than 5 hydrogen bond donors, no more than 10 hydrogen bond acceptors, a molecular weight (MW) less than 500 daltons, and a logP value that does not exceed 5. These criteria should be taken more into account as a guide rather than a rule.

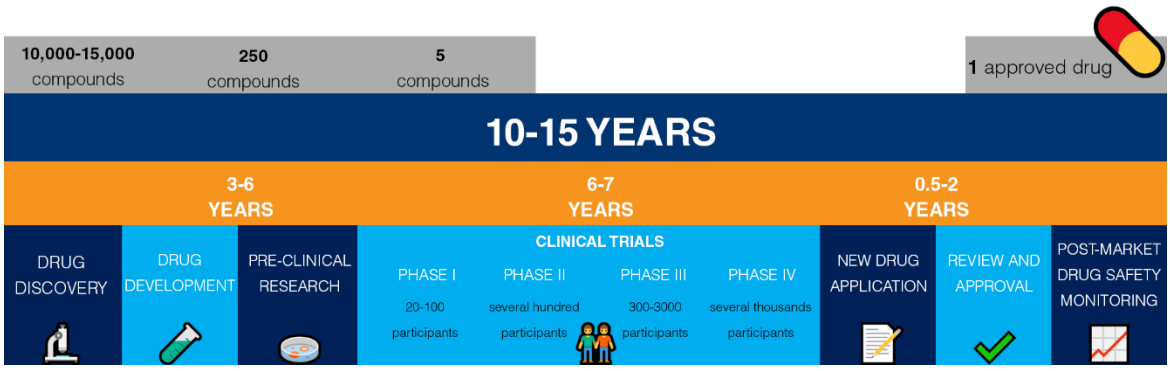


Figure 2 Drug discovery pipeline

The drug discovery and development is a long and costly process. A drug, in order to get approved, must fulfil two main requirements: produce the desired response (efficacy) with minimal side-effects (safety), and attempt to be better than existing therapies. In the drug discovery pipeline (Figure 2), after target identification and validation, the first two steps in many drug discovery programs consist in the identification of hits and lead molecules. The hits are compounds with some reproducible biological activity of interest. Lead series are usually the improvement of the hit. They comprise a set of related molecules that share some common structural features and show better or worse activity than the initial hit along with an improvement on its drug-like properties such as MW and logP.

The launch of a new drug to the market costs between 800 and 1600 million GBP and it takes between 10 to 15 years to develop.[23] In a classical pipeline, 1 out of 40,000 hits will become a commercial drug. Thanks to CADD techniques, nowadays 1 out of 10,000 hits end up in the market.[23]

Molecular modelling is a field of research focused on the application of fundamental laws of physics and chemistry to the study of molecules and biological macromolecules. Molecular modelling techniques have been developed as a product of, or in conjunction with, some new technological advance such as combinatorial chemistry, high throughput screening (HTS), or fast graphical processing units (GPU). In the case of drug discovery, the principal aim is to create models and simulations that can predict the properties of molecules and their interactions. A correct implementation of these techniques means a considerable saving of time and money and an increase of the successful rate on the development of a new drug. In some cases there is evidence that the use of computer aided drug design (CADD) techniques led to an increase of the hit rates in comparison to HTS.[24–26] Furthermore, CADD has already been used in the discovery of several compounds that have passed clinical trials and become novel therapeutics of a variety of diseases. Examples include the following: carbonic anhydrase inhibitor dorzolamide, approved in 1995[27]; the angiotensin-converting enzyme (ACE) inhibitor captopril, approved in 1981 as an antihypertensive drug[28]; three therapeutics for the treatment of human immunodeficiency virus (HIV): saquinavir (approved in 1995), ritonavir and

indinavir (both approved in 1996)[25]; and tirofiban, a fibrinogen antagonist approved in 1998[29]. More recently HIV integrase inhibitor raltegravir[30] and human renin inhibitor aliskiren[31] (both approved in 2007) were discovered through the means of CADD. Nevertheless, CADD is still not fully integrated in the drug discovery pipeline as most approved drugs did not require crucial involvement of CADD.

CADD techniques can be based in two general types of approach, applied independently or in conjunction (Figure 3). Structure based drug design (SBDD) analyses the three-dimensional structures of a given target of interest, usually a protein, with the objective to identify potential binding sites and/or interactions that are key for its respective biological functions. Ligand-based drug design (LBDD) relies on known ligands for a target, in order to establish a structure-activity relationship (SAR) between the physicochemical properties and their activity. This information is then used to optimise the already known drugs or ease the path to a new set of ligands with improved activity.[32]

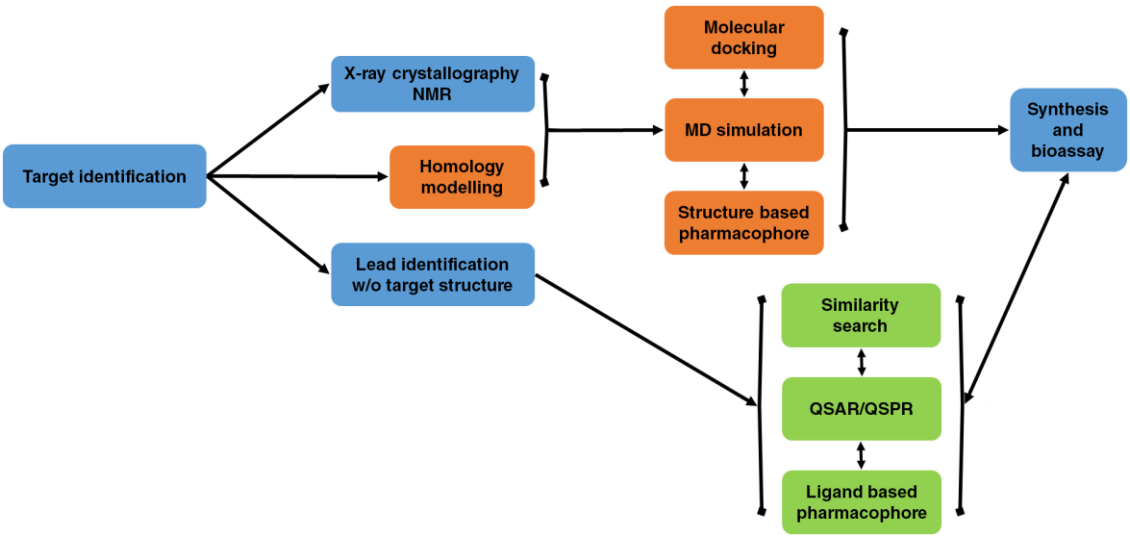


Figure 3 Basic CADD workflow in drug discovery. Wet-lab methods are coloured blue, SBDD techniques in orange and LBDD ones in green[32]**.**

1.2.1 Predicting allosteric sites

Allostery is a very promising phenomenon for drug design, especially in the case of proteins for which the design of orthosteric inhibitors has failed. The problem relies on the correct identification of the aforementioned allosteric sites. These are usually determined by X-ray crystallography: if the target of interest can be crystallised in a *holo* state (ligand-bound), then the binding site can be easily characterised. With a known binding site, usually a pocket of a cleft, known, it is relatively straightforward to design novel ligands through the means of SBDD techniques. For targets with no reported allosteric sites, the situation is more challenging. New allosteric ligands cannot be designed if there is no binding site mapped.

To facilitate the mapping of potentially functional allosteric binding sites, a series of techniques have been developed in the last years. Some of these techniques are outlined below.

1.2.1.1 Experimental prediction methods

Allosteric sites have been identified mainly by the means of high throughput screening (HTS), which means the screening of thousands or even hundreds of thousands of ligands for a given protein target. Techniques based on solution Nuclear Magnetic Resonance (NMR) or X-ray crystallography are research-favorites due to their ability to detect ligand binding, even at low affinity levels (K_d values up to 10 mM). In the case of X-ray crystallography, a handful of information is provided, since it not only reveals the location of the binding site, but also which are the specific protein-ligand interactions. Once a binding site is identified, the hit-to-lead process follows to design a series of suitable candidates. Although the outcomes of X-ray crystallography are very enlightening for the determination of binding sites, there are many limitations related to it. In many cases, due to the protein's behaviour or any other factors such as expression or purification, it is close to impossible to crystallise the product of interest. Crystal growth may be a slow and tedious process, and even after the protein has been successfully crystallised, there is no guarantee that the particular crystal form will be suitable for X-ray diffraction.[33] Nowadays, the use of novel and promising techniques such as cryogenic electron microscopy (CryoEM), and microcrystal electron diffraction

(MicroED)[34], help to address the need for a fast and reliable structure determination. CryoEM changed completely the structure determination game, providing an effective and fast tool to determine protein targets that in a previous time were near to impossible to crystallise.[33,35] The breakthrough of MicroED opened the door for the development of a high-throughput MicroED screening, in were hundreds of datasets can be collected in a short period of time.[36]

Techniques such as Multiple Solvent Crystal Structures (MSCS),[37] alanine scanning[38], and structure activity relationship by nuclear magnetic resonance (SAR by NMR)[39] have been employed to identify 'hotspots' in a number of proteins. While MSCS has lost popularity in the last years[40], SAR by NMR has facilitated the development of several compounds and it continues to be a popular technique for fragment based drug design[41–46].

For an identification of a binding site with a higher level of confidence through this method, screening compounds would require of a certain size (MM>200 Da) as well as complementarity of shape and pharmacophoric interactions[47,48]. Furthermore, the observation of molecule binding to a pocket does not necessarily mean that this pocket would have the desirable properties, such as size or presence of residues that establish polar interactions with the ligand (especially H-bonds), to develop a potential drug based on this site.[49] Following this idea, Wood and coworkers developed FragLites.[50]

This method claims to find potential interacting binding sites through X-ray crystallisation with a library of designed small halogenated compounds called FragLites. These are defined as small (≤13 heavy atoms) compounds bearing a pharmacophore doublet (combination of two functionalities capable of forming polar protein-ligand interactions, especially hydrogen bonds) and a heavy halogen atom because of their minimal size, maximal simplicity, and high visibility in X-ray crystallography due to anomalous scattering of the halogen atom. These ligand features give a degree of aqueous solubility that allows them to be used at high concentrations in crystallographic and other assay conditions.

The variation from the small, low affinity ligands to FragLites helps to map our targets of interest with more insight. Nevertheless, there is the possibility that a

protein might not crystallise. In those cases, the use of computational tools seems more feasible for the identification of allosteric hotspots.

1.2.1.2 Computational prediction methods

Although the experimental methods mentioned before yield a considerable success in the identification of allosteric sites, the toll to pay is high. X-ray crystallography can be a “Russian roulette”: determining the optimal conditions for protein crystallisation can take a considerable amount of time and resources. Furthermore, for protein crystallisation it is required to have the latter expressed (bought or expressed in-house) along with the synthesised or purchased screening compounds. Approaches involving machine learning[6] and Multiple Copy Simultaneous Search (MCSS)[51], a method that involves a fragment-based design approach to identify energetically favourable positions in a pre-specified binding site of interest[52], among others. Methods involving machine learning rely on experimental data, i.e. cryptic pockets solved by X-ray crystallography, whose number is very limited[53]. In MCSS probes do not interact with one another, which results in the loss of any possible cooperativity in their binding. Another limitation lies in the static structure of the protein target analysed: any ligand-induced conformational changes cannot be observed, which precludes its applicability to identification of cryptic and transient pockets. MCSS is a very “rigid” method that neglects protein flexibility.

Nevertheless, via the means of computational tools one can study the behaviour of a protein and even identify new potential binding sites in a shorter time. Through last years, thanks to the increasing interest in the matter, a series of computational tools and techniques have been developed to identify allosteric sites, therefore paving the way for the computational discovery of novel binding sites (Table 1).

Table 1 Some of the most widely used allosteric pocket detection webservers that are nowadays available

Name	Reference	Method	Web server available
FTMap	54	**Fast fourier transformation**	https://ftmap.bu.edu/login.php
Cryptosite	55	**Machine learning**	https://modbase.compbio.ucsf.edu/cryptosite/
AlloPred	56	**Normal Mode Analysis + Machine Learning**	http://www.sbg.bio.ic.ac.uk/allopred/home
AllosMod	57	**Molecular Dynamics**	http://modbase.compbio.ucsf.edu/allosmod
PARS	58,59	**Normal Mode Analysis**	http://bioinf.uab.cat/pars
SPACER	60,61	**Normal Mode Analysis**	http://allostery.bii.a-star.edu.sg
fpocket	62,63	**Voronoi tessellation**	http://bioserv.rpbs.univ-paris-diderot.fr/services/fpocket/

Webservers are the most popular among medicinal chemists. Due to their ease of use and fast production of results, researchers use them to identify potential binding hotspots. In pocket detection webservers the user inputs a PDB structure and waits a short amount of time (minutes to days) to obtain results. One of the most popular pocket detection webservers is the FTMap[64]. It employs a fast, easy to use method based on the sampling of several probe molecules on a densely space grid. It uses sixteen different probe molecules which include ethanol, isopropanol, isobutanol, acetone, acetaldehyde, dimethyl ether, cyclohexane, ethane, acetonitrile, urea, methylamine, phenol, benzaldehyde, benzene, acetamide, and N,N-dimethylformamide. This method is very fast because the sampling is achieved by an energy function that is evaluated using a fast-Fourier transform. FTMap's energy function incorporates cavity terms to reward

hydrophobic enclosure and a statistical, knowledge-based, pair-wise potential to account for solvation effects.[54]

There are other similar tools that share comparable goals and/or methods available such as Cryptosite and Allosite (Table 1). Even if all these methods predict allosteric sites based on protein structure, the underlying approach varies from the normal mode analysis (NMA) employed in tools such as SPACER and AlloPred, to Machine Learning methods in AlloSite and CryptoSite. These tools attempt to identify different features, such as structural fluctuations of the protein, the effect of different perturbations in the active site or at other specific positions or using sequence and structural features. Though webservers achieve a remarkable agreement with experimental data[54,64], they present some caveats. Mainly, the lack of a longer sampling through dynamics affecting the overall cleft formation, which restricts its ability to identify new cryptic binding sites. Neglecting protein dynamics and flexibility does not allow the user to discover the conformational changes of a protein that could lead to the identification of novel binding hotspots.

One way to overcome the challenge presented by rigid of crystal structures is to employ molecular dynamics (MD) simulations. Briefly, MD consists of calculating the trajectory of a system by the application of Newtonian mechanics. Via the evolution through the time of the system of interest, its intrinsic dynamics and conformational changes can be assessed. This allows for the identification of binding sites previously hidden in the obtained crystal structure.

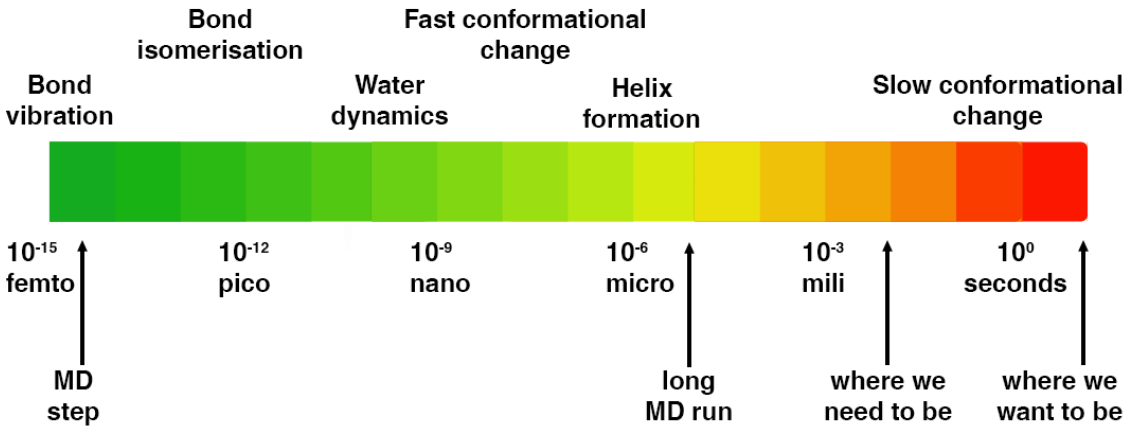


Figure 4 Nowadays proteins can be simulated for hundreds of nanosecond or several microseconds but allosteric conformational changes tend to happen around the millisecond timescale, where most computers struggle to sample in a considerable amount of time.

To observe a pocket opening, a conformational change needs to occur: such conformational changes often happen in the millisecond time scales (Figure 4). The computer power required for simulating these time scales is too demanding. Furthermore, biologically relevant events tend to present rough energy landscapes, with many local minima separated by high energy barriers. An unbiased MD simulation may easily get stuck in a local minimum that is non-functional or irrelevant for the binding event. Recent studies have demonstrated that, in long simulations, proteins can get trapped in irrelevant conformations without returning to the original, functionally relevant conformation. It is quite common that a protein remains in a local energy minimum for a large fraction of a simulation time, and therefore the sampling process is inefficient[65–67].

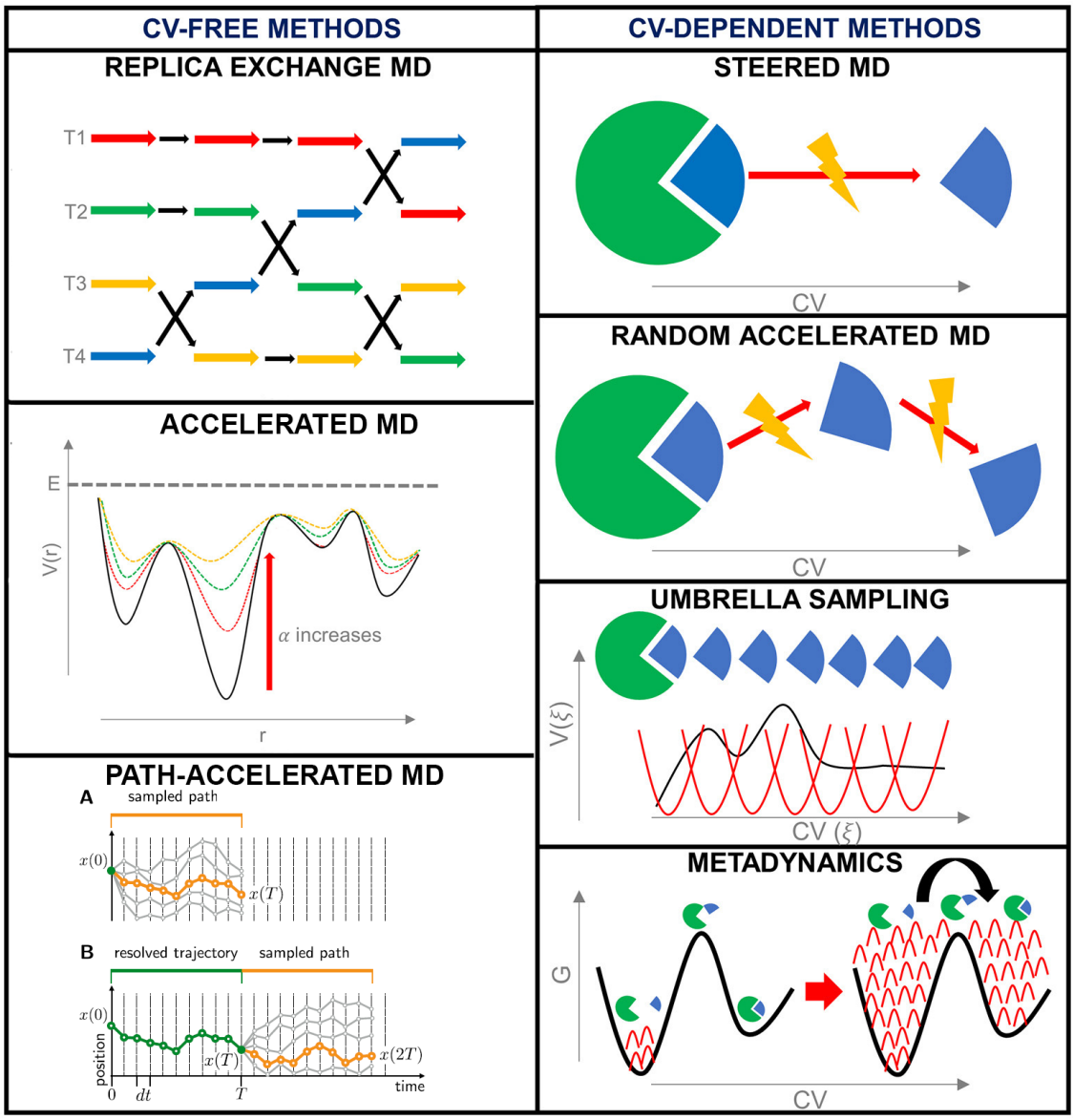


Figure 5 Summary of some of the most relevant enhanced sampling techniques[67,68]

One strategy to overcome this issue is the use of enhanced sampling techniques. These methods add a bias force/potential to the system to overcome the energy barrier that grounds the system in a local minimum, providing an acceleration in conformational sampling, and therefore allow to study processes such as ligand binding or unbinding. Enhanced sampling techniques consist of methods that make use of collective variables to introduce the bias, and methods that do not, such as replica-exchange MD[69]. A collective variable (CV) represents degrees of freedom of interest or “reaction coordinates” of the system under investigation. CV dependent enhanced sampling methods add a bias along the CVs during the simulation to observe the process of interest. This process will reduce the energy

penalty required to sample and/or observe a conformational change or a binding/unbinding event (Figure 5). Some of the most commonly used enhanced sampling techniques would be replica exchange MD[70], umbrella sampling[71] or metadynamics[72]. Especially the last two methods are used for the prediction of binding or unbinding events in ligand-protein complexes.

The main issue regarding the use of enhanced sampling techniques to identify allosteric binding sites is that, since these hotspots are unknown, one does not know what bias to apply. In most cases there, is no evidence of the conformational changes occurring and therefore the use of enhanced sampling might not be the best technique to use in the studies of allosteric events. Nevertheless, a series of collective variables for the exploration of protein druggability have been proposed with promising results. This approach, that goues under the name JEDI (Just Exploring the Druggability at Protein Interfaces) features a druggability potential that is made of a combination of different empirical descriptors[73]

1.3 Cosolvent MD

To avoid the bias of enhanced sampling, an alternative approach was proposed to overcome the problem of trapping in a local minimum: namely, the cosolvent molecular dynamics. In the cosolvent MD, the protein of interest is simulated in a mixture of water and small molecule drug-like probes (cosolvents). The competition of the probes with water to interact with different protein regions helps to map potential allosteric binding sites (Figure 6). Binding of a cosolvent probe to the protein may induce conformational changes in the protein in a relatively short time scale, thus overcoming one of the aforementioned issues. Also, this alternative mapping strategy might be able to unveil cryptic sites that would not be discovered by conventional approaches[74,75]. Due to the novelty of the method there is still no evidence in the literature that relates to the discovery of novel allosteric or cryptic binding sites with this technique. Nevertheless, different cosolvent MD approaches have been benchmarked with successful results[74], indicating that this technique is expected to be successful.

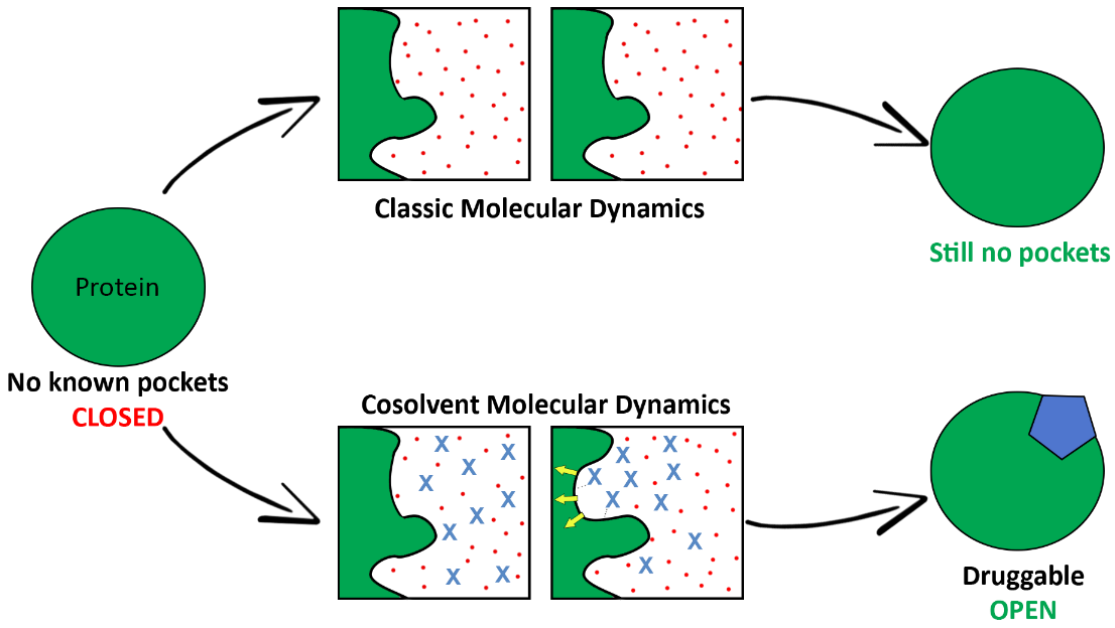


Figure 6 Small molecular fragments could have a different with the protein than water, that can lead to the formation of new binding hotspots unvisited in classical MD conditions

1.3.1 Cosolvent methods today

As with other methods, cosolvent MD can have many flavours, starting from the simulation conditions to the scoring energy function that identifies and ranks the potential hotspots. As it is a relatively novel method, there is still no consensus regarding the best simulation conditions for it (Table 2).

An important factor to consider while carrying out cosolvent MD is the appropriate simulation time. A lot of groups using cosolvent MD choose to perform many replicas of short MD (10-30 ns), rather than performing longer (100-1000 ns) simulations. Relatively short simulations are chosen to avoid the phase separation between water and more hydrophobic cosolvent probes, which would lead to unrealistic results. The phase separation problem is related to cosolvent concentration: first cosolvent MD attempts were performed at high probe concentration (50% v/v) and suffered from a prompt phase separation. Although it is important to maximise the number of probes interacting with the protein, oversaturation of probe molecules in the system might cause clustering and poor sampling. It is recommended to have fewer probes that freely interact with the

protein than a higher concentration that may induce artefacts. If, a higher concentration is desired, a way to avoid probe clustering and other artefacts is to a repulsion potential to the probes, similar to what MacKerell's group does in SILCS for benzene.[75]

Another important consideration in cosolvent MD is the probe selection. Currently, there is a consensus regarding this aspect. Most cosolvent techniques use a series of different probes, individually tested, which share a number of features: they should be small drug-like fragments and present a range of different functional groups that comprise different degrees of polarity, aromaticity, charge and shape diversity. The main idea is to use molecules that achieve the same kind of interactions potential drugs could. With the use of small drug-like fragments it is easier to extensively map the target of interest, as small molecules could interact with different pockets from a potential binding region. By pooling a different range of chemical features in the tested cosolvents it would be possible to identify which pockets would be more prone for one or another chemical entity in the final drug to be designed. The use of Mackerell's SILCS[75] has led to the design of novel ligands in proteins such as Mcl-1 or the B-cell lymphoma 6 (BCL6) BTB domain (BCL6BTB)[76,77]. Cosolvent MD does not only help to identify new binding sites, but also give a head start on the structure-based ligand design.

Considering the improvements and evaluations of the existing cosolvent MD methods, there is a consensus regarding the simulation conditions. Simulations should not be long to avoid unrealistic phase separation. A range of 10-50 ns seems to be the optimal one to obtain enough sampling of the tested system. Mixture concentration should not exceed 10%, especially in larger systems that would require a larger number of probe molecules. It is preferential to perform several short MD simulations than a long one. Cosolvent probes must be small-sized (fragments), drug-like, and range a broad spectrum of polarities. Mixture of several cosolvents with water is still not well established and more validation is required.

Table 2 Some of the established cosolvent MD techniques used to identify hotspots and binding sites on protein surfaces

Developer (method)	Cosolvent probes	Proteins
Barril (MDmix)[78,79]	Isopropanol, ethanol, acetanitrile, methanol, acetamide	Thermolysin, p53, elastase. MDM2, LFA-1/ICAM-1, PTP1B, p38 MAPK, AR, HEWL, Hsp90, HIVp
MacKerell (SILCS)[75,80]	Benzene, propane, water (as a hydrogen-bonding probe), acetonitrile, methanol, formamide, acetaldehyde, methylammonium, acetate, imidazole	BCL-6, trypsin, a-thrombin, HIVp, FKBP, Fxa, NadD, Rnase A, IL-2, p38 MAPK, DHFR, FGFr1 kinase, adenosine deaminase, ERα, AmpC, β-lactamase, T4-L99A, AR, PPARγ, mGluR5, β2AR
Carlson (MixMD)[74,81,82]	Acetonitrile, isopropanol, pyrimidine, imidaole, N-methylacetamide, acetate, methylammonium	HEWL, elastase, p53, Rnase A, thermolysin, HIVp, ABL kinase, AR, CHK1 kinase, glucokinase, PDK1 kinase, PTP1B, farnesyl pyrophosphate synthase
Gervasio (SWISH)[83,84]	Benzene, imidazole, indole, pyrimidine, pyridine, tetrahydropyran	TEM-1, IL2, PLK1, NPC2, p38α, LfrR, hPNMT

Despite the considerable progress in the development of cosolvent techniques achieved in the past years, it is still challenging for the common user to employ these methods. Although an experienced computational chemist may conduct cosolvent MD with no difficulties, they may need to use some of the scoring and analysis methods reported in the literature to evaluate the simulations. Sadly, the codes for these analysis tools are usually not available and when they are, one

could face several unexpected difficulties upon installation of the code. In some cases, the tool is included as a plugin from another molecular visualisation suite such as PyMol,[81] forcing the user to perform the analysis with this specific software.

Chapter 2 Signal transducer and activator of transcription 3

Signal transducer and activator of transcription 3 (STAT3) protein is a transcription factor with the ability to transduce signals from the cell membrane to the nucleus to activate gene transcription, thus bypassing the involvement of secondary messengers. STAT3 over-expression or inhibition plays an important role in processes such as inflammation, cellular proliferation, survival, apoptosis, transformation, angiogenesis, invasion and metastasis of cancer[85,86].

2.1 Influence of STAT3 on biological functions

In its resting state, STAT3 is predominantly localised in the cytosol. STAT3 is activated in response to ligand binding to cytokine receptors, including growth hormone (GH), prolactin (Prl), and erythropoietin (Epo), as well as several growth factor receptors (EGF, insulin, IL-6 and others). When STAT3 is activated, it translocates to the nucleus and regulates the expression of certain genes. The activation is rapid and transient under normal conditions, but in most cancers STAT3 is excessively activated and phosphorylated. Although STAT3 monomers tend to be located in the cytoplasm, they can also be found in the mitochondria.[87] Constitutive STAT3 activation results in dysregulation of cell cycle control and apoptosis genes. STAT3 has been shown to be constitutively activated or overexpressed in breast, lung, prostate, ovarian, colon, gastric and head and neck cancers as well as melanoma, leukaemia, multiple myeloma and lymphoma[88] (Table 3).

STAT3 activation confers resistance to some conventional therapies that promote apoptosis to eliminate tumour cells. In fact, STAT3 drives the expression of proliferation and survival genes, like *c-my, bcl-XL* and *mcl-1*[89–91].STAT3's ability to inhibit inflammation has been related to tumour development. The blocking of STAT3 signalling in tumour cells leads to the production of inflammatory signals, which in turn activates innate immune cells against tumour cells[85,86].

Blocking of STAT3 signalling leads to apoptosis of tumour cells, it also prevents the transformation of normal into tumour cells. Therefore, STAT3 is an attractive therapeutic target due to its over-expression in tumour cells and the fact that it can regulate the expression of a number of genes involved in oncogenesis.

Table 3 Status of STAT3 in various cancers

Cancers Characterized by Elevated STAT3 Expression or Activity	Poor Prognosis Linked to High STAT3 Levels	Upstream/Downstrem Abnormalities of STAT3 Signaling	Xenograft Models Responisve to Inhibition of STAT3
Leukemia	Renal cell carcinoma	Elevated EFGR expression	Head and neck squamous cell carcinoma
Lymphomas	Colorectal cancer	Constitutively activated EGFR-RTK	Gliobastoma
Multiple myeloma	Ovarian carcinoma	Overexpression of SFKs	Myeloproliferative neoplasms
Breast cancer	Gastric carcinoma	Hyperactivated JAKs	Renal cell carcinoma
Prostate carcinoma	Intestinal-type gastric adenocarcinoma	Elevated TGFa/IL-6	Breast cancer
Lung cancer (non-small-cell)	Cervical squamous-cell carcinoma	-	Lung adenocarcinoma
Renal cell carcinonma lung cancer	Osteosarcoma	-	Acute lymphoblastic leukemia
Hepatocellular carcinoma	Epithelial ovarian carcinoma	-	-
Cholangiocarcinoma	-	-	-
Ovarian carcinoma	-	-	-
Pancreatic adenocarcinoma	-	-	-
Melanoma	-	-	-
Head and neck squamous cell carcinoma	-	-	-

2.2 STAT family

Seven mammalian STAT family members have been identified: STAT1, STAT2, STAT3, STAT4, STAT5 (STAT5A and STAT5B) and STAT6 (Table 4). STAT2, STAT4 and STAT6 are only activated in normal human cells. However, STAT1, STAT3 and STAT5 play an important role in cancer development. STAT1 acts as tumour suppressor, while STAT3 and STAT5 act as oncogenes.[85,86] The main difference between STAT isoforms is their binding sequence specificity.[92]

2.3 Structural features of the STAT family

STAT proteins have a modular structure that include six different domains: N-terminal coiled-coil domain, DNA binding domain,linker-domain, SH2 domain, and C-terminal transactivation domain (Figure 7).[93] Each of these domains are important for the physiological functions of STAT proteins.[94] The N-terminal domain is involved in STAT dimerisation and tetramerisation. While a STAT dimer is required to bind to DNA, STATs tetramerisation contributes to stability of that binding by interaction with low-affinity STAT binding sites and therefore increasing transcriptional activity.[95] The DNA-binding domain (residues 320-494 in human STAT3) forms complexes between STAT and DNA.[96] The DNA-binding domain binds to DNA as a homodimer, adopting an immunoglobulin-fold structure. Between residues 500-585 (493-583 in human STAT3) there is an α-helix linker domain followed by a SH2-domain[97], which spans between residues 600-700 (583-688 in human STAT3). The SH2-domain is essential for the binding of STATs to phosphorylated receptors and for the dimerisation between two activated STAT monomers. The dimerisation is enhanced by phospho-tyrosine/SH2-domain interactions. The C-terminal transactivation domain (723-770 in human STAT3) is natively unfolded and forms structure only upon binding with interacting partners and is involved in communication with transcriptional complexes.[97] It contains two residues crucial for STAT activation: Y705 and S727. These are essential for the activation and dimerisation of STATs.

Table 4 Comparison of the human STAT family members

STAT Isoform	Structural features	Amino-acid stretch	pTyr	pSer	STAT Isoform	Structural features	Amino-acid stretch	pTyr	pSer
STAT1	N-terminal domain	1-136	Y701	S727	**STAT5A**	N-terminal domain	1-144	Y694	S726
	Coiled-coil domain	136-315				Coiled-coil domain	145-331		
	DNA-binding domain	316-487				DNA-binding domain	332-497		
	Linker domain	488-576				Linker domain	498-572		
	SH2 domain	577-682				SH2 domain	592-684		
	Transactivation domain	712-750				Transactivation domain	706-794		
STAT2	N-terminal domain	1-138	Y690	-	**STAT5B**	N-terminal domain	1-144	Y699	S731
	Coiled-coil domain	139-315				Coiled-coil domain	145-331		
	DNA-binding domain	316-485				DNA-binding domain	332-497		
	Linker domain	486-574				Linker domain	498-572		
	SH2 domain	575-680				SH2 domain	592-684		
	Transactivation domain	698-851				Transactivation domain	711-787		
STAT3	N-terminal domain	1-136	Y705	S727	**STAT6**	N-terminal domain	1-123	Y641	-
	Coiled-coil domain	137-319				Coiled-coil domain	124-272		
	DNA-binding domain	320-493				DNA-binding domain	273-441		
	Linker domain	494-582				Linker domain	442-591		
	SH2 domain	583-688				SH2 domain	592-685		
	Transactivation domain	723-770				Transactivation domain	711-787		
STAT4	N-terminal domain	1-136	Y693	S721					
	Coiled-coil domain	137-315							
	DNA-binding domain	316-483							
	Linker domain	484-571							
	SH2 domain	572-677							
	Transactivation domain	705-748							

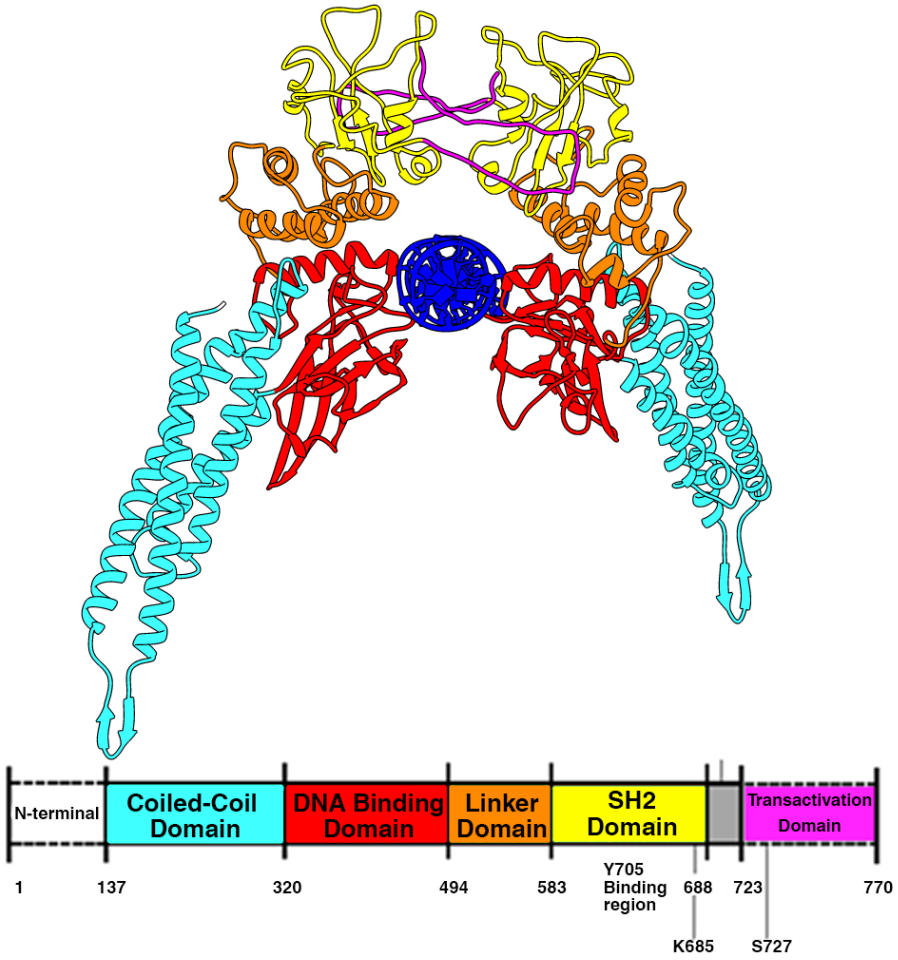


Figure 7 Crystal of STAT3-DNA complex (PDB code: 1BG1) displayed in cartoon representation. Different domains are colour coded. DNA duplex (blue) is located between the two monomers

2.4 STAT3 pathway and activation

STAT3 is involved in the Janus Kinase (JAK) pathway. It can be activated via the tyrosine phosphorylation cascade after ligand binding and stimulation of the cytokine receptor-kinase complex and growth factor complex like epidermal growth factor receptors (EGFRs), interleukin-6 (IL-6), fibroblast growth factor receptors (FGFRs), vascular endothelial growth factor receptors (VEGFRs) and more[98] (Figure 8). These receptors will induce the phosphorylation of tyrosine residues on specific sites at the cytoplasmic domain of the receptors. STAT3 is phosphorylated at two sites, Y705 and S727. The phosphorylation of STAT3 induces its dimerisation via (pTyr)-SH2 domain interactions. Next, the homodimer

translocates to the nucleus and activates specific target genes that promote the transcription of DNA.

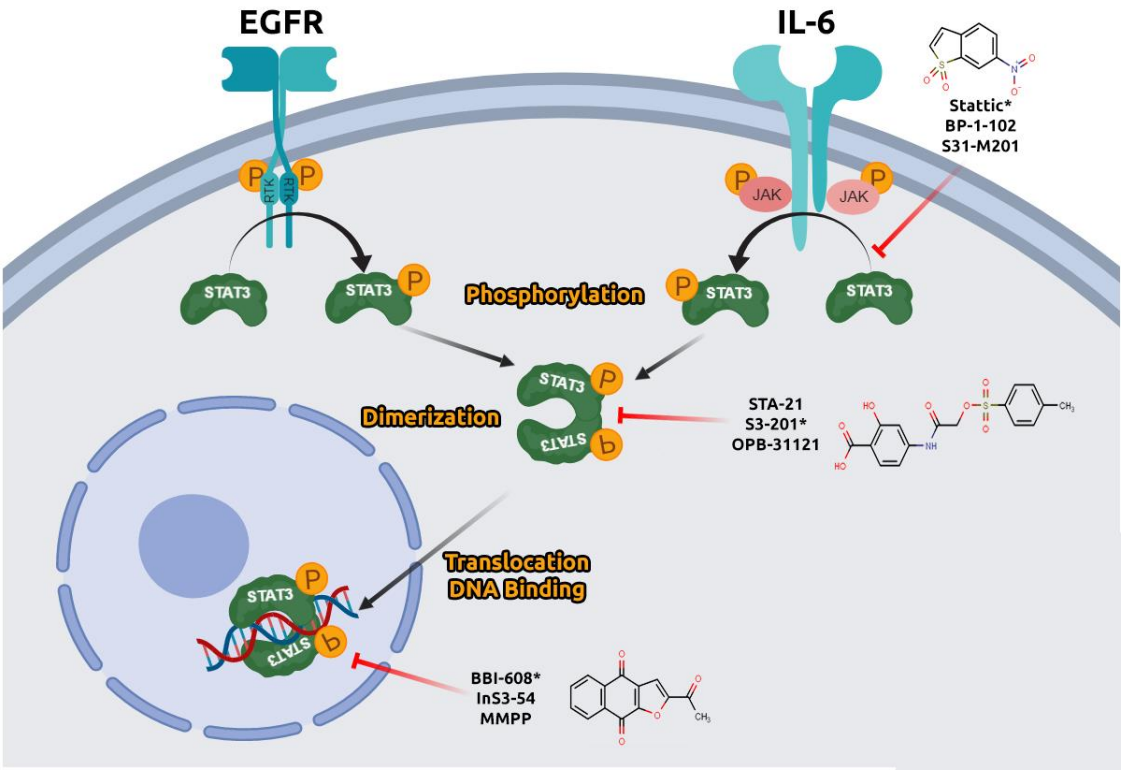


Figure 8 Graphical representation of the STAT3 pathway and small molecule inhibitors proposed to act at different stages

The formation of the dimer is crucial for STAT3 activation. Previously it was thought that by suppressing the phosphorylation of STAT3 the pathway would be switched off, but recently it has been described that unphosphorylated STAT3 is able to form homodimers, via reciprocal interaction of their SH2 domains and translocation to the nucleus, where they bind to specific DNA sequences.[99]

2.5 Crystal structures of STAT3

For the structure-guided development of new STAT3 inhibitors, it is required to gather information about the protein and more specifically the binding sites. Therefore, it is crucial to access the crystal structure for structure-guided studies including drug design. Currently, there are only five crystal structures of STAT3 deposited in the Protein Data Bank:[100]

- **1BG1:**[101] Phosphorylated STAT3 homodimer bound to DNA (resolution: 2.25 Å) (organism: Mus musculus) (residues 127-722) (Figure 9A)
- **3CWG:**[102] Unphosphorylated STAT3 core fragment (resolution: 3.05 Å) (organism :Mus musculus) (residues 127-688) (Figure 9B)
- **4E68:**[99] Unphosphorylated STAT3 core fragment bound to DNA (resolution: 2.59 Å) (organism: Mus musculus, synthetic construct) (residues 127-722) (Figure 9C)
- **4ZIA:**[103] Crystal structure of STAT3 N-terminal domain (resolution: 2.70 Å) (organism: Mus musculus) (residues 1-127) (Figure 9D)
- **6QHD:**[104] Lysine acetylated and tyrosine phosphorylated STAT3 in complex with DNA (resolution 2.85 Å) (organism: Homo sapiens) (residues 127-722) (Figure 9E)

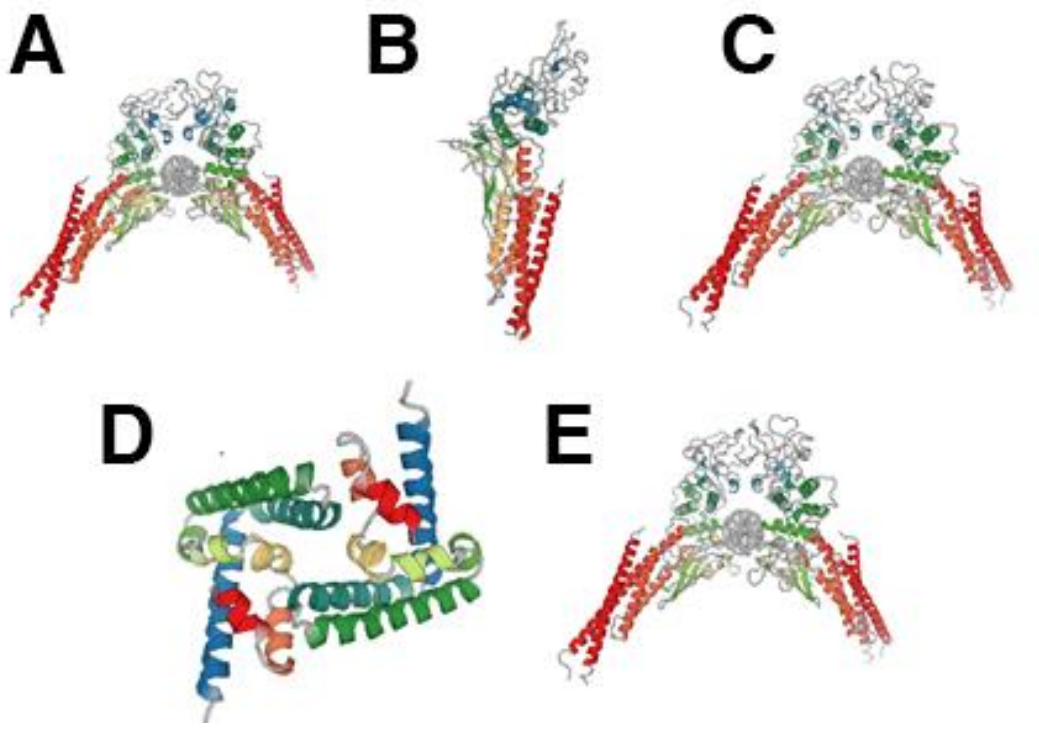


Figure 9 Crystallographic STAT3 structures available to date. A) 1BG1, B) 3CWG, C) 4E68, D) 4ZIA and E) 6QHD[100]

All these structures have some regions of tertiary structure missing in the protein-protein interaction domains. A flexible loop of residues containing the tyrosine residue Y705, which is enhancing dimer stabilisation and interaction surface, is missing. This loop interacts with the partner SH2 domain, stabilises the

association, and binds their phosphorylated Tyr into a specific binding site on the partner SH2 domain. This uncertainty of the structural arrangement makes the application of computational drug design techniques very challenging. The lack of crystal structures with a bound ligand makes the prediction of potential binding sites even more challenging.

2.6 STAT3 “druggability” studies and reported inhibitors

STAT3 can be inhibited directly or indirectly. Indirect inhibition of STAT3 can be achieved by the blockage of upstream tyrosine kinases or other factors involved in the STAT3 pathway. Indirect inhibitors are characterised by a non-selective mechanism of actions, which increases the likelihood of undesirable toxicity and other adverse off-target effects.

The problem with this strategy is that the STAT3 pathway may not be effectively blocked by a single compound and that compounds may inhibit other downstream targets. Furthermore, in cells in which proliferation results from the inactivation of negative regulators of signaling the inhibition of upstream signaling pathways may have little effect[86,105]. The nonspecific mechanism of action of indirect inhibitors highlights the restriction of this approach.

Direct inhibition can be achieved by targeting one of the three structural STAT3 domains: SH2 domain, DNA binding domain, and N-terminal domain. Direct inhibition should block one or more processes related to STAT3 signaling, including STAT3 phosphorylation, dimerisation, DNA binding, and STAT3-induced expression of genes.

Many authors focused on inhibiting the SH2 domain, considering its key role in STAT3 activation[105–108]. The SH2 domain is responsible for the interaction with phosphorylated Tyrosine residues within the cytoplasmic portion of the upstream receptors. It is also involved in the dimerisation of STAT3. Therefore, inhibition of SH2 domain by small molecules is expected to suppress the phosphorylation and activation of STAT3, as well as inhibit STAT3-DNA interaction.[109]

Several compounds have been reported as binders that effectively compete with phosphorylated STAT3 monomers for the pTyr-binding site. Some examples of

these inhibitors are the salicylic acid derivatives[105], S31-201[110], stattic[111] and S31-1757[112] (Figure 8). They all have been described as SH2 inhibitors, but there is no conclusive evidence that their mechanism of action is by binding in the pTyr-binding site.

Several authors showed discrepancies regarding the SH2 sub-pockets that are targeted by small molecule inhibitors. There is no consensus: some authors describe three sub-pockets, which do not match in most cases, and some others describe only two sub-pockets (Figure 10). There is mostly discrepancy in which is the third sub-pocket.

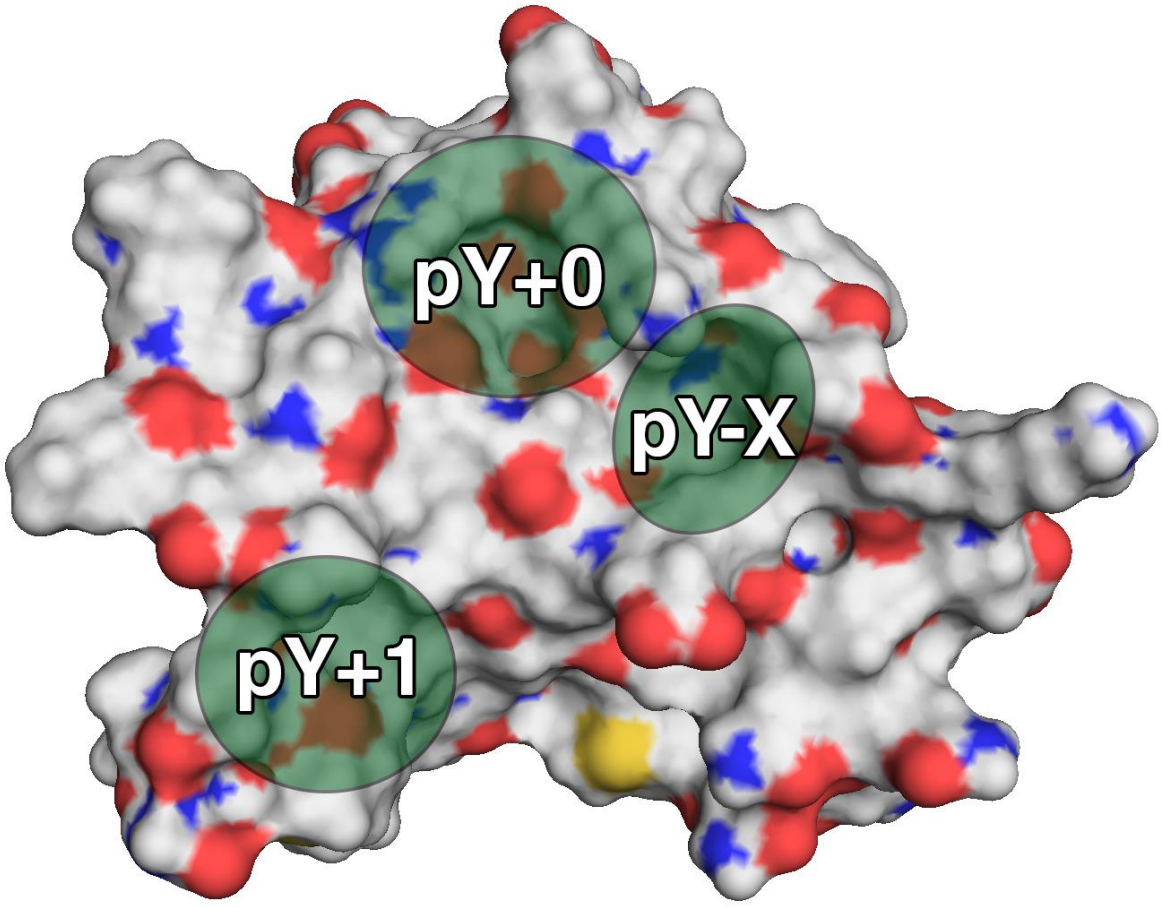


Figure 10 Most reports agree that STAT3 SH2 binding site is formed by three pockets: pY+0, pY-X and pY+1

Table 5 STAT3 binding sites defined by different authors

Park et al.[113]			***Poli et al.***[114]		***Chiao et al.***[115]	
Pocket pY+0	**Pocket pY-X**	**Pocket pY+1**	**Pocket pY**	**Pocket pY+1**	**Pocket pY**	**Pocket pY-X**
K591	M586	S636	K591	S636	K591	M586
R609	G587	Q635	R609	W623	R609	G587
S611	F588	T620	S611	Q635	S611	F588
E612	I589	K626	E612	V637	E612	I589
S613	S590		S613	Y657	S613	S590
Shahani et al.[105]			***Siddiquee et al.***[106]		***Fletcher et al.***[110]	
Pocket A	**Pocket B**	**Pocket C**	**Pocket A**	**Pocket B**	**Pocket A**	**Pocket B**
K591	I634	W623	K591	K592	K591	K592
R609	R595	V637	R609	R595	R609	R595
S611	E594	I659	S611	I597	S611	R595
S613	I597	F716	S613	I634	S613	I597
		K626				I634

Pallandre et al.[116]				***Zhang et al.***[105]		
Pocket A	**Pocket B**	**Pocket C**		**Pocket A**	**Pocket B**	**Pocket C**
K591	R595	Q635	T620	K591	R595	I659
R609	I634	S636	F621	E594	I634	W623
S611		W623	P639	R609		V637
S613		V637	Y640	E612		E638
S614		E638	Y657			
			I659			

Table 5 summarises some of the different descriptions by several authors. As showed, in most cases the pockets A or pTyr (Figure 10) are identical or highly similar, containing K591 and R609 in all descriptions. Residues like R595, W623 and S636, are present in most identifications, but in some studies completely

different pockets are annotated[115,117]. The lack of a STAT3 co-crystallised with ligands hampers the identification of STAT3 binding sites.

Another issue in the STAT3 binding site annotation is that these binding sites have been identified solely by means of molecular docking. Considering the number of drug design tools available to date that would be able to validate these findings, relying solely on molecular docking results renders those predictions not trustworthy and would require further validation studies.

2.6.1 Classification of direct inhibitors

The **direct inhibitors** can be divided into different categories:

- **Peptides** were the first compounds designed for inhibiting STAT3 in a direct approach. They block aberrant activity of STAT3 in cancer cells via preventing protein dimerisation. Their specific target is the SH2 domain, at the pY705 level. Starting from the sequence around pTyr705, the first phosphopeptide inhibitor PpYLKTK was developed.[118] This phosphopeptide inhibits STAT3 activity in tumour cell lines, induces cell death and has a high affinity and specificity for STAT3.[118]
  While peptide-based inhibitors can bind to STAT3 with high affinities, they suffer from low metabolic stability and the lack of cellular permeability due to their molecular nature and the negative charges on the phosphotyrosine group. However, they provided an excellent starting point for the development of more cell-permeable peptidomimetics[105,109,119].
- **Peptidomimetics** are inhibitors that mimic pYX_1X_2Q motif and inhibits STAT3 dimerisation by competitive binding to the SH2 domain.
- **Natural compounds**. Only a small number of natural compounds were found to directly target STAT3 protein like cryptotanshinone, a natural compound extracted from the root of *Salvia Miltiorrhiza Bunge*.[120,121] Previously curcumin was thought to be a relevant STAT3 inhibitor but recent studies have classified this molecule both as a PAINS (panassay interference compounds) and an IMPS (invalid metabolic panaceas), proving that curcumin is an unstable, reactive, nonbioavailable compound and, therefore, a highly improbable lead.[122]

- **Small molecules**. Non-peptide small-molecule inhibitors are cell-permeable, but most of the reported compounds bind STAT3 with weak affinities (IC_{50} values in the micromolar range) and the cellular activity cannot be clearly attributed to STAT3 targeting[105,119,123].

  The non-peptide small molecules represent a more attractive approach for inhibiting STAT3 directly, compared to peptides and peptidomimetics, given their cell permeability and physicochemical properties. However, optimisation of their structures and increasing the binding affinity is required to improve their in vivo activity and efficacy.
- **Oligonucleotides** represent a new, highly selective and low toxic class of drugs for targeting STAT3, which has shown promising results *in vivo* on nude mice xenografts[119].

2.6.2 SH2 domain inhibitors

Table 6 shows the most relevant STAT3 SH2 domain inhibitors reported up to date in the literature. Structures of these compounds are shown in Figure 11.

Table 6 SH2 domain inhibitors described to date*.*

Inhibitor Name	Reference
PpYLKTK	118
SPI	124
S31-201	110
S31-M2001	125
S31-1757	88
FLL32	126
Cryptotashinone	120
STA-21	107
Stattic	127

Celecoxib	[128]
Niclosamide	[129]
LLL-12	[130]
LLL-3	[131]
TPCA-1	[132]
BBI-608/Napabucasin	

Peptidomimetic inhibitors

- **S31-M2001** (Figure 11) is a novel oxazole-based peptidomimetic that selectively disrupts STAT3 dimerisation and therefore inhibits STAT3 transcription and migration in both human and mouse cells.[118]

Natural compounds.

- **Cryptotanshinone** (Figure 11) is a natural compound that binds to the SH2 domain and inhibits the formation of STAT3 dimers. Cryptotanshinone inhibits the STAT3 phosphorylation and decreases the expression of STAT3 downstream target genes involved in cell survival.[113]

Small molecules. Some small compounds have been developed to inhibit STAT3's SH2 domain by the means of computational drug design techniques such as virtual screening and QSAR.[98,101] These include STA-21, stattic and S31-201.

- **S31-201** (Figure 11) is a low-molecular-weight salicylic acid derivate that blocks STAT3 dimerisation through SH2 domain binding. Furthermore, S31-201 induces apoptosis in malignant cell by suppressing STAT3-dependent expression of cyclin D1, Bcl-XL, and surviving.[103]

- **Stattic** (Figure 11) inhibits selectively the dimerisation and DNA binding of STAT3, by preventing activating enzymes to the STAT3 SH2 domain. As a result, stattic induces apoptosis in breast cancer cell lines.[104]
- **STA-21** (Figure 11) is an antibiotic that specifically binds to SH2 domain inhibiting the STAT3 dimerisation and nuclear translocations. It has been reported that STA-21 inhibits breast cancer cell growth and survival.[100]
- **Celecoxib** (Figure 11) also binds to the SH2 domain of STAT3 and inhibit the binding of the native peptide. It has been reported that this inhibitor reduces cell viability and migration in human rhabdomyosarcoma cells.[121]
- **LLL12** (Figure 11) inhibits STAT3 phosphorylation, DNA-binding and induces apoptosis in various cell lines.[123]
- **TPCA-1** (Figure 11) blocks STAT recruitment to upstream kinases by docking into SH2 domain. Is an effective inhibitor of STAT3 phosphorylation, DNA binding, and transactivation *in vivo*.[125]

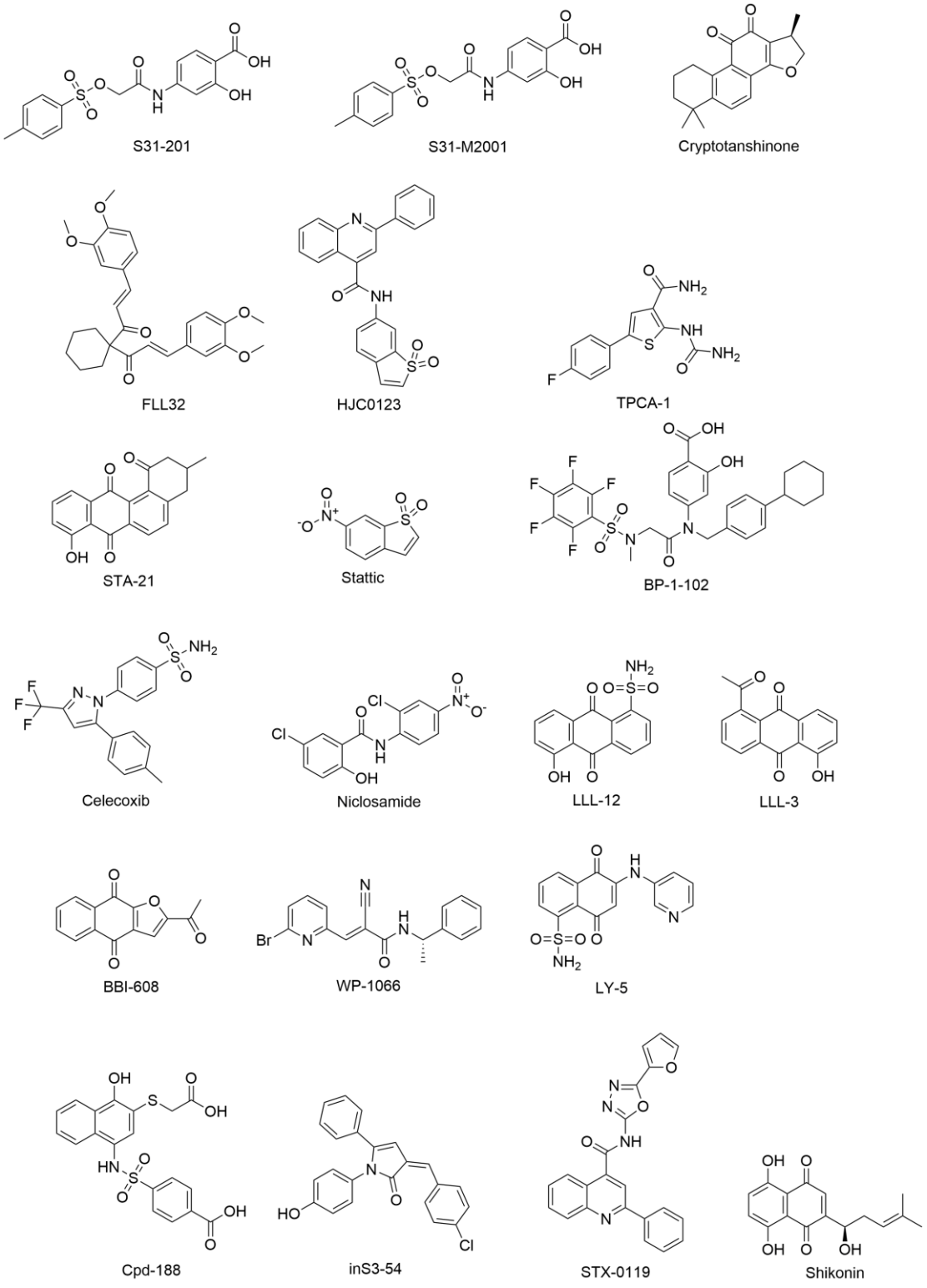


Figure 11 Chemical structures of described STAT3 inhibitors

2.6.3 Targeting the DNA binding domain BBI-608

Obtaining a ligand that effectively targets the SH2 domain is a hard quest. Most of the described inhibitors did not succeed in the drug design development

pipeline due to their poor activity, selectivity and toxicity issues. The lack of a co-crystallised STAT3 structure impedes a structure-guided drug design, as the most favourable protein conformation for binding is unknown. Although many reported STAT3 inhibitors are considered to target SH2, none of them has made it to clinic. The proposed mode of action (MoA) of those inhibitors remains elusive as it is not supported by structural biology data and relies solely on molecular docking calculations.

At the time of writing of this dissertation, only three direct STAT3 inhibitors are undergoing clinical trials. OPB-51602 and OPB-31121 (Otsuka Pharmaceuticals) have reached early phase clinical trials in both advanced solid malignancies[127,128]. Although signals of efficacy were observed in EGFR inhibitor-resistant non-small cell lung cancer (NSCLC) and gastrointestinal malignancies, the further development of these compounds was limited by concerns over their unpredictable PK profiles and potentially severe toxicities[129]. A plausible explanation for these side-effects is the ubiquitous expression of STAT3 within the body and its diverse physiological roles, including the modulation of mitochondrial metabolism and the immune system[130]. Second-generation OPB compounds with more favourable toxicity profiles have been identified and are currently being evaluated in early phase clinical trials[128]. OPB ligands are claimed to bind in the SH2 binding site, but in an allosteric position close to the canonical pTyr site. TTI-101 (Tivardi Therapeutics), which is another STAT3 inhibitor, is currently being evaluated in Phase I clinical trials for a range of advanced cancers, including breast cancer[128].TTI-101 is an antisense oligonucleotide and its mechanism is completely different from small molecule inhibitors. Napabucasin/BBI-608 is a first-in-class cancer stemness inhibitor that targets STAT3[131] , which is being tested (Phase 3) as a treatment in advanced colorectal cancer[132]. The BBI-608 patent documents contain a solved crystal structure of drug-STAT3 complex, with the BBI-608 bound in a pocket between the linker and DNA binding domain (Figure 12). The structure has not been deposited to the PDB Data Bank, though [133]. The community has not been eager on targeting the DBD due to the belief that targeting DBD of transcription factors has potentially limited selectivity[134–136]. Hence, DBDs has been considered “undrugabble”. This consensus has been challenged by the solution of BBI-608

structure bound to the DBD, which serves as “proof-of-concept” for direct inhibition of STAT3 DBD by small molecules. BBI-608 has been originally described as a SH2 domain inhibitor, and its derivatives are considered as such in the literature (again based on insufficient *in silico* methods such as virtual screening). In the recent years, interest in targeting STAT3 DBD has grown, and there is already a handful of described inhibitors claim to bind in the DBD based on their studies.[134,135,137] Small molecule ligands shikonin and the inS family of compounds (Figure 11) have been reported as effective STAT3 DBD inhibitors as the protein is dimerised upon inhibition[135].

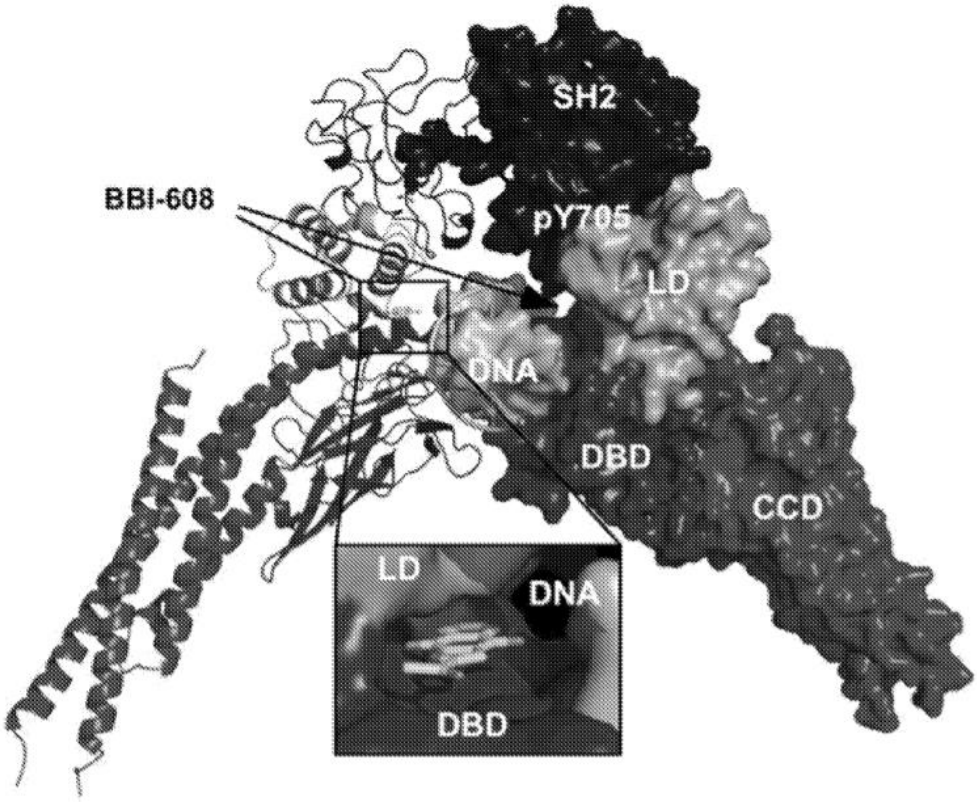


Figure 12 BBBI-608 binding conformation as per *Ji et al*[133]

2.7 Inter-domain mutations affect STAT3 activity

It has been recently reported by *Mertens et al.*[138] that mutations in the linker domain strongly suggest contacts between this domain and both the DNA binding and SH2 domains. These are likely to cause changes that strongly affect STAT3 activity. Residues within the linker domain (Figure 13), which form inter-domain

interactions, are involved in the hydrogen bonding and are presumably crucial for maintaining STAT3 biological activity. W564 forms the H-bond with SH2 domain S611, and W546 and K551 form an H-bond with E434. Alanine scanning of these residues showed a significantly drop in STAT3 DNA-binding and phosphorylation compared to the wild type protein. These destabilising effects could provide a plausible mechanism for STAT3 disruption[138].

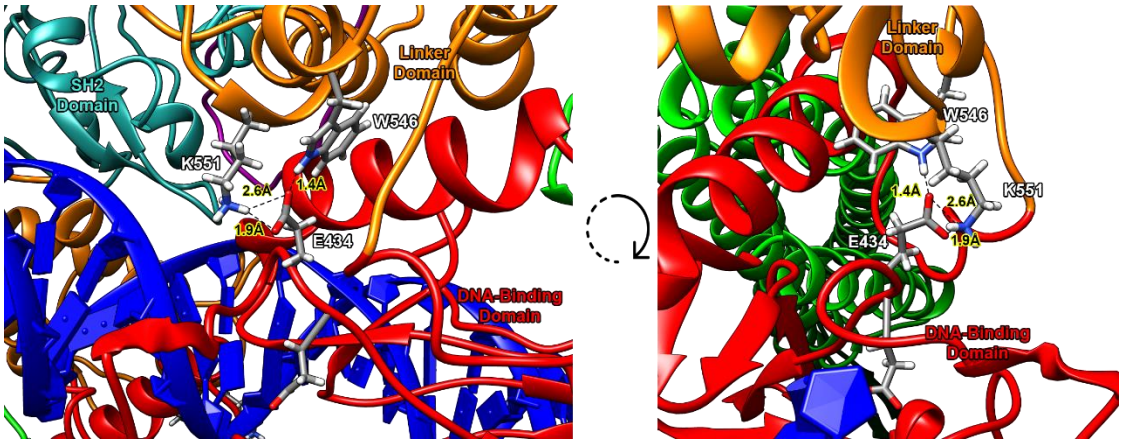


Figure 13 Location of mutated residues in the STAT3 linker domain structure. View of mouse STAT3 in complex with DNA (PDB ID: 1BG1). DNA binding domain (red), linker domain (orange) and SH2 domain (cyan), are highlighted. Hydrogen bonds are highlighted by yellow dashes and distances labelled in angstroms.

There also several reports that STAT3 undergoes allosteric communication across domains.[139] Solution NMR analysis shows that during pTyr binding and dimer formation (which occurs via the SH2 domain) considerable chemical shift perturbations are observed in the linker domain residues which are not directly involved in this processes.[139] One of the residues with the largest chemical shift perturbation upon binding of p-Tyr is I568, an interfacial residue between SH2 and linker domain, which alongside V572 is involved in hydrophobic interactions with F610 in SH2. These results show importance of allosteric sites within STAT3, and suggest focusing beyond the SH2 domain as a viable strategy for structure-guided development of novel STAT3 inhibitors. Potential novel allosteric sites at STAT3, which could be mapped, could then be rendered as alternative binding sites for small molecule inhibitors.

Chapter 3 Objectives

The main goal of this work was to develop a robust computational method to identify new, transient and/or cryptic “druggable” sites on hard-to-target proteins, such as STAT3. To achieve this goal, the following objectives were proposed:

1. Develop a workflow to identify potentially “druggable” allosteric and/or cryptic binding sites through the means of cosolvent molecular dynamics simulations.
2. Develop a robust, easy-to-use, and open source platform to analyse and interpret the results of cosolvent MD simulations, in terms of scoring and ranking of the identified “druggable” binding sites.
3. Apply the developed method in objective 2 to identify potential STAT3 allosteric binding sites, which could not be identified via conventional computational approaches.
4. Characterise STAT3 SH2 domain “druggability” in terms of the behaviour of the activation peptide binding site by means of equilibrium atomistic MD simulations and molecular docking studies, in order to assess the local flexibility and transient cavities. Molecular docking would be performed to evaluate the conservation of the pTyr binding site and to identify any potential binding pockets unreported to date.
5. Assess the effect of reported inter-domain mutations on STAT3 structure, dynamics, and ligand binding by means of umbrella sampling simulations of the STAT3-DNA complexes. These results would enable understanding of STAT3 allostery at the atomistic level.
6. Evaluate the binding mode and the mechanism of action of the STAT3 inhibitor BBI-608 by means of molecular docking, equilibrium MD, and umbrella sampling simulations. The calculated structure of BBI-608 bound to STAT3 would facilitate the structure-guided design of potent and selective allosteric STAT3 inhibitors.

Chapter 4 Methodology

The lack of a detailed understanding of how difficult to target proteins such as STAT3 interact with their ligands remains a major roadblock in advancing drug discovery efforts and uncover allosteric regulatory mechanisms. Reliable mapping of novel binding sites is essential for designing specific inhibitors, thus to develop new therapeutics in structure-guided manner.

This chapter covers the methods applied in this project. A need to include protein flexibility throughout binding events and druggability assessment, combined with the proteins' large size, requires usage of methods based on classical molecular mechanics, which are relatively fast and proven successful in structure-based drug discovery efforts. These comprise molecular dynamics simulations (equilibrium as well as enhanced sampling techniques, including cosolvent dynamics) and molecular docking/virtual screening.

4.1 Classical molecular mechanics

Computational methods applied in structure-based studies of biomolecular systems can be divided into two main groups based on their levels of theory. Quantum mechanics (QM) describes electrons explicitly, and is used to describe the process that involves their movement in atoms (and molecules), like breaking or formation of covalent bonds.[140] QM methods are very accurate yet they requires a high amount of computational resources and time, therefore application of these methods is limited to small systems. In classical molecular mechanics (MM), which utilises Born-Oppenheimer approximation (Equation 1), the electrons are treated implicitly and each atom is treated as a single particle. The Born-Oppenheimer approximation dictates that the nucleus motions not affected by the electronic movement, given the fact that the nucleus is thousand times heavier than the electron. Therefore, the electron-nucleus motion is decorrelated and the solution of the problem in the Schrödinger equation can be narrowed down to electronic motion only. This allows to study the structure and dynamics of large systems such as protein-ligand complexes, but it does not permit any covalent bond breaking or making.

$$\hat{H}\Psi(q_i, Q_I) = E\Psi(q_i, Q_I) \qquad \text{Equation 1}$$

$$\hat{H} = \hat{H}_{el} + \hat{T}_N$$

$$\Psi(q_i, Q_I) \approx \Psi^{el}(q_i; Q_I)\Psi^{NUC}(Q_I)$$

Both MM and QM levels of theory permits to:[140]

- Calculate the free binding energy changes associated with the formation of non-covalent protein-ligand complexes and determine their properties (QM and MM levels of theory)
- Model the chemical reactions (QM only)
- Perform conformational analysis of small molecules (QM and MM)
- Identify the near-native structure of the protein-ligand complex (molecular docking) and rank the set of small molecule ligands based on their calculated binding affinity to a given protein target (QM and MM; virtual screening and hit identification)
- Suggest the changes in the ligand molecule to improve the binding affinity (QM and MM; hit-to-lead and lead optimisation)
- Assess the flexibility and conformational changes within the system of interest (MM; classical molecular dynamics)

MM relies on three basic principles. The first one, also known as the Anfinsen dogma of protein folding, is the thermodynamic hypothesis. It states that a (macro)molecule driven by thermodynamic forces will change its conformation from the structure that represents a high energy state to a native structure which represents the global energy minimum state in a reversible fashion.[140,141] The second principle, the additive assumption, states that the total potential energy (*V*) of a system can be written as a sum of different potentials with simple physical interpretations (bond stretching, angle bending, Coulombic interactions, dispersion forces, etc.). The third principle, which is the transferability, is based on the assumption that parameters derived from small molecules such as bond lengths and angles can be transferred to larger, more complex, macromolecular systems, such as proteins and nucleic acids. Therefore, systems of different sizes can be studied using the same physical model (the force field).

4.1.1 Force field

The force field is a core concept in classical molecular mechanics, which approximates the potential energy of a system with a combination of bonded (intramolecular) and non-bonded (intra- and intermolecular) contributions (Equation 2).

$$V_{total} = V_{bond} + V_{anglel} + V_{dihedral} + V_{Coulomb} + V_{LJ} \qquad \text{Equation 2}$$

The harmonic terms describing the distortions from equilibrium positions in bond-stretching and angle-bending can be calculated using Equation 3 and Equation 4 respectively.[140]

$$V_{bond} = \sum_{i=1}^{N_b} \frac{1}{2} k_i^b (r_i - r_{0,i})^2 \qquad \text{Equation 3}$$

$$V_{angle} = \sum_{i=1}^{N_\Theta} \frac{1}{2} k_i^\Theta (\Theta_i - \Theta_{0,i})^2 \qquad \text{Equation 4}$$

The term constructed to describe the torsional motion of dihedral angles can be calculated from (Equation 5).

$$V_{dihedral} = \sum_{i=1}^{N_\phi} \frac{1}{2} k_i^\phi (1 + \cos{(n_i \phi_i - \delta)^2}) \qquad \text{Equation 5}$$

All the above interactions are represented by harmonic potentials for the bond lengths r_i, bond angle Θ_i, dihedral angle ϕ_i and phase angle δ that takes values of either 0º or 180º. The k^b, k^Θ, k^ϕ denote the force constants for the bond-stretching, angle-bending and dihedral angle terms.

The non-bonded interactions are more distant and not connected by covalent bonds. These can be divided in to short-range and long-range. The short-range interactions correspond to the van der Waals interactions and describe the

repulsion of two atoms due to overlapping valence electrons and attraction due to induction and dispersion forces. These interactions are commonly approximated by the 12-6 Lennard-Jones potential (Equation 6).[142] The distance dependence of the repulsion term is proportional to $r_{ij}{}^{-12}$ inter-atomic distance mimicking the exponential soft-wall behaviour, and proportional to $r_{ij}{}^{-6}$ with regards to the attraction. The 12-6 Lennard-Jones potential on a given particle i due to particles j in a system is described by:

$$V_{LJ}^{ij} = \sum_{i \neq j} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \qquad \text{Equation 6}$$

where ϵ_{ij} denotes the depth of the potential well, or region surrounding a local minimum of potential energy, and r_{ij} is a finite distance at which the inter-particle potential is zero.

The long-range interactions are the consequence of electric charges in the system. The individual atoms of a molecule are charged, allowing for the use of Coulomb's law to describe the mutual interactions between two (partial) atomic charges, and providing multipoles for molecules or individual charge groups. The Coulomb potential on particle i due to particle j reads (Equation 7):

$$V_{Coulomb} = \sum_{i \neq j} \frac{1}{4\pi\epsilon_0\epsilon_{rel}} \frac{q_i q_j}{r_{ij}} \qquad \text{Equation 7}$$

where ϵ_{rel} is dielectric constant of the medium. The electrostatic interactions decrease as $r_{ij}{}^{-1}$ increases, making them longer ranged than the van der Waals interactions.

The most widely used biomolecular force fields include the AMBER[143,144], CHARMM[145], GROMOS[146] and OPLS[147] force fields. In this work, AMBER force field (AMBERFF99 SB-ILDN[148]) has been used. These force fields share similar mathematical functional forms and they differ in the parameters that describe the various energetic components and in the methods used to obtain these

parameters. Recent force fields were in the most part defined by fitting parameters to data obtained from quantum-level calculations or experiments on small molecules thought to mimic the properties of proteins.[149] Most of these force fields have not changed in some time, but a few parameters such as a few torsion angles have been refined to improve their accuracy for proteins and peptides. Although, standard and commonly used force fields like Amber ff99-SBILDN tend to sample globular proteins reasonably well, their performance is subpar with intrinsically disordered proteins (IDPs) and other very flexible regions of a protein.[150] Therefore, there have been force fields specifically designed to properly model the folding of IDPs such as Amber ff03ws, the counterpart was that these force fields are so fitted to IDPs that tend to sample the behavior of folded proteins. Recently, a rewrite of a99SB by Robutstelli et al (a99SB-disp) also helped improve the modelling of IDPs along with an accurate description of folded protein properties.[151] In this work the Amber ff99-SBILDN was solely used for MD simulation as the studied systems correspond to folded proteins. A99SB-disp could have been used to better sample very flexible regions of the studied proteins such as STAT3's transactivation domain but sadly these corrections were not available upon the time the simulations were performed.

4.2 Molecular dynamics

Molecular dynamics (MD) consist in a computer simulation technique that predicts the time evolution of a system of interacting particles (atoms, molecules, beads, etc). An outcome of an MD simulation is a trajectory that specifies how the positions and velocities of the particles in the system vary with time. Analysis of the trajectory for a given biomolecular system can provide valuable information concerning molecular geometries and energies; mean atomic fluctuations; local fluctuations (like formation/breakage of hydrogen bonds, water/solute/ion interaction patterns, or nucleic-acid backbone torsion or motions); enzyme/substrate binding; free energies and even large-scale conformational changes of macromolecules such as small protein folding.[140]

The principle behind the classical MD is Newton's second law of motion, used to calculate the dynamics of the system (Equation 8).

$$\vec{F} = m\vec{a}$$ Equation 8

If the mass of each particle in the system is known and the forces are derived from the interactions with surrounding particles using the force field (Chapter 4.1), the acceleration of each particle can be calculated. Then, the instantaneous velocity and the displacement can be calculated by the numerical integration from Equation 9 and Equation 10:

$$\frac{\vec{F}_i(t)}{m_i} = \vec{a}_i(t) = \frac{d\vec{v}_i(t)}{dt}$$ Equation 9

$$\vec{v}_i(t) = \frac{d\vec{r}_i(t)}{dt}$$ Equation 10

The force acting on each atom i in the system is given by the negative gradient of potential energy function V, which depends on the coordinates of all other atoms in the system (Equation 11)[140]

$$\vec{F}_i(t) = \frac{-\delta V(\vec{r}_1(t), \vec{r}_2(t), \ldots, \vec{r}_N(t)}{\delta \vec{r}_i(t)}$$ Equation 11

If the potential energy of the system is known and the coordinates for a starting structure and a set of velocities are given, then the force acting on each atom can be calculated and a new set of coordinates is generated by advancing the simulation in a short span of time called timestep (δt), from which new forces are calculated. These integration cycles are usually calculated via a leapfrog integration method. This method defines the positions and velocities as time-dependent Taylor series, which can be integrated to obtain its related primitive function. Repetition of this procedure will generate a trajectory corresponding to the evolution of the system in time (Figure 14).[140] In case the velocities are unknown (e.g. first equilibration step) these are calculated from a Maxwell distribution for a set temperature.

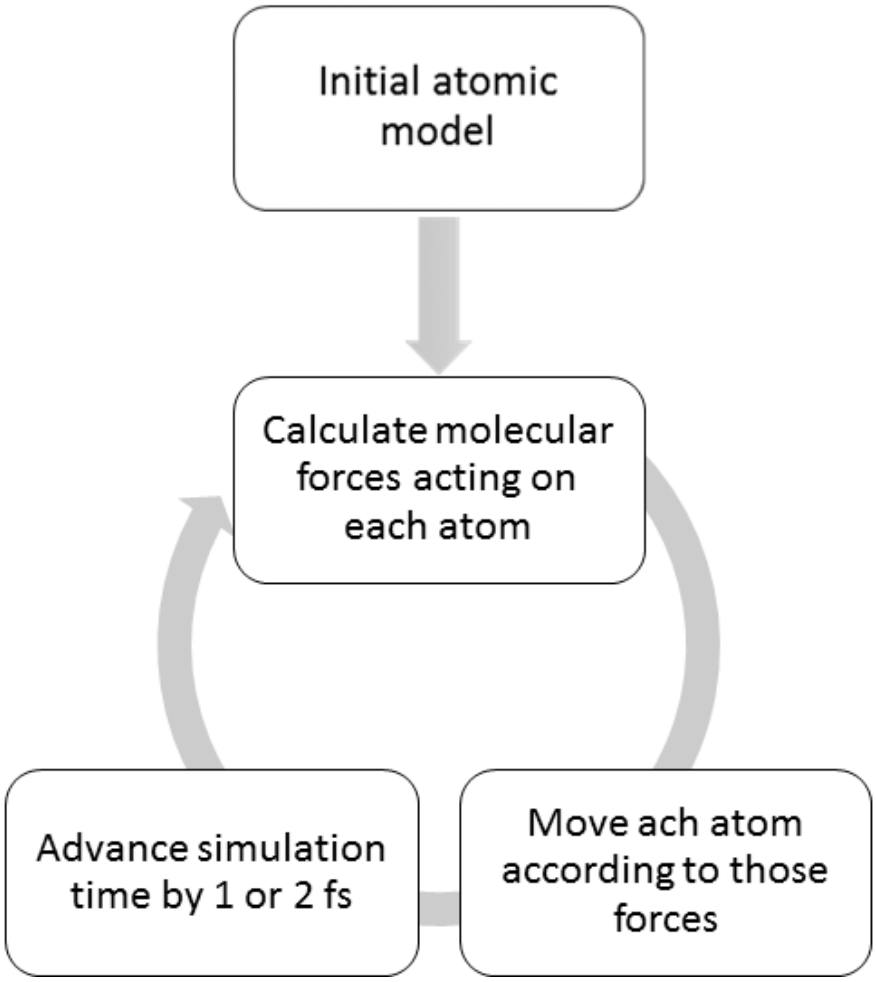


Figure 14 An overall scheme of the molecular dynamics simulation

In order to reproduce the behaviour of real molecules in motion, the force field terms are parameterised to fit quantum-mechanical calculations and experimental data. Parameterisation includes identifying the ideal stiffness and lengths of the springs that describe chemical bonding and atomic angles, determining the most appropriate partial atomic charges used for calculating electrostatic-interaction energies, identifying the proper Van der Waals atomic radii, and more.[152]

Application of molecular dynamics simulations to ligand-protein interaction are still limited by two major roadblocks: inaccuracies in the force fields applied, and an insufficient sampling problem, which is related to time scales accessible (up to microseconds, due to high computational costs of longer simulations).[152]

4.2.1 Classic/canonical MD

MD simulations with the atomistic resolution are well-established and deliver a generous amount of details and insights for the studied system allowing investigation of intra and inter-molecular processes. In the atomistic description, each atom in the system is defined as a single interaction centre, and the forces acting on it are evaluated every time step. The time step is limited by the fastest vibrations in the system (C-H) that corresponds to 1-2 fs. A time step of this size allows for stable numerical integration, however, limits the accessible time and a total length for biomolecular simulations. Thus, studies of conformational changes of large proteins are out of reach for conventional atomistic simulations at equilibrium. Therefore, the use of full atomistic models is restricted to relatively small systems or short time scales (nanoseconds to microseconds). The clear advantage of the atomistic simulation is that the results are detailed and enable to study the phenomena which are difficult to access by experimental methods, e.g. the lifetime of a single hydrogen bond.

4.2.2 Conditions in molecular dynamics simulations

Simulations are usually carried out in the explicit solvent (water), and the water molecules are added to fully immerse the system in the box (Figure 15). There are many water models available for MD simulations, three-point TIP3P water model being the most popular and the one used in this work. The system should have periodic boundary conditions, meaning that molecules that exit one side of the system will wrap to the other side of it. This is to enable the constant number of particles in the simulation box, which is required by statistical mechanical ensemble (NVT or NPT), and to avoid finite-size effects. It is important that the periodic box would be large enough to embed the whole molecular system.

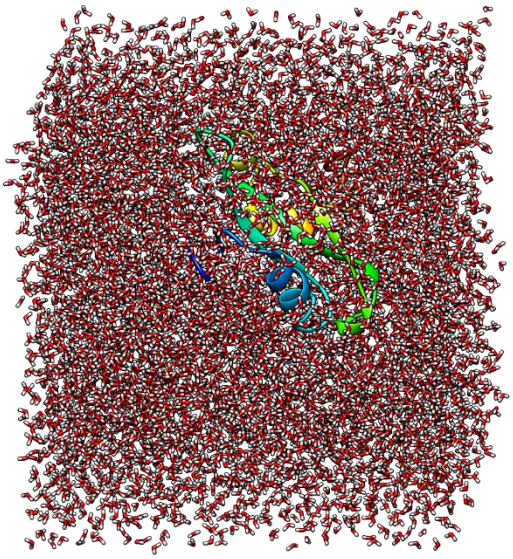


Figure 15 Simulated protein in a cubic TIP3P water box

Three input files are needed to start the simulations. These are the topology file, the coordinate file, and the force field. The topology file contains all the information about the structure and connectivity between atoms in the system.

A typical MD simulation consists of several subsequent stages, which are outlined in Table 7.

Table 7 Different stages of an MD dynamics performed

Stage	Purpose
Energy minimisation	Adjust the structure to the force field, particular distribution of solvent molecules and relaxation of possible steric clashes
Heating (NVT)	Linear heating of the system from 0K to 300K
Equilibration (NPT)	Equilibration at constant pressure. Used to equilibrate kinetic and potential energies
Production	Sample structural and dynamics characteristics

4.2.2.1 Energy minimization

The methods to obtain the protein structure tend to present several atomic clashes. Therefore, in order to assure the structural stability of the system, energy minimisation is required. If the system energy is too high, the resulting force vector will have a momentaneous high intensity, disrupting the simulation box and crashing the integration cycle. Energy minimisation consists in an approach to reduce the probability of the aforementioned crash. One of the most popular methods to reach an energy minimum in molecular mechanics are the steepest descent algorithm. Nevertheless, it should be considered that the system might get trapped in a local energy well, therefore a more global method to sample the global energy minimum might be required.

The steepest descent (SD)[153] algorithm is a method based on the derivative of the potential energy. For every minimisation cycle, the 3N dimension position vector r_{n+1} can be calculated following equation 12.

$$r_{n+1} = r_n \frac{F_n}{MAX(F_n)} h_n \qquad \text{Equation 12}$$

Where r_n is the starting position, F_n is the force applied in that atom, $MAX(F_n)$ is the maximum force applied in any atom and h_n corresponds to the atomic displacement for that cycle. For this reason, the process goes through all system atoms, and the convergence criteria is either a predefined number of cycles or an upper threshold of the system highest force.

SD is a simple method that tends to be quick. Due to its use of only orthogonal gradients, it is prone to get trapped in a local energetic well[153]. This cause might be attenuated by modifying the parameter h (maximum allowed displacement per cycle), with a progressive decrease in a series of cycles to improve the final configuration. From a biomolecular point of view, the protein's starting configuration tends to be close to the global minimum, since in most cases corresponds to a representation of the native state. Nevertheless, issues with the experimental data upon resolving its structures may arise erroneous sidechain

configurations that energy minimisation should solve. Once the system is energetically minimised, the thermodynamic variables need to be defined.

4.2.2.2 Reaching thermodynamic equilibrium

In principle, the model obtained after the energy minimisation procedure has a temperature of 0K, as there is no dynamical atomic motion assigned to it. Therefore, the system temperature must be resolved. A specific condition of the experiment should temperature increase. In the case of system heating and NVT ensemble is used.

The NVT ensemble is a statistical ensemble that represents the probability of accessible states in a predefined configuration. For this case, the number of particles (N), the system volume (V) and the temperature (T) are set as constant. The probabilities assigned to each microstate of the system follow equation 13:

$$\rho = \frac{e^{\frac{-E}{kT}}}{Z}$$ Equation 13

Where k corresponds to the Boltzmann constant, T to the temperature, E is the state energy and Z or partition function is (Equation 14):

$$Z = \sum_{1}^{n} e^{\frac{-E}{kT}}$$ Equation 14

As the probabilities in this ensemble are not dependant on any other variable (i.e. pressure), this approach can be used for heating the system. Position restraints are applied to ensure that this process does not affect the starting structural conformation. These restraints tend to be applied through the addition of a harmonic potential on selected protein atoms.[140]

Next, in order to reach a microstate temperature T in a restrained configuration, a distribution of velocities is applied to the atoms (Equation 15).

$$\sum_{i=1}^{N} m_i |V_i|^2 /_2 = \frac{kT}{2}(3N - Nc)$$ Equation 15

Where $m_i$is the mass and $|V_i|^2$ corresponds to the average velocity of atom i, N to the total number of particles and Nc to the number of restrained components. This corresponds to a Boltzmann-Maxwell distribution of velocities that reaches the desired temperature[154]. To keep the temperature updated through the integration timesteps, a thermostat algorithm is applied to the system. A simple method to change the temperature corresponds to rescaling the velocity for every new step to the temperature T. Following this principle Equation 15 turns into Equation 16

$$\sum_{i=1}^{N} m_i |V_i|^2 /_2 \rightarrow \sum_{i=1}^{N} m_i \gamma |V_i|^2 /_2$$ Equation 16

Where γ is (Equation 17)

$$\gamma = \sqrt{\frac{T}{T_i}}$$ Equation 17

And T_i is the temperature of step i. These methods do not actually allow thermal fluctuations through the system due to the fact that the temperature rescales directly with the velocity. Based on the aforementioned feature, the Berendsen thermostat was devised. This is an algorithm that assumes that the system is weakly coupled to a heating bath, which updates the average temperature. Because of the weak coupling, the temperature does not scale directly with velocity.[154,155] Therefore, γ for a Berendsen thermostat follows Equation 18.

$$\gamma = \sqrt{1 + \frac{\Delta t}{\tau}\left(\frac{T}{T_i} - 1\right)}$$ Equation 18

Where t corresponds to a coupling term named “rise time”. This term controls how strong the system feels the temperature bath. Because scaling methods

scale the velocity directly, they do not allow stochastic variations. Other thermostats, such as the Nosé-Hoover thermostat address this issue. Usually these thermostats demand more computational resources but are able to simulate a proper canonical ensemble. An alternative between both methods would be velocity rescaling.[156] This method, implanted in Gromacs, adds a wiener stochastic function to the γ term. Therefore, the velocity scaling becomes randomised, sampling a full canonical ensemble. This thermostat was the one used throughout this work.

4.2.2.3 Pressure equilibration

Once the system has reached thermal equilibration, the volume configuration needs to be set. In an NVT ensemble, the box volume is constant, and this might not be the most accurate volume conditions for the box in question. Therefore, equilibration is needed to set up the remaining macro thermodynamic variables, such as pressure. Furthermore, the experiments we aim to model tend to happen in a constant pressure regime.

In this second stage of equilibration we use an NPT or isothermal-isobaric ensemble with N – number of particles, P – pressure and T- temperature as constants. For this case the system is coupled to a pressure control as well as a temperature bath. The two most typical ways are the weak Berenden coupling[153] and the Parrinello-Rahman barostat[157]. The Berendsen coupling barostat works in a similar manner to its thermostat, as it scales the box volume through time to achieve a predefined pressure. This barostat belongs to a class called isotropic scaling as it does not change the overall shape of the box but equally modifies the size of the box in all dimensions. On the other hand, the Parrinello-Rahman barostat performs an anisotropic scaling.

4.2.2.4 Production simulation

Once the thermodynamic macro variables are defined and the system is equilibrated, the dynamical ensemble can be calculated. As mentioned before, the integration cycle drives the calculations. After the equilibration stages, the atomic harmonic restraints are disable so the protein system can sample in a

more realistic manner. In most cases, the parameters used for a production run are the same that for an NPT configuration.

4.2.3 Variations in classic conditions

The main aim of this project is to detect new transient or cryptic pockets that may be difficult to identify by the use of classical computer-aided drug design (CADD) techniques. Modifications in the typical simulation conditions could be exploited in the identification of new “druggable” cavities.

4.2.3.1 The cosolvent approach

The discovery of novel binding sites is not always coupled to the local minima retrieved from the simulated landscape, or the highest-populated clusters. The use of small, drug-like “probes” as cosolvents (Figure 16) can be used to assess which functional groups would best complement the surface of the binding site of interest (either already known or newly identified).

Several cosolvent dynamics approaches have been developed in recent years. Most of these approaches use very small probes such as: ethanol, isopropanol, methanol, acetonitrile or acetamide.[158] In the present work, the cosolvent based framework has been extended to usage of drug-like fragments derived from known drugs/inhibitors. The low probe concentration permits to perform longer simulations before observing phase separation without the use of repulsive potentials. Longer trajectories permit the ligands to search for more cavities, evaluate a longer protein-probe stability and therefore determine its affinity.

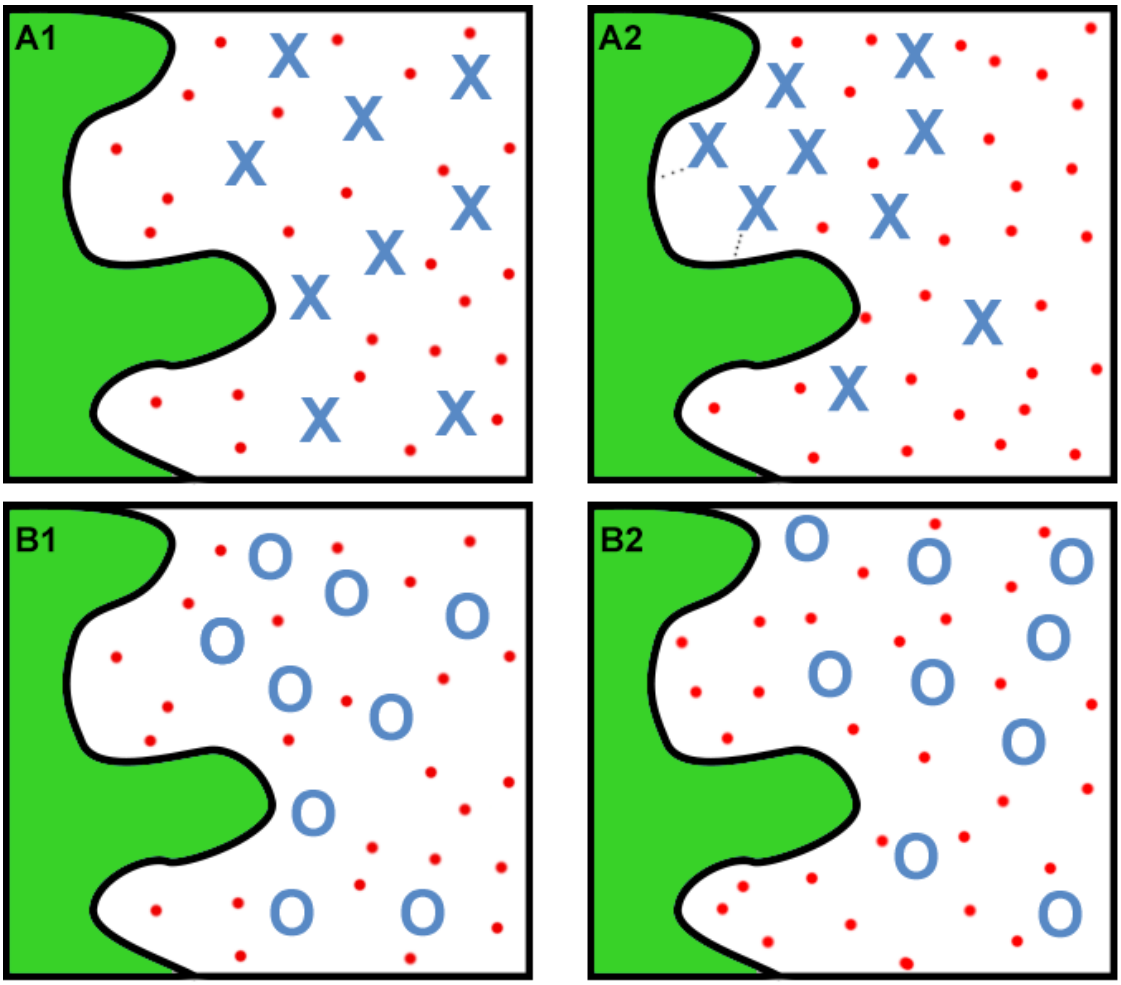


Figure 16 Cartoon depiction of a cosolvent simulation. In step 1 cosolvents A (crosses) and B (circles) are randomly in a simulation box put in a protein (green) – water (red dots) system. After the simulation (step 2) molecules A show a higher affinity to one cavity of the protein and interact with it, while B molecules do not interact with the protein at all.

4.2.3.2 Enhanced sampling techniques

The energy landscape of a protein is characterised by a series of metastable states separated by high energy barriers. Since atomistic MD is limited to a few femtoseconds integration timesteps, it is difficult to go into the millisecond timescale and beyond with nowadays machines, where these new conformational states could be visited.[159] Only a few special machines, such as Anton or the use of the Folding@home, are able to reach extensive timescales.[160,161]According to the transition state theory, the relationship between the timescale of state transition and the height of an energy barrier is exponential.[162–164] This means that most conformational events of interest, such as the folding process of a protein or the binding/unbinding process of a

substrate, often occur at a larger timescale.[165] In an attempt to overcome this issue, a number of enhanced sampling techniques have been developed over the past decades to allow for fast thermodynamics and/or kinetics calculations. In the next section I will discuss the umbrella sampling (US), since it is the method that was used in this project.

4.2.3.2.1 Umbrella sampling

Umbrella sampling, developed by Torrie and Valleau[64], consists on the application of a bias, an additional energy term, to the system to ensure efficient sampling along a reaction coordinate. A reaction coordinate (ξ) is a continuous parameter that provides a distinction between two thermodynamic states. Generally, ξ appears to be defined on geometric grounds, such as distance, torsion, or the difference between the root mean square deviations from two reference states. If the reaction coordinate of choice is good enough to differentiate distinct states, the free energy between these would be calculated. This is aimed in different simulations (windows), the distributions of which overlap.[166] Window i bias potential ω_i is an additional energy term that only depends on the reaction coordinate (Equation 12).

$$E^b(r) = E^u(r) + \omega_i(\xi) \qquad \text{Equation 12}$$

The superscript 'b' denotes biased quantities, while the superscript 'u' denotes unbiased quantities. Quantities without superscripts are always unbiased.

The reaction coordinate is split into a number of windows to ensure an optimal sampling. A bias function is implemented in each of these windows to keep the system close to window i reference point. A simple harmonic bias of strength K is often used (Equation 13).

$$\omega_i(\xi) = K/2(\xi - \xi_i^{ref})^2 \qquad \text{Equation 13}$$

To obtain unbiased free energy $A_i(\xi)$, the unbiased distribution of the reaction coordinate must be obtained, as showed in Equation 14.

$$P_i^u(\xi) = \frac{\int \exp[-\beta E(r)]\,\delta[\xi'(r) - \xi] d^N r}{\int \exp[-\beta E(r)]\, d^N r} \qquad \text{Equation 14}$$

The unbiased probability $P_i^u(\xi)$can be determined by Equation 15.

$$P_i^u(\xi) = P_i^b(\xi) \exp[\beta \omega_i(\xi)] \langle \exp[-\beta \omega_i(\xi)] \rangle \qquad \text{Equation 15}$$

Biased probability can be obtained from the simulation of each window, and the free energy of every window could be calculated by Equation 16.

$$A_i(\xi) = -\left(\frac{1}{\beta}\right) ln P_i^b(\xi) - \omega_i(\xi) + F_i \qquad \text{Equation 16}$$

Where $F_i = -\left(\frac{1}{\beta}\right) ln\langle \exp[-\beta \omega_i(\xi)] \rangle$ as long as one window covers the entire range of ξ to be examined. If the free energy curves are to be combined into a global one (Figure 17), F_i has to be calculated with methods such as the weighted histogram analysis method (WHAM)[167,168].

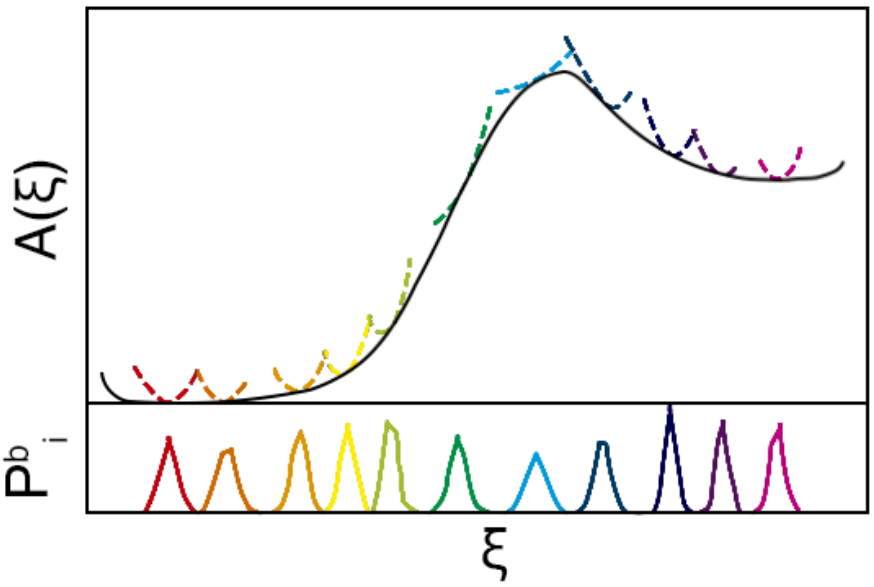


Figure 17 Global free energy (black solid curve) and the contributions A_i of some of the windows (dashed curves). Only every third window is shown for clarity. At the bottom: the biased distributions P_i^b as obtained from a simulation are shown (coloured solid curves). Relatively few bins (100) have been used to generate this scheme[166]

WHAM is applied to minimise the statistical error of $P^u(\xi)$. The global distribution is calculated by a weighted average of the distributions of the umbrella windows (Equation 17):

$$P^u(\xi) = \sum_{i}^{windows} p_i(\xi) \mathrm{P}_i^u(\xi) \qquad \text{Equation 17}$$

The weights p_i are chosen in order to minimise the statistical error of P^u (Equation 18):

$$\frac{\partial \sigma^2(P^u)}{\partial p_i} = 0 \qquad \text{Equation 18}$$

Under the condition $\sum E p_i = 1$. This leads to (Equation 19)[167,168]:

$$p_i = \frac{a_i}{\sum_j a_j}, a_i(\xi) = N_i \exp[-\beta \omega_i(\xi) + \beta F_i] \qquad \text{Equation 19}$$

With N_i being the total number of steps sampled for window i. Equation 20 calculates F_i.

$$\exp(-\beta F_i) = \int P^u(\xi) \exp[-\beta \omega_i(\xi)] \, d\xi \qquad \text{Equation 20}$$

Since P^u enters Equation 20 and F_i enters Equation 17 via Equation 19, these have to be iterated until convergence.

For an efficient umbrella sampling run, an overlap between windows is required for WHAM analysis. Good sampling is essential for a proper choice of the reaction coordinate. If the reaction coordinate misses important structural changes, it can lead to artificial reduction or increase of the energy barriers from the results obtained by umbrella sampling.

4.2.4 Data analysis

MD simulations produces trajectories, which are series of sequential snapshots of the simulated molecular system which represent atomic coordinates at specific time periods. This generates large amounts of data, which must be processed and analysed. The type of analysis performed can vary substantially depending on the question(s) and hypothesis posed before even carrying out the simulation. In the next section, I will outline the analytical techniques applied in this work.

4.2.4.1 Root-mean squared deviations and root-mean squared fluctuations

Root-mean squared deviation (RMSD) is the calculation of the average distance of certain atoms in a system from a reference structure, r_i^{ref}, as showed in Equation 21:

$$RMSD(t) = \sqrt{\left(\frac{1}{M}\sum_{i=1}^{N} m_i |r_i(t) - r_i^{ref}|^2\right)} \qquad \text{Equation 21}$$

Where $M = \Sigma_i m_i$ and $r_i(t)$ is the position of atom i at time t after the structure is fitted to the reference state. RMSD is calculated to evaluate the stability of the simulated system. If the obtained RMSD plot shows severe deviations through time, then the system has not reached energy convergence, meaning that further simulations would be required.[140] An example of the RMSD plot is showed in Figure 18.a.

Root-mean square fluctuation (RMSF) is a measure of the difference between the position of particle i and some reference position (Equation 22).

$$RMSF_i = \sqrt{\frac{1}{T}\sum_{t_j=1}^{T} |r_i(t_j) - r_i^{ref}|^2} \qquad \text{Equation 22}$$

Where T is the time over which one wants to average and r_i^{ref} is particle i reference location. RMSF is distinguished from RMSD by giving a value for each

particle or group, such as an amino acid residue, over time. This means that the fluctuation of every particle (residue, atom, chain, etc.) can be evaluated, and thus the regions of the system showing large conformational changes can be identified.[140] An example of the RMSF plot is showed in Figure 18.b.

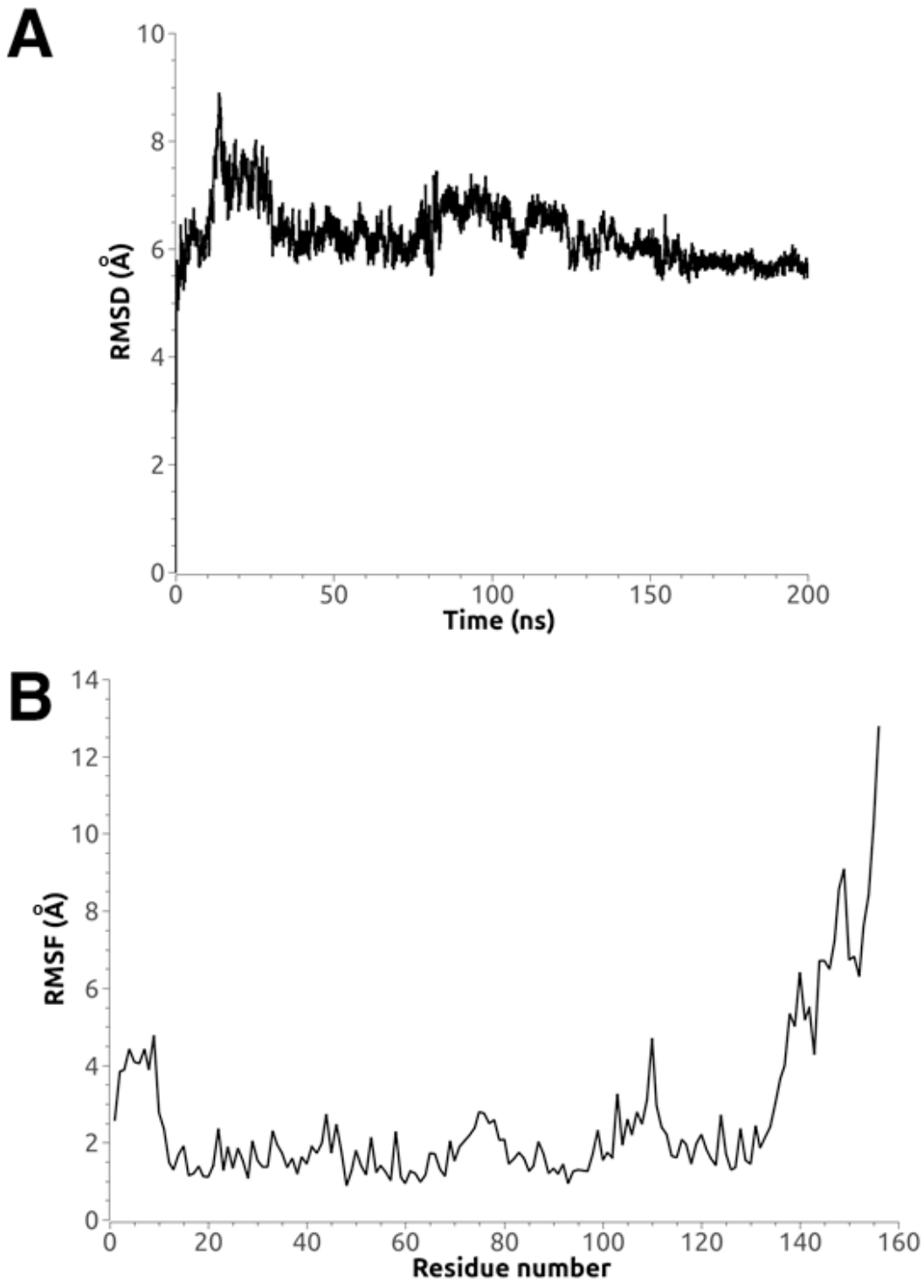


Figure 18 RMSD (A) and RMSF (B) plots

4.2.4.2 Geometric clustering

MD simulations can generate thousands of snapshots (conformations) to be analysed. In some cases, these conformations can be very similar, and it is of interest to reduce the set of conformations for subsequent analysis. Geometric clustering is an analysis technique performed to classify different structure samples during an MD simulation (Figure 19). There are several clustering methods available, but in this work user approached the one developed by Daura and coworkers[169] (via the *gmx cluster* module), which is based on the mutual RMSD between all conformations sampled during the MD simulation (for a specified RMSD cut-off). The generated clusters are mutually exclusive, meaning that a structure can only be a member of a single cluster. Geometrical clustering is often used to describe the various conformational changes in a protein. These structures can also be used for further studies such as virtual screening (VS) or umbrella sampling (US) simulations.

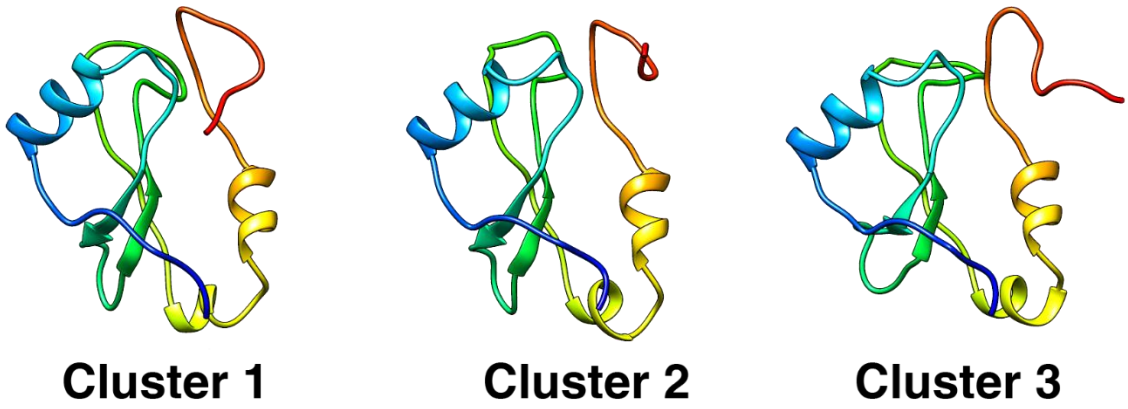


Figure 19 Top three clusters from an MD simulation

4.2.4.3 Principal component analysis (PCA)

Following the idea of the previous section the number of variables that are used to describe each conformation is very large. These variables may be correlated. Principal component analysis (PCA) is a commonly used method to eliminate these correlations and to reduce the dimensionality of the data set. In general, a principal component (PC) is a linear combination of the variables. The first PC of a data set corresponds to that linear combination of the variables which give the

“best fit” straight line through the data if plotted in a n-dimensional space, meaning that it has the largest possible variance (Figure 20). The second and succeeding principal components have the highest variance possible with data not already accounted by previous principal components. Each PC corresponds to an axis in a n-dimensional space and is orthogonal to all other PC. There can be as many dimensions as variables the original data set provides, but in an optimal situation only a few (3-5) principal components may be required to explain most of the data.[140] The modules *gmx covar* and *gmx anaeig* have been used to perform PCA analysis.

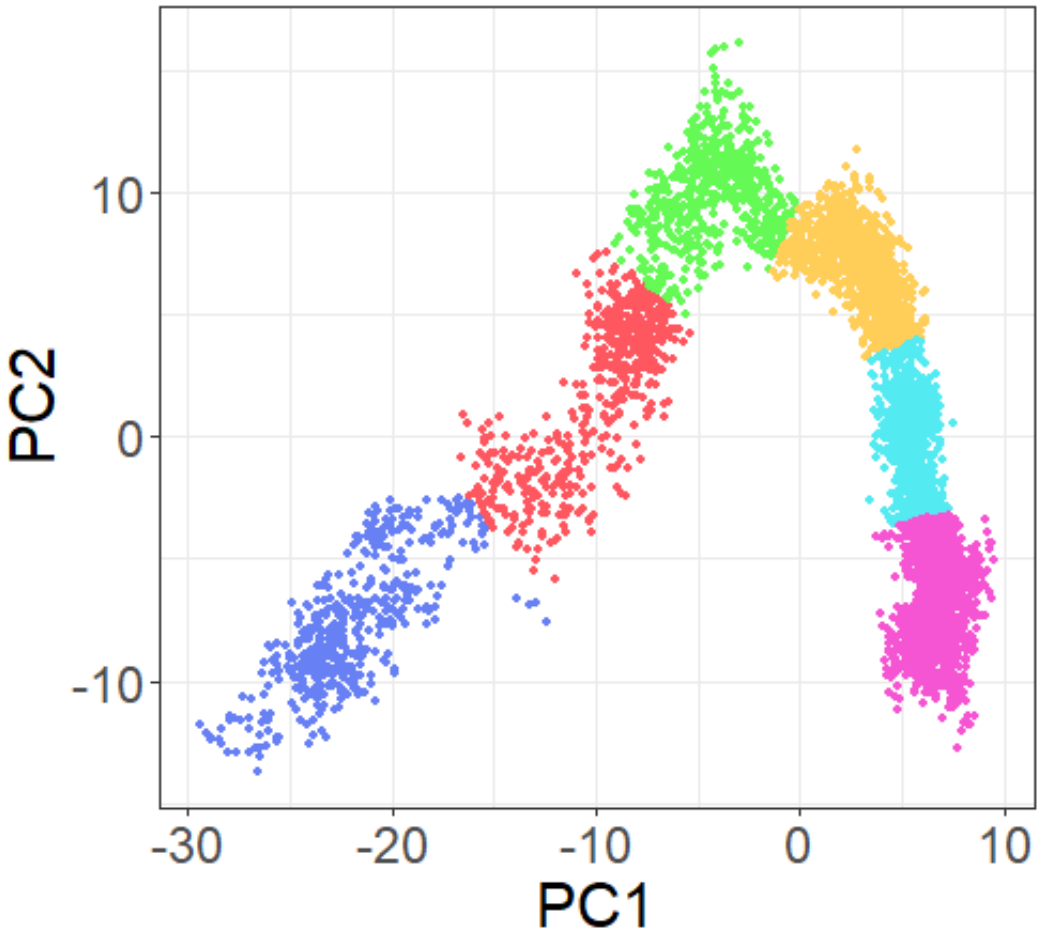


Figure 20 Two dimensional PCA plot from an MD simulation. Different clusters of the simulation are differentiated by colour

4.2.4.4 Distances

Monitoring of distance between two groups of atoms is another way to obtain useful information about the system. H-bond formation (or breaking), shifting of sidechain unveiling cavities or distance between protein monomers are some of the several outcomes that distance measure can provide.

4.2.4.5 Solvent accessible surface area (SASA)

In the context of cosolvent MD, it is a matter of interest to evaluate if the participation of the cosolvent probes is indeed promoting the formation of new binding regions. One way to evaluate this feature may be to calculate the solvent accessible surface area (SASA). As its name states, SASA is the surface area of a biomolecule that is accessible to a solvent. In this work, the SASA of the systems were computed with the *gmx sasa* module, which applies the double cubic lattice method[170].

4.3 Molecular docking

Molecular docking is a technique used to predict the interaction between a ligand and a protein binding site. This technique calculates the most optimal binding geometries (poses) and binding energies, by placement of the ligand in different orientations and conformations within the binding site, which can be considered completely rigid or semi-flexible. Molecular docking attempts to mimic the natural course of interactions between the ligand and its cognate receptor.[20,140]

There are three important applications of the molecular docking. One is the determination of the binding mode (geometry) of a ligand bound to a protein. Molecular docking generates hundreds of thousands of putative ligand binding orientations/conformations at the defined binding site within the protein target.[171] A scoring function is used to rank these ligand conformations by evaluating the approximate binding energy of each of the putative complexes.

The second application is to identify the potential hits for a given protein target by searching large ligand databases, i.e. the virtual screening.[172] A reliable scoring function should be able to distinguish binder and non-binders and to rank known binders the highest.

The third application of molecular docking is to predict the absolute binding affinity between the protein target and the given conformation of a ligand. This is particularly important to the hit-to-lead and lead optimisation.[173] An accurate scoring greatly increases the optimisation efficiency and saves costs by correctly predicting the binding affinities of the modified ligands before the much more expensive step of ligand synthesis and experimental testing.

Most modern molecular docking methodologies consider the backbone of the protein target rigid, with partially flexible chains. The more flexibility is considered in the process, the more computational time and resources are required. The docking procedure involves sampling over many degrees of freedom.

For each ligand, a certain number of different conformations is generated, oriented, fitted, energy minimised, and energy scored. This number is user-defined, but it may vary from hundreds to many thousands. The number of orientations is also user-defined and typically within thousands. The resulting binding poses with lower energy scores (more favourable binding energy) are selected for further studies.

To evaluate the energies of binding poses resulting from the docking, scoring functions are used. Scoring functions are simplified descriptors of free binding energy. The appropriate scoring function would rank the experimentally determined binding modes the highest (lowest energy, most favourable).

Scoring functions can be grouped into three classes: force field-based, knowledge-based, and empirical scoring functions. Force field based scoring functions are developed based on the classical molecular mechanical force fields, i.e. physical atomic interactions,[174] including van der Waals interactions, electrostatic interactions and bond stretching/bending/torsional forces. Empirical scoring functions estimate the binding affinity of a complex on the basis of a set of weighted energy terms obtained empirically.[175] Compared to the force field scoring functions, the empirical scoring functions are much faster to calculate due to their simplified energy terms. A third kind of scoring functions are knowledge-based scoring functions, which employ energy potentials that are derived from the structural information embedded in experimentally determined atomic

structures.[176] Moreover, quantum mechanical (QM) and semi-empirical QM (SQM) based scoring functions have been recently designed to capture the binding affinity trend and native pose identification.[177,178]

To improve the quality of predictions, sometimes a combination of different scoring functions is used and weighed to give a new scoring value or rescore previous results. (Figure 21).

Docking is divided into stages, as illustrates in the diagram (Figure 21). For each stage, multiple methods are available.

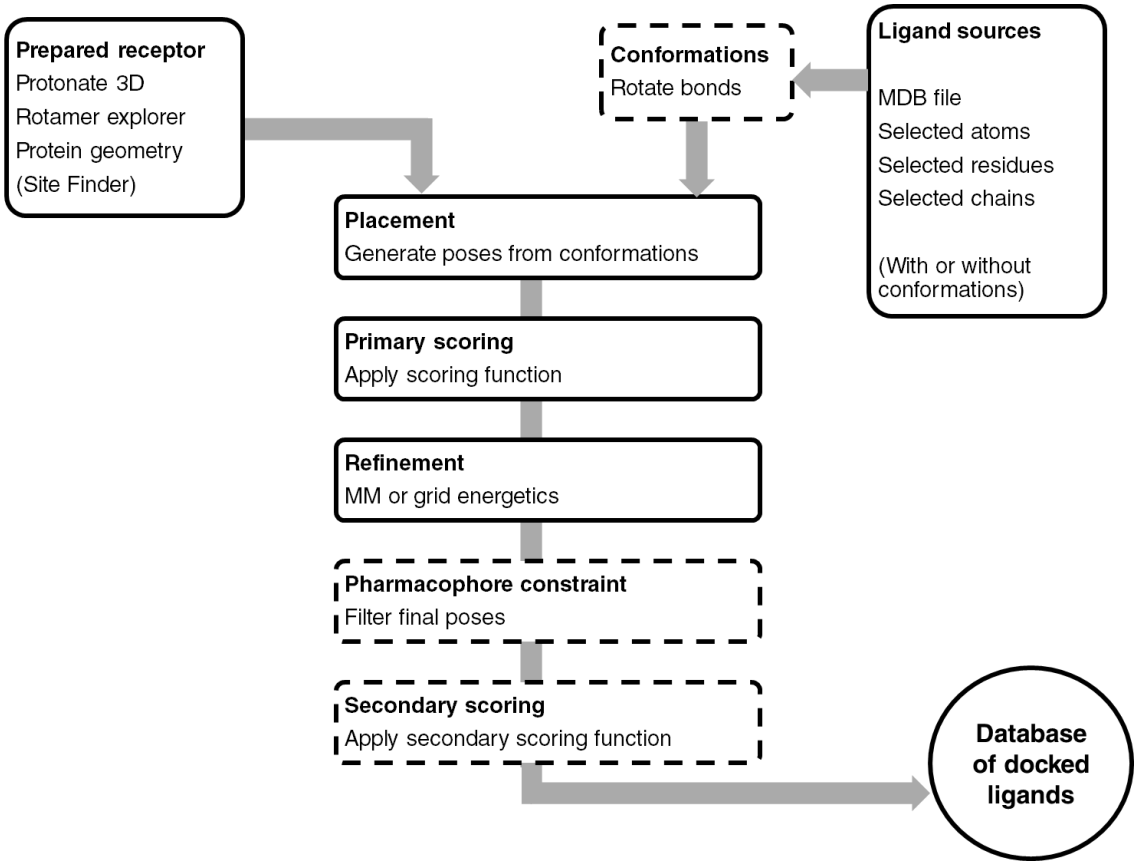


Figure 21 Molecular docking protocol to follow for a regular virtual screening. Dashed boxes represent optional steps considering the purpose of the procedure or the software used.

Despite a large number of comparative studies, it is still impossible to determine which programme and protocol are the best. Many studies have shown that success in molecular docking depends heavily on a number of factors such as the scoring function, the nature of the studied target, input docking and/or the metrics used to determine the study success. Comparisons between studies may

result in contradictory conclusions.[179] In this work, three different docking software packages were used: MOE-Dock[180], UCSF DOCK[181] and AutoDock4[182]. These three packages are a selection of some of the most popular applications. Each docking software has implemented its own placement methodology and scoring functions which differs from the other to a greater or lesser extent. With respect to the most commonly used package applied in this work (MOE-Dock), the docking procedure is divided into four main components: ligand-conformation generation, optional pharmacophore filtering, ligand placement and scoring in the pocket and flexible receptor and ligand refinement with re-scoring. The generation of ligand conformations is accomplished by supplying a collection of prepared ligand conformations generated using the Conformation Import application to the docking engine. The maximum number of outputted conformations is set to 10 000 by default. Then, using the MMF94x force field[183–187] the resulting ensemble is energy minimised, and partial charges are assigned to the atoms. Using the Triangle Matcher protocol, which defines the active site using α-spheres[188] similar to the spheres generated in UCSF DOCK (SPHGEN), ligand placement takes place. For AutoDock, a Lamarckian Genetic Algorithm (LGA) approach is typically used for globe pose sampling.[189] The top 1000 poses produced from placement are then scored using the London ΔG scoring function[190] (Equation 23)[180]:

$$\Delta G_{LdG} = c + E_{flex} + \sum_{h-bonds} c_{hb} f_{hb} + \sum_{metal-lig} c_m f_m + \sum_{atomsi} \Delta D_i \qquad \text{Equation 23}$$

c, c_{hb} and c_m are constants that have been trained over 400 protein ligand complexes. E_{flex} is a topological estimate of ligand entropy. Both f_{hb} and f_m are measures of geometric imperfections of protein-ligand and metal-ligand interaction. ΔD_i is the desolvation energy term which is approximated using a volume integral London dispersion. The top conformations (number defined by the user) are kept and minimised using MMF94x within a rigid receptor. The resulting poses are the scored using the generalized-Born volume integral/weighted surface area (GBVI/WSA dG) scoring function[180] in a flexible receptor (optional) (Equation 24).

$$\Delta G_{Binding}^{Calc.} = \alpha\left(\frac{2}{3}\left(E_{Inter}^{Coul.} + \Delta G_{Bind}^{R}\right) + E_{Inter}^{vdW} + \Delta G_{Bind}^{npsol}\right) + c \qquad \text{Equation 24}$$

$E_{Inter}^{Coul.}$ and E_{Inter}^{vdW} correspond to the coulombic and van der Waals contribution to binding respectively. The electrostatic solvation contribution, ΔG_{Bind}^{R}, is the change in reaction field energy upon binding. Reaction field energies are calculated using the generalized Born/volume integral implicit solvent model (GB/VI)[191], which estimates the free energy of hydration as a classical electrostatic energy plus a cavitation energy using a volume integral London dispersion energy. The ΔG_{Bind}^{npsol} term represents the change in non-polar solvation (van der Waals and cavitation cost) upon binding.[180] Instead of the double-scoring process, UCSF DOCK and AutoDock rely on a single scoring process with different procedures. In both cases the accessory program GRID[192] is used to pre-compute the energy interaction between a dummy probe atom and all receptor atoms on a 0.3 Å resolution grid within the area of study. Afterwards, every ligand pose is evaluated with each own scoring function. AutoDock uses a semiempirical free binding energy force field scoring function, while DOCK 6 uses a force field based one. The default conditions applied for each methodology are summarised in Table 8.

Table 8 Standard conditions employed in the different docking programs

	Placement methodology	**Scoring Function I**	**Scoring Function II**
MOE	Triangle Matcher	London dG	GBI/WSA dG
AutoDock	Lamarckian Genetic Algorithm	AutoDock 4 Scoring Function (semiempirical free energy force field)	
DOCK6	Sphere Generation	DOCK 6 Scoring Function	

4.3.1 Analysis of the molecular docking results

4.3.1.1 Cluster analysis

Hundreds or even thousands of poses can be calculated for a single ligand in every docking calculation. In the case of a blind docking experiment, where the

binding site is unknown and the whole protein is selected as a receptor, the analysis of the molecular docking results can be challenging. Not only conformations are evaluated but also their location of binding. A frequently used resource to facilitate the analysis is to clusterise the scored conformations within an RMSD value (2.0-4.0 Å)[182]. Most populated clusters mean that the visited cavity could be more druggable than its counterparts.

4.3.1.2 Ligand interactions

Ligand-receptor interactions are analysed to determine which could be the key residues for a good interaction with the receptor, or if the docked ligand shows the same interactions as the crystallised one. Such interactions include hydrogen bonds, hydrophobic interactions and solvent interactions.[193,194] For an easier visualisation, some docking software packages like MOE provide a two-dimensional diagram of the ligand interactions with the receptor residues (Figure 22).

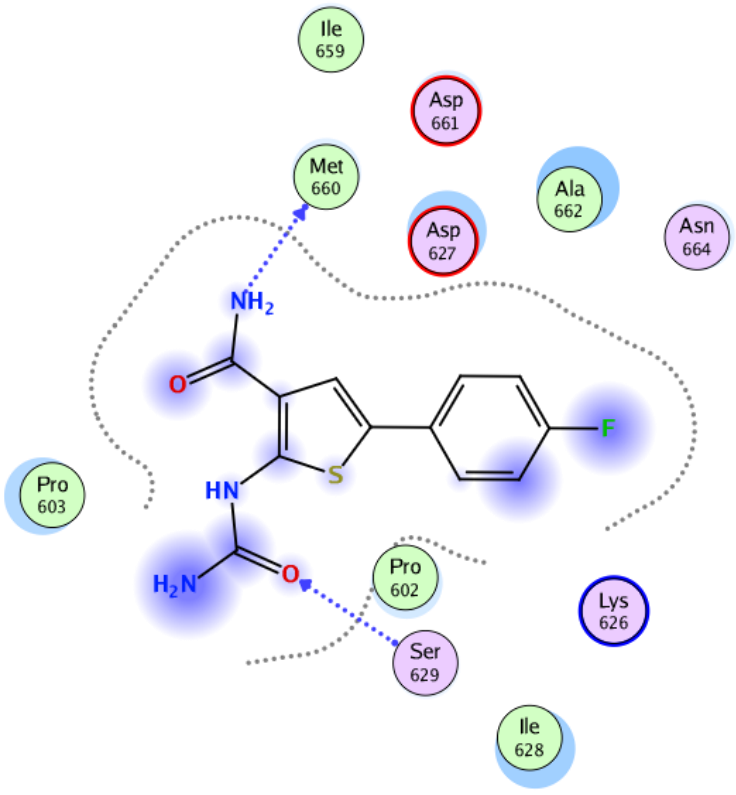


Figure 22 Ligand interaction map for a TPCA-1 docked conformation

4.3.2 Validation of the molecular docking results

As mentioned before, the number of molecular docking packages available to date is very extensive. The user has many flavours to choose which could lead to very different results. Before performing any virtual screening on a protein target (with a known binding site) it is general practice to validate the package of choice. Molecular docking is performed with the same crystallised ligand in order to replicate the crystallographic conformation. If that conformation is between the top scored ones, that is an indicator that the applied docking package is likely to provide trustworthy results. Decoys (molecules known to not bind in the region of study) are often used to evaluate the ratio of false positives that the used docking programme could encounter[195].

4.4 MM-PBSA

The molecular mechanics – Poisson-Boltzmann surface area (MM-PBSA) approach is used to calculate the free energy difference between two states, typically the bound and unbound state of two solvated molecules, or to compare the free energy of two different solvated conformations of the same molecule.[196]

The overall objective is to calculate the absolute binding free energy for the non-covalent association of any two molecules, A and B, in solution (Equation 25):

$$[A]_{aq} + [B]_{aq} \Leftrightarrow [A^{\neq}B^{\neq}]^{\pm}_{aq} \qquad \text{Equation 25}$$

$[A]_{aq}$ refers to the ensemble of molecule A free in solution, $[B]_{aq}$ refers to the ensemble of molecule B free in solution, and $[A^{\neq}B^{\neq}]^{\pm}_{aq}$ represents the complex formed from molecules A and B, considering any structural changes and the solvent reorganisation ($aq^{\pm}$) that may occur upon the complex formation.

Free energy is a state function, meaning that the free-energy difference associated with a given event like a drug binding to its protein target is determined only by the energy prior to that event and the energy following it. In other words, to calculate the binding free energy of a ligand to a protein, the ligand-protein complex, needs to be “compared” to both the ligand and the protein as separate entities in the solution.

Ideally the free energy of binding would be calculated directly as shown in Figure 23.

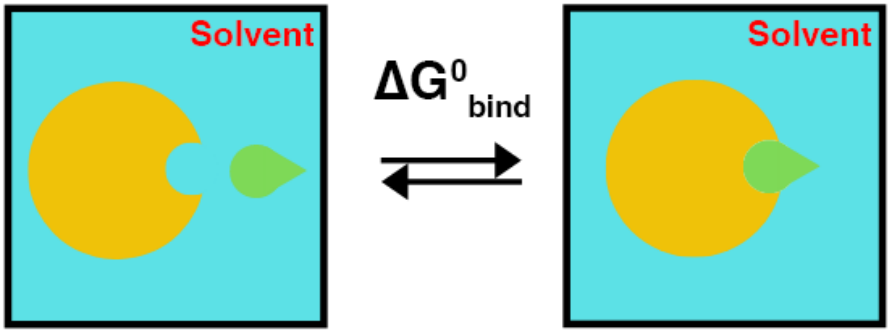


Figure 23 Adapted diagram of binding free energy between of a ligand to a protein[196]

However, calculating free energies can usually only be done using small steps according thermodynamic cycle (Figure 24)[143]:

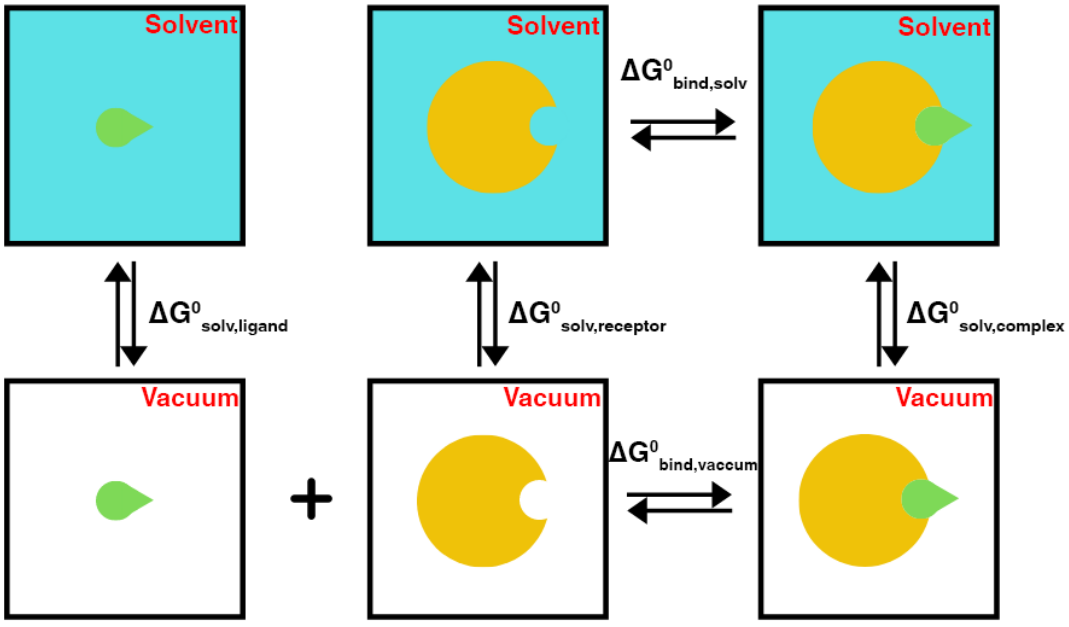


Figure 24 Adapted diagram of the thermodynamic cycle used to calculate the binding free energy[196]

From this diagram (Figure 24), the binding free energy can be calculated by Equation 26.

$$\Delta G^0_{bind,solv} = \Delta G^0_{bind,vaccum} + \Delta G^0_{solv,complex} - \left(\Delta G^0_{solv,ligand} + \Delta G^0_{solv,receptor}\right) \quad \text{Equation 26}$$

In the MM-PBSA approach, different contributions to the binding free energy are calculated in the following ways:

- Solving the linearised Poisson-Boltzmann or Generalised Born equation for each of the three states and adding an empirical term for hydrophobic contributions, solvation free energies are calculated (Equation 27)[143]:

$$\Delta G^0_{solv} = G^0_{electrostatic,\in=80} - G^0_{electrostatic,\in=1} + \Delta G^0_{hydrophobic} \quad \text{Equation 27}$$

- Obtaining ΔG^0_{vaccum} by calculating the average interaction between a protein and a ligand. If necessary, the entropy change upon binding is taken into account[143] (Equation 28):

$$\Delta G^0_{vaccum} = \Delta E^0_{molecular\ mechanics} - T \cdot \Delta S^0_{normal\ mode\ analysis} \quad \text{Equation 28}$$

Often, entropy contributions are neglected, if only a comparison of states of similar entropy is desired, such as two similar ligands binding to the same protein. The reason is that the entropy calculations using normal mode analysis or quasi-harmonic approach are computationally expensive and are associated with large errors that introduce significant uncertainty to the results. The average ligand-protein interaction energies are usually obtained by performing calculations on a group of uncorrelated snapshots collected from equilibrated MD simulation trajectories.[143] Although MM-PBSA is a popular approach to estimate the free binding energy of small ligands to biological macromolecules its accuracy is not excellent. This method is very sensitive to the solute dielectric constant and contain several questionable approximations such as lack of conformational entropy and information about the number and free energy of water molecules in the binding site.[197] There are several benchmarks assessing the performance of

this method.[198,199] In this work, MM-PBSA has been calculated using GROMACS via the *g_mmpbsa* module.[196,200]

Chapter 5 Development of the Cosolvent Analysis Toolkit (CAT)

In this part of the work, the focus was on the development of an analysis tool that helps the identification of binding hotspots resulting from cosolvent MD. The main goal was to design a platform able to identify newly formed sites with a user-friendly analysis. The result is the Cosolvent Analysis Toolkit (CAT). CAT has been designed as an open-source analytical platform, compatible with commonly used molecular graphics software packages such as UCSF Chimera and VMD[201,202]. CAT incorporates two types of analysis: identification and ranking of the entire 'hotspots', and identification and ranking of the molecular fragments suitable for targeting those 'hotspots'. The former serves as a general detector and can be readily used to guide structural biology experimental efforts, while the latter brings useful information about the inhibitor/ligand design from the structure-guided standpoint.

5.1 Scoring function development

To create a robust analytical method for reliable detection of molecular hotspots, the development of a scoring function was required. From a molecular interaction standpoint, such scoring function should include three characteristics: calculation of the intrinsic interaction energy between the protein and a cosolvent (probe) molecule, and two normalizing factors: retention time of the probe at the binding site, and the overall depth of the binding site relatively to the protein surface. With this selection of features, we attempt to identify regions with better probe-protein interaction as well as other geometrical features that would deem the detected hotspot as "druggable". Therefore, the scoring function per residue can be written as follows (Equation 29):

$$S_{Residue} = S_{Interaction} S_{Stability} S_{Depth} \qquad \text{Equation 29}$$

To calculate the interaction scoring part per residue in the protein, CAT defines a sphere surrounding the geometric centre of each residue (dashed blue circle in

Figure 25). Hence, the interaction energy between the protein and every probe inside the sphere is calculated. To avoid atomic clashes, softcore potentials[203] were used, as described in Equation 30.

$$\Delta E_i = E_{LJ} + E_{coulumb} \qquad \text{Equation 30}$$
$$= 4\varepsilon\left[\left(\frac{\sigma}{(r+\delta_{lj})}\right)^{12} - \left(\frac{\sigma}{(r+\delta_{lj})}\right)^{6}\right] + \frac{Kq_iq_j}{(r+\delta_{elec})}$$

Where r is the interatomic distance, ε corresponds to the depth of the Lennard-Jones potential, σ is the finite distance to the zero potential, K is the Coulombic constant in kcal/mol , δ_{lj} and δ_{elec} are the softcore deltas for the Lennard-Jones potential and Coulombic potential respectively[204].

Within the assigned sphere, the average number of cosolvent molecules <M> inside the sphere can be calculated. From a simulation trajectory with N frames, $S_{interaction}$ can be calculated as the ratio between the average interaction energy though the trajectory and the average number of molecules inside the sphere (Equation 31):

$$S_{interaction} = \frac{1}{< M >}\sum_{1}^{N}\frac{E_i}{N} \qquad \text{Equation 31}$$

For the stability score, which quantifies the retention time of the probe at the binding site, the RMSD of the total number of cosolvent molecules $\sqrt{\Delta M^2}$ inside the sphere was used in (Equation 32):

$$S_{stability} = \frac{(1-\sqrt{\Delta M^2})}{(< M > -\sqrt{\Delta M^2})} \qquad \text{Equation 32}$$

$S_{stability}$ values range from 0 to 1, allowing the highest values for low variance, representing more stable interactions and molecules being retained for a longer time.

For the third scoring term, which describes the overall depth of the binding site relatively to the protein surface, CAT counts the number of protein atoms ($J_{Contacts}$) inside each residue sphere (Figure 25), assigning to it a volumetric score S_{Depth}. Afterwards, it is normalised to the highest scored residue, to set the range between 1 and 0, as showed in Equation 33:

$$S_{Depth} = \frac{< J_{Contacts} >}{MAX < J_{Contacts} >} \qquad \text{Equation 33}$$

To define the regions, dummy atoms are created for each residue in its corresponding centre of geometry, with its respective $S_{interaction}$ (Equation 27) assigned to it.

To define binding regions, CAT systematically scans through the protein backbone, defining a new spherical region (Figure 25) which clusterises the dummy atoms. This "CAT cluster" has a S_{Region} assigned as (Equation 34):

$$S_{Region} = \frac{1}{N_{Residues}^{Inside}} \sum_{Inside} S_{Residue} \qquad \text{Equation 34}$$

CAT outputs a PDB file with dummy atoms highlighting the areas of interest regarding the $S_{residue}$ per residue and S_{Region} per region, ranked from the best (most likely) to the worst.

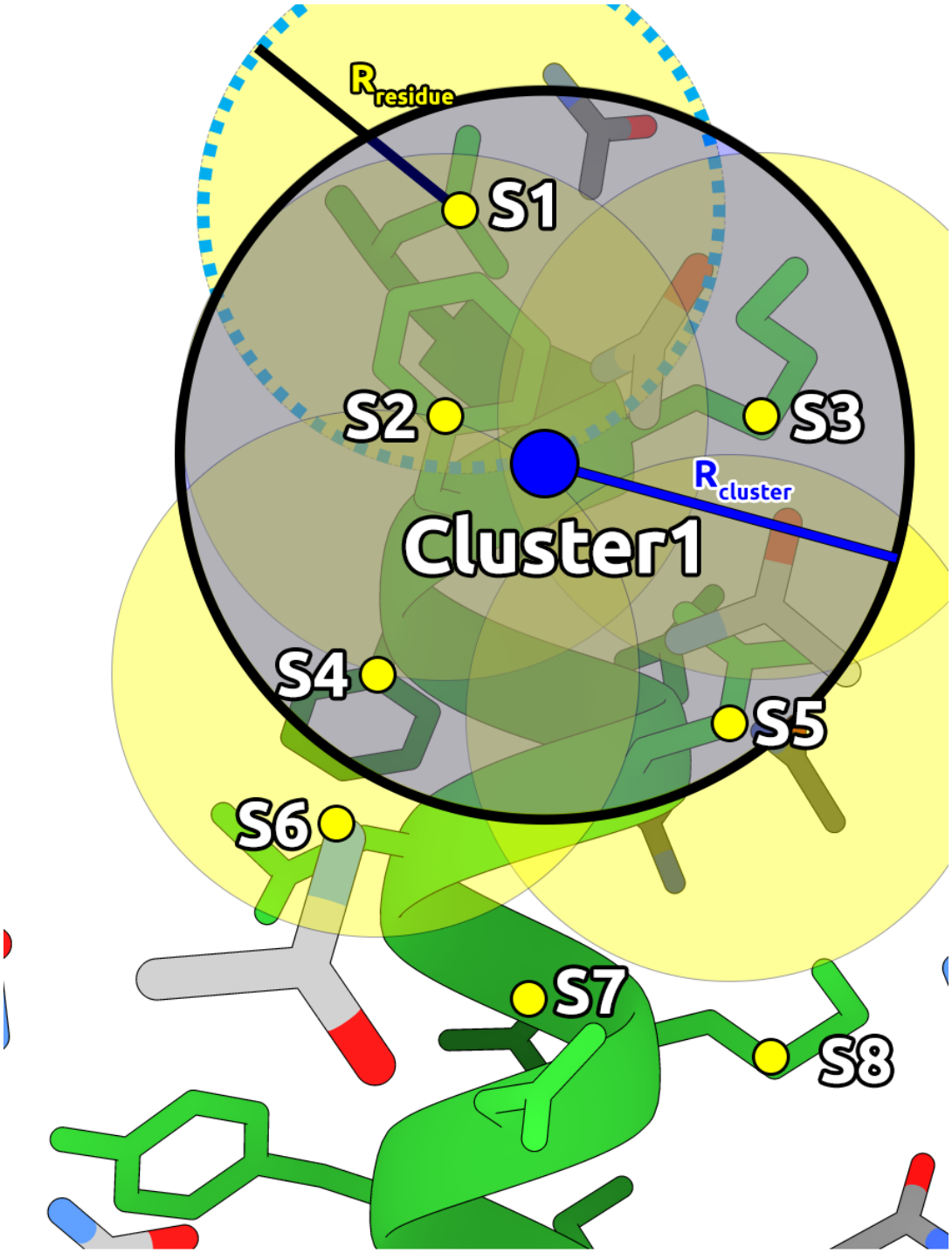


Figure 25 Clustering scheme of CAT: A sphere is generated per residue, which encapsulates shells of interacting comolecules (yellow circular regions defined by the variable Rresidue). Afterwards, a secondary clustering region (blue shaded area, defined by the variable RCluster) defines close side-chains centres of geometry, resulting in a series of representative clusters of interest.

In this study, values for the electrostatic and Lennard-Jones softcore delta were scanned (Appendix). The best result was attained with deltas set to 1 Å. The

sphere radius for the residue–cosolvent interaction was set at 8Å, to incorporate approximately 3 solvation shells. The clustering sphere radius was set to 5 Å, which encapsulated inter Cα distances for different secondary structure motifs.

5.2 Probe selection

Five probe molecules: acetamide, benzene, acetanilide, imidazole, and isopropanol were chosen based on three criteria. First, the set has a broad range of solubility characteristics, going from highly hydrophobic molecules (benzene) to more hydrophilic molecules (acetamide). Second, all probes are widely used as crystallisation co-factors, probes employed in fragment-based drug discovery (FBDD) efforts, and as moieties present in known small molecule ligands. Third, it is a set validated in previously reported studies on allosteric hotspot mapping[47,74,205,206].

5.3 Benchmark

To benchmark the method, the aim was to select proteins with reported crystallographic structures of their orthosteric binding site with more than one reported allosteric site. Furthermore, proteins that have been studied in benchmarks including cosolvent methodologies have been taken into special account[207]. After a careful curation, four structurally diverse targets were selected: the ligand-binding domain of androgen receptor (AR-LBD), protein-tyrosine phosphatase 1B (PTP1B), GTPase HRas and cyclin-dependent kinase 2 (CDK2) with novel allosteric sites recently described[43]. Benchmark structures with PDB codes are listed in Table 9.

Table 9 PDB codes of the crystal structures used for our benchmarking, codes highlighted in bold correspond to the structures used for the cosolvent simulations

Molecule	Starting structure	Benchmark structures
AR ligand binding domain (AR-LDB)	**2PIO**	2PIQ, 2PIR, 2PIT, 2PIU, 2PIV, 2PIW, 2PIX, 2PKL
Protein-tyrosine phosphatase 1B (PTP1B)	**1XBO**	1T4J[208], 1T48, 6B95[209]
GTPase HRas (HRas)	**1P2S**	1P2T, 1P2U, 1P2V, 3K8Y, 3K9L, 3K9N, 3RRZ, 3RS0, 3RS2, 3RS3, 3RS4, 3RS5, 3RS7
Cyclin-dependent kinase 2 (CDK2)	**4EK3**	6Q3C, 6Q3B, 6Q3F, 6Q49, 6Q48,6Q4B, 6Q4A, 6Q4C, 6Q4D, 6Q4F, 6Q4E, 6Q4J, 6Q4I, 6Q4H, 6Q4G, 6Q4K.

5.4 Results

An in-depth study of the molecules used to test the accuracy of the CAT scoring function and its corresponding ranking has been done in this study. The obtained results were directly compared to the FTMap webserver, a robust, powerful and widely popular 'hotspot' detecting tool[47,57]. The comparison concluded that the explicit solvent/cosolvent interactions and MD sampling were crucial for the right assessment of cryptic binding sites, and CAT scoring function reliably detected, filtered and correctly ranked "druggable" regions.

5.4.1 Androgen receptor ligand binding domain (AR-LBD)

The androgen receptor is a multimeric DNA-binding transcription factor that regulates expression of genes critical for the development and maintenance of the male sexual phenotype.[210] Through its ligand binding domain (LBD) it binds to male steroid hormones such as testosterone, androsterone, or dihydrotestosterone; the binding event occurs at the internal ligand binding pocket (LBP).[207] Furthermore, the presence of auxiliary allosteric binding sites has been reported in two of the solvent exposed regions of the protein: at the activation function 2 (AF-2) between helices 3 and 4 and at the binding function 3 (BF-3) close to helix 9.[207] (Figure 26)

The average structure with its respective CAT clusters have been superimposed to a series of experimentally-solved structures with bound ligands in orthosteric and allosteric regions (Table 9).[207] Binding poses of these ligands and their corresponding interactions with protein residues have been considered in the analysis. CAT detected both allosteric regions with fragments interacting with some of the key residues interacting with the crystallised ligands, as shown in Figure 26. At the allosteric AF-2 binding site (Figure 26), several highly ranked clusters were mapped mainly in helix 3(H3), including key residues K720 and V716 involved in hydrophobic interactions with the ligand, as depicted Figure 26.c. This was consistent with the values of energy scores, as highest-ranked clusters in that area corresponded to fragments with hydrophobic/aromatic probes such as benzene and acetanilide. The smaller polar probes also detected the H3 area, albeit with a cluster rank (Table 10). Probes interacted with R726 and N727, two very flexible residues that enclosed or open the allosteric pocket. As assessed by visual inspection and covariance analysis (Appendix), the shape of the pocket considerably varied, tuned by the behaviour of these two residues, which acted as gatekeepers. Considering the small size of this binding site, the success of detection of this area as a potential “druggable” hotspot was very encouraging. CAT identified regions that not only interact with a couple of helix 3 residues (Figure 26), but with the majority of the residues within this site and surrounding sidechains that could contribute to the further pocket opening. The

interaction of the probes with helix 4 was not as favourable as with other areas, as no highly ranked clusters were found close to it.

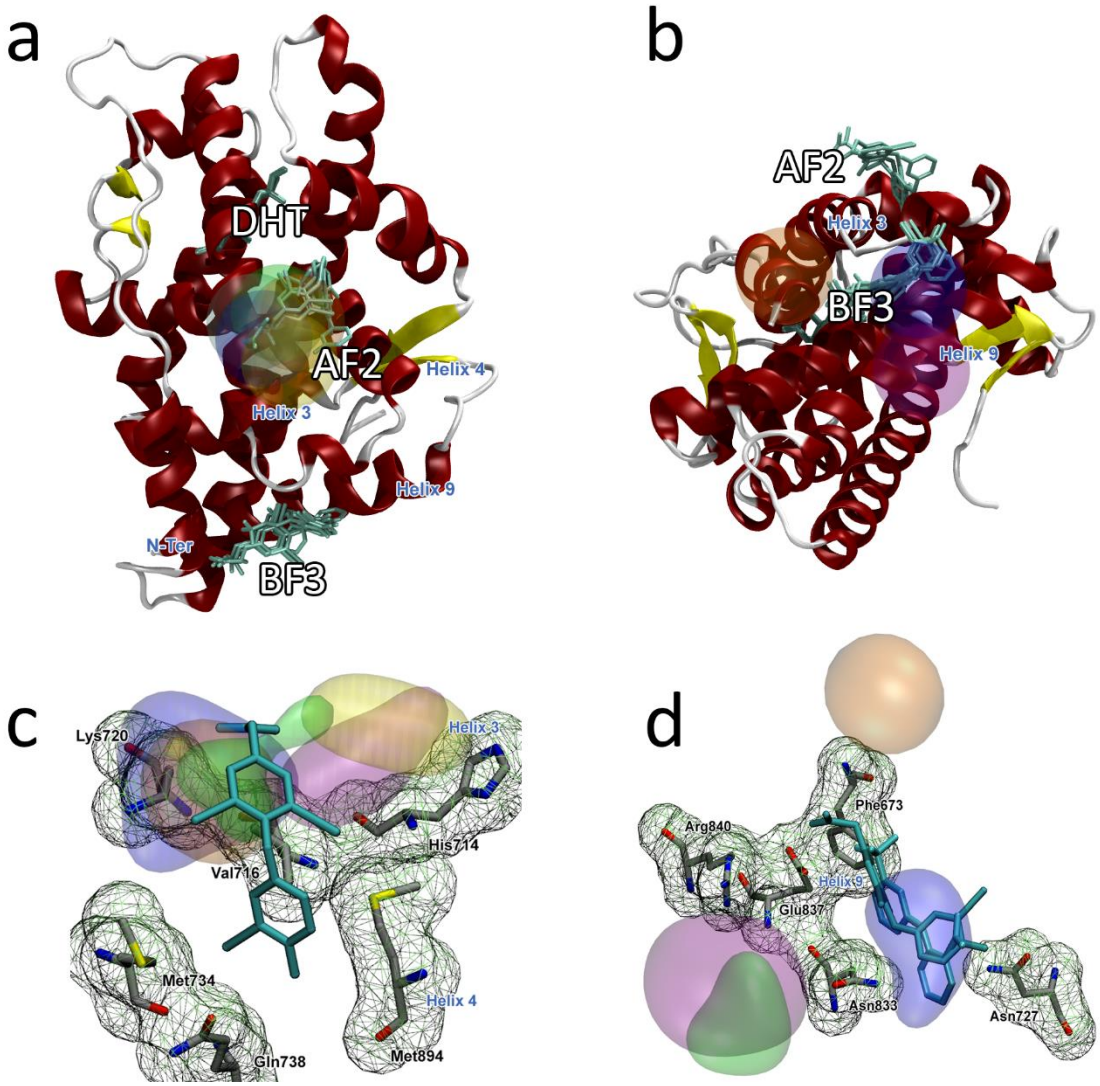


Figure 26 Androgen receptor LBD hotspots found by CAT. Clusters have the following colours assigned: acetamide – blue, benzene – purple, acetanilide – orange, imidazole – yellow, and isopropanol – green. The crystallographic ligand is coloured cyan. A) Panoramic representation of LBD domain centred on the AF-2 site compromised around the H3 and the respective top cluster given by CAT; B) Panoramic representation centred around the BF-3 region and the respective CAT clusters. Simulations with all five probes found the site with a high rank, as described in Table 10. For the second site, only acetamide and benzene show high ranks. C) AF-2 site and its key residues; K720, V716 and H714, that form part of H3, are detected by simulations with all five probes. D) BF-3 and its key residues; simulations with acetamide detected N833 and N727 as key residues for the site, but with a lower ranking than the clusters found in AF-2 site.

For the BF-3 allosteric binding site (Figure 26), CAT also gave satisfying results when compared with the experimental data. Although there was a higher number of CAT clusters in this area, especially around H9, the scoring rank of them was worse than that obtained for the AF-2 site. Key residues contributing to the binding site architecture, such as R840, Y834, G829 and F826, presented a CAT cluster (Figure 26.d). In this case, acetamide was the fragment with the highest affinity to the area. Enclosing the site, CAT also detected interactions with the N-terminal area at F673 and the “gatekeeper” residues from AF-2 site: R726 and N727. To summarise, all clusters detected within AF-2 included the majority of the residues from this binding site. These residues are listed in Table 10.

Table 10 AR-LBD CAT results and comparison with FTMap

Target	Binding Site	Protein Contacts	Cosolvent	CAT Rank	Found by FTMap?
AR-DBD	Orthosteric	E706, V746, R752, F764, H874, F878	Not found	-	✓
	AF-2 Allosteric	I672, F673, V716, K720, P723, G724, N727, K734	Acetamide	9	X
			Acetanilide	1,2,9	
			Benzene	1	
			Imidazole	4	
			Isopropanol	2	
	BF-3 Allosteric	F826, E829, Y834, E,833, R840, E897	Acetamide	1,3	X
			Acetanilide	6,7	
			Benzene	10	
			Isopropanol	8	

The crystal structure of the AR dimer has been recently reported,[211] where the interactions between the AR monomers could be observed (PDB code: 5JJM). These interactions are crucial for the DNA binding and disrupting them could be a novel way to inhibit the protein. Interestingly, parts of the region involved in protein-protein interactions were detected by CAT along the dimerisation interface. This validates the applicability of CAT in mapping of novel and unique superficial interaction hotspots, which are very challenging to be detected by established methods, such as FTMap.

The only reported AR binding site that was not detected by CAT was the orthosteric one. This site, which is a deep pocket binding dihydrotestosterone (DHT), was too buried inside the protein core and shielded from the surface for the cosolvent molecules to detect it. The opening of this pocket would require large conformational changes and thus simulations longer than performed in this study. It is very likely that in the timescales required some of the probes would undergo phase separation, which is not desirable in CAT analysis and may lead to observing artefacts. We believe that the use of repulsive potentials for probes in combination with longer simulations would improve the identification of this binding spot.

Comparison between CAT and FTMap outcomes showed some interesting results, as the latter covers what CAT misses. FTMap identified the orthosteric binding site as a potential druggable hotspot (Table 10), as the highest populated clusters were mapped to that site. However, FTMap failed to identify the allosteric binding sites: only one sparsely populated cluster is placed at the AF-2 site and none at the BF-3 site (Figure 26). Moreover, unlike CAT, FTMap did not identify any dimer-forming regions of AR as potential hotspots. Therefore, an apparent strength of CAT is to reliably detect the hotspots that are challenging to FTMap.

5.4.2 PTP1B

Tyrosine-protein phosphatase non-receptor type 1 (PTP1B) is a negative regulator of the insulin signalling pathway. It has emerged as a promising drug target for obesity and type II diabetes mellitus[212]. Numerous potent PTP1B inhibitors have been discovered during last years, unfortunately nearly all

medicinal chemistry efforts have been hampered by lack of selectivity and inhibition of related proteins, especially T-cell protein tyrosine phosphatase (TCPTP)[212].

PTP1B orthosteric binding site is formed by three loops: the WPD with W179, P180 and D181, a phosphotyrosine (pTyr) loop including Y46, and a Q loop with G262[213]. An allosteric site (BB site) has been discovered by X-ray crystallography[208], which has paved a new path to design selective PTP1B inhibitors. This site is located between helices 3 and 6, and it includes residues L192, A193, F196, E276 and F280.[208] (Figure 27). Allosteric sites and other binding events have been identified by the means of multitemperature crystallography, fragment screening, and covalent tethering[209]. This last study included more than hundred crystal structures and different binding events. In this study we focused focus in the two newly identified allosteric sites: the allosteric 197 site, close to the previously known BB allosteric site, and the loop 16 (L16) site (Figure 27).

CAT analysis for the cosolvent MD simulations in the apo/open state (PDB code: 1XBO) identified both binding sites: orthosteric and allosteric. For the orthosteric site, all probe molecules tested interacted with various regions of the site. As showed in Figure 27.a and Figure 27.b, imidazole mapped all regions of interest: WPD- , pTyr- , and Q- loops. Isopropanol interacted preferentially with the WPD loop, while acetamide, acetanilide, and benzene interacted with the pTyr loop residues. For the BB allosteric site, CAT placed clusters for all probes except imidazole, with clusters centred at the binding site (Figure 27.c). Helix 3 was mapped in its entirety, as it was the helix 4 region that comprised the pocket along with its key residues. The close proximity of the 197 site to the BB site might have induced some bias to to CAT clusters, as both pockets share residues. Although both pockets might be included in the same cluster, the 197 site was mapped by CAT, mainly by acetanilide and benzene. Interestingly, most of the clusters from this pocket included K197, the mutated residue reported by Keedy and coworkers in their study on the “drugabbility” of this pocket.[209] Regarding the L16 site, CAT placed a series of highly ranked clusters close to the binding site, but in direct

contact with just one, two, or no pocket residues. Nevertheless, the level of mapping was sufficient to determine the area as a potentially “druggable”.

Table 11 PTP1B CAT results and comparison with FTMap

Target	Binding Site	Protein Contacts	Cosolvent	CAT Rank	Found by FTMap?
PTP1B	Orthosteric	Y46, W179, P180, D181, G262	Acetamide	3	✓
			Benzene	5,6	
			Imidazole	6	
			Isopropanol	2	
	Allosteric	L192, A193, F196, E276, F280	Imidazole	1,5,7	X
			Acetanilide	1,3,10	
			Benzene	2	
			Isopropanol	3,6,7	
	197	R105, D148, K150, Y152, Y153, E157, N193, K197	Acetanilide	4,5,7	X
			Benzene	2,6	
			Isopropanol	5,6	
	L16	K237, K239, S242, I281	Acetanilide	1	X
			Benzene	1	
			Imidazole	10	

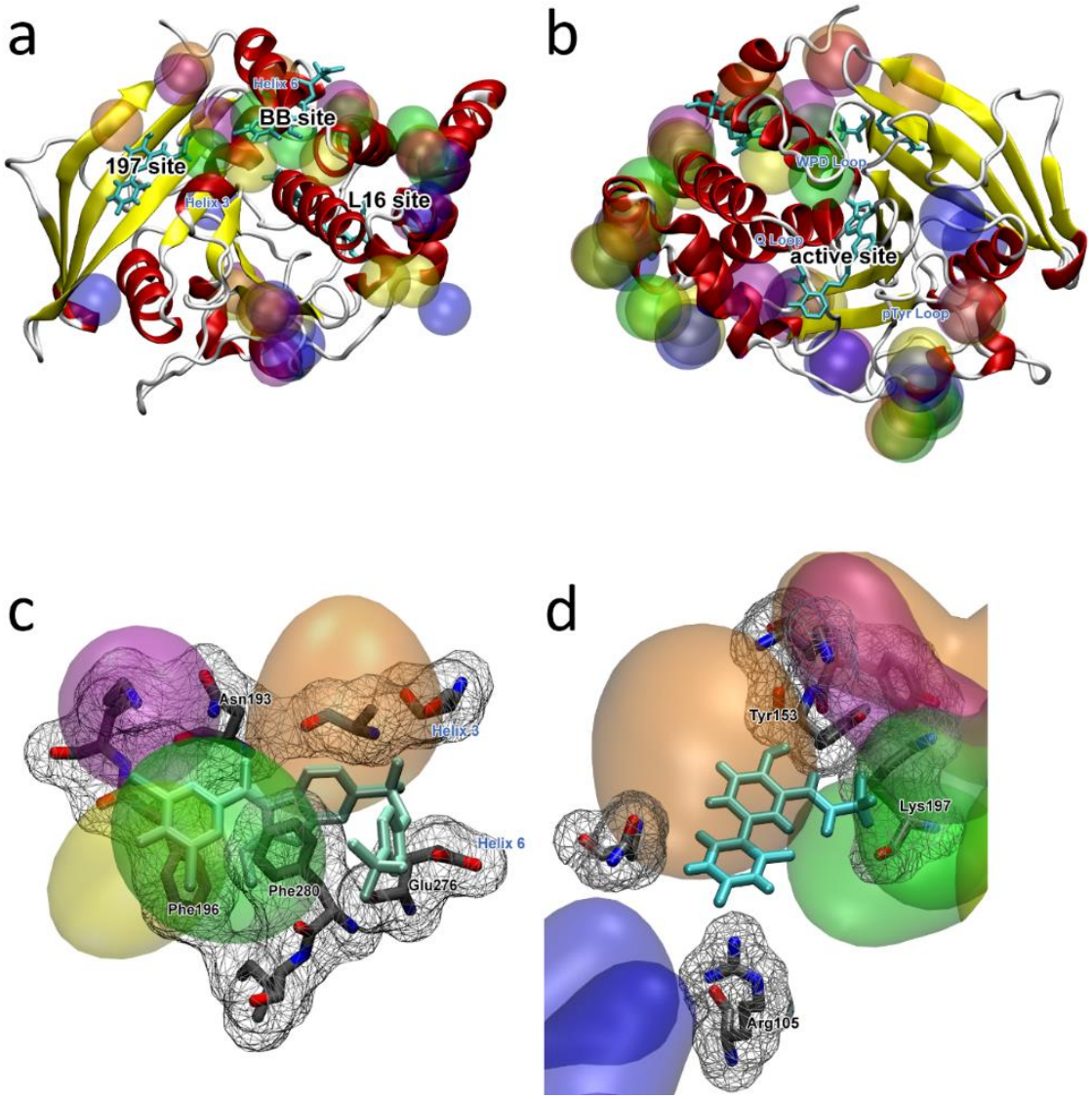


Figure 27 PTP1B hotspots found by CAT. Clusters have the following colours assigned: acetamide – blue, benzene – purple, acetanilide – orange, imidazole – yellow, and isopropanol – green. The crystallographic ligand is coloured cyan. A) Panoramic view centred on the allosteric binding sites; B) View centred on the orthosteric binding site. CAT performs well finding and scoring the binding site for PTP1B, since 4 out of the 5 probes are able to interact with the site residues. Unfortunately, only isopropanol and benzene find the orthosteric binding site, and acetamide interact with neighbour key residues. C) BB allosteric binding site and its main residues; all probes but acetamide rank cluster in the allosteric binding site, principally isopropanol, which shows interactions with N193, F196 and F280. D) 197 site recently identified by Keedy and coworkers[209] CAT mapped the whole site, including K197.

As showed in Table 11, FTMap has not been able to identify the allosteric binding site, which further validated CAT as an appropriate toolkit to detect the allosteric hotspots that are challenging to established methods such as FTMap. The

comparison between CAT and FTMAP shows a remarkable performance and robustness of the scoring function and the clustering method implemented in CAT. The drug-like small molecule bound at the allosteric PTP1B site reported by Wiessmann and coworkers using X-ray crystallography[208] showed that this binding site is a bona fide "druggable" site which could be used as starting point for a structure-guided design, and which has been validated in the follow-up drug discovery efforts[208]. As showed in Figure 27, CAT ranked the clusters at the orthosteric site high, yet it was not biased towards deep pockets, being able to report all experimentally detected pockets in the top-ranked 10 CAT clusters, which included the allosteric site undetected by FTMap.

5.4.3 Fragment hotspot screening – H-ras GTPase

The main difference of the three isoforms of the human Ras proteins, H-Ras, K-Ras and N-Ras, lies within the primary sequence of the hypervariable region and its post-translational modifications[214]. The catalytic G-domains of the three respective Ras proteins are highly conserved, with only a 10% average difference in primary sequence identity in the C-terminal lobe (residues 87 to 171)[215]. The N-terminal lobe 1 carries the catalytic binding site with all the G-domains switches[216] (Figure 28).

The "effector lobe" contains the small molecule binding sites of Ras, including the allosteric site consisting of residues R97, D107 and Y137 (denoted as the allosteric lobe)[217]. This allosteric site is connected to the active site in H-Ras by helix 3 (H3), one edge of the inter-lobe linker, and one of the switches of the N-terminal lobe at the other. This is showed in Figure 28.

Due to the sensitivity regarding the conformational changes of the H-Ras, the cosolvent MD simulations prior to the CAT analysis were run only in the "off" state, to enable the direct comparison with the reported experimental MSCS (Multiple Solvent Crystal Structure) results on the H-Ras[46,218]. The MSCS showed several hotspots formed in different regions of the protein in the "off" conformation. The CAT analysis detected several of these hotspots in highly-ranked clusters.

Two major ‘hotspots’ were identified for H-Ras: one found in the inter-lobe linker region, and another one in the allosteric lobe (Figure 28). Both hotspots involved H3 helix, but each of them was situated on either side of the helix. Cluster 1, as numbered in the study by Buhrman and coworkers[219], was located near to the active site, between H3 and switch II, showing R68 and Y96 as the major contributors. Several highly-ranked CAT clusters interacted with cluster 1 residues, mainly in helix H3. All probes but acetamide interacted with the key residues R68 and Y96. Although acetamide did not interact with these amino acids, it placed its highest-ranking cluster around a large region of H3. Cluster 2, found between helices H3 and H4, mapped to one of the largest hotspots. In this case, CAT interacted with both helix 3 and 4, with residues I93 and H94 from helix 3 and virtually all residues from helix 4. There were no acetamide clusters found around the pocket, which indicated that this region had a low affinity for highly polar moieties. Cluster 4 consisted of a pocket in the inter-lobe linker region very close to the nucleotide substrate binding site. It was comprised by D30 and K147; the latter being a target for ubiquitination on Ras-GTP[220]. There were only two CAT clusters that interacted with the residues from this pocket. Acetamide interacted with D30, while imidazole did with K147. Remaining clusters mapped to the pockets that overlapped with sites occupied by effector Ras binding (RBD) or cysteine-rich (CRD) domains and RasGAP[220].

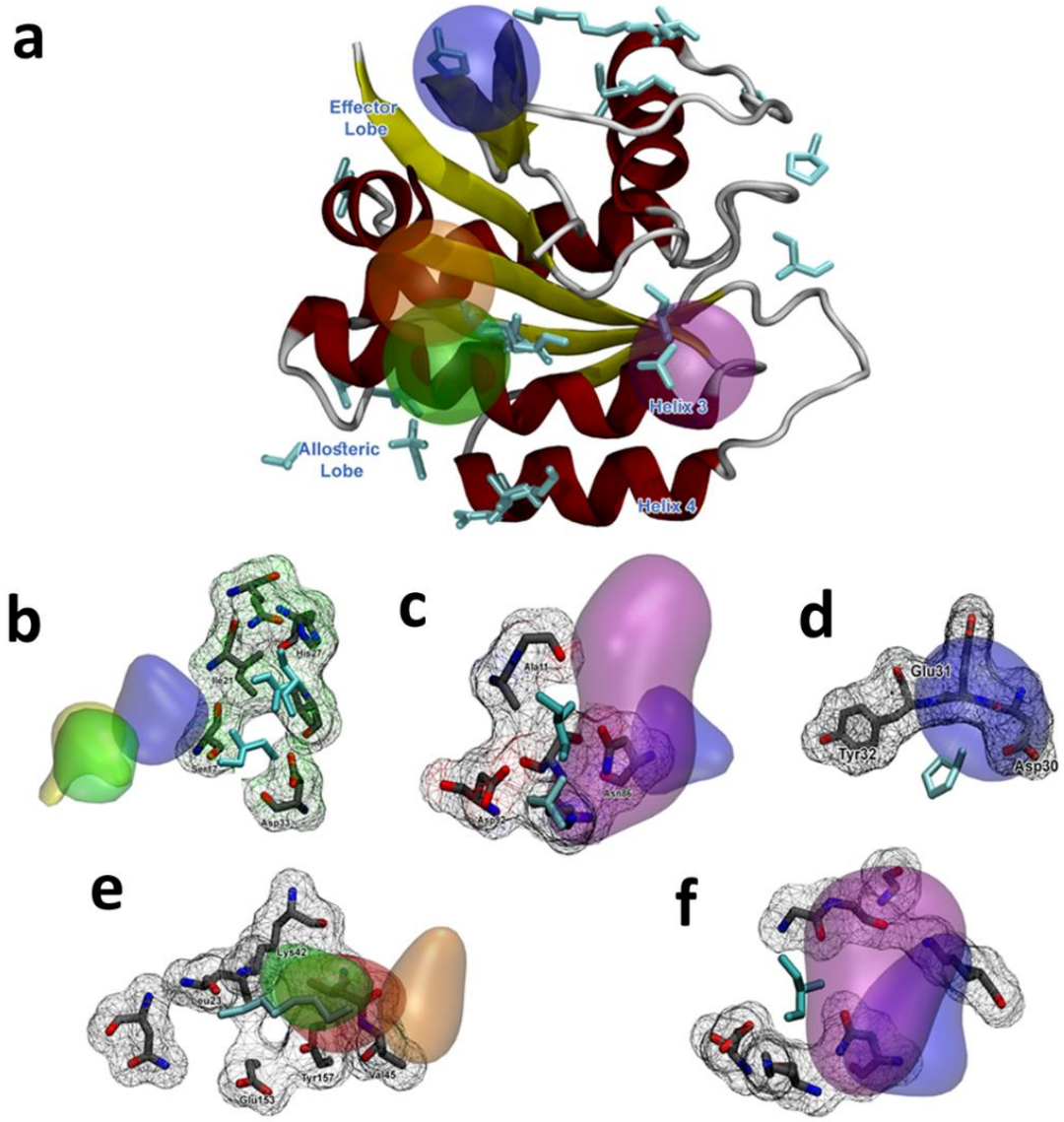


Figure 28 H-ras hotspots found by CAT. The clusters are coloured as follows: acetamide – blue, benzene – purple, acetanilide – orange, imidazole – yellow, and isopropanol – green. The crystallographic fragment is coloured cyan. A) Panoramic view of the H-ras and the highest ranked cluster for each cosolvent molecule. A) Depiction of Site 3, B) Site 5, C) Site 6 D) Site 7 and E) Site 8, Following the naming and numbering from Buhrman and coworkers [219]. As shown, acetamide and benzene performed better than the other three probes, but the combination of the five different probes found most of the superficial binding sites and CAT score found the interacting residues to different crystallised molecular fragments.

At the inter-lobe linker region and at the region overlapping with Raf-CRD, CAT has mapped cluster 7. Clusters 3 and 6 overlapped with the RasGAP binding site. Although CAT mapped all experimentally detected sites, its performance for the lower-ranking clusters was worse than for the first two hotspots. Not all the probes

interacted with these binding sites. Interestingly, binding sites mapped by MD simulations using our most polar probe, acetamide, did not overlap with hotspots detected by other fragments (and *vice versa*). This suggests that putative hotspots detected by acetamide might not be druggable, or that they may be very small hence not amenable for fragment growth and structure-based ligand design.

FTMap detected only hotspots marked by clusters 1 and 2; both being among highly-ranked FTMap clusters (Table 12). On the other hand, FTMap detected the calcium acetate binding site[221] whereas neither CAT nor MSCS succeeded.

Table 12 H-Ras CAT results and comparison to FTMap

Target	Binding Site	Protein Contacts	Cosolvent	CAT Rank	Found by FTMap?
HRAS	Site 1	R68, Q95, Y96, Q99, D92	Acetamide	1	✓
			Acetanilide	2,7	
			Benzene	1,5	
			Imidazole	1,4,5	
			Isopropanol	2,4	
	Site 2	H94, L133, S136, Y137	Acetanilide	1,5,9	✓
			Benzene	8	
			Imidazole	4,7,10	
			Isopropanol	2,5,6	
	Site 3	S17, I21, Q,25, H27, V29, D33, T35, D38, Y40	Acetamide	3,9	X
			Imidazole	8	

	Site 4	F28, D30, K147	Acetamide	3	X
			Imidazole	9	
	Site 5	A11, G12, N86, K88, S89, D92	Acetanilide	9	X
			Benzene	3	
			Imidazole	4	
			Isopropanol	5,6	
	Site 6	D30, E31, Y32	Acetamide	3	X
	Site 7	L23, N26, K42, V44, V45, R149, E153, Y157	Acetanilide	6,8	X
			Benzene	4,6	
			Isopropanol	7	
	Site 8	G13, Y32, N86, K117	Acetamide	3	X
			Benzene	2	

5.4.4 Novel sites prediction on CDK2

Cyclin-dependent kinase 2 (CDK2) is a serine/threonine kinase that interacts with several different cyclins[222]. It is comprised by two regions known as C and N lobes, connected by a hinge, with a significant role in the cell cycle, in the transcription regulation[223]. CDK2 directly acts on the protein expression related to the transition from the G1 to S phase of the cell cycle. Hence, it is an interesting protein target for cancer drugs. Functionally, CDK2 goes through a consisting series of conformational changes to reach an active state. The interlobe region interacts with cyclins (preferably A and E), shifting the activation loop (located between residues A149 to T165) and subsequently revealing the ATP binding site. This allows the phosphorylation of the threonine located in the active site, reaching a final active configuration.

Recently, CDK2 was used in a novel experimental approach for the identification of binding sites called Fraglite[43]. Wood and coworkers experimentally mapped a series of CDK2 allosteric sites using halogenated fragments expressing paired hydrogen-bonding motifs, improving the assessment on its allosteric "druggability" and tractability. The method reliably identifies drug-like interactions, which are detected by X-ray crystallography, exploiting the anomalous scattering of the halogen substituent. The study reported a set of five regions with known fragment binding.

To further assess the capacities of the CAT scoring function, cosolvent dynamics runs were carried out with all five previously used probes. This methodology could be used with the FragLite probes, CAT analysis can be performed with any kind of small molecule probe, but the purpose of this methodology is also to design a series of optimal conditions for an orthogonal workflow. Since most of these sites were only recently discovered, these serve as the evidence for non-biasing of the scoring function developed in this study. CAT analysis ranked all 5 novel fragment binding sites[43] along with the ATP binding site. Detected hotspots are shown in Table 13.

Table 13 CDK2 CAT results and comparison to FTMap

Target	Binding Site	Protein Contacts	Cosolvent	CAT Rank	Found by FTMap?
CDK2	Orthosteric	E12, G13, Q131, N132, D86, L134, L134, D145	Acetamide	1	✓
			Imidazole	4,8	
	Site 1	K33, K34, Y77, K6, Y19, L32, K75, K34, H71L	Acetanilide	1,9	X
			Benzene	1,3	
			Isopropanol	2,3	
	Site 2	T160, H161, R157, T158	Acetamide	7	X
			Imidazole	6	

			Benzene	5	
			Isopropanol	4	
	Site 3	L124, R150, G147, H125, R126, C177, J178, Y179	Acetamide	3	✓
			Acetanilide	2,4,5	
			Benzene	5,8	
			Imidazole	9	
			Isopropanol	8	
	Site 4	T221, P222, D223, L219, R245, L267, Y262	Acetanilide	2,4	X
			Imidazole	10	
	Site 5	R199, T198, M192, T97, I104	Acetamide	3	X
			Acetanilide	2,8,10	
			Isopropanol	7	

Site 2 was located exactly in the activation loop. CAT highlighted all residues comprising this loop, including T160, with overlapping clusters of several different probes (Figure 29). T160 goes through a phosphorylation event, being one of the main contributors of binding site stabilisation[224]. The region along site 3 represented the dimerisation area where the binding of cyclin occurs, being found by CAT with different cluster ranks. Protein-protein areas were commonly highlighted by CAT scoring function, given the calculated energetic aspect which can filter highly favourable interactions within shorter residence periods of probe molecules.

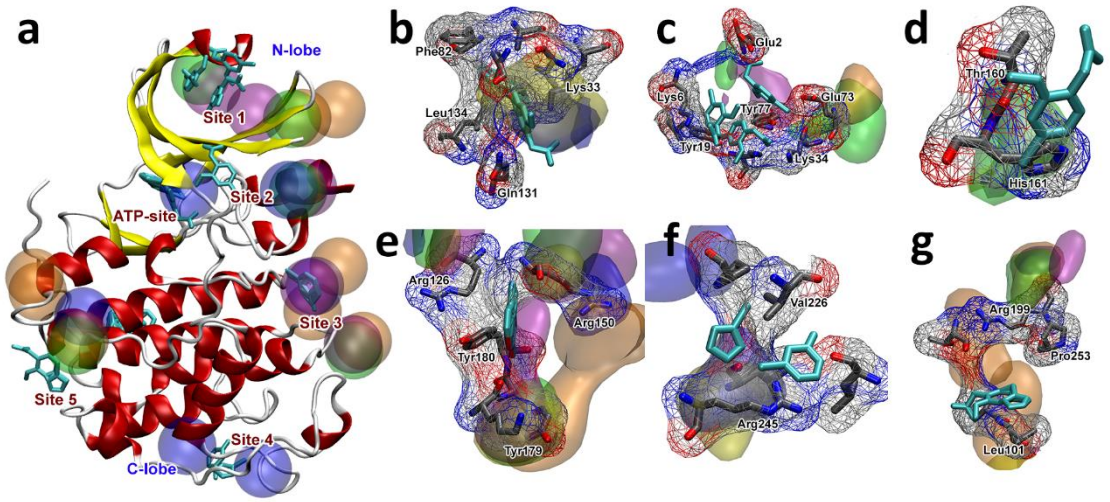


Figure 29 CDK2 hotspots found by CAT. The clusters are coloured as follows: acetamide – blue, benzene – purple, acetanilide – orange, imidazole – yellow, and isopropanol – green. The crystallographic fragment is coloured cyan. A) Panoramic view of CDK2 and the highest ranked cluster for each cosolvent molecule. A) Depiction of CDK2 and highest scored clusters, B) Orthosteric site, C) Site 1 D) Site 2 E) Site 3 F) Site 4 G) Site 5. As shown, acetamide and acetanilide performs better than the other 3 cosolvent molecules, given the nature of the experimental X-ray mapped crystallographic binding regions. Site 4 and 5 in specific shows high ranked clusters for these 2 probes, given by the high polarity of the site sidechains.

Sites 4 and 5 were located in the C-Lobe region. Site 4 was directly related to the C-lobe loops and it is a novel binding site for CDK2. It interacted with polar residues (such as T221 and R245), which explains its high affinity for acetanilide. When constrained, this region can change the dynamics of the semi-unstructured T221-D247 C-Lobe loop, which is related to the cyclin dimerisation stabilisation event, resulting in a plausible site for structured based drug design.

Site 5 was found at the end of the α-helical bundle that comprised most of the C-lobe sequence (Figure 29). It was highly ranked in CAT, particularly for highly polar probes, such as acetamide and acetanilide. As described by Wood and coworkers, fragments used in their study should be tailored to accurately find a specific binding region by the usage of fragments prone to form hydrogen-bonding interactions. Hence, the used structures should represent highly specific interactions, resonating with the results given by CAT, which ranked polar probes in the same manner.

Chapter 6 STAT3

Signal transducer activator of transcription 3 (STAT3) protein has emerged as a prominent target in tumour progression due to its pivotal role in cell signalling. The activation of STAT3 has been related to drug resistance[225], the expression of anti-apoptotic proteins[226], and the inflammatory processes in tumour development, among others[80,227,228]. In spite of its importance in cancer progression, the pharmacological targeting of STAT3 by small molecule inhibitors is still in infancy. Due to its tendency to aggregate, STAT3 structure determination is a major hurdle that prevents structure-guided design based on STAT3 structure in both monomeric and dimeric forms, as well as bound to an inhibitor[92,94,95]. Although many strategies have been described in literature to inhibit STAT3, a few inhibitors are still going through clinical trials (e.g. TTI-101 [ClinicalTrials.gov Identifier: NCT03195699] or napabucasin (BBI-608)[126,131,132][ClinicalTrials.gov Identifier: NCT03647839]) and STAT3 has become one of the most challenging cancer-related protein to target by small molecule, due to its inconclusive ligand binding nature. Gaining insights into the atomistic level structure and dynamics of STAT3 permits the identification of small molecule binding sites and structure-guided development of novel therapeutic strategies targeting STAt3 and modulating its oncogenic pathways.

The SH2 domain has traditionally been the main target for drug design, mostly accompanied by computational studies relying on molecular docking calculations or similar structure-based approaches[100,101,108,110,113,118,124,229,230], despite no crystallographic data available up to date to support them. These ligands attempt – albeit with limited success - to compete with p-Y705 at the binding site known for the binding of the phosphorylated residue.[111,231]OPB-31121[127] and OPB-5160[231], are at the time of writing this dissertation, the only two ligands described as SH2 inhibitors that bind in a different pocket than p-Y705[127,227]. Furthermore, STAT3 can undergo other post-translational modifications besides Y705 phosphorylation such as S727 phosphorylation[232,233] or K685 acetylation[234] and it has been experimentally demonstrated that unphosphorylated STAT monomers can dimerise and bind to DNA.[234] These allow STAT3 to overcome

inhibition targeting SH2 domain, and contributes the explanation of limited success of ligands binding to the SH2 domain.

This chapter describes the evaluation of STAT3 “druggability” with special interest in the SH2 and DBD domains. MD simulations and molecular docking were performed for the SH2 domain to evaluate its stability and conservation, that would deem it as an optimal binding site. The application of umbrella sampling (simulations), relationships between inter-domain mutations and binding of a potent ligand, BBI-608, in order to decipher the mode of action of the ligand.

6.1 Is SH2 the ideal site to target?

For the past years, STAT3 inhibition has been focused on its dimerisation process. STAT3 monomers are activated via a peptide (PY*LKTK)[111] that induces the dimer formation and, for the past years, the modus operandi for STAT3 ligand design relied on the search of small molecules competing with this peptide. Several candidates have been proposed, but only a few of them have made it through the preclinical testing. Low specificity and activity was the main issue for these compounds to not go further. This raises the questions on (i) whether STAT3 SH2 domain binding site is the best site to target, (ii), whether alternative, allosteric binding sites exist in the SH2 domain, and (iii) whether other STAT3 domains would be feasible for targeting by small molecules.

6.1.1 Molecular dynamics of the SH2 domain and its “druggability”

A series of equilibrium molecular dynamics (MD) simulations of SH2 domain have been carried out, to study the intrinsic dynamics of this domain, with the focus on the pTyr binding site (residues K591, R609, S611, E612, S613). Five 100 ns replicas were performed.

Simulation data showed that SH2s pTyr binding site is highly flexible and conformationally adaptive, as key residues from the β-sheet that form the SH2 pocket are displaced, therefore changing its conformation (Figure 30).

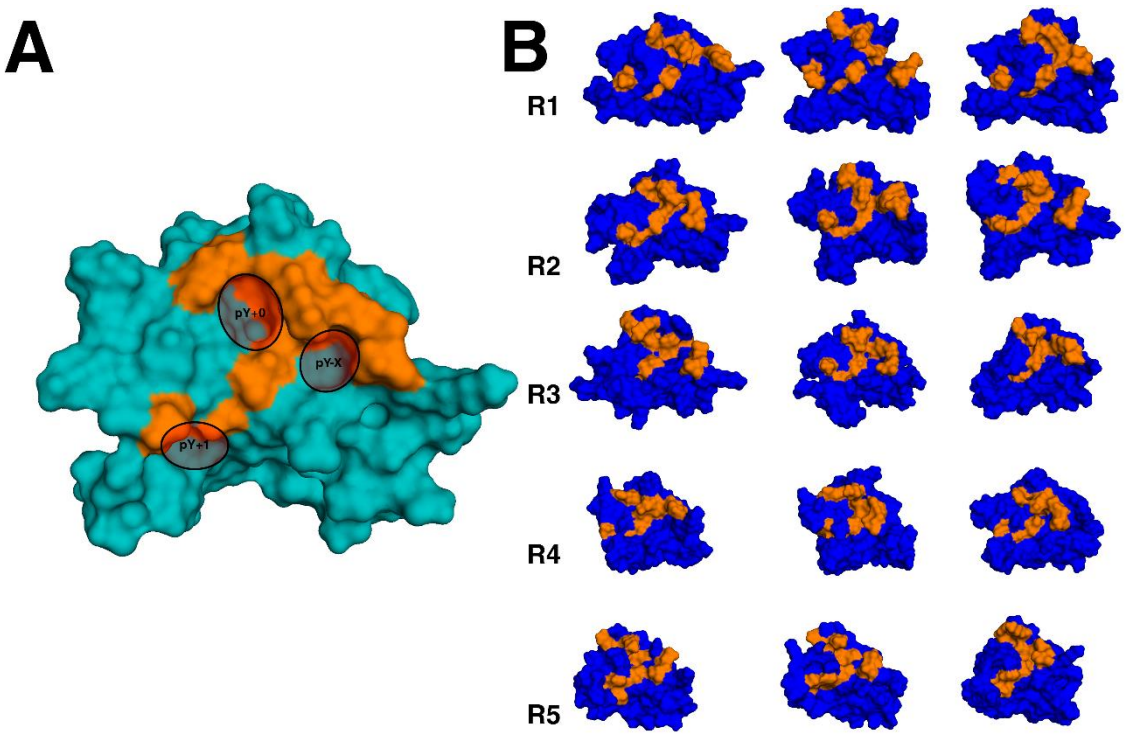


Figure **30** Surfaces of the SH2 domain. A) corresponds to the SH2 crystal structure while B) shows the three most populated clusters for each of the replicas simulated

To assess the “druggability” of the obtained SH2 ensemble, FTMap, a well-established pocket detection tool is used. The most populated cluster for each replica has been selected in order to identify the stable binding pockets. Interestingly, the modified cavity generated around the phosphorylated pTyr residue has not been identified as “druggable” by FTMap (Figure 31) in favour of other regions of the domain. These results indicate that the pTyr binding site might not be as “druggable” as previously thought. Due to the lack of ligand-bound crystal structure, CADD approaches focusing on the pTyr binding site could be wrongly biased and massively improved if more stable conformations obtained by MD are used for drug discovery methods such as virtual screening. Since there is no experimental evidence for the direct binding of most of the described SH2 inhibitors, there is the possibility that they bind to another cavity within the SH2 domain (or elsewhere). Collectively, MD simulations show the need of the further study in search of new STAT3 binding sites.

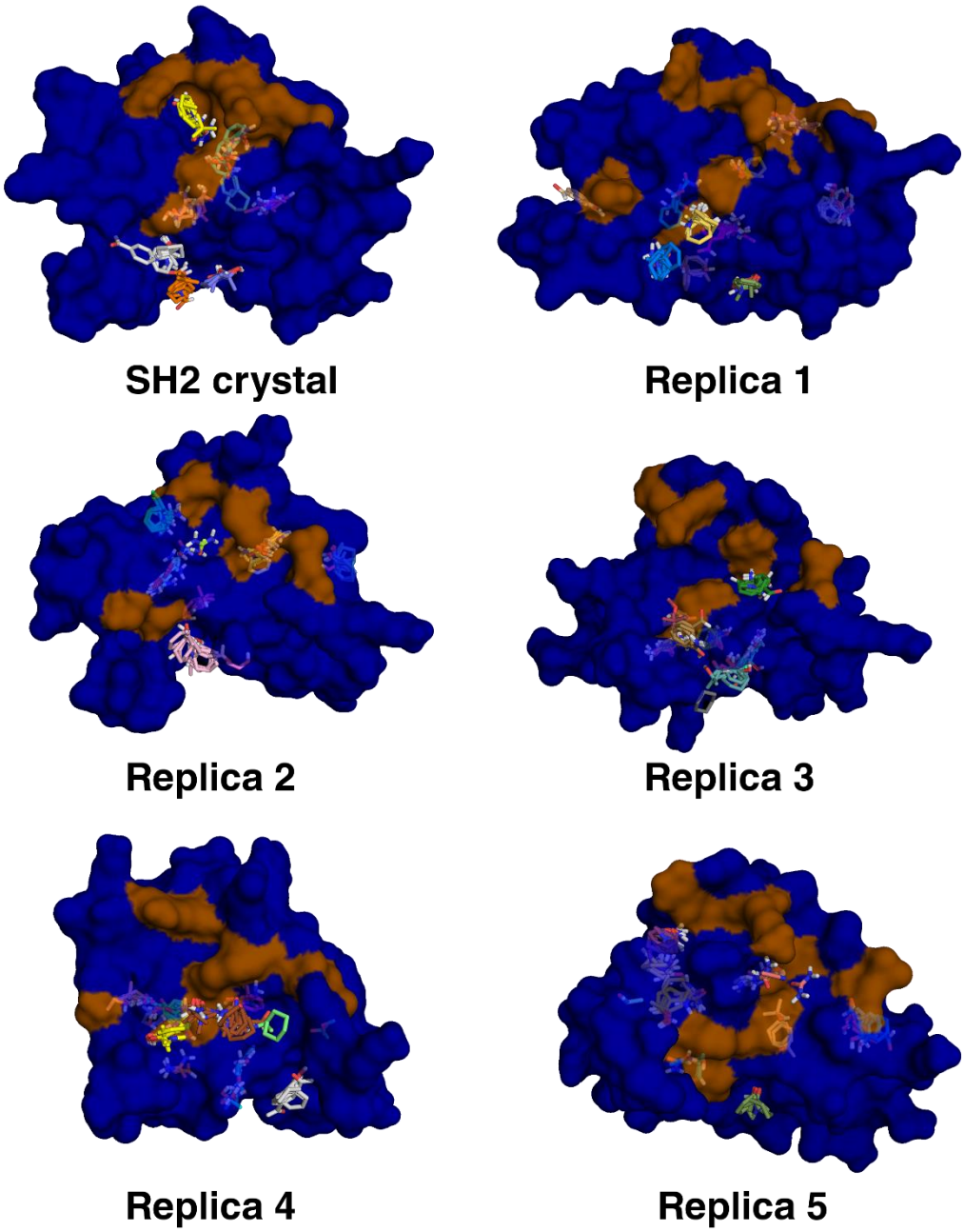
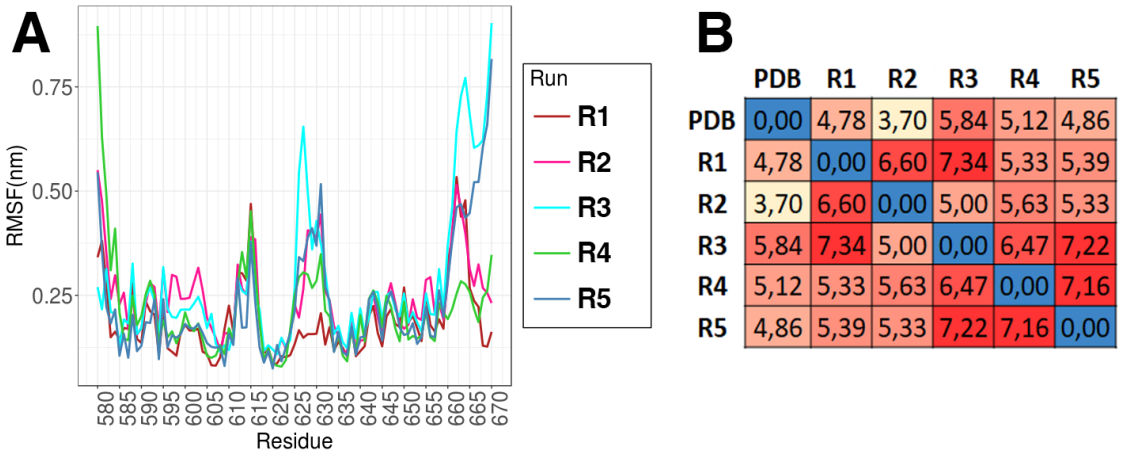


Figure 31 FTMap results for the SH2 domain. The crystal structure and five different MD replicas have been calculated. Fragments in sticks correspond to the mapped areas by FTMap. In orange, the surface region corresponding to the pTyr site residues. It can be seen how after MD, FTMap does not deem the pTyr site as a “druggable” pocket in favour of other regions of the domain.

Some of the sites predicted by FTMap agree with the sites mapped by AutoDock4 and DOCK6 from the molecular docking benchmark (see section 4.2.1.2). The site predicted by AutoDock4 (Site A) is situated just below the P site between residues I634 and P639 for one pocket and I 652 and I660 for the second

while the one predicted by DOCK6 (Site B) is situated just behind the P site, interacting with residues like Y584, E593 and L606. The proximity of these cavities to the P site might be causing the movement of near residues, closing the P site and therefore inhibiting its activation. Backbone RMSD between the crystal structure and MD clusters are calculated, showing a considerable difference between the initial and simulated structures. Furthermore, RMSD between the simulation clusters are also considerably high (>5Å in most cases), indicating the flexibility of the domain (Figure 32.b). RMSF analysis showed high fluctuation of pTyr. Upon comparing all runs, some of the residues with a higher difference between each other are the ones that form pTyr binding site or close companions (Figure 32.a).



	PDB	R1	R2	R3	R4	R5
PDB	0,00	4,78	3,70	5,84	5,12	4,86
R1	4,78	0,00	6,60	7,34	5,33	5,39
R2	3,70	6,60	0,00	5,00	5,63	5,33
R3	5,84	7,34	5,00	0,00	6,47	7,22
R4	5,12	5,33	5,63	6,47	0,00	7,16
R5	4,86	5,39	5,33	7,22	7,16	0,00

Figure 32 A) RMSF per residue for every SH2 MD simulation B) RMSD comparison between the SH2 domain crystal structure and the main cluster for every MD simulation

6.1.2 Molecular docking

Several of the STAT3 inhibitors described in the literature have been discovered via virtual screening (VS). This means that a database of tens of thousands of compounds has been tested computationally to a target via molecular docking. Typically, validation is performed to verify the applicable scoring function for molecular docking. This means that, if possible, molecular docking is performed with the crystallised ligand to reproduce the native binding mode, which serves as the direct validation of the docking procedure. In the case of STAT3, there is no ligand-bound crystal structure available, meaning that the direct validation of the docking procedure is not possible. In most cases, it cannot be certain, though,

whether the reported STAT3 inhibitors truly bind at the pTyr site, given the lack of experimental data for *holo* STAT3. Usage of the pTyr activation peptide could serve in the validation, unfortunately, peptide docking requires parameters not available in scoring functions developed for the most commonly used small molecule docking packages.

In order to validate the STAT3 docking results another, indirect approach was used in this study. A virtual screening of the series of STAT3 inhibitors denoted as “direct inhibitors” (believed to bind to the pTyr site) was performed, using three different docking programs, employing different scoring functions: MOE-Dock, AutoDock4 and UCSF DOCK6. Rather than focusing solely on the pTyr binding site, the whole SH2 domain was considered to be the target. Convergence of the results obtained by different scoring functions would be a viable strategy for boosting the confidence in the results, in the absence of structural data.

The obtained results were heavily dependent on the docking software and the scoring function used, as each of the three docking packages identified different regions of the SH2 domain as the main binding site (Figure 33). Only MOE identified the pTyr binding site as the most populated and highest scoring cluster. While AutoDock4 selected a region adjacent to the pTyr site, formed by residues Q635 to E638, T714 to T717, K626, I658 and V667 DOCK6 picked the back of the SH2 domain as the most likely “druggable” site. These results indicated that the outcomes of virtual screening are heavily biased by the method chosen (e.g. scoring function), and defining the mechanism of action (i.e. binding mode) of STAT3 inhibitors solely via the molecular docking results is not recommended.

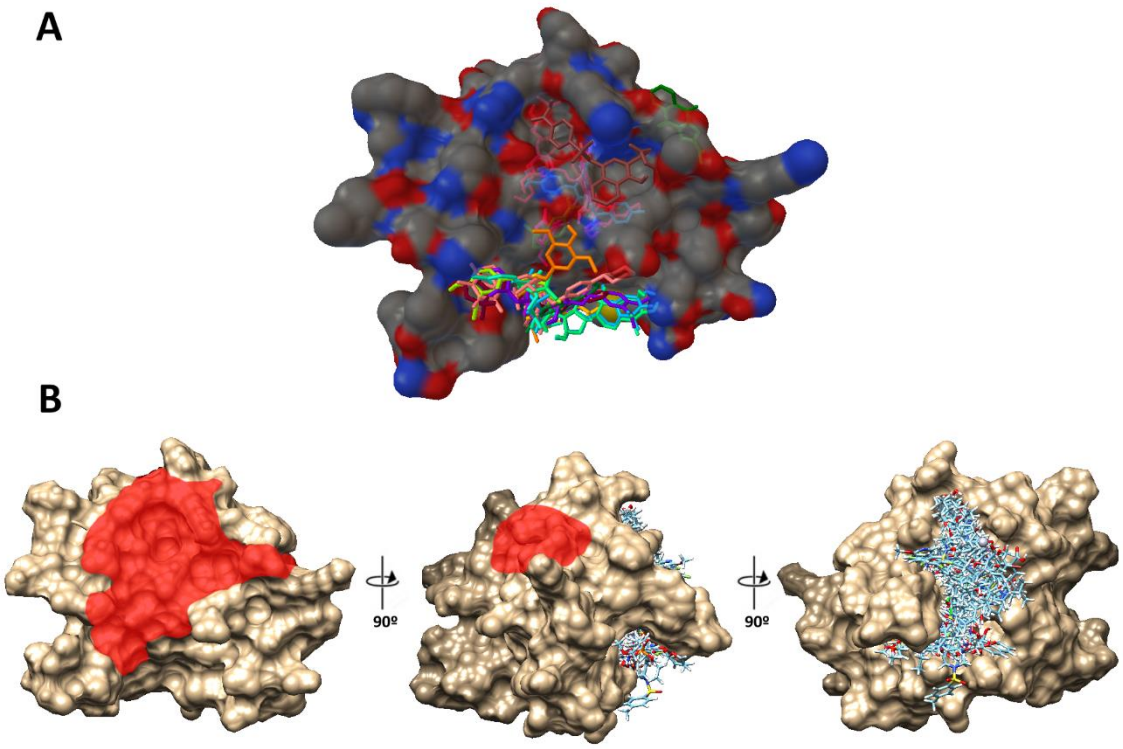


Figure 33 Molecular docking results vary depending on the software used. From a set of described inhibitors A) Autodock4 identified a region below the pTyr site while B) UCSF DOCK6 conformations bound preferentially at the back of the domain (pTyr site highlighted in red)

Considering the high flexibility of SH2 domain and large conformational changes within it observed during MD simulations, molecular docking has been subsequently attempted, using the newly obtained conformations. The most populated cluster of each simulation run (0,20 nm cut-off) from section 6.1.1 has been used as its corresponding receptor.

Figure 33 shows poor “druggability” ofthe pTyr site, and a strong preference of other regions of the domain. Unlike in the previous calculations, the results obtained by all docking programmes and different conformation were highly consistent. The highest scoring position and most populated clusters were the same for all the used scoring functions employed. The procedure identified two putative binding sites (Figure 34).

The results indicate that these two sites have different preferences regarding the chemical structure of the ligand. Smaller and planar ligands, containing two or more aromatic rings, bind preferentially in an area below the pTyr site (Figure 34).

This site, denoted as Site L, forms a deep cavity due to the loops that comprise it. High flexibility of these loops (observed during MD simulations) make this site transiently open and close. Larger and more flexible ligands bind preferentially to the site denoted as Site P, located the opposite side of the SH2 domain, where interactions with the linker domain may occur (Figure 34). Potency of these ligands is poor (micromolar level), and none of them have made it to clinic. These ligands were specifically designed to bind in to the pocket which they could not interact with, resulting in a poor affinity.

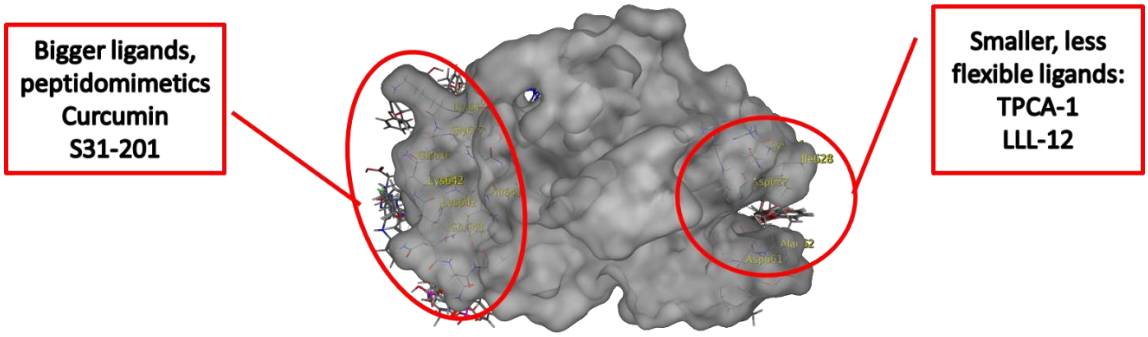


Figure 34 After MD simulation molecular docking binds known inhibitors in the depicted two areas instead of the pTyr site

6.2 If you cannot win them, join them

Experimental data have demonstrated that point mutations in the linker domain suggest contacts with both the DNA-binding and SH2 domains, which could cause structural changes that severely affect STAT3 activity[138]. Alanine scanning demonstrated that the modification of interdomain hydrogen bonds can produce a significant decrease (i.e., K551A, W546A) or increase (D570K) in the STAT3−DNA-binding compared to that in the wild-type protein[138]. Understanding the effect of point mutations on STAT3 activity at the atomistic level could provide significant information about novel binding sites, unveiling new ways to target STAT3 by small-molecule ligands.

6.2.1 Equilibrium MD simulations

In the study by Mertens and coworkers, several mutations were indicated as crucial to control DNA retention time within its respective binding cleft at STAT3[138]. These mutations occurred either in DNA-binding domain, or in the inter-domain

region. To evaluate the DNA-binding in the mutated-STAT3, three replicas of 50 ns MD simulations were performed to equilibrate the STAT3-DNA complexes prior to umbrella sampling (US) simulations, and to assess any differences between WT and mutants, in respect to their structures and dynamics. Systems have been studied for that amount of time based on work by Husby and coworkers, which claimed to achieve an energetically conserved and stable simulation.[235] The results obtained in this work agreed with the published data[235], as one of the STAT3 monomers was more flexible than another monomer. RMSD variations were pronounced mainly in the loops of SH2 and Coiled-Coil (CC).

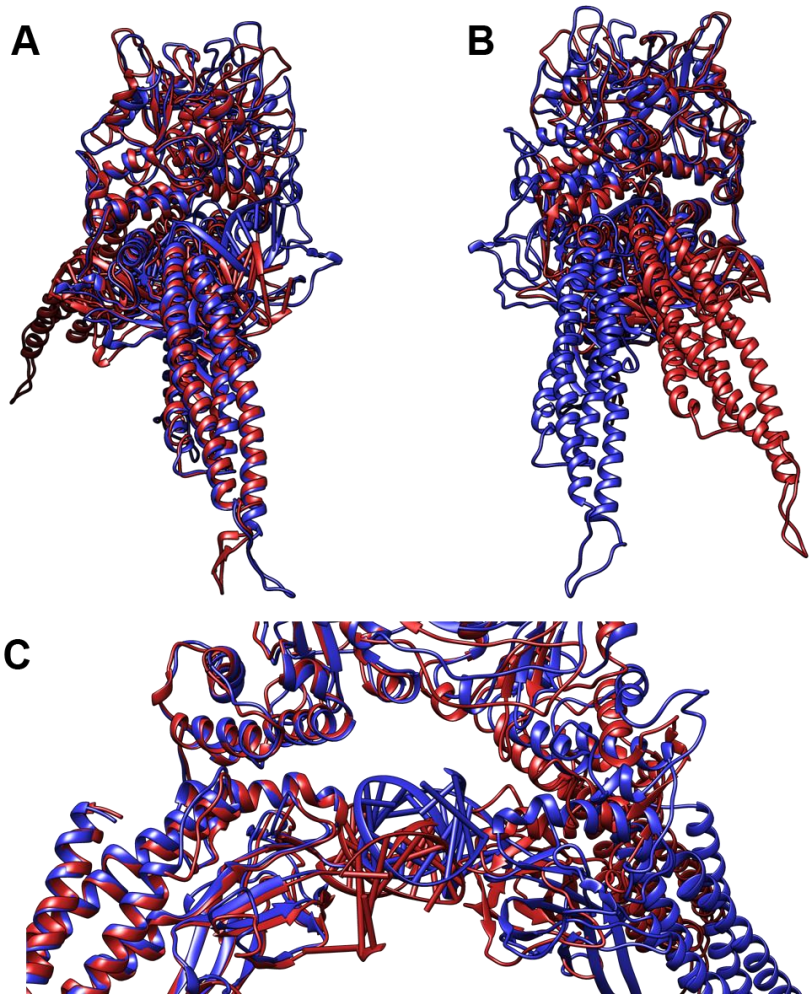


Figure 35 MD simulations show conformational changes between WT (blue) and D570K (red) STAT3 dimers systems. In the D570K mutant, one of the monomers is shifted (B), changing the conformational landscape of the dimer. Panel C) shows how the position of the DNA duplex is shifted downwards in D570K mutant compared to WT.

One of the most significant configurational changes occurred within D570K mutant, as the DNA double helix shifted downwards (Figure 35). This was most

likely caused by the electrostatic effects at the residue located in the interface between linker and DNA-binding domain. The modification of the side chain charge from negative (D) to positive (K) increased favourable protein interaction with the negatively charged nucleic backbone. This tightened the DNA binding, resulting in a higher average DNA RMSD when compared to the crystal structure. It strongly indicates that the end-point configurations of the protein-DNA complexes play a significant role in their binding free energy, since the protein-nucleotide interactions change significantly between different mutants.

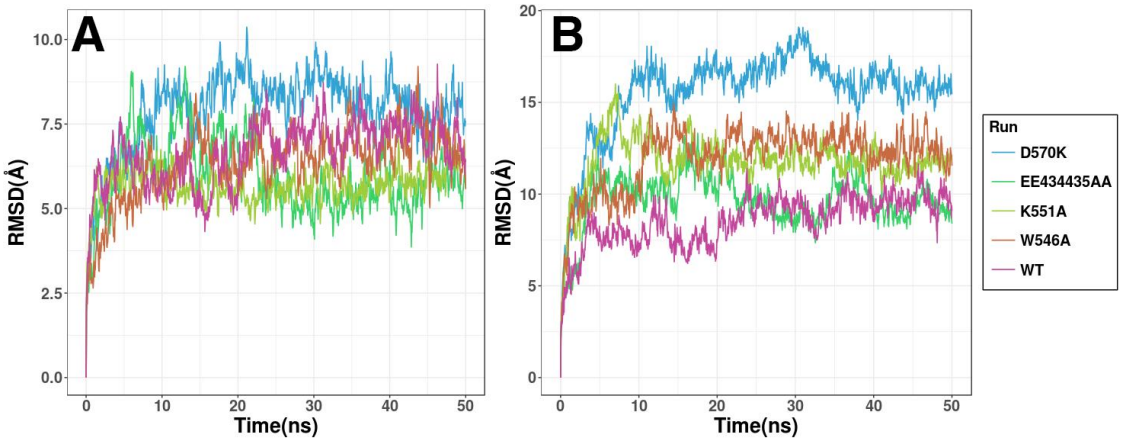


Figure 36 Protein A) and DNA B) RMSD after 50 ns of MD simulation. D570K (brown) mutation shows higher RMSD in both protein and DNA counterparts, compared to the other mutations and WT-STAT3

Next, I assessed whether the conformational changes induced by D570K mutation were observed in other mutations. Figure 36 shows root-mean-square deviation (RMSD) plots of all STAT3 considered in this study as well as WT protein. The obtained RMSD values are considerably high, but it should be taken into account that the STAT3 dimer is modelled. This leads to the observation of the displacement of one monomer, reason of that high RMSD along with the modelled loops in the CC domain. The STAT3 dimer is thought to be a mirrored image of every monomer, but these simulations indicate a different result. Except D570K, there were no large differences in protein RMSD between the mutated STAT3 dimers and WT. This gap between D570K and other mutants was likely to arise from the combination of electrostatic and steric effects (all other mutations replaced large and polar residue with smaller and apolar alanine), which affects intrinsic dynamics of the CC domain. Hence, the dynamics of the CC domain

might “tune” DNA-STAT3 interactions by allowing adjacent SH2 and DBD domains to improve their structural “fit” to the DNA.

6.2.2 Umbrella Sampling (US) simulations

To follow up on the effects of the mutations which control DNA retention at the STAT3 on the structure, dynamics, and energetics of STAT3-DNA complexes, I carried out a set of umbrella sampling (US) simulations, where DNA has been pulled from STAT3 dimer. This process was carried out using a series of US windows and simulated for 10ns each. We understand that the simulation time per window is short for this technique and it will struggle at the moment of evaluating convergence and presumably provide a series of energy values biologically impossible. Our main interest for this analysis is the comparison between the different mutated systems and the profiles they can provide. At no point we presume that the obtained values will be representative.

The potentials of mean force (PMF) calculated via weighted histogram analysis method (WHAM) were consistent with the results reported by Mertens and coworkers[14] in most cases. Experimental results showed a drop in DNA-binding for the tested inter-domain mutations (EE434/435AA, W546A and K551A) through time and an extraordinary high retention time for D570K, with a 100% DNA-binding even after two hours.[138] The WT STAT3 had a higher energy barrier to reach its unbound state in comparison to W546A mutant (Table 14). The mutant showed a lower retention time, which indicates that inter-domain interactions between mutated residues and E434 are crucial for STAT3-DNA binding. Therefore, disrupting these interactions could represent an attractive strategy to target STAT3 by small molecules. K551A shows a similar profile than W546A but presented a higher PMF value than the other mentioned and a few kcal/mol more than WT (Table 14).

Consistently with the results of the equilibrium MD simulations, PMF showed that D570K binding affinity to DNA was more favourable than of any other mutant, and more than WT STAT3 (Table 14). This indicates that this mutation promotes a very tight binding between STAT3 and DNA, with a higher energy gap for DNA release upon pulling (Figure 37). The PMF curve showed that DNA pulling from D570K required a higher energy gap to release DNA from the STAT3 dimer.

Experimental retention time correlated with the simulations when compared to WT STAT3. The data showed that DNA binding to the D570K was persistent through time, it did not drive transcription and resisted dephosphorylation, thus prevented STAT3 to exert its function[138].

Collectively, these results indicate that D570K promotes a very tight DNA binding, so much that the bound duplex stays “locked” between the dimers, which effectively inhibits STAT3 by preventing it from releasing DNA and exerting its function as a transcription factor.

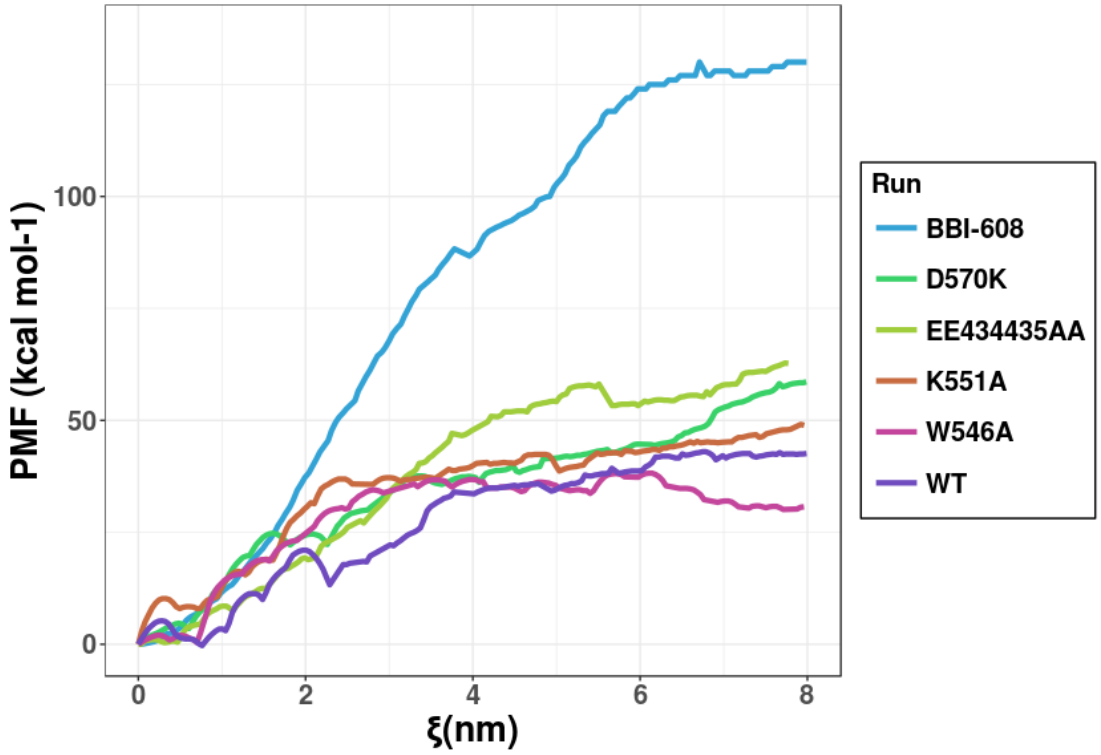


Figure 37 Potential of mean force (PMF) of dissociation of DNA from the STAT3 dimer. Both K551A (orange) and W546A (pink) mutants showed a lower PMF than WT (violet). EE434435AA (light green) would have displayed similar results, but the interaction between DNA duplex and DBD of one STAT3 monomer in the latest sampling windows resulted in higher PMF value than expected. In comparison, D570K (marine green) showed much higher PMF value than WT, indicating that DNA-protein interaction is more favourable in this mutant, relatively to WT. The BBI-608 binding (blue) showed similar effect to D570K mutation.

Table 14 Free energy change calculated for all umbrella sampling (US) calculations

Run	ΔG_{US} (kcal/mol)
WT	-50.0
EE43435AA	-65.0
W546A	-35.7
K551A	-52.7
D570K	-60.0
BBI-608	-130.0

The only discrepancy between these results and experimental data was observed for the EE434/435AA double mutant. In the simulations, the mutant showed a higher PMF value than WT, which indicated that its DNA binding affinity should be higher, while experimental data showed that its behaviour resembled that of K551A and W546A mutants, which have shown a considerable drop of DNA-binding through time. Analysis of the final US windows indicated that the middle of the DNA duplex interacted favourably with the DNA-binding domain of one STAT3 monomer, but not another. Therefore, the US curve of EE434435AA mutant displayed higher values arising from these interactions (DNA-STAT3 monomer) rather than from favourable interactions DNA-STAT3 dimer, as it was for D570K mutant.

The results by Mertens and coworkers[138] are based on 60 to 100 min experiments that evaluated the percentage of DNA-binding of STAT3. This process would include several STAT3 molecules that would most likely go through several mechanistic cycles. Therefore, it is plausible that I did not sample the conformations that are contributing to these results. As expected, the obtained energy values are far from what it should be expected for this system. But the size of this system along with the available computer power and time impedes the desired further sampling for it. Therefore, as mentioned before, we pretend to interpret these values more as a comparison/estimation between the

studied systems rather than an absolute indicative of the energy profile of the system.

All simulations indicated significant conformational changes of the arginine R414, which were required to release DNA (or to allow the DNA binding to the STAT3 dimer). R414 has been shown as one of the main residues for DNA identification and binding[235], but it was not described as a “gatekeeper” residue. R414 is at close distance from DNA and its initial position did not allow the DNA exit from the dimer. By acting as a “gatekeeper” of DNA binding (Figure 38), R414 exerted a key role in controlling opening and closing of STAT3 dimer, as well tuned the dynamics of STAT3 monomers by modulating intra-domain DBD-SH2, CC-SH2, DBD-LD, and CC-LD interactions.

Figure 38 Conformational changes of arginine R414 observed during the simulations. Evolution of the DNA duplex and R414 position over the simulation time is depicted by colours, from red to white. Along the DNA pulling pathway from the STAT3 dimer the R414 sidechain rotates, allowing the dissociation to occur.

6.2.3 Inhibition of STAT3 by napabucasin (BBI-608)

The results of simulations of *apo*STAT3 (WT and mutants) highlighted a set of inter-domain residues, explaining their effect on STAT3 behaviour and function at the atomistic level of detail. These observations may pave the way to novel strategies for STAT3 inhibition using small molecule ligands.

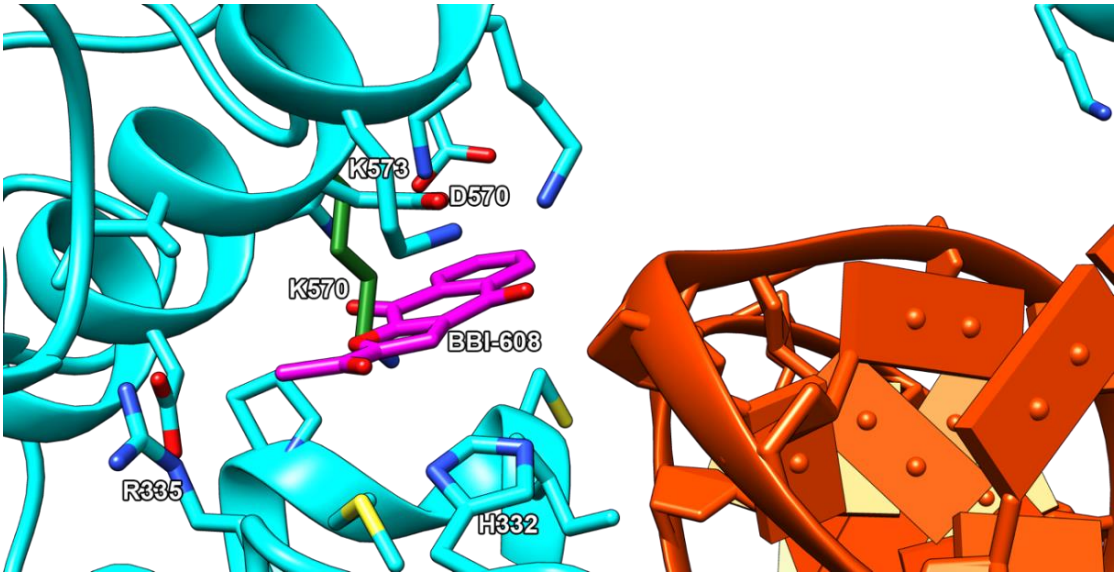


Figure 39 The binding pose of BBI-608, according to the data reported by Ji and coworkers[133]. Since the crystal structure has not been released, I have modelled the most plausible binding mode by molecular docking. The side chain of the mutated K570 residue is displayed and coloured green – it is overlapping with the plausible location of BBI-608.

Recently, Ji and coworkers reported that napabucasin (BBI-608), which is a STAT3 inhibitor in advanced clinical trials (Phase 3), binds to a small pocket between the linker and DNA binding domain in a STAT3 crystal structure[133]. Since the crystal structure has not been released to the public domain, I have assessed the druggability of this segment of STAT3, identified the putative pocket, and subsequently built the model of BBI-608 bound to STAT3 using molecular docking approach, and subsequently validated the obtained binding mode by the atomistic MD and US simulations.

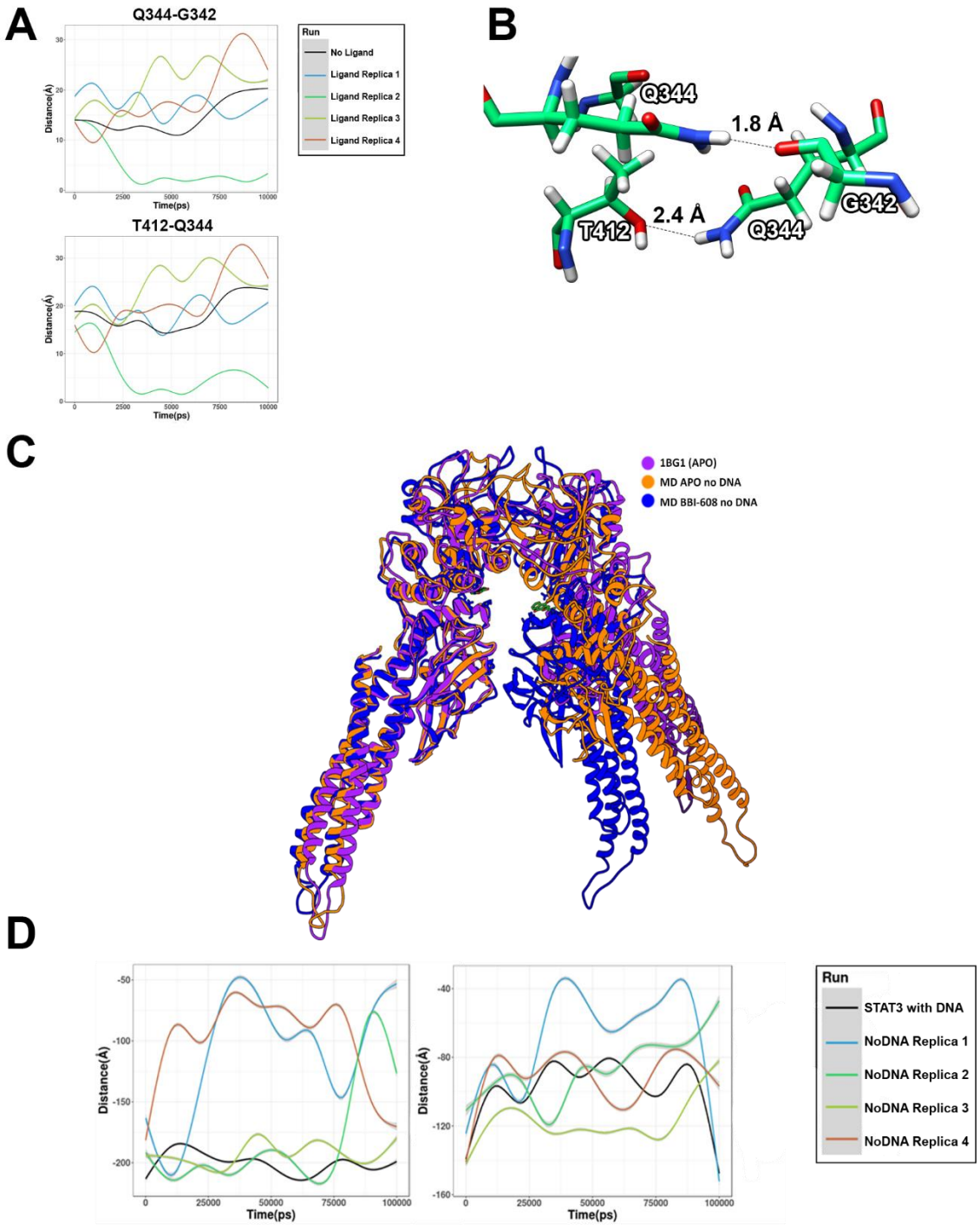


Figure 40 A) Interatomic distances between the Cα of residues Q344 and G432 and residues T412 and Q344, calculated along a 50ns MD simulation of ligand-STAT3 complexes. Replica 1 (blue) consists in the dissociated system and both Replicas 2 and 3 (orange and green) keep their ligand bound through the whole simulation B) Close-up and C) panoramic view as STAT3 dimer closes once the BBI-608 molecule is bound D) Protein-ligand energy interaction for both BBI-608 molecules interacting with each monomer along a 50 ns MD simulation (three replicas: blue, red, and green). STAT3 dimer bound to DNA (black) is showed as the reference.

Molecular docking was performed by MOE for each STAT3 monomer separately, in an attempt to generate the most plausible conformation relying the limited data available. Both blind (whole monomer) and targeted (residues of the identified pocket) docking calculations resulted in a set of conformations with favourable energy scores and highly-populated cluster located within the DBD site pocket, in close contact with residues H332, P333, R335, K573 and D570 (Figure 39). Two conformations, matching the published data, were found: both were assessed and validated.

To validate the binding mode of BBI-608, MD simulations of WT STAT3-DNA-BBI-608 complex were performed for 100 ns in triplicate. Subsequently, the ligand affinity has been calculated. The ligand docked either of STAT3 monomers remained bound through the whole simulation. Interaction energies, calculated by MMPBSA analysis (*g_mmpbsa*[200] module) resulted in -18.1 ± 2.6 kcal/mol, showing a favourable binding.

US simulations, performed using the same protocol as for STAT3-DNA complexes, started the pull from the most populated cluster. The calculated binding affinity has been severely overestimated (-160 kcal/mol), nevertheless it showed very tight binding. It is obvious that these values are completely unthinkable and therefore the experiment was replicated with similar results. Much longer sampling is required for this system to reach convergence and obtain a proper picture of the energetic profile of the system, but we believe that even so there is a correlation between the studied systems. Compared with the results obtained for protein-DNA complexes described in the previous section, it implies that the presence of BBI-608 enhances DNA binding with a similar effect to D570K mutation. As such, BBI-608 inhibits the function of STAT3 in a similar manner to D570K mutation, which does not drive transcription and resists phosphorylation[12]. Since D570 has been annotated by Ji and coworkers[133] as the BBI-608 binding site residue, we concluded that BBI-608 binding to WT STAT3 generated a similar DNA-protein interaction pattern and retention time than D570K mutation.

Next, I performed MD simulations of BBI-608-bound WT STAT3 dimer without DNA, to study the influence of the ligand on the protein behaviour in the absence of DNA. BBI-608 was bound to each STAT3 monomers and four 100 ns replicas

showed a variation of results. In only one out of the four replicas, both BBI-608 molecules remained bound in their pockets through the whole trajectory, while DBD domains of both STAT3 monomers moved closer to each other, reaching the point of forming inter-domain hydrogen bonds. In one simulation both ligand molecules dissociated from the pockets, which caused an opening of the STAT3 dimer and in the other two simulations one of the ligands left the binding cavity at the 35 and 70ns of MD simulations, while another remained bound to STAT3.

To assess ligand-induced conformational changes within the DNA entry through both DBD domains, distances between some of the residues involved in H-bonding (e.g. Q344-G342 and T412-Q344) were measured and analysed in all four replicas and compared to the simulation of the WT STAT3 dimer without DNA and/or BBI-608 bound (Figure 40).

STAT3 dimer closed further down in the presence of BBI-608. This was particularly pronounced in one of the replicas, in which the distance between monomers reduced to <5 Å. These results indicated that BBI-608 binding to *apo*STAT3 is likely to trigger conformation changes that would prevent DNA from binding. In the replica simulation, where both ligands dissociated from STAT3, the distance between monomers increased upon ligand dissociation, as both ligands exit via the gap formed between DBD domains of STAT3 monomers.

Protein-ligand interaction energy was calculated and a correlation between ligand dissociation, dimer separation and poor ligand affinity was observed.

The simulations also indicated that ligand binding to one of the STAT3 monomers is more favourable than binding to another one. Although interaction energies are favourable for both monomers, one showed an interaction energy value twice as favourable as for another monomer. Although the allosteric effects within STAT3 were beyond the scope of this study, these results strongly suggest that such effects may occur in STAT3 dimers and contribute to the modulation of STAT3 by inhibitors. Another explanation could be that the model was not optimal. Although the used docking poses have an extraordinary resemblance to Ji's data,[133] only one is shown in the patent. Since this dimer is asymmetric, DNA interactions with both cavities would be different and could imply a different conformation and/or location for the ligand that does not correspond to the model

one. Furthermore, the MD results strongly indicate that one monomer is much more mobile than the other, and this is likely to affect protein−ligand affinity.

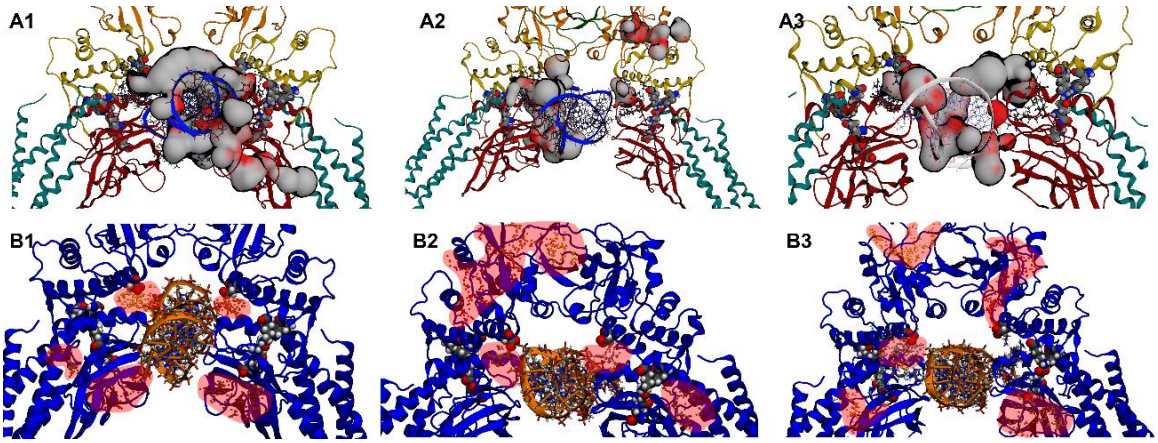


Figure 41 "Druggability" of STAT3 dimer. Sitefinder A) and fpocket B) were used to identify new potential pockets for structure-based drug design. In both cases the BBI-608 DBD site was identified along with novel DBD pockets.

6.2.4 Identification of a novel "druggable" binding site

Experimental results[12] combined by the simulations strongly indicated that targeting the interface between STAT3 monomers may trigger similar response to the inhibition by ligands binding to the DBD domain, and therefore be explored in structure-based ligand design efforts. With most STAT3 ligands being designed for the SH2 domain and just a few for the DBD domain, the identification of new "druggable" pockets for STAT3 inhibition is of a great interest.

As such, I scanned STAT3 for the presence of potential binding sites with two pocket detection tools: fpocket[55] and MOE's Site Finder. Upon selection of the dimer model and different clusters from its MD simulation trajectories I confirmed the binding site identified for BBI-608[133], which has been identified by both tools as their top-ranked site. Interestingly, that site was detected by both fpocket and SiteFinder for all analysed structures (Figure 41). In addition, a new pocket within DNA binding domain (close contact with E434 and E435) was identified. The main difference between results obtained by both tools is that fpocket was more prone to detect SH2 domain sites as pockets (Figure 41, panels B2 and B3) while SiteFinder identified a novel "druggable" pocket close to R414 (Figure 41, panel A1). Ligand binding to the R414 pocket could result in a possible DNA release/binding impediment.

6.3 New site, new opportunities

The identification of the DBD groove as a binding site “druggable” by small molecules provided two outcomes relevant for the STAT3 structure-based drug design: (1) ligand binding outside of the SH2 domain is possible, and (2) there is a specific pocket to be scrutinised. Most of the computationally designed STAT3 inhibitors were intended to target the pTyr site at the SH2 domain. Targeting the new DBD pocket would require a set of different features in order to optimise the potency and selectivity. Since there is only one ligand identified to bind to that pocket (BBI-608), I used SBDD techniques to explore this pocket and to discover new inhibitors.

6.3.1 Drug repurposing

Since *de novo* design is inherently time consuming and its application would be beyond the scope of this project, I screened a set of FDA approved drugs (repurposing). For the validation, both a blind docking including linker and DBD domains, and a DBD-targeted docking have been performed. Considering the success in application of MOE in deconvoluting of the BBI-608 binding mode, I applied the same protocol in the repurposing study. Out of the whole database[236] (1930 molecules), a series of non-steroidal anti-inflammatory drugs (NSAIDs) scored the highest in meeting the selection criteria based on the binding energy (score) and ability to reproduce protein-ligand interaction (K573) and conformation observed for BBI-608 (Figure 42).

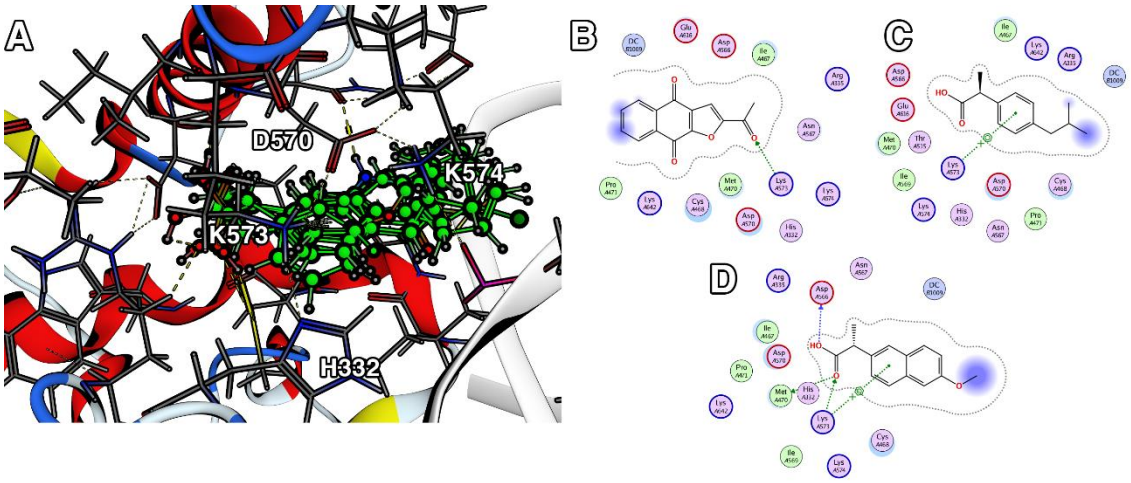


Figure 42 A) Overlap between most favourable conformations, ligand interactions maps of BBI-608 (B), ibuprofen (C) and naproxen (D)

From the series, six ligands were selected for the further evaluation: carprofen, fenoprofen, flurbiprofen, ibuprofen, naproxen and suprofen (Table 15).

To validate the predicted binding modes, molecular docking calculations were followed-on by equilibrium MD simulations. Simulations were performed for 100 ns in triplicates, and the binding affinity was calculated using MM-PBSA. Fenoprofen and suprofen had poor interaction energies and they dissociated from the binding site during the simulation. The other ligands formed stable complexes and their interaction energies calculated by MM-PBSA (*g_mmpbsa*[200]) showed an excessively favourable binding, with predicted affinity higher than BBI-608 (Table 15). Like in the case of US, we take these values more into account as comparison between rather than a prediction. Otherwise they would be fairly strong binders which we believe is not the case. These values can be related directly to the method, MM-PBSA. Other techniques such as free energy perturbation (FEP) could have been employed to determine the ligand-binding affinity.

Table 15 Energy interaction studied compounds calculated by MM-PBSA

Ligand	MMPBSA (kcal/mol)
Carprofen	17,9 ± 6,6
Flurbiprofen	29,9 ± 6,5
Ibuprofen	23,9 ± 4,8
Naproxen	34,9 ± 5,4
BBI-608	18,1 ± 2,6

6.4 CAT analysis of STAT3

As mentioned in previous sections STAT3 is a hard to target protein. A lack of structures co-crystallised with small molecule ligands hinders proof of its canonical binding site. Although there is evidence on its allosteric behaviour, most studies on the protein hotspot identification are inconclusive due to their lack of experimental validation. Therefore, in an attempt to identify potential binding sites, I performed cosolvent MD and its further analysis with CAT for STAT3.

The same protocol as the CAT benchmark was followed. This includes selection of the same probes (acetamide, acetanilide, benzene, imidazole, isopropanol), at the same concentration, 10% (m/m), and that three times 50 ns replicas per cosolvent were performed for the CAT analysis. The STAT3 monomer, except the CC domain, was simulated instead of the dimer.

Taking into account the data gathered in this work, it can be assumed that STAT3 presented at least two binding regions: the pTyr site in the SH2 domain, and the DBD groove identified by Ji and coworkers[133]. CAT identified the pTyr site but struggled to properly map the buried DBD pocket (Figure 43). At the pTyr site several highly ranked clusters were mapped, mainly in the helix that comprises the pY+0 pocket including key residues K591 and R595 (Figure 43.b). Three different probes identified the aforementioned region: benzene, imidazole and isopropanol. Interestingly, other cosolvent clusters (mainly acetamide) were placed mainly in two other regions from the SH2 domain. These correspond to the preferential hotspots for AutoDock4 and UCSF DOCK6 mentioned in section

4.2.1.3 (Figure 43.c). Regarding the DBD site, no clusters identified the internal residues of the pocket, but all probes identified at different levels of preference the external face of the pocket (K573, T515 and D334) (Figure 43.d). Acetanilide seemed to have a different behaviour, as it does not map any region from the SH2 or Linker domain, meaning that it has a considerable preference for the DNA-Binding Domain. Its affinity is very favourable on the region close to R414, residue that as a gatekeeper for DNA exit as mentioned in 6.2.2 (Figure 43.e).

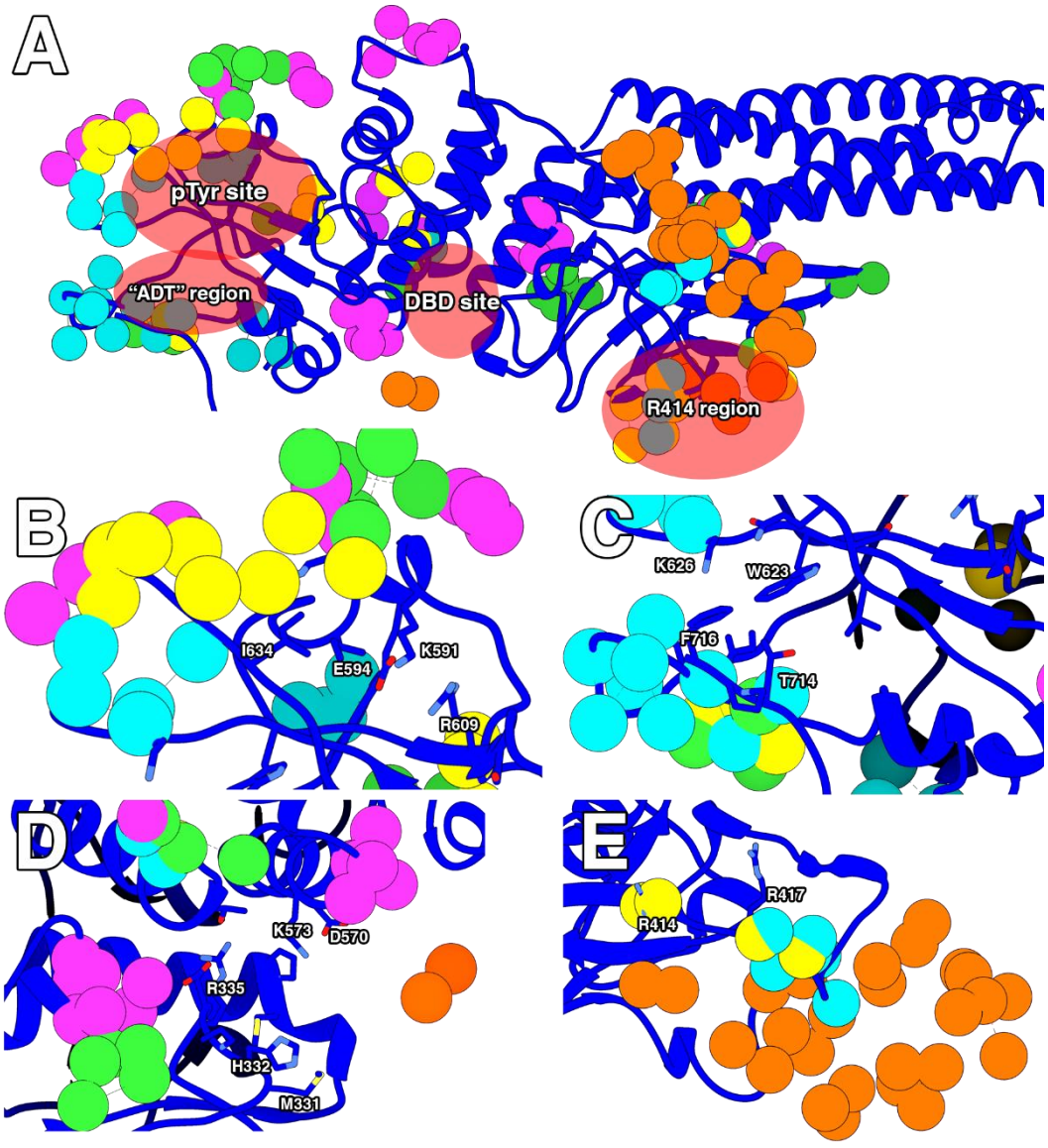


Figure 43 STAT3 hotspots found by CAT. Clusters have the following colours assigned: acetamide as cyan, benzene as magenta, acetanilide as orange, imidazole as yellow, and isopropanol as green. A) Panoramic view of the STAT3 monomer and the respective top regions identified by CAT. B) The loop (K591-E594) that forms the pY+0 pocket from the pTyr site is

identified by different cosolvents with high ranks. C) The L site, identified previously with AutoDock molecular docking is also mapped by CAT clusters as well as the newly identified DBD pocket (D) and a pocket close to the gatekeeper R414 (E)

Performing cosolvent MD and further CAT analysis with the STAT3 system is a challenge. STAT3 is a multidomain protein with scarce evidence around its "druggability". Cosolvent MD has been mainly tested with small proteins and/or single domains. This means that the bigger the system, the larger number of probes to be found in the system leading to a potential clustering and/or phase separation. Nevertheless, these results were deemed as satisfactory. The main goal was to identify both the pTyr and DBD binding sites, and while we succeed with the first, CAT struggles to map the DBD pocket due to its "hidden" nature. Again, it is likely that the binding site would be mapped more accurately if longer simulations were performed, but the risk to undergo phase separation is too big. As mentioned in section 1.3 one way to overcome this issue could have been to use of repulsion potentials in combination with longer simulations. Furthermore, regions that were classified as the preferential SH2 domain binding site for AutoDock4 and UCSF DOCK6 are also mapped. I believe that these regions should be taken into higher account as for example OPB-31121[127], one of the few inhibitors going through clinical trials, is believed to bind in the "AutoDock region". It is also interesting how CAT maps an area close to R414, the gatekeeper residues for the DNA binding/unbinding process. If ligand binding were to be achieved in that region, it could lead to a new mechanism of inhibition.

Table 16 STAT3 CAT results

Target	Binding Site	Protein Contacts	Cosolvent	CAT Rank
STAT3	pTyr site	M586, G587, I589, K591, R595, R609, E612, W623, K626, S636	Benzene	1,4
			Imidazole	1
			Isopropanol	3,10
			Acetamide	5

	DBD pocket	M331, H332, P333, D334, R335, P470, M471, T515, D570, K573	Benzene	2,4
			Imidazole	4
			Isopropanol	6

Chapter 7 Conclusions

The correct assessment of the structural changes within the protein target is crucial for the right evaluation of possible time-dependent binding sites. As such, an accurate tool is pivotal for selecting possible contact regions to be further studied. While standard analysis of hotspot mapping quantifies primarily the volume of the binding region, cosolvent MD simulation followed by CAT analysis focuses on the cosolvent-induced conformational changes, to map, assess, and rank the putative 'hotspots', via an empirical scoring function. This characteristic gives the algorithm presented herein a high level of robustness and reliability in searching and ranking hotspots, as shown by the comparison with experimental data and FTMAP predictions. The scoring function implemented in CAT makes it unique and distinct from computational methodologies reported in the literature.

In the present work, it has been developed, tested and validated the applicability of CAT analysis to detect several potentially druggable allosteric sites, which were detected by X-ray crystallography studies. The usage of five different cosolvent molecules demonstrated, at the same time, a broad sample space regarding interacting molecules and provides an insight on the chemical nature of the putative ligand moieties that would preferentially bind to the respective site. CAT is robust yet versatile: the analysis can be performed on cosolvent trajectories using any cosolvent molecule of choice.

The major shortcoming of CAT observed so far was its inability to map some deep buried pockets. This could be attributed to insufficient sampling during MD simulation, however FTmap performs very well on this task. Although this issue may be easily sorted by longer MD simulation in water prior to cosolvent MD simulations, a combination of both tools could be an interesting approach. I understand that the principle in which FTMap is based is not the same, although they share some features. The main goal while choosing this tool as a comparison with CAT relied in the ease of use and fast results one could get. The use of cosolvent tools such as CAT can give more insights in the dynamics and crypticity of the target in comparison to FTMap.

In future works, I would aim to explore CAT analysis applied to multi-cosolvent trajectories and to address the sampling problem which underlies the sub-optimal performance in mapping the buried pockets.

Using computational approaches based on atomistic molecular dynamics simulations, enabled me to understand the effects of specific STAT3 mutations, which were described in the literature, and to explain their modulation of the STAT3 activity. Consistently with Mertens and coworkers[138], D570K mutation exerted its effect by enhancing interactions between STAT3 and DNA, which interfered with the DNA release by the STAT3 dimer and thus inhibited the protein's function by not driving transcription and resisting dephosphorylation. Subsequently, recent identification of a plausible binding site for small molecule STAT3 inhibitor nababucasin (BBI-608) helped me to deconvolute its inhibition mechanism, with resembled the effect exerted by D570K mutation. The identification of the putative binding site for BBI-608 around the DNA binding domain may contribute to novel potent and selective STAT3 inhibitors, similarly as the binding site shown in Ji's work[133]. The accuracy and similarity between the model and the available patent data, raising the question on why this pocket was not identified before. Mutated inter-domain residues E435, W546 and K551 unveil a poor binding to DNA, leading to another way of targeting STAT3 by disrupting theses residues, pointing towards possible novel allosteric binding sites. Structure-based ligand design targeting these novel pockets, coupled with novel methodologies, such as employing recently the developed FragLites[237], is likely to expand a set of chemotypes active towards STAT3 and contribute to the development of novel inhibitors of this important yet very challenging drug target.

Appendix A

A.1 Cosolvent MD simulation protocol

Structure preparation

The crystal structures used as starting conformations for the cosolvent MD simulations were in the *apo* state, whenever available (PDB codes are listed in Table 9). Structures were stripped of water molecules and any present cofactors and/or ligands. For structures with missing loops, the MODELLER[238] interface in UCSF Chimera[239] was used to rebuild the missing fragments. The best ZDOPE scored loops were selected to complete the model.

Incomplete side chains were replaced using the Dunbrack rotamer library[240], implemented in UCSF Chimera[21]. For side chains with multiple locations, the highest occupancy conformations have been selected. Structural hydrogens were added and the following protein parametrisation was performed using the Gromacs 2016.03[241] suite with AMBERFF99SB-ILDN[148] force field. A cubic box was centred around the protein target with 1 nm distance between the protein extreme to the edge. A pre-defined number of molecular probes (cosolvent molecules) were randomly inserted into to the system, ensuring that after the following solvation with TIP3P waters there was a 10% (m/m) probe concentration in water in order to avoid phase separation and/or probe clustering. Each simulation used a single type of cosolvent molecule. The probe selection criteria consisted of using a series of drug-like small molecular fragments with a broad range of relevant properties, including hydrophilicity/hydrophobicity, aromaticity, and number of hydrogen-bonding acceptors/donors, with a diverse range of logP values. The following molecules were used: acetamide, benzene, acetanilide, imidazole and isopropanol. To diminish the effect of phase separation and $\pi - \pi$ stacking of aromatic and very hydrophobic cosolvent molecules such as benzene, an approach similar to Mackerell and colleagues[242] was chosen, which relied on placing a dummy atom with a negligible negative charge (e=-0.01) in the centre of the 6-membered ring. All probes were parametrised using GAFF[243] with AM1-BCC[244] charges assigned by ACPYPE/ANTECHAMBER[245].

MD simulation protocol

To maintain the charge neutrality of the simulated unit, sodium and chloride ions were added to a concentration of 0.1 M. Bonds were constrained using the LINCS[246] algorithm, with a 2 fs time step. The electrostatic interactions were calculated using the particle-mesh Ewald method, with a non-bonded cut-off set at 0.1 nm. All structures were minimised via the steepest descent algorithm for 20000 steps was stopped when the maximum force fell below 1000kJ/mol/nm using the Verlet cutoff scheme. After minimisation, heating via NVT ensemble was performed for 100 ps with a time step of 2 fs with position restraint (1000 kJ/mol.mn^2 in all three dimensions) applied to the backbone. The temperature coupling was set between the protein and the non-protein entities by using a Berendsen thermostat, with a time constant of 0.1 ps and the temperature set to reach 300 K with the pressure coupling off. Sequentially, a pressure NPT ensemble equilibration was performed followed by 100 ps, and three NPT ensemble production run replicas of 50 ns, totalling 150 ns for each different combination of protein and cosolvents, including the control simulations that are comprised of only protein-water systems. All production runs were unrestrained simulations.[247] The temperature was set constant at 300 K by using a modified Berendsen thermostat (τ = 0.1 ps) [156]. Pressure was kept constant at 1 bar by Parinello-Rahman isotropic coupling (τ = 2.0 ps) to a pressure bath.

Data analysis has initially been done within the Gromacs package. For each data set, the analysis involved calculating root-mean-square deviation (RMSD), root-mean-square fluctuations (RMSF), the covariance matrices and principal component analysis (PCA) and solvent accessible surface area (SASA) to analyse convergence of the runs. Afterwards, CAT analysis was employed for every dataset to identify any potentially druggable hotspots.

A.2 CAT supplementary information

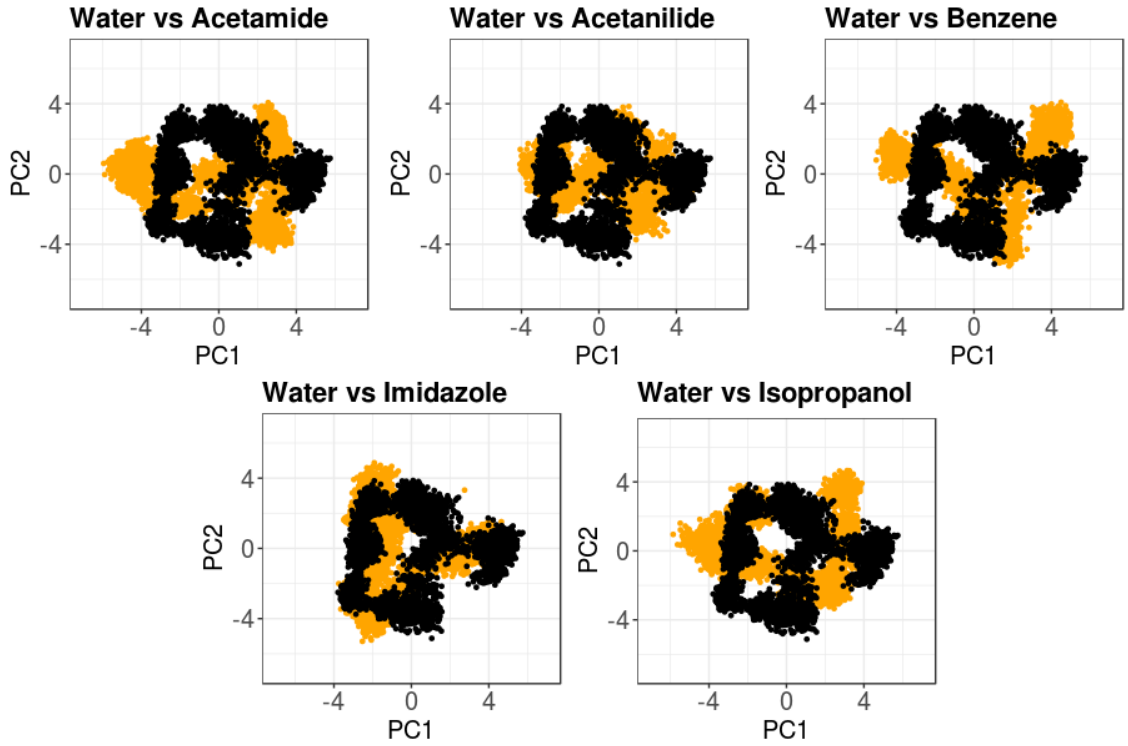


Figure A.1 PCA distribution of control simulations (protein-water) with cosolvent systems in HRAS GTPase. Black dots correspond to the control simulation, and orange dots to the cosolvent MD.

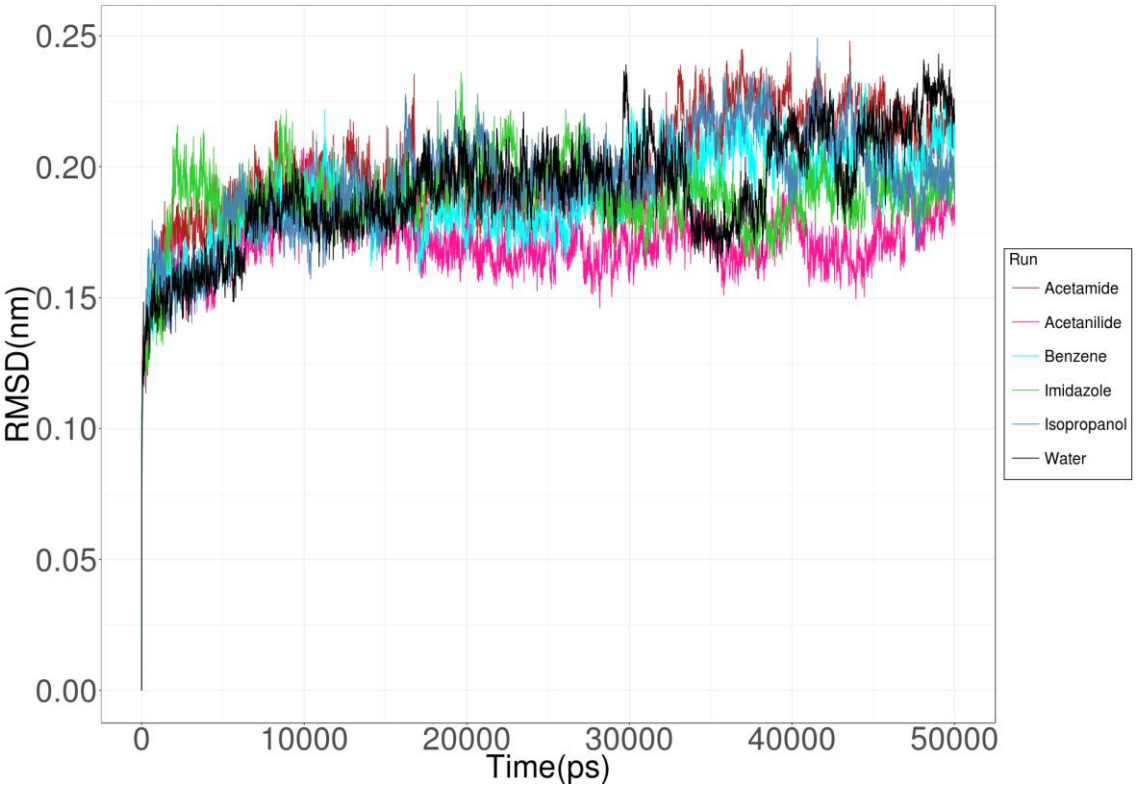


Figure A.2 RMSD plots for HRAS 5 cosolvent runs and water simulation.

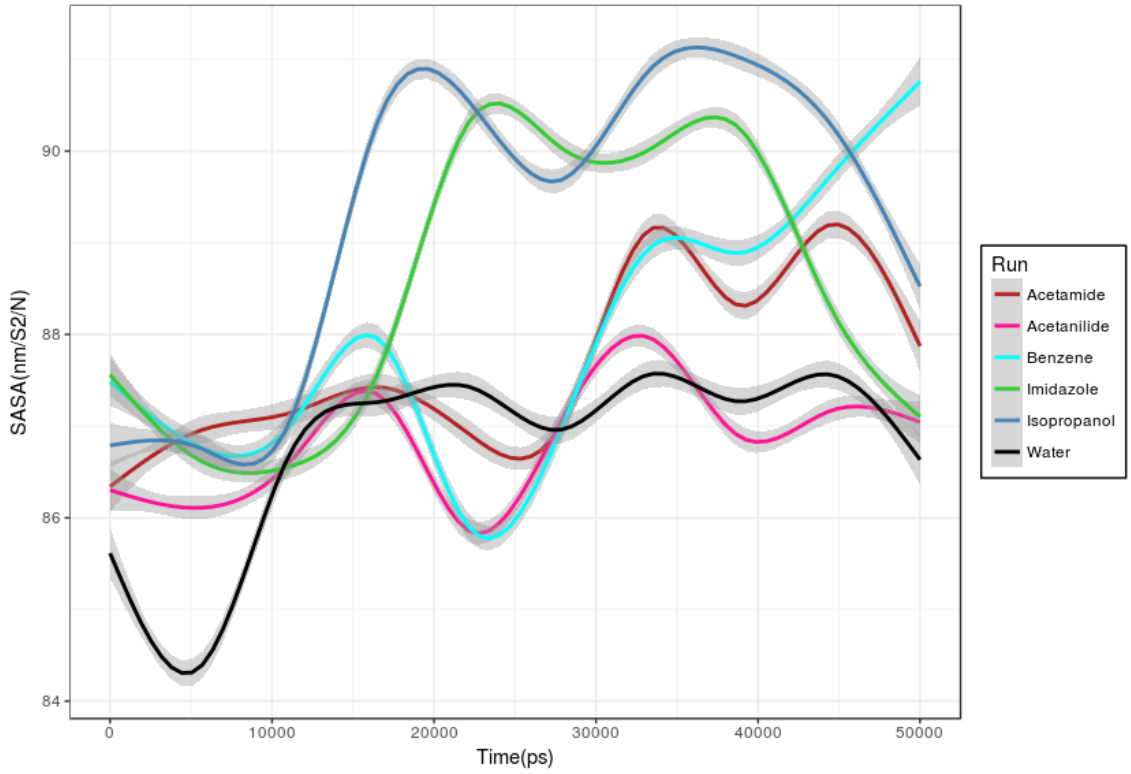


Figure A.3 SASA plots for HRAS 5 cosolvent runs and water simulation.

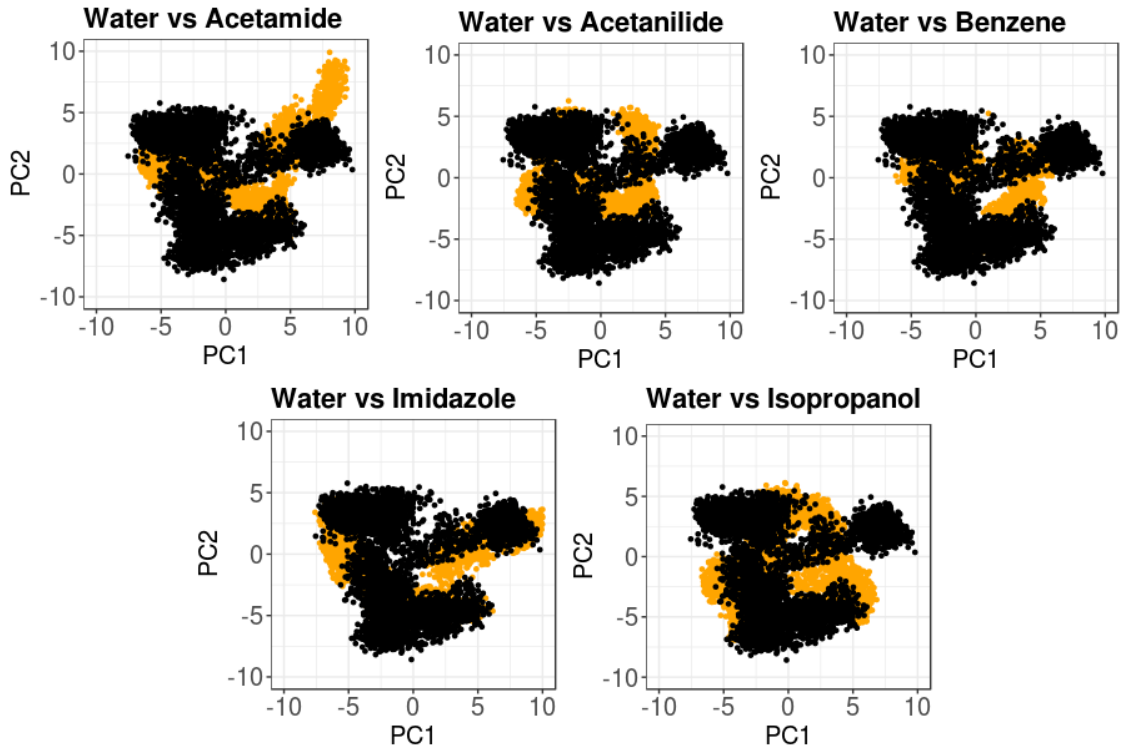


Figure A.*4* PCA distribution of control simulations (protein-water) with cosolvent systems in PTP1B. Black dots correspond to the control simulation, and orange dots to the cosolvent MD.

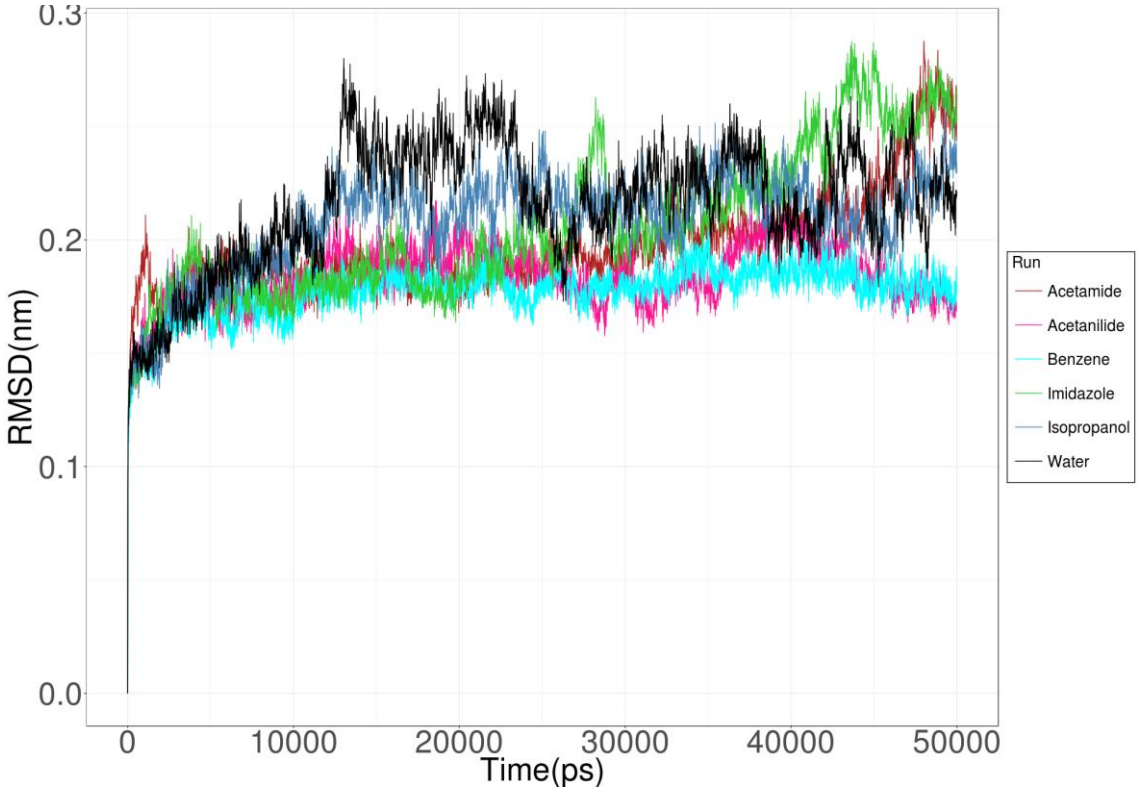


Figure A.5 RMSD plots for PTP1B for all 5 cosolvent runs and water simulation.

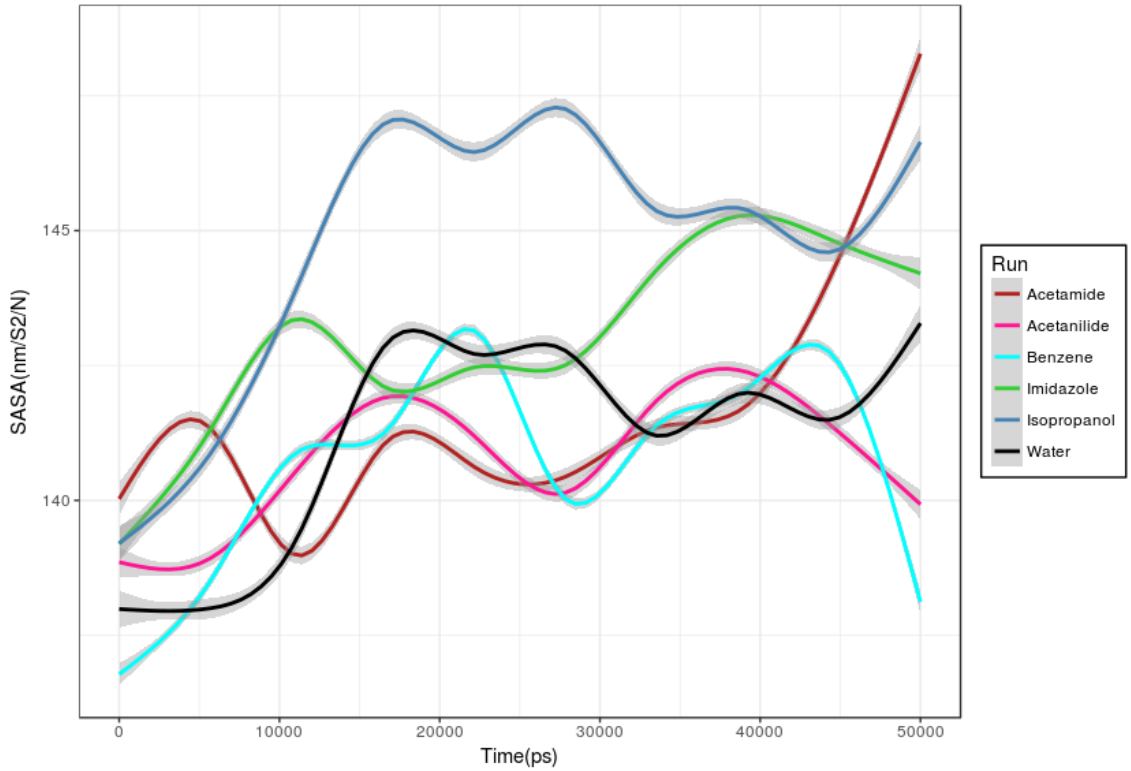


Figure A.6 SASA plots for HRAS 5 cosolvent runs and water simulation.

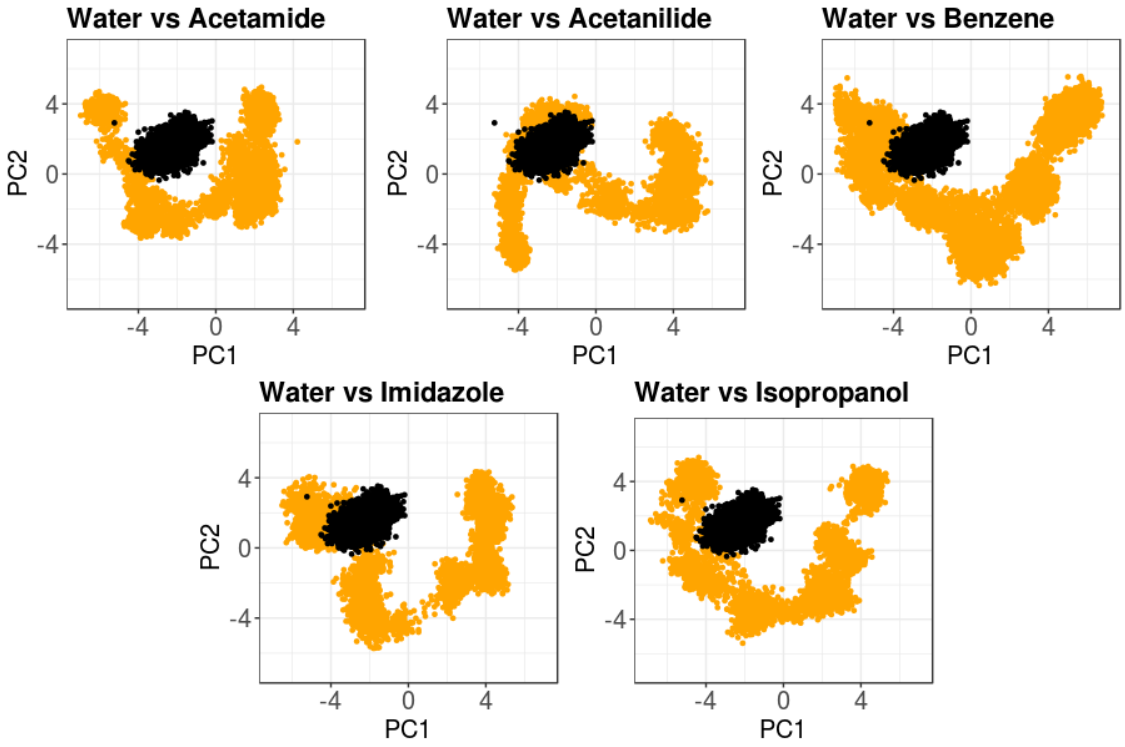


Figure A.7 PCA distribution of control simulations (protein-water) with cosolvent systems in AR ligand-binding domain. Black dots correspond to the control simulation and orange dots to the cosolvent MD.

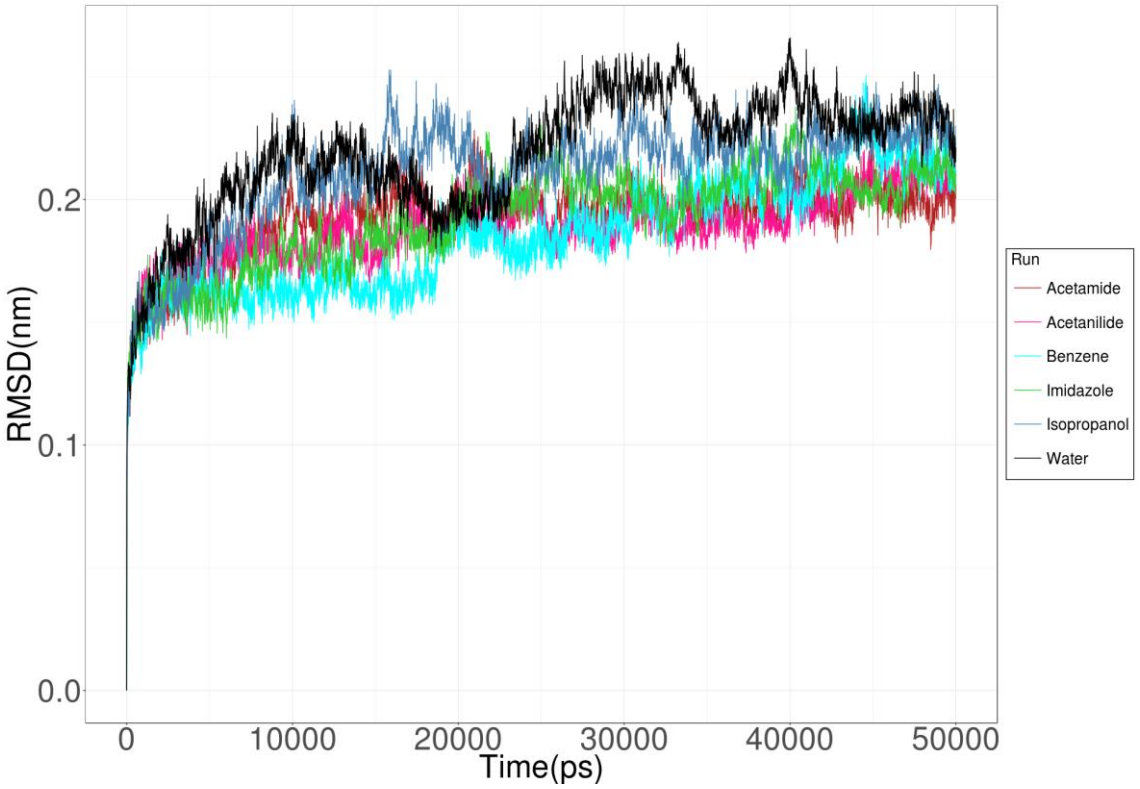


Figure A.8 RMSD plots for AR-LBD cosolvent runs and water simulation.

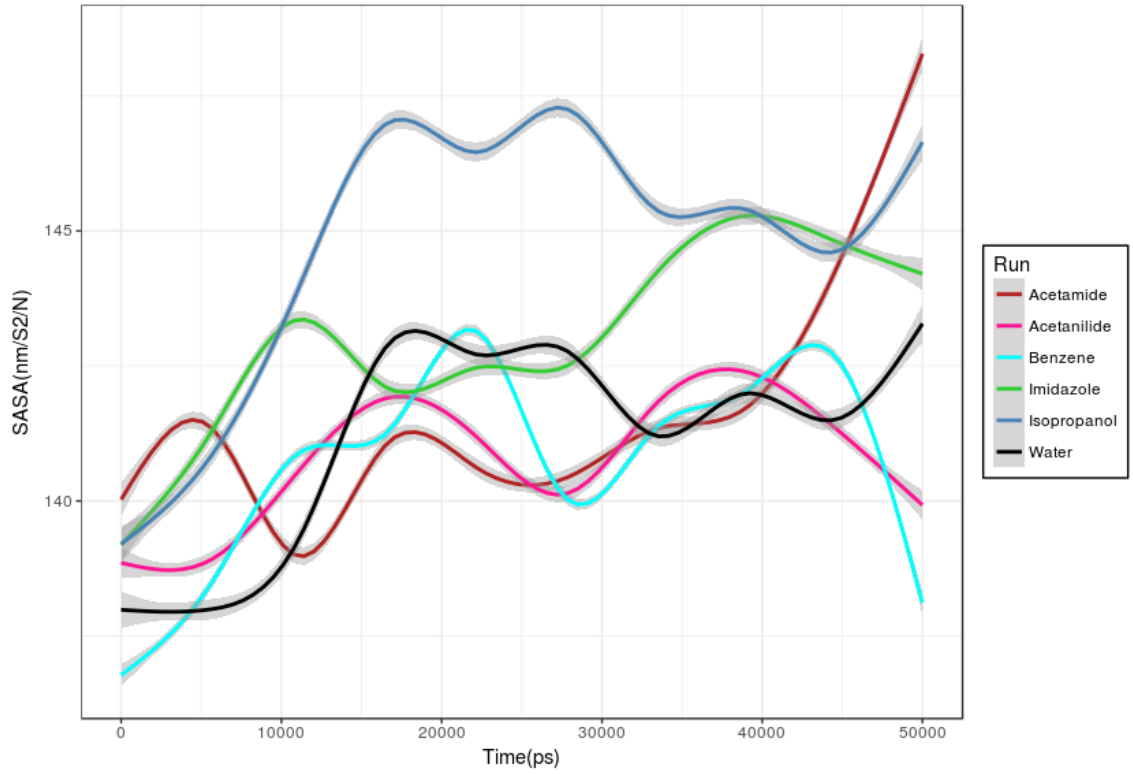


Figure A.9 RMSD plots for AR-LBD cosolvent runs and water simulation.

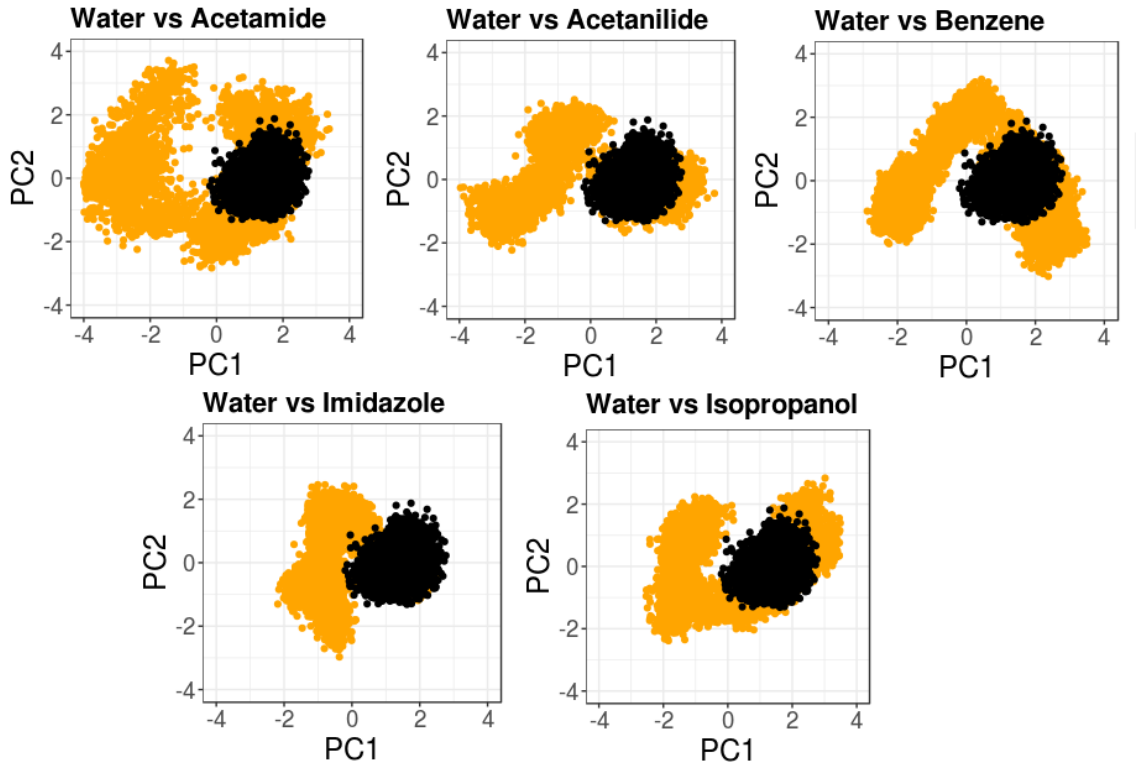


Figure A.10 PCA distribution of control simulations (protein-water) with cosolvent systems for CDK2. Black dots correspond to the control simulation and orange dots to the cosolvent MD.

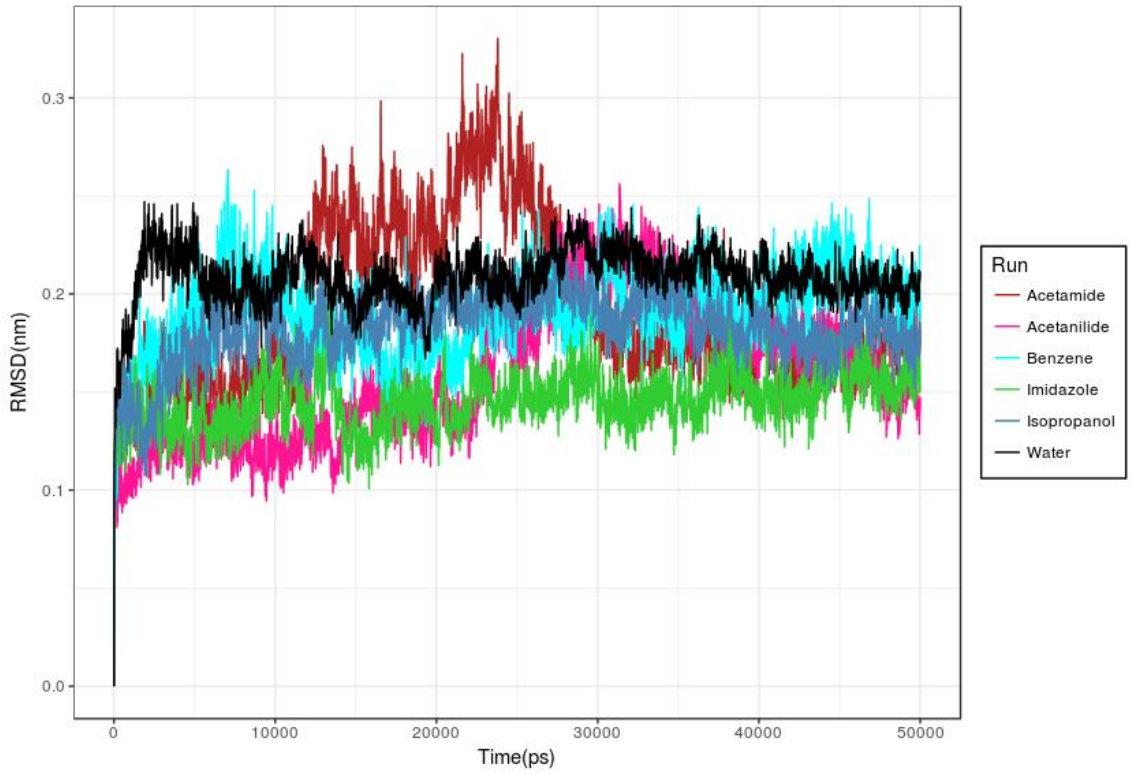


Figure A.11 RMSD plots for CDK2 cosolvent runs and water simulation.

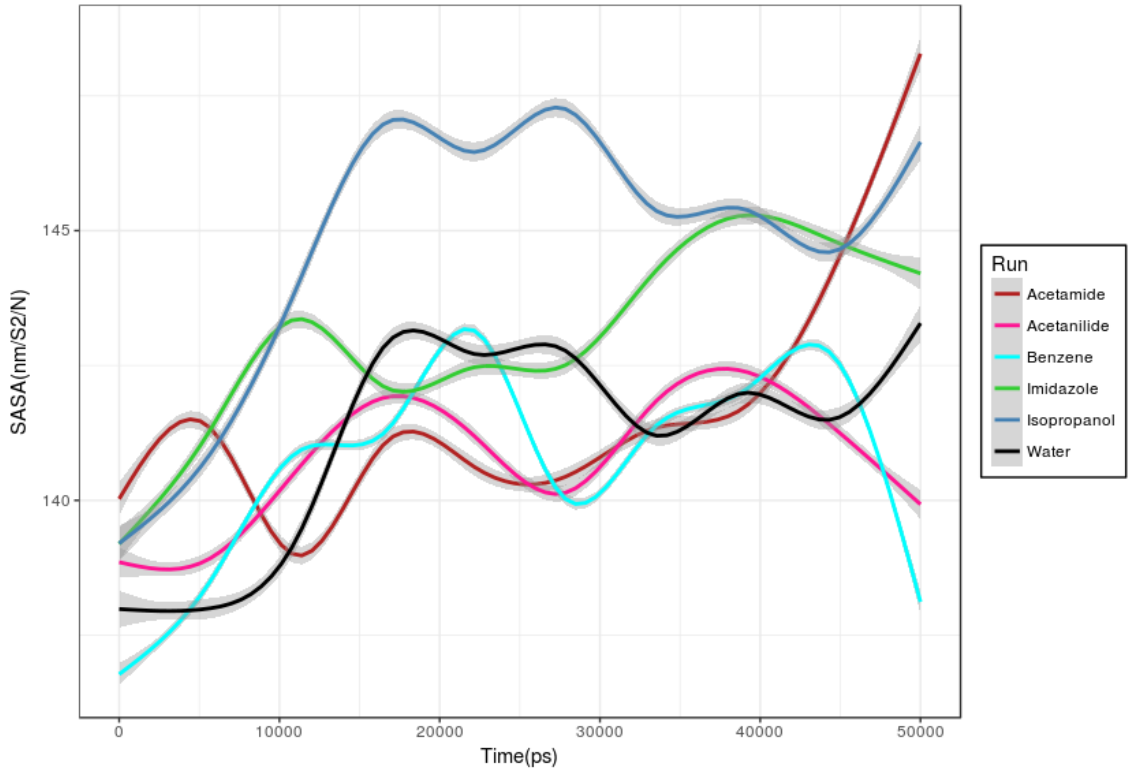


Figure A.12 SASA plots for CDK2 5 cosolvent runs and water simulation.

Table A.*1* CAT scores for all targets and the top 10 clusters

AR-LBD					
Rank	**Acetamide**	**Benzene**	**Isopropanol**	**Acetanilide**	**Imidazole**
1	-0,65	-0,68	-0,93	-1,15	-1,73
2	-0,61	-0,64	-0,84	-0,98	-1,49
3	-0,56	-0,54	-0,81	-0,83	-1,29
4	-0,51	-0,45	-0,71	-0,74	-1,13
5	-0,46	-0,35	-0,68	-0,62	-1,02
6	-0,43	-0,35	-0,62	-0,54	-0,94
7	-0,43	-0,33	-0,56	-0,54	-0,87

8	-0,41	-0,27	-0,5	-0,49	-0,84
9	-0,41	-0,24	-0,42	-0,47	-0,82
10	-0,4	-0,24	-0,42	-0,45	-0,81
PTP1B					
Rank	**Acetamide**	**Benzene**	**Isopropanol**	**Acetanilide**	**Imidazole**
1	-2,3	-0,47	-1,65	-0,88	-1,68
2	-1,08	-0,47	-0,92	-0,53	-1,02
3	-0,8	-0,42	-0,58	-0,44	-0,78
4	-0,68	-0,31	-0,57	-0,4	-0,77
5	-0,67	-0,3	-0,49	-0,4	-0,74
6	-0,54	-0,28	-0,43	-0,34	-0,73
7	-0,53	-0,26	-0,42	-0,29	-0,65
8	-0,52	-0,24	-0,42	-0,29	-0,62
9	-0,51	-0,22	-0,41	-0,29	-0,58
10	-0,51	-0,22	-0,4	-0,28	-0,57
HRas					
Rank	**Acetamide**	**Benzene**	**Acetanilide**	**Imidazole**	**Isopropanol**
1	-0,64	-0,32	-0,4	-0,83	-0,55
2	-0,64	-0,19	-0,33	-0,64	-0,5
3	-0,57	-0,17	-0,3	-0,59	-0,49
4	-0,52	-0,17	-0,28	-0,57	-0,49

5	-0,52	-0,16	-0,22	-0,56	-0,42
6	-0,45	-0,15	-0,21	-0,51	-0,41
7	-0,43	-0,15	-0,2	-0,49	-0,41
8	-0,39	-0,14	-0,19	-0,45	-0,34
9	-0,37	-0,14	-0,19	-0,44	-0,32
10	-0,34	-0,08	-0,19	-0,41	-0,31
CDK2					
Rank	**Acetamide**	**Benzene**	**Acetanilide**	**Imidazole**	**Isopropanol**
1	-0,74	-0,94	-0,89	-0,92	-1,21
2	-0,67	-0,78	-0,87	-0,78	-1,01
3	-0,61	-0,68	-0,71	-0,75	-0,74
4	-0,57	-0,52	-0,63	-0,62	-62
5	-0,56	-0,46	-0,59	-0,58	-0,62
6	-0,55	-0,43	-0,57	-0,55	-0,53
7	-0,5	-0,41	-0,56	-0,55	-0,47
8	-0,5	-0,35	-0,54	-0,51	-0,46
9	-0,48	-0,35	-0,53	-0,51	-0,45
10	-0,45	-0,32	-0,52	-0,5	-0,44

Appendix B

B.1 Chapter 6 Simulation/Docking protocol

Molecular modelling of human STAT3 dimer.

Initial models of dimeric human STAT3 (wild type and mutants) in complex with DNA were created using crystal structure of unphosphorylated mouse STAT3B (PDB code: 4E68), which spans residues 136-716. The N-terminal domain has been excluded from the structures subjected to the simulations. Loops spanning the residues 184-194 and 688-702 were modelled using MODELLER interface[248,249] in UCSF Chimera[250]. The DNA double strand bounded to the model was designed based on 4E68, with the 5'-3' strand sequence as TGCATTTCCCGTAATC. The final model was subjected to 20000 cycles of steepest descent energy minimisation.

The STAT3 mutations (Figure 13), which were selected following the study by Mertens and coworkers[138], were introduced in UCSF Chimera by swapping side chains to the target residues and adjusting new conformations using Dunbrack rotamer library integrated within UCSF Chimera.

Modelling of ligand-bound STAT3. The crystal structure of ligand-bound STAT3 is yet not available, therefore we built the most similar model possible with the accessible data. Starting from the dimer model of wild-type STAT3 described in the previous section, molecular docking calculations were performed with MOE[251], using napabucasin (BBI-608) as the ligand. To ensure the scoring function accuracy with the target and the best possible fit both blind (full dimer) and targeted docking (pocket described) were performed. 200 different conformations of the ligand were scored per each run using Triangle Matcher, and London dG for the first scoring function. Thereafter, the top 100 conformations were rescored using Induced Fit and GBVI/WSA (Generalized-Born volume integral/weighted surface area) score [251]. From the final poses obtained, one for each monomer was selected based on score, interactions and consistency with the experimental structure shown in Ji's work[133] which shows BB1-608 bound to a pocket located within the DNA-STAT3 interface.

Molecular dynamics and umbrella sampling simulations. In order to study the dynamics effects of mutations in the STAT3 dimer, each of the listed mutation were created in UCSF Chimera[250]. Structural hydrogens were added and the following protein parametrisation was performed using the Gromacs 2016.03[252] suite with AMBERFF99SB-ILDN[253] force field. Before pulling the DNA from the complex, the systems were relaxed with a short equilibrium MD production run. Hence, a 1 nm cubic box was centred on the structure and the system is solvated with TIP3P waters. Sodium and chloride ions were added to a concentration of 0.1 M resulting in systems with more than seven hundred thousand atoms. Bonds were constrained using the LINCS[254] algorithm. The electrostatic interactions were calculated using particle-mesh Ewald method, with a non-bonded cut-off set at 0.1 nm. All structures were minimised via the steepest descent algorithm for 20,000 steps of 0.02 nm, and minimisations were stopped when the maximum force fell below 1000kJ/mol/nm using the Verlet cut-off scheme[255]. After minimisation, temperature equilibration was performed for 100ps with a time step of 2fs with position restraints applied to the backbone using an NVT. The temperature coupling was set between the protein and the non-protein entities by using a Berendsen thermostat[256], with a time constant of 0.1 ps, and the temperature set to reach 300K with the pressure coupling off. Sequentially, a pressure NPT equilibration was performed followed by 100ps of an NVT equilibration, the following 100ps of NPT equilibration, and a production run of 100 ns. Temperature was set constant at 300K by using a modified Berendsen thermostat (τ = 0.1 ps)[256]. Pressure was kept constant at 1 bar by Parinello-Rahman isotropic coupling (τ = 2.0 ps) to a pressure bath[257].

For the umbrella sampling simulation, a 50 ns pre-umbrella equilibration was made, with the complex rotate its principal axis to align with the z-axis of the simulation box. A pull sampling was used using a constant force approach (k = 1000 kJ/mol/nm, with a rate of 0.01 nm) between the centres of masses of SH2 domain and the DNA double helix, along the described path shown in Figure 2. From each corresponding pull simulation, a series of conformations have been selected in order to sample the process of entering-exiting the DNA-binding site. Each of the 25 selected umbrella windows has been through a 1ns NPT

equilibration run, followed by a 5ns NPT distance restrained production run, totalizing per system, 135ns of simulation time, using the previously described protocol and parameters. Afterwards, the potential of mean force (PMF) curve of the studied scenario has been calculated with the Weighted Histogram Analysis Method (WHAM) tool from Gromacs[258], and associated errors was calculated using both a convergence criteria and the implemented bootstrap method in gromacs WHAM. All calculations for the analysis were made using the gromacs tools.

B.2 STAT3 supplementary information

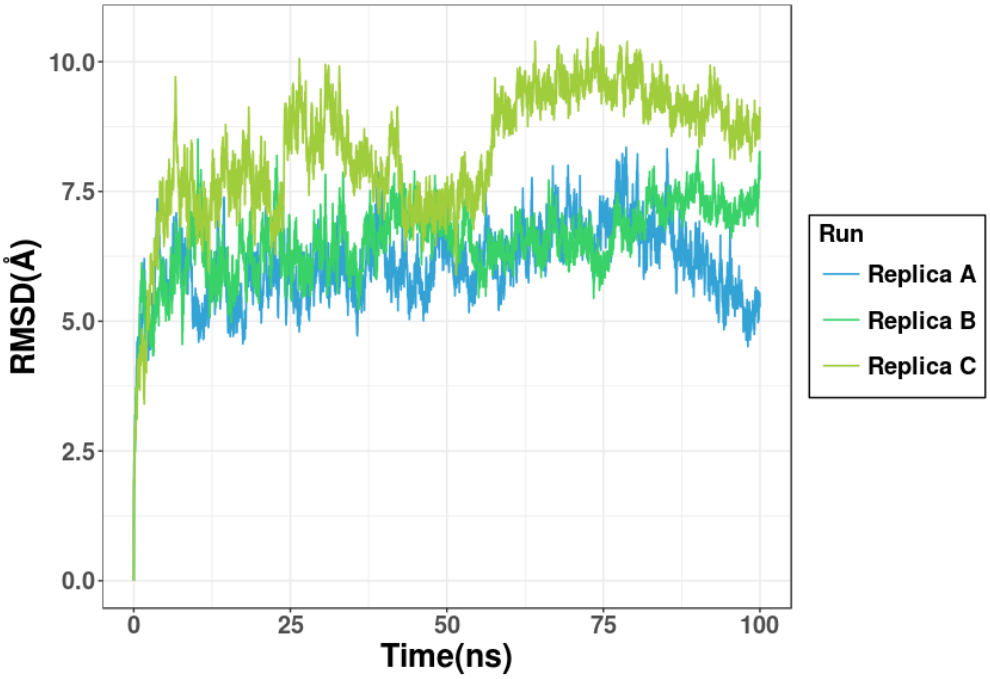


Figure A.13 RMSD values for the three performed simulations of STAT3-BBI608 complexes. For all three replicas, the systems attained convergence. However, the replica C had shown some loop fluctuations which increased the RMSD values.

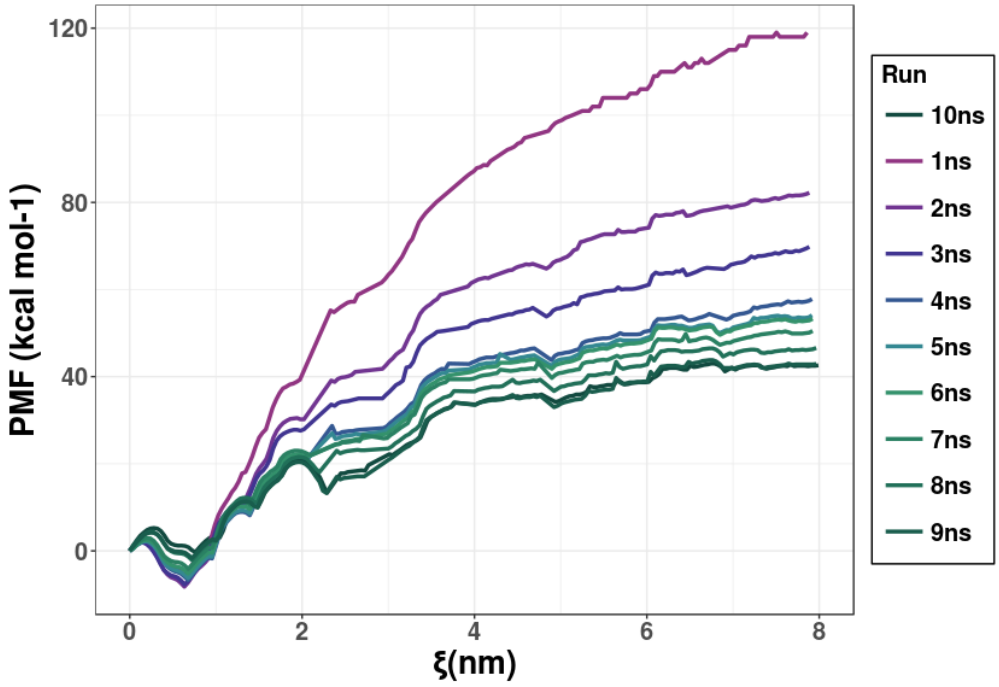


Figure A.14 PMF per ns for the WT apo STAT3, where ξ is the DNA pulling coordinate. We can see a clear convergence after 5 ns. The variations are negligible after 7 ns.

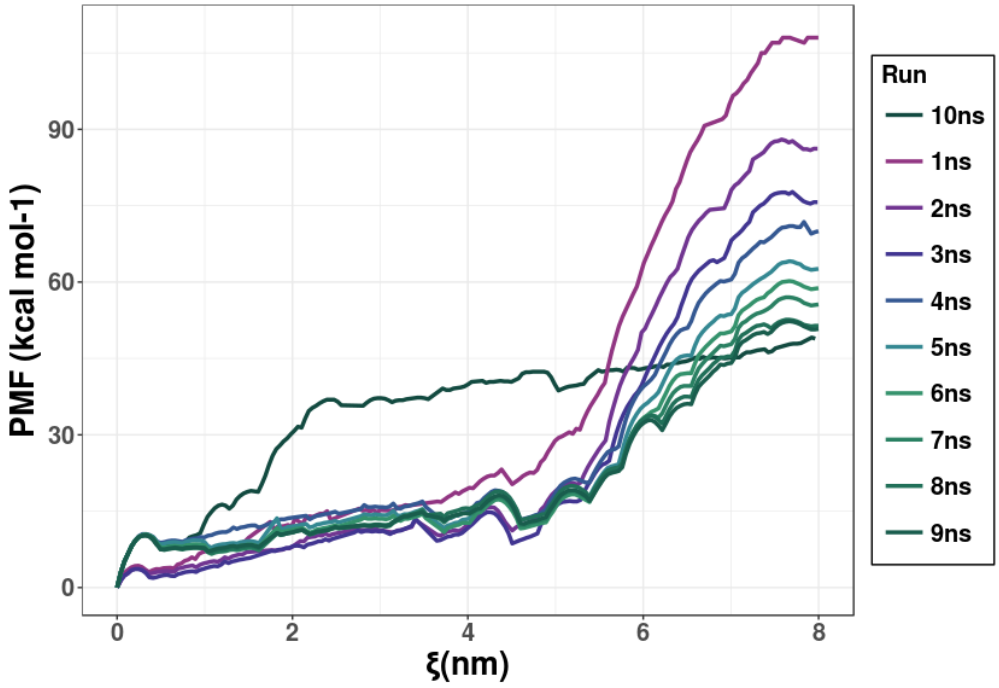


Figure A.15 PMF per ns for the K551A STAT3 mutant, where ξ is the DNA pulling coordinate. After 7 ns, the overall energy of unbinding has converged. However, after 9 ns, a new energetic state has been sampled. This does not change the ΔG but does change the PMF landscape.

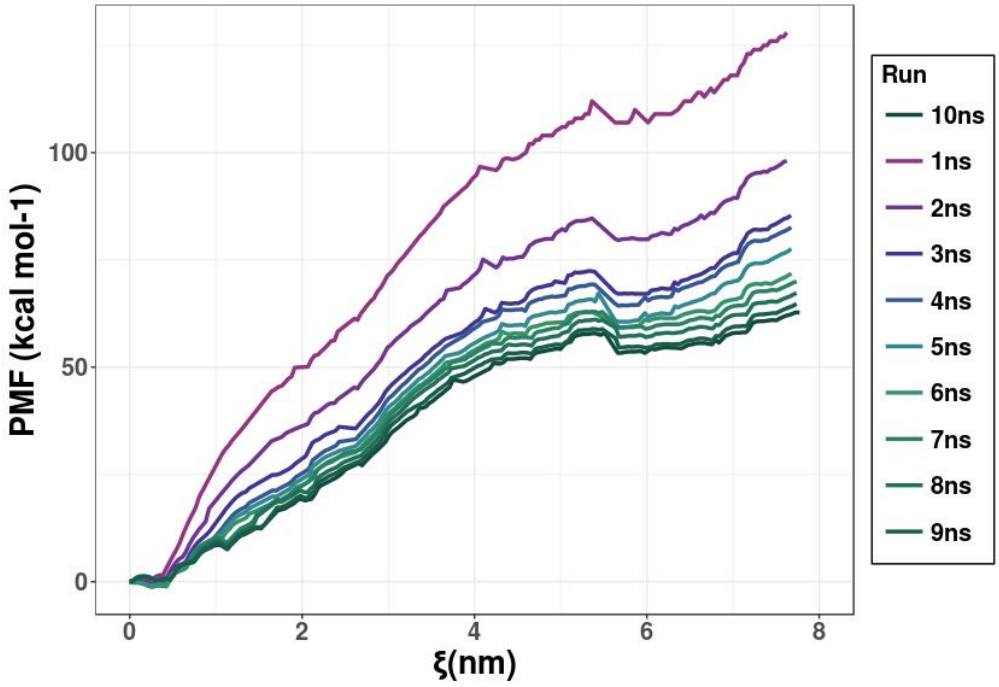


Figure A.16 PMF per ns for the EE434435AA STAT3 double mutant, where ξ is the DNA pulling coordinate. The system has a steady convergence after 5 ns, with negligible fluctuations after 7 ns.

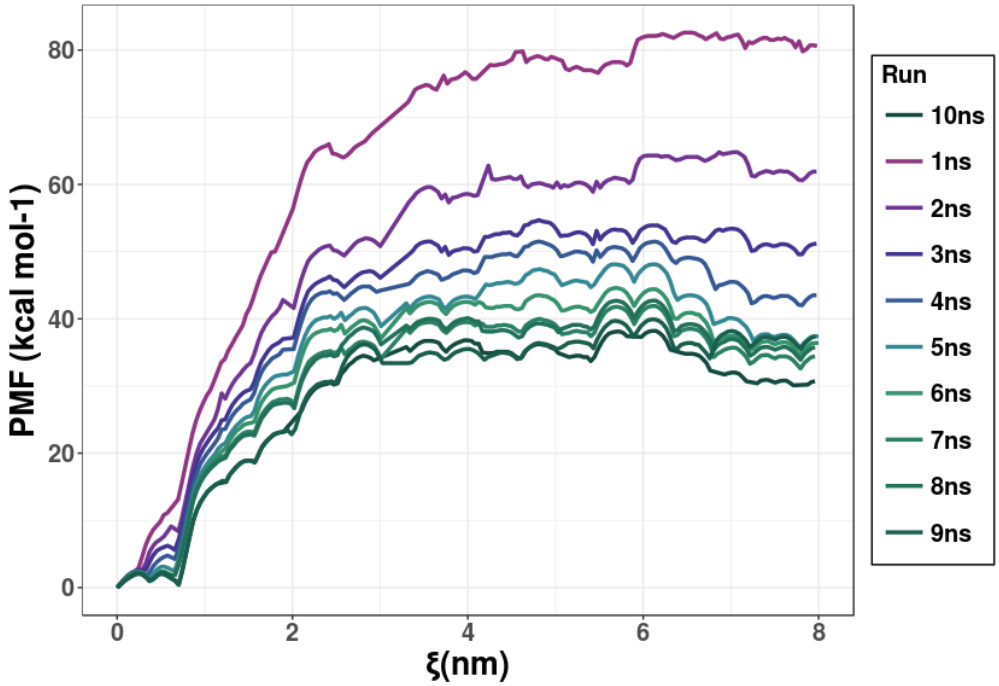


Figure A.17 PMF per ns for the W546A STAT3 mutant, where ξ is the DNA pulling coordinate. The system has a steady convergence after 5 ns, with negligible fluctuations after 7 ns

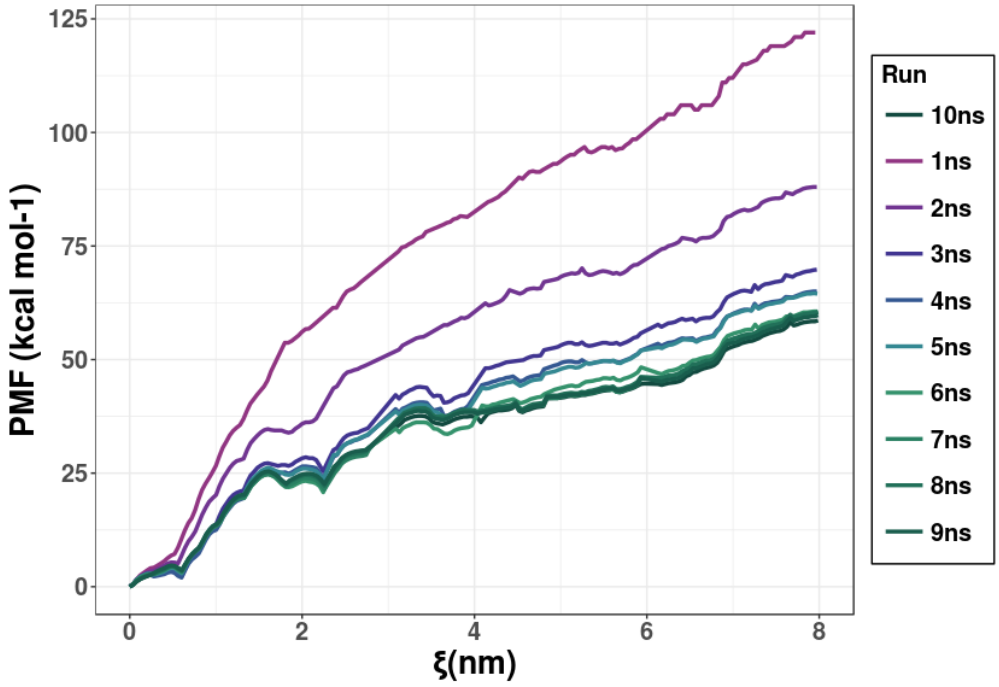


Figure A.18 6 PMF per ns for the D570K STAT3 mutant, where ξ is the DNA pulling coordinate. The system has a steady convergence after 3 ns, with negligible fluctuations after 5 ns.

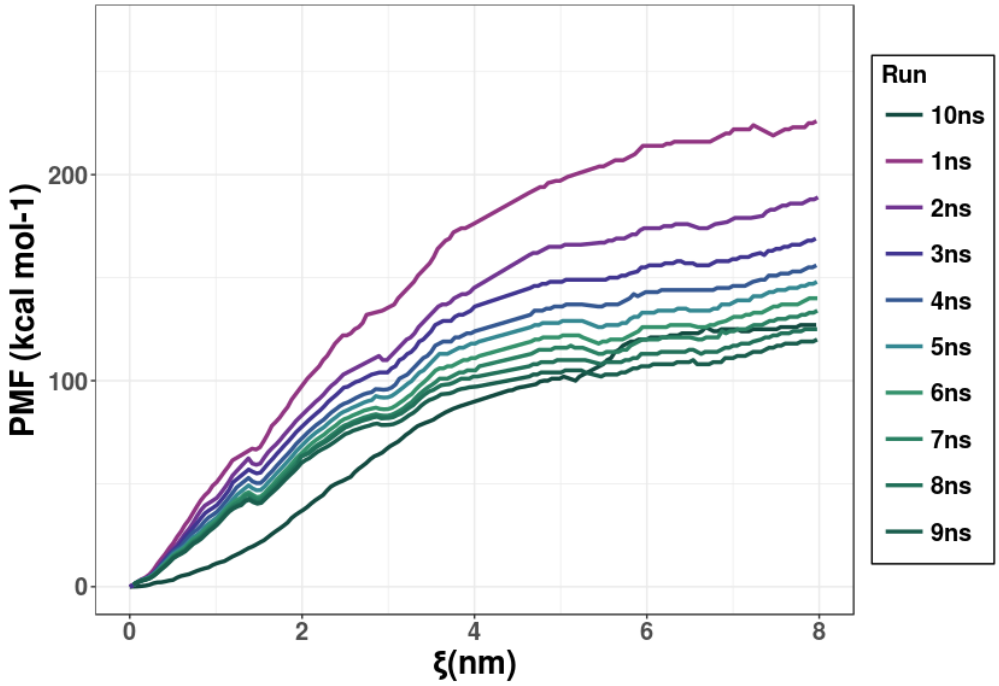


Figure A.19 PMF per ns for the BBI608-STAT3-DNA complex. where ξ is the DNA pulling coordinate. The system has a steady convergence after 5 ns. Regardless of the PMF landscape change in the 10ns run, the overall ΔG has been achieved convergence after 7 ns.

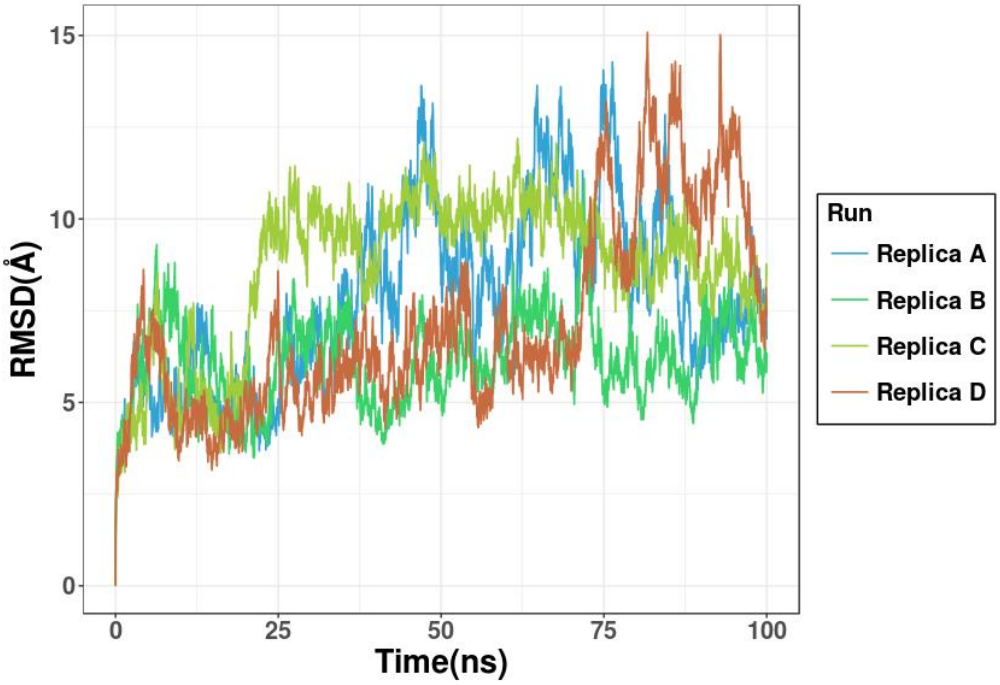


Figure A.20 8 RMSD values for the performed STAT3-BBI608 simulations without DNA. Given the lack of DNA structure to stabilize the dimer, the RMSD fluctuates significantly more in comparison to the DNA-bound dimer.

Appendix C

C.1 Equipment

Computational methods and procedures described in this work were accomplished with in-house systems and high performance computing resources (HPC):

C.1.1 In-house equipment

Local machine:

- 1 Intel Core E3-1200 processor (4.00 GHz, 8 cores, X MB cache)
- 1 GeForce GTX 1070 GPU
- 16 GB memory
- 6 TB SATA Disk

C.1.2 HPC resources

Rocket (Newcastle University)

110 standard nodes, each with:

- 2 Intel Xeon E5-2699 v4 processors (2.2 GHz, 22 cores, 55 MB cache)
- 44 cores (2 processors * 22 cores), totalling 4840 across the standard nodes
- 128 GB memory - (8 DDR4 RDIMMs, each with 16GB) – ie. 2.9 GB per core
- 600 GB SAS disk (469 GB scratch space)

6 medium (M) nodes:

- 2 Intel Xeon E5-2699 v4 processors (2.2 GHz, 22 cores, 55 MB cache)

- 44 cores (2 processors * 22 cores), totalling 264 across the medium nodes
- 512 GB memory - (16 DDR4 RDIMMs, each with 32GB) – ie. 11.6 GB per core
- 1200 GB SAS disk (1.1 TB scratch space)

4 large (L) nodes:

- 2 Intel Xeon E5-2699 v4 processors (2.2 GHz, 22 cores, 55 MB cache)
- 44 cores (2 processors * 22 cores), totalling 176 across the large nodes
- 512 GB memory - (16 DDR4 RDIMMs, each with 32GB) – ie. 11.6 GB per core
- 2 * 4000 GB SAS disks (7.2 TB scratch space)

2 extra-large (XL) nodes:

- 4 Intel Xeon E7-4830 v4 processors (2.0 GHz, 14 cores, 35 MB cache)
- 56 cores (4 processors * 14 cores), totalling 112 across the extra-large nodes
- 1536 GB memory - (48 DDR4 RDIMMs, each with 32GB) – ie. 27.4 GB per core
- 8 * 1200 GB SAS disks (8.7 TB scratch space)

1 GPU node:

- 2 * 16 core IBM POWER9 processors (2.6GHz, 3.09 GHz Turbo)
- 4 NVIDIA Tesla V100 GPUs, each with 16GB VRAM and NVLink2 interconnect
- 256 GB memory

- 893 GB scratch space

C.2 Software used

Name	Version	Reference
Data visualization and analysis		
VMD	1.9.3	202
Pymol	2.0	259
UCSF Chimera	1.12	239
RStudio	1.2.5001	260
Docking		
MOE	2016.01	190
AutoDock	4.2.6	182
AutoDock Tools	1.5.6	182
Raccoon	1.0b	261
UCSF DOCK	6.9	181
Molecular dynamics simulation and analysis		
Gromacs	2016.1	241

Appendix D

D.1 Digital repositories

Digital data such as molecular docking results, modelled systems or molecular dynamics trajectories can be found at:

http://tiny.cc/SabanesData

or

The code for the Cosolvent Analysis Toolkit as well as tutorials can be found on github:

http://tiny.cc/CAT_MD

or

References

Las actualizaciones automáticas de citas están deshabilitadas. Para ver la bibliografía, haz clic en Actualizar en la pestaña de Zotero.(1) Kenakin, T. P. ‘7TM Receptor Allostery: Putting Numbers to Shapeshifting Proteins. *Trends Pharmacol. Sci.* **2009**, *30* (9), 460–469. https://doi.org/10.1016/j.tips.2009.06.007.

(2) Kenakin, T.; Miller, L. J. Seven Transmembrane Receptors as Shapeshifting Proteins: The Impact of Allosteric Modulation and Functional Selectivity on New Drug Discovery. *Pharmacol. Rev.* **2010**, *62* (2), 265–304. https://doi.org/10.1124/pr.108.000992.

(3) Fenton, A. W. Allostery: An Illustrated Definition for the ‘Second Secret of Life.’ *Trends Biochem. Sci.* **2008**, *33* (9), 420–425. https://doi.org/10.1016/j.tibs.2008.05.009.

(4) Bridges, T. M.; Lindsley, C. W. G-Protein-Coupled Receptors: From Classical Modes of Modulation to Allosteric Mechanisms. *ACS Chem. Biol.* **2008**, *3* (9), 530–541. https://doi.org/10.1021/cb800116f.

(5) Wenthur, C. J.; Gentry, P. R.; Mathews, T. P.; Lindsley, C. W. Drugs for Allosteric Sites on Receptors. *Annu. Rev. Pharmacol. Toxicol.* **2014**, *54* (1), 165–184. https://doi.org/10.1146/annurev-pharmtox-010611-134525.

(6) Melancon, B. J.; Hopkins, C. R.; Wood, M. R.; Emmitte, K. A.; Niswender, C. M.; Christopoulos, A.; Conn, P. J.; Lindsley, C. W. Allosteric Modulation of Seven Transmembrane Spanning Receptors: Theory, Practice, and Opportunities for Central Nervous System Drug Discovery. *J. Med. Chem.* **2012**, *55* (4), 1445–1464. https://doi.org/10.1021/jm201139r.

(7) Fang, Z.; Grütter, C.; Rauh, D. Strategies for the Selective Regulation of Kinases with Allosteric Modulators: Exploiting Exclusive Structural Features. *ACS Chem. Biol.* **2013**, *8* (1), 58–70. https://doi.org/10.1021/cb300663j.

(8) Meanwell, N. A.; Kadow, J. F. Maraviroc, a Chemokine CCR5 Receptor Antagonist for the Treatment of HIV Infection and AIDS. *Curr. Opin. Investig. Drugs Lond. Engl. 2000* **2007**, *8* (8), 669–681.

(9) Olsen, R. W. GABAA Receptor: Positive and Negative Allosteric Modulators. *Neuropharmacology* **2018**, *136* (Pt A), 10–22. https://doi.org/10.1016/j.neuropharm.2018.01.036.

(10) Sigel, E.; Baur, R. Allosteric Modulation by Benzodiazepine Receptor Ligands of the GABAA Receptor Channel Expressed in Xenopus Oocytes. *J. Neurosci. Off. J. Soc. Neurosci.* **1988**, *8* (1), 289–295.

(11) Hu, X.; Tian, X.; Guo, X.; He, Y.; Chen, H.; Zhou, J.; Wang, Z. J. AMPA Receptor Positive Allosteric Modulators Attenuate Morphine Tolerance and Dependence. *Neuropharmacology* **2018**, *137*, 50–58. https://doi.org/10.1016/j.neuropharm.2018.04.020.

(12) Mitchell, N. A.; Fleck, M. W. Targeting AMPA Receptor Gating Processes with Allosteric Modulators and Mutations. *Biophys. J.* **2007**, *92* (7), 2392–2402. https://doi.org/10.1529/biophysj.106.095091.

(13) Ramanoudjame, G.; Du, M.; Mankiewicz, K. A.; Jayaraman, V. Allosteric Mechanism in AMPA Receptors: A FRET-Based Investigation of Conformational Changes. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103* (27), 10473–10478. https://doi.org/10.1073/pnas.0603225103.

(14) Kim, K. B.; Kefford, R.; Pavlick, A. C.; Infante, J. R.; Ribas, A.; Sosman, J. A.; Fecher, L. A.; Millward, M.; McArthur, G. A.; Hwu, P.; et al. Phase II Study of the MEK1/MEK2 Inhibitor Trametinib in Patients with Metastatic BRAF-Mutant Cutaneous Melanoma Previously Treated with or without a BRAF Inhibitor. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **2013**, *31* (4), 482–489. https://doi.org/10.1200/JCO.2012.43.5966.

(15) Roskoski, R. Allosteric MEK1/2 Inhibitors Including Cobimetanib and Trametinib in the Treatment of Cutaneous Melanomas. *Pharmacol. Res.* **2017**, *117*, 20–31. https://doi.org/10.1016/j.phrs.2016.12.009.

(16) Nadendla, R. R. Molecular Modeling: A Powerful Tool for Drug Design and Molecular Docking. *Resonance* **2004**, *9* (5), 51–60. https://doi.org/10.1007/BF02834015.

(17) Anderson, A. C. The Process of Structure-Based Drug Design. *Chem. Biol.* **2003**, *10* (9), 787–797. https://doi.org/10.1016/j.chembiol.2003.09.002.

(18) Syme, N. R.; Dennis, C.; Bronowska, A.; Paesen, G. C.; Homans, S. W. Comparison of Entropic Contributions to Binding in a “Hydrophilic” versus “Hydrophobic” Ligand−Protein Interaction. *J. Am. Chem. Soc.* **2010**, *132* (25), 8682–8689. https://doi.org/10.1021/ja101362u.

(19) EFPIA. The Pharmaceutical Industry in Figures. Key Data 2017 https://www.efpia.eu/media/219735/efpia-pharmafigures2017_statisticbroch_v04-final.pdf.

(20) Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E. W. Computational Methods in Drug Discovery. *Pharmacol. Rev.* **2014**, *66* (1), 334–395. https://doi.org/10.1124/pr.112.007336.

(21) Van Drie, J. H. Computer-Aided Drug Design: The next 20 Years. *J. Comput. Aided Mol. Des.* **2007**, *21* (10), 591–601. https://doi.org/10.1007/s10822-007-9142-y.

(22) Doman, T. N.; McGovern, S. L.; Witherbee, B. J.; Kasten, T. P.; Kurumbail, R.; Stallings, W. C.; Connolly, D. T.; Shoichet, B. K. Molecular Docking and High-Throughput Screening for Novel Inhibitors of Protein Tyrosine Phosphatase-1B. *J. Med. Chem.* **2002**, *45* (11), 2213–2221. https://doi.org/10.1021/jm010548w.

(23) Vijayakrishnan, R. Structure-Based Drug Design and Modern Medicine. *J. Postgrad. Med.* **2009**, *55* (4), 301. https://doi.org/10.4103/0022-3859.58943.

(24) Talele, T. T.; Khedkar, S. A.; Rigby, A. C. Successful Applications of Computer Aided Drug Discovery: Moving Drugs from Concept to the Clinic. *Curr. Top. Med. Chem.* **2010**, *10* (1), 127–141. https://doi.org/10.2174/156802610790232251.

(25) Hartman, G. D.; Egbertson, M. S.; Halczenko, W.; Laswell, W. L.; Duggan, M. E.; Smith, R. L.; Naylor, A. M.; Manno, P. D.; Lynch, R. J. Non-Peptide Fibrinogen Receptor Antagonists. 1. Discovery and Design of Exosite Inhibitors. *J. Med. Chem.* **1992**, *35* (24), 4640–4642. https://doi.org/10.1021/jm00102a020.

(26) Schames, J. R.; Henchman, R. H.; Siegel, J. S.; Sotriffer, C. A.; Ni, H.; McCammon, J. A. Discovery of a Novel Binding Trench in HIV Integrase. *J. Med. Chem.* **2004**, *47* (8), 1879–1881. https://doi.org/10.1021/jm0341913.

(27) Cohen, N. C. Medicine Pipeline: Structure-Based Drug Design and the Discovery of Aliskiren (Tekturna®): Perseverance and Creativity to Overcome a R&D Pipeline Challenge‡. *Chem. Biol. Drug Des.* **2007**, *70* (6), 557–565. https://doi.org/10.1111/j.1747-0285.2007.00599.x.

(28) Yu, W.; MacKerell, A. D. Computer-Aided Drug Design Methods. *Methods Mol. Biol. Clifton NJ* **2017**, *1520*, 85–106. https://doi.org/10.1007/978-1-4939-6634-9_5.

(29) Jones, C. G.; Martynowycz, M. W.; Hattne, J.; Fulton, T. J.; Stoltz, B. M.; Rodriguez, J. A.; Nelson, H. M.; Gonen, T. The CryoEM Method MicroED as a Powerful Tool for Small Molecule Structure Determination. *ACS Cent. Sci.* **2018**, *4* (11), 1587–1592. https://doi.org/10.1021/acscentsci.8b00760.

(30) de la Cruz, M. J.; Hattne, J.; Shi, D.; Seidler, P.; Rodriguez, J.; Reyes, F. E.; Sawaya, M. R.; Cascio, D.; Weiss, S. C.; Kim, S. K.; et al. Atomic-Resolution Structures from Fragmented Protein Crystals with the CryoEM Method MicroED. *Nat. Methods* **2017**, *14* (4), 399–402. https://doi.org/10.1038/nmeth.4178.

(31) Purdy, M. D.; Shi, D.; Chrustowicz, J.; Hattne, J.; Gonen, T.; Yeager, M. MicroED Structures of HIV-1 Gag CTD-SP1 Reveal Binding Interactions with the Maturation Inhibitor Bevirimat. *Proc. Natl. Acad. Sci.* **2018**, *115* (52), 13258–13263. https://doi.org/10.1073/pnas.1806806115.

(32) de la Cruz, M. J.; Martynowycz, M. W.; Hattne, J.; Gonen, T. MicroED Data Collection with SerialEM. *Ultramicroscopy* **2019**, *201*, 77–80. https://doi.org/10.1016/j.ultramic.2019.03.009.

(33) Allen, K. N.; Bellamacina, C. R.; Ding, X.; Jeffery, C. J.; Mattos, C.; Petsko, G. A.; Ringe, D. *An Experimental Approach to Mapping the Binding Surfaces of Crystalline Proteins †*; 1996.

(34) Bogan, A. A.; Thorn, K. S. Anatomy of Hot Spots in Protein Interfaces. *J. Mol. Biol.* **1998**, *280* (1), 1–9. https://doi.org/10.1006/jmbi.1998.1843.

(35) Shuker, S. B.; Hajduk, P. J.; Meadows, R. P.; Fesik, S. W. Discovering High-Affinity Ligands for Proteins: SAR by NMR. *Science* **1996**, *274* (5292), 1531–1534.

(36) Miranker, A.; Karplus, M. Functionality Maps of Binding Sites: A Multiple Copy Simultaneous Search Method. *Proteins Struct. Funct. Genet.* **1991**, *11* (1), 29–34. https://doi.org/10.1002/prot.340110104.

(37) Schubert, C. R.; Stultz, C. M. The Multi-Copy Simultaneous Search Methodology: A Fundamental Tool for Structure-Based Drug Design. *J. Comput. Aided Mol. Des.* **2009**, *23* (8), 475–489. https://doi.org/10.1007/s10822-009-9287-y.

(38) Mattos, C.; Bellamacina, C. R.; Peisach, E.; Pereira, A.; Vitkup, D.; Petsko, G. A.; Ringe, D. Multiple Solvent Crystal Structures: Probing Binding Sites, Plasticity and Hydration. *J. Mol. Biol.* **2006**, *357* (5), 1471–1482. https://doi.org/10.1016/J.JMB.2006.01.039.

(39) Hann, M. M.; Leach, A. R.; Harper, G. Molecular Complexity and Its Impact on the Probability of Finding Leads for Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (3), 856–864. https://doi.org/10.1021/ci000403i.

(40) Leach, A. R.; Hann, M. M. Molecular Complexity and Fragment-Based Drug Discovery: Ten Years On. *Curr. Opin. Chem. Biol.* **2011**, *15* (4), 489–496. https://doi.org/10.1016/j.cbpa.2011.05.008.

(41) Hunter, C. A. Quantifying Intermolecular Interactions: Guidelines for the Molecular Recognition Toolbox. *Angew. Chem. Int. Ed.* **2004**, *43* (40), 5310–5324. https://doi.org/10.1002/anie.200301739.

(42) Wood, D. J.; Lopez-Fernandez, J. D.; Knight, L. E.; Al-Khawaldeh, I.; Gai, C.; Lin, S.; Martin, M. P.; Miller, D. C.; Cano, C.; Endicott, J. A.; et al. FragLites—Minimal, Halogenated Fragments Displaying Pharmacophore Doublets. An Efficient Approach to Druggability Assessment and Hit Generation. *J. Med. Chem.* **2019**, acs.jmedchem.9b00304. https://doi.org/10.1021/acs.jmedchem.9b00304.

(43) Kozakov, D.; Grove, L. E.; Hall, D. R.; Bohnuud, T.; Mottarella, S. E.; Luo, L.; Xia, B.; Beglov, D.; Vajda, S. The FTMap Family of Web Servers for Determining and Characterizing Ligand-Binding Hot Spots of Proteins. *Nat. Protoc.* **2015**, *10* (5), 733–755. https://doi.org/10.1038/nprot.2015.043.

(44) Cimermancic, P.; Weinkam, P.; Rettenmaier, T. J.; Bichmann, L.; Keedy, D. A.; Woldeyes, R. A.; Schneidman-Duhovny, D.; Demerdash, O. N.; Mitchell, J. C.; Wells, J. A.; et al. CryptoSite: Expanding the Druggable Proteome by Characterization and Prediction of Cryptic Binding Sites. *J. Mol. Biol.* **2016**, *428* (4), 709–719. https://doi.org/10.1016/j.jmb.2016.01.029.

(45) Greener, J. G.; Sternberg, M. J. AlloPred: Prediction of Allosteric Pockets on Proteins Using Normal Mode Perturbation Analysis. *BMC Bioinformatics* **2015**, *16*. https://doi.org/10.1186/s12859-015-0771-1.

(46) Weinkam, P.; Pons, J.; Sali, A. Structure-Based Model of Allostery Predicts Coupling between Distant Sites. *Proc. Natl. Acad. Sci.* **2012**, *109* (13), 4875–4880. https://doi.org/10.1073/pnas.1116274109.

(47) Panjkovich, A.; Daura, X. PARS: A Web Server for the Prediction of Protein Allosteric and Regulatory Sites. *Bioinformatics* **2014**, *30* (9), 1314–1315. https://doi.org/10.1093/bioinformatics/btu002.

(48) Panjkovich, A.; Daura, X. Exploiting Protein Flexibility to Predict the Location of Allosteric Sites. *BMC Bioinformatics* **2012**, *13* (1), 273. https://doi.org/10.1186/1471-2105-13-273.

(49) Mitternacht, S.; Berezovsky, I. N. Binding Leverage as a Molecular Basis for Allosteric Regulation. *PLoS Comput. Biol.* **2011**, *7* (9), e1002148. https://doi.org/10.1371/journal.pcbi.1002148.

(50) Goncearenco, A.; Mitternacht, S.; Yong, T.; Eisenhaber, B.; Eisenhaber, F.; Berezovsky, I. N. SPACER: Server for Predicting Allosteric Communication and Effects of Regulation. *Nucleic Acids Res.* **2013**, *41* (Web Server issue), W266-272. https://doi.org/10.1093/nar/gkt460.

(51) Le Guilloux, V.; Schmidtke, P.; Tuffery, P. Fpocket: An Open Source Platform for Ligand Pocket Detection. *BMC Bioinformatics* **2009**, *10*, 168. https://doi.org/10.1186/1471-2105-10-168.

(52) Schmidtke, P.; Le Guilloux, V.; Maupetit, J.; Tufféry, P. Fpocket: Online Tools for Protein Ensemble Pocket Detection and Tracking. *Nucleic Acids Res.* **2010**, *38* (Web Server issue), W582-9. https://doi.org/10.1093/nar/gkq383.

(53) Brenke, R.; Kozakov, D.; Chuang, G.-Y.; Beglov, D.; Hall, D.; Landon, M. R.; Mattos, C.; Vajda, S. Fragment-Based Identification of Druggable "hot Spots" of Proteins Using Fourier Domain Correlation Techniques. *Bioinforma. Orig. Pap.* **2009**, *25* (5), 621–62710. https://doi.org/10.1093/bioinformatics/btp036.

(54) Tiwary, P.; van de Walle, A. A Review of Enhanced Sampling Approaches for Accelerated Molecular Dynamics. In *Multiscale Materials Modeling for Nanomechanics*; Weinberger, C. R., Tucker, G. J., Eds.; Springer Series in Materials Science; Springer International Publishing: Cham, 2016; pp 195–221. https://doi.org/10.1007/978-3-319-33480-6_6.

(55) Yang, Y. I.; Shao, Q.; Zhang, J.; Yang, L.; Gao, Y. Q. Enhanced Sampling in Molecular Dynamics. *J. Chem. Phys.* **2019**, *151* (7), 070902. https://doi.org/10.1063/1.5109531.

(56) Salmaso, V.; Moro, S. Bridging Molecular Docking to Molecular Dynamics in Exploring Ligand-Protein Recognition Process: An Overview. *Front. Pharmacol.* **2018**, *9*. https://doi.org/10.3389/fphar.2018.00923.

(57) De Vivo, M.; Masetti, M.; Bottegoni, G.; Cavalli, A. Role of Molecular Dynamics and Related Methods in Drug Discovery. *J. Med. Chem.* **2016**, *59* (9), 4035–4061. https://doi.org/10.1021/acs.jmedchem.5b01684.

(58) Sugita, Y.; Okamoto, Y. Replica-Exchange Molecular Dynamics Method for Protein Folding. *Chem. Phys. Lett.* **1999**, *314* (1–2), 141–151. https://doi.org/10.1016/S0009-2614(99)01123-9.

(59) Torrie, G. M.; Valleau, J. P. Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling. *J. Comput. Phys.* **1977**, *23* (2), 187–199. https://doi.org/10.1016/0021-9991(77)90121-8.

(60) Laio, A.; Parrinello, M. Escaping Free-Energy Minima. *Proc. Natl. Acad. Sci.* **2002**, *99* (20), 12562–12566. https://doi.org/10.1073/pnas.202427399.

(61) Ghanakota, P.; Carlson, H. A. Driving Structure-Based Drug Discovery through Cosolvent Molecular Dynamics. *J. Med. Chem.* **2016**, *59* (23), 10383–10399. https://doi.org/10.1021/acs.jmedchem.6b00399.

(62) Faller, C. E.; Raman, E. P.; MacKerell, A. D.; Guvench, O. Site Identification by Ligand Competitive Saturation (SILCS) Simulations for Fragment-Based Drug Design. In *Methods in molecular biology (Clifton, N.J.)*; 2015; Vol. 1289, pp 75–87. https://doi.org/10.1007/978-1-4939-2486-8_7.

(63) Cheng, H.; Linhares, B. M.; Yu, W.; Cardenas, M. G.; Ai, Y.; Jiang, W.; Winkler, A.; Cohen, S.; Melnick, A.; MacKerell, A.; et al. Identification of Thiourea-Based Inhibitors of the B-Cell Lymphoma 6 BTB Domain via NMR-Based Fragment Screening and Computer-Aided Drug Design. *J. Med. Chem.* **2018**, *61* (17), 7573–7588. https://doi.org/10.1021/acs.jmedchem.8b00040.

(64) Lanning, M. E.; Yu, W.; Yap, J. L.; Chauhan, J.; Chen, L.; Whiting, E.; Pidugu, L. S.; Atkinson, T.; Bailey, H.; Li, W.; et al. Structure-Based Design of N-Substituted 1-Hydroxy-4-Sulfamoyl-2-Naphthoates as Selective Inhibitors of the Mcl-1 Oncoprotein. *Eur. J. Med. Chem.* **2016**, *113*, 273–292. https://doi.org/10.1016/j.ejmech.2016.02.006.

(65) Alvarez-Garcia, D.; Barril, X. Molecular Simulations with Solvent Competition Quantify Water Displaceability and Provide Accurate Interaction Maps of Protein Binding Sites. *J. Med. Chem.* **2014**, *57* (20), 8530–8539. https://doi.org/10.1021/jm5010418.

(66) Seco, J.; Luque, F. J.; Barril, X. Binding Site Detection and Druggability Index from First Principles. https://doi.org/10.1021/jm801385d.

(67) Guvench, O.; MacKerell, A. D. Computational Fragment-Based Binding Site Identification by Ligand Competitive Saturation. *PLoS Comput. Biol.* **2009**, *5* (7), e1000435. https://doi.org/10.1371/journal.pcbi.1000435.

(68) Graham, S. E.; Leja, N.; Carlson, H. A. MixMD Probeview: Robust Binding Site Prediction from Cosolvent Simulations. *J Chem Inf Model* **2018**, *58*, 5. https://doi.org/10.1021/acs.jcim.8b00265.

(69) Ghanakota, P.; Carlson, H. A. Moving Beyond Active-Site Detection: MixMD Applied to Allosteric Systems. *J. Phys. Chem. B* **2016**, *120* (33), 8685–8695. https://doi.org/10.1021/acs.jpcb.6b03515.

(70) Oleinikovas, V.; Saladino, G.; Cossins, B. P.; Gervasio, F. L. Understanding Cryptic Pocket Formation in Protein Targets by Enhanced Sampling Simulations. *J. Am. Chem. Soc.* **2016**, *138* (43), 14257–14263. https://doi.org/10.1021/jacs.6b05425.

(71) Comitani, F.; Gervasio, F. L. Exploring Cryptic Pockets Formation in Targets of Pharmaceutical Interest with SWISH. *J. Chem. Theory Comput.* **2018**, *14* (6), 3321–3331. https://doi.org/10.1021/acs.jctc.8b00263.

(72) Masciocchi, D.; Gelain, A.; Villa, S.; Meneghetti, F.; Barlocco, D. Signal Transducer and Activator of Transcription 3 (STAT3): A Promising Target for Anticancer Therapy. *Future Med. Chem.* **2011**, *3* (5), 567–597. https://doi.org/10.4155/fmc.11.22.

(73) Yang, J.-L. STAT3 Inhibition, a Novel Approach to Enhancing Targeted Therapy in Human Cancers (Review). *Int. J. Oncol.* **2012**, 1181–1191. https://doi.org/10.3892/ijo.2012.1568.

(74) Wegrzyn, J.; Potla, R.; Chwae, Y.-J.; Sepuri, N. B. V.; Zhang, Q.; Koeck, T.; Derecka, M.; Szczepanek, K.; Szelag, M.; Gornicka, A.; et al. Function of Mitochondrial Stat3 in Cellular Respiration. *Science* **2009**, *323* (5915), 793–797. https://doi.org/10.1126/science.1164551.

(75) Zhang, X.; Sun, Y.; Pireddu, R.; Yang, H.; Urlam, M. K.; Lawrence, H. R.; Guida, W. C.; Lawrence, N. J.; Sebti, S. M. A Novel Inhibitor of STAT3 Homodimerization Selectively Suppresses STAT3 Activity and Malignant Transformation. *Cancer Res.* **2013**, *73* (6), 1922–1933. https://doi.org/10.1158/0008-5472.CAN-12-3175.

(76) Yu, H.; Kortylewski, M.; Pardoll, D. Crosstalk between Cancer and Immune Cells: Role of STAT3 in the Tumour Microenvironment. *Nat. Rev. Immunol.* **2007**, *7* (1), 41–51. https://doi.org/10.1038/nri1995.

(77) Lee, H.; Herrmann, A.; Deng, J.-H.; Kujawski, M.; Niu, G.; Li, Z.; Forman, S.; Jove, R.; Pardoll, D. M.; Yu, H. Persistently Activated Stat3 Maintains Constitutive NF-KappaB Activity in Tumors. *Cancer Cell* **2009**, *15* (4), 283–293. https://doi.org/10.1016/j.ccr.2009.02.015.

(78) Grivennikov, S.; Karin, E.; Terzic, J.; Mucida, D.; Yu, G.-Y.; Vallabhapurapu, S.; Scheller, J.; Rose-John, S.; Cheroutre, H.; Eckmann, L.; et al. IL-6 and Stat3 Are Required for Survival of Intestinal Epithelial Cells and Development of Colitis-Associated Cancer. *Cancer Cell* **2009**, *15* (2), 103–113. https://doi.org/10.1016/j.ccr.2009.01.001.

(79) Ehret, G. B.; Reichenbach, P.; Schindler, U.; Horvath, C. M.; Fritz, S.; Nabholz, M.; Bucher, P. Ehret et al.: DNA Binding Specificities of STAT1, STAT5 and STAT6 -1- DNA Binding Specificity of Different STAT Proteins: Comparison of in Vitro Specificity with Natural Target Sites Running Title: DNA Binding Specificities of STAT1, STAT5 and STAT6. *JBC Pap. Press Publ. Oct.* **2000**, *26*.

(80) Levy, D. E.; Darnell, J. E. Stats: Transcriptional Control and Biological Impact. *Nat. Rev. Mol. Cell Biol.* **2002**, *3* (9), 651–662. https://doi.org/10.1038/nrm909.

(81) Germain, D.; Frank, D. A. Targeting the Cytoplasmic and Nuclear Functions of Signal Transducers and Activators of Transcription 3 for Cancer Therapy. *Clin. Cancer Res.* **2007**, *13* (19), 5665–5669. https://doi.org/10.1158/1078-0432.CCR-06-2491.

(82) John, S.; Vinkemeier, U.; Soldaini, E.; Darnell, J. E.; Leonard, W. J. The Significance of Tetramerization in Promoter Recruitment by Stat5. *Mol. Cell. Biol.* **1999**, *19* (3), 1910–1918.

(83) Horvath, C. M. STAT Proteins and Transcriptional Responses to Extracellular Signals. *Trends Biochem. Sci.* **2000**, *25* (10), 496–502.

(84) Shuai, K. The STAT Family of Proteins in Cytokine Signaling. *Prog. Biophys. Mol. Biol.* **1999**, *71* (3–4), 405–422.

(85) Abell, K.; Watson, C. J. The Jak/Stat Pathway: A Novel Way to Regulate PI3K Activity. *Cell Cycle Georget. Tex* **2005**, *4* (7), 897–900. https://doi.org/10.4161/cc.4.7.1837.

(86) Nkansah, E.; Shah, R.; Collie, G. W.; Parkinson, G. N.; Palmer, J.; Rahman, K. M.; Bui, T. T.; Drake, A. F.; Husby, J.; Neidle, S.; et al. Observation of Unphosphorylated STAT3 Core Protein Binding to Target DsDNA by PEMSA and X-Ray Crystallography. *FEBS Lett.* **2013**, *587* (7), 833–839. https://doi.org/10.1016/j.febslet.2013.01.065.

(87) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

(88) Becker, S.; Groner, B.; Muller, C. W. Three-Dimensional Structure of the Stat3beta Homodimer Bound to DNA. *Nature* **1998**, *394*, 145–151.

(89) Ren, Z.; Mao, X.; Mertens, C.; Krishnaraj, R.; Qin, J.; Mandal, P. K.; Romanowski, M. J.; McMurray, J. S.; Chen, X. Crystal Structure of Unphosphorylated STAT3 Core Fragment. *Biochem.Biophys.Res.Commun.* **2008**, *374*, 1–5.

(90) Hu, T.; Yeh, J. E.; Pinello, L.; Jacob, J.; Chakravarthy, S.; Yuan, G. C.; Chopra, R.; Frank, D. A. Impact of the N-Terminal Domain of STAT3 in STAT3-Dependent Transcriptional Activity. *Mol.Cell.Biol.* **2015**, *35*, 3284–3300.

(91) Belo, Y.; Mielko, Z.; Nudelman, H.; Afek, A.; Ben-David, O.; Shahar, A.; Zarivach, R.; Gordan, R.; Arbely, E. Unexpected Implications of STAT3 Acetylation Revealed by Genetic Encoding of Acetyl-Lysine. *Biochim. Biophys. Acta BBA - Gen. Subj.* **2019**, *1863* (9), 1343–1350. https://doi.org/10.1016/j.bbagen.2019.05.019.

(92) Shahani, V. M.; Yue, P.; Haftchenary, S.; Zhao, W.; Lukkarila, J. L.; Zhang, X.; Ball, D.; Nona, C.; Gunning, P. T.; Turkson, J. Identification of Purine-Scaffold Small-Molecule Inhibitors of Stat3 Activation by QSAR Studies. *ACS Med. Chem. Lett.* **2011**, *2* (1), 79–84. https://doi.org/10.1021/ml100224d.

(93) Siddiquee, K.; Zhang, S.; Guida, W. C.; Blaskovich, M. a; Greedy, B.; Lawrence, H. R.; Yip, M. L. R.; Jove, R.; McLaughlin, M. M.; Lawrence, N. J.; et al. Selective Chemical Probe Inhibitor of Stat3, Identified through Structure-Based Virtual Screening, Induces Antitumor Activity. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104* (18), 7391–7396. https://doi.org/10.1073/pnas.0609757104.

(94) Song, H.; Wang, R.; Wang, S.; Lin, J. A Low-Molecular-Weight Compound Discovered through Virtual Database Screening Inhibits Stat3 Function in Breast Cancer Cells. *Proc Natl Acad Sci U A* **2005**, *102* (13), 4700–4705. https://doi.org/10.1073/pnas.0409894102.

(95) Matsuno, K.; Masuda, Y.; Uehara, Y.; Sato, H.; Muroya, A.; Takahashi, O.; Yokotagawa, T.; Furuya, T.; Okawara, T.; Otsuka, M.; et al. Identification of a New Series of STAT3 Inhibitors by Virtual Screening. *ACS Med. Chem. Lett.* **2010**, *1* (8), 371–375. https://doi.org/10.1021/ml1000273.

(96) Xiong, A.; Yang, Z.; Shen, Y.; Zhou, J.; Shen, Q. Transcription Factor STAT3 as a Novel Molecular Target for Cancer Prevention. *Cancers* **2014**, *6* (2), 926–957. https://doi.org/10.3390/cancers6020926.

(97) Fletcher, S.; Page, B. D. G.; Zhang, X.; Yue, P.; Li, Z. H.; Sharmeen, S.; Singh, J.; Zhao, W.; Schimmer, A. D.; Trudel, S.; et al. Antagonism of the Stat3-Stat3 Protein Dimer with Salicylic Acid Based Small Molecules. *ChemMedChem* **2011**, *6* (8), 1459–1470. https://doi.org/10.1002/cmdc.201100194.

(98) McMurray, J. S. A New Small-Molecule Stat3 Inhibitor. *Chem. Biol.* **2006**, *13* (11), 1123–1124. https://doi.org/10.1016/j.chembiol.2006.11.001.

(99) Zhang, X.; Yue, P.; Page, B. D. G.; Li, T.; Zhao, W.; Namanja, A. T.; Paladino, D.; Zhao, J.; Chen, Y.; Gunning, P. T.; et al. Orally Bioavailable Small-Molecule Inhibitor of Transcription Factor Stat3 Regresses Human Breast and Lung Cancer Xenografts. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (24), 9623–9628. https://doi.org/10.1073/pnas.1121606109.

(100) Park, I.-H.; Li, C. Characterization of Molecular Recognition of STAT3 SH2 Domain Inhibitors through Molecular Simulation. *J. Mol. Recognit. JMR* **2011**, *24* (2), 254–265. https://doi.org/10.1002/jmr.1047.

(101) Poli, G.; Gelain, A.; Porta, F.; Asai, A.; Martinelli, A.; Tuccinardi, T. Identification of a New STAT3 Dimerization Inhibitor through a Pharmacophore-Based Virtual Screening Approach. **2015**, *6366* (MD), 1–7. https://doi.org/10.3109/14756366.2015.1079184.

(102) Chiao, J. W.; Melikian, M.; Han, L.; Xue, C.; Tsao, A.; Wang, L.; Mencher, S. K.; Fallon, J.; Solangi, K.; Bertho, G.; et al. Interaction of a Small Molecule Natura-α and STAT3-SH2 Domain to Block Y705 Phosphorylation and

Inhibit Lupus Nephritis. *Biochem. Pharmacol.* **2016**, *99*, 123–131. https://doi.org/10.1016/j.bcp.2015.11.018.

(103) Borg, C.; Rognan, D.; Boibessot, T. European Journal of Medicinal Chemistry Novel Aminotetrazole Derivatives as Selective STAT3 Non-Peptide Inhibitors Jean-Ren e. **2015**, *103*, 163–174.

(104) Pallandre, J.-R.; Borg, C.; Rognan, D.; Boibessot, T.; Luzet, V.; Yesylevskyy, S.; Ramseyer, C.; Pudlo, M. Novel Aminotetrazole Derivatives as Selective STAT3 Non-Peptide Inhibitors. *Eur. J. Med. Chem.* **2015**, *103*, 163–174. https://doi.org/10.1016/J.EJMECH.2015.08.054.

(105) Turkson, J.; Ryan, D.; Kim, J. S.; Zhang, Y.; Chen, Z.; Haura, E.; Laudano, A.; Sebti, S.; Hamilton, A. D.; Jove, R. Phosphotyrosyl Peptides Block Stat3-Mediated DNA Binding Activity, Gene Regulation, and Cell Transformation. *J. Biol. Chem.* **2001**, *276* (48), 45443–45455. https://doi.org/10.1074/jbc.M107527200.

(106) Debnath, B.; Xu, S.; Neamati, N. Small Molecule Inhibitors of Signal Transducer and Activator of Transcription 3 (Stat3) Protein. *J. Med. Chem.* **2012**, *55* (15), 6645–6668. https://doi.org/10.1021/jm300207s.

(107) Shin, D.-S.; Kim, H.-N.; Shin, K. D.; Yoon, Y. J.; Kim, S.-J.; Han, D. C.; Kwon, B.-M. Cryptotanshinone Inhibits Constitutive Signal Transducer and Activator of Transcription 3 Function through Blocking the Dimerization in DU145 Prostate Cancer Cells. *Cancer Res.* **2009**, *69* (1), 193–202. https://doi.org/10.1158/0008-5472.CAN-08-2575.

(108) Chen, L.; Chen, L.; Wang, H.-J.; Wang, H.-J.; Xie, W.; Xie, W.; Yao, Y.; Yao, Y.; Zhang, Y.-S.; Zhang, Y.-S.; et al. Cryptotanshinone Inhibits Lung Tumorigenesis and Induces Apoptosis in Cancer Cells in Vitro and in Vivo. *Mol. Med. Rep.* **2014**, *9* (6), 2447–2452. https://doi.org/10.3892/mmr.2014.2093.

(109) Nelson, K. M.; Dahlin, J. L.; Bisson, J.; Graham, J.; Pauli, G. F.; Walters, M. A. The Essential Medicinal Chemistry of Curcumin. *J. Med. Chem.* **2017**, *60* (5), 1620–1637. https://doi.org/10.1021/acs.jmedchem.6b00975.

(110) Chen, J.; Bai, L.; Bernard, D.; Nikolovska-Coleska, Z.; Gomez, C.; Zhang, J.; Yi, H.; Wang, S. Structure-Based Design of Conformationally Constrained, Cell-Permeable STAT3 Inhibitors. *ACS Med. Chem. Lett.* **2010**, *1* (2), 85–89. https://doi.org/10.1021/ml100010j.

(111) Zhao, W.; Jaganathan, S.; Turkson, J. A Cell-Permeable Stat3 SH2 Domain Mimetic Inhibits Stat3 Activation and Induces Antitumor Cell Effects in Vitro. *J. Biol. Chem.* **2010**, *285* (46), 35855–35865. https://doi.org/10.1074/jbc.M110.154088.

(112) Siddiquee, K. A. Z.; Gunning, P. T.; Glenn, M.; Katt, W. P.; Zhang, S.; Schrock, C.; Schroeck, C.; Sebti, S. M.; Jove, R.; Hamilton, A. D.; et al. An Oxazole-Based Small-Molecule Stat3 Inhibitor Modulates Stat3 Stability and Processing and Induces Antitumor Cell Effects. *ACS Chem. Biol.* **2007**, *2* (12), 787–798. https://doi.org/10.1021/cb7001973.

(113) Lin, L.; Deangelis, S.; Foust, E.; Fuchs, J.; Li, C.; Li, P.-K.; Schwartz, E. B.; Lesinski, G. B.; Benson, D.; Lü, J.; et al. A Novel Small Molecule Inhibits STAT3 Phosphorylation and DNA Binding Activity and Exhibits Potent Growth Suppressive Activity in Human Cancer Cells. **2010**. https://doi.org/10.1186/1476-4598-9-217.

(114) Schust, J.; Sperl, B.; Hollis, A.; Mayer, T.; Berg, T. Stattic: A Small-Molecule Inhibitor of STAT3 Activation and Dimerization. *Chem. Biol.* **2006**.

(115) Reed, S.; Li, H.; Li, C.; Lin, J. Celecoxib Inhibits STAT3 Phosphorylation and Suppresses Cell Migration and Colony Forming Ability in Rhabdomyosarcoma Cells. *Biochem. Biophys. Res. Commun.* **2011**, *407* (3), 450–455. https://doi.org/10.1016/j.bbrc.2011.03.014.

(116) Li, R.; Hu, Z.; Sun, S.-Y.; Chen, Z. G.; Owonikoko, T. K.; Sica, G. L.; Ramalingam, S. S.; Curran, W. J.; Khuri, F. R.; Deng, X. Niclosamide Overcomes Acquired Resistance to Erlotinib through Suppression of STAT3 in Non-Small Cell Lung Cancer. *Mol. Cancer Ther.* **2013**, *12* (10), 2200–2212. https://doi.org/10.1158/1535-7163.MCT-13-0095.

(117) Lin, L.; Hutzen, B.; Li, P.-K.; Ball, S.; Zuo, M.; Deangelis, S.; Foust, E.; Sobo, M.; Friedman, L.; Bhasin, D.; et al. A Novel Small Molecule, LLL12, Inhibits STAT3 Phosphorylation and Activities and Exhibits Potent Growth-Suppressive Activity in Human Cancer Cells 1,2. https://doi.org/10.1593/neo.91196.

(118) Fuh, B.; Sobo, M.; Cen, L.; Josiah, D.; Hutzen, B.; Cisek, K.; Bhasin, D.; Regan, N.; Lin, L.; Chan, C.; et al. LLL-3 Inhibits STAT3 Activity, Suppresses Glioblastoma Cell Growth and Prolongs Survival in a Mouse Glioblastoma Model. *Br. J. Cancer* **2009**, *100*, 106–112. https://doi.org/10.1038/sj.bjc.6604793.

(119) Nan, J.; Du, Y.; Chen, X.; Bai, Q.; Wang, Y.; Zhang, X.; Zhu, N.; Zhang, J.; Hou, J.; Wang, Q.; et al. TPCA-1 Is a Direct Dual Inhibitor of STAT3 and NF-KB and Regresses Mutant EGFR-Associated Human Non–Small Cell Lung Cancers. https://doi.org/10.1158/1535-7163.MCT-13-0464.

(120) Li, Y.; Rogoff, H. A.; Keates, S.; Gao, Y.; Murikipudi, S.; Mikule, K.; Leggett, D.; Li, W.; Pardee, A. B.; Li, C. J. Suppression of Cancer Relapse and Metastasis by Inhibiting Cancer Stemness. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112* (6), 1839–1844. https://doi.org/10.1073/pnas.1424171112.

(121) Brambilla, L.; Genini, D.; Laurini, E.; Merulla, J.; Perez, L.; Fermeglia, M.; Carbone, G. M.; Pricl, S.; Catapano, C. V. Hitting the Right Spot: Mechanism of Action of OPB-31121, a Novel and Potent Inhibitor of the Signal Transducer and Activator of Transcription 3 (STAT3). *Mol. Oncol.* **2015**, *9* (6), 1194–1206. https://doi.org/10.1016/j.molonc.2015.02.012.

(122) Wong, A. L. A.; Hirpara, J. L.; Pervaiz, S.; Eu, J.-Q.; Sethi, G.; Goh, B.-C. Do STAT3 Inhibitors Have Potential in the Future for Cancer Therapy? *Expert Opin. Investig. Drugs* **2017**, *26* (8), 883–887. https://doi.org/10.1080/13543784.2017.1351941.

(123) Wong, A. L.; Soo, R. A.; Tan, D. S.; Lee, S. C.; Lim, J. S.; Marban, P. C.; Kong, L. R.; Lee, Y. J.; Wang, L. Z.; Thuya, W. L.; et al. Phase I and Biomarker Study of OPB-51602, a Novel Signal Transducer and Activator of Transcription (STAT) 3 Inhibitor, in Patients with Refractory Solid Malignancies. *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.* **2015**, *26* (5), 998–1005. https://doi.org/10.1093/annonc/mdv026.

(124) Myers, M. G. Cell Biology. Moonlighting in Mitochondria. *Science* **2009**, *323* (5915), 723–724. https://doi.org/10.1126/science.1169660.

(125) Hubbard, J. M.; Grothey, A. Napabucasin: An Update on the First-in-Class Cancer Stemness Inhibitor. *Drugs* **2017**, *77* (10), 1091–1103. https://doi.org/10.1007/s40265-017-0759-4.

(126) Jonker, D. J.; Nott, L.; Yoshino, T.; Gill, S.; Shapiro, J.; Ohtsu, A.; Zalcberg, J.; Vickers, M. M.; Wei, A. C.; Gao, Y.; et al. Napabucasin versus Placebo in Refractory Advanced Colorectal Cancer: A Randomised Phase 3 Trial. *Lancet Gastroenterol. Hepatol.* **2018**, *3* (4), 263–270. https://doi.org/10.1016/S2468-1253(18)30009-8.

(127) Li, Chiang, J.; Yang, A.; Rogof, H. METHOD OF TARGETING STAT3 AND OTHER NON-DRUGGABLE PROTEINS. WO/2017/023866, February 10, 2017.

(128) Huang, W.; Dong, Z.; Wang, F.; Peng, H.; Liu, J.-Y.; Zhang, J.-T. A Small Molecule Compound Targeting STAT3 DNA-Binding Domain Inhibits Cancer Cell Proliferation, Migration, and Invasion. *ACS Chem. Biol.* **2014**, *9* (5), 1188–1196. https://doi.org/10.1021/cb500071v.

(129) Huang, W.; Dong, Z.; Chen, Y.; Wang, F.; Wang, C. J.; Peng, H.; He, Y.; Hangoc, G.; Pollok, K.; Sandusky, G.; et al. Small-Molecule Inhibitors Targeting the DNA-Binding Domain of STAT3 Suppress Tumor Growth, Metastasis and STAT3 Target Gene Expression in Vivo. *Oncogene* **2016**, *35* (6), 783–792. https://doi.org/10.1038/onc.2015.215.

(130) Caboni, L.; Lloyd, D. G. Beyond the Ligand-Binding Pocket: Targeting Alternate Sites in Nuclear Receptors. *Med. Res. Rev.* **2013**, *33* (5), 1081–1118. https://doi.org/10.1002/med.21275.

(131) Shih, P.-C.; Yang, Y.; Parkinson, G. N.; Wilderspin, A.; Wells, G.; Shih, P.-C.; Yang, Y.; Parkinson, G. N.; Wilderspin, A.; Wells, G.; et al. A High-Throughput Fluorescence Polarization Assay for Discovering Inhibitors Targeting the DNA-Binding Domain of Signal Transducer and Activator of Transcription 3 (STAT3). *Oncotarget* **2018**, *9* (66), 32690–32701. https://doi.org/10.18632/oncotarget.26013.

(132) Mertens, C.; Haripal, B.; Klinge, S.; Darnell, J. E. Mutations in the Linker Domain Affect Phospho-STAT3 Function and Suggest Targets for Interrupting STAT3 Activity. *Proc. Natl. Acad. Sci.* **2015**, *112* (48). https://doi.org/10.1073/pnas.1515876112.

(133) Namanja, A. T.; Wang, J.; Buettner, R.; Colson, L.; Chen, Y. Allosteric Communication across STAT3 Domains Associated with STAT3 Function and Disease-Causing Mutation. **2016**. https://doi.org/10.1016/j.jmb.2016.01.003.

(134) Leach, A. G. *Molecular Modeling, Principles and Applications*; Pearson Education: Edinburgh, 2001.

(135) Jones, J. E. *On the Determination of Molecular Fields. II. From the Equation of State of a Gas*; Proc. R. Soc. London, Ser. A, 1924.

(136) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117* (19), 5179–5197. https://doi.org/10.1021/ja00124a002.

(137) Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S.; Weiner, P. A New Force Field for Molecular Mechanical Simulation of Nucleic Acids and Proteins. *J. Am. Chem. Soc.* **1984**, *106* (3), 765–784. https://doi.org/10.1021/ja00315a051.

(138) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; et al. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins †. *J. Phys. Chem. B* **1998**, *102* (18), 3586–3616. https://doi.org/10.1021/jp973084f.

(139) Schuler, L. D.; Daura, X.; van Gunsteren, W. F. An Improved GROMOS96 Force Field for Aliphatic Hydrocarbons in the Condensed Phase. *J. Comput. Chem.* **2001**, *22* (11), 1205–1218. https://doi.org/10.1002/jcc.1078.

(140) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, *118* (45), 11225–11236. https://doi.org/10.1021/ja9621760.

(141) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved Side-Chain Torsion Potentials for the Amber Ff99SB Protein Force Field. *Proteins Struct. Funct. Bioinforma.* **2010**, NA–NA. https://doi.org/10.1002/prot.22711.

(142) Durrant, J. D.; McCammon, J. A.; Feynman, R.; Fischer, E.; Teague, S.; Ma, B.; Kumar, S.; Tsai, C.; Nussinov, R.; Kumar, S.; et al. Molecular Dynamics Simulations and Drug Discovery. *BMC Biol.* **2011**, *9* (1), 71. https://doi.org/10.1186/1741-7007-9-71.

(143) Ghanakota, P.; Carlson, H. A. Driving Structure-Based Drug Discovery through Cosolvent Molecular Dynamics. *J. Med. Chem.* **2016**, *59* (23), 10383–10399. https://doi.org/10.1021/acs.jmedchem.6b00399.

(144) Tuckerman, M.; Berne, B. J.; Martyna, G. J. Reversible Multiple Time Scale Molecular Dynamics. *J. Chem. Phys.* **1992**, *97* (3), 1990–2001. https://doi.org/10.1063/1.463137.

(145) Arrhenius, S. Über Die Reaktionsgeschwindigkeit Bei Der Inversion von Rohrzucker Durch Säuren. *Z. Für Phys. Chem.* **1889**. https://doi.org/10.1515/zpch-1889-0416.

(146) Eyring, H. The Activated Complex in Chemical Reactions. *J. Chem. Phys.* **1935**, *3* (2), 107–115. https://doi.org/10.1063/1.1749604.

(147) Evans, M. G.; Polanyi, M. Some Applications of the Transition State Method to the Calculation of Reaction Velocities, Especially in Solution. *Trans. Faraday Soc.* **1935**, *31* (0), 875–894. https://doi.org/10.1039/TF9353100875.

(148) Kubelka, J.; Hofrichter, J.; Eaton, W. A. The Protein Folding 'Speed Limit.' *Curr. Opin. Struct. Biol.* **2004**, *14* (1), 76–88. https://doi.org/10.1016/j.sbi.2004.01.013.

(149) Kästner, J. Umbrella Sampling. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1* (6), 932–942. https://doi.org/10.1002/wcms.66.

(150) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. THE Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules. I. The Method. *J. Comput. Chem.* **1992**, *13* (8), 1011–1021. https://doi.org/10.1002/jcc.540130812.

(151) Souaille, M.; Roux, B. Extension to the Weighted Histogram Analysis Method: Combining Umbrella Sampling with Free Energy Calculations. *Comput. Phys. Commun.* **2001**, *135* (1), 40–57. https://doi.org/10.1016/S0010-4655(00)00215-0.

(152) Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; Gunsteren, W. F. van; Mark, A. E. Peptide Folding: When Simulation Meets Experiment. *Angew. Chem. Int. Ed.* **1999**, *38* (1–2), 236–240. https://doi.org/10.1002/(SICI)1521-3773(19990115)38:1/2<236::AID-ANIE236>3.0.CO;2-M.

(153) Eisenhaber, F.; Lijnzaad, P.; Argos, P.; Sander, C.; Scharf, M. The double cubic lattice method: Efficient approaches to numerical integration of surface area and volume and to dot surface contouring of molecular assemblies. *J. Comput. Chem.* **1995**, *16* (3), 273–284. https://doi.org/10.1002/jcc.540160303.

(154) Rajamani, R.; Good, A. C. Ranking Poses in Structure-Based Lead Discovery and Optimization: Current Trends in Scoring Function Development. *Curr. Opin. Drug Discov. Devel.* **2007**, *10* (3), 308–315.

(155) Seifert, M. H. J.; Kraus, J.; Kramer, B. Virtual High-Throughput Screening of Molecular Databases. *Curr. Opin. Drug Discov. Devel.* **2007**, *10* (3), 298–307.

(156) Shoichet, B. K.; McGovern, S. L.; Wei, B.; Irwin, J. J. Lead Discovery Using Molecular Docking. *Curr. Opin. Chem. Biol.* **2002**, *6* (4), 439–446.

(157) Niu Huang; Chakrapani Kalyanaraman; John J. Irwin, and; Jacobson*, M. P. Physics-Based Scoring of Protein−Ligand Complexes: Enrichment of Known Inhibitors in Large-Scale Virtual Screening. **2005**.

(158) Böhm, H. J. The Development of a Simple Empirical Scoring Function to Estimate the Binding Constant for a Protein-Ligand Complex of Known Three-Dimensional Structure. *J. Comput. Aided Mol. Des.* **1994**, *8* (3), 243–256.

(159) Tanaka, S.; Scheraga, H. A. Medium- and Long-Range Interaction Parameters between Amino Acids for Predicting Three-Dimensional Structures of Proteins. *Macromolecules 9* (6), 945–950.

(160) Raha, K.; Merz, K. M. Large-Scale Validation of a Quantum Mechanics Based Scoring Function: Predicting the Binding Affinity and the Binding Mode of a Diverse Set of Protein−Ligand Complexes. *J. Med. Chem.* **2005**, *48* (14), 4558–4575. https://doi.org/10.1021/jm048973n.

(161) Pecina, A.; Meier, R.; Fanfrlík, J.; Lepšík, M.; Řezáč, J.; Hobza, P.; Baldauf, C. The SQM/COSMO Filter: Reliable Native Pose Identification Based on the Quantum-Mechanical Description of Protein–Ligand Interactions and

Implicit COSMO Solvation. *Chem. Commun.* **2016**, *52* (16), 3312–3315. https://doi.org/10.1039/C5CC09499B.

(162) Wang, Z.; Sun, H.; Yao, X.; Li, D.; Xu, L.; Li, Y.; Tian, S.; Hou, T. Comprehensive Evaluation of Ten Docking Programs on a Diverse Set of Protein–Ligand Complexes: The Prediction Accuracy of Sampling Power and Scoring Power. *Phys. Chem. Chem. Phys.* **2016**, *18* (18), 12964–12975. https://doi.org/10.1039/C6CP01555G.

(163) Corbeil, C. R.; Williams, C. I.; Labute, P. Variability in Docking Success Rates Due to Dataset Preparation. *J. Comput. Aided Mol. Des.* **2012**, *26* (6), 775–786. https://doi.org/10.1007/s10822-012-9570-1.

(164) Allen, W. J.; Balius, T. E.; Mukherjee, S.; Brozell, S. R.; Moustakas, D. T.; Lang, P. T.; Case, D. A.; Kuntz, I. D.; Rizzo, R. C. DOCK 6: Impact of New Features and Current Docking Performance. *J. Comput. Chem.* **2015**, *36* (15), 1132–1156. https://doi.org/10.1002/jcc.23905.

(165) Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J. Comput. Chem.* **2009**, *30* (16), 2785–2791. https://doi.org/10.1002/jcc.21256.

(166) Halgren, T. A. Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization, and Performance of MMFF94. *J. Comput. Chem.* **1996**, *17* (5–6), 490–519. https://doi.org/10.1002/(SICI)1096-987X(199604)17:5/6<490::AID-JCC1>3.0.CO;2-P.

(167) Halgren, T. A. Merck Molecular Force Field. II. MMFF94 van Der Waals and Electrostatic Parameters for Intermolecular Interactions. *J. Comput. Chem.* **1996**, *17* (5–6), 520–552. https://doi.org/10.1002/(SICI)1096-987X(199604)17:5/6<520::AID-JCC2>3.0.CO;2-W.

(168) Halgren, T. A. Merck Molecular Force Field. III. Molecular Geometries and Vibrational Frequencies for MMFF94. *J. Comput. Chem.* **1996**, *17* (5–6), 553–586. https://doi.org/10.1002/(SICI)1096-987X(199604)17:5/6<553::AID-JCC3>3.0.CO;2-T.

(169) Halgren, T. A.; Nachbar, R. B. Merck Molecular Force Field. IV. Conformational Energies and Geometries for MMFF94. *J. Comput. Chem.* **1996**, *17* (5–6), 587–615. https://doi.org/10.1002/(SICI)1096-987X(199604)17:5/6<587::AID-JCC4>3.0.CO;2-Q.

(170) Halgren, T. A. Merck Molecular Force Field. V. Extension of MMFF94 Using Experimental Data, Additional Computational Data, and Empirical Rules. *J. Comput. Chem.* **1996**, *17* (5–6), 616–641. https://doi.org/10.1002/(SICI)1096-987X(199604)17:5/6<616::AID-JCC5>3.0.CO;2-X.

(171) Edelsbrunner, H. Weighted Alpha Shapes. *Tech. Pap. Dep. Comput. Sci. Univ. Ill. Urbana-Champaign Urbana Ill.*

(172) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated Docking Using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function. *J. Comput. Chem. 19* (14), 24.

(173) Chemical Computing Group. *Molecular Operating Environment (MOE)*; 2016.

(174) Labute, P. The Generalized Born/Volume Integral Implicit Solvent Model: Estimation of the Free Energy of Hydration Using London Dispersion Instead of Atomic Surface Area. *J. Comput. Chem.* **2008**, *29* (10), 1693–1698. https://doi.org/10.1002/jcc.20933.

(175) Meng, E. C.; Shoichet, B. K.; Kuntz, I. D. Automated Docking with Grid-Based Energy Evaluation. *J. Comput. Chem.* **1992**, *13* (4), 505–524. https://doi.org/10.1002/jcc.540130412.

(176) MOE Tutorial.

(177) Clark, A. Ligand Interaction Diagram https://www.chemcomp.com/journal/ligintdia.htm (accessed May 28, 2016).

(178) Mysinger, M. M.; Carchia, M.; Irwin, John. J.; Shoichet, B. K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* **2012**, *55* (14), 6582–6594. https://doi.org/10.1021/jm300687e.

(179) AMBER Home Page http://ambermd.org/ (accessed May 31, 2016).

(180) Kumari, R.; Kumar, R.; Lynn, A.; Lynn, A. *G_mmpbsa* —A GROMACS Tool for High-Throughput MM-PBSA Calculations. *J. Chem. Inf. Model.* **2014**, *54* (7), 1951–1962. https://doi.org/10.1021/ci500020m.

(181) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera?A Visualization System for Exploratory Research and Analysis. *J. Comput. Chem.* **2004**, *25* (13), 1605–1612. https://doi.org/10.1002/jcc.20084.

(182) Humphrey, W.; Dalke, A.; Schulten, K. {VMD} -- {V}isual {M}olecular {D}ynamics. *J. Mol. Graph.* **1996**, *14*, 33–38.

(183) Ferrari, A. M.; Wei, B. Q.; Costantino, L.; Shoichet, B. K. Soft Docking and Multiple Receptor Conformations in Virtual Screening. *J. Med. Chem.* **2004**, *47*, 5076–5084. https://doi.org/10.1021/jm049756p.

(184) Ferrari, A. M.; Wei, B. Q.; Costantino, L.; Shoichet, B. K. Soft Docking and Multiple Receptor Conformations in Virtual Screening. *J. Med. Chem.* **2004**, *47* (21), 5076–5084. https://doi.org/10.1021/jm049756p.

(185) Sayyed-Ahmad, A.; Gorfe, A. A. Mixed-Probe Simulation and Probe-Derived Surface Topography Map Analysis for Ligand Binding Site Identification. *J. Chem. Theory Comput.* **2017**, *13* (4), 1851–1861. https://doi.org/10.1021/acs.jctc.7b00130.

(186) Kimura, S. R.; Hu, H. P.; Ruvinsky, A. M.; Sherman, W.; Favia, A. D. Deciphering Cryptic Binding Sites on Proteins by Mixed-Solvent Molecular Dynamics. *J. Chem. Inf. Model.* **2017**, *57* (6), 1388–1401. https://doi.org/10.1021/acs.jcim.6b00623.

(187) Estébanez-Perpiñá, E.; Arnold, L. A.; Arnold, A. A.; Nguyen, P.; Rodrigues, E. D.; Mar, E.; Bateman, R.; Pallai, P.; Shokat, K. M.; Baxter, J. D.; et al. A Surface on the Androgen Receptor That Allosterically Regulates Coactivator Binding. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104* (41), 16074–16079. https://doi.org/10.1073/pnas.0708036104.

(188) Wiesmann, C.; Barr, K. J.; Kung, J.; Zhu, J.; Erlanson, D. A.; Shen, W.; Fahr, B. J.; Zhong, M.; Taylor, L.; Randall, M.; et al. Allosteric Inhibition of Protein Tyrosine Phosphatase 1B. *Nat. Struct. Mol. Biol.* **2004**, *11* (8), 730–737. https://doi.org/10.1038/nsmb803.

(189) Keedy, D. A.; Hill, Z. B.; Biel, J. T.; Kang, E.; Rettenmaier, T. J.; Brandão-Neto, J.; Pearce, N. M.; von Delft, F.; Wells, J. A.; Fraser, J. S. An Expanded Allosteric Network in PTP1B by Multitemperature Crystallography, Fragment Screening, and Covalent Tethering. *eLife* **2018**, *7*. https://doi.org/10.7554/eLife.36307.

(190) Vanderschueren, D.; Laurent, M. R.; Claessens, F.; Gielen, E.; Lagerquist, M. K.; Vandenput, L.; Börjesson, A. E.; Ohlsson, C. Sex Steroid Actions in Male Bone. *Endocr. Rev.* **2014**, *35* (6), 906–960. https://doi.org/10.1210/er.2014-1024.

(191) Nadal, M.; Prekovic, S.; Gallastegui, N.; Helsen, C.; Abella, M.; Zielinska, K.; Gay, M.; Vilaseca, M.; Taulès, M.; Houtsmuller, A. B.; et al. Structure of the Homodimeric Androgen Receptor Ligand-Binding Domain. *Nat. Commun.* **2017**, *8*, 14388. https://doi.org/10.1038/ncomms14388.

(192) Feldhammer, M.; Uetani, N.; Miranda-Saavedra, D.; Tremblay, M. L. PTP1B: A Simple Enzyme for a Complex World. *Crit. Rev. Biochem. Mol. Biol.* **2013**, *48* (5), 430–445. https://doi.org/10.3109/10409238.2013.819830.

(193) Barford, D.; Flint, A. J.; Tonks, N. K. Crystal Structure of Human Protein Tyrosine Phosphatase 1B. *Science* **1994**, *263* (5152), 1397–1404.

(194) Bourne, H. R.; Sanders, D. A.; McCormick, F. The GTPase Superfamily: Conserved Structure and Molecular Mechanism. *Nature* **1991**, *349* (6305), 117–127. https://doi.org/10.1038/349117a0.

(195) Gorfe, A. A.; Grant, B. J.; McCammon, J. A. Mapping the Nucleotide and Isoform-Dependent Structural and Dynamical Features of Ras Proteins. *Structure* **2008**, *16* (6), 885–896. https://doi.org/10.1016/j.str.2008.03.009.

(196) Milburn, V. M.; Tong, L.; deVos, A. M.; Brünger, A.; Yamaizumi, Z.; Nishimura, S.; Kim, S. H. Molecular Switch for Signal Transduction: Structural Differences between Active and Inactive Forms of Protooncogenic Ras Proteins. *Science* **1990**, *247* (4945), 939–45.

(197) Abankwa, D.; Gorfe, A. A.; Inder, K.; Hancock, J. F. Ras Membrane Orientation and Nanodomain Localization Generate Isoform Diversity. *Proc. Natl. Acad. Sci.* **2010**, *107* (3), 1130–1135. https://doi.org/10.1073/pnas.0903907107.

(198) Dechene, M.; Wink, G.; Smith, M.; Swartz, P.; Mattos, C. Multiple Solvent Crystal Structures of Ribonuclease A: An Assessment of the Method. *Proteins Struct. Funct. Bioinforma.* **2009**, *76* (4), 861–881. https://doi.org/10.1002/prot.22393.

(199) Buhrman, G.; O'Connor, C.; Zerbe, B.; Kearney, B. M.; Napoleon, R.; Kovrigina, E. A.; Vajda, S.; Kozakov, D.; Kovrigin, E. L.; Mattos, C. Analysis of Binding Site Hot Spots on the Surface of Ras GTPase. *J. Mol. Biol.* **2011**, *413* (4), 773–789. https://doi.org/10.1016/j.jmb.2011.09.011.

(200) Sasaki, A. T.; Carracedo, A.; Locasale, J. W.; Anastasiou, D.; Takeuchi, K.; Kahoud, E. R.; Haviv, S.; Asara, J. M.; Pandolfi, P. P.; Cantley, L. C. Ubiquitination of K-Ras Enhances Activation and Facilitates Binding to Select Downstream Effectors. *Sci. Signal.* **2011**, *4* (163), ra13. https://doi.org/10.1126/scisignal.2001518.

(201) Buhrman, G.; Kumar, V. S. S.; Cirit, M.; Haugh, J. M.; Mattos, C. Allosteric Modulation of Ras-GTP Is Linked to Signal Transduction through RAF Kinase. *J. Biol. Chem.* **2011**, *286* (5), 3323–3331. https://doi.org/10.1074/jbc.M110.193854.

(202) Satyanarayana, A.; Kaldis, P. Mammalian Cell-Cycle Regulation: Several Cdks, Numerous Cyclins and Diverse Compensatory Mechanisms. *Oncogene* **2009**, *28* (33), 2925–2939. https://doi.org/10.1038/onc.2009.170.

(203) Tsai, L.-H.; Harlow, E.; Meyerson, M. Isolation of the Human Cdk2 Gene That Encodes the Cyclin A- and Adenovirus E1A-Associated P33 Kinase. *Nature* **1991**, *353* (6340), 174–177. https://doi.org/10.1038/353174a0.

(204) Morris, M. C.; Gondeau, C.; Tainer, J. A.; Divita, G. Kinetic Mechanism of Activation of the Cdk2/Cyclin A Complex KEY ROLE OF THE C-LOBE OF THE Cdk. *J. Biol. Chem.* **2002**, *277* (26), 23847–23853. https://doi.org/10.1074/jbc.M107890200.

(205) Real, P. J.; Sierra, A.; de Juan, A.; Segovia, J. C.; Lopez-Vega, J. M.; Fernandez-Luna, J. L. Resistance to Chemotherapy via Stat3-Dependent Overexpression of Bcl-2 in Metastatic Breast Cancer Cells. *Oncogene* **2002**, *21* (50), 7611–7618. https://doi.org/10.1038/sj.onc.1206004.

(206) Zhao, Y.; Zeng, C.; Tarasova, N. I.; Chasovskikh, S.; Dritschilo, A.; Timofeeva, O. A. A New Role for STAT3 as a Regulator of Chromatin Topology. *Transcription 4* (5), 227–31.

(207) Genini, D.; Brambilla, L.; Laurini, E.; Merulla, J.; Civenni, G.; Pandit, S.; D'Antuono, R.; Perez, L.; Levy, D. E.; Pricl, S.; et al. Mitochondrial Dysfunction Induced by a SH2 Domain-Targeting STAT3 Inhibitor Leads to Metabolic Synthetic Lethality in Cancer Cells. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114* (25), E4924–E4933. https://doi.org/10.1073/pnas.1615730114.

(208) Lee, H.-J.; Zhuang, G.; Cao, Y.; Du, P.; Kim, H.-J.; Settleman, J. Drug Resistance via Feedback Activation of Stat3 in Oncogene-Addicted Cancer Cells. *Cancer Cell* **2014**, *26* (2), 207–221. https://doi.org/10.1016/j.ccr.2014.05.019.

(209) Lin, L.; Hutzen, B.; Li, P.-K.; Ball, S.; Zuo, M.; Deangelis, S.; Foust, E.; Sobo, M.; Friedman, L.; Bhasin, D.; et al. A Novel Small Molecule, LLL12, Inhibits STAT3 Phosphorylation and Activities and Exhibits Potent Growth-Suppressive Activity in Human Cancer Cells 1,2. https://doi.org/10.1593/neo.91196.

(210) Xu, X.; Kasembeli, M. M.; Jiang, X.; Tweardy, B. J.; Tweardy, D. J.; Schmidt, H. H. H. W. Chemical Probes That Competitively and Selectively Inhibit Stat3 Activation. **2009**. https://doi.org/10.1371/journal.pone.0004783.

(211) Masciocchi, D.; Villa, S.; Meneghetti, F.; Pedretti, A.; Barlocco, D.; Legnani, L.; Toma, L.; Kwon, B.-M.; Nakano, S.; Asai, A.; et al. Biological and Computational Evaluation of an Oxadiazole Derivative (MD77) as a New Lead for Direct STAT3 Inhibitors. *MedChemComm* **2012**, *3* (5), 592–599. https://doi.org/10.1039/C2MD20018J.

(212) Yu, H.; Pardoll, D.; Jove, R. STATs in Cancer Inflammation and Immunity: A Leading Role for STAT3. *Nat. Rev. Cancer* **2009**, *9* (11), 798–809. https://doi.org/10.1038/nrc2734.

(213) Yu, H.; Lee, H.; Herrmann, A.; Buettner, R.; Jove, R. Revisiting STAT3 Signalling in Cancer: New and Unexpected Biological Functions. *Nat. Rev. Cancer* **2014**, *14* (11), 736–746. https://doi.org/10.1038/nrc3818.

(214) Cocchiola, R.; Romaniello, D.; Grillo, C.; Altieri, F.; Liberti, M.; Magliocca, F. M.; Chichiarelli, S.; Marrocco, I.; Borgoni, G.; Perugia, G.; et al. Analysis of STAT3 Post-Translational Modifications (PTMs) in Human Prostate Cancer with Different Gleason Score. *Oncotarget* **2017**, *8* (26), 42560–42570. https://doi.org/10.18632/oncotarget.17245.

(215) Husby, J.; Todd, A. K.; Haider, S. M.; Zinzalla, G.; Thurston, D. E.; Neidle, S. Molecular Dynamics Studies of the STAT3 Homodimer:DNA Complex: Relationships between STAT3 Mutations and Protein–DNA Recognition. *J. Chem. Inf. Model.* **2012**, *52* (5), 1179–1192. https://doi.org/10.1021/ci200625q.

(216) Douguet, D. Data Sets Representative of the Structures and Experimental Properties of FDA-Approved Drugs. *ACS Med. Chem. Lett.* **2018**, *9* (3), 204–209. https://doi.org/10.1021/acsmedchemlett.7b00462.

(217) Wood, D. J.; Lopez-Fernandez, J. D.; Knight, L. E.; Al-Khawaldeh, I.; Gai, C.; Lin, S.; Martin, M. P.; Miller, D. C.; Cano, C.; Endicott, J. A.; et al. FragLites—Minimal, Halogenated Fragments Displaying Pharmacophore Doublets. An Efficient Approach to Druggability Assessment and Hit Generation. *J. Med. Chem.* **2019**, acs.jmedchem.9b00304. https://doi.org/10.1021/acs.jmedchem.9b00304.

(218) Šali, A.; Blundell, T. L. Comparative Protein Modelling by Satisfaction of Spatial Restraints. *J. Mol. Biol.* **1993**, *234* (3), 779–815. https://doi.org/10.1006/JMBI.1993.1626.

(219) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera - A Visualization System for Exploratory Research and Analysis. *J. Comput. Chem.* **2004**. https://doi.org/10.1002/jcc.20084.

(220) Shapovalov, M. V.; Dunbrack, R. L. A Smoothed Backbone-Dependent Rotamer Library for Proteins Derived from Adaptive Kernel Density Estimates and Regressions. *Structure* **2011**, *19* (6), 844–858. https://doi.org/10.1016/j.str.2011.03.019.

(221) Abraham, M. J.; Murtola, T.; Schulz, R.; PÃ¡ll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, *1–2*, 19–25. https://doi.org/10.1016/j.softx.2015.06.001.

(222) Guvench, O.; MacKerell, A. D. Computational Fragment-Based Binding Site Identification by Ligand Competitive Saturation. *PLoS Comput. Biol.* **2009**, *5* (7), e1000435. https://doi.org/10.1371/journal.pcbi.1000435.

(223) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25* (9), 1157–1174. https://doi.org/10.1002/jcc.20035.

(224) Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: II. Parameterization and Validation. *J. Comput. Chem.* **2002**. https://doi.org/10.1002/jcc.10128.

(225) Sousa, A. W.; Vranken, W. F. Open Access ACPYPE - AnteChamber PYthon Parser InterfacE. **2012**, 1–8.

(226) Hess, B. P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *J. Chem. Theory Comput.* **2008**, *4* (1), 116–122. https://doi.org/10.1021/ct700200b.

(227) Lexa, K. W.; Carlson, H. A. Full Protein Flexibility Is Essential for Proper Hot-Spot Mapping. *J. Am. Chem. Soc.* **2011**, *133* (2), 200–202. https://doi.org/10.1021/ja1079332.

(228) Bussi, G.; Donadio, D.; Parrinello, M. Canonical Sampling through Velocity Rescaling. *J. Chem. Phys.* **2007**, *126* (1), 014101. https://doi.org/10.1063/1.2408420.

(229) Yang, Z.; Lasker, K.; Schneidman-Duhovny, D.; Webb, B.; Huang, C. C.; Pettersen, E. F.; Goddard, T. D.; Meng, E. C.; Sali, A.; Ferrin, T. E. UCSF Chimera, MODELLER, and IMP: An Integrated Modeling System. *J. Struct. Biol.* **2012**, *179* (3), 269–278. https://doi.org/10.1016/j.jsb.2011.09.006.

(230) Huang, C. C.; Meng, E. C.; Morris, J. H.; Pettersen, E. F.; Ferrin, T. E. Enhancing UCSF Chimera through Web Services. *Nucleic Acids Res.* **2014**, *42* (W1), W478–W484. https://doi.org/10.1093/nar/gku377.

(231) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera- A Visualization System for Exploratory Research and Analysis. *J. Comput. Chem.* **2004**, *25* (13), 1605–1612. https://doi.org/10.1002/jcc.20084.

(232) Corbeil, C. R.; Williams, C. I.; Labute, P. Variability in Docking Success Rates Due to Dataset Preparation. *J. Comput. Aided Mol. Des.* **2012**, *26* (6), 775–786. https://doi.org/10.1007/s10822-012-9570-1.

(233) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations through Multi-

Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, *1–2*, 19–25. https://doi.org/10.1016/J.SOFTX.2015.06.001.

(234) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved Side-Chain Torsion Potentials for the Amber Ff99SB Protein Force Field. *Proteins Struct. Funct. Bioinforma.* **2010**, *78* (8), NA-NA. https://doi.org/10.1002/prot.22711.

(235) Hess*, B. P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. **2007**. https://doi.org/10.1021/CT700200B.

(236) Verlet, L. Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Phys. Rev.* **1967**, *159* (1), 98–103. https://doi.org/10.1103/PhysRev.159.98.

(237) Bussi, G.; Donadio, D.; Parrinello, M. Canonical Sampling through Velocity Rescaling. *J. Chem. Phys.* **2007**, *126* (1), 014101. https://doi.org/10.1063/1.2408420.

(238) Parrinello, M.; Rahman, A. Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method. *J. Appl. Phys.* **1981**, *52* (12), 7182–7190. https://doi.org/10.1063/1.328693.

(239) Hub, J. S.; de Groot, B. L.; van der Spoel, D. G_wham—A Free Weighted Histogram Analysis Implementation Including Robust Error and Autocorrelation Estimates. *J. Chem. Theory Comput.* **2010**, *6* (12), 3713–3720. https://doi.org/10.1021/ct100494z.

(240) Schrödinger, LLC. The PyMOL Molecular Graphics System, Version 2.0, 2017.

(241) RStudio Team. *RStudio: Integrated Development Environment for R*; RStudio, Inc.: Boston, MA, 2015.

(242) Forli, S.; Huey, R.; Pique, M. E.; Sanner, M. F.; Goodsell, D. S.; Olson, A. J. Computational Protein–Ligand Docking and Virtual Drug Screening with the AutoDock Suite. *Nat. Protoc.* **2016**, *11* (5), 905–919. https://doi.org/10.1038/nprot.2016.051.