Conformational Variability: Implications for Biomolecular Activity and in Drug Design

Mousumi Bhattacharyya

UNIVERSITAT DE BARCELONA

Facultat de Farmacia

Departament de Fisicoquimica

Conformational Variability:
Implications for Biomolecular Activity and in Drug Design

Mousumi Bhattacharyya

2012

UNIVERSITAT DE BARCELONA

Uninersitat de Barcelona
Facultat de Farmàcia
Departament de Fisicoquimica
Programa de Doctorado de Biomedicina

Conformational Variability:
Implications for Biomolecular Activity and in Drug Design

Memòria presentada per Mousumi Bhattacharyya per optar al títol de doctor per la universitat de Barcelona

Director de tesis

Dr Xavier Barril Alonso

Doctorando

Mousumi Bhattacharyya

Mousumi Bhattacharyya

2012

To my parents, Mamata and late Kohinur Bhattacharyya, my sister Moumita, and my better half, Aniruddha.

Gratitude

At last, it has come to the final stage of a journey with the end of this thesis. It has certainly been a long and colourful venture of my life and I like to take this opportunity to convey my heartiest gratitude to all of those without whose support, patience, help, companionship, love and care it wouldn't have been possible. I am grateful to many people who provided the supervision, direction and assistance to enable me to reach this destination.

First of all I would like to thank my supervisor Dr. Xavier Barril (or Xavi) for giving me the opportunity to work in this group. His constant support and help made this thesis to see the light of the day. Other than learning his way of working I have also learnt how to be a calm and cool headed person. I would like to convey my sincere gratitude to Prof Javier Luque (or Javi) for his scientific inputs in my process of learning, for extending his help whenever I needed any, scientific to non-scientific. This journey has given me the opportunity to come across people like Xavi and Javi who are not only great scientists in their fields but also very good human beings.

I owe heartfelt gratitude to Oscar, Ignacio, Axel, Jesus and Damian. In my very initial days in Barcelona when I was finding everything very hard in settling down they have made it easy with their helping hands. Muchas gracias a Flavio i Assumpta, los Secretario i Presidente del COPPO (Comision Organizadora de Pica-Pica Oficiales) de la grupo. Due to this pica-pica I got to see the much-heard "Jamon feast". Thanks also to Flavio and Lurdes for making my initial stay comfortable and friendly with them.

Ah, ara Ramon,...moltes gracies a ti, Ramon. His funny and humorous philosophical view towards life was very interesting as well as enjoyable. Gran Gracias a Ana for being very friendly throughout my stay, Jordi, Carles, Salome for allowing me to do "cachondeo" by pulling their legs. Special thanks a Patricio for all those lemon-pies and in the pica-pica and also for introducing me to the Chilean desserts. I had the privilege to work in this group which is truly international and diverse.

Here comes the opportunity to say something about our closely knitted group. I feel lucky to have beside me Jesus, Peter, Montse, Dani, Sergi and in the end of my stay Laura. My deepest gratitude goes with them. Thanks to Jesus for those very interesting discussion on Mc Donald. Ah, Peter...thanks a lot for all of your support, encouragement, and help. Thanks for creating all those laughter moments in the lab…for the discussions, sometimes we used to have, on almost everything under the sun, ranging from movie to politics to music to food. They have eventually helped me gaining knowledge on various grounds of European culture, from classical music to movies to cuisines (especially, with the Christmas cakes and meat balls). Taking this opportunity I thank to Cecile also for her friendship and spending time with me whenever she was in Barcelona. Moltes moltes gracies a Montse -our one and only "Joana"- for her kind, and helpful companionship throughout and, most importantly, for her enormous enthusiasm in organizing a truly scientific session viz "after-work beer" on every Friday evening. Hope she will maintain this passion to keep the session alive with the help of samosa-amigo Xavi…gracias a ti tambien, Xavi. Big big thanks to Dani for his help and being there whenever I needed anything. My deepest gratitude to Sergi, who, I have always felt, has a very brotherly character. Thanks for his kindness and friendly attitude towards me. Thanks a lot to Laura for her encouragement and generosity.

Although we do not share research interests with Pep's (Prof. Josep Abril) group, we do share the lab. This has brought to me a great opportunity to get to know Pep, Marta and Natalia. In spite of being busy in various academic programs Pep was always available for any kind of help I asked for, from scientific to technical consultancies. My heartfelt gratitude goes to Pep for everything, including the "merienda donuts". I would like to express my deepest gratitude to Marta "Mahasweta" and Natalia for extending their help whenever I needed. Thanks for your companionship, and constant encouragement. Same goes to Marc also. Gracias a Guss tambien.

I would like to thank Prof B Jayaram for giving me opportunity to work in his group. I would also like to convey my gratitude to Gautam, Ashutosh and others from the same group for their hospitability.

Finally I would like thank my first scientific research guru, Prof. Somdatta Sinha. She has been a very supportive, and encouraging, and motherly character in my early days of this field.

This section won't be complete unless I acknowledge to my Indian friends in Barcelona. I would like to thank Gautam, Andrea, Aditidi, Ujjwal da, Chaitanya, Joyee, Saurabhab, and lately Kanchan. They all have made my stay in Barcelona enjoyable.

Last but not the least, I would like to thank my family, my mother, sister, and Aniruddha, my uncles and aunts for being there always. I am indebted to them for their constant support, patience, and unconditional love.

Finally, I would like to thank Generalitat de Catalunya for providing me the financial assistance in accomplishing my thesis work.

Moltes moltes gracies a Barcelona.

INDEX

Chapter 1. Introduction

Knowing the flexibility of biomolecules is very critical for understanding how they function in the cellular environment (Mamonova *et al.*, 2005, Bhalla *et al.*, 2006, Fischer and Verma, 1999). Most functions carried out by the molecular machinery of the cell namely, enzyme catalysis, molecular recognition or association, cellular transport, allosteric regulation etc., require reorganizing its native structure through a time dependent process. Therefore understanding, predicting and modulating the functions of proteins and other bio-macromolecules require not only knowledge about their structures, but also a profound understanding of their flexibility at the molecular level.

In contrast to bio-macromolecules, which have an activity of their own, small bio-molecules or ligands usually are active only when they bind to a target (Gohlke *et al.*, 2003b). The process of binding of a ligand to its binding partner is generally not a straightforward event, as it involves a set of recognition mechanisms, in terms of interactions, between two entities where both adapt, through locally or globally, mutual conformational arrangements to accommodate each other. This association is a complex phenomenon due to the large number of degrees of freedom associated with the 3-dimensional structures of both molecules in addition to the water molecules and other components present in the cellular environment. As a large number of stable conformations may exist for a flexible molecule and the energetic differences between them may be relatively small, understanding flexibility is a very challenging pursuit. But it is also a necessary one in the rational drug discovery and development ventures where predicting the binding mode and binding affinity of a bioactive molecule towards its target is imperative or rather inseparable component.

Increasing development in the techniques of the X-ray crystallography during the last few decades has given access to the substantial amount of three-dimensional atomistic structures of the biological molecules. The solved crystallographic structures have unquestionably boosted up our understanding about the structural complexities of biomolecules. However, a crystal structure of such a molecule (proteins, nucleic acids or ligands) represents only the final conformation that has been trapped into a particular crystalline environment. Information about the inherent dynamic behavior the molecule is largely lost in the structures obtained from x-ray crystallography, as it is limited to the approximated estimation of uncertainty or atomic fluctuations encapsulated in the B-factors. Though very useful in certain field or study, B-factors cannot explain the vibrant or energetic behavior of the molecule into a crowded cellular system. On the other hand, there is limited availability of purely experimental techniques able to follow the detailed dynamics of proteins in time. Light and electron microscopy allows insights into the behavior of complex structures of living cells and tissues in real time but with poor resolution (Cardoso *et al.*, 2012). Hence, various computational approaches are being developed to bridge the gap between the structure and dynamics and to correlate with various functions accomplished by the bio-molecules. While using very fine set of mathematical formalisms of Quantum Mechanical (QM) theory it is now possible to accurately predict the energetics of the small molecules or ligands, Classical Molecular Dynamics (MD) simulations allow us to investigate the time-dependent behavior of the macromolecules, like proteins or nucleic acids in a cell-like simulated environment. At present, MD simulations are extensively used, either alone or in combination with other

experimental methods, to understand the functionality of biological systems at the molecular level.

1.2 Scientific Domains visited in this Thesis

In a broad sense this thesis work has been carried out to investigate and understand the relationship between the structure and function (or biological activity) of the biomolecules. More specifically, the study is divided into two parts. In the first part we explore the conformational flexibility and energetic behavior of small drug-like molecules upon binding to their respective targets using state-of-the art sophisticated quantum mechanical approaches. In the second part of this thesis, we investigate the intrinsic structural flexibility and dynamics of a biologically important protein Histone deacetylase 8 (HDAC8) to understand allosteric regulation and to propose a plausible mechanism to control activity through phosphorylation at a residue widely separated from its catalytic site.

1.3 Energetics of the bioactive conformers

It is a well established fact that small molecules that bind to their biological targets (e.g. proteins) change their conformations during the course of binding to become complementary to the flexible target binding site (Fig 1.1). Thus, the bound conformation (or bioactive conformation) of a molecule, often, does not correspond to the global energy minimum (Bostrom *et al.*, 1998b, Perola and Charifson, 2004a) or even local energy

minimum conformation of its unbound form (Butler *et al.*, 2009b) (Fig 1.2). This leads to an added penalty to the total free energy of binding of the ligand to its target that comes from the internal energy of the conformer. Before going more into our study that explains the effects of this observable fact or bio-chemical phenomenon, a brief definition of the terms used in this study to distinguish different conformational states of a molecule is given below.

Global Energy Minimum: A flexible molecule can exist in more than one conformational state due to internal degrees of freedom associated with the internal coordinates. The collection of all possible conformations of a molecule is referred as ensemble of conformation. However not all conformations are energetically favorable due to the high intra-molecular strain. In the free state of a molecule, amongst all energetically favorable conformations, the one that has the lowest energy is known as "global energy minimum" (Fig. 1.2).

Bioactive Conformation: Among the energetically favorable set of conformations only one (or few) can bind to their respective biological target. The reduction in the conformational space always results in an entropic term that opposes binding. The particular conformation that is preferred and accepted by the target is referred as the bioactive conformation. Bioactive conformations can be anywhere on the potential energy surface (Fig 1.2), but unless they correspond to the global energy minimum, the

conformational selection involves an energetic component that opposes binding.

Local Energy Minimum: In the free state of a molecule each energetically stable conformation is defined as local energy minima. Since the bioactive conformation is chosen or fitted to the binding site by its biological target, that particular conformation may not correspond to a minima in the unbound form. If the bioactive conformation is allowed to relax freely, the resulting structure is known as the local energy minimum of the bioactive.

The energetic cost due to the conformer selection in the binding site can significantly affect to the affinity of the ligand: each 1.4 kcal/mol of increased conformational energy of a bioactive conformation leads to a 10 fold decrease in its binding affinity (Bostrom *et al.*, 1998b). This itself explains the importance of properly estimating the internal energy or strain energy of the ligands in the drug discovery field. Computer-aided drug design methods depend on software programs to pick up a potential drug-like molecule from a pool of compounds. When those programs make use of three-dimensional representations of molecules, they rely on existing force-fields to predict the internal energy of each molecule's conformer. This ends up adding a significant fraction of error in the predicted outcome, with uncertainties in the prediction up to 10kcal/mol (Tirado-Rives and Jorgensen, 2006a). Perola and Charifson, using molecular mechanics force fields designed for small molecules, found that in approximately 40% of the molecules studied the bioactive conformations had strain energy above 5kcal/mol (Perola and Charifson, 2004b). Considering the relationship between strain energies and binding affinities, this would involve a loss of affinity greater than 35-fold, which seems exaggerated for optimized ligands.

The impact of using different computational approaches was investigated by Butler et al., observing that two thirds of the bioactive conformers were within 0.5 kcal/mol of a local energy minimum when calculated at a high level of theory (QM), 50% of them are above 3 kcal/mol when calculated at the MM level with a basic solvation model (Butler *et al.*, 2009b). Here we have extended the work done by Butler et al. to focus on the bioactive conformation from the global scenario of energy surface by comparing the energetics (calculated using different QM approaches) between the bioactives and a set of conformers that represent the global set of conformers. We also compare different methods used to compute the energy of the drug-like molecules using molecular mechanics (MM) force fields.

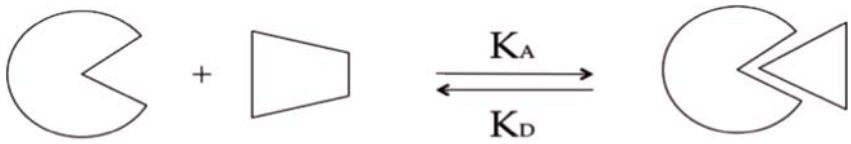


Fig 1.1: Binding of a small molecule to the target through a process of conformational selection and molecular recognition.

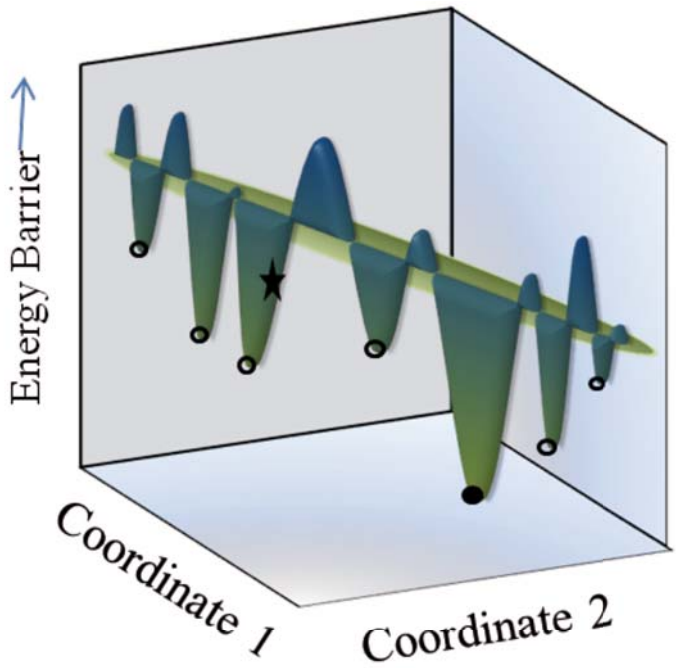


Fig 1.2: Representation of an "Energy Surface" of a flexible molecule which shows that the bioactive conformation (marked as star) of this particular molecule does not even correspond to its local minimum (marked as circle). Filled surface represents Global minima.

1.4 Allosterism in HDAC8

Lee H et al. found that the phosphorylation by Protein Kinase A, as a course of posttranslational modifications, leads to a decreased deacetylase activity of HDAC8 (Lee *et al.*, 2004) shown that the phosphorylated site is a non-catalytic, non-conserved N-terminal residue Ser39 (Fig. 1 of chapter 4.2) (Somoza *et al.*, 2004). The relation observed between the phosphorylation

and activity led us to assume that allosterism might have a role to play in HDAC8.

Allosterism is a well known regulation mechanism to various cellular functions involving proteins where the functionality of a protein is controlled by binding of an effector molecule at a site away from the active site [Fig. 1A]. More often the binding of the effector molecule is attributed with two functional states: i) active state i.e. making the host protein active upon binding to it and ii) inactive state which does the opposite by making the host protein inactive (Daily and Gray, 2007). At a molecular level, binding of the effector molecule induces changes in the protein three-dimensional structure to adapt a suitable conformation through a continuous process of structural changes. Researches, both theoretically and experimentally, have pointed that allosterism is coupled with the dynamics and changes between different conformational ensembles (Kumar *et al.*, 2000, Luque *et al.*, 2002, Gunasekaran *et al.*, 2004). Therefore to have an insight about a system that might turn on allosterism due to a natural modification, a profound knowledge about its dynamic behavior at the molecular level would be helpful.

Hence, we used classical molecular dynamics (MD) simulations as our primary tool to generate the time dependent ensemble of conformations of the Wt-HDAC8, its phosphorylated form and other mutants to analyze them. Biased metadynamics simulations have also been used to address some issues arose during the course of our analysis. By developing a statistical approach the differences evolved in the dynamics of Wt-HDAC8 and its phosphorylated form were detected. Following this we proposed a path of allosterism on HDAC8 that could have been followed to control the catalytic

activity upon its phosphorylation. We validated our hypothesis and results with experimental techniques.

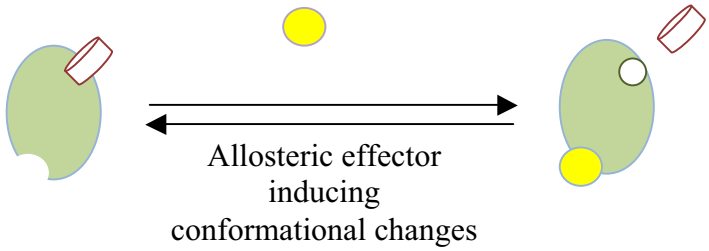


Fig 1.3. Binding of an effector molecule at a protein site far from its active site induces conformational changes which in turn influences the binding of a ligand at its binding/active site.

Chapter 2. Objectives

In this thesis work we intend to explore the association of conformational behavior or conformational selection with the exerted functions of the bio-molecules at molecular level. We have ventured into two different types of molecules, a macro molecule or protein which is histone deacetylase 8 and a set of drug like compounds.

Understanding conformational flexibility and energetic in small drug like molecules

Although it is well known that the bioactive conformation rarely corresponds to the global minimum which have been identified with the current small-molecule empirical force fields, there is no consensus yet about how frequently this occurs in reality or about the maximal conformational penalty that a ligand can attain (Perola and Charifson, 2004b, Tirado-Rives and Jorgensen, 2006b). As mentioned before also, this is an important aspect in the drug discovery field. Knowing accurate range of conformational selection or fit cost to properly accommodate into the target binding site will narrow down the search and generation of drug-like compounds (Figure 2.1).

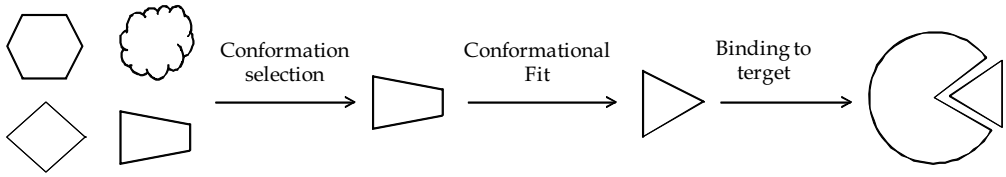


Figure 2.1: for a flexible molecule the costs involved to acquire the conformation close to the bound form and then fit it into the target may interfere with the binding process. If the costs are high binding affinity may be affected.

1. To accurately estimate the strain energy or internal energy of a drug molecule, that can be tolerated during the process of binding.
2. To evaluate the impact of different levels of QM theory in estimating both the intrinsic energetics and the effect of aqueous solvation.
3. To assess the currently available empirical force fields for the small-molecules that are used extensively in the drug discovery projects in industry as well as in academia.

The ultimate goal of this research line is to identify a computational protocol offering the best balance between accuracy and computational cost, which could be used to supplement docking calculations with accurate estimates of the internal energy of drug candidates under study.

Understanding conformational flexibility and functionality in HDAC8

While we know the response towards the activity of Histone Deacetylase 8 as a result of the phosphorylation at a particular residue (Ser 39) that is well separated from its catalytic site (Lee *et al.*, 2004, Somoza *et al.*, 2004, Sengupta and Seto, 2004) it is not known how the response is generated and propagated to the enzyme catalytic site. This might be one of such examples where local signals are transmitted to a distant site in a manner that involves a shift in the conformational ensemble, as it happens usually in the allosterically motivated signal transductions in the proteins. We presume that phosphorylation turns on the long range communicating

mechanisms through conformational changes that may be inferred from molecular dynamics simulations. The objectives are to systematically investigate:

1) Conformational and dynamic changes undergone by HDAC8 upon phosphorylation at Ser 39.
2) Identify putative pathways in the allosteric mechanism that may explain how a local change affects the global properties of the system.
3) Validate the proposed pathways and allosteric mechanisms by means of point mutations that are investigated both computationally and experimentally.

While histone deacetylase isoforms from class I and II have been associated to different cancerous states of the cells, the specific role and, particularly, the mechanism of each isoform in cancer formation is yet to be defined clearly (Bieliauskas and Pflum, 2008). Systematic study of HDAC8 using molecular dynamics simulations will facilitate understanding, predicting and further modulating its functions. Doing so we hope this work would shade light on the future development of the HDAC8 specific inhibitors.

Chapter 3. Methods

A theoretical model for any complex process is an “approximate but well-defined” mathematical procedure of simulations.

---- John A.Pople in his nobel lecture

3.1 Theory Precedes Experiment:

The need of using theoretical models in interpreting and/or discovering facts in biology can be explained with the forgotten story of the methylene molecule (SCHAEFER III, 1986, Ma and Nussinov, 2004) that started with the Foster and Boys’ pioneering work in 1960 on predicting the structure of methylene. Until 1970 it was universally agreed fact that the methylene molecule was linear in its triplet ground state, as concluded by the father of modern spectroscopy Gerhard Herzberg from his experiments (Herzberg and Shoosmith, 1959) ignoring the fact of bent methylene structure calculated and proposed by Foster and Boys with computational methodologies. However, using rigorous quantum mechanical approaches Bender and Schaefer in 1970 showed that the bond angle of methylene is 135° which was later emboldened by the experimental evidences including a reinterpretation of the spectroscopic studies confirming the bent geometry predicted by theory. This inspiring story along with the computational advancements encourages us to take up the complex challenges of the biological problems to solve them rationally with computational methods.

This chapter essentially presents a review-in-short of different computational methods that we have used, directly or indirectly, to accomplish our investigations. Two fundamental lines of computational techniques have been used vastly: Quantum mechanical and Classical mechanics.

3.2 Quantum mechanics

3.2.1 Schrodinger Equations

During the second half of the 20th century, noteworthy improvements have been made in translating quantum mechanical methods to the complex biological problems including the vibrant drug discovery area to predict the accurate structures of small bio-molecules and their energetic. Basics of Quantum mechanics (QM) start with the popular Schrödinger equation (Schrödinger, 1926) as in eq (3.1) that is used to calculate the energetics of a system:

$$\hat{H}\psi = E\psi \tag{3.1}$$

where E is the total molecular energy and $\hat{H}$ is the Hamiltonian which operates on the wavefunction Ψ to return the system energy E. However, for a many-electron system the simple Hamiltonian becomes as complex as equation (3.2):

$$H = -\sum_i \frac{\hbar^2}{2m_e}\nabla_i^2 - \sum_k \frac{\hbar^2}{2m_k}\nabla_k^2 - \sum_i \sum_k \frac{e^2 Z_k}{r_{ik}} + \sum_{i<j} \frac{e^2}{r_{ij}} + \sum_{k<l} \frac{e^2 Z_k Z_l}{r_{kl}} \tag{3.2}$$

where i and j represent the electrons, k and l to the nuclei, ħ is the Planck's constant divided by 2π, m_e is the mass of the electron, m_k is the mass of nucleus k, ∇^2 is the Laplacian operator, e is the charge on the electron, Z is an atomic number, and the distances between two nuclei, electrons, and electron from nuclei are represented by r_{ab}. Therefore, the system energy E from eq. (3.1), obtained using the Hamiltonian, is the composite form of kinetic and potential energy. The first two terms of the eq. (3.2) are kinetic energy (electron and nuclear kinetic energy) of the system and the last three terms represent the potential energies arising from the coulombic attraction and repulsion between particles: nuclei-electrons (third term of eq. (3.2)), electrons-electrons (fourth term of eq.(3.2)), and nuclei-nuclei (the last term) respectively, which in words becomes like,

$$\hat{H} = KE_{el} + KE_N + PE_{el\text{-}N} + PE_{el\text{-}el} + PE_{N\text{-}N} \qquad (3.3)$$

(KE: Kinetic Energy, PE: Potential Energy)

3.2.2 **Born-Oppenheimer Approximation**:

Considering the fact that the movement of each particle is dependent on the motion of other particles present in the system makes equation 2 difficult and unrealistic to solve for a bio-molecule of minimum size. Several approximations to this equation have been made, as a result, to simplify the use of electronic structure methods on the practical applications. They can be roughly classified into: semiempirical, *ab initio*, and density functional theory (DFT). First such approximation came from Oppenheimer and Born (Born and Oppenheimer, 1927), where they considered the motion of the nuclei constant compared to the electron because of the fact that mass of the

electron is several hundred fold lighter than that of the nuclei. This has greatly reduced number of terms to be calculated in the eq. (3.2) by eliminating the attractive electron–nuclear potential energy term and considering the repulsive nuclear–nuclear potential energy (PE_{N-N}) term as a constant for a given geometry.

Therefore Born-Oppenheimer approximation of $\hat{H}$ in eq. (3.2) and (3.3) gets reduced to:

$$\hat{H} = (\hat{H}_{el} + \hat{H}_N) \tag{3.4}$$

And the Schrödinger equation (eq. 3.1) can be represented as:

$$(\hat{H}_{el} + \hat{H}_N)\, \Psi_{el} = E_{el}\, \Psi_{el} \tag{3.5}$$

where,

$$\hat{H}_{el} = \ KE_{el} + PE_{el\text{-}N} + \ PE_{el\text{-}el} \text{ and } \hat{H}_N = PE_{N\text{-}N} \tag{3.6}$$

However, the electron-electron repulsive term i.e. PE_{el-el} is also very computationally demanding that requires knowing the distance between each of the electron pairs which is by all means difficult to known prior to calculation.

3.2.3 The Hartree-Fock SCF Method

The most common type of *ab initio* method is the Hartree-Fock (HF) where the approximation primarily applied on the electron-electron repulsion

or $PE_{el\text{-}el}$term of the Schrödinger equation (eq. 3.2, 3.5 and 3.6) (Hartree, 1928) of a poly-electronic system. The HF method is, based on the approximation that decouples the correlated motions of electrons and considers that each electron is moving independently in an average electron-field compositely created by the effect of the other electrons present and by the nuclei.. However, this average inclusion of repulsive electronic ($PE_{el\text{-}el}$) effect does not work well for systems where electron-electron correlation is important, including transition metals and, to a lesser degree, conjugated organic molecules. The Hartree-Fock equations are represented as:

$$f_i \chi_i = \varepsilon_i \chi_i \qquad (3.7)$$

χ_i is the spin orbital of electron i and f is the Fock operator acting on χ_i. ε_i is the orbital energy. The Fock operator is the one-electron Hamiltonian that is applied on a particular electron in the poly-electronic system and it is represented as:

$$f_i = -(\tfrac{1}{2})\nabla_i^2 - \sum_l Z_l/r_{il} + PE_i^{HF} \qquad (3.8)$$

where, PE_i^{HF} is the Hartree-Fock potential of an n-electron system. PE_i^{HF} defines the average repulsive potential energy obtained by each electron due to the rest (n-1) electrons and the nuclei. However the primary problem still persists since solving HF-potential requires the prior knowledge of the spin orbitals of the rest (n-1) of electrons. To tackle this issue an iterative approach, known as Self-Consistent field (SCF), is applied to solve HF equations.

Self-Consistent field

This approach follows a repetitive process to get the final solution of Fock product in terms of spin orbitals (χ_i) for PE_i^{HF} from an initial guess. From the initial set of parameters a new set of parameters were calculated and compared to get a better solution for PE_i^{HF} by minimization approach. This process continues until the convergence criteria are contented i.e. until two successive potentials are identical i.e. self-consistent in itself.

3.2.4 Møller-Plesset Perturbation Method:

Møller and Plesset proposed the perturbation theory (Møller and Plesset, 1934, Head-Gordon *et al.*, 1988), known as MPi, to deal with the multi-electron-correlation Hamiltonian to calculate the system energy accurate enough. i in MPi stands for the order of correction considered in the electron correlation. Møller and Plesset (MP) theory is based on the Rayleigh–Schrödinger perturbation theory (RS-PT) in which the Hamiltonian ($\hat{H}$) of the Schrödinger equation (eq. 3.1) has been represented as the sum of an unperturbed Hamiltonian ($\hat{H}_0$) and a perturbation (V_P) i.e. the correlation potential times a real parameter λ (eq. 3.9):

$$\hat{H} = \hat{H}_0 + \lambda V_P \qquad (3.9)$$

λ varies between 0 and 1. In MPi, at $\lambda = 0$, $\hat{H}$ comes from the contribution of $\hat{H}_0$ and becomes the sum of the one-electron Fock operators over the n electrons (eq. 3.10).

$$\hat{H}_0 = \sum_i^n f_i \qquad (3.10)$$

When $\lambda = 1$, $\hat{H}$ gets the values that comprise the contribution of the approximated correlation potential arises from the electron-electron repulsion ($PE_{el\text{-}el}$). Based on this the energy of a system in MPi, that we are only interested on here, can be expressed as the power of λ as:

$$Ei = Ei^0 + \lambda^1 Ei^1 + \lambda^2 Ei^2 + \ldots \qquad (3.11)$$

where, Ei^1,Ei^2etc are the first-order, second-order correction of electronic repulsion potential. The first-order Møller-Plesset (MP1) is basically the Hartree-Fock energy corrected through first-order in Møller-Plesset perturbation theory. As MP1 properties are close to the HF and not substantially better than that second-order Møller-Plesset (MP2) perturbation theory is widely accepted for the *ab initio* calculations for maintaining the compromise between the computational-cost and the results..

3.2.5 Density Functional Theory

Development of the Density Functional Theory (DFT) started as early as in 1964 by Hohenberg and Kohn (Hohenberg and Kohn, 1964). Their quantum mechanical approach is based on the electron density, ρ, instead of the wave function for calculating the ground state molecular properties. The logic behind using the electron density came from the fact that the Hamiltonian in the Schrodinger equation (eq. 3.1&2) depends upon the total

number of electrons present in the system and that this property (of the Hamiltonian) can be successfully replaced by the electron density calculated over the electronic volume (Andrew R. Leach, 2001, Cramer, 2004). Therefore the total number of electrons should be obtainable by integrating the electron density over the same volume as stated by eq. 3.12:

$$N = \int \rho(r) dr \tag{3.12}$$

The potential energy originated due to the electron-nuclei attractive force ($PE_{el\text{-}N}$, also termed as external potential (V_{ext})), of the Hamiltonian (eq. 3.2&3) was also explored with the fundamentals of electron density. It explained that the position of the nuclei should be where the local maxima of the respective electron density exist and the nuclear atomic numbers can be derived by the formalism given as in equation 3.13:

$$\partial \rho(r_i) / \partial r_i \mid r_{i=0} = -2 Z_i \rho(r_i) \tag{3.13}$$

where Z is the atomic number of i, r_i is the distance from i, and ρ is the electron density averaged over radius r_i.

However, instead of applying the electron density dependent variables directly into the Schrodinger equation (eq. 3.1&2), the simplified or approximated forms were adopted in DFT by Kohn and Sham in 1965 (Kohn *et al.*, 1965) in the Kohn and Hohenberg formalism (eq. 3.14) to calculate the system energy (eq. 3.15):

$$E[\rho(r_i)] = \int V_{ext}(r)\, \rho(r) dr + V_{sum}[\rho(r)] \tag{3.14}$$

$$V_{sum}\,[\rho(r)] = KE[\rho(r)] + PE_{el\text{-}N}[\rho(r)] + PE_{el\text{-}el}[\rho(r)] + PE_{ec}[\rho(r)] \quad (3.15)$$

where $V_{sum}\,[\rho(r)]$ is a sum of different energy contributions as depicted in the equation 3.15. The potential energy terms $PE_{el\text{-}el}[\rho(r)]$ and $PE_{el\text{-}N}$ are the contributions originated from the inter-electronic interactions and electron-nuclear interaction. On the other hand KE $[\rho(r)]$ is the kinetic energy of a hypothetical system that are composed of non-interacting electrons but with an electron density $\rho(r)$ similar to the real system whereas $PE_{ec}[\rho(r)]$ is the contribution arises due to the electron exchange and correlation and added to compensate the $KE[\rho(r)]$. All the energetic terms are derived using the following formalisms::

$$KE[\rho(r)] = (-½) \sum_{i=1}^{N} \int \Psi_i(r)\, \nabla_i^2\, \Psi_i(r)\, dr \quad (3.16)$$

$$PE_{el\text{-}el}[\rho(r)] = (½) \iint [\rho(r_1)\, \rho(r_2)/d_{12}\,]\, dr_1\, dr_2 \quad (3.17)$$

$$d_{12} = |\, r_1 - r_2|$$

$$PE_{el\text{-}N}[\rho(r)] = - \sum_{a=1}^{A} \int (Z_a/d_{ia})\, \rho(r)\, dr \quad (3.18)$$

$d_{ia} = |r_i - r_a|$, where, i stands for electron, and a for atom

Basis Sets

The electron exchange-correlation term $PE_{ec}[\rho(r)]$ can be derived from several functional forms which are based on Gaussian functions. These sets of functions which are used to represent the molecular orbitals are known as basis sets. The widely used basis set is the B3LYP. B stands for the

exchange functional derived using the Becke's (Becke, 1993) approach whereas LYP stands for Lee, Yang and Parr (Lee *et al.*, 1988) approach to calculate the correlation part. Altogether B3LYP is represents a hybrid formalism in which the Becke's exchange functional is combined with the exact energy from Hartree–Fock theory where three fitted parameters (hence B3LYP) define the amount of hybridization of the exact energy with the exchange part.

3.3 Classical Mechanics:

Classical mechanics, in a broad sense, is used to describe the dynamicity and corresponding potential energy of a system under the external forces. When methodologies based on Quantum Mechanics (QM) calculations are used to calculate and predict the relative energies and the energy barriers between different conformations Classical Mechanics or Molecular Mechanics/ Molecular Dynamics (MM/MD) is used to search or sample the conformational space to get access to different conformations of the molecules. Classical mechanics or MM/MD evolved from fundamentally different perspectives than that of the QM. In contrast to the QM, where a molecular property is represented as a function of its electronic properties, Classical mechanics uses atomic representation of a molecule to calculate its energetics as well as dynamics. Doing so, it reduces a large number of degrees of freedom required to calculate the energetics of a molecule. Since molecular mechanics uses much simpler energy functions to depict a state of a molecule in terms of energy, it has been used widely in solving biological problems, which is generally associated with macromolecules (proteins,

DNA, etc.) containing several hundred thousand atoms, that are otherwise not quite feasible to deal using QM. However, the success of MM lies on the functional representation of different parameters associated to an atom or group of bonded or non-bonded interacting atoms to properly describe the potential energy. Collectively, the definitions of intra- or inter- molecular properties in terms of parameters and forces to describe the energy functions are known as force fields.

3.3.1 Force Fields and Potential Energy Function:

Although different force fields (FF) use different approaches in defining the potential energy functions the fundamental approximations remain identical. All such FFs use general assumptions like:

i) each atom is represented as a spherical entity with certain radius and net charge
ii) two consecutive atoms are bonded with a specific equilibrium bond length (covalent or 1-2 interactions)
iii) three consecutive bonded atoms are related with certain equilibrium value of angle (1-3 interactions), and
iv) four consecutive bonded atoms (1-4 interactions) are expressed with a torsional terms or dihedral angle
v) definition of non-bonded atom pairs close in space,

The potential energy (PE) function of a molecule in a specific state or conformation is expressed as the summation of energetic terms derived from the above bonded and non-bonded or covalent and non-covalent interactions:

$$PE = PE_{bonded} + PE_{non\text{-}bond} \tag{3.19}$$

$$PE_{bonded} = PE_{bond} + PE_{angle} + PE_{torsion} + PE_{improper} \quad (3.20)$$

$$PE_{non\text{-}bond} = PE_{electrostatic} + PE_{vdw} \quad (3.21)$$

where each energetic contribution evolves due to the distortion from the equilibrium conditions between two or more atoms. A short review on each term is stated below. The parameters used to define each bonded and non-bonded terms vary from each other.

3.3.2 Definitions of Bonded Contributions:

As molecules are at constant movement due to the external forces e.g. temperature, pressure, etc., bonds, angles and torsions are also subjected to change from their stable equilibrium form. This is what causes the bonds between atoms to stretch or vibrate, angles to rotate, or change in the dihedrals angle and ends up with changing in the potential energy.

3.3.2.1 Bond Vibration Energy:

The potential energy due to the bond length distortion follows Hooke's law where the energetic penalty is proportional to the square of equilibrium length distortion due to bond vibration. Hence PE_{bond} can be expressed as:

$$PE_{bond} = \sum_i^n k_b (b - b_{eq})^2 \quad (3.22)$$

where, b_{eq} is the equilibrium bond length, b is the new bond length, k_b is the force constant. The energy function is applied to all the bonds in a molecule.

A typical plot of the bond length dependant energy change has been shown in Fig B below.

3.3.2.2 Angle Bending Energy:

Potential energy change due to the change of angle (ø) between three consecutive atoms from its equilibrium value (øeq) also often defined with the harmonic energy function derived from Hooke's law (eq. 3.23).

$$PE_{angle} = \sum_i^n k_A (ø - ø_{eq})^2 \qquad (3.23)$$

However the significant energy and structural changes in molecules are not very often due to the above two terms (due to the covalent nature), rather dominated by the torsional and non-bonded interactions.

3.3.2.3 Torsional Energy:

Torsion or dihedral angle is formed by four consecutively bonded atoms and the measure of the torsion is the angle formed due to the rotation about the bond formed by two central atoms [Fig A] or more precisely between the two planes formed by three of the four consecutively bonded. This rotation is responsible for the major conformational changes in a flexible molecule along with the non-bonded interactions.

Torsion potentials are most commonly represented as Fourier series expansion of cosine:

$$PE_{tor} = \sum_i V_n [\cos(n\omega_i - \Phi) + 1] \qquad (3.24)$$

where, ω is the torsion angle, Φ is the phase angle, V_n is the barrier height required to rotate with torsion angle ω about the central bond. n is the number that determines the periodicity or number of occurrences of maximum/ minimum in the energy profile in the range of 0 to 360°. n varies mostly from 1 to 4 or more if required to describe a particular torsion. i represents all the torsions present in the molecule or in the system.

3.3.2.4 Improper Torsion Potential

Other than above mentioned different bonded energetic terms that are used to maintain a favorable molecular conformation, improper torsion energy potential is also needed to maintain the planarity among the atoms or to treat the distortion due to the out-of-plane conformation. In contrast to the torsion angle that considers four consecutively bonded atoms and a central bond by two central atoms improper torsion potential is calculated involving a central atom that is bonded to the other three atoms. Energy potential for improper dihedral is calculated using the expression below:

$$PE_{tor} = \sum_i k_{imp}(1 - \cos 2\theta_i)] \qquad (3.25)$$

θ is the out-of-plane angle and that is maintained at either at 0° or 180° to maintain the planarity.

3.3.3 Non-Bonded Energetic Contributions:

Non-bonded interactions between atoms close into the space account for the most important contribution in the estimation of total potential energy of a molecule. As mentioned above (eq. 3.21) the majority of the force fields use the two common independent non-bonded terms, electrostatics and Van

der Waals (vdw) interactions. A very short review on this account is given below.

3.3.3.1 Electrostatic Interactions

Electrostatics interaction is a common and key phenomenon between two charged atoms. The interaction potential, due to the electrostatic attraction or repulsion forces, is a function of the charge of the atoms, distance between them and the medium in which the atoms are interacting. Classically it is assumed that the total charge of a molecule is distributed among its atoms according to their respective electro-negativity in contrast to the Quantum Mechanics where the atomic charge is a function of the electron distribution around the nuclei. Therefore, in electrostatic interaction two atoms interact with the respective partial charges. In molecular mechanics force fields the potential energy due to the electrostatic intervention into each pair of non-bonded atoms is estimated using the Coulomb's law:

$$PE_{ele} = \sum_i \sum_j \frac{q_i q_j}{4\pi\varepsilon r_{ij}} \qquad (3.26)$$

q_i and q_j are the partial charges on atoms i and j and r_{ij} is the distance between them whereas ε is the dielectric constant of the medium to mimic the real medium of the interactions. Commonly, particularly in biological systems, molecules are surrounded by a solvent, ions, or other components. Therefore, in the ideal case, the solvent would be explicitly simulated and the interactions with the solvent would be computed making it unnecessary to use a dielectric constant (ε = 1). However, this approach is computationally

expensive and a more efficient alternative consists in treating the solvent implicitly by approximating the interference effect of the bulk-medium in the interactions. There are several approximations that are widely used to simulate the continuum effect of bulk solvent: dielectric constant, distance dependent dielectric constant, Poisson-Boltzmann analytical equation, or Generalized Born methods among others.

3.3.3.2 van der Waals Repulsion

The van der Waals forces between a pair of atoms are most commonly computed with the Lennard-Jones potential function. It accounts for the energy due to repulsion between atoms at very small distance and weak attraction at larger distances. This potential arises due to the fact that atoms have a finite size that could be assumed as a hard sphere of certain radius, which restricts two approaching atoms to a certain distance. Small change in this separation highly increases the energy of the system. Lennard-Jones potentials, hence, are used to prevent the collisions between atoms in the molecules and therefore make an important contributing factor in the total energy. While electrostatic interactions dominate the interactions between two charged atoms, van der Waals interactions explain the energetic contributions due to the interactions between two approaching neutral atoms.

The commonly used Lennard-Jones function to compute the van der Waals interactions is expressed as:

$$PE_{vdw} = 4\varepsilon \sum_i \sum_j \left(\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right) \tag{3.27}$$

where r is the distance between the atoms, ε is the depth of the potential well of each interacting pair, σ is the equilibrium distance of that particular atom pair at which the potential is zero.

Each equilibrium parameter and constant in the force fields normally obtained either from experimental analysis or from the higher order quantum chemical calculations. Therefore, contrary to the *ab initio* methods, molecular mechanics force fields are empirical in nature.

3.3.4 Solvent models:

As mentioned before, biomolecules are surrounded by the fluid containing water molecules, ions, or other components in the cell. Hence to obtain accurate or close to accurate energetics of them it is necessary to treat the molecules in an ideal cell-like environment in presence of solvent. In addressing regular queries of biological problems explicit treatment of water is the most common choice among the solvents. However treating solvent implicitly is also common in the molecular mechanics. Therefore two general approaches when dealing with solvent are used: explicit water models and implicit solvent models..

3.3.4.1 Explicit solvent model:

In explicit solvent treatment energy contribution from each water molecule is added to the total system. Commonly, the water molecule is modeled as a rigid body with fixed OH-bond distance as well as the H-O-H

angle. The interaction energy between two water molecules or with another non-water molecule is computed using Coulombic (electrostatic) and Lennard-Jones (van der Waals) potential functions.

Different rigid water models with varying approach is commonly used in simulating the presence of water as solvent. They are classified based on the number of assigned interaction sites on them. Some of the examples are SPC, SPC/E, (Berendsen *et al.*, 1981) TIP3P, TIP4P, TIP5P etc. The most frequently used water model is TIP3P (Jorgensen *et al.*, 1983, Mahoney and Jorgensen, 2000) and, as the name suggests, it contains three interaction sites i.e. each atom of the water molecule with partial point charge (two hydrogen and the oxygen) is able to interact with electrostatic interactions with other molecules while the van der Waals interactions are ignored for the hydrogen atoms. This is widely used model for the molecular simulations due to the simplicity in representation and computational time efficiency. SPC model is also a three-site model but uses slightly different parameters than the TIP3P model.

Unlike TIP3P TIP4P or other TIPXP (X > 3) models, as mentioned above, use additional sites or dummy atoms with certain partial point charge to obtain better charge distribution around the shell of an water molecule. TIP4P delivers the best quality results, (Beauchamp *et al.*, 2012) but the additional interaction point decreases the computational performance, compared to TIP3P.

3.3.4.2 Implicit solvent

In the implicit solvent treatment a layer of continuous medium with a specific dielectric property is represented to mimic the bulk solvent effect on the solute. Hence it is known as continuum model of solvent (Fig 3.1). A wide range of continuum solvent models are used in to compute the free energy of solvation in the molecular mechanical calculations. Among various methods surface area based approaches e.g. numerically solvable Poisson-Boltzmann (PB) and simplified Generalized Born (GB) equations are widely used. Solving the complex equations of PB numerically though produces the most accurate data; the computational cost is prohibitive for most applications. GB is the simpler approximation of the PB that uses empirically derived parameters. Therefore accuracy of the results is highly dependent on the parameters in use. Besides, in both PB and GB, the hydrophobic or entropic effect, which is a non-additive term due to the interaction of solvent molecules with the solute, is not specifically taken into account. This is compensated with the incorporation of the approximated measure of solvent accessible surface area of the molecule as used, for example, in PB or GBSA methods.

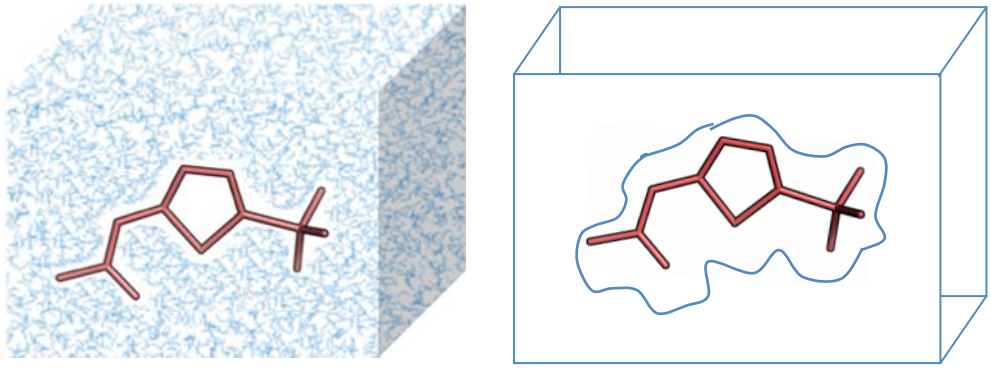


Fig. 3.1: Left: a schematic representation of a cross section of a box of solvent (water) treatment applied on a molecule. Right: similar representation but here with a continuum solvent treatment applied on the same molecule.

Other than the simple rigid body explicit solvent water models or various implicit solvent models, efforts are being taken to include explicit polarization/ionization effect in the current water models to more accurately deal with the charge transfer in the solute-solvent interface or from interface to core of the solute effect of presence of ions.

3.3.5 Popular Force Fields:

There are a large number of force fields that are used with specific purpose in the classical mechanics calculations on small molecules and macromolecules. However the widely used force fields are:

AMBER (Assisted Model Building and Energy Refinement) force fields: ff99SB (Hornak *et al.*, 2006), ff03 (Duan *et al.*, 2003) for protein simulations and ff99bsc0 (Pérez *et al.*, 2007) for the nucleic acid simulations. CHARMM (Chemistry at HARvard Molecular Mechanics), GROMOS (GROningen MOlecular Simulation) (Brooks *et al.*, 1983) package force field is one of the most popular and widely used force fields for simulating protein and nucleic acid and organic molecule whereas OPLS (Optimized Potentials for Liquid Simulations) (Jorgensen and Tirado-Rives, 1988), MMFF (Merck Molecular Force Field) (Halgren, 1996b), GAFF (General AMBER Force Field) (Wang *et al.*, 2004) specifically designed for small organic molecules and are commonly used in the rational drug design ventures.

3.3.6 Force fields based methods

3.3.6.1 Optimization or Minimization of Molecules

The most popular application of the empirical force fields is to search for a minimum energy conformation of a molecule. The process of finding the lowest energy geometry of a particular conformation of a molecule is known as energy minimization or optimization. That particular geometry of the molecule might or might not be the global energy-minimum conformation of the molecule. Search for the global energy minimum might expose to several other conformations that are lowest in energy among their neighboring conformations (hence local energy minima) and are separated from each other by higher energy conformations. This scenario of distribution of different energetics of a particular molecule could be assumed as a surface landscape with many ups and downs and hence it is known as potential energy surface (Fig 3.2).

Different algorithms are used to find the points that are located at the downhill i.e. the local minimum of the energy surface. Most common among them are (Andrew R. Leach, 2001):

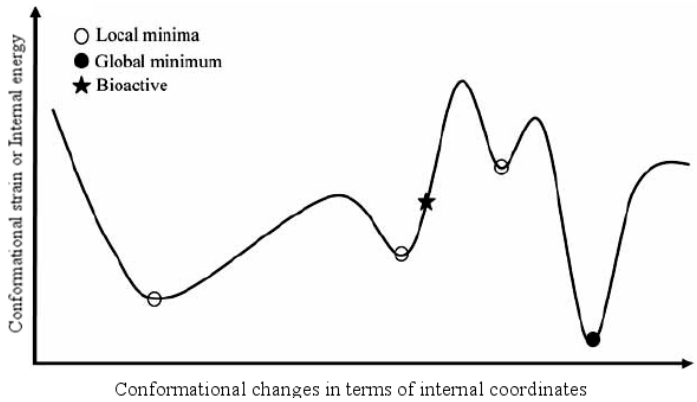


Fig. 3.2: Potential energy surface with respect to a reaction coordinate. Figure adopted from the review article by Foloppe et al. Current Medicinal Chemistry, 2009

1) Simplex Method: Although computationally very inefficient finding a nearest minimum with this algorithm is obvious. This is a very useful optimization method when the starting structure is very strained and thus high in energy (Andrew R. Leach, 2001).
2) Steepest descent and Conjugate gradient: These are the currently most popular methods for optimizing both small and macro molecules. They are computationally very efficient and find the nearest minimum point with a faster rate (Andrew R. Leach, 2001).
3) Newton-Raphson: Although it finds the nearest minimum point on the energy surface at a fastest rate among the currently popular methods it is computationally very inefficient due to the huge requirement of the memory. It is mostly used to minimize a small molecule (Andrew R. Leach, 2001).

3.3.6.2 Molecular Docking:

Another usefulness of the force fields is the molecular docking approaches where formation of intermolecular complexes, primarily between a flexible protein and a flexible inhibitor is predicted. The accomplishment of docking implies correctly predicting an orientation of a small molecule that fits into its macromolecular receptor i.e. molecular recognition. Accurate prediction of an event of molecular recognition and a complex selection i.e. by filtering out all the incorrect binding poses in terms of a score is proportionally related with the measure of binding affinity of a ligand with its corresponding receptor. However the success of docking lies in its scoring

function that is used to score a complex formation that compliments its degree of acceptance.

Among different approach based scoring functions (e.g. empirical scoring functions, knowledge based, consensus scoring functions, etc) force fields based scoring functions are also commonly used in the drug discovery projects (Sousa *et al.*, 2006). The well recognized limitations in the force field based scoring functions include approximation in the estimation of solvation energy (mimicked using implicit solvent) and exclusion of the entropic cost upon binding. These are still active fields of research..

3.3.6.3 Molecular Dynamics:

Estimation of potential energy of a locally minimized configuration of a molecule provides insight only to a particular static state of its potential energy surface. However, due to its inherent flexibility, biomolecules can access a huge number of conformations at room temperature, which are largely influenced by the constant interactions with solvent molecules, ions or other biomolecules. Therefore to obtain physically meaningful features of the biomolecules or to calculate different thermodynamic properties it is imperative to generate energetically favorable ensemble of structures of a biomolecular system in presence of solvent and ions. Widely used methods for this are molecular dynamics (MD) and Monte-Carlo (MC) simulations. However Monte-Carlo simulations do not produce a time evolution of the systems, instead it gives an ensemble of energetically favorable different conformations that might exist in the conformational space of the molecule.

In contrast to that MD simulations generate time-dependant ensembles of conformations and, hence, an ideal method of approximation to investigate the dynamic behavior of the system that are with constant external forces.

Classical molecular dynamics simulations obey the famous Newton's laws of motion with specific use of its second law of motion. According to the Newton's second law the rate of change of velocity i.e. Δv over time or $\frac{d}{dt}\left(\frac{dx_i}{dt}\right)$ or the acceleration (a) of a particle *i* of mass=m is equal to the applied force F. i.e..

$$\mathrm{F}x_i = m_i \frac{d}{dt}\left(\frac{dx_i}{dt}\right) \tag{3.28}$$

The right hand side represents the change of velocity or acceleration where x_i is the initial position of the particle I, dx_i is the change in the initial position and t is the time period of applied force $F_{x.}$ in a particular direction times the mass of the particle..

In equation 3.28,

$$a = (d^2x/dt^2) \tag{3.29}$$

$$v = dx/dt \tag{3.30}$$

Integrating equation 3.29 and 3.30 they can be deduced to the equation 3.31:

$$x = a.t^2 + v.t + x_0 \tag{3.31}$$

x_0 is the initial position and x is the position after time t.

On the other hand it is known that the potential energy (PE) is a function of mass (m), acceleration (a) acting on a particle, and distance (x) from the initial position which could be written as in equation 3.32.:

$$dE = m.a.(dX) \qquad (3.32)$$

Basically, in its most simplistic form, MD algorithm calculates the energy of the system and the trajectory at each Δt time from the initial positions of the atoms and randomly assigned velocity on them. This process is used iteratively to get a time dependent trajectory of a system. Each biomolecular system is composed of several thousands of atoms (N) and hence 3N degrees of freedom which make the integration of the equations of motion impossible to solve analytically. Usually a proper numerical algorithm is chosen to solve those equations.

Some Important Features of MD Simulations:

3.3.6.3.1 Periodic boundary condition:

Usually simulating a biomolecule surrounded with solvent layers is carried out within a box of specific shape and dimension (primarily a cubic box). The most commonly used box is truncated octahedron. However using a fixed sized box is not a realistic model to mimic the behavior of the cellular environment. Besides, a significant proportion of the solvent molecules, if confined within the box, would be facing the boundary wall of the box and therefore affected by the surface properties of the boundary. To avoid this surface phenomenon the box is allowed to replicate its own image in all three

dimensions in a periodic array such that solvent molecules facing the boundary feel the properties of the bulk fluid. This also helps in treating the self diffusion of the molecule in the bulk solvent i.e. during the simulation if part of the molecule goes out of the original simulation box it reappears from its opposite side crossing the boundary of the image box. The calculations then carried out on the approaching image in the central box and the effects are replicated in all other image particles. This way the number of particles, or atoms here, is kept constant in the central box. Mathematically, if an atomic position is a(x, y, z) the coordinate of its image (a′) can be obtained by adding the length of the cubic box (say, d):

$$a' = a\,(x, y, z) + d$$

where, d is the length of the cubic box. The new coordinate position is function of the direction in which the atom reenters.

3.3.6.3.2 Long-Range Forces and Cut-offs Distances:

A biological system contains thousands of atoms that are under the constant inter-atomic interaction forces. While the bonded interactions are relatively easy to handle with limited number of interactions the non-bonded interactions that accounts for the electrostatic potential between two atoms requires computation of a large number of potential interactions, which by its definition (eq. 3.26) is equal to N x (N-1)/2 if all pairs are being considered.. Here N is the number of atoms present in the system, including the solvent. This is how it becomes computationally the most demanding process in the molecular simulations. One way to treat this is using a cut-off radius around each atom (normally between 8 Å to 15 Å for the electrostatic, and shorter

for the Lennard-Jones term) within which all pair-atomic potentials are computed and atom-pairs separated with this distance are discarded assuming they don't feel the attraction/repulsion forces. However this method does not significantly improve the speed of the process distances between each pair must be calculated anyway. Moreover truncating interactions might add up some significant errors in the total calculation. The most regularly used method to treat the long range interactions with periodic boundary conditions is the Ewald summation (Ewald, 1921) method with particle-mesh modification (Darden *et al.*, 1993) (hence PME, particle-mesh Ewald) on charge distribution around each atom. In PME the assignment of the point charge is replaced by the grid based continuous charge distribution that reduces significantly the cost of computation.

3.3.6.3.3 Constant Temperature, Pressure, and Volume Conditions in MD

Pressure control:

To reproduce the experimental conditions in the simulation process the widely accepted approach is to use constant temperature and pressure (NPT) with invariable number of particles in simulating a system of biomolecules. However due to the additional volume scaling factor (λ_P in eq (3.33)) to adjust the volume of the box and maintain the density at each time step t in the constant pressure dynamics i.e. in NPT it becomes computationally more expensive than simulating with constant temperature and volume (NVT). In NPT the volume is maintain with following expression:

$$\lambda_P = 1 - \kappa\,(\delta t/\,\tau_P)[P_t - P_{bath}] \qquad (3.33)$$

where τ_P is the pressure coupling constant (similar to the Berendsen temperature control) that couples the system to an external pressure bath of a desired constant pressure P_{bath} to adjust the system pressure. P_t is the pressure of the system at time t.

Although NVT is widely used for the MD production run it is recommended to simulate an NPT simulation to get the proper volume and density of the box.

Temperature control:

The temperature is maintained using different algorithms, among which the most popular ones are Berendsen thermostat and Langevin dynamics. Although Langevin thermostat ensures even heating of the total system it can introduce artifacts of similar sequence of events in long runs, which could be hypothesized as the dynamic behavior if overlooked. The artifacts arise from the fact that Langevin algorithm uses pseudorandom numbers as forces to modify the velocity of the particles and if repetitive series of pseudorandom numbers are generated by the machine or by the program itself, the simulating system becomes susceptible towards producing recurring dynamic events (Cerutti *et al.*, 2008). Therefore, Langevin thermostat coupling would be best to use with proper random seeds or a time dependent random number.

Berendsen thermostat (Berendsen *et al.*, 1984) uses a weak coupling constant to connect the system to an external heat bath of desired temperature. However Berendsen temperature control algorithm only

maintains the conserved kinetic energy of the system by scaling the velocities of the particles. This is done using the following mathematical expression:

$$\lambda_T = \text{sqrt}(1+ (\delta t/ \tau_T)[(T_{bath}/T_t) - 1] \quad (3.34)$$

Again λ_T is the velocity scaling factor and τ_T is the coupling constant that decides the degree of coupling between the system and the external heat bath.

3.3.6.4 Metadynamics:

Molecular dynamics simulation is a time dependent equilibrium process that evaluates the microscopic properties of molecules, especially of biological interest, with high resolution. Biological process or events of interest might take place in longer time scale and to obtain or sample those events either a very long simulation or a number of replicas needs to run, both of which are computationally demanding procedures. Various approaches have been developed to accelerate the simulation pace to get access to the rare biological events within the realistic time frame. Such an approach is Metadynamics (MTD) which we have used in our work to address certain issues.

It is a powerful algorithm that can be used for reconstructing the free energy profile, to obtain potential energy barrier between different states or other rare events in the system of interests. In MTD, it assumes that a system can be represented with fewer reaction coordinates or collective variables, as known commonly, in space. MTD explores the conformational space of those collective variables and keeps a memory on the path or trajectory

visited. This has been accomplished by using a history-dependent biasing potential that restricts it to visit and explore the same path repeatedly and helps to escape being trapped into the local minima. The sum of the total history-dependent potential is used later to construct the potential of mean force (PMF) or potential energy surface. (Laio and Parrinello, 2002)

Chapter 4. Results and Discussion

4.2 Conformational Flexibility of Small Molecules

The energy of bioactive conformations of drug-like compounds: assessment of its magnitude and of the methods used to compute it

*Mousumi Bhattacharyya, Shuai Liang‡, F. Javier Luque, Adrian Roitberg‡ and Xavier Barril†**

Institució Catalana de Recerca i Estudis Avançats (ICREA), Institut de Biomedicina de la Universitat de Barcelona (IBUB), Departament de Fisicoquímica, Facultat de Farmàcia, Universitat de Barcelona, Barcelona, Spain and Quantum Theory Project and Department of Chemistry, University of Florida, Gainesville, Florida 32611, United States of America.

xbarril@ub.edu

RECEIVED DATE (to be automatically inserted after your manuscript is accepted if required according to the journal that you are submitting your paper to)

* To whom correspondence should be addressed: X. Barril, Dept. Fisicoquímica, Fac. Farmàcia, Av. Joan XXIII s/n, 08028 Barcelona, Spain. Telephone: +34-934029002. Fax: +34-934035987. E-mail: xbarril@ub.edu

‡ University of Florida

† ICREA

ABSTRACT

Formation of a protein-ligand complex is opposed by the loss of rotational, translational and conformational degrees of freedom, but if the ligand binds in a conformation that does not corresponds to the global minimum, an energetic penalty should also be expected. Although it is well known that the bioactive conformation rarely corresponds to the global minimum identified with current small-molecule force fields, there is no consensus yet about how frequently this occurs in reality or about the maximal conformational penalty that a ligand can attain. Here we investigate this aspect of molecular recognition, using a diverse set of 92 drug-like ligands from the PDB. The global minimum is obtained minimizing a diverse set of conformations for each ligand, while the bioactive conformation results from a restrained minimization of the crystallographic structure. Optimization at B3LYP/6-31D(d) level of theory in the presence of a reaction field (PCM solvation), followed by single point RI-MP2//aug-cc-pVDZ provides average penalties of 2.5 kcal/mol, and only 15% of structures have energies above 5 kcal/mol. Using these values as reference, common force fields and simpler solvation models are evaluated. Depending on the choice of force field, the corresponding values range between 10 and 5 kcal/mol and between 85% and 40%, respectively. Interestingly, an inverse relationship is found between the level of theory and the average internal energy. Contrary to previous reports, we find that the poor performance obtained with molecular mechanics force-fields is more frequently due to inaccuracies in the free energy of solvation provided by the Generalized Born solvation method than to the in vacuum internal energy provided by the force field. This study provides the most accurate estimate of the conformational

penalties of bioactive conformations to date, as well as a benchmark to assess the accuracy of theoretical formalisms.

INTRODUCTION

The formation of a complex between a ligand and its macromolecular target involves loss of conformational freedom for both partners, with entropic and energetic consequences that are difficult to quantify. Quantitative prediction of this magnitude is important in drug design because reduction of this unfavourable term can lead to increased binding affinities, as demonstrated by the classical strategy of introducing conformational restrictions. (Mann, 2003) Correct predictions of conformational free energies (i.e. inclusion of the entropic component) requires a detailed knowledge of the conformational landscape of the ligand that can only be attained with labour and computationally intensive calculations coupled with sampling techniques such as molecular dynamics or Monte-Carlo.(Kirschner *et al*., 2007, Forti *et al*.,) However, and from a practical point of view, predicting the relative conformational energy of the bioactive conformation would already be very useful. This is a far less ambitious goal that involves identifying the lowest energy conformer and comparing it with the bioactive conformer. Considering the small size of drug-like molecules and their relative rigidity (e.g. oral drugs have only 5.4 rotatable bonds on average (Vieth *et al*., 2004)), this may seem an easy task. Far from it, the standard methods used to calculate this magnitude carry large errors. For instance, Tirado-Rives and Jorgensen estimated that conformational energies determined at the ab initio or DFT levels carry uncertainties of around 5 kcal/mol, but this value increases up to 10 kcal/mol when current force fields

are used.(Tirado-Rives and Jorgensen, 2006a) This might explains the apparent contradiction between expected and calculated energies of bioactive conformations. Considering that an energetic penalty of 1.4 kcal/mol would lead to a 10-fold decrease in binding affinity, reasonably potent ligands are unlikely to adopt conformations that are penalized by more than 3 kcal/mol. However, most studies have found that bioactive conformations may have much higher energies. Initial studies in the 1990's investigated small sets of relatively simple compounds, but energy penalties above 5 kcal/mol were frequently identified, already noting the limitations in the energy functions and the fundamental role of solvation. (Nicklaus *et al*., 1995, Bostrom *et al*., 1998a, Vieth *et al*., 1998) More recently, Perola and Charifson created a much larger and representative set, formed by drug-like and diverse ligands of multiple protein classes, finding energy penalties above 5 kcal/mol in 40% of the cases and above 9 kcal/mol in 10% of the cases.(Perola and Charifson, 2004a) Such disheartening results present a serious problem for medicinal chemistry in general and computer-aided drug design in particular. Firstly, they suggest that obtaining reliable estimates for the energy of bioactive conformations is not possible, whereas reduction of the conformational strain is a fundamental strategy in hit and lead optimization. In this study we set out to investigate what is the real conformational strain in bioactive conformations of drug-like compounds. To do that we have investigated several levels of theory (Table 1), providing a methodological recipe that can be used when accurate determination of the bioactive conformation is needed. The second major problem posed by the uncertainties in the conformational energies is that, as stated by Tirado-Rives and Jorgensen, they "fundamentally undermine accurate evaluation of absolute free energies of binding". (Tirado-Rives and Jorgensen, 2006a) In fact, given that the

methods used to obtain conformational energies underpin all three-dimensional (3D) computer-aided drug design methods, their poor performance seriously compromises the outcome of some of the most widely used techniques in pharmaceutical research, including molecular docking,(Brooijmans and Kuntz, 2003) pharmacophore-based methods (Langer and Hoffmann, 2006) or 3D-QSAR.(Kubinyi, 1993) Our results provide important indications of how current methods should be supplemented to improve their performance.

Table 1. Summary of the computational methods used.

Method		Description
E_{inter} [1]	MMFF	Merck Molecular Force-Field, introduced in 1996 (Halgren, 1996a) and updated in 1999,(Halgren, 1999) is one of the most popular force-fields for organic molecules.
	OPLS	Optimized Potentials for Liquid Simulations. Developed in Jorgensen's lab for the purpose of simulating organic molecules in solvation, particularly in combination with the TIP3P or TIP4P water models.(Kaminski *et al*., 2001) Version 2.0 has recently been developed by Schrödinger.
	GAFF	General AMBER force field. Develpedto facilitate simulations of drugs and small molecule ligands in conjunction with biomolecules.(Wang *et al*., 2004)
	B3LYP	Becke, three-parameter, Lee-Yang-Parr. Hybrid functional in density functional theory (DFT). Together with the 6-31G* basis set provides a good balance between speed and accuracy for many applications and is a common 'entry-point' level for quantum mechanical calculations.(Riley *et al*., 2007)
	RI-MP2	Resolution of Identity MP2. Albeit 5-20 times faster than conventional MP2 (second order Møller–Plesset perturbation theory), it identically reproduces relative energies.(Weigend *et al*., 1998) In conjunction with the aug-cc-pVDZ basis set, is the most accurate methods to calculate internal energies of molecules.(Riley *et al*., 2007)
ΔG_{solv} [2]	GBSA	Generalized Born model augmented with the hydrophobic solvent accessible Surface Area term. It is one of the most commonly used implicit solvent models, particularly in molecular mechanics.(Chen and Brooks, 2008a) Many different parameterizations exist.
	PCM	Polarizable Continuum Model. Implicit solvation model commonly available in several quantum chemical computational packages.(Tomasi *et al*., 1999) PCM models are generally parametrized for the most common QM theory levels (e.g. B3LYP/6-31G*).
	MST	The Miertus-Scrocco-Tomasi solvation model is a particular implementation of PCM used here.(Soteras *et*

		al., 2005)

[1]Methods used to calculate the intrinsic energy of the molecule (internal energy). [2]Methods to calculate the effect of aqueous solvation.

MATERIALS AND METHODS

Dataset. As in our previous work (Butler *et al*., 2009a) we have used the public part of the dataset of drug-like molecules compiled by Perola and Charifson at Vertex, (Perola and Charifson, 2004a) which consists of 100 target-bound ligands of known binding affinity. The previously analyzed structures were further checked for their temperature factors and occupancies, excluding some molecules containing atoms with multiple occupancies or lacking proper B-factor data (identical temperature factors for all atoms). The final set consists of 92 chemically diverse and drug-like molecules. As the positions of hydrogen atoms are not directly observed, they were generated from the connection tables. PDB files with the experimental B-factors and MDL SDF files, where connectivity is unambiguously assigned, are provided as Supporting Information.

Defining the bioactive conformation.

Several restraining schemes are discussed in the results section to obtain a bioactive conformation compatible with each method. The corresponding formulae are displayed below. In all cases, restraints are only applied to non-hydrogen atoms.

1) Harmonic restraints with force constant derived from the crystallographic B-factors. These are the T/B restraints introduced in

our previous work, (Butler *et al*., 2009a) where atoms are subjected to harmonic constraints with a force constant (κ) dependent on the atomic B factor (B) and the temperature at which the diffraction pattern was obtained (T):

$$\kappa = 4\,\pi^2\,k_B\,T/B \qquad (1)$$

where k_B is the Boltzmann constant.

2) Harmonic restraints with force constant derived from rescaled B-factors. Since the absolute B-factors can be meaningless, an option to ensure that all ligands have similar average constratins while still allowing different relative mobility to individual atoms is to rescale B factors to a reference value (B_{ref}):

$$B' = B \cdot (B_{ref} / B_{avg}) \qquad (2)$$

where B_{avg} is the average B-factor of all atoms in the ligand. In doing so, the temperature dependence in eq. 1 is no longer meaningful and a common value should be used instead. We have tested several B_{ref} values in combination with a fixed reference temperature (T_{ref}=300K). The final choice was B_{ref}=10Å^2, in which case the force constant for the harmonic constraint takes this form:

$$\kappa = 4\,\pi^2\,k_B\,30\,B_{avg} / B \qquad (3)$$

The choice of B_{ref} was motivated by the fact that, on average, it provided similar results to those obtained on our previous work.(Butler *et al*., 2009a) It should also be noted that for B=10Å^2, the maximum theoretical

displacement from the crystallograpfic coordinates is 0.5Å, which corresponds to the half-width of the flat-bottom restraints used by Perola and Charifson. However, the use of harmonic constraints results in bioactive conformations more similar to the crystallographic ones because any displacement from the initial coordinates is penalized.

Generation of the local energy minimum.

The conformation obtained from the PDB i.e. the target-bound conformation was minimized directly with each of the methods investigated. We refer to this conformation as the local minimum.

Generation of a conformational ensemble.

A set of conformers were generated with the Confgen program of the Schrödinger suit(Watts *et al*., 2010), using the default parameters. An average of 16 different conformations were generated per compound, ranging from 48 (1UVS) to only one conformation (4 ligands). Each conformation was minimized with the various levels of quantum chemical theory and molecular mechanic approaches. The local minimum conformation was also added to this set. The conformer with the lowest energy in the whole set is termed as the global minimum.

Molecular Mechanics – GAFF Force Field.

All the parametrizations of the bound states and their conformers of each ligand molecule are performed with the Antechamber suite of

AmberTools1.5, (Case *et al*., 2010) in combination with GAFF(Wang *et al*., 2004)and AM1 BCC charge methodology.(Jakalian *et al*., 2002) GAFF has been designed to be compatible with AMBER force field for proteins and nucleic acids.(Ponder and Case, 2003)Followed by the parametrizations of these molecules, we conducted the energy minimizations to these ligand molecules using the Sander module of the AMBER11 package.(Case *et al*., 2010) The energy minimizations are performed in both vacuum and solution, where the solution model is using an advanced Generalized Born implicit water model (igb = 5 in Sander).(Onufriev *et al*., 2004)The truncated Newton conjugate gradient (TNCG) method is used with the convergence criterion for the energy gradient being set as 10^{-4} kcal·mol^{-1}·Å^{-1}. Under several cases, the minimizations with TNCG method cannot reach the predefined energy convergence criterion. These cases are further equilibrated by short MD simulations to help the system to escape from local energy traps. The energy minimizations by TNCG method are then conducted again to these molecules. The MD simulations and the TNCG minimizations occasionally need to be performed iteratively for the energy convergence criterion to be reached. This procedure effectively works for 90 out of the 92 ligands and their conformers. Two ligands, 1M48 and 3STD are excluded in this study since the energies of several conformers do not converge well by the current energy minimization procedure. Since the generation of partial charges depends on the tree-dimensional structure of the ligand, they were generated using the bound conformations from the PDB and transferred to all other initial conformations.

Molecular Mechanics – MMFF and OPLS Force Fields.

All calculations using the MMFF (MMFF94 (Halgren, 1996a) and MMFF94s (Halgren, 1999)), OPLS2005(Kaminski *et al*., 2001)and OPLS2.0 force fieldswere carried out with MacroModel, part of the Schrödinger suit.The coulombic term was calculated with a dielectric constant1 (vacuo) or with the Generalized Born method (GB/SA) implemented in the program. For GB/SA, the internal dielectric was set to 1 and the external to 80, corresponding to aqueous solvation. In OPLS2.0, partial charges were calculated for the bioactive conformation and transferred to all other initial conformations with a script kindly provided by E. Harder (Schrödinger Inc.).

Quantum Mechanics

Quantum mechanical (QM) calculations were performed with the Gaussian09package (Frisch *et al*., 2009) with the B3LYP density functional method and 6-31G* basis set. Initially, optimizations were carried out using the default choice ofpolarizable continuum model (PCM) parameters for aqueous solvation. However, after initial results showed proton transfers between polar groups and a tendency to form collapsed structures, a hybrid approach was used. The default integral equation formalism (IEF-PCM) of Gaussian-09 was preserved as it is much faster and more stable than the previous implementation. (Tomasi *et al*., 1999)However, the radii to construct the cavities and the scaling factors for charged atoms were taken from the MST solvation model that has been developed and parameterized in our group.(Soteras *et al*., 2005)As this set-up does not allow to smooth the

molecular surface with extra spheres, a separate single-point energy calculation was done on the optimized geometries using the MST model. Correlation between IEF-PCM with MST radii and the full-fledged MST is sufficiently high to support the strategy followed here (r^2=0.993).

The internal energy of each optimized conformation was further computed by a single point calculation using the RI-MP2 (Resolution-of-the-identity approximation of MP2) (Weigend *et al*., 1998) method with Dunning's augmented correlation consistent basis sets (aug-cc-pVDZ) as implemented in Turbomole.(Ahlrichs *et al*., 1989)For selected conformers, the energies were calculated with MP2 andthe same basis set. As expected, relative energies were identical in spite of the fact that RI-MP2 is computationally much more efficient.

RESULTS AND DISCUSSION

Defining the bioactive conformation.

Considering the experimental uncertainties associated with the determination of three-dimensional structures(Davis *et al*., 2003) and the sensitivity of internal energies to geometric variations, it is important to allow the molecules to relax into a conformation that satisfies the experimental structure but also agrees with the energy function of choice. Optimally, this can be done by refining the structure while simultaneously minimizing the internal energy of the ligand by means of QM methods.(Ryde and Nilsson, 2003, Li *et al*., 2009) However, experimental structure factors are not always available and, as long as the deviation from the initial

structure is small, the use of constraints provides a much simpler approach that is already implemented in existing molecular mechanics and quantum mechanics computer programs. As shown below, the average root mean squared deviations (RMSD) across all structures is approximately 0.2±0.1Å, a deviation that can comfortably be considered within the experimental error for the set of structures used in this study. In fact, the estimated error in atomic coordinates for our average resolution(2.1Å)is 0.35Å.(Cachau and Podjarny, 2005)

The actual nature of the constraints that should be used to refine the bioactive conformation is a matter of debate. Most studies published to date make use of flat-bottom constraints to ensure that the optimized geometries do not deviate significantly from the ones derived from the electron densities. However, we have shown that the strain energies obtained with this approach largely depend on the rather arbitrary choice of the half-width parameter. With this limitation in mind, we introduced the use of harmonic constraints with a force constant dependent on the crystallographic B-factor (T/B constraints).(Butler *et al*., 2009a) In principle, this is a rigorous procedure to derive suitable harmonic constraints. However, in practice, B-factors can contain noise from the refinement process and their absolute values are often meaningless.(Kleywegt and Jones, 1997) Relative B-factors, on the other hand, are very informative because they candistinguish which parts of the ligands that are better resolved (have more localized electron densities) and which are comparatively more disordered.A key feature of T/B constraints is that poorly resolved regions are allowed more freedom of movement than well defined ones, and we wanted to retain it. But considering the limitations associated with absolute B-factors, we have rescaled them to a common

average value before converting them to the force constants used in harmonic constraints (see Methods). Although this can only be seen as an approximation, it should be noted that comparison with flat-bottom restraints provide highly correlated results ($r^2 \approx 0.9$; Figure 1), indicating that the general trends are independent of the choice of restraints, although the results for individual structures may vary significantly. Other options to represent the bioactive conformation might be to use the unrestrained minimized geometry closest to the crystallographic structure (Tirado-Rives and Jorgensen, 2006a) or to place the restraints on the dihedral angles instead of on the Cartesian coordinates.(Sitzmann *et al*., 2012) We find that the former strategy leads to geometries that deviate from the initial structure significantly more than justified by the experimental uncertainty (1.0±0.7Å) and, in consequence, significantly lower energies. Applying the restraints over internal coordinates also leads to larger displacements and lower energy penalties, which, together with the poor correlation with the results obtained with Cartesian restraints (Figure 1), sheds some doubts on the validity of such strategy.

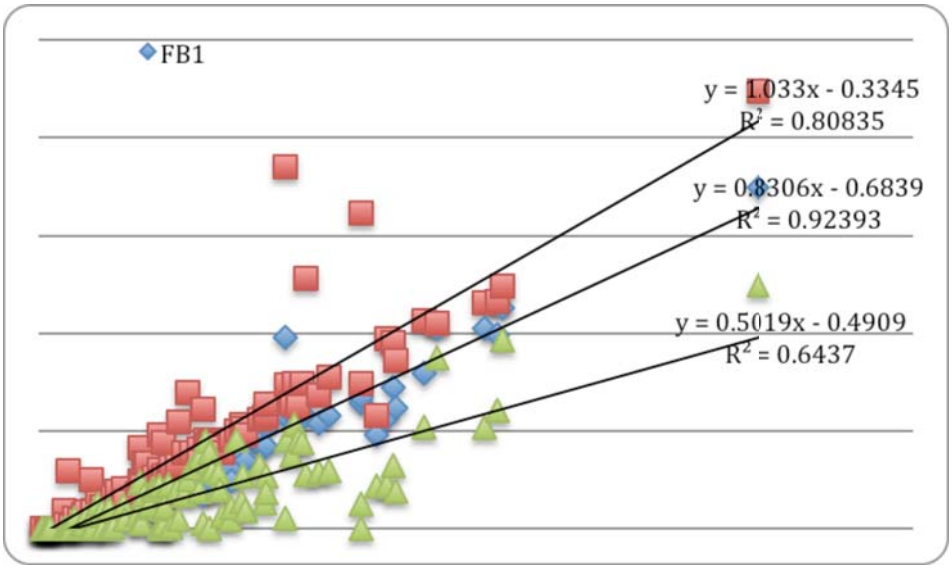


Figure 1. Correlations of energetic penalties of the bioactive conformation (relative to the local minimum) obtained with different restraining schemes. The values obtained with harmonic restraints derived from re-scaled B-factors are used as reference, and compared with those obtained with Flat-bottom restraints (red squares: ½ width = 0.5Å; blue diamonds: ½ width = 0.3Å) or frozen dihedral angles (green triangles).

Different methods provide equivalent geometrical sampling

In order to ensure a fair comparison between methods, it is necessary tocheck that the minimizations proceeded similarly. This is particularly important because they have been carried out with different programs and minimization algorithms. It is also conceivable that some energy functionals might provide a more roughed energy surface, which might induce premature termination of the optimization protocol at shallow energy wells near the

starting configuration. As shown in Table 2, the average unrestrained minimization leads to structures that are approximately 0.85Å away from the initial structure (for the diverse set of ConfGen conformers) or 0.75Å when starting from the crystallographic conformation. Only the AMBER-optimized structures display somewhat larger RMSD values (0.95Å and 1.1Å, for X-ray and ConfGen structures, respectively). The likely cause of this different behaviour is the optimization protocol used in this case, which combines minimization with short MD simulations to ensure that the energy convergence criterion is met (see Methods).

In all cases, the conformational variability decreases after minimization, with approximately 40% of the ConfGen structures converging into pre-sampled minima (i.e. only 60% of the optimized geometries differ from all the rest by more than 1.0Å). This suggests that most minima are identified and sampling is reasonably complete. A similar indication comes from the fact that the local minimum (i.e. the result of optimizing the X-ray structure) is also found when starting from the diverse set of structures generated with ConfGen in 50% of the cases. Perhaps more importantly, on average 4.3 structures are found within one kcal/mol of the predicted global minimum, indicating that the probability of finding the global minimum or an energetically similar structure is very high with our approach.

Although the sampling protocol does not provide an exhaustive exploration of the conformational space, the preceeding analysis indicates that: 1) sampling is similar for all methods, ensuring a fair comparison between them; and 2) the lowest energy conformation corresponds to the global minimum or, at the very least, is energetically very close to it,

enabling us to obtain a good estimate of the relative energy of the bioactive conformation.

The energy of the bioactive conformation strongly depends on the level of theory.

The energy difference between the bioactive conformation and the global minimum is presented in Table 3 and Figure 2. For all force-field based simulations, a large proportion of the molecules are predicted to have unrealistically highstrains on their bioactive conformations. Inclusion of the solvation effect ameliorate the situation significantly, but not sufficiently: whereas in vacuo approximately 50% of the bioactive conformations havea penalty of 8 kcal/mol and the value averages at 11kcal/mol for all conformations; with GBSA these values drop to 4 and 5 kcal/mol, respectively (Table 3). It should be noted that a strain of 4 kcal/mol would result in nearly 1000-fold loss in potency, and is not reasonable to expected that half of the ligands sustain such energetic penalties. These results are, nevertheless, in line with previous reports. In fact, the average strain energies obtained with MMFF and OPLS (5.1 and 4.8 kcal/mol, respectively) are in close agreement with those reported by Perola and Charifson (4.0 and 5.1), for a larger set of 150 ligands.(Perola and Charifson, 2004a) The somewhat smaller penalties for MMFF may be due to the softer retraints used by these authors in the refinement of the bioactive conformation. We suspect that this is not observed in OPLS because the harder constraints may be compensated by the use of an improved version of OPLS (2005 vs 1996). Judging from the strain energies in vacuo (Table 3), the performance of the force fields

increases in this order: GAFF < MMFF < OPLS2005 < OLPS2.0. Inclusion of the GBSA term mantains the ranking, except for the two versions of OPLS, which now perform on par. A possible explanation for this behaviour is that the GBSA formalism may need to be adapted to the partial charges of OPLS 2.0, which are derived in a different manner than OPLS2005. Although the relative performance of the different force fieldsappears reasonable if one considers theirrespective history (size and composition of the training set, version number, specific objective for which they were developed, etcetera) the large effect of the solvation term precludes any conclusive statement about the relative merits of the force fields investigated. Indeed, with the exception of MMFF, all the force fields investigated where specifically developed to carry out simulations with explicit solvation and, in spite of their popularity, GBSA models fail the reproduce the conformational ensembles obtained with explicit water molecules.(Chen and Brooks, 2008a, Zhou, 2003)The large difference between solvated and vacuum calculations clearly indicates that bioactive conformations resemble more the water-solvated state, where the polar solvent area is exposed, than the gas-phase state, where polar groups tend to collapse. This is in agreement with previous findings and emphasizes the importance of polar interactions for protein-ligand interactions.(Schmidtke *et al*., 2010) However, it poses a major problem for accurate determination of the energy of bioactive conformations, particularly if one considers that, unless a very careful and force field specific parameterization is carried out, GBSA models can over stabilize polar interactions.(Chen and Brooks, 2008b)

Table 2. Geometric parameters of the optimized structures.

	RMSD (Xray-bio)	RMSD (Xray-local)	RMSD (Xray-global)	RMSD (Initial-final)	% Distinct Minima[1]
MMFF+GB	0.17 ± 0.06	0.77 ± 0.38	1.74 ± 0.93	0.82 ± 0.32	58.5 ± 22.1
GAFF+GB	0.21 ± 0.12	0.94 ± 0.56	1.87 ± 1.01	1.13 ± 0.42	61.5 ± 20.2
OPLS+GB	0.16 ± 0.06	0.71 ± 0.36	1.74 ± 1.00	0.86 ± 0.36	60.0 ± 21.8
OPLS2.0+GB	0.16 ± 0.06	0.74 ± 0.38	1.81 ± 0.89	0.90 ± 0.38	60.3 ± 21.9
DFT+PCM/MST	0.23 ± 0.10	0.72 ± 0.41	1.61 ± 0.91	0.85 ± 0.32	58.7 ± 21.7

[1]RMSD cut-off 1.0Å

The solvation model is, however, only one possible source of errors. The development of accurate and complete force fields for drug-like compounds is seriously hampered by the exhuberant chemical diversity that they can display. The relative responsibility of those two terms in the failure to quatify the energy of bioactive conformations will be discussed later on. In any case, it is obvious that more rigorous calculations are needed to obtain correct estimates. Considering the size and flexibility of the molecules as well as the importance of the solvation effect, we devised a procedure that, albeit computationally very costly, should be able to confidently identify the global minimum and accurately determine the relative energy of the bioactive conformation. As graphically shown in Figure 3 and explained in detail in the Methods section, it consists in optimizing a diverse set of conformations at a quantum mechanical level (B3LYP/6-31D(d)), taking into account the solvation effect with the polarizable continuum method (PCM). Note that non-standard parameters and minor approximations had to be used (see Methods). Each one of the final geometries is then used to calculate the MST free energy of water solvation and the internal energy at the RI-MP2/aug-cc-pVDZ. With unsigned errors of 6.8%, this is the best method to calculateconformational energies.(Riley *et al*., 2007) By comparison, B3LYP/6-31G* produces errors in the region of ~17%.(Riley *et al*., 2007) Addition of the MST solvation free energies and MP2/aug-cc-pVDZinternal energies enables us to obtain the best possible estimate of the relative energies of bioactive conformations in aqueous media.The results obtained at both QM levels (DFT & MP2) are strickingly different from the MM results: the average penalty is approximately 2.5 kcal/mol and only 14% (DFT) or 24% (MP2) of the bioactive conformations exceeds 4 kcal/mol (the respective MM values are, at best, 4.7 kcal/mol and 50%). Considering the

large differences between methods, a critical assessment of their relative strengths and limitations is needed before proceeding to extract any definite conclusion.

Table 3. Energy distribution of the bioactive conformations, calculated with different molecular mechanics force-fields.

	GAFF		MMFF		OPLS2005		OPLS2.0	
	Vac.	GBSA	Vac.	GBSA	Vac.	GBSA	Vac.	GBSA
% > 0	96.7	98.9	100.0	100.0	100.0	100.0	100.0	100.0
% > 1	93.3	91.1	92.4	89.1	92.4	90.2	92.4	88.0
% > 2	88.9	81.1	89.1	83.7	87.0	78.3	88.0	79.3
% > 3	84.4	72.2	82.6	73.9	81.5	67.4	81.5	63.0
% > 4	81.1	61.1	71.7	56.5	72.8	51.1	73.9	50.0
% > 5	77.8	53.3	66.3	45.7	66.3	38.0	65.2	39.1
% > 6	68.9	44.4	63.0	33.7	60.9	27.2	55.4	26.1
% > 7	61.1	35.6	58.7	25.0	55.4	22.8	51.1	17.4
% > 8	57.8	32.2	53.3	16.3	51.1	14.1	44.6	14.1
% > 9	51.1	21.1	47.8	13.0	43.5	13.0	41.3	10.9
% > 10	47.8	13.3	44.6	6.5	41.3	8.7	34.8	9.8
Avg.	13.0	6.1	11.4	5.1	11.1	4.8	10.2	4.7
St Dev	15.3	4.4	12.0	3.6	11.7	3.5	9.7	3.5

[1]Optimization with solvent effect included, no extra espheres. [2]Single-point calculation with extra spheres. [3]Internal energy of the solvent-optimized geometries. [4]Addition of the MST term to the RI-MP2 internal energy.

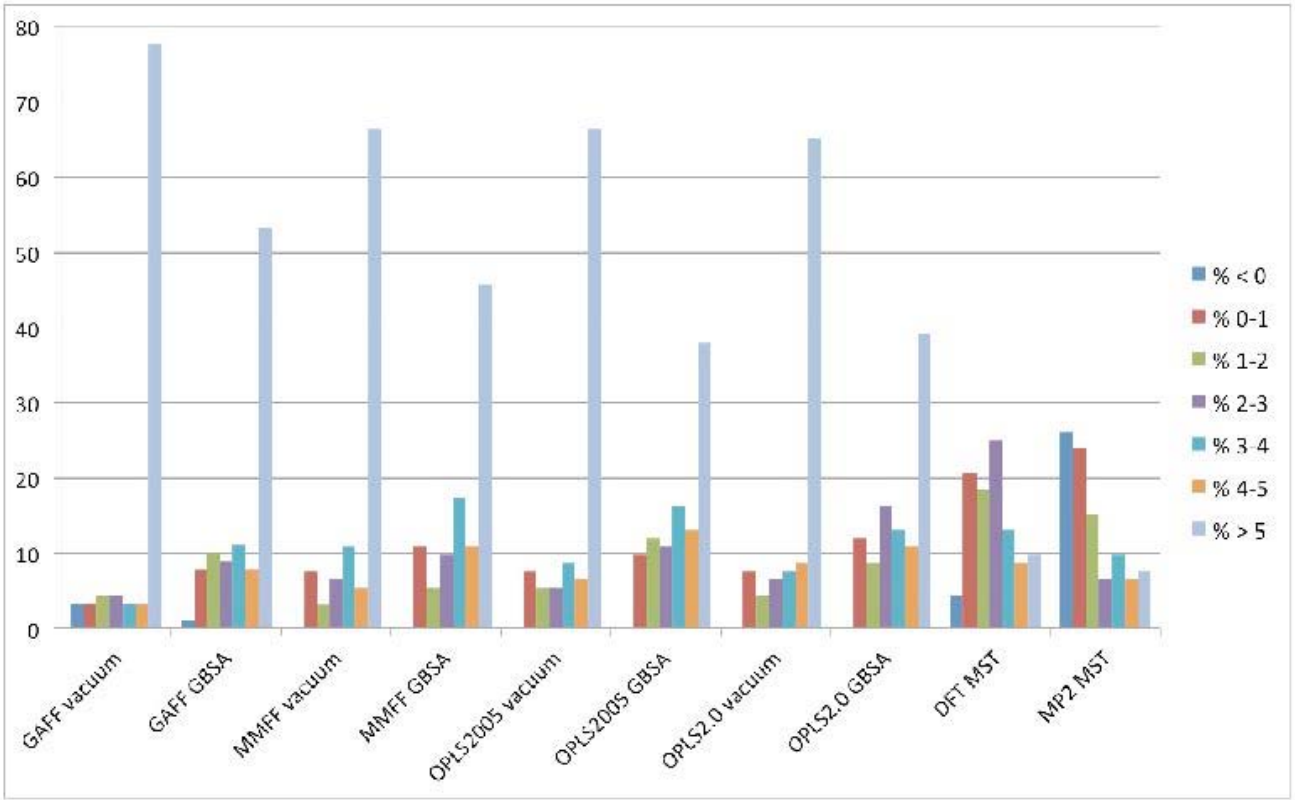


Figure 2. Distribution of energies of the bioactive conformation obtained at various levels of theory.

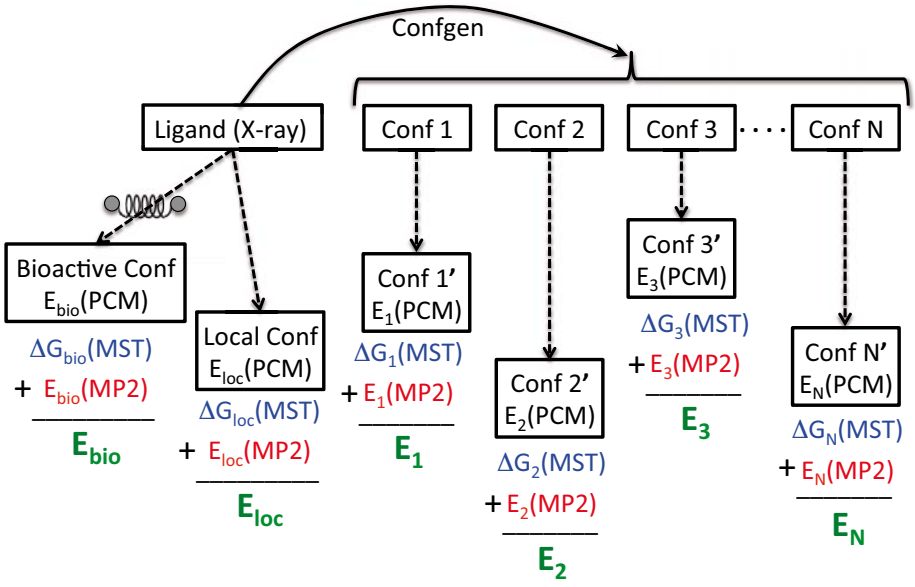


Figure 3. Graphical representation of the calculation procedure

Assessment of QM results andconsistency across methods.

Given that the molecules in the test set are charged and flexible, and as already noted from the MM results, it was clear that optimization had to be performed with a method able to simulate the response of aqueous media. PCM methods offer such a possibility, but they can only be used with those levels of theory on which they have been parameterized. At the same time, considering the number of geometry optimizations that had to be carried out for this test set (~1700), the DFT method B3LYP/6-31G* was the highest level of theory that could be afforded, and it provides a very good balance between speed and accuracy.(Riley *et al*., 2007) Single point calculations in vacuo of the final geometries show that a significant proportion of DFT/PCM

global minima have, in fact, high internal energies (Table 4), providing further evidence that inclusion of the solvations effect during minimization is necessary.

In order to increase the quality of the results, we calculated the energy of the global DFT minimum at an optimal level of theory (MP2/aug-cc-pVDZ).(Riley *et al*., 2007) As shown on Table 4, the number of high-energy bioactive conformations decreasesat this level(e.g. only 29% of molecules have energies above 2kcal/mol, vs. 50% for DFT). However, this is largely due to the fact that 25% of compounds appear to have negative energies for the bioactive conformation (i.e. the DFT global minimum turns out to be more penalized than the bioactive conformation). This is a clear warning that the energy landscape can be very different for the DFT and MP2 methods. Geometric parameters are very well predicted at the B3LYP/6-31G* level,(Riley *et al*., 2007) and the position of the minima is not expected to change at the higher level of theory. The relative height of the minima, on the other hand, can change significantly and it it becomes necessary to calculate the energy of every conformerat the MP2 level. The grayed-out part of Table 4 shows the corresponding results. The global MP2/MST energy minimum is oflower energy than the bioactive conformation in 92% of the cases, while for the remaining 8% we must assume that the MP2 global minimum has not been found because it does not correspond to a low DFT minimum. It is striking that the proportion of high-energy bioactive conformations is higher for MP2 than DFT: 37% and 16% of conformers have predicted penalties above 3 and 5 kcal/mol respectivel, whereas the corresponding DFT values are 25% and 8%. Since part of this effect may be due to lack of correspondence between the energy landscapes at both levels of theory, we

have selected a subset of 41 compounds showing a good correlation between the relatives energies obtained with DFT and MP2 (r^2> 0.8 and a minimum of 6 conformers). For this set, the DFT results should be of similar quality to the high-level calculations, and we find that the proportion of strained bioactiove conformations is much decreased, with only 5% (i.e. 2 compounds) displaying strain energies above 4 kcal/mol (Table 5). This corresponds to the 1UVS and 1FJS ligands.The first one was already identified as having an unrealistically strain in a previous publication, (Butler *et al*., 2009a) due to an incorrect model of an amide moiety in the PDB structure. The source of the high strain in 1FJS can also be traced to an incorrect tridimensional structure, not supported by the actual electron density (Figure 4). Thus we can conclude that, when calculated with sufficiently accurate levels of theory, ligands adopt conformations with strain energies below 4 kcal/mol. Higher values are strong indication of an incorrect representation of the ligand in the crystallographic structure.

Table 4. Energy distribution of the bioactive conformations, calculated at different QM levels.

	DFT-PCM Global Min.					MP2-MST Global Min.[1]			
	DFT PCM	DFT MST	DFT Vac.	MP2 Vac.	MP2 MST	MP2 MST	MP2 Vac.	DFT Vac.	DFT MST[1]
% > 0	100	96.7	95.7	83.7	75.0	92.4	84.8	87.0	95.7
% > 1	76.1	73.9	76.1	68.5	44.6	68.5	72.8	67.4	63.0
% > 2	50.0	55.4	66.3	46.7	29.3	51.1	56.5	55.4	44.6
% > 3	25.0	30.4	57.6	40.2	22.8	37.0	46.7	43.5	29.3
% > 4	13.0	17.4	46.7	34.8	13.0	22.8	38.0	32.6	20.7
% > 5	7.6	8.7	33.7	26.1	7.6	16.3	29.3	22.8	13.0
% > 6	4.3	7.6	29.3	22.8	5.4	10.9	25.0	20.7	12.0
% > 7	3.3	5.4	21.7	18.5	4.3	7.6	15.2	15.2	6.5
% > 8	2.2	4.3	16.3	15.2	4.3	7.6	14.1	12.0	5.4
% > 9	2.2	2.2	13.0	14.1	3.3	5.4	14.1	12.0	2.2
% > 10	0.0	1.1	12.0	13.0	1.1	2.2	13.0	10.9	2.2
Avg.	2.4	2.5	4.9	3.8	1.3	2.6	4.5	4.2	2.4
St Dev	1.9	2.2	5.7	5.8	2.8	2.8	6.9	7.0	2.6

[1]Re-evaluation of all DFT/PCM-optimized geometries at the MP2-MST level.

Table 5. Energy distribution of the bioactive conformations, calculated at different QM levels for the subset of compounds displaying good correlation between the methods used in optimizations and the methods used to re-evaluate the energy (N=41).

	DFT-PCM Global Min.				
	DFT PCM	DFT MST	DFT Vac.	MP2 Vac.	MP2 MST
% > 0	100	92.7	92.7	92.7	75.6
% > 1	75.6	73.2	80.5	78.0	43.9
% > 2	43.9	48.8	65.9	51.2	26.8
% > 3	22.0	24.4	58.5	46.3	17.1
% > 4	4.9	14.6	56.1	43.9	12.2
% > 5	4.9	7.3	43.9	34.1	9.8
% > 6	4.9	7.3	41.5	34.1	7.3
% > 7	4.9	7.3	34.1	29.3	4.9
% > 8	4.9	4.9	26.8	24.4	4.9
% > 9	4.9	4.9	24.4	24.4	4.9
% > 10	0.0	2.4	22.0	24.4	2.4
Avg.	2.3	2.4	6.3	5.5	1.5
St Dev	2.0	2.4	7.4	7.2	2.7

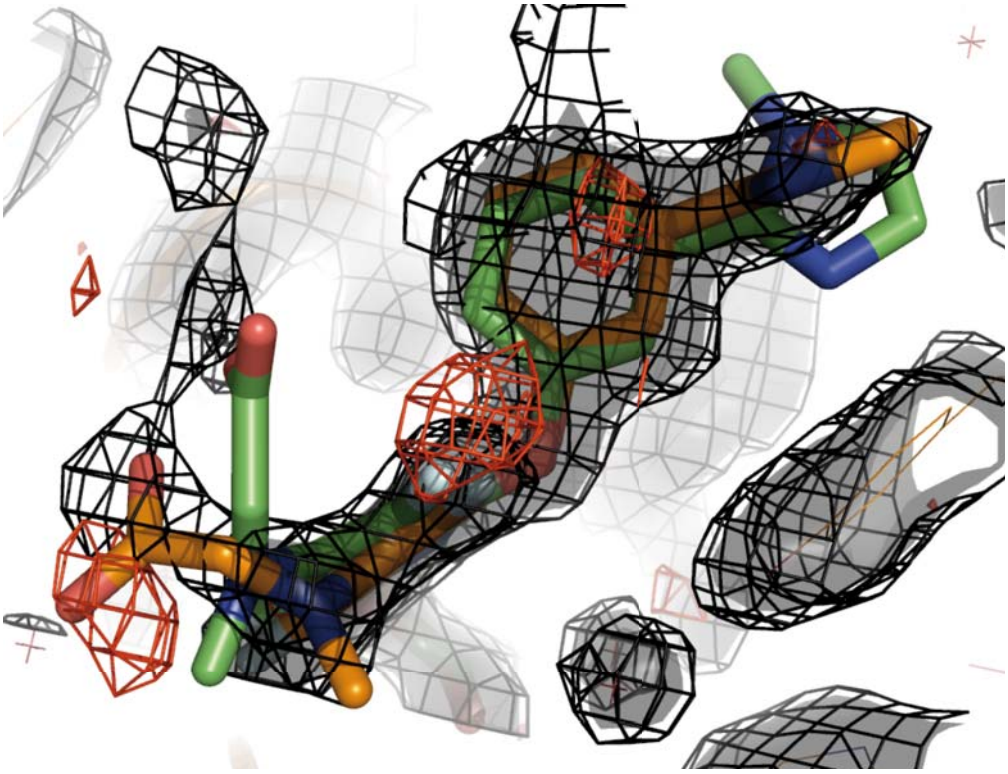


Figure 4. 1FJS ligand in the crystallographic (orange) and local minimum (green) geometries. Electron density of the 2Fo-Fc map is shown at 1 sigma (black mesh) and 1.5 sigma (transparent surface). The difference map Fo-Fc is shown at 2.5 sigma (red mesh). The conformational change occurs in an area where the electron density is ill defined.

Note: Conclusion of this chapter has been included separately -after the results section- in the sub-section 5.1 of conclusion chapter-5 of this thesis .

4.2 Allosteric Regulation of HDAC8

Background

It is well established fact that the chromatin structure is one of the key determinants of gene expression or regulation in the cell cycle; therefore impairment of the chromatin structure might lead to the abnormal cell proliferation (Grunstein, 1997). Histone proteins that are surrounded and compacted by the DNA form the building blocks of the chromatin structures. Hence, maintenance of the chromatin structures substantially depends on the histone subunits which are also subjected to various post-translational modifications, from acetylation to methylation, phosphorylation, ubiquitylation and sumoylation (Peterson and Laniel, 2004).

Histone deacetylases (HDACs) are the class of enzymes that act as the post-translational modifier of histone proteins. These proteins catalyze the deacetylation reaction by removing the acetyl groups from ε-N-acetyl lysine on histone tails which then modulate the transcription activity and consequently the cell cycle regulation (Luger *et al.*, 1997). It has been shown that over-expression of HDAC enzymes leads to gene silencing and hence proved to be critical in cell-cycle regulation, cell proliferation, differentiation, and in the development of human cancers (Sengupta and Seto, 2004, Jacobson and Pillus, 1999, Kouzarides, 1999). Role in neurodegenerative diseases, such as Parkinson's and Alzheimer's disease has also been evidenced recently (Choudhary *et al.*, 2009, Johnstone, 2002). Due to the fact that HDACs play roles in the key events in the onset and progression of cancer and neurodegenerative diseases, they are potential drug targets for various treatments. In fact two HDAC inhibitors, suberoylanilide hydroxamic acid (SAHA) and valproic acid have been approved for the treatment of cancer and neurodegenerative diseases. A search for HDAC

isoform-specific inhibitors is now on. Mammalian HDACs have been classified structurally into three classes: Class-I (HDAC 1, 2, 3 & 8), Class II (HDAC 4, 5, 6, 7, 9 & 10) and Class III (Sirt1 – Sirt7) (Yang and Seto, 2008, Haberland *et al.*, 2009). HDAC8 belongs to class-1 histone deacetylases and has been attributed with the T-cell-derived tumor cell proliferation (Balasubramanian *et al.*, 2008) and neuroblastoma tumorigenesis (Oehme *et al.*, 2009). In addition to these chemo-therapeutic findings, the distinct behaviour that HDAC8 shows towards the post-translational modification has made it an important and appealing target to study. Studies by Lee H et al revealed that deacetylase activity of HDAC8 is regulated by posttranslational modifications and that the phosphorylation at the non-catalytic, non-conserved N-terminal residue Ser39 by PKA leads to the decreased deacetylase activity (Lee *et al.*, 2004). Remarkably, this phenomenon is not observed in other HDACs (Sengupta and Seto, 2004).

The 3D structure described by Vannini (Vannini *et al.*, 2004) and Somoza et al (Somoza *et al.*, 2004) reveals that the phosphorylation site of HDAC8, Ser39, is a non-catalytic site residue and located on α-helix at the N-terminal site that forms the solvent exposed surface of the Zn^{+2}dependant active site. Ser39 is far from the catalytic site cavity and facing opposite to it (Fig 4.2.1). Considering the structural organization of Ser39 we presume that deacetylase activity of HDAC8 might be regulated allosterically upon its phosphorylation.

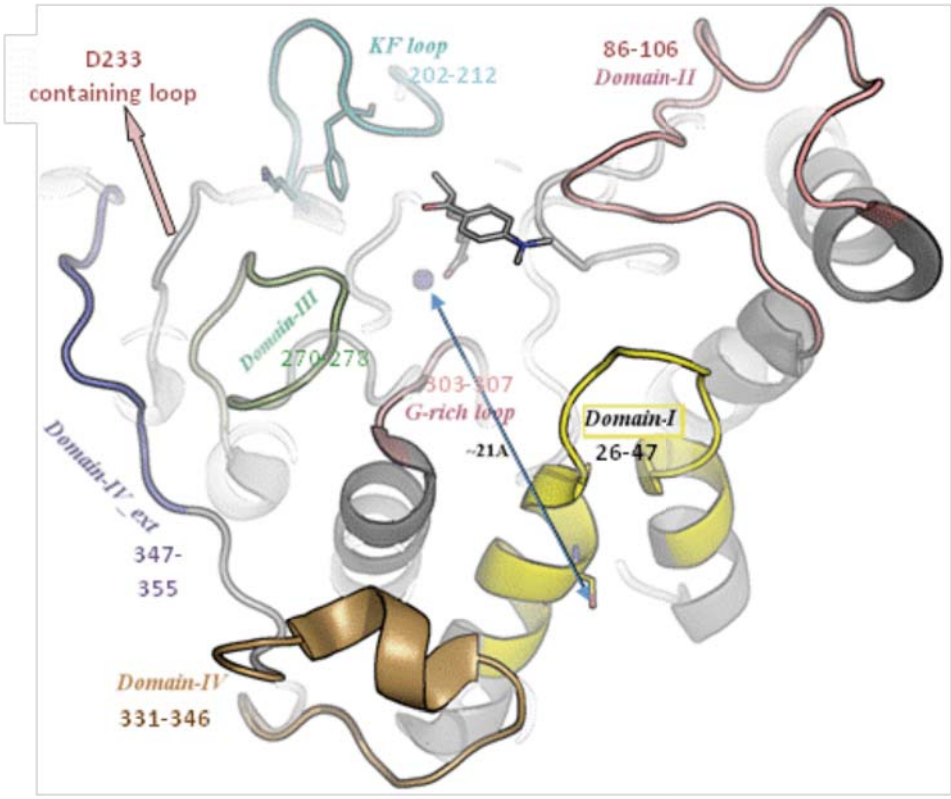


Figure 4.2.1. Crystal structure (1T64) showing the spatial distance between the Zn^{+2} (in sphere) binding-catalytic site and the Ser39 (in stick). 3D structure of the HDAC8 is subdivided into different domains. Each domain given an identifying name in addition to the residue range.

A more detail analysis around the phosphorylation site reveals that Ser39 is surrounded by the hydrophobic residues Val25, Phe336 and the hydroxyl group of the Ser39 is flanked by the negatively charged Glu335 and Asp29. A stacking stabilization between the long hydrophobic side chain of the Glu335 and pi-ring of the Phe336 has been also seen in the crystallographic structure of HDAC8 (1T64). Therefore, it was postulated that introduction of the negatively charged phosphate group in this

environment could lead to a major structural disruption (Fig. 4.2.2) (Somoza *et al.*, 2004). It has also been shown by Lee et al that mutation of Ser39 to Ala enhances the deacetylase activity of HDAC8 in contrast to the mutation of Ser39 to Glu, which mimics phosphorylation and further decreases HDAC8's enzymatic activity (Lee *et al.*, 2004). Recent studies reveal that interactions of a conserved Arg37 residue (close to Ser39) with its spatially neighbouring residues might be responsible for the release of the by-product acetate deacetylase reaction through a 14Å solvent facing channel (Wang *et al.*, 2005, Haider *et al.*, 2011). This provides some support for the hypothesis that a direct mechanism (i.e. structural disruption due to electrostatic repulsion of the phosphate with neighbour residues) may cause the loss of enzymatic activity. However, a more indirect allosteric mechanism involving a global conformational adaptation cannot be ruled out.

Sequence alignment of the residues of the exit tunnel shows the highly conservative sequence profile of the Zn-binding catalytic site in class I and II HDACs in humans (Somoza *et al.*, 2004, Vannini *et al.*, 2004) and sequence conservation analysis on only class-I HDACs shows that not only the catalytic core but also the overall deacetylase domain is very conserved in nature (Fig 4.2.2). With the background of this conserved nature of the deacetylase domain in different HDACs it would be undoubtedly interesting to understand how phosphorylation makes HDAC8 to behave like the “odd man out”among the class I HDACs. To date, as per our knowledge, no other study from this perspective has been addressed. Additionally, knowledge about the mechanism involved in inhibition by phosphorylation may provide new venues to regulate the activity of HDAC8 with drug-like compounds through allosteric modulation.

Allosteric regulation is a well established fact in biochemistry where an enzymatic activity of a protein is regulated upon binding of an effector molecule at the protein's allosteric site (a site other than the protein's catalytic site). Allosteric modulation is a very common molecular mechanism of action of drugs, e.g. benzodiazepine binding to GABA-A receptor and regulation of GPCRs (Jensen and Spalding, 2004), which may offer a distinct pharmacological profile compared to direct competition with the substrate.

To understand at the molecular level how phosphorylation acts as a regulator of HDAC8 activity, we studied molecular dynamics of the same with and without phosphorylation on Ser39. Contrary to the initial expectations, introduction of the negatively charges phosphate does not result in a major structural rearrangement on the local environment. However, we observed that a loop distantly placed from the phospho-acceptor Ser39 in the 3D space of HDAC8 opens and de-associates itself from the deacetylase core in Wt within the simulation time span of 80ns, but this change does not occur when the protein is phosphorylated. The loop is held in place by the interactions between two conserved residues (Fig 4.2.2, residues with the black stars), a pi-ring of Phe207 and the hydrophobic side chain of Lys202 (hence we call it KF-loop) (Fig 4.2.1) and remains closed until 500ns. From this perspective we postulated that the decreased deacetylase activity obtained in the phosphorylated form of HDAC8 could be due to the lack of structural flexibility in this loop-region, which is involved in substrate recognition and binding (Vannini *et al.*, 2004). Additionally, the pathway involved in the allosteric transmission of the structural perturbation from the phosphorylation site to the KF-loop was investigated using the statistical Kullback-Leibler divergence methodology. To further test the relationship

between enzymatic activity and mobility of the KF-loop, as well as to validate the putative allosteric pathway, several point mutations were designed and investigated both computationally and experimentally.

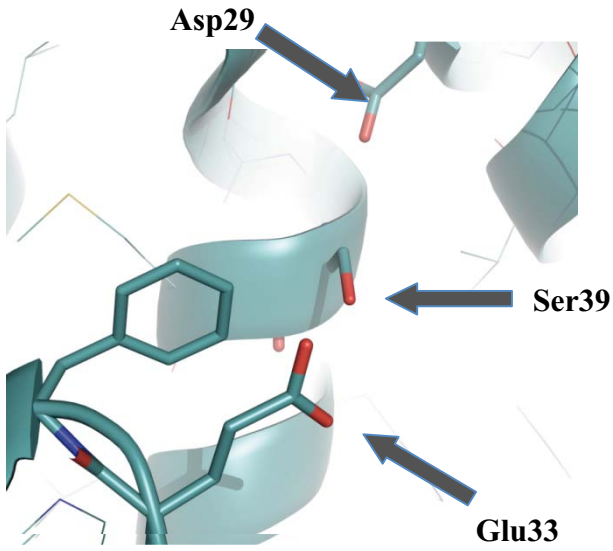


Figure 4.2.2: The two negatively charged flanking residues of Ser39 are shown in sticks. The long hydrophobic side chain of Glu335 is seen being stabilized with the pi ring of the Phe336.

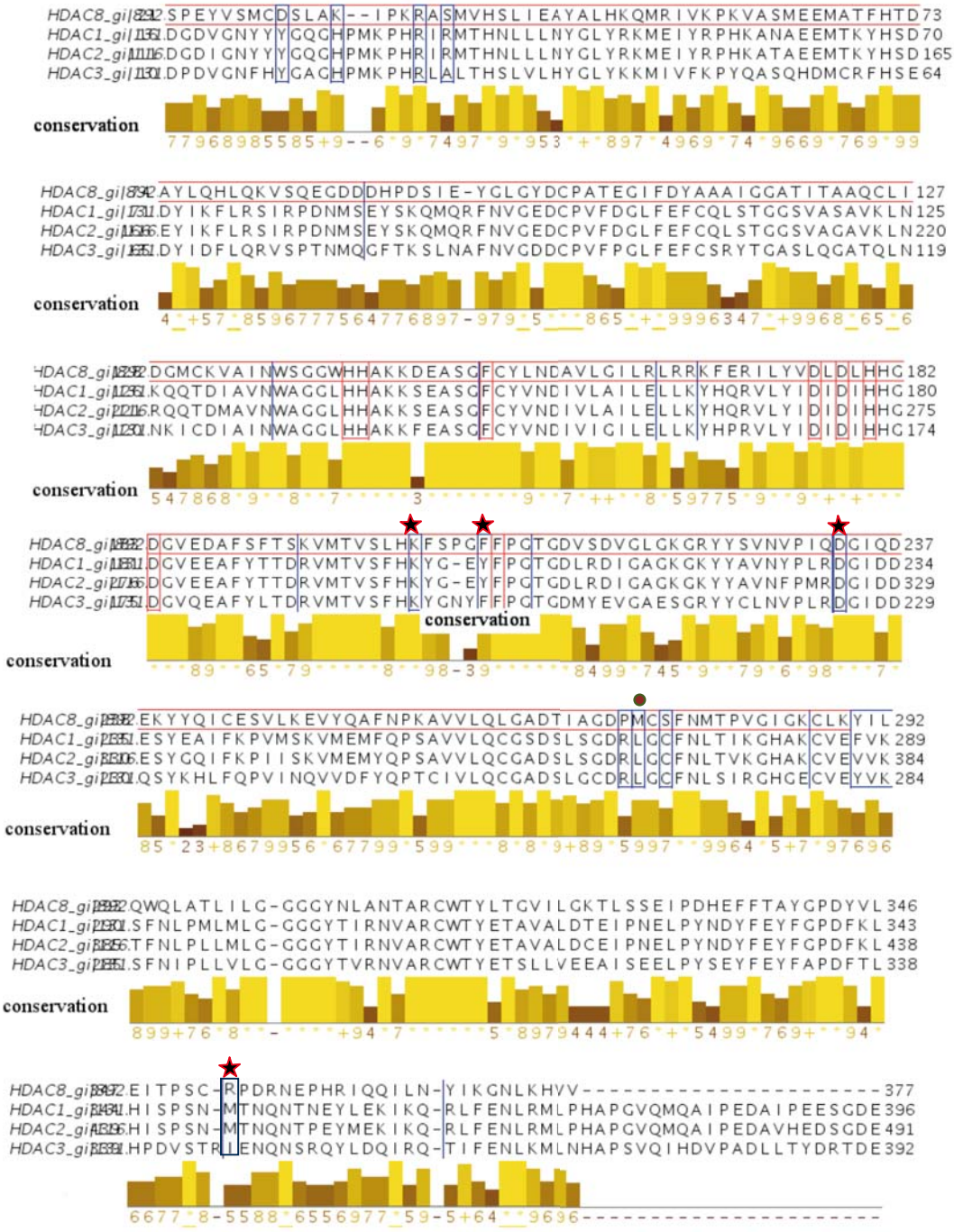


Fig. 4.2.3: Multiple sequence alignment of all class-I HDACs. Alignment shows the conserved sequence profile in all human class-I HDACs. Residues in the red boxes are involved in the catalytic activity. Residues marked with stars are those accountable for holding the KF-loop.

Simulation set-up

MD simulations were carried out on a wild-type apo-form of HDAC8, its computationally generated phosphorylated form (pHDAC8), and on different single-point mutants, both in the phosphorylated and non-phosphorylated forms. The mutations were chosen based on primary analysis of Wt and pHDAC8 simulations. The initial structure was taken from PDB (pdb id: 1T64; resolution 1.9Å). The crystallographic structure contains two Trichostatin-A molecules that were removed to get the apo-structure. To date no apo-structure of HDAC8 has been elucidated. The only HDAC structure that has been solved in apo-form is the class-II HDAC4 catalytic domain (2VQW) (Bottomley *et al.*, 2008). Structural alignment between HDAC8 (with bound ligand) and HDAC4 (ligand free) shows only ~0.6 Å root mean square deviation with respect to the whole structure. The Wt structure was then phosphorylated at Ser39 by introducing a phosphate group with the help of leap programs of Amber molecular modeling package (Pearlman *et al.*, 1995). The parameter for the phosphoserine was taken from the Amber Parameter Database (http://www.pharmacy.manchester.ac.uk/bryce/amber).

Deacetylase activity of HDAC8 is not only Zn^{+2} dependent (Vannini *et al.*, 2004), also two monovalent K+ ions observed in X-ray crystallographic structures have been proved to be essential for its activity (Gantt *et al.*, 2010, Wu *et al.*, 2010). Therefore, our simulations include them. Zn parameters were taken from the Stote & Karplus model. (Stote and Karplus, 1995) The Zn coordination in the active site was maintained by applying a soft harmonic constraint between the protein atom His168 and a Zn^{+2} ion. Protonation state of the His side chains in the active site were assigned according to the previously reported studies (Wu *et al.*, 2010,

Corminboeuf *et al.*, 2006). The systems were solvated with TIP3P water molecules (Jorgensen *et al.*, 1983) and neutralized with K+ and Cl- ions in a truncated octahedral box spanning 13 Å further from the protein. After a short equilibration in constant pressure and temperature (NPT) the production simulations were carried out in constant volume and temperature (NVT) with periodic boundary conditions and particle-mesh Ewald methods (grid spacing of 1Å) to treat the long-range electrostatic interactions (Essmann *et al.*, 1995). The simulations were carried out with the Amber 11 package (Case *et al.*, 2010).

Tracking region of differences in two different MD simulations:

A protein is a complex unit composed of different proportion of amino acid units. Therefore the dynamic behavior of proteins is the result of the composite dynamics of each unit. Understanding the behavior of such unit or group of units in the time course would, consequently, facilitate our understanding about the dynamics of the protein as a whole. Therefore, to understand the sequence of events that motivate two different simulations (here, Wt and its phosphorylated form) to behave differently we applied KL-divergence measure on the probability distribution of different properties (observables) of an amino acid residue that were sampled in the conformational space during the course of simulation. We use observables related to the internal coordinates of the residues and protein intermolecular interactions, as they might reflect most on the dynamic behavior of a protein. Residue dihedral or torsion angle is the one that dominates most the conformational changes compared to the localized effect of bond stretching

or angle bending vibrations (Lee *et al.*, 2007, Amadei *et al.*, 1993). Inter-residue interaction energy, on the other hand, captures the energetically favourable or unfavourable non-covalent interactions between a pair of residues. We applied KL-divergence on these two variables to estimate the variability evolved due to the phosphorylation of the HDAC8 enzyme.

Kullback-Leibler measure is a statistical divergence (or KL-distance) estimate that has been used since long in various fields of studies to capture the divergence between two independent distributions with recent successful implementation in the image processing algorithm to track and quantify the regions-of-interest in terms of dissimilarity between two images taken over a time span (Boltz *et al.*, 2009). This well-known and widely applied method has been also used before in the biological systems to understand the similarities between two ensembles of structures (Lindorff-Larsen and Ferkinghoff-Borg, 2009, Ming and Wall, 2005). Here we use this method to evaluate divergence between two independent distributions of a variable taken from the two MD simulations. This is a very simple yet efficient and elegant model to estimate differential behavior in two systems.

For a discrete set of points for a variable the KL-distance is calculated using the equation as below::

$$D_{KL}(p,q) = \sum_i p_i \cdot \ln(p_i / q_i) \qquad (4.2.1)$$

where, for a variable i, p and q are the probability distribution of that variable in reference and target respectively. The quantity is always a non-negative number. It becomes positive in case $p_i \neq q_i$, or zero when the two distributions are equal. However, KL divergence is not a symmetric measure

between two distributions. Given the above equation, it measures the divergence of distribution p from the distribution q.

Therefore,

$$D_{KL}(p,q) \neq D_{KL}(q,p)$$

To get the symmetric form out of it we used the well accepted simple form of the method where,

$$D_{KL}(sym) = (D_{KL}(p,q) + D_{KL}(q,p)) / 2 \qquad (4.2.2)$$

The added advantage of this method is that the distribution does not have to be perfectly normal.

As mentioned before, we used the later form of the distance-metric to quantify the differences and capture the regions-of-interest that might be accountable for the propagation of allosteric effect in the phosphorylated form with respect to the two reaction-coordinates (as stated before, dihedral angle and pair-wise interaction energy). The joint-probability distributions of all possible combinations of dihedral angle (phi, psi, and all chi torsional angles) (with 30° grid spacing) that a particular residue can sample in the conformational space were calculated.

Therefore,

p and q in eq (4.2.1) and (4.2.2) = joint-probability distribution of all torsion angle (phi, psi, and all chi).

The dihedral angles were computed using Gromacs trajectory analysis tool g_chi (Lindahl *et al.*, 2001) on 8000 uniformly spanned trajectories taken from the equilibrated Wt and pHDAC8 simulations. All possible residue-pair ((number-of-residues) X (number-of-residues-1)) interaction energy was computed for 4000 structures, extracted again from the equilibrated part of the simulations, using Generalized Born approach of MM-PB(GB)SA programs of Amber package (Tsui and Case, 2000, Gohlke *et al.*, 2003a).

Conformational Analysis reveals differential Dynamic behavior of a distant loop in Wt and pHDAC8:

We extended the production simulations of Wt HDAC8 and its phosphorylated form (pHDAC8) initially until 200ns and then extended for the pHDAC8 form up to 500ns. We performed k-means clustering on those two simulations to analyze the conformational changes visually. The clustering analysis of the last 50ns of the total 200ns run reveals two distinct conformational states adopted by the Wt and pHDAC8. While all the clusters of the Wt are crowded with an open conformation, the pHDAC8 remains in a closed inhibited-like starting structure (Fig 4.2.4) in addition to the expected local changes observed in the neighbouring domains of the Ser39 in pHDAC8 conformations (Fig 4.2.4). Particularly it was noticed that a major part of long C-terminal loop consisting residues 202-214 (of residues 202-219) dissociates away from the centre of deacetylase core and remains open throughout (Fig 4.2.4). This loop is not only sequentially far from the phosphorylation site (Ser39) but also situated spatially at a distant site, more

precisely, opposite end of a hypothetical diagonal joining it with the phosphor-acceptor Ser39 (Fig 4.2.1). Visual investigation on different conformations around this loop region reveals that the loop was mainly seized by the interactions between the long hydrophobic side chain of Lys202 and the phenyl ring of Phe207 (hence, the loop is KF-loop). In the Wt simulations this interaction was broken which in turn leads to the opening of the loop (Fig 4.2.4). In order to corroborate the significance of this observation, 2 more independent MD replicas (100ns and 400 ns) were carried out and checked for each phosphorylation state and one 100ns replica for the non-phosphorylated form, observing loop opening in 2 of the non-phosphorylated replicas and none of the phosphorylated ones. We hypothesized that the observed opening of the loop in Wt is related to its catalytic activity, and that the absence of loop opening in the phosphorylated form is a regulatory mechanism to inhibit the enzyme. The hypothesis is reinforced by the presence of Phe208 in this loop (neighbour to the above-mentioned Phe207), which has been shown to be important for exerting hydrophobic or pi-stacking stabilization to hold the hydrophobic side chain of the acetylated lysine of the substrate peptide in the only peptide-bound crystal structure of HDAC8 (2V5W) (Vannini *et al.*, 2007). This residue has also been shown to have similar structural effect in all available crystal structures of HDAC8 that contains a bound ligand (Fig 4.2.11). Loop opening exposes the hydrophobic side-chains of Phe207 and Phe208, which may help recruitment of the substrate peptide. Loop flexibility would, according to our hypothesis, be necessary to allow the substrate to get accommodated in the catalytic core and the rigidity of the loop induced by phosphorylation of Ser39 would be one of the mechanisms to decrease catalytic activity.

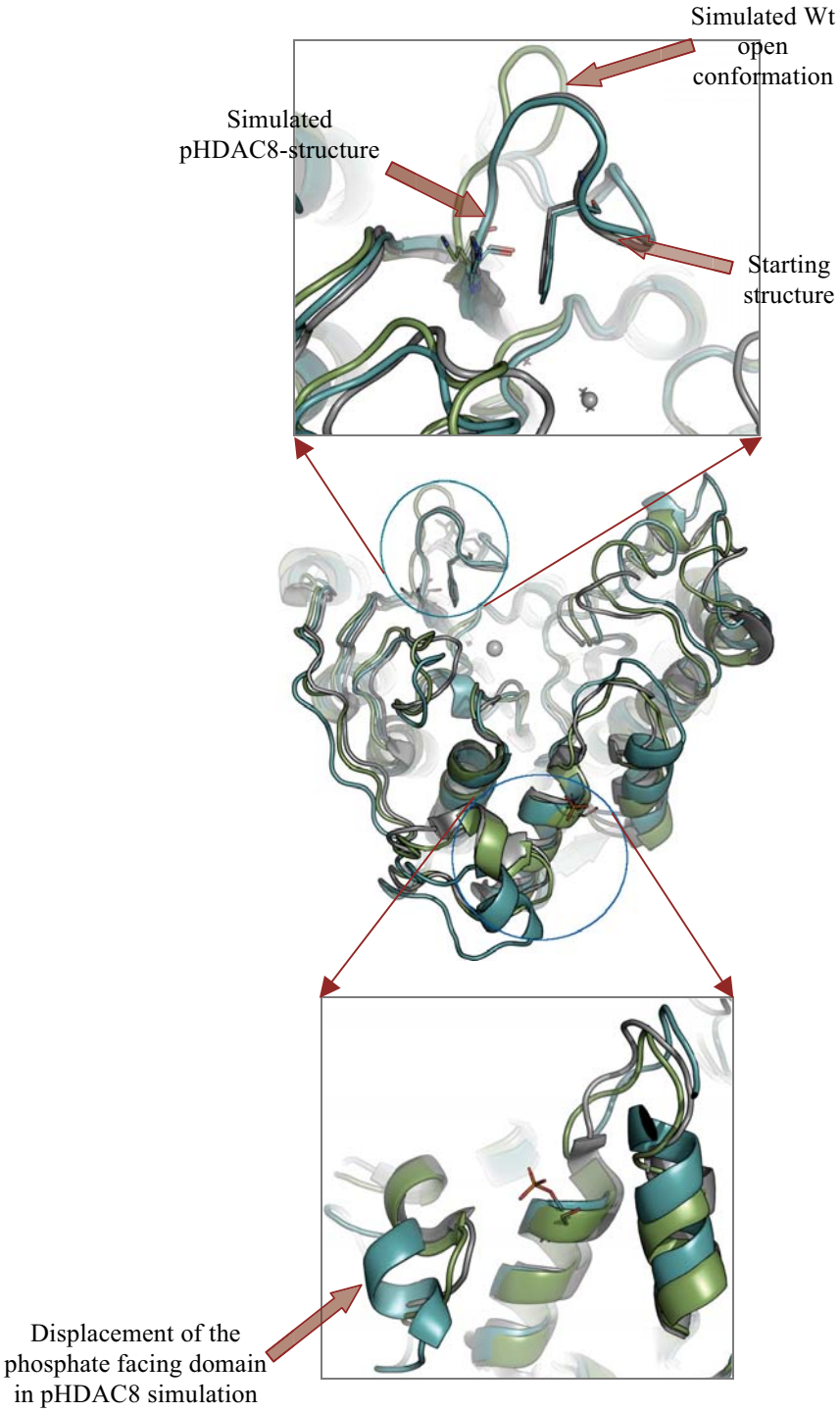

Simulated Wt
open
conformation
Simulated
pHDAC8-structure
Starting
structure
Displacement of the
phosphate facing domain
in pHDAC8 simulation

Figure 4.2.4. Middle Structural superposition of the representative structures from the most populated cluster of Wt (in green cartoon) and phosphorylated form (in cyan) with the initial starting structure (in grey) from PDB. Upper) The long KF-loop in pHDAC8 simulations is seen replicating the initial closed structure in contrast to the Wt. Lower) Displacement of the phosphate group facing domains that contain negatively charged residues Asp29 and Glu325 (Fig 2) in pHDAC8 simulations.

Backbone heavy atom RMSD with respect to the initial structure of the KF-loop consisting residues 202-214 shows sharp increase of the loop RMSD for Wt which coincides with the increased distance between the centre of mass of the Lys202 side chain and the pi-ring of the Phe207 in Wt (Fig 4.2.5A). In the phosphorylated form the same distance and RMSD remain stabilized through the500 ns simulated (Fig 4.2.5A, B).

As shown in Figure 4.2.4A, the aromatic ring of F207 stacks against the hydrophobic portion of the K202 side-chain (aliphatic carbons in positions β to ε), while the charged amino group is partly solvent exposed and also forms internal hydrogen bonds with D233 and S276. Tracking the distance over the course of time-evolution between these two side-chains we see that loop opening occurs right after an elongation of the K202-F207 distance, (Fig 4.2.5A) suggesting that the interaction break plays a causal role in the loop opening. Indeed, the ampiphilic character of the side-chain of Lys means that, while the initial conformation is perfectly stable, a small conformational change leading to exposure of the phenyl ring of F207 to the charged amino group would result in an unfavorable interaction (it is worth noting that, while cation-pi interactions can be intrinsically favorable (Cubero *et al.*, 1998), the desolvation cost of ammonium NH groups by far outweighs the energy gain of a direct NH-π interaction (Bissantz *et al.*, 2010,

Salonen *et al.*, 2009). As noted before, the loop does not open in the phosphorylated form, which can be explained by the increased stability of the network of hydrogen bonds surrounding K202. In fact, the effect originated due to the introduction of the phosphate group appears to have diffused allosterically to this remote part of the protein to strengthen the ionic triad involving K202, D233 and R353 (Table 4.2.1 and Fig. 4.2.6). We calculated the distances for these pairs of residues with the evolution of time using ptraj tool from Amber package and checked the proximity of these residue-pairs by counting the number of occurrences a pair of residues come closer than 3Å. It shows that the pair R353-D233 (Fig. 4.2.6) is at H-bonding distance ~32% times in Wt-closed-states (before the loop opening starts), in contrast to the phosphorylated form where number is ~80% (Table 4.2.1). Cooperativity has previously been described for similar solvent-exposed salt-bridge triads, (Olson *et al.*, 2001, Lebbink *et al.*, 2002) and it seems likely that the increased stability of the interaction between R353 and D233 in the phosphorylated form strengthens the D233-K202 pair. R353 also forms a much higher proportion of hydrogen bonds with the backbone of C275 and S276, further contributing to the stabilization to this part of the protein in the phosphorylated form. By contrast, in the non-phosphorylated form, loss of the R353-D233 interaction allows the latter residue to rotate its side chain freely (Fig 4.2.7), resulting in the increased mobility of–and weaker interaction with– K202, ultimately leading to the rupture of the salt-bridge and subsequent opening of the loop. It is worth noting that residues K202, F207 and D233 are identical or have very conservative mutations amongst human class-I HDACs, however R353 is not conserved in the class, providing further support for its possible involvement in the regulatory

mechanism induced by phosphorylation (Fig 4.2.3; the residues marked with the filled stars).

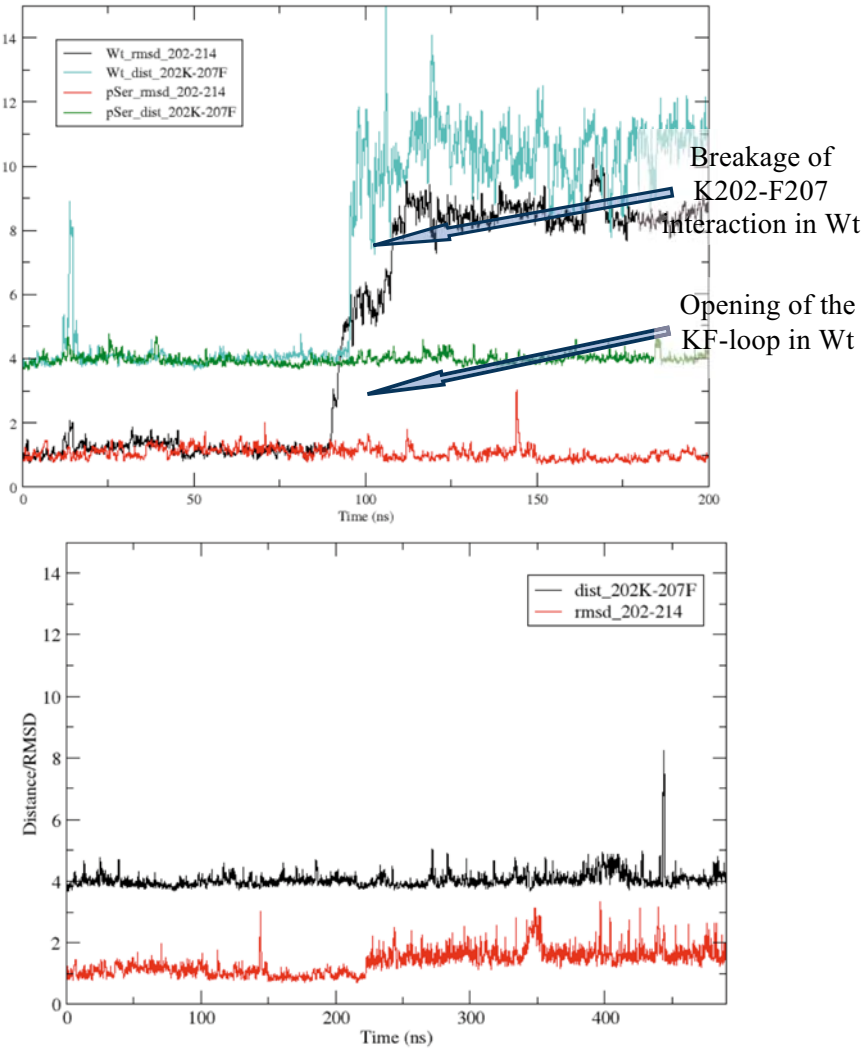


Fig 4.2.5A: Evolution of RMSD of the KF-loop and the distance between the side chain of Lys202 and pi-ring of Phe207 in Wt (black and cyan respectively) and in its phosphorylated form (red and green respectively) with respect to the initial structure until 200ns. B) Plot of the same in the extended simulation (500ns) in pHDAC8 shows the distant loop maintaining the closed structure.

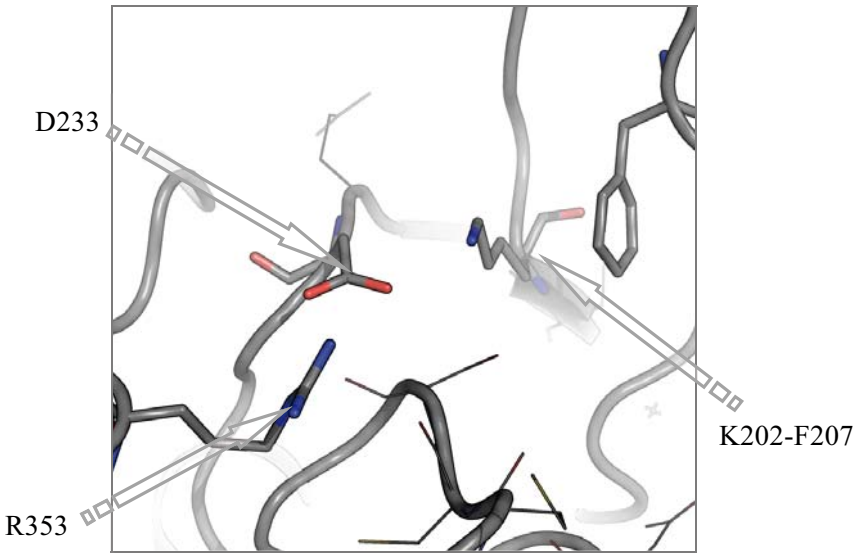
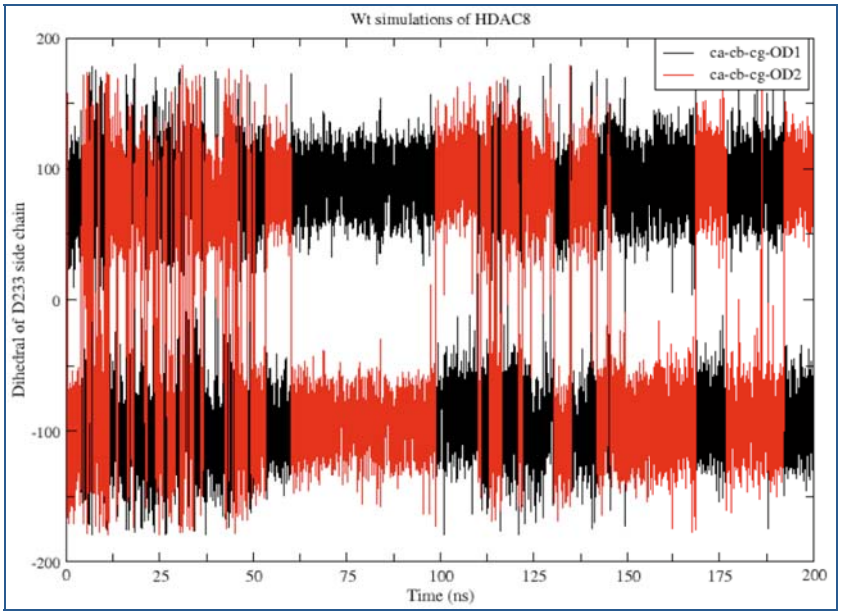


Fig 4.2.6: The interaction networks around the Lys202-Phe207 involving residues Asp233, Lys202, and Arg353.

Table 4.2.1: Hydrogen-bond frequency for selected pairs of atoms.

Interacting pair	Frequency at which a pair comes closer than 3Å	
	Wt (closed state)	pHDAC8
K202Nζ-D233Oδ	81%	78%
K202Nζ-S276Oγ	66%	74%
R353 Nη-S276O (R353Nη-C275O)	17% (37%)	55% (67%)
D233Oδ-R353Nη	32%	80%



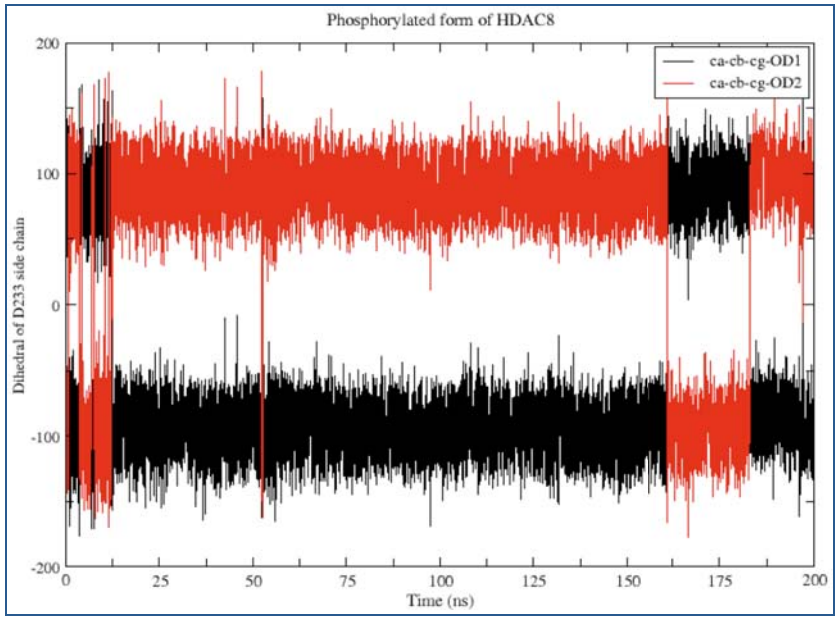


Fig 4.2.7: Side chain dihedral angle profile of D221 in the simulations of Wt (left) and phosphorylated form (right) of HDAC8.

Exploring the Allosteric path in HDAC8:

The results above were provocative enough to prompt us to question how the effect of the local perturbation due to the introduction of the phosphate group induces the global effect, and track the possible path through which the effect diffused or transmitted to the distant loop, which might be worth knowing for further modulation of its functionality. As visual inspection of different conformations of a complex system like protein is not only tedious but also susceptible of error and lacks the quantitative measure of similarity or differences in two different systems, we have used the statistical method, KL-divergence metric (eq. 4.2.1), to quantify the dissimilarities at residue level originated due to the different dynamic behavior of these two systems of Wt and pHDAC8. As mentioned above we have used two independent observables: 1) dihedral angle distribution 2) pair-wise residue interaction energy (PRIE) to get the maximum recovery of differences in the dynamic and energetic behavior of the two systems.

Since our concern was to identify the sequence of events that drove the Wt system to adopt an open conformation we divided the trajectories of Wt simulations into two states, closed (until the KF-loop opens up) and closed+open (or global dynamics) and carried out analysis on both of them with the focus mostly on the analysis with the closed states of the Wt. For the dihedral angle analysis, we have calculated the joint-probability densities of all possible combinations of 1) only backbone phi and psi angles (backbone dihedrals) and, 2) all-dihedral i.e. backbone as well as side chain dihedrals (phi, psi and all chi angles).

From this analysis we found that the residues that have contributed most to the variability in the backbone dihedral space (Fig. 4.2.8) in two simulations are, not unexpectedly, from the surface loop regions that form the wall of the active site and the neighboring residues of the phosphor-acceptor whereas the deacetylase core shows negligible or no differences. The all-dihedral angle analysis shows similar trends again. In addition to the residues that were found responsible in causing the significant differential backbone movement, some new residues have got selected in all-dihedral analysis. These differences are therefore the results emerged due to the side chain movement for those residues. The obtained top scoring residues reflect on the systematic behavior of different part of the HDAC8 system in two different simulations. Overall these results also reflect on the promiscuous path of signal transmission from one part to another part in the protein (Fig 4.2.8). Gly151, Phe152, Arg37 among them are experimentally found to be structurally and functionally important (Vannini *et al.*, 2004).

However, the path of communications were understood well when we analyzed the results of the pair-residue interaction energy (PRIE) analysis based on closed-state trajectories of Wt simulations by calculating the distances in individual probability distribution of each residue-pair interaction energy and compared the same with the global-dynamics. Projection of the top scoring residues on the 3-dimentional structure of protein (Fig 4.2.9) depicts a clear picture of it. We can assume that the effect started with the residues that are spatially and sequentially close to the phospho-acceptor Ser39. Introduction of the phosphate group to the Ser39 destabilizes the spatial organization of its neighboring residues such as Asp29, Lys33, Pro35, Lys36, Arg37, Ala38, Met40 of domain I; Asp88,

Ser93, Ala104 of domain-II, His334, Glu335, Phe336 of domain-IV -that is facing the phospho-acceptor containing domain-I -, Gly305, Tyr306, Asn307 of Glycine rich or G-rich loop (adjacent to the catalytic binding site) and Asn310, Arg313 of the alpha-helix located behind the phosphor-acceptor domain (or domain-I) (Fig. 4.2.1). This local perturbation is further transmitted to the neighboring domains, such as, to the residues Asp343, Val345, Leu346, and Arg353, of the extended loop of domain-IV, and Met274 of domain-III (a neighboring domain to the KF-loop). Focusing on the extended loop of domain-IV, we observe that deviation from the crystal structure (as measured by calculated RMSD values) is significantly larger in the phosphorylated state, allowing the side-chain of the of the residue Arg353 to adopt a more comfortable interaction with D233 and the backbone of domain III (C275 & S276), ultimately, this appears to cause the opening of the KF-loop.

Other structural changes induced by phosphorylation

Analyzing the first 40 top scoring individual interaction-pairs filtered out of some ~50000 interaction-pairs hints toward a promiscuous path of signal propagation for allosterism in pHDAC8 and other (complementary) inactivation mechanisms. Introduction of the phosphate group at Ser39, which is surrounded by the hydrophobic residues V25, A38, F336 and negatively charged residues D29, E335 (Fig. 4.2.1), drives this environment to undergo a major structural rearrangement. This reorganization destabilizes the interaction between domain-I and the Glycine rich (G-rich) loop located next to the catalytic site (Fig 4.2.1). In particular, the hydrogen bond between

the side chain of R37 and the backbone of G305 breaks upon phosphorylation. The G-rich loop is adjacent to the metal coordinating residues and lines the bottom of the active site, while R37 (a highly conserved residue in all HDACs) forms the wall of the internal cavity perpendicular to the binding tunnel. This internal cavity has been postulated to be involved in the diffusion of the reaction by-product acetate with the solvent (Vannini *et al.*, 2004, Haider *et al.*, 2011, Finnin *et al.*, 1999). Distance calculation between Nζ of R37 and the backbone O of G305 shows this pair comes closer than 3.5Å in 89% of the cases in the unphosphorilated form, but only in 12% of the cases in pHDAC8 (Table 4.2.2). The interaction with G303 is largely maintained (Table 4.2.2), but the loss of an interaction point enables R37 to partly occupy the internal cavity, which could be associated with an impaired passing of acetate in pHDAC8.

Table 4.2.2: Interactions of R37 with G-rich loop (residues 302-305)

Interacting pair	Frequency at which a pair comes closer than 3.5Å	
	Wt closed states	pHDAC8
R37-G305	~89% (~61% throughout)	~12%
R37-G303	~56% (~27% throughout)	~57%

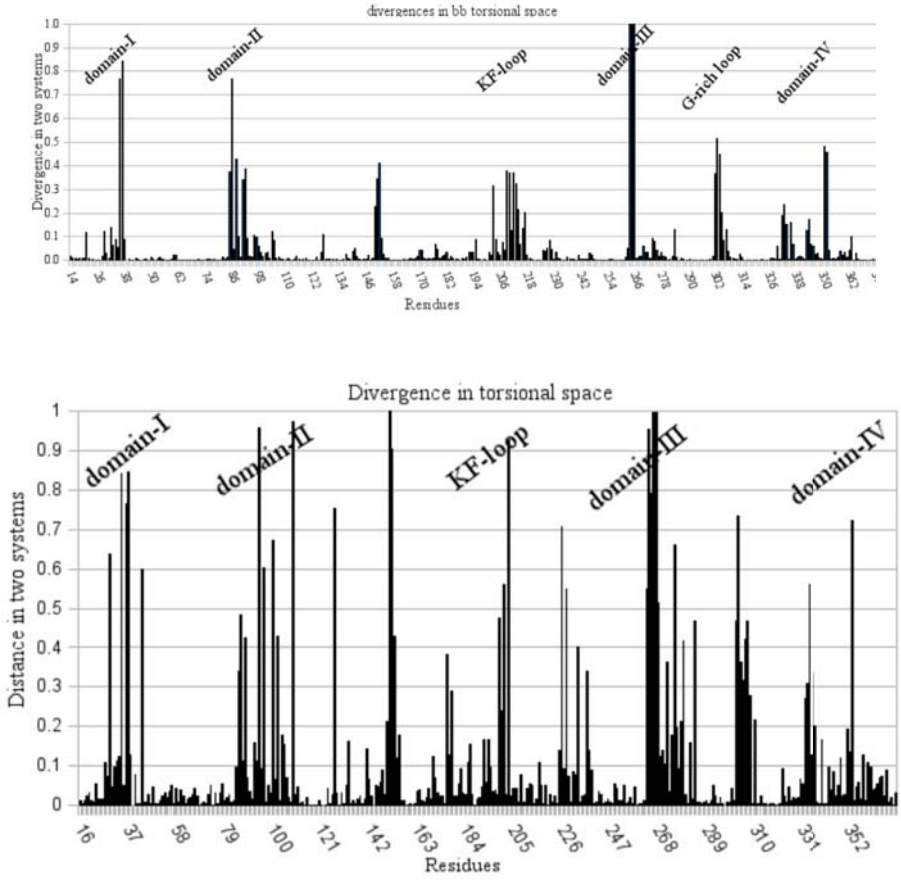
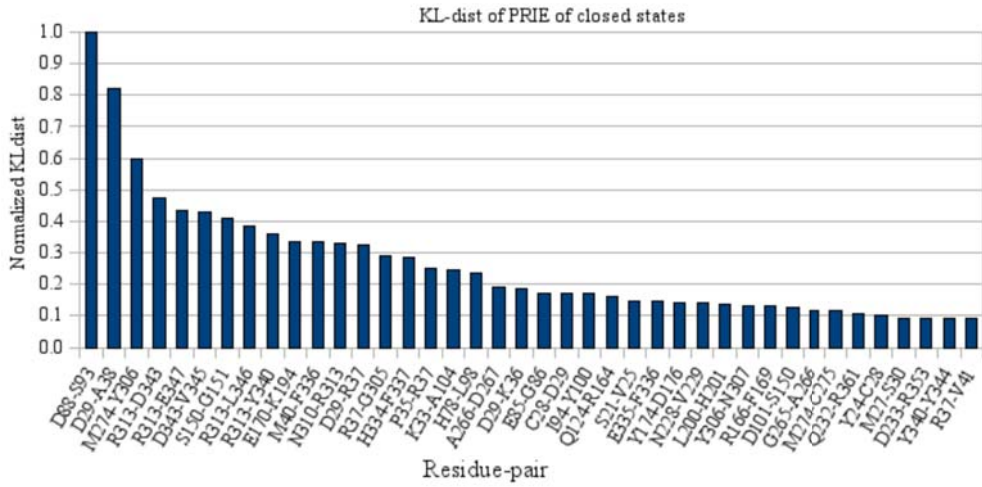


Figure 4.2.8. Dihedral angle distribution analysis. Top: analysis on the backbone dihedral angle distribution. Below: rotamer distribution analysis

Other observations derived from the KL analysis include:

1. A top-scoring residue pair Met274 of domain III and Tyr306 of G-rich loop–interacting with a favorable hydrophobic stacking interactions (Fig

4.2.9)- has been found to be on average ~2kcal more stable in pHDAC8 simulations than the unphosphorylated form. This can be the consequence of several other interaction behavior observed.



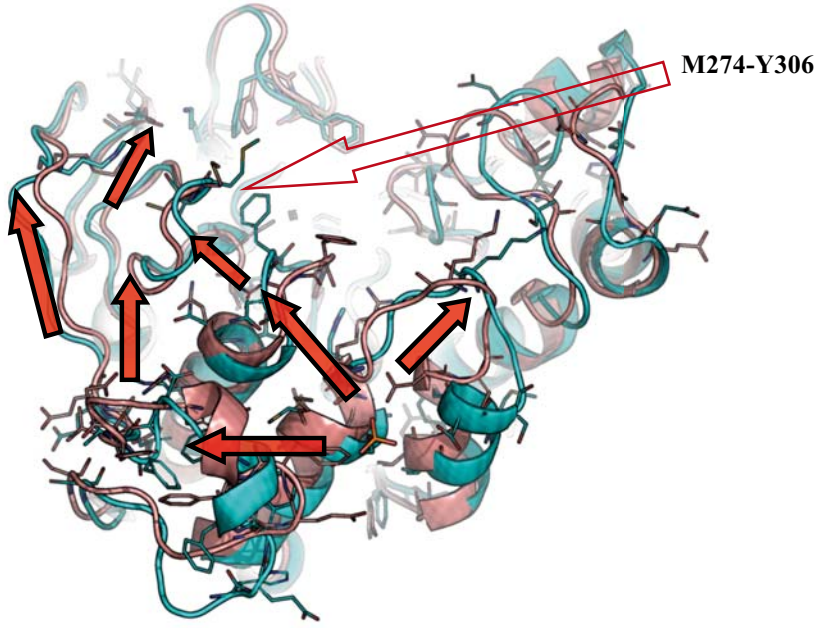


Fig 4.2.9: Pair-residue interaction energy analysis. Projection of the top scoring (in the KL-divergence analysis) interacting-pairs on Wt (light pink) and pHDAC8 form (cyan).

The alpha-helix that holds the C-terminal end of the G-rich loop (Fig 4.2.1) contains the residue R313 which has several favorable interactions with the residues E347, L346, Y340 of the extended loop of domain-IV (Fig.4.2.1 & 9). Due to the probable repulsion effect between pSer39, Glu335, Phe336 and Asp29 (Fig. 4.2.1) domain-IV moves away from domain-I to adopt a stable conformation. This possibly has facilitated the interactions between R313 and E347, L346, Y340 and consequently imparted a rear-stabilization in the domain-III (Fig 4.2.1) through the favorable stacking interactions between the hydrophobic sidechain of Met274 and phenyl-ring of Tyr306 (Tatko and Waters, 2004). Domain-III is surrounded by the G-rich loop, KF-loop (the loop that dissociates from the core structure), Asp-233 containing loop, and the extended loop of domain-IV. Apparently stabilization in the domain-III through the interacting pair Met274-Tyr306 imposed a rigidness on the neighboring domain i.e. on the KF-loop and its interaction network, e.g. with D233 and R353 as explained before in the section of conformational analysis (Fig 4.2.6 & 7).

2. Structural arrangement of R353 is such that it is spatially close to the KF-loop with only one loop between (loop containing residues 232-236) (Fig 4.2.6) and connected with it through interactions with the negatively charged residue D233. PRIE-analysis reveals that D233-R353 interaction is stabilized in the simulations of phosphorylated form by average 3.5kcal/mol and very likely that this plays a role in exerting an added stabilizing effect on the KF-loop in

phosphorylated simulations as hypothesized and explained in the previous section.

3. Domain-II (residue 86-106), predominated with mostly negatively charged amino acid residues is facing 4 residues of domain-I which are mostly hydrophobic with one exception of Lys33. Dihedral distribution as well as PRIE distribution analysis revealed that the behavior of this huge domain is significantly different in the two simulations with energetically favorable intra-domain residue-pairs D88-S93, H78-L98, E85-G86, I94-Y100, D92-S93 and an inter-domain (domain-I & domain-II) residue pair K33-A104. Analyzing inter-domain pair K33-A104 (Fig. 4.2.9), obtained from PRIE analysis we have found that this interacting pair is energetically favorable in phosphorylated simulations by around 3kcal/mol. Distance calculation shows that these two residues come closer than 3.5Å in more than 76% times in the simulations of phosphorylated system in contrast to the Wt where it is less than 4%. All these results indicate that in domain-II the extent of flexibility got reduced and becomes stable in the pHDAC8 than in Wt simulations due to inter- and intra- domain favorable interactions. Interestingly study has found that the optimum deacetylase activity of HDAC8 was obtained with a substrate-peptide that contains a C-terminal distal sequence consisting positively charged residues –KRHR where the acetylated-lysine is at the N-terminal side [table 1 in Gurard-Levin ZA, 2008] (Gurard-Levin and Mrksich, 2008). The substrate bound X-ray 3D structure shows the acetylated-lysine is held at the active site by the hydrophobic stacking from the KF-loop and with the electrostatic

interactions from domain-II (Vannini *et al.*, 2007). Form our analysis and the evidences obtained we propose a substrate binding mechanism in HDAC8 [Fig 4.2.10].

We could conclude from the above analysis that, in addition to the KF-loop, induced rigidity in the domain-II [Fig 4.2.1] due to phosphorylation might also be affecting in the substrate recognition and binding and makes it a poor-acceptor for its substrate.

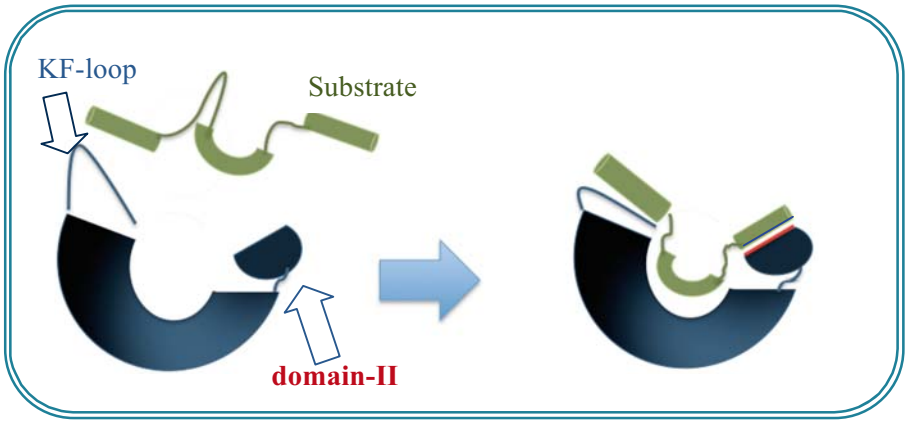


Fig 4.2.10: Possible substrate recognition mechanism involving KF-loop and Domain-II.

Does deacetylase activity of the HDAC8 KF-loop dependant?

To establish our hypothesis on the role of KF-loop in the catalytic activity we proposed some mutations based on the conformational analysis that has been explained in the previous sections. Based on the interaction network around the pair K202-F207 (Fig 4.2.6) the mutations were designed

in such a way that could either facilitate or inhibit breaking of the favourable interactions between hydrophobic side chain of Lys202 and pi-ring of the Phe207 to correlate with the catalytic activity. The list of mutations is given in the Table 4.2.3 and the rationales behind the selection of the mutants have been described below. The mutational studies were carried out both computationally and in wet-lab experiments.

A Metadynamics (MTD) simulation was carried out to explore the energetic barrier required just to open the KF-loop in each mutant for the comparative analysis with the Wt system. Metadynamics is used here in order to get the energetic cost which is required to keep the loop closed and to hold the K202-F207 pair instead of exploring the conformational space. For each system we carried out two different 2-dimensional MTD simulations to get the consensus estimate of the energetic barrier. Each system was run with two reaction coordinates:

1) The measure of root mean square deviation (RMSD) of the KF-loop with respect to the equilibrated closed structure (rmsd from closed conformation) and Wt open structure (rmsd from open conformation).
2) The measure of root mean square deviation (RMSD) of the KF-loop (rmsd) and distance between K202 and F207 with respect to the equilibrated closed structure.

The first mutant study was carried out on residue 202 where Lys was mutated to Met to stabilize the interaction with F207 with a more favorable hydrophobic interaction (Tatko and Waters, 2004) between sidechain of Met and pi-ring of Phe207. This would rigidify the KF-loop, inactivating the

protein. At the same time, the loss of the salt-bridge switch mechanism should render the protein less sensitive to regulation by phosphorylation.

A 100ns MD simulation of this mutant form did not generate the open-loop structure. The average distance between Met side chain and the pi-ring remains around 4.5Å throughout the simulations. We used Metadynamic simulation method on the equilibrated structure of this mutant to obtain the energetic cost required to break the hydrophobic interaction and open the loop. As expected a higher penalty, nearly 6kcal/mol, compared to the Wt (4.5kcal/mol) was required for the breaking of interaction and open the loop. Experimental analysis has found that this mutant is 3 fold less active than the Wt (Figure 4.2.12). Phosphorylation of this mutant further decreases its activity, indicating that this is not the only mechanism involved in protein inactivation, but possibly the most important one.

To investigate whether the loop-open states are energetically disfavored than the loop-close state in the mutant K202M we calculated the average energy of the ensemble of structures for the closed and open states using molecular mechanics/Poisson-Boltzmann solvent accessible surface area (MM/PBSA). Since the unbiased MD of K202M (100ns) did not generate any open conformation, the collection of open conformations were taken from the final geometry of the Metadynamics run of K202M, which was further subjected to 50ns Molecular Dynamics simulation run. It is worthy of note that, while the closed state is only ~11kcal more stable than its open state in the Wt form, the energy gap increases to ~50kcal in the K202M mutant K202M (Table 4.2.4). We also investigated the same on the phosphorylated mutant i.e. on pSer39- K202M. The results (Table 4.2.4) again were very much in agreement with the analysis and hypothesis.

The second mutation investigated is Phe → Ala in residue 207. This change was designed to forbid the hydrophobic interaction with K202, thus facilitating opening of the loop. The expected outcome was uncertain beca Table 4.2.3: list of mutants studied ning could be countered by a loss of affinity if the residue is involved in the capture of the substrate. However, we expected to see decreased response to phosphorylation of the mutant compared to Wt HDAC8.100ns molecular dynamics run of this mutant confirm the de-association of the loop from the deacetylase core of HDAC8 in the very beginning of each simulation. The experimental assay (Figure 4.2.12) shows that F202A is slightly less active than the Wt, suggesting a functional role of this residue other than regulation by phosphorylation. As noted before, this is not surprising because the residue is conserved in all Class I HDACs. Although the phosphorylated form of the mutant shows less activity than the non-phosphorylated form, the relative change is significantly lower than for the Wt form (22% vs 35%), confirming a role of this loop in regulation by phosphorylation.

Table 4.2.3: list of mutants studied

Residue ID	Residue	Mutation	Reasons
202	Lys	Met	To study the involvement of KF-loop in the deacetylase activity
207	Phe	Ala	
233	Asp	Ala	To study the allosteric path on HFAC8.
353	Arg	Ala	

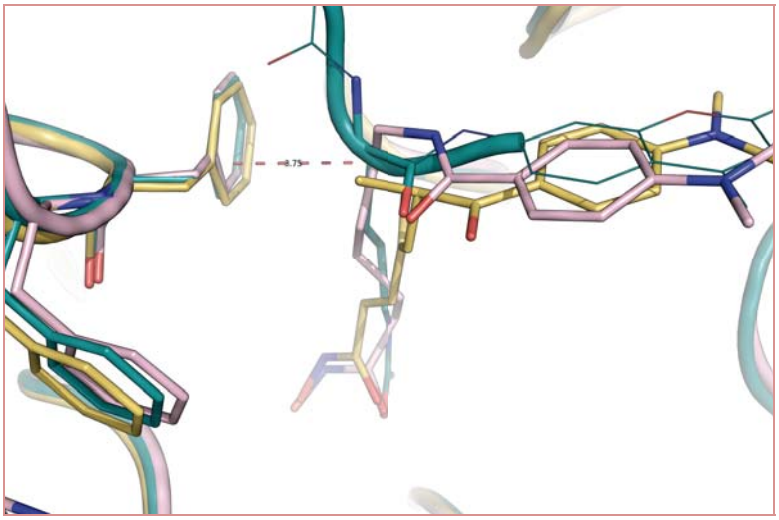


Figure 4.2.11: Structural alignment of different 3D structures of HDAC8 from PDB (PDB Ids: 1T64, 1T67, 2V5W) reveals the hydrophobic stacking interactions exerted by F208 to hold the substrate and ligand

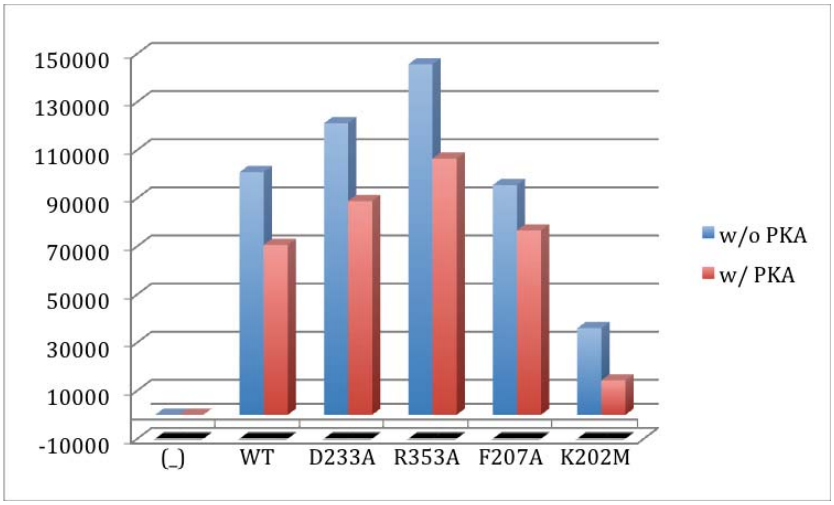


Figure 4.2.12: Resultant Activity assay plot, once the results were normalized by western blot quantification and negative control was substracted. In blue: without PKA values, in red: with PKA values. Mean values (from left to right, without/with PKA): 0/0; 100.825 (97736, 103915) / 70.572 (69859, 71286) ; 121.127 (120132, 122121) / 88.828 (86257, 91491) ; 145.665 (141585, 149745) / 106.448 (104337, 108560) ; 95.487 (88491, 102484) / 76.602 (69466, 83740)

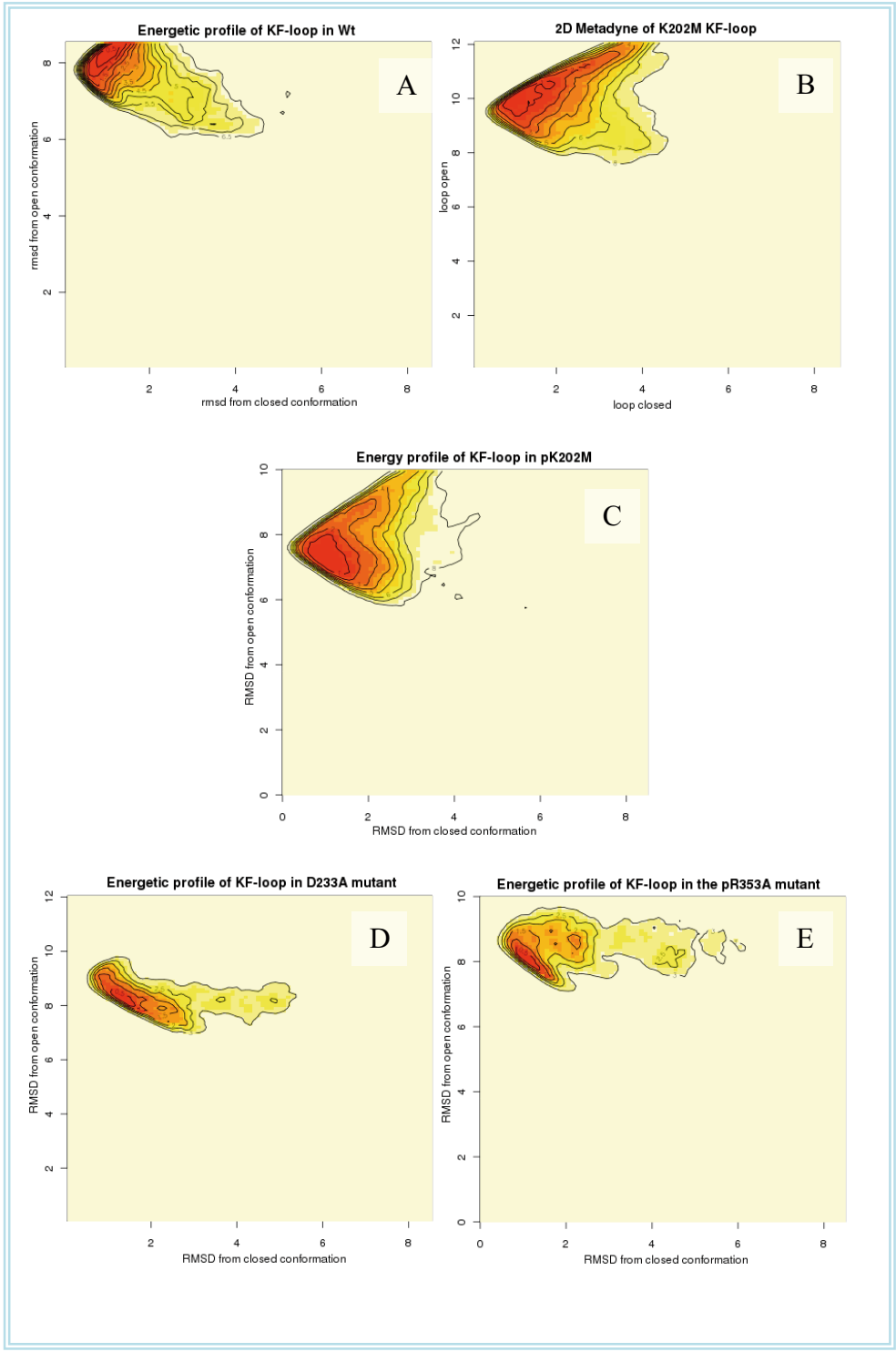
The third mutant investigated involves Asp233, which was replaced by Ala. Again, this mutant should facilitate the opening of the loop because the carboxylate of this residue forms a salt bridge with K202 in the closed conformation (Fig 4.2.6). Similarly, but in a more indirect manner, the mutation of Arg 353 to Ala should also facilitate loop opening. In this case, de-stabilization of the closed form should result from the loss of the D233-R353 salt-bridge, which cooperates with the K202-D233 salt bridge to form a stable network of ionic hydrogen bonds in the closed conformation. Similar to F207A, the opening of the loop should result in increased activity, unless the mutated residues have other functional roles beyond the regulatory one. Like in all other mutations investigated here, they were expected to decrease the response to phosphorylation.

100 ns of unbiased MD simulations did not afford the open conformation of the KF-loop, but Metadynamics runs predicted a barrier for the opening of the loop significantly lower than in the Wt form (2.5 vs. 4.5 kcal/mol). These results are in accord with the experimental results, which show an increase in deacetylase activity (Figure 4.2.12). Unexpectedly, phosphorylation reduced the activity in similar proportion to the Wt. Like in the case of K202M, we must conclude that opening of the KF-loop is an important, but not the only molecular mechanism to regulate enzymatic activity.

Table 4.2.4: Results of the mutational study on the assessment of the association of KF-loop in the deacetylase activity obtained from both computational and experimental activity assay (also can refer figure 4.2.12).

System of study	$E_{mm\text{-}pbsa}$ of closed states	$E_{mm\text{-}pbsa}$ of open states	$\Delta E_{mm\text{-}pbsa}$ (close - open)	loop open-closed rmsd (Flexible protein)	loop open rmsd-KFdist (Flexible protein)	Deacetylase activity profile
Wt	-4992.95	-4982.01	-10.94	4.5 (Fig 13A)	3.5 (Fig 13F)	Active
pHDAC8	-5160.26	-5148.92	-11.34	5.0	-	Less active that Wt
K202M	-5008.37	-4959.22	-49.15	6 (Fig 13B)	4.5 (Fig 13G)	Inactive
pK202M	-5276.14	-5220.94	-55.20	>6 (Fig 13C)	-	Inactive

System of study	$E_{mm\text{-}pbsa}$ of closed states	$E_{mm\text{-}pbsa}$ of open states	$\Delta E_{mm\text{-}pbsa}$ (close - open)	loop open-closed rmsd (Flexible protein)	loop open rmsd-KFdist (Flexible protein)	Deacetylase activity profile
F202A	No closed conformation was obtained in the equilibrated structure.					Less active than Wt
pF202A	Same as above					More active than the Wt-phosphorylated enzyme.
D233A	-4913.92	-4961.20	47.28	2.5 (Fig 13D)	2.0 (Fig 13H)	More active than Wt
p R353A	-5036.66	-5015.74	-20.92	2.5 (Fig 13E)	2.5 (Fig 13I)	Wt-type activity profile


Energetic profile of KF-loop in Wt
A
rmsd from open conformation
rmsd from closed conformation
2D Metadyne of K202M KF-loop
B
loop open
loop closed
Energy profile of KF-loop in pK202M
C
RMSD from open conformation
RMSD from closed conformation
Energetic profile of KF-loop in D233A mutant
D
RMSD from open conformation
RMSD from closed conformation
Energetic profile of KF-loop in the pR353A mutant
E
RMSD from open conformation
RMSD from closed conformation

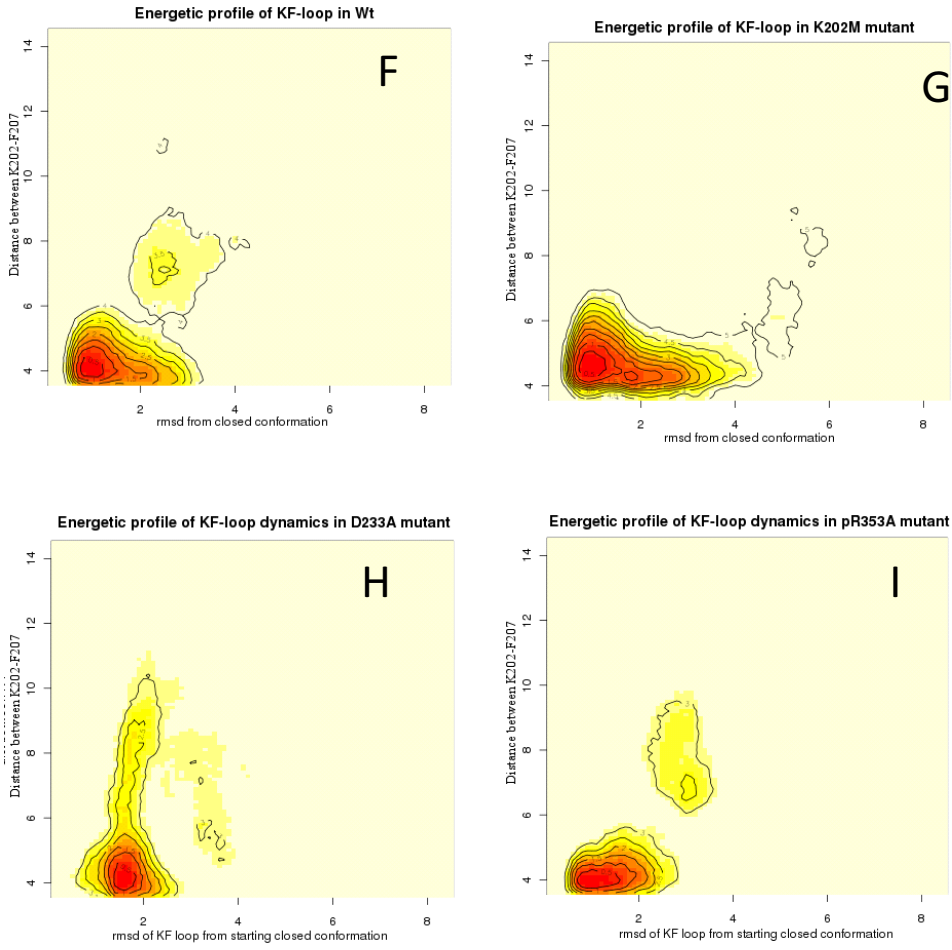


Fig 4.2.13: Energetic cost required to break the interactions between K202-F207 and obtain the barrier between the closed conformation and open conformation.

Note: The conclusions that we have drawn from this study has been included separately in the sub-section 5.2 of the conclusion chapter-5.

Chapter 5. Conclusions

5.1. Conformational Flexibility of Small Molecules

This study represents the most ambitious effort to date to accurately calculate the conformational energy of ligands upon binding. We have used, not only the molecular mechanics force-fields typically used in computer-aided drug design, but also the best quantum mechanical methods available, on a set of 92 diverse ligands. The main conclusions that can be extracted from the results presented in Chapter 4.1 of this thesis are:

- A practical issue in calculating the energy of a bioactive conformation is how to perform a restrained minimization that allows the molecule to reach a geometry that is simultaneously relaxed and in agreement with the experimental data. While the optimal solution consists in coupling QM methods with crystallographic refinement software, we propose a much simpler scheme based on harmonic restraints. This is found to provide results in line with the more common flat-bottom restraints. Torsional restraints, on the other hand, are not recommended.
- A computational protocol has been defined to ensure that the global energy conformation is found, making it possible to obtain a value of reference for the bioactive conformation, even at high QM levels.
- The most commonly used force-fields in computer-aided drug design (GAFF, MMFF & OPLS) have been evaluated. In all cases a large proportion of ligands are predicted to be strained by more than 5 kcal/mol: 40% or more if GBSA is used to simulate the effect of solvation, 65% onwards if calculations are carried out in vacuo.

- Optimization with the hybrid DFT functional B3LYP/6-31G* and PCM solvation produces much more realistic values, with only 8% of ligands reaching strains of 5 kcal/mol.
- Single-point calculations of the DFT-optimized geometries at the best level of theory to calculate conformational energies (MP2/aug-cc-pVDZ) can be used to further increase the quality of the results, but lack of correlation between the respective potential energy surfaces makes interpretation of the results difficult.
- On a subset of 41 ligands for which DFT and MP2 are highly correlated, we find that none of the ligands adopts conformations above 4kcal/mol. Setting an upper threshold for the bioactive conformations that, for the first time, is in line with logical estimates. Bioactive conformations overcoming this threshold are indicative of errors in the crystallographic refinement process.
- The computational approach here described can be used to accurately determine the energy of bioactive conformations, which is useful in lead optimization.

Our dataset can be used as a benchmark to evaluate less expensive computational approaches that, for instance, could be combined with 3D virtual screening methods (pharmacophoric searches, docking).

5.2 Allosterism in HDAC8

With this study we have thoroughly investigated the dynamic behavior of the HDAC8. Unlike the previous studies, which concentrated more in exploring the cavity around the catalytic site (Vannini *et al.*, 2004, Haider *et al.*, 2011), we strive to understand the molecular mechanisms involved in regulation of its activity by phosphorylation at position S39. The following conclusions can be extracted from the results presented in Chapter 5 of this thesis:

- Phosphorylation of HDAC8 has numerous structural consequences, both at the backbone and side-chain level, including domains I, II, IV and on the extended loop of domain IV (Fig 4.2.1)
- The Kullback-Leibler divergence method (KL), in combination with the use of descriptors related to the conformational adaptation (dihedral angles and pair-wise residue interaction energies)is adequate to identify in an unsupervised manner the most significant changes.
- Allosterism in HDAC8 involves multiple pathways and cannot be uniquely attributed to a sequential set of events.
- The loop containing residues 202-212 (or the KF-loop, Fig 4.2.1) shows a preference for the closed conformation observed in the crystal structure when the protein is phosphorylated, but has a marked tendency to open in the non-phosphorylated form.
- Loop opening exposes the aromatic residues F207 and F208 to solvent. The importance of the latter residue for substrate binding indicates that this conformational transition is functionally important.

We postulate that loop opening is necessary for recruitment of the substrate.

- Loop opening is controlled by a conformational switch involving F207 and the ionic triad K202, D233 and R353. We postulate that these three residues act cooperatively, and in the absence of any of them, the network of charged hydrogen bonds breaks apart, resulting in repulsive interactions between the charged amino group of K202 and the phenyl ring of F207, causing the loop to open.
- Single-point mutations were designed to test the postulated mechanism, and the enzymatic activity of the proteins was in agreement with the expectations:
    - K202M precludes the repulsive interaction with F207, trapping the protein in an inactive (closed loop) form.
    - F207A makes the closed conformation unstable. Although this residue is located next to the binding site and conserved in Class I HDACs (i.e. necessary for non-phosphorylatable HDACs), the mutant remains active and becomes less sensitive to regulation by PKA.
    - D233A promotes loop opening and increases the enzymatic activity by ~20%.
    - R353A, the only non-conserved member of the ionic triad, also promotes loop opening and increases the enzymatic activity by ~40%.
    - With the partial exception of D233A, all mutants are still down-regulated by phosphorylation, indicating that opening of the KF-loop in only one of the mechanism of inactivation.

- MD simulations and KL-divergence analysis point to several other complementary mechanisms, such as:
    - Blockage of the internal cavity involved in release of the reaction product acetate.
    - Re-arrangement of Domains III and IV may lead to alteration in substrate or protein-protein interactions.
    - Increased rigidity of Domain II upon phosphorylation prevents substrate recognition.

The structural insights provided by this study might assist in the design of subtype specific inhibitors, which till date is unavailable (Bieliauskas and Pflum, 2008).

References

Ahlrichs, R, Bar, M, Haser, M, Horn, H & Kolmel, C (1989). Electronic structure calculations on workstation computers: The program system turbomole. *Chemical Physics Letters* **162**: 165-9.

Amadei, A, Linssen, AB & Berendsen, HJ (1993). Essential dynamics of proteins. *Proteins* **17**: 412-25, 10.1002/prot.340170408.

Andrew R. Leach (2001). *Molecular modelling: principles and applications.* Addison-Wesley Longman Ltd.

Balasubramanian, S, Ramos, J, Luo, W, Sirisawad, M, Verner, E & Buggy, JJ (2008). A novel histone deacetylase 8 (HDAC8)-specific inhibitor PCI-34051 induces apoptosis in T-cell lymphomas. *Leukemia* **22**: 1026-34, 10.1038/leu.2008.9.

Beauchamp, KA, Lin, YS, Das, R & Pande, VS (2012). Are Protein Force Fields Getting Better? A Systematic Benchmark on 524 Diverse NMR Measurements. *Journal of Chemical Theory and Computation* **8**: 1409-14.

Becke, AD (1993). A new mixing of Hartree–Fock and local density‐functional theories. *J Chem Phys* **98**: 1372.

Berendsen, HJC, Postma, JPM, Van Gunsteren, WF, DiNola, A & Haak, J (1984). Molecular dynamics with coupling to an external bath. *J Chem Phys* **81**: 3684.

Berendsen, H, Postma, J, Van Gunsteren, W & Hermans, J (1981). Interaction models for water in relation to protein hydration. *Intermolecular forces* **331**.

Bhalla, J, Storchan, GB, MacCarthy, CM, Uversky, VN & Tcherkasskaya, O (2006). Local flexibility in molecular function paradigm. *Mol Cell Proteomics* **5**: 1212-23, 10.1074/mcp.M500315-MCP200.

Bieliauskas, AV, Pflum, MK (2008). Isoform-selective histone deacetylase inhibitors. *Chem Soc Rev* **37**: 1402-13, 10.1039/b703830p.

Bissantz, C, Kuhn, B & Stahl, M (2010). A medicinal chemist's guide to molecular interactions. *J Med Chem* **53**: 5061.

Boltz, S, Debreuve, E & Barlaud, M (2009). High-dimensional statistical measure for region-of-interest tracking. *IEEE Trans Image Process* **18**: 1266-83, 10.1109/TIP.2009.2015158.

Born, M, Oppenheimer, R (1927). Zur quantentheorie der molekeln. *Annalen der Physik* **389**: 457-84.

Bostrom, J, Norrby, PO & Liljefors, T (1998a). Conformational energy penalties of protein-bound ligands. *J Comput Aided Mol Des* **12**: 383-96.

Bostrom, J, Norrby, PO & Liljefors, T (1998b). Conformational energy penalties of protein-bound ligands. *J Comput Aided Mol Des* **12**: 383-96.

Bottomley, MJ, Lo Surdo, P, Di Giovine, P, Cirillo, A, Scarpelli, R, Ferrigno, F*, et al.* (2008). Structural and functional analysis of the human HDAC4 catalytic domain reveals a regulatory structural zinc-binding domain. *J Biol Chem* **283**: 26694-704, 10.1074/jbc.M803514200.

Brooijmans, N, Kuntz, ID (2003). Molecular recognition and docking algorithms. *Annu Rev Biophys Biomol Struct* **32**: 335-73, 10.1146/annurev.biophys.32.110601.142532 [doi]; 110601.142532 [pii].

Brooks, BR, Bruccoleri, RE, Olafson, BD, Swaminathan, S & Karplus, M (1983). CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry* **4**: 187-217.

Butler, KT, Luque, FJ & Barril, X (2009a). Toward accurate relative energy predictions of the bioactive conformation of drugs. *J Comput Chem* **30**: 601-10, 10.1002/jcc.21087.

Butler, KT, Luque, FJ & Barril, X (2009b). Toward accurate relative energy predictions of the bioactive conformation of drugs. *J Comput Chem* **30**: 601-10, 10.1002/jcc.21087.

Cachau, RE, Podjarny, AD (2005). High-resolution crystallography and drug design. *J Mol Recognit* **18**: 196-202, 10.1002/jmr.738.

Cardoso, MC, Schneider, K, Martin, RM & Leonhardt, H (2012). Structure, function and dynamics of nuclear subcompartments. *Curr Opin Cell Biol* **24**: 79-85, 10.1016/j.ceb.2011.12.009.

Case, D, Darden, T, Cheatham, T, Simmerling, C, Wang, J, Duke, R*, et al.* (2010). AMBER 11. *University of California, San Francisco*.

Cerutti, DS, Duke, R, Freddolino, PL, Fan, H & Lybrand, TP (2008). A vulnerability in popular molecular dynamics packages concerning Langevin and Andersen dynamics. *Journal of chemical theory and computation* **4**: 1669-80.

Chen, J, Brooks, CL,3rd (2008a). Implicit modeling of nonpolar solvation for simulating protein folding and conformational transitions. *Phys Chem Chem Phys* **10**: 471-81, 10.1039/b714141f.

Chen, J, Brooks, CL,3rd (2008b). Implicit modeling of nonpolar solvation for simulating protein folding and conformational transitions. *Phys Chem Chem Phys* **10**: 471-81, 10.1039/b714141f.

Choudhary, C, Kumar, C, Gnad, F, Nielsen, ML, Rehman, M, Walther, TC*, et al.* (2009). Lysine acetylation targets protein complexes and co-regulates major cellular functions. *Science* **325**: 834-40, 10.1126/science.1175371.

Corminboeuf, C, Hu, P, Tuckerman, ME & Zhang, Y (2006). Unexpected deacetylation mechanism suggested by a density functional theory QM/MM study of histone-deacetylase-like protein. *J Am Chem Soc* **128**: 4530-1, 10.1021/ja0600882.

Cramer, CJ (2004). *Essentials of computational chemistry: theories and models.* John Wiley & Sons Inc.

Cubero, E, Luque, FJ & Orozco, M (1998). Is polarization important in cation-pi interactions? *Proc Natl Acad Sci U S A* **95**: 5976-80.

Daily, MD, Gray, JJ (2007). Local motions in a benchmark of allosteric proteins. *Proteins* **67**: 385-99, 10.1002/prot.21300.

Darden, T, York, D & Pedersen, L (1993). Particle mesh Ewald: An N· log (N) method for Ewald sums in large systems. *J Chem Phys* **98**: 10089.

Davis, AM, Teague, SJ & Kleywegt, GJ (2003). Application and limitations of X-ray crystallographic data in structure-based ligand and drug design. *Angew Chem Int Ed Engl* **42**: 2718-36, 10.1002/anie.200200539 [doi].

Duan, Y, Wu, C, Chowdhury, S, Lee, MC, Xiong, G, Zhang, W*, et al.* (2003). A point‐charge force field for molecular mechanics simulations of proteins based on condensed‐phase quantum mechanical calculations. *Journal of computational chemistry* **24**: 1999-2012.

Essmann, U, Perera, L, Berkowitz, ML, Darden, T, Lee, H & Pedersen, LG (1995). A smooth particle mesh Ewald method. *J Chem Phys* **103**: 8577.

Ewald, PP (1921). Die Berechnung optischer und elektrostatischer Gitterpotentiale. *Annalen der Physik* **369**: 253-87.

Finnin, MS, Donigian, JR, Cohen, A, Richon, VM, Rifkind, RA, Marks, PA*, et al.* (1999). Structures of a histone deacetylase homologue bound to the TSA and SAHA inhibitors. *Nature* **401**: 188-93.

Fischer, S, Verma, CS (1999). Binding of buried structural water increases the flexibility of proteins. *Proc Natl Acad Sci U S A* **96**: 9613-5.

Forti, ,Flavio, Cavasotto, C,N., Orozco, ,Modesto, Barril, ,Xavier & Luque, F,Javier**A Multilevel Strategy for the Exploration of the Conformational Flexibility of Small Molecules**.

Frisch, M, Trucks, G, Schlegel, H, Scuseria, G, Robb, M, Cheeseman, J*, et al.* (2009). Gaussian 09, Revision A. 02, Gaussian. *Inc.: Wallingford, CT* **115**.

Gantt, SL, Joseph, CG & Fierke, CA (2010). Activation and inhibition of histone deacetylase 8 by monovalent cations. *J Biol Chem* **285**: 6036-43, 10.1074/jbc.M109.033399.

Gohlke, H, Kiel, C & Case, DA (2003a). Insights into protein-protein binding by binding free energy calculation and free energy decomposition for the Ras-Raf and Ras-RalGDS complexes. *J Mol Biol* **330**: 891-913.

Gohlke, H, Kiel, C & Case, DA (2003b). Insights into protein-protein binding by binding free energy calculation and free energy decomposition for the Ras-Raf and Ras-RalGDS complexes. *J Mol Biol* **330**: 891-913.

Grunstein, M (1997). Histone acetylation in chromatin structure and transcription. *Nature* **389**: 349-52, 10.1038/38664.

Gunasekaran, K, Ma, B & Nussinov, R (2004). Is allostery an intrinsic property of all dynamic proteins? *Proteins* **57**: 433-43, 10.1002/prot.20232.

Gurard-Levin, ZA, Mrksich, M (2008). The activity of HDAC8 depends on local and distal sequences of its peptide substrates. *Biochemistry* **47**: 6242-50, 10.1021/bi800053v.

Haberland, M, Montgomery, RL & Olson, EN (2009). The many roles of histone deacetylases in development and physiology: implications for disease and therapy. *Nat Rev Genet* **10**: 32-42, 10.1038/nrg2485.

Haider, S, Joseph, CG, Neidle, S, Fierke, CA & Fuchter, MJ (2011). On the function of the internal cavity of histone deacetylase protein 8: R37 is a

crucial residue for catalysis. *Bioorg Med Chem Lett* **21**: 2129-32, 10.1016/j.bmcl.2011.01.128.

Halgren, TA (1999). MMFF VI. MMFF94s option for energy minimization studies. *Journal of Computational Chemistry* **20**: 720-9.

Halgren, TA (1996a). Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *Journal of Computational Chemistry* **17**: 490-519.

Halgren, TA (1996b). Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *Journal of computational chemistry* **17**: 490-519.

Hartree, D.R. (1928). The wave mechanics of an atom with a non-coulomb central field. part I. theory and methods. **24**, 89-110.

Head-Gordon, M, Pople, JA & Frisch, MJ (1988). MP2 energy evaluation by direct methods. *Chemical physics letters* **153**: 503-6.

Herzberg, G, Shoosmith, J (1959). Spectrum and structure of the free methylene radical. *Nature* **183**: 1801-2.

Hohenberg, P, Kohn, W (1964). Inhomogeneous electron gas. *Physical Review* **136**: B864.

Hornak, V, Abel, R, Okur, A, Strockbine, B, Roitberg, A & Simmerling, C (2006). Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins: Structure, Function, and Bioinformatics* **65**: 712-25.

Jacobson, S, Pillus, L (1999). Modifying chromatin and concepts of cancer. *Curr Opin Genet Dev* **9**: 175-84, 10.1016/S0959-437X(99)80027-6.

Jakalian, A, Jack, DB & Bayly, CI (2002). Fast, efficient generation of high‐quality atomic charges. AM1‐BCC model: II. Parameterization and validation. *Journal of computational chemistry* **23**: 1623-41.

Jensen, AA, Spalding, TA (2004). Allosteric modulation of G-protein coupled receptors. *Eur J Pharm Sci* **21**: 407-20, 10.1016/j.ejps.2003.11.007.

Johnstone, RW (2002). Histone-deacetylase inhibitors: novel drugs for the treatment of cancer. *Nat Rev Drug Discov* **1**: 287-99, 10.1038/nrd772.

Jorgensen, WL, Chandrasekhar, J, Madura, JD, Impey, RW & Klein, ML (1983). Comparison of simple potential functions for simulating liquid water. *J Chem Phys* **79**: 926.

Jorgensen, WL, Tirado-Rives, J (1988). The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *J Am Chem Soc* **110**: 1657-66.

Kaminski, GA, Friesner, RA, Tirado-Rives, J & Jorgensen, WL (2001). Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *The Journal of Physical Chemistry B* **105**: 6474-87.

Kirschner, KN, Lexa, KW, Salisburg, AM, Alser, KA, Joseph, L, Andersen, TT*, et al.* (2007). Computational design and experimental discovery of an antiestrogenic peptide derived from alpha-fetoprotein. *J Am Chem Soc* **129**: 6263-8, 10.1021/ja070202w.

Kleywegt, GJ, Jones, TA (1997). Model building and refinement practice. *Methods Enzymol* **277**: 208-30.

Kohn, W, Sham, LJ & CALIFORNIA UNIV SAN DIEGO LA JOLLA. (1965). *Self-consistent equations including exchange and correlation effects.* APS.

Kouzarides, T (1999). Histone acetylases and deacetylases in cell proliferation. *Curr Opin Genet Dev* **9**: 40-8.

Kubinyi, H (1993). *3D Qsar in Drug Design: theory, methods and applications.* Kluwer Academic Pub.

Kumar, S, Ma, B, Tsai, CJ, Sinha, N & Nussinov, R (2000). Folding and binding cascades: dynamic landscapes and population shifts. *Protein Sci* **9**: 10-9, 10.1110/ps.9.1.10.

Laio, A, Parrinello, M (2002). Escaping free-energy minima. *Proc Natl Acad Sci U S A*. 99: 12562.

Langer, T., Hoffmann, R.D. (2006). Pharmacophores and Pharmacophore Searches. In: Mannhold R., Kubinyi H. and Folkers G. (eds). *Methods and Principles in Medicinal Chemistry* Weinheim: Wiley-VCH, pp 375.

Lebbink, JHG, Consalvi, V, Chiaraluce, R, Berndt, KD & Ladenstein, R (2002). Structural and thermodynamic studies on a salt-bridge triad in the NADP-binding domain of glutamate dehydrogenase from Thermotoga maritima: cooperativity and electrostatic contribution to stability. *Biochemistry (N Y)* **41**: 15524-35.

Lee, C, Yang, W & Parr, RG (1988). Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Physical Review B* **37**: 785.

Lee, H, Rezai-Zadeh, N & Seto, E (2004). Negative regulation of histone deacetylase 8 activity by cyclic AMP-dependent protein kinase A. *Mol Cell Biol* **24**: 765-73.

Lee, SH, Palmo, K & Krimm, S (2007). A comparative study of molecular dynamics in Cartesian and in internal coordinates: dynamical instability in the latter caused by nonlinearity of the equations of motion. *J Comput Chem* **28**: 1107-18, 10.1002/jcc.20627.

Li, X, He, X, Wang, B & Merz, K,Jr (2009). Conformational variability of benzamidinium-based inhibitors. *J Am Chem Soc* **131**: 7742-54, 10.1021/ja9010833.

Lindahl, E, Hess, B & Van Der Spoel, D (2001). GROMACS 3.0: a package for molecular simulation and trajectory analysis. *Journal of Molecular Modeling* **7**: 306-17.

Lindorff-Larsen, K, Ferkinghoff-Borg, J (2009). Similarity measures for protein ensembles. *PLoS One* **4**: e4203, 10.1371/journal.pone.0004203.

Luger, K, Mader, AW, Richmond, RK, Sargent, DF & Richmond, TJ (1997). Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389**: 251-60, 10.1038/38444.

Luque, I, Leavitt, SA & Freire, E (2002). The linkage between protein folding and functional cooperativity: two sides of the same coin? *Annu Rev Biophys Biomol Struct* **31**: 235-56, 10.1146/annurev.biophys.31.082901.134215.

Ma, B, Nussinov, R (2004). From computational quantum chemistry to computational biology: experiments and computations are (full) partners. *Phys Biol* **1**: P23-6, 10.1088/1478-3967/1/4/P01.

Mahoney, MW, Jorgensen, WL (2000). A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. *J Chem Phys* **112**: 8910.

Mamonova, T, Hespenheide, B, Straub, R, Thorpe, MF & Kurnikova, M (2005). Protein flexibility using constraints from molecular dynamics simulations. *Phys Biol* **2**: S137-47, 10.1088/1478-3975/2/4/S08.

Mann, A. (2003). Conformational restriction and/or steric hindrance in medicinal chemistry. In: Wermuth C. (ed). *The practice of medicinal chemistry (3rd edition)* London, United Kingdom: Academic Press,.

Ming, D, Wall, ME (2005). Quantifying allosteric effects in proteins. *Proteins* **59**: 697-707, 10.1002/prot.20440.

Møller, C, Plesset, MS (1934). Note on an approximation treatment for many-electron systems. *Physical Review* **46**: 618.

Nicklaus, MC, Wang, S, Driscoll, JS & Milne, GW (1995). Conformational changes of small molecules binding to proteins. *Bioorg Med Chem* **3**: 411-28.

Oehme, I, Deubzer, HE, Wegener, D, Pickert, D, Linke, JP, Hero, B*, et al.* (2009). Histone deacetylase 8 in neuroblastoma tumorigenesis. *Clin Cancer Res* **15**: 91-9, 10.1158/1078-0432.CCR-08-0684.

Olson, CA, Spek, EJ, Shi, Z, Vologodskii, A & Kallenbach, NR (2001). Cooperative helix stabilization by complex Arg–Glu salt bridges. *Proteins: Structure, Function, and Bioinformatics* **44**: 123-32.

Onufriev, A, Bashford, D & Case, DA (2004). Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins* **55**: 383-94, 10.1002/prot.20033.

Pearlman, DA, Case, DA, Caldwell, JW, Ross, WS & Cheatham, TE (1995). AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Comput Phys Commun* **91**: 1-41.

Pérez, A, Marchán, I, Svozil, D, Sponer, J, Cheatham III, TE, Laughton, CA*, et al.* (2007). Refinement of the AMBER force field for nucleic acids: improving the description of [alpha]/[gamma] conformers. *Biophys J* **92**: 3817-29.

Perola, E, Charifson, PS (2004a). Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding. *J Med Chem* **47**: 2499-510, 10.1021/jm030563w.

Perola, E, Charifson, PS (2004b). Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding. *J Med Chem* **47**: 2499-510, 10.1021/jm030563w.

Peterson, CL, Laniel, MA (2004). Histones and histone modifications. *Curr Biol* **14**: R546-51, 10.1016/j.cub.2004.07.007.

Ponder, JW, Case, DA (2003). Force fields for protein simulations. *Adv Protein Chem* **66**: 27-85.

Riley, KE, Op't Holt, BT & Merz, KM,Jr (2007). Critical Assessment of the Performance of Density Functional Methods for Several Atomic and Molecular Properties. *J Chem Theory Comput* **3**: 407-33, 10.1021/ct600185a.

Ryde, U, Nilsson, K (2003). Quantum chemistry can locally improve protein crystal structures. *J Am Chem Soc* **125**: 14232-3, 10.1021/ja0365328.

Salonen, LM, Bucher, C, Banner, DW, Haap, W, Mary, JL, Benz, J*, et al.* (2009). Cation–π Interactions at the Active Site of Factor Xa: Dramatic Enhancement upon Stepwise N‐Alkylation of Ammonium Ions. *Angewandte Chemie International Edition* **48**: 811-4.

SCHAEFER III, HF (1986). Methylene: A paradigm for computational quantum chemistry. *Science* **231**: 1100-7.

Schmidtke, P, Souaille, C, Estienne, F, Baurin, N & Kroemer, RT (2010). Large-scale comparison of four binding site detection algorithms. *J Chem Inf Model* **50**: 2191-200, 10.1021/ci1000289.

Schrödinger, E (1926). An undulatory theory of the mechanics of atoms and molecules. *Physical Review* **28**: 1049-70.

Sengupta, N, Seto, E (2004). Regulation of histone deacetylase activities. *J Cell Biochem* **93**: 57-67, 10.1002/jcb.20179.

Sitzmann, M, Weidlich, IE, Filippov, IV, Liao, C, Peach, ML, Ihlenfeldt, WD*, et al.* (2012). PDB ligand conformational energies calculated quantum-mechanically. *J Chem Inf Model* **52**: 739-56, 10.1021/ci200595n.

Somoza, JR, Skene, RJ, Katz, BA, Mol, C, Ho, JD, Jennings, AJ*, et al.* (2004). Structural snapshots of human HDAC8 provide insights into the class I histone deacetylases. *Structure* **12**: 1325-34, 10.1016/j.str.2004.04.012.

Soteras, I, Curutchet, C, Bidon-Chanal, A, Orozco, M & Luque, FJ (2005). Extension of the MST model to the IEF formalism: HF and B3LYP parametrizations. *J Mol Struct (Theochem)* **727**: 29-40.

Sousa, SF, Fernandes, PA & Ramos, MJ (2006). Protein–ligand docking: current status and future challenges. *Proteins: Structure, Function, and Bioinformatics* **65**: 15-26.

Stote, RH, Karplus, M (1995). Zinc binding in proteins and solution: a simple but accurate nonbonded representation. *Proteins* **23**: 12-31, 10.1002/prot.340230104.

Tatko, CD, Waters, ML (2004). Investigation of the nature of the methionine-pi interaction in beta-hairpin peptide model systems. *Protein Sci* **13**: 2515-22, 10.1110/ps.04820104.

Tirado-Rives, J, Jorgensen, WL (2006a). Contribution of conformer focusing to the uncertainty in predicting free energies for protein-ligand binding. *J Med Chem* **49**: 5880-4, 10.1021/jm060763i.

Tirado-Rives, J, Jorgensen, WL (2006b). Contribution of conformer focusing to the uncertainty in predicting free energies for protein-ligand binding. *J Med Chem* **49**: 5880-4, 10.1021/jm060763i.

Tomasi, J, Mennucci, B & Cancès, E (1999). The IEF version of the PCM solvation method: an overview of a new method addressed to study

molecular solutes at the QM ab initio level. *J Mol Struct (Theochem)* **464**: 211-26.

Tsui, V, Case, DA (2000). Theory and applications of the generalized Born solvation model in macromolecular simulations. *Biopolymers* **56**: 275-91, 2-E.

Vannini, A, Volpari, C, Filocamo, G, Casavola, EC, Brunetti, M, Renzoni, D*, et al.* (2004). Crystal structure of a eukaryotic zinc-dependent histone deacetylase, human HDAC8, complexed with a hydroxamic acid inhibitor. *Proc Natl Acad Sci U S A* **101**: 15064-9, 10.1073/pnas.0404603101.

Vannini, A, Volpari, C, Gallinari, P, Jones, P, Mattu, M, Carfi, A*, et al.* (2007). Substrate binding to histone deacetylases as shown by the crystal structure of the HDAC8-substrate complex. *EMBO Rep* **8**: 879-84, 10.1038/sj.embor. 7401047.

Vieth, M, Hirst, JD & Brooks, CL,3rd (1998). Do active site conformations of small ligands correspond to low free-energy solution structures? *J Comput Aided Mol Des* **12**: 563-72.

Vieth, M, Siegel, MG, Higgs, RE, Watson, IA, Robertson, DH, Savin, KA*, et al.* (2004). Characteristic physical properties and structural fragments of marketed oral drugs. *J Med Chem* **47**: 224-32, 10.1021/jm030267j.

Wang, DF, Helquist, P, Wiech, NL & Wiest, O (2005). Toward selective histone deacetylase inhibitor design: homology modeling, docking studies, and molecular dynamics simulations of human class I histone deacetylases. *J Med Chem* **48**: 6936-47.

Wang, J, Wolf, RM, Caldwell, JW, Kollman, PA & Case, DA (2004). Development and testing of a general amber force field. *Journal of computational chemistry* **25**: 1157-74.

Watts, KS, Dalal, P, Murphy, RB, Sherman, W, Friesner, RA & Shelley, JC (2010). ConfGen: a conformational search method for efficient generation of bioactive conformers. *J Chem Inf Model* **50**: 534-46, 10.1021/ci100015j.

Weigend, F, Häser, M, Patzelt, H & Ahlrichs, R (1998). RI-MP2: optimized auxiliary basis sets and demonstration of efficiency. *Chemical physics letters* **294**: 143-52.

Wu, R, Wang, S, Zhou, N, Cao, Z & Zhang, Y (2010). A proton-shuttle reaction mechanism for histone deacetylase 8 and the catalytic role of metal ions. *J Am Chem Soc* **132**: 9471-9, 10.1021/ja103932d.

Yang, XJ, Seto, E (2008). The Rpd3/Hda1 family of lysine deacetylases: from bacteria and yeast to mice and men. *Nat Rev Mol Cell Biol* **9**: 206-18, 10.1038/nrm2346.

Zhou, R (2003). Free energy landscape of protein folding in water: explicit vs. implicit solvent. *Proteins: Structure, Function, and Bioinformatics* **53**: 148-61.