Computational engineering of co-substrate specificity in protein kinases

Inauguraldissertation

zur
Erlangung der Würde eines Doktors der Philosophie
vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultat der
Universitat Basel

von
Valentina Romano
aus Italien

Basel, 2016

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät

auf Antrag von:

Prof. Dr. Torsten Schwede
Prof. Dr. Anna Tramontano

Basel 23 February 2016 Prof. Dr. Jörg Schibler (Dekan)

creative commons

Namensnennung-Keine kommerzielle Nutzung-Keine Bearbeitung 2.5 Schweiz

---

Sie dürfen:

das Werk vervielfältigen, verbreiten und öffentlich zugänglich machen

Zu den folgenden Bedingungen:

Namensnennung. Sie müssen den Namen des Autors/Rechteinhabers in der von ihm festgelegten Weise nennen (wodurch aber nicht der Eindruck entstehen darf, Sie oder die Nutzung des Werkes durch Sie würden entlohnt).

Keine kommerzielle Nutzung. Dieses Werk darf nicht für kommerzielle Zwecke verwendet werden.

Keine Bearbeitung. Dieses Werk darf nicht bearbeitet oder in anderer Weise verändert werden.

- Im Falle einer Verbreitung müssen Sie anderen die Lizenzbedingungen, unter welche dieses Werk fällt, mitteilen. Am Einfachsten ist es, einen Link auf diese Seite einzubinden.
- Jede der vorgenannten Bedingungen kann aufgehoben werden, sofern Sie die Einwilligung des Rechteinhabers dazu erhalten.
- Diese Lizenz lässt die Urheberpersönlichkeitsrechte unberührt.

Die gesetzlichen Schranken des Urheberrechts bleiben hiervon unberührt.

Die Commons Deed ist eine Zusammenfassung des Lizenzvertrags in allgemeinverständlicher Sprache: http://creativecommons.org/licenses/by-nc-nd/2.5/ch/legalcode.de

Haftungsausschluss:
Die Commons Deed ist kein Lizenzvertrag. Sie ist lediglich ein Referenztext, der den zugrundeliegenden Lizenzvertrag übersichtlich und in allgemeinverständlicher Sprache wiedergibt. Die Deed selbst entfaltet keine juristische Wirkung und erscheint im eigentlichen Lizenzvertrag nicht. Creative Commons ist keine Rechtsanwaltsgesellschaft und leistet keine Rechtsberatung. Die Weitergabe und Verlinkung des Commons Deeds führt zu keinem Mandatsverhältnis.

Quelle: http://creativecommons.org/licenses/by-nc-nd/2.5/ch/ Datum: 3.4.2009

To my dad.

I love you daddy, wherever you are.

If science teaches us anything, it teaches us to accept our failures, as well as our successes, with quiet dignity and grace.

Dr. F. Frankenstein

Abstract

Protein kinases are key regulators of most biochemical pathways and their involvement in different diseases is extensively documented. To identify the protein substrates of kinases is therefore of great importance for elucidating their functional role in the cell and to develop disease-specific therapies. However, the identification of specific kinase substrates is highly challenging due to the large number of protein kinases in cells, their substrate specificity overlap and the lack of absolute specificity of inhibitors. In the late 90s, Shokat and coworkers developed a protein engineering-based method addressing the question of identification of substrates of protein kinases. The approach was based on the mutagenesis of a specific residue to enlarge the ATP binding pocket of the target kinase to accommodate a chemically modified ATP as co-substrate, which would not bind to the native kinase. One of the challenges in applying this method to other kinases is to identify the optimal combination of kinase binding pocket mutations and ATP analogues such that the ATP analogue acts as specific co-substrate for the engineered kinase. Furthermore, the engineered kinases have to remain catalytically active.

This work aims to develop a computational protocol for the engineering of protein kinases. We predict which residues within the binding pocket of the target kinase could be mutated to change its co-substrate specificity from ATP to an ATP analogue. The protocol explores pairings of potential mutations and ATP analogues and can be used as prescreening test in the wider experiment for identifying specific substrates of protein kinases.

The protocol was tested on different tyrosine and serine/threonine protein kinases from the scientific literature where Shokat's method was applied and experimental data were available. The method correlates well with published experimental data available for the tested protein kinases. Subsequently, we applied the computational protocol to the *Mycobacterium tuberculosis* protein kinase G, *Mtb* PknG. *Mtb* is a pathogenic bacterium and is the causative agent of tuberculosis. Tuberculosis is a widespread infectious disease which causes around two million deaths per year. PknG plays a key role in the survival of *Mtb* within the host

organism. Since its specific downstream substrates as well as its mechanism of action are still unknown, PknG is an attractive target for our computational approach. Our protocol allowed us to design a number of pairs of PknG mutants and ATP analogues. The most promising pairs were tested *in vitro*, in our laboratory. All *in vitro* tests were performed by Mohamed-Ali Mahi. The most interesting pair was then used in follow-up *ex vivo* experiments, performed by the group of Prof. J. Pieters at Biozentrum.

Contents

1 Introduction

1.1 Protein engineering

Proteins perform a vast array of functions in cells. They act as enzymatic catalysts, transport materials across cell walls and have structural, sensory and regulatory functions. Protein engineering is a widely used tool in many fields. It allows for investigation of protein functions, for the construction of proteins with new and/or improved functions and for increasing protein stability and functionality. In the last two decades, computational methods for protein engineering have been developed and used to achieve significant findings in different fields such as pharmaceutics, synthetic biology and industrial production [1-5]. They are used to design novel biocatalysts, such as the O_2-dependent phenol oxidase able to catalyze a phenol oxidase reaction [6], to design proteins with an improved binding affinity and specificity, like the case of a series of antibody Fc variants with optimized affinity and specificity for cell surface Fc receptors [7], and to design proteins able to bind non-natural cofactors, such as a four-helix bundle protein that selectively binds to the nonbiological DPP-Fe(III) cofactor [8]. Computational protein engineering methods are organized into two main categories, rational design and *de novo* design [3] (Figure 1.1). The rational design approach requires sequence and structural information. Multiple sequence alignments (MSA), three-dimensional (3D) structures and 3D structure predictions are the best tools to extract significant information such as key residues, functionally sites and correlated mutations. All those tools can be used individually or in combination to generate engineered proteins (Figure 1.1, left part). *De novo* design refers to the generation of novel protein folds and/or enzymatic activities. Such approaches consist of four steps, modelling of the theozyme (that is a computational model of the transition state, TS, of a specific reaction including key amino acids), searching for a protein scaffold, energy minimization to remove possible TS-catalytic residues clashes and selection of design models for experimental validation (different factors can be used for selection, such as ligand binding energy) (Figure 1.1, right part). In the field of *de novo* design, the

ROSETTA enzyme design protocol developed in Baker's laboratory [9] is a milestone and is the most widely used tools for *de novo* protein engineering.

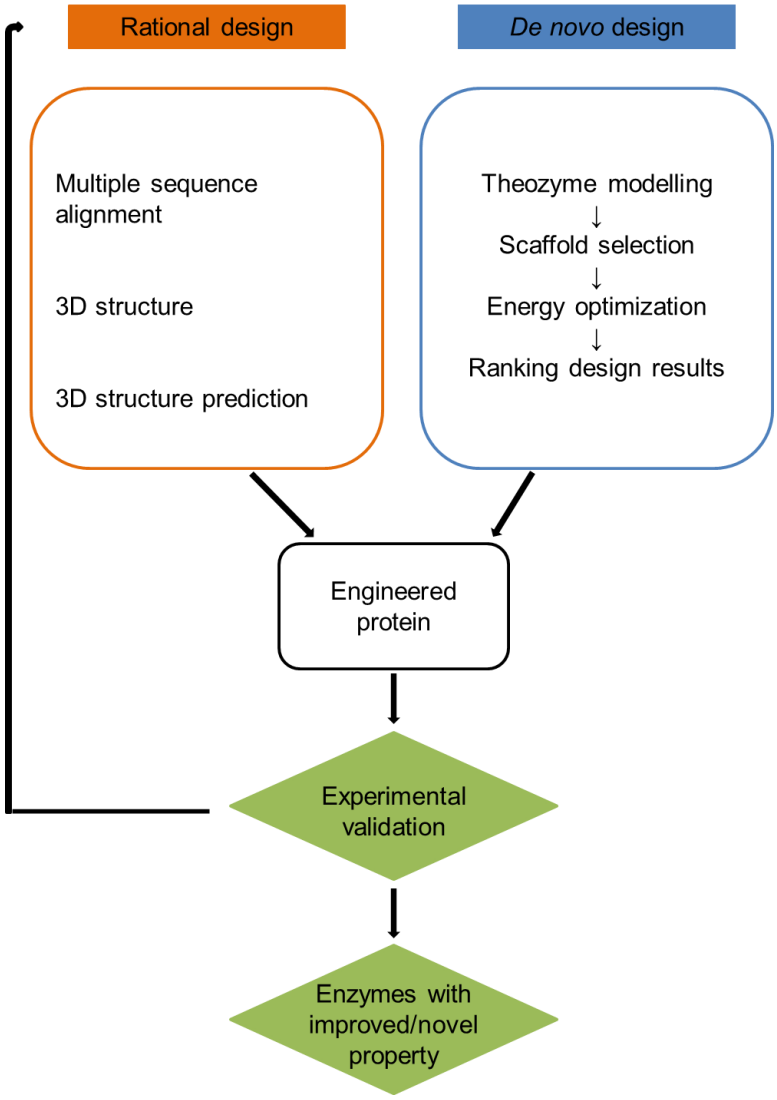

Figure 1.1. The two main approaches in protein engineering, rational design and *de novo* design.

Recent work has demonstrated that computational protein engineering methods can generate active catalysts [10-12]. For instance, Jiang and coworkers designed new enzymes able to catalyze retro-aldol reactions [11]. To evaluate the accuracy of the design models, they solved their structures by x-ray

crystallography. The crystal structure of the retro-aldolase 22, RA22, shows that the catalytic residues (Lys159, His233 and Asp53) superpose well on the design model and the rest of the active site is nearly identical to that one of the design model (Figure 1.2). To evaluate the functionality of design models, they monitored the retro-aldolase activity via a fluorescence-based assay and the design RA22 shows a retro-aldolase activity.

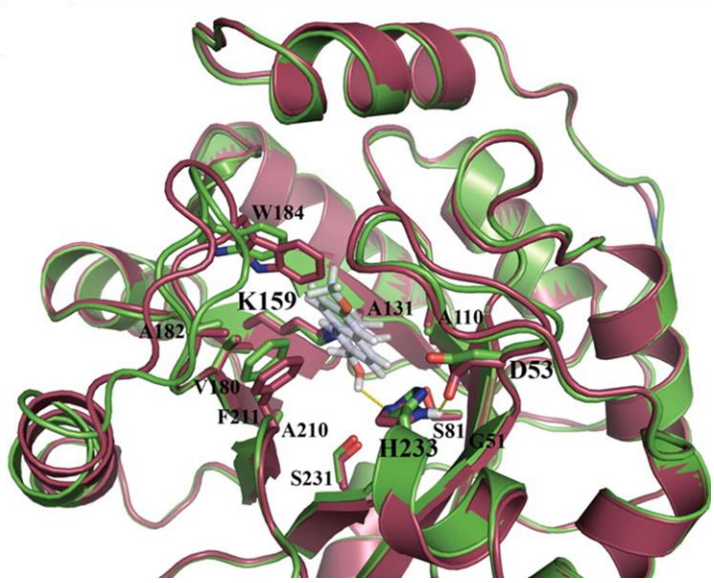

Figure 1.2. (From [11]) Superposition of the binding site residues of the crystal structure (green, PDB: 3HOJ) and designed model (purple) of the RA22 in complex with its substrate (4-hydroxy-4-(6-methoxy-2-naphthyl)-2-butanone, gray stick). The Cα root-mean-square deviation (rmsd) is 0.62 Å.

Nevertheless, there are some aspects of computational methods that need to be improved, such as the prediction of protein-ligand binding affinities. The predicted binding affinities are generally modest or even undetectable when measured experimentally [13, 14]. Schreier and coworkers [13] worked on designing models of variants of the arabinose-binding protein (ABP). One of those models was reported to bind to serotonin (Stn) and was called Stn.A2. They solved the structure of Stn.A2 bound to Stn and compared it with the computational model. The binding pockets of the structure and the model show high similarity with an overall atom rmsd of 0.79 Å (Figure 1.3 A). Although the conformation of the

protein resembles the model, the binding of the ligand to the protein was not experimentally confirmed. Isothermal titration calorimetry, ITC, was used to probe the ligand binding affinity of the designed model. It works by measuring the heat that is released during a binding process. While ABP binds to its specific ligand, no significant change in heat upon addition of Stn could be detected in the case of Stn.A2 (Figure 1.3 B). The analysis performed by Schreier and coworkers shows the importance of direct validation of the predicted protein-ligand interactions as well as the significant role of the experimental testes as instrument to improve computational methods.

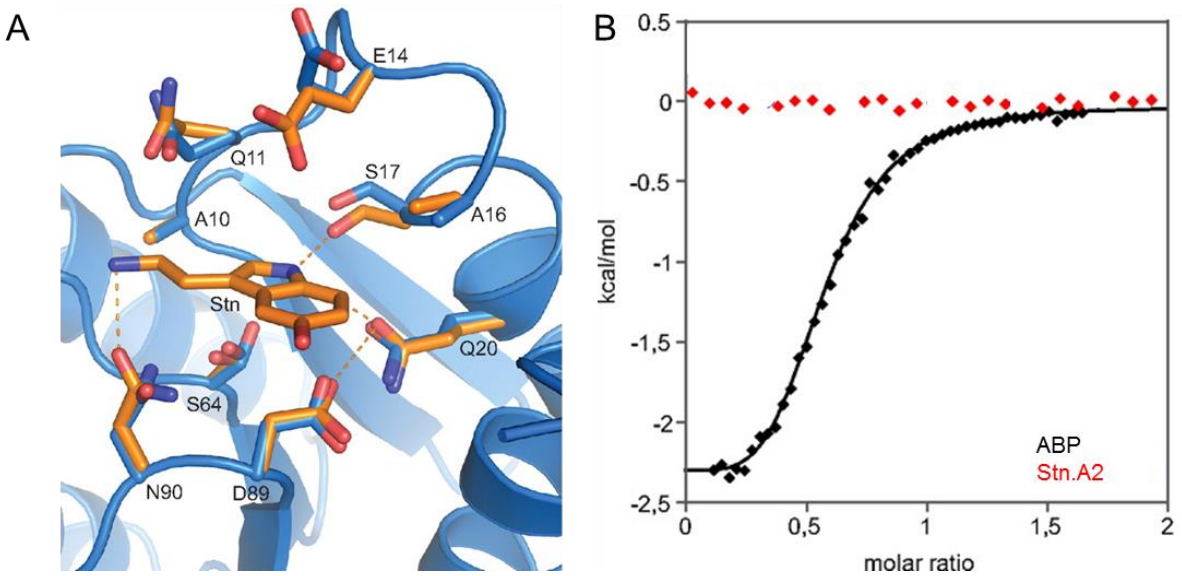

Figure 1.3. (From [13]) A) Superposition of binding site residues of the crystal structure (blue, PDB: 5ABP) and the design model (orange) of Stn.A2 in complex with Stn. B) ITC measurements for ABP and Stn.A2. ITC shows that binding occurs for AP but not for the designed Stn.A2.

1.2 A protein engineered-based method for protein kinases

The identification of the direct substrates of protein kinases is of great importance for elucidating the functional roles of these enzymes in cells. However, the identification of specific kinase substrates is highly challenging due to the large number of protein kinases in cells, their substrate specificity overlap and the lack

of absolute specificity of inhibitors [15, 16]. In 1997, Shokat and coworkers developed a protein engineering-based method to solve this important issue in molecular biology [17]. They used the prototypical viral proto-oncogene tyrosine protein kinase Src (v-Src) and engineered its ATP binding pocket by mutating residue Ile338 into Gly (residues are numbered as in PDB structures). The single point mutation enlarged the binding pocket making a hydrophobic region behind the ATP binding pocket accessible to ATP-competitive ligans with non-polar groups at the N6 position of the adenine base. The engineered v-Src preferentially used the N6-(benzyl) ATP as phosphodonor. The use of a N6-(benzyl) ATP with a radiolabeled γ phosphate, γ-^{32}P, resulted in the v-Src substrates being specifically radiolabeled and identified in presence of other protein kinases and all other kinase substrates (Figure 1.4) [18, 19]. This approach allowed the identification of cofilin and calumenin as specific substrates of v-Src [20]. The residue that controls the access to the hydrophobic region beyond the ATP binding pocket, Ile338 in v-Src, is called ‘gatekeeper’ residue. The Shokat method is based on the ‘bump-and-hole’ model [21, 22]. The gatekeeper residue is substituted with a small amino acid generating a ‘hole’ within the ligand-binding site that can accept ligands with bulky substituent groups, ‘bumps’, that sterically clash with the wild-type binding pocket. The conservation of the ATP binding site between different protein kinases makes the approach widely applicable for identifying specific kinase substrates. In a similar approach, kinases were engineered to bind specifically modified inhibitors [23-29].

1.3 Serine/threonine and tyrosine protein kinases: structure and mechanism

The activity of many proteins in cell is regulated by phosphorylation, a reversible covalent modification. Protein kinases are the enzymes that catalyze phosphorylation reactions and are key regulators of most biochemical pathways by phosphorylating a single protein or several closely related proteins in cells [30].

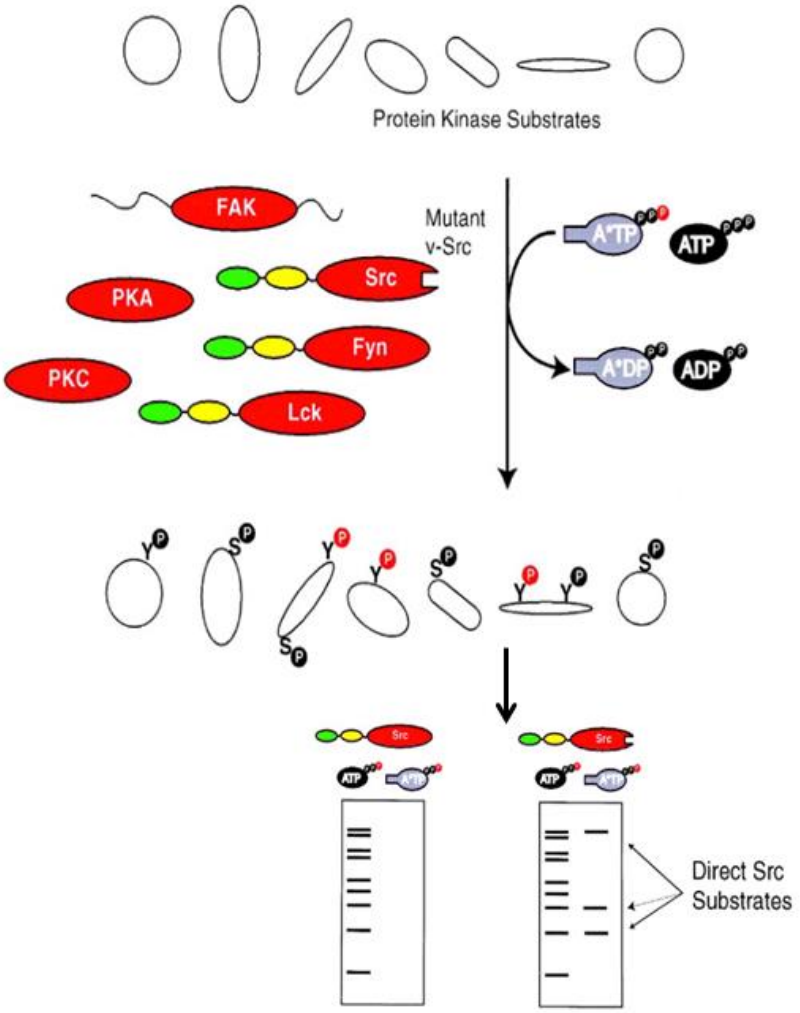

Figure 1.4. (From [19]). Schematic representation of the method developed by Shokat and coworkers in 1997. Empty ovals represent substrates of protein kinases, red ovals represent kinase domains and other colored ovals represent regulatory domains. A*TP is the [γ-^{32}P] N6-(benzyl) ATP (red P is γ-^{32}P). Y and S are tyrosine and serine that are phosphorylated by ATP (black P) or ATP analogue (red P).

They constitute one of the largest protein families known (more than 100 proteins in yeast and more than 550 in humans). The majority of serine/threonine and tyrosine kinases share a bilobal kinase domain fold [15, 31]. The N-lobe is formed by five β-strands and a single α-helix whereas the C-lobe is predominantly α-helical (Figure 1.5) The C-lobe contains the activation segment, typically

composed of 20-30 residues, and the catalytic loop. The activation segment contains the activation loop, that activates protein kinase when a specific residue (usually Tyr or Thr) is phosphorylated, and the loop that is involved in substrate binding [32, 33]. The catalytic loop contains a highly conserved Asp that has a significant role in the phosphorylation reaction. It acts as catalytic base to free up the hydroxyl oxygen of a Ser, Thr or Tyr on the protein substrate. The deprotonated residue is involved in a nucleophilic attack on the terminal phosphoryl group (PO_3^{2-}) of ATP [15].

The kinase domain exists in two main conformations, active and inactive (Figure 1.6). In the active conformation the two lobes are close to each other and the activation loop is phosphorylated in an open and extended conformation that allows substrate binding.

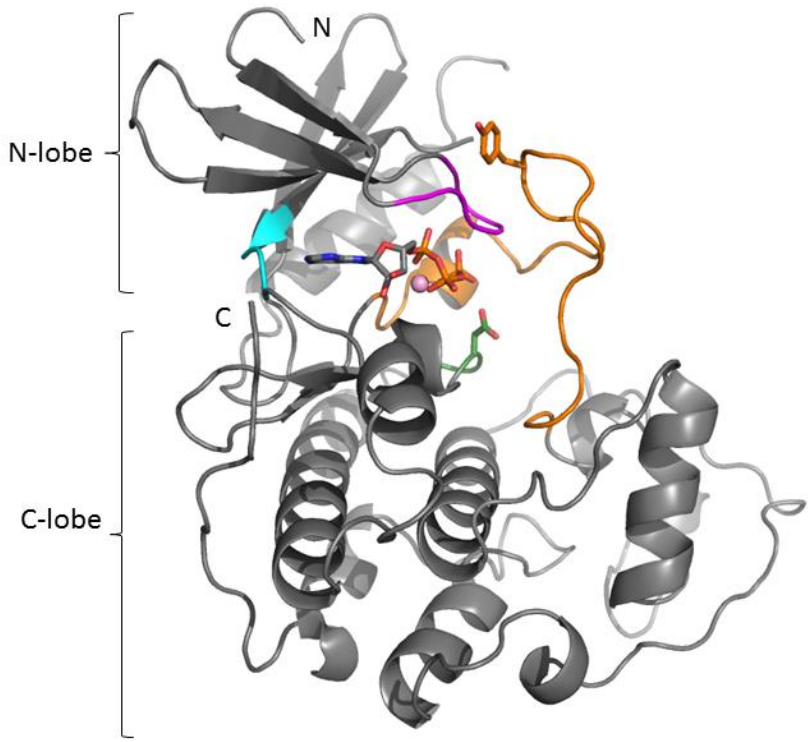

Figure 1.5. Ribbon representation of the inactive kinase domain of the human cyclin-dependent kinase 2 (Cdk2, PDB: 1HCK). ATP is represented as stick and the Mg^{2+} ion as a pink sphere. The activation segment is represented in orange, the catalytic loop in green, the P loop in magenta and the hinge region in cyan. The Tyr belonging to the activation loop and the Asp belonging to the catalytic loop are represented as sticks.

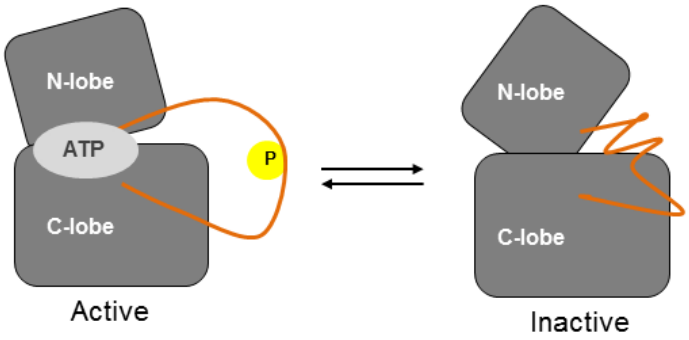

Figure 1.6. Schematic representation of the equilibrium between active and inactive conformations of a kinase domain. Activation loop is represented in orange.

In the inactive state the two lobes are far apart, the activation loop is unphosphorylated and in a closed conformation that sometimes folds into a short helix [34, 35]. The ATP is bound in a cleft between the two lobes that are connected by a short segment called the hinge region [36]. When ATP occupies the ligand-binding site, the phosphates are in part coordinated by the glycine-rich loop that is also known as the phosphate binding loop (P loop). The P loop contains the conserved motif GXGXΦG where G is Gly, X is any amino acid and Φ is usually Tyr or Phe. Glycines make the P loop flexible allowing it to approach the phosphates of ATP and to bind them via backbone interactions [33, 37]. As previously mentioned, ATP acts as phosphodonor in diverse biochemical pathways catalyzed by protein kinases (Figure 1.7 A) and the ATP binding site is made up of five areas (Figure 1.8 A) [38, 39]. The adenine ring of ATP occupies the adenine region and makes favorable hydrophobic interactions with hydrophobic residues below and above the adenine plane. In addition, the adenine base contains an H-bond donor at position N6 and an H-bond acceptor at position N1. In serine/threonine and tyrosine kinases those two atoms are involved in two H-bonds generally described by a tri-residue N_i-O_{i-2} motif. The atom N1 forms an H-bond to the backbone N of the *i*th residue and the atom N6 forms an H-bond to the backbone O of the (*i*-2)th residue [40, 41]. Sometimes, the C2 position of adenine acts as H-bond donor and is involved in a weak interaction

with a hinge region residue (Figure 1.8 A) [40]. The N6-N_i and N1-O_{i-2} H-bonds are highly conserved in almost all serine/threonine and tyrosine protein kinases. Due to their highly directional nature, the H-bonds between the protein kinase and the adenine moiety strongly influence the position and the orientation of the planar adenine ring within the adenine region.

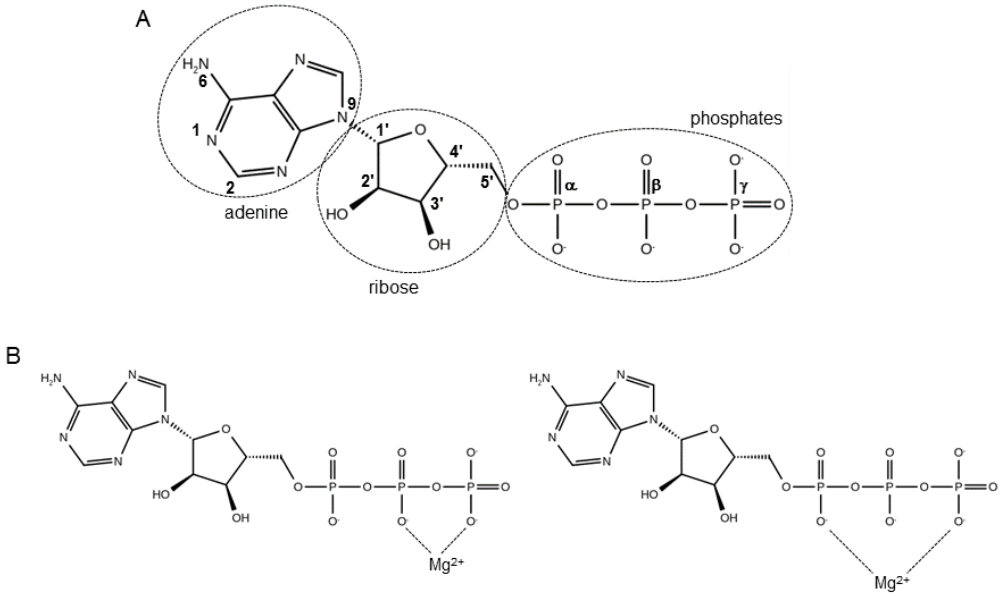

Figure 1.7. A) Two dimensional (2D) structure of ATP. B) Two isomeric forms of the ATP-Mg^{2+} complex. The divalent cation can be coordinated by β and γ phosphates (structure on the left) or by α and γ phosphates (structure on the right).

The ribose ring occupies the sugar region that is mostly polar. Either ribose 2' or 3' hydroxyls groups (OH) are involved in one H-bond within the sugar region. Babor and coworkers had analyzed the ribose-protein polar interactions in a dataset of ATP, ADP and FAD in complex with poteins concluding that water molecules play a crucial role in those interactions [42]. Generally, 2'OH or 3'OH interact with the oxygen atom of a water molecule which, in turn, interacts with a protein residue within the sugar region. In Cdk2 the 2'OH forms an H-bond with a water molecule that, in turn, interacts with the side chain of Asp86 (Figure 1.8 B). The ribose ring is not planar, quite flexible and can assume different conformations. In addition, the orientation is not conserved because of diverse torsion angle values at bonds

connecting it to phosphates and adenine (5'CH2–C4' and C1'–N9, Figure 1.7 A) [43].

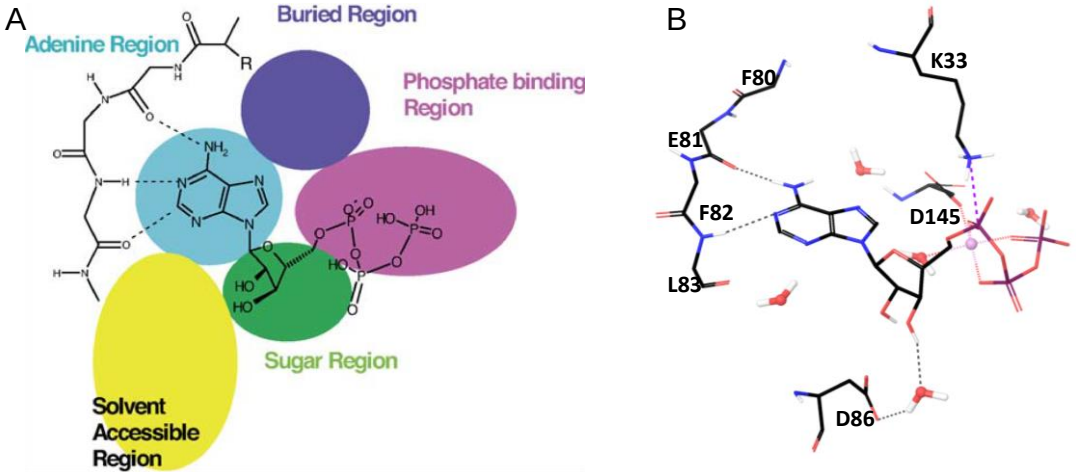

Figure 1.8 A) (From [39]) Schematic representation of ATP binding pocket regions. Dashed black lines represents hydrogen bonds. The five regions are: adenine region (cyan), sugar region (green), phosphates region (violet), buried region (blue) and solvent accessible region (yellow). B) Cdk2 binding site in complex with ATP-Mg^{2+} (PDB: 1HCK). For simplicity only hinge region residues (80-83), amino acids involved in interactions with the ligand (K33, D145 and D86) and water molecules belonging to the binding site are represented. Mg^{2+} is represented as pink sphere and red dashed lines represent interactions of the bivalent cation with ATP and protein. Black dashed lines represent H-bonds and the purple dashed line the interaction between Lys33 and the α phosphate.

The phosphates region contains the P loop and two residues (a Lys and an Asp) that play an important role in the catalytic activity and are conserved in almost all serine/threonine and tyrosine protein kinases [44]. Lys interacts with the α phosphate or with both the α and β phosphates. Its role is facilitating the transfer of the PO_3^{2-} group without influencing the binding of ATP. In Cdk2 the catalytic Lys is Lys33 that interacts with the α phosphate (Figure 1.8 B). To understand the role of the conserved Asp in the phosphates region, it is firstly essential to illustrate the function of divalent metal ions, such as Mg^{2+} or Mn^{2+}, in the catalytic activity of protein kinases. Kinetic studies had revealed that protein kinases are essentially inactive in absence of divalent cations within the binding site [30]. The divalent ion is not part of the binding pocket but the complex of a metal ion and

ATP is the true substrate of the protein kinase. The divalent ion, often Mg^{2+}, chelates the β and γ phosphates or the α and γ phosphates of ATP (Figure 1.7 B). It is important for diverse reasons. First of all, it neutralizes the negative charge of the phosphates limiting electrostatic repulsions. Moreover, the interaction between the ion and the phosphates hold the nucleotide in a well-defined conformation with the terminal phosphate correctly placed for the transfer to the substrate. The strictly conserved Asp interacts with the essential divalent ion assuming a significant role in the kinase catalytic activity [30, 44, 45]. In Cdk2 the catalytic Asp is Asp145 and it interacts with Mg^{2+} (Figure 1.8 B). The phosphate moiety can assume different conformations within the phosphates region because of diverse torsion angle values at bonds connecting phosphate and oxygen atoms. Moreover, the presence of one or two divalent cations can also influence the conformation assumed by the phosphates. The solvent accessible area is a hydrophobic slot open to the solvent and it is not occupied by ATP. The buried region is a hydrophobic region located in the back of the ATP pocket and it is not used by ATP. The size and the shape of the buried region are controlled by the first amino acid of the hinge region. In 73% of human kinases a hydrophobic amino acid with a bulky side chain (Met, Phe or Leu) is observed at that position, 22% have a small residue, such as Thr or Val and the remaining 5% has one of the other amino acids [38, 39, 46, 47]. This amino acid acts as a ‘molecular gate’ controlling the accessibility to the buried region, indeed a residue with a large side chain effectively ‘closes the gate’ making the buried region inaccessible. For that reason, this residue has been termed the ‘gatekeeper’ residue [19, 48-50].

1.4 Protein-ligand interaction

Most biological processes rely on the mutual recognition of proteins with their specific ligands. Selective protein-ligand binding is governed by two main factors: geometry and chemistry. Geometry implies the shape complementarity between the protein and the ligand. The ligand binds to the protein with a specific geometry, where the geometry is defined by the location, the orientation and the

conformation of the ligand within the protein binding site. Chemistry implies the occurrence of specific and favorable protein-ligand non-covalent interactions (Figure 1.9) [51, 52]. In the unbound state, protein and ligand are separately solvated and do not interact, whereas in the bound state both partners are partially desolvated and form non-covalent interactions between each other.

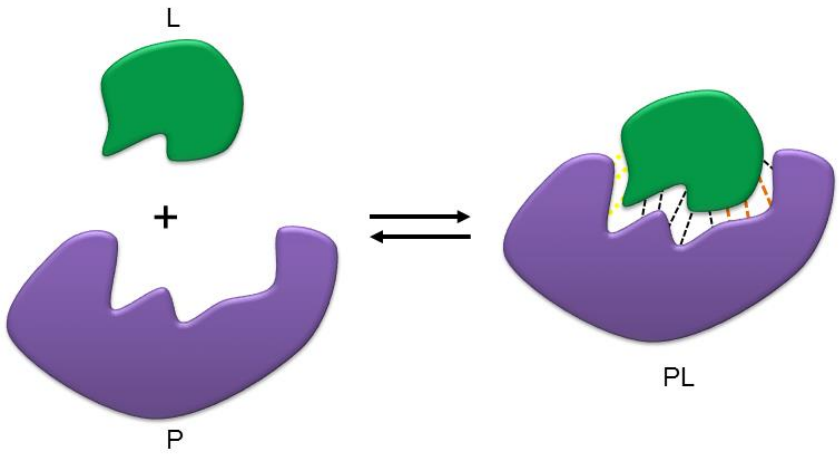

Figure 1.9 A schematic overview of the geometry and chemistry contributions to protein-ligand interaction. The protein (P, in violet) and the ligand (L, in green) interact to form the protein-ligand complex (PL). Non-covalent interactions are represented as dashed lines, where are represented H-bonds in yellow, van der Waals interactions in black and electrostatic interactions in orange.

The interaction of a protein, P, with a ligand, L, to form a protein-ligand complex, PL, can be written as follows:

$$P + L \leftrightarrow PL \quad (1.1)$$

It is common practice to describe the equilibrium represented in equation 1.1 by the dissociation constant, K_d:

$$K_d = \frac{[P][L]}{[PL]} \quad (1.2)$$

where K_d has the dimensions of a concentration (mol/L) and represents the protein-ligand binding affinity. The smaller the K_d value, the more strongly the ligand binds to the protein [51, 53]. A non-covalent association of a protein and a ligand is governed by general thermodynamics and it occurs only when it is characterized by a negative Gibb's energy, ΔG:

$$\Delta G = \Delta H - T\Delta S \quad (1.3)$$

The enthalpic contribution, ΔH, reflects the strength of the non-covalent interactions between the protein and the ligand. The entropic contribution, ΔS, relates to changes in the order of both the protein and the ligand in the complex formation process and of the solvent. T is the temperature of the system [51, 54]. The relationship between the Gibbs energy and the binding affinity is given by equation 1.4:

$$\Delta G = RT \ln K_d \quad (1.4)$$

where R is the gas constant, T the temperature and K_d the dissociation constant previously described [51, 54, 55]. At 'room temperature' (T equal to 298 K and R equal to 8.314 J K^{-1} mol^{-1}) and using 2.303 as conversion factor between natural logarithm ($\ln$) and logarithm to the base 10 ($\log_{10}$), equation 1.4 becomes:

$$\Delta G = -1.4 \log_{10} K_d \quad (1.5)$$

This means that each free energy change of 1.4 kcal/mol will lead to a 10 fold change in K_d.[56]. Generally, biologically important non-covalent interactions have dissociation constants that range from picomolar (~$1*10^{-12}$) for the tightest interactions to millimolar (~$1*10^{-3}$) for the weakest ones. These correspond to free energy of binding ranging from ~-17 kcal/mol to ~-4 kcal/mol [53, 55]. The typical non-covalent interactions found in protein-ligand complexes are reported in Table 1.1.

Table 1.1 Common protein-ligand interactions with relative enthalpic contributions in kcal/mol.

Interactions	**Example**	**Energy (kcal/mol)**
van der Waals	Alkyl groups	0.1-1
Hydrogen bond	X-H ---Y[a]	2-30
Electrostatic	δ^+--- δ^- [b]	1-20
Hydrophobic	Non polar groups	<10

a) X is the H-bond donor and Y the H-bond acceptor

b) It represents an ion-dipole interaction

The van der Waals (vdW) interactions are both attractive and repulsive interactions. Attractive vdW involve two induced dipoles, atoms or molecules that are at a given distance and are not covalently bound. Repulsive vdW interactions occur when the two induced dipoles become too close to each other. The vdW interactions are described by the Lennard-Jones potential (LJ potential):

$$V_{LJ} = 4\varepsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right] \quad (1.6)$$

where ε is the depth of the potential, σ is the distance at which the interparticle potential is zero and r the distance between the particles. The r^{-12} term is the repulsive short-range term and the r^{-6} is the attractive long-range term [54, 57]. Those interactions are very weak compared to other non-covalent interactions (Table 1.1) and to covalent bonds (e.g. the covalent C-C bond has an energy of ~83 kcal/mol [58]).

The H-bonds result from an electrostatic interaction between one hydrogen atom covalently bound to an electronegative atom called ‘donor’, D, and an electronegative atom called ‘acceptor’, A (Figure 1.10). H-bonds are highly directional and generate interatomic distances shorter than the sum of the van der Waals radii of the involved atoms [54, 55]. Typically, the distance between H-bond donors and acceptors ranges from 2.5 to 3.2 Å and the D-H---A angles have values between 130° and 180° [59]. In biological systems, such as protein-ligand

complexes, H-bonds follow strict geometric rules (their orientations, lengths and angles) and that makes those interactions very specific [56].

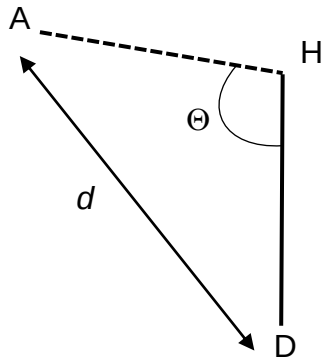

Figure 1.10. Schematic representation of an H-bond. A is the acceptor, D the donor and H the hydrogen atom. *d* represents the donor-acceptor distance while Θ is the D-H---A angle.

Electrostatic interactions are long-range interactions and can be attractive or repulsive. They are classified in three types, charge-charge (between charged groups), charge-dipole (that normally occur between ionized amino acid side chains and the dipole of the ligand or a water molecule) and dipole-dipole. All types of electrostatic interactions are described by Coulomb's law:

$$V = k\frac{q_1 q_2}{\varepsilon r_{12}} \quad (1.7)$$

where q_1 and q_2 are the interacting particles, ε is the dielectric constant of the medium in which particles are placed, r_{12} is the distance between q_1 and q_2 and k is the Coulomb's constant [54, 55].

Hydrophobic interactions occur between non-polar ligands and hydrophobic side chains protein residues. Generally, the hydrophobic residues and hydrophobic ligands repel water molecules resulting in a net non-polar attraction. The protein-ligand binding displaces water molecules from the protein interaction interface and from the interaction interface of the ligand to the bulk solvent. The

release of water molecules to the solvent results in a loss of enthalpy due to the disruption of protein-water and ligand-water interactions. The enthalpy loss is compensated by a gain of entropy, indeed water molecules are transferred from an organized network (around protein and ligand) to the bulk solvent (Figure 1.11) [54, 56, 60].

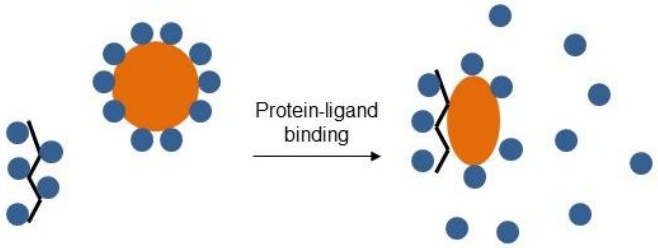

Figure 1.11. Schematic representation of protein-ligand hydrophobic interactions. Orange circle represent the protein, black crooked line the ligand and blue circles water molecules.

1.5 Computational approaches to compute protein-ligand affinity

Molecular recognition of proteins with specific ligands is central to biology. It is the basis of many processes such as hormonal control or enzymatic catalysis, just to cite a few. It is also the foundation for exogenous control of biological systems and many medications act by binding specific macromolecular targets. Indeed, the experimental identification of a ligand that specifically binds to a target protein is of great importance to elucidate the functional role of the protein and is a significant step in drug discovery. However, the ligand identification is a major and costly challenge. Computational methods can speed up this process and the calculated binding affinities can reduce the number of *in vitro* and *in vivo* experiments to perform [61, 62]. The theoretical treatment of protein-ligand binding requires the consideration of all species involved in the binding process (protein, ligand, protein-ligand complex, water and counterions) and of all possible interactions [51]. Moreover, both protein and ligand are flexible and have many

degrees of freedom and thus the exploration of all potential relevant conformations is a considerable computational task [61]. To date, there are numerous available computational methods that differ in their accuracy, complexity and speed. In this work, we focus on scoring functions and free energy methods.

1.5.1 Scoring functions

Protein-ligand scoring functions usually take into account only one protein-ligand complex structure and do not consider the unbound state of the binding partners. They are based on the assumption that the complex conformation used during computation is the only one that is significantly occupied. Scoring functions use one complex conformation, a simplified energy model and a simplified solvent model for the purpose of computational speed. In addition, they are system dependent and thus different methods perform better with different systems [61]. The most common scoring functions are organized in three main classes: knowledge-based, empirical and force field-based scoring functions.

Knowledge-based scoring functions employ energy potentials that are derived from structural information of experimentally determined protein-ligand structures (available in databases such as the Protein Data Bank, PDB [63], and the Cambridge Structural Database, CSD [64]). The energy of a complex is calculated as the combination of energy potential terms for all pairwise contacts. Pairwise contacts are converted into energy potential, $w(i)$, by the inverse Boltzman's law:

$$w(i) = -k_b T \ln\left(\frac{\rho(i)}{\rho_{ref}}\right) \quad (1.8)$$

where k_b is the Boltzman constant, T is the absolute temperature of the system, $\rho(i)$ is a state-dependent density function and ρ_{ref} is the density function of the reference state. Knowledge-based scoring functions offer a good balance between accuracy and speed. A disadvantage is that the set of protein-ligand

structures needed to derive distance information is limited [65-67]. Examples of such scoring functions are DrugScore [68], PMF [69] and SMoG [70].

Empirical scoring functions estimate the binding affinity of a protein-ligand complex by summing up a set of weighted energy terms:

$$\Delta G = \sum_i W_i \Delta G_i \quad (1.9)$$

where ΔG_i represents diverse energy terms (vdW energy, electrostatic energy, H-bond energy, hydrophobic terms and etc.) and W_i is a weighted coefficient determined by regression analysis. The analysis uses experimental binding affinity data of a training set of protein-ligand complexes with a known 3D structure. The interesting feature of empirical scoring functions is the simple functional form. This implies that the methods are quite fast. On the other hand, the regression analysis needed to determine the weighted coefficients depends on the data set used [65-67]. GlideScore [71], X-Score [72], LUDI [73] and F-Score [74] are some examples of empirical scoring functions.

Force field-based scoring functions are based on the decomposition of the protein-ligand binding affinity into individual interaction terms (e.g. vdW interactions, electrostatic interactions, bond stretching, bond bending and torsional energies). The main feature is that such scoring functions avoid specific parameterization using a set of parameters derived by well-established force fields [65-67]. One of the most significant limitations of these methods is the exclusion of the solvent although recent implementations include models to treat the solvent implicitly or explicitly [75]. Examples of force field-based scoring functions are GOLD [76], AutoDock [77] and DOCK [78].

Each scoring function has its advantages and limitations. To take advantages of different scoring functions and to balance errors, consensus scoring functions have been introduced. They combine information from different scores [65]. An example of a consensus scoring function is X-Score.

1.5.2 Free energy methods

Free energy methods use conformational sampling to generate thermodynamic averages. The use of averages is an advantage because it removes sensitivity to the details of a single conformation, as is the case in scoring functions. On the other hand, the use of conformational sampling requires more computational time to generate converged results [61]. It is estimated that to compute the binding affinity of a single ligand with a target protein, free energy methods need about two days versus one minute with scoring function approaches [62]. Conformational sampling can be performed by molecular dynamics (MD) simulations. They are generally used to get successive conformations of a given system and the ensemble of sequential time-dependent conformations, called a MD trajectory, is used to calculate diverse properties of the system of interest. The potential energy of a studied system is calculated using a force field that is given by a functional form and a set of parameters (that can be derived from experimental works or quantum mechanical calculations). Given a force field and the initial position of the system of interest, it is possible to perform a MD simulation and calculate the trajectory. A commonly used force field is AMBER [79, 80] whose functional form is:

$$V(r) = \sum_{bonds} K_b (b - b_0)^2 + \sum_{angles} K_\theta (\theta - \theta_0)^2 + \sum_{torsions} \frac{1}{2} K_\phi [1 + \cos(n\phi - \delta)] + \sum_{i>j} \sum_{i>j} \left(4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \right) \quad (1.10)$$

The equation 1.10 represents the potential energy, V, as function of the system structure r. It is separated into the internal terms, including bond, angle and torsion contributions, and the nonbonded terms that include vdW and electrostatic terms. The parameters b_0 and θ_0 represent equilibrium bond and angle terms, n is the periodicity of the dihedral term and δ its phase. The parameters b, θ and ϕ are bonds, angles and dihedrals that define the structure r. K_b, K_θ and K_ϕ are bond, angle and dihedral force constants. The nonbonded terms are the vdW term described by the Lennard-Jones potential and the electrostatic term

described by Coulomb's law. Another commonly used force field the Optimized Potentials for Liquid Simulations, OPLS [81, 82]. Its functional form is based on AMBER and the nonbonded interaction parameters have been developed from extensive Monte Carlo (MC) simulations of small molecules whereas in AMBER they come from experimental data. Thus, OPLS is expected to better describe a system were nonbonded interactions are particularly important (such as a protein-ligand complex) [83].

Free energy methods are organized into two classes, 'alchemical' methods and 'end-point' methods.

'Alchemical' methods employ unphysical ('alchemical') transformations to estimate the free energies of several physical processes such as protein-ligand binding. In the case of a ligand that binds to a protein, the alchemical transformation is the conversion of a ligand into another, unphysical ligand, within the binding site and in solution [61]. Free energy perturbation (FEP) and thermodynamic integration (TI) are two 'alchemical' free energy methods that use a thermodynamic cycle like that one represented in Figure 1.12.

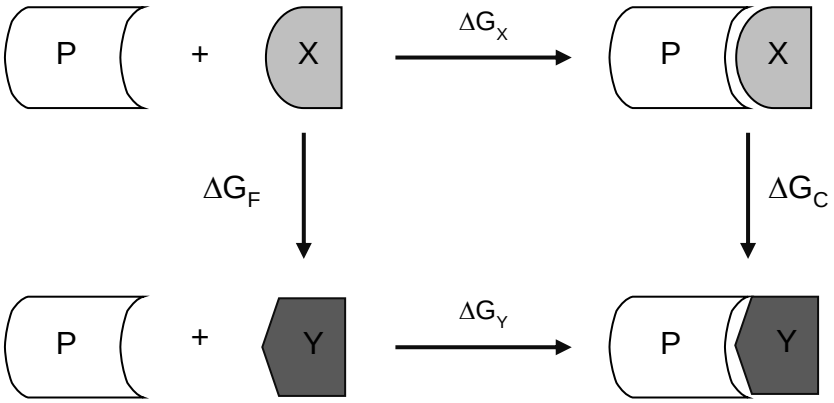

Figure 1.12. Schematic representation of the thermodynamic cycle used by FEP and TI methods, from [62]. P is the protein, X and Y are ligands. The ΔG_X and ΔG_Y are the change in free energy for the formation of the complexes PX and PY, respectively. ΔG_F and ΔG_C are the change in free energy for the transformation of ligand X into Y in solution and within the protein binding site, respectively.

The thermodynamic cycle in Figure 1.12 can be written as:

$$\Delta G_X + \Delta G_Y - \Delta G_Y - \Delta G_F = 0 \quad (1.11)$$

ΔG_Y is the change in free energy when the unphysical ligand Y binds to the protein. Since an unphysical ligand is not able to interact with the solvent or the protein, this quantity is always equal to zero. Equation 1.11 becomes:

$$\Delta G_X = \Delta G_F - \Delta G_C \quad (1.12)$$

Therefore, the free energy of binding of the ligand X to the protein P is given by the difference of the change in free energy for the transformation of the ligand X into the ligand Y in solution and within the binding site of the protein P.

FEP and TI employ long MD or MC simulations and an explicit treatment of the solvent. They are time consuming methods that give a good estimation of binding energy, with errors of about 1 to 2 kcal/mol [61]. A limitation of such methods is that the alchemical transformation cannot be too drastic. This restricts the diversity of the ligands that can be treated and also the possibility of examining the effect of significant protein mutations on the binding of given ligands since usually protein mutations are considered large perturbations [84].

'End-point' methods compute the change in free energy only considering the initial and finale states of a given process [61]. A first 'end-point' method is the linear interaction energy (LIE) method. In the case of protein-ligand binding it involves running two MD simulations, one for the ligand in solution and another for the ligand within the protein binding site [85]. An ensemble of conformations obtained for the initial and the final states is used to compute the average electrostatic (E_{elec}) and average vdW (E_{vdW}) interaction energies of the ligand within its environment in the initial state, the free state, and in the bound state. The free energy of binding is estimated as follows:

$$\Delta G_{bind} \approx \alpha\left(\langle E_{elec}\rangle_{bound} - \langle E_{elec}\rangle_{free}\right) + \beta\left(\langle E_{vdW}\rangle_{bound} - \langle E_{vdW}\rangle_{free}\right) \quad (1.13)$$

where the angle brackets indicate the averages for the interaction energy terms. The factor α and β accounts for changes in the internal energy of the solvent and the protein in response to the interaction with the ligand and they are determined empirically [85]. A drawback of LIE is that it is not really universal since α and β are system dependent. They have to be determined for each case of study and requires available experimental data [84]. A second 'end-point' method is the molecular mechanics-Poisson Boltzmann surface area (MM-PBSA) method together with its generalized Born (GB) variant (MM-GBSA). Generally, those methods use MD simulations of the free protein, the free ligand and their complex (three-trajectory approach) to obtain conformation ensembles that, in turn, are used to compute the average energy terms that contribute to the free energy of binding [60, 61]. The solvent is treated implicitly using either PBSA [86] or GBSA [87] models. The nonpolar contribution is assumed to be proportional to the solvent accessible surface area (SASA) and the electrostatic contribution is given by the continuum-electrostatics models PB or GB (that is an approximation of the exact PB equation). Figure 1.13 shows the thermodynamics cycle used in MM-PBSA and MM-GBSA methods.

The free energy of binding is given by the following formula:

$$\Delta G_{bind,solv} = \Delta G_{comp,solv} - \left[\Delta G_{prot,solv} - \Delta G_{lig,solv}\right] \quad (1.13).$$

The MD simulation can also be performed using a protocol known as single-trajectory approach. In that case, a MD simulation is performed only for the protein-ligand complex and results converge faster than using the three-trajectory approach [61].

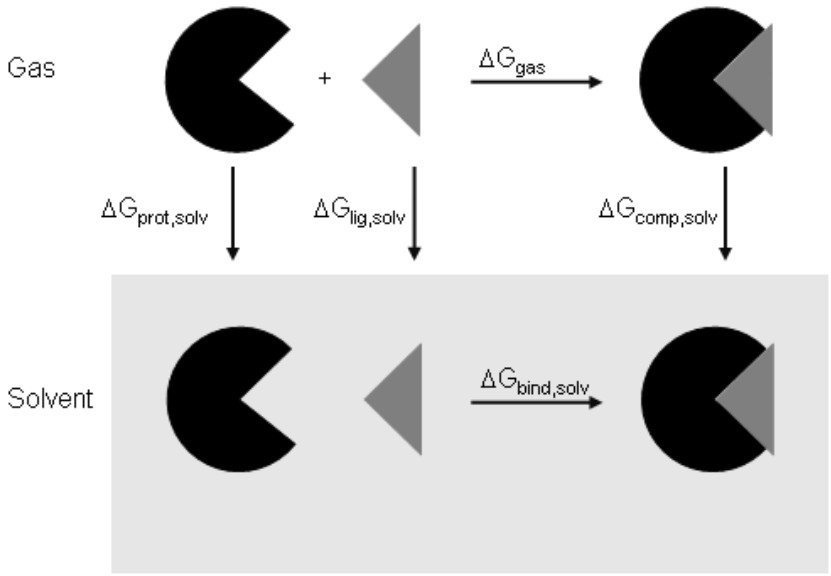

Figure 1.13. Schematic representation of the thermodynamic cycle used for MM-PBSA and MM-GBSA calculations. ΔG_{gas} and $\Delta G_{bind,solv}$ are the free energy of binding for the formation of the complex *in vacuum* and in solution. $\Delta G_{prot,sol}$, $\Delta G_{lig,solv}$ and $\Delta_{Gcomp,solv}$ are ΔGs for the transition of the protein, the ligand and complex from gas to the solvent.

1.6 *Mycobacterium tuberculosis* protein kinase G, an attractive target

We applied the computational protocol developed in this work to a specific kinase, the *Mycobacterium tuberculosis* (*Mtb*) serine/threonine protein kinase G (PknG). *Mtb* is a pathogenic bacterium and is the causative agent of tuberculosis (TB). TB is a widespread infectious disease and causes around two million deaths per year [88]. It generally affects the lungs and is spread through the air when infected people cough or sneeze [89]. Nowadays there is a global increase in drug-resistant TB cases and therefore there is an urgent need to develop new therapies to combat this infectious disease [90].

Mtb belongs to the family of Mycobacteriaceae. The virulence of *Mtb* is related to its capacity to survive within the host alveolar macrophage. In general the first barrier pathogens come across when infecting a multicellular organism is the

immune defense system. A key cell of the immune system is the macrophage that is a phagocyte which recognizes microbes and engulfs them into vacuoles called phagosomes. Phagosomes then fuse with lysosomes, called phagolysosome biogenesis, resulting in the degradation of pathogens (Figure 1.14, right side). In the case of *Mtb*, the mycobacteria are picked-up by macrophage but they survive and replicate intracellularly causing the infection. Diverse studies have shown that, upon *Mtb* infection, the secretion of PknG within the macrophage cytosol prevent the phagolysosome biogenesis promoting the survival of the parasite within the host organism [91, 92] (Figure 1.14, left side).

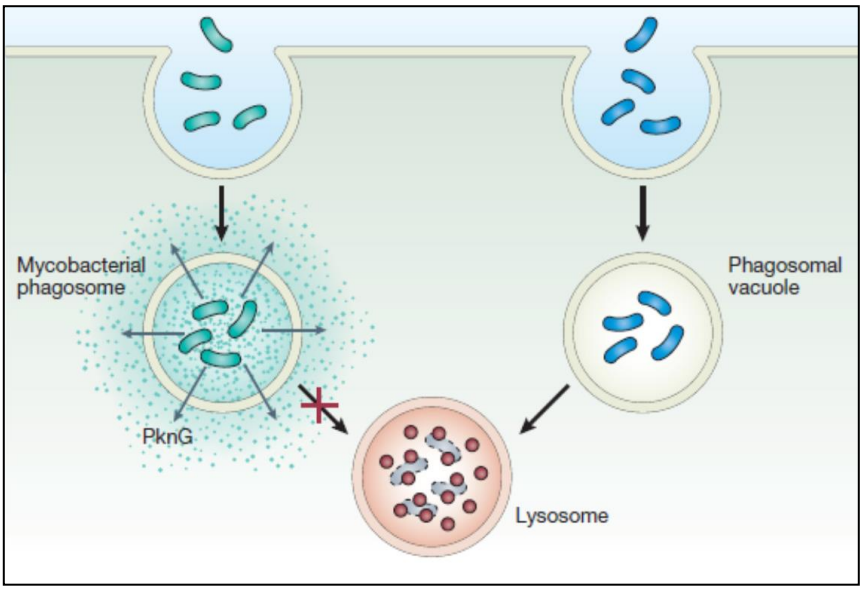

Figure 1.14. (From [93]) Immune system response upon pathogenic mycobacteria infection. Blue mycobacteria (on the right side) represent generic pathogens, green mycobacteria (on the left side) represent *Mtb*.

Mtb genome includes genes encoding 11 serine/threonine protein kinases, PknA, PknB, PknD, PknE, PknF, PknG, PknH, PknI, PknJ, PknK and PknL. Except for PknG and PknK, which are soluble proteins, all other serine/threonine kinases are transmembrane proteins [94, 95]. Inactivation of PknG gene resulted in decreased viability of *Mtb* both *in vitro* and *in vivo* (mice) [96], and the blocking of PknG kinase activity by tetrahydrobenzothophene (AXX) results in bacterium degradation [97]. Although experimental evidence supports the significant role of

PknG in mycobacterium survival, its downstream substrates involved in pathway mediating infectivity and its precise mode of action remain unknown.

The PknG kinase domain has the classical bilobal fold, with the ATP binding site located within the cleft between the N-lobe and the C-lobe. The kinase domain is sandwiched between the rubredoxin (Rdx) domain, probably involved in PknG activity regulation, and a tetratricopeptide repeat (TPR) domain, probably involved in mediating protein-protein interactions [97] (Figure 1.15).

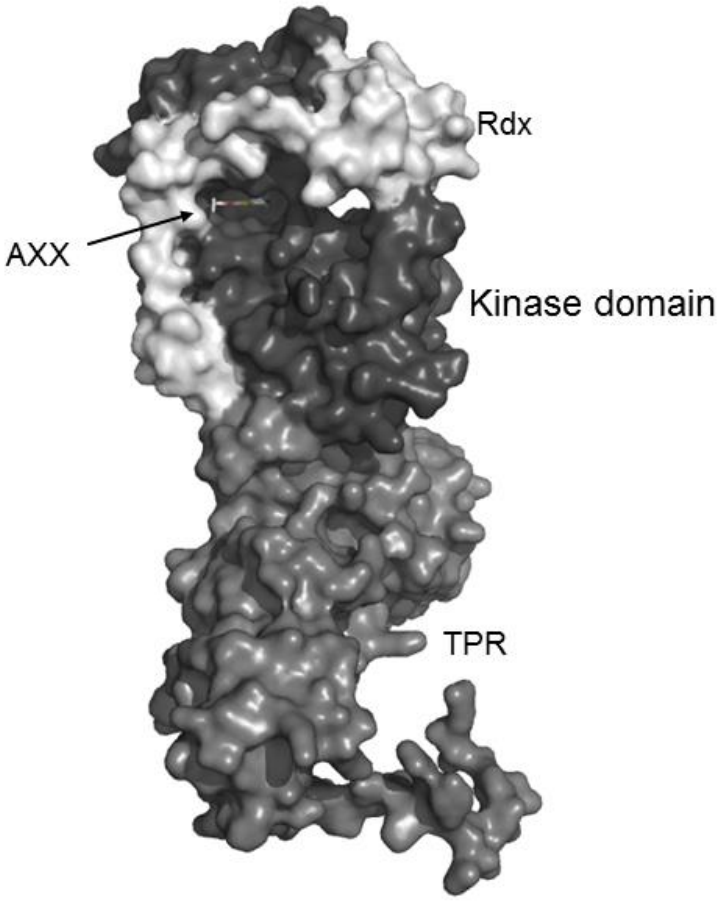

Figure 1.15. Surface representation of PknG (PDB: 2PZI). The kinase domain is sandwiched between the Rdx and the TPR domains. AXX occupies the ATP binding cleft.

To date, it is known that GarA is a physiological substrate of PknG in *Mtb* [98]. GarA is a forkhead associated (FHA) protein of 162 residues. The FHA domain folds into an 11-stranded β sandwich [99] and is preceded by an N-terminal

peptide extension of about 50 residues (Figure 1.16). The N-terminal peptide contains a highly conserved ETTS motif which, in turn, contains the residue phosphorylated by PknG, Thr21 (Figure 1.16). GarA controls glycogen degradation and glutamate metabolism [100, 101].

Figure 1.6 Schematic representation of GarA. The motif ETTS contains the phosphorylation site, Thr21 (red).

1.7 Objectives

This work aims to develop a computational protocol for the engineering of protein kinases. We intend to predict which residues within the binding pocket of the target kinase could be mutated to change its co-substrate specificity from ATP to an ATP analogue. The protocol explores pairings of potential mutations and ATP analogues and might be part of a wider experiment for identifying specific substrates of protein kinases and better understanding role of these enzymes in cell pathways.

We validated the computational protocol on different tyrosine and serine/threonine protein kinases from the scientific literature where Shokat's method was applied and experimental data were available. Subsequently, we applied our protocol to the *Mycobacterium tuberculosis* protein kinase G, PknG. PknG plays a key role in the survival of *M.tuberculosis* within the host organism and its specific downstream substrates as well as its mechanism of action are unknown. The *in silico* protocol allowed us to design a number of pairs of PknG mutants and ATP analogues and the designed pairs were tested *in vitro*.

2 Methods

This chapter describes the method developed and used in this thesis. Figure 2.1 shows a schematic representation of the computational protocol that is organized in two main parts. The first part is an algorithm to predict residues to mutate within the ligand-binding site of a kinase of interest to generate an engineered kinase. The second part consists in the evaluation of the interaction between an engineered kinase and a ligand analogue.

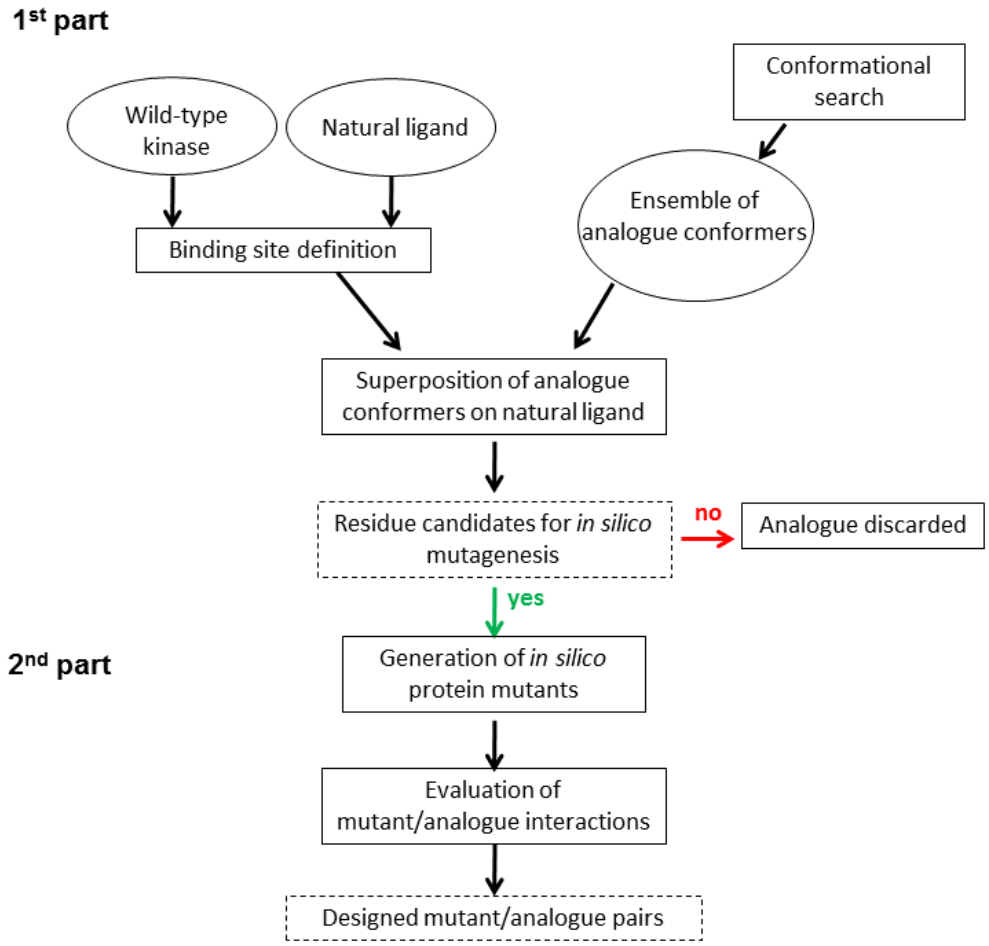

Figure 2.1. Workflow of the computational protocol. The entire protocol is organized in two parts, the first part identifies residues to mutate and the 2nd part evaluates mutant-analogue interactions. The specific inputs are depicted in circles, steps of the workflow are shown in rectangles and outputs are depicted in rectangles with dashed lines. In case no residues are identified for the *in silico* mutagenesis, the analogue is assumed to act as substrate for the wild-type protein and thus is discarded (red arrow).

The protocol was tested on a literature-based test set containing 7 wild-type kinase proteins and 15 kinase mutants to which the Shokat's method was applied and for which experimental data were available. Afterwards, the protocol was applied to the *Mycobacterium tuberculosis* protein kinase G.

2.1 Input structures

We collect the kinase structures as well as structures of natural ligands to use as input structures. The kinases structures are X-ray structures from the Protein Data Bank (PDB, [63]) and natural ligands come from known PDB structures. Table 2.1 shows wild-type kinases, PDB entries, engineered kinases and natural ligands used in this work.

Table 1.1. Kinase proteins, kinase mutants, natural ligands and PDB entries used in our work.

Kinases	**PDB**	**Kinase mutants**	**Natural ligands**
v-Src [18]	2SRC*****	v-SrcI338A v-SrcI338G	ANP
JNK [27]	1JNK	JNKM108GL168A	ANP
v-Src [102]	2SRC*****	v-SrcI338A v-SrcI338G v-SrcI338F v-SrcI338M v-SrcI338S v-SrcI338T v-SrcI338V v-SrcI338C	PP1 pyrazolepirimidine core
Fyn	2DQ7	FynT338A	
Abl	2G1T	AblT334A	
CamKII	2VZ6	CamKIIF89G	
Cdk2	1HCK	Cdk2F80G	
P38	1DI9	P38T106A P38T106G	
PknG	2PZI	PknGM232G, PknGV211G PknGM232H, PknGM232S PknGM232T, PknGV235G PknGY234G	ATP

*****A model for v-Src was built based on the structure of c-Src whose PDB is 2SRC.

Unless stated otherwise, *in silico* mutagenesis was performed using Maestro (version 9.5, Schrödinger, LLC, New York, NY, 2013) and structures were prepared with the Protein Preparation Wizard tool [103]. Residues are numbered as in PDB structures.

To date, the crystal structure of the cellular proto-oncogene tyrosine-protein kinase Src (c-Src) in complex with ANP (an ATP analogue with an amino group in place of the oxygen between β and γ phosphates) has been solved (*Homo sapiens*, PDB: 2SRC, resolution 1.50 Å [45]). The kinase domain of the viral proto-oncogene tyrosine-protein kinase Src (v-Src) differs from that of c-Src at the position 338 within the ligand-binding pocket. It is an Ile (Ile338) in v-Src and a Thr (Thr338) in c-Src. The rest of the sequences are identical. To obtain the model of v-Src bound to ANP, we mutated *in silico* Thr338 of c-Src into Ile. The model of v-Src in complex with ANP was then prepared as follows: first, we added hydrogen atoms that are generally not visible in X-ray structures [104]. Then, we optimized the hydrogen bonding network and the orientation of the amide groups of Asn, Gln and of the imidazole ring of His. This optimization allowed for improved interactions between charged groups as well as hydrogen bonds within the structure. The optimization was performed at pH 7.0. Finally, a minimization step allowed the model to relax. The Optimized Potentials for Liquid Simulations (OPLS_2005) [81, 82] was used as force field and the termination criterion was based on the root mean square deviation (rmsd) of the heavy atoms relative to their initial location (rmsd ≤ 0.30 Å). The v-SrcI338A and v-SrcI338G mutants were obtained in the same way.

The crystal structure of JNK bound to ANP and Mg^{2+} was solved in 1998 (*Homo sapiens*, PDB: 1JNK, resolution 2.30 Å [105]). The M108GL168A mutant in complex with ANP was obtained by mutating *in silico* Met108 to Gly and Leu168 to Ala. The structure of wild-type JNK in complex with ANP and the model of the engineered JNK in complex with ANP were prepared as described above.

The pyrazolopyrimidine core of PP1 (1-tert-butyl-3-(4-methylphenyl)-1H-pyrazolo[3,4-d]pyrimidin-4-amine) mimics the adenine ring of ATP in its binding within the nucleotide-binding pocket of a kinase protein (Figure 2.2). To obtain a model of v-Src in complex with the pyrazolopyrimidine core of PP1, we proceeded

in the following way. The model of v-Src bound to ANP was superposed onto the structure of the hematopoietic cell kinase Hck, a homologous protein, in complex with PP1 (*Homo sapiens*, PDB: 1QCF, resolution 2.00 Å [106]). The superposition was performed using residues belonging to the hinge regions (residues 338 to 341 in both v-Src and Hck) and the coordinates of the PP1 core were copied into the v-Src binding pocket. The same procedure was used for the other protein kinases, proto-oncogene c-Fyn (Fyn, *Homo sapiens*, PDB: 2DQ7, resolution 2.80 Å [107]), Abelson murine leukemia viral oncogene homolog 1 (Abl, *Homo sapiens*, PDB: 2G1T, chain D, resolution 1.80 Å [108]), Calcium/calmodulin-dependent protein kinase type II subunit alpha (CamKII, *Homo sapiens*, PDB: 2VZ6, chain B, resolution 2.30 Å [109]), Cyclin-dependent kinase 2 (Cdk2, *Homo sapiens*, PDB: 1HCK, resolution 1.90 Å [110]), and Mitogen-activated protein kinase p38 alpha (P38, *Homo sapiens*, PDB: 1DI9, resolution 2.60 Å [111]).

The structure of the *Mycobacterium tuberculosis* protein kinase G, Mtb PknG, bound to AXX (2-[(cyclopropylcarbonyl)amino]-4,5,6,7-tetrahydro-1-benzothiophene-3-carboxamide) had been solved (PDB: 2PZI, resolution 2.40 Å [97]). The kinase domain of the PknG structure (residue 151 to 396) contained 2 gaps, 1 residue missing between amino acid 304 and 306 and 4 residues missing in the C-lobe between position 241 and 246. To build the missing parts, we built a homology model of PknG using Modeller 9.10 [112]. Two proteins were used as templates, PknG itself and the homologous serine/threonine protein kinase B (PknB) from the same organism (28.40% sequence identity, PDB: 3ORO, resolution 1.90 Å [113]). The second structure was used as template for the missing parts of PknG. The obtained model and the crystal structure of PknG were superposed using backbone atoms (rmsd 0.017 Å) and the gaps within the crystal structure were replaced by the corresponding parts in the model. The quality of the model of PknG was estimated by the QMEAN scoring function (good model with a QMEAN score = 0.727, [114]). Once the structure of PknG was fixed, it was superposed onto the structure of PknB in complex with AGS (an ATP analogue with a sulfur atom bound to γ phosphate). Superposition was made using hinge regions and the AGS coordinates were copied within the PknG ligand-binding site. All PknG mutants were prepared as described before.

Ligand analogues used in this study (N6-(benzyl) ATP, N6-(2-phenythyl) ATP, N6-(cyclopentyl) ATP, N6-(1-methylbutyl) ATP, PP1, N6-(methyl) ATP, 8-Azido ATP, 2-(phenylethynyl) ATP (PF9) and 7-deaza-7propargylamino-ATP (7d7p ATP)) are created in the Maestro graphical interface and are represented in Figure 2.2.

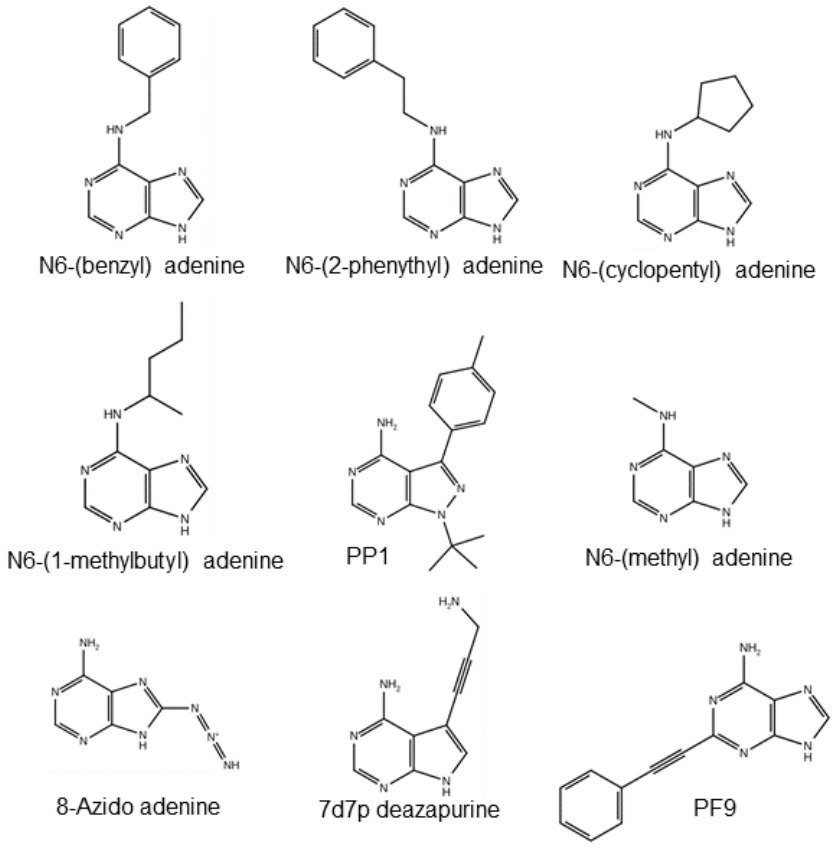

Figure 2.2. Chemical structures of ATP analogues used in this work. For simplicity, only the adenine base is represented.

For each molecule, an ensemble of low energy conformers was generated by performing an *in vacuum* conformational search keeping the adenine base, the ribose ring, the phosphates and the pyrazolopyrimidine core of PP1 fixed and allowing the bonds of each substituent group to rotate freely. We used the Monte Carlo multiple minimum (MCMM) method [115] for 10 000 steps, the OPLS_2005 force field, and a threshold value of 100 Kj/mol. During the conformational search,

new structures generated were retained if they exhibited conformational energies lower than 100 kJ/mol. Moreover, to obtain an ensemble of unique structures and eliminate redundant conformers, each conformer was compared with the previous ones and only retained if the root mean square deviation (rmsd) for all atoms exceeds 0.5 Å. The conformational search was performed with the MacroModel module implemented in the Schrödinger suite (version 10.1, Schrödinger, LLC, New York, NY, 2013).

2.2 Computational protocol

The first part of the protocol (Figure 2.1) was written in Python 2.5.4 and contained functions from the OpenStructure software framework [116]. The structures of target kinases and the natural ligands were used to define the ligand-binding site. The pocket was defined by all residues within 5 Å from the atom position on the natural ligand at which the substituent group was attached. For example, if the natural ligand was ATP and the ligand analogue was N6-(methyl) ATP, the N6 ATP atom defined the binding pocket of the target protein kinase. For each analogue, the ensemble was superposed onto the adenine moiety of the native ligand within the binding pocket of the reference protein. The identification of residues to mutate was based on a distance criterion. If the distance between an atom of a protein residue and any atom of the substituent group of a ligand analogue is shorter than the sum of their van der Waals (vdW) radii, the corresponding residue is a potential candidate for mutagenesis. The vdW radii for each atom come from the Cambridge Structural Database, CSD [64] (Appendix A.1). If no residues were identified for point mutations it means that the analogue could potentially act as substrate for the native target and not only for the engineered protein and is thus discarded.

One of the most important contributions to the stabilization of protein-ligand complexes is given by hydrogen bonds (H-bonds). In case of non-halogenated ligands, two atoms are generally involved in protein-ligand H-bonds, nitrogen (N) and oxygen (O). When N and O are involved in H-bonds, the distance between

them is smaller than the sum of their vdW radii. For instance, if N acts as hydrogen donor the N-O distance becomes 3.04 Å instead of 3.07 Å (sum of their vdW radii) and if O acts as H-bond donor the O-O distance becomes 2.70 Å instead of 3.04 Å and O-N becomes 2.88 Å instead of 3.07 Å [30, 117]. To take into account the distances between atoms that could be involved in H-bonds, we introduced a tolerance value. The tolerance value was calculated as the difference between the sum of the vdW radii of two atoms and their distance in case they are involved in H-bonds. For example, in case of the O-O distance the tolerance value is given by 3.04 Å (sum of vdW radii) minus 2.70 Å (donor-acceptor distance in case of H-bond). The tolerance values were computed for all atom pairs that can be involved in protein-ligand H-bonds and a value of 0.5 Å was chosen as general tolerance value (since it included all possible calculated values). The distance criterion was thus given by the formula:

$$d < sumvdW - tv \qquad (2.1)$$

If the distance d between a protein-ligand atom pair is smaller than the sum of their vdW radii minus the tolerance value, tv, the corresponding residue is a potential candidate for mutagenesis.

In the second part of the computational protocol, residues identified as possible candidates for mutagenesis were mutated *in silico* to generate kinase mutants (Table 2.1). The kinase mutant-ligand conformer pairs were evaluated and ranked by the empirical protein-ligand scoring function GlideScore as implemented in the Scrödinger suite [71] (Figure 2.1). GlideScore is given by equation 2.2:

$$GScore = Lipo + Hbond + Metal + RotB + Site + Coul + vdW + BuryP \quad (2.2)$$

$Lipo$ is the lipophilic term and rewards favorable receptor-ligand hydrophobic interactions. It is a function of the receptor-atom–ligand-atom distances:

$$Lipo = C_{lipo} \sum f\left(r_{lr}\right) \quad (2.3)$$

$Hbond$ represents the hydrogen bonds term. It is separated into three weighted components that depend on whether the donor and the acceptor are both neutral, one neutral and one charged, or both charged. It depends on hydrogen bonding distances and angles:

$$Hbond = C_{hbond} \sum g(\Delta r) h(\Delta \alpha) \quad (2.4)$$

where r is the distance between the donor and acceptor and α is the donor-acceptor angle. $Metal$ is the metal term and describes the interaction between ligand atoms and metal ions eventually present in the receptor:

$$Metal = C_{metal} \sum f(r_{lr}) \quad (2.5)$$

$RotB$ is a penalty term for frozen rotable bonds. Bonds are considered frozen if atoms on both sides of a rotable bond of the ligand (any sp^3-sp^3 and any sp^2-sp^3 bond) are in contact with receptor. $Site$ is a term that accounts for polar interactions. It rewards for polar group found in hydrophilic regions. $Coul$ and vdW account for Coulomb energy and van der Waals energy, respectively. $Coul$ is described by the Coulomb's law and vdW by the Lennard-Jones potential (equations 2.6 and 2.7).

$$Coul = \sum^{ligand} \sum^{receptor} k \frac{q_l q_r}{r_{lr}} \quad (2.6)$$

$$vdW = \sum^{ligand} \sum^{receptor} \left[\left(\frac{r_{lr,0}}{r_{lr}} \right)^{12} - 2 * \left(\frac{r_{lr,0}}{r_{lr}} \right)^{6} \right] \quad (2.7)$$

The quantities q_l and q_r are the charges of ligand atoms and receptor atoms, k is the Coulomb's constant, $r_{lr,0}$ is the ligand-receptor distance at which the potential reaches its minimum and r_{lr} is the ligand-receptor distance. Both terms are computed with reduced net ionic charges on groups with formal charges (such

as acetate or guanidinium) to avoid an over reward of charge-charge interactions with respect to charge-dipole and dipole-dipole interactions. Finally, $BuryP$ is a penalty term for buried polar groups.

All terms in equation 2.2 are multiplied by coefficients determined by reproducing binding affinities of a training set of protein-ligand complexes with known three-dimensional structures. The $Lipo$ and vdW terms account for nearly 80% of the total score. The most promising pair of kinase mutant and analogue conformer was the one with the best GlideScore, and for this the Glide Energy was taken into account. The Glide energy is the sum of the $Coul$ and the vdW terms and represents a protein-ligand interaction energy. The correlation between experimental IC_{50} values and predicting energies of interaction (Glide energies) shows that the Glide energy is more suitable to describe and compare protein-ligand binding affinities [118].

For all pairs of kinase mutants and ATP analogues, to be independent from the various ribose ring and phosphates conformations, position and orientations within the binding site (see Introduction, paragraph 1.3), only the adenine base and the substituent group were scored by GlideScore and not the rest of the molecule (Appendix A.2).

2.3 Methods to score protein-ligand interactions

Besides GlideScore, we evaluated other two protein-ligand scoring functions, X-Score [72] and DrugScore eXtended (DSX, [68]). We also tested the Molecular Mechanics-Generalized Born Surface Area, MM-GBSA, method [119]). We checked the ability of the scoring functions and the free energy method to reproduce experimental data available for v-Src and an ATP derivative.

2.3.1 X-Score

X-Score is an empirical consensus protein-ligand scoring function. It is the arithmetical average of three empirical scoring functions:

$$X-Score=\left(HCScore+HMScore+HSScore\right)/3 \quad (2.8)$$

All terms in equation 2.8 are calculated summing up five terms that account for protein-ligand interactions:

$$HCScore=VDW+HB+RT+HC \quad (2.9)$$

$$HMScore=VDW+HB+RT+HM \quad (2.10)$$

$$HSScore=VDW+HB+RT+HS \quad (2.11)$$

All quantities in these three scoring functions are calculated using identical algorithms except for the hydrophobic effect terms (HC, HM and HS) that are calculated by three different algorithms. Each term in equations 2.9-2.11 is multiplied by a coefficient determined by fitting experimental binding affinities of a dataset of protein-ligand complexes. VDW denotes the van der Waals interaction energy and it is described by a 'softer' Lennard-Jones potential, an 8-4 equation instead of the standard 12-6 equation:

$$VDW=\sum_{i}^{ligand}\sum_{j}^{protein}\left[\left(\frac{d_{ij,0}}{d_{ij}}\right)^{8}-2*\left(\frac{d_{ij,0}}{d_{ij}}\right)^{4}\right] \quad (2.12)$$

The term d_{ij} denotes the distance between the ligand atom i and the protein atom j and $d_{ij,0}$ is the sum of the vdW radii of atom i and j. HB accounts for hydrogen bonding interaction between the ligand and the protein and it is calculated by summing up all hydrogen bonds:

$$HB=\sum_{i}^{ligand}\sum_{j}^{protein}f\left(d_{ij}\right)*f\left(\theta_{1,ij}\right)*f\left(\theta_{2,ij}\right) \quad (2.13)$$

The distance function $f(d_{ij})$ and the angular functions $f(\theta_{1,ij})$ and $f(\theta_{2,ij})$ are the geometric descriptors of each ligand-protein hydrogen bond. d_{ij} is the distance between the donor (D) and the acceptor (A), $\theta_{1,ij}$ is the donor angle and $\theta_{2,ij}$ is the acceptor angle. RT represents the deformation effect, that is the loss of entropy when the ligand and the protein are in a bound state with respect to their free states in the solvent. The deformation effect of the ligand is estimated by considering all rotable bonds (RT) that became frozen after the binding whereas the deformation effect of the protein is neglected:

$$RT = \sum_{i}^{ligand} RT_i \quad (2.14)$$

Finally, as previously mentioned, the hydrophobic effect is represented by three different algorithms. HC stands for hydrophobic contact algorithm, and this effect is calculated by summing up the hydrophobic atom pairs between the ligand and the protein:

$$HC = \sum_{i}^{ligand} \sum_{j}^{protein} f(d_{ij}) \quad (2.15)$$

HM stands for hydrophobic matching algorithm. This term describes as favorable to the binding process each hydrophobic ligand atom placed within a hydrophobic place of the protein:

$$HM = \sum_{i}^{ligand} HM_i \quad (2.16)$$

HM_i is equal to 1 when a hydrophobic atom is placed in a hydrophobic environment and is equal to 0 when it is placed in a hydrophilic environment.

HS stands for hydrophobic surface algorithm. The hydrophobic effect is assumed to be proportional to the solvent-accessible surface (SAS) area of the ligand:

$$HS = \sum_{i}^{ligand} SAS_i \quad (2.17)$$

2.3.2 DrugScore eXtended

The knowledge-based scoring function DrugScore eXtended (DSX) consists of distance-dependent pair potentials, torsional potentials and solvent accessible surface-dependent potentials. It is given by the equation 2.18:

$$score_{total} = score_{pair} + score_{tors} + score_{SR} \quad (2.18)$$

Each term $score$ is the classical equation for knowledge-based scoring functions:

$$score(i) = \ln\left(\frac{\rho(i)}{\rho_{ref}}\right) \quad (2.19)$$

where $\rho(i)$ is a state-dependent density function and ρ_{ref} is the density function of the reference state. The term $score_{pair}$ represents the distance-dependent pair potential and is a function of the distance between a protein atom and a ligand atom. The knowledge-bases used to derive the DSX pair potential are the PDB and CSD. In the PDB, only structures with a resolution less than 2.4 Å and containing at least one ligand are taken into account. Moreover, only contacts between atoms with B-factors ≤ 40 Å and occupancies ≥ 0.5 were considered. In the CSD, only structures with R-factors ≤ 0.075 are considered. The term $score_{tors}$ is the DSX torsional potential. It is a function of the four atoms being part of the torsion and of the torsion angle. This term aims to penalize unlikely torsion angles. In that case, the knowledge-base is the CSD already used for $score_{pair}$. Finally, the term $score_{SR}$ accounts for the desolvation effect that is based on the solvent-accessible surface (SAS). It is a function of the SAS ratio, represented by the ration of SAS for an atom in complexed state over SAS for an atom in uncomplexed state. The PDB is used as the knowledge-base.

2.3.3 MM-GBSA method

The MM-GBSA approach allows for computing the free energy of binding of protein-ligand complexes. It requires, at least, one trajectory for the complex of interest that is produced by a molecular dynamics (MD) simulation. The MD simulations were performed using AMBER 12 [120] and the AMBER ff99SB all-atom force field [79, 80]. The starting coordinates for MD simulations were the coordinates of the protein-ligand complexes previously described. The ligands were automatically parameterized using Antechamber [121]. Each protein-ligand complex had a negative charge that was neutralized by adding the proper number of sodium ions. Each of the neutralized protein-ligand complexes was solvated with a rectangular box of TIP3P water molecules.

Prior to running the production MD, we performed minimization, heating, and equilibration steps described as follows. First of all, water molecules and ions were minimized with the steepest descent method (SD) for 2500 steps keeping the protein-ligand complex fixed (partial minimization) and then the entire system was minimized for other 1000 steps (full minimization). In both partial and full minimizations, periodic boundaries based on the Partial Mesh Ewald (PME) [122] method were used for calculation of non-bonded interactions, with a cutoff of 9 Å. Afterwards, the system was heated from 0 K to 300 K for 50 ps keeping the protein-ligand complex fixed and using the NVT ensemble. All bonds involving hydrogens were constrained using the SHAKE algorithm [123]. Finally, the system was equilibrated for 500 ps switching from the constant volume of the heating step to a constant pressure (NPT ensemble). This allowed the solvent to equilibrate around the protein-ligand complex. After the minimization, heating, and equilibration steps had completed, a long MD simulation was carried out using NVT ensemble at a temperature of 300 K for 20 ns with a time step of 2 fs.

The free energy of binding was calculated using the MM-GBSA approach [124]. The binding free energy is calculated by subtracting the free energy of the unbound protein and ligand from the free energy of the protein-ligand complex (equation 2.20):

$$\Delta G_{bound,solvated} = \Delta G_{complex,solvated} - \left[\Delta G_{protein,solvated} - \Delta G_{ligand,solvated}\right] \quad (2.20)$$

Each term of the right-hand side of equation 2.20 is estimated as follows:

$$\Delta G_{solvated} = \langle \Delta E_{gas} \rangle + \langle \Delta G_{solvation} \rangle - \langle T\Delta S \rangle \quad (2.21)$$

The brackets indicate averages of each term (ΔE_{gas}, $\Delta G_{solvation}$ and ΔS) calculated from an ensemble of representative complex structures obtained from the MD trajectory.

The gas-phase energy (ΔE_{gas}) is the molecular mechanical (MM) energy from the force field. It comprises internal energy terms (bond, angle and torsion energies) and the van der Waals and the electrostatic interaction energies between the protein and the ligand. Here, we used the single trajectory protocol (STP) approach that means we only used the trajectory of the protein-ligand complex to generate three ensembles of representative frames, one for the unbound ligand, one for the unbound protein and one for the complex [125]. A STP approach does not produce any difference between the internal energy terms of the complex and the unbound protein and ligand. Thus, those terms cancel out.

The solvation free energy ($\Delta G_{solvation}$) is divided into electrostatic and nonpolar contributions. The electrostatic contribution is given by a continuum (implicit) solvent model and here we used the generalized Born (GB) model [126, 127]. The use of the GB model allows for an increase in the calculation speed compared to numerically solving the Poisson equation when the electrostatic contribution is estimated by the Poisson-Boltzman approach [128]. The nonpolar contribution is proportional to the solvent accessible surface area (SASA) [129].

The entropic contribution ($T\Delta S$) is given by three terms, the translational, the rotational, and the vibrational entropies. The entropic contribution is required to calculate absolute free energies of binding. In case of related systems, such as two ligands binding to the same protein or vice versa, it is assumed that the solute entropy is the same for each system and thus the entropy contribution can be neglected. That approximation was applied in our case, allowing the computation

of the relative binding free energies [130]. Thus, in this work the equation 2.21 became:

$$\Delta G_{solvated} = \langle \Delta E_{vdw} \rangle + \langle \Delta E_{electrostatic} \rangle + \langle \Delta G_{solvation,elec} \rangle + \langle \Delta G_{solvation,nonpolar} \rangle \quad (2.22)$$

In that case, frames from the long MD simulation (20 ns) were extracted every 10 ps resulting in 2000 frames used to apply the MM-GBSA method already described. We used the program MMPBSA.py [119].

2.4 Data comparison

All plots reported in this work were made using the Matplotlib and NumPy packages [131] [132].

The plots of v-Src, v-SrcI338A and v-SrcI338G in complex with ATP and N6-(benzyl) ATP were created by comparing the experimental catalytic efficiency (kcat/Km) and the predicted interaction energies (Glide energies). To compare experimental kcat/Km with computed interaction energies, we calculated the logarithm to base 10 (log) of the kcat/Km ratio, in order to use the corresponding small integer values, and multiplied the predicted interaction energies by -1, to obtain positive values.

For the engineered JNK (JNKM108GL168A) and the N6-(substituent) ATPs, we compared the observed percentage of phosphorylation with the computed interaction energies. The predicted quantities were scaled between 0 and 100 to fit the same range of observed values. Predicted data were transformed using the Min-Max normalization:

$$v' = \frac{v - \min_A}{\max_A - \min_A}(new\_\max_A - new\_\min_A) + new\_\min_A \quad (2.23)$$

where, $[\min_A, \max_A]$ is the initial range, $[new\_\min_A, new\_\max_A]$ is the new range and v is the value to transform. The lowest Glide energy was set to 0 and the highest to 100.

The plots of Src tyrosine kinases and serine/threonine kinases in complex with PP1 were made measuring the linear correlation between the predicted interaction energies and the experimental measured pIC_{50} ($-\log(IC_{50})$). For each family, the Pearson correlation coefficient was computed.

2.5 Experimental methods

We applied our computational protocol for protein engineering to *Mtb* PknG and five purchasable ATP analogues. The predictions made were experimentally tested. The experiments were performed using PknG and diverse PknG mutants as target proteins, GarA as substrate and three purchasable ATP-competitive ligands that were identified as good candidates as cofactors for PknG mutants (N6-(benzyl) ATP, PF9 and 7d7p ATP).

All experiments were performed by Mohamed-Ali Mahi[i].

2.5.1 Proteins expression

In all experiments the PknGΔN mutant was used instead of the full length protein. PknGΔN lacks the first 73 residues (N-terminus) and was highly soluble and stable compared to the full length protein [97].

For expression of His-tagged PknGΔN, the expression vector pET15b-PknGΔN was transformed in competent *Escherichia coli* cells (*E. coli* BL21(DE3)). A pET expression vector essentially contains an isopropyl β-D-1-thiogalactopyranoside (IPTG) inducible promoter, a ribosome binding site (RBS), the polilinker where the foreign gene is inserted, a termination site and a gene for ampicillin resistance. The *E. coli* cells were recovered in Lysogeny Broth (LB) medium, a nutrient-rich

[i] Mohamed-Ali Mahi, Biozentrum University of Basel, Switzerland

medium for allowing rapid and robust growth of bacteria, and then were spread on LB-ampicillin (LB-Amp) plates. Each LB-ampicillin plate contained the LB medium and the antibiotic ampicillin. Consequently, only cells that were successfully transformed and thus contained the expression vector with the gene for ampicillin resistance survived in LB-Amp. Once the OD_{600} indicated that the bacterial cultures were in the exponential growth phase (increased bacterial cell concentration), the PknGΔN expression was induced by IPTG.

The same procedure was used for each PknGΔN variant and the substrate GarA (Table 2.2).

Table 2.2. PknGΔN variants and expression vectors used.

PknGΔN variants	**Expression vectors**
PknGΔN-M232G	pET15b-PknGΔN-M232G
PknGΔN-V235G	pET15b-PknGΔN-V235G
PknGΔN-M232S	pET15b-PknGΔN-M232S
PknGΔN-M232T	pET15b-PknGΔN-M232T
GarA	pET15b-GarA

2.5.2 Protein purification

The bacterial pellet, containing the His-tagged PknGΔN, was suspended in a suitable buffer, blended by a homogenizer and the lysate was cleared by subsequent centrifugation steps. The supernatant was filtered and applied to a HisTrap column to perform an immobilized metal affinity chromatography to purify the protein of interest (His-tagged PknGΔN). The His-tag protein was eluted from the column with an elution buffer containing imidazole. The collected fractions were tested by SDS-PAGE. The gel was stained with 'Instant Blue-like' solutions. Fractions contained high amount of His-tagged PknGΔN were further purified by size exclusion chromatography to remove imidazole and to obtain pure protein. Collected fractions containing pure protein were again tested by SDS-PAGE. Protein concentration was determined by direct UV Absorbance at 280 nm (UV A280nm).

The same protocol was used to purify PknGΔN mutants and the substrate GarA.

2.5.3 Kinase assays

The kinase activity of PknGΔN and PknGΔN mutants was checked by performing *in vitro* kinase assays. PknGΔN and PknGΔN mutants did not have autophosphorylation activity since the cluster of autophosphorylation sites (Thr23, Thr32, Thr63 and Thr64) was located in the removed N-terminus [133]. This deletion allows avoiding kinase autophosphorylation signals on the gel. The deletion of the N-terminus did not affect the kinase activity of PknG [97].The *in vitro* phosphorylation was carried out preparing two reaction mixes for each protein kinase. Each reaction mix contained the protein kinase of interest (PknGΔN or a PknGΔN mutant), the substrate GarA, the natural ATP or an ATP analogue and $MnCl_2$. The presence of manganese chloride was necessary because divalent cations (in that case Mn^{2+}) are required from kinase proteins to be active and to phosphorylate a specific substrate. Reactions were quenched by adding EDTA to the reaction mix. Each reaction mix was loaded on a SDS-PAGE gel and phosphorylated and non-phosphorylated substrates were separated on the gel. Proteins were detected by staining the gel with 'Instant Blue-like' solution.

3 Results

To develop the computational protocol used in this thesis and to test its performance, we used different protein kinases from the scientific literature to which the Shokat's method was applied. These kinases made up our data set containing 7 wild-type kinase proteins and 15 engineered kinases (Table 3.1). The ATP analogues were four N6-(substituent) ATPs with bulky hydrophobic groups at the N6 position of the adenine ring and the pyrazolopyrimidine PP1, as shown in Figure 3.1.

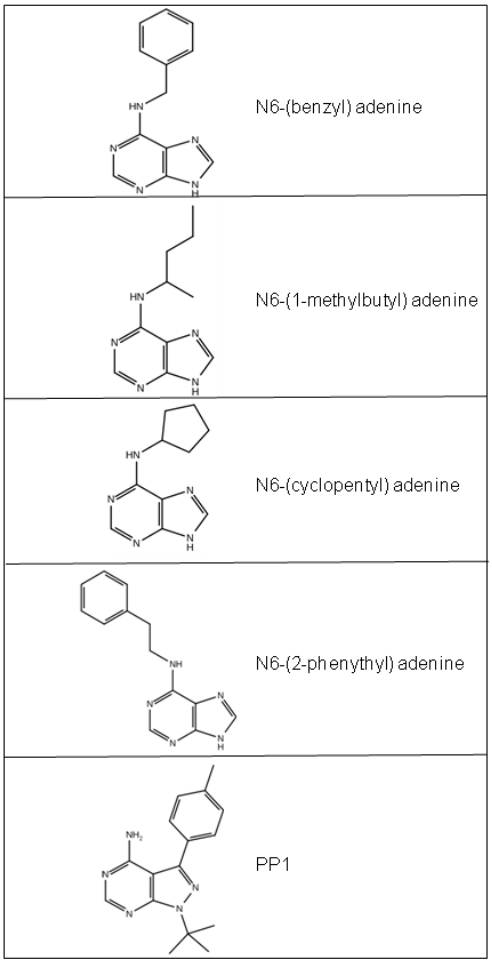

Figure 3.1. Chemical structures of N6-(substituent) ATPs and PP1. For simplicity, only the structures of the adenine ring and the hydrophobic groups are reported.

Table 3.1. Percentage of substrate phosphorylation by ATP, catalytic efficiency (kcat/Km), IC_{50} and predicted interaction energy for each protein-ligand pair.

Kinases	**Ligands**	**Experimental data**	**Predicted interaction energies (kcal/mol)**
v-Src tyrosine kinase [18]		**k_{cat}/K_m($min^{-1}M^{-1}$)**	
v-Src	ATP	$1.6 * 10^5$	-21.35
v-SrcI338A		$1.4*10^4$	-19.91
v-SrcI338G		$1 * 10^4$	-19.01
v-Src	N6-(benzyl) ATP	0	0 [a)]
v-SrcI338A		$2.5 * 10^4$	-14.92
v-SrcI338G		$4.0 * 10^4$	-29.17
JNK kinase [27]		**% substrate phosphorylation by ATP**	
JNK	N6-(benzyl)	99	0 [a)]
	N6-(2-phenythyl)	98	0 [a)]
	N6-(cyclopentyl)	97	0 [a)]
	N6-(1-methylbutyl)	93	0 [a)]
JNKM108GL168A	N6-(benzyl)	62	-17.34
	N6-(cyclopentyl)	59	-14.42
	N6-(1-methylbutyl)	47	-20.42
	N6-(2-phenythyl)	8	-33.0
Tyrosine and serine/threonine kinases [102]		**IC_{50}(μM)**	
v-Src	PP1	5 ± 2	0 [a)]
v-SrcI338F		8 ± 2	0 [a)]
v-SrcI338M		8 ± 1	0 [a)]
v-SrcI338S		0.4 ± 0.05	-28.78
v-SrcI338T		0.1 ± 0.02	-33.32
v-SrcI338V		0.1 ± 0.02	-27.41
v-SrcI338C		0.07 ± 0.02	-27.44
v-SrcI338A		0.005 ± 0.002	-39.26
v-SrcI338G		0.005 ± 0.002	-38.56
Fyn		0.05 ± 0.02	-36.81
FynT339A		0.005 ± 0.002	-36.21
Abl		0.3 ± 0.03	-32.93
AblT334A		0.03 ± 0.005	-33.86
CamKII		80 ± 10	0 [a)]
CamKIIF89G		0.5 ± 0.1	-15.74
Cdk2		50 ± 10	0 [a)]
Cdk2F80G		0.16 ± 0.03	-24.85
P38		0.82 ± 0.2	-34.61
P38T106A		0.0027 ± 0.005	-33.43
P38T106G		0.0027 ± 0.005	-32.78

[a)] positive interaction energies approximated to 0.

We used our *in silico* protocol to explore pairings of potential mutations and ligand analogues by identifying which residues within the binding pocket of the target protein could to be mutated to accommodate a specific ATP analogue. The predicted interactions between engineered kinases and ATP analogues were compared with the published data.

3.1 v-Src and N6-(benzyl) ATP

To develop and test the performance of our approach, we firstly used it with the system studied by Shokat and coworkers in 1997, v-Src and the N6-(benzyl) ATP [18]. Shokat and coworkers engineered the v-Src ligand-binding pocket to generate a kinase mutant that preferentially used N6-(benzyl) ATP as phosphodonor instead of the natural nucleotide, ATP. They showed that the wild-type protein had a substrate preference for ATP over the ATP analogue ($1.6 * 10^5$ $min^{-1}M^{-1}$ vs 0) whereas the I338G mutant preferentially used N6-(benzyl) ATP as nucleotide over the natural ATP (the kcat/Km ratio was 4 to 1). The key finding was that residue at position 338, Ile338, limited the ability of v-Src to bind to N6-(benzyl) ATP and the mutation of this residue into one with a smaller side chain (such as Gly) conferred on v-Src the capability to utilize the ATP analogue as substrate.

We applied our protocol to that system. The wild-type v-Src and ANP (an ATP analogue with an amino group in place of the oxygen between β and γ phosphates) were used as input structures. An ensemble of 1040 N6-(benzyl) ATP conformers, obtained by a conformational search (see Methods), was analyzed in the context of the binding pocket of wild-type v-Src. Residues overlapping with the benzyl group at the N6 position of the adenine base were considered as potential candidates for single-point mutations. None of the conformers occupied the wild-type binding pocket without clashes and a collection of 43 conformers overlapped with single residues. The analysis of these conformers showed that they were organized in 2 main groups, a first one (32 conformers) that clashed with Ile338 and a second one (11 conformers) that clashed with Leu393 (Figure 3.2). Residues Ile338 and Leu393 were computationally mutated into Gly in order to obtain v-SrcI338G and v-SrcL393G, respectively. We analyzed the interaction between v-Src mutants and N6-(benzyl) ATP conformers using three protein-ligand scoring functions, GlideScore [71], X-Score [72], and DSX [68]. Table 3.2 and 3.3 represent values obtained for v-SrcI338G and v-SrcL393G, respectively. For v-SrcI338G and the 32 N6-(benzyl) ATP conformers we got different results from the 3 scoring functions used.

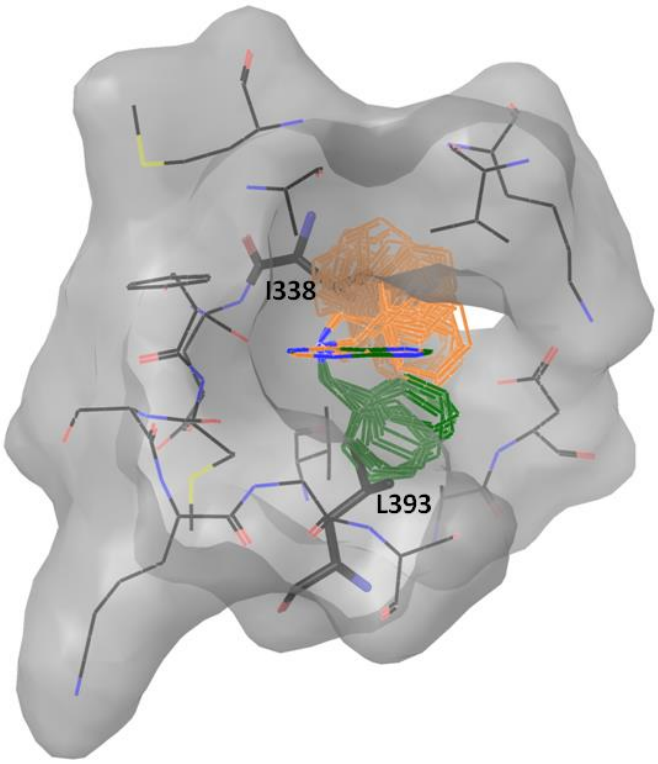

Figure 3.2. The v-Src ligand-binding pocket (represented as a molecular surface) and the two main groups of clashing conformers. The conformers that clash with Ile338 and Leu393 are represented in orange and green, respectively.

According to GlideScore only 4 conformers bound to the engineered kinase (negative interaction energies), for X-Score all conformers bound to v-SrcI338G and DSX identified 27 conformers having favorable interaction energies with the kinase mutant (Table 3.2). To understand the different results obtained, we visually inspected the 32 conformers within the engineered binding site of v-SrcI338G. Figure 3.3 represents a N6-(benzyl) ATP conformer within the binding pocket of v-SrcI338G. As we can see, the benzyl ring clashed with backbone atoms of Gly338 and Glu339. The GlideScore for that mutant-analogue pair is 10 000 kcal/mol, the X-Score is 5.25 and the DSX is -22.945 kcal/mol. Only GlideScore identified the conformer as a bad one in agreement with the observed clashes between the protein and the ligand. Similarly, for v-SrcL393G and the 11

N6-(benzyl) ATP conformers we got different results from the three different scoring functions (Table 3.3).

Table 3.2. GlideScore, X-Score, and DSX values for v-SrcI338G and N6-(benzyl) ATP conformers.

	v-SrcI338G overlapping conformers (32 conformers)	
	1 to 4	5 to 32
GlideScore (kcal/mol)	-8.67 to -6.84	10 000
	1 to 32	-
X-Score (pk_d units)	5.61 to 5.01	-
	1 to 27	28 to 32
DSX (kcal/mol)	-75.03 to -22.95	> 0

Table 3.3. GlideScore, X-Score, and DSX values for v-SrcL393G and N6-(benzyl) ATP conformers.

	v-SrcL393G overlapping conformers (11 conformers)	
	1 to 11	-
GlideScore (kcal/mol)	10 000	-
	1 to 11	-
X-Score (pk_d units)	5.25 to 5.05	-
	1 to 11	-
DSX (kcal/mol)	-49.90 to -12.56	-

For GlideScore none of the 11 conformers can be accommodated by the engineered kinase binding site whereas for both X-Score and DSX all conformers can occupy the binding pocket. We visually inspected the 11 conformers within the v-SrcL393G ligand-binding pocket. Figure 3.4 represents a N6-(benzyl) ATP conformer within the binding pocket of v-SrcL393G. The benzyl ring clashed with

Ile338 and Ala403. Only GlideScore identified that conformer as a bad one (GlideScore was equal to 10 000 kcal/mol) whereas both X-Score and DSX were unable to detect clashes between engineered protein and ligand analogue, assigning favorable scores (5.18 and -45.90 kcal/mol, respectively).

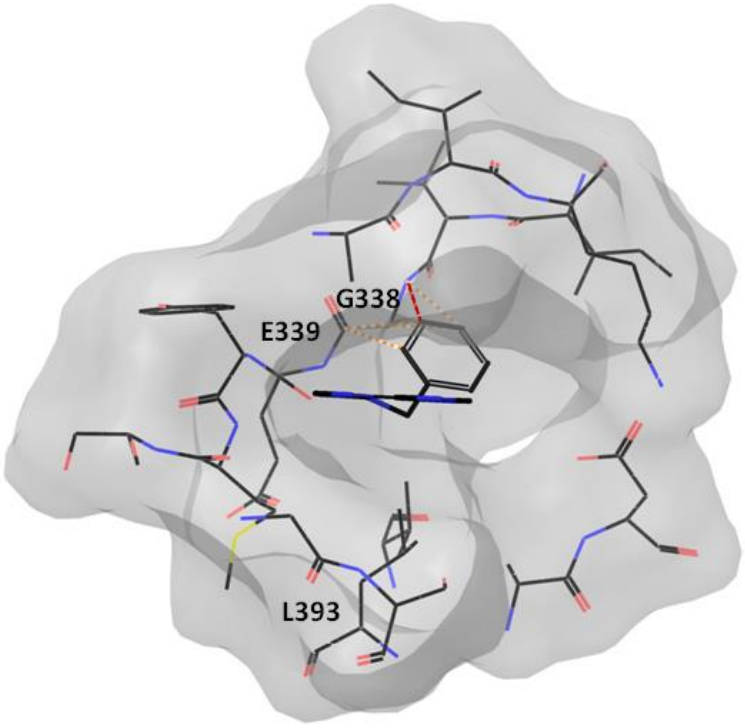

Figure 3.3. The v-SrcI338G ligand-binding pocket (represented as a molecular surface) and a N6-(benzyl) ATP conformer (represented as stick). The orange and red dashed lines represent “bad and ugly” protein-ligand contacts, respectively (Appendix A3).

The analysis already described suggested us to use GlideScore as scoring function to evaluate and rank the kinase mutant-ligand conformer pairs. In that specific case, GlideScore provided favorable interaction energies for complexes of v-SrcI338G and four N6-(benzyl) ATP conformers (Table 3.2). At that point, we computed the relative free energies of binding for all complexes of v-SrcI338G and ligand conformers by using the MM-GBSA method. Here we reported the free energy of binding only for four complexes of v-SrcI338G and ligand conformers for which we previously got favorable GlideScores and two complexes of the

engineered v-Src and ligand conformers for which we got unfavorable GlideScores. In the last two complexes the N6-(benzyl) ATP conformers clashed with the engineered binding sites.

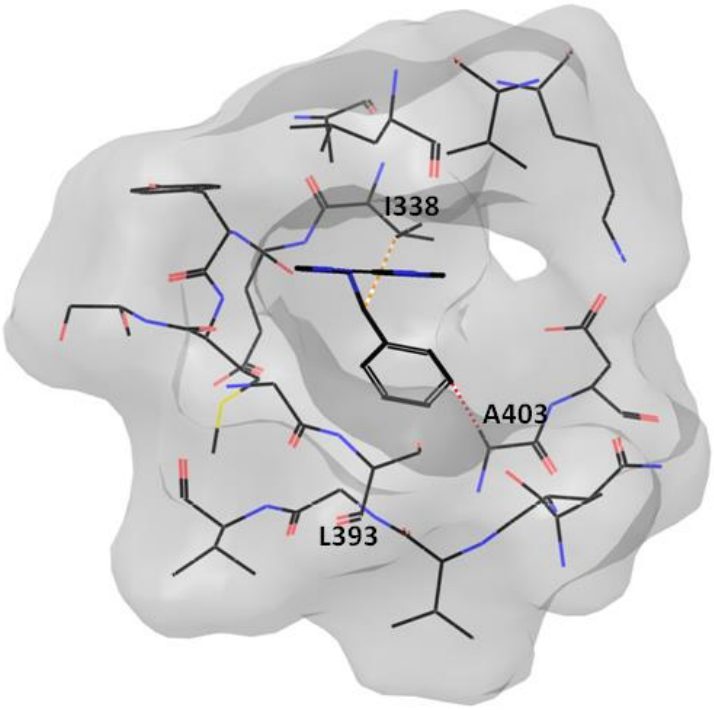

Figure 3.4. The v-SrcL393G ligand-binding pocket (represented as a molecular surface) and a N6-(benzyl) ATP conformer (represented as stick). The orange and red dashed lines represent bad and ugly protein-ligand contacts, respectively (Appendix A3).

Each complex was subjected to a molecular dynamics (MD) protocol with a production phase of 20 ns. To check the stability of the mutant-analogue complexes, for each of them the rmsd of the Cα atoms relative to the initial complex structure was monitored. Figure 3.5 represents the structures and the rmsd plots for a complex with a favorable GlideScore (A) and for a complex where the ligand clashes with the protein and for which we previously obtained an unfavorable GlideScore (B). The single trajectories obtained via MD simulations were used to compute the free energy of binding (ΔG) for each mutant-analogue complex. Those values were compared to GlideScore values previously obtained (Table 3.4).

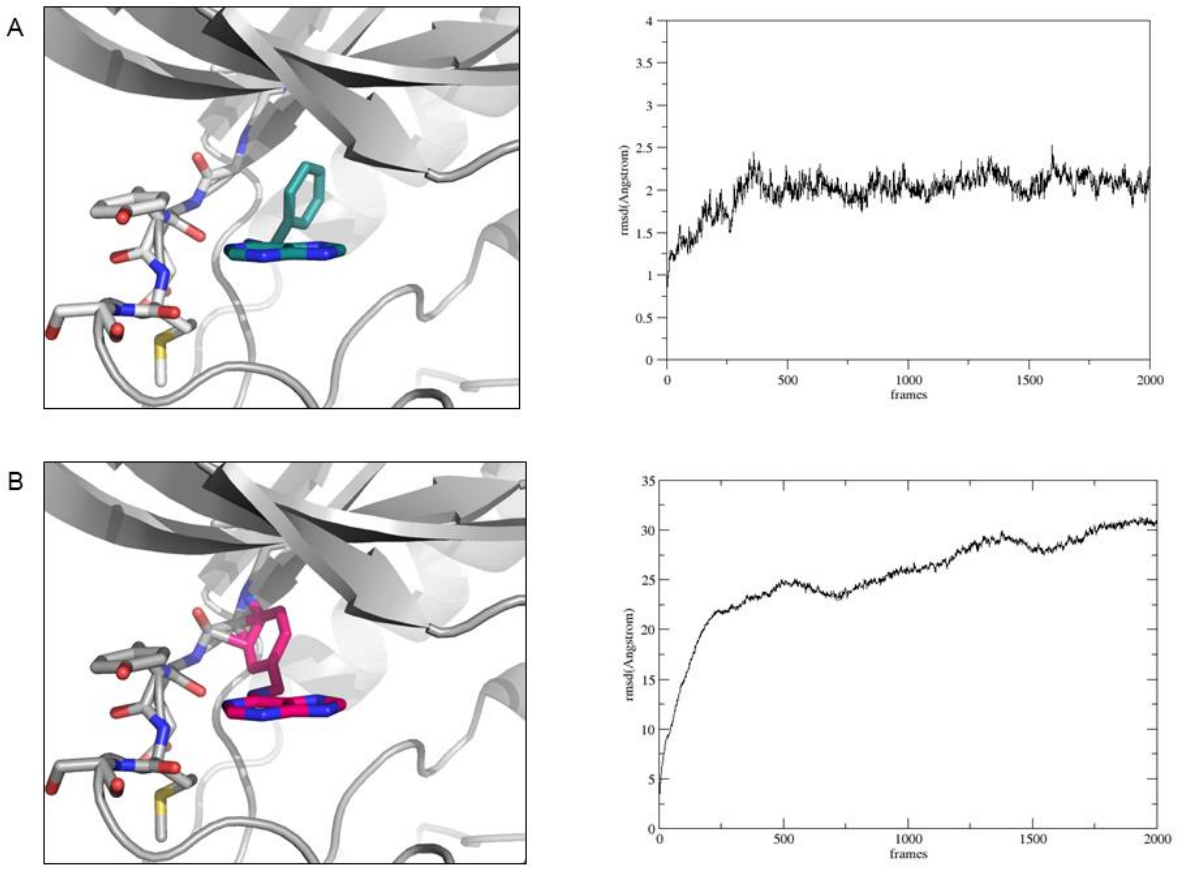

Figure 3.5. Molecular dynamics trajectory plots correlating rmsd deviation from the initial v-SrcI338G–N6-(benzyl) ATP complex coordinates, over a simulation time of 20 ns (2000 frames). A) v-SrcI338G–N6-(benzyl) ATP with no protein-ligand clashes and the best GlideScore; B) v-SrcI338G–N6-(benzyl) ATP with strong protein-ligand clashes and the worse GlideScore. In both complexes, only adenine moieties are represented.

Table 3.4. GlideScore and ΔG values for six v-SrcI338G–N6-(benzyl) ATP complexes. The complexes are numbered from 1 to 6 according to GlideScore rank.

Mutant-analogue complexes	**GlideScore(kcal/mol)**	**ΔG(kcal/mol)**
v-SrcI338–N6-(benzyl)1	-7.67	-24.02 ± 2.43
v-SrcI338–N6-(benzyl)2	-7.58	-25.71 ± 2.40
v-SrcI338–N6-(benzyl)3	-7.38	-24.29 ± 2.37
v-SrcI338–N6-(benzyl)4	-6.84	-24.47 ± 2.77
v-SrcI338–N6-(benzyl)5	10 000	4.07 ± 2.30
v-SrcI338–N6-(benzyl)6	10 000	31.29 ± 2.25

The rmsd plots showed that the complex with a favorable GlideScore was stable during the production phase of the MD simulation (Figure 3.5 A). The rmsd showed a convergence after 2-4 ns (frames 200-400) in a range between 2 to 2.5 Å. In that case the ligand did not significantly change its pose within the binding site. For the complex where the ligand overlaps the protein, after few nanoseconds the rmsd increased steeply (from ~ 5Å to ~20Å) and it continues to increase during the production phase (Figure 3.5 B). The ligand has moved away from the binding site and this might be a response to the initial unfavorable superposition between the protein and the ligand. The single trajectories obtained via MD simulations were used to compute the free energy of binding (ΔG) for each mutant-analogue complex. Those values were compared to GlideScore values previously obtained (Table 3.4). The relative free energies of binding provided the same trend of the GlideScore values. The MM-GBSA results validated the ability of GlideScore to correctly predict the binding of a ligand to its target protein. GlideScore was therefore selected as protein-ligand scoring function in our computational protocol.

As previously mentioned, in their first papers Shokat and coworkers found that the mutation of the gatekeeper residue Ile338 into Ala or Gly conferred to v-Src the ability to accommodate N6-(benzyl) ATP within the binding pocket [18, 134]. By applying our protocol, we identified Ile338 as a candidate for single-point mutation and evaluated the interaction of v-SrcI338A and v-SrcI338G with an ensemble of N6-(benzyl) ATP conformers. The most promising pair of kinase mutant and analogue conformer was the one with the best GlideScore. For that pair the Glide energy was taken into account and compared to the available kinetic data. The wild-type v-Src and both I338A and I338G mutants are predicted to bind to ATP with almost equal interaction energies (Figure 3.6 A). The wild-type v-Src does not accommodate N6-(benzyl) ATP because of the steric overlapping between Ile338 and the benzyl ring at the N6 position (predicted interaction energy is 0 kcal/mol). The I338A and I338G mutants have an enlarged binding pocket and both mutants show favorable interaction with N6-(benzyl) ATP (Figure 3.6 B).

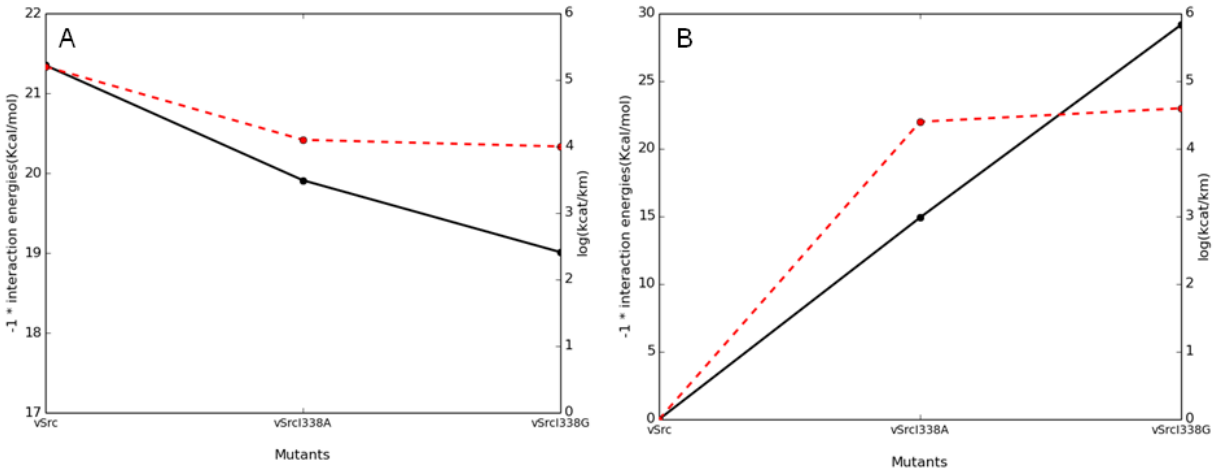

Figure 3.6 Comparison of the catalytic efficiency and predicted interaction energy for v-Src, v-SrcI338A, and v-SrcI338G with ATP (A) and with N6-(benzyl) ATP (B). Shown on the x-axis are the wild-type protein and the two mutants. The primary y-axis (on the left) is the predicted interaction energies multiplied by -1 and the secondary y-axis (on the right) is the log of the kcat/Km ratio. The red dashed lines represent the experimental data and the solid black lines represent predicted interaction energies.

3.2 JNK and N6-(substituent) ATPs

Habelhah and coworkers modified the JNK ATP binding site so that it bound to N6-(substituent) ATPs that cannot be accommodated by the wild-type binding pocket. The designed pair of engineered JNK and ATP analogue allowed for the identification of novel JNK substrates [27]. They identified two key residues to mutate within the JNK ligand-binding pocket, Met108 and Leu168. Those residues were mutated into Gly and Ala, respectively. To determine the ATP analogue with the highest affinity for the engineered JNK, they compared four N6-(substituent) ATPs, N6-(benzyl) ATP, N6-(2-phenythyl) ATP, N6-(cyclopentyl) ATP and N6-(1-methylbutyl) ATP (Figure 3.1). Their efficiency as phosphodonor was tested by measuring their ability to prevent phosphorylation of substrates by ATP when they are added in excess with respect to ATP. For JNK and the ATP analogues the percentage of substrate phosphorylation ranged from 99% to 93%, showing the inability of the wild-type kinase to accommodate any of the four ATP

analogues. On the other hand, the JNKM108GL168A mutant was able to accommodate N6-(substituent) ATPs and N6-(2-phenythyl) was the ligand with the highest affinity to the mutant (the percentage of substrate phosphorylation was 8%) (Table 3.1).

We applied our computational protocol on the system studied by Habelhah and coworkers. We used wild-type JNK and ANP as input structures - Figure 3.7 represents the JNK ligand-binding pocket. For each of the four ATP analogues we performed a conformational search (as described in Methods) to obtain ensembles of low-energy conformers. The ensembles were analyzed within the binding site of wild-type JNK and results are shown in Table 3.5. All conformers of each ATP analogue clashed with diverse residues within the JNK ligand-binding pocket (Met108, Glu109, Leu110, Met111, Ala53, Ile86, and Leu206) implying that they cannot fit in the wild-type binding site. For each ATP analogue we identified diverse pairs of residues to be mutated into smaller residues to enlarge the binding site (Appendix A4).

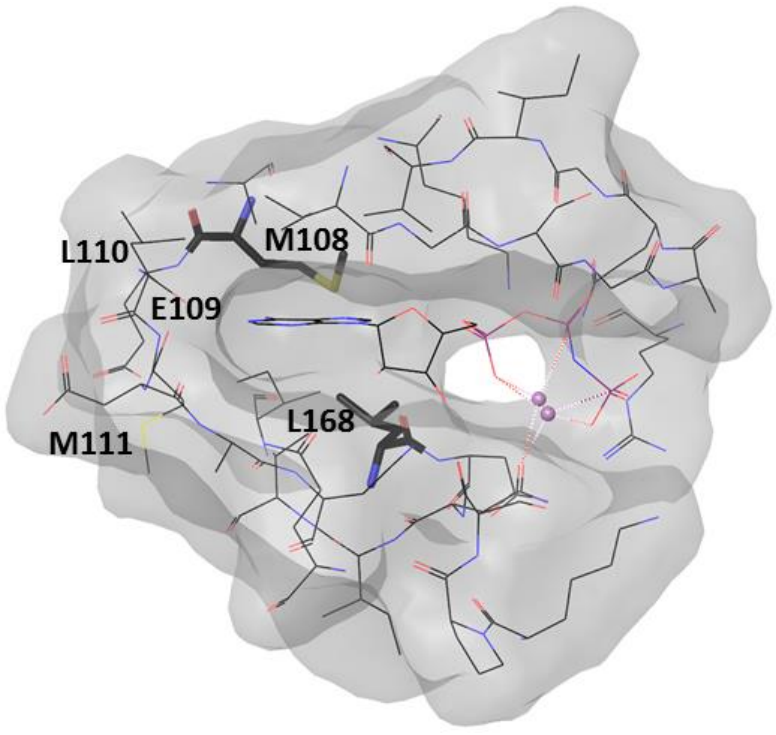

Figure 3.7. Wild-type JNK ligand-binding pocket (represented as a molecular surface), ANP (represented as line) and Mg^{2+} ions (represented as pink spheres). Residues Met108 and Leu168 are represented as sticks and occupy the upper and the lower part of the binding site, respectively. Amino acids from 108 to 111 make up the hinge region.

Table 3.5. Classification of conformers of N6-(benzyl) ATP, N6-(2-phenythyl) ATP, N6-(cyclopentyl) ATP and N6-(1-methylbutyl) ATP within the JNK ligand-binding pocket.

	Total number of conformers	Conformers with 0 overlaps	Conformers overlapping with 1 res.	Conformers overlapping with 2 res.	All others
N6-(benzyl)	1040	0	27	269	744
N6-(2-phenythyl)	965	0	22	227	716
N6-(cyclopentyl)	380	0	1	70	309
N6-(1-methylbutyl)	971	0	6	121	844

We identified three double mutants that can accommodate the N6-(substituent) ATPs, M108GL168A, M108GI86A, and M108GA53G. Since experimental data were only available for the M108GL168A mutant, we focused on it and compared our findings with the percentage of substrate phosphorylation. We reproduced the relative ranking of the four ATP analogues as substrates for the engineered JNK (Figure 3.8).

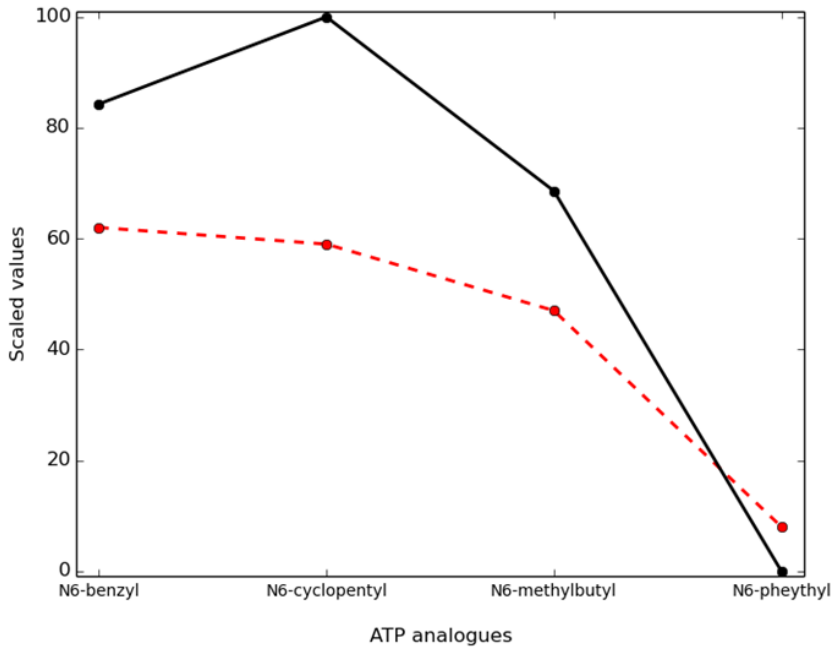

Figure 3.8. The percentage of substrate phosphorylation by ATP in presence of ATP analogues and the scaled predicted interaction energies for the engineered JNK and each of the four ATP analogues. The dashed red line represents the experimental data and the solid black line the predicted data.

3.3 Tyrosine and serine/threonine protein kinases and PP1

PP1 is a potent inhibitor of Src protein kinases [135]. The first evidence that the ability of PP1 as Src kinases inhibitor was related to the nature of the gatekeeper residue came from the discovery that PP1 was a good inhibitor (0.7 μM) of c-Src, where the gatekeeper residue is a Thr, and was a poor inhibitor (5 μM) of v-Src, where the gatekeeper is an Ile. That finding suggested that the size of the side chain of the gatekeeper amino acid might be essential for the kinase inhibition from PP1. Liu and coworkers decided to test the relationship between the nature of the gatekeeper residue and the PP1 potency in two families of protein kinases, Src tyrosine and serine/threonine kinases [102]. Proteins studied are reported in Table 3.7.

Table 3.7. Protein kinases studied by Liu and coworkers [102].

Src family protein kinases	Serine/threonine protein kinases
v-Src, Fyn, Abl	CamKII, Cdk2, P38

The gatekeeper amino acid is Ile338 in v-Src, Thr339 in Fyn, Thr334 in Abl, Phe89 in CamKII, Phe80 inCdk2, and Thr106 in kinase P38 (Figure 3.9). It was shown that in both families residues equal to or larger than Ile, such as Phe and Met, made PP1 a poor inhibitor whereas residues smaller than Ile, such as Ser, Thr, Val, Cys and especially Ala and Gly increased the potency of PP1.

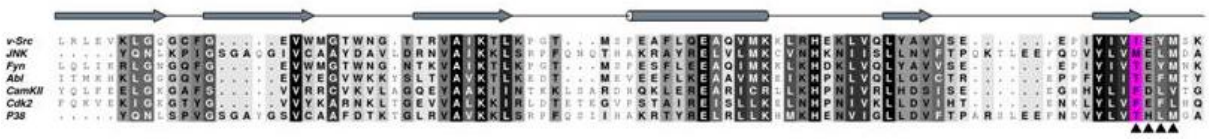

Figure 3.9. Sequence alignment of the N-lobe and hinge region of the seven wild-type protein kinases belonging to our data set. The alignment was build using the T-Coffee web server [136]. Residues are colored by percentage identity. The hinge region is designated with black triangles and the gatekeeper residues are colored in magenta. Secondary structure elements are represented as follows: β strands as arrows, α helices as cylinders, and coils as lines.

Table 3.8 shows results obtained using our computational protocol with the seven kinases and PP1.

Table 3.8. Classification of the PP1 conformers within the binding sites of v-Src, Fyn, Abl, CamKII, Cdk2 and P38

Wild-type kinase and gatekeeper residue	Total number of conformers	Conf. with 0 overlaps	Conf. overlapping with 1 residue	All others
v-Src (I338)	1000	0	1000	0
Fyn (T339)	1000	1000	0	0
Abl (T334)	1000	1000	0	0
CamKII (F89)	1000	0	717	283
Cdk2 (F80)	1000	0	1000	0
P38 (T106)	1000	1000	0	0

Wild-type v-Src, CamKII and Cdk2 cannot accommodate PP1 within their binding pocket. All 1000 conformers overlapped with a single residue within the binding site. Those residues are Ile338 of v-Src, Phe89 of Cdk2 and Phe80 of Cdk2 that clashed with the hydrophobic ring of all 1000 PP1 conformers. Those results agreed with IC_{50} values > 1 μM (Table 3.1).

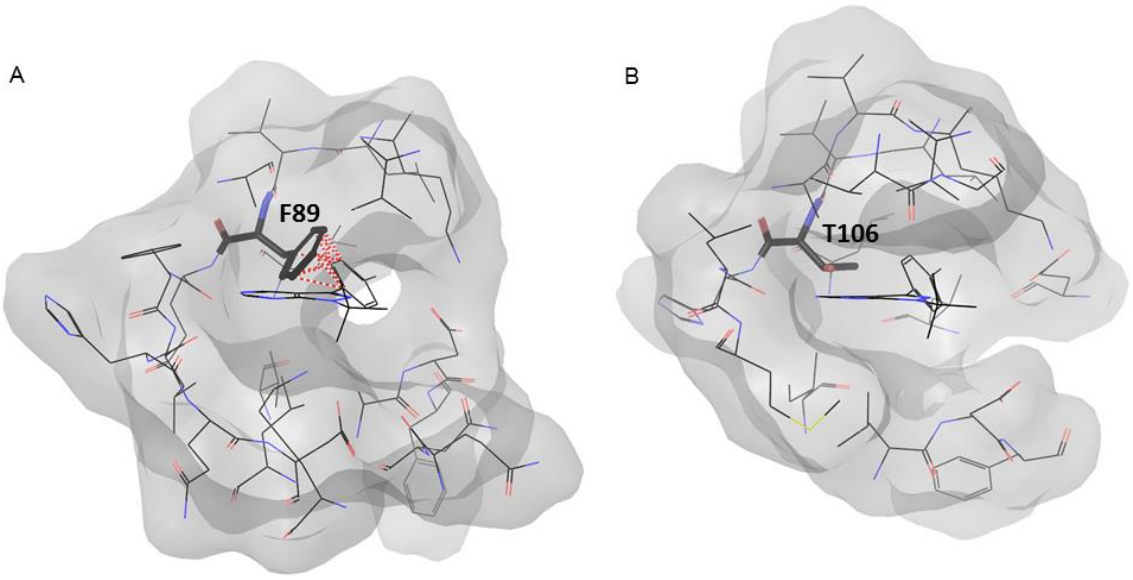

Figure 3.10. A) Wild-type Cdk2 ligand-binding pocket (represented as a molecular surface) and a PP1 conformer (represented as line). B) Wild-type P38 ligand-binding pocket (represented as molecular surface) and a PP1 conformer (represented as line). Phe89 and Thr106 are represented as sticks and protein-ligand clashes as dashed red lines (Appendix A3).

Figure 3.10 A represents a PP1 conformer within the Cdk2 binding site. The side chain of Phe89 and the tolyl ring of PP1 clashed to each other. On the other hand, the Thr naturally presents in the wild-type Fyn, Abl and P38 did not clash with any PP1 conformer. PP1 can be accommodated within the binding site of the three kinases and that finding agreed with IC_{50} values ranging from 0.05 μM to 0.82 μM (Table 3.1). Figure 3.10 B represents a PP1 conformer within the P38 binding site. Thr106 and the hydrophobic ring of PP1 did not overlap.

For each of the six protein kinases, we computationally mutated the gatekeeper residues in order to obtain the 14 mutants studied in Liu's work. For both kinases families the predicted energies of interaction between proteins and inhibitor estimated the trend of the inhibitor potency (Table 3.1). Moreover, we obtained a positive correlation between the experimental –log(IC_{50}), the pIC_{50}, and the predicted interaction energies, with a Pearson correlation of 0.85 for Src tyrosine kinases and of 0.75 for serine/threonine kinases (Figure 3.11 A and B).

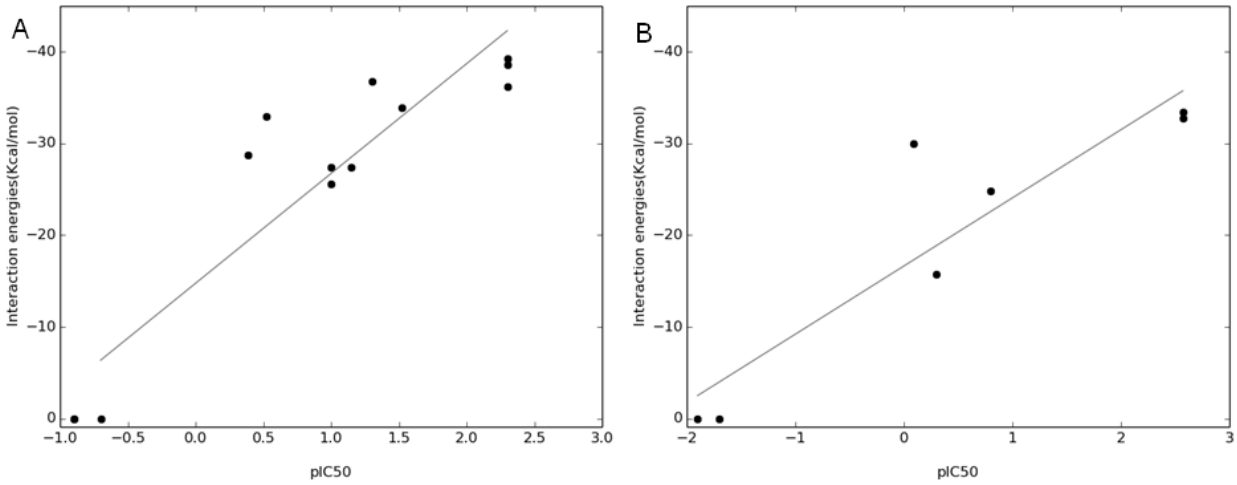

Figure 3.11. Correlation plots of the predicted interaction energies versus the experimental pIC_{50}. (A) Src family tyrosine kinases with PP1. (B) Serine/threonine kinases with PP1 (B).

3.4 Experimental application

We apply our computational protocol to the *Mycobacterium tuberculosis* serine/threonine protein kinase G (*Mtb* PknG). Nowadays, several studies have proved the significant role of PknG in the survival of *Mtb* within the host macrophage. Nonetheless, PknG downstream substrates involved in the infection pathway are still unknown. These open questions make PknG an attracting target to which apply our computational approach. Figure 3.12 represent the kinase domain of PknG in complex with ATP.

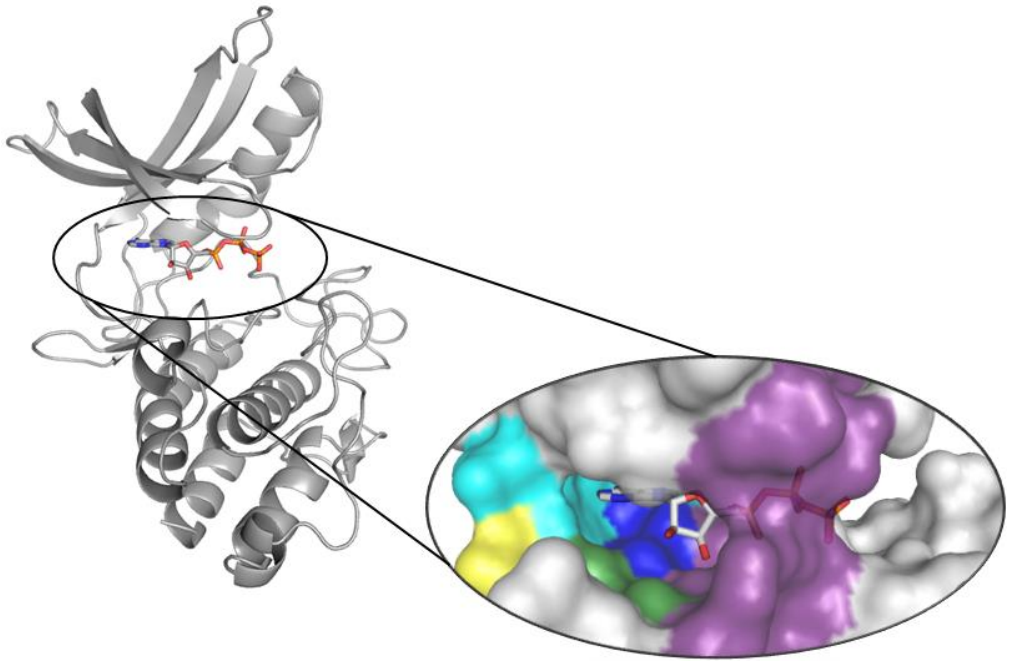

Figure 3.12. Structure of the PknG kinase domain (gray ribbon) in complex with ATP (gray stick). The structure is shown with a zoom-in into the binding site. It is represented as a surface where the adenine region is colored in cyan, the sugar region in green, the phosphates region in violet, the buried region in blue and the solvent accessible region in yellow.

We aim to identify residues to mutate within the PknG ligand-binding site to generate mutants that are still catalytically active and can specifically use ATP analogues as phosphodonor. We used five purchasable ATP analogues, two with

a conserved chemistry with respect to the analogues used by Shokat and coworkers (hydrophobic groups at the N6 position of the adenine ring, N6-(methyl) ATP and N6-(benzyl) ATP) (Figure 3.13 A) and three with different chemistry (groups on different positions of the adenine ring, 8-Azido ATP, 2-(phenylethyl) ATP, (PF9) and 7-deaza-7propargylamino-ATP (7d7p ATP)) (Figure 3.13 B). We applied our computational protocol on PknG and the five ATP analogues, and the predictions made were experimentally tested.

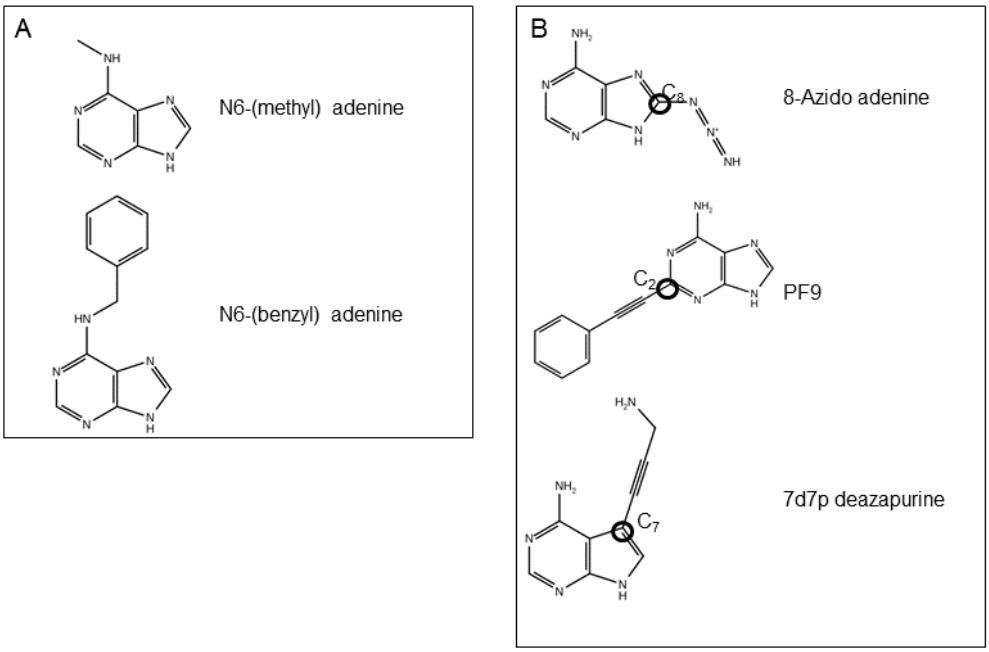

Figure 3.13 A and B. Chemical structures of the ATP analogues used with PknG. For simplicity, only the adenine base is represented. Black circles highlight positions at which substituent groups are attached.

3.4.1 PknG and N6-(methyl) ATP

The ensemble of N6-(methyl) ATP conformers, 890 conformers, was analyzed within the wild-type PknG binding site. Of the total ensemble, a group of 101 conformers occupied the ligand-binding site of wild-type PknG without any clash (Figure 3.14). The methyl group of the ATP analogue was small enough to accommodate the wild-type binding pocket without overlapping with any residue. N6-(methyl) ATP was predicted to bind to PknG and would not be a specific ATP-

competitive ligand for an engineered PknG. These results made N6-(methyl) ATP an inadequate candidate for our aim and it was discarded.

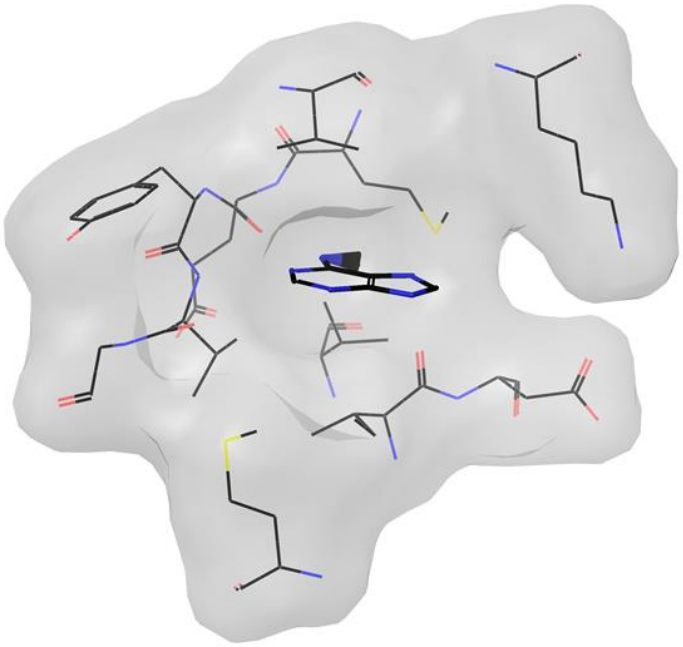

Figure 3.14. The wild-type PknG ligand-binding pocket (represented as a molecular surface) and N6-(methyl) ATP conformers that do not overlap with any binding-site residues (represented as stick). For simplicity, only the adenine rings are represented.

3.4.2 PknG and N6-(benzyl) ATP

An ensemble of N6-(benzyl) ATP conformers (1040 conformers) was analyzed within the binding pocket of wild-type PknG. We investigated if the wild-type protein could accommodate some conformers without any clash or it needed to be engineered. The PknG binding pocket was not large enough to accommodate an ATP analogue with a benzyl group at the N6 position of the adenine base. A number of 53 conformers, over the total ensemble, overlapped 2 residues within the ligand-binding pocket. A group of 52 conformers clashed with Met232 placed in the upper part of the binding site and the remaining conformer clashed with Val211 situated in the lower part of the pocket (Figure 3.15).

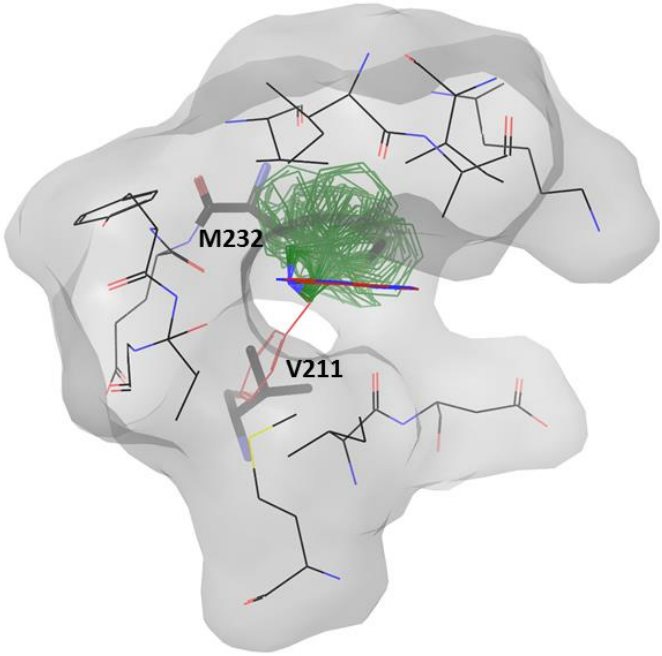

Figure 3.15. PknG ligand-binding pocket (represented as a molecular surface) and the two groups of overlapping conformers (represented as lines). Conformers that clash with Met232 and Val211 are represented in green and red, respectively.

We engineered the PknG binding site by mutating Met232 and Val211 into Gly in order to obtain PknGM232G and PknGV211G mutants. Even though the mutagenesis of Val211 into a smaller Gly introduced an additional space in the lower part of the binding site, it was still not large enough to accommodate the N6-(benzyl) ATP conformer. The hydrophobic group of the ATP analogue clashed with backbone atoms of Gly211 and Gln212 (Figure 3.16). Additionally, the analogue clashed with the side chain of Met232 suggesting the important role of that amino acid in defining the size of the binding site and in regulating the access of N6-(benzyl) ATP. The energy of interaction of PknGV211G and N6-(benzyl) ATP was unfavorable, at $18.04*10^{7}$ kcal/mol.

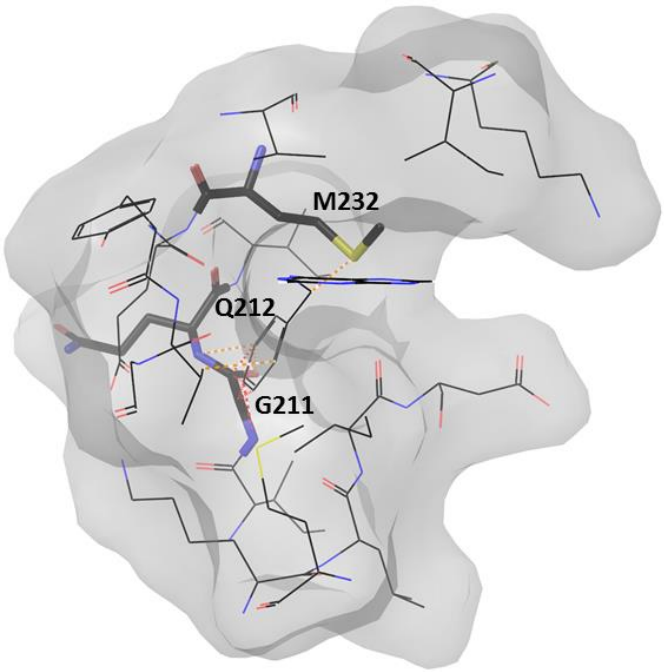

Figure 3.16. PknGV211G ligand-binding pocket (represented as a molecular surface) and N6-(benzyl) adenine conformer (represented as lines). The orange and red dashed lines represent “bad and ugly” protein-ligand contacts, respectively (Appendix A3).

The M232G mutant interacted with one N6-(benzyl) ATP conformer with a negative GlideScore (-6.50 kcal/mol) and thus a favorable interaction energy (-24.20 kcal/mol). The mutagenesis of Met232 into Gly created an additional pocket otherwise occupied by the polar side chain of Met. The enlarged binding site enabled the accommodation of N6-(benzyl) ATP. The adenine base was involved in a hydrogen bond (H-bond) with the Val235 whereas the hydrophobic group was involved in hydrophobic interactions with four non polar residues, Val179, Val211, Ile213 and Ile230 (Figure 3.17).

Our prediction was experimentally tested. An *in vitro* kinase assay was performed to assess the ability of PknGM232G to bind N6-(benzyl) ATP and use it as co-substrate. The assay was performed at different concentrations of cofactors (from 200 μM to 10 μM) (Figure 3.18).

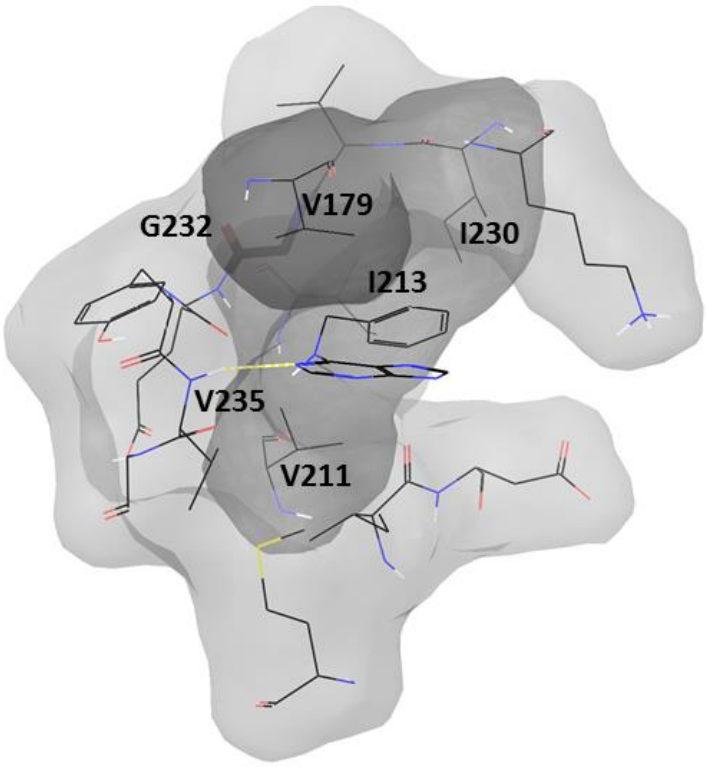

Figure 3.17. PknGM232G ligand-binding pocket (represented as a molecular surface) and N6-(benzyl) adenine conformer (represented as lines). Polar hydrogen are colored in white. The yellow dashed line represents the protein-ligand H-bond. Residues involved in hydrophobic interactions are represented as darker molecular surface (Val179, Ile230, Ile 213 and Val211) (Appendix A3).

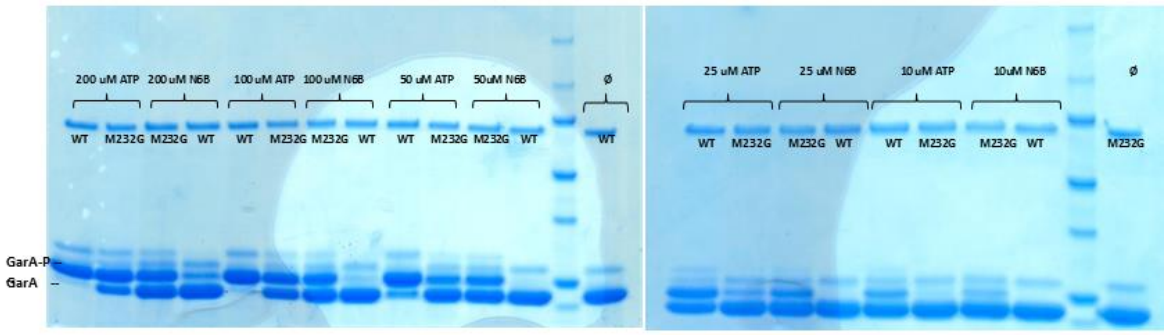

Figure 3.18. *In vitro* kinase assay (SDS-PAGE) using GarA as substrate, wild-type PknGΔN (WT) or PknGΔN-M232G (M232G) and ATP or N6-(benzyl) ATP (N6B) as cofactors. GarA-P is the phosphorylated GarA. The symbol Ø means that neither ATP nor N6B are used in those lanes.

The gel showed that PknGM232G was catalytically active and the single point mutation did not affect its kinase activity. Both PknG and PknGM232G phosphorylated GarA using ATP as phosphodonor and the activity of the mutant was lower than that one of the wild-type, especially at low concentrations of ATP. At concentrations of 200 μM and 100 μM, PknGM232G showed a significant activity with the ATP analogue. In the range from 50 μM to 10 μM PknGM232G used N6-(benzyl) ATP as phosphodonor to phosphorylate GarA and its kinase activity with N6-(benzyl) ATP was higher than with natural ATP. The kinase assay was repeated twice and both experiments showed the same results.

3.4.3 PknG and 8-Azido ATP

The 8-Azido ATP is an ATP analogue with an azide ion at the position C8 of the adenine ring (Figure 3.13). An ensemble of 1000 8-Azido ATP conformers was analyzed within the ligand-binding pocket of wild-type PknG. A group of conformers accommodated the wild-type ligand-binding site without any overlap meaning that 8-Azido ATP is expected to act as substrate in wild-type kinase (Figure 3.19). The 8-Azido ATP was not a suitable ligand for our goal.

The analysis of the 8-Azido ATP conformers within the PknG binding site allowed for a more general remark. For all conformers, azide ions occupy the region of the ATP binding site generally involved in the binding of phosphates, ‘phosphatases region’. In more detail, 356 conformers clashed with Lys181 and a group of 340 conformers overlapped Ile292 and Asp293. Lys181 and Asp293 are highly conserved in all tyrosine and serine/threonine protein kinases, they specifically interact with phosphates and have a significant role in the kinase catalytic activity [44]. The mutation of one of them into another residue would disrupt the ability of the protein kinase to phosphorylate its specific substrate resulting in an inactive engineered kinase. Since we aimed to change the kinase co-substrate specificity without affecting its catalytic activity, Lys181 and Asp293 would not be good candidate for mutagenesis. Those observations pointed out that it is of great importance always check the role of residues identify as potential candidate for mutagenesis before transforming them.

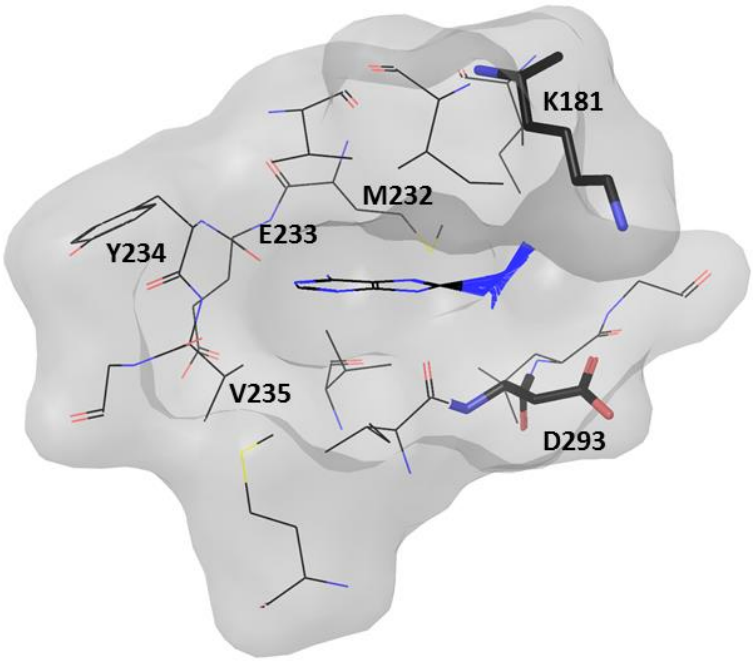

Figure 3.19. The wild-type PknG ligand-binding pocket (represented as a molecular surface) and 8-Azido ATP conformers without overlaps (represented as lines). Residues 232-235 formed the hinge region involved in the binding of the adenine ring. Lys181 and Asp293 (catalytic residues) are represented as stick.

3.4.4 PknG and PF9

The 2-(phenylethynyl) ATP (PF9) is an ATP analogue characterized by the presence of the non-polar phenylethynyl group at the position C2 of the adenine ring (Figure 3.13). An ensemble of 1000 low energy conformers of PF9 was analyzed within the ligand-binding pocket of the wild-type PknG. A group of 15 conformers clashed with 2 residues within the binding pocket, Val235 which overlapped with 14 conformers and Tyr234 which overlapped with 1 conformer. Val235 and Tyr234 were mutated *in silico* to Gly to obtain the V235G and Y234G mutants, respectively. The predicted interaction energies of the PknG mutants and the PF9 conformers are shown in Table 3.8.

Table 3.8. GlideScore values of PknGV235G and PknGY234G with PF9 conformers.

	PknGV235G conformers (14 conformers)		PknGY234G conformer (1 conformer)
	1, 2	3 to 14	1
GlideScore (kcal/mol)	-5.94, -4.9	10 000	10 000

The mutagenesis of Tyr234 into Gly did not allow the accommodation of PF9 within the engineered binding site. On the other hand, among the group of 14 conformers 2 occupied the V235G binding pocket with favorable GlideScores. The mutant-analogue pair with the best GlideScore had an interaction energy of -23.3 kcal/mol. The phenylethynyl group of PF9 was involved in favorable hydrophobic interactions with the side chain of Ile157 and the aromatic ring of Tyr234 (Figure 3.20).

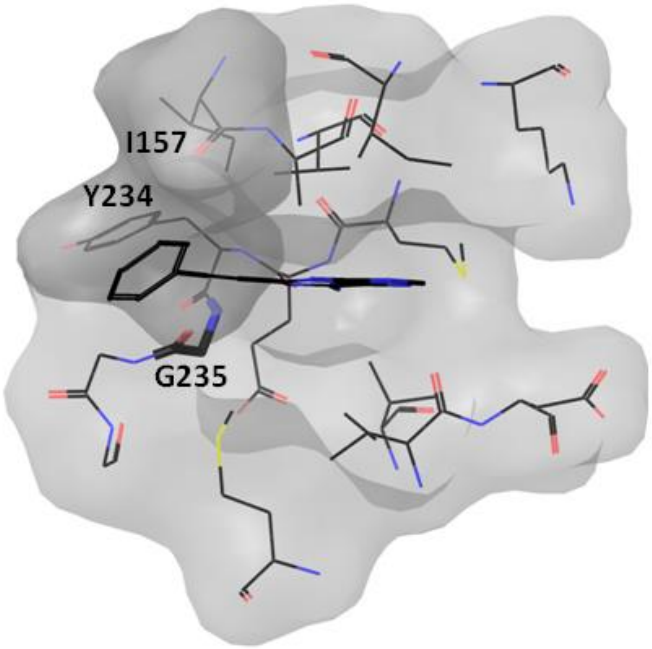

Figure 3.20. The PknGV235G ATP binding pocket (represented as a molecular surface) and the PF9 conformer with best GlideScore (represented as stick). Residues involved in hydrophobic interactions are represented as darker molecular surface (Tyr234 and Ile157). Gly235 (the mutated residue) is represented as stick.

The kinase assay was performed to experimentally test our prediction. The gel shows the phosphorylation activity of both PknG and PknGV235G in presence of native ATP or PF9 (Figure 3.21).

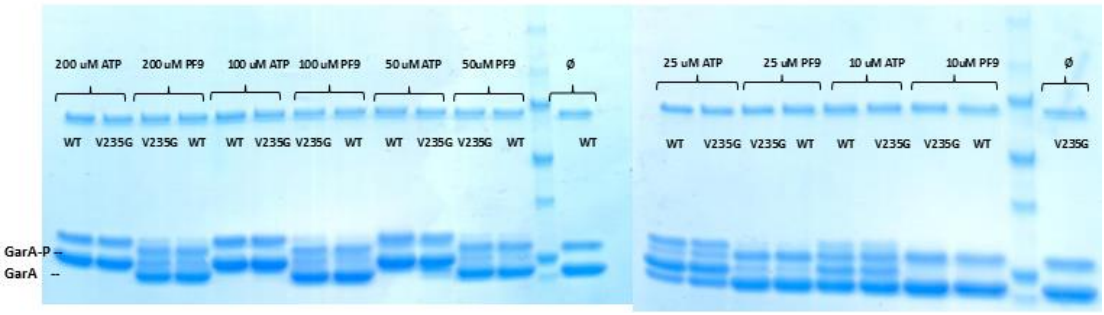

Figure 3.21. *In vitro* kinase assay (SDS-PAGE) using GarA as substrate, wild-type PknGΔN (WT) or PknGΔN-V235G (V235G) and ATP or PF9 as cofactors. GarA-P is the phosphorylated GarA. The symbol Ø means that neither ATP nor PF9 are used in those lanes.

The assay was performed using several cofactor concentrations (200 μM to 10 μM). The mutagenesis of Val235 into Gly did not affect the catalytic activity of the protein kinase. The engineered PknGV235G was able to phosphorylate GarA in presence of ATP at all concentrations. It was not possible to discriminate between the activity of PknG and PknGV235G in presence of PF9, both proteins had an insignificant phosphorylation activity in presence of PF9. In addition, at all concentrations PknGV235G had a high kinase activity in presence of native ATP, comparable to the phosphorylation activity of the wild-type protein in presence of ATP.

3.4.5 PknG and 7d7p ATP

The 7-deaza-7propargylamino ATP (7d7p ATP) has a propargylamine group at position C7 of the deazapurine ring where a nitrogen atom (N7) is replaced by a carbon atom (C7) (Figure 3.13). An ensemble of 1000 low energy conformers of 7d7p ATP clashed with only one residue within the PknG ligand-binding pocket, Met232. Figure 3.22 shows the ensemble of 7d7p ATP conformers within the

wild-type PknG binding pocket. All propargylamines pointed to the buried region of the ligand-binding pocket that is placed on the back of the ATP pocket and is generally not occupied by the ATP [39].

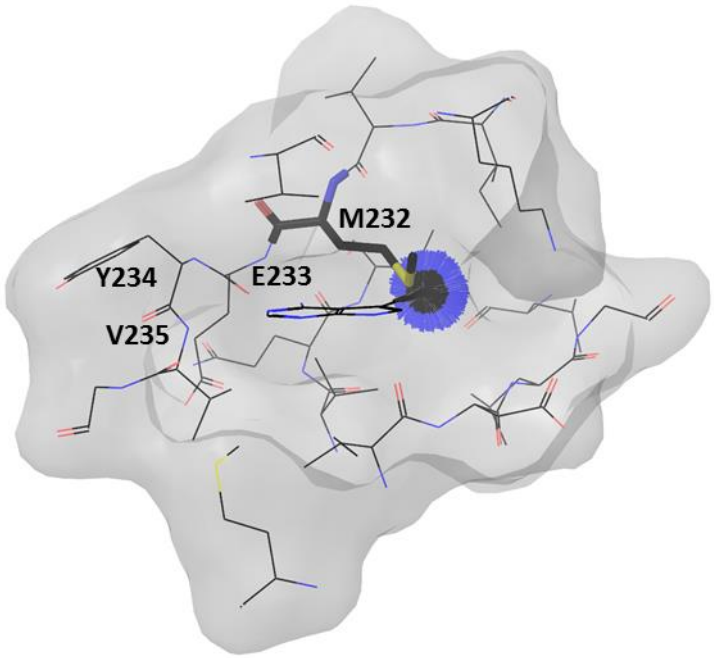

Figure 3.22. The wild-type PknG ligand-binding pocket (represented as a molecular surface) and 7d7p ATP conformers (represented as lines). Residues 232-235 formed the hinge region. Met232 is represented as stick.

The 7d7p ATP contains a $–NH_2$ group which is an H-bond donor and acceptor. We decided to mutate Met232 into residues that could be specifically involved in H-bonds. Met232 was changed into His, Ser and Thr in order to obtain three PknG mutants, M232H, M232S and M232T. The imidazole ring of a neutral His can act as H-bond donor (with the polar hydrogen) and H-bond acceptor (with the basic nitrogen) and hydroxyl groups of both Ser and Thr can act as donor and acceptor in H bonds (Figure 3.23). All pairs of PknGM232H and 7d7p ATP had positive GlideScores indicating unfavorable protein-ligand interaction. The mutagenesis of Met232 into His means changing a polar side chain into another that can be involved in H-bonds. On the other hand, Met and His side chain are almost equal in size (171 $Å^3$ vs 167$Å^3$) [137]. His at position 232 blocks the

access to the buried region preventing all 7d7p ATP conformers to occupy the PknGM232H ligand-binding site. Figure 3.24 represents the 7d7p ATP conformer with the worst interaction energy within the PknGM232H binding site ($13.63*10^7$ kcal/mol).

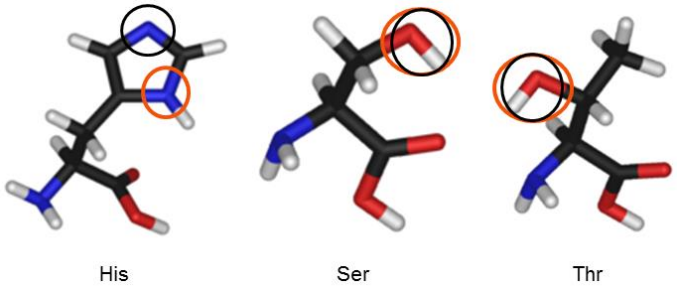

Figure 3.23. Three-dimensional structures of His, Ser and Thr. H-bond donor and H-bond acceptor are highlighted by black and orange circles, respectively.

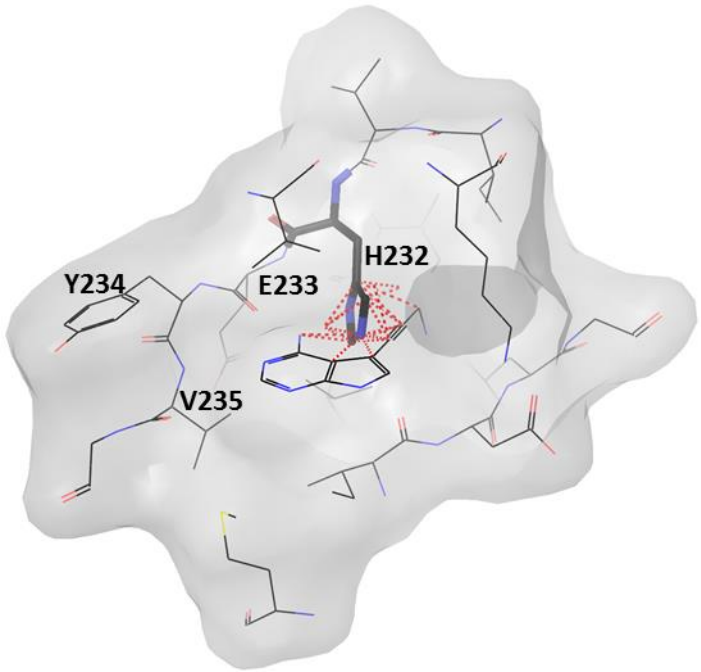

Figure 3.24. The PknGM232H ligand-binding pocket (represented as a molecular surface) and the 7d7p ATP conformer with the worst interaction energy (represented as lines). His232 is represented as stick and residues 232-235 represent the hinge region. Red dashed lines represent “ugly” protein-ligand contacts (Appendix A3).

Table 3.9 shows the results of the analysis of protein-ligand interactions for PknGM232S and PknGM232T with the 7d7p ATP conformers. The engineered PknGM232S had favorable GlideScores for 562 conformers whereas PknGM232T showed favorable interactions with 19 conformers.

Table 3.9. GlideScore values of PknGM232S and PknGM232T with 7d7p ATP conformers.

	PknGM232S conformers (1000 conformers)		**PknGM232T conformer (1000 conformer)**	
	1 to 562	others	1 to 19	others
GlideScore (kcal/mol)	-4.92 to -3.95	10 000	-5.18 to -1.42	10 000

Figure 3.25 A and B represent the complexes of PknGM232S and 7d7p ATP and PknGM232T and 7d7p ATP with the best GlideScores (energies of interaction equal to -10 kcal/mol and -7.03 kcal/mol, respectively).

Within the PknGM232S-7d7p ATP complex, Ser232 modelled to be involved in an H-bond with the deazapurine ring (nitrogen N6) together with Glu233 and Val235 (with N6 and N1, respectively). The primary amine of 7d7p ATP was predicted to be involve in two H-bonds with backbone nitrogens of Leu294 and Gly295 (Figure 3.25 A). The mutagenesis of Met232 into the smaller Ser allowed some conformers to access the buried region of the ligand-binding pocket. On the other hand, Ser232 modelled to be involved in a specific H-bond with the deazapurine ring and not with the propargylamine of the ATP analogue. Among the 562 conformers that favorably interacted with the engineered PknGM232S, some of them were characterized by the presence of H-bonds between Ser232 side chain and the propargylamine. Figure 3.25 C represents the complex of PknGM232S and 7d7p ATP with a specific H-bond between Ser232 and the propargylamine and the more favorable GlideScore (-4.52 kcal/mol). The interaction energy for such complex was -6.61 kcal/mol, a bit lower than the energy of interaction between the mutant and the conformer with the best GlideScore (-10 kcal/mol).

The mutation of Met232 into Thr enlarged the ligand-binding pocket making the buried region accessible to some 7d7p ATP conformers. Figure 3.24 B shows the complex of the engineered PknGM232T in complex with a 7d7p ATP conformer

for which we got the best GlideScore. The hydroxyl group of Thr232 acted as H-bond donor interacting with N6 on the deazapurine ring. The $–NH_2$ group of 7d7p ATP formed H-bonds with the backbone nitrogen of Leu294 and the backbone oxygen of Ile292 (Figure 3.24 B). Analyzing all mutant-analogue pairs with favorable GlideScores, we found that Thr232 is always involved in H-bonds with the deazapurine moiety and not the propargylamine of 7d7p ATP.

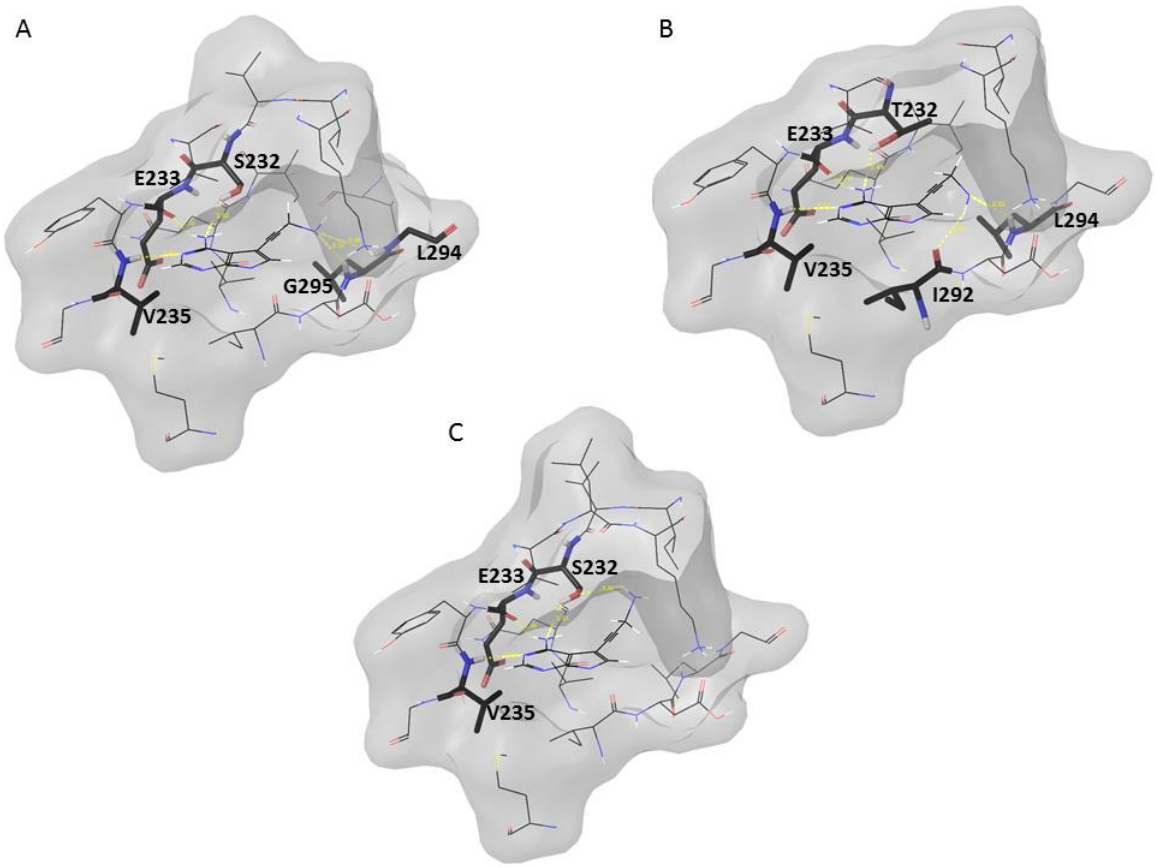

Figure 3.25. A) The PknGM232S ligand-binding pocket (represented as a molecular surface) and the 7d7p ATP conformer with the best GlideScore (represented as line). B) The PknGM232T ligand-binding pocket (represented as a molecular surface) and the 7d7p ATP conformer with the best GlideScore. C) The PknGM232S ligand-binding pocket (represented as a molecular surface) and the 7d7p ATP conformer where the propargylamine interacts with Ser232 (represented as line). Residues involved in the H-bonds are represented as sticks. Polar hydrogens are colored in white. Yellow dashed lines represent protein-ligand H-bonds (Appendix A3). For each conformer only the deazapurine ring is represented.

The ability of PknGM232S and PknGM232T to use 7d7p ATP as cofactor was experimentally tested (Figure 3.26 A and B). Similar results were obtained for PknGM232S and PknGM232T. The mutagenesis of Met232 into Ser or Thr did not affect the kinase activity of the protein as both mutants showed a phosphorylation activity. At all concentrations 7d7p ATP acted as co-substrates for PknGM232S and PknGM232S but not for wild-type PknG. While 7d7p ATP specifically bound to the two engineered PknG and not to the wild-type kinases, PknGM232S and PknGM232T can use both native and analogue ATP as phosphodonor. Their phosphorylation ability was almost the same with the native ATP and the ATP analogue, even at low ligand concentrations. The experiments were repeated twice and the second one confirmed the first result obtained.

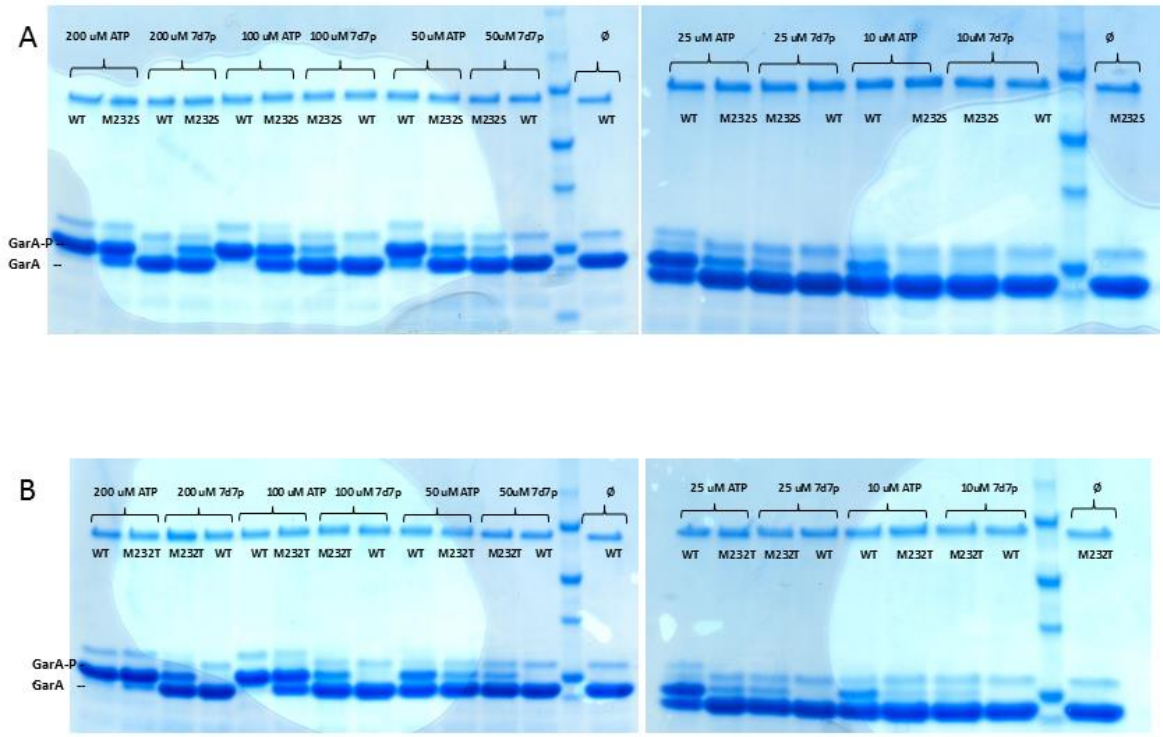

Figure 3.26. *In vitro* kinase assay (SDS-PAGE) using wild-type PknGΔN (WT), PknGΔN-M232S (M232S) (A) and PknGΔN-M232T (M232T) (B), GarA as substrate and ATP or 7d7p ATP (7d7p) as cofactors. GarA-P is the phosphorylated GarA. The symbol Ø means that neither ATP nor 7d7p are used in those lanes.

In the method developed by Shokat and coworkers, the gatekeeper residue is always converted into a small amino acid, generally Gly or Ala, to enlarge the

ligand-binding site that can accept ligands with bulky substituent groups. Applying our computational protocol on PknG and 7d7p ATP, we identified Met232, the gatekeeper residue, as potential candidate for single-point mutagenesis. We mutated Met232 into Ser and Thr in order to enlarge the ligand-binding site and to also take the advantage of their ability to participate to H-bonds. We compared the PknGM232S–7d7p ATP and PknGM232T–7d7p ATP complexes to the complex of a Shokat-like mutant, PknGM232G, and 7d7p ATP. Figure 3.27 A, B and C represent the interaction of PknGM232S, PknGM232T and PknGM232G with 7d7p ATP conformers for which we got the best GlideScores. In both PknGM232S and PknGM232T, Ser and Thr are involved into H-bonds with the deazapurine ring. Moreover, the methyl group of Thr232 is involved in a hydrophobic interaction with the triple bond of the propargylamine group. In the PknGM232G–7d7p ATP complex, Gly232 is not involved in any specific interaction. The mutagenesis of Met232 into Gly just enlarges the binding pocket allowing 7d7p ATP to fit in the engineered pocket. That analysis showed that PknGM232S and PnkGM232T were involved into specific interactions with 7d7p ATP with respect to the Shokat-like mutant, PknGM232G. Those outcomes were tested by *in vitro* kinase assays (Figure 3.28 A and B). The three mutants (PknGM232G, PknGM232S and PknGM232T) used ATP as cofactor to phosphorylate GarA. At concentration values ranging from 200 μM to 50 μM, PknGM232S and PknGM232T used 7d7p ATP as co-substrate whereas PknGM232G did not show any catalytic activity in presence of the ATP-competitive ligand.

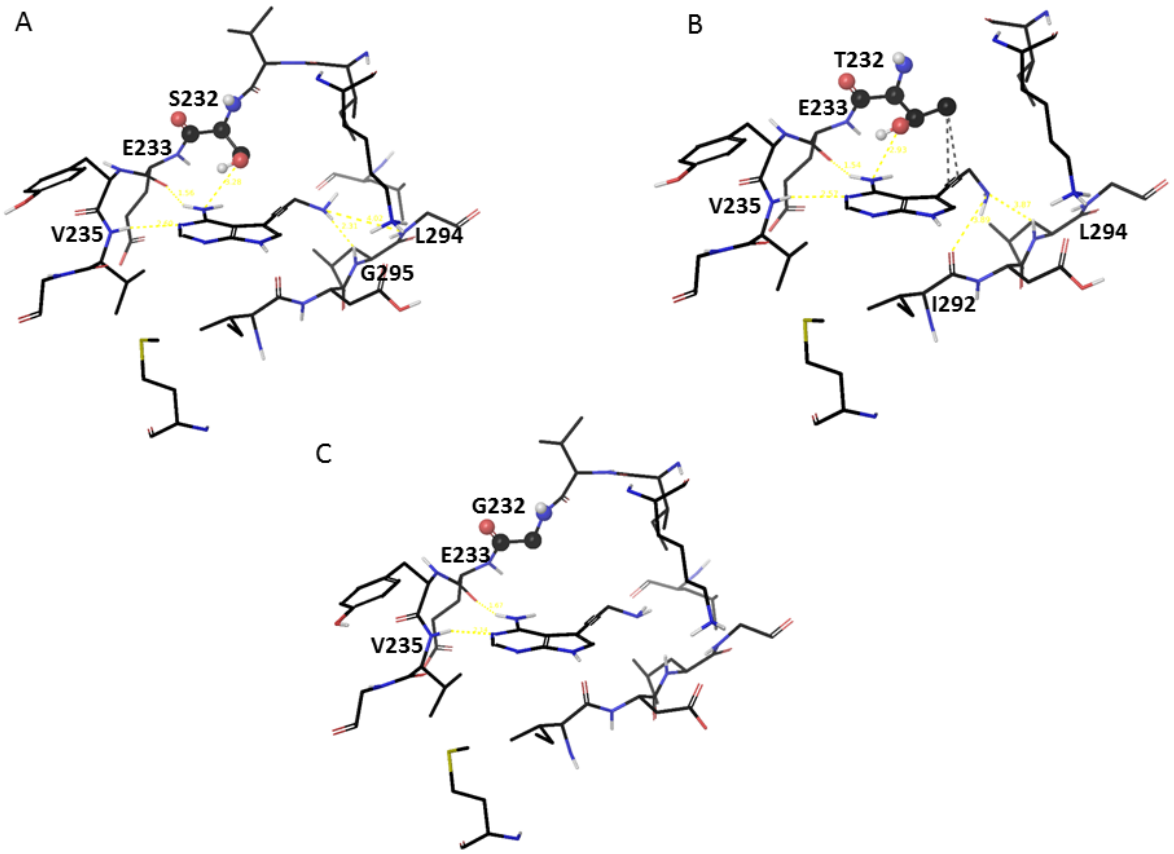

Figure 3.27. A) The PknGM232S ligand-binding pocket resides and the 7d7p ATP conformer with the best GlideScore (represented as lines). Ser232 is represented in balls and sticks. B) The PknGM232T ligand-binding pocket resides and the 7d7p ATP conformer with the best GlideScore (represented as lines). Thr232 is represented in balls and sticks. C) The PknGM232G ligand-binding pocket resides and the 7d7p ATP conformer with the best GlideScore (represented as lines). Gly232 is represented in balls and sticks. Polar hydrogens are colored in white. Yellow dashed lines represent protein-ligand H-bonds and gray dashed lines represent protein-ligand hydrophobic interactions (Appendix A3). Only the deazapurine ring is represented.

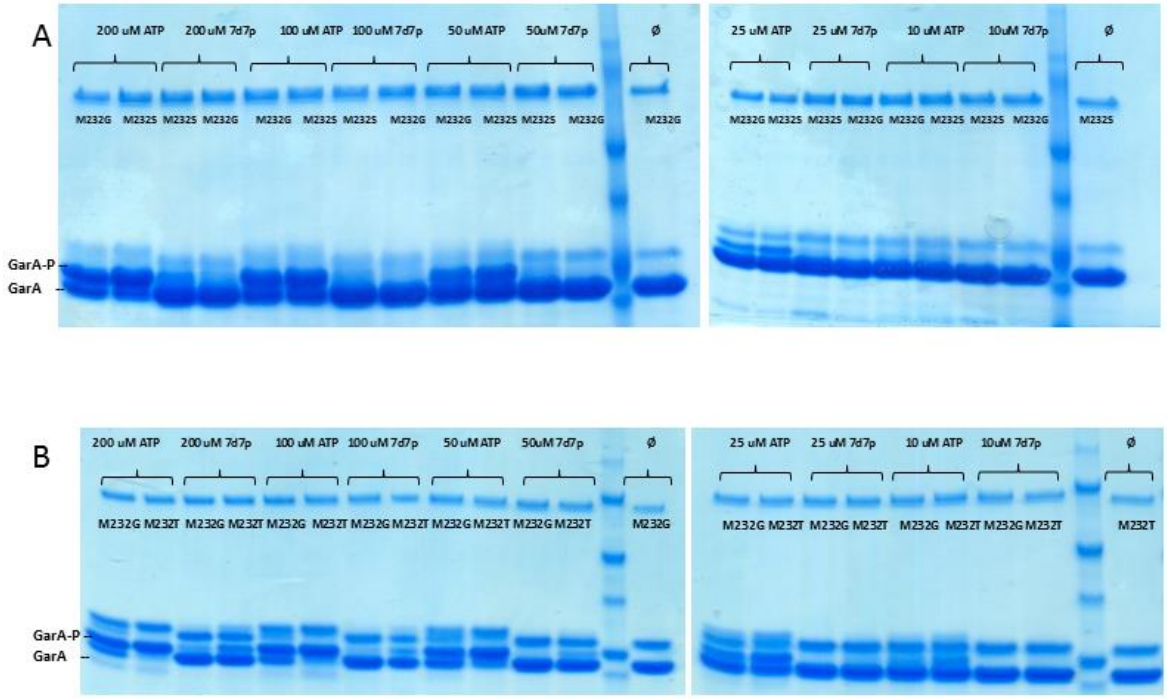

Figure 3.28. *In vitro* kinase assay (SDS-PAGE) using PknGΔN-M232S (M232S), PknGΔN-M232G (M232G) (A) and PknGΔN-M232T (M232T), PknGΔN-M232T (M232T) (B), GarA as substrate and ATP or 7d7p ATP (7d7p) as cofactors. GarA-P is the phosphorylated GarA. The symbol Ø means that neither ATP nor 7d7p are used in those lanes.

4 Discussion

We use the method developed by Shokat and coworkers [18, 134] as starting point to develop a computational protocol for protein engineering. Our protocol aims to identify binding-site residues of the target kinase that could be mutated to accommodate a specific ATP analogue as co-substrate, without interfering with the catalytic activity of the kinase protein. It could be used as prescreening step in the context of the complex process of kinase protein substrate identification.

4.1 Evaluation of the protein-ligand interaction

The protocol is organized in two parts. The first part consists in the prediction of residues to mutate within the ligand-binding site of the protein kinase of interest to generate kinase mutants. The second part is the evaluation of the interaction between kinase mutants and ligand analogues by a protein-ligand scoring function (Figure 2.1).

To identify the best protein-ligand scoring function for our protocol, we tested three different scoring functions, GlideScore, X-Score and DSX. The system studied by Shokat and coworkers in their first papers, v-Src and N6-(benzyl) ATP, was used to test the three scoring functions. We identified v-SrcI338G being a potential candidate for v-Src engineering. The interaction of v-SrcI338G with a group of N6-(benzyl) ATP conformers was evaluated by the three scoring functions. Only GlideScore was able to distinguish between conformers that occupied the engineered binding pocket without any clash and those clashing with residues belonging to the engineered binding site. We think that the reason why X-Score and DSX did not detect some protein-ligand clashes is due to their Van der Waals (vdW) interactions terms as compared to the GlideScore vdW term. In GlideScore, the vdW term is given by the Lennard-Jones 12-6 equation (equation 2.7) and both heavy and hydrogen atoms contribute to the term. In X-Score, the vdW term is described by a ‘softer’ Lennard-Jones 8-4 potential (equation 2.12) and only heavy atoms contribute to this term. Moreover, to avoid huge repulsion

due to overlapping atom pairs there is an upper limit for the vdW term. For all pairs of atoms that exceed that limit the vdW term is cut flat to the upper limit value. For that reason X-Score is not particularly good in identifying protein-ligand clashes [72]. Lastly, the DSX distance-dependent pair potential is characterized by the fact that in case of low contact distances and overlaps the density function of the reference state is not well defined because no structural data are available in this distance range neither in PDB nor in CSD. For that reason, DSX is much better in rescoring poses obtained by docking programs that already identify and eliminate poses with protein-ligand clashes [68]. We found that the analysis of the vdW terms in the three different scoring functions well explain the different results obtained and justify the choice of GlideScore as scoring function in the evaluation part of the protocol.

In the case of v-Src and N6-(benzyl) ATP, GlideScore allows for the identification of four N6-(benzyl) ATP conformers that favorably interact with the engineered v-SrcI338G (Table 3.4). We then performed MM-GBSA calculations to further confirm those findings. For each v-SrcI338G–N6-(benzyl) ATP pair, GlideScore focusses on a single bound conformation whereas the MM-GBSA approach uses MD simulations to generate an ensemble of bound conformations. There is an agreement between GlideScore and MM-GBSA results (Table 3.4). In general, MM-GBSA has a lower computational costs compared to FEP and TI and provides a more sophisticated computation of free energy compared to common scoring functions. Simple scoring functions may neglect significant contribution to the final energy [138]. Moreover, it has been proven that the best results are obtained when MM-GBSA is applied to an ensemble of ligands with high structural similarity, such as ligand conformational isomers [139-141]. N6-(benzyl) ATP conformers are isomers that can be interconverted by rotation around single bonds of the substituent group and thus they are suitable ligands to which apply MM-GBSA calculations. By applying MM-GBSA on such an ensemble, we can distinguish conformers that can accommodate the ligand-binding site from those that do not fit within the binding pocket. Those results confirm the GlideScore findings, supporting our choice of using GlideScore as protein-ligand scoring function in our computational strategy.

4.2 Application of the computational protocol

In the first part of the thesis, the protocol was tested on different tyrosine and serine/threonine protein kinases from the scientific literature where Shokat's method was applied and experimental data were available. The strategy well distinguishes residues that prevent the access of ATP-competitive ligands within the binding site, such as Ile338 of v-Src whose side chain preclude the accommodation of N6-(benzyl) ATP, from residues that do not interfere with the accommodation of analogues, such as Thr339 of Fyn which allow the binding of PP1 without being mutated. The method correlates well with published experimental data available for the tested protein kinases. The predicted interaction energies of v-Src, v-SrcI338A and v-SrcI338G with ATP and N6-(benzyl) ATP have the same trend as the kcat/Km ratio (Figure 3.6 A and B). Given N6-(benzyl) ATP, v-SrcI338G is identified as the best binder to the ATP analogue in agreement with experimental findings of Shokat [142]. Wild-type JNK does not accommodate any of the four N6-(substituent) ATPs and Met108 and Leu168 within JNK binding pocket are identified as potential candidates for double mutagenesis. Our method allows for the selection of N6-(2-phenythyl) ATP as the best ATP analogue to bind the engineered JNKM108GL168A (Figure 3.8) and that finding agrees with the experimental identification of N6-(2-phenythyl) ATP as the analogue with the highest affinity to the engineered JNK [27]. For both Src tyrosine and serine/threonine kinase families we obtain a good correlation between the experimental pIC_{50} and the predicted interaction energies, with correlation coefficient of 0.85 and 0.75, respectively (Figure 3.11 A and B). The predicted energies of interaction sufficiently estimate the trend of the inhibitor potency. In the case of PP1, the strategy discriminates well between proteins or mutants that are inhibited by the ligand analogue, such as Fyn or CamKIIF89, and proteins or mutants that are not inhibited, like the cases of Cdk2 or v-SrcI338F (Table 3.1).

In the second part of the thesis, we apply our computational protocol on PknG and five purchasable ATP analogues (N6-(methyl) ATP, N6-(benzyl) ATP, 8-Azido ATP, PF9 and 7d7p ATP). We aimed to identify residues to mutate within the

PknG ligand-binding site to generate mutants that are catalytically active and that specifically and preferentially use ATP-competitive ligands as phosphodonors. Table 4.1 shows the results obtained.

Table 4.1. Computational predictions and results of *in vitro* kinase assays for PknG and five purchasable ATP analogues. For each analogue is reported if it is a good candidate as ATP analogue, which PknG mutants are tested, results of computational prediction and results of experimental assays.

ATP analogues	Good candidate	PknG mutants	Predicted interaction	Experimentally tested	Experimental validation
N6-(methyl) ATP	X	-	-	-	-
N6-(benzyl) ATP	√	V211G	X	-	-
		M232G	√	√	√
8-Azido ATP	X	-	-	-	-
PF9	√	Y234G	X	-	-
		V235G	√	√	X
7d7p ATP	√	M232H	X	-	-
		M232S	√	√	√
		M232T	√	√	√

An ATP analogue is a good co-substrate for an engineered kinase if it does not bind to the ligand-binding pocket of the wild-type kinase but just to that of the engineered kinase.

Among the group of five analogues examined, N6-(methyl) ATP and 8-Azido ATP are inadequate candidates as ATP analogues as both ligands fit into the wild-type PknG ATP binding site without any clash. In N6-(methyl) ATP the methyl group is small enough to occupy the adenine region without any overlap whereas in 8-Azido ATP the azido ion has access to the phosphates region (Figure 3.14 and 3.19).

PF9, N6-(benzyl) ATP and 7d7p ATP are good candidates as cofactors for PknG mutants. PF9 fits into the binding site of PknGV235G and the complex has a favorable predicted interaction energy (-23.3 kcal/mol). The *in vitro* kinase assay shows that PknGV235G has no significant activity with PF9 whereas its phosphorylation activity is very significant with native ATP, and this is true at all cofactor concentrations (Figure 3.21). The experimental results do not confirm our prediction showing that PknGV235G preferentially uses native ATP as

phosphodonor. To understand such disagreement, we analyzed in more details the structure of PknGV235G in complex with PF9 (Figure 4.1 A and B). The mutation of the hydrophobic Val235 into Gly allows the accommodation of the ligand analogue within the engineered binding pocket. Even though the phenylethynyl group is involved in hydrophobic interactions with Tyr234 and Ile157, the benzyl ring occupies the solvent accessible region of the ATP binding pocket that generally is not occupied by ATP and is opened to the solvent (Figure 4.1 B). In evaluating the interaction between engineered kinases and ATP-competitive ligands, we do not take into account the solvent. No models are used to simulate the solvent and its role in the protein-ligand interaction. This approximation could explain the favorable interaction energy we get for the PknGV235G–PF9 complex. The scoring function only takes into consideration the favorable protein-ligand hydrophobic interactions and does not consider the unfavorable placement of the hydrophobic benzyl ring in a hydrophilic environment.

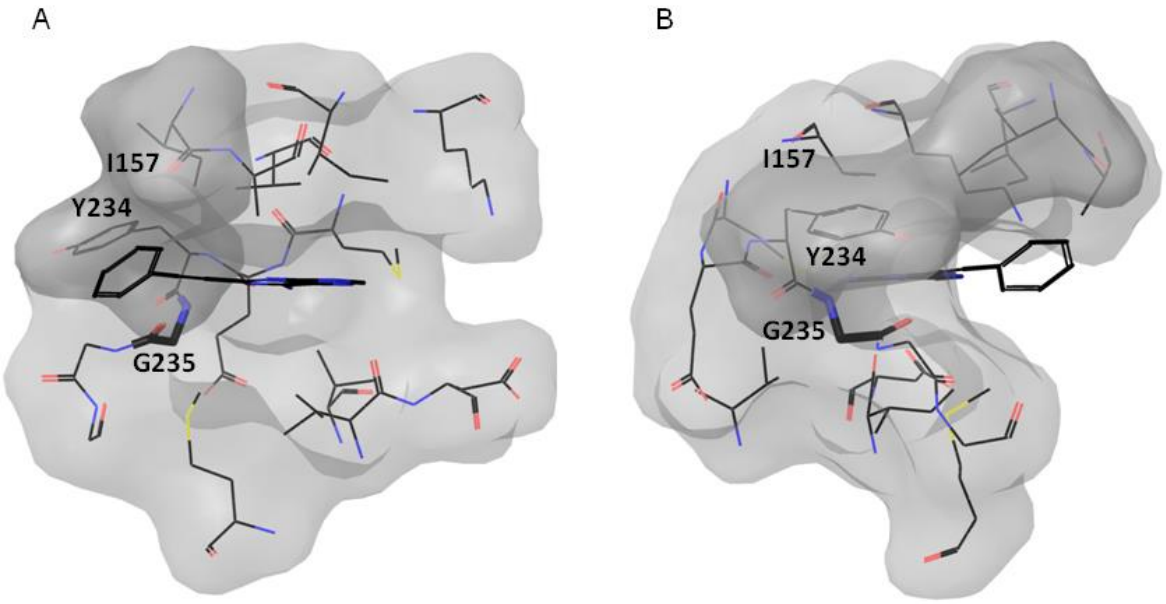

Figure 4.1. A) The PknGV235G ATP binding pocket (represented as a molecular surface) and the PF9 conformer with best GlideScore (represented as lines). Residues involved in hydrophobic interactions are represented as darker molecular surface (Tyr234 and Ile157). Gly235 (the mutated residue) is represented as stick. B) The same image rotated 90° with respect to A.

N6-(benzyl) ATP reveals a favorable interaction with PknGM232G (predicted energy is -24.20 kcal/mol). The *in vitro* kinase assay shows that at cofactor concentrations ranging from 50 μM to 10 μM there is a discrimination between the phosphorylation activity of PknGM232G with ATP and with N6-(benzyl) ATP. The engineered PknG preferentially uses the ATP-analogue as cofactor to phosphorylate GarA. On the other hand, wild-type PknG preferentially uses native ATP and this occurs at all cofactor concentrations (Figure 3.18). The experimental assay confirmed the *in silico* prediction made and those results suggested that PknGM232G and N6-(benzyl) ATP are good candidates for follow-up *in vivo* experiments.

Favorable interaction energies were predicted for 7d7p ATP in complex with PknGM232S and PknGM232T (-10 kcal/mol and -7.03 kcal/mol, respectively). The gatekeeper Met232 is mutated into the smaller Ser and Thr allowing the ATP analogue to access the buried region of the ligand-binding pocket. The mutagenesis of Met into Ser and Thr not only creates an additional space within the binding pocket but also introduce two amino acids that can be involved in specific interactions (H-bonds) with the ligand analogue. The *in vitro* kinase assay shows the ability of PknGM232S and PknGM232T to use 7d7p ATP as co-substrate in contrast to PknGM232G that preferentially uses the native ATP (Figure 3.28 A and B). Despite the agreement between the computational predictions and *in vitro* results for PknGM232S and PknGM232T, both mutants show the same preference to interact with native ATP and ATP analogue (Figure 3.26 A and B). Those two designed pairs are thus not suitable for following *in vivo* assays.

Our computational analysis together with the *in vitro* kinase assays recognize the PknGM232G–N6-(benzyl) ATP pair as an interesting combination for following experimental tests. The *in vitro* experiment shows that PknGM232G is catalytically active and preferentially uses N6-(benzyl) ATP as cofactor. Therefore, using this designed pair might result in the experimental identification of the specific substrates involved in the PknG infective pathway. A similar approach provided the identification of cofilin and calumenin as specific v-Src substrates [20].

Our computational approach can be used as prescreening test in the experimental procedure for the kinase substrates identification. The advantage of a computational prescreening is that it reduces the number of experimental assays to perform resulting in a significant reduction of the time and the cost of the whole experiment. In the specific case of PknG, we computationally tested a group of five ATP analogues and three of them were identified as good candidates as ATP-competitive ligands. Of the all possible PknG mutants that would be obtained by mutating residues within the ligand-binding site, seven were predicted as potentially engineerable PknGs and only four of them are identified as interesting targets to experimentally test (Table 4.1). In conclusion, only four designed pairs were considered promising and further experimentally tested.

4.3 Advantages and limitations of the computational protocol

The main goal of our study is to engineer a protein kinase so that it is catalytic active and specifically uses an ATP-analogue as phosphodonor. To reach that goal, we used a single conformation of the target kinase derived from the crystal structure of a complex containing either the target protein bound to its native ligand or a molecule mimicking it. In case the solved structures were not available for the kinase of interest, a model could be built using as template the structure of a homologue with the features described above. At the same time, we were interested in ATP analogues that act as phosphodonors for engineered protein kinases. In short, not only ATP analogues had to adequately fit into the engineered binding sites but they also had to bind to those same binding sites in the right geometry. Each ATP derivative was modelled on the adenine, ribose and phosphates geometry of the native ATP within the kinase ligand-binding site, and for each analogue we did not take into account the total flexibility. The conformational space was not completely sampled. The search was performed by keeping fixed the adenine moiety, the ribose ring and the phosphates and by leaving the bonds of each substituent group free to rotate. The sample of the entire conformational space could provide conformers with more suitable energies

but catalytically useless conformations. As we were specifically looking for ATP analogues that bind to a given kinase and act as cofactors, the assumption of not considering the total flexibility for each analogue allowed the inspection of significant ensemble of conformers that were all potentially active as phosphodonors. Inactive conformations of ATP analogues were not analyzed.

Our computational protocol explores pairings of potential mutations and ligand analogues. The Shokat's method has been extensively used. Ubersax *et al* modified the gatekeeper Phe88 of the cyclin-dependent kinase CdK1 into Gly allowing the engineered kinase to bind to N6-(benzyl) ATP. The designed pair permitted to identify several Cdk1 substrates, such as the forkhead transcription factor Fhk2 [143]. In a similar approach, Eblen and coworkers used the mutant at the gatekeeper position of the extracellular signal-regulated kinase ERK2 ERK2Q103G, and N6-(cyclopentyl) ATP to identify the nucleoporin translocated promoter region (TPR) as one ERK2 specific substrate [29]. In these different studies, the method always involved the mutagenesis of the gatekeeper position into Gly and N6-(substituent) ATPs with bulky hydrophobic substituent groups. With respect to Shokat's method, our computational strategy did not exclusively focus on gatekeeper residues but instead inspected diverse residues belonging to the ligand-binding site to identify potential candidates for single-point mutations. On one hand, this allows for the possibility to generate diverse mutants for a target kinase. On the other hand, it permits testing ATP analogues with substituent groups attacked on various positions of the adenine ring and that have disparate chemical features. However, once a non-gatekeeper residue is identified as potential candidate for mutagenesis other aspects have to be considered. For instance, if the potential candidate belongs to the phosphates region, it is necessary to verify its role in the kinase catalytic activity before changing it into another residue. Catalytic residues, such as the highly conserved Lys and Asp, are not appropriate candidates considering that their mutagenesis would result in the disruption of the catalytic activity of the protein of interest, therefore generating an inactive engineered kinase.

When the Shokat method is mainly based on the 'bump-and-hole' model, in our approach we also considered the chemical complementarity between the mutated

residue and the ATP derivative. In the case of PknG and 7d7p ATP, the *in vitro* experiments show that PknGM232S and PknGM232T bind to 7d7p ATP, whereas the pair PknGM232G–7d7p ATP, designed only by using the 'bump-and-hole' approach, does not work. These results support the idea of using both shape and chemical complementarity in designing combinations of engineered kinases and chemical derivatives of ATP. Nevertheless, the fact that PknGM232S and PknGM232T exhibit the same phosphorylation activity in the presence of ATP and 7d7p ATP shows that the two mutants equally bind to native ATP and ATP analogue. That finding suggests that the use of shape and chemical complementarities should be further explored. It would be interesting to design a pair characterized by a specific interaction between the side chain of the mutated residue and the substituent group of the ATP derivative. This would allow to test if the design of a precise protein-ligand interaction confers to the engineered protein the ability to specifically bind to the ATP derivative and not to the native ATP.

A limitation of the method developed in this thesis is that it is a manual approach. Once a residue is identified as potential candidate for mutagenesis, we visually inspected the binding site to identify the amino acid that at that position could well interact with a given ATP analogue. The chemical space is not totally sampled as not all twenty amino acids are tested at a given position of the binding site. This prevented the possibility of exploring all possible promising mutant-ligand pairs as it would be performed by a systematic automated approach.

Another drawback of our approach is that in modeling and evaluating pairs of engineered kinase and ATP analogue, we did not include water and its effects in protein-ligand binding. Water molecules play a key role in protein-ligand binding and to neglect them could affect the binding process, for instance by excluding interactions in which water molecules are involved or by considering accessible to the ligand an area of the binding site that indeed is involved in interaction with the solvent. This is the case, for instance, of the complex of PknGV235G and PF9 when ignoring water results also in overlooking the placement of a hydrophobic group, the phenylethynyl within a polar environment. Therefore, we predicted a favorable energy of interaction that was not confirmed by *in vitro* experiments. A way to overcome such limitation might be to use theoretical approaches for

solvent modeling. One example is the method developed by Bottoms and coworkers in 2006. This method identifies hydration sites (sites where water molecules are located) in protein binding sites using the conservation of those sites across protein families [144]. Similarly, Michel and coworkers developed a procedure called ‘Just Add Water Molecules’ (JAWS) that allows to find the number and position of water sites in protein binding sites [145].

5 Summary

In this thesis we have successfully developed a computational protocol for designing pairs of engineered kinases and ATP-competitive ligands. Our method can be used as prescreening test in the wider experimental procedure for the kinase substrates identification. The reliability of the protocol was tested on different tyrosine and serine/threonine protein kinases from the scientific literature where Shokat's method was applied and experimental data were available. On one side the protocol provided a relative rank of suitable ATP analogues for a given engineered kinase, on the other hand it provided a way to evaluate for a given ATP-competitive ligand which mutations within the kinase binding site would be compatible. The method successfully correlated with published experimental data and this suggested the possibility to apply it on a new system for which no data were available.

We applied our protocol to the *Mtb* PknG. The identification of *Mtb* PknG downstream substrates would address the open questions concerning its mode of action and its role in the survival of the pathogen within the host organism. We tested five purchasable ATP analogues. Four pairs were identified being promising combinations of mutation in the PknG binding site and ATP derivative and were further tested in *in vitro* assays. The ability of a PknG mutant to preferentially use an ATP analogue as phosphodonor, was confirmed only by one pair, PknGM232G–N6-(benzyl) ATP. That pair is the analogue to the Shokat classic pair, where the gatekeeper residue is mutated to Gly and the ATP analogue has a hydrophobic group at position N6 of the adenine ring. Different pairs computationally identified (PknGV235G-PF9, PknGM232S–7d7p ATP and PknGM232T–7d7p ATP) were not confirmed promising by *in vitro* testes PknGM232G–N6-(benzyl) ATP is currently tested in an *ex vivo* context by the group of the Prof. J. Pieters, at Biozentrum. With results obtained in this study, we hope to clarify the pathway regulated by PknG and help in the development of new therapies to combat such infectious diseases.

References

1. Zhang, W.H., G. Otting, and C.J. Jackson, *Protein engineering with unnatural amino acids.* Current Opinion in Structural Biology, 2013. **23**(4): p. 581-587.
2. Saven, J.G., *Computational protein design: engineering molecular diversity, nonnatural enzymes, nonbiological cofactor complexes, and membrane proteins.* Curr Opin Chem Biol, 2011. **15**(3): p. 452-7.
3. Li, X., Z. Zhang, and J. Song, *Computational enzyme design approaches with significant biological outcomes: progress and challenges.* Comput Struct Biotechnol J, 2012. **2**: p. e201209007.
4. Tiwari, M.K., et al., *Computational approaches for rational design of proteins with novel functionalities.* Comput Struct Biotechnol J, 2012. **2**: p. e201209002.
5. Li, Y. and P.C. Cirino, *Recent advances in engineering proteins for biocatalysis.* Biotechnology and Bioengineering, 2014. **111**(7): p. 1273-1287.
6. Kaplan, J. and W.F. DeGrado, *De novo design of catalytic proteins.* Proc Natl Acad Sci U S A, 2004. **101**(32): p. 11566-70.
7. Lazar, G.A., et al., *Engineered antibody Fc variants with enhanced effector function.* Proc Natl Acad Sci U S A, 2006. **103**(11): p. 4005-10.
8. Cochran, F.V., et al., *Computational de novo design and characterization of a four-helix bundle protein that selectively binds a nonbiological cofactor.* Journal of the American Chemical Society, 2005. **127**(5): p. 1346-1347.
9. Richter, F., et al., *De Novo Enzyme Design Using Rosetta3.* PLoS One, 2011. **6**(5).
10. Rothlisberger, D., et al., *Kemp elimination catalysts by computational enzyme design.* Nature, 2008. **453**(7192): p. 190-5.
11. Jiang, L., et al., *De novo computational design of retro-aldol enzymes.* Science, 2008. **319**(5868): p. 1387-91.
12. Siegel, J.B., et al., *Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction.* Science, 2010. **329**(5989): p. 309-13.
13. Schreier, B., et al., *Computational design of ligand binding is not a solved problem.* Proceedings of the National Academy of Sciences of the United States of America, 2009. **106**(44): p. 18491-18496.
14. Baker, D., *An exciting but challenging road ahead for computational enzyme design.* Protein Sci, 2010. **19**(10): p. 1817-9.

15. Hanks, S.K. and T. Hunter, *Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification.* FASEB journal : official publication of the Federation of American Societies for Experimental Biology, 1995. **9**(8): p. 576-96.
16. Cheng, H.C., et al., *Regulation and function of protein kinases and phosphatases.* Enzyme Res, 2011. **2011**: p. 794089.
17. Shah, K., et al., *Engineering unnatural nucleotide specificity for Rous sarcoma virus tyrosine kinase to uniquely label its direct substrates.* Proceedings of the National Academy of Sciences of the United States of America, 1997. **94**(8): p. 3565-70.
18. Liu, Y., et al., *Engineering Src family protein kinases with unnatural nucleotide specificity.* Chem Biol, 1998. **5**(2): p. 91-101.
19. Liu, Y., et al., *A molecular gate which controls unnatural ATP analogue recognition by the tyrosine kinase v-Src.* Bioorganic & Medicinal Chemistry, 1998. **6**(8): p. 1219-26.
20. Shah, K. and K.M. Shokat, *A chemical genetic screen for direct v-Src substrates reveals ordered assembly of a retrograde signaling pathway.* Chem Biol, 2002. **9**(1): p. 35-47.
21. Cravatt, B.F., *Kinase chemical genomics--a new rule for the exceptions.* Nature Methods, 2005. **2**(6): p. 411-2.
22. Bucci, M., C. Goodman, and T.L. Sheppard, *A decade of chemical biology.* Nature Chemical Biology, 2010. **6**(12): p. 847-854.
23. Dephoure, N., et al., *Combining chemical genetics and proteomics to identify protein kinase substrates.* Proc Natl Acad Sci U S A, 2005. **102**(50): p. 17940-5.
24. Kraybill, B.C., et al., *Inhibitor scaffolds as new allele specific kinase substrates.* Journal of the American Chemical Society, 2002. **124**(41): p. 12118-12128.
25. Larochelle, S., et al., *Dichotomous but stringent substrate selection by the dual-function Cdk7 complex revealed by chemical genetics.* Nat Struct Mol Biol, 2006. **13**(1): p. 55-62.
26. Hindley, A.D., et al., *Engineering the serine/threonine protein kinase Raf-1 to utilise an orthogonal analogue of ATP substituted at the N6 position.* FEBS letters, 2004. **556**(1-3): p. 26-34.
27. Habelhah, H., et al., *Identification of new JNK substrate using ATP pocket mutant JNK and a corresponding ATP analogue.* J Biol Chem, 2001. **276**(21): p. 18090-5.

28. Juris, S.J., et al., *Identification of otubain 1 as a novel substrate for the Yersinia protein kinase using chemical genetics and mass spectrometry.* Febs Letters, 2006. **580**(1): p. 179-183.
29. Eblen, S.T., et al., *Identification of novel ERK2 substrates through use of an engineered kinase and ATP analogs.* Journal of Biological Chemistry, 2003. **278**(17): p. 14926-14935.
30. Berg, J.M., J.L. Tymoczko, and L. Stryer, *Biochemistry*. 5th ed. 2002, New York: W.H. Freeman. xxxviii, 974, [72] p.
31. Knighton, D.R., et al., *Crystal-Structure of the Catalytic Subunit of Cyclic Adenosine-Monophosphate Dependent Protein-Kinase.* Science, 1991. **253**(5018): p. 407-414.
32. Scheeff, E.D., et al., *Structure of the Pseudokinase VRK3 Reveals a Degraded Catalytic Site, a Highly Conserved Kinase Fold, and a Putative Regulatory Binding Site.* Structure, 2009. **17**(1): p. 128-138.
33. Huse, M. and J. Kuriyan, *The conformational plasticity of protein kinases.* Cell, 2002. **109**(3): p. 275-82.
34. Hubbard, S.R. and J.H. Till, *Protein tyrosine kinase structure and function.* Annu Rev Biochem, 2000. **69**: p. 373-98.
35. Boggon, T.J. and M.J. Eck, *Structure and regulation of Src family kinases.* Oncogene, 2004. **23**(48): p. 7918-27.
36. Fang, Z., C. Grutter, and D. Rauh, *Strategies for the selective regulation of kinases with allosteric modulators: exploiting exclusive structural features.* ACS Chem Biol, 2013. **8**(1): p. 58-70.
37. Cheek, S., H. Zhang, and N.V. Grishin, *Sequence and structure classification of kinases.* J Mol Biol, 2002. **320**(4): p. 855-81.
38. Traxler, P. and P. Furet, *Strategies toward the design of novel and selective protein tyrosine kinase inhibitors.* Pharmacol Ther, 1999. **82**(2-3): p. 195-206.
39. Vulpetti, A. and R. Bosotti, *Sequence and structural analysis of kinase ATP pocket residues.* Farmaco, 2004. **59**(10): p. 759-65.
40. Mao, L., et al., *Molecular determinants for ATP-binding in proteins: a data mining and quantum chemical analysis.* J Mol Biol, 2004. **336**(3): p. 787-807.
41. Cappello, V., A. Tramontano, and U. Koch, *Classification of proteins based on the properties of the ligand-binding site: the case of adenine-binding proteins.* Proteins, 2002. **47**(2): p. 106-15.

42. Babor, M., V. Sobolev, and M. Edelman, *Conserved positions for ribose recognition: importance of water bridging interactions among ATP, ADP and FAD-protein complexes.* Journal of molecular biology, 2002. **323**(3): p. 523-32.
43. Klebe, G. and T. Mietzner, *A fast and efficient method to generate biologically relevant conformations.* J Comput Aided Mol Des, 1994. **8**(5): p. 583-606.
44. Adams, J.A., *Kinetic and catalytic mechanisms of protein kinases.* Chem Rev, 2001. **101**(8): p. 2271-90.
45. Xu, W., et al., *Crystal structures of c-Src reveal features of its autoinhibitory mechanism.* Mol Cell, 1999. **3**(5): p. 629-38.
46. Zuccotto, F., et al., *Through the "gatekeeper door": exploiting the active kinase conformation.* Journal of medicinal chemistry, 2010. **53**(7): p. 2681-94.
47. Huang, D., et al., *Kinase selectivity potential for inhibitors targeting the ATP binding site: a network analysis.* Bioinformatics, 2010. **26**(2): p. 198-204.
48. Noble, M.E.M., J.A. Endicott, and L.N. Johnson, *Protein kinase inhibitors: insights into drug design from structure.* Science (New York, N Y), 2004. **303**(5665): p. 1800-5.
49. Azam, M., et al., *Activation of tyrosine kinases by mutation of the gatekeeper threonine.* Nat Struct Mol Biol, 2008. **15**(10): p. 1109-18.
50. Elphick, L.M., et al., *Using chemical genetics and ATP analogues to dissect protein kinase function.* Acs Chemical Biology, 2007. **2**(5): p. 299-314.
51. B©œhm, H.-J. and G. Schneider, *Protein-ligand interactions from molecular recognition to drug design*. 2003, Weinheim: Wiley-VCH. xx, 242 p.
52. Tsai, N., Wolfson, Maizel and ussinov, *Protein–Ligand Interactions: Energetic Contributions and Shape Complementarity.* Encyclopedia of life sciences, 2001.
53. Kuriyan, J., B. Konforti, and D. Wemmer, *The molecules of life : physical and chemical principles*. xxii, 1008 pages.
54. Bronowska, A.K., *Thermodynamics of Ligand-Protein Interactions: Implications for Molecular Design* H.I.f.T.S. Heidelberg, Editor. 2011.
55. Gohlke, H. and G. Klebe, *Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors.* Angewandte Chemie (International ed in English), 2002. **41**(15): p. 2644-76.
56. Bissantz, C., B. Kuhn, and M. Stahl, *A medicinal chemist's guide to molecular interactions.* J Med Chem, 2010. **53**(14): p. 5061-84.
57. SanjavNilapwar, *Characterization and Exploitation of Protein Ligand Interactions for Structure Based Drug Design*. 2009, University College London

58. Lodish, H.F., et al., *Molecular cell biology*. Seventh edition, International edition / ed. xxxiii, 1154, G-26, I-31 pages.
59. Jeffrey, G.A., *An introduction to hydrogen bonding*. Topics in physical chemistry. 1997, New York ; Oxford: Oxford University Press. vii, 303 p.
60. Dubey, K.D., R.K. Tiwari, and R.P. Ojha, *Recent advances in protein-ligand interactions: molecular dynamics simulations and binding free energy.* Curr Comput Aided Drug Des, 2013. **9**(4): p. 518-31.
61. Gilson, M.K. and H.X. Zhou, *Calculation of protein-ligand binding affinities.* Annu Rev Biophys Biomol Struct, 2007. **36**: p. 21-42.
62. Jorgensen, W.L., *The many roles of computation in drug discovery.* Science, 2004. **303**(5665): p. 1813-8.
63. Berman, H.M., et al., *The Protein Data Bank.* Acta Crystallographica Section D-Biological Crystallography, 2002. **58**: p. 899-907.
64. Allen, F.H., *The Cambridge Structural Database: a quarter of a million crystal structures and rising.* Acta Crystallogr B, 2002. **58**(Pt 3 Pt 1): p. 380-8.
65. Kitchen, D.B., et al., *Docking and scoring in virtual screening for drug discovery: methods and applications.* Nat Rev Drug Discov, 2004. **3**(11): p. 935-49.
66. Huang, S.-Y., S.Z. Grinter, and X. Zou, *Scoring functions and their evaluation methods for protein-ligand docking: recent advances and future directions.* Physical chemistry chemical physics : PCCP, 2010. **12**(40): p. 12899-908.
67. Huang, S.Y. and X. Zou, *Advances and challenges in protein-ligand docking.* Int J Mol Sci, 2010. **11**(8): p. 3016-34.
68. Neudert, G. and G. Klebe, *DSX: a knowledge-based scoring function for the assessment of protein-ligand complexes.* J Chem Inf Model, 2011. **51**(10): p. 2731-45.
69. Muegge, I. and Y.C. Martin, *A general and fast scoring function for protein-ligand interactions: a simplified potential approach.* J Med Chem, 1999. **42**(5): p. 791-804.
70. DeWitte, R.S. and E.I. Shakhnovich, *SMoG: de Novo design method based on simple, fast, and accurate free energy estimates .1. Methodology and supporting evidence.* Journal of the American Chemical Society, 1996. **118**(47): p. 11733-11744.
71. Friesner, R.A., et al., *Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy.* J Med Chem, 2004. **47**(7): p. 1739-49.

72. Wang, R., L. Lai, and S. Wang, *Further development and validation of empirical scoring functions for structure-based binding affinity prediction.* J Comput Aided Mol Des, 2002. **16**(1): p. 11-26.
73. Bohm, 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): p. 243-56.
74. Goodsell, D.S., et al., *Automated docking in crystallography: analysis of the substrates of aconitase.* Proteins, 1993. **17**(1): p. 1-10.
75. Shoichet, B.K., A.R. Leach, and I.D. Kuntz, *Ligand solvation in molecular docking.* Proteins, 1999. **34**(1): p. 4-16.
76. Verdonk, M.L., et al., *Improved protein-ligand docking using GOLD.* Proteins-Structure Function and Genetics, 2003. **52**(4): p. 609-623.
77. Morris, G.M., et al., *Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function.* Journal of Computational Chemistry, 1998. **19**(14): p. 1639-1662.
78. Ewing, T.J., et al., *DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases.* J Comput Aided Mol Des, 2001. **15**(5): p. 411-28.
79. Lindorff-Larsen, K., et al., *Improved side-chain torsion potentials for the Amber ff99SB protein force field.* Proteins, 2010. **78**(8): p. 1950-8.
80. Cornell, W.D., et al., *A second generation force field for the simulation of proteins, nucleic acids, and organic molecules (vol 117, pg 5179, 1995).* Journal of the American Chemical Society, 1996. **118**(9): p. 2309-2309.
81. Jorgensen, W.L. and J. Tiradorives, *The Opls Potential Functions for Proteins - Energy Minimizations for Crystals of Cyclic-Peptides and Crambin.* Journal of the American Chemical Society, 1988. **110**(6): p. 1657-1666.
82. Kaminski, G.A., et al., *Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides.* Journal of Physical Chemistry B, 2001. **105**(28): p. 6474-6487.
83. Ponder, J.W. and D.A. Case, *Force fields for protein simulations.* Protein Simulations, 2003. **66**: p. 27-+.
84. Aqvist, J., V.B. Luzhkov, and B.O. Brandsdal, *Ligand binding affinities from MD simulations.* Acc Chem Res, 2002. **35**(6): p. 358-65.
85. Aqvist, J., C. Medina, and J.E. Samuelsson, *A new method for predicting binding affinity in computer-aided drug design.* Protein engineering, 1994. **7**(3): p. 385-91.

86. Srinivasan, J., et al., *Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate - DNA helices.* Journal of the American Chemical Society, 1998. **120**(37): p. 9401-9409.
87. Gohlke, H., C. Kiel, and D.A. Case, *Insights into protein-protein binding by binding free energy calculation and free energy decomposition for the Ras-Raf and Ras-RalGDS complexes.* Journal of molecular biology, 2003. **330**(4): p. 891-913.
88. Yuk, J.-M. and E.-K. Jo, *Host immune responses to mycobacterial antigens and their implications for the development of a vaccine to control tuberculosis.* Clinical and experimental vaccine research, 2014. **3**(2): p. 155-67.
89. Perkins, M., *Development, testing and introduction of new diagnostic tools for tuberculosis*, in *Tuberculosis*. 2014, Henry Stewart Talks: London. p. 1 online resource (1 streaming video file (35 min.).
90. Abubakar, I., et al., *Tuberculosis 2013:5 Drug-resistant tuberculosis: time for visionary political leadership.* Lancet Infectious Diseases, 2013. **13**(6): p. 529-539.
91. Walburger, A., et al., *Protein kinase G from pathogenic mycobacteria promotes survival within macrophages.* Science, 2004. **304**(5678): p. 1800-1804.
92. Nguyen, L. and J. Pieters, *The Trojan horse: survival tactics of pathogenic mycobacteria in macrophages.* Trends in cell biology, 2005. **15**(5): p. 269-276.
93. Warner, D.F. and V. Mizrahi, *The survival kit of Mycobacterium tuberculosis.* Nature Medicine, 2007. **13**(3): p. 282-284.
94. Cole, S.T., et al., *Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence.* Nature, 1998. **393**(6685): p. 537-+.
95. Av-Gay, Y. and M. Everett, *The eukaryotic-like Ser/Thr protein kinases of Mycobacterium tuberculosis.* Trends Microbiol, 2000. **8**(5): p. 238-44.
96. Cowley, S., et al., *The Mycobacterium tuberculosis protein serine/threonine kinase PknG is linked to cellular glutamate/glutamine levels and is important for growth in vivo.* Molecular microbiology, 2004. **52**(6): p. 1691-1702.
97. Scherr, N., et al., *Structural basis for the specific inhibition of protein kinase G, a virulence factor of Mycobacterium tuberculosis.* Proc Natl Acad Sci U S A, 2007. **104**(29): p. 12151-6.
98. O'Hare, H.M., et al., *Regulation of glutamate metabolism by protein kinases in mycobacteria.* Molecular microbiology, 2008. **70**(6): p. 1408-1423.
99. Durocher, D. and S.P. Jackson, *The FHA domain.* FEBS Lett, 2002. **513**(1): p. 58-66.

100. England, P., et al., *The FHA-containing protein GarA acts as a phosphorylation-dependent molecular switch in mycobacterial signaling.* Febs Letters, 2009. **583**(2): p. 301-307.
101. Barthe, P., et al., *Dynamic and Structural Characterization of a Bacterial FHA Protein Reveals a New Autoinhibition Mechanism.* Structure, 2009. **17**(4): p. 568-578.
102. Liu, Y., et al., *Structural basis for selective inhibition of Src family kinases by PP1.* Chem Biol, 1999. **6**(9): p. 671-8.
103. Sastry, G.M., et al., *Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments.* J Comput Aided Mol Des, 2013. **27**(3): p. 221-34.
104. Kirchmair, J., et al., *The Protein Data Bank (PDB), its related services and software tools as key components for in silico guided drug discovery.* Journal of medicinal chemistry, 2008. **51**(22): p. 7021-40.
105. Xie, X., et al., *Crystal structure of JNK3: a kinase implicated in neuronal apoptosis.* Structure, 1998. **6**(8): p. 983-91.
106. Schindler, T., et al., *Crystal structure of Hck in complex with a Src family-selective tyrosine kinase inhibitor.* Mol Cell, 1999. **3**(5): p. 639-48.
107. Kinoshita, T., et al., *Structure of human Fyn kinase domain complexed with staurosporine.* Biochem Biophys Res Commun, 2006. **346**(3): p. 840-4.
108. Levinson, N.M., et al., *A Src-like inactive conformation in the abl tyrosine kinase domain.* PLoS Biol, 2006. **4**(5): p. e144.
109. Rellos, P., et al., *Structure of the CaMKIIdelta/calmodulin complex reveals the molecular mechanism of CaMKII kinase activation.* PLoS Biol, 2010. **8**(7): p. e1000426.
110. Schulze-Gahmen, U., H.L. De Bondt, and S.H. Kim, *High-resolution crystal structures of human cyclin-dependent kinase 2 with and without ATP: bound waters and natural ligand as guides for inhibitor design.* Journal of medicinal chemistry, 1996. **39**(23): p. 4540-6.
111. Shewchuk, L., et al., *Binding mode of the 4-anilinoquinazoline class of protein kinase inhibitor: X-ray crystallographic studies of 4-anilinoquinazolines bound to cyclin-dependent kinase 2 and p38 kinase.* J Med Chem, 2000. **43**(1): p. 133-8.
112. Sali, A. and T.L. Blundell, *Comparative protein modelling by satisfaction of spatial restraints.* J Mol Biol, 1993. **234**(3): p. 779-815.

113. Lombana, T.N., et al., *Allosteric activation mechanism of the Mycobacterium tuberculosis receptor Ser/Thr protein kinase, PknB.* Structure, 2010. **18**(12): p. 1667-77.
114. Benkert, P., S.C. Tosatto, and D. Schomburg, *QMEAN: A comprehensive scoring function for model quality assessment.* Proteins, 2008. **71**(1): p. 261-77.
115. Chang, G., W.C. Guida, and W.C. Still, *An Internal Coordinate Monte-Carlo Method for Searching Conformational Space.* Journal of the American Chemical Society, 1989. **111**(12): p. 4379-4386.
116. Biasini, M., et al., *OpenStructure: an integrated software framework for computational structural biology.* Acta Crystallogr D Biol Crystallogr, 2013. **69**(Pt 5): p. 701-9.
117. Poznanski, J., A. Poznanska, and D. Shugar, *A Protein Data Bank survey reveals shortening of intermolecular hydrogen bonds in ligand-protein complexes when a halogenated ligand is an H-bond donor.* PLoS One, 2014. **9**(6): p. e99984.
118. De Colibus, L., et al., *More-powerful virus inhibitors from structure-based analysis of HEV71 capsid-binding molecules.* Nature structural & molecular biology, 2014. **21**(3): p. 282-8.
119. Miller, B.R., et al., *MMPBSA.py: An Efficient Program for End-State Free Energy Calculations.* Journal of Chemical Theory and Computation, 2012. **8**(9): p. 3314-3321.
120. D.A. Case, T.A.D., T.E. Cheatham, III, C.L. Simmerling, J. Wang, R.E. Duke, R., et al., *AMBER 12*, S.F. University of California, Editor. 2012.
121. Wang, J., et al., *Automatic atom type and bond type perception in molecular mechanical calculations.* J Mol Graph Model, 2006. **25**(2): p. 247-60.
122. Darden, T., D. York, and L. Pedersen, *Particle Mesh Ewald - an N.Log(N) Method for Ewald Sums in Large Systems.* Journal of Chemical Physics, 1993. **98**(12): p. 10089-10092.
123. Ryckaert, J.P., G. Ciccotti, and H.J.C. Berendsen, *Numerical-Integration of Cartesian Equations of Motion of a System with Constraints - Molecular-Dynamics of N-Alkanes.* Journal of Computational Physics, 1977. **23**(3): p. 327-341.
124. Gohlke, H. and D.A. Case, *Insights into protein-protein binding by binding free energy calculation and free energy decomposition using a generalized born model.* Abstracts of Papers of the American Chemical Society, 2003. **225**: p. U791-U791.

125. Wang, J.M., T.J. Hou, and X.J. Xu, *Recent Advances in Free Energy Calculations with a Combination of Molecular Mechanics and Continuum Models.* Current Computer-Aided Drug Design, 2006. **2**(3): p. 287-306.
126. Jayaram, B., D. Sprous, and D.L. Beveridge, *Solvation free energy of biomacromolecules: Parameters for a modified generalized born model consistent with the AMBER force field.* Journal of Physical Chemistry B, 1998. **102**(47): p. 9571-9576.
127. Onufriev, A., D. Bashford, and D.A. Case, *Exploring protein native states and large-scale conformational changes with a modified generalized born model.* Proteins-Structure Function and Bioinformatics, 2004. **55**(2): p. 383-394.
128. Kollman, P.A., et al., *Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models.* Accounts of Chemical Research, 2000. **33**(12): p. 889-897.
129. Hermann, R.B., *Theory of Hydrophobic Bonding .2. Correlation of Hydrocarbon Solubility in Water with Solvent Cavity Surface-Area.* Journal of Physical Chemistry, 1972. **76**(19): p. 2754-&.
130. Massova, I. and P.A. Kollman, *Combined molecular mechanical and continuum solvent approach (MM-PBSA/GBSA) to predict ligand binding.* Perspectives in Drug Discovery and Design, 2000. **18**: p. 113-135.
131. Hunter, J.D., *Matplotlib: A 2D graphics environment.* Computing in Science & Engineering, 2007. **9**(3): p. 90-95.
132. van der Walt, S., S.C. Colbert, and G. Varoquaux, *The NumPy Array: A Structure for Efficient Numerical Computation.* Computing in Science & Engineering, 2011. **13**(2): p. 22-30.
133. Scherr, N., et al., *Survival of pathogenic mycobacteria in macrophages is mediated through autophosphorylation of protein kinase G.* Journal of bacteriology, 2009. **191**(14): p. 4546-54.
134. Shah, K., et al., *Engineering unnatural nucleotide specificity for Rous sarcoma virus tyrosine kinase to uniquely label its direct substrates.* Proc Natl Acad Sci U S A, 1997. **94**(8): p. 3565-70.
135. Hanke, J.H., et al., *Discovery of a novel, potent, and Src family-selective tyrosine kinase inhibitor - Study of Lck- and FynT-dependent T cell activation.* Journal of Biological Chemistry, 1996. **271**(2): p. 695-701.

136. Notredame, C., D.G. Higgins, and J. Heringa, *T-Coffee: A novel method for fast and accurate multiple sequence alignment.* Journal of Molecular Biology, 2000. **302**(1): p. 205-217.
137. Simpson, R.J., *Proteins and Proteomics: A Laboratory Manual*.
138. Homeyer, N. and H. Gohlke, *Free Energy Calculations by the Molecular Mechanics Poisson-Boltzmann Surface Area Method.* Molecular Informatics, 2012. **31**(2): p. 114-122.
139. Okimoto, N., et al., *High-performance drug discovery: computational screening by combining docking and molecular dynamics simulations.* PLoS Comput Biol, 2009. **5**(10): p. e1000528.
140. Rastelli, G., et al., *Fast and Accurate Predictions of Binding Free Energies Using MM-PBSA and MM-GBSA.* Journal of Computational Chemistry, 2010. **31**(4): p. 797-810.
141. Tian, C., et al., *The stereoselectivity of CYP2C19 on R- and S-isomers of proton pump inhibitors.* Chem Biol Drug Des, 2014. **83**(5): p. 610-21.
142. Liu, Y., et al., *Engineering Src family protein kinases with unnatural nucleotide specificity.* Chemistry & biology, 1998. **5**(2): p. 91-101.
143. Ubersax, J.A., et al., *Targets of the cyclin-dependent kinase Cdk1.* Nature, 2003. **425**(6960): p. 859-864.
144. Bottoms, C.A., T.A. White, and J.J. Tanner, *Exploring structurally conserved solvent sites in protein families.* Proteins, 2006. **64**(2): p. 404-21.
145. Michel, J., J. Tirado-Rives, and W.L. Jorgensen, *Prediction of the water content in protein binding sites.* J Phys Chem B, 2009. **113**(40): p. 13337-46.
146. Finn, R.D., et al., *Pfam: the protein families database.* Nucleic Acids Res, 2014. **42**(Database issue): p. D222-30.
147. Roderick E Hubbard, M.K., *Hydrogen Bonds in Proteins: Role and Strength.* Encyclopedia of life sciences, 2010.
148. Israelachvili, J. and R. Pashley, *The Hydrophobic Interaction Is Long-Range, Decaying Exponentially with Distance.* Nature, 1982. **300**(5890): p. 341-342.

A Appendix

A.1 Sum of vdW radii

Table A.1. Sum of vdW radii for the most common protein-ligand atom pairs in case of non-halogenated ligands.

Atom pairs	Sum of vdW radii (Å)
C-C	3.4
C-N	3.25
C-S	3.5
C-O	3.22
C-P	3.50
N-N	3.1
N-S	3.35
N-O	3.07
N-P	3.35
O-S	3.32
O-O	3.04
O-P	3.32
S-P	3.60
S-S	2.03

A.2 Ribose and phosphates conformations

The evaluation of the interaction between kinase mutants and ATP-competitive ligands was performed by GlideScore. We decided to score only the substituted adenine base of each analogue and not the ribose ring and the phosphates within the binding pocket of an engineered kinase. This approximation is based on the fact that the position and the orientation of the planar adenine moiety within the binding site is conserved among almost all serine/threonine and tyrosine protein kinases. This is due to the highly conserved H-bonds between the adenine ring and two protein residues. On the other hand, both the ribose ring and the phosphates can assume various conformations in different protein kinases. This is due to their intrinsic flexibility and to the protein environment, such as the presence of water molecules within the binding site or of divalent ions and the way of interaction with specific residues. To verify the accuracy of the approximation, we analyzed the geometries of ATPs and ATPs-like (ADP, AMP, ACP, AGS and ANP) within the ligand-binding pocket of serine/threonine and tyrosine protein kinases for which crystal structures had been solved. We built a dataset containing 258 entries (Table A.2.1). Each entry represents the crystal structure of the kinase domain of a serine/threonine or a tyrosine protein kinase bound to ATP or an ATP-like ligand. Two databases were used to build our dataset, the Protein families database, Pfam [146] and the Protein Data Bank, PDB [63]. In Pfam we performed two searches, one using the accession code PF00069 (that identify protein kinase domains) and the other one using the accession code PF07714 (that identify protein tyrosine kinases). Those searches identified 2157 entries. In the PDB we performed four searches, two using the Pfam codes PF00069 and PF07714 and two using the keywords ‘serine/threonine protein kinase’ and ‘tyrosine protein kinase’. In that manner we identified 1229 structures. Then, we compared entries obtained from Pfam and PDB in order to eliminate redundant structures. Then, we only consider structures of kinase domains bound to ATP or ATP-like ligands together with at least one divalent ion and not to other molecules (such as kinase inhibitors). We superposed the ligand binding sites of all 258 structures to check the geometries of the diverse cofactors within the binding

pockets. Figure A.2.1 represents the binding site superposition of 10 structures belonging to the dataset.

Figure A.2.1. Structural superposition of the binding sites of 10 protein kinase structures (PDBs: 1ATP, 1B38, 1B39, 1CSN, 1HCK, 1PHK, 1QL6, 3E8N, 3Q53 and 3DKC).

The structural superposition shows that the position and the orientation of the adenine moieties within the different binding sites are highly conserved. On the other hand, the ribose rings and the phosphates assume diverse conformations, positions and orientations within the diverse binding sites. Figure A.2.2 A and B represents the binding site of the *Mus musculus* cAMP-dependent protein kinase in complex with ATP and Mn^{2+} (PDB: 1ATP) and of the human cyclin-dependent kinase 2 in complex ATP and Mg^{2+} (PDB: 1B38). The position and the orientation of the two adenine moieties is conserved in both kinases where they are involved in H-bonds with protein residues. The ribose rings show two different conformations and phosphates shows two different geometries as consequence of different interactions. In cAMP-dependent protein kinase α and γ phosphates coordinate the Mn^{2+} and the catalytic Lys interacts with α and γ phosphates. In cyclin-dependent kinase 2 the Mg^{2+} is coordinated by the α and β phosphates and the catalytic Lys interacts with the α phosphate.

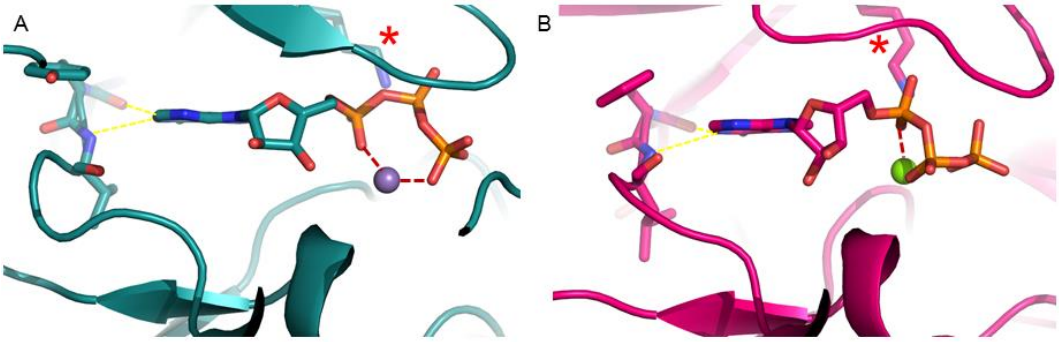

Figure A.2.2. ATP binding site of cAMP-dependent protein kinase (PDB: 1ATP) and cyclin-dependent kinase 2 (PDB: 1B38). Yellow dashed lines represent protein-ligand H-bonds and red dashed lines represent interaction between phosphates and divalent ions, Mn^{2+} in violet and Mg^{2+} in green. Red stars indicate catalytic Lys.

Table A.2.1. PDB entries of our dataset.

1ATP	2QO9	3U87	2W4K
1B38	2QOC	3UIM	2W5A
1B39	2QOQ	3VJN	2WQE
1CSN	2R5T	3VJO	2WQN
1FIN	2SRC	4DFX	2XUU
1FQ1	2V55	4DG0	3A7J
1GOL	2VWB	4EKK	3BRB
1GY3	2VWI	1DS5	3C4Z
1H1W	2WTK	1YXU	3C50
1HCK	2XRW	2IVT	3C51
1JST	2XS0	2YAB	3D5W
1OL6	2Y9Q	3C0H	3DLS
1PHK	2ZV8	3DLZ	3EH9
1Q24	3A99	3DQX	3EQG
1Q97	3ALN	3GC0	3EQH
1QL6	3ALO	3NYO	3EQI
1QMZ	3BEG	1MRU	3F5G
1RDQ	3COK	2G2F	3F61
1S9I	3DAK	2PMI	3G2F
1S9J	3DFC	2W5B	3GU6
1U5R	3EHA	3DQW	3GU7
1UA2	3F5U	3EQC	3LJ0
1ZYD	3FHI	3EQD	3MBL
2BIY	3FXX	3FZP	3NIZ
2CCH	3G51	3ORI	3P23
2CCI	3GT8	3ORK	3QC9
2CJM	3GU4	3ORL	3QHR
2IJM	3GU5	3ORM	3QHW
2P55	3HKO	3ORO	3TNQ
2PHK	3HX4	3ORP	1I44
2Y4I	3IDB	3ORT	1K3A
2YAA	3IDC	1J1C	1O6Y
3A7H	3IGO	1JBP	1U54
3A8W	3IS5	1L3R	2PVF
3BLQ	3JUH	1MP8	2PVY
3BU5	3KN5	1MQ4	2PWL
3C4W	3KU2	1NY3	2PY3
3C4X	3LIJ	1OL5	2PZ5
3DKC	3LLT	1OL7	2PZP
3DV3	3MFR	1PKG	2PZR
3DY7	3MFS	1Q8Y	2Q0B
3E7E	3MFU	1WBP	2Z7Q
3E8N	3MIA	2B9F	3CLY
3EQB	3NIE	2B9H	3GQI
3FJQ	3NSZ	2B9I	3KEX
3GNI	3O7L	2B9J	3LMG
3HMN	3ORN	2CN5	3PLS
3HRC	3PFQ	2F9G	1JPA
3HRF	3PVB	2G2I	2HEN
3KMW	3Q4Z	2QUR	3LCT
3OS3	3Q5I	2RIO	
3PP1	3SLS	2W4J	

A.3 Visualization of protein-ligand complexes

All protein-ligand complexes of this thesis are visually inspected in Maestro (version 9.5, Schrödinger, LLC, New York, NY, 2013). Protein-ligand contacts are identified using the following formula:

$$C = \frac{D_{12}}{R_1 + R_2} \quad \text{(A.1)}$$

where D_{12} is the distance between atoms 1 and 2, $R_{1,}$ and R_2 are the van der Waals radii of atoms 1 and 2. Protein-ligand contacts are good when distances are within the attractive range of potential and as atoms get closer the quality of their contacts gets worse. When the distance between two atoms is lower than 0.89 Å the protein-ligand contact is defined as ‘bad’ whereas if the distance is lower than 0.75 Å the contact is defined as ‘ugly’.

The H-bonds are identified by the acceptor-donor (A-D) distance and angle. The cut off for the A-D distance is 4.0 Å and the A-D angle must be greater than 120 ° [147].

The hydrophobic interactions are identified by measuring the distance between two carbon atoms and 4.0 Å is the max distance of carbon atoms for an hydrophobic interaction [148].

A.4 Pairs of residues within JNK binding site

Table A.4.1 shows all pairs of residues within the JNK ligand-binding site identified as good candidates for double mutagenesis to enlarge the binding site.

Table A.4.1. Pairs of JNK residues identified as good candidates for double mutagenesis to allow the accommodation of four ATP analogues (N6-(benzyl) ATP, N6-(2-phenythyl) ATP, N6-(cyclopentyl) ATP and N6-(1-methylbutyl) ATP). In red are residue pairs mutated in the work of Habelhah and coworkers [27].

ATP analogues	**Residues pairs within JNK ligand-binding site**
N6-(benzyl) ATP	Met111-Ile86, Met111-Glu109, Met108-Leu168, Ile86-Leu168, Ile86-Glu109, Ile86-Met111, Val158-Leu168, Glu109-Ala53, Glu109-Met111, Ala53-Met111
N6-(2-phenythyl) ATP	Met111-Ile86, Met108-Leu168, Ile86-Val158, Ile86-Leu168, Ile86-Glu109, Ile86-Met111, Val158-Leu168, Glu109-Leu110, Glu109-Ala53 Glu147-Met111, Leu110-Ala53, Ala53-Met111
N6-(cyclopentyl) ATP	Met111-Ile86, Met108-Leu168, Ile86-Glu109, Ile86-Met111, Ala53-Met111
N6-(1-methylbutyl) ATP	Met111-Ile86, Met111-Glu109, Met108-Leu168, Ile86-Val158, Ile86-Leu168, Ile86-Glu109, Ile86-Met108, Glu109-Met108, Ala53-Met108

Valentina Romano

Address: Avenue de Bâle, 68300 Saint Louis, France
Date of birth: 13 May 1982
E-mail: romano.vale@gmail.com

WORK EXPERIENCE

Research Assistant (October 2011 – February 2016)
Biozentrum Department, University of Basel, Switzerland

Internship as Research Assistant (April 2011 – September 2011)
Department of Physics, "Sapienza" University, Rome, Italy

EDUCATION

Doctoral studies (October 2011 - Defense expected in February 2016)
Biozentrum Department, University of Basel, Switzerland
Thesis: "*Computational engineering of co-substrate specificity in protein kinases*"

Master of Science in Chemistry (March 2011)
University of Naples "Federico II", Italy
Thesis: "*Study of structural and binding properties of TGF-β growth factors*"
Final Grade: 110/110 cum Laude (top grade, with honors)

Bachelor of Science in Chemistry (July 2008)
University of Naples "Federico II", Italy
Thesis: "*NMR conformational analysis of SDF-1 protein derived peptide*"
Final Grade: 110/110 (top grade)

LANGUAGES

English – Full working proficiency
Italian – mother tongue
French – Proficient in speaking, intermediate in writing

RESEARCH PAPERS

Computational protocol to evaluate protein mutants in the kinase gatekeeper position Romano V, de Beer T, Schwede T. (submitted)

Toward a better understanding of the interaction between TGF-b family members and their ALK receptors Romano V*, Raimondo D*, Calvanese L, D'Auria G, Tramontano A and Falcigno L (2012), J.Mol.Model, 18(8): 3617-3625 * Authors contributed equally