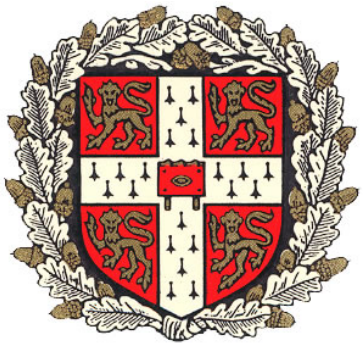
Extending the Boundaries of the Usage of NMR Chemical Shifts in Deciphering Biomolecular Structure and Dynamics

Aleksandr B. Sahakyan

A thesis submitted for the degree of
Doctor of Philosophy



Department of Chemistry
University of Cambridge
Darwin

9 May, 2012

To my parents and teachers

Acknowledgements

I would like to express my immense gratitude to Prof. Michele Vendruscolo for changing my life and myself. He encouraged me to apply to the University of Cambridge; an idea that surely would not freely circulate in my mind without a powerful stimulus that his kind letter was. Subsequently, years of productive discussion with him and his thorough guidance have most directly impacted the quality of research reflected in this thesis. But first of all, it has been his highly contagious enthusiasm, which he is so generous to "infect" us with, that was of an enormous aid in keeping the project and me going.

The project has directly benefited from the very productive collaboration with Dr. Wim F. Vranken from the Vrije Universiteit Brussel, who has kindly provided re-referenced extract of experimental chemical shift measurements. I am thankful to Catherine Pitt for her patience and quick resolution to many IT issues and computer crashes that I caused.

My presence here has become a reality owing to the Herchel Smith PhD Fellowship for which I wholeheartedly acknowledge the Herchel Smith Fund.

The path that has eventually laid my way to Cambridge started from my junior school days owing to my crossing with many people who have had key roles in my life. I am grateful to my junior school principal Hamlet Karamyan and class principal Ashkhen Burnazyan for their mind reforming teaching and for allowing me to conduct experiments in the back-side of the chemistry and physics classrooms, twice leading to mass evacuation in the school 185, Yerevan. Silva Kirakosyan and Hamlet Pirumyan had a profound role in encouraging deep thinking in chemistry and physics correspondingly. I am grateful

to Dr. Robert Ghazaryan, who has introduced me to porphyrin chemistry at the Yerevan State Medical University. The meeting that has significantly shifted my research interests was the one with Drs. Aleksan Shahkhatuni and Henry Panosyan, Astghik Shahkhatuni, Suren Mamyan and Aleksandr Piroyan, who have introduced me to the borderless opportunities of the NMR spectroscopy and had a profound role in the development of my scientific mentality. The personal schedule kindly granted to me by Dean Prof. Hasmik Hasratyan in my undergraduate years made it possible for me to allocate significant amount of time to research, working in the above-mentioned NMR laboratory. I am grateful to Dr. Ad Bax, for the chance of spending an inspiring month in his laboratory, as well as his time and tolerance to somewhat hilariously brave 19 y.o. me. The consecutive years of training and experience have left no trace of the previous confidence.

I am grateful to all the people whom I have met in Chemistry Department, Cambridge, within my PhD period, in particular to the dwellers of the office 145, for many interesting discussions and friendly atmosphere that has created home far from the home. In alphabetical order, I am particularly grateful to Dr. Carlos Bertoncini, Dr. Benedetta Bolognesi, Aditi Borkar, Dr. Carlo Camilloni, Dr. Andrea Cavalli, Prajwal Ciryam, Dr. Tomasso Eliseo, Dr. Farah El Turk, Dr. Eline Esbjörner, Dr. Celine Galvagnion, Dr. Shang-Te Danny Hsu, Hoi-Tik Alvin Leung, Francisco Newby, Dr. Edward O'Brien, Pietro Sormanni, Dr. Gian Gaetano Tartaglia, Dr. Giulia Tomba, Dr. Gergely Toth and Dr. Mark Tsechansky.

I am grateful to Darwin College for electing me as a Schlumberger Interdisciplinary Research Fellow to be inducted from October 2012, facilitating the continuation of my research in this exciting environment.

Last but not least, I am grateful to my parents, Karine and Babken, for their over-caring nature and absolute devotion that have left no life of their own.

Abstract

NMR chemical shifts have an extremely high information content on the behaviour of macromolecules, owing to their non-trivial dependence on myriads of structural and environmental factors. Although such complex dependence creates an initial barrier for their use for the characterisation of the structures of protein and nucleic acids, recent developments in prediction methodologies and their successful implementation in resolving the structures of these molecules have clearly demonstrated that such barrier can be crossed. Furthermore, the significance of chemical shifts as useful observables in their own right has been substantially increased since the development of the NMR techniques to study low populated "excited" states of biomolecules. This work is aimed at increasing our understanding of the multiple factors that affect chemical shifts in proteins and nucleic acids, and at developing high-quality chemical shift predictors for atom types that so far have largely escaped the attention in chemical shift restrained molecular dynamics simulations. A general approach is developed to optimise the models for structure-based chemical shift prediction, which is then used to construct CH3Shift and ArShift chemical shift predictors for the nuclei of protein side-chain methyl and aromatic moieties. These results have the potential of making a significant impact in structural biology, in particular when taking into account the advent of recent techniques for specific isotope labelling of protein side-chain atoms, which make large biomolecules accessible to NMR techniques. Through their incorporation as restraints in molecular dynamics simulations, the chemical shifts predicted by the approach described in this work create the opportunity of studying the structure and dynamics of proteins in a wide range of native and non-native states in order to characterise the mechanisms underlying the function and dysfunction of these molecules.

Disclaimer

This work has been done at the Department of Chemistry, University of Cambridge, within the period of October 2009 - April 2012. The initial extracts of the experimental chemical shift measurements, which are used to develop the described computational techniques, are provided by Dr. Wim F. Vranken, Structural Biology Brussels, Vrije Universiteit Brussel. The thesis represents original research done by the author, unless stated otherwise, and does not exceed the word-count limitation defined by the Degree Committee.

Abbreviations

COSMO - conductor-like screening model
CSP - chemical shift predictor
B3LYP - Becke's three-parameter exchange functional with the Lee, Yang and Parr correlation functional
BMRB - biological magnetic resonance bank
DFT - density functional theory
IEFPCM - integral equation formalism polarisable continuum model
MCSCF - multi-configurational self-consistent field
MD - molecular dynamics
MM - molecular mechanics
MP2 - correlated second order Møller-Plesset perturbation theory
NMR - nuclear magnetic resonance
NOE - nuclear Overhauser effect
PDB - protein data bank
QM - quantum mechanics
RAS - restricted active space
RDC - residual dipolar coupling
REMD - replica exchange molecular dynamics
RMSD - root mean squared deviation
SCF - self-consistent field

Contents

List of Figures

Introduction

The development of the current chemical shift predictors [Kohlhoff *et al.*, 2009; Lehtivarjo *et al.*, 2009; Meiler, 2003; Neal *et al.*, 2003; Shen & Bax, 2007; Wishart *et al.*, 1997; Xu & Case, 2001] for protein backbone atoms has largely benefited from the availability of substantial amount of protein backbone chemical shift measurements and associated high-resolution structures in publicly accessible databases [Berman *et al.*, 2000; Ulrich, 2007]. However, the importance and ease of protein chemical shift measurements is increasing in NMR community. The measurements are extending to protein side-chain atoms owing to the recently developed specific isotope labelling and NMR techniques that facilitate the precise measurements of chemical shifts from biomolecules, including those over-sized and/or invisible for conventional NMR techniques [Goto & Kay, 2000; Kainosho *et al.*, 2006; Tugarinov *et al.*, 2006]. It is thus highly desirable to have structure-based chemical shift predictors for protein side-chains, that will facilitate atomic-resolution research in biomolecular NMR, based on chemical shift measurements only. However, the initial attempts that were making use of the existing models for empirical chemical shift prediction appeared not to be successful in producing models that are sufficiently accurate for protein side-chain (and nucleic acid) atoms. The number of experimental measurements for side-chain chemical shifts, which is not high enough to train reasonable parameters for the existing models, largely determines the failure. This necessitates either a better definition of different terms that define the chemical shift values, or a substantial revision and optimisation of the models, so that an accurate prediction methodology can be devised for individual chemical shift types.

The work described in this thesis adopts both approaches by studying electric

field, solvent and ring current effects on chemical shifts, as well as by proposing a new chemical shift model definition. The work has already resulted in applicable chemical shift predictors for protein side-chain methyl CH3Shift and aromatic ArShift nuclei.

Chapter 1 reviews the concepts used in the thesis, aiming to briefly introduce the essential background, rather than to review all the published research relevant to this work. Such thoroughness is attempted in all the consecutive chapters, while discussing the motivation and place of the work-piece in the general tree of science. The chapter goes on by describing the results of the investigation of the non-specific solvent effects on biomolecular chemical shifts, as studied using a peptide backbone model. The presence of both solvent-exposed and buried residues in biomolecules promises a gain in extra precision if we account the solvent effects in the chemical shift prediction model. The results obtained here, are also of general importance on their own, further confirming the universality of electric field effects in mediating the non-specific solvent-solute interactions. Unfortunately, the recommendations inferred from this study is not yet feasible to be incorporated in chemical shift predictors, since the separate consideration of solvent exposed and buried nuclei splits the already-sparse chemical shift data for further parametrization of the models.

Chapter 2 adopts the strategy of smart model definition and optimisation that would not lead to overfitting problems while parametrizing the chemical shift predictors via an experimental data set with relatively low number of entries. The study is carried out by developing a structure-based and mathematically differentiable chemical shift predictor for protein side-chain methyl groups, that is now available as CH3Shift web server and a stand-alone program.

Chapter 3 modifies the same technique developed in Chapter 2, to derive a structure-based chemical shift predictor for the nuclei of aromatic side-chains, ArShift. Both Chapters 2 and 3 present the results with a thorough demonstration of the applicability of the predictors in assessing biomolecular structure and dynamics.

Chapter 4 proposes a new methodology for protein structure validation from the side-chain perspective. The method intrinsically takes into account the prediction errors, hence resulting in more robust values that will facilitate the high-throughput protein structure determination and validation based only on chemical shift measurements.

Chapter 5 presents the preliminary studies towards the development of chemical shift predictors for nucleic acids. The suitability and cross-dependence of ring current and electric field terms in nucleic acids are examined, as those are the major factors in defining the chemical shift values in highly conjugated and charged systems.

Chapter 6 touches the overall future prospects of biomolecular chemical shift predictions and very briefly introduces StarCore, an automatic **str**uctural **cor**relator that has a potential of generalising the development of structure-based predictors for any experimental parameter, suitable for an immediate application in restrained molecular dynamics simulations.

Bread has to mold in order to get penicillin.

Jerry Boatz

1 Background

1.1 Briefly on nuclear shielding constants and chemical shifts

NMR spectroscopy is one of the most useful analytical techniques in physical chemistry. It has a number of measurable parameters which provide unique insight about the structure and dynamics of molecules. As responses of the most internal molecular components, nuclei, NMR observables reflect fine trends in the interactions involving the atomic nuclei and the surrounding electron cloud [Levitt, 2006]. Historically, the first NMR observable with acceptable precision of measurements was chemical shift, which determines the overall position of resonance signals in conventional NMR spectra. Chemical shifts have been and are continuing to stay at the focus of scientific interest, which is a reflection of the complexity of chemical shifts and their non-trivial dependence on myriads of molecular and environmental parameters [Case, 1998, 2000; Jameson, 1996].

In NMR spectrometer, a nucleus in the molecule is exposed to an externally applied magnetic field, $\mathbf{B}_0$. However, the applied field affects the electron cloud around the nucleus, inducing electric currents with certain pattern. The electron currents, in turn, generate induced magnetic field, $\mathbf{B}_{ind}$, which alters the effective local field, $\mathbf{B}_{loc} = \mathbf{B}_0 + \mathbf{B}_{ind}$, to be sensed by the nucleus. Although the induced field will usually be around the order of 10^{-4} of the external $\mathbf{B}_0$ field, the resulting shift in the spin precession frequency is clearly reflected in NMR spectra and is measurable with high precision. The magnitude of that shift depends on the strength of the applied field and therefore on the ω_0 Larmor frequency of the nucleus, as well as on the selection of the reference precession frequency. To this end, in order to have a normalised measure of the alteration caused by the induced secondary fields, named electron screening or shielding effect, the following expression defining the δ chemical shift is used (Equation 1.1).

$$\delta = \frac{\omega - \omega_{ref}}{\omega_0} = \frac{\nu - \nu_{ref}}{\nu_0} \tag{1.1}$$

There, ω_0 represents the Larmor frequency of the given type of nucleus in magnetic field of a given strength, ν_0 is the operating frequency of the spectrometer, ω_{ref} and ν_{ref} hold the same meaning, but for a nucleus in a reference compound, and, ω and ν are the observed ones from the molecule under investigation. The obtained field-independent value is called a chemical shift, and is further scaled multiplying by 10^6, to convert the value into a more convenient form known as part per million, ppm.

Chemical shifts are very sensitive to the characteristics, such as spatial density distribution, shape, symmetry etc., of the electron cloud surrounding the given nucleus, and are thus susceptible to electric fields stemming from the environment and the nearby moieties of the molecule. This multi-dependence determines the potential of chemical shifts as parameters to differentiate nuclei under different molecular and environmental conditions.

The shielding ability of electrons is described via $\boldsymbol{\sigma}\mathbf{B}_0$ magnetic shielding of the nucleus, where the $\boldsymbol{\sigma}$ is the absolute nuclear shielding constant (shielding relative to a bare nucleus). Therefore the local magnetic field felt by the nucleus is $\mathbf{B}_{loc} = \mathbf{B}_0(1 - \boldsymbol{\sigma})$. The nuclear shielding constant $\boldsymbol{\sigma}$ is a second-rank tensor,

the elements of which can be represented as a 3×3 matrix. In isotropic solutions, where the molecule uniformly samples all the orientations, one can only observe the effects of the trace of such tensor, $1/3(\sigma_{xx} + \sigma_{yy} + \sigma_{zz})$, which results in isotropic nuclear shielding constant. It is often useful to break down the change in isotropic shielding constant into its different short- and long-range contributors (Equation 1.2):

$$\Delta\sigma = \Delta\sigma^{loc} + \Delta\sigma^{B} + \Delta\sigma^{vdW} + \Delta\sigma^{Hbond} + \Delta\sigma^{EF} + \Delta\sigma^{solv} \tag{1.2}$$

where, $\Delta\sigma^{loc}$ contains the local determinants of nuclear shielding, holding mostly the effects from substituents directly bonded to the atom of interest. Electron currents of the neighbouring groups are also capable of imposing additional magnetic fields at the nuclear site, which give rise to $\Delta\sigma^{B}$ term. That term is also the holder of ring current effects, if conjugated aromatic systems are present in vicinity. Hydrogen bonding, $\Delta\sigma^{Hbond}$, and van der Waals, $\Delta\sigma^{vdW}$, effects are generated by non-covalently interacting neighbours. Hydrogen bonding can cause a significant change ($\Delta\sigma^{Hbond}$) in shielding of the involved nuclei. Despite the clear theoretical and experimental evidence showing stronger interactions with electron delocalization effects present in hydrogen bonds [Stevens & Coppens, 1980], many aspects of hydrogen bond effects [Legon & Millen, 1987; Masunov *et al.*, 2001], including its influence on nuclear shielding constants [Oldfield, 2002], were shown to be mostly of electrostatic character. Thus, in some cases, this term can be merged with $\Delta\sigma^{EF}$. Electric field effects ($\Delta\sigma^{EF}$) from both the neighbouring and distant groups are proven to be important in formation of chemical shift values [Buckingham, 1960; Buckingham & Pople, 1963; Pearson *et al.*, 1993]. Solvent effects on nuclear shielding constants reflected in the $\Delta\sigma^{solv}$ term explain the solvent dependence of the observed chemical shifts, which along with $\Delta\sigma^{EF}$ term determine the sensitivity of nuclear shielding constants towards long-range interactions.

Chemical shifts are merely the referenced values of absolute nuclear shielding constants, as determined via the Equation 1.3:

$$\delta = \delta_{ref} + \sigma_{ref} - \sigma \tag{1.3}$$

where σ_{ref} and δ_{ref} are the absolute nuclear shielding constant and the assumed chemical shift value for the reference molecule. Therefore all the relations for the absolute shielding constants and their tensor properties hold true for chemical shifts as well.

1.2 Brief overview on chemical shift calculations

The existence of multiple complex factors contributing to the values of chemical shifts, make their calculations a non-trivial task.

In general, all the methods suggested for the estimation of chemical shifts can be classified into three main groups: a) first principle *ab initio* or density functional theory (DFT) approaches where no *a priori* knowledge or experimental observation is needed [Helgaker *et al.*, 1999; Oldfield, 1995]; b) methods based on empirical evaluation of shielding constants through a number of approximations accompanied by a proper parametrization of the methods against the experimental measurements [Wishart & Nip, 1998]; and c) the methods based on databases, where, for example in case of protein chemical shifts, a sequence homology is used to identify and assign chemical shifts from the closest structure in the underlying experimental data set, mostly using machine learning techniques [Shen & Bax, 2007; Wishart *et al.*, 1997]. The latter two classes can be considered in the group of empirical methods.

1.3 Quantum mechanical calculations of chemical shifts

The fact that the nuclear shielding phenomenon can be fully attributed to electron cloud properties surrounding the given nucleus, enables the calculation of chemical shifts via quantum chemical methods within the frames of the Born-Oppenheimer approximation. Here, the external magnetic field $\mathbf{B}_0$ is treated as a perturbation and is expressed by the magnetic vector potential $\mathbf{A}$ calculated

from an electron density (Equation 1.4) [Helgaker *et al.*, 1999].

$$\mathbf{B}_0 = \bigtriangledown \mathbf{A} \tag{1.4}$$

The same $\mathbf{B}_0$ can be described via many choices for $\mathbf{A}$ with a single selection denoted as the "gauge" of the magnetic vector potential. A number of methods have been developed to obtain the quantum chemical descriptors of the nuclear shielding, that would be independent of the selection of magnetic vector potential [Helgaker *et al.*, 1999; Jameson, 1996]. Of those suggestions, the most widely used approach is the *gauge-invariant atomic orbital* (GIAO) method [Ditchfield, 1974; Wolinski *et al.*, 1990], which assigns an exponential factor containing the $\mathbf{B}_0$ field to each of the atomic orbitals in the used basis set. The GIAO can easily be used with the known *ab initio* and DFT model chemistries coupled with the conventional basis sets. The σ_{ij} elements of the nuclear shielding tensor are then evaluated as follows (Equation 1.5):

$$\sigma_{ij} = \left[\frac{\partial^2 E}{\partial B_i \partial \mu_j}\right] = \sum_{\alpha\beta} D_{\alpha\beta} \frac{\partial^2 h_{\alpha\beta}}{\partial B_i \partial \mu_j} + \sum_{\alpha\beta} \frac{\partial D_{\alpha\beta} \partial h_{\alpha\beta}}{\partial B_i \partial \mu_j} \tag{1.5}$$

where B_i is the i^{th} component of the external magnetic field and μ_j is the j^{th} component of the nuclear magnetic moment. $D_{\alpha\beta}$ and $h_{\alpha\beta}$ are from the GIAO basis set and represent the generic elements of the one-electron density and Hamiltonian matrices respectively (as adapted in [Benzi *et al.*, 2004]).

Further, in order to obtain chemical shifts, one should also calculate and subtract the absolute nuclear shielding constant of the given nucleus in a reference compound (TMS - tetramethylsilane for protons and carbons).

The absolute nuclear shielding constants and their tensor properties are very attractive objects for studies on their own, and, can be directly compared to the experimental shielding constants obtained using conventional chemical shift measurements combined with thorough experimental estimations for the absolute shielding constants of small reference molecules. However, if the influence of a certain factor on chemical shift is of interest, only the calculation of the change in nuclear shielding constants would be the way to complete the task using quantum chemistry, as the extra referencing of the values for obtaining chemical shifts will

impose an additional error arising from the absolute shielding calculations in another (reference) molecular model.

1.4 Empirical calculations of chemical shifts

Despite the demonstrated precision of the *ab initio* and DFT methods for the evaluation of chemical shifts [Adamo & Barone, 1998; Helgaker *et al.*, 1999], their usage is still limited by the size of the molecular system. For a correct NMR parameter calculation, the first principle approaches require a usage of not only a good level of theory, but also a sophisticated basis set for a sufficient freedom assigned to electron density. Therefore the calculations quickly become very demanding in terms of necessary computational resources as we move from small molecules with tens of atoms to bigger systems.

To this end, the empirical methods of chemical shift evaluation are of great importance, especially taking into account their applicability to biomolecular systems. The methods emerged as a number of patterns correlating different descriptors of biomolecular structure with chemical shifts were observed. The acquisition of a sufficient amount of structural and NMR data and their systematic deposition in the biomolecular databases (PDB - Protein Data Bank for biomolecular structures [Berman *et al.*, 2000], and BMRB - Biological Magnetic Resonance Bank for NMR parameters [Ulrich, 2007]), made possible the emergence of clear trends, which, after further refinement and implementation, formed the first empirical chemical shift predictors.

Methods appeared that rely on databases of experimental measurements (see for instance Talos [Cornilescu *et al.*, 1999], ShiftY [Wishart *et al.*, 1997], PROSHIFT [Meiler, 2003] and Sparta [Shen & Bax, 2007]) or DFT calculations performed on small polypeptide segments of different conformation and composition (ShiftS [Xu & Case, 2001]). However, one of the major drawbacks of the database based approaches is the incomplete coverage of all the possible conformational and chemical (in terms of neighbouring residue combination) variations. Here is where the usefulness of empirical chemical shift prediction methods based on parametrized equations that only operate on the local atomic-scale geometric arrangement at the vicinity of the query nucleus is obvious. Of all the suggested variations,

ShiftX [Neal *et al.*, 2003] has become the most widely used one. The method breaks down chemical shifts into different contributions similar to Equation 1.2 and uses a set of parametrized equations for chemical shift dependence on a number of structural descriptors and phenomenological terms to derive estimates of different chemical shift contributions.

A relatively more recent method of reliable chemical shift prediction, CamShift [Kohlhoff *et al.*, 2009], is developed based prevalently on interatomic distances as descriptors of structure. The latter specialty determines the computational efficiency of the method and enables its usage in restrained molecular dynamic simulations [Robustelli *et al.*, 2010], where frequent and numerous chemical shift calculations are required.

1.5 Molecular dynamics simulations in structural biology

Increasing amount of evidence supports the idea that not only is the molecular world (with its complex behaviour) a result of a unique structural organisation, but is also a consequence of specific dynamics. As we increase the complexity of the studied molecular systems, their dynamics becomes crucial for the correct description of molecular properties. To this end, the molecular dynamics (MD) simulations have become an important part in structural chemistry and biology, where the dynamics of biomolecules is proven to be one of the key determinants in the complex biological regulation processes (see for example [Kern & Zuiderweg, 2003]) that propagates into the phenomenon of life.

MD simulations *in silico* reproduce both the short timescale fluctuations and long timescale rearrangements in molecules by solving the Newton's equation of motion (Equation 1.6):

$$\frac{\partial^2 x_i}{\partial t^2} = \frac{F_{x_i}}{m} \tag{1.6}$$

where m is the mass of a particle, x_i is one of the Cartesian coordinates and F_i is the force acting along the selected coordinate axis. Thus, the major problem in MD simulations becomes the correct representation of forces, which an atom

would be encountered to given all the present intra and intermolecular interactions.

Thorough parametrization of the molecular mechanics (MM) force fields and their increasing complexity, supported by the exponentially growing efficiency of computational resources, have converted MD simulations into an essential tool for filling the time resolution gap in the available experimental techniques [Karplus & McCammon, 2002; Schaeffer *et al.*, 2008; Shea & Brooks, 2001]. It has also been demonstrated that if the force fields are appropriately modified to account for experimental observables, resulting restrained simulations can almost become experimental techniques, outputting realistic results regardless the type of the used molecular mechanics (MM) component [Camilloni *et al.*, 2012; De Simone *et al.*, 2009b].

1.6 Chemical shift restrained molecular dynamics simulations

NMR spectroscopy has long been used for protein structure determination, prevalently based on nuclear Overhauser effect as the source of distance restraints [Wüthrich, 1986]. The inclusion and popularity of the anisotropic NMR parameters further increased the applicability and robustness of the NMR-based structural methods [Bax & Grishaev, 2005; Blackledge, 2004]. However, chemical shifts still remain one of the most available NMR parameters, which not only can be measured with a high precision, but can also be retrieved from the highly dynamic and partially disordered macromolecules, where the conventional NMR methods fail to obtain a sufficient number of distance restraints for solving the structure.

Although it has been increasingly clear that chemical shifts are very sensitive to structures of proteins, they have not been used in 3-dimensional structure determination, because of the lack of obvious laws governing their structural dependence unlike, for example, the case of vicinal J-couplings [Hoch *et al.*, 1985]. Therefore a full exploitation of chemical shifts was an open problem, prevalently determined by the absence of reliable and fast chemical shift estimation algo-

rithms. The situation has substantially been changed when the reliable predictors eventually appeared. The development of the ShiftX predictor [Neal *et al.*, 2003] has been soon followed by its application in conformational search for protein structure determination using chemical shift data via the suggested Cheshire procedure [Cavalli *et al.*, 2007] and Monte Carlo simulations [Robustelli *et al.*, 2009]. Furthermore, those simulations became even more feasible as the fast and portable predictor CamShift had been developed [Kohlhoff *et al.*, 2009] for protein backbone atoms, which was immediately and successfully tested within MD simulations to add chemical shift restraints [Robustelli *et al.*, 2010]. Hence, new chemical shift restrained molecular dynamics simulations have appeared that combine the power of experimentally determined chemical shifts and the available state-of-the-art MM force fields to determine the 3-dimensional structures of biomolecules, as well as to infer their dynamics.

The method introduces the restraints in the molecular dynamic simulations via a chemical shift penalty function E^{CS} as expressed in Equation 1.7, so that the resulting potential energy E of the system becomes $E = E^{FF} + E^{CS}$, where E^{FF} denotes the potential energy from the molecular mechanics force field.

$$E^{CS} = \alpha \sum_i \sum_j E_{ij}^{CS} \tag{1.7}$$

In the Equation 1.7, E_{ij}^{CS} is the penalty component for the amino acid residue i and the chemical shift type j. In case CamShift, that supports only the predictions for backbone chemical shifts, is used as the chemical shift prediction engine in the restrained simulations, j is one of the $^{1}H^{\alpha}$, $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, $^{1}H^{N}$ and ^{15}N nuclei. The multiplier α defines the weight of the penalising energy in modifying the total energy E of the studied system.

$$E_{ij}^{CS} = \begin{cases} 0 & \text{if } |\delta_{clc}^{ij} - \delta_{exp}^{ij}| \leq n\epsilon_j \\ \left(\frac{|\delta_{clc}^{ij} - \delta_{exp}^{ij}| - n\epsilon_j}{\beta_j}\right)^2 & \text{if } n\epsilon_j < |\delta_{clc}^{ij} - \delta_{exp}^{ij}| < x_0 \\ \left(\frac{(x_0 - n\epsilon_j)}{\beta_j}\right)^2 + \gamma \times \tanh\left(\frac{2x_0(x_0 - n\epsilon_j)(|\delta_{clc}^{ij} - \delta_{exp}^{ij}| - x_0)}{\gamma\beta_j^2}\right) & \text{if } x_0 \leq |\delta_{clc}^{ij} - \delta_{exp}^{ij}| \end{cases} \tag{1.8}$$

The E_{ij}^{CS} component itself is a function of the $\delta_{clc}^{ij} - \delta_{exp}^{ij}$ difference between the calculated and experimental chemical shifts, and is defined in Expression 1.8 with its schematic representation drawn in Figure 1.1.

The Expression 1.8 implies a flat bottom (zero penalty) if the $\delta_{clc}^{ij} - \delta_{exp}^{ij}$ difference for the chemical shift is less than or equal to the standard error ϵ_j of the chemical shift prediction for the given j type of nucleus multiplied by the tolerance coefficient n. Otherwise, the energy penalty increases harmonically till the x_0 cutoff value for the $|\delta_{clc}^{ij} - \delta_{exp}^{ij}|$ difference, after which the dependence obeys a harmonic tangent function, with γ set to 20.

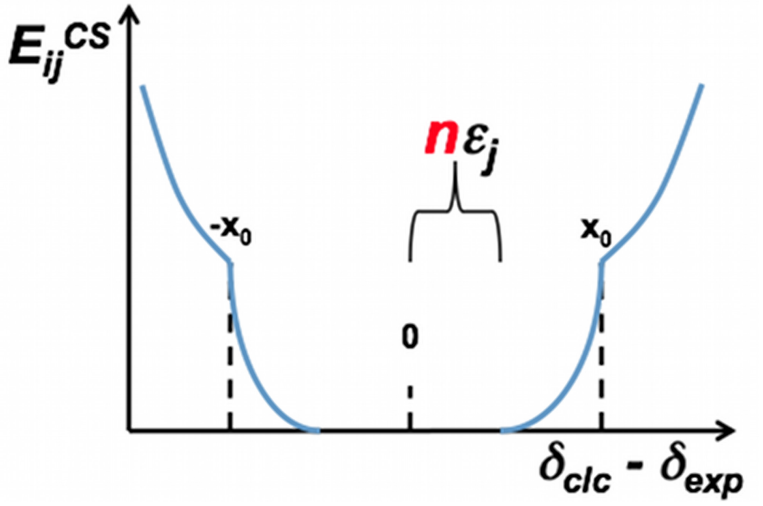


Figure 1.1: *A schematic representation of the chemical shift penalising energy function used in the chemical shift restrained molecular dynamics simulations. The E_{ij}^{CS} is the penalty component for the amino acid residue i and chemical shift type j. The horizontal axis represents the difference between the calculated and experimental chemical shifts. The function E_{ij}^{CS} has a flat bottom to account for the standard error ϵ_j in the CamShift estimation of chemical shifts of type j. The multiplier n represents the coefficient of tolerance toward the error. Outside the flat bottom, the penalty increases harmonically until the indicated x_0 cutoff value for the $\delta_{clc}^{ij} - \delta_{exp}^{ij}$ chemical shift difference. The figure is adapted from [Robustelli et al., 2009].*

The value of x_0 is usually set to 4 for protons and 20 for carbon and nitrogen nuclei. β_j is a weight factor determined by the variation of the chemical shifts

of type j in the BMRB database. The E_{ij}^{CS} chemical shift penalising energies are then used to calculate F_{ij}^{CS} chemical shift-based forces acting on each nucleus for which chemical shift data have been included in simulations and aim to move atoms in directions that minimise the $\delta_{clc}^{ij} - \delta_{exp}^{ij}$ difference. The force is the negative gradient of energy, as expressed in Equation 1.9.

$$\mathbf{F}_{ij}^{CS} = - \bigtriangledown \mathbf{E}_{ij}^{CS} \tag{1.9}$$

The restrained molecular dynamics simulations are then performed using the hybrid force that combines both the MM-force-field-based and restraining $\mathbf{F}_{ij}^{CS}$ forces for plugging into Newton's equation of motion (Equation 1.6). The use of hybrid forces ensures that a new driving force is involved in the simulations to improve the agreement between the calculated and experimental chemical shifts and to navigate the conformational search toward the natively populated ensembles.

1.7 Theoretical studies of dielectric permittivity effects on protein backbone chemical shifts

The advent and development of computational techniques enable the studies on a wider range of idealised problems and situations *in silico*, which would have been impossible to perform otherwise by purely experimental means. Recent advances in biomolecular NMR spectroscopy have greatly enhanced the computational studies, where NMR parameters can easily be incorporated as restraints in molecular dynamics simulations, increasing the accuracy of the results from simulations up to almost experimental precision [Cavalli *et al.*, 2007; Kohlhoff *et al.*, 2009; Robustelli *et al.*, 2009, 2010; Sahakyan *et al.*, 2011a,b; Shen & et al, 2008]. To further increase our fundamental understanding of chemical shifts, the effects of dielectric permittivity or non-specific solute-solvent interactions on chemical shifts of biological importance are targeted, making use of a model peptide in different conformations and the capabilities of computational chemistry.

1.7.1 The studied reduced molecular model

As a molecular model for the quantum chemical studies, the bi-capped L-alanine is used (Figure 1.2) with the acetylic (*Ace*) and N-methyl amide (*Nme*) groups added at the L-alanine amino and carboxylic termini respectively. The resulting structure (N-acetyl-L-alanyl-N-methylamide) is the simplest stereoactive model of a peptide and is commonly referred as alanine dipeptide.

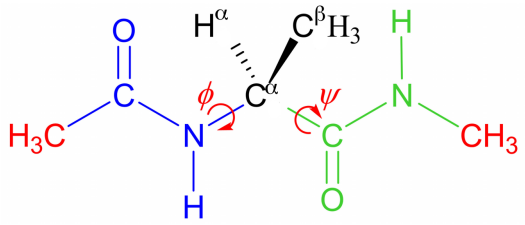


Figure 1.2: *The structure of Ace-Ala-Nme, the model for quantum chemical investigation of dielectric permittivity dependence of nuclear shielding constants relevant to biomolecular NMR. The two peptide moieties are highlighted (blue and green) along with the backbone ϕ and ψ dihedral angles.*

Hereafter, the notation *Ace-Ala-Nme* will be adopted throughout the discussion. As a reduced peptide model, *Ace-Ala-Nme* has been used in a number of studies for both costly *ab initio* calculations that target biomolecular problems [Han *et al.*, 1998; Head-Gordon *et al.*, 1991; Wang & Duan, 2004], and for experimental studies [Mehta *et al.*, 2004; Weise & Weisshaar, 2003].

1.7.2 The implemented scheme for quantum mechanical calculations

Hybrid density functional theory [Kohn & Sham, 1965] (DFT) is used with the Becke's three-parameter exchange functional and the Lee, Yang and Parr correlation functional [Becke, 1993; Lee *et al.*, 1988; Miehlich *et al.*, 1989] (B3LYP) for all the quantum mechanical (QM) calculations in this work. DFT, and hybrid methods in particular, have gained significant popularity owing to the indirect accounting for electron correlation effects and the lower-scale dependence on the size of the studied system [Sousa *et al.*, 2007], therefore holding a great promise for

biomolecular research. DFT-based QM calculations are methods of choice where multi-configurational self-consistent field (MCSCF) approaches are not applicable, and usually provide results that are more precise than the standard *ab initio* Hartree-Fock calculations.

The B3LYP is one of the most accurate model chemistries with an expanding record of applications to a wide range of problems, including NMR parameter calculations [Barone, 1995]. A good agreement between the B3LYP and *ab initio* MP2 (correlated second order Møller-Plesset perturbation) levels of theory is noted in correctly describing the energies and vibrational frequencies of peptide models [Jalkanen *et al.*, 2004]. Furthermore, examples of an excellent reproduction of nuclear shielding constants [Le & Oldfield, 1996; Xu & Case, 2002] additionally motivate the choice of the B3LYP hybrid functional in this study.

The split-valence 6-311G(d,p) basis set [Krishnan *et al.*, 1980] is used for geometry optimisation, and the TZVP basis set of Ahlrichs and coworkers [Schäfer *et al.*, 1994] for single point calculations. The latter basis set is specifically optimised for DFT methods and is recommended for evaluations of both nuclear shielding [Helgaker *et al.*, 1999] and indirect spin-spin coupling constants [Sahakyan *et al.*, 2008a], where TZVP, in combination with B3LYP model chemistry, outperforms the restricted active space (RAS) multi-configurational methods.

The non-specific environmental (solvent) effects are accounted via the IEF-PCM integral equation formulation of the polarisable continuum model [Cancès *et al.*, 1997], where the solvent is modelled as a continuum with a uniform dielectric constant, and, the solute molecule is situated in a cavity constructed by the default atomic radii from the united atom topological model.

Gauge-invariant atomic orbitals (GIAO) [Ditchfield, 1974; Wolinski *et al.*, 1990] are used for the calculation of NMR nuclear shielding constants. All the QM calculations are done using the Gaussian 03 suite of programs [M. J. Frisch et al, 2004].

1.7.3 Calculation of nuclear shielding constants as a function of dielectric permittivity of media

For a systematic investigation of the chemical shift versus dielectric constant dependence, representative conformations for the model structure (Figure 1.2) are selected corresponding to α-helical, β-parallel, β-antiparallel and collagen structures as represented in Figure 1.3.

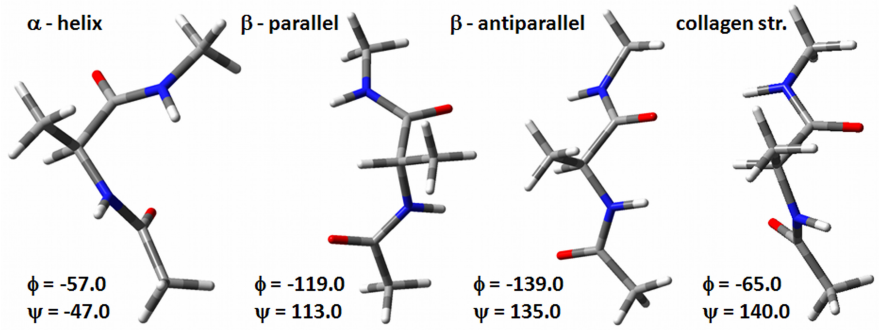


Figure 1.3: *The gas-phase optimised structures of the selected representative conformations of Ace-Ala-Nme with the fixed ϕ and ψ angles and the corresponding secondary structure types indicated on the figure.*

The structures are geometry optimised with the B3LYP/6-311G(d,p) level of theory. All the geometric parameters, except ϕ and ψ angles, are optimised and stationary geometries are found. The influence of dielectric permittivity on nuclear shielding constants is studied via the followed single point GIAO B3LYP/TZVP calculations with IEFPCM solvation model on the obtained geometries. The profiles of the dielectric constant, ϵ, dependence of the nuclei used in biomolecular NMR are obtained by varying the ϵ of solvent from 1.5 to 80 with 0.5-10 step size, as we move from lower to higher dielectric constants. The data points for the lowest dielectric constant, 1, was obtained from gas-phase GIAO B3LYP/TZVP calculations.

Although the subsequent discussion will mainly use the calculated absolute nuclear shielding constants and their changes, the values can easily be referenced to obtain chemical shifts, as the absolute nuclear shielding constants for the reference compounds (water for ^{17}O, ammonia for ^{15}N, and tetramethylsilane for

^{1}H and ^{13}C) are also computed. The resulting values for the reference absolute nuclear shielding constants are (in ppm) 31.92 for ^{1}H, 184.30 for ^{13}C, 264.70 for ^{15}N, and 331.04 for ^{17}O. The referenced values from the calculations are close to experimental chemical shifts, further demonstrating the suitability of the chosen scheme for calculations.

1.7.4 Calculation of nuclear shielding surfaces over ϕ and ψ dihedral angles at different ϵ dielectric constants

GIAO B3LYP/TZVP calculations with IEFPCM solvation are done varying the ϕ and ψ angles, so as to sample all the combinations from -180^o to 180^o with 12^o resolution for each of ϕ and ψ angles. The B3LYP/6-311(d,p) optimised geometry of the α-helical conformation of *Ace-Ala-Nme* is used with only the backbone dihedral angles varied without further geometry optimisation.

The calculations are carried out at two different dielectric constants resembling water ($\epsilon = 78.39$) and protein interior ($\epsilon = 4$). Overall, 1922 calculations ($2 \times 31 \times 31$) are performed, of which the ones for the ϕ/ψ combination around the centre of the Ramachandran plot are discarded (white areas in Figure 1.6 and Appendix A), since the united atom topological model fails to assign atomic radii to the hydrogen atom found to be close to more than one heavy atoms in the molecule.

1.7.5 The effects of dielectric permittivity on nuclear shielding constants of biomolecular importance

Despite the long history of studies on reduced peptide models, a systematic theoretical investigation of solvent dependence of the calculated spectroscopic parameters has not been done with coverage for NMR parameters. The conformational propensities of *Ace-Ala-Nme* and solvent effects on its potential energy landscape has been explored relatively recently, where a good representation of the experimentally observed trends was noted by *ab initio* MP2 [Wang & Duan, 2004] level of theory.

The coupling between the conformation and environmental effects on nuclear

shielding constants is investigated. For modelling the environmental (solvent) effects, Tomasi's polarisable continuum model (PCM or an extended version IEFPCM) [Cancès *et al.*, 1997] was proven to be of reasonable accuracy. The continuum models are suitable for modelling the bulk solvent effects, as they account only the non-specific solute-solvent interactions. Therefore if specific and long-lived solute-solvent interactions, such as hydrogen bonding, exist, the PCM description of the solvent effects will normally fail to provide observations consistent with experiment [Pecul & Sadlej, 1998]. However, for the ^{17}O nucleus, which is known for its large chemical shift deviation from one solvent to another, it was shown that the best result for polar solvents can be achieved by the inclusion of both continuum model, to account the long-range interactions, and a few explicit solvent molecules, to correctly describe the specific interactions if present [Cossi & Crescenzi, 2004]. For aprotic solvents, the continuum methods alone still perform very well. Moreover, in the chosen *Ace-Ala-Nme* molecule, a reliable description of solvent polarisation effects by implicit PCM solvation is shown [Wang & Duan, 2004].

The theoretical results in this work present a systematic investigation of non-specific environmental effects, or specifically, dielectric permittivity effects, on nuclear shielding constants of all the protein backbone nuclei used in biomolecular NMR studies or holding a great promise for such studies (^{17}O NMR [Zhu *et al.*, 2010]). It is also a more complete representation of the nuclear shielding behaviour over the Ramachandran space for the backbone nuclei other than $^{13}C^{\alpha}$ and $^{13}C^{\beta}$, which have been extensively studied by Oldfield and coworkers [de Dios *et al.*, 1993; Havlin *et al.*, 1997; Oldfield, 2002].

At first, let us examine the results from the ϵ-dependence of the distinct representatives of different protein secondary structures (Figure 1.3) modelled via *Ace-Ala-Nme* (Figure 1.4).

As can be seen from Figure 1.4, all types of nuclear shielding constants show a characteristic dependence on ϵ with abrupt change at lower dielectric constants and gradual saturation after $\epsilon \approx 20$. A similar type of dependence was noted before, while studying nitrogen nuclear shielding constants in small organic molecules with COSMO (conductor-like screening model) [Ksiazek *et al.*, 2009] and PCM [Zahedi *et al.*, 2009] implicit solvation models, as well as for

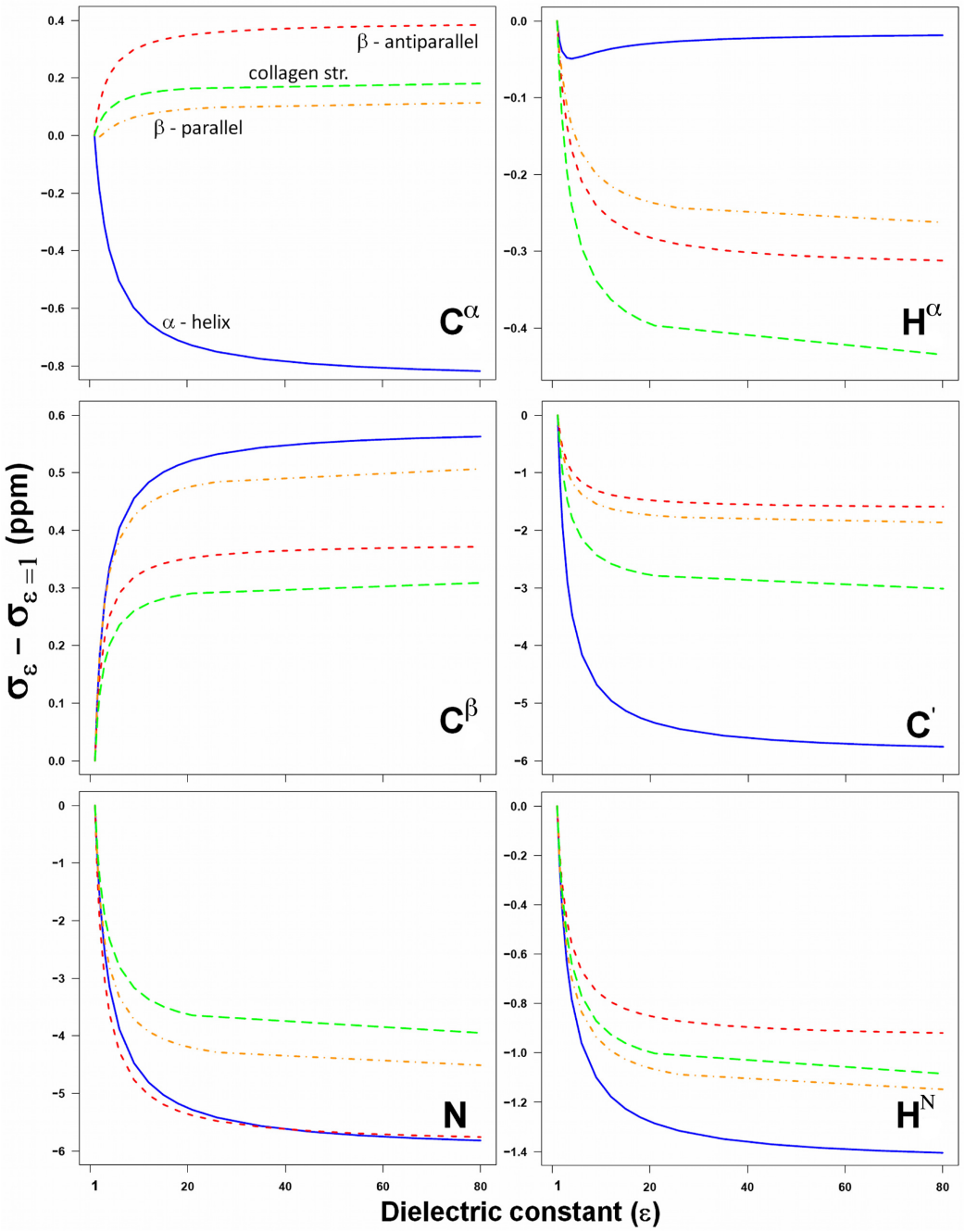


Figure 1.4: *The changes in nuclear shielding constants (in ppm) of the backbone nuclei relevant to biomolecular NMR against the dielectric constant of the medium (from 1 to 80). The blue, green, red and orange colours indicate the data from α-helix, collagen, β-antiparallel and β-parallel structures respectively.*

studies on solvent dependence of NMR one-bond J-coupling constants [Sahakyan *et al.*, 2008a]. Such dependence is also reflected in numerous experimental measurements of chemical shifts in small molecules dissolved in different solvents or binary solvent mixtures with varying composition (see for example [Becconsall & Hampson, 1965; Senthilnathan & Singh, 1974; Takayama *et al.*, 1989]).

The found shape of dependence, which holds true for all the backbone nuclei of interest, can be rather influential on the chemical shift values inside proteins. In particular, at the protein interior, where the effective dielectric constant is rather low, chemical shifts will be especially sensitive to small changes in dielectric permittivity. This increased sensitivity can contribute to the variation of chemical shifts inside proteins, and, taking into account the substantial magnitude of the observed changes, the neglect of the effective ϵ-dependence might contribute errors in the performance of empirical chemical shift predictors. Previously, an attempt was done to consider the buried and solvent-exposed residues separately for the development of the recent side-chain chemical shift predictors [Sahakyan *et al.*, 2011a,b], however, the current size of the databases does not allow obtaining reasonable prediction model based on split data. With the continuous growth of the number of publicly available chemical shift measurements, such separate consideration that would account for solvent exposure, as well as different secondary structure, types could be one of the immediate steps to try for the improvement of the accuracy of chemical shift predictors.

Table 1.1 presents the chemical shift differences from idealised water (protein surface) to protein interior and vacuum conditions to show the magnitude of the dielectric permittivity effects on the backbone chemical shifts. Please note, that the difference in chemical shift is the negative of the difference in nuclear shielding. Both Figure 1.4 and Table 1.1 demonstrate changes in nuclear shielding sensitivities towards ϵ, when different secondary structures are considered. In all the examined nuclei, except $^{1}H^{\alpha}$, the α-helix conformation is more sensitive with its dependence being even reverse for the $^{13}C^{\alpha}$ nucleus (see Figure 1.4). To explain the conformational dependence of the $\sigma(\epsilon)$ function, let us consider a molecule situated in a solvent or a given environment with dielectric constant ϵ. A solute molecule inside the given environment polarises and/or reorients the neighbouring molecules in vicinity, which gives rise to, so called, solvent reaction field, F_{RF},

Table 1.1: *The chemical shift difference for the studied nuclei of Ace-Ala-Nme in different secondary structures, when comparing the results from $\epsilon = 80$ (water) to $\epsilon = 1$ (vacuum) and $\epsilon = 4$ (protein) dielectric constants.*

	$\delta_{\epsilon=80} - \delta_{\epsilon=1}(ppm)$				$\delta_{\epsilon=80} - \delta_{\epsilon=4}(ppm)$			
Nucl.	α-helix	β-paral.	β-antipar.	collagen	α-helix	β-paral.	β-antipar.	collagen
^{15}N	5.819	4.512	5.758	3.950	2.662	1.694	2.132	1.617
^{1}H^{N}	1.405	1.148	0.920	1.085	0.621	0.449	0.361	0.446
^{1}H$^{\alpha}$	0.019	0.262	0.312	0.435	-0.031	0.125	0.144	0.194
^{13}C$^{\alpha}$	0.818	-0.114	-0.384	-0.181	0.424	-0.091	-0.174	-0.088
^{13}C$^{\beta}$	-0.563	-0.507	-0.372	-0.309	-0.226	-0.179	-0.120	-0.107
^{13}C'	5.758	1.863	1.590	3.013	2.269	0.688	0.595	1.212
^{17}O	-50.676	-37.958	-38.067	-38.509	-19.702	-13.686	-13.844	-14.608

an electric field along the direction of the solute dipole moment. The magnitude of the reaction field, in its simplest case, can be determined using the Onsager model [Onsager, 1936], where the solute molecule is simplified as a dipole μ inside a spherical cavity with a radius r (Equation 1.10).

$$F_{RF} = \frac{\mu}{4\pi\varepsilon_0 r^3}\frac{2\varepsilon - 2}{2\varepsilon + 1} \tag{1.10}$$

The ϵ_0 in the equation above is the dielectric permittivity of the vacuum. A linear interdependence of nuclear shielding at the atom X, which is only connected to the atom Y, and electric field projection along the X-Y bond, $F_{\parallel}$, is noted by Buckingham [Buckingham, 1960], and can be expressed by the following expression:

$$\Delta\sigma = -aF_{\parallel} - bF_{\parallel}^2 \tag{1.11}$$

where a and b are parameters that depend on the bond type and describe the electronic polarisability and hyperpolarisability respectively, against the applied electric field. The second term in the equation is negligible in most cases, resulting in a linearity of the outlined dependence. Therefore the major environmental factor which actually affects the nuclear shielding is the $(2\epsilon - 2)/(2\epsilon + 1)$ ratio. The slope factor (strength) of the σ versus $(2\epsilon - 2)/(2\epsilon + 1)$ linear dependence is determined by the angle between the polarisability axis of the bond involving the studied nucleus and the solvent reaction field vector, F_{RF}, or the dipole moment vector of the solute, μ. In fact, the dipole moment in α-helix conformation is nearly parallel to the N-H and C=O bonds, which clarifies the stronger sensitivity of the ^{15}N, ^{1}H$^{\alpha}$ and ^{13}C' nuclear shielding constants to the dielectric permittivity of the medium.

Further extension of the Onsager model, to account for solute-solvent quadrupolar interaction, results in the $(6\epsilon - 6)/(3\epsilon + 2)$ multiplier that defines the solvent reaction field [van Pelt *et al.*, 1981]. Figure 1.5 demonstrates the interrelation between the nuclear shielding constants and the extended $(6\epsilon - 6)/(3\epsilon + 2)$ term for the studied nuclei.

Indeed, the dependence becomes close to linear with deviation at higher values of dielectric constant. That violation of the linear dependence can be explained

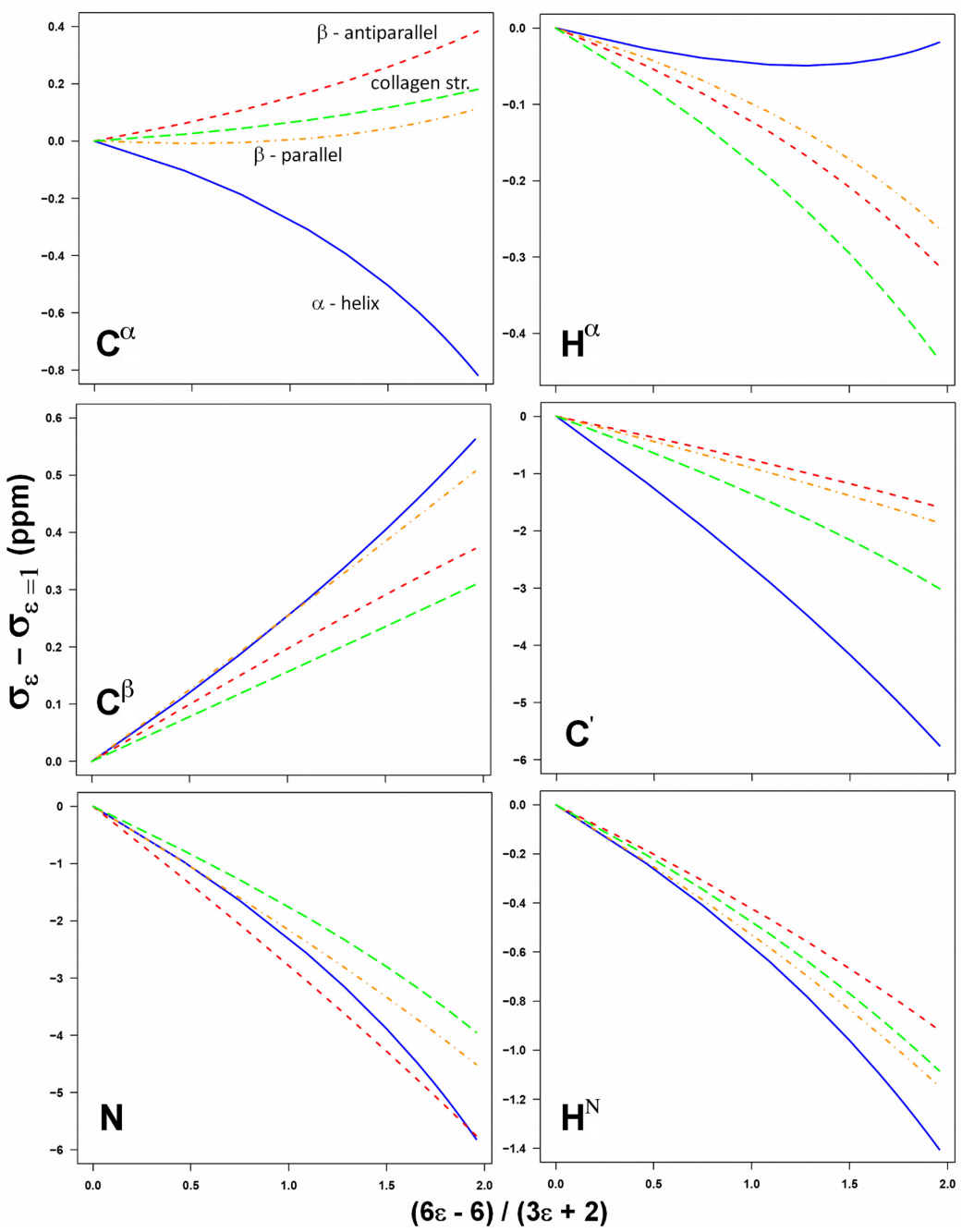


Figure 1.5: *The changes in nuclear shielding constants (in ppm) of the backbone nuclei used in biomolecular NMR studied against the* $(6\epsilon - 6)/(3\epsilon + 2)$ *function of the dielectric constant of the medium (with* ϵ *varying from 1 to 80). The blue, green, red and orange colours indicate the data from* α*-helix, collagen,* β*-antiparallel and* β*-parallel structures respectively.*

by more significant changes in the electron density of the solute molecule in a highly polarising environment, capable of affecting the local interaction terms in the Equation 1.2.

It should be noted that all the calculations in this work evaluate the direct solvent or environmental effects on nuclear shielding constants. However, the geometries of molecules also change depending on the dielectric constant of the surrounding medium. In particular, a gradual elongation of the exposed N-H and C=O groups and shortening of the buried peptide bonds were observed [Wang & Duan, 2004]. This gives rise to indirect solvent effects on the nuclear shielding, mediated by geometry changes. The bond length and angular dependencies of the $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ nuclei are well studied [de Dios *et al.*, 1993] and proven to be significant. However, as the structural difference from gas phase to $\epsilon = 80$ is very small, with bond length difference being less than 0.01 Å, the indirect contribution to the nuclear shielding constants is usually less than 10 % [Sahakyan *et al.*, 2008a,b; Zahedi *et al.*, 2009].

1.7.6 Dielectric permittivity in proteins: an evaluation based on chemical shifts

Dielectric permittivity is a macroscopic parameter that describes how the electric fields are affected by a given dielectric medium, and, is determined by the ability of a material to polarise in response to the field. It plays an active role in modulating a variety of molecular processes, because of the importance of long-range electrostatic interactions. Directly scaling the electrostatic interactions, dielectric permittivity is particularly important in proteins and has important contributions to the processes ranging from enzyme catalysis [Warshel, 2003] to signal transduction and molecular recognition [Biot *et al.*, 2003; Varma & Jakobsson, 2004]. Although the physical basis behind the dielectric permittivity is rather clear and the implications are obvious in modelling solvents and small molecules [Scaife, 1989], for proteins and other macromolecular objects, the value of this physical quantity is still a matter of debate.

Initial estimations for dielectric permittivity in proteins have come from the measurements in dry protein powders and resulted in dielectric constant values

ranging from 2 to 5 [Pethig, 1979]. However, further investigations have clearly demonstrated that for proteins, as polyelectrolytes with certain structure and dynamics, the dielectric constant and its interpretation is highly affected by the surrounding solvent composition and the examined timescale of protein dynamics. In particular, the measurements of the ionisable and buried amino acid side-chain pK_a values, which have been considered as indirect detectors of dielectric permittivity, implies an increased dielectric constant inside proteins [Fitch *et al.*, 2002; Garcia-Moreno *et al.*, 1997]. The main computational techniques used for the electrostatics calculations on protein systems are based on the Poisson-Boltzmann equation, where the dielectric constants of both solvent and protein are *a priori* assumed [Honig & Nicholls, 1995]. By varying the dielectric constant of the protein interior to get the best fit between the estimated and experimental pK_a values, those calculations also suggest the view that the polarisation of the buried groups are underestimated, and result in dielectric constant evaluations of 10-40, sometimes reaching up to 60 depending on the studied protein [Schutz & Warshel, 2001; Sharp, 1998]. However, the estimations were also shown to be highly model-dependent [Schutz & Warshel, 2001].

Computer simulations of protein dielectric constant have been done using the interrelation between the $\mathbf{D}$ total dipole moment of the system and its ϵ dielectric constant, as depicted in the Fröhlich-Kirkwood model [Fröhlich, 1958; Kirkwood, 1937]. In particular, the dielectric constant of the system is represented as a function of probability distribution of the total dipole moment via the following equation:

$$\frac{\langle \mathbf{D}^2 \rangle - \langle \mathbf{D} \rangle^2}{3\varepsilon_0 V k_B T} = \frac{(2\varepsilon_{env} + 1)(\varepsilon - 1)}{(2\varepsilon_{env} + \varepsilon)} \tag{1.12}$$

where ϵ_0 is the dielectric permittivity of vacuum, ϵ_{env} is the external dielectric constant, V is the solute volume and T is the simulation temperature. The mentioned evaluations via molecular dynamics (MD) trajectories of different proteins show that charged residues and their dynamics play primary roles in increasing the internal dielectric constant of proteins [Pitera *et al.*, 2001; Raha & Merz, 2007]. Although those evaluations resulted in values from 10 to 41 for the dielectric constant at the protein interior, estimations excluding the charged residues

imply ϵ to be between 2 and 4 [Pitera *et al.*, 2001]. Substantial differences in evaluations are noted, depending on pH, solvents and temperature via affecting the ionisation and mobility of the charged side-chains. However, all the studies so far have included the solvent molecules and counter-ions in the simulations only, with **D** being evaluated solely based on protein atoms. Thus, outlining the primary importance of the charged residues, those studies ignore possible screening and neutralisation of the protein charges by solvent dipoles and counter ions. Hence, a better estimation which will most probably result in a lower value for the interior dielectric constant should be expected from the studies where the first solvation shell is involved in the calculations of dipole moment distribution. Another important observation from the study with an MD approach but partial charges evaluated from quantum mechanics (QM) calculations is the strong dependence of the dielectric constant estimation on the studied region in proteins [Raha & Merz, 2007]. Even for the proteins with buried charged residues, the calculations, where only the charges of the interior regions are included, show that the dielectric constant abruptly drops down from as high as 80 to as low as 1-5 if we approach the core region.

The lower dielectric permittivity assumption has been further revived by a study based on the backbone amide hydrogen exchange rates, which are consistent with the Poisson-Boltzmann evaluations when the dielectric constant of 6 is used [LeMaster *et al.*, 2007]. The backbone anions have much shorter lifetime, which facilitates the report of the dielectric shielding from a sub-nanosecond snapshot of the studied protein.

Thus, all the studies while diverging in their evaluations for the protein dielectric constant, converge on the idea that the parameter is highly dynamic and depends on many factors. However, myriads of computational models still use an empirical and *a priori* assumption of the uniformity of ϵ for protein electrostatics modelling, therefore increasing the representativeness and consistence of that single value with the experimental observables is still of utmost relevance.

The ϵ-dependence profiles of chemical shifts obtained in this study can be used to evaluate the dielectric constant inside proteins. Averaging the data from 1010 proteins with 103084 residues, Avbelj and coworkers found approximately 0.4 ppm difference between the $^{13}C^{\alpha}$ chemical shifts in solvent exposed and buried

residues of α-helical structure [Avbelj *et al.*, 2004]. We can approximate, that the residues with solvent accessibility are exposed to the environment with $\epsilon \approx 80$. Considering that all the specific interactions are averaged out over the 103084 residues and assuming that statistically the average change in chemical shift reflects only the influence of dielectric permittivity, a crude estimation of about 4-5 can be inferred from the $\sigma(\epsilon)$ function for the $^{13}C^{\alpha}$ nucleus plotted in Figure 1.4. This is not the first evaluation of dielectric constant at the protein interior by using the chemical shift data. In a recent work [Hass *et al.*, 2008], the authors used the nuclear shielding polarisabilities of the backbone $^{1}H^{N}$ and ^{15}N calculated on N-methyl acetamide, to estimate the local electric field at the N-H backbone moiety from the measured chemical shift perturbations. Using the field evaluations from a point charge model and from chemical shift perturbations, the authors calculated the effective dielectric constant as a factor that scales the electric field. The resulting 2.7-3.2 values support the assumption of the low dielectric constant at the protein interior once more.

1.7.7 Nuclear shielding surfaces over ϕ and ψ dihedral angles at different ϵ dielectric constants

After examining the behaviour of the $\sigma(\epsilon)$ function for representative secondary structures, the next step is the investigation of nuclear shielding $\sigma(\phi, \psi)$ surfaces, which contain the complete set of *Ace-Ala-Nme* backbone conformations. Similar studies have been done before, prevalently on $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ nuclei (see [Oldfield, 2002] and references therein). However, besides providing results for all the backbone nuclei used in conventional biomolecular NMR, the current study performs nuclear shielding calculations at two, $\epsilon = 4$ and $\epsilon = 78.39$, dielectric constants. Moreover, a new $\Delta\sigma^{w-p}(\phi, \psi)$ dependence is investigated (Figure 1.6, Appendix A), where $\Delta\sigma^{w-p}$ is the difference between the nuclear shielding constants in water ($\epsilon = 78.39$), w, and protein interior ($\epsilon = 4$), p.

Taking into account the similar shape of $\sigma(\epsilon)$ functions with saturating behaviour at higher dielectric constants for all the studied nuclei (Figure 1.4), the $\Delta\sigma^{w-p}$ difference can be considered as a measure of sensitivity to the non-specific environmental effects. Therefore $\Delta\sigma^{w-p}(\phi, \psi)$ surfaces describe the conforma-

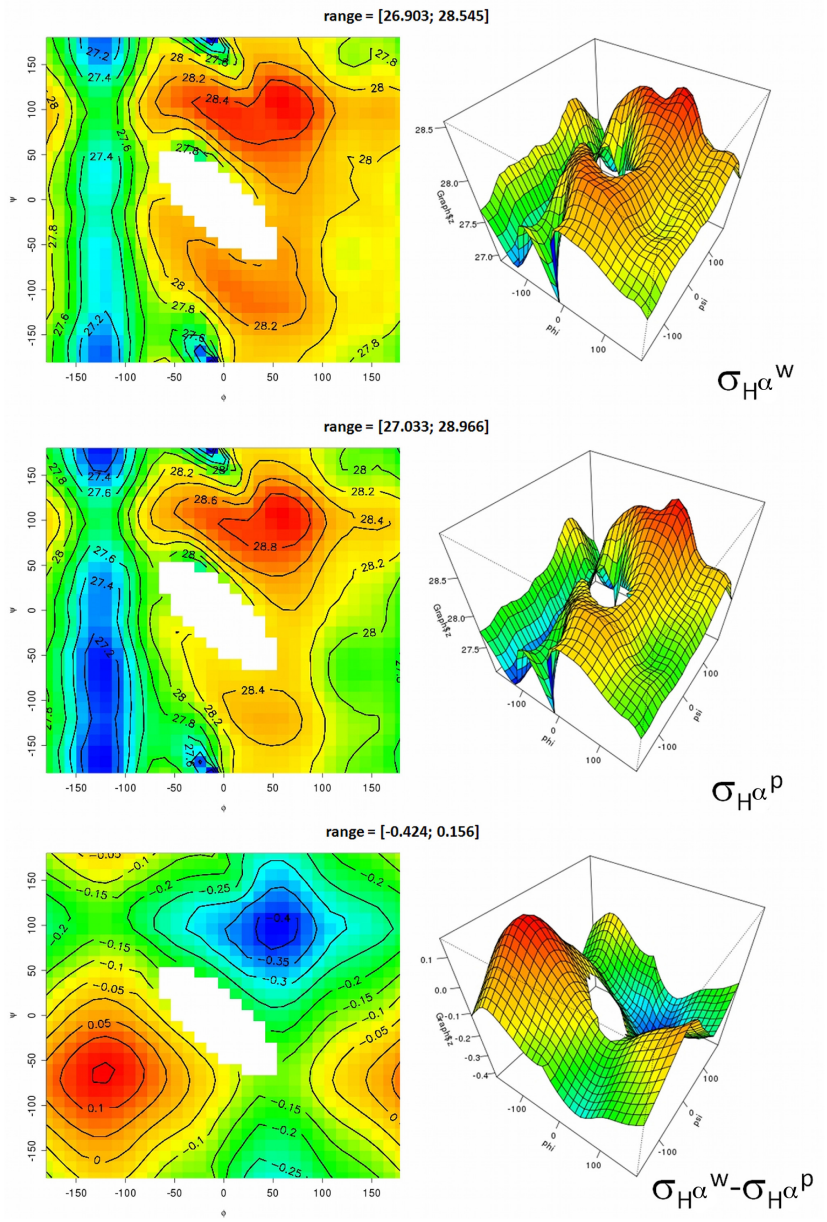


Figure 1.6: *The projected and 3-dimensional representation of the* $^1H^\alpha$ *nuclear shielding surfaces over* ϕ/ψ *dihedral angles in Ace-Ala-Nme molecule. The calculations are done in* $\epsilon = 78.39$ *(water, w, top) and* $\epsilon = 4$ *(protein interior, p, middle) conditions. The surface at the bottom shows the difference in nuclear shielding constants from water to protein interior across the Ramachandran space. Similar results for the other nuclei are presented in Appendix A.*

tional dependence of such sensitivities. The obtained surfaces are in a very good agreement with the previous experimental measurements. In particular, the known relation between the $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ shifts and the backbone torsion angles [Case, 1998] is clearly visible from the $\sigma_{^{13}C^{\alpha}}(\phi,\psi)$ and $\sigma_{^{13}C^{\beta}}(\phi,\psi)$ surfaces. The $\Delta\sigma^{w-p}$ difference is equal to the $\Delta\delta^{p-w}$, in terms of chemical shifts, and is directly comparable with the $\delta^{protein}-\delta^{rand.coil}$ secondary shifts. The comparisons show a good agreement with the observations from real proteins [Avbelj *et al.*, 2004] and can be used in future for a better parametrization of the existing empirical chemical shift predictors to account for non-specific solvent interactions. It is interesting to explicitly outline the highly symmetrical nature of the $\Delta\sigma^{w-p}_{^{1}H^{\alpha}}(\phi,\psi)$ surface (Figure 1.6), which is also in an inversely proportional agreement with $\Delta\sigma^{w-p}_{^{13}C^{\alpha}}(\phi,\psi)$ (Appendix A). The latter inverse dependence explains the published inversely proportional linear dependence of $^{1}H^{\alpha}$ and $^{13}C^{\alpha}$ chemical shifts, regardless of the solvent exposure of the involved residues, which is not the case for other nuclei [Vranken & Rieping, 2009]. Furthermore, the previously observed 4-5 ppm increase in $^{13}C^{\alpha}$ shielding constants of β-sheet fragments over α-helical ones (see [Havlin *et al.*, 1997] and references therein) is very well reproduced in the $\sigma^{w}_{^{13}C^{\alpha}}(\phi,\psi)$ and $\sigma^{p}_{^{13}C^{\alpha}}(\phi,\psi)$ surfaces, where the values rise from below 128 to over 132 ppm.

1.7.8 Conclusions

The major outcome of this study is the obtained universal behaviour of chemical shift versus dielectric permittivity dependence for protein backbone atoms, with large changes at the lower dielectric constant media and levelling off at the higher values. Such dependence holds true for all the nuclei of interest in biomolecular NMR, and becomes linear when solvent reaction field is considered instead of the dielectric permittivity. The magnitude of the dependence is shown to be highly dependent on the backbone conformation, to which a reasonable explanation is outlined, based on the solvent induced electric fields. The special danger of the observed shape of dependence for the chemical shift evaluations for the nuclei at protein interior is emphasised. Thoroughly screening the nuclear shielding sensitivity towards the dielectric permittivity over all the ϕ/ψ combinations, we

now have a better view on the coupling between the non-specific environmental/solvent effects on chemical shifts and the backbone conformation. Combining the obtained profiles for ϵ-dependence of nuclear shielding constants in different secondary structures and the observed average change in backbone $^1H^\alpha$ chemical shifts of solvent exposed and buried α-helical structures [Avbelj *et al.*, 2004], this work suggests an effective dielectric constant of $\approx$ 4-5 for protein interior, thus increasing the weight of the low-ϵ hypothesis for the dielectric permittivity evaluations inside proteins.

Perfection is achieved not when there is nothing more to add, but when there is nothing left to take away.

Antoine de Saint-Exupery

2

Chemical Shifts of Protein Side-Chain Methyl Groups

2.1 Summary

Protein methyl groups have recently been the subject of much attention in NMR spectroscopy because of the opportunities that they provide in obtaining information about the structure and dynamics of proteins and protein complexes. With the advent of selective labelling schemes, methyl groups are particularly interesting in the context of chemical shift based protein structure determination, an approach that to date has exploited primarily the mapping between protein structures and backbone chemical shifts. This chapter describes the development of CH3Shift method of performing structure-based predictions of methyl chemical shifts. The terms considered in the predictions take account of ring current, magnetic anisotropy, electric field, rotameric type, and dihedral angle effects, which are considered in conjunction with polynomial functions of interatomic distances.

The CH3Shift method achieves an accuracy in the predictions that ranges from 0.133 to 0.198 ppm for ^{1}H chemical shifts for Ala, Thr, Val, Leu and Ile methyl groups. The use of the CH3Shift method is illustrated by assessing the accuracy of side-chain structures in structural ensembles representing the dynamics of proteins.

2.2 Motivation

Despite the fact that chemical shifts are the most readily and accurately measurable observables in protein NMR spectroscopy, their complex dependence on a myriad of molecular and environmental factors [Jameson, 1996; Oldfield, 1995] has represented a major obstacle for their direct use in protein structure determination. Recent advances in experimental and computational techniques, however, are starting to make it possible to use them to obtain structures of proteins [Cavalli *et al.*, 2007; Korzhnev *et al.*, 2010; Raman *et al.*, 2010; Shen & et al, 2008] and protein complexes [Das *et al.*, 2009; Montalvao *et al.*, 2008], both in solution and in the solid states [Robustelli *et al.*, 2008; Shen *et al.*, 2009]. As the protocols that have been introduced so far for using chemical shifts in structure determination [Cavalli *et al.*, 2007; Shen & et al, 2008; Wishart, 2011] require the ability of predicting them based on protein structures, a number of methods for performing such predictions have been developed in the last several years [Kohlhoff *et al.*, 2009; Lehtivarjo *et al.*, 2009; Meiler, 2003; Neal *et al.*, 2003; Shen & Bax, 2007; Wishart *et al.*, 1997; Xu & Case, 2001]. Although these methods have so far been mainly concerned with backbone chemical shifts, further progress can be expected in establishing fully reliable methods for protein structure determination using side-chain chemical shifts as well. This idea has been supported by a series of recent studies that reported quantitative relationships between the rotameric states of side-chain methyl groups and the corresponding chemical shift values [Hansen *et al.*, 2010; Mulder, 2009]. These developments are particularly interesting since proteins are rich in methyl-bearing amino acids and therefore methyl chemical shifts provide excellent opportunities to probe their structures and dynamics [Baldwin *et al.*, 2010; Gelis *et al.*, 2007; Hsu *et al.*, 2009; Sheppard *et al.*, 2010; Tugarinov *et al.*, 2005b]. Furthermore, optimised NMR experiments

to measure chemical shifts and new schemes for efficient and highly-specific isotope labelling of side-chain methyl groups [Goto & Kay, 2000; Kainosho *et al.*, 2006; Otten *et al.*, 2010; Tugarinov *et al.*, 2006] are enabling their use to characterise the structure and dynamics of large protein complexes, and are making methyl chemical shifts an ever-growing component in the Biological Magnetic Resonance Data Bank (BMRB) [Ulrich, 2007].

2.3 Structure-based prediction of methyl chemical shifts

Most of the current state-of-the-art methods for performing structure-based predictions of chemical shifts [Kohlhoff *et al.*, 2009; Lehtivarjo *et al.*, 2009; Meiler, 2003; Neal *et al.*, 2003; Shen & Bax, 2007; Wishart *et al.*, 1997; Xu & Case, 2001] are based on the use of a combination of many factors [Jameson, 1996], including ring current [Haigh & Mallion, 1972, 1980], magnetic anisotropy [McConnell, 1957] and electric field [Buckingham, 1960; Buckingham & Pople, 1963] effects. In addition, it has also been shown recently that predictions of similar accuracy can be obtained by expressions that capture the relationship between structures and chemical shifts by writing formally the chemical shifts as polynomial functions of atomic coordinates [Kohlhoff *et al.*, 2009]. Although this approach provides less insight into the physical effects that determine the chemical shifts, it has the advantage of being computationally efficient and of generating structural restraints to be used in molecular dynamics simulations because the polynomial functions that give the chemical shifts are readily calculable and differentiable.

To enable the usage of structure-based chemical shift predictions for protein methyl groups, in this work the CH3Shift method is introduced, which expresses the chemical shift δ of a given nucleus as a combination of phenomenological terms and distance-based terms, that are further optimised via the developed automatic and robust technique. Analogous to the Equation 1.2, chemical shifts can be expressed via the sum of the following terms:

$$\delta = \delta^{rc}_{rot} + \Delta\delta_{dih} + \Delta\delta_{ring} + \Delta\delta_{ma} + \Delta\delta_{EF} + \Delta\delta_{dist} \tag{2.1}$$

where δ^{rc}_{rot}, $\Delta\delta_{dih}$, $\Delta\delta_{ring}$, $\Delta\delta_{ma}$, $\Delta\delta_{EF}$ and $\Delta\delta_{dist}$ are, the rotameric, dihedral, ring current, magnetic anisotropy, electric field and the distance-based contributions respectively. For fitting the parameters against these various terms, a database of experimental methyl chemical shifts (kindly provided by Dr. Wim F. Vranken) and corresponding high-resolution X-ray structures are used. For defining the distance-based terms, atoms in the region between a smaller sphere of 1.8 Å radius and a larger sphere of 6.5 Å radius are considered around each of the methyl groups, centred on the methyl carbon nucleus (Figure 2.1). The smaller sphere includes the methyl group itself and the preceding carbon or sulphur (for methionine) atoms, and the arrangement within that region can be considered constant regardless of the structural environment and the side-chain conformation. The 6.5 Å cutoff radius is rather safe for capturing all the effects,

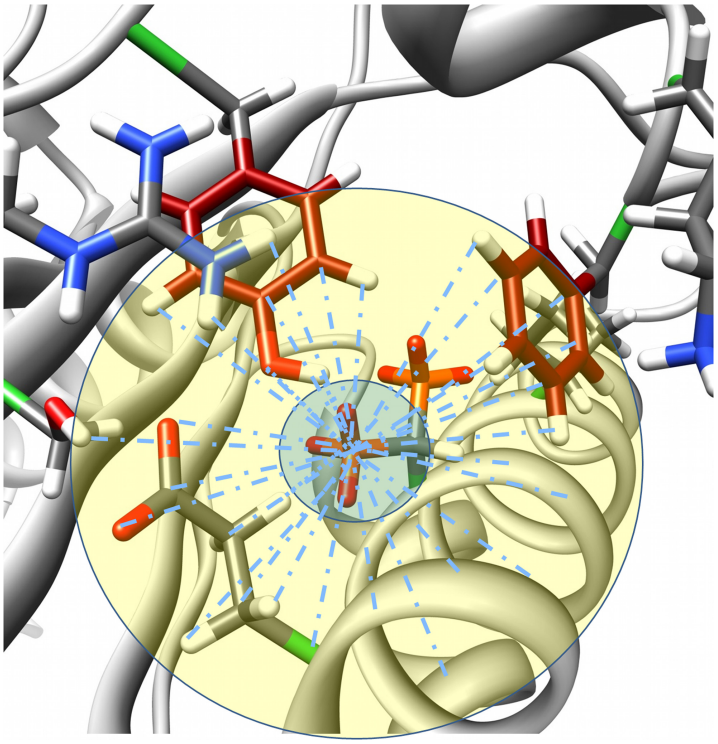


Figure 2.1: *Illustration of a methyl bearing side-chain with a representation of the active (yellow) and neutral (blue) regions defined by 6.5 and 1.8 Å cutoff radii from the methyl carbon nucleus. Some of the side-chains having significant contributions to the methyl group chemical shifts are explicitly indicated.*

since the most significant, ring current, effect is shown to become negligible at distances longer than 5.5 Å [Case, 1995]. The weaker electric field effect on chemical shifts is also rapidly decaying over the distance, and, taking into account the extremely small chemical shift polarisability coefficients, can be safely ignored at distances longer than 6.5 Å.

2.4 Database analysis and filtering criteria

In order to parametrize the CH3Shift method, the CH3Shift-DB database is constructed. The initial re-referenced extract of chemical shifts was created and kindly provided by Dr. Wim F. Vranken. The chemical shift information was retrieved from the BMRB [Ulrich, 2007] and converted into CCPN projects [Vranken & Rieping, 2009; Vranken *et al.*, 2005]. The referencing of the chemical shifts was corrected, when required, using VASCO [Rieping & Vranken, 2010], a method to correct and validate protein chemical shift values in relation to the coordinates of the corresponding nuclei.

Upon obtaining the re-referenced extract of chemical shifts, a number of additional filtering steps have been done. In particular, only the chemical shift entries with stereospecific assignment for Val and Leu residues are considered. Cases for which chemical shifts were flagged as stereospecifically assigned but the difference between the two methyl chemical shifts was zero, were discarded. When multiple BMRB records were present, the median of the chemical shift values were taken from all the entries corresponding to the same nuclei in the same protein. This type of averaging ensures that outlying data entries, which can be attributed to various types of artefacts that can arise in the experiments or in the spectra interpretation, have minimal impact on the final compilation of the data. Only the chemical shift entries corresponding to structures determined by X-ray crystallography were considered. Of the total 750 protein structures, each with a unique PDB (Protein Data Bank [Berman *et al.*, 2000]) identifier of an X-ray structure, 26 structures were discarded since they were related to protein-nucleic acid complexes; in this way we decrease the possibility of the chemical shift data being modulated by non-protein contacts and ring current effects. 43 other structures were discarded for containing porphyrinic rings, iron or cobalt

atoms, in order to filter out any non-standard ring current and paramagnetic effects. The above mentioned filtering criteria resulted in the removal of 1558 chemical shift entries out of the initial 19431. The compiled data set thus contained 17873 residue-specific chemical shift records, which are distributed over the amino acid residue types as 5965 Ala, 3147 Thr, 2243 Val, 2750 Leu, 3126 Ile, and 642 Met (Figure 2.2).

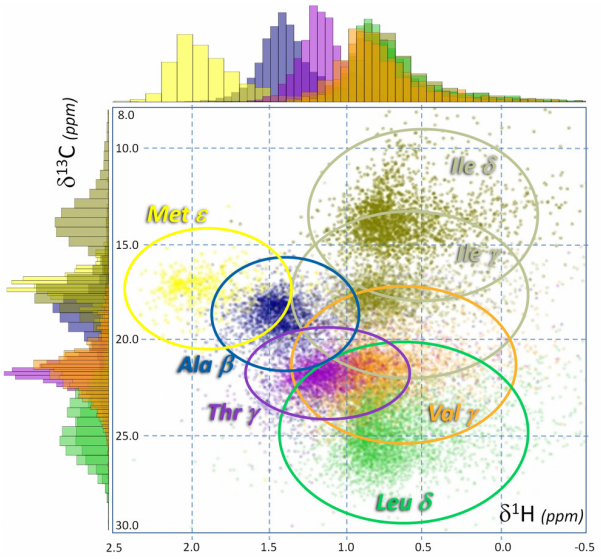


Figure 2.2: *HSQC-like correlation graph of the methyl group* ^{13}C *and* ^{1}H *chemical shift distributions in the CH3Shift-DB database, which shows the different chemical shift propensities for different types of residues. The circles indicate the substantial overlap between the chemical shifts of different methyl group types.*

The significant overlap in the methyl chemical shifts represents the main obstacle in the efficient assignment of the experimental spectra of the methyl group region. The representation in Figure 2.2 clearly illustrates the importance of the recent advances in the assignment of the NMR spectra, in particular for large protein complexes [Ruschak & Kay, 2010; Sheppard *et al.*, 2009; Sprangers & Kay, 2007; Xu *et al.*, 2009].

The crystallographic Rfree factor was not used in the filtering procedure because 125 of the 681 PDB files in the initial database did not include information on Rfree and the values that were available had an average of 0.243, first quartile

of 0.222 and third quartile of 0.266, indicating that there are only small variations in these values. It would therefore be difficult to use the Rfree value for protein structure selection. Also was left unused the information about sequence homology for filtering. For the development of chemical shift predictors, the inclusion of similar sequences (and structures) in the database is likely to be advantageous to some extent. Since chemical shift values are very sensitive to the local environment, small changes in homologous structures can result in relatively large differences in actual chemical shift values. However for completeness, the extent of sequence similarity is calculated between the PDB entries used for generating the CH3Shift-DB database using the PISCES server [Wang & Dunbrack, 2003] which generates a list of non-redundant PDB entries from an input list of PDB IDs. A total of 218 entries had a sequence identity of more than 25% with one of the entries in a non-redundant subset. Upon increasing the cutoff, the numbers were: 91 entries at 40%, 72 entries at 50%, 55 entries at 60%, 39 entries at 70%, 35 entries at 80% and 31 entries at 90% sequence identity; thus very similar sequences (more than 80%) only account for about 5% of the total number of entries.

The X-ray structures were preprocessed by the addition of hydrogen atoms followed by 1000 steps of hydrogen-only geometry optimisation, using the Almost all-atom molecular simulations toolkit (`http://www.open-almost.org`, accessed in April, 2010) and the Amber03 force field [Duan *et al.*, 2003]. Finally, the database was further optimised by considering only the chemical shifts falling within a window of 2.5 standard deviation for each specific nucleus and residue type, and for which an X-ray structure at 2.0 Å resolution or better was present. The removal of the most uncommon experimental chemical shift values was necessary to avoid the presence of the erroneous data or data from measurements in non-standard conditions. This procedure was also useful to avoid the complications associated with considering chemical shifts strongly affected by the close vicinity of aromatic rings or charged groups, which are highly sensitive to the dynamics and the exact geometric arrangement of the source nucleus and the strong affector moieties.

2.5 Rotameric terms

Since effects from the spatial neighbourhood and the conformation of the residue that holds the methyl group alter the chemical shifts of the methyl nuclei from the value determined by the covalently linked local environment, we can separate the neighbourhood-independent core component of the chemical shift from the rest. This was done for Ala by allowing the fitting procedure to generate an intercept along with the optimised parameters for the other factors discussed below. For the other residue types, the observation of significant differences between the average chemical shifts in different rotameric states (see Appendix B) suggested the possibility to also account for the rotamer-specific shifts through the intercept. Therefore, for the residue types with a side-chain χ_1 dihedral angle, the following expression is considered (Equation 2.2):

$$\delta_{rot}^{rc} = k_1 R_1 + k_2 R_2 + k_3 R_3 \tag{2.2}$$

where the R_1, R_2 and R_3 factors classify the rotameric state and are equal to 1 for $-120 < \chi_1 \leq 0$, $0 < \chi_1 \leq 120$ and $(120 < \chi_1 \leq 180) \cup (-180 \leq \chi_1 \leq -120)$ conditions for R_1, R_2 and R_3 correspondingly, with 0 values otherwise. The mentioned windows of χ_1 angle well separate the most common three χ_1-based rotameric states and allow treating different rotameric classes separately.

2.6 Dihedral angle terms

In these terms included are the backbone ϕ, ψ and all the available side-chain χ_i (with $i = 1, ..., 5$) dihedral angles. The effects from each of those angles (if present) were modelled via four polynomial and ten cosine terms (see Appendix C). The ten cosine terms were selected from the analysis of about hundred cosine, sine and mixed terms. All the geometric terms from the existing dihedral angles are calculated in the database structures. Further, a cross correlation matrix was created for the geometric terms along all the functions to identify the functions among the set that were correlating with each other. The Pearson correlation coefficient value of 0.7 was used to eliminate strongly correlated ones. The final ten functions were then chosen from the remaining ones according to their simplicity.

Different sets of functions were tried, but the results indicate that as long as there is a sufficiently large number of geometric terms that are not strongly correlated (in this case ten cosine functions and four polynomials), the fitting procedure for the coefficient optimisation finds values for the coefficients resulting in models of comparable performance.

2.7 Ring current terms

Ring current effects on chemical shifts arising from the aromatic rings of Phe, Tyr, His, Trp-5 and Trp-6 (5- and 6-membered tryptophan rings) are accounted by the inclusion of $G(\overrightarrow{\mathbf{r}})$ geometric factors from the model by Haigh and Mallion [Haigh & Mallion, 1972, 1980] (Equation 2.3):

$$\Delta\delta_{ring} = k_{ring} G(\overrightarrow{\mathbf{r}}) = k_{ring} \sum_{ij} S_{ij} \left(\frac{1}{r_i^3} + \frac{1}{r_j^3} \right) \tag{2.3}$$

where S_{ij} is the algebraic (signed) triangle area formed by the O' projection of the query point O onto the ring plane and the ring atoms i and j. Defining $\mathbf{T}_{O'i}$ and $\mathbf{T}_{ij}$ as vectors joining O' to the ring atom i and ring atom i to j respectively, the sign of the triangle is positive if the vector product $\mathbf{T}_{O'i} \times \mathbf{T}_{ij}$ has the same direction as the ring normal with ring atoms counted in $i \rightarrow j$ direction. r_i and r_j are the distances between O and atoms i and j respectively. k_{ring} is a proportionality constant. The summation goes over all the adjacent ij atom pairs forming the ring, that is over the number of bonds in the conjugated ring. The ring current effects on chemical shifts are thoroughly reviewed in Chapter 5 for the detailed assessment of different models to be used in the development of chemical shift predictors for nucleic acids.

All the aromatic rings that have at least two of their non-hydrogen atoms at the vicinity of the methyl carbon nucleus, within the active region, are included. For tryptophan residues, if one of the two rings satisfy the above mentioned criterion, the second ring is included as well. The safe 6.5 Å cutoff radius was chosen because the ring current effects are negligible at distances longer than approximately 5.5 Å [Case, 1995]. As a query point O, the methyl carbon and the geometric centre of the three methyl hydrogens are taken for ^{13}C and ^{1}H

chemical shifts, respectively.

2.8 Magnetic anisotropy terms

Magnetic anisotropy effects are incorporated into the calculations by following the method used to account for the peptide group anisotropy effects on backbone ^{1}H chemical shifts by Case and coworkers [Ösapay & Case, 1991]. The method uses the McConnell formulation [McConnell, 1957] of the magnetic anisotropy contribution to the chemical shifts, reduced by an assumption of axial symmetry for the source of the anisotropy. In this case, the distant group magnetic anisotropy contribution to the chemical shift value can be approximated as (Equation 2.4):

$$\Delta\delta_{ma} = \frac{\Delta\chi}{3N_A} \times \frac{3cos^2\theta - 1}{r^3} \tag{2.4}$$

where $\Delta\chi$ is the magnetic susceptibility anisotropy, N_A is the Avogadro number, r is the distance between the nucleus and a point defined in the anisotropic moiety, θ is the angle between the $\mathbf{r}$ vector and the normal of the plane of that group. The second factor in Equation 2.4 can be considered as a geometric term for the magnetic anisotropy effects and be included in the modelling of the chemical shifts.

Protein backbone peptide groups, as well as the carboxylic, amide and guanidinium moieties of Asp, Asn, Glu, Gln, and Arg side-chains are considered as sources of magnetic anisotropy. In case of peptide moieties, the optimal placement of the origin on the plane for calculation of $\mathbf{r}$ is approximately at the centre of the NCO group [Ösapay & Case, 1991]. By generalising this finding, the geometric centres of the COO and CON atoms were used as origins for the carboxylic and amide planes respectively. For arginine side-chains, the carbon centre of the guanidinium group was used.

2.9 Electric field terms

Electric fields alter the chemical shifts by polarising the local electronic distributions. For an atom X that is connected only to another atom Y, this dependence

was shown to be approximated by the chemical shift polarisability constant multiplied by the electric field projection along the X-Y axis [Buckingham, 1960; Buckingham & Pople, 1963]. Here, the electric field effect was accounted for by following Coulomb's law and reducing the electrostatic effects of the atoms to the simple electric monopole interactions. Amber03 charges [Duan *et al.*, 2003] were used and only the atoms within the active region were considered. The electric field along the local symmetry axis of the methyl group was calculated, i.e. along the H_3C-C or H_3C-S (for methionine) bond. Thus, the implemented electric field term is (Equation 2.5):

$$\Delta\delta_{EF} = k_{EF} \sum_i \frac{q_i cos\theta}{r_i^2} \tag{2.5}$$

where q_i is the partial charge of the i^{th} atom in the active region, θ is the angle between the local symmetry axis of the methyl group and the vector $\mathbf{r}$ with length r_i that joins the methyl nucleus with the i^{th} atom. k_{EF} is the proportionality constant for the electric field term.

2.10 Distance-based terms

The distance-based terms used in CH3Shift are modified from the scheme implemented in the CamShift method for the backbone nuclei [Kohlhoff *et al.*, 2009]. Here, fewer types of distance, but included in a greater number of polynomial terms is used (Equation 2.6).

$$\Delta\delta_{dist} = \sum_{i\in\{-1,1,3,6\}} k_i r^{-i} \tag{2.6}$$

Besides the r and r^{-3} terms, which are used for all the atoms, r^{-1} and r^{-6} terms are also added. The inclusion of the r^{-6} term has been implemented in chemical shift predictors for small molecules to treat the weak interaction between atoms [Abraham *et al.*, 2001]. The combination of the r, r^{-1} and r^{-3} terms effectively takes into account the electrostatic interactions, given the presence of screening effects that can alter the dielectric constant of the surrounding medium with the strength linearly proportional to the distance from the NMR active nucleus.

After extracting all the atom-specific distances between the given methyl site and the atoms in the active sphere, the distances that join the methyl carbon and exactly the same type of atoms (for example Arg-C^{γ} atoms, if more than one Arg residue is found in neighbourhood) are summed before applying the power operation (Equation 2.6).

As a further optimisation of distance-based terms from their previous form [Kohlhoff *et al.*, 2009], a procedure in which distances are merged, i.e. they are summed after the corresponding power operation, is used. Besides the backbone N, C', H^N, C^{α}, H^{α} and C^{β} atoms, which are essentially always present in the proximity of side-chain methyl groups and allow parameter fitting with high statistical significance, the rest of the distances are treated jointly. For example, all the geometric terms (after the above power operation) that stem from the distances between the query nucleus and the sp^3 hybridised carbon atoms of any amino acid residue, are summed into a single term (separately for each i in r^i terms), which will then allow to fit a single coefficient for all the sp^3 hybridised carbons in a given r^i category.

The list of distances treated in a merged way includes those between the given query nucleus and a) sp^3 hybridised carbons, b) hydrogen atoms attached to a sp^3 hybridised carbons, c) sp^2 hybridised carbons (in aromatic rings), d) hydrogens attached to a sp^2 hybridised carbons, e) sulphur atoms, f) hydroxylic oxygens, g) hydroxylic and thiolic hydrogens, h) other carbons (side-chain carboxylic, amide), i) other hydrogens atoms (imino, amino, guanidinium), j) other oxygen atoms (side-chain carboxylic, amide) and k) other nitrogen atoms (heterocyclic, amide, guanidinium, lysine amino). The optimal types of merged distances and terms were found by multiple trials, paying a particular attention to measures for avoiding overfitting.

Since accounting for the correct protonation state is very challenging, the most common protonation states are enforced for all the relevant amino acids during hydrogen addition to the structures in the database. All acidic residues were considered as deprotonated, lysine and cysteine as protonated, and histidine as protonated only at the δ position. The importance of considering explicitly in the parametrisation the exact protonation states is decreased by the joint treatment of the distances, which is adopted to avoid overfitting problems because the database

that is used includes a relatively low number of instances of any particular type of internuclear distance. An accurate assessment of the effects stemming from the different protonation states will become possible with the growth of the number of structures and associated chemical shift data.

2.11 Parameter fitting, optimisation and over-fitting control

Least squares fitting procedure is used to determine the coefficients for all the used terms for describing the chemical shifts. All the calculations as well as data filtering and manipulations were done in the *R* statistical programming language [R Development Core Team, 2011].

In order to decrease the number of parameters and increase the statistical significance of the predictions, the model optimisation was done by a Monte Carlo procedure in the space of the possible combinations of the used terms. In this approach, all the terms were set as adjustable (i.e. present or absent), except the ring current and magnetic anisotropy terms, as they were statistically significant even when the full model was used for fitting. For each nucleus and residue type, 70000 Monte Carlo steps were performed; at each step a randomly selected term was switched on or off with an acceptance probability defined by the Metropolis criterion. As the pseudo-energy in the Monte Carlo procedure, the fitting quality from the leave-one-out tests after each fitting step was used. The pseudo-temperature factor was defined to obtain about 60-70% acceptance rate, and thus sample parameter space efficiently. The final model was selected as the one resulting in the best agreement between the predicted and experimental chemical shifts from the leave-one-out tests (see Table 2.1).

As typical of phenomenological approaches, there is an overlap between different terms in the procedure followed here, which can account for a given effect in more than one way. For instance, the anisotropy and ring current effects are modelled by both special geometric factors and the distances joining the atoms of the aromatic rings or magnetically anisotropic molecular moieties to the methyl nuclei. The electric field effect, which is included as a direct evaluation based

Table 2.1: *Summary of the results of the* CH3Shift *model optimisation. The ratios of the standard deviation of experimental chemical shifts used for model fitting and the standard error of the predictions in the fitted data (not from the leave-one-out test) are shown. All optimised models have offsets in their equations; the offsets for Thr-*$^1H^{\gamma 2}$*, Val-*$^1H^{\gamma 2}$*, Leu-*$^1H^{\delta 1}$ *and Ile-*$^1H^{\gamma 2}$ *nuclei are rotamer-specific. All the* Ω_i *terms, which denote the ten cosine functions that we used (see Appendix C), as well as the* θ^i *terms, operate on each of the four dihedral angles* ϕ, ψ, χ_1 *and* χ_2*. Therefore the absence (−) of any of them results in the reduction of the number of parameters by four. Likewise, the absence of any of the* ϕ, ψ, χ_1 *or* χ_2 *terms in the final model means a reduction of the number of parameters by 14 (four for* θ^i *and ten for* Ω_i*). All the models also include the terms for ring current and magnetic anisotropy effects from conjugated rings, peptide moieties and anisotropic side-chain moieties, which were always set present and non-adjustable.*

Res.	Nucl.	offs.	rot.	F_{EF}	r	$1/r$	$1/r^3$	$1/r^6$	ϕ	ψ	χ_1	χ_2	θ	θ^2	θ^3	θ^4	Ω_1	Ω_2	Ω_3	Ω_4	Ω_5	Ω_6	Ω_7	Ω_8	Ω_9	Ω_{10}	SD_{train}/SE_{pred}
Ala	$^{13}C^{\beta}$	+	-	-	+	+	+	+	+	+	-	-	+	+	+	+	+	-	+	-	-	-	+	-	+	-	1.873
Ala	$^{1}H^{\beta}$	+	-	+	+	-	+	+	+	+	+	+	-	+	-	-	+	+	+	-	+	-	-	+	-	-	1.545
Thr	$^{1}H^{\gamma 2}$	+	+	+	+	+	+	+	+	-	-	-	-	+	-	-	-	+	-	-	-	+	-	-	+	+	1.375
Val	$^{1}H^{\gamma 1}$	+	-	+	+	+	+	+	+	+	-	+	-	-	+	-	-	+	-	-	-	-	-	-	-	-	1.362
Val	$^{1}H^{\gamma 2}$	+	+	-	-	+	-	+	-	+	-	+	-	-	-	-	-	-	-	-	-	+	-	+	-	-	1.433
Leu	$^{1}H^{\delta 1}$	+	+	-	-	-	+	+	+	+	-	+	-	-	-	+	-	-	-	-	-	-	-	-	-	-	1.252
Leu	$^{1}H^{\delta 2}$	+	-	+	-	+	+	+	+	-	+	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	1.421
Ile	$^{1}H^{\gamma 2}$	+	+	+	-	+	+	+	-	-	+	-	+	+	-	+	+	-	-	-	+	+	+	+	-	-	1.413
Ile	$^{1}H^{\delta 1}$	+	-	+	+	+	-	+	+	-	-	-	-	-	+	-	+	+	-	-	-	-	-	-	+	-	1.496

on partial charges, is also covered by the distance terms. This double-counting makes it difficult to provide a physical interpretation of the individual coefficients resulting from the fitting procedure. Therefore extensive tests are performed to check the consistency of the prediction performance, looking for possible abrupt changes in the prediction qualities from one trial to another, or from one compilation of the training data to another, which would have suggested the presence of an overfitting problem. Two types of tests are done to assess the quality of the fits. The first was the standard leave-one-out test, in which any single prediction is done while that particular chemical shift entry with the corresponding structural parameters is excluded from the training set used to optimise the coefficients. For the second test, the compiled chemical shift data with the associated structural factors were randomly split into training and test sets with the percentage of data in the test set varying from 5 to 30% of the whole set. The calculations were run for each of the residue and nucleus types separately, and, each of the random splitting of the data were replicated 250 times. The fitting quality is assessed by examining the dependence of the standard errors of prediction in the training and test sets (with all the 250 trials) against the percentage of the whole data used to optimise the coefficients. The cases of overfitting are characterised by an artificial improvement in the quality of the predictions in the training set associated by a decrease in the quality in the test set, when the percentage of data used for training was decreased (for an example, see Appendix D). The cases reported in this work are those for which no behaviour characteristic of overfitting was found. In other cases, however, e.g. for methionine ^{1}H and ^{13}C chemical shifts, overfitting could not be avoided, a result mainly determined by the fewer number of currently available experimental chemical shift data for methionine.

2.12 The CH3Shift software program and web server

The developed structure-based chemical shift predictor for protein methyl groups is available as a software program. Besides the stand-alone implementation, CH3Shift web server is created. Given the structure file of a protein in PDB

format, the program returns the predicted methyl group ^{1}H and ^{13}C chemical shifts. In addition, it has multiple functionalities, such as comparison of the results to the experimental data, re-referencing of the results based on the provided experimental chemical shifts via a least squares optimisation and various plotting options. The program is available through `http://www-sidechain.ch.cam.ac.uk/CH3Shift` web address. The graphical user interface is developed via *Rwui*, a web application to create user friendly interfaces for R scripts [Newton & Wernisch, 2007].

2.13 Analysis of the differences in the methyl group chemical shifts of Val, Leu and Ile

The differences of the ^{13}C chemical shifts of the two methyl groups in Val, Leu and Ile residues have recently been shown to be useful for deriving structural information [Hong *et al.*, 2009; London *et al.*, 2008; Mulder, 2009]. These chemical shift differences depend on the rotameric states of the side-chains, an observation strengthened by the finding that ^{13}C chemical shifts and vicinal J-couplings are correlated [Mulder, 2009]. The initial analysis of the CH3Shift-DB database outlines an interdependence of some types of chemical shifts from different methyl groups of Val, Leu and Ile residues (Figure 2.3).

A significant correlation is present between the ^{1}H chemical shifts of Val and Leu residues regardless of the rotameric states of the residue (Figure 2.3). The reason for the correlations observed among ^{1}H nuclei, but not among ^{13}C nuclei, can be the more pronounced sensitivity of proton chemical shifts on the long-range environmental interactions that are correlated at the two methyl sites of the same residue. These results demonstrate that the magnitudes of the chemical shift alterations from the non-bonded interactions are approximately of the same order at two methyl sites of the same residue. On the contrary, the ^{13}C chemical shifts, besides the sensitivity towards the non-bonded effects, are also sensitive to the core effects as supported by the observation of their strong dependence on the dihedral angles defining side-chain conformation [Pearson *et al.*, 1997]. Hence, taking the difference of carbon chemical shifts minimises the contribution

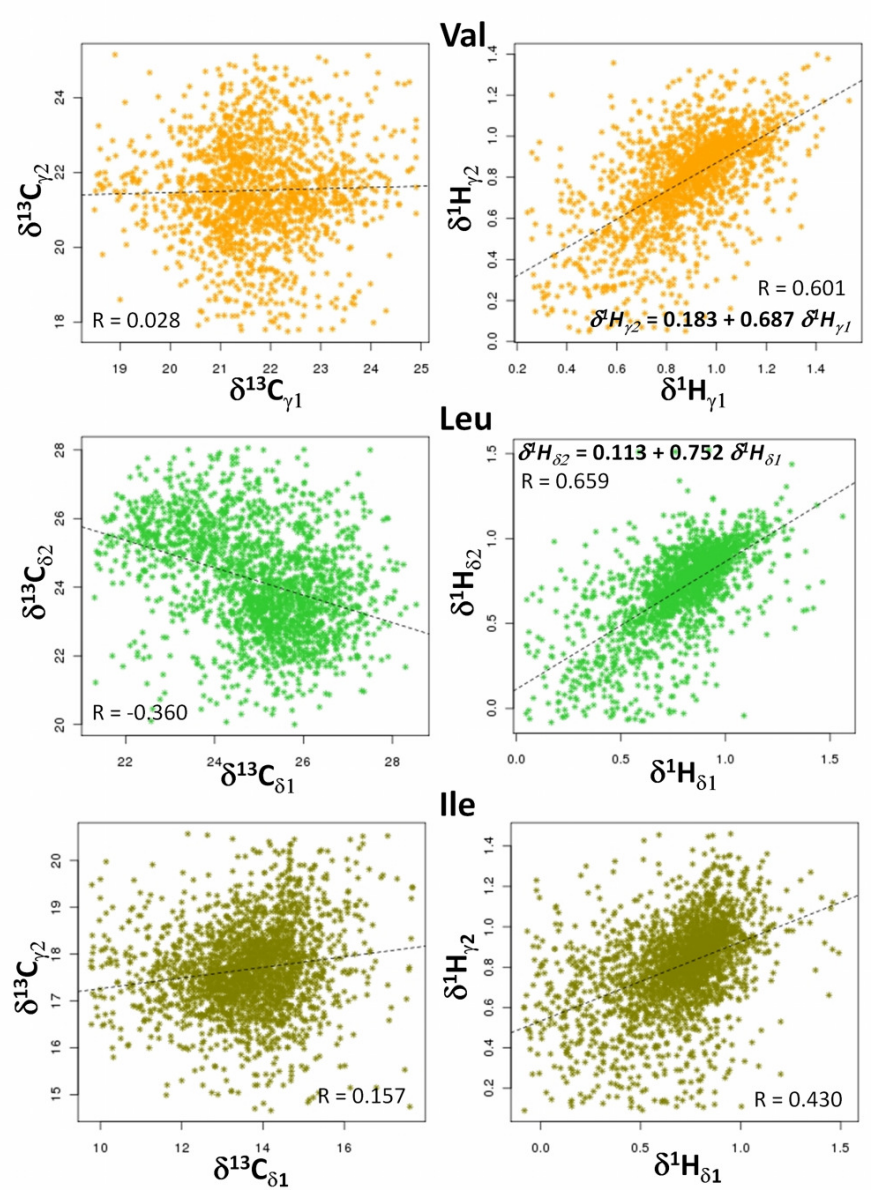


Figure 2.3: *Correlation between the methyl chemical shifts of the amino acid residues in the CH3Shift-DB database that contain two methyl groups. The correlation coefficients and the linear equations are shown.*

from the long-range effects, leaving only the core effects which clearly correlate with the χ dihedral angles.

2.14 Challenges in the structure-based predictions of methyl chemical shifts

Despite the recent advances in the structure-based predictions of backbone chemical shifts [Kohlhoff *et al.*, 2009; Lehtivarjo *et al.*, 2009; Meiler, 2003; Neal *et al.*, 2003; Shen & Bax, 2007; Wishart, 2011; Xu & Case, 2001], the extension of these methods to side-chains has been very challenging for a series of reasons. The first is that the number of methyl chemical shift records in the BMRB is still small when compared to the number of entries for protein backbone nuclei. Thus, the fitting of the parameters for methyl chemical shift predictors can be done based on just a few thousands of experimental data for each methyl type, as opposed to tens of thousand experimental chemical shift entries available for each backbone nucleus. This scarcity of experimental data restricts the number of factors that can be included in the fitting, in order to avoid overfitting.

The second reason is that our current knowledge of the structure and dynamics of the side-chain conformations, for which methyl group chemical shifts are measured, is often limited. Protein side-chains tend to be rather dynamic, and their positions can be variable because of rotameric jumps. Furthermore, even small uncertainties in the determined average χ_i dihedral angles for the residues, where the methyl is joined to the backbone by a longer chain, result in a more substantial distortion of the methyl group position from its average value. These uncertainties are especially relevant for methyl groups close to aromatic rings because the geometric factor for describing ring current effects is very sensitive to small fluctuations in the geometry. Although the dynamics of buried methyl groups were shown to be comparable in solid and solution states of proteins [Agarwal *et al.*, 2008; Reif *et al.*, 2006] because of the prevalence of the protein hydrophobic core methyl groups that are well separated from the solvent and preserve their microenvironment regardless the phase of the system, such dynamics are expected to be non-negligible [DeGortari *et al.*, 2010]. Moreover, solvent-

exposed methyl groups, which are likely to be even more dynamic, comprise a substantial proportion of the filtered database, since the high quality NMR and X-ray investigations are mostly done on relatively smaller proteins for which the ratio of the surface and core methyl groups is greater than the average. Therefore, overall in the CH3Shift-DB database, the average structures of the methyl groups from the X-ray studies can vary from the solution state and can negatively affect the quality of the predictions. In an attempt to avoid these problems we filtered out the surface methyl groups from the training database. The solvent accessible surface area was calculated for each methyl carbon in the database, and the residues were classified as buried if all its methyl carbons had zero solvent accessible surface area. The percentages of the solvent exposed residues in the database was 73.6% for Ala-β, 86.5% for Thr-γ2, 44.2% for Val-γ1, 43.0% for Val-γ2, 39.0% for Ile-γ2, 38.2% for Ile-δ1, 39.4% for Leu-γ1, 38.3% for Leu-γ2, 66.0% for Met-ϵ. The reduction of the number of entries, however, led to overfitting problems and thus this approach was not implemented. Furthermore, the existing predictor, which is trained on the database with both buried and exposed residues, did not show an improvement of the performance when only the buried residues were used in leave-one-out tests. On the contrary, a slight decrease of performance was noted for all the tested nuclei, pointing out that, overall, the high-resolution protein structures used in the fitting procedure resulted in a model that is close to the maximum possible performance one can expect from the current state of the database and the difference between the buried and exposed residues can be accounted only after having a substantial improvement of the quality and quantity of data in the CH3Shift-DB database.

Many of the geometric factors in Equation 2.1 are very sensitive to the dynamics of the methyl groups and the surrounding residues. Moreover, the dependence is not linear, thus short and long-range structural fluctuations are crucial in determining the actual values of the structural factors. Ideally, instead of using a single structure for each of the selected proteins, an ensemble of conformations should be analysed to retrieve and average out all the structural factors. However, although feasible for protein backbone atoms [Lehtivarjo *et al.*, 2009], the ensemble version of the CH3Shift parametrization is yet to benefit from the increasing quality of molecular mechanics force fields for side-chains [Lindorff-

Larsen *et al.*, 2010]. The complex effects that the dynamics has on the chemical shifts are also indicated by the result that the changes in the absolute errors in the ^{1}H chemical shift predictions calculated from the X-ray structure were not correlated with the S^2 order parameter over different methyl groups in ubiquitin (Appendix E). Although a special attention is paid to the processing and filtering steps, some remaining uncertainties in referencing and stereospecific assignment can still be an issue in the compiled chemical shift data. The fraction of those uncertainties will certainly be reduced with time, owing to increasingly standardised experiments and efficient stereospecific assignment techniques.

Finally, perhaps the biggest problem in developing a protein methyl chemical shift predictor is the small variance of the experimental chemical shift values observed in methyl ^{1}H and ^{13}C chemical shifts, as compared to the variance of the chemical shifts of backbone nuclei. Thus, for an acceptable predictive power, the model here is required to produce results that have much smaller standard errors as compared to the backbone chemical shift predictors, for the errors to be smaller than the already small standard deviations of the corresponding experimental chemical shift values in BMRB.

2.15 Random coil methyl chemical shifts

As noted above, methyl chemical shifts of proteins tend to have a small variance compared to other types of chemical shifts, as clearly indicated by the BMRB statistics [Ulrich, 2007]. This observation can be explained by the dynamical nature of the side-chains bearing methyl groups and the absence of specific interactions, such as hydrogen bonding, that involve or are close to the sites of the side-chain methyl groups. A smaller electronic polarisability at the methyl sites in comparison to that at the diatomic moieties of the protein backbone can also be the reason for the smaller methyl chemical shift variance, as the electron distribution at the methyl sites and the corresponding nuclear shieldings are expected to be less affected by environmental and non-bonded effects.

Thus, methyl chemical shifts are fairly close to their random coil values. For a quantitative investigation of this phenomenon, the extracted and re-referenced chemical shift data are further analysed to derive random coil values for the

Table 2.2: *Comparison of the random coil chemical shifts for the ^{13}C and ^{1}H nuclei of the protein side-chain methyl groups with the corresponding average chemical shift values for the α-helical and β-strand structures. The standard deviations (SD) and the number of entries (N) in the corresponding data sets are shown.*

^{13}C	Ala-β	Thr-γ2	Val-γ1	Val-γ2	Leu-δ1	Leu-δ2	Ile-γ2	Ile-δ1	Met-ϵ
$\overline{\delta}_{rc}$	19.015	21.673	21.231	20.955	24.684	23.794	17.567	13.457	17.285
SD_{rc}	1.341	0.638	0.895	1.191	1.326	1.300	0.844	1.305	0.906
N_{rc}	721	367	134	95	177	125	126	128	37
$\overline{\delta}_{\alpha}$	18.199	21.695	22.115	22.372	24.785	24.015	17.599	13.663	17.010
SD_{α}	0.927	0.759	1.051	1.205	1.389	1.535	0.923	1.247	0.789
N_{α}	1520	271	341	308	641	509	439	445	128
$\overline{\delta}^{\beta}$	21.552	21.565	21.499	21.281	24.957	24.832	17.825	13.878	17.317
SD^{β}	1.660	0.860	0.960	1.287	1.549	1.517	0.961	1.296	1.014
N^{β}	494	339	532	375	394	267	537	529	58
^{1}H	Ala-β	Thr-γ2	Val-γ1	Val-γ2	Leu-δ1	Leu-δ2	Ile-γ2	Ile-δ1	Met-ϵ
$\overline{\delta}_{rc}$	1.356	1.177	0.903	0.834	0.844	0.742	0.846	0.748	1.911
SD_{rc}	0.163	0.152	0.165	0.216	0.180	0.242	0.216	0.244	0.299
N_{rc}	515	496	136	102	171	141	165	152	52
$\overline{\delta}_{\alpha}$	1.439	1.190	0.949	0.835	0.783	0.707	0.790	0.676	1.827
SD_{α}	0.189	0.155	0.206	0.257	0.220	0.249	0.231	0.260	0.283
N_{α}	954	332	338	306	599	501	505	509	150
$\overline{\delta}^{\beta}$	1.272	1.078	0.823	0.732	0.760	0.631	0.758	0.660	1.820
SD^{β}	0.200	0.162	0.208	0.230	0.223	0.270	0.235	0.237	0.341
N^{β}	338	443	528	429	366	285	645	595	75

methyl ^{13}C and ^{1}H chemical shifts. Here, for a given type of nucleus and amino acid, the random coil chemical shift is defined as the average value of all the recorded experimental chemical shifts that come from solvent accessible residues which, along with the adjacent two residues, have ϕ/ψ dihedral angle combinations characteristic to either turns or coils. This definition is analogous to that used in the CamCoil method, which has been shown to provide accurate predictions of backbone random coil chemical shifts [De Simone *et al.*, 2009a]. The resulting values are summarised in Table 2.2 along with the standard deviation (SD) and the number (N) of chemical shift entries that fulfilled the above mentioned filtering criteria.

For the comparison of the derived random coil values and the associated statistical data with those from structured regions of proteins, a similar filtering of data was done to derive average α-helical and β-strand chemical shift values. As can be inferred from Table 2.2, chemical shifts from the structured regions do not differ much from their random coil values. The only exception is alanine, for which the methyl group is of C^{β} type, thus is strongly influenced by the backbone conformation. Overall, the data confirms that the development of a protein methyl chemical shift predictor concerns relatively small deviations from random coil chemical shift values.

2.16 Performance of the developed CH3Shift method

In order to assess the performance of the CH3Shift predictor, the correlations are presented between the predicted and experimental chemical shifts, along with standard errors (defined as the standard deviation of the prediction errors in ppm) and correlation coefficients indicated on the plots (Figure 2.4, left).

The correlations are obtained from leave-one-out tests, so that the data tested are not used in the parametrization of the method for that particular prediction. The corresponding distributions of the prediction errors are presented in Figure 2.4, right. Only those nuclei and residue types are presented and discussed herein, for which the prediction accuracy is substantial.

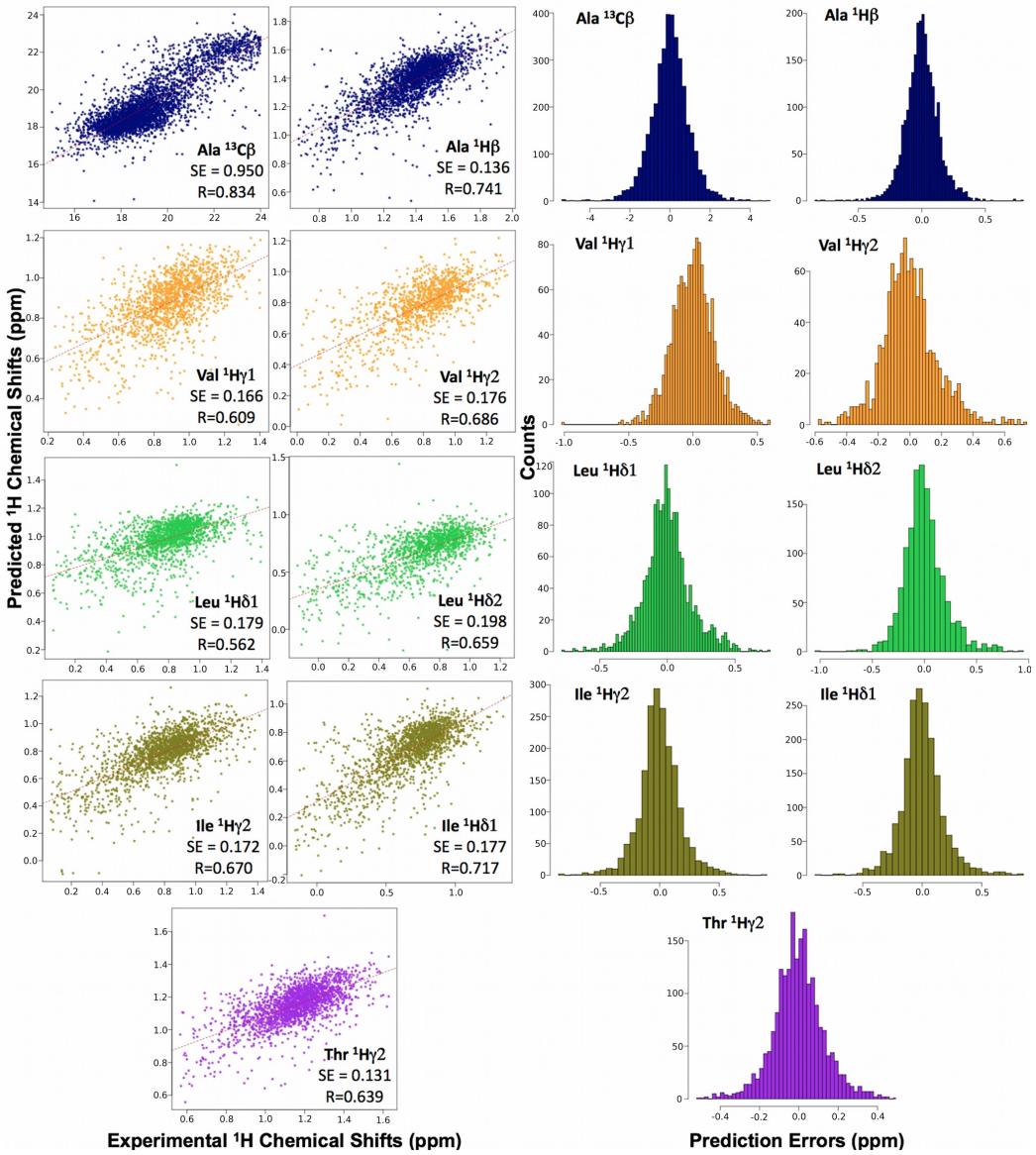


Figure 2.4: *Correlation between predicted and experimental chemical shifts for all the types of methyl* ^{1}H *and Ala* ^{13}C *nuclei (left) in the CH3Shift-DB database. Predictions are obtained from leave-one-out tests, with standard errors given in ppm; the Pearson correlation coefficients are also shown. The histograms of the error distributions for each of the discussed nucleus and residue types are shown at the right side.*

Except Ala, predictions for ^{13}C nuclei do not provide a significant improvement over those based on the average values derived from the BMRB database (Appendix F). The reason for this situation is most probably the neglect of the strong isotope effects on ^{13}C nuclei caused by the immediately attached hydrogen. It will perhaps become possible to account for these effects in the parametrization step by considering a database that includes additional information about the isotopic state of the attached hydrogen atoms (-CD3, -CHD2, -CH2D, -CH3).

Figure 2.5 (green bars) shows the standard errors of the CH3Shift chemical shift predictions and compares them with the standard deviations of the corresponding chemical shifts in the BMRB repository. Overall, the prediction quality is the best for alanine (Figures 2.4 and 2.5). Not unexpectedly, a decrease in the performance of predictor can be noted as the side-chain length grows (Figure 2.5). This effect can be attributed to the structural and dynamical uncertainties associated with the increase in the number of dihedral angles defining the system.

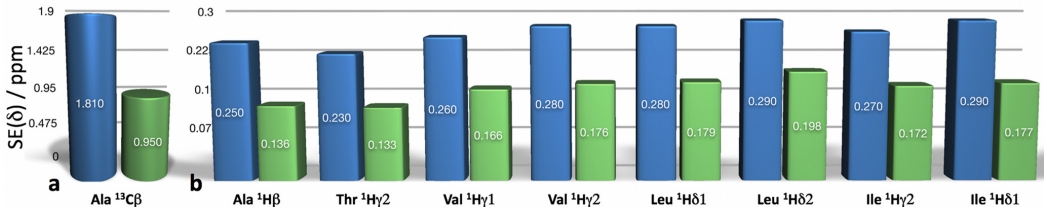


Figure 2.5: *Histogram of the standard errors (in ppm) of the methyl chemical shift predictions in different types of protein side-chain methyl groups for which a good accuracy is achieved. The green bars show the standard errors of the* CH3Shift *predictor, the blue bars show the standard deviations of the corresponding chemical shifts as inferred from BMRB.*

2.17 Applicability of the CH3Shift method for protein structure determination

The CH3Shift method was designed to provide methyl chemical shift predictions that can be incorporated in protein structure determination methods. The initial tests indicated that, despite the associated errors in predictions of the methyl

chemical shifts in the current implementation of the CH3Shift method, such predictions can be used to correctly rank protein structures in terms of their overall distance from the reference conformation. To test the possibility for such usage of the developed predictor, the 2NR2 dynamical ensemble of ubiquitin [Richter *et al.*, 2007] is analysed with CH3Shift. The chemical shifts were calculated for the methyl group nuclei for each of the 144 conformers in the ensemble. The outcome of this trial demonstrates that, for a given methyl group, the structures that result in better predictions have local environments closer to that in the reference X-ray structure (1UBQ, [Vijay-Kumar *et al.*, 1987]) of ubiquitin (Figure 2.6).

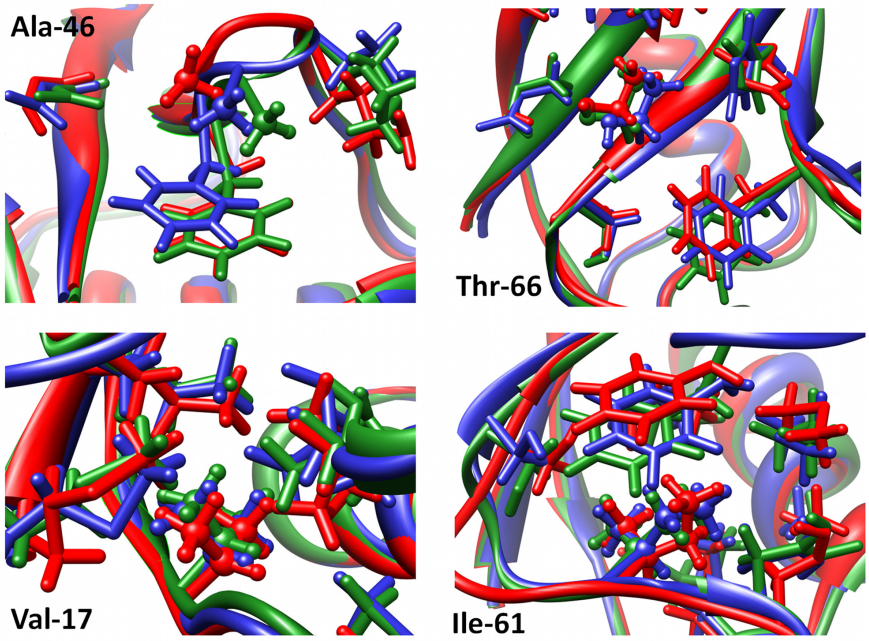


Figure 2.6: *Methyl chemical shift analysis of the 2NR2 dynamical ensemble of ubiquitin. The X-ray structure (green) is compared with the best (blue) and the worst (red) structure in the 2NR2 ensemble in terms of agreement between experimental and calculated methyl chemical shifts. The methyl containing target residues are highlighted as ball-and-stick representations, and the notable residues in vicinity are shown as stick representations.*

The green model corresponds to the X-ray structure of ubiquitin, whereas the blue and red models to the structures with the best and worst agreement, respec-

tively, of the methyl group chemical shift prediction results with the experimental values. For each of the methyl groups, the best local structure is selected from 144 conformations as the one with the best predicted ^{1}H chemical shifts and the ^{13}C predictions in the top ten. This scheme reduces the importance of the carbon chemical shifts, because of the current overall lower prediction quality for methyl carbons. For Ala-46 (Figure 2.6), although the neighbouring phenylalanine ring position of the worst agreement structure is closer to that in the X-ray one, the methyl group is shifted with a significant deviation of its position relative to the ring. On the contrary, the structure of best agreement, which is altered by the loop movement, keeps the relation between the side-chain positions close to the arrangement in the X-ray structure. In Thr-66, an excellent match between the best-agreement and X-ray structures is found, whereas the structure of worst agreement suffers from significantly distorted phenylalanine and histidine ring positions. For Val-16, the overall positions of all the influential moieties around the methyl groups are closer between the X-ray and best-agreement structures. An interesting case is the Ile-61, for which not only the tyrosine ring is substantially distorted in the worst-agreement structure, but also the rotameric type of the isoleucine side-chain itself is different. These results thus indicate that refinement strategies based on methyl chemical shifts have the potential of increasing the accuracy of the side-chain positions.

Next, the 2K39 [Lange *et al.*, 2008] ensemble and the 1D3Z [Cornilescu *et al.*, 1998] set of structures in comparison to the 2NR2 ensemble and the 1UBQ X-ray structure are analysed. Unlike 1D3Z, which contains 10 structures that individually fit to the NOE, J-coupling and RDC data, the 2K39 and 2NR2 ensembles (with 116 and 144 structures respectively) are the results of a treatment of NMR data aimed at reflecting the dynamics of the protein. A recent model free analysis (MFA) of the NMR restraints for the ubiquitin methyl side-chains has shown [Farès *et al.*, 2009] that the 2NR2 ensemble agrees best with the RDCs derived from spherical harmonics according to the Pearson correlation coefficient, but the 2K39 ensemble exhibits a better RMSD (in ppm). Therefore additional comparisons of these two ensembles using different approaches can be important for a further assessment of the methodologies to derive protein dynamics from NMR data. The quality of the back-calculated CH3Shift chemical shifts for methyl

^{1}H nuclei is assessed for the various ensembles of ubiquitin against representing the experimental values. Average RMSDs (in ppm) of the methyl ^{1}H chemical shift prediction errors in 2K39 (116 structures, red), 2NR2 (144 structures, blue) and 1D3Z (10 structures, grey) ensembles, as compared to the prediction errors from the 1UBQ X-ray structure of ubiquitin (green) are shown in Figure 2.7. If the residue contains two methyl groups, the data from both methyl moieties are used for the RMSD calculations. The whiskers indicate the standard deviation of RMSDs over the constituent conformers. The worse RMSDs are not directly related to the solvent accessibility of the residue, as can be seen from the colour-coded band at the bottom of the figure. The observed large RMSDs for the Ala-46 and Leu-50 residues are likely to be connected to the effects of the Phe-45 and Tyr-59 aromatic rings at the vicinity. For a clearer view of the correspondence between the calculated and experimental chemical shifts, the individual correlation plots are shown in Figure 2.8. The best agreement is found for the X-ray structure (Figures 2.7 and 2.8). Although this result could simply be a consequence of the fact that only X-ray structures of proteins were used to parametrize the CH3Shift predictor, it may be also possible that the NMR ensembles, which were derived using other NMR parameters (S^2 order parameters and RDCs), may not represent very accurately the specific population weights that would result in better estimates of the chemical shifts.

As a further assessment of the quality of the ensembles, the leucine ^{13}C chemical shift differences were estimated via the equation [Mulder, 2009] $\Delta\delta^{13}C(\delta_1 - \delta_2) = -5 + 10p_{tr}$ and compared to the experimental values. The p_{tr} is the fraction of the leucine side-chain trans (by χ_2) rotamer during the course of the dynamics and is estimated here based on all the constituent conformers in each of the ubiquitin ensembles. The results are summarised in Figure 2.9.

The data coming from 1D3Z should be interpreted considering that this ensemble is not meant to represent the dynamics of the protein, but rather to provide a high-resolution representation of its average structure. It should also be noted that, in the case of the structural ensembles considered here, the overall correspondence between the experimental ^{13}C chemical shift difference for leucine and the corresponding values predicted through Mulder's equation is comparable to that of the standard deviation of the experimental chemical shifts (1.59 ppm

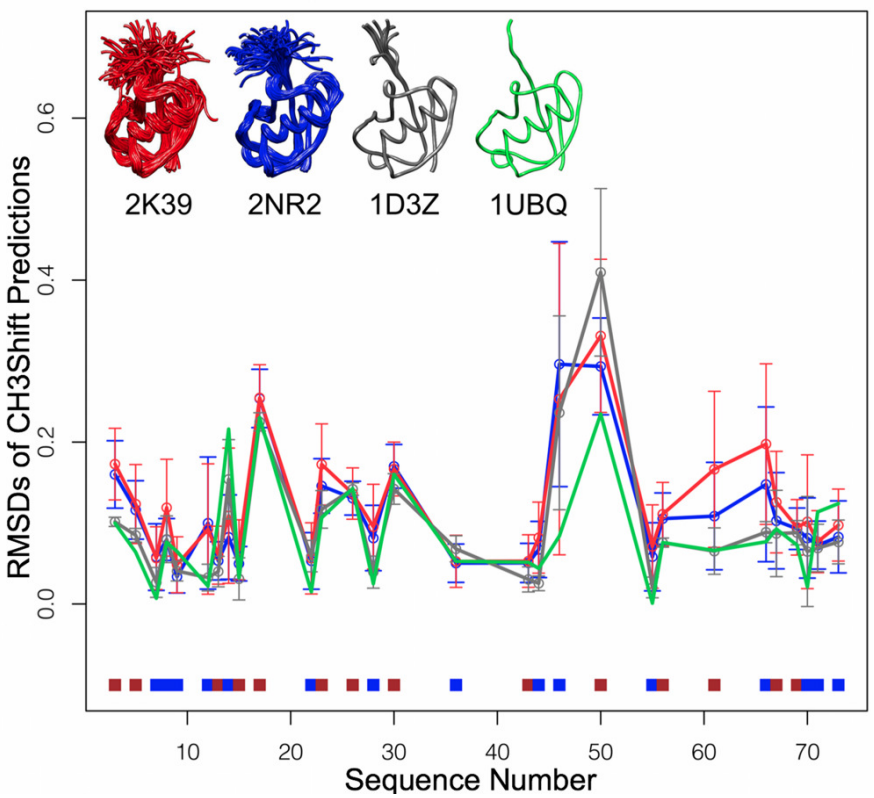


Figure 2.7: *The RMSDs (in ppm) of the average* CH3Shift *predictions (chemical shifts predicted and averaged across all the conformers in a given ensemble) of methyl* ^{1}H *chemical shifts for the 2K39 (116 structures, red), 2NR2 (144 structures, blue) and 1D3Z (10 structures, grey) ensembles. For comparison, the corresponding RMSDs are shown for an X-ray structure of ubiquitin (1UBQ, green). Standard deviations of the RMSD values calculated for the individual conformers are shown as whiskers. The colour-coded band at the bottom indicates the residue-specific solvent accessibility with the blue colour for the solvent-exposed methyl groups and brown colour for the buried ones.*

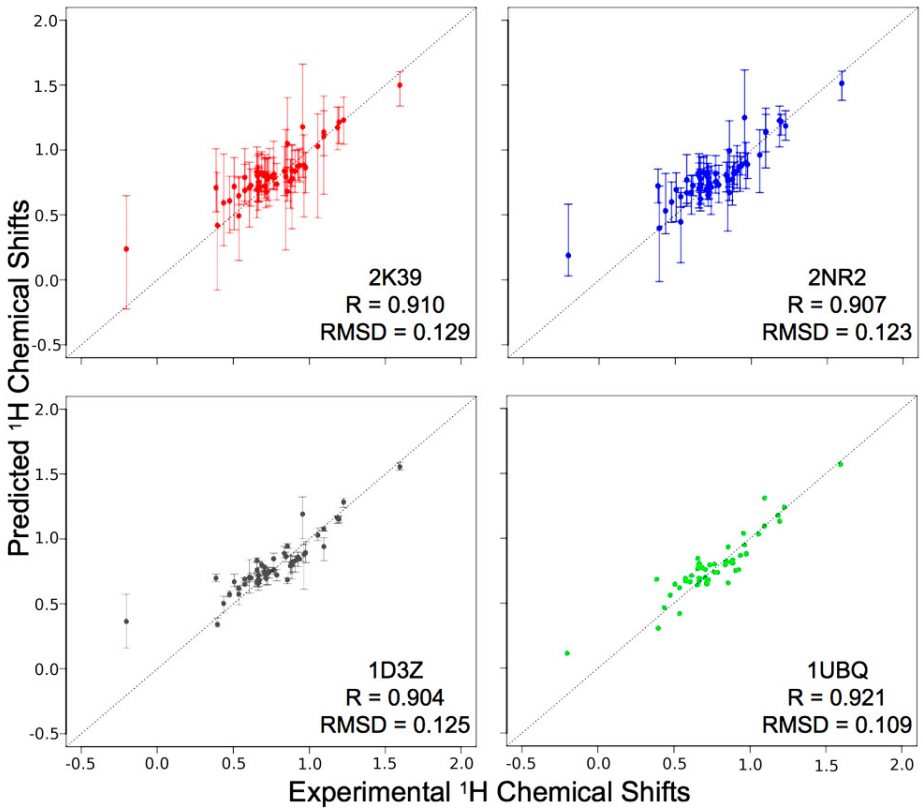


Figure 2.8: *Correlation between the predicted and experimental* ^1H *chemical shifts for the methyl groups in three ubiquitin ensembles (2NR2, 2K39, 1D3Z) and one X-ray structure (1UBQ). The whiskers show the range of the predicted chemical shifts over the multiple conformers where available. The Pearson correlation coefficients and RMSDs (in ppm) are shown. The outlier point with a negative experimental chemical shift value is from an atom strongly exposed to ring current effects, where the prediction quality is more sensitive to the flaws in representation of the correct dynamics of the corresponding locus in the protein.*

for $C^{\delta 1}$ and 1.68 ppm for $C^{\delta 2}$). The examination of the χ_1/χ_2 rotamer distribution for the 2NR2 ensemble indicates a strong correlation of the two side-chain dihedral angles with a prevalent population of two rotameric states in most of the cases. This result, although in contrast to the similar examination of the 2K39 ensemble, is in a good agreement with previous observations on the usual behaviour of leucine side-chains [Hansen *et al.*, 2010; London *et al.*, 2008; Mulder, 2009].

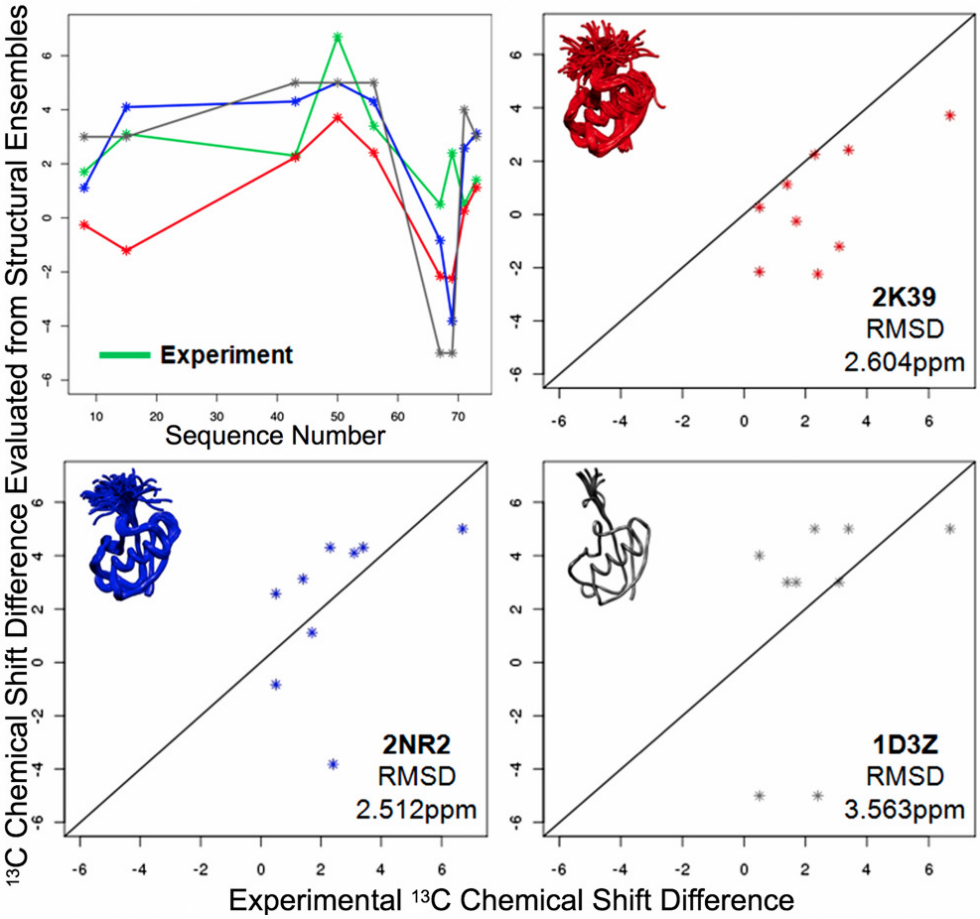


Figure 2.9: *Differences (in ppm) in the methyl chemical shifts of leucine side-chains in three ubiquitin ensembles (2K39 - red, 2NR2 - blue and 1D3Z - grey) as predicted through the formula proposed by Mulder [Mulder, 2009]. Residue-specific predictions are compared with the corresponding experimental values (green).*

2.18 Conclusions

The CH3Shift method for the structure-based prediction of protein methyl chemical shifts is presented. The predictions are performed by using a combination of polynomial functions of interatomic distances with well-characterised phenomenological terms that describe effects of ring currents, magnetic anisotropies, electric fields, rotameric types, and dihedral angles. The performance of the CH3Shift method for Ala, Thr, Val, Leu and Ile methyl groups provides an opportunity for the use of the CH3Shift method to assess the quality of protein structures. Furthermore, it will be possible to continuously improve the quality of the predictions with the growth in the number of methyl chemical shift data deposited in the BMRB.

The solution to a problem changes the problem.

John Peers

3

Chemical Shifts of Protein Side-Chain Aromatic Groups

3.1 Summary

A method for the structure-based prediction of side-chain aromatic ^{1}H chemical shifts of proteins is presented. The ability of the developed predictor to differentiate correct structural models from incorrect ones is also demonstrated, together with its use to detect differences caused by cofactor or ligand binding, or by sequence alterations between structures.

3.2 Motivation

Side-chains play crucial roles in determining the conformational properties of protein surfaces and interior cavities, which in most cases define the specificity of biomolecular interactions. Aromatic side-chains in particular, are capable of

forming interactions with a variety of chemical groups through hydrophobic, π-π stacking, π-anion and π-cation attractions, and, often comprise the hot spots of protein-protein [Crowley & Golovin, 2005] and protein-ligand [Bissantz *et al.*, 2010] complex formation, and protein folding [Frank *et al.*, 2002]. Furthermore, aromatic side-chains, as sources of ring current effects, substantially influence the chemical shifts of other nuclei, including the highly exploited backbone ones. However, although ring current terms are frequently included in chemical shift predictions of backbone nuclei, aromatic chemical shifts are not normally used in turn to define the geometry of the aromatic rings themselves. Recent advances on specific labelling technologies for aromatic side-chains [Kainosho *et al.*, 2006; Lundström *et al.*, 2009b] will soon increase the number of assigned aromatic chemical shifts, thus adding new prospects to the established tradition of aromatic chemical shift measurements [Redfield *et al.*, 1982]. The incorporation of chemical shifts of aromatic side-chains in structure determination algorithms, in addition to the backbone atoms, makes it possible to extend the use of chemical shifts in structure determination studies. To achieve this, one needs to develop chemical shift prediction method for aromatic side-chain nuclei that is based solely on the configurations of proximal atoms. These types of predictors, at variance with other currently available chemical shift predictors that provide chemical shift evaluations for side-chain nuclei [Han *et al.*, 2011; Lehtivarjo *et al.*, 2009; Meiler, 2003; Xu & Case, 2001], are readily differentiable with respect to the atomic coordinates, and thus enable the calculation of biasing forces to integrate into the equations of motions within a molecular dynamics scheme. Predicting aromatic side-chain chemical shifts via differentiable equations opens new opportunities to monitor a range of important processes, and will increase the scope of chemical shift usage in determining the structures of biomolecular complexes and complex biomolecular systems [Das *et al.*, 2009; Montalvao *et al.*, 2008].

3.3 Database analysis and filtering

To address the challenges described above, ArShift, a chemical shift predictor for protein side-chain aromatic ^{1}H nuclei, is developed. The ArShift predictions, like CH3Shift, are based on known phenomenological terms that describe the

effects of ring current [Haigh & Mallion, 1980], magnetic anisotropy [Ösapay & Case, 1991] and electric field [Buckingham, 1960] terms, which are accompanied by a set of dihedral angle terms and distance-based polynomials [Kohlhoff *et al.*, 2009].

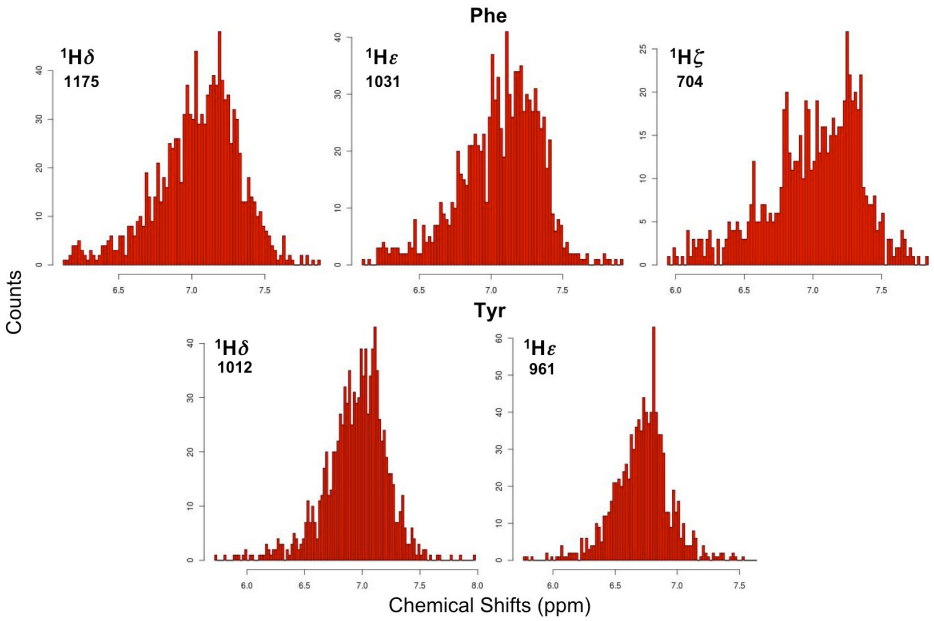


Figure 3.1: *Distribution of the experimental* ^1H *chemical shifts of the Phe and Tyr aromatic side-chains used for parametrizing the* ArShift *predictor. The number of the re-referenced* ^1H *chemical shifts that met all the filtering criteria are also shown.*

A comprehensive analysis of the aromatic chemical shift assignments available from the BMRB database [Ulrich, 2007] is used after filtering and re-referencing steps [Rieping & Vranken, 2010] to reduce the number of inaccurate entries. Out of a total of 502 proteins with a unique PDB [Berman *et al.*, 2000] X-ray structure identifier, 21 structures were discarded because they were relative to protein-nucleic acid complexes, and 29 other structures were discarded for containing porphyrin rings, iron or cobalt atoms. This filtering eliminated non-protein contacts, as well as non-standard ring current and paramagnetic effects, at the cost of removing 336 amino acid residues out of the initial 3630 available ones. Thus, the compiled data set holds 3294 residue-specific chemical shift records (1796

Phe and 1498 Tyr) from 452 protein PDB files. All the X-ray structures were processed by adding hydrogens and doing 1000 steps of hydrogen-only geometry optimisation with the Amber 03 force field [Duan *et al.*, 2003]. The database was further trimmed by considering only the chemical shifts within the 3 times standard deviation window for each of the specific nucleus and residue types, and, for which an X-ray structure with a 2.0 Å resolution or better was present. Finally, the single most outlying entry after these data processing steps was removed for each nucleus type. The final distribution and numbers of chemical shift records of different types are reported in Figure 3.1.

3.4 Intercept, dihedral angle, ring current, magnetic anisotropy, electric field and distance terms

To identify the component of the chemical shift independent from the neighbouring amino acids, an intercept was generated by the fitting procedure along with the optimised parameters for the other factors. The approach here is analogous to the one used in CH3Shift development. The presence of significant differences among the average chemical shifts in different rotameric states (for the distribution of χ angles see Figure 3.2) is noted, necessitating the need to account for rotamer-specific terms within the intercept. To this end, a $k_1R_1 + k_2R_2 + k_3R_3$ term was included in the fitting, where k_i are parameters, and R_1, R_2 and R_3 define the rotameric state; these latter terms are equal to 1 for $-120 < \chi_1 \leq 0$, $0 < \chi_1 \leq 120$ and $(120 < \chi_1 \leq 180) \cup (-180 \leq \chi_1 \leq -120)$ for R_1, R_2 and R_3, respectively, and 0 otherwise.

The inclusion of dihedral angle effects was done by considering the backbone ϕ, ψ, and the side-chain χ_1 and χ_2 dihedral angles, via the same approach and equations (Appendix C) described for CH3Shift.

Ring current effects on the Phe and Tyr aromatic proton chemical shifts from the neighbouring Phe, Tyr, His, Trp-5 and Trp-6 (5- and 6-membered tryptophan rings) aromatic rings were accounted for through the factor from the empirical quantum mechanical model by Haigh and Mallion [Haigh & Mallion, 1972, 1980].

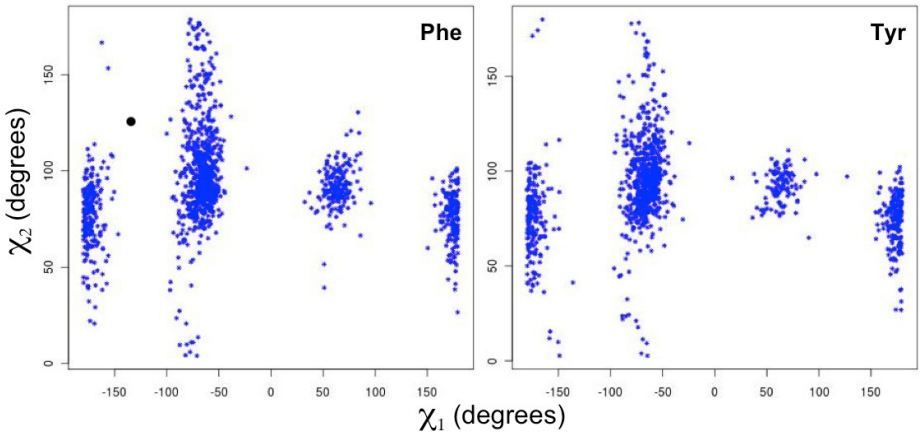


Figure 3.2: *The distribution of χ_1 and χ_2 dihedral angles in Phe and Tyr aromatic side-chains.*

All the aromatic rings that had at least two of their non-hydrogen atoms at the vicinity of the examined aromatic proton were accounted. For Trp residues, if one of the constituent two rings satisfied the above mentioned criterion, the second ring was also included regardless of its presence inside the defined active region.

Magnetic anisotropy effects were included in the model following the method to account the peptide group anisotropy effects on backbone ^{1}H chemical shifts used by Case and coworkers [Ösapay & Case, 1991].

Electric field effects[Buckingham, 1960], as before, were accounted via the simple Coulomb law. Amber03 point charges [Duan *et al.*, 2003] were used and only the atoms within the active region were considered. Electric fields acting on aromatic protons along the corresponding C-H bonds were used.

For each aromatic proton, a region was defined that included all the neighbouring nuclei within a 6.5 Å distance. Only the atoms in that region were considered for the derivation of structural terms influencing the aromatic ^{1}H side-chain chemical shifts. The atoms of the own aromatic ring were neglected, since their relative position remains constant over different conformations and vicinity of the given aromatic side-chain. Hence, their influence on chemical shifts can be safely accounted within the random coil (intercept) term of the model.

The distance-based terms were modified from the CamShift scheme for backbone nuclei [Kohlhoff *et al.*, 2009], with substantially fewer types of distance used.

Besides the r and r^{-3} terms, which were used for all the atoms regardless of the immediate or distant connectivity, r^{-1} and r^{-6} terms were also added. The inclusion of the r^{-6} term has been implemented in chemical shift predictors used for small molecules to model the weak interaction between atoms [Abraham *et al.*, 2001]. Distances involving backbone N, C', H^N, C^α, H^α and C^β atoms around the aromatic groups were considered individually, since they are present in relatively large numbers. The rest of the distances were instead treated jointly, as described for CH3Shift.

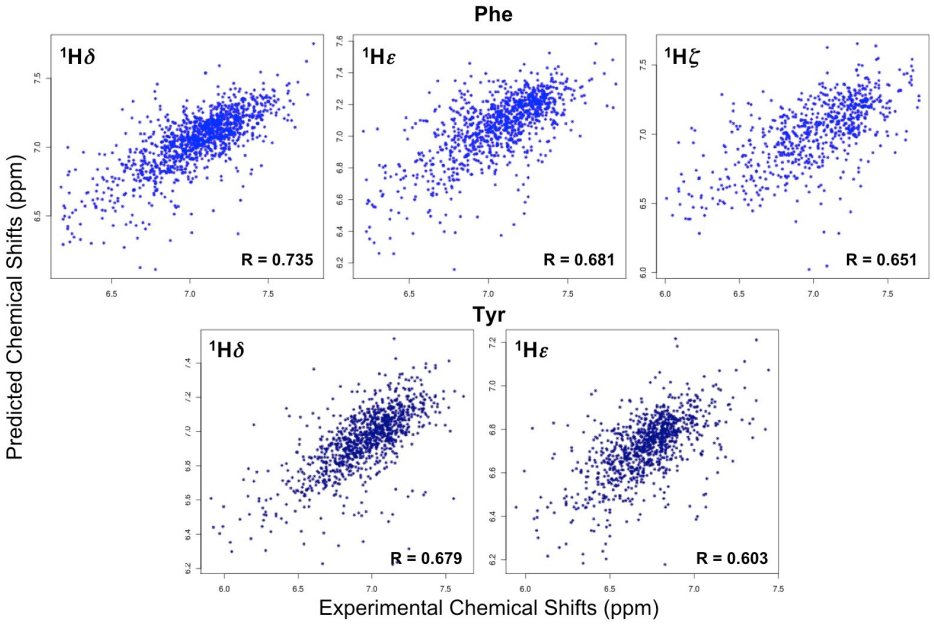


Figure 3.3: *Comparison between predicted and experimental chemical shifts for all types of Phe and Tyr aromatic* ^1H *nuclei. Predictions are obtained from leave-one-out tests. The Pearson correlation coefficients are also shown.*

3.5 Averaging of the geometric factors

The aromatic proton chemical shifts of the $^1H^{\delta_1}/^1H^{\delta_2}$ and $^1H^{\epsilon_1}/^1H^{\epsilon_2}$ pairs tend to appear as a single resonance signal in NMR spectra owing to frequent flips of the aromatic heads of Phe and Tyr residues within the NMR timescale. To this end, the HD1/HD2 and HE1/HE2 naming convention in the PDB files of proteins is

arbitrary, which is accounted in generation of the Figure 3.2. Dictated by the natural averaging of the above mentioned chemical shift types, all the geometric factors that are dependent on the specific position of the NMR nucleus, unlike dihedral angle terms which are common for all the nuclei of the same amino acid residue, were averaged across HD1/HD2 and HE1/HE2 pairs.

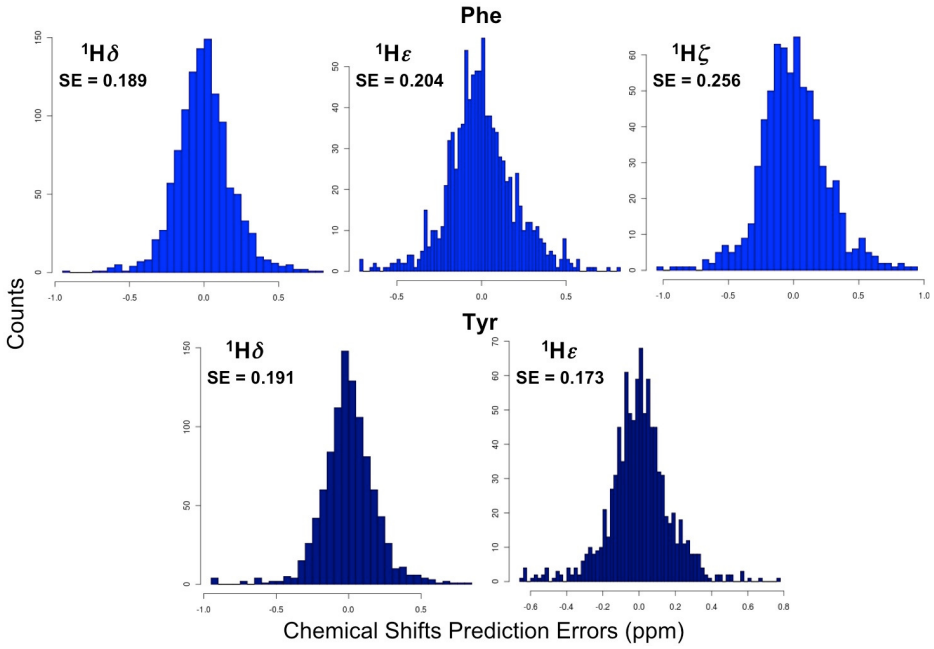


Figure 3.4: *Histograms of the error distributions (in ppm) in the predictions for different types of aromatic side-chain* ^1H *chemical shifts from leave-one-out tests. The standard errors of predictions are also shown.*

3.6 Model optimisation and fitting

The least squares fitting procedure was used to define the coefficients for the different terms contributing to the predictions of the chemical shifts, as implemented in the *R* programming environment [R Development Core Team, 2011]. All the mentioned terms from the complete model, which require further optimisation in order to decrease the number of parameters and increase the statistical weights of the constituent coefficients, are obtained from the fitting procedure. The model

Table 3.1: *Summary of the results for the terms that were set as adjustable (i.e. present or absent) in the Monte Carlo procedure that is adopted to select the best combination of factors for predicting aromatic side-chain proton chemical shifts. The ratios of the standard deviation of experimental chemical shifts used for model fitting and the standard error of the predictions in the fitted data (not from the leave-one-out test) are shown. All optimised models have offsets in their equations of which the offset for the Phe-$^1H^\zeta$ nucleus is rotamer-specific. Attention should be paid to the interconnection between different terms. All the Ω_i terms that denote the 10 cosine functions that we used, as well as the θ^i terms, operate on each of the four (ϕ, ψ, χ_1 and χ_2) dihedral angles. Therefore the absence ($-$) of any of them results in the reduction of the number of parameters by 4. Similarly, the absence of any of the ϕ, ψ, χ_1 or χ_2 terms in the final model means a reduction of the number of parameters by 14 (4 θ^i and 10 Ω_i). All the models also include the terms for ring current and magnetic anisotropy effects from conjugated rings, peptide moieties and anisotropic side-chain moieties that were always set as present and non-adjustable.*

Res.	Nucl.	offs.	rot.	F_{EF}	r	$1/r$	$1/r^3$	$1/r^6$	ϕ	ψ	χ_1	χ_2	θ	θ^2	θ^3	θ^4	Ω_1	Ω_2	Ω_3	Ω_4	Ω_5	Ω_6	Ω_7	Ω_8	Ω_9	Ω_{10}	SD_{train}/SE_{pred}
Phe	$^1H^\delta$	+	-	+	-	-	+	+	-	-	+	+	-	-	-	-	-	-	+	-	-	-	+	+	+	-	1.475
Phe	$^1H^\epsilon$	+	-	-	+	+	-	+	-	-	+	+	+	-	-	-	-	-	-	-	-	-	-	-	+	-	1.376
Phe	$^1H^\zeta$	+	+	+	+	+	-	-	+	-	+	+	-	-	-	+	-	+	+	+	+	+	-	-	-	-	1.399
Tyr	$^1H^\delta$	+	-	+	-	-	+	+	+	-	+	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	1.397
Tyr	$^1H^\epsilon$	+	-	-	-	-	+	+	-	-	-	+	-	+	-	+	-	-	-	-	-	-	-	-	-	-	1.296

optimisation was done by setting all the terms as adjustable (i.e. present or absent), except the ring current and magnetic anisotropy terms, which resulted in statistically significant parameters even when the full model was used for fitting. A Monte Carlo scheme was used and 70000 trial combinations were explored for each of the nucleus and residue types, where a randomly selected factor in a model was switched on or off with an acceptance probability defined by the Metropolis criterion.

The fitting quality from the leave-one-out tests after each fitting step was used as pseudo-energy in the Monte Carlo procedure. The temperature factor was arbitrarily defined to obtain an acceptance rate of about 60-70% in order to sample extensively the parameter space. The optimised model was selected as the one resulting in the best agreement between the predicted and experimental chemical shifts from the leave-one-out tests (see Table 3.1, Figures 3.3 and 3.4).

3.7 The ArShift web server

A web server is available to enable users to carry out predictions using the ArShift method. By uploading a protein structure in PDB format onto this web server the user obtains, as an output, the predicted aromatic side-chain ^{1}H chemical shifts. The program is available at the `http://www-sidechain.ch.cam.ac.uk/ArShift` web address. The GUI is developed via Rwui, a web application to create user friendly interfaces for R scripts [Newton & Wernisch, 2007].

3.8 Performance of the ArShift web server: prospects for protein structure quality assessment

As mentioned above, the accuracy of the prediction method is assessed by using leave-one-out tests, where the predictions are performed individually for all the chemical shift entries used for deriving the coefficients. The standard deviations of the residual errors (denoted here as standard errors) for the models implemented

in the ArShift package are 0.189, 0.204, 0.256, 0.191 and 0.173 ppm for Phe-$^{1}H^{\delta}$, Phe-$^{1}H^{\epsilon}$, Phe-$^{1}H^{\zeta}$, Tyr-$^{1}H^{\delta}$ and Tyr-$^{1}H^{\epsilon}$ nuclei, respectively (Figures 3.3 and 3.4). The comparison of the ArShift standard errors and the standard deviations of the corresponding chemical shift types in the BMRB database are presented in Figure 3.5:

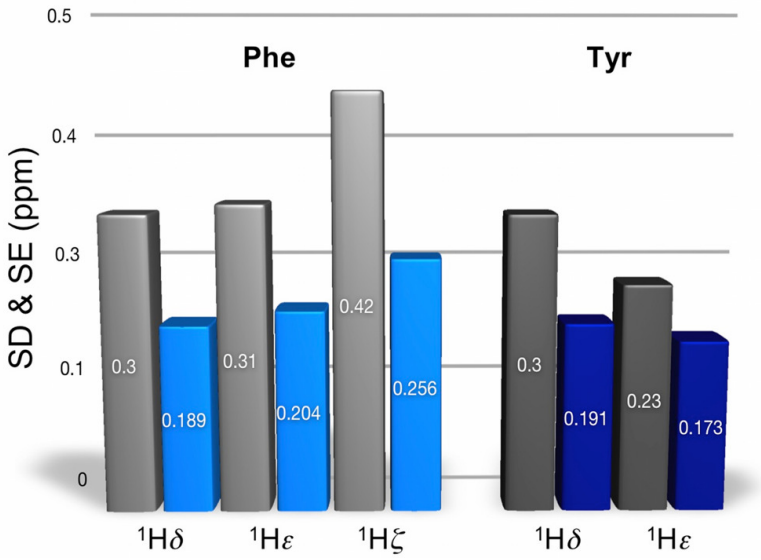


Figure 3.5: *Performance of the ^{1}H chemical shift predictions for different types of protein aromatic side-chain nuclei. The coloured bars (blue for Phe and dark blue for Tyr) show the standard errors in ppm of the* ArShift *predictor. The grey bars show the standard deviations of the corresponding chemical shifts in the BMRB database.*

where the prediction results for ^{13}C nuclei are not presented because they do not provide a significant improvement over the average values derived from the BMRB database. The reason for this situation is most probably the neglect of the stronger isotope effects on ^{13}C nucleus caused by the immediately attached nuclei. It will become perhaps possible to account for these effects in the parametrization step by considering a database that, besides the chemical shift values, includes information about the isotopic state of the attached hydrogen atoms (deuterated or not).

Next, a protein-based leave-one-out test is performed, in which repeatedly individual protein entries were removed from the model development data set with

subsequent parametrization and prediction of the corresponding chemical shifts. The protein-based RMSD of ArShift calculated in this way is 0.178 ± 0.065 ppm.

In order to increase the accuracy of the predictions, a self-consistent approach is used, in which the model optimisation and parametrization was done twice. After the initial model generation, the examination of the RMSD distribution from the protein-based leave-one-out test (Figure 3.6, top) revealed the existence of a high-RMSD shoulder next to the normal distribution of RMSD values centred at around 0.171 ppm.

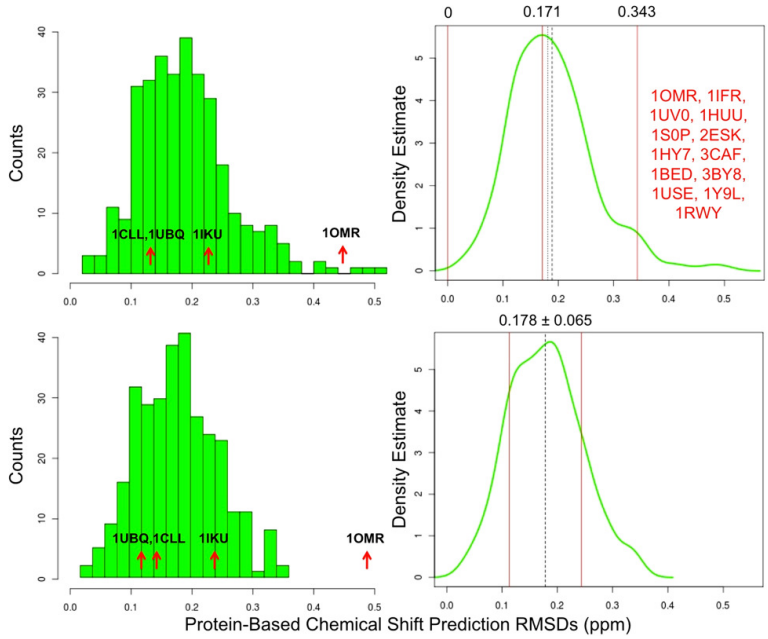


Figure 3.6: *Accuracy of the* ArShift *predictions in terms of RMSD distributions (in ppm) from the protein-based leave-one-out tests. Results before (top) and after (bottom) the exclusion in the parametrization of 13 outlier structures out of total 452 are shown.*

To this end, all the PDB IDs falling outside two standard deviations were further examined, revealing that in all these cases the X-ray structures were substantially different from the NMR ones, because of significant conformational changes upon Ca^{2+} or ligand binding, or sequence alterations (Figure 3.7).

Some X-ray structures were also lacking peptide segments that were present

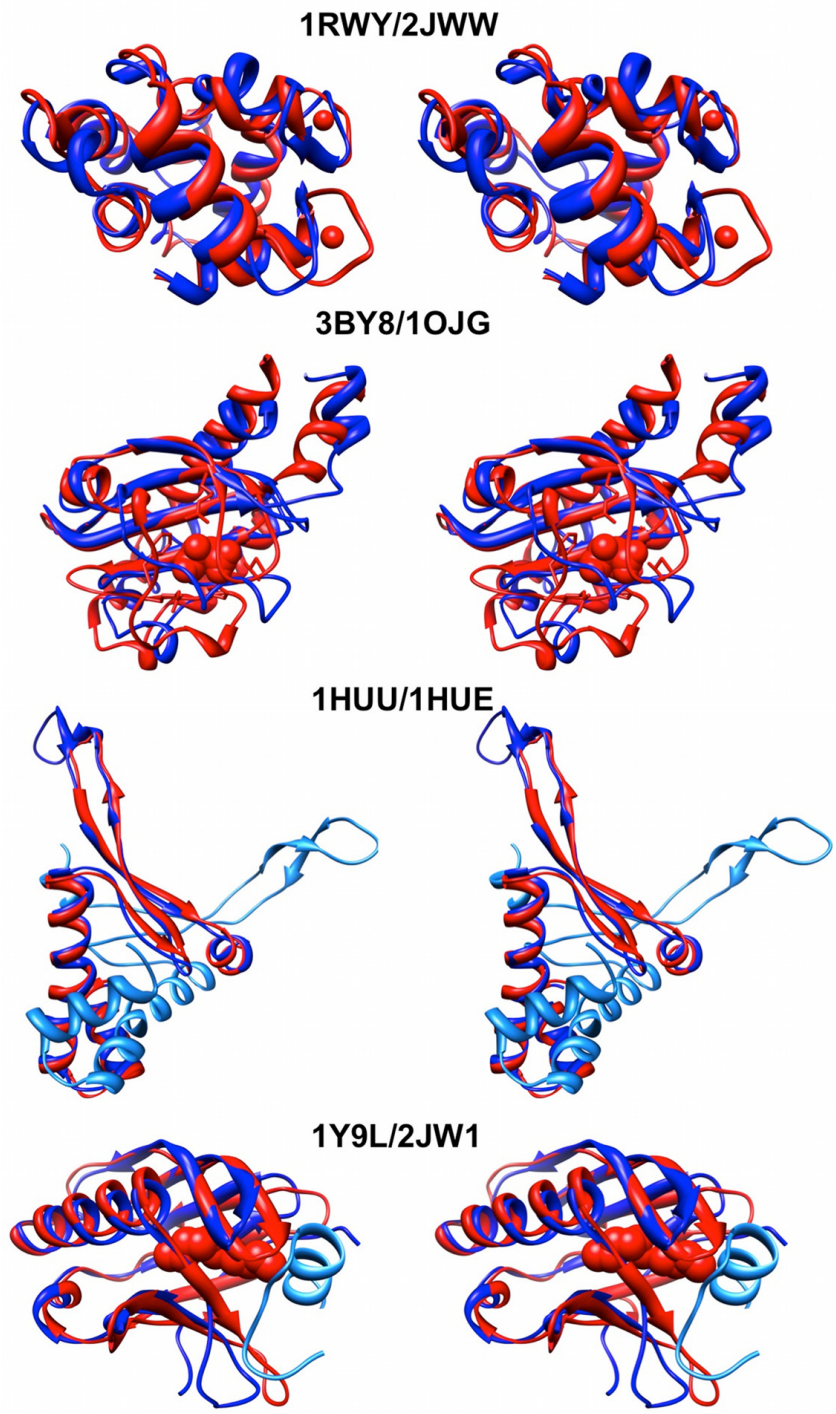


Figure 3.7: *Stereo view of representative cases identified by* ArShift *in which X-ray (red) and NMR structures (blue) differ significantly, for example because of* Ca^{2+} *or ligand binding, or missing segments.*

in the corresponding NMR structures (light blue moieties in Figure 3.7). Therefore, even though all the structures used in the parametrization process were determined in the crystal form, the first iteration of the model generation process resulted in a predictor that self-diagnosed the cases where the crystal structures did not match those in solution for which chemical shifts had been measured. This finding demonstrates that the high-resolution X-ray structures, used for the development of the predictor, do train coefficients that are not biased towards crystal structures.

After the removal of 13 proteins for which the predictions detected mismatches between X-ray and NMR structures, a second iteration of model optimisation and parametrization was done with the remaining 439 high-resolution X-ray structures, to generate the final predictor.

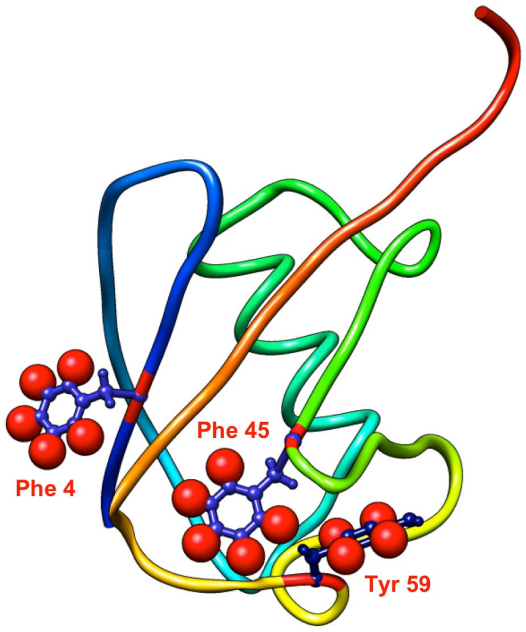


Figure 3.8: *Constituent Phe and Tyr aromatic side-chains in the structure of ubiquitin (1UBQ). The ^1H chemical shifts of these side-chains can be used to characterise the quality of the structure through the* ArShift *method.*

To further illustrate the applicability of the ArShift predictor, the 2K39 [Lange *et al.*, 2008], 2NR2 [Richter *et al.*, 2007] ensembles and the 1D3Z [Cornilescu

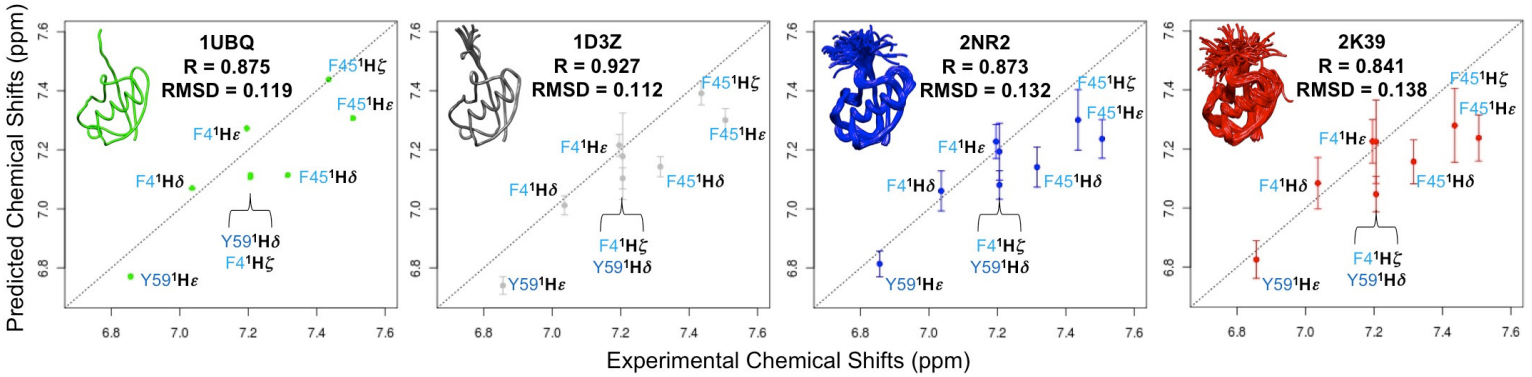


Figure 3.9: *Annotated plots representing the correlation between the predicted and experimental* ^1H *chemical shifts for the Phe and Tyr side-chains in three NMR ensembles and one X-ray structure of ubiquitin. The whiskers show the standard deviations of the predicted chemical shift values over the multiple conformers. The Pearson correlation coefficients and RMSDs are also shown.*

et al., 1998] set of structures are analysed in comparison to the 1UBQ [Vijay-Kumar *et al.*, 1987] X-ray structure of ubiquitin (Figures 3.8, 3.9, 3.10 and 3.11).

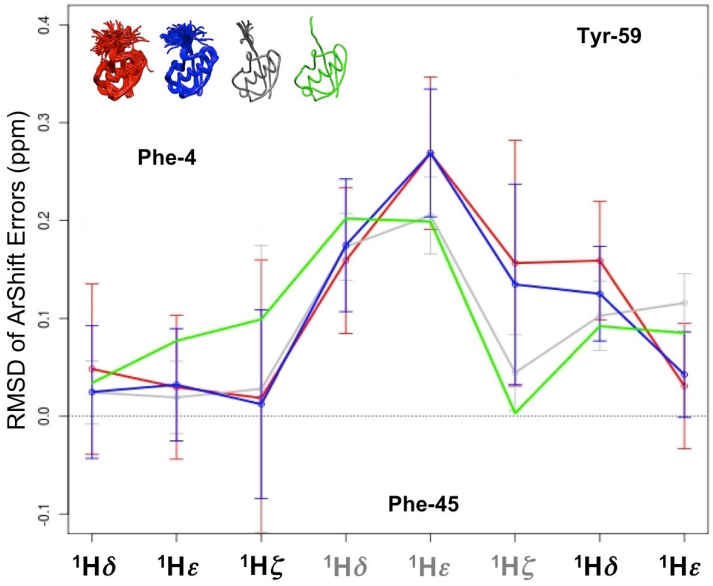


Figure 3.10: *Analysis of the differences (in ppm) between calculated and experimental aromatic side-chain* ^{1}*H chemical shifts for three NMR ensembles and one X-ray structure of ubiquitin: 2K39 (116 structures, red), 2NR2 (144 structures, blue), 1D3Z (10 structures, grey) and 1UBQ (green). The RMSDs of the average chemical shift prediction is shown with the whiskers indicating the standard deviations of the predicted chemical shift values over the conformers in the individual ensembles.*

The results indicate that the 1D3Z structure is the most consistent with the experimental aromatic side-chain ^{1}H chemical shifts, followed by 1UBQ, 2NR2 and 2K39 (Figures 3.9, 3.10 and 3.11). The 1D3Z set of NMR structures is not meant to represent the ensemble dynamics of the protein. However, the reason for the observed hierarchy of agreement could be that 1D3Z still contains structures, the aromatic residues of which are in geometric arrangement that altogether describe the system better than the 2NR2 and 2K39 ensembles. The current dynamical ensembles for ubiquitin, although covering the conformations present in 1D3Z, most probably do not populate the constituent conformations with the correct weights to reproduce experimental chemical shifts.

A similar test for a calmodulin X-ray structure (1CLL [Chattopadhyaya *et al.*,

1992]) and solution state ensemble (1X02 [Kainosho *et al.*, 2006]) highlights the overall good quality of the former as an average representation of the structure of this protein (Figures 3.12 and 3.13), as the aromatic side-chain ^{1}H chemical shifts back-calculated from the 1CLL structure are in very good agreement with the experimental chemical shifts [Kainosho *et al.*, 2006].

We also found that averaging the predicted aromatic chemical shifts over the 20 conformers in the 1X02 NMR ensemble improves the agreement between the predicted and experimental chemical shift values. An obvious exception from this trend is Phe-89, suggesting the presence of a possible imprecision in the structure or in the dynamics of this particular residue in the 1X02 ensemble.

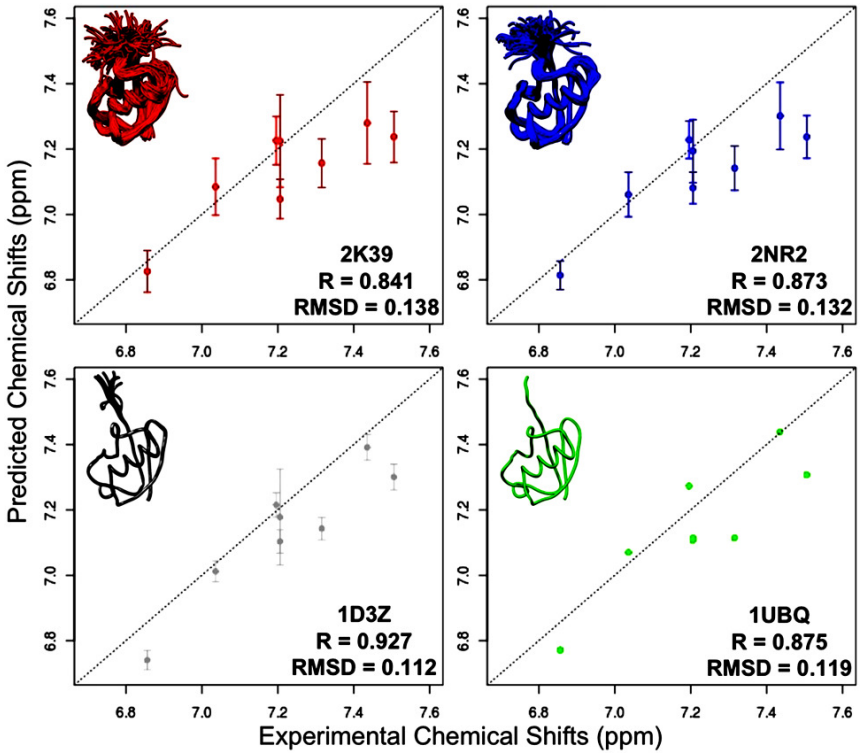


Figure 3.11: *Correlation between predicted and experimental ^{1}H chemical shifts for the Phe and Tyr side-chains in three NMR ensembles (2K39, 2NR2 and 1D3Z) and an X-ray structure of ubiquitin (1UBQ). Standard deviations of the predicted chemical shift values over multiple conformers are shown as error-bars. The Pearson correlation coefficients (R) and RMSDs (in ppm) are reported on the plots. The annotated version of this plot is presented in Figure 3.9.*

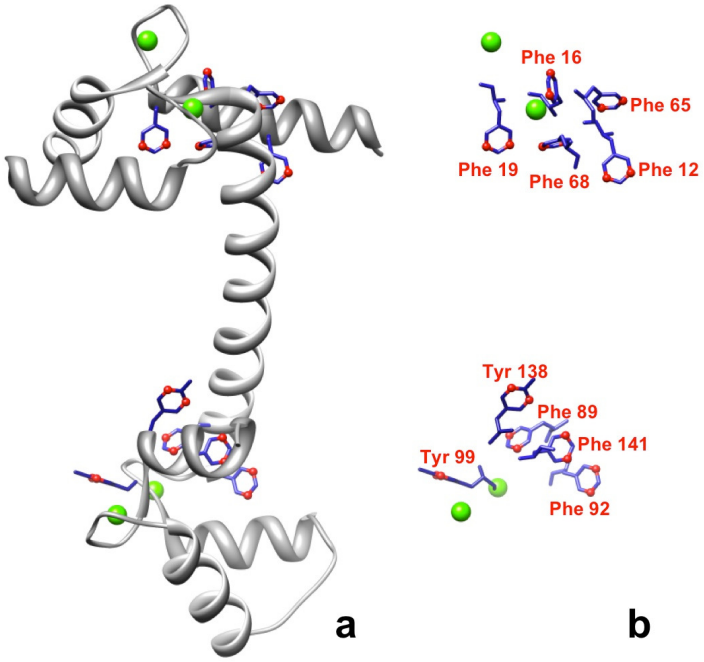


Figure 3.12: *X-ray structure of Ca^{2+}-bound calmodulin (1CLL, a). All the constituent Phe and Tyr side-chains are highlighted. The ϵ positions, for which ^{1}H chemical shifts have been measured through the SAIL labelling technique [Kainosho* et al.*, 2006], are coloured in red (b).*

A comparison with other existing prediction methods [Han *et al.*, 2011; Lehtivarjo *et al.*, 2009; Meiler, 2003; Xu & Case, 2001] illustrates the excellent performance of ArShift (Figures 3.14, 3.15 and 3.16). A test on recoverin [Tanaka *et al.*, 1995; Weiergräber *et al.*, 2003] in its Ca^{2+}-bound and free states, which substantially differ in their conformations, indicate that ArShift is more sensitive towards structural imperfections than the other methods that we considered (Figure 3.15).

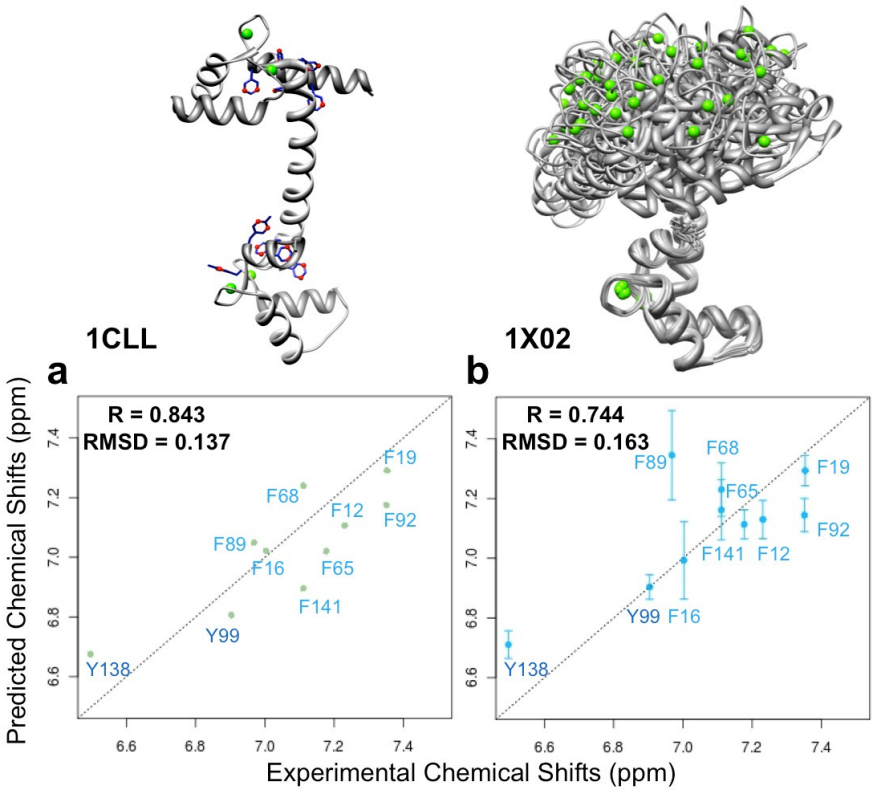


Figure 3.13: *Correlation between predicted and experimental aromatic* $^1H^\epsilon$ *chemical shifts for the 1CLL crystal structure and the 1X02 NMR ensemble of calmodulin. Standard deviation of the corresponding predicted chemical shift values over the constituent conformers in the ensemble are shown as error-bars. The Pearson correlation coefficients (R) and RMSDs (in ppm) are shown on the plots.*

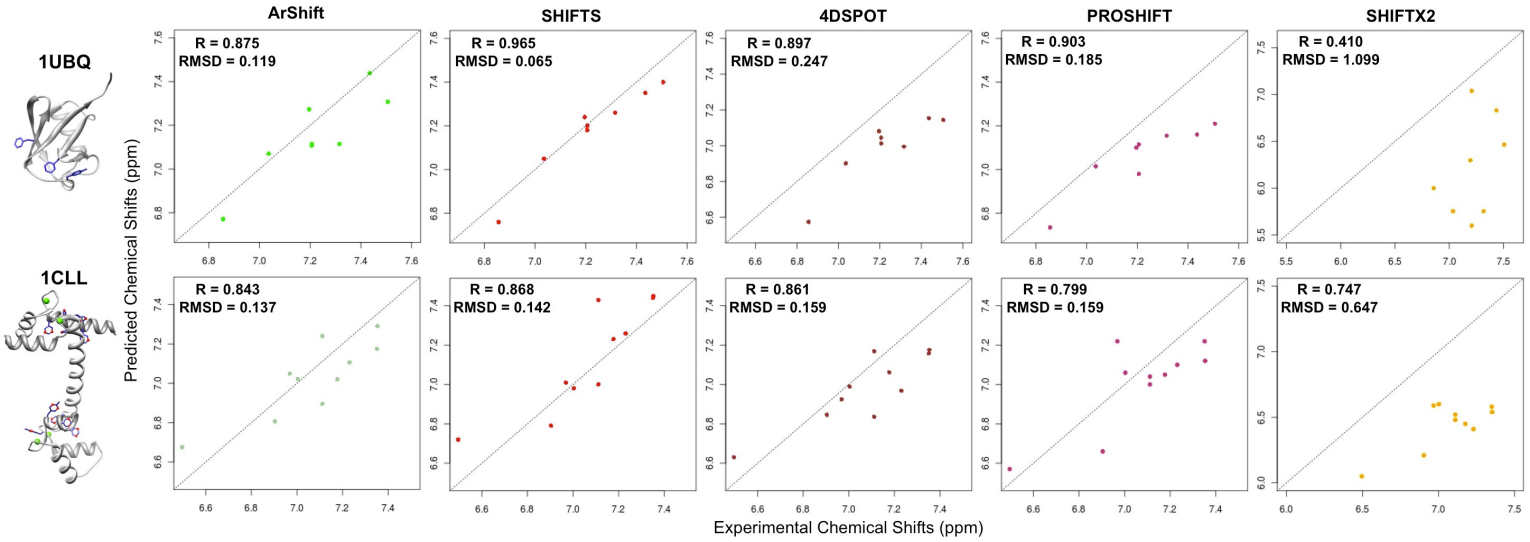


Figure 3.14: *Comparison between the* ArShift *aromatic side-chain* ^1H *chemical shift predictions and those of other existing methods: ShiftS, 4DSpot, PROSHIFT and ShiftX2. Two X-ray structures, 1UBQ of ubiquitin and 1CLL of calmodulin, are used in this example. The Pearson correlation coefficients and RMSDs (in ppm) are shown.*

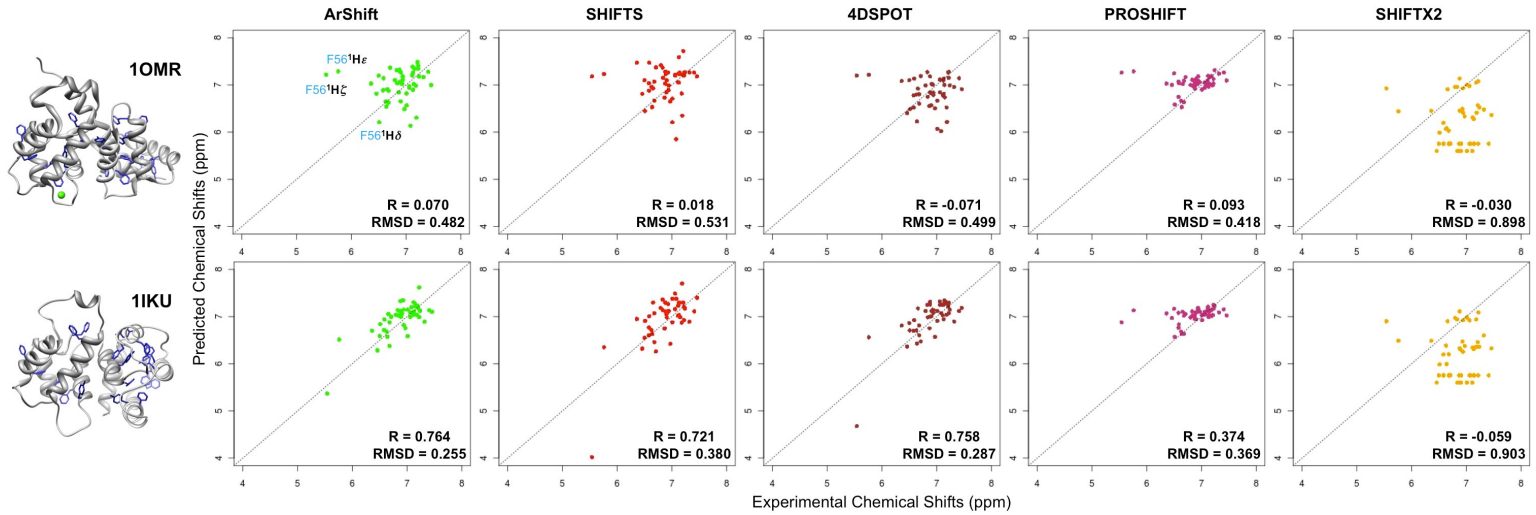


Figure 3.15: *Comparison between experimental and predicted aromatic side-chain ^1H chemical shifts of recoverin. In addition to* ArShift*, we considered four other existing prediction methods: ShiftS, 4DSpot, PROSHIFT and ShiftX2. Results are shown for the solution NMR structure (1IKU) of recoverin in the Ca^{2+}-free state and for the X-ray crystal structure (1OMR) in the Ca^{2+}-unbound state. The* ArShift *method differentiates the Ca^{2+}-bound and Ca^{2+}-free states more accurately than the other methods.*

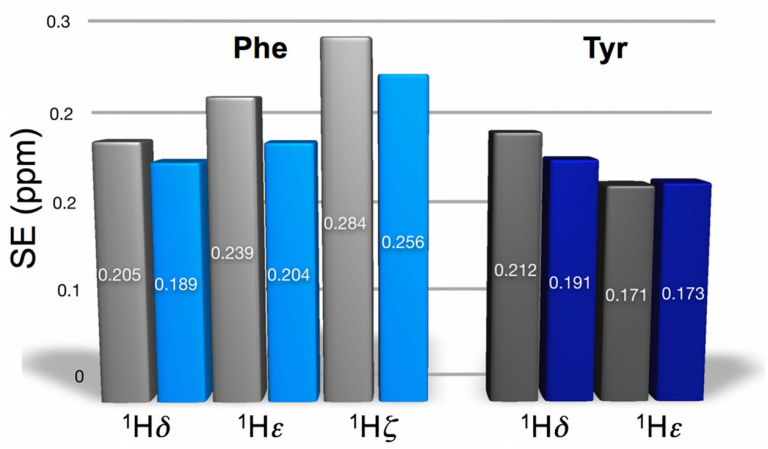


Figure 3.16: *Comparison of the* ^1H *chemical shift prediction performance of* ArShift *(blue for Phe and dark blue for Tyr residues) and ShiftS (grey). The bars show the standard errors of predictions in ppm.*

3.9 Testing of the usefulness of ArShift predictor in re-scoring molecular dynamics trajectories

To directly demonstrate that the ArShift predictor is sensitive towards structural imprecision and the aromatic chemical shifts can indeed be used to restrain molecular dynamics simulations for determining protein native ensembles, the 124-residue DNA binding domain of SV40 T-antigen is studied. The latter contains 10 Phe and 7 Tyr residues, of which 37 aromatic ^1H chemical shifts are available [Luo *et al.*, 1996]. The 2FUF X-ray structure [Meinke *et al.*, 2006], for which ArShift results in predictions with 0.161 ppm RMSD (Figure 3.17), has been used as a starting point.

17 ns molecular dynamics simulation is done in a non-native temperature range to unfold the structure (Figure 3.18).

Amber ff99SB force field [Hornak *et al.*, 2006] is used as implemented in the GROMACS package [van der Spoel *et al.*, 2006]. The 124-residue protein is then processed by adding hydrogens, resulting in a system with 2026 atoms. For the explicit solvation, an octahedron box with 9 Å minimum distance between the so-

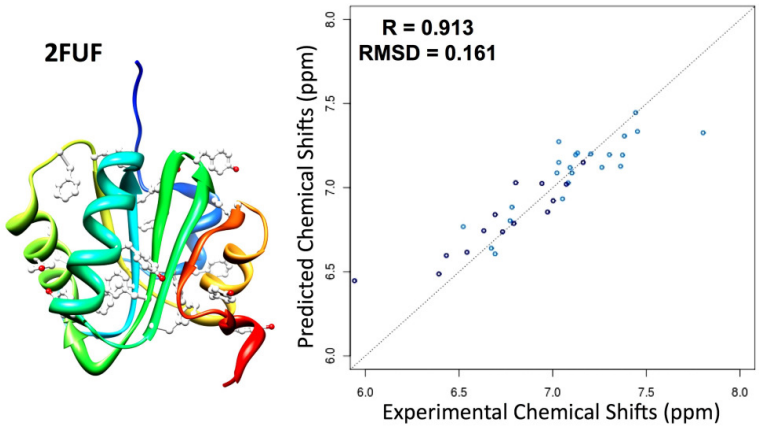


Figure 3.17: *The 2FUF crystal structure of the DNA binding domain of SV40 T-antigen and the performance of* ArShift *in predicting the experimental chemical shifts. The dark blue points indicate the Tyr residues, while the blue ones are coming from Phe residues. The Pearson correlation coefficient and RMSD are shown on the plot.*

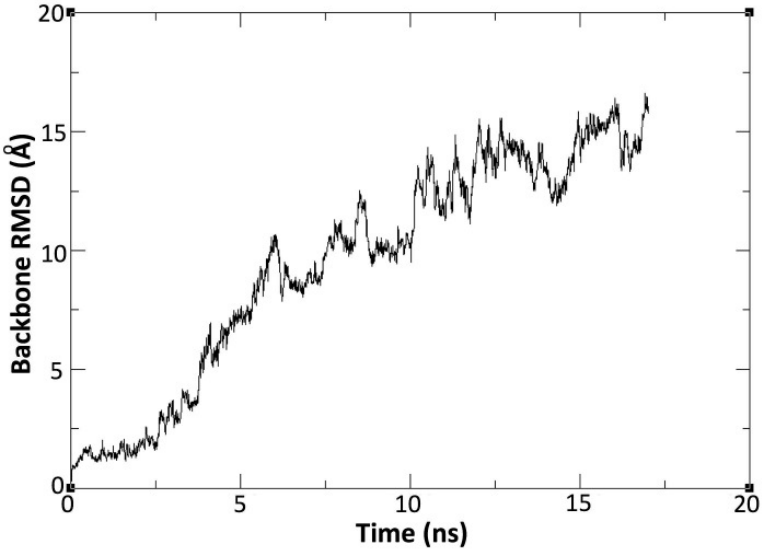


Figure 3.18: *The evolution of the backbone RMSD (in Å) of the DNA binding domain of SV40 T-antigen during the 17 ns unfolding simulation.*

lute and the box is used with 6440 TIP3P [Jorgensen *et al.*, 1983] water molecules and 4 Cl^- ions. A 9 Å cutoff distance is set for all non-bonded interactions. Particle mesh Ewald with 0.12 nm grid spacing is used. The system is then stabilised by 2000 steps of steepest descent geometry optimisation and 200 ps of position restrained simulation at a temperature of 298.15 K. The protein, water and ions have been separately coupled to v-rescale thermostat [Bussi *et al.*, 2007], that uses velocity rescaling with a stochastic term.

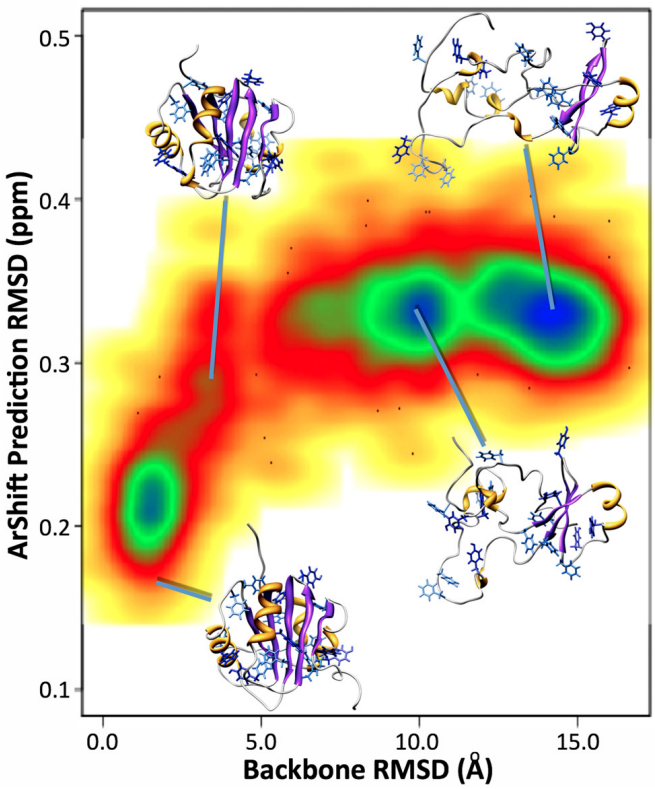


Figure 3.19: *The* ArShift *prediction RMSDs plotted against the backbone RMSDs of structures from the unfolding simulation of DNA-binding domain of SV40 T-antigen. Overall, 2430 structures have been analysed along the trajectory. The colour indicates the density of the data points on the graph. 25 points from the lowest density areas are explicitly shown.*

For the 17 ns production run (Figure 3.18), the temperature has been linearly increased from 298.15 K to 500 K during the initial 4 ns of simulation, with the continuation done at a constant 500 K temperature.

From the resulting trajectory, 2430 structures (every 7 ps) are then analysed using ArShift. The graph showing the connection between the chemical shift prediction RMSDs and the closeness of structure to its native state (Figure 3.19) is highly funnelled. Hence, this sensitivity can be used to bias molecular simulations and to score different protein structures in accordance to their quality.

3.10 Conclusions

A chemical shift predictor for ^{1}H atoms of Tyr and Phe side-chain aromatic moieties has been developed. The model is differentiable with respect to atomic coordinates and can thus be used in restrained molecular dynamics simulations. The performance of the prediction method is benchmarked against other chemical shift predictors for different model proteins. The usefulness of ArShift in scoring the quality of structures is demonstrated, pointing out its usefulness as a collective variable for exploring molecular dynamics trajectories in comparison to experimental chemical shifts. The ArShift parameters will be constantly improved as more experimental chemical shift measurements become available.

Be like a postage stamp. Stick to one thing until you get there.

Josh Billings

4

Validation of Protein Structures Using Side-Chain Chemical Shifts

4.1 Summary

A method of assessing the quality of the structures of proteins based on the use of side-chain NMR chemical shifts is presented. As these parameters are very accurate reporters of side-chain positions and are highly sensitive to tertiary structure and packing, they are particularly useful for structure validation. In order to analyse a given structure, a quality score, Q_{CS}, is defined that compares the chemical shifts calculated from such a structure with the corresponding experimental values in a way that takes account of the errors in the predictions. The results illustrate the advantages in the examination of the quality of protein structures from the perspective of side-chains.

4.2 Motivation

Owing to recent advances in genome sequencing [International Human Genome Sequencing Consortium, 2001; Venter *et al.*, 2001], the rate at which new protein-encoding genes are identified is far faster than the rate at which the structures of the corresponding proteins are determined. It is therefore important to develop methods to speed up the process of protein structure determination. Indeed, one of the major aims of structural genomics initiatives is to determine at least one representative three-dimensional structure for all known protein families [Burley *et al.*, 1999].

Although X-ray crystallography has a major role in these efforts, there is an interest in developing methods based on nuclear magnetic resonance (NMR) spectroscopy [Bax, 1994; Bax & Grishaev, 2005; Wüthrich, 2003] because they can be applied in the solution state, which closely resembles the conditions under which proteins carry out their functions, and because often proteins cannot be readily crystallised. Great advances in this direction have been made in the last 15 years, resulting in an increase in the precision and type of NMR measurements [Baldwin & Kay, 2009; Brutscher, 2001; Korzhnev *et al.*, 2002, 2010; Lundström *et al.*, 2009a; Tjandra & Bax, 1997], and in the size of proteins that can now be studied [Ruschak & Kay, 2010]. In this context, the introduction of the novel isotope labelling techniques [Goto & Kay, 2000; Kainosho *et al.*, 2006; Tugarinov *et al.*, 2006] is of particular importance since the knowledge of the chemical shifts of side-chain methyl and aromatic nuclei provides access to the solution-state structure and dynamics of super-molecular complexes [Ruschak & Kay, 2010].

In order to increase the role of NMR spectroscopy in structural genomics, it is very important to develop automated methods of data acquisition and processing. With such methodologies, the rate-limiting step of the structure determination procedure would become the preparation of samples [Heinemann *et al.*, 2001]. Standard NMR techniques for determining protein structures consist of multiple stages, some of which can require a substantial amount of time. The stages directly linked to NMR data acquisition and analysis include data recording, assignment of the spectra, interpretation of NOE (nuclear Overhauser enhancement) signals, and structure calculation and validation. To shorten the time required by

these stages, one approach is to substantially decrease the amount of data needed to resolve the structures of proteins. To this end, recent developments in methods that use chemical shifts to determine high-resolution protein structures [Cavalli *et al.*, 2007; Montalvao *et al.*, 2008; Robustelli *et al.*, 2010; Shen & et al, 2008] can become very helpful in increasing the throughput of NMR strategies, since the data acquisition is minimised to the most basic experiments and interpretation procedures needed to assign the resonance signals.

Regardless the nature of the strategies for speeding up NMR-based structure determination, the role of the structure validation increases substantially upon automation of the structure determination workflow. This aspect is particularly important since NMR spectroscopy, unlike X-ray crystallography, currently lacks consensus intrinsic measures of structural quality. Moreover, NMR structures are usually obtained by the aid of molecular mechanics force fields, as NMR measurements alone are generally not sufficient by themselves to completely define the three-dimensional structure of a protein. NMR data interpretation and processing are thus potentially prone to errors. Cases are known in which the misinterpretation of even a small number of NOE cross-peaks resulted in incorrect structures [Clore *et al.*, 1995; Lambert *et al.*, 2004]. For instance, it was demonstrated that the lack of knowledge about the oligomeric state of a protein may result in a misinterpretation of the spectra, so that homo-oligomeric protein complexes can be considered as monomeric structures [Nabuurs *et al.*, 2006].

4.3 Chemical shift based structural quality score

The approach and the applications reported here are based on recently developed structure-based predictors of protein side-chain chemical shifts [Sahakyan *et al.*, 2011a,b]. However, the same approach can be used with any chemical shift prediction engine. In the used predictors, chemical shifts are represented as a combination of phenomenological terms that report on the influence of dihedral angle, electric field, magnetic anisotropy and ring current effects on nuclear shielding [Sahakyan *et al.*, 2011a,b] and non-phenomenological distance-based terms that complete and increase the performance of the model [Kohlhoff *et al.*, 2009; Sahakyan *et al.*, 2011a,b].

An important aspect that should be considered in establishing a general structure validation method is that any structure-based chemical shift predictor (CSP) is associated with a certain error, which is defined here as the absolute difference between the predicted and experimental chemical shifts of the given query nucleus. A key component of the proposed method is that, having also the full profile of the CSP performance from the leave-one-out tests on a large database of proteins for each atom type, we can estimate the probability of the predictor to result in the observed error. Such probability estimates can be calculated by binning the absolute error scale and calculating the fraction of instances when the CSP results in an error within each bin-range. This kind of binning would, however, decrease the number of available entries for calculating the probability estimates and thus the statistical significance of the resulting numbers. To this end, the probability of the predictor to result in an error larger than the observed error is calculated, rather than the one within a certain bin. The resulting Q_{CS} score shows the probability that the prediction error is caused by the CSP rather than inaccuracies in the protein structure under analysis. A low value of the Q_{CS} score indicates the possible presence of problems in the structure. To further increase the statistical significance of the test, multiple chemical shifts are used from a given residue to extract joint probabilities. For instance, if two methyl ^{1}H chemical shift measurements are available from a valine residue (from $H^{\gamma 1}$ and $H^{\gamma 2}$ atoms), or if two or three signals are assigned for a phenylalanine residue belonging to any of the H^{δ}, H^{ϵ} and H^{ζ} atoms, then we can calculate the joint probabilities of all the NMR resolved nuclei in the given side-chain to end up in prediction errors larger than the observed ones. For a residue in a protein with measured experimental chemical shifts for the nuclei $i, j, ...$, the chemical shift based structural quality factor, Q_{CS}, can thus be presented as (Equation 4.1):

$$Q_{CS} = P(|\delta_i^{exp} - \delta_i^{calc}| \geq \sigma_i^{calc} \cap |\delta_j^{exp} - \delta_j^{calc}| \geq \sigma_j^{calc} \cap ...) \tag{4.1}$$

where $|\delta_i^{exp-calc}|$ is the absolute error of the chemical shift prediction for the nucleus i of the given residue, σ_i^{calc} is the standard error of the CSP in reproducing the experimental chemical shifts for the type of nucleus i in the given residue type, and the $\cap$ symbols for joint probabilities signify that all the conditions on the

different nuclei should be considered simultaneously.

The probabilities that are calculated here are fairly accurate because of the presence of relatively large database on which the chemical shift predictions are benchmarked via leave-one-out tests, as described in the previous chapters. In the current implementation, the carefully filtered database for aromatic side-chain protons uses 1796 entries for phenylalanine and 1498 entries for tyrosine hydrogen atoms coming from 452 proteins [Sahakyan *et al.*, 2011b]. The methyl group database behind the CH3Shift predictor uses 17873 chemical shift entries from proteins corresponding to 682 unique PDB identifiers [Sahakyan *et al.*, 2011a].

4.4 Examples of the Q_{CS} score application to validate protein structures

Chemical shifts are routinely measured during the initial stage of NMR data processing. Moreover, chemical shifts can often be measured even from very problematic systems, such as protein aggregates [Nielsen *et al.*, 2009] and intrinsically disordered proteins [Ágoston *et al.*, 2011]. The advances over the past decades that have been made to better understand the nature of chemical shifts [Jameson, 1996; Oldfield, 1995] and the developments of fast and efficient structure-based chemical shift prediction methods [Kohlhoff *et al.*, 2009; Neal *et al.*, 2003; Sahakyan *et al.*, 2011a,b; Xu & Case, 2001], have made it possible to substantially increase the scope of chemical shifts in structural biology. It is thus timely to extend the use of chemical shifts to protein structure validation.

Chemical shifts are extremely sensitive to specific structural features of protein conformations. Any change in the atomic environment of a given nucleus significantly alters its observed chemical shift value. Therefore any imprecision of the structure in the vicinity of the query atom or in the position of the query atom itself will become evident from the structure-based chemical shift predictions for that nucleus. Hence, if one can clearly differentiate between the errors that are normally expected from the given chemical shift prediction from the apparent errors that the prediction produces, a measure of structural imprecision at the given site of the protein structure can be devised, as described above. Previously,

protein backbone chemical shifts have been used to assess structural qualities of proteins by a comparison between experimental chemical shifts and those back-calculated from the protein structures under consideration using parametrizations based either on first principles [Vila *et al.*, 2009] or empirical [Berjanskii *et al.*, 2010] methods. The quality score for chemical shifts (Q_{CS}), described in this work, takes into account the errors intrinsic to the predictor and thus exploits chemical shifts for protein structure validation in a quantitative way. In addition, side-chain ^{1}H chemical shifts are particularly suitable for protein validation purpose, since they are strongly dependent on tertiary contacts, and, unlike backbone atoms, side-chains are not shielded from the surrounding by the other moieties of the same amino acid residue.

A low value of the Q_{CS} score indicates a possible structural imprecision, because the discrepancies between the experimental and calculated chemical shifts are larger than the intrinsic errors in the calculations of the chemical shift themselves; this method of course assumes that there are no assignment errors in the NMR spectra.

In the following, a series of applications is presented that demonstrates the usefulness of such structure validation approach. This method was prompted by the initial observation that the analysis of aromatic proton chemical shifts over 452 proteins in a database of high-resolution X-ray structures identified several proteins for which the prediction qualities were rather poor as assessed by the comparison to the experimental NMR measurements [Sahakyan *et al.*, 2011b]. As the predictions were performed through a protein-based leave-one-out tests, the predictor was not biased toward a particular protein because of that protein being involved in the parametrization. Examination of all the poorly performing structures revealed that all of them were different in conformation from the corresponding NMR structures obtained from the solution-state experiments from which the chemical shifts were measured [Sahakyan *et al.*, 2011b]. The reason for the difference was either a substantial conformational rearrangement upon Ca^{2+} ion or ligand binding or the presence of missing or extra peptide segments in either the solid or solution states. These observations clearly indicated that some errors resulting from the structure-based chemical shift predictions reveal actual structural inaccuracies in the structural model or a mismatch between the

experimental data and the structure that is evaluated against the experimental data.

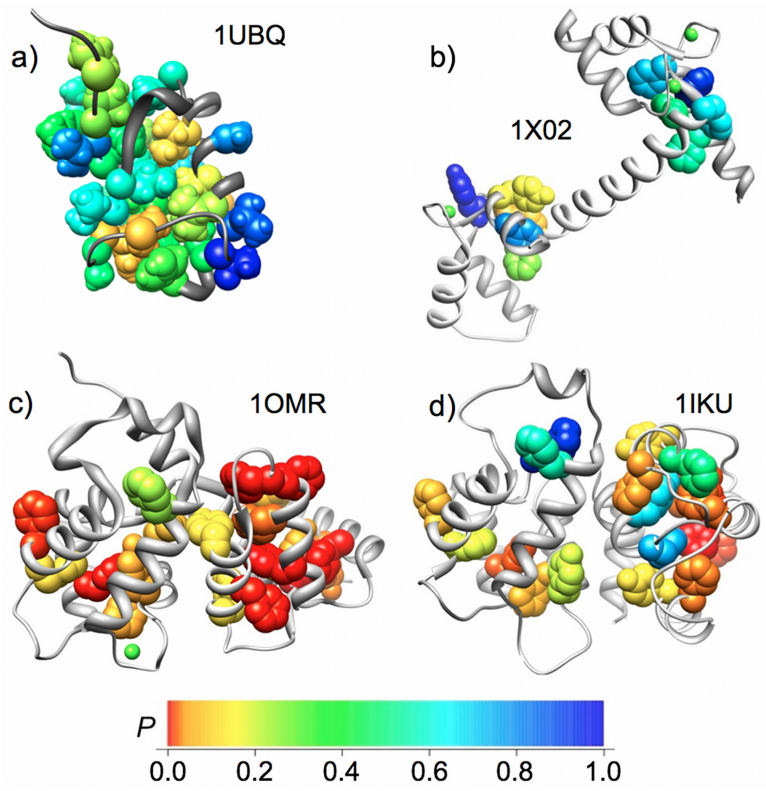


Figure 4.1: *Examples of protein structure validation based on side-chain chemical shifts. Side-chains bearing methyl or aromatic groups are shown in space-filling representation and coloured according to their Q_{SC} scores: (a) Ubiquitin (1UBQ); (b) Calmodulin (1X02); (c) X-ray structure (1OMR) and (d) NMR solution-state structure (1IKU) of Ca^{2+}-bound recoverin.*

The analyses of the Q_{CS} scores for four protein structures (Figure 4.1): ubiquitin (PDB id - 1UBQ, [Vijay-Kumar *et al.*, 1987] Figure 4.1a), calmodulin (1X02, [Kainosho *et al.*, 2006] Figure 4.1b) and recoverin in its Ca^{2+}-bound state obtained from X-ray crystallography (1OMR, [Weiergräber *et al.*, 2003] Figure 4.1c) and in its Ca^{2+}-free myristoyl-bound state obtained from solution state NMR (1IKU, [Tanaka *et al.*, 1995] Figure 4.1d) are presented. The residues that show low Q_{CS} scores are frequently clustered together, indicating the presence of a local problem in the structure. For instance, none of the NMR structural ensembles and the

X-ray structure of ubiquitin, that are analysed before reproduce the experimental ^{1}H chemical shifts for the aromatic Phe-45 ring [Sahakyan *et al.*, 2011b]. A map of the Q_{CS} scores obtained by analysing the X-ray 1UBQ structure [Vijay-Kumar *et al.*, 1987] (Figure 4.1a) shows that Leu-50, which is in the vicinity of Phe-45, is also showing low Q_{CS} scores as judged from the methyl group ^{1}H chemical shifts. This result suggests that Phe-45 may be undergoing complex conformational fluctuations that affect its own chemical shifts, as well as those of the sites close to it, and neither the X-ray structure of ubiquitin nor the NMR ensembles fully represent the dynamics of that residue. It is also interesting that the 2K39 ensemble of ubiquitin [Lange *et al.*, 2008] shows the presence of several rotameric states for Phe-45, but does not appear to capture the correct weights of different states since, like in the other ensembles, the average predictions of the aromatic protons of Phe-45 do not agree well with the experimental chemical shifts [Sahakyan *et al.*, 2011b].

A similar patchy behaviour (Figure 4.1b) of low Q_{CS} scores is found for the calmodulin 1X02 ensemble [Kainosho *et al.*, 2006], which indicates the possible presence of structural inaccuracies in the corresponding sites. In particular, the conformational fluctuations of the spatially neighbouring Tyr-138 and Phe-89 residues might not be fully represented by the 20 structures that comprise the ensemble [Sahakyan *et al.*, 2011b]. Another interesting case is that of recoverin [Sahakyan *et al.*, 2011b], for which two structural models are available, obtained from solid [Weiergräber *et al.*, 2003] and solution [Tanaka *et al.*, 1995] states. These structures are fairly dissimilar in conformation, primarily because the X-ray structure represents the Ca^{2+}-bound state. Since the NMR chemical shift measurements had been carried out for the Ca^{2+}-free solution state of recoverin, the solid-state structure is expected to result in an overall lower Q_{CS} score spread for the aromatic residues with available ^{1}H chemical shift measurements (Figures 4.1c and d). In addition, some imprecision in Q_{CS} scores even in Ca^{2+}-free state of recoverin can be explained by not accounting the unconventional (myristoylated) moiety of the protein by the chemical shift predictor.

As described in the previous chapters, the development of CSPs is based on a database of chemical shifts measured in solution state NMR and average structures compiled from a dataset of high-resolution X-ray structures [Sahakyan

et al., 2011a,b]. It is true that in many cases the side-chain chemical shifts are an outcome of complex dynamics between different rotameric states of side-chains, where a small population of such states with extreme values of chemical shifts can significantly alter the measurable average values of chemical shifts. However, it has been already demonstrated that the mass action of the large database of average high-resolution structures does train a good model with coefficients really reflecting the pure structure-to-chemical shift translation [Sahakyan *et al.*, 2011b]. By calculating the chemical shifts from a single, frozen, protein structure, we can rely on the non-biased nature of the used CSP and on the calculated chemical shifts as being the ones really corresponding to the supplied structure. Therefore the violations, observed while comparing the calculated and experimental chemical shift values from a particular side-chain, indicate that the used average structure is not a good representation of the state of a protein, which itself can be either because the real average structure is different, or because that particular side-chain possesses a complex dynamics. In fact, the latter explanation is most probably what we observe for the moiety surrounding/including the Phe-45 residue in ubiquitin, since one of the available NMR ensembles does capture different rotameric states [Lange *et al.*, 2008]. Hence, to really refine the population of such states, the grand proposal and the aim of all the developments in this thesis is the incorporation of such CSP-engines in restrained molecular dynamics simulations, through which more realistic ensembles capturing the invisible states of proteins can be obtained.

To directly demonstrate that the proposed quality score is sensitive to actual structural quality of proteins, the same unfolding trajectory of a protein with known X-ray structure [Sahakyan *et al.*, 2011b] is used, generated via a high temperature MD simulation. Since the unfolding is done *in silico* and we have the snapshots of the structures along the trajectory and hence the structural RMSDs relative to the crystallographic structure, we can check whether the worsening of the structural quality upon unfolding is also accompanied by the worsening of the Q_{SC}. The results are presented in Figure 4.2, where the negative sum of all the aromatic side-chain chemical shift based quality scores in the examined protein, $(-\sum Q_{SC})$, is plotted against the side-chain structural root-mean-squared deviations (RMSD in Å), indicating that Q_{SC} indeed reports on the structural

precision.

It is noteworthy that the deviation of Q_{SC} at the lower structural RMSD region is relatively greater as compared to its spread in the region corresponding to the completely unfolded structures (Figure 4.2), which might be an indication that accounting the structural dynamics is important at the native states and it is the averaged observable over the dynamical ensemble that results in trustworthy values.

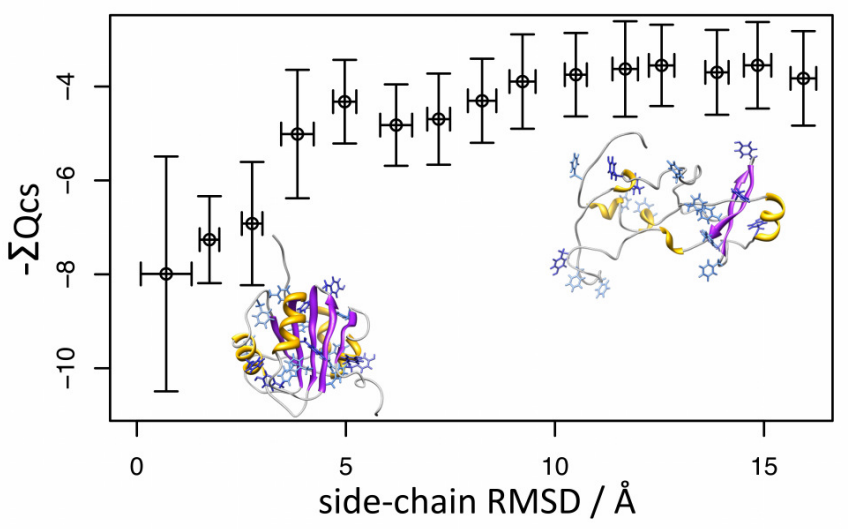


Figure 4.2: *The sum of all the aromatic side-chain chemical shift based quality scores (Q_{SC}) plotted against the side-chain structural root-mean-squared deviation (RMSD in Å) along the unfolding pathway of DNA-binding domain of SV40 T-antigen. The unfolding trajectory is obtained via a 17 ns high-temperature molecular dynamics simulation as described before [Sahakyan et al., 2011b]. Structural snapshots extracted at 7 ps intervals are analysed. The negative sign for the $\sum Q_{SC}$ is used to make the figure comparable to the similar landscapes in the original publication [Sahakyan et al., 2011b]. The data are obtained by averaging all the quality scores within 1.1 Å bins of structural RMSD. The whiskers indicate the standard deviations of both the structural RMSD (x-axis) and $-\sum Q_{SC}$ (y-axis) within the 1.1 Å bins of structural RMSD.*

The methods for calculating side-chain chemical shifts, as well as those that exploit such calculations for structure validation purposes, will be particularly useful to examine structural models of large proteins. In these cases, side-chain

chemical shifts are among the few NMR parameters that can be measured, in particular considering the recent advances in selective isotope labelling techniques for side-chains.[Goto & Kay, 2000; Kainosho *et al.*, 2006; Tugarinov *et al.*, 2006]. As a test on large systems for the validation technique introduced in this work, the largest single-chain protein studied by NMR spectroscopy, the 723-residue malate synthase G (MSG), is analysed, accounting for all the available structural models. Two models, 1P7T [Anstrom *et al.*, 2003] and 1D8C [Howard *et al.*, 2000], determined by X-ray crystallography at about 2.0 Å resolution, have been considered along with the 1Y8B ensemble of 10 NMR structures [Tugarinov *et al.*, 2005a] and the 2JQX solution structure [Grishaev *et al.*, 2008] refined against NMR and small-angle X-ray scattering (SAXS) data. Two structures from the 1P7T PDB entry that comprise the elementary cell have been considered separately for validation (1P7T_a and 1P7T_b). The missing segments in the X-ray structures have been modelled using the Modeller program [Fiser *et al.*, 2000] with 100 different structural variants created for the missing loops of each X-ray structure. By using multiple variants for the modelled loops, the influence of structural uncertainties arising from the *in silico* addition of missing segments in the X-ray structures is assessed.

Figure 4.3 shows the correlation graphs between the experimental [Sheppard *et al.*, 2009; Tugarinov & Kay, 2003] and calculated chemical shifts for all the available structures of MSG. As multiple structures are available, whiskers are included in the graphs to indicate the range of the calculated chemical shift variations in addition to the triangles that show the average chemical shift values.

The uncertainties in the modelled loop conformations of the X-ray structures of MSG affect only few chemical shifts (Figure 4.3), for which the whiskers indicate a fairly small variance because of the residues with available methyl group chemical shift measurements being distant from the modelled loops. Pearson correlation coefficients (R) and root-mean-squared deviations (RMSD) are shown on the graphs for each structural model of MSG. These results indicate that further experimental information will be required to improve the accuracy of the Ala, Val and Leu side-chain conformations, including the fine details in their three-dimensional packing. A plot of the Q_{CS} scores along the sequence of MSG, as well as their cumulative sum, report on the overall structural quality of the different

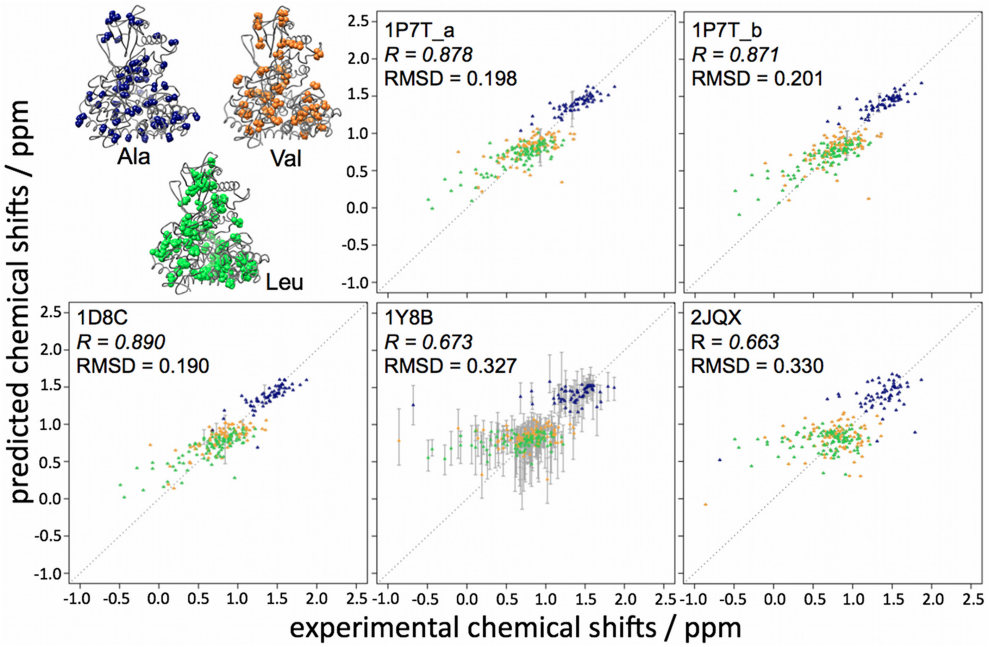


Figure 4.3: *Comparison between predicted and experimental chemical shifts (in ppm) for side-chain methyl hydrogen atoms of alanine (dark blue), valine (orange) and leucine (green) residues of the available structures and structural ensembles of malate synthase G determined by X-ray crystallography and NMR spectroscopy. PDB codes, Pearson correlation coefficients (R) and root-mean-squared deviations (RMSD) are shown for each case. The whiskers indicate the range of the predicted chemical shifts for the models comprised of multiple structures.*

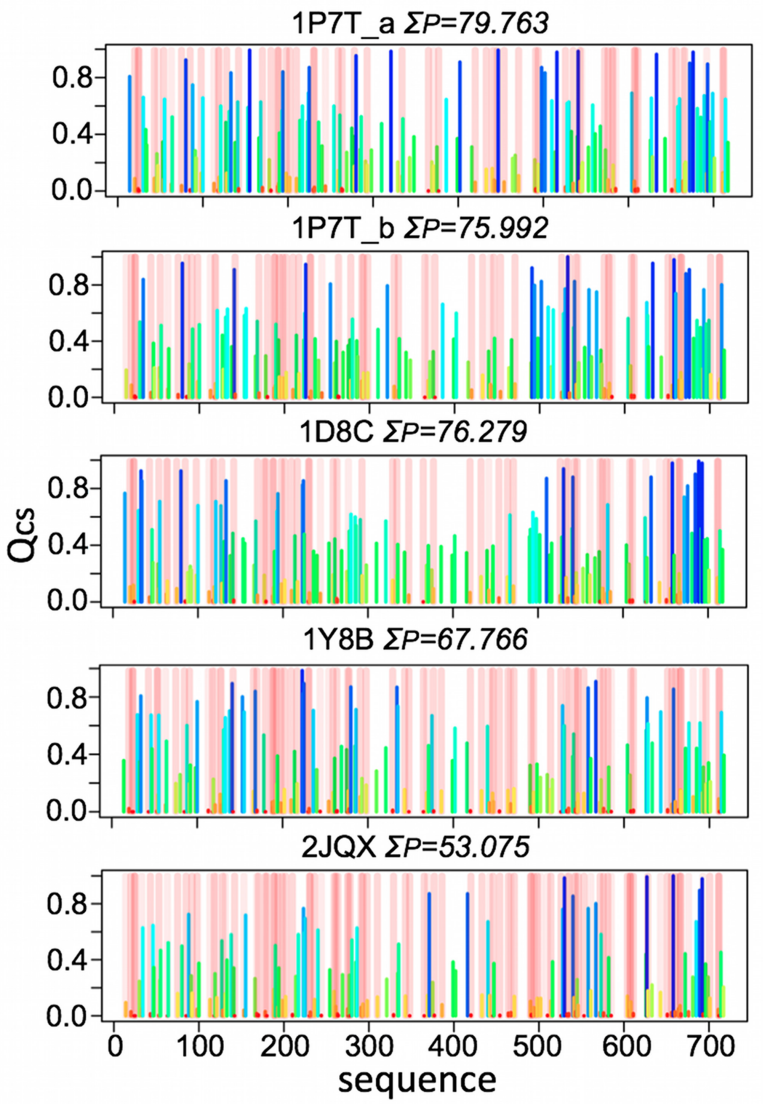


Figure 4.4: *Q_{SC} scores plotted against the sequence index of the methyl bearing amino acid residue in the X-ray structures and NMR ensembles of malate synthase G. The colours of the bars follow the same scale used in Figure 4.1. Transparent red bands identify the regions of the sequence for which the validation method predicts structural imprecision with high confidence. PDB codes and the total Q_{SC} scores are shown for each plot.*

structural models (Figure 4.4).

The heights of the bars are proportional to the Q_{CS} scores for the different residues, while the colours are from the same scale displayed in Figure 4.1. The transparent red bands indicate the regions where the confidence in structural inaccuracies is higher. Both the X-ray and NMR structures show inaccuracies in the representation of side-chain geometries and tertiary contacts (Figure 4.4). However, the NMR structures tend to give lower Q_{CS} scores in the tests of the structural quality as assessed from the side-chain perspective. Hence, the inclusion of side-chain NMR data in structure determination protocols is capable of improving the situation by increasing the accuracy of NMR for side-chains that form protein interior and exterior surfaces so important to study protein-ligand and protein-protein interactions.

4.5 Conclusions

A method of using chemical shifts to validate protein structures and identify regions of possible structural inaccuracies is described. Although the method can be readily used with any atom type for which a structure-based chemical shift predictor exists, we have focused our attention on side-chain proton chemical shifts because their values exhibit a strong dependence on tertiary interactions and spatial effects. By contrast, backbone or side-chain carbon chemical shifts are prevalently determined by backbone conformation, rotameric states and covalent interactions. Validation methods based on exclusively side-chain chemical shifts can exploit recent advances in labelling techniques, which are making it possible to measure side-chain chemical shifts for very large proteins and protein complexes by NMR [Goto & Kay, 2000; Kainosho *et al.*, 2006; Tugarinov *et al.*, 2006].

The availability of a chemical shift based approach for protein structure determination may offer several opportunities to the NMR community:

a) The method is based on NMR parameters, hence the complications involved in the use of other experimental techniques and measurements is not required.

b) The method uses NMR parameters that are generally measured, but not often used directly in NMR structure calculation. Indeed, NMR resonance signal assignment is the crucial first step in obtaining other parameters, such as RDCs

and NOE intensities, which are used in standard methods of protein structure determination. Thus all the measured RDC- and NOE-data can be used in structure determination, since we would not require some of them to be left out for further usage in structure validation.

c) Protein structures can be analysed from two different perspectives, those of backbone and side-chain atoms. Since backbone chemical shifts are more sensitive to the core effects and the conformation of peptide moieties, they report on the quality of the overall fold. On the contrary, side-chain chemical shifts, especially the ^{1}H ones, are very sensitive to the weak interactions between spatially adjacent atoms, hence being more sensitive towards the fine details of the three-dimensional packing.

d) As chemical shifts are the most basic parameters measured in NMR, protein validation methods based on these parameters can become the method of choice for future high-throughput protein structure determination protocols. This strategy will decrease the number of measurements and reduce the time required for the automatic analysis of spectra and structure determination. To this end, protein structure determination and validation methods based solely on backbone and side-chain chemical shifts can broaden the scope of NMR spectroscopy.

Truth, in science, can be defined as the working hypothesis best fitted to open the way to the next better one.

Konrad Lorenz

5

Towards the Structure-Based Chemical Shift Predictors for Nucleic Acids

5.1 Summary

Ring current effects are one of the most influential factors defining NMR chemical shifts. Thorough studies of that effect are particularly important for the future development of accurate structure-based predictors of chemical shifts for nucleic acids, where the ring current effects dominate the local magnetic fields and the conjugated systems populate largely variable relative arrangements. In this study, the classical Pople and, derived from empirical quantum mechanics, Haigh-Mallion models for ring current effects are compared from the viewpoint of their usage and applicability in restrained molecular dynamics simulations of biomolecules. X-ray structures of ribonucleic acids (RNAs) are analysed and a database (DiBaseRNA) of three-dimensional arrangements of nucleic acid base pairs is generated, ready for further non-empirical quantum mechanical studies.

The database is also of importance for force field parametrizations and studies on hydrogen bonding in nucleic acids. DiBaseRNA is then used to calculate chemical shifts via a hybrid density functional theory approach. First principle studies are performed for a thorough and contemporary parametrization of the ring current and electric field effects in nucleic acid bases, making use of the simplified Pople and Haigh-Mallion frameworks for ^{1}H, ^{13}C, ^{15}N and ^{17}O nuclei. The coupling of the electric field and ring current effects is studied for all the nuclei using linear model fitting with joint electric field and ring current, as well as only electric field and only ring current approximations for different interring arrangements found in RNA bases. The interdependence of ring current and electric field geometric factors is outlined, proven to be especially important for non-hydrogen atoms. A web server, RingPar, is generated for accessing and exploring all the coefficients from such fittings. The new parameters from some of the fitting schemes quantum mechanically eliminate the electric field influence, take into account the structural variance and are biased towards the interring arrangements found in RNA structures.

5.2 Motivation

Although the chemical shift based protein structure determination has become an important tool for biomolecular studies during the past decade [Case, 1998; Cavalli *et al.*, 2007; Kohlhoff *et al.*, 2009; Oldfield, 2002; Robustelli *et al.*, 2010; Sahakyan *et al.*, 2011a,b; Shen & et al, 2008], the same approach is unfortunately not widely applied to nucleic acids. This can be because of the absence of structure-based and differentiable chemical shift predictors for nucleic acids that would also possess a sufficient precision in predicting structure-induced chemical shift variation for individual atom types in RNAs and DNAs. Previous studies have demonstrated, however, that the established ideas and dependencies of chemical shifts are transferrable to nucleic acids [Cromsigt *et al.*, 2001; Lam & Chi, 2010; Wijmenga & van Buuren, 1998; Wijmenga *et al.*, 1997], and further developments can potentially result in a better precision of such predictors. The major reason for the lack of the robust prediction models so far was the relatively modest presence of high quality structural and NMR data in corresponding pub-

lic databases. Therefore, to better operate on such sparse data and increase the precision of nucleic acid chemical shift predictors for direct structural studies, it is of utmost importance to have a deep understanding of the fine details in the factors that modulate chemical shifts in nucleic acids. The most powerful factor affecting chemical shift values is known to be the, so called, ring current effect, which is even more pronounced in nucleic acids, since they are particularly full of conjugated rings. Therefore, the logical first step would be to revisit the study of the ring current effect in nucleic acids, this time also investigating its influence on non-hydrogen atoms and examining the coupling with different conformational and electric field effects. The latter modulator of chemical shifts is also expected to be more relevant to nucleic acids, taking into account the highly charged nature of polynucleotide chains.

Biomolecular ring current effect on ^{1}H chemical shifts has been thoroughly studied by Case and coworkers [Case, 1995, 1998; Ösapay & Case, 1991] and parametrized for nucleic acid bases [Case, 1995] through the Haigh-Mallion [Haigh & Mallion, 1972, 1980] and Waugh-Fessenden-Johnson-Bovey [Johnson & Bovey, 1958; Waugh & Fessenden, 1957] models. The current utilisation of chemical shift predictors for direct structural studies dictates the suitability of Pople point dipole [Pople, 1956] and Haigh-Mallion models in empirically describing ring currents, with the preference being towards Pople model owing to its simplicity and efficiency of implementation in molecular dynamics codes.

In this work, the ring current effect is revisited by thoroughly comparing the Pople and Haigh-Mallion models. Equations are proposed for easily migrating from one model to another, in order to either increase the computational efficiency of existing implementations that already use the more complex Haigh-Mallion model, or to increase the accuracy of simpler point dipole moment. We then focus on nucleic acids, this time extending the study on the heavy nuclei (^{13}C, ^{15}N and ^{17}O that has a prospective importance) and investigating the interdependence of ring current and electric field terms in modulating chemical shift values. As model systems for such studies, real interring arrangement in nucleic acid bases are considered by generating a di-base atlas of base positions for all the possible pairs found in a high resolution RNA structure database [Murray *et al.*, 2003].

Taking into account the varying sign convention used in different works where

ring current effects are discussed, the most widely used ring current models are reviewed below in a unified notification and sign convention, to prevent any possible confusion in future works. With the majority of the results being transferrable to any 6- and 5-membered rings, this study can be useful for solving wide range of problems that involve the analysis of ring current contributions in nuclear shielding phenomenon.

5.3 Methods

All the quantum mechanical calculations are done using Gaussian 03 suite of programs [M. J. Frisch et al, 2004]. All the scripting and the linear model fitting in the work are done using the *R* programming language for statistical computing [R Development Core Team, 2011]. The RingPar web server for accessing ring current and electric field parameters is created using *Rwui*, an interface to generate web servers based on *R*-scripts [Newton & Wernisch, 2007]. The server and the generated DiBaseRNA database can be accessed via the following address: `http://www-sidechain.ch.cam.ac.uk/RingPar`.

Further details on the performed calculations and database analyses are presented along with the discussion in the sections below.

5.4 Ring current models

Since the successful estimation of diamagnetic anisotropy of crystalline benzene by Lonsdale's and Pauling's assumption of π-electron precession along the ring atoms [Lonsdale, 1937; Pauling, 1936], the ring current concept has become one of the most discussed aspects of NMR spectroscopy. Subsequently, theoretical models of ring current estimation emerged with both classical and quantum mechanical approaches [Haigh & Mallion, 1980], which were also accompanied by empirical look-up tables that widened the usage of ring current evaluations at around simple conjugated systems [Haigh & Mallion, 1972; Johnson & Bovey, 1958]. Of the numerous theoretical models [Haigh & Mallion, 1980], three have received considerable attention owing to the ease of their implementation [Pople, 1956] and the availability of the derived empirical tables [Haigh & Mallion, 1972;

Johnson & Bovey, 1958]. The majority of suggested methods are reviewed in detail elsewhere [Haigh & Mallion, 1980]. However, taking into account the confusing and frequent mismatch of signs observed in different publications as the convention of notations of nuclear shielding constant and chemical shifts have been changing over decades, below is the brief unified description of the three most implemented frameworks, stated in order of their appearance and completeness.

The assumption of the secondary magnetic field, created by the electric current circulating in benzene ring [Lonsdale, 1937; Pauling, 1936], was soon followed by the simplification of the magnetic field description [Pople, 1956], where the source of the magnetic field is approximated by a magnetic dipole at the ring centre. The magnetic dipole holds a magnitude of $ne^2a^2B_0/4\pi mc^2$, where n is the number of circulating electrons, a is the radius of the ring (usually, 1.39 Å is taken for benzene ring which equals to the C-C bond length), B_0 is the applied uniform magnetic field, e, m and c hold their conventional meaning and in part emerge from the expression for precession frequency ($\omega_L = -eB_0/2mc$). The secondary or induced magnetic field B_{ind} at any point around that dipole is determined by the expression $B_{ind} = ne^2a^2B_0(1 - 3cos^2\theta)/4\pi mc^2r^3$ with θ being the angle between the query point and the ring normal, and r being the distance from the ring centre (Figure 5.1a). Taking into account that $B_{ind} = -\sigma B_0$, where σ is the isotropic nuclear shielding constant, the expression for the change in σ_{ring} isotropic nuclear shielding constant in ppm originated by the ring current effect according to the Pople point dipole model can be written as (Equation 5.1):

$$\Delta\sigma^P_{ring} = 10^6 \times \frac{ne^2a^2}{4\pi mc^2} \times \frac{3cos^2\theta - 1}{r^3} \tag{5.1}$$

Please note, that the chemical shift is the negative of the nuclear shielding constant, therefore the geometric factor $(1 - 3cos^2\theta)/r^3$, especially when geometric terms of different models are outlined in comparison, should be used only for chemical shifts, not for the nuclear shielding constants. Furthermore, the interconnection between the local or effective B_{loc}, applied B_0 magnetic fields and the isotropic nuclear shielding constant is given by the expression $B_{loc} = B_0(1 - \sigma)$. On the other hand, $B_{loc} = B_0 + B_{ind}$. In many summaries and explanations, an

expression $B_{loc} = B_0 - B_{ind}$ can be found, because the induced magnetic field usually opposes the external one. However, only the first expression, where the negative sign is embedded within B_{ind}, is in correspondence with the sign and concept of nuclear shielding constant, hence is advised to be used for derivations.

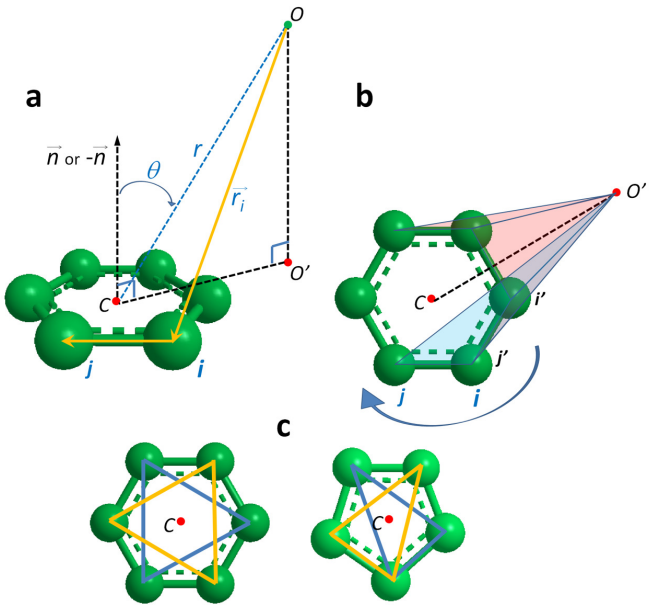


Figure 5.1: *Schematic representation of the geometric concepts used in the Pople point dipole (a) and Haigh-Mallion (a, b) models of ring currents, along with our suggested way of determining the ring normals (c) for 6- and 5-membered rings, that will result in geometric factors more stable and robust against out-of-plane geometric fluctuations of the constituent ring atoms. For further details, see the text.*

A more detailed classical description of the ring currents was suggested by Waugh and Fessenden [Waugh & Fessenden, 1957], then corrected and parametrized for benzene by Johnson and Bovey [Johnson & Bovey, 1958]. Here, the complete classical description of the electric current circulating in a loop of radius a is considered. The ring current model was also extended to account for the nature of the π-orbitals by assigning two loops, above and below the ring plane and separated by $\Delta z = z_2 - z_1$. In this case, each of the two loops will possess $n/2$ circulating electrons, thus the general form of the equation for $\Delta\sigma_{ring}$ in ppm is

given below (Equation 5.2):

$$\Delta\sigma_{ring}^{WFJB} = 10^6 \times \frac{ne^2}{12\pi mc^2 a} \times \sum_{p=1}^{2} \left(\frac{1}{\sqrt{(1+\rho)^2 + z_p^2}} \left(K(k) + \frac{1-\rho^2-z_p^2}{(1-\rho)^2+z_p^2} E(k) \right) \right) \tag{5.2}$$

where cylindrical coordinate system centered at the ring center is used with z and ρ expressed in the units of loop radius a. K and E are the complete elliptic integrals of the argument k, which is defined by the expression $\sqrt{4\rho/[(1+\rho)^2+z_p^2]}$. A 1.28 Å ($0.918a$, with radius a taken to be equal to the benzene ring radius) separation between the loops was found to be the optimal to represent the hydrogen shielding in benzene [Johnson & Bovey, 1958]. To escape possible confusion, it should be noted that, although chemical shift δ notation was used in the original articles [Johnson & Bovey, 1958; Waugh & Fessenden, 1957], the expression reflects the $\Delta\sigma_{ring}$ change in isotropic nuclear shielding constant and a minus sign should be used to get $\Delta\delta_{ring}$. In addition, the H_{eff} notation in the original publications denote the induced field B_{ind}, not the effective local field B_{loc}.

The third popular ring current model was proposed by Haigh and Mallion, based on the London and McWeeny approximations and Hückel molecular orbital theory [Haigh & Mallion, 1972, 1980]. Thus, the method can be regarded as empirical quantum mechanical (QM) model. In its simplified representation, the secondary magnetic field obeys the $B_{ind} \sim -J_{ring} \sum_{ij}[S_{ij}(1/r_i^3 + 1/r_j^3)]$ proportionality. There, J_{ring} is a quantum mechanical quantity calculated from the Coulson bond orders P_{ij} and, so called, mutual bond-bond polarisabilities $\overline{\pi}_{(ij)(kl)}$ within the Hückel formalism. S_{ij} is the algebraic (signed) triangle area formed by the O' projection of the query point O onto the ring plane and the ring atoms i and j (Figure 5.1a and b). Denoting $\mathbf{T}_{O'i}$ and $\mathbf{T}_{ij}$ as vectors joining O' to the ring atom i and ring atom i to j respectively, the sign of the triangle is positive if the vector product $\mathbf{T}_{O'i} \times \mathbf{T}_{ij}$ has the same direction as the ring normal with ring atoms counted in $i \rightarrow j$ direction. r_i and r_j are the distances between O and atoms i and j respectively. The summation goes over all the adjacent ij

atom pairs forming the ring, thus with number of constituents being equal to the number of bonds in the conjugated ring. An expression for the ring current contribution to the change of the isotropic nuclear shielding constant follows (Equation 5.3):

$$\Delta\sigma_{ring}^{HM} = 10^6 \times KJ_{ring} \times \sum_{ij} S_{ij} \left(\frac{1}{r_i^3} + \frac{1}{r_j^3} \right) \tag{5.3}$$

where K is a proportionality constant, and, the minus sign is discarded to follow the $\sigma = -B_{ind}/B_0$ definition for the nuclear shielding constant. The minus sign was present in the reference [Haigh & Mallion, 1972], which was canceled by the negative parameter, calculated for benzene. The comparative mismatch of the signs continues in the reference [Neal *et al.*, 2003], where the Haigh-Mallion geometric factor, in a form consistent with nuclear shielding constant, was used to calculate chemical shift difference ($\Delta\delta = -\Delta\sigma$). However, the further reverse definition of the algebraic sign of the S_{ij} triangle areas changed the sign of the expression making consistent with chemical shifts.

Paying attention to the last factors in Equations 5.1, 5.2 and 5.3, one can see that in each of the described M models, the $\Delta\sigma_{ring}^{M}$ can be represented as a geometric factor $G^M(\overrightarrow{\mathbf{r}})$ multiplied by a proportionality constant K^M. The geometric factor itself only describes the geometric arrangement of the query point relative to the conjugated ring, where the shielding needs to be evaluated. An exception is the geometric factor $G^{WFJB}(\overrightarrow{\mathbf{r}})$ of the Waugh-Fessenden-Johnson-Bovey model, which also includes an adjustable parameter Δz, describing the separation between the two loops with circulating electrons above and below the ring plane [Johnson & Bovey, 1958; Waugh & Fessenden, 1957].

In a more general case of the conjugated system being comprised of multiple cycles, the ring current effect on nuclear shielding constant can be represented as a sum of the effects from each cycle c (Equation 5.4):

$$\Delta\sigma_{ring}^{M} = \sum_{c} K_c^M G_c^M(\overrightarrow{\mathbf{r}}) \tag{5.4}$$

Please note, that the geometric factor in Equation 5.4 has the same sign as the change in the nuclear shielding constant $\Delta\sigma_{ring}^{M}$ and the reverse sign of the

change in chemical shifts $\Delta\delta^{M}_{ring}$. The $G^{M}_{c}(\overrightarrow{\mathbf{r}})$ is determined by the corresponding geometric factors in Equations 5.1, 5.2 and 5.3.

5.5 Comparative analysis of Pople and Haigh-Mallion ring current models on benzene

Taking into account the general intention in studying the ring current models that are intrinsically convenient for implementation in restrained molecular dynamics simulations, the discussion is continued for Pople point dipole [Pople, 1956] and Haigh-Mallion [Haigh & Mallion, 1972, 1980] models only. Although the Waugh-Fessenden-Johnson-Bovey [Johnson & Bovey, 1958; Waugh & Fessenden, 1957] model is differentiable, its geometric factor, unlike the factors from the other models, contains an adjustable parameter that should be optimised for different conjugated systems. The Haigh-Mallion model has become the most applied framework for ring current induced chemical shift change evaluation, even though a study demonstrates that, if thoroughly parametrized, the precision of the point dipole model is comparable to Waugh-Fessenden-Johnson-Bovey model and is only slightly worse than the performance of the Haigh-Mallion model [Moyna *et al.*, 1998].

Nucleic acids are literally constructed of different conjugated rings with conformational distributions covering highly shielded and deshielded regions, from very close, stacked, to planar, hydrogen bonded, states. Therefore the success of the development of structure-based chemical shift predictors for nucleic acids is expected to be highly dependent on the quality of the ring current description. We can thus return to the same problem of the comparison of the ring current models, this time also asking if errors are to be expected, which interring arrangements are the most error prone.

The Pople and Haigh-Mallion models are implemented following the convention described above. Our experience shows that the main reason for instabilities in such implementations, for either chemical shift prediction or restrained molecular dynamics code, is the high level of fluctuations of the ring normal, which should be calculated in both models. The fluctuations of the ring normal ori-

entation occur because of slight out-of-plane displacements of the ring atoms. To this end, it would be better to adopt the usage of two planes defined by non-interconnected set of atoms for each plane (Figure 5.1c), so that two ring normals are inferred in order to take the average of the two vectors.

We can now begin from the simplest and the most studied case, the benzene molecule. The molecular structure of benzene is geometry optimised using hybrid-DFT (density functional theory [Kohn & Sham, 1965]) with B3LYP exchange-correlation functional [Becke, 1993; Lee *et al.*, 1988; Miehlich *et al.*, 1989] and 6-311+G(d,p) split-valence basis set [Krishnan *et al.*, 1980]. The geometry optimisation is done with D_{6h} symmetry constraints and tight convergence criteria. This has resulted in a C-C bond length, further taken as ring radius a, equal to 1.3946 Å, which is quite close to the widely accepted zero-point average C-C distance for benzene, 1.395 Å [Tamagawa *et al.*, 1976].

The ring current geometric factors, from both Pople and Haigh-Mallion models, are then calculated for 100000 points uniformly distributed around the benzene ring in the orthogonal plane corresponding to $\phi = 0$ within the cylindrical coordinates ranging from 0 to 4 units of benzene radius a for both z and ρ coordinates relative to the benzene ring (Figure 5.2).

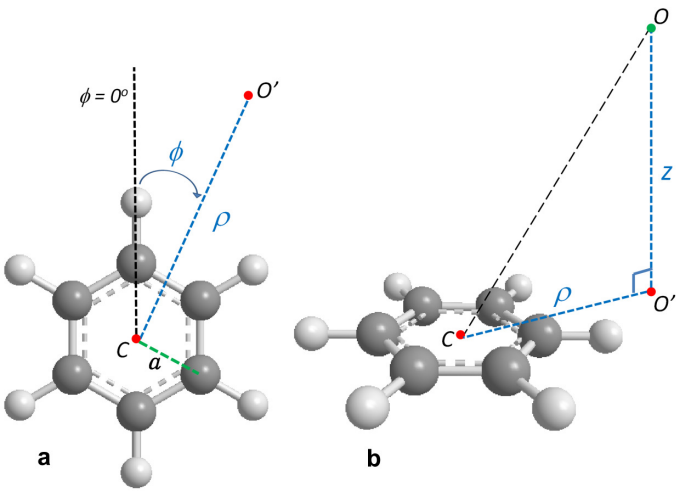


Figure 5.2: *The cylindrical coordinates and the notation associated with any O point around the benzene ring.*

The resulting maps are shown for both Pople (Figure 5.3a) and Haigh-Mallion (Figure 5.3b) geometric factors. The white-coloured area in the maps (Figure 5.3a and b) depicts the discarded region with extremely large absolute values of the geometric factor.

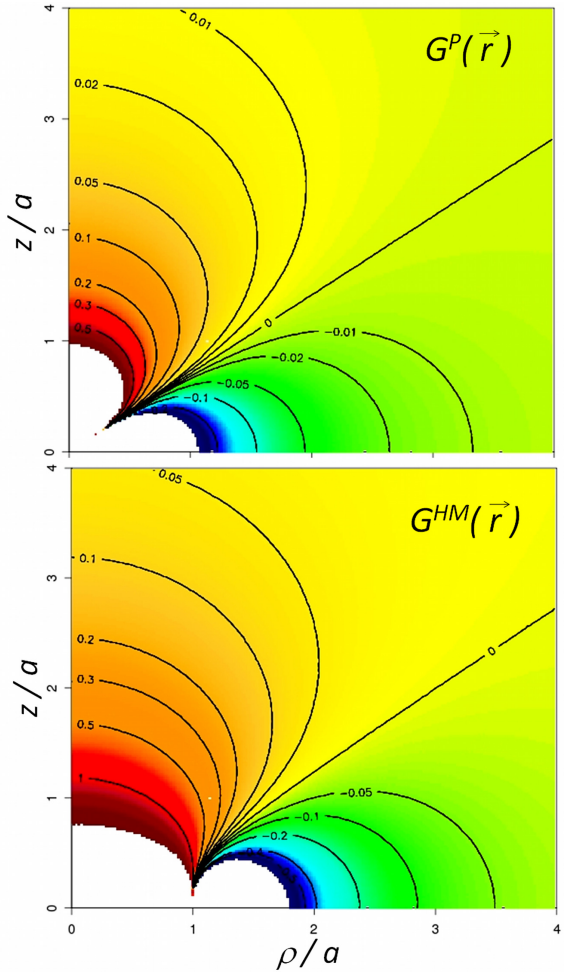


Figure 5.3: *Maps of the geometric factors around the benzene ring as defined by Pople point dipole (a) and Haigh-Mallion (b) models.*

The direct correlation between the Pople and Haigh-Mallion geometric factors are shown in Figure 5.4, where the data are broken down into four different regions of the proton position around the benzene ring. If drawing an analogy with di-base arrangement in nucleic acids, A and B would correspond to the

stacking interring arrangement, C is for the diagonal and D is for hydrogen-bonded planar arrangements. The border specification of the zones A, B, C, and D are $\rho \leq (1.7 + a)$ Å and $2.9 > z > 1.7$ Å, $\rho \leq (1.7 + a)$ Å and $z \geq 2.9$ Å, $\rho > (1.7 + a)$ Å and $z > 1.7$ Å, $\rho > (1.7 + a)$ Å and $z \leq 1.7$ Å respectively, where 2.9 Å is the sum of the carbon and hydrogen van der Waals radii and 1.7 Å is the van der Waals radius of the hydrogen atom.

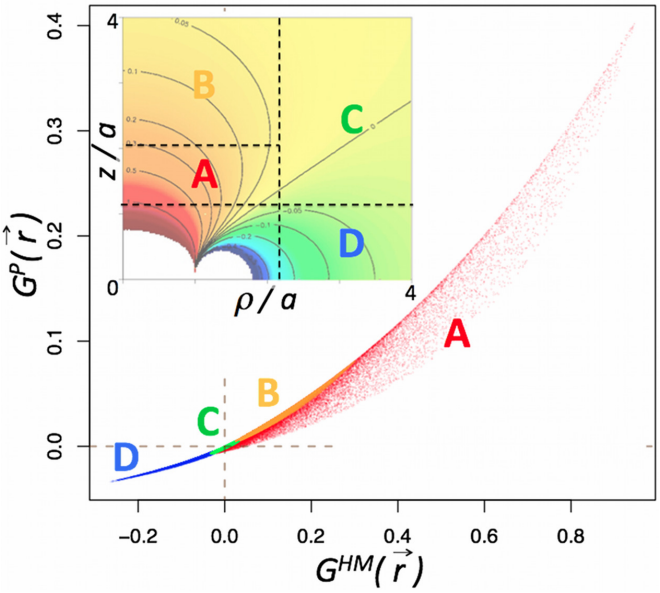


Figure 5.4: *Interconnection between the geometric factors of the Haigh-Mallion (x-axis) and Pople (y-axis) models for ring current effects. The correlations corresponding to four different spatial regions around the ring are differentiated by red, orange, green and red colours, also denoted by letter A, B, C and D and clarified in the built-in graph. In case of two nucleic acid bases, the regions A and B correspond to the stacked arrangement, C is for the diagonal and D is for coplanar, hydrogen-bonded, arrangements. For the full specification of the borders for the separate spatial regions, see the text.*

The correlation, in general, has a consistent shape, pointing out that the geometric factors can be easily inter-translated, hence eliminating even the small underperformance of the simple point dipole model as compared to the Haigh-Mallion one [Moyna *et al.*, 1998]. The only exception is the region A (Figure 5.4, red). It should be noted however, that in practice the distance between a proton

of one nucleic acid base and the ring centre in another nucleic acid base are almost always greater than 3.0 Å for the stacked arrangement and more than $1.7+a$ Å far for the coplanar arrangement when the protons as close as the ones participating in hydrogen bonding are considered. Therefore the divergence between the Pople and Haigh-Mallion models at the A region can only be influential for very strong stacking interactions, perhaps happening in nucleic acid and intercalative drug complexes.

The equations of the interconversion from Pople to Haigh-Mallion geometric factors and vice versa is determined below, using the compiled database of di-base arrangements in RNAs to allow the usage of simple and intuitive point dipole model while enhancing its performance up to the level of the Haigh-Mallion model.

5.6 Generation of the DiBaseRNA database of interring arrangements in RNAs

To study the ring current effects in molecular structures that closely resemble nucleic acids, a structural database is generated that reflects the observed inter-base arrangement in the high-resolution X-ray structures of RNAs. In general, RNAs possess conformations that are more variable, thus, at this stage, focusing on RNAs rather than DNAs provides an opportunity to study the ring current effects in a much wider scale, accounting for a broader conformational space. The initial RNA structures are taken from the RNA05 database of Richardson and coworkers [Murray *et al.*, 2003], which contains 171 coordinate files of RNA X-ray structures with 9486 nucleotide content and 3.0 Å or better resolution. Then, all the structures with equal to or better than 1.8 Å resolution are scanned and all the possible di-base arrangements between any pairs among the conjugated rings of adenine (A), guanine (G), cytosine (C) and uracil (U) bases (Figure 5.5a) are retrieved.

The di-base geometries are taken by classifying them into three - adjacent (ADJ), spatial (SPT) and hydrogen bonded (HBD) arrangements. Here, ADJ indicates that in the XY arrangement of the X and Y bases, the conjugated rings belong to the neighbouring nucleotides within the same chain. The adjacent

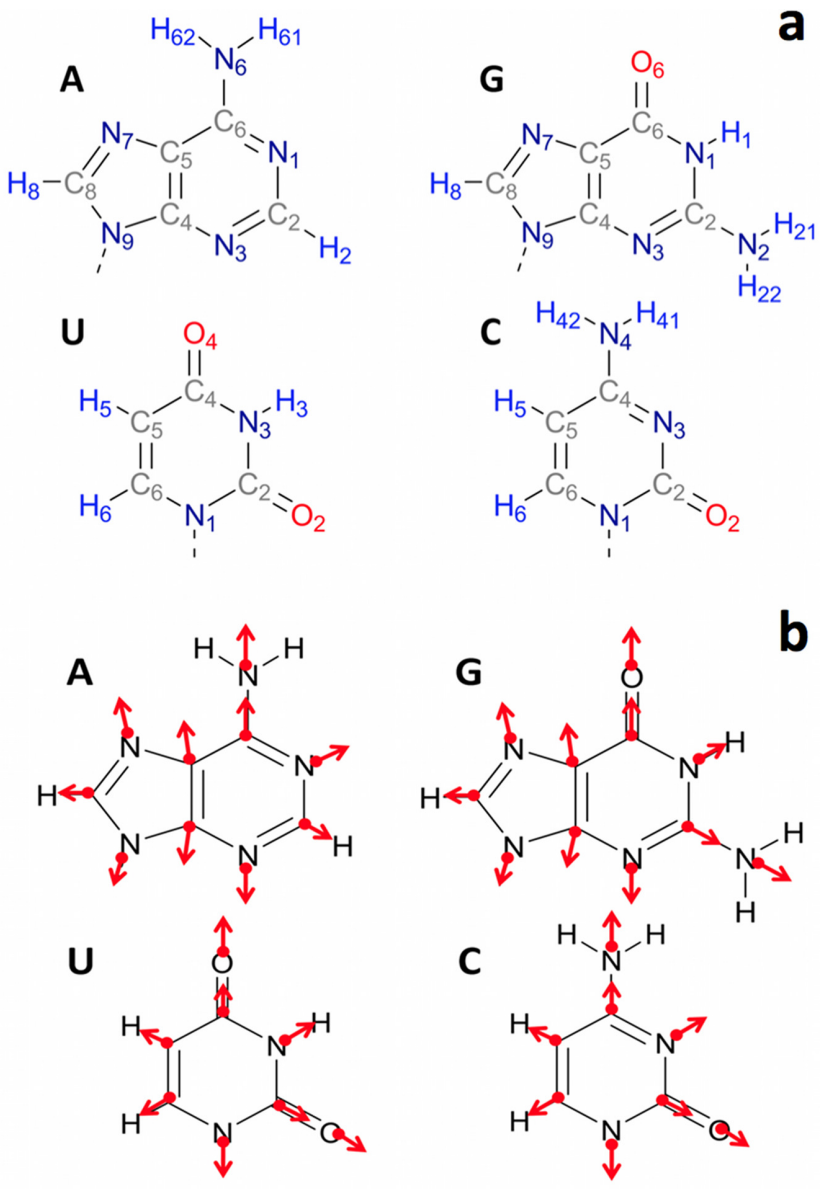


Figure 5.5: *The four nitrogen bases of RNAs with the numbering scheme (a) and the outline of the points where the electric field values generated by the second base is calculated, with arrows showing the direction for the considered projections (b).*

di-bases are scanned in both 5' to 3' and 3' to 5' directions for the retrieval. In cases, where one of the di-base members is common in both directions (for instance if we search for the adjacent arrangement of GA in the GAG sequence, A will be common in GA moieties retrieved from both directions), the fragment from the 3' to 5' scan is discarded. HBD is the arrangement, where the bases are nearly coplanar with hydrogen bonds (either canonical Watson-Crick, or of other types) between them and can belong to the same or different polynucleotide chains. The SPT arrangement is defined as the one, where the two bases are not coplanar and belong either to different chains or the same chain but separated by at least 3 nucleotides. They would usually represent the diagonal arrangement of the base-rings for the situations where the RNA molecule is self-assembled into a local double helical structures. Further, hydrogens are added to the N9 and N1 positions (Figure 5.5a) of the purine and pyrimidine rings as a replacement of the glycosidic bond, so that the resulting coordinate files correspond to a complete and closed-shell systems of two bases, ready for quantum chemical calculations. An example of the resulting geometries for the GG pair is shown in Figure 5.6, with the full set presented in Appendix G.

The generated database is further refined by a partial geometry optimisation with frozen core (non-hydrogen) atoms via the semiempirical AM1 Hamiltonian [Dewar *et al.*, 1985] as implemented in the MOPAC2009 package [Stewart, 2008]. The AM1 Hamiltonian is selected due to the published data of its performance in accurately representing the amide bond lengths [Stewart, 2007] and the geometries of amino groups attached to conjugated systems [Yatsenko & Paseshnichenko, 1999]. Furthermore, the known issue of nitrogen pyramidality overestimation is the least pronounced for AM1 [Stewart, 2007] in NDDO-type (neglect of diatomic differential overlap) semiempirical methods. In order to remove the residual pyramidality, the N-H bonds have been further frozen to be within the same plane as the corresponding conjugated ring. The resulting coordinate files in PDB format represent the variant 1 of the proposed DiBaseRNA database that features experimentally determined positions of the core atoms of each of the conjugated systems. However, small structural variation in the experimental bond lengths within the ring, caused by the experimental errors, is inevitable. Hence another variant of the DiBaseRNA database is generated (variant 2), where only

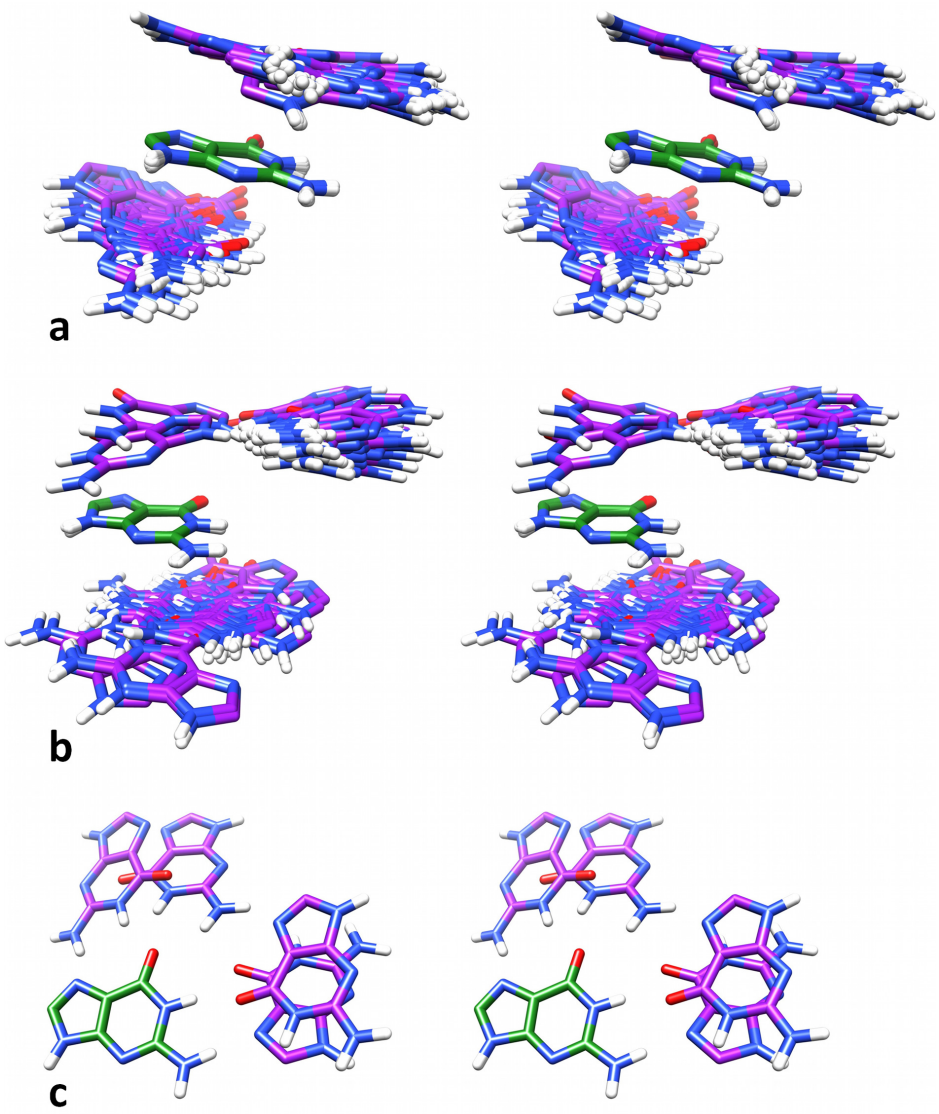


Figure 5.6: *An example of the interring arrangement pattern from the DiBaseRNA database. Guanine-guanine (GG) di-bases are presented in their adjacent (a, ADJ), spatial (b, SPT) and hydrogen bonded (c, HBD) states. For the explanation of the meaning of the used classification for the arrangements, please see the text.*

the relative arrangement of the two rings are learned from the X-ray data and the final files are constructed by rotating and translating the standard geometries of the constituent bases to match the experimental arrangement. To obtain the standard geometries, A, G, C, and U bases are geometry optimised without any constraints with tight convergence criteria. Hybrid density functional theory [Kohn & Sham, 1965] via the Becke three-parameter exchange functional and the Lee, Yang and Parr correlation functional (B3LYP) [Becke, 1993; Lee *et al.*, 1988; Miehlich *et al.*, 1989] is used with the split-valence 6-311+G(2d,p) basis set [Krishnan *et al.*, 1980]. After the geometry optimisation, amino groups in adenine, guanine and cytosine do appear to be slightly out of plane of the conjugated system, however, other calculations proved that a certain level of the out-of-plane displacement is not an artefact of the chosen method of calculation and does represent the reality that is observed by both experimental and more sophisticated theoretical methods [Šponer & Hobza, 1994; Sychrovsky *et al.*, 2009].

The content and the number of structures, present in each variant of the database, are summarised in Table 5.1.

5.7 Interconversion between Pople and Haigh-Mallion ring current geometric factors

A consistent shape is observed for the correlation between the geometric factors of the Pople and Haigh-Mallion models at around benzene 6-membered ring (Figure 5.4), if only the regions around the ring that are populated in usual biomolecular structures are accounted for. This shows that simple equations

Table 5.1: *The number of entries in the generated database of the di-base arrangement as observed in high resolution X-ray structures of RNAs.*

Base pairs	AA	AC	AG	AU	GC	GG	GU	CC	CU	UU
Adjacent	37	39	64	24	79	114	55	72	39	23
H-bonded	-	3	6	38	95	4	9	5	-	-
Spatial	21	21	24	7	20	81	26	10	13	4

can be devised for the interconversion of the Pople and Haigh-Mallion geometric factors. Such equations can be useful for either adding an extra precision to the Pople model, or for converting the Haigh-Mallion geometric factors into the Pople ones, since Haigh-Mallion model is the most implemented one in the existing chemical shift predictors [Neal *et al.*, 2003; Sahakyan *et al.*, 2011a,b], whereas Pople model is much simpler and convenient for implementing as restraints in molecular dynamics simulations or geometry optimisation routines.

To develop the conversion equations, the ring current geometric factors (both Pople and Haigh-Mallion) calculated for all the hydrogen atoms of all the di-base entries in DiBaseRNA database are used. Only the geometric factors originated from the neighbouring ring in the di-base couple for each DiBaseRNA entry is accounted. The 5- and 6-membered rings are considered separately.

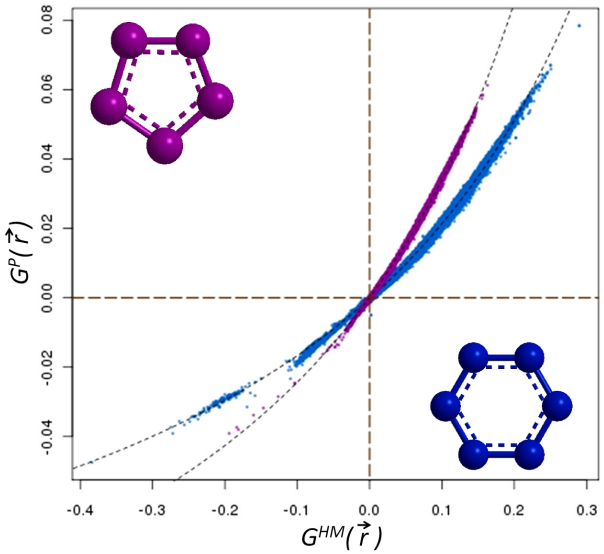


Figure 5.7: *Correlation between the Pople and Haigh-Mallion ring current geometric factors for 5- (violet points) and 6-membered (dark blue points) rings in the RNA structures of the DiBaseRNA database. The fitted correlations shown as dotted lines.*

Using the DiBaseRNA database, ring current geometric factors are calculated for all the hydrogen atoms induced by the coupled ring in each of the di-base entry. The obtained factors are thus for only the regions around the 5- and 6-membered rings that are populated in RNA structures. A simple mathematical model is

found using the *Eureqa* automatic technique to search for hidden dependencies in data [Schmidt & Lipson, 2009]. The resulting equations are (Equations 5.5 and 5.6):

$$G^{P,6}(\overrightarrow{\mathbf{r}}) = \frac{(0.325964 G^{HM,6}(\overrightarrow{\mathbf{r}}) - 0.000466)}{(1.743379 - 2.391111 G^{HM,6}(\overrightarrow{\mathbf{r}}))} \tag{5.5}$$

$$G^{P,5}(\overrightarrow{\mathbf{r}}) = \frac{(1.000243 G^{HM,5}(\overrightarrow{\mathbf{r}}) - 0.000146)}{(3.525744 - 5.873933 G^{HM,5}(\overrightarrow{\mathbf{r}}))} \tag{5.6}$$

where $G^{M,6}(\overrightarrow{\mathbf{r}})$ and $G^{M,5}(\overrightarrow{\mathbf{r}})$ are geometric factors for ring current models M and 6- and 5-membered rings correspondingly. The $G^{P}(\overrightarrow{\mathbf{r}})$ versus $G^{HM}(\overrightarrow{\mathbf{r}})$ dependences determined by the Equation 5.5 and 5.6 are shown on Figure 5.7 as dotted lines.

5.8 Density functional theory calculations of the ring current and electric field effects on nuclear shielding constants of nucleic acid bases

The PBE1PBE [Adamo & Barone, 1998] density functional theory [Kohn & Sham, 1965] is used to calculate nuclear shielding constants. The PBE1PBE functional is parameter-free and it was proven to be the best so far for studying NMR shielding of a wide range of nuclei, in many cases outperforming the results from the low-order perturbation studies, such as the ones using the correlated second order Møller-Plesset perturbation method [Adamo & Barone, 1998]. The split valence 6-311+G(2d,p) basis set [Krishnan *et al.*, 1980] with gauge-invariant atomic orbital (GIAO) method [Ditchfield, 1974; Wolinski *et al.*, 1990] is used taking into account prior research outlining its superiority [Cheeseman *et al.*, 1996] for nuclear shielding calculations. All the calculations are done using the Gaussian03 suite of programs [M. J. Frisch et al, 2004] with increased self-consistent field convergence criteria using the built-in *tight* keyword.

The DiBaseRNA database is used to calculate the ring current induced nuclear shielding changes and the electric fields originated by the neighbouring nitrogen

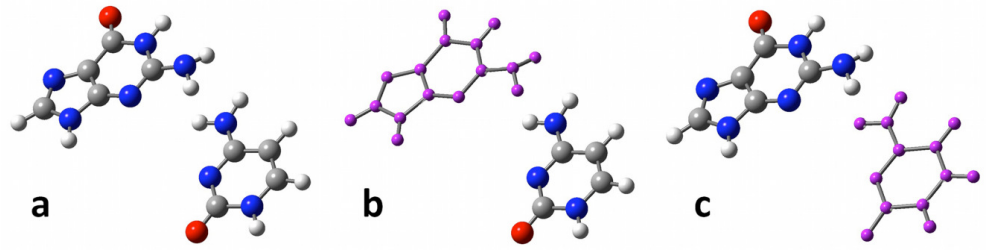


Figure 5.8: *An example geometry breakdown for the three DFT calculations done for each entry of the DiBaseRNA database.*

base on the given base. For each of the entries in the DiBaseRNA database, three calculations are done; one is for the complete pair, to infer the nuclear shielding constant values of the system, the constituent atoms of which are already under the influence if ring currents and electric field effects (Figure 5.8a), and, one calculation for each of the two nucleic acid base in isolation, where we still keep the positions of the neighbouring base by changing its atoms into dummy atoms (Figure 5.8b and c), hence enabling further retrieval of the electric field values generated by the isolated base but acting on the positions where the atoms of the second base are located in the complete complex. The electric field projections are taken along the local symmetry axes at each of the atom locations as defined in Figure 5.5b.

5.9 The influence of structural fluctuations on ^{1}H, ^{15}N, ^{13}C and ^{17}O chemical shifts of nucleic acid bases

At first, nuclear shielding calculations are done on the variant 1 of the DiBaseRNA database, where the core structures come from the X-ray structural dataset with better than or equal to 1.8 Å resolution. The hydrogen atom positions in this DiBaseRNA variant are optimised via semiempirical quantum chemistry in a planar constraint. Those structures have been additionally geometry optimised, this time with B3LYP/6-31G(d,p) model chemistry, again allowing only the hydrogen

atoms to change their position. Then, the nuclear shielding computations are carried out as described above. The results indicate that the small structural fluctuations of the core of the bases, that are in place for even the high-resolution X-ray structures, cause significant fluctuations of ^{1}H, ^{15}N, ^{13}C and ^{17}O chemical shifts, most probably because of the changes in the aromaticity of the conjugated ring. The fluctuation histograms for N1, H1, C6 and O6 atoms in guanine base are shown in Figure 5.9 as examples.

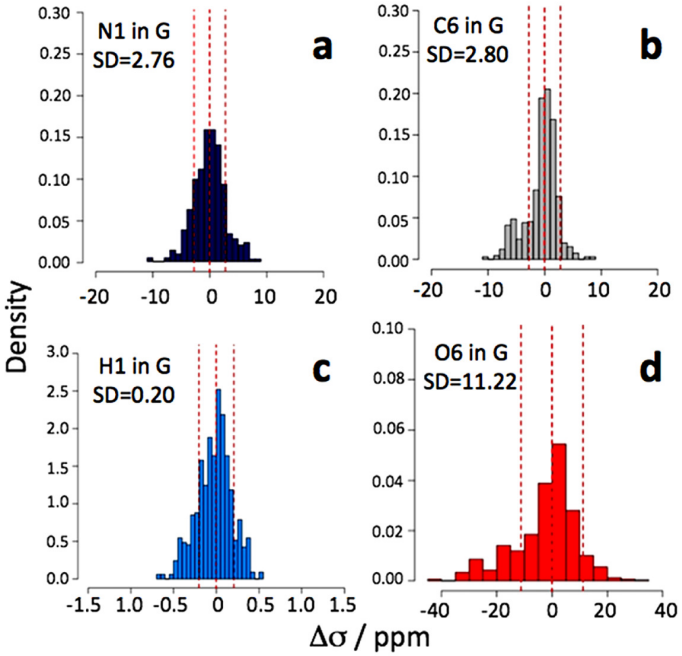


Figure 5.9: *The fluctuation histograms for the nuclear shielding constants of several ^{15}N, ^{13}C, ^{1}H and ^{17}O nuclei in guanine. The fluctuations are referenced by the median value of the nuclear shielding constant of each type. The standard deviations of the fluctuations are shown.*

Since the variation of nuclear shielding constants upon small structural fluctuations is greater than the changes caused by ring current or electric field effects, this result outlines the necessity of using the dataset that is constructed via the fixed standard structures of individual nucleic acid bases, placed in the same dibase arrangement patterns (variant 2 of the DiBaseRNA database) as observed in the fragments extracted from the X-ray structure database. The real dynamical

fluctuations of the bond lengths mostly occur around their equilibrium values with their influence, whether linear or non-linear, averaged in the observed chemical shifts.

5.10 The hierarchy of ring current and electric field effects for hydrogen and heavy nuclei in RNA bases

The DFT-calculated electric fields and changes in nuclear shielding constants that are imposed by the spatially neighbouring base-rings (by going from Figure 5.8c or b to a) are used to perform a linear model fitting using various frameworks for deriving different models and associated coefficients. The fittings have been performed in multiple schemes, by trying both Pople and Haigh-Mallion ring current models with and without accounting for the electric field effects. The fittings have been done for both hydrogen and non-hydrogen atoms, treating both the whole set of geometries in the DiBaseRNA database, and only concentrating on different classes of geometric arrangements of bases. Such multi-stratification of the accounted physical phenomena and geometric classes allows investigating the hierarchy of different effects in explaining chemical shift changes for each type of base-base interaction. Examples of the exploratory plots are presented in Figure 5.10 for all the types of ^{1}H and ^{13}C atoms, except the ones that directly take part in hydrogen bonding (i.e. excluding A, H, B and C from all the molecular moieties capable of forming A-H...B-C hydrogen bonds). The complete set of the results is presented in Appendix H, including the results for hydrogen bonded atoms. In case of the joint treatment of ring current effects, it is assumed that the coefficient of the ring current geometric term from the given ring-type is the same for all the atoms of single type (for all ^{1}H, ^{13}C, ^{15}N, separately). In case of the separate treatment of the ring current effects, the coefficients are assumed to be different for the individual atom types (for instance, individual ring current coefficients for adenine-C5, adenine-C6, guanine-C5, uracil-C2 etc. for ^{13}C nuclei), even if the ring is of the same type (for instance, adenine). Although physically the ring current-induced change in the local magnetic field value solely depends

on the position of the query point O and the type of the conjugated ring, in cases where the ring current geometric factor also covers electric and other types of interactions, the coefficients from such fittings will also contain the response of the nuclear shielding constant of the query atom towards such interactions. This response will be dependent on the electronic environment for each atom type, thus by enabling individual coefficients for ring current geometric term acting on each query atom type, we separately account the expected different responses of the query atoms towards the outer changes. And indeed, in the case of separate treatment of the ring current effects, improvement in the agreement between the ring induced nuclear shielding constants predicted from the fitted model and calculated via hybrid-DFT can be noted, which becomes even more apparent for non-hydrogen atoms.

For ^{1}H nuclei, the plots demonstrate that the ring induced chemical shift alterations in the hydrogen-bonded complexes are prevalently explained by the electric-field-only treatment. In general, by looking at all the exploratory graphs, it seems the ring current geometric terms can, in many cases, replace the required hydrogen-bonding geometric terms for nucleic acid chemical shift predictor development. This will largely simplify the models of the chemical shift prediction for base atoms that are involved in hydrogen bonding, where the ring current geometric factor can simultaneously account for both ring current and hydrogen bonding effects via a single joint coefficient. It is also clear, that the SPT and ADJ arrangements of the conjugated rings affect the ^{1}H chemical shifts via mostly the ring current effect.

Conjugated ring effects on ^{13}C nuclei are almost always explained by electric field effects, regardless of the interring arrangement type. However, it is interesting to note that the ring current geometric terms can capture electric field effects if the ring current coefficients are trained separately for each carbon atom type. The same is true for ^{15}N nuclei, but not for ^{17}O, for which ring current geometric terms or their addition to the electric field terms improve only the description of the oxygen chemical shifts that take part in hydrogen bonding. For the other (SPT, ADJ and their combination) interring arrangements, the electric field term is the dominant factor affecting ^{17}O nuclear shielding constants. ^{17}O nuclei possess a remarkably high sensitivity to electric field effects, which can change the

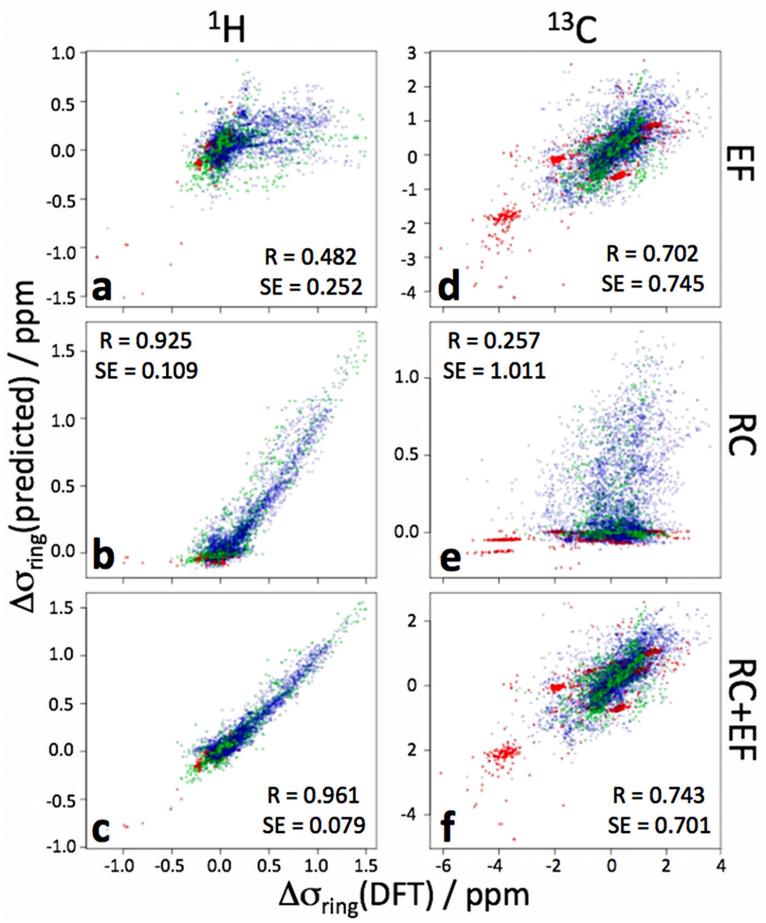


Figure 5.10: *Linear model fitting results for all the* ^{1}H *(a, b, c) and* ^{13}C *(d, e, f) nuclei, that do not directly participate in hydrogen bonding, from all RNA bases. The plots represent the correlations between the change in nuclear shielding constants predicted by the fitted model and the ones from the hybrid-DFT calculations. Three different models are fitted, using only electric field terms (EF, a and d), only ring current terms (RC, b and e) and both effects (RC+EF, c and f). Here, the Pople point dipole model is used in the joint treatment scheme, where, for the given ring type, the coefficients for its ring current geometric factor are assumed to be the same for all the atoms of single type (for all H, all C, all N and all O atoms), regardless their chemical state. Blue, green and red points come from the interring arrangements of the ADJ, SPT and HBD classes in the DiBaseRNA database (see the text). The Pearson correlation coefficients and the standard errors of the predictions are shown on the plots. The complete set of the fitting results for all the nuclei and model variants is presented in Appendix H.*

chemical shift values by up to 80 ppm. This indeed makes ^{17}O a highly sensitive probe if the ^{17}O NMR becomes routine for biomolecules, for which the initial signs are already visible [Zhu *et al.*, 2010].

The usage of Haigh-Mallion geometric term, instead of Pople model, has almost always slightly improved the quality of the linear models, however with the difference still being not essential.

5.11 Conclusions

A comprehensive benchmarking of the Pople point dipole model for ring current effects against the more precise Haigh-Mallion model is performed. Equations, general for all the 5- and 6-membered rings are proposed to convert the geometric factors of one model to another, either for migrating towards the Pople model in case the chemical shift predictor already uses the Haigh-Mallion one, or to increase the accuracy of the Pople model, still keeping its simple core, easy to implement in restrained molecular dynamics simulations. A database of di-base arrangements, observed from high-resolution RNA structures, is generated on which hybrid-DFT calculations are done to estimate the change in local electric fields and nuclear shielding constants upon the presence of conjugated rings in vicinity. Besides the widely explored ^{1}H, this study also extends to ^{13}C, ^{15}N and ^{17}O nuclei, with the latter quadrupolar nucleus just starting to enter the area of biomolecular NMR [Zhu *et al.*, 2010]. Then, a series of linear model fittings is performed to derive ring current and electric field based models for explaining the chemical shift changes induced by the neighbouring ring. The multi-stratified fitting enabled the assessment of the hierarchy of ring current and electric field effects on both hydrogen and non-hydrogen nuclei. In particular, it is directly demonstrated that the chemical shift changes of non-hydrogen atoms are mostly determined by electric field, rather than ring current, effects. On the other hand, hydrogen bonding-induced electric field effects are very well captured by the geometric factors of ring current models - a property that will be useful to account for modelling hydrogen bonding effects on chemical shifts in nucleic acid bases. A server, RingPar, is created that enables users to extract the fitting coefficients and resulting correlation plots, after defining the type of fitting to be done and

the structural classes to be used. Besides nuclear shielding constants, RingPar also reports models and coefficients for the nuclear shielding anisotropies, useful for future developments of models for chemical shift anisotropies. This is the first step towards the further development of accurate chemical shift predictors for nucleic acid bases, that will work for both hydrogen and non-hydrogen nuclei.

Almost anything is easier to get into than out of.

Agnes Allien's law from Paul Dickson's "The Official Rules"

6

Prospects and Future Work

Chemical shifts, owing to their high information content and extremely non-linear structural dependence, are capable of providing a wealth of information on fine details of protein structure and dynamics. The study described in this thesis extends the boundaries of the chemical shift usage in bimolecular research by providing possibilities for accurate chemical shift prediction for protein side-chain methyl and aromatic groups. Since these types of side-chain chemical shifts are the only available NMR observables from large proteins and protein assemblies, the proposed predictors will soon facilitate the extension of the protein size-limitation that exists in NMR spectroscopy for performing studies in atomic detail. Furthermore, certain physical effects on chemical shifts are thoroughly examined to enable the future increase in accuracy and applicability of the structure-based chemical shift predictors by including solvent exposure information, and, more importantly, by extending the methodology for nucleic acids. The latter step is expected to power structural biology with new structural and dynamical insights on nucleic acids and protein-nucleic acid complexes.

As can be inferred from the chapters devoted to CH3Shift and ArShift, the major breakthrough in enabling the chemical shift predictors to result in sufficient accuracy for side-chain atoms was the optimisation of the defined complete model for the chemical shift prediction for individual atom types. This clearly hints that a fully automatic method can be constructed, which, for a given X experimental parameter (whether NMR chemical shift, pK_a dissociation constant, S^2 generalised order parameter or anything else that is expected to have a hidden structural dependence), can develop a structure-based and mathematically differentiable predictor, ready for its immediate usage in X-restrained molecular dynamics simulations. Being bored of multiple fittings and trial/error cycles within the period of my PhD research, and, trying to oppose the Agnes Allien's law brought here as the opening phrase of the chapter, I have thus decided to *develop a developer* of such structural dependencies that is capable of saving years of graduate students' brain-time spent on deriving models for any experimental parameter that can have a structure dependence.

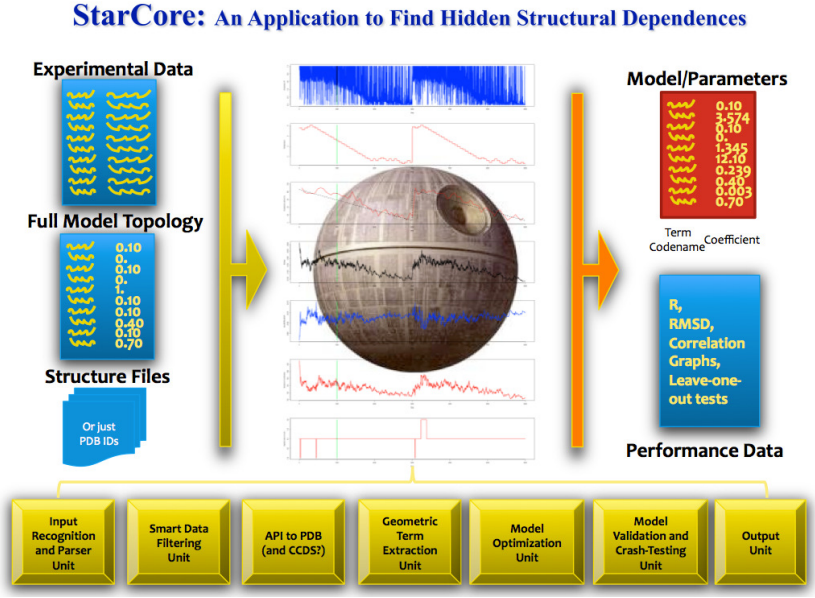


Figure 6.1: *Schematic representation of the* StarCore *workflow that operates on a database of any X experimental observable (from biomolecules), looks for hidden structural dependencies and develops a structure-based and differentiable predictor of X.*

The alpha version of the program (StarCore, **str**uctural **cor**relator) is already ready, supporting only protein structures in its present form. The workflow of the program is presented in Figure 6.1. It contains a library of thousands of geometric terms that link the position of a certain geometric point (an atomic position, geometric centre of several atoms etc.) to the neighbouring atomic arrangement in biomolecules. For its operation, StarCore requires a file that contains a preferably large set of experimental parameters and the addresses (protein chain, amino acid sequence number, PDB atom name or names, if the geometric centre of several atoms should be accounted) of each measurement pointing to its structural locus. Also should be provided a set of PDB structures, which can be omitted in case the related structures are already in the central PDB database. The user needs to edit the model topology file, where the whole set of geometric factors are listed with their codenames that define both individual and collective (in case all the geometric factors of certain type should be controlled at once) terms. The editing stage assumes the alteration of the probability values associated to each geometric factor-controlling codename. The value 1 enforces StarCore to always use the corresponding term in the generated models, and, the 0 value sets the term as always absent from the model. Any value in between 0 and 1 defines how frequently the term will be altered from *present* into *absent* or vice versa, in biased random model generation steps. StarCore has a sophisticated built-in Mont Carlo optimiser with pseudo-temperature bath, simulated annealing module and a selection of pseudo-energies (trade-offs between the goodness of fit and the number of coefficients/terms in the model) that can be used to guide the model optimisation process. Provided that all the required inputs are supplied and the controls are set, StarCore reads in all the experimental data, automatically filters them to remove the detected apparent outliers, and, for each experimental measurement, extracts all the possible geometric terms (from the associated structure file) with non-0 value in the model topology file. Then, the model optimisation unit starts to work from the complete model (all the non-0 terms switched on) and optimises the equation through altering the constituent terms, one at a time, by setting them as *present* or *absent*. For each altered model, StarCore does a complete linear model fitting via least squares optimisation of the coefficients and calculates the pseudo-energy score for the fit. The

decision, whether the move should be accepted or rejected, is based on controllable Metropolis criterion-type logical unit. The outcome of StarCore is an optimal model file with the codenames and parametrized coefficients for only the terms that are present in the found optimal model.

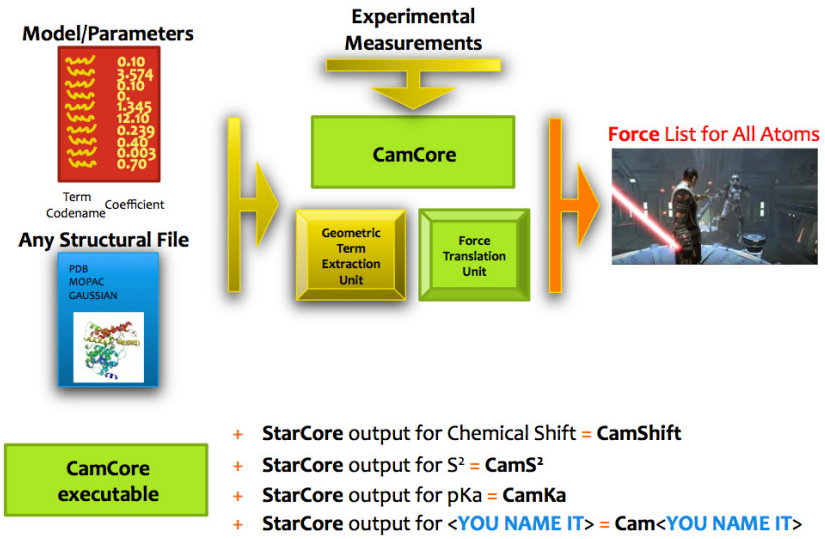


Figure 6.2: *Schematic representation of the* CamCore *engine that takes a snapshot of biomolecular structure (within the workflow of molecular dynamics simulations) and calculates the restraining forces based on the model file prior developed with* StarCore.

For a given experimental measurement type X, the model file from StarCore-search can then be supplied to another powerful engine, CamCore (under development), which contains the whole set of functions that StarCore uses, in addition holding their differentiated forms (Figure 6.2). CamCore infers the types of geometric terms that are present in the model/parameter file (StarCore output). It then extracts the required geometric factors from the supplied structural snapshot of a biomolecule, calculates the X parameters, compares the calculated values with the provided experimental measurements for the biomolecule under study and calculates the restraining forces via the protocol described in Chapter 1. With the output being the list of restraining forces for a given structural snapshot, CamCore can become the universal engine for any molecular dynamics package to enable restrained simulations that use experi-

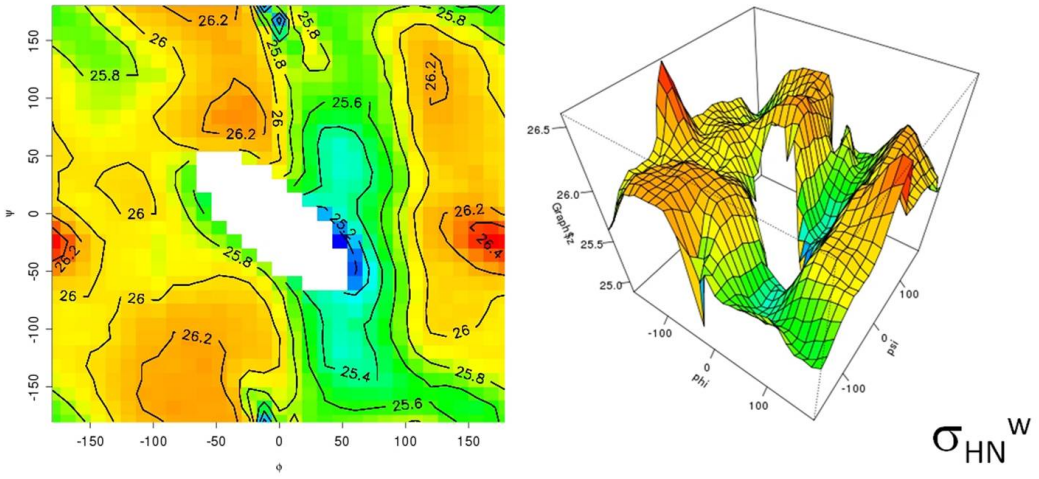
mental parameters for which StarCore has found a reasonable model and has written that model into a CamCore compatible output file.

These results are preliminary, however, the current state of StarCore is already under the use for the development of a new version of chemical shift and a novel structure-based pK_a dissociation constant predictor for proteins, so that they can be used for future restrained molecular dynamics simulations through the CamCore engine. Such a generalised approach in developing and directly implementing structure-based differentiable predictors of experimental parameters automatically optimises both the model coefficients and the model itself, thus holding a great promise in merging the power of experimental and computational techniques into a single and easy-to-use unified platform that shall indeed become an essential tool for structural biology.

Appendix A

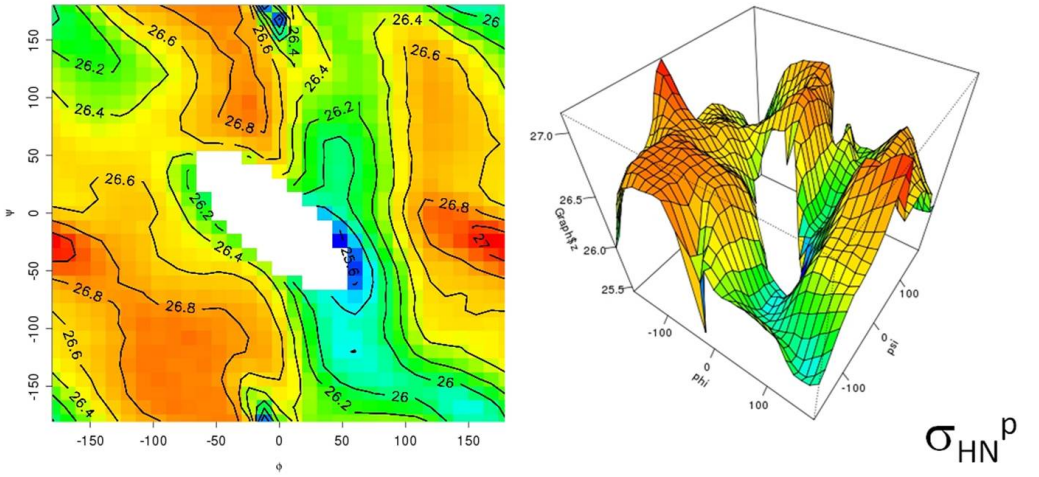
6 figures representing the projected and 3-dimensional plots of the surfaces of backbone $^{1}H^{N}$, $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, ^{15}N and ^{17}O nuclear shielding constants over ϕ/ψ dihedral angles in *Ace-Ala-Nme* molecule. The calculations are done in $\epsilon = 78.39$ (water, w, top plots) and $\epsilon = 4$ (protein interior, p, middle plots) conditions. The surfaces at the bottom show the difference in nuclear shielding constants from water to protein interior across the Ramachandran space.

range = [24.861; 26.590]



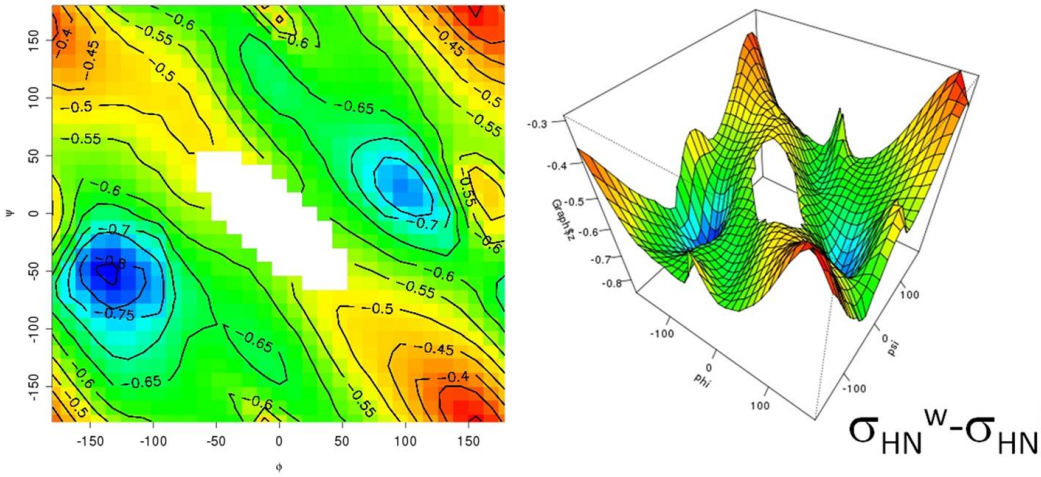
σ_{HN}^{w}

range = [25.439; 27.132]



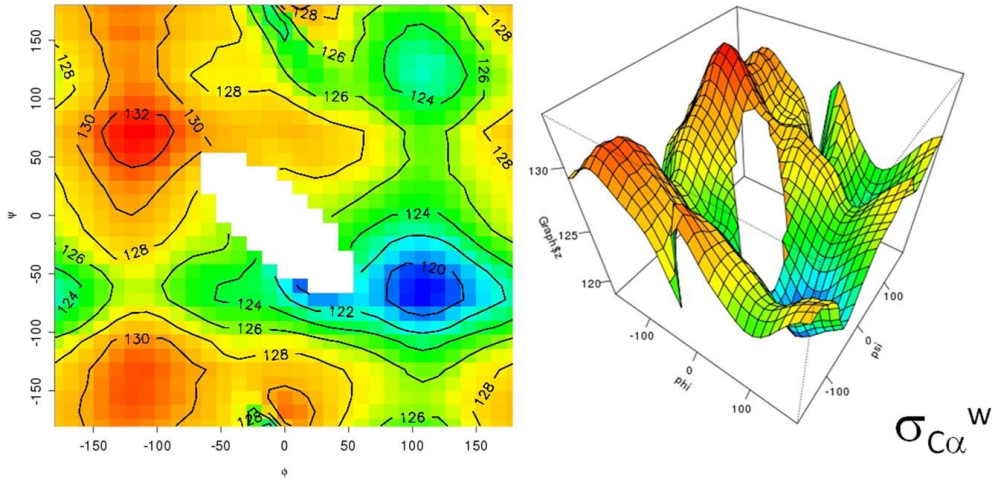
σ_{HN}^{p}

range = [-0.859; -0.289]



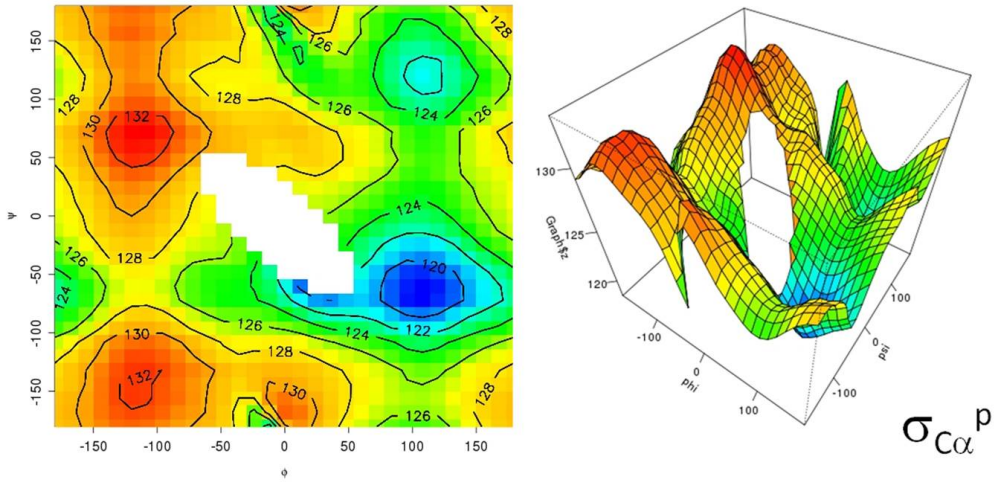
$\sigma_{HN}^{w}-\sigma_{HN}^{p}$

range = [118.437; 133.294]



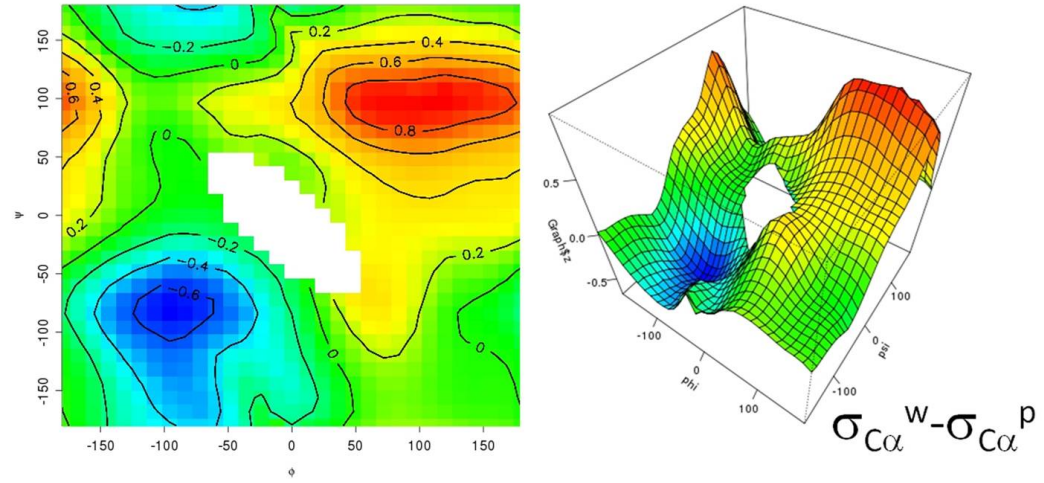
$\sigma_{C\alpha}{}^{w}$

range = [118.275; 133.273]



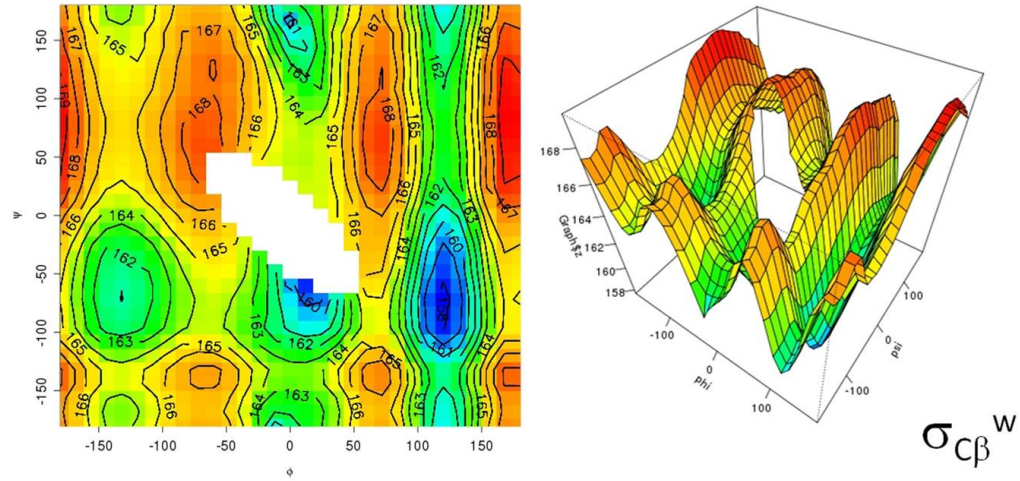
$\sigma_{C\alpha}{}^{p}$

range = [-0.699; 0.963]



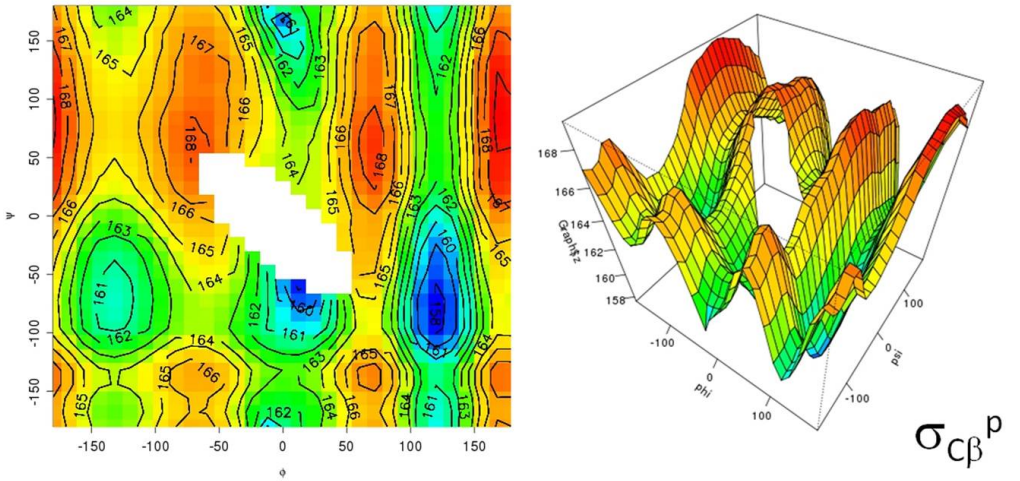
$\sigma_{C\alpha}{}^{w}$-$\sigma_{C\alpha}{}^{p}$

range = [157.661; 169.630]



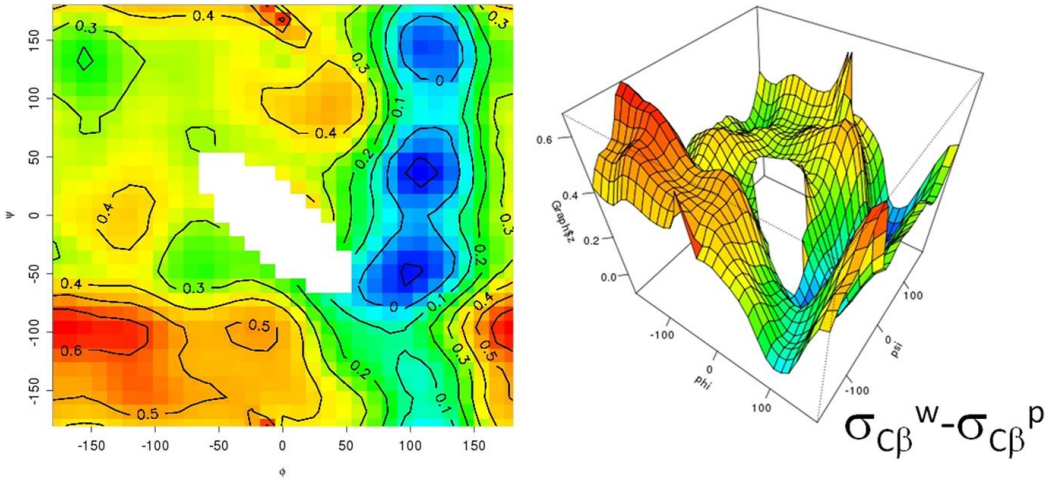
$\sigma_{C\beta}{}^{w}$

range = [157.592; 169.314]



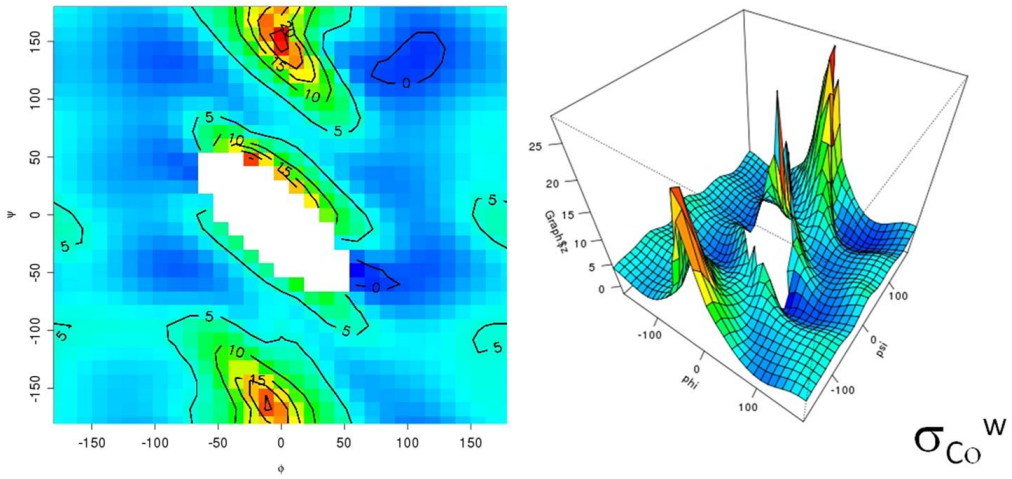
$\sigma_{C\beta}{}^{p}$

range = [-0.118; 0.671]



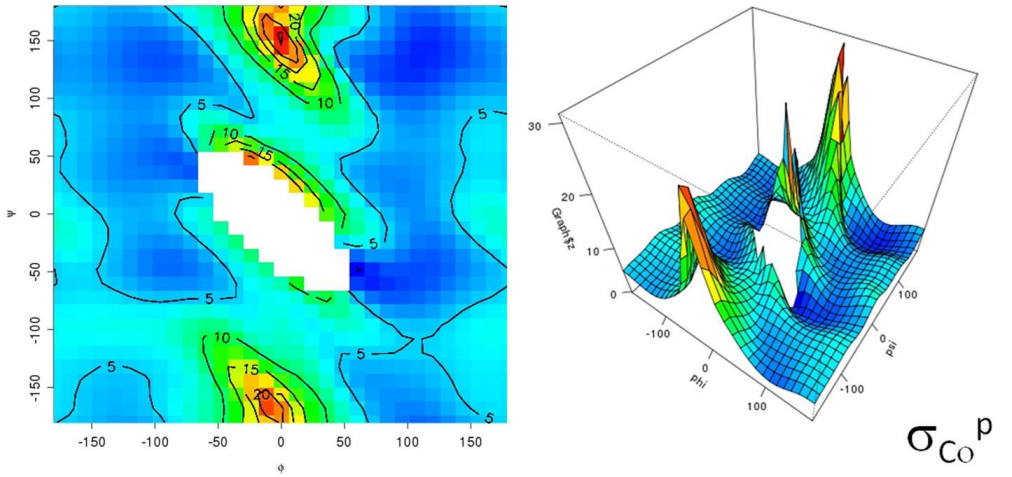
$\sigma_{C\beta}{}^{w}-\sigma_{C\beta}{}^{p}$

range = [-1.870; 28.277]



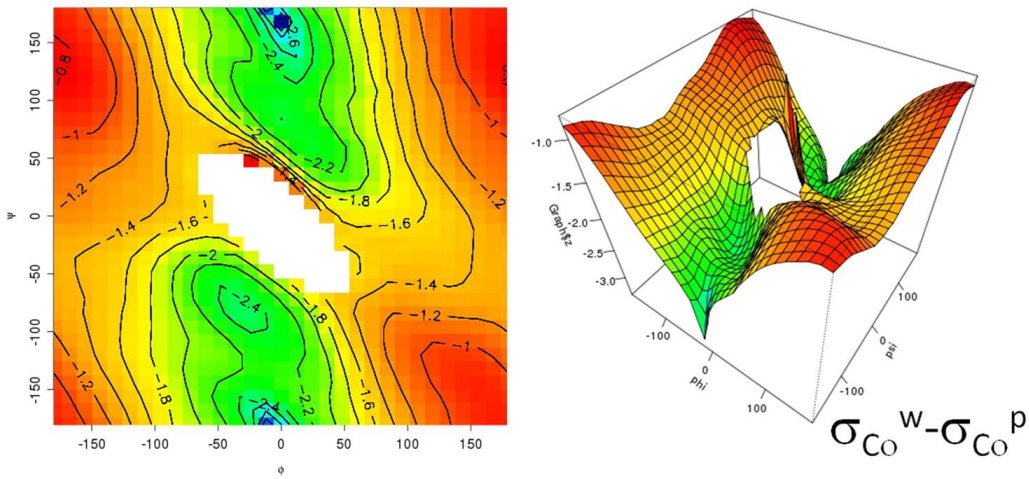
$\sigma_{Co}{}^{w}$

range = [-0.249; 31.100]



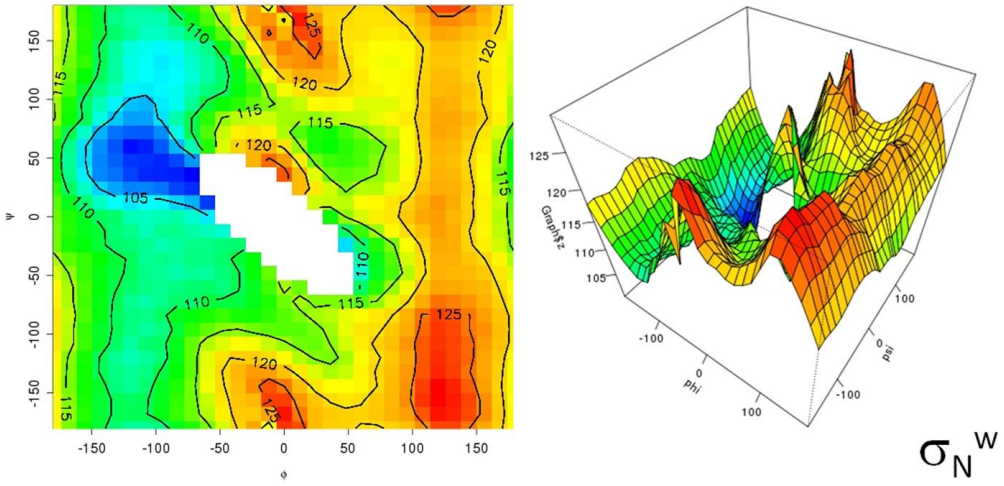
$\sigma_{Co}{}^{p}$

range = [-3.343; -0.758]



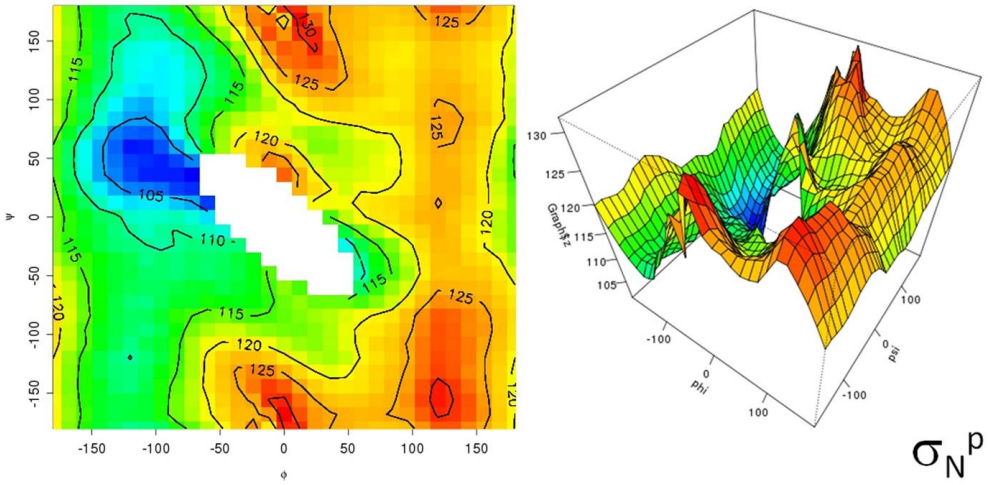
$\sigma_{Co}{}^{w}$-$\sigma_{Co}{}^{p}$

range = [100.011; 129.349]



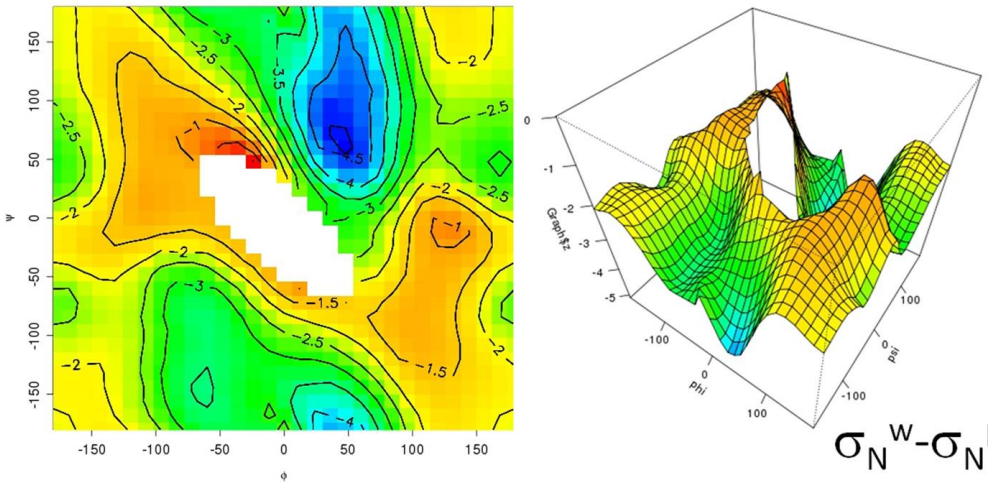
$\sigma_N{}^w$

range = [101.215; 131.766]



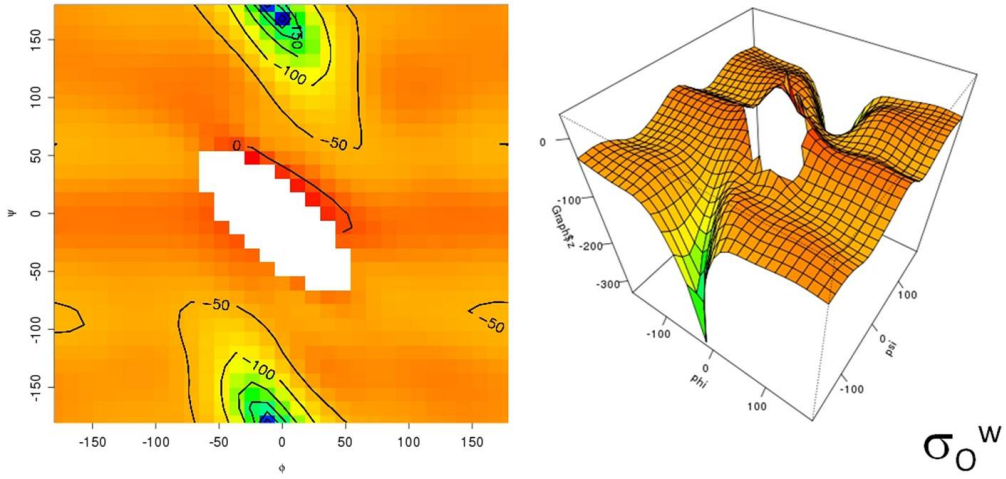
$\sigma_N{}^p$

range = [-5.108; 0.016]



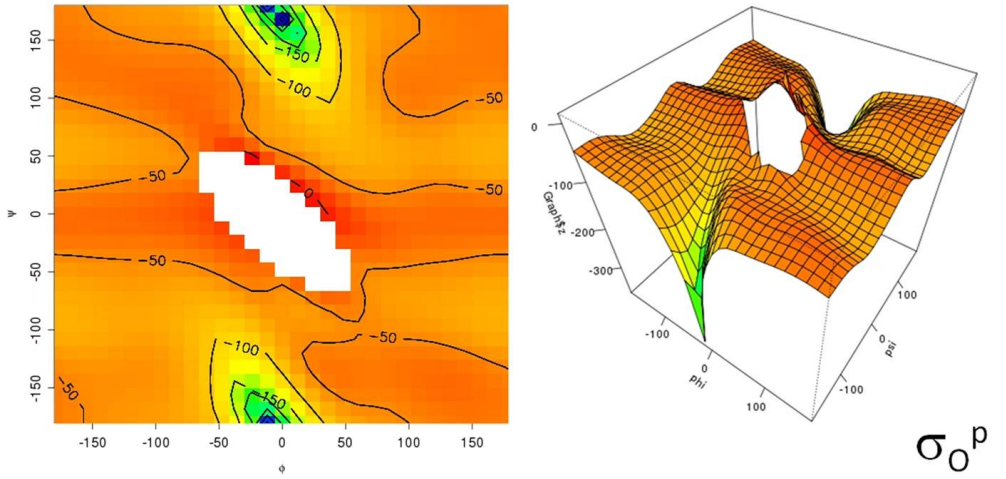
$\sigma_N{}^w - \sigma_N{}^p$

range = [-335.315; 35.418]



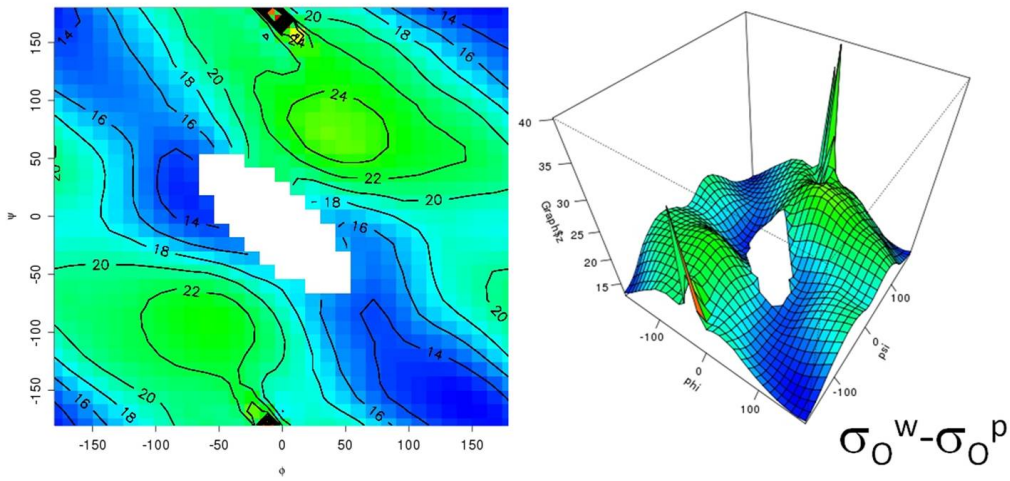
$\sigma_O{}^W$

range = [-375.477; 15.031]



$\sigma_O{}^P$

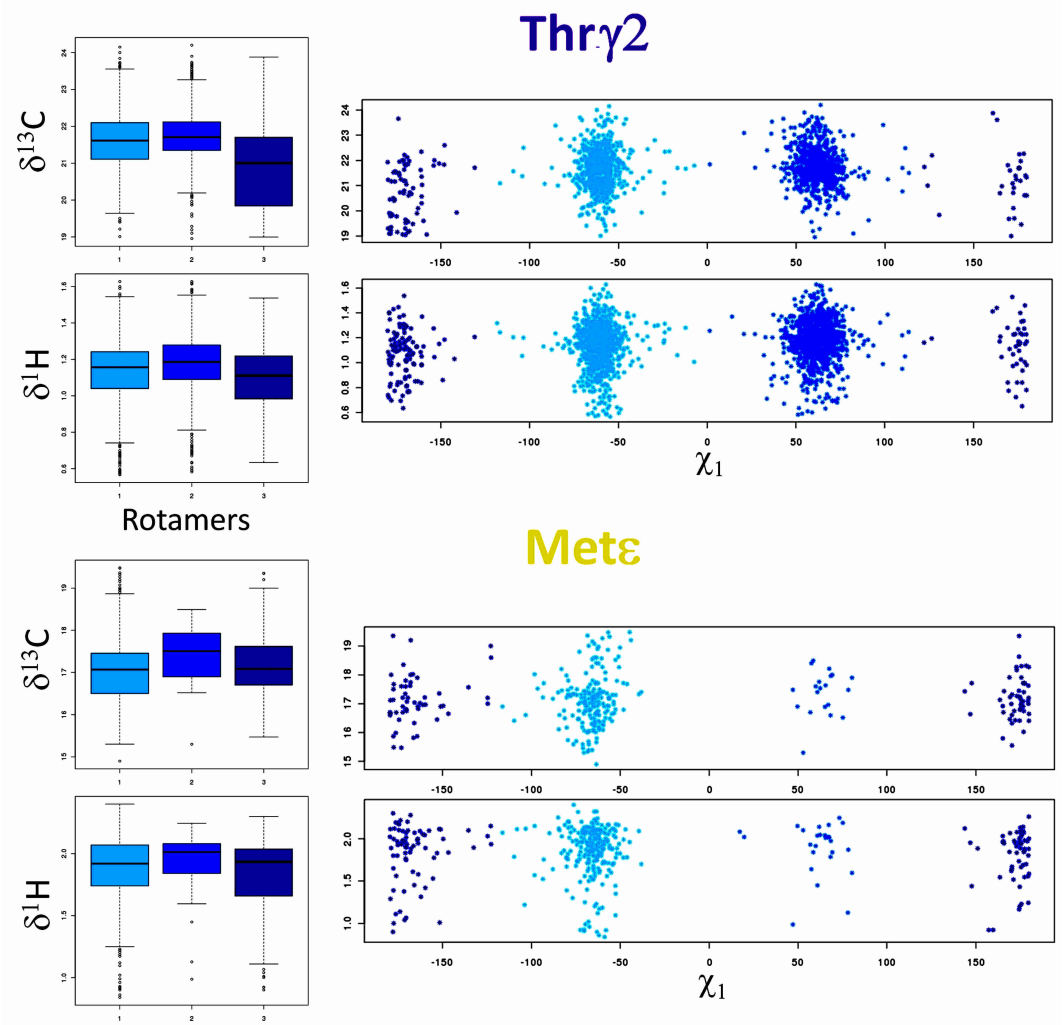
range = [12.286; 40.162]

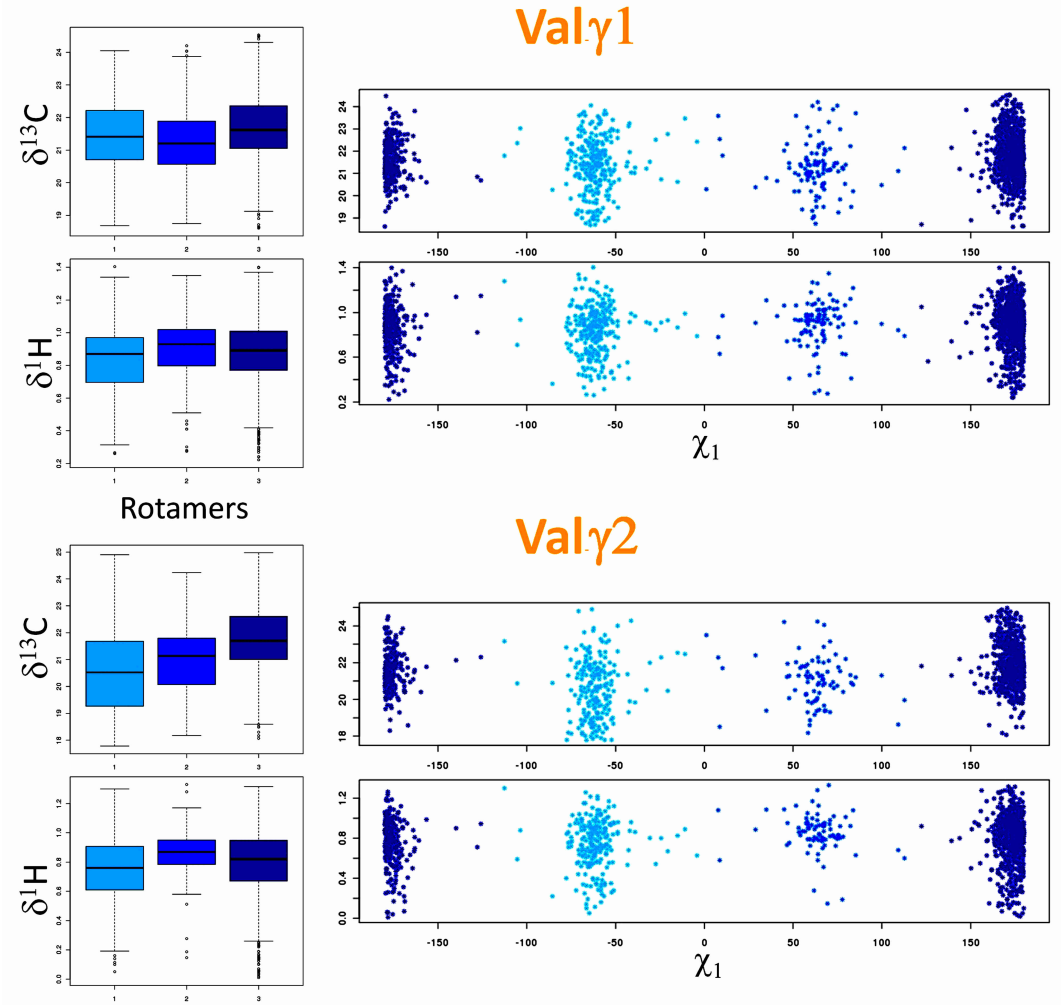


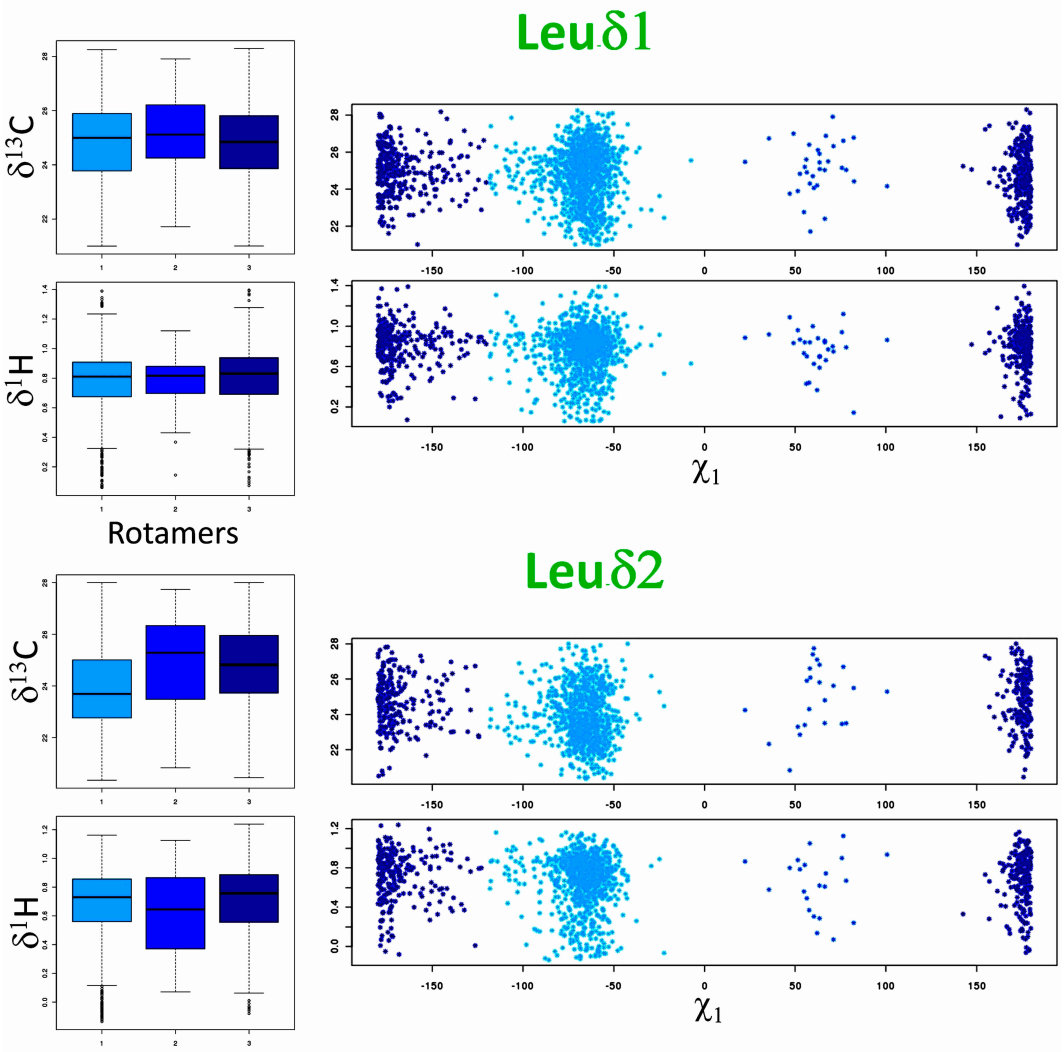
$\sigma_O{}^W$-$\sigma_O{}^P$

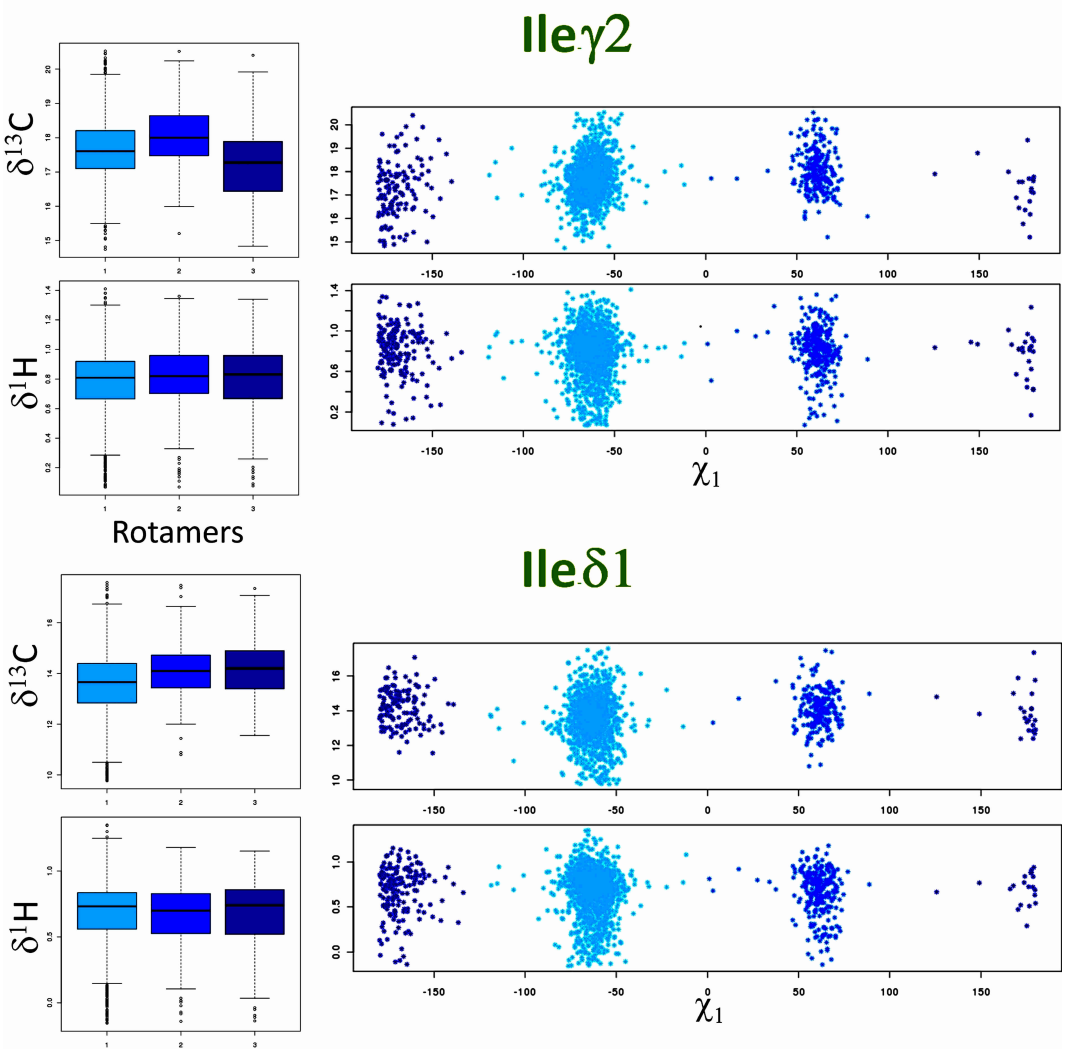
Appendix B

Distribution of the experimental methyl chemical shifts (in ppm) for the different types of rotamers defined by χ_1 angle (in degrees). The boxplots at the left side of the figures illustrate the statistical properties of the chemical shift distributions for the three major χ_1 rotameric states of side-chains. Boxes are constructed via the median, first and third quartiles of the distribution. The whiskers show the range of values that are within the 1.5 times IQR (interquartile range). Individual points indicate the outliers.


Thrγ2
δ13C
δ1H
Rotamers
χ1
Metε
δ13C
δ1H
χ1


Val γ1
δ13C
δ1H
Rotamers
χ1
Val γ2
δ13C
δ1H
χ1


Leu δ1
δ13C
δ1H
Rotamers
χ1
Leu δ2
δ13C
δ1H
χ1


Ileγ2
δ13C
δ1H
Rotamers
χ1
Ileδ1
δ13C
δ1H
χ1

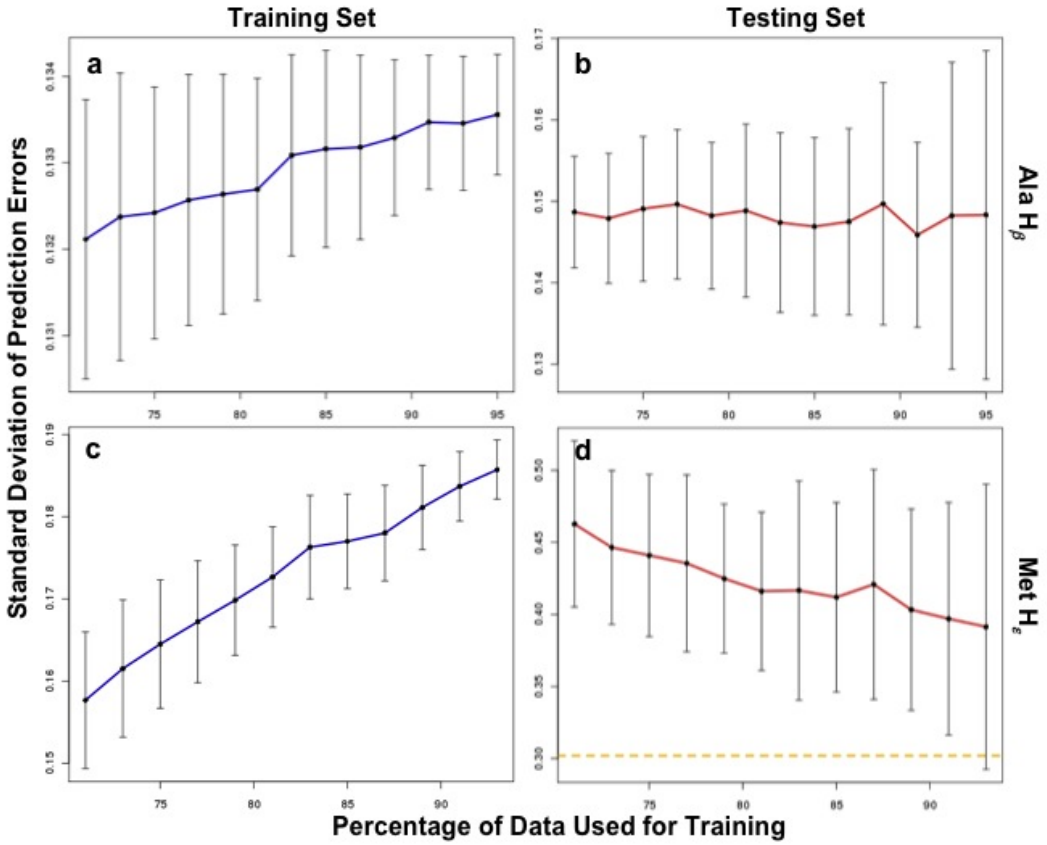
Appendix C

Full expression of the dihedral angle term used for the development of CH3Shift. The symbol θ indicate any of the available dihedral angles from the ϕ, ψ, χ_{1-5} set and the coefficients k_i are optimised by a least squares optimisation of the compiled data.

$$\begin{aligned}\Delta\delta_{dih} = k_1^{dih}\theta + k_2^{dih}\theta^2 + k_3^{dih}\theta^3 + k_4^{dih}\theta^4 + k_5^{dih}\cos\theta + k_6^{dih}\cos 3\theta + k_7^{dih}\cos 5\theta + \\ k_8^{dih}\cos(\theta + \pi/2) + k_9^{dih}\cos(2\theta + \pi/3) + k_{10}^{dih}\cos^2\theta + k_{11}^{dih}\cos^2 2\theta + \\ k_{12}^{dih}\cos^2 3\theta + k_{13}^{dih}\cos^3(\theta + \pi/2) + k_{14}^{dih}\sin\theta\cos\theta\end{aligned}$$

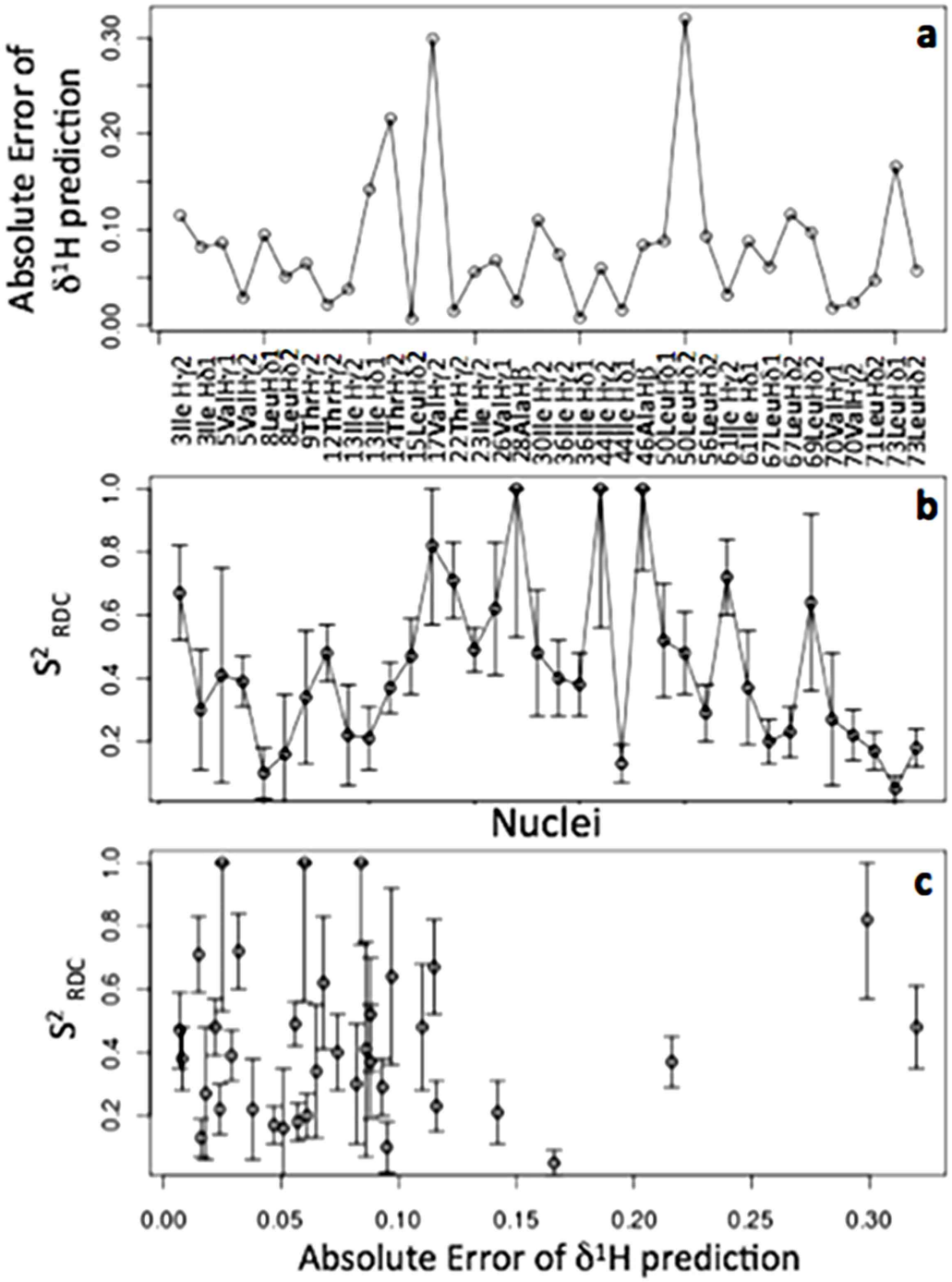
Appendix D

Examples of normal (Ala H^{β}, panels a and b) and overfitted (Met H^{ϵ}, panels c and d) models. The whiskers indicate the standard deviation of the prediction errors over the 250 different replicas of the test, where the partition over the training and test sets are performed randomly. For Met, an artificial improvement of the prediction quality in the training set (c) can be noted, associated to a decrease of the quality in the test set (d), when the percentage of data used for training is decreased. This is an indication of overfitting, further confirmed by a close to 1 correlation coefficient between the predicted and experimental chemical shifts in the leave-one-out test. In the test sets, the errors for the Ala H^{β} are always lower than the standard deviation of the corresponding chemical shift type used for training the model, whereas for the Met H^{ϵ} nucleus, the errors are higher than the experimental data dispersion (d, orange dashed line).


Training Set
Testing Set
a
b
c
d
Standard Deviation of Prediction Errors
Percentage of Data Used for Training
Ala H_β
Met H_ε

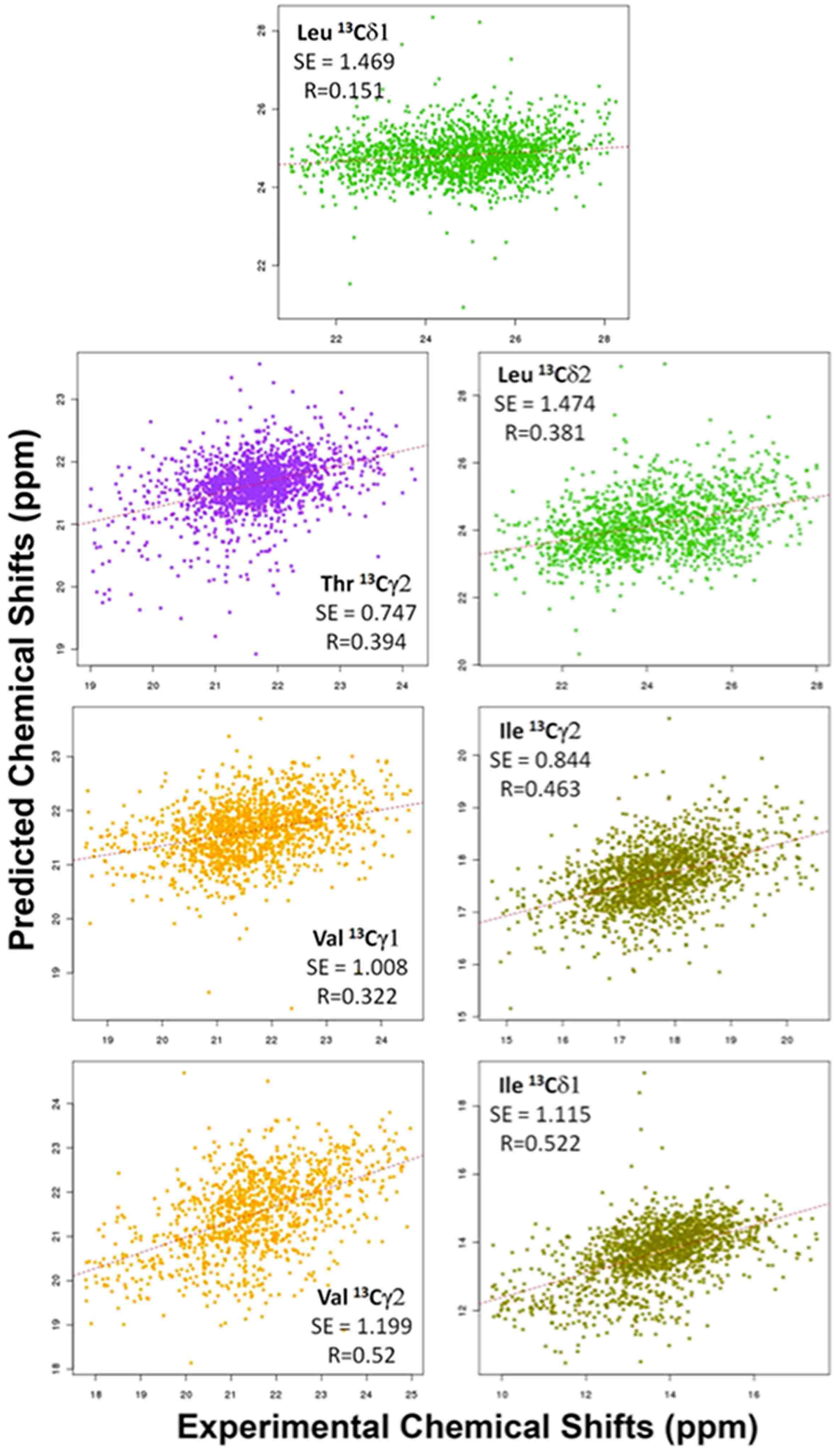
Appendix E

Changes in the absolute errors of ^{1}H chemical shift predictions calculated from the X-ray structure (a) and S^2 generalised order parameter (b) [Farès *et al.*, 2009] over different methyl groups in ubiquitin. The absence of correlation between these two parameters is illustrated in panel c.


a
Absolute Error of δ¹H prediction
0.00
0.10
0.20
0.30
3Ile Hγ2
3Ile Hδ1
5Val Hγ1
5Val Hγ2
8Leu Hδ1
8Leu Hδ2
9Thr Hγ2
12Thr Hγ2
13Ile Hγ2
13Ile Hδ1
14Thr Hγ2
15Leu Hδ2
17Val Hγ2
22Thr Hγ2
23Ile Hγ2
26Val Hγ1
28Ala Hβ
30Ile Hγ2
36Ile Hγ2
36Ile Hδ1
44Ile Hγ2
44Ile Hδ1
46Ala Hβ
50Leu Hδ1
50Leu Hδ2
56Leu Hδ2
61Ile Hγ2
61Ile Hδ1
67Leu Hδ1
67Leu Hδ2
69Leu Hδ2
70Val Hγ1
70Val Hγ2
71Leu Hδ2
73Leu Hδ1
73Leu Hδ2
b
S² RDC
0.2
0.4
0.6
0.8
1.0
Nuclei
c
S² RDC
0.2
0.4
0.6
0.8
1.0
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Absolute Error of δ¹H prediction

Appendix F

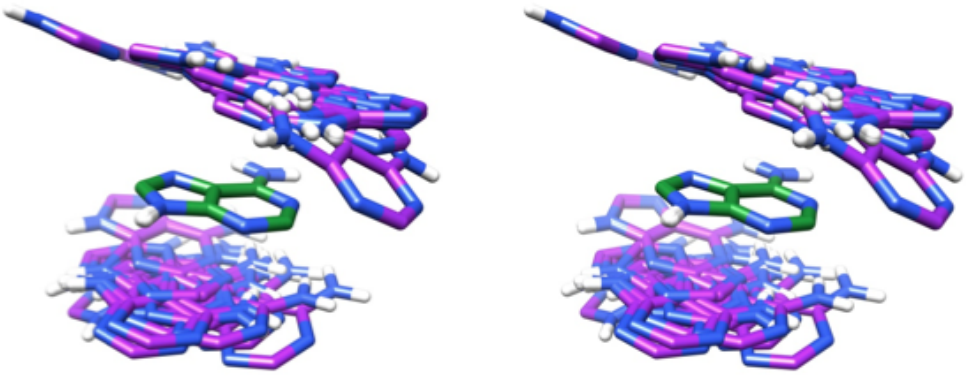
For all the methyl ^{13}C nuclei in CH3Shift-DB database, except Ala C$^{\beta}$, weak correlations are obtained between predicted and experimental chemical shifts. Predictions are done from leave-one-out tests, with standard errors given in ppm. The Pearson correlation coefficients are also shown.


Leu ^{13}Cδ1
SE = 1.469
R=0.151
Leu ^{13}Cδ2
SE = 1.474
R=0.381
Thr ^{13}Cγ2
SE = 0.747
R=0.394
Ile ^{13}Cγ2
SE = 0.844
R=0.463
Val ^{13}Cγ1
SE = 1.008
R=0.322
Ile ^{13}Cδ1
SE = 1.115
R=0.522
Val ^{13}Cγ2
SE = 1.199
R=0.52
Predicted Chemical Shifts (ppm)
Experimental Chemical Shifts (ppm)

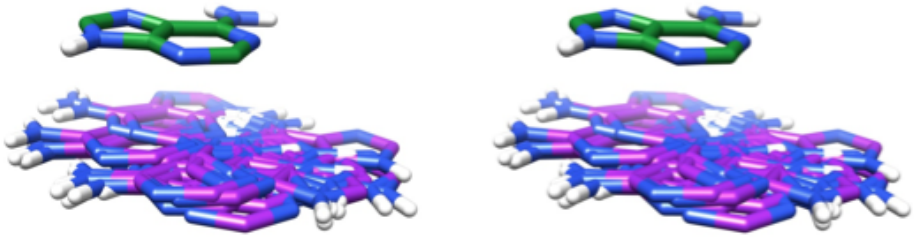
Appendix G

The interring arrangement pattern of all the base pairs in the DiBaseRNA database. Their adjacent (ADJ), spatial (SPT) and hydrogen bonded (HBD) states are shown.

AA

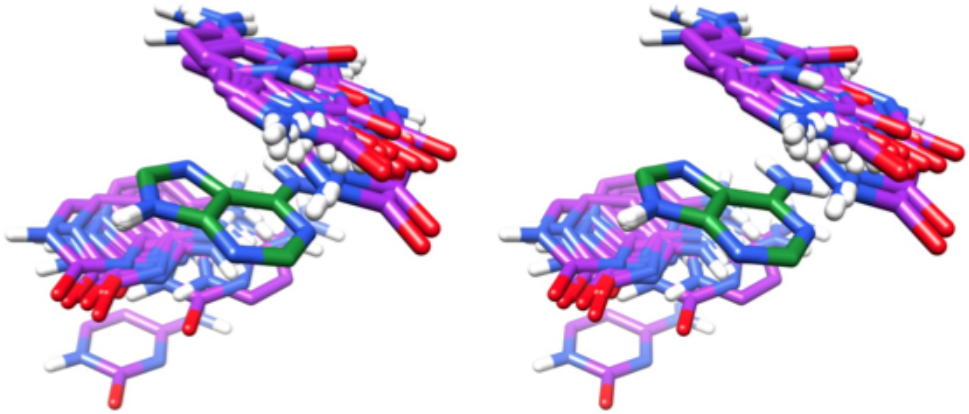


ADJ: superimposed 37 structures of bases in covalently linked adjacent RNA nucleotides.

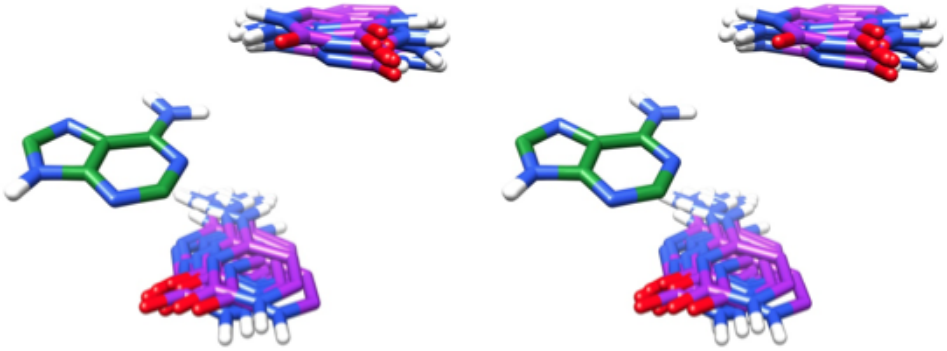


SPT: superimposed 21 structures of bases in only spatially close but not hydrogen bound nucleotides.

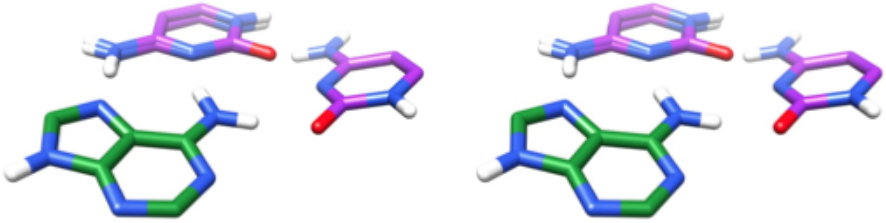
AC



ADJ: superimposed 39 structures of bases in covalently linked adjacent RNA nucleotides.

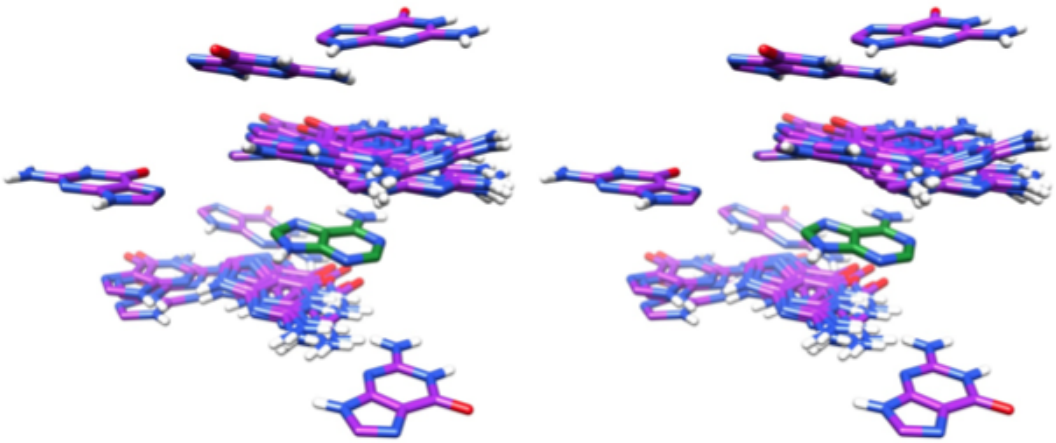


SPT: superimposed 21 structures of bases in only spatially close but not hydrogen bound RNA nucleotides.

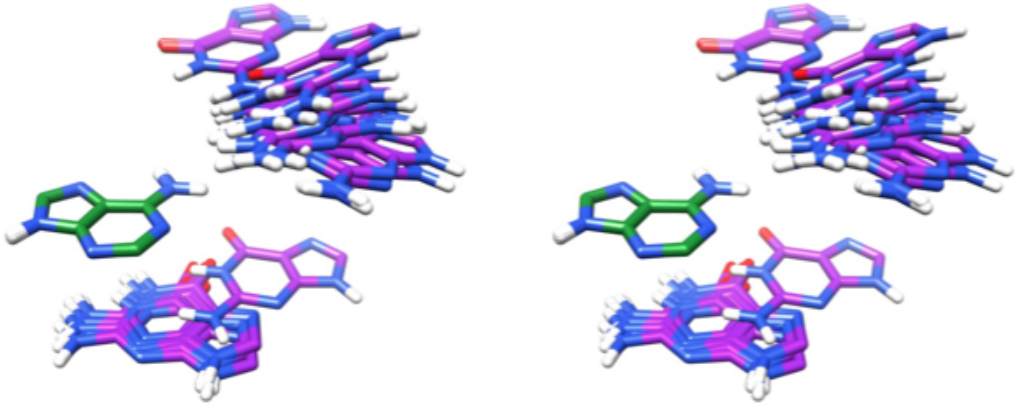


HBD: superimposed 3 structures of bases in hydrogen bonded RNA nucleotides.

AG



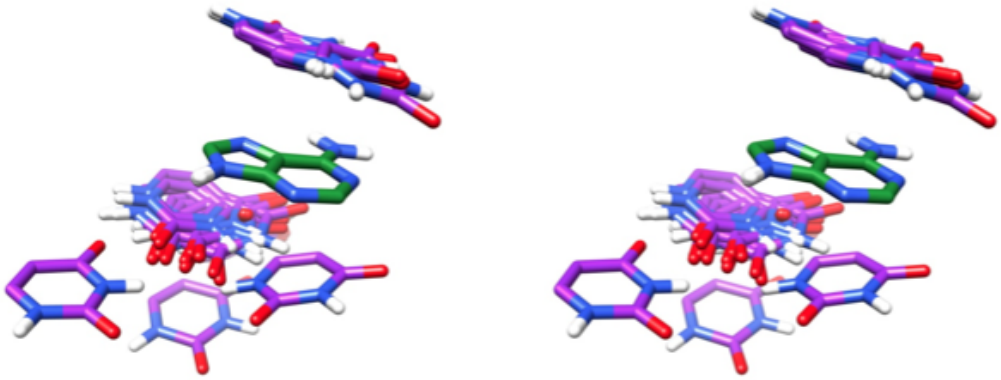
ADJ: superimposed 64 structures of bases in covalently linked adjacent RNA nucleotides.



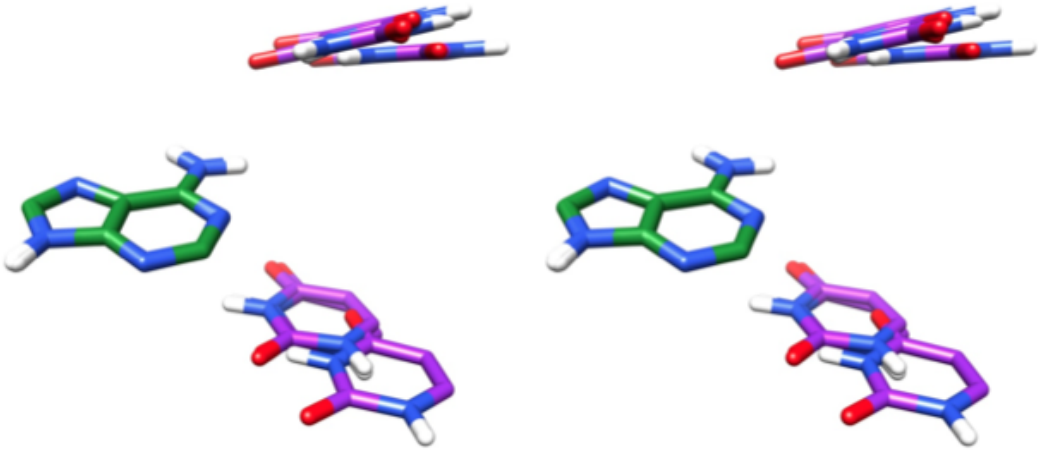
SPT: superimposed 24 structures of bases in only spatially close but not hydrogen bound RNA nucleotides.

HBD: superimposed 6 structures of bases in hydrogen bonded RNA nucleotides.

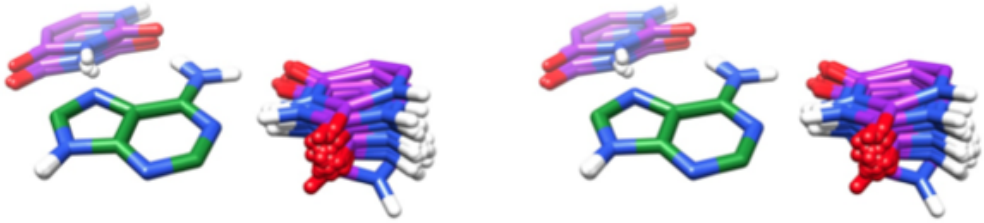
AU



ADJ: superimposed 24 structures of bases in covalently linked adjacent RNA nucleotides.

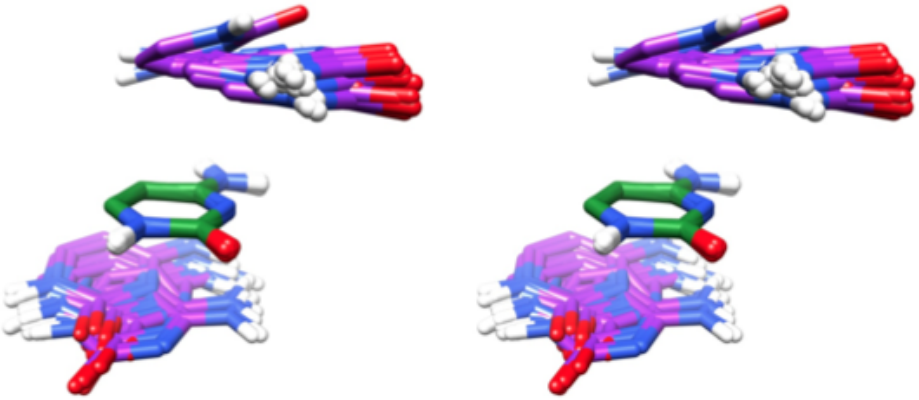


SPT: superimposed 7 structures of bases in only spatially close but not hydrogen bound RNA nucleotides.

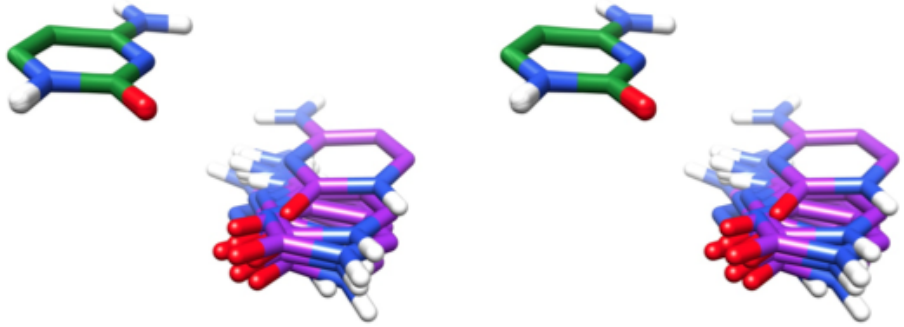


HBD: superimposed 38 structures of bases in hydrogen bonded RNA nucleotides.

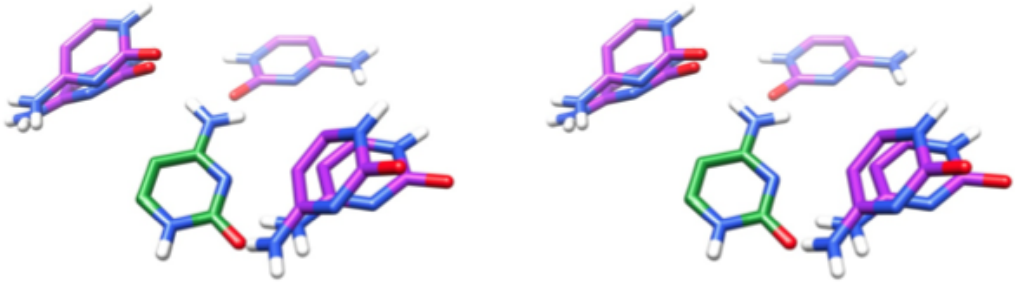
CC



ADJ: superimposed 72 structures of bases in covalently linked adjacent RNA nucleotides.

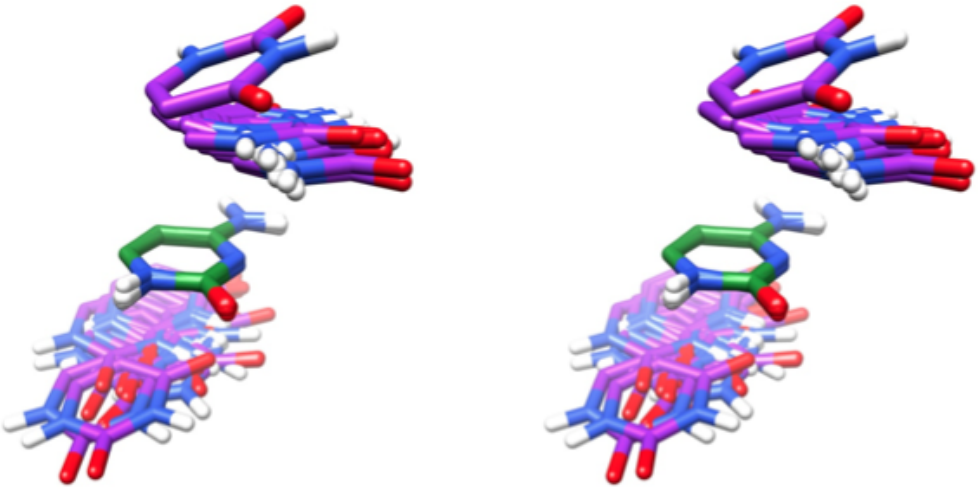


SPT: superimposed 10 structures of bases in only spatially close but not hydrogen bound RNA nucleotides.

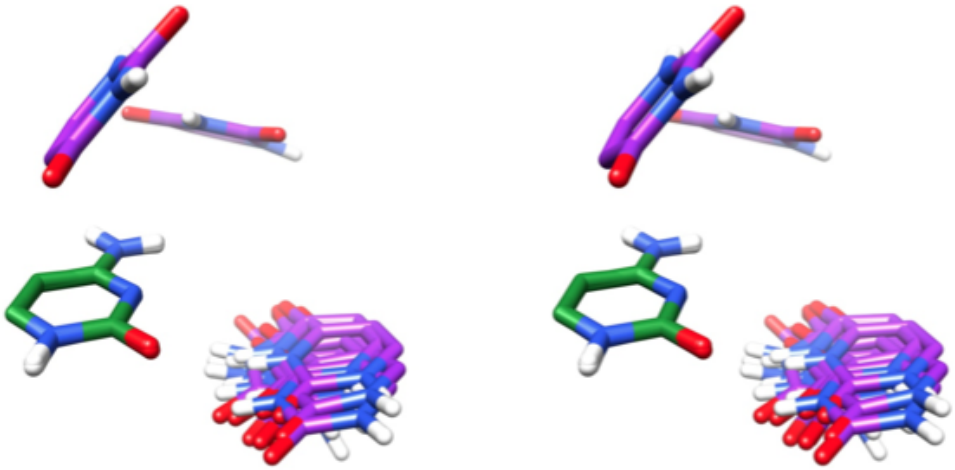


HBD: superimposed 5 structures of bases in hydrogen bonded RNA nucleotides.

CU

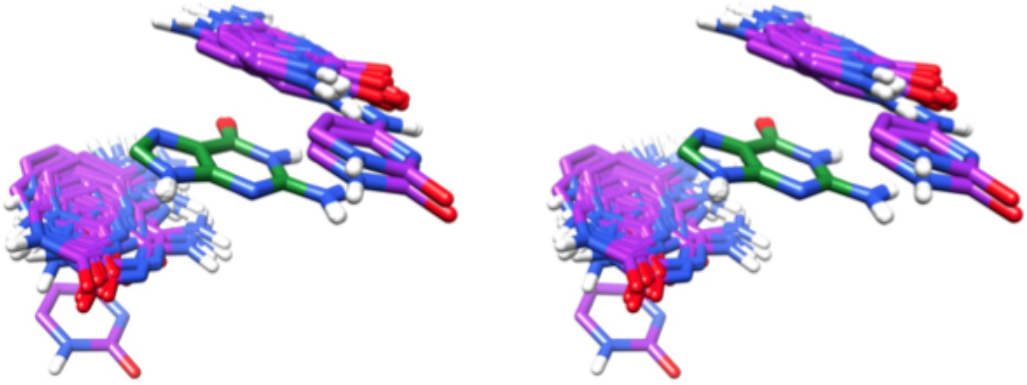


ADJ: superimposed 39 structures of bases in covalently linked adjacent RNA nucleotides.

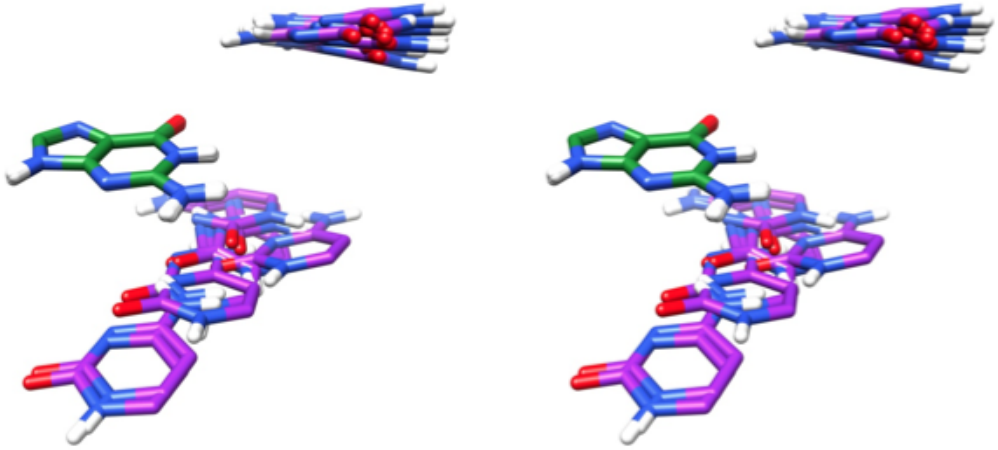


SPT: superimposed 13 structures of bases in only spatially close but not hydrogen bound RNA nucleotides.

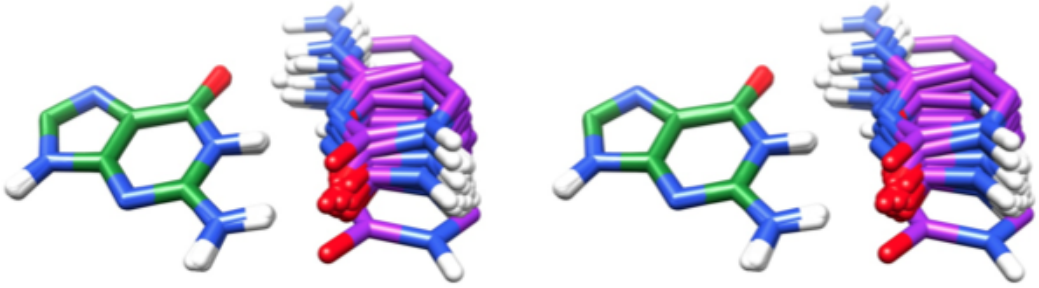
GC



ADJ: superimposed 79 structures of bases in covalently linked adjacent RNA nucleotides.

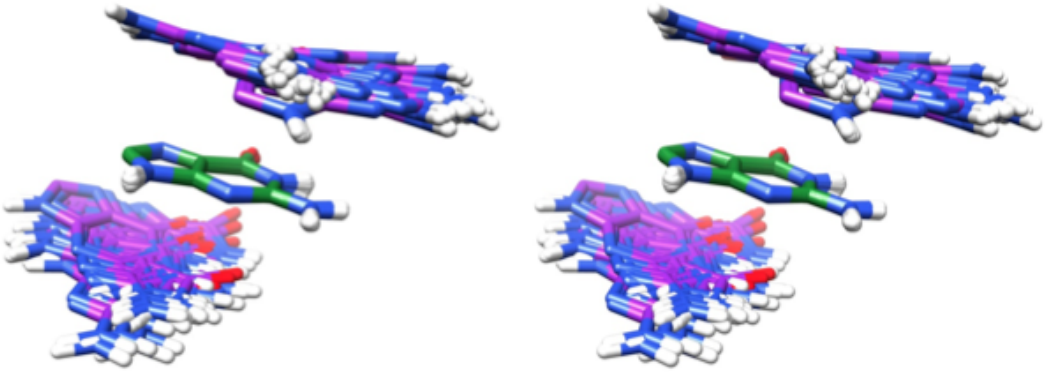


SPT: superimposed 20 structures of bases in only spatially close but not hydrogen bound RNA nucleotides.

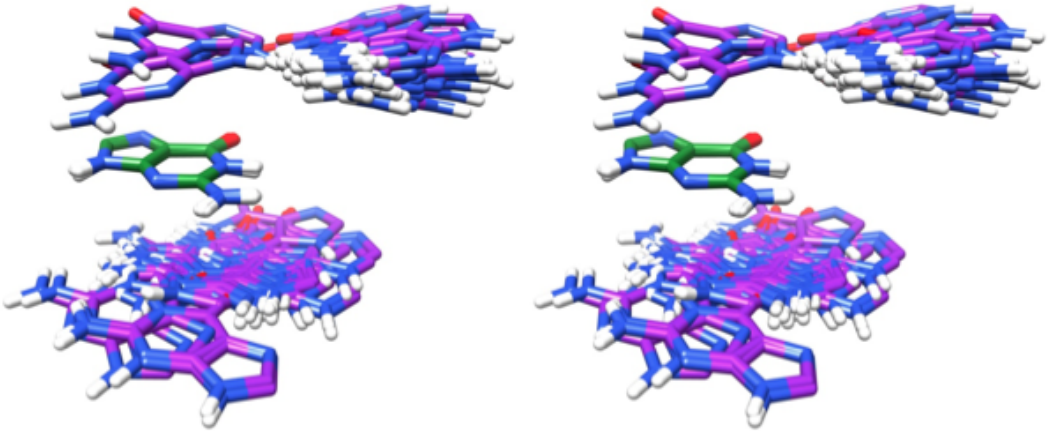


HBD: superimposed 95 structures of bases in hydrogen bonded RNA nucleotides.

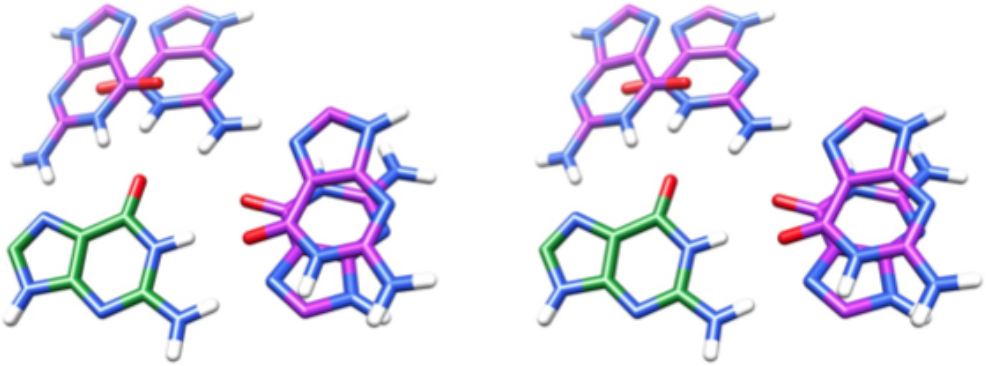
GG



ADJ: superimposed 115 structures of bases in covalently linked adjacent RNA nucleotides.

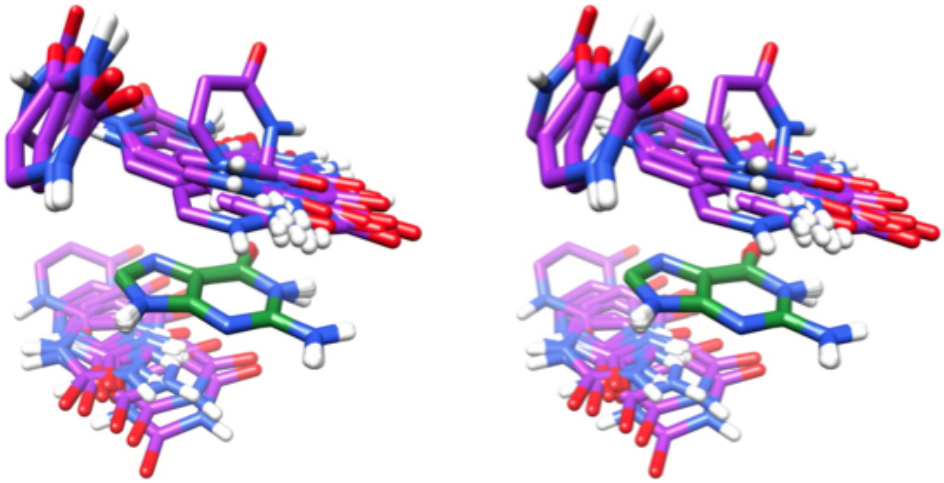


SPT: superimposed 81 structures of bases in only spatially close but not hydrogen bound RNA nucleotides.

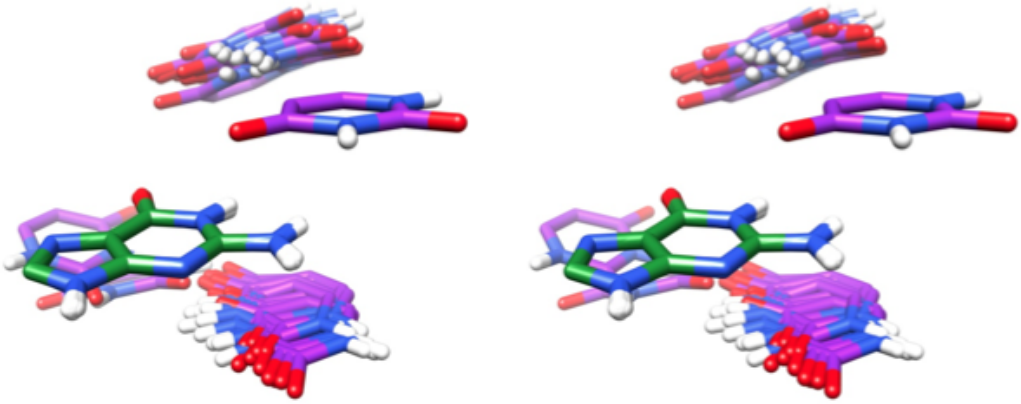


HBD: superimposed 4 structures of bases hydrogen bonded RNA nucleotides.

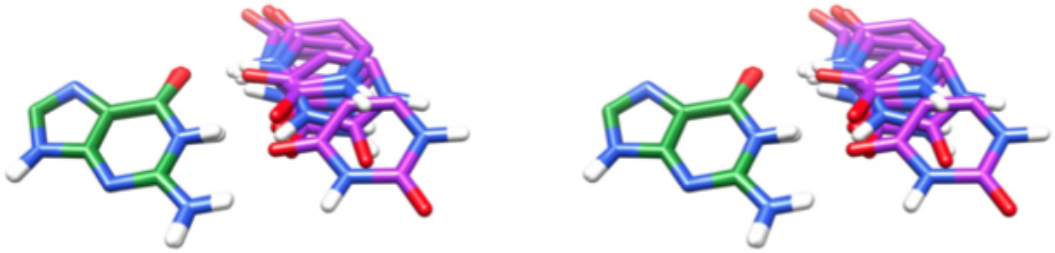
GU



ADJ: superimposed 55 structures of bases in covalently linked adjacent RNA nucleotides.

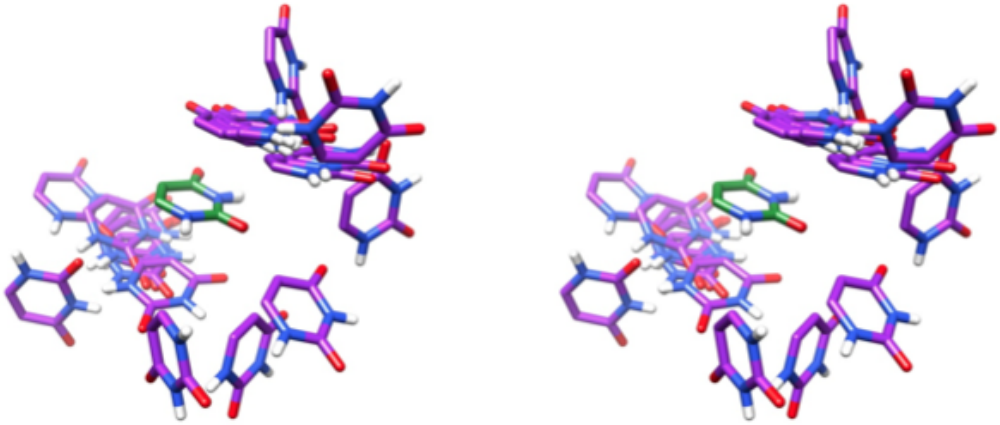


SPT: superimposed 26 structures of bases in only spatially close but not hydrogen bound RNA nucleotides.

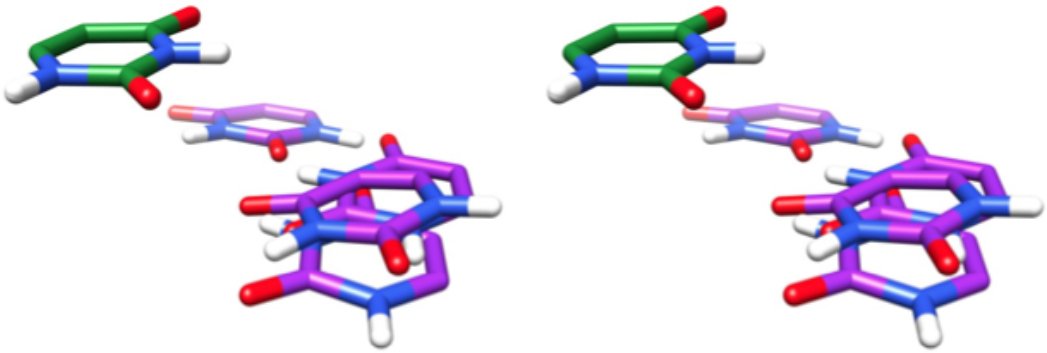


HBD: superimposed 9 structures of bases in hydrogen bonded RNA nucleotides.

UU



ADJ: superimposed 23 structures of bases in covalently linked adjacent RNA nucleotides.

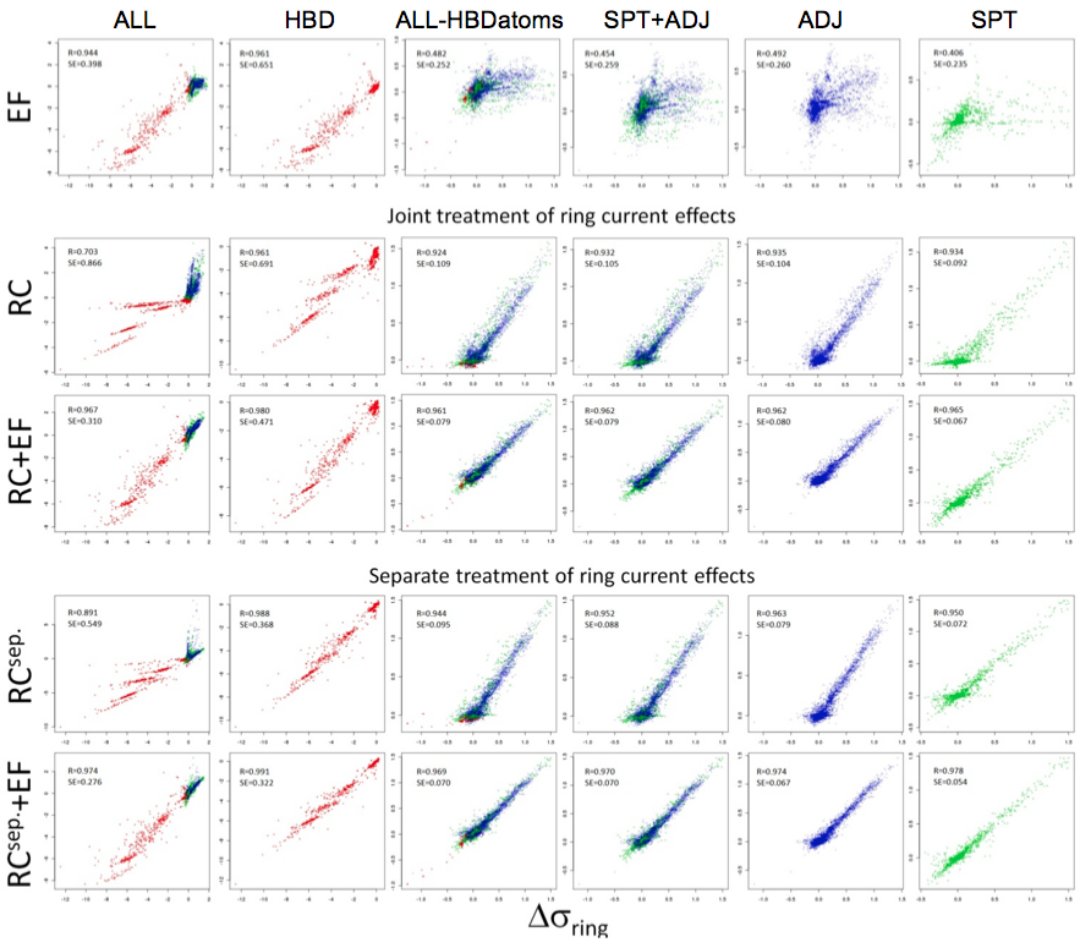


SPT: superimposed 4 structures of bases in only spatially close but not hydrogen bound RNA nucleotides.

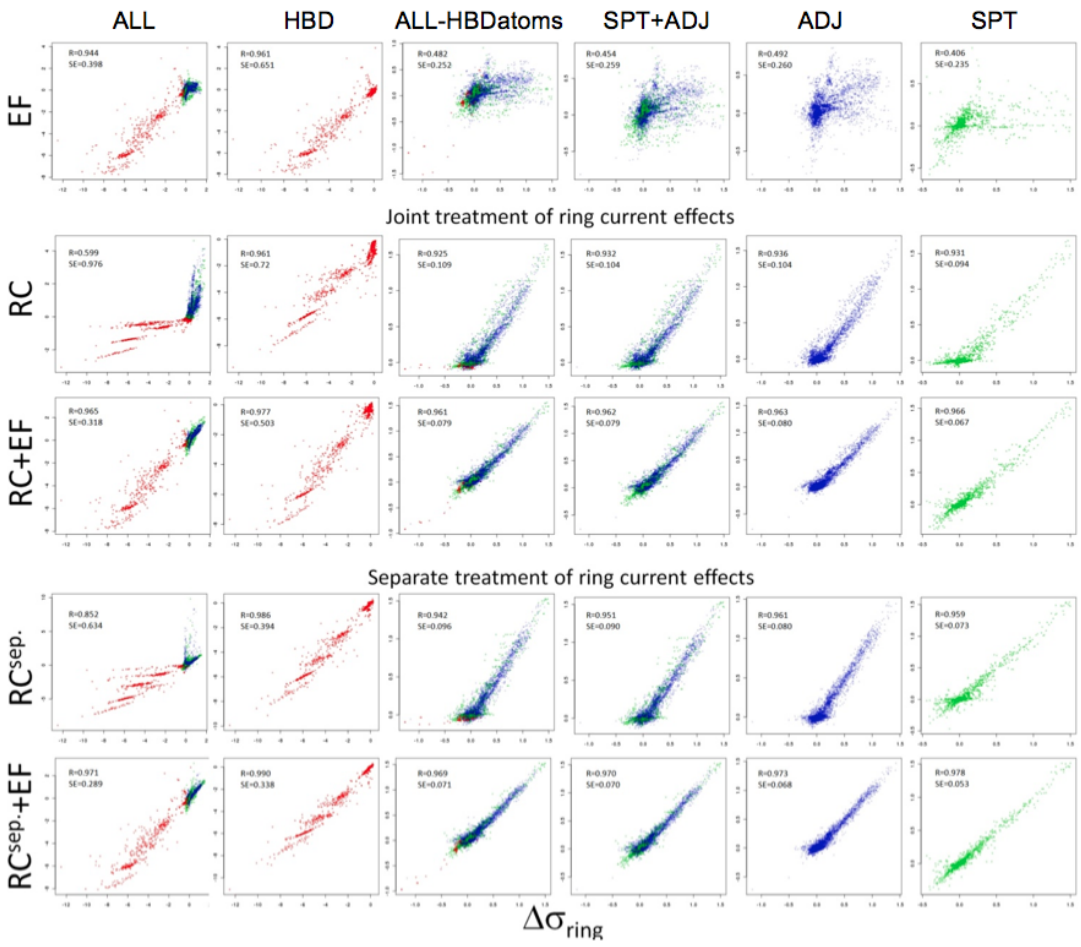
Appendix H

The correlation graphs resulting from the fitting of the calculated ring current induced change in nuclear shielding constants through the electric field (EF), ring current (RC) and ring current + electric field (RC+EF) models. The header of the graph shows the considered nucleus-type and the ring current model (Haigh-Mallion or Pople) used for modelling the dependence. The columns correspond to the treatment of data coming from all di-base arrangement geometries (ALL), only the planar hydrogen-bonded arrangements (HBD, red), from all the geometries, but excluding only the atoms that take part in hydrogen bonding (for A-H...B-C, bond, excluding A, H, B and C, ALL-HBDatoms), from the spatial and adjacent arrangements (SPT+ADJ), from only the adjacent (ADJ, blue) and spatial (SPT, green) arrangements. All the coefficients from such fittings can be obtained via the RingPar server, which does on-fly fitting and reports the results for bot nuclear shielding constants and the anisotropy of nuclear shielding constants. The coefficients correspond to the geometric factors and electric field directions discussed in detail in the associated article.

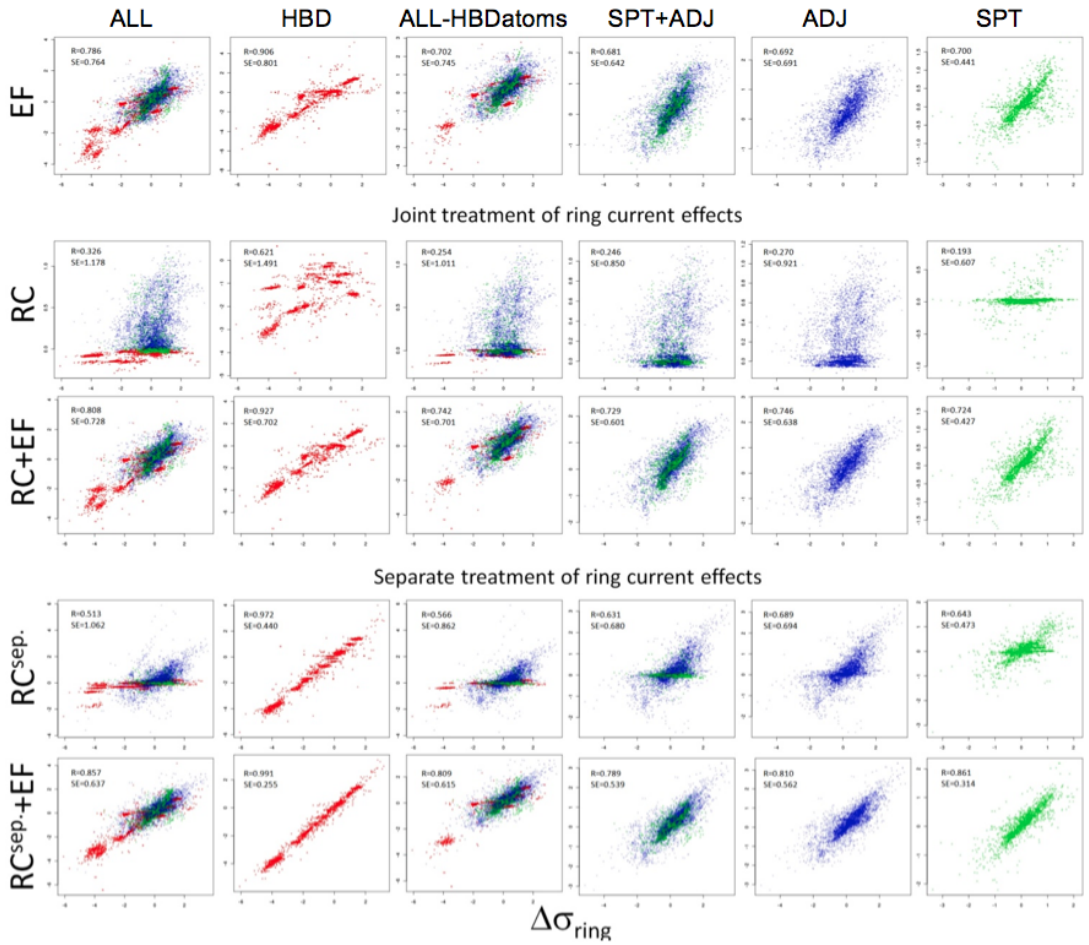
^{1}H/Haigh-Mallion



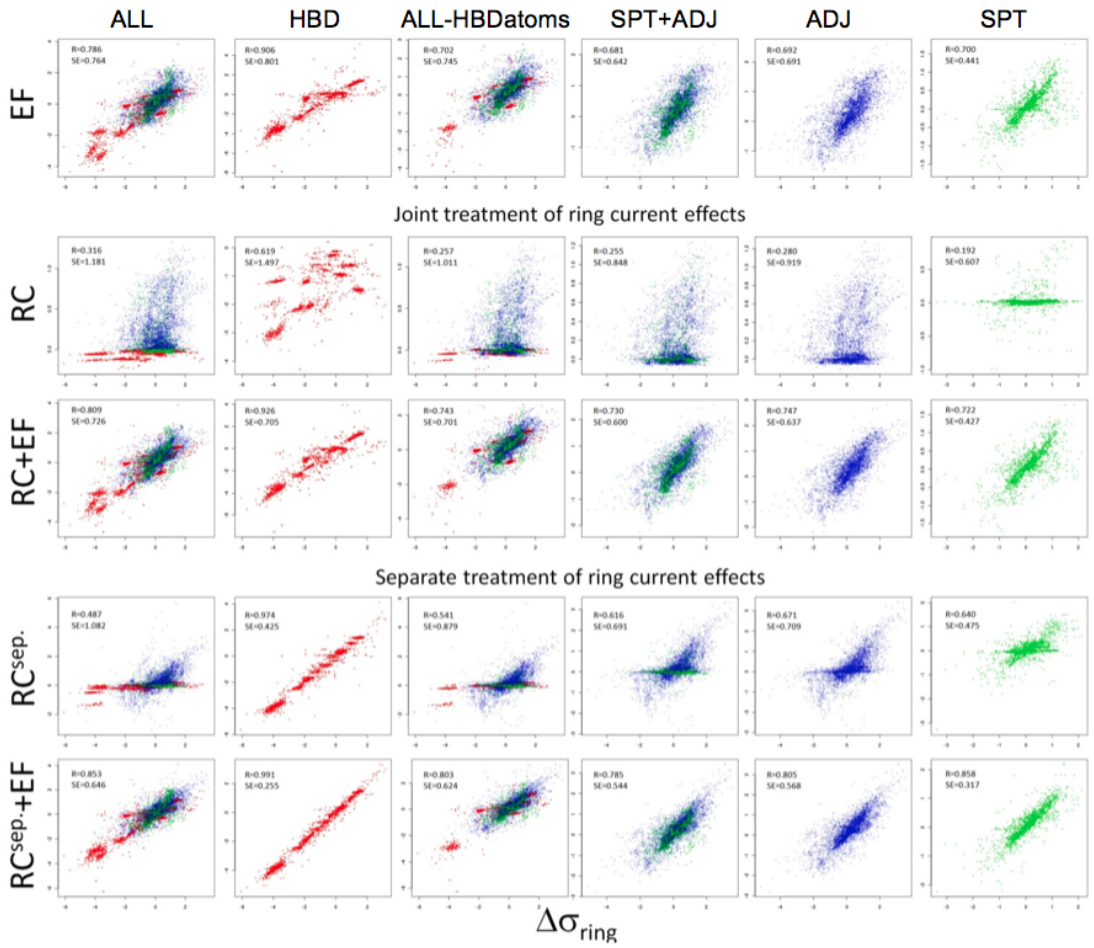
^{1}H/Pople


ALL
HBD
ALL-HBDatoms
SPT+ADJ
ADJ
SPT
EF
R=0.944 SE=0.398
R=0.961 SE=0.651
R=0.482 SE=0.252
R=0.454 SE=0.259
R=0.492 SE=0.260
R=0.406 SE=0.235
Joint treatment of ring current effects
RC
R=0.599 SE=0.976
R=0.961 SE=0.72
R=0.925 SE=0.109
R=0.932 SE=0.104
R=0.936 SE=0.104
R=0.931 SE=0.094
RC+EF
R=0.965 SE=0.318
R=0.977 SE=0.503
R=0.961 SE=0.079
R=0.962 SE=0.079
R=0.963 SE=0.080
R=0.966 SE=0.067
Separate treatment of ring current effects
RC$^{sep.}$
R=0.852 SE=0.634
R=0.986 SE=0.394
R=0.942 SE=0.096
R=0.951 SE=0.090
R=0.961 SE=0.080
R=0.959 SE=0.073
RC$^{sep.}$+EF
R=0.971 SE=0.289
R=0.990 SE=0.338
R=0.969 SE=0.071
R=0.970 SE=0.070
R=0.973 SE=0.068
R=0.978 SE=0.053
$\Delta\sigma_{ring}$

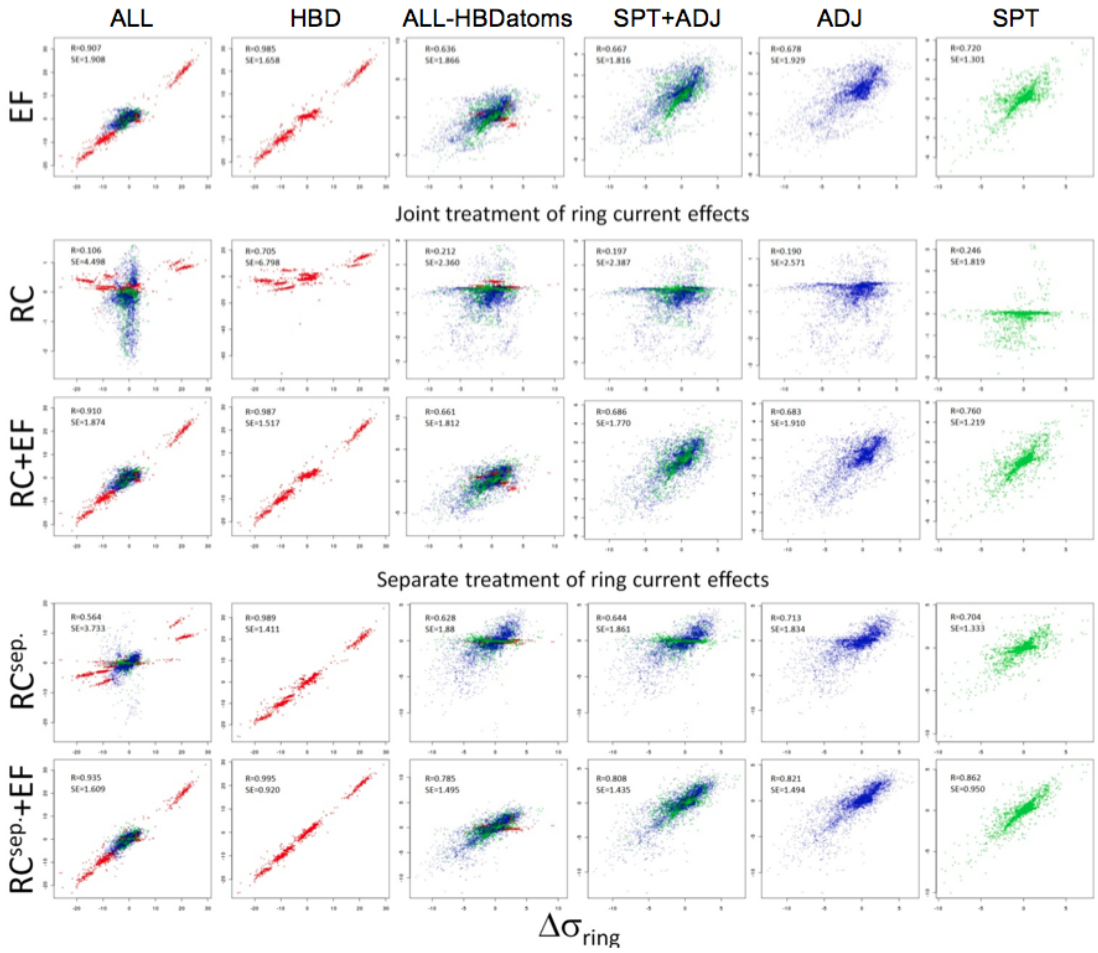
^{13}C/Haigh-Mallion



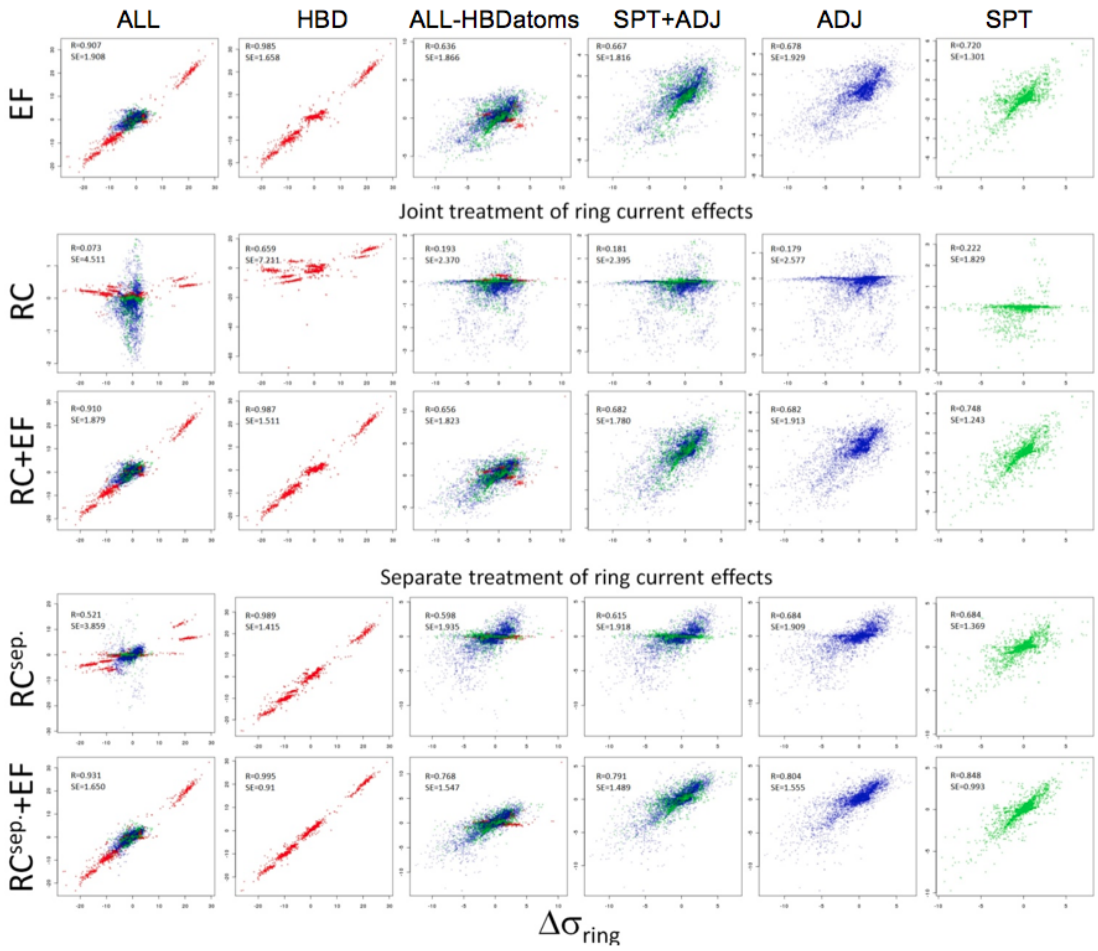
^{13}C/Pople


ALL
HBD
ALL-HBDatoms
SPT+ADJ
ADJ
SPT
EF
Joint treatment of ring current effects
RC
RC+EF
Separate treatment of ring current effects
RC^sep.
RC^sep.+EF
Δσ_ring

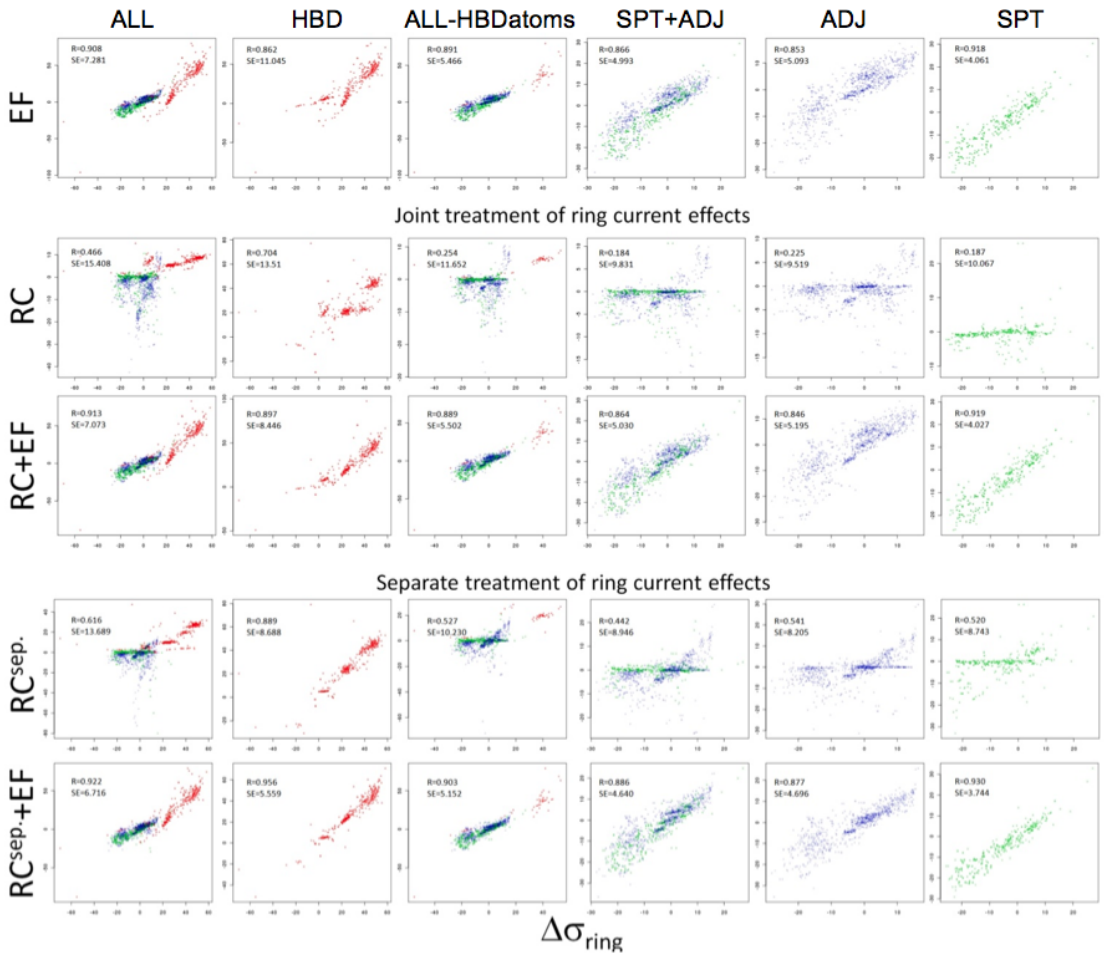
^{15}N/Haigh-Mallion



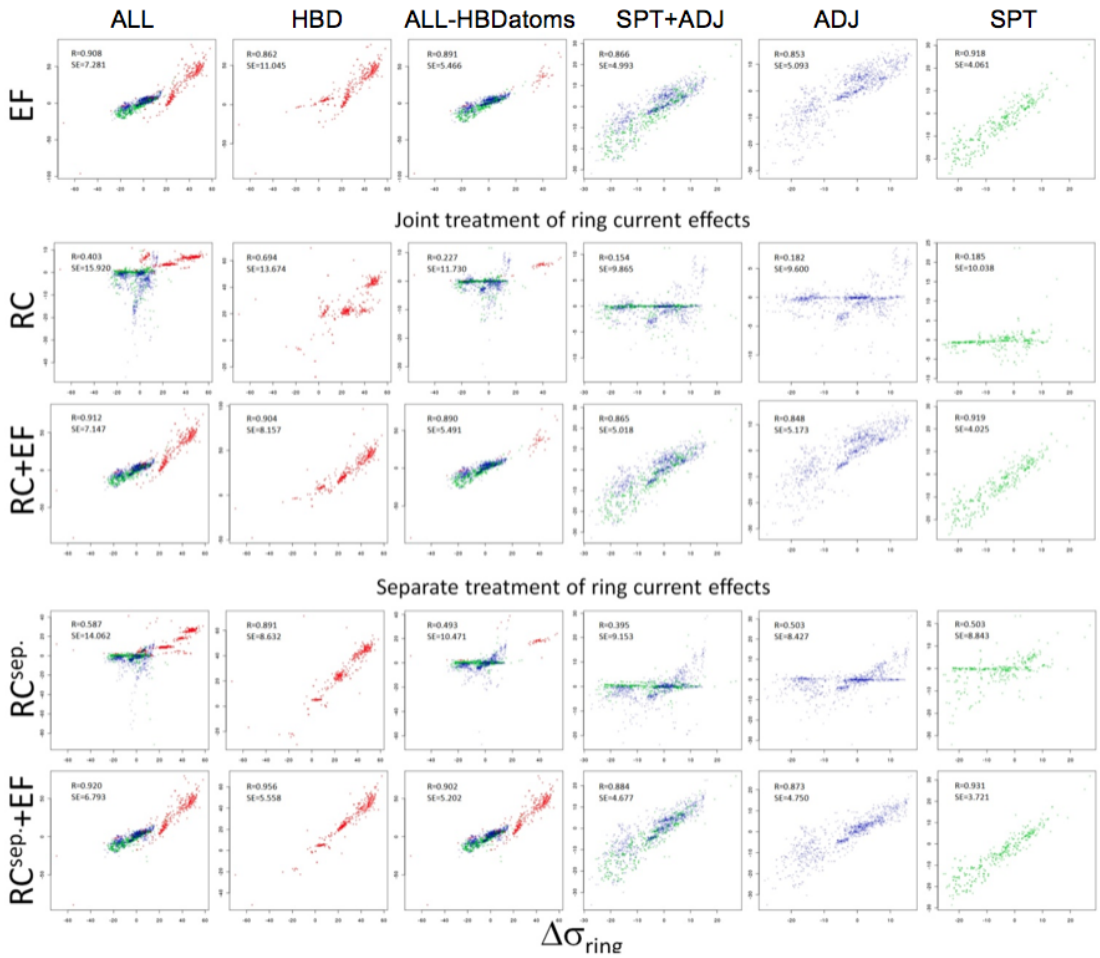
^{15}N/Pople


ALL
HBD
ALL-HBDatoms
SPT+ADJ
ADJ
SPT
EF
R=0.907 SE=1.908
R=0.985 SE=1.658
R=0.636 SE=1.866
R=0.667 SE=1.816
R=0.678 SE=1.929
R=0.720 SE=1.301
Joint treatment of ring current effects
RC
R=0.073 SE=4.511
R=0.659 SE=7.231
R=0.193 SE=2.370
R=0.181 SE=2.395
R=0.179 SE=2.577
R=0.222 SE=1.829
RC+EF
R=0.910 SE=1.879
R=0.987 SE=1.511
R=0.656 SE=1.823
R=0.682 SE=1.780
R=0.682 SE=1.913
R=0.748 SE=1.243
Separate treatment of ring current effects
RCsep.
R=0.521 SE=3.859
R=0.989 SE=1.415
R=0.598 SE=1.935
R=0.615 SE=1.918
R=0.684 SE=1.909
R=0.684 SE=1.369
RCsep.+EF
R=0.931 SE=1.650
R=0.995 SE=0.91
R=0.768 SE=1.547
R=0.791 SE=1.489
R=0.804 SE=1.555
R=0.848 SE=0.993
Δσring

^{17}O/Haigh-Mallion



^{17}O/Pople


ALL
HBD
ALL-HBDatoms
SPT+ADJ
ADJ
SPT
EF
R=0.908 SE=7.281
R=0.862 SE=11.045
R=0.891 SE=5.466
R=0.866 SE=4.993
R=0.853 SE=5.093
R=0.918 SE=4.061
Joint treatment of ring current effects
RC
R=0.403 SE=15.920
R=0.694 SE=13.674
R=0.227 SE=11.730
R=0.154 SE=9.865
R=0.182 SE=9.600
R=0.185 SE=10.038
RC+EF
R=0.912 SE=7.147
R=0.904 SE=8.157
R=0.890 SE=5.491
R=0.865 SE=5.018
R=0.848 SE=5.173
R=0.919 SE=4.025
Separate treatment of ring current effects
RCsep.
R=0.587 SE=14.062
R=0.891 SE=8.632
R=0.493 SE=10.471
R=0.395 SE=9.153
R=0.503 SE=8.427
R=0.503 SE=8.843
RCsep.+EF
R=0.920 SE=6.793
R=0.956 SE=5.558
R=0.902 SE=5.202
R=0.884 SE=4.677
R=0.873 SE=4.750
R=0.931 SE=3.721
Δσring

References

Abraham, R., Canton, M. & Griffiths, L. (2001). Proton chemical shifts in nmr: Part 17. chemical shifts in alkenes and anisotropic and steric effects of the double bond. *Magn. Reson. Chem.*, **39**, 421–431. 42, 68

Adamo, C. & Barone, V. (1998). Toward chemical accuracy in the computation of nmr shielding: the pbe0 model. *Chem. Phys. Lett.*, **298**, 113–119. 9, 120

Agarwal, V., Xue, Y., Reif, B. & Skrynnikov, N.R. (2008). Protein side-chain dynamics as observed by solution- and solid-state nmr spectroscopy: a similarity revealed. *J. Am. Chem. Soc.*, **130**, 16611–16621. 49

Ágoston, B.S., Kovács, D., Tompa, P. & Perczel, A. (2011). Full backbone assignment and dynamics of the intrinsically disordered dehydrin erd14. *Biomol. NMR Assign.*, **5**, 189–193. 91

Anstrom, D.M., Kallio, K. & Remington, S.J. (2003). Structure of the escherichia coli malate synthase g:pyruvate:acetyl-coenzyme a abortive ternary complex at 1.95 å resolution. *Protein Sci.*, **12**, 1822–1832. 97

Avbelj, F., Kocjan, D. & Baldwin, R.L. (2004). Protein chemical shifts arising from α-helices and β-sheets depend on solvent exposure. *Proc. Natl. Acad. Sci. USA*, **101**, 17394–17397. 28, 30, 31

Baldwin, A.G., Religa, T.L., Hansen, D.F., Bouvignies, G. & Kay, L.E. (2010). $^{13}chd_2$ methyl group probes of millisecond time scale exchange

in proteins by ^{1}h relaxation dispersion: an application to proteasome gating residue dynamics. *J. Am. Chem. Soc.*, **132**, 10992–10995. 33

Baldwin, A.J. & Kay, L.E. (2009). Nmr spectroscopy brings invisible protein states into focus. *Nat. Chem. Biol.*, **5**, 808–814. 88

Barone, V. (1995). *Recent Advances in Density Functional Methods*, vol. 1. World Scientific, Singapore. 16

Bax, A. (1994). Multidimensional nuclear magnetic resonance methods for protein studies. *Curr. Opin. Structu. Biol.*, **4**, 738–744. 88

Bax, A. & Grishaev, A. (2005). Weak alignment nmr: a hawk-eyed view of biomolecular structure. *Curr. Opin. Struct. Biol.*, **15**, 563–570. 11, 88

Becconsall, J.K. & Hampson, P. (1965). Solvent effects in proton and ^{13}c n.m.r. chemical shifts of polar compounds. *Mol. Phys.*, **10**, 21–32. 21

Becke, A.D. (1993). Density functional thermochemistry. iii. the role of exact exchange. *J. Chem. Phys.*, **98**, 5648–5652. 15, 111, 118

Benzi, C., Crescenzi, O., Pavone, M. & Barone, V. (2004). Reliable nmr chemical shifts for molecules in solution by methods rooted in density functional theory. *Magn. Reson. Chem.*, **42**, S57–S67. 8

Berjanskii, M., Liang, Y., Zhou, J., Tang, P., Stothard, P., Zhou, Y., Cruz, J., MacDonell, C., Lin, G., Lu, P. & Wishart, D.S. (2010). Prosess: a protein structure evaluation suite and server. *Nucl. Acids Res.*, **38**, W633–W640. 92

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. & Bourne, P.E. (2000). The protein data bank. *Nucl. Acids Res.*, **28**, 235–242. 1, 9, 36, 65

Biot, C., Buisine, E. & Rooman, M. (2003). Free-energy calculations of protein-ligand cation-π and amino-π interactions: from vacuum to protein like environments. *J. Am. Chem. Soc.*, **125**, 13988–13994. 25

Bissantz, C., Kuhn, B. & Stahl, M. (2010). A medicinal chemist's guide to molecular interactions. *J. Med. Chem.*, **53**, 5061–5084. 64

Blackledge, M. (2004). Recent progress in the study of biomolecular structure and dynamics in solution from residual dipolar couplings. *Prog. Nucl. Magn. Reson. Spec.*, **46**, 23–61. 11

Brutscher, B. (2001). Accurate measurement of small spin-spin couplings in partially aligned molecules using a novel j-mismatch compensated spin-state-selection filter. *J. Magn. Reson.*, **151**, 332–338. 88

Buckingham, A.D. (1960). Chemical shifts in the nuclear magnetic resonance spectra of molecules containing polar groups. *Can. J. Chem.*, **38**, 300–307. 6, 23, 34, 42, 65, 67

Buckingham, A.D. & Pople, J.A. (1963). High-resolution n.m.r. spectra in electric fields. *Trans. Faraday Soc.*, **59**, 2421–2430. 6, 34, 42

Burley, S.K., Almo, S.C., Bonanno, J.B., Capel, M., Chance, M.R., Gaasterland, T., Lin, D., Sali, A., Studier, F.W. & Swaminathan, S. (1999). Structural genomics: beyond the human genome project. *Nat. Genet.*, **23**, 151–157. 88

Bussi, G., Donadio, D. & Parrinello, M. (2007). Canonical sampling through velocity rescaling. *J. Chem. Phys.*, **126**, 014101. 85

Camilloni, C., Robustelli, P., De Simone, A., Cavalli, A. & Vendruscolo, M. (2012). Characterisation of the conformational equilibrium between the two major substates of rnase a using nmr chemical shifts. *J. Am. Chem. Soc.*, **134**, 3968–3971. 11

Cancès, M.T., Mennucci, B. & Tomasi, J. (1997). A new integral equation formalism for the polarizable continuum model: Theoretical background and applications to isotropic and anisotropic dielectrics. *J. Chem. Phys.*, **107**, 3032–3041. 16, 19

Case, D.A. (1995). Calibration of ring-current effects in proteins and nucleic acids. *J. Biomol. NMR*, **6**, 341–346. 36, 40, 104

Case, D.A. (1998). The use of chemical shifts and their anisotropies in biomolecular structure determination. *Curr. Opin. Struct. Biol.*, **8**, 624–630. 4, 30, 103, 104

Case, D.A. (2000). Interpretation of chemical shifts and coupling constants in macromolecules. *Curr. Opin. Struct. Biol.*, **10**, 197–203. 4

Cavalli, A., Salvatella, X., Dobson, C.M. & Vendruscolo, M. (2007). Protein structure determination from nmr chemical shifts. *Proc. Natl. Acad. Sci. USA*, **104**, 9615–9620. 12, 14, 33, 89, 103

Chattopadhyaya, R., Meador, W.E., Means, A.R. & Quiocho, F.A. (1992). Calmodulin structure refined at 1.7 å resolution. *J. Mol. Biol.*, **228**, 1177–1192. 77

Cheeseman, J.R., Trucks, G.W., Keith, T.A. & Frisch, M.J. (1996). A comparison of models for calculating nuclear magnetic resonance shielding tensors. *J. Chem. Phys.*, **104**, 5497–5509. 120

Clore, G.M., Omichinski, J.G., Sakaguchi, K., Zambrano, N. & et al, H.S. (1995). Interhelical angles in the solution structure of the oligomerization domain of p53: correction. *Science*, **267**, 1515–1516. 89

Cornilescu, G., Marquardt, J.L., Ottiger, M. & Bax, A. (1998). Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase. *J. Am. Chem. Soc.*, **120**, 6836–6837. 57, 75

Cornilescu, G., Delaglio, F. & Bax, A. (1999). Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J. Biomol. NMR*, **13**, 289–302. 9

Cossi, M. & Crescenzi, O. (2004). Solvent effects on ^{17}o nuclear magnetic shielding: N-methylformamide in polar and apolar solutions. *Theor. Chem. Acc.*, **111**, 162–167. 19

Cromsigt, J.A.M.T.C., Hilbers, C.W. & Wijmenga, S.S. (2001). Prediction of proton chemical shifts in rna. *J. Biomol. NMR*, **21**, 11–19. 103

Crowley, P.B. & Golovin, A. (2005). Cation-π interactions in protein-protein interfaces. *Proteins: Struct. Funct. Bioinf.*, **59**, 231–239. 64

Das, R., Andre, I., Shen, Y., Wu, Y.B., Lemak, A., Bansal, S., Arrowsmith, C.H., Szyperski, T. & Baker, D. (2009). A transient and low-populated protein-folding intermediate at atomic resolution. *Proc. Natl. Acad. Sci. USA*, **106**, 18978–18983. 33, 64

de Dios, A.C., Pearson, J.G. & Oldfield, E. (1993). Chemical shifts in proteins: an ab initio study of carbon-13 nuclear magnetic resonance chemical shielding in glycine, alanine, and valine residues. *J. Am. Chem. Soc.*, **115**, 9768–9773. 19, 25

De Simone, A., Cavalli, A., Hsu, S.T.D., Vranken, W. & Vendruscolo, M. (2009a). Accurate random coil chemical shifts from an analysis of loop regions in native states of proteins. *J. Am. Chem. Soc.*, **131**, 16332–16333. 53

De Simone, A., Richter, B., Salvatella, X. & Vendruscolo, M. (2009b). Toward an accurate determination of free energy landscapes in solution states of proteins. *J. Am. Chem. Soc.*, **131**, 3810–3811. 11

DeGortari, I., Portella, G., Salvatella, X., Bajaj, V.S., van der Wel, P.S., Yates, J.R., Segall, M.D., Pickard, C.J., Payne, M.C. & Vendruscolo, M. (2010). Time averaging of nmr chemical shifts in the mlf peptide in the solid state. *J. Am. Chem. Soc.*, **132**, 5993–6000. 49

Dewar, M.J.S., Zoebisch, E.G., Healy, E.F. & Stewart, J.J.P. (1985). Am1: a new general purpose quantum mechanical model. *J. Am. Chem. Soc.*, **107**, 3902–3909. 116

Ditchfield, R. (1974). Self-consistent perturbation theory of diamagnetism. 1. gauge-invariant lcao method for nmr chemical shifts. *Mol. Phys.*, **27**, 789–807. 8, 16, 120

Duan, Y., Wu, C., Chowdhury, S., Lee, M.C., Xiong, G., Zhang, W., Yang, R., Cieplak, P., Luo, R., Lee, T., Caldwell, J., Wang, J. &

Kollman, P. (2003). A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.*, **24**, 1999–2012. 38, 42, 66, 67

Farès, C., Lakomek, N.A., Walter, K.F.A., Frank, B.T.C., Meiler, J., Becker, S. & Griesinger, C. (2009). Accessing ns-μs side chain dynamics in ubiquitin with methyl rdcs. *J. Biomol. NMR*, **45**, 23–44. 57, 148

Fiser, A., Do, R.K. & Sali, A. (2000). Modeling of loops in protein structures. *Protein Sci.*, **9**, 1753–1773. 97

Fitch, C.A., Karp, D.A., Lee, K.K., Stites, W.E., Lattman, E.E. & Garcia-Moreno, B. (2002). Experimental pk_a values of buried residues: analysis with continuum methods and role of water penetration. *Biophys. J.*, **82**, 3289–3304. 26

Frank, B.S., Vardar, D., Buckley, D.A. & McKnight, C. (2002). The role of aromatic residues in the hydrophobic core of the villin headpiece subdomain. *J. Protein. Sci.*, **11**, 680–687. 64

Fröhlich, H. (1958). *Theory of Dielectrics*. Clarendon, Oxford, UK. 26

Garcia-Moreno, B., Dwyer, J.J., Lattman, E.E., Spencer, D.S. & Stites, W.E. (1997). Experimental measurement of the effective dielectric in the hydrophobic core of a protein. *Biophys. Chem.*, **64**, 211–224. 26

Gelis, I., Bonvin, A.M., Keramisanou, D., Koukaki, M., Gouridis, G., Karamanou, S., Economou, A. & Kalodimos, C.G. (2007). Structural basis for signal-sequence recognition by the translocase motor seca as determined by nmr. *Cell*, **131**, 756–769. 33

Goto, N.K. & Kay, L.E. (2000). New developments in isotope labeling strategies for protein solution nmr spectroscopy. *Curr. Opin. Struct. Biol.*, **10**, 585–592. 1, 34, 88, 97, 100

Grishaev, A., Tugarinov, V., Kay, L.E., Trewhella, J. & Bax, A. (2008). Refined solution structure of the 82-kda enzyme malate synthase g

from joint nmr and synchrotron saxs restraints. *J. Biomol. NMR*, **40**, 95–106. 97

Haigh, C.W. & Mallion, R.B. (1972). New tables of ring current shielding in proton magnetic resonance. *Org. Magn. Reson.*, **4**, 203–228. 34, 40, 66, 104, 105, 108, 109, 110

Haigh, C.W. & Mallion, R.B. (1980). Ring current theories in nuclear magnetic resonance. *Prog. Nucl. Magn. Res. Spectrosc.*, **13**, 303–344. 34, 40, 65, 66, 104, 105, 106, 108, 110

Han, B., Liu, Y., Ginzinger, S.W. & Wishart, D.S. (2011). Shiftx2: significantly improved protein chemical shift prediction. *J. Biomol. NMR*, **50**, 43–57. 64, 80

Han, W.G., Jalkanen, K.J., Elstner, M. & Suhai, S. (1998). Theoretical study of aqueous n-acetyl-l-alanine n'-methylamide: structures and rama, vcd, and roa spectra. *J. Phys. Chem. B*, **102**, 2587–2602. 15

Hansen, D.F., Neudecker, P., Vallurupalli, P., Mulder, F.A.A. & Kay, L.E. (2010). Determination of leu side-chain conformations in excited protein states by nmr relaxation dispersion. *J. Am. Chem. Soc.*, **132**, 42–43. 33, 61

Hass, M.A.S., Jensen, M.R. & Led, J.J. (2008). Probing electric fields in proteins in solution by nmr spectroscopy. *Proteins: Struc. Func. Bioinf.*, **73**, 333–343. 28

Havlin, R.H., Le, H., Laws, D.D., de Dios, A.C. & Oldfield, E. (1997). An ab initio quantum chemical investigation of carbon-13 nmr shielding tensors in glycine, alanine, valine, isoleucine, serine, and threonine: comparisons between helical and sheet tensors, and the effects of χ_1 on shielding. *J. Am. Chem. Soc.*, **119**, 11951–11958. 19, 30

Head-Gordon, T., Head-Gordon, M., Frisch, M.J., Brooks, C.L. & Pople, J.A. (1991). Theoretical study of blocked glycine and alanine peptide analogues. *J. Am. Chem. Soc.*, **113**, 5989–5997. 15

Heinemann, U., Illing, G. & Oschkinat, H. (2001). High-throughput three-dimensional protein structure determination. *Curr. Oppin. Biotech.*, **12**, 348–354. 88

Helgaker, T., Jaszuński, M. & Ruud, K. (1999). Ab initio methods for the calculation of nmr shielding and indirect spin-spin coupling constants. *Chem. Rev.*, **99**, 293–352. 7, 8, 9, 16

Hoch, J.C., Dobson, C.M. & Karplus, M. (1985). Vicinal coupling constants and protein dynamics. *Biochemistry*, **24**, 3831–3841. 11

Hong, M., Mishanina, T.V. & Cady, S.D. (2009). Accurate measurement of methyl ^{13}c chemical shifts by solid-state nmr for the determination of protein side chain conformation: the influenza a m2 transmembrane peptide as an example. *J. Am. Chem. Soc.*, **131**, 7806–7816. 47

Honig, B. & Nicholls, A. (1995). Classical electrostatics in biology and chemistry. *Science*, **268**, 1144–1149. 26

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 Struct. Funct. Gen.*, **65**, 712–725. 83

Howard, B.R., Endrizzi, J.A. & Remington, S.J. (2000). Crystal structure of escherichia coli malate synthase g complexed with magnesium and glyoxylate at 2.0 å resolution: mechanistic implications. *Biochemistry*, **39**, 3156–3168. 97

Hsu, S.T.D., Cabrita, L.D., Fucini, P., Christodoulou, J. & Dobson, C.M. (2009). Probing side-chain dynamics of a ribosome-bound nascent chain using methyl nmr spectroscopy. *J. Am. Chem. Soc.*, **131**, 8366–8367. 33

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921. 88

Jalkanen, K.J., Elstner, M. & Suhai, S. (2004). Amino acids and small peptides as building blocks for proteins: comparative theoretical and spectroscopic studies. *J. Mol. Struct. Theochem*, **675**, 61–77. 16

Jameson, C.J. (1996). Understanding nmr chemical shifts. *Annu. Rev. Phys. Chem.*, **47**, 135–169. 4, 8, 33, 34, 91

Johnson, C.E. & Bovey, F.A. (1958). Calculation of nuclear magnetic resonance spectra of aromatic hydrocarbons. *J. Chem. Phys.*, **29**, 1012–1014. 104, 105, 106, 107, 108, 109, 110

Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W. & Klein, M.L. (1983). Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, **79**, 926–935. 85

Kainosho, M., Torizawa, T., Iwashita, Y., Terauchi, T., Mei, M.O. & Guntert, P. (2006). Optimal isotope labeling for nmr protein structure determinations. *Nature*, **440**, 52–57. xvi, 1, 34, 64, 78, 79, 88, 93, 94, 97, 100

Karplus, M. & McCammon, J.A. (2002). Molecular dynamics simulations of biomolecules. *Nat. Struc. Biol.*, **9**, 646–652. 11

Kern, D. & Zuiderweg, E.R.P. (2003). The role of dynamics in allosteric regulation. *Curr. Opin. Struct. Biol.*, **13**, 748–757. 10

Kirkwood, J.G. (1937). The dielectric polarizability of polar liquids. *J. Chem. Phys.*, **7**, 911–919. 26

Kohlhoff, K.J., Robustelli, P., Cavalli, A., Salvatella, X. & Vendruscolo, M. (2009). Fast and accurate predictions of protein nmr chemical shifts from interatomic distances. *J. Am. Chem. Soc.*, **131**, 13894–13895. 1, 10, 12, 14, 33, 34, 42, 43, 49, 65, 67, 89, 91, 103

Kohn, W. & Sham, L.J. (1965). Self-consistent equations including exchange and correlation effects. *Phys. Rev. A*, **140**, 1133–1138. 15, 111, 118, 120

Korzhnev, D.M., Skrynnikov, N.R., Millet, O., Torchia, D.A. & Kay, L.E. (2002). An nmr experiment for the accurate measurement of heteronuclear spin-lock relaxation rates. *J. Am. Chem. Soc.*, **124**, 10743–10753. 88

Korzhnev, D.M., Religa, T.L., Banachewicz, W., Fersht, A.R. & Kay, L.E. (2010). A transient and low-populated protein-folding intermediate at atomic resolution. *Science*, **329**, 1312–1316. 33, 88

Krishnan, R., Binkley, J.S., Seeger, R. & Pople, J.A. (1980). Self-consistent molecular orbital methods. xx. a basis set for correlated wave functions. *J. Chem. Phys.*, **72**, 650–654. 16, 111, 118, 120

Ksiazek, A., Borowski, P. & Wolinski, K. (2009). Theoretical analysis of solvent effects on nitrogen nmr chemical shifts in oxazoles and oxadiazoles. *J. Magn. Reson.*, **197**, 153–160. 19

Lam, S.L. & Chi, L.M. (2010). Use of chemical shifts for structural studies of nucleic acids. *Prog. Nucl. Magn. Res. Spectrosc.*, **56**, 289–310. 103

Lambert, L.J., Schirf, V., Demeler, B., Cadene, M. & Werner, M.H. (2004). Flipping a genetic switch by subunit exchange. *EMBO J.*, **23**, 3186. 89

Lange, O.F., Lakomek, N.A., Farès, C., Schröder, G.F., Walter, K.F.A., Becker, S., Meiler, J., Grubmüller, H., Griesinger, C. & de Groot, B.L. (2008). Recognition dynamics up to microseconds revealed from an rdc-derived ubiquitin ensemble in solution. *Science*, **320**, 1471–1475. 57, 75, 94, 95

Le, H. & Oldfield, E. (1996). Ab initio studies of amide ^{15}n chemical shifts in dipeptides: applications to protein nmr spectroscopy. *J. Phys. Chem.*, **100**, 16423–16428. 16

Lee, C., Yang, W. & Parr, R.G. (1988). Development of the colle-salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B*, **37**, 785–789. 15, 111, 118

Legon, A.C. & Millen, D.J. (1987). Angular geometries and other properties of hydrogen-bonded dimers: a simple electrostatic interpretation of the success of the electron-pair model. *Chem. Soc. Rev.*, **16**, 467–498. 6

Lehtivarjo, J., Hassinen, T., Korhonen, S.P., Perakyälä, M. & Laatikainen, R. (2009). 4d prediction of protein ^{1}h chemical shifts. *J. Biomol. NMR*, **45**, 413–426. 1, 33, 34, 49, 50, 64, 80

LeMaster, D.M., Anderson, J.S. & Hernández, G. (2007). Spatial distribution of dielectric shielding in the interior of pyrococcus furiosus rubredoxin as sampled in the subnanosecond timeframe by hydrogen exchange. *Biophys. Chem.*, **129**, 43–48. 27

Levitt, M.H. (2006). *Spin dynamics: basics of nuclear magnetic resonance*. Willey, Chichester. 4

Lindorff-Larsen, K., Piana, S., Palmo, K., Maragakis, P., Klepeis, J.L., Dror, R.O. & Shaw, D.E. (2010). Improved side-chain torsion potentials for the amber ff99sb protein force field. *Proteins*, **78**, 1950–1958. 50

London, R.E., Wingad, B.D. & Mueller, G.A. (2008). Dependence of amino acid side chain ^{13}c shifts on dihedral angle: application to conformational analysis. *J. Am. Chem. Soc.*, **130**, 11097–11105. 47, 61

Lonsdale, K. (1937). Magnetic anisotropy and electronic structure of aromatic molecules. *Proc. R. Soc. A*, **159**, 149–161. 105, 106

Lundström, P., Hansen, D.F., Vallurupalli, P. & Kay, L.E. (2009a). Accurate measurement of alpha proton chemical shifts of excited protein states by relaxation dispersion nmr spectroscopy. *J. Am. Chem. Soc.*, **131**, 1915–1926. 88

Lundström, P., Vallurupalli, P., Hansen, D.F. & Kay, L.E. (2009b). Isotope labeling methods for studies of excited protein states by relaxation dispersion nmr spectroscopy. *Nature Prot.*, **4**, 1641–1648. 64

Luo, X., Sanford, D.G., Bullock, P.A. & Bachovchin, W.W. (1996). Solution structure of the origin dna-binding domain of sv40 t-antigen. *Nat. Struct. Biol.*, **3**, 1034–1039. 83

M. J. Frisch et al (2004). *Gaussian 03, revision E.01*. 16, 105, 120

Masunov, A., Dannenberg, J.J. & Contreras, R.H. (2001). C-h bond-shortening upon hydrogen bond formation: influence of an electric field. *J. Phys. Chem. A*, **105**, 4737–4740. 6

McConnell, H.M. (1957). Theory of nuclear magnetic shielding in molecules. 1. long-range dipolar shielding of protons. *J. Chem. Phys.*, **27**, 226–229. 34, 41

Mehta, M.A., Fry, E.A., Eddy, M.T., Dedeo, M.T., Anagnost, A.E. & Long, J.R. (2004). Structure of the alanine dipeptide in condensed phases determined by ^{13}c nmr. *J. Phys. Chem. B*, **108**, 2777–2780. 15

Meiler, J. (2003). Proshift: protein chemical shift prediction using artificial neural networks. *J. Biomol. NMR*, **26**, 25–37. 1, 9, 33, 34, 49, 64, 80

Meinke, G., Bullock, P.A. & Bohm, A. (2006). Crystal structure of the simian virus 40 large t-antigen origin-binding domain. *J. Virol.*, **80**, 4304–4312. 83

Miehlich, B., Savin, A., Stoll, H. & Preuss, H. (1989). Results obtained with the correlation energy density functionals of becke and lee, yang and parr. *Chem. Phys. Lett.*, **157**, 200–206. 15, 111, 118

Montalvao, R., Cavalli, A., Salvatella, X., Blundell, T.L. & Vendruscolo, M. (2008). Structure determination of protein-protein complexes using nmr chemical shifts: the case of an endonuclease colicin-immunity protein complex. *J. Am. Chem. Soc.*, **130**, 15990–15996. 33, 64, 89

Moyna, G., Zauhar, R.J., Williams, H.J., Nachman, R.J. & Scott, A.I. (1998). Comparison of ring current methods for use in molecular modelling refinement of nmr derived three-dimensional structures. *J. Chem. Inf. Comput. Sci.*, **38**, 702–709. 110, 113

Mulder, F.A.A. (2009). Leucine side-chain conformation and dynamics in proteins from ^{13}c nmr chemical shifts. *ChemBioChem*, **10**, 1477–1479. xiv, 33, 47, 58, 61

Murray, L.J.W., Arendall, W.B., Richardson, D.C. & Richardson, J.S. (2003). Rna backbone is rotameric. *Proc. Natl. Acad. Sci. USA*, **100**, 13904–13909. 104, 114

Nabuurs, S.B., Spronk, C.A.E.M., Vuister, G.W. & Vriend, G. (2006). Traditional biomolecular structure determination by nmr spectroscopy allows for major errors. *PLoS Comput. Biol.*, **2**, e9. 89

Neal, S., Nip, A.M., Zhang, H. & Wishart, D.S. (2003). Rapid and accurate calculation of protein ^{1}h, ^{13}c and ^{15}n chemical shifts. *J. Biomol. NMR*, **26**, 215–240. 1, 10, 12, 33, 34, 49, 91, 109, 119

Newton, R. & Wernisch, L. (2007). Rwui: a web application to create user friendly web interfaces for r scripts. *R News*, **7**, 32–35. 47, 71, 105

Nielsen, J.T., Bjerring, M., Jeppesen, M.D., Pedersen, R.O., Pedersen, J.M., Hein, K.L., Vosegaard, T., Skrydstrup, T., Otzen, D.E. & Nielsen, N.C. (2009). Unique identification of supramolecular structures in amyloid fibrils by solid-state nmr spectroscopy. *Angew. Chem. Int. Ed.*, **48**, 2118–2121. 91

Oldfield, E. (1995). Chemical shifts and 3-dimensional protein structures. *J. Biomol. NMR*, **5**, 217–225. 7, 33, 91

Oldfield, E. (2002). Chemical shifts in amino acids, peptides, and proteins: from quantum chemistry to drug design. *Annu. Rev. Phys. Chem.*, **53**, 349–378. 6, 19, 28, 103

Onsager, L. (1936). Electric moments of molecules in liquids. *J. Am. Chem. Soc.*, **58**, 1486–1493. 23

Ösapay, K. & Case, D.A. (1991). A new analysis of proton chemical shifts in proteins. *J. Am. Chem. Soc.*, **113**, 9436–9444. 41, 65, 67, 104

Otten, R., Chu, B., Krewulak, K.D., Vogel, H.J. & Mulder, F.A. (2010). Comprehensive and cost-effective nmr spectroscopy of methyl groups in large proteins. *J. Am. Chem. Soc.*, **132**, 2952–2960. 34

Pauling, L. (1936). The diamagnetic anisotropy of aromatic molecules. *J. Chem. Phys.*, **4**, 673–677. 105, 106

Pearson, J.G., Oldfield, E., Lee, F.S. & Warshel, A. (1993). Chemical shifts in proteins: a shielding trajectory analysis of the fluorine nuclear magnetic resonance spectrum of the eshcherichia coli galactose binding protein using a multipole shielding polarizability-local reaction field-molecular dynamics approach. *J. Am. Chem. Soc.*, **115**, 6851–6862. 6

Pearson, J.G., Le, H., Sanders, L.K., Godbout, N., Havlin, R.H. & Oldfield, E. (1997). Predicted chemical shifts in proteins: structure refinement of valine residues by using ab initio and empirical geometry optimizations. *J. Am. Chem. Soc.*, **119**, 11941–11950. 47

Pecul, M. & Sadlej, J. (1998). Solvent effects on nmr spectrum of acetylene calculated by ab initio methods. *Chem. Phys.*, **234**, 111–119. 19

Pethig, R. (1979). *Dielectric and Electric Properties of Biological Materials*. Wiley, New York. 26

Pitera, J.W., Falta, M. & van Gunsteren, W.F. (2001). Dielectric properties of proteins from simulation: the effects of solvent, ligands, ph, and temperature. *Biophys. J.*, **80**, 2546–2555. 26, 27

Pople, J.A. (1956). Proton magnetic resonance of hydrocarbons. *J. Chem. Phys.*, **24**, 1111–1111. 104, 105, 106, 110

R Development Core Team (2011). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. 44, 69, 105

Raha, K. & Merz, K. (2007). *Structural Basis of Dielectric Permittivity of Proteins: Insights from Quantum Mechanics; Proceedings of the International School of Physics "Enrico Fermi"*. IOS Press, Amsterdam. 26, 27

Raman, S., Lange, O.F., Rossi, P., Tyka, M., Wang, X., Aramini, J., Liu, G., Ramelot, T.A., Eletsky, A., Szyperski, T., Kennedy,

M.A., Prestegard, J., Montelione, G.T. & Baker, D. (2010). Nmr structure determination for larger proteins using backbone-only data. *Science*, **327**, 1014–1018. 33

Redfield, C., Poulsen, F.M. & Dobson, C.M. (1982). Nmr structure determination for larger proteins using backbone-only data. *Eur. J. Biochem.*, **128**, 527–531. 64

Reif, B., Xue, Y., Agarwal, V., Pavlova, M.S., Hologne, M., Diehl, A., Ryabov, Y.E. & Skrynnikov, N.R. (2006). Protein side-chain dynamics observed by solution- and solid-state nmr: comparative analysis of methyl ^{2}h relaxation data. *J. Am. Chem. Soc.*, **128**, 12354–12355. 49

Richter, B., Gsponer, J., Varnail, P., Salvatella, X. & Vendruscolo, M. (2007). The mumo (minimal under-restraining minimal over-restraining) method for the determination of native state ensembles of proteins. *J. Biomol. NMR*, **37**, 117–135. 56, 75

Rieping, W. & Vranken, W.F. (2010). Validation of archived chemical shifts through atomic coordinates (vasco). *Proteins*, **78**, 2482–2489. 36, 65

Robustelli, P., Cavalli, A. & Vendruscolo, M. (2008). Determination of protein structures from solid-state nmr chemical shifts. *Structure*, **16**, 1764–1769. 33

Robustelli, P., Cavalli, A., Dobson, C.M., Vendruscolo, M. & Salvatella, X. (2009). Folding of small proteins with chemical shift restrained monte carlo simulations without the use of molecular fragment replacement or structural homology. *J. Phys. Chem. B*, **113**, 7890–7896. xi, 12, 13, 14

Robustelli, P., Kohlhoff, K., Cavalli, A. & Vendruscolo, M. (2010). Using nmr chemical shifts as structural restraints in molecular dynamics simulations of proteins. *Structure*, **18**, 923–933. 10, 12, 14, 89, 103

Ruschak, A. & Kay, L.E. (2010). Methyl groups as probes of supra-molecular structure, dynamics and function. *J. Biomol. NMR*, **46**, 75–87. 37, 88

Sahakyan, A.B., Shahkhatuni, A.A., Shahkhatuni, A.G. & Panosyan, H.A. (2008a). Dielectric permittivity and temperature effects on spin-spin couplings studied on acetonitrile. *Magn. Reson. Chem.*, **46**, 63–68. 16, 21, 25

Sahakyan, A.B., Shahkhatuni, A.G., Shahkhatuni, A.A. & Panosyan, H.A. (2008b). Electric field effects on one-bond indirect spin-spin coupling constants and biomolecular perspectives. *J. Phys. Chem. A*, **112**, 3576–3586. 25

Sahakyan, A.B., Vranken, W.F., Cavalli, A. & Vendruscolo, M. (2011a). Structure-based prediction of methyl chemical shifts in proteins. *J. Biomol. NMR*, **50**, 331–346. 14, 21, 89, 91, 94, 103, 119

Sahakyan, A.B., Vranken, W.F., Cavalli, A. & Vendruscolo, M. (2011b). Using side-chain aromatic proton chemical shifts for a quantitative analysis of protein structures. *Angew. Chem. Int. Ed.*, **50**, 9620–9623. xvii, 14, 21, 89, 91, 92, 94, 95, 96, 103, 119

Scaife, B.K.P. (1989). *Principles of Dielectrics*. Oxford University Press, New York. 25

Schaeffer, R.D., Fersht, A. & Daggett, V. (2008). Combining experiment and simulation in protein folding: closing the gap for small model systems. *Curr. Opin. Struct. Biol.*, **18**, 4–9. 11

Schäfer, A., Huber, C. & Ahlrichs, R. (1994). Fully optimized contracted gaussian basis sets of triple zeta valence quality for atoms li to kr. *J. Chem. Phys.*, **100**, 5829–5835. 16

Schmidt, M. & Lipson, H. (2009). Distilling free-form natural laws from experimental data. *Science*, **324**, 81–85. 120

Schutz, C.N. & Warshel, A. (2001). What are the dielectric "constants" of proteins and how to validate electrostatic models. *Proteins: Struct. Funct. Genet.*, **44**, 400–417. 26

Senthilnathan, V.P. & Singh, S. (1974). Chemical shift studies in binary solvent mixtures. *Proc. Indian Nat. Sci. Acad.*, **40**, 199–208. 21

Sharp, K. (1998). Calculation of electron transfer reorganization energies using the finite difference poisson-boltzmann model. *Biophys. J.*, **74**, 1241–1250. 26

Shea, J.E. & Brooks, C. (2001). From folding theories to folding proteins: A review and assessment of simulation studies of protein folding and unfolding. *Annu. Rev. Phys. Chem.*, **52**, 499–535. 11

Shen, Y. & Bax, A. (2007). Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology. *J. Biomol. NMR*, **38**, 289–302. 1, 7, 9, 33, 34, 49

Shen, Y. & et al (2008). Consistent blind protein structure generation from nmr chemical shift data. *Proc. Natl. Acad. Sci. USA*, **105**, 4685–4690. 14, 33, 89, 103

Shen, Y., Vernon, R., Baker, D. & Bax, A. (2009). De novo protein structure generation from incomplete chemical shift assignments. *J. Biomol. NMR*, **43**, 63–78. 33

Sheppard, D., Guo, C. & Tugarinov, V. (2009). 4d ^{1}h - ^{13}c nmr spectroscopy for assignments of alanine methyls in large and complex protein structures. *J. Am. Chem. Soc.*, **131**, 1364–1365. 37, 97

Sheppard, D., Sprangers, R. & Tugarinov, V. (2010). Experimental approaches for nmr studies of side-chain dynamics in high-molecular-weight proteins. *Prog. NMR Spectrosc.*, **56**, 1–45. 33

Sousa, S.F., Fernandes, P.A. & Ramos, M.J. (2007). General performance of density functionals. *J. Phys. Chem. A*, **111**, 10439–10452. 15

Šponer, J. & Hobza, P. (1994). Nonplanar geometries of dna bases. ab initio second-order møller-plesset study. *J. Phys. Chem.*, **98**, 3161–3164. 118

Sprangers, R. & Kay, L. (2007). Quantitative dynamics and binding studies of the 20s proteasome by nmr. *Nature*, **445**, 618–622. 37

Stevens, E.D. & Coppens, P. (1980). Experimental electron density distributions of hydrogen bonds. high-resolution study of α-oxalic acid dihydrate at 100 k. *Acta Cryst.*, **B36**, 1864–1876. 6

Stewart, J.J.P. (2007). Optimisation of parameters for semiempirical methods v: modification of nddo approximations and application to 70 elements. *J. Mol. Modeling*, **13**, 1173–1213. 116

Stewart, J.J.P. (2008). *MOPAC2009*. 116

Sychrovsky, V., Foldynova-Trantirkova, S., Spackova, N., Robeyns, K., Meervelt, L.V., Blankenfeldt, W., Vokacova, Z., Sponer, J. & Trantirek, L. (2009). Revisiting the planarity of nucleic acid bases: pyramidalization at glycosidic nitrogen in purine bases is modulated by orientation of glycosidic torsion. *Nucl. Acids Res.*, **37**, 7321–7331. 118

Takayama, T., Ando, I. & Asakura, T. (1989). A study of dielectric solvent effect on silicon-20 nmr chemical shifts of some chlorosilanes. *Bull. Chem. Soc. Japan*, **62**, 1233–1236. 21

Tamagawa, K., Iijima, T. & Kimura, M. (1976). Molecular structure of benzene. *J. Mol. Struct.*, **30**, 243–253. 111

Tanaka, T., Ames, J.B., Harvey, T.S., Stryer, L. & Ikura, M. (1995). Sequestration of the membrane-targeting myristoyl group of recoverin in the calcium-free state. *Nature*, **376**, 444–447. 80, 93, 94

Tjandra, N. & Bax, A. (1997). Measurement of dipolar contributions to $^1j_{CH}$ splittings from magnetic-field dependence of j modulation in two-dimensional nmr spectra. *J. Magn. Reson.*, **124**, 512–515. 88

Tugarinov, V. & Kay, L.E. (2003). Ile, leu, and val methyl assignments of the 723-residue malate synthase g using a new labelling strategy and novel nmr methods. *J. Am. Chem. Soc.*, **125**, 13868–13878. 97

Tugarinov, V., Choy, W.Y., Orekhov, V.Y. & Kay, L.E. (2005a). Solution nmr-derived global fold of a monomeric 82-kda enzyme. *Proc. Natl. Acad. Sci. USA*, **102**, 622–627. 97

Tugarinov, V., Ollerenshaw, J.E. & Kay, L.E. (2005b). Probing side chain dynamics in high molecular weight proteins by deuterium nmr spin relaxation: an application to an 82-kda enzyme. *J. Am. Chem. Soc.*, **127**, 8214–8225. 33

Tugarinov, V., Kanelis, V. & Kay, L.E. (2006). Isotope labeling strategies for the study of high-molecular-weight proteins by solution nmr spectroscopy. *Nat. Protoc.*, **1**, 749–754. 1, 34, 88, 97, 100

Ulrich, E.L. (2007). Biomagresbank. *Nucl. Acids Res.*, **36**, D402–D408. 1, 9, 34, 36, 51, 65

van der Spoel, D., Lindahl, E., Hess, B., Kutzner, C., van Buuren, A.R., Apol, E., Meulenhoff, P.J., Tieleman, D.P., Sijbers, A.L.T., Feenstra, K.A., van Drunen, R. & Berendsen, H.J.C. (2006). Gromacs user manual, version 4.0. *The GROMACS development team*. 83

van Pelt, J.F.J.M., Bronduk, J.J., Claessen, V.W.M. & Biemond, J. (1981). Electric reaction field of a molecular quadrupole and solvent chemical shift. *J. Chem. Soc. Faraday Trans.*, **77**, 1789–1794. 23

Varma, S. & Jakobsson, E. (2004). Ionization states of residues in ompf and mutants: effects of dielectric constant and interactions between residue. *Biophys. J.*, **86**, 690–704. 25

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A. & et al (2001). The sequence of the human genome. *Science*, **291**, 1304–1351. 88

Vijay-Kumar, S., Bugg, C.E. & Cook, W.J. (1987). Structure of ubiquitin refined at 1.8 å resolution. *J. Mol. Biol.*, **194**, 531–544. 56, 77, 93, 94

Vila, J.A., Arnautova, Y.A., Martin, O.A. & Scheraga, H.A. (2009). Quantum-mechanics-derived $13c_\alpha$ chemical shift server (cheshift) for protein structure validation. *Proc. Natl. Acad. Sci. USA*, **106**, 16972–16977. 92

Vranken, W.F. & Rieping, W. (2009). Relationship between chemical shift value and accessible surface area for all amino acid atoms. *BMC Struc. Biol.*, **9**, 20. 30, 36

Vranken, W.F., Boucher, W., Stevens, T.J., Fogh, R.H., Pajon, A., Llinas, M., Ulrich, E.L., Markley, J.L., Ionides, J. & Laue, E.D. (2005). The ccpn data model for nmr spectroscopy: development of a software pipeline. *Proteins*, **59**, 687–696. 36

Wang, G. & Dunbrack, R.L. (2003). Pisces: a protein sequence culling server. *Bioinformatics*, **19**, 1589–1591. 38

Wang, Z.X. & Duan, Y. (2004). Solvation effects on alanine dipeptide: a mp2/cc-pvtz//mp2/6-31g** study of (ϕ, ψ) energy maps and conformers in the gas phase, ether, and water. *J. Comput. Chem.*, **25**, 1699–1716. 15, 18, 19, 25

Warshel, A. (2003). Computer simulations of enzyme catalysis: methods, progress, and insights. *Annu. Rev. Biophys. Biomol. Struct.*, **32**, 425–443. 25

Waugh, J.S. & Fessenden, R.W. (1957). Nuclear resonance spectra of hydrocarbons: the free electron model. *J. Am. Chem. Soc.*, **79**, 846–849. 104, 107, 108, 109, 110

Weiergräber, O.H., Senin, I.I., Philippov, P.P., Granzin, J. & Koch, K.W. (2003). Impact of n-terminal myristoylation on the ca^{2+}-dependent conformational transition in recoverin. *J. Biol. Chem.*, **278**, 22972–22979. 80, 93, 94

Weise, C.F. & Weisshaar, J.C. (2003). Conformational analysis of alanine dipeptide from dipolar couplings in a water-based liquid crystal. *J. Phys. Chem. B*, **107**, 3265–3277. 15

Wijmenga, S.S. & van Buuren, B.N.M. (1998). The use of nmr methods for conformational studies of nucleic acids. *Progr. Nucl. Magn. Res. Spectrosc.*, **32**, 287–387. 103

Wijmenga, S.S., Kruithof, M. & Hilbers, C.W. (1997). Analysis of ^{1}h chemical shifts in dna: assessment of the reliability of ^{1}h chemical shift calculations for use in structure refinement. *J. Biomol. NMR*, **10**, 337–350. 103

Wishart, D.S. (2011). Interpreting protein chemical shift data. *Prog. Nucl. Magn. Reson. Spectrosc.*, **58**, 62–87. 33, 49

Wishart, D.S. & Nip, A.M. (1998). Protein chemical shift analysis: a practical guide. *Biochem. Cell. Biol.*, **76**, 153–163. 7

Wishart, D.S., Watson, M.S., Boyko, R.F. & Sykes, B.D. (1997). Automated h-1 and c-13 chemical shift prediction using biomagresbank. *J. Biomol. NMR*, **10**, 329–336. 1, 7, 9, 33, 34

Wolinski, K., Hinton, J.F. & Pulay, P. (1990). Efficient implementation of the gauge-independent atomic orbital method for nmr chemical shift calculations. *J. Am. Chem. Soc.*, **112**, 8251–8260. 8, 16, 120

Wüthrich, K. (1986). *NMR of proteins and nucleic acids*. Wiley, New York. 11

Wüthrich, K. (2003). Nmr studies of structure and function of biological macromolecules (nobel lecture). *J. Biomol. NMR*, **27**, 13–39. 88

Xu, X.P. & Case, D.A. (2001). Automated prediction of ^{15}n, $^{13}c_{\alpha}$, $^{13}c_{\beta}$ and ^{13}c' chemical shifts in proteins using a density functional database. *J. Biomol. NMR*, **21**, 321–333. 1, 9, 33, 34, 49, 64, 80, 91

Xu, X.P. & Case, D.A. (2002). Probing multiple effects on ^{15}n, $^{13}c_{\alpha}$, $^{13}c_{\beta}$, and ^{13}c' chemical shifts in peptides using density functional theory. *Biopolymers*, **65**, 408–423. 16

Xu, Y., Liu, M., Simpson, P.J., Isaacson, R., Cota, E., Marchant, J., Yang, D., Zhang, X., Freemont, P. & Matthews, S. (2009). Automated assignment in selectively methyl-labeled proteins. *J. Am. Chem. Soc.*, **131**, 9480–9481. 37

Yatsenko, A.V. & Paseshnichenko, K.A. (1999). On the suitability of am1 for the modeling of molecules containing amino groups. *J. Mol. Struct. Theochem*, **492**, 277–283. 116

Zahedi, E., Aghaie, M., Zare, K. & Aghaie, H. (2009). Solvent effects on stability and ^{15}n nmr shielding of 5-methylcytosine tautomers: a theoretical approach. *J. Mol. Struct. Theochem*, **899**, 94–97. 19, 25

Zhu, J., Ye, E., Terskikh, V. & Wu, G. (2010). Solid-state ^{17}o nmr spectroscopy of large protein-ligand complexes. *Angew. Chem. Int. Ed.*, **49**, 8399–8402. 19, 126