DISSERTATIONS IN
HEALTH
SCIENCES

JUUSO LEHTIVARJO

Predicting NMR Parameters from the Molecular Structure

PUBLICATIONS OF THE UNIVERSITY OF EASTERN FINLAND
Dissertations in Health Sciences

UNIVERSITY OF
EASTERN FINLAND

Predicting NMR Parameters from the Molecular Structure

JUUSO LEHTIVARJO

Predicting NMR Parameters from the Molecular Structure

To be presented by permission of the Faculty of Health Sciences, University of Eastern Finland for public examination in the Auditorium CA100 in Canthia Building of The University of Eastern Finland, Kuopio, on Thursday, August 21st 2014, at 12 noon

Publications of the University of Eastern Finland
Dissertations in Health Sciences
Number 241

School of Pharmacy, Faculty of Health Sciences, University of Eastern Finland
Kuopio
2014

Juvenes Print
Tampere, 2014

Series Editors:
Professor Veli-Matti Kosma, M.D., Ph.D.
Institute of Clinical Medicine, Pathology
Faculty of Health Sciences

Professor Hannele Turunen, Ph.D.
Department of Nursing Science
Faculty of Health Sciences

Professor Olli Gröhn, Ph.D.
A.I. Virtanen Institute for Molecular Sciences
Faculty of Health Sciences

Professor Kai Kaarniranta, M.D., Ph.D.
Institute of Clinical Medicine, Ophthalmology
Faculty of Health Sciences

Lecturer Veli-Pekka Ranta, Ph.D. (pharmacy)
School of Pharmacy
Faculty of Health Sciences

Distributor:
University of Eastern Finland
Kuopio Campus Library
P.O.Box 1627
FI-70211 Kuopio, Finland
http://www.uef.fi/kirjasto

ISBN (print): 978-952-61-1517-7
ISBN (pdf): 978-952-61-1518-4
ISSN (print): 1798-5706
ISSN (pdf): 1798-5714
ISSN-L: 1798-5706

Author's address: School of Pharmacy
University of Eastern Finland
KUOPIO
FINLAND

Supervisors: Professor Reino Laatikainen, Ph.D.
School of Pharmacy
University of Eastern Finland
KUOPIO
FINLAND

Docent Mikael Peräkylä, Ph.D.
School of Pharmacy
University of Eastern Finland
KUOPIO
FINLAND

Reviewers: Professor Mark Johnson, Ph.D.
Department of Biosciences
Åbo Akademi
TURKU
FINLAND

Wouter Boomsma, Ph.D.
Department of Biology
University of Copenhagen
COPENHAGEN
DENMARK

Opponent: Professor Aatto Laaksonen, Ph.D.
Department of Materials and Environmental Chemistry
Stockholm University
STOCKHOLM
SWEDEN

Lehtivarjo, Juuso
Predicting NMR Parameters from the Molecular Structure
University of Eastern Finland, Faculty of Health Sciences
Publications of the University of Eastern Finland. Dissertations in Health Sciences 241. 2014. 55 p.

ISBN (print): 978-952-61-1517-7
ISBN (pdf): 978-952-61-1518-4
ISSN (print): 1798-5706
ISSN (pdf): 1798-5714
ISSN-L: 1798-5706

ABSTRACT

The spectral parameters of the nuclear magnetic resonance (NMR) spectra are dependent on the chemical environment around the nuclei, making NMR spectroscopy a powerful method for studying molecular structure and dynamics at the atomic level. Conversely, the spectral parameters can be calculated if one knows the molecular structure. The spectral parameter prediction plays a key role in many applications of computational NMR. This thesis presents two NMR parameter prediction approaches for two different purposes.

Chemical shifts are the key parameters of the NMR spectrum. In the field of protein NMR, the use of chemical shifts in protein structural studies has been increasing in the last years, driven by improvements in the chemical shift prediction methods. In addition to the protein structure, chemical shifts are dependent on protein dynamics. In order to account for the dynamic effects, a four-dimensional approach for protein chemical shift prediction was developed. Here, the 4th dimension is time and it is mapped by molecular dynamics simulations. The conformational space was further expanded by starting the MD simulations from different conformations of the same protein. From the structural parameters averaged over the conformational space of the MD simulations, chemical shifts prediction models for all backbone and most side chain nuclei were built with principal component regression. In comparison with the non-dynamic models, the dynamic models achieved 13 % lower root-mean-square (RMS) errors for different backbone nuclei, underlining the importance of dynamics in reproducing experimental protein chemical shifts. An additional outcome of the project is the prediction program 4DSPOT, which is freely available for the protein NMR community (www.uef.fi/4dspot).

NMR spectra can be simulated and iteratively analyzed with quantum mechanical principles if the parameters are known. The scalar coupling constants are the parameters that give rise to the fine structure of the NMR signals. The second prediction method presented in this thesis targets small molecule couplings to be used in automatic spectrum analysis. Coverage and speed are emphasized in the design of the method. Thus, the method is based on a lightweight hash dictionary search, followed by a k Nearest Neighbors regression to resolve the substituent and conformational dependencies.

Despite the growth of databases, there are still many situations when experimental data is too sparse to permit prediction model building. However, recently the accuracy of quantum-mechanical calculation of NMR parameters has greatly improved. Therefore, the use of quantum chemistry as a source of teaching data is discussed and some preliminary results are shown.

National Library of Medicine Classification: QU 25
Medical Subject Headings: Nuclear Magnetic Resonance, Biomolecular; Spectrum Analysis; Molecular Conformation; Proteins; Molecular Dynamics Simulation
Library of Congress Subject Headings: Nuclear magnetic resonance spectroscopy; Proteins - - Structure

Lehtivarjo, Juuso
NMR-parametrien mallintaminen molekyylirakenteesta
Itä-Suomen yliopisto, terveystieteiden tiedekunta
Publications of the University of Eastern Finland. Dissertations in Health Sciences 241. 2014. 55 s.

ISBN (print): 978-952-61-1517-7
ISBN (pdf): 978-952-61-1518-4
ISSN (print): 1798-5706
ISSN (pdf): 1798-5714
ISSN-L: 1798-5706

TIIVISTELMÄ

Ydinmagneettisen resonanssispektroskopian (NMR) spektriparametrit ovat riippuvaisia atomiydinten kemiallisesta ympäristöstä, minkä vuoksi NMR on tehokas menetelmä molekyylin rakenteen ja liikkeiden tutkimiseen atomitasolla. Spektriparametrit voidaan myös mallintaa laskennallisesti molekyylirakenteen perusteella. Tätä mahdollisuutta käytetään hyväksi monessa laskennallisen NMR:n sovelluksessa. Tässä väitöskirjassa on esitelty kaksi spektriparametrien mallinnusmenetelmää kahta eri tarkoitusta varten.

Kemialliset siirtymät ovat NMR-spektrin tärkeimpiä parametreja. Niiden käyttö proteiini-NMR:ssä on lisääntynyt viime vuosina kemiallisen siirtymän mallinnusmenetelmien parantumisen myötä. Proteiinin rakenteen lisäksi kemialliset siirtymät ovat riippuvaisia proteiinin liikkeistä. Tämän vaikutuksen huomioon ottamiseksi väitöskirjatyössä kehitettiin neliulotteinen menetelmä proteiinien kemiallisen siirtymän mallinnukseen. Neljäs ulottuvuus tässä yhteydessä on aika, ts. molekyylin liikkeet kartoitetaan molekyylidynamiikkasimulaatioiden avulla. Simulaatioiden kattamaa liikeavaruutta laajennettiin aloittamalla simulaatiot saman proteiinin eri konformaatioista. Liikkeiden suhteen keskiarvoistetuista rakenneparametreista tehtiin pääkomponenttiregressiomallit proteiinin pääketjun ja useimpien sivuketjujen ytimien kemiallisille siirtymille. Verrattaessa liikkumattomista proteiineista tehtyyn malliin, dynaamisen mallin RMS-virheet (keskineliövirheen neliöjuuret) olivat noin 13 % pienempiä, mikä kuvastaa liikkeiden tärkeyttä kemiallisen siirtymän kuvaamisessa. Projektin yhteydessä syntyi myös vapaasti käytettävissä oleva kemiallisen siirtymän mallinnusohjelma 4DSPOT (www.uef.fi/4dspot).

NMR-spektri voidaan simuloida ja analysoida iteratiivisesti kvanttimekaanisia periaatteita käyttäen kunhan spektriparametrit tiedetään. Kytkentävakiot ovat parametreja, jotka aiheuttavat NMR-signaalien hienorakenteen. Väitöskirjan toisessa osassa mallinnettiin pienmolekyylien kytkentävakioita automaattista spektrianalyysiä varten. Menetelmä on kehitetty painottaen kattavuutta ja nopeutta, minkä vuoksi se perustuu hajautustaulupohjaiseen tietokantahakuun. Substituenttien ja rakenteen vaikutus kytkentävakioihin mallinnetaan tämän jälkeen k:n lähimmän naapurin menetelmällä.

Tietokantoihin tallennetun kokeellisen NMR-parametritiedon lisääntymisestä huolimatta osalle molekyylirakenteista sitä on saatavilla liian vähän mallien rakentamista varten. Kvanttimekaanisten spektriparametrilaskujen tarkkuus on sen sijaan parantunut viime aikoina huomattavasti. Tästä johtuen väitöskirjassa on alustavasti tutkittu myös kvanttikemian käyttöä mallinnusmenetelmien opetustiedon lähteenä.

Luokitus: 543.429.23; 543.384
Yleinen Suomalainen asiasanasto: NMR-spektroskopia; spektrianalyysi; rakenneanalyysi; proteiinit; mallintaminen; dynamiikka

Acknowledgements

The present study was initiated in the University of Kuopio, Department of Biosciences in 2008 and finished in the University of Eastern Finland, School of Pharmacy in 2014. During these years, some parts of the work have also been conducted in the University of Cambridge, Department of Chemistry and in PERCH Solutions Ltd. I want to acknowledge the departments and the company for providing me with the research facilities. The financial support from The National Doctoral Programme of Informational and Structural Biology and The Emil Aaltonen Foundation is greatly appreciated.

My deepest gratitude is expressed to my principal supervisor Professor Reino Laatikainen, who took me in as an intern into his group in 2005 and later proposed that I tackle this PhD thesis project. I have much enjoyed his wealth of ideas and our long discussions about science (and fishing). I'm also much obliged to my second supervisor Docent Mikael Peräkylä for his excellent guidance and support throughout the project.

I express my gratitude to Professor Mark Johnson and Dr. Wouter Boomsma for reviewing this thesis, as well as Dr. Ewen MacDonald for correcting the language.

I am deeply grateful to all the people who have participated in this thesis project in one way or another. I thank Tommi Hassinen who has originally written much of the program code used in my projects and thus saved me a vast amount of time; and Kari Tuppurainen for his help with mathematics. Continuing the list, I am thankful to Dipl. Chem. Matthias Niemitz for providing me with the opportunity to work and undertake research in PERCH Solutions Ltd. In the company, Dr. Samuli-Petrus Korhonen has been a fantastic supervisor, with answers always available on anything I would ever dare to ask. I also thank Professor Michele Vendruscolo for providing me with a research visit position in his group, and Dr. Aleksandr Sahakyan for guiding me there into the world of quantum chemistry. The winter of 2012-2013 spent in Cambridge was one of the best experiences of my whole life.

I am proud I have been given the possibility to begin my academic journey in the Department of Chemistry of University of Kuopio. Regardless of the location of the coffee room, or whatever the name of the department, the faculty or even the university has been over the years, the extraordinary spirit of the old “Kemian laitos” has remained the same. I thank all the wonderful people I have met and worked with during my years in the department. Janne, Jussi and Pekka, you have made coming to work feel more like coming to a kind of academic kindergarten, which is one of the reasons I have enjoyed staying there for so many years.

I am eternally thankful to my parents Tarja and Seppo who have always trusted in me in whatever I have decided to do and giving me such a happy childhood.

In the foreword of my M.Sc. thesis I thanked Kaisa for marrying me. Now, six years later, I thank her for giving birth to our two lovely children Martti and Taimi. Since you three have become my family, my life has been so happy and content.

List of the original publications

This dissertation is based on the following original publications:

I Lehtivarjo J, Hassinen T, Korhonen S-P, Peräkylä M and Laatikainen R. 4D prediction of protein ^{1}H chemical shifts. *Journal of Biomolecular NMR 45: 413-426, 2009.*

II Lehtivarjo J, Tuppurainen K, Hassinen T, Laatikainen R and Peräkylä M. Combining NMR ensembles and molecular dynamics simulations provides more realistic models of protein structures in solution and leads to better chemical shift prediction. *Journal of Biomolecular NMR 52: 257-267, 2012.*

III Lehtivarjo J, Niemitz M and Korhonen S-P. Universal J-coupling prediction. *Journal of Chemical Information and Modeling 54: 810-817, 2014.*

The publications were adapted with the permission of the copyright owners. Some unpublished results are also presented.

Contents

Abbreviations

Ac	acetyl
ACA	Automated Consistency Analysis
ASNN	associative neural network
BMRB	Biological Magnetic Resonance Bank
CASE	computer-assisted structure elucidation
COSY	correlation spectroscopy
CPMG	Carr-Pursell-Meiboom-Gill
DFT	density functional theory
DSO	diamagnetic spin-orbital
DSS	2,2-dimethyl-2-silapentane-5-sulfonic acid
EXCY	exchange spectroscopy
FC	Fermi contact
FID	free induction decay
FT	Fourier transform
H/D	hydrogen/deuterium
HMBC	heteronuclear multiple bond correlation
HSQC	heteronuclear single quantum coherence
IDP	intrinsically disordered protein
kNN	k nearest neighbors
MD	molecular dynamics
MMS	Molecular Modelling System
NMe	N-methyl
NMR	nuclear magnetic resonance
NMRE	NMR ensemble
NOE	nuclear Overhauser effect
pAA	phosphorylated amino acid
PC	principal component
PCR	principal component regression
PDB	Protein Data Bank
pSer	phosphoserine
PSO	paramagnetic spin-orbital
QM	quantum mechanics
RDC	residual dipolar coupling
REMD	replica-exchange molecular dynamics
RMS	root mean square
SD	spin-dipole
TMS	tetramethylsilane
TROSY	transverse-relaxation optimized spectroscopy

1 Introduction

Nuclear magnetic resonance (NMR) spectroscopy is one of the few methods which can provide information about molecules at the atomic level. The nuclear magnetic resonances of atoms are affected by the chemical surroundings of the atoms, thus containing information about the molecular structure. In comparison with X-ray crystallography, NMR may not be the most straightforward tool for determination of a molecular structure. Instead, its strengths lie in different areas, for example the possibility to study molecular motions. Overall, NMR spectroscopy may be the most versatile tool for studying molecular structures of different sizes.

This thesis belongs to the field of computational NMR, and its overall aim was to develop methods for predicting NMR spectral parameters from the molecular structure. The prediction i.e. back-calculation plays a key role in many applications of computational NMR. Moreover, the relationship between molecular structure and NMR parameters still remains unresolved in many respects, and NMR parameter prediction is one very useful way study these correlations.

Protein NMR is a subtopic of NMR spectroscopy, originating from the possibility to convert NMR parameters into interatomic distances. Although the determination of protein structure with NMR has been mostly based on the nuclear Overhauser effect (NOE) signals, scalar coupling constants and residual dipolar couplings (RDC), the use of chemical shifts in protein structural studies has increased rapidly in the last 10 years. In addition to the protein structure, NMR parameters are also dependent upon molecular motions. The major part of this thesis deals with protein chemical shift prediction. The outcome is the chemical shift predictor 4DSPOT, which was the first predictor to take account of the molecular dynamics (MD).

NMR spectroscopy is the only type of spectroscopy in which the whole spectrum can be quantum mechanically simulated. For example, this enables NMR spectra to be analyzed in a highly automatic manner. However, the simulation of NMR spectrum requires that the spectral parameters can be predicted from the structure. The prediction of NMR coupling constants of small molecules forms another part of this thesis.

Before moving on to reviewing the literature concerned with prediction studies, this chapter will provide a brief introduction to NMR spectroscopy and the two particular parameters, chemical shifts and coupling constants.

1.1 THE NMR PHENOMENON AND SPECTRAL PARAMETERS

The physics behind NMR signals is based on the observation that atomic nuclei have a quantum mechanical property called *Nuclear spin angular momentum,* or simply *spin*. Certain atomic nuclei (those with an odd mass number, or an even mass number but an odd atomic number) have a non-zero spin leading to *nuclear magnetic moments* μ, parallel to the spin angular momentum. When placed under the external magnetic field $\boldsymbol{B_0}$ (usually presented along z-axis) of the spectrometer magnet, the NMR active nuclei start to precess with the magnetization vectors μ oriented either parallel or antiparallel to $\boldsymbol{B_0}$. The precession happens with a *Larmor frequency* ν, related to the field strength as in Eq. [1.1].

$$2\pi\nu = \gamma\mathbf{B_0} \qquad [1.1]$$

Here, γ is the magnetogyric ratio, a natural constant for the nucleus in question, meaning that also the Larmor frequencies differ for different nuclear isotopes. The parallel or antiparallel vectors determine the *spin states* of the given nucleus. The spin states lie at different energy levels and thus the sum of the vectors (net magnetization vector ***M***, Fig. 1a) is non-zero and parallel to the magnetic field B_0. This is the observable magnetization detected in NMR spectroscopy. Since it is relative to the difference between spin state populations (only about $1/10^5$), ***M*** is small in magnitude compared with B_0. In the modern Fourier Transform (FT) -NMR devices the vector ***M*** is altered by exciting the system with an external radiofrequency pulse. The return of ***M*** to its ground state, i.e. emission, is then followed.

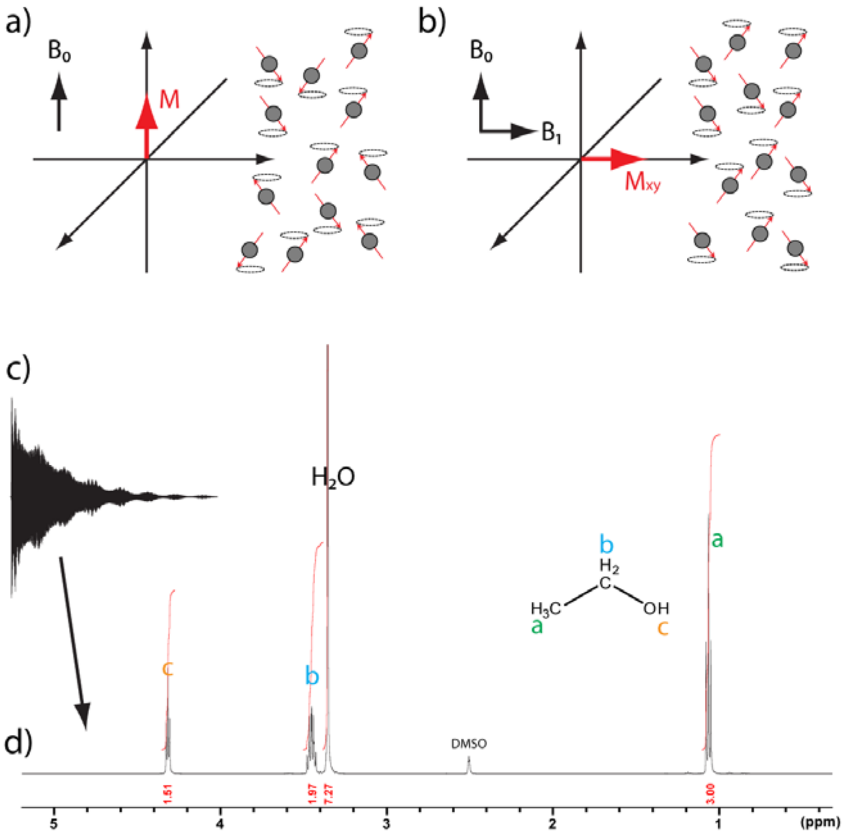


Figure 1. a) Ground state of sample with spin ½ nuclei in NMR magnet under a magnetic field $\mathbf{B_0}$. b) Excited state with radiofrequency pulse $\mathbf{B_1}$ applied c) Free induction decay measured along x-axis d) NMR spectrum after Fourier transformation of FID.

In the most basic NMR experiment, the signal is created by applying an external radiofrequency pulse B_1 along the x-axis. A pulse of suitable strength and length pulse flips the vector ***M*** from the z-axis to the y-axis by forcing all the nuclear spins to the same phase (Fig. 1b), creating so-called transverse magnetization M_{xy}. In addition, the magnetization parallel to the z-axis is momentarily cancelled as the pulse equalizes the populations of the spin states. After the pulse, the vector M_{xy} precesses around the z-axis as the individual spins precess at their Larmor frequencies. With the passage of time, the transverse magnetization decays as the spins will dephase due to spin-spin-interactions. This is called the transverse relaxation (T_2). Simultaneously, the original magnetization in the z-axis direction recovers as the alignment of individual spins returns to the ground state population with a mechanism called longitudinal relaxation (T_1). For the spin ½ nuclei, typical relaxation times vary between 0.1 to 10 s, with $T_2 < T_1$. The transverse magnetization recorded with a receiver coil in the xy-plane shows as a curve called *free induction decay* (FID, Fig. 1c), showing frequencies from 0 to about 5000 Hz. Subsequently, the FID is Fourier transformed to frequency signals, shown as peaks in the NMR spectrum (Fig. 1d).

The example spectrum in (Fig. 1d) is from basic ^{1}H NMR experiment of ethanol and water solvated in dimethylsulfoxide. It already contains four measurable NMR parameters: 1) the *chemical shifts* arising from the different chemical environments of the nuclei; 2) the signal areas providing information about the amount of similar nuclei in the molecule; 3) the *scalar (J-) coupling constants* arising from the interactions of nuclei with each other; and 4) the line widths affected by the relaxation times. Additionally, one may use more special NMR experiments to obtain more NMR parameters such as NOE signals arising from through-space interaction of nuclei; residual dipolar couplings, intentionally not averaged out in aligning medium, providing information about the distance between nuclei and their orientation against the external magnetic field; and the relaxation times T_1 and T_2 and other relaxational parameters derived from those, such as the S^2 order parameters. This thesis, however, deals only with chemical shifts and scalar coupling constants. The next two chapters will provide an introduction to these parameters and the underlying NMR phenomena.

1.1.1 Chemical shifts

The *chemical shift* is the most fundamental parameter one can extract from the nuclear magnetic resonance (NMR) spectra. The chemical shifts are the locations of the NMR signals in the spectrum: they show where the signals are *shifted*, due to their *chemical* environment, compared with the signal of some reference compound.

The physics behind the chemical shift arises from the electrons. The external magnetic field $\boldsymbol{B_0}$ causes the movement of electrons around the nucleus. Consequently, the moving current of these electrons produces a local magnetic field, which can be either opposite (diamagnetic, from s orbital currents) or parallel (paramagnetic, from p orbital currents) to $\boldsymbol{B_0}$. This effect is called *shielding* (or *deshielding* in the paramagnetic case), and it will change the local magnetic field experienced by the nucleus ($\boldsymbol{B_L}$) as follows:

$$\mathbf{B_L} = \mathbf{B_0}\,(1 - \sigma) \qquad [1.2]$$

The strength of shielding is *anisotropic*, i.e. it varies at different directions relative to $\boldsymbol{B_0}$. Therefore, the shielding constant σ is actually a tensor. However, in *isotropic* conditions, e.g. in the liquid phase, the anisotropy of the shielding constant averages out and results in a scalar shielding constant. The shielding constant effectively consists of three factors:

$$\sigma = \sigma_{\text{local dia}} + \sigma_{\text{local para}} + \sigma' \qquad [1.3]$$

In Eq. 1.3, the terms $\sigma_{\text{local dia}}$ and $\sigma_{\text{local para}}$ are the effects arising from the electron cloud of the nucleus itself, and σ' is the effect of electrons in neighboring atoms and groups of the molecule, which also affect the magnetic field experienced by the nucleus. Moreover, it is also illustrative to break up the σ' term into more detailed contributions of different physical origins. For example, for protons in general one may write Eq. 1.4 [1]:

$$\Delta\sigma = \Delta\sigma_{\text{local dia}} + \Delta\sigma_{\text{magn}} + \Delta\sigma_{\text{el}} + \Delta\sigma_{\text{vdw}} + \Delta\sigma_{\text{solvent}} \qquad [1.4]$$

Here, the local paramagnetic effect is not taken into account since the paramagnetic effect is negligible for protons due to the missing p orbitals. The local diamagnetic effect is altered by the electronegativity of the neighboring atoms, causing changes in the electron density of the proton. For example, an inductive group such as the nitro group $-NO_2$ or a positively charged carbon will withdraw electrons from the proton and cause deshielding. In the opposite example, negatively charged carbons will donate electrons to the proton, and shielding will occur.

The neighboring group term σ' is divided into four new terms. The most important of these contributions is the magnetic term ($\Delta\sigma_{magn}$) arising from the electron movement around the neighboring atoms and bonds, both covalently bound and spatial. With respect to protons, these effects are generally larger than the other neighboring group terms of Eq. 1.4. The effect arises from the magnetic anisotropy ($\Delta\chi$) of the neighboring groups, and it can be either positive or negative, depending on the sign of $\Delta\chi$ and the molecular geometry by the McConnell equation [2],

$$\Delta\sigma = \frac{1}{3}\Delta\chi(1 - 3\cos^2\theta)/4\pi R^3 \qquad [1.5]$$

in which θ is the angle between the probe nucleus and the neighboring group, giving rise to the cone-shaped functions (Fig. 2). For example, a proton lying in the C=C plane, e.g. an ethylene proton, is deshielded, whereas those protons that are perpendicular to the plane are shielded. Generally, the magnetic anisotropy is larger in π-systems than in single bonds. An especially strong effect is seen in aromatic rings, where delocalized π-electrons of the conjugated bonds circulate around the ring and cause a local magnetic field (Fig. 2). The protons above or below the ring experience strong shielding, whereas those protons located in the ring plane are deshielded. Widely used equations for the ring current shielding include the Pople model, expressing the aromatic ring as a point-dipole in the midpoint of the ring [3,4], and the semi-empirical Haigh-Mallion model [5].

The remaining terms have minor effects on the proton chemical shifts. The electric field term $\Delta\sigma_{el}$ arises when the magnetic dipoles of the adjacent polar groups will polarize the chemical bonds containing the probe nucleus, thus affecting the electron density and the experienced shielding. The most famous equation for the electric field term, presenting $\Delta\sigma_{el}$ of ^1H nuclei as a function of the electric field and the component of the electric field parallel to the H-X bond, was found by Buckingham [6]. Next, the nuclei may sometimes come so close to each other that their electron clouds will overlap and deform, giving rise to the van der Waals term $\Delta\sigma_{vdw}$. Last, $\Delta\sigma_{solvent}$ accounts for the effect of various interactions between the solute and solvent.

The origins of chemical shifts of the ^{13}C isotope are somewhat different than those of protons. Generally, the chemical shifts of ^{13}C nuclei and other heavy elements are mainly determined by the local paramagnetic term, which can arise due to the availability of low-lying p-orbitals. Therefore, the ^{13}C shifts can be described well with substituent effects, and the magnetic contributions of the neighboring atoms (σ') have only a minor influence.

The magnetization experienced by the nucleus ($\mathbf{B}_L$) could be used to derive the absolute resonance frequency ν of the nucleus (Eq. 1.1). However, this would require a value for $\mathbf{B}_0$, which is difficult to measure accurately. Therefore, absolute values of ν are not used to express the resonance frequency of the nuclei of interest. Instead, they are given relative to those of a reference compound ν_{ref}, resulting in the chemical shift δ (Eq. 1.6).

$$\delta = \frac{(\nu - \nu_{ref})}{\nu_0} \qquad [1.6]$$

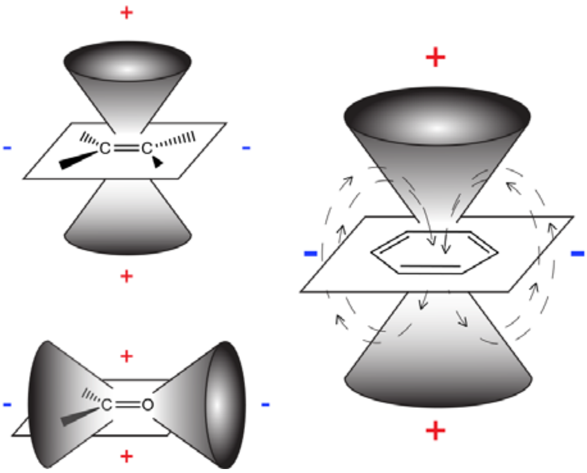


Figure 2. Examples of magnetic anisotropy of chemical bonds and aromatic ring currents. Plus signs denote shielding (chemical shift value δ is decreased) and minus signs denote deshielding (δ is increased).

Furthermore, as NMR spectrometers operate at different field strengths and frequencies, chemical shifts are measured relative to the operating frequency of the spectrometer ν_0. As a result of Eq. 1.5, the chemical shift is a unitless quantity. However, as the values ν and ν_{ref} are of the order of Hz and ν_0 of the order of MHz, the values of chemical shift are expressed as parts per million (ppm).

With respect to small molecules, the most commonly used reference compounds are tetramethylsilane (TMS) for referencing the ^{1}H and ^{13}C shifts and liquid ammonia or nitromethane to reference the ^{15}N shifts. However, for proteins, the recommended compound to reference the ^{1}H chemical shifts is a water soluble and a pH-insensitive 2,2-dimethyl-2-silapentane-5-sulfonic acid (DSS) [7]. Moreover, it is recommended that reference frequencies for other nuclear isotopes are derived from the DSS ^{1}H signal by applying pre-determined frequency ratios derived from magnetogyric ratios, thus avoiding the need for multiple reference compounds [7].

1.1.2 Coupling constants

In NMR, coupling in general means interactions between nuclei that show up as perturbations in the signal appearance: otherwise, an NMR signal of a single nucleus would always appear as a single peak, which is not the case (see the ethanol signals in Fig. 1d). Two types of coupling are observed in NMR experiments, namely 1) scalar coupling (J-coupling), with a range of several hundred Hz, and 2) direct dipolar or quadrupolar coupling, with values up to tens of kHz. From these, the latter are dependent on the orientation of molecular frame versus $\boldsymbol{B_0}$, and thus mostly averaged out in isotropic liquid samples where the molecules can tumble around freely. This is usually beneficial, because dipolar and quadrupolar couplings may cause strong distortions of the NMR signals and complicates the interpretation of the spectra. Nonetheless, they have their uses, such as the possibility to derive distance-angle restraints from dipolar couplings, e.g. available from solid state samples or by measuring the samples in partially aligning media such as liquid crystalline bicelles. Dipolar and quadrupolar couplings arise from the direct magnetic interactions between the nuclei, whereas scalar coupling, also called indirect dipole-dipole coupling, is carried over electrons. Scalar couplings are visible in standard NMR spectra unless a decoupling experiment is performed. They contain information about molecular topology and structure (see chapter 2.2.1). A brief introduction about the origins of the parameter will follow.

Since the electrons also possess a spin and thus a magnetic moment, they are affected by the magnetic moments of the nuclei. In turn, the nuclei are affected by the magnetic moments of the electrons. In other words, one nucleus induces spin polarization (along with other mechanisms, see below) in the bonding electrons, which is then experienced by the neighboring nuclei. Let us consider a CH_2 system with protons resonating at different frequencies and denote those protons as A and B (Fig. 3). From the viewpoint of proton A, the bonding electrons are aligned based on the spin state of nucleus B, obeying the rules of Pauli (the electron spins in the same molecular orbital, here the H-C bond, must be antiparallel) and Hund (the electron spins of different molecular orbitals, here the carbon sp^3 orbitals, are parallel). The energetically favorable combination of the spins of proton A and its own electrons is when they are aligned in an antiparallel configuration, because then the magnetic moments to become parallel. Therefore, the transition from $\alpha\alpha$ to $\beta\alpha$ requires slightly more energy than that from $\alpha\beta$ to $\beta\beta$. As a consequence, the NMR signal of A is split into two signals. The same phenomenon happens also from the viewpoint of nucleus B, which is also split. The line splitting, i.e. the coupling constant J_{AB}, is equal in both signals. By definition, the sign of J_{AB} here is negative, since the low-energy conformation has a parallel alignment of nuclear magnetic moments. The opposite scenario, usually found in 1- and 3-bond couplings, would make the sign positive.

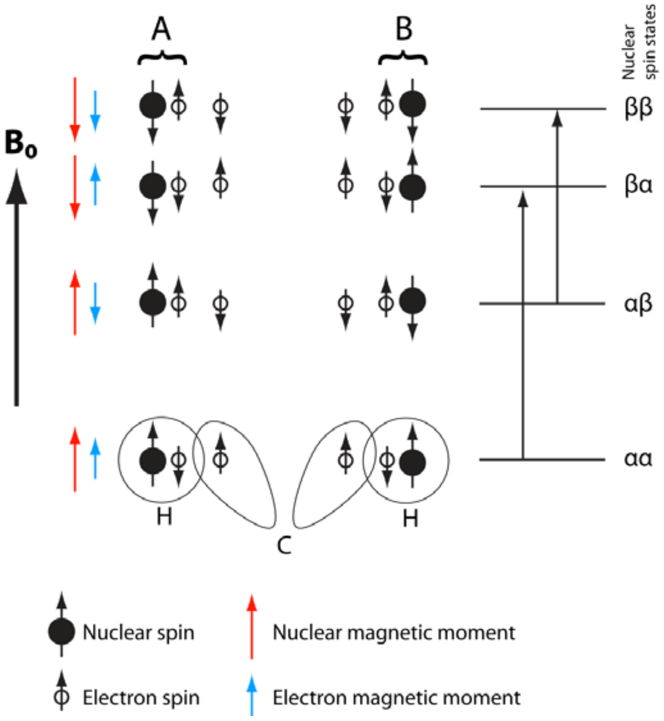


Figure 3. Schematic presentation of the origin of J-coupling between protons A and B in a CH_2 system.

The induction of nuclear magnetic moments to electrons can be described by four different physical mechanisms, known as the Ramsey terms [8]. First, two mechanisms describe how the nuclei induce the spin polarization of electrons. These are the 1) Fermi Contact (FC) mechanism, which arises from the contacts of electrons and nuclei directly at the nuclear site, and 2) the spin-dipole (SD) mechanism, which is the polarization happening elsewhere in the electron cloud, where the electrons see the nucleus as a magnetic dipole.

In addition to spin polarization, coupling is affected also by the orbital current mechanism, in which the magnetic field of a nucleus induces electronic currents in the orbitals. Like all electron currents, also these possess a magnetic moment, which is then felt by the neighboring nucleus. To continue the list of the Ramsey terms, these effects are 3) the diamagnetic spin-orbital (DSO) mechanism, in which new orbital currents are induced in the presence of $\boldsymbol{B_0}$, and 4) paramagnetic spin-orbital (PSO) mechanisms, in which the applied $\boldsymbol{B_0}$ modifies the existing orbital currents of non-σ orbitals in a way that the net current is not cancelled out [9].

The magnitude of a coupling constant is the sum of the above mechanisms, i.e. they are proportional to how much the neighboring nucleus induces spin polarization and orbital currents to the electron system of the bond [9,10]. However, the FC mechanism often dominates the values of scalar coupling constants [10]. For example, proton-proton J-couplings can be quantum mechanically calculated with excellent accuracy by accounting only for the FC mechanism [11].

Scalar coupling constants are denoted by the convention $^{n}J_{XY}$, where n is the number of bonds between the nuclei, and X and Y are the coupled nuclei. The value of a J-coupling interaction is independent of $\boldsymbol{B_0}$ and thus reported in Hz. However, it is dependent upon the magnetogyric ratios of the coupled nuclei and thus the magnitude of couplings between different elements varies from several Hz up to several hundred Hz. Since J-couplings are carried by bonding electrons, their magnitude is largest for ^{1}J and fades quickly according to the number of bonds between the coupled nuclei.

The collection of atoms coupled to each other is called a spin system. As more atoms participate in one spin system, more complex signal patterns arise. Basically, the signal of a spin ½ nuclei with n coupled neighbors is split into n+1 peaks. For example, the proton A and B signals of the abovementioned CH_2 system will be revealed as two peaks (a doublet) each, and some other proton with a methyl neighbor will show as four peaks (a quartet). A nucleus with two coupled neighbors with dissimilar J values also appears as four peaks, but now as a doublet of a doublet. However, the difference of chemical shifts of the coupled nuclei also plays a significant role in signal pattern appearances, and the above splitting rules only hold in weakly coupled (first-order) cases, where the coupling is much smaller than the chemical shift difference ($\nu A - \nu B >> J_{AB}$). When the opposite is true, the energy levels are closer to each other and start to mix, which leads to more allowed transitions and NMR signals. This is called the second-order effect, appearing as distortions in the multiplets. The analysis of second-order multiplets is often impossible without computational methods and quantum mechanical rules [12]. In first-order patterns, coupling signs do not have any effect on the spectral appearance, whereas in second order cases, they may have significant effects [1,13].

2 Review of the literature

2.1 PROTEIN CHEMICAL SHIFT PREDICTION

Protein NMR developed in 1980s when distance restraints based on the NOE signals were introduced, soon leading to the first protein structures being determined by NMR. The current state of the field is briefly reviewed in chapter 2.1.1. However, the first observation of protein structural effects on NMR parameters was that the chemical shifts of the unfolded protein are different than the native one [14]. Although this finding was not directly used in protein structure determination until recently, in one sense it enables whole protein NMR since in folded proteins, the nuclei of the same atom type effectively resonate at different frequencies. If this would not be the case, the NMR signals would heavily overlap; for example, the analysis of NOE signals would be impossible. Later, numerous studies have unravelled the correlations between chemical shifts and different structural factors, reviewed in chapter 2.1.2. The dynamic nature of proteins also has a significant contribution to chemical shifts, elaborated in chapter 2.1.3.

Today, the use of the chemical shift information is attracting interest in all aspects of protein studies. This has been achieved with the growth of databases, especially the Biological Magnetic Resonance Bank (BMRB) [15], which stores the NMR parameters of biomolecules. In order to exploit the vast amount of structural information hidden in chemical shifts (see previous chapters), one needs a reliable prediction (i.e. back-calculation) of chemical shifts based on protein tertiary structures, (i.e. atom coordinates as input). The atom coordinate based prediction methods can be roughly divided into 1) empirical methods, using experimental chemical shift data as the base of information, and 2) quantum mechanical (QM) methods, where the chemical shifts are calculated *ab initio* without any prior knowledge. These two families of methods are reviewed in chapters 2.1.4 and 2.1.5, respectively. Both approaches have found their place in the various applications of protein NMR, as reviewed in chapter 2.1.6.

The prediction of chemical shift can also be done based on protein sequences by homology modelling [16,17]. Naturally, these approaches are of no use for most protein structure applications as they cannot distinguish one conformation from another and perhaps consequently, not much development on these methods has occurred recently. However, they are still useful for certain applications such as chemical shift assignment [18] in cases when no tertiary structure is available.

2.1.1 Brief introduction to protein NMR

Protein structure determination by NMR is a procedure with many steps, starting from isotope labeling of the sample, continuing to sequential assignment and acquisition of the parameters required to derive the structure restraints, and ending to structure calculation and validation. Since the 1980s, continuous development of all of these steps, along with the use of higher magnetic fields offered by present-day NMR magnets, has raised the method to its current status as the only widely used protein structure determination method in addition to X-ray crystallography. Whereas the procedure initially relied on NOEs and J-coupling information, nowadays more and more structures are refined with RDCs, introduced in 1997 by Tjandra and Bax [19]. Another remarkable methodological advance is the transverse-relaxation optimized spectroscopy (TROSY) [20], pushing up the molecular size limit of protein NMR. Today, there has been more emphasis placed on the development of the dynamic nuclear polarization methods, which have vastly increased the quality of solid state protein NMR spectra [21].

In addition to the determination of protein structure and dynamics as themselves, there are numerous other topics in protein NMR, many of those revealing data unobtainable with any other method. In order to emphasize the potential and versatility of NMR, some of them are listed here (with references to recent review papers), including 1) solid state protein NMR [22,23], 2) studies of membrane proteins [24,25]; 3) protein-ligand interaction mapping [26], 4) studies of intrinsically disordered proteins (IDP) [27], 5) in-cell NMR [28], and 6) studies of protein folding and intermediates [29]. Furthermore, in addition to proteins, NMR is commonly used with other biomolecules too. The burial of the concept of so-called 'junk DNA' [30] and the resulting interest on functional RNA molecules [31] has recently attracted much attention to NMR studies of nucleic acids [32,33].

Advances have also occurred in the computational side of the protein NMR workflow, including automation of both the sequential assignment phase and the actual structure calculations [34]. In order to provide a basis for some aspects of this thesis, the basic workflow of structure calculation is briefly introduced. Basically, structure calculation is molecular modelling, either in Cartesian or torsion-angle space, with the experimental structure restraints applied. In order to map the conformational space as extensively as possible, one can use different methods such as simulated annealing [35,36] and different setups of Monte Carlo-simulations [37–40]. In addition to the traditional NOE-based distance restraints and J-coupling based torsion angle restraints, a variety of other experimental restraints are available, including chemical shifts based torsion angle and distance restraints (see chapter 2.1.6), RCD-based orientation restraints and hydrogen bond restraints from amide hydrogen/deuterium (H/D) exchange rates.

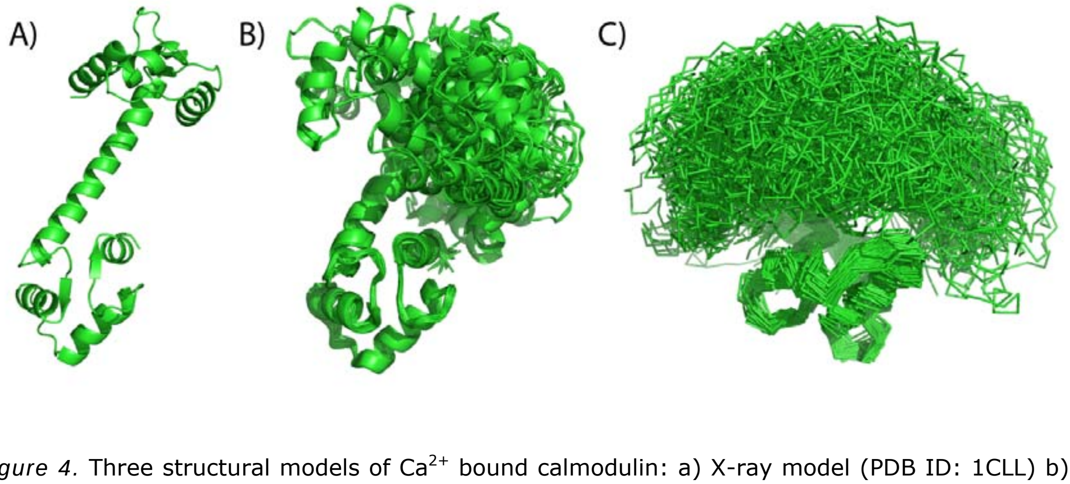


Figure 4. Three structural models of Ca^{2+} bound calmodulin: a) X-ray model (PDB ID: 1CLL) b) NMR ensemble (PDB 1X02) c) dynamic ensemble (PDB 2K0E)

The structure calculation protocol returns multiple structures consistent with the restraints. Therefore, in contrast to X-ray structures (Fig. 4a), NMR structures are usually published as an ensemble of conformations, known as *NMR ensembles,* typically containing about 20 conformations (Fig. 4b), selected by some criteria such as the lowest energy or the least restraint violations. It must be emphasized that these ensembles do not follow Boltzmann statistics, and they represent molecular motions only indirectly by allowing more flexibility in moieties with less restraints. In order to map real dynamics instead, many approaches determining *dynamic ensembles* (Fig. 4c) have been presented recently (see chapter 2.1.3).

2.1.2 Structural information of protein chemical shifts

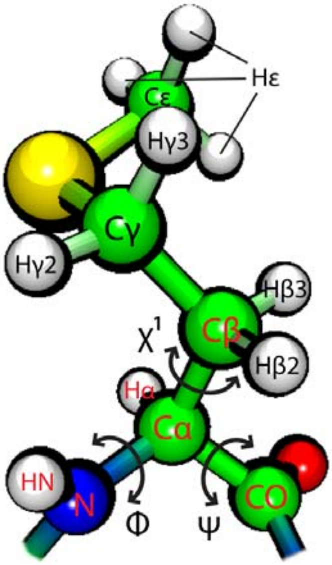


Figure 5. ^{1}H, ^{13}C and ^{15}N atom types of the methionine residue. Backbone chemical shifts, common for most amino acids, are shown in red. The side chain atom types, here specific for methionine, are shown in black. The figure also shows the backbone torsion angles Φ and Ψ and first side chain torsion angle χ1.

Since proteins are polymeric structures with limited chemical diversity originating from only 20 different standard amino acids, their chemical shifts have several characteristics. First, this allows the nuclei in proteins to be classified into residue-specific *atom types* (Fig. 5), following a standardized nomenclature [7]. Second, for each of these atom types, a certain reference value, called the *random coil shift,* can be given. The random coil shifts are measured in short peptides unable to adopt secondary structure conformation [41–43], and thus resemble the chemical shifts without any external structural effects. When combined with the nearest-neighbor effects, i.e. the effect from the preceding and next residue in the amino acid chain [44], all substituent effects are covered. This can be juxtaposed with the $\Delta\sigma_{\text{local dia}}$ and $\Delta\sigma_{\text{local para}}$ shielding terms of Eq. 1.3. The total observed chemical shift (δ_{obs}) can then be expressed with Eq. 2.1,

$$\delta_{obs} = \delta_{rc} + \Delta\delta \qquad [2.1]$$

in which δ_{rc} is the random coil shift. The remaining part ($\Delta\delta$) is the *secondary chemical shift*, which features the effects arising from protein folding to its secondary and tertiary structures, and relates to the structural effects term (σ') of Eq. 1.3. The secondary chemical shift contains all of the structural information embedded in the protein chemical shifts and is therefore an extremely useful measure in many applications (see chapter 2.1.6). The source of $\Delta\delta$ can be both local, arising from sequentially neighboring residues, or more distant, carried by through-space effects as a result of tertiary folding. For most backbone nuclei, the variance of $\Delta\delta$ is larger than the variance between δ_{rc} of different amino acids [45]. Examples of $\Delta\delta$ distributions for some atom types are shown in Fig. 6.

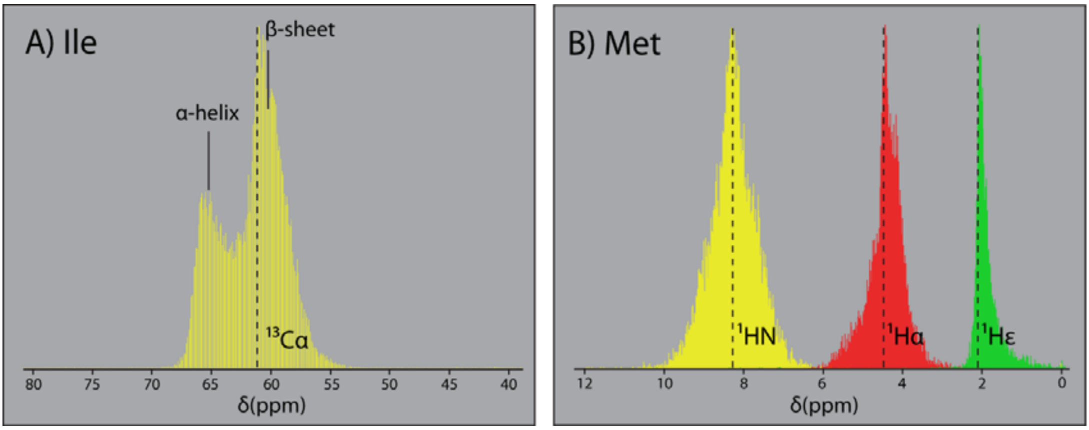


Figure 6. Chemical shift histograms and random coil shifts [41] (broken lines) of a) Ile ^{13}Cα, showing distinct regions for α-helix and β-sheet structures; b) Met ^{1}Hα, ^{1}HN and ^{1}Hε nuclei, each showing different amount of variance. Histograms are made with BMRB chemical shift distribution visualizer.

The secondary shift can be subdivided into more detailed contributions with analogy to Eq. 1.4, but specifically for the major effects seen in proteins. For example, Eq. 2.2 can be written for the backbone nuclei [46].

$$\Delta\delta = \delta_{anis} + \delta_{ring} + \delta_{HB} + \delta_{el} + \delta_{side} + \delta_{misc} \qquad [2.2]$$

Here, δ_{anis} corresponds to the bond anisotropy modulated by peptide backbone torsion angles, δ_{ring} is the ring current contribution, δ_{HB} is the hydrogen bonding effect, δ_{el} is the electric field effect similarly as in Eq. 1.4, and δ_{side} is the side chain contribution. Finally, δ_{misc} is the mostly unknown residual containing all the minor effects e.g. those arising from the sample conditions. The terms of Eq. 2.2 are of structural rather than physical origin. For example, δ_{anis}, δ_{ring} and δ_{side} physically contribute to the magnetic shielding term $\Delta\sigma_{\mathrm{magn}}$ of Eq. 1.4. The hydrogen bonding effect δ_{HB} is mainly an electric field effect, but it is often found to be beneficial to separate this term from δ_{el} [47]. Unfortunately, many of these terms overlap, which complicates undertaking an the explicit analysis of the origins of protein chemical shifts. For example, the explicit hydrogen bonding and aromatic ring effects are distinguishable from other contributions only via QM calculations [48–52]. Estimates of the importance of the contributions for the different protein atom classes are given in Table 1.

Table 1. Structural contributions to the protein chemical shifts. '++' denotes a key contribution, '+' is a moderate effect and '-' denotes contributions that are mainly negligible for this class of atoms.

Contribution	**^{1}Hα**	**^{1}HN**	**^{13}Cα**	**^{13}Cβ**	**^{13}CO**	**^{15}N**	**^{1}H(sc)**	**^{13}C(sc)**
δ_{anis} (backbone torsion angles Φ and Ψ)	++	+	++	++	++	++	+	+
δ_{side} (side chain torsion angles χ^{n})	-	-	+	+	+	+	++	++
δ_{ring} (aromatic ring currents)	++	++	-	-	-	-	++	-
δ_{HB} (hydrogen bonds)	-	++	-	-	+	+	-	-
δ_{el} (electric fields)	+	+	-	-	-	-	+	-

sc = side chain

The majority of the backbone secondary chemical shift arises from bond anisotropy (δ_{anis}), which can be explained with torsion angles. This is a consequence of two aspects. First, the magnetic anisotropy of chemical bonds is one of the largest contributions to the chemical shifts, and the torsion angles effectively describe the orientation of the magnetic anisotropy tensor for the perturbed nuclei. Second, in the stable secondary structures of native proteins the backbone torsion angles have definable values, which permits a correlation between the torsion angles and shifts to be determined, initially conducted in the study of Dalgarno et al. [53]. Later, the backbone torsion angle effects were found to account for about half of the $\Delta\delta$ of the ^{1}Hα, ^{13}C and ^{15}N nuclei [45,47,54,55]. Reversed, this opens up the possibility to determine protein secondary structures (often distinctly revealed in the chemical shift distribution, Fig. 6a) and torsion angles from chemical shifts (see chapter 2.1.6). The ^{13}Cα and ^{13}Cβ chemical shifts are especially important since, when compared to protons, they are less prone to ring currents and hydrogen bonding effects [46,55], and thus have a more straightforward relationship with respect to the backbone dihedral angles. In addition, certain side chains, especially β-branched and/or charged ones, are able to induce significant anisotropy effects (δ_{side}) into the backbone $\Delta\delta$ [45,48,54–56].

In contrast to all other backbone nuclei, the ^{1}HN shifts are less reproducible with torsion angle data only. Instead, the hydrogen bonding effects (δ_{HB}) are of greater importance for ^{1}HN, accounting for about 25% of the total $\Delta\delta$ [45,55]. The hydrogen bond effect for ^{1}HN is not only exponentially dependent on the hydrogen bond length, but it is dependent also on the bond and torsion angles of the hydrogen bond system [49,55,57]. Moreover, the secondary hydrogen bonding (the hydrogen bonding of adjacent groups, e.g. the carbonyl of the same residue) have substantial effects [51]. These sensitive contributions, as well as the fact that ^{1}HN is an exchangeable proton prone to effects arising from sample conditions, show as a broad $\Delta\delta$ distribution (Fig. 6b) and make the interpretation of the ^{1}HN shifts considerably more difficult than the other backbone nuclei. In addition, the ^{13}CO and ^{15}N nuclei are also affected by the presence or the absence of the hydrogen bond [48,58–60].

The aromatic ring current term (δ_{ring}) is the single strongest contribution to the proton shifts. The relation is long known [3–5] and widely studied [52,61–63]. Like δ_{HB}, the effect is very sensitive to the molecular geometry, which makes it difficult to reproduce if the structural resolution is inadequate. For heavy nuclei, which are effectively buried under protons, the aromatic ring effect is mostly negligible [45,55].

While backbone nuclei motions are restricted by the tertiary packing, side chain nuclei enjoy greater conformational freedom within the NMR time scale. This is reflected in their $\Delta\delta$ having a smaller variance than the backbone shifts (Fig. 6b). However, the side chain shifts are dependent on the side chain torsion angles and are therefore good parameters for estimating the rotameric conformations of side chains [64]. Furthermore, ^{1}H side chain shifts are prone to aromatic ring currents and electric field effects, and thus can be used as probes for through-space contacts [65,66].

2.1.3 Dynamic information of protein chemical shifts

In addition to the three spatial dimensions of the atom coordinates, NMR is dependent upon the *fourth dimension* i.e. the motions of the studied molecule. A variety of experiments can be used to study the dynamics of different time scales of protein dynamics (Fig. 7). The widely used S^{2} order parameters [67] are derived from the nuclear relaxation rates R_1 (=$1/T_1$), R_2 (=$1/T_2$) and heteronuclear NOEs and they provide information about the angular motions of bond vectors on the sub-ns timescale. Continuing towards slower dynamics, rotating frame relaxation dispersion experiment [68] map μs dynamics by measuring R_1 in the rotating frame ($R_{1\rho}$); and the Carr-Pursell-Meiboom-Gill (CPMG) relaxation dispersion experiment reveal chemical exchange on the ~ms timescale via R_2 rates [69].

Finally, on the slow end of the timescale are those experiments that lie outside the chemical shift window, i.e. different conformations give rise to separate signals. These include Exchange spectroscopy (EXCY), probing dynamics happening during T_1 relaxation [70]; and real-time NMR [71], following dynamics that are slow enough to show as changing chemical shifts and intensities in sequential experiments, e.g. protein folding and H/D exchange. Between fast and slow exchange, line shape is heavily affected (Fig. 7) and can be analyzed to map ms scale dynamics [70].

In addition to the above-mentioned relaxational parameters, all other NMR parameters that are observed via chemical shifts are averaged over the time-scale of the chemical shift measurement (Fig. 7, fast exchange region), roughly up to the millisecond time-scale. This also enables the use of these parameters to measure protein dynamics. In this sense, the RDC-derived S^2_{RDC} parameters [72] are especially useful since they are analogous to S^2 but map the motions over a much wider ps to ms timescale.

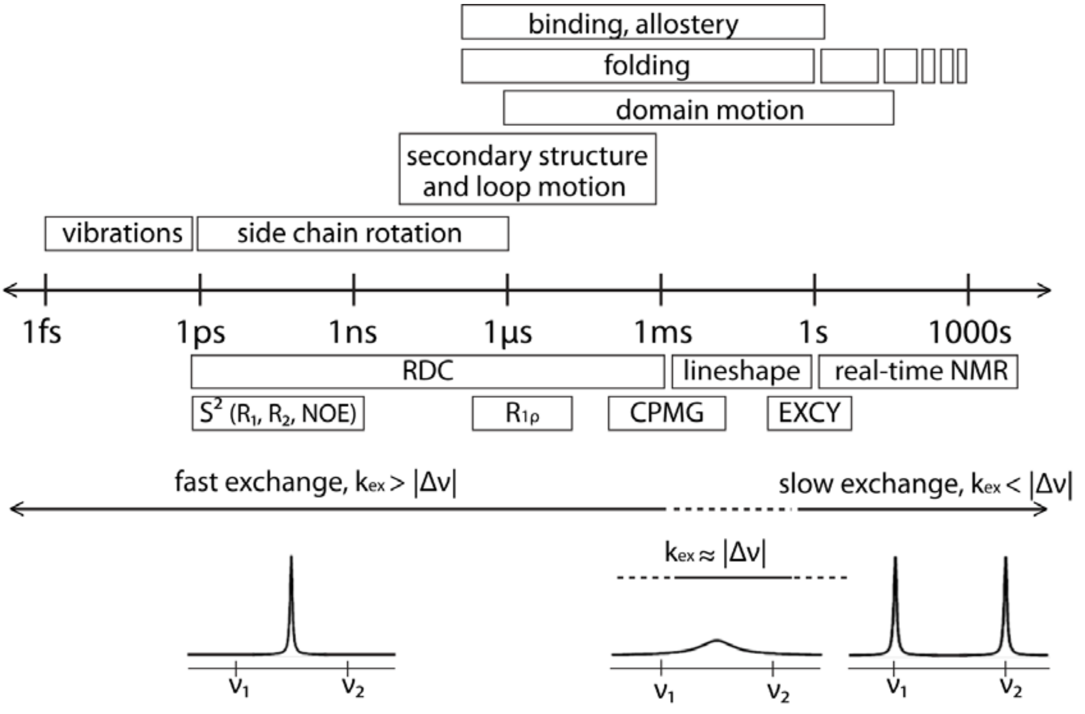


Figure 7. NMR experiments for mapping the different time scales of protein dynamics. On the bottom, signal outlook dependence on exchange rate k_{ex} and chemical shift difference |Δv| are shown with approximate time-scales. Figure combined from refs. [70,71,73]

Basically all time-averaged NMR parameters can be used to map dynamics by the generation of dynamic ensembles (Fig 4c). The largest flaw encountered with conventional NMR ensembles is that they try to fulfill all structural restraints in each conformation. This is in contrast to the dynamic nature of proteins, existing in multiple conformations constantly exchanging with each other; and the observed NMR parameters are averaged over those conformations. The more flexible the protein, the more important this becomes, and finally with intrinsically disordered proteins either single conformer or typical NMR ensemble representations would make no sense. [74] Therefore, much effort has been aimed towards achieving dynamic ensembles, which in principle represent the experimental data as an ensemble average. The most widely used approach to achieve this goal is the replica-exchange molecular dynamics (REMD) [75]. In REMD, the same structure is modeled in several replicas simultaneously and experimental restraints are always applied for the whole ensemble. This confers more freedom on the protein to adopt different conformations.

In REMD simulations, the typically used parameters include NOEs [76–78] and RDCs [78,79]. It is possible also to use restraints with dynamic nature that are not feasible in conventional structure calculations, such as the S^2 order parameters [76,77,80]. The time-scales of the used restraints will determine also the time-scale of the produced ensemble [74]. Chemical shifts restraints [81] have also been added into the REMD restraint tool box recently and used in several case studies. By spanning a broad time-scale, they have been shown to be able to model the interdomain [82,83] and loop [84] motions. Compared with non-restrained MD, the chemical shift restrained simulations have been shown to be better able to reproduce not only the experimental chemical shifts [85], but also other observable parameters [82,84].

In addition to the REMD simulations, it is possible to determine ensembles without any explicit dynamic exchanges but still reproducing the time-averaged parameters. These methods rely on generating large pools of conformations from which the representative ensemble is then selected based on the experimental parameters. [74] Chemical shift information, via prediction error, has been used as one of the experimental parameters in these kinds of methods [86,87]. The major issue with this approach is the ensemble selection problem being heavily underdetermined, i.e. the number of possible conformations is much larger than the number of observed parameters and no single solution is possible [74,86,88]. Different attempts have been made to overcome the issue include e.g. maximizing the entropy of the ensemble [86,89] or finding the minimum amount of conformations required for the representation of the observables [88,90]. In addition to IDPs [86,87], these methods are important for multidomain proteins with flexible linker regions, usually populating multiple conformations with equally low free energy [88].

Chemical shifts can also be used to probe dynamics in a more direct manner. Berjanskii and Wishart have demonstrated how chemical shift can be used to estimate S^2 order parameters and root mean square fluctuations [91]. Subsequently, they extended this work also for side chains [92]. Chemical shift prediction can also be exploited. The study by Robustelli et al. [93] showed how chemical shift predictions of MD trajectory snapshots can be used for revealing local dynamics in a simple manner by comparing the experimental chemical shifts to the distributions of shift predictions of moieties undergoing a conformational exchange during the MD simulation. The same approach has been recently shown able to capture the motions occurring in different time scales [94]. In addition, side chain rotamer populations have been determined directly from the side chain methyl shifts [95,96].

2.1.4 Empirical protein chemical shift prediction methods

Empirical chemical shift prediction methods rely on databases of experimental chemical shifts and the corresponding protein structures tertiary structures, usually by the means of atom coordinates. The level of abstraction varies among the methods. For example, the effect of bond anisotropy from the carbonyl C=O bonds can be taken into account directly using the classical McConnell equation [2] (Eq. 1.5), or in a more abstract manner by modelling the chemical shift as a function of backbone torsion angles. In addition, the methods vary in the extent to which the empirical data is used, starting from regression models (that describe the chemical shifts as a single function, often analogous to Eq. 2.2) and ending in database search methods, which employ different approaches to retrieve similar chemical shifts directly from the database. In fact, most of the recent methods are combinations of two or more different approaches.

The history of protein chemical shift prediction dates back to 1980s. Immediately after the first complete assignment of protein ^{1}H spectrum [97], the first correlations were observed between chemical shifts and structural parameters. The first of these correlations were the influence of hydrogen bond length [98] and backbone dihedral angle ψ [53] to ^{1}H shifts. At the same time as more assignments became available, these developments lead to the first programs to calculate ^{1}H chemical shifts in the early 1990s [99,100]. Simultaneously, the strong dependency between backbone dihedral angles and ^{13}C shifts was noted, yielding the first chemical shift hypersurfaces [101], soon to be followed by the same observation for ^{15}N shifts [102]. These studies not only created the foundation still used today by empirical prediction methods, but they also attracted interest in chemical shifts as a source of structural information, which has been driving the field forwards.

In addition to 4DSPOT, a number of empirical prediction approaches have been published this millennium (Table 2). Certainly the most famous is SHIFTX [47], which advanced the prediction accuracy to a level where it still remains today. SHIFTX was also one of the first methods to feature predictions for all backbone nuclei in a single program. Other widely used methods include SPARTA [103], its successor SPARTA+ [60], and CamShift [104].

In general, the best structure-based predictors report RMS errors of ~0.25 for ^{1}Hα nuclei, ~0.45 ppm for ^{1}HN nuclei, ~1.0 ppm for ^{13}C nuclei and ~2.5 ppm for ^{15}N nuclei; the corresponding $\Delta\delta_{obs}$ vs. $\Delta\delta_{pred}$ R correlation coefficients vary between 0.7 and 0.9 (^{1}HN < ^{15}N ≈ ^{13}CO ≈ ^{13}Cβ < ^{1}Hα < ^{13}Cα). Most of the reports also present accuracy comparisons to other methods. Naturally, these comparisons are subjective on the test set chosen, and consequently they are more or less favorable to the currently presented programs. With the exception of one comparison for the predictions of solid state structures [105], no neutral comparisons have been published in the literature. Moreover, the differences in the methodology of the predictors cause more issues. Since some of the predictors are designed for some particular use, the good prediction accuracy, assessed by root mean square (RMS) error or R correlation coefficient, might not always be the only desirable property. For example, the SHIFTX2 method has shown that by combining sequence-based and structure-based prediction methods, the RMS errors are halved [55]. However, this obviously has no extra benefit for comparing structures of the same sequence (i.e. different conformations of the same protein), which is the case in most applications. Similarly, the possible use of dynamics in prediction complicates the situation. For fair comparisons, the input should be the same for all compared methods. On the other hand, no prediction model should be evaluated based on test data of different dynamics distinct from the underlying teaching set. Therefore, if one predictor is designed for dynamic protein models as the input and others are not, direct comparison may be difficult.

For all the above reasons, predictor comparisons should be interpreted with caution. After all, the recent empirical chemical shift predictors do not differ extensively from each other by the means of RMS error as an indicator of the prediction accuracy. Instead, when choosing which predictor to use, a more important aspect should be the target for which the method is to be used. For example, CamShift may lose some accuracy by using differentiable parameters, required for deriving MD restraints [104]. Similarly, SHIFTX2 is the tool of choice when a homologous structure is already available [55]; and for prediction of dynamic ensembles, one should use a predictor that is taught against averaged descriptors such as PPM [106] or 4DSPOT (Papers I and II).

Table 2. Recent protein chemical shift prediction methods.

Name	Predicted nuclei [a]	Method in brief	Applications
Empirical methods			
SHIFTX [47]	bb, ^{1}H sc	Hypersurfaces and classical equations	Structure validation servers [107–109] CS23D structure generation [110]
SHIFTX2 [55]	bb, ^{1}H sc, ^{13}C sc	SHIFTX combined with sequence homology search	
SPARTA [103] SPARTA+ [60]	bb	Amino acid triplet search for similar torsion angles, combined with classical equations. SPARTA+ adds neural networks.	CS-ROSETTA structure generation [111,112]
CamShift [104] CH3Shift [66] ArShift [65]	bb, ^{1}H methyls, aromatic ^{1}H	Differentiable functions, mainly interatomic distances	Distance restraints [81]
PROSHIFT [113]	bb, ^{1}H sc, ^{13}C sc	Neural network model using a large number of structural parameters	
PRSI [56,114]	bb	Residue-specific torsion angle hypersurfaces	
shAIC [115]	bb	Parameters selected based on Akaikes Information Criterion[116] for robustness outside the structural space of the teaching set	
PPM [106]	bb, ^{1}H methyls	Parameters averaged over 100 ns MD simulations	
HASH [117]	^{1}Hα	Prediction of ^{1}Hα shifts from already known other backbone shifts	Sequential assignment
QM methods			
SHIFTS [48,49,118]	bb	Chemical shift hypersurfaces built from DFT calculated shifts of 1335 peptide conformations. Accompanied by equations for hydrogen bonding effects.	
CheShift [119]	^{13}Cα	Chemical shift hypersurfaces built from DFT calculated shifts of almost 700 000 peptide conformations.	Structure validation [120]
ProCS [50]	^{1}HN	QM-derived equation containing additive terms for torsion angle effects, hydrogen bonding and ring currents.	Structure refinement (PHAISTOS)[37]
Direct QM calculation [121–128]	any	Automatic fragmentation of proteins for QM calculation. May be combined with MD simulations and explicit solvent.	

[a] bb = backbone, sc = side chain

2.1.5 Quantum-mechanical methods

As chemical shifts are dependent on the surrounding electron density, they can be calculated with quantum-mechanical principles. After decades of development in the methods as well as improvements in the available computational power, the accuracy has greatly improved and rather well established "golden standards" for QM calculation of small molecule chemical shifts are now available [129,130]. In proteins, many attempts have been recently made to calculate chemical shifts with QM methods using fragmentation schemes [121–126], in which the protein is chopped up into small enough pieces that QM methods can crunch in a reasonable time. Most of these approaches are based on density functional theory (DFT) level of calculations, which offer a reasonable tradeoff between accuracy and speed. The QM-based approaches can handle any desired protein system, for example protein-ligand or protein-DNA complexes, but the accuracy is greatly dependent upon the method and the level of theory being used. In some cases, the accuracy can reach the level of empirical methods, but for nuclei more prone to electrostatic or solvent effects, like ^{15}N and ^{1}HN, the results are less reliable [124]. In addition, as with experimental data, the dynamics must be accounted for in order to achieve sufficient accuracy. To overcome these limitations, efforts have been made to apply MD simulations [127,128], even quantum-mechanical MD [131], and explicit solvent models [127,128] to the methods.

Despite the abovementioned advances, the direct QM calculation of protein chemical shifts is still rather time-consuming, usually requiring several days of computation on supercomputers. However, QM calculations can also be exploited in an indirect manner by calculating the chemical shifts of some model systems in order to create artificial chemical shift databases, which can be then used to train prediction models similarly as done in empirical methods (Table 2).

SHIFTS [118] was the first program to use this approach. In the method, DFT chemical shifts for backbone ^{13}C and ^{15}N nuclei were calculated for 1335 peptide conformations. From this data, chemical shift hypersurfaces to model torsion angle effects were built, accompanied with equations for hydrogen bonding effects. The method was later extended to cover ^{1}HN shift prediction using a more detailed hydrogen bonding model [49]. At present, SHIFTS is the only QM-based chemical shift predictor covering all backbone nuclei.

A similar approach was used by the CheShift server [119] for ^{13}Cα shift prediction, intended for protein structure validation. Although the prediction RMS error is about twice as large compared to the rival empirical approaches, this method has been shown to be more sensitive to detect structural differences. This is possible since the QM-calculated shifts are always consistent with structure and the inaccuracy of experimental protein models is not limiting the sensitivity.

The most recent QM-derived chemical shift model is ProCS [50], which extends the earlier work of Parker et al. [51] to predict the ^{1}HN shifts. The study shows that the derived function for the ^{1}HN shifts accurately reproduces the QM-calculated shifts, but due to inaccuracies and the lack of dynamics of the experimental X-ray structures, the results do not correlate with the experimental chemical shifts. Interestingly, the method was still perfectly able to be used in structure refinement, achieving clearly more sharp profiles for hydrogen bond geometries.

Due to the high computational cost and originally poor accuracy, QM chemical shift prediction methods have long been shadowed by empirical ones. However, in the last four years, many new QM methods have emerged in the literature. The development in prediction accuracy has been much faster than in the empirical methods, which in contrast seem to be closing the limit determined by the precision of the experimental data. Furthermore, since the QM methods have also been shown to have greater sensitivity against structural differences [50,119,132], it can be expected that rather soon QM methods will replace empirical methods in many applications.

2.1.6 Applications

Although the main emphasis of this chapter is on applications based on chemical shift prediction, there is an abundance of structural information available from chemical shifts directly. The simplest such application is the secondary structure determination, which relies on the typical upfield and downfield backbone shifts of α-helices and β-sheets, originally performed based only on ^{1}Hα shifts [133]. Today, this application is well established in a number of programs, combining information from all backbone shifts to allow prediction of secondary structures probabilities [134–136], or secondary structure populations of disordered proteins[137], with good confidence. The next obvious step forward was to determine torsion angles based on chemical shifts, since the correlation between these parameters is strong and well known (see chapter 2.1.2). This was first achieved in the program TALOS [138], predicting torsion angles from a database of amino acid triplets and their chemical shifts. The determination of as torsion angle restraints for structure calculation, by TALOS or some of its successors such as TALOS+ [139], PREDITOR [140] or DANGLE [141], is, based on bibliometrics, the most widely used application of chemical shifts in NMR structure determination.

The applications exploiting chemical shift prediction have mostly emerged in this millennium, after the accuracy of empirical predictors improved to be sufficient for such purposes (Table 2). Generally, most of the applications use the prediction error (δ_{obs}-δ_{calc}) in some way as a probe for structural correctness. This measure can be readily used in structure validation, the final and crucial step in NMR structure determination. Basically, any chemical shift predictor can be used for detecting discrepancies between shifts and structure and the possible reasons for these discrepancies (for example, see Appendix IV).

Many free programs and web servers have been published to help in structure validation. These include CheShift [119] and its updated version CheShift-2 [120], based on the QM-derived ^{13}Cα shift prediction (see chapter 2.1.5). Three other web servers, PROSESS [107] and CING [108], combine multiple metrics from a variety of programs to provide both global and residue-specific scores for structure validity. In both programs, comparison of experimental chemical shifts to SHIFTX [47] predictions forms a part of the given score. In addition, the CoNSEnsX server [109] uses a similar approach but stresses the importance of dynamic ensembles by assessing the parameters in an ensemble averaged manner. Sahakyan and co-workers also demonstrated the usability of side chain chemical shifts in structure validation [142], making their approach especially sensitive to three-dimensional packing.

An application closely related to validation is also the ensemble generation i.e. selecting conformations to represent the experimental data. This has been implemented in the ENSEMBLE program [143] by combining information of multiple NMR parameters, one again being the chemical shift prediction error of SHIFTX [47]. As the problem of conformer selection is highly degenerate, Bayesian statistics is used for calculating the conformational weights [86].

In addition to the assessment of protein structures, also the chemical shift datasets often require validation [46]. Detection of possible offsets and re-referencing can be done by statistical analysis of chemical shifts distribution with [144] or without [145,146] knowledge of the structural coordinates, or by applying structure-based shift predictions [147]. Furthermore, it is always necessary to check all of the chemical shift outliers e.g. to detect possible errors arising from assignment phase. The Vivaldi server [148] compares the experimental chemical shifts with the secondary structure and solvent accessibility-dependent chemical shift distributions of the VASCO database [144] and reports severe outliers. Recently, as a way of stressing the importance of proper assessment of all phases of NMR structure determination, the Worldwide Protein Data Bank has recently established a task force to standardize and further develop NMR structure validation protocols [149].

In addition to torsion angle restraints (see above), chemical shifts can be turned into structure calculations restraints via prediction. The first such approaches added universal penalty terms to structure calculations based on prediction error [150]. However, this "direct chemical shift refinement" approach was never widely used since most of the larger conformational changes lead to an increase in the chemical shift penalty term. Therefore, these methods are unable to refine anything other than very small structural changes (local minima problem), which, on the other hand, are not sufficiently accounted for in the chemical shift predictors [46].

Recently, several other attempts have emerged. Robustelli and co-workers introduced the chemical shift based distance restraints [81], based on differentiable prediction parameters of the CamShift chemical shift predictor [104]. The method was demonstrated to be able to derive correct protein folds starting from partially folded states. Later the approach has been used in replica-averaged MD studies of several proteins [82–84] (see chapter 2.1.3). In addition to MD simulations, proteins structures can also be determined and refined with Monte Carlo simulations, in which moving from one conformation to another is conducted based on probabilities of the moves. These simulations can also use chemical shift information as the source of information. This approach has been recently suggested by implementing the chemical shift predictions of the ProCS method [50] in the PHAISTOS framework [37].

Several methods have been developed for generating protein structures based solely on chemical shift information in an automatic manner [110,111,151], also from incomplete chemical shift assignments [112]. The basic workflow of these methods includes homology modelling, usually by the Rosetta method [152], to create a number of structure proposals, for which the chemical shifts are predicted. The structures are then scored based on the observed-calculated shift difference. However, the reliability of these methods is still questionable [153]. In attempts to improve the confidence of the resulting structures [153], some of these methods offer versions complemented with automatically assigned NOE information [154,155]. Overall, due to the high cost of protein structure determination, automation of the procedure is a highly desirable goal.

2.2 SMALL MOLECULE J-COUPLING PREDICTION

2.2.1 Structural information of coupling constants

At first glance, scalar coupling constants are the parameters that mess up the otherwise clean NMR signals and that can often make the analysis of NMR spectra painfully difficult. Fortunately, the information embedded in those fine structures is invaluable and therefore worthwhile analyzing. Most importantly, J-couplings contain direct information about molecular topology i.e. how the atoms are connected to each other. This enables the use of NMR for elucidating unknown structures.

In addition to the J-couplings interpretable from basic 1D spectra, indirect spin-spin coupling plays an important role in many other NMR experiments in the determination of atom connectivity. In these multidimensional correlation experiments, the magnetization created in one nucleus is transferred via J-coupling to another, in which the spectrum is recorded. There are popular experiments e.g. COSY (Correlation Spectroscopy) for ^{1}H-^{1}H correlations; and HSQC (Heteronuclear Single Quantum Coherence) and HMBC (Heteronuclear Multiple Bond Correlation) experiments in which one- and multibond connectivities, respectively, are resolved via magnetization transferred from a proton to a heteronucleus (e.g. ^{1}H-^{13}C) and back. [156]

Coupling constants are also probes of the molecular structure. Since the spin polarization of J-couplings is carried over bond electrons, any perturbation to these electron clouds will have an effect on the coupling values. These effects can be divided into two main groups, 1) the effects arising from the interaction of molecular orbitals modulated by the geometry of the molecule, and 2) effects from electron donating or withdrawing groups as neighboring substituents. [1] Ranges of J-coupling values of some common coupling classes encountered in typical small molecule NMR are shown in Fig. 8. Obviously, the absolute coupling values are larger for the shorter coupling paths, and ^1J couplings commonly reach several hundred Hz.

The $^2J_{HH}$ couplings are the shortest commonly seen couplings in typical ^1H spectra. Although they are basically dependent on the bond angle between the coupled atoms, these angles do not typically have much variation, at least within the same hybridization of the central atom. Instead, these couplings are prone to the electric effects of nearest substituents, arising both from their electronic effects and the geometry against the coupled atoms in question. The substituents with σ-electron withdrawing or π-electron-donating character have positive effects on the $^2J_{HH}$ couplings. Sometimes the usually large negative $^2J_{HH}$ couplings over sp^3 hybridized carbons can be close to zero or even positive. These effects also apply *vice versa*, so σ-electron donors or π-electron acceptors have a negative effect on the couplings. [157] The negative effect of adjacent π-orbitals is largest when they are oriented in parallel to either of the coupled protons [1].

The most important couplings in all small molecule studies are the ^3J couplings for two reasons. First, the path is long enough to reveal more topological and structural features, and on the other hand the J values are still large enough to permit reliable evaluation. In general, the ^3J couplings follow the Karplus dependence [158], i.e. they are heavily dependent on the torsion angle determined by the four atoms (H^1-A-B-H^2, Fig. 8). This makes them invaluable when analyzing saturated aliphatic (*trans/gauche*) and olefinic (*cis/trans*) stereochemistry. However, it has to be noted that $^3J_{HH}$ couplings over freely rotating bonds are averaged over the conformations and have much less variance than those in rigid aliphatic ring systems. The electronic effects of the substituents also have their role and they have been accounted for in the subsequent refinements of the Karplus equation [159,160] by introducing additional electronegativity terms.

In aromatic rings, where the dihedral geometry is always unambiguous, the ortho $^3J_{HH}$ couplings are solely dependent on the atomic charges of the system. The divergent charge distributions of heterocycles, such as pyridines, have a major influence on the otherwise rather invariable coupling values. For olefinic and aromatic $^3J_{HH}$ couplings in general, the ring size has a significant effect arising from variations in H^1-A-B and A-B-H^2 valence angles. When these angles are small (e.g. in 8-membered rings), the coupling values are largest. [1]

J_{HH} coupling paths longer than four bonds are visible only when the molecular orbitals lie in a suitable geometry capable of transporting enough spin polarization. The most common such systems are the aromatic meta ($^4J_{HH}$) and para ($^5J_{HH}$) couplings, in which the spin polarization is carried via the conjugated π-systems. Similar couplings are seen in other conjugated π-systems such as alkanes, alkynes and allenes [1]. In aliphatic paths, the $^4J_{HH}$ couplings are visible only in particular W-shaped systems (both torsions exist in the trans conformation) [161]. These couplings are best seen in strained bicyclic systems and are largest when there are multiple paths between the same coupling pair [1].

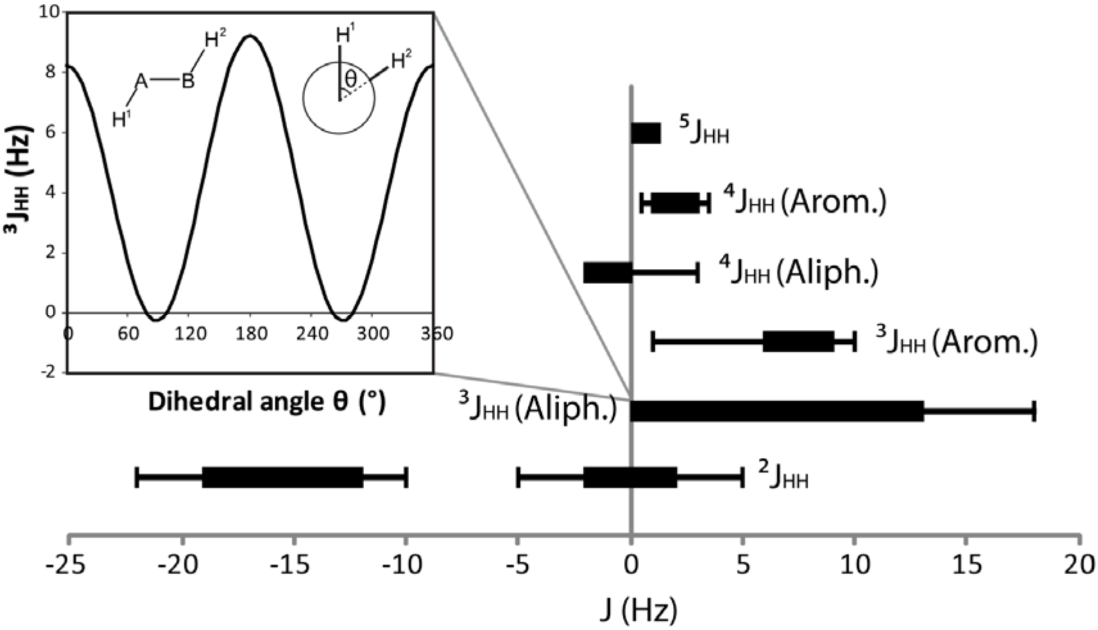


Figure 8. Ranges of J_{HH}-couplings often encountered in ^{1}H NMR spectroscopy. Thick bars denote the commonly seen ranges and thin bars the extreme cases. The figure is based on the Juniper database (Paper III) and Refs. [1,157,162]. In the inset, the "Karplus curve", the dependency of a $^{3}J_{HH}$ coupling value of the torsion angle θ, is shown [158].

In addition to J_{HH}, a very important coupling family in small molecule NMR are those of ^{19}F and ^{31}P, both being spin ½ nuclei with 100 % natural abundance. As with hetero couplings in general, these couplings usually appear as first order patterns due to large chemical shift difference. In approximate terms, J_{FH} couplings are at least twice as large as the corresponding J_{HH}. [162,163] Both $^{3}J_{FH}$ and $^{3}J_{PH}$ follow the Karplus dependence in a similar manner as $^{3}J_{HH}$ [164] and electronegativity-corrected equations have been developed [163].

The $^{1}J_{PH}$ couplings have been found to be dependent on phosphorus electron density, which varies extensively since phosphorus can exist in many oxidation states and ionic forms. Consequently, the $^{1}J_{PH}$ couplings have an especially wide range, from 50 Hz up to 1000 Hz of positively charged phosphorus [165]. For longer paths, the values decrease to a similar range as J_{FH}. However, whereas usually longer path couplings are smaller, in J_{PH} it is often the case that $^{3}J_{PH}$ is larger than $^{2}J_{PH}$. J_{PH} couplings are also strong enough to be carried over the oxygen atom of a phosphate group. [162]

When one considers the typical ^{13}C NMR spectroscopy, then J-couplings are usually of minor importance since due to low natural abundance only $^{1}J_{CH}$ is typically visible and, on the other hand, most ^{13}C spectra are measured with proton decoupling to improve signal intensity. However, if desired, they are available through inverse detection (e.g. HSQC) and $^{2}J_{CH}$ and $^{3}J_{CH}$ have also been found to be useful in structural studies [166,167]. In addition, J_{PC} and J_{FC} through 1 to 3 bonds are often visible also in standard spectra and they cause large splitting, especially in the case of $^{1}J_{FC}$. [162]

2.2.2 Coupling constant prediction methods and applications

Starting from 1960s, [158,168–170] a vast number of equations for different J-couplings have been derived. These equations are widely applied in structural studies for example to reveal cyclic, acyclic and olefinic stereochemistry [157,162,166,167] or rotamer populations [171]. These applications do not require actual J-coupling calculation since the equations can be used in reverse to predict the structure. Perhaps for this reason, only a few general prediction methods e.g. covering all typical couplings of ^1H spectrum are available. Most of these are commercial programs embedded within some NMR software [172–174], and the predictions of some of them are only qualitative. These predictors are mainly intended for spectral analysis by either manual or automated workflow (see below).

The only published more general coupling constant predictor is SPINUS [175], based on the approach and molecular parameters of the chemical shift prediction previously published by the same group [176]. SPINUS applies Associative Neural Networks (ASNN [177]), in which the prediction error remaining from neural network is corrected with the k Nearest Neighbor (kNN) search in the additional memory of experimental data. In the coupling constant version of the approach, the additional memory contains the coupled proton pairs with their experimental coupling values. With a database of 618 coupling constants, the coverage of SPINUS is obviously limited. For example, the $^3J_{HH}$ couplings of aliphatic ring systems had to be treated with the Karplus equation due to the lack of data with reliable stereochemistry. In internal Leave-One-Out tests, SPINUS was reported to predict $^2J_{HH}$ and $^3J_{HH}$ couplings with a mean average error below 1.0 Hz.

Similar to chemical shifts, coupling constants can be calculated with quantum-mechanics [9,10] and nowadays this option has been widely implemented in many QM software. The QM calculated J-couplings have been shown to be a great help in stereochemical determinations [178]. Recently, Bally and Rablen have reported that J_{HH} can be calculated with excellent accuracy using relatively light DFT methods. Unfortunately, even the lightest feasible DFT methods take several minutes for very small molecules such as chloromethane [11], preventing their use in applications where a high throughput is required. Moreover, for coupled atoms with lone pairs (e.g. fluorine) the DFT methods often fail and electron correlated wave-function methods are required [10].

Prediction of NMR parameters, not only chemical shifts but also J-couplings, has a key role in automatic NMR-based *structure verification*. Briefly, these methods answer the questions "does this NMR spectrum correspond to this compound", or more explicitly, "Is this compound purchased from this company what it should be?" or "Was the proposed synthesis successful?". The answer to these questions is given by analyzing the consistency of observed and predicted NMR parameters. Usually, the used parameters are chemical shifts combined with e.g. signal multiplicity from single ^1H spectra [179] or atom connectivities from 2D correlation experiments such as HSQC [180,181] and HMBC [182]. The analysis can be done also by comparing the entire observed and simulated spectra [12] (see below).

The *structure elucidation* methods [183,184] take this procedure one step further by elucidating unknown structures based on NMR spectra (often abbreviated as CASE for Computer-assisted structure elucidation). The CASE workflow starts by generating a number of structure proposals, based on some basic information such as molecular mass or formula and, if available, some complementary knowledge e.g. about the presence of certain functional groups [185]. In this step, atom connectivities e.g. from HMBC experiments are widely used [183,184] but alternatively, fragment databases can be employed [186]. In the following step, the NMR parameters predicted from the proposals are compared with the observed parameters to find the best match [185]. A fully automatic and reliable structure elucidation would be one of the 'holy grails' of small molecule NMR, but due to many uncertainties in the parameter extraction and prediction steps this goal has still not been achieved. Nonetheless, many successful case studies have been published, e.g. revealing incorrectly solved structures from the literature [187].

From the viewpoint of this thesis, the most important application for J-coupling prediction is for the structure verification. The Automated Consistency Analysis (ACA) program [12,188] performs the full spectrum analysis by simulating the spectrum of the proposed molecule from QM principles and comparing it to the observed spectrum. The fitting is done iteratively by changing the parameters of the calculated spectrum until the calculated spectrum matches the observed one. In the ACA workflow, J-coupling prediction is required for two reasons. First, coupling constant values are needed for the simulation of the initial spectrum i.e. the starting point of the iteration. Second, since ACA analysis also yields the values of the observed chemical shifts and J-couplings, the difference between observed and predicted J-couplings can be used as one of the parameters for the final score of the consistency evaluation (Match Index). This differs from the approach described by Golotvin et al. [179,180] in which J-coupling prediction is used only to enable the comparison of signal multiplicities with the actual J-coupling values not being evaluated.

3 Aims of the study

The working title of this study, as given at the time when applying to the doctoral program, was “4D Prediction of Protein Chemical Shifts and Applications”. In addition to the development of protein chemical shift prediction exploiting molecular dynamics, the ambitious study plan described possible applications for protein structure refinement and ligand interaction analysis. However, the fourth dimension drove me into other projects and collaborations, and as usual, the aims of the study were rewritten at the end of the thesis project. The background and aims of the two major projects are described in the following two chapters.

3.1 4D PROTEIN CHEMICAL SHIFT PREDICTION

The development of NMR chemical shift predictors started about 15 years ago in the research group of Prof. Laatikainen. Small molecule ^{1}H and ^{13}C chemical shift predictors have been a part of PERCH NMR Software [188] for many years and a paper describing the latest advances, including the efficient random forest regression, has been published recently [189].

Since the very beginning, the small molecule chemical shift prediction has been done in four dimensions, i.e. three spatial dimensions and time, in contrast to several competitors who conduct shift predictions from non-dynamic 3D structures or even two-dimensional structures (i.e. only from the molecular topology). The 4D approach has been proven to be essential for certain flexible molecules, but very often molecules have only one main conformer. Thus, on average, the 4D improvement is rather small [189], and chemical shifts of most small molecules can be well predicted using non-dynamic structures or Monte Carlo conformational mapping. This is the case especially for ^{13}C nuclei, which are less prone to through-space effects. On the other hand, proteins are naturally flexible, and the dynamic effects on the shifts are known to be significant [45]. Since I was already involved in the small molecule shift prediction project [189], the use of the 4D approach for protein ^{1}H chemical shift prediction formed as my master’s thesis project in 2007 and this continued later into this PhD study.

The principal aim of the project was to modify the small molecule chemical shift predictor for proteins, covering all ^{1}H, ^{13}C and ^{15}N nuclei including side chain atoms, and to assess the effect of the fourth dimension on the prediction results. The project was expected to yield a sensitive method for studying and understanding protein dynamics and to improve the 4D protein models; and to further prove the “dynamic hypothesis” i.e. the protein function is largely connected to its correlated motions [190]. Especially the shifts of ^{1}HN and ^{13}CO nuclei, which are associated with correlated secondary structure motions via hydrogen bonding, should reflect the validity of the dynamics of protein models.

Generally, the connection between molecular motion and function has been a long-term research subject in the Laatikainen group, with one topic being the design of flexible ligands [191] and the analysis of their interactions based on NMR chemical shifts. From this background, the 4D protein chemical shift prediction, in conjunction with the already established 4D shift prediction for small molecules, was expected to be valuable also for ligand interaction studies, which is one of the most used applications of protein NMR in drug industry.

Later, additional interesting study questions have arisen. Most importantly, the use of NMR derived proteins structures as teaching data was intriguing since this had not been attempted previously. As most of the applications for protein chemical shift prediction are aimed to NMR structures anyway, they were a natural choice for teaching data. The concept of using whole NMR ensembles as teaching data was largely initiated by the appearance of the report by Baskaran [192], showing that averaging of the single conformation prediction results of the ensembles achieved an improved prediction accuracy.

3.2 J-COUPLING PREDICTION

An even older research topic of the Laatikainen group has been the iterative NMR spectrum analysis based on integral-transform iteration and total-lineshape fitting [13]. In 2001, this development was commercialized by PERCH Solutions Ltd. resulting as PERCH NMR Software for automatic and manual spectrum analysis. Since an initial guess of NMR parameters is required for the iterative analysis workflow (see chapter 2.2.2), NMR parameter predictors have a rather important role in this program. The comprehensive chemical shift predictor [189], already mentioned above, has been a part of the program since its early stage. However, until now, the coupling constant prediction was based on hard-coded equations and values for the different coupling paths. Today, the in-house spectral database is substantial and simultaneously the throughput of the Automatic Consistency Analysis (ACA) program of PERCH NMR Software has been improved, yielding a large database of experimental coupling constants. Thus, the time was right for development of data driven method for J-coupling prediction, which became the second project of my PhD studies, carried out in collaboration with PERCH Solutions Ltd.

The intended use of the coupling constant predictor in automatic spectrum analysis induces several requirements. First, as the chemical diversity of small molecules is practically infinite, the *coverage* of the method plays a key role. Due to the iterative nature of the automatic analysis method, it cannot create new coupling constants ‘from scratch’, and thus the missing coupling constant predictions often lead to failed or at least ambiguous analysis results. Moreover, even when only considering ^{1}H spectra, there are several other nuclei (e.g. fluorine and phosphorus) that cause visible couplings to the spectra, and thus the method cannot be targeted only for J_{HH} couplings. The coverage requirement is closely linked to the next necessity, which is the need for the method to be easily maintainable. In reality, this means that prediction database should be easily extended, not only by its developers but also by users, if missing or too inaccurate predictions are encountered. Finally, due to the intended use of ACA as a high throughput method, the J-coupling prediction method should also be rapid. Emphasizing these considerations, the J-coupling predictor project was initiated.

4 Methods

This chapter briefly describes the data, methods and algorithms used in this thesis. For a full description of the methods, see the original papers.

4.1 DATA

In both of these prediction methods, the teaching data consists of two parts: 1) the molecular structures presented as atom coordinates in Cartesian space and 2) the observed data of the parameters to be predicted, either chemical shifts or coupling constants.

4.1.1 Protein chemical shift data

In the protein chemical shift prediction studies (Papers I and II), the data sources used were the public databases Protein Data Bank (PDB) [193], containing the 3D protein structures, and Biological Magnetic Resonance Bank (BMRB) [15] containing the observed chemical shifts. In Paper I, a database containing 40 protein molecules and about 21 000 ^{1}H chemical shifts was built. By the time of Paper II, the database had been extended to contain 94 protein molecules and about 50 000, 36 000 and 9 000 ^{1}H, ^{13}C and ^{15}N chemical shifts, respectively. In contrast to the mixed database of X-ray and NMR derived proteins structures utilized in Paper I, in Paper II the database was solely built from NMR structures, as the emphasis had shifted towards the use of the method with NMR ensembles. Recently, the database has been further extended (see chapter 5.1.4).

The molecular dynamics simulations for the proteins in the 4DSPOT teaching database were conducted using the AMBER molecular dynamics program [194], versions 9 (Paper I) and 10 (Paper II). The ff99SB force field [195] was used to perform the simulations for protein conformations solvated with TIP3P water molecules in periodic solvent boxes. Before the production simulations, an equilibration procedure was applied, which included the following steps 1) energy minimization and an 11.25 ps heating simulation to 300 K at a constant volume with position restraints on protein heavy atoms, followed by 2) energy minimization and an 11.25 ps equilibration simulation at 300 K and constant pressure with position restraints on protein backbone atoms. The production phases were 150 ps and 100 ps in Paper I and II, respectively; however, in Paper I, only the last 100 ps was used for averaging the chemical shift descriptors. During the production simulation, structures were saved with 0.375 and 1.0 ps intervals in Papers I and II, respectively. Throughout the simulations, a time step of 1.5 ps was used and the bonds to hydrogen atoms were constrained with the SHAKE algorithm.

4.1.2 J-coupling data

In the J-coupling prediction study (Paper III), the used molecular structures are geometry-optimized small molecules. The observed coupling constants originate from two sources. First, there is a literature-derived section, containing about 1 300 couplings. However, the vast majority of the data, about 40 000 couplings, has been gathered by automatic spectrum analysis performed in PERCH Solutions Ltd. using the ACA program of PERCH NMR Software [12]. The molecular structures are from the same sources, either the literature or the automatic analyses, as the observed coupling constants. One prerequisite for the structures is that the stereochemistry should be known. The used conformations are the anticipated minimum energy conformations obtained with a Metropolis Monte Carlo search, subsequently geometry optimized in the Molecular Modelling System –program of PERCH NMR Software using a modified version of the MMFF94 force field [196].

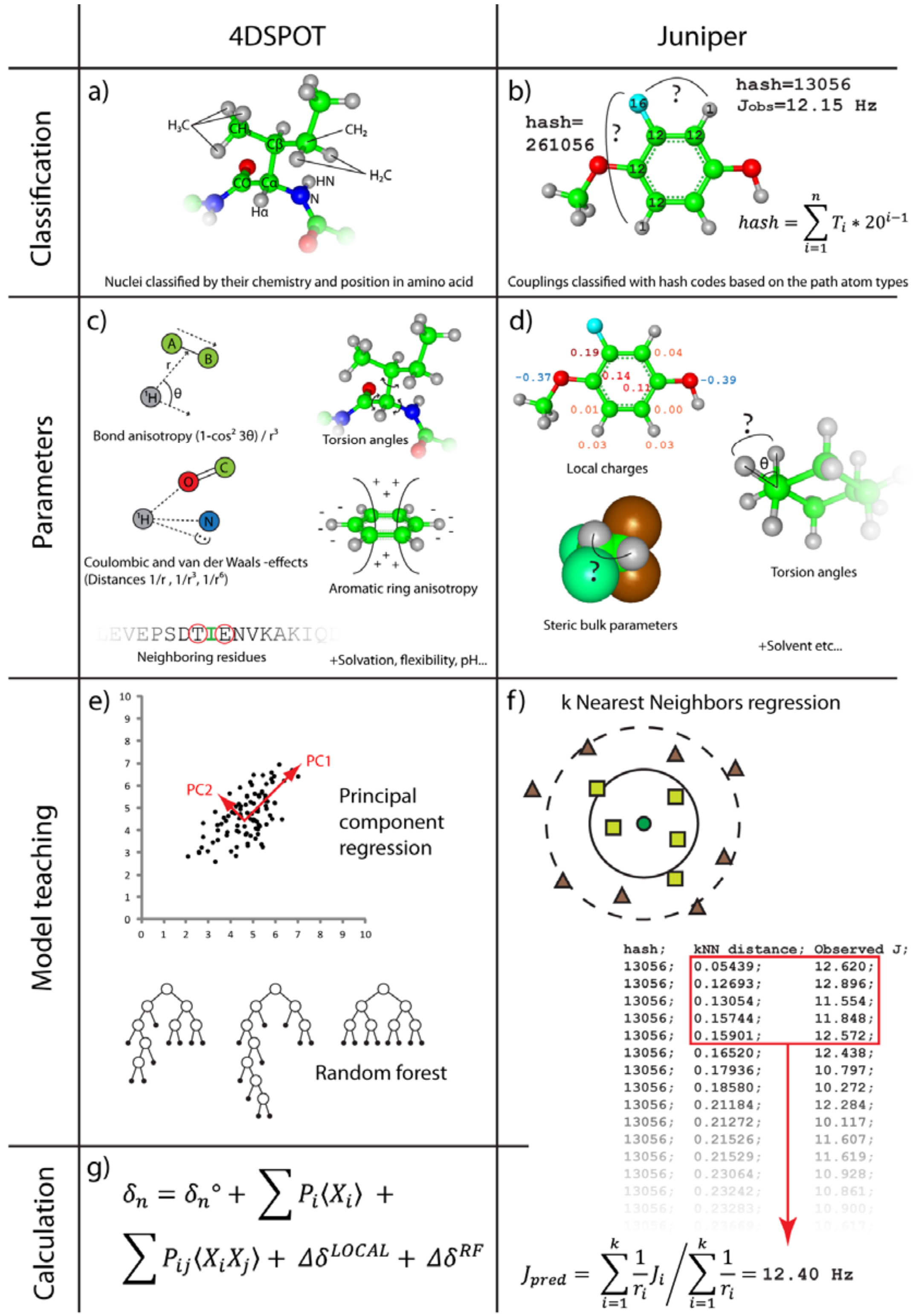
4.2 PREDICTION ALGORITHMS

Despite being targeted for different NMR parameters and different molecular families, both predictors also share common elements in their workflow. In both approaches, the prediction algorithm can be roughly divided into four main steps (Fig. 9), 1) classifying the nuclei or coupling paths to sub-classes, 2) creating the prediction parameters to describe the NMR parameters as a function of the molecular structure, 3) teaching the prediction models and 4) calculating the prediction results.

Since similar chemical environments usually have similar NMR parameters, the predicted systems are classified to homogenous sub-classes in both methods. This increases certainty of the prediction as non-relevant data points are omitted, allows the use of simple regression methods as homogenous datasets are easier to model, and importantly, it allows the regression methods to operate with a smaller amount of descriptors [197]. Dividing the data into subsets also speeds up the calculations. With respect to protein chemical shifts, the classification is rather simple due to the symmetry of the amino acid chain, and nuclei are classified according to their backbone classes ($^{1}H\alpha$, ^{1}HN, $^{13}C\alpha$, ^{13}CO, $^{13}C\beta$ and ^{15}N) and several side chains classes, such as ^{1}H and ^{13}C in CH_2 or CH_3 group (Fig. 9a). In contrast, the classification in Juniper plays a key role in the method. Since there is a vast chemical diversity to be covered, the classification must be carried out in a general manner. In the method, the coupling path atoms are classified into 20 different types according to their element and chemistry (for example, hydrogen is of type 1, carbon in an aromatic ring is 12 and fluorine is 16, see Fig. 9b). Based on these types T of the coupling path atoms i, a *hash code* is calculated with the equation shown in Fig. 9b. In the prediction, a query coupling with a certain hash code only sees the database couplings with the same hash code, thus sharpening the focus to relevant data points only.

The main task in both methods is to calculate the NMR spectral parameters as a function of the molecular structure. For this purpose, a set of structural parameters (also referred to as *descriptors*), many of those with some physical background (see chapters 2.1.2 and 2.2.1), are introduced. The parameters used to describe the chemical shift and the coupling constants are somewhat different, but they also share common terms, such as torsion angles and partial charge parameters. For both methods, the main classes of prediction parameters are shown in Fig. 9(c-d).

Finally, the goal of calculating NMR parameters as a function of structural parameters would be impossible without mathematical methods to derive the function. As the study questions of 4DSPOT and Juniper differ, so do also the mathematical methods, which will be elaborated in the following chapter.



$$hash = \sum_{i=1}^{n} T_i * 20^{i-1}$$

```
hash;    kNN distance; Observed J;
13056;   0.05439;      12.620;
13056;   0.12693;      12.896;
13056;   0.13054;      11.554;
13056;   0.15744;      11.848;
13056;   0.15901;      12.572;
13056;   0.16520;      12.438;
13056;   0.17936;      10.797;
13056;   0.18580;      10.272;
13056;   0.21184;      12.284;
13056;   0.21272;      10.117;
13056;   0.21526;      11.607;
13056;   0.21529;      11.619;
13056;   0.23064;      10.928;
13056;   0.23242;      10.861;
13056;   0.23283;      10.900;
13056;   0.23669;      10.617;
```

$$\delta_n = \delta_n{}^{\circ} + \sum P_i \langle X_i \rangle + \sum P_{ij} \langle X_i X_j \rangle + \Delta\delta^{LOCAL} + \Delta\delta^{RF}$$

$$J_{pred} = \sum_{i=1}^{k} \frac{1}{r_i} J_i \Big/ \sum_{i=1}^{k} \frac{1}{r_i} = 12.40 \text{ Hz}$$

Figure 9. Schematic presentation of the main steps in the workflow of 4DSPOT and Juniper.

4.3 MATHEMATICAL METHODS

4.3.1 Principal component regression

In 4DSPOT, the search for the function between the chemical shift and the descriptors relies on principal component regression (PCR). The result of PCR is a standard linear model of the chemical shift as a function of the original descriptors, but the coefficients of the model are solved through the principal components (PC) of the descriptor space. Briefly, PCs are created by transforming the coordinate axes of the system so that maximum variance of the data can be explained by PC1, the next largest variance by PC2, which is orthogonal to PC1, and so on. The 2-dimensional example of PCR is shown in Fig. 9e. In *n*-dimensional space, the result will be *n* principal components. However, usually a much smaller amount of PCs will suffice to achieve an adequate representation of the dependent variable. This is due to the collinearity of parameters i.e. multiple parameters describe more or less the same effect and are thus explained on the same PC. Consequently, the last PCs contain mostly noise. On the other hand, some of the parameters of the model may be irrelevant for the current subset of the data and thus they are without any explanatory value. Thus, a threshold value is used to remove spurious descriptors of the model. The ability to reduce the dimensionality of the model is the reason why PCR is so valuable as a method for handling large number of descriptors. However, the approach is not without its drawbacks, as it has been shown that occasionally the first PCs do not contribute to the observable values at all [198], and thus the used PCs need to be carefully selected [199,200].

In the 4DSPOT workflow, PCR is applied in several phases. In the first phase, the prediction is conducted using all atom classes and the result is used to remove severe outliers which would cause uncertainty in the PCR. The second phase strives to resolve the non-linearity of the model by searching correlation parameters X_iX_j and adding relevant ones to the descriptor matrix. In the third phase, PCR is applied locally for the different atom classes, using the same parameters and correlation parameters. In this phase, the remaining error from previous phases is used as a dependent variable.

The complete equation for a chemical shift calculation of 4DSPOT is shown in Fig. 9g, in which $\delta_n°$ is the basic value for the current atom type and P_i are the coefficients, solved by PCR, for the conformationally averaged parameters X_i. Similarly, P_{ij} are the coefficients for the correlation parameters X_iX_j. The local correction from the third phase is applied in the term $\Delta\delta^{LOCAL}$. Finally, $\Delta\delta^{RF}$ is the random forest correction (see next chapter).

4.3.2 Random forest

A decision tree is a non-linear regression or classification method, which works by splitting the data based on the given parameters in such a way that the prediction error is reduced as much as possible. Splitting the data yields two new datasets called nodes. The algorithm then continues splitting the nodes as long as the benefit is larger some predetermined given threshold values, or the number of data points in a node is less than desired. The final non-split node is called a leaf, and it contains the prediction e.g. as an average of the data points in the leaf (analogous to kNN, see next chapter). Random forest [201] is an ensemble implementation of a decision tree method. In the method, multiple decision trees are grown by randomly dividing the data into teaching and test sets, thus internally taking care of cross-validation. The prediction result is given as an average of all of the trees.

In 4DSPOT, random forest is used as the last step of the prediction protocol to estimate and correct the remaining error from PCR (Fig. 9g, correction term $\Delta\delta^{RF}$). This is based on the assumption that if one has similar sub-structures, then the prediction errors should be also be similar. In the search for these sub-structures, non-parametric methods are feasible since they can distinguish two clusters of data without deriving any continuous function between them. For example, if one considers the secondary structures for a given amino acid: it is highly possible that the prediction errors between α-helices and β-sheet differ, and even though the secondary structures can be distinguished based on torsion angles, it might be difficult to derive any linear function between those forms. Instead, a decision tree can decide in a manner that the data should be split using some torsion angle value as its cut-off. The combination of linear (PCR) and non-linear (random forest) regression has been used with success also in small molecule chemical shift prediction [189].

4.3.3 k Nearest neighbors

The Juniper method calculates the coupling constants with k Nearest Neighbor (kNN) regression [202]. Similar to the random forest, kNN does not actually derive any function between the observable and the structural parameters. Instead, it is another non-linear regression method, perhaps the most simple of those techniques. Briefly, the method searches the k closest matching data points (i.e. the nearest neighbors), by means of distances in the parameter space, and returns the prediction as an average of those values. Although the idea is simple, the appropriate use of kNN usually requires parameter scaling and result weighting in order to obtain reasonable results. In addition, too high dimensionality is a known problem in kNN, as it will effectively collapse the distances to resemble each other [197,203]. In Juniper, kNN was found to be a practical method for the regression since the database is large and structural dependencies are simpler than those affecting chemical shifts, thus it is possible to model with a fewer number of descriptors. Moreover, low dimensionality is achieved by effectively splitting the data into sufficiently small subsets with hash codes, which also speeds up the calculation considerably.

The equation and illustrative example of the use of kNN to calculate the coupling constant in the Juniper method are shown in Fig. 9f. In the equation, J_i is the observed coupling values in the database and r_i represents the Euclidean distances to query coupling in the parameter space. The final value for the prediction is the average of J_i, weighted by the inverses of r_i.

4.4 PROGRAMMING

The 4DSPOT chemical shift predictor is based on the predictor code of the previous small molecule chemical shift predictor [189], programmed in FORTRAN language. The program reads in the geometric parameters of the molecule, creates the actual chemical shifts descriptors and applies the prediction models (PCR and random forest) to calculate the chemical shifts. In the development stage, the prediction model building is undertaken in program 4DSMOB, also programmed in FORTRAN.

The molecular modelling framework for both presented prediction programs is based the code of Molecular Modelling System (MMS) of PERCH NMR Software [188], programmed in C++ language. MMS contains many functions for molecule input and analysis, thus greatly facilitating the development of the current predictors. In the 4DSPOT program (Papers I and II), the MMS code has been modified to handle protein molecules; the resulting program is called 4DLIB. In 4DSPOT workflow, 4DLIB is used as an external module to input the protein structure and the MD simulation trajectory and to output the geometric parameter file.

The J-coupling predictor Juniper (Paper III) is programmed directly on top of the official MMS program in C++ language and thus available in the PERCH NMR Software package. It is also available as a web server at www.perchsolutions.com/juniper.html. In addition, both projects have involved a considerable amount of scripting, done in FORTRAN, Python, Perl and R languages.

5. Results and discussion

This chapter presents the main findings and advances of this thesis. In addition, the 4DSPOT program and some unpublished results considering protein chemical shift prediction are presented.

5.1 4D PREDICTION OF PROTEIN CHEMICAL SHIFTS

5.1.1 Effect of dynamics in protein chemical shift prediction

In Paper I, the groundwork for 4D chemical shift prediction was laid by introducing the approach to predict protein ^{1}H shifts with conformationally averaged structural parameters (i.e. the chemical shift descriptors). This is not to be confused with averaging the structure itself, which would lead to unrealistic local structures. As compared with ^{1}Hα and ^{1}HN shift prediction of single conformations, the improvement gained with the 4D approach was about 6-7 %, although for certain proteins it was up to 28 %. Already the initial version was found to be sensitive enough to detect structural errors (Appendix IV).

The main concept in Paper II was to extend the conformational space explored in Paper I by introducing more conformational freedom from NMR ensembles. In contrast to 100 ps MD simulations, which are only capable of mapping local fluctuations, the conformations of NMR ensembles implicitly represent the longer time-scale motions of a protein in solution, e.g. side chain rotation and random coil movement. Since the averaging over NMR ensemble conformations was already shown to achieve about 9 % improvement in chemical shift prediction accuracy [192], and since the approach presented in Paper I was also successful, it was expected that by combining the approaches, it should be possible to further improve the prediction. Indeed, this was the case, reflected on average as 13 % lower RMS errors (Table 3) of the combined NMR ensemble and MD (NMRE+MD) model, compared with the non-dynamic model. As $\Delta\delta_{obs}$ vs. $\Delta\delta_{pred}$ R correlation coefficients this is, e.g. for the poorly predicted ^{1}HN nuclei, a notable increase from 0.61 to 0.72. Paper II also introduced the ^{13}C and ^{15}N predictions of 4DSPOT. Furthermore, the random forest correction was established, yielding additional 2-4 % decrease of backbone nuclei RMS errors (Table 3).

In principle, it is possible to exploit molecular dynamics in chemical shift prediction also by performing multiple predictions e.g. for the snapshots of a MD trajectory and then averaging the prediction results. This approach has been tested in several studies [192,204,205]. Although the approach works, it has a drawback: in those shift prediction methods that parameterize the chemical shift descriptors from static (usually X-ray) structures, the conformational flexibility is already implicitly accounted in the descriptors. This leads to accounting the dynamics twice in the above approach [106]. In addition, the dynamic improvement concerns the query protein only and not the prediction model. Both approaches (chemical shift descriptor averaging vs. prediction result averaging) were evaluated with 4DSPOT in Paper II. The outcome was clear, achieving 18 % smaller errors in the fully dynamic approach (structural parameters of the query protein and prediction model proteins both averaged). Later, the approach postulated in Papers I and II was successfully used also in the PPM predictor [106], where considerably longer MD simulations were performed for 35 X-ray protein structures. The prediction of the whole conformational ensemble in one run should be useful also in evaluating and restraining the replica-averaged MD simulations [82,84,85].

Table 3. Chemical shift prediction RMS errors (ppm) for different prediction models of 4DSPOT.

Teaching set	Test set	^{1}Hα	^{1}HN	^{1}H(sc)	^{13}Cα	^{13}CO	^{13}Cβ	^{13}C(sc)	^{15}N[b]
ND	n/a[a]	0.29	0.50	0.29	1.12	1.22	1.22	1.04	2.75
NMRE	n/a[a]	0.27	0.48	0.28	1.05	1.19	1.13	1.02	2.58
MD	n/a[a]	0.27	0.47	0.28	1.10	1.19	1.17	1.03	2.67
NMRE+MD	n/a[a]	0.24	0.43	0.26	0.97	1.14	1.03	1.00	2.41
NMRE+MD (NW)[c]	n/a[a]	0.24	0.43	0.26	0.98	1.14	1.03	1.00	2.41
NMRE+MD (NORF)[d]	n/a[a]	0.25	0.44	0.26	1.00	1.15	1.06	1.00	2.50
ND2014	ND	0.28	0.48	0.28	1.07	1.14	1.18	1.01	2.61
NMRE2014	NMRE	0.27	0.46	0.27	1.01	1.11	1.09	0.98	2.48
ND2014	n/a[a]	0.27	0.45	0.28	1.04	1.00	1.12	1.06	2.57
NMRE2014	n/a[a]	0.26	0.44	0.26	0.98	0.96	1.05	1.04	2.46
GB 100ps	n/a[a]	0.28	0.48	0.28	1.12	1.22	1.15	1.04	2.74
GB 1ns	n/a[a]	0.28	0.47	0.29	1.14	1.22	1.16	1.05	2.73

ND = non-dynamic model
NMRE = NMR ensemble model
MD = Molecular dynamics model
NMRE+MD = combined NMRE and MD model
GB = Generalized Born implicit solvent MD model
[a] Leave-One-Out cross validation
[b] Backbone only
[c]"No-water" model, without descriptors derived from explicit solvent. Compare to "NMRE+MD".
[d] Without random forest correction ($\Delta\delta^{RF}$). Compare to "NMRE+MD".

5.1.2 Other notable findings

In order to reach the best prediction accuracy, the majority of empirical protein chemical shift predictors have relied on high-quality X-ray derived structures. However, this approach now seems to be facing a resolution barrier [46], i.e. the quality of X-ray structures has reached its limits and in chemical shift prediction, no further improvement can be achieved. Another concern is the difference between solution and crystal structures, which has been reported, by the means of backbone root mean square distance (RMSD), to be on average 1.4 Å [206] or, in another similar study, to vary from 1.5 to 2.5 Å [207]. Even more importantly, significant differences were noted in hydrogen bond contacts, which have obvious implications on the structural studies [208]. For these reasons, 4DSPOT has relied on NMR derived structures from its very onset.

Commonly, X-ray structures are regarded as being more accurate than NMR structures [209], and they can achieve atomic-scale resolution (1.0 Å). However, it is not straightforward to compare the quality of NMR and X-ray structures. Whereas the resolution of X-ray structures can be derived experimentally, no such statistic is directly available for the NMR structures. Therefore, several statistical measures have been developed, such as the equivalent resolution (e-resolution) [209]. Although the study shows the average e-resolution is still better for X-ray structures, it also reveals that the quality of NMR structures is constantly improving. This further promotes the use of NMR ensembles as teaching data. Paper I compared the predictions of several proteins that had both X-ray and NMR structures available, and confirmed that the results for X-ray structures showed better (6 % smaller RMS error) prediction results, evidently due to their more accurate local structures. However, Paper II indicated that by using NMR structures, which enable ensemble averaging, at least as good chemical shift prediction results are achieved. In addition, consistency is gained when both the observed chemical shifts and the protein structures originate from the same study. Due to these benefits, one can suggest that there is no reason to avoid using NMR structures as teaching data, although the structural accuracy may be slightly worse.

The plots of prediction errors versus the observed secondary chemical shifts are strongly biased (Fig. 10). Prediction error is not uniform but instead correlates with the secondary shift i.e. errors are larger at the extreme ends of the observed chemical shift distribution. In other words, the 4DSPOT predictions do not cover the whole range of the secondary shift. This problem has been observed also in other studies [50,115,132] and can be expected to be uniform for all empirical predictors. The correlation is especially strong for ^{1}HN shifts prone to sensitive through-space effects such as hydrogen bonding and aromatic ring currents. This gives rise to the doubt that the currently available protein models are not realistic enough to allow modelling of the finest interactions required for reproduction of the extreme values of chemical shifts. This is further evidenced since there is a slight reduction in the bias when the more realistic four-dimensional protein models are used.

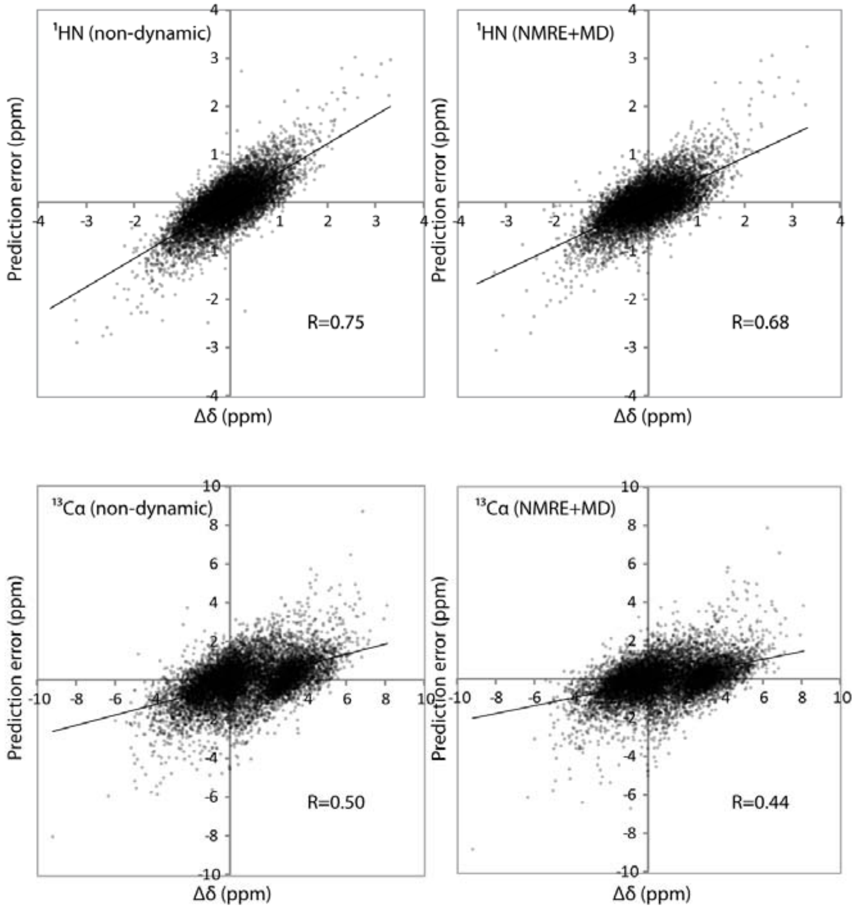


Figure 10. ^{1}HN and ^{13}Cα prediction errors of 4DSPOT as a function of the observed secondary shift (Δδ) using the non-dynamic and the four-dimensional NMRE+MD models.

Recent benchmarks have revealed the improvement of force fields as validated against NMR observables [210,211]. During the 4DSPOT project, several different force fields were tested in MD simulations in order to improve the representation of protein structures and dynamics. In contrast to the more traditional force fields, which have performed well in the abovementioned benchmarks, such as the torsion angle refined successors [212,213] of the ff99SB force field [195], our attempts were to utilize the polarizable force fields, such as ff02 [214], and ff02EP that also includes electron pairs [215]. However, no significant improvement in prediction results was seen, probably due to the very short MD runs used. Regardless, the poor accuracy e.g. of the ^{1}HN and ^{15}N chemical shift predictions clearly shows that the representation of the hydrogen bond contacts must be incorrect, and thus, further improvement could be attained by employing force fields that use additional features to account for the electrostatic effects. One particularly interesting such force field is AMOEBA [216,217], which uses quadrupoles to model the polarization effects. However, building a chemical shift prediction model with AMOEBA would be computationally demanding, and thus it was left for the future.

5.1.3 The prediction program 4DSPOT

In addition to the results presented in the previous chapter, another outcome of the project is the protein chemical shift prediction program 4DSPOT, which is freely available for academic use. The first version of the program was published June 2009. After the version 1.1 (appearing at the time when Paper II was published), several improvements have been made, with the most important of these being the support for multiple protein chains, enabling the chemical shift prediction of protein complexes. Along with some other minor changes, program version 1.2 was published in January 2014. The program packages and example files can be downloaded from 4DSPOT web site at www.uef.fi/4dspot. Both Windows and Linux versions are available.

On top of the actual shift prediction, the 4DSPOT program offers several supporting features e.g. the possibility to input observed chemical shifts in BMRB or VASCO [144] formats and print statistics about the observed-predicted difference. The shift reference corrections can also be applied. The program is also capable of adding missing hydrogens and removing ions and ligands. When dynamic prediction models are used, it is recommended to follow the same MD methodology as has been used to build the 4DSPOT teaching data; otherwise the published prediction accuracy cannot be guaranteed. Therefore, the 4DSPOT package contains also the tools to prepare the PDB files for the MD simulations, and the scripts for the AMBER molecular dynamics program [194] to run the simulations.

There is some usage of 4DSPOT reported in the literature. Mainly, it has been compared with other predictors [65,106]. In the study of Kannan et al. [84] 4DSPOT was used to evaluate the conformational ensembles refined with methyl and backbone chemical shift restraints. Since the study dealt with side chain shifts and whole NMR ensembles, 4DSPOT fitted for the purpose especially well, and it seems to be reasonable to promote the further use of 4DSPOT in similar studies.

5.1.4 New prediction models

After Paper II, several new prediction models have been tested. Due to public requests, dynamic prediction models not requiring explicit solvent molecules (the so-called “no-water” models) have been added to the package. This is not to be confused with implicit solvation models: instead, the “no-water” models refer to explicit solvent MD simulations from which the solvent molecules are removed afterwards. It was noted that the use of explicit solvent based chemical shift descriptors did not provide any extra benefits to the prediction results of MD and NMRE+MD models (Table 3), but they did require longer calculation times to create. Thus, the “no-water” models are now suggested for default use.

The VASCO database [144] holds validated and reference corrected chemical shifts, paired with corresponding structures, for almost 5000 proteins [144]. In the 4DSPOT version 1.2, this wealth of data was used to build new non-dynamic and NMRE models (called ND2014 and NMRE2014, respectively) using a four times larger dataset than before (398 proteins). As this model has not been published in any paper, it is not included in the 4DSPOT package but can be downloaded from 4DSPOT web site. The model has been tested with the same methods as those models in Paper II and cross-tested with the original non-dynamic and NMRE models. In comparison with the original models, the new models achieved 5-10 % improvement in prediction accuracy (Table 3; for more details see the report in 4DSPOT web site). In particular, the improvement of ^{13}CO prediction is rather notable and about twice as large as for the other nuclei. It is not likely that this improvement originates from the better representation of ^{13}CO nuclei in the protein structures; instead, it is thought to arise from improved chemical shift measurement and referencing. In addition, it was noted that the chemical shift re-referencing approaches LACS [145], used in Papers I and II, and VASCO are not fully compatible with each other. Again, this was seen especially in the ^{13}CO shifts, which have known referencing issues [45]. Therefore, when the original data was used as a test set for the 2014 models, the observed shifts were imported from the VASCO archive.

It has been proposed that the correlated motions are revealed better in implicit solvent MD simulations [190]. The studies regarding correlated motions of proteins are ongoing in our research group, and it was desired to further analyze the possible correlations between correlated motions and chemical shift prediction results. Therefore, prediction models using Generalized Born implicit solvent modeling in MD simulations (100 ps and 1 ns) were built and tested. However, the chemical shift prediction accuracy was not as good as with the explicit models (Table 3). The situation was worst in the ^{13}Cα nuclei results, which were not any better than those of non-dynamic models, and the longer (1 ps) simulation impaired the results even more. This suggests that the protein backbone is not stable in these implicit solvent MD simulations.

5.2 CHEMICAL SHIFT PREDICTION MODELS FROM QUANTUM MECHANICAL CALCULATIONS

Building reliable empirical chemical shift prediction models requires a rather large amount of high quality data, at least several hundreds of data points for each atom type. Therefore, despite the continuously growing databases, there are many situations where the available structures and/or experimental chemical shift measurements are too sparse to allow the generation of empirical prediction models. These situations include e.g. amino acids with post-translational modifications, the effects of small molecule ligands, and nucleic acids with a lack of high-resolution structures. In these situations, the only option is to use quantum mechanical methods to predict the chemical shifts *ab initio*. Especially interesting is to utilize the indirect approach (see chapter 2.1.3), where prediction models are built in a similar manner as any empirical approach, but using QM calculated shifts as teaching data. This approach would overcome the limitations posed by the computational cost of QM, crucial to applications requiring recurrent predictions such as ROSETTA-based structure generation [110,111,151] and chemical shifts restrained MD [81]. In addition, it might be that in some cases, such as the ^{1}HN shifts, experimental data will never be accurate enough due to conformational averaging, and thus the single-point QM calculations are the only way to model the sensitive effects. In collaboration with the group of Prof. Vendruscolo in the University of Cambridge, two such projects were initiated. The backgrounds and preliminary results of these projects are presented in this chapter.

5.2.1 QM-derived chemical shift model for methionine

The methionine methyl, being available with a ^{13}C isotope label and with relaxational properties resulting in narrow line widths [218], has been proven to be a useful probe for elucidating protein structure and dynamics [219–222]. However, the chemical shifts of the methionine methyl are not adequately predicted, or not predicted at all, in most of the protein chemical shift prediction programs, preventing the use of applications based on chemical shift prediction. Therefore, a DFT-calculated model for methionine methyl was proposed.

The chemical shifts of methionine were quantum mechanically calculated in two different model systems. First, a two molecule system was used, always containing one methionine residue with neutral N-methyl (NMe) and acetyl (Ac) caps. In an attempt to model the through-space effects, another amino acid residue, again with neutral caps, was added in close proximity in a random orientation (Fig. 11a). The torsional conformations of the amino acids were derived from experimental structures of the Dunbrack rotamer library [223] but their spatial orientation against each other was randomly sampled. Second, in order to capture also the close contact effects of neighboring residues, a set of dipeptide models were built (Fig. 11b). Sequences of Ac-Met-X-NMe and Ac-X-Met-NMe, where X is one of the 20 natural amino acids, were used in order to capture both the preceding and following residue effects for the methionine. For a total of 20 000 model systems, the chemical shifts were calculated with DFT. These results were then used as the teaching data of the Camshift [104] and CH3Shift [142] predictors.

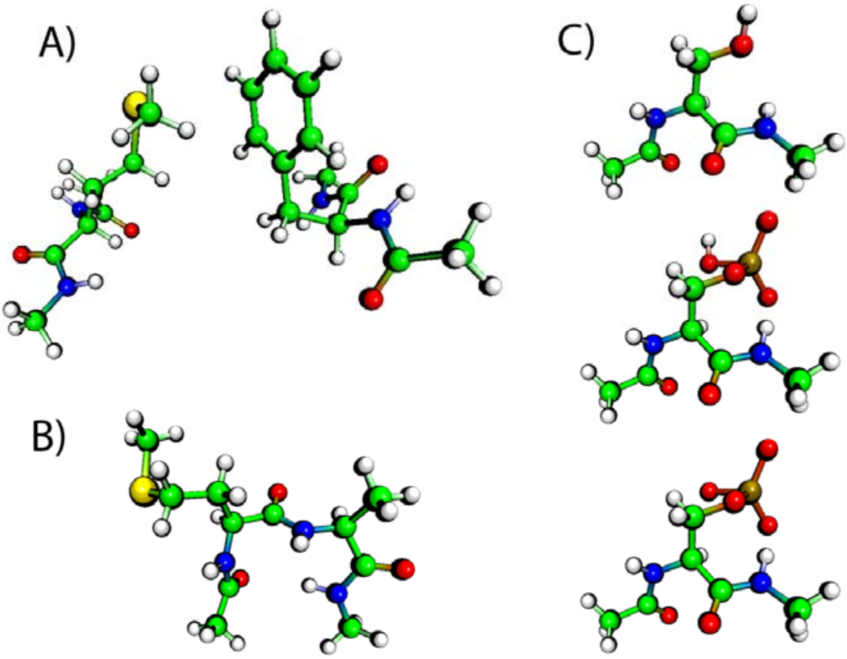


Figure 11. Model systems used in DFT calculations of the projects deriving chemical shift models from ab initio data. a) model for through-space effects for methionine (here with a neighboring Phe residue) b) model for neighboring residue effects for methionine (here with following Ala residue) c) the models systems for phosphoserine phosphorylation effect (Ser and -1 and -2 ionic forms of pSer in same torsional conformation)

The preliminary results of the project were both promising and discouraging at the same time. Using the full set of descriptors (distances, torsion angles, H-bonding, aromatic ring currents), CamShift was able to reproduce the DFT chemical shifts with good accuracy (^{1}Hα RMS=0.12 ppm, Fig. 12a). However, when evaluated against the experimental data, the ^{13}Cα prediction worked well but the ^{1}Hα, ^{1}Hε and ^{13}Cε predictions were poor (Fig. 12b). Several explanations were proposed. First of all, ^{13}Ca shifts can be mainly described by torsion angles and the atom moiety is stable enough to be modelled without dynamics; this has been previously shown by Vila et al. [119]. On the other hand, the methyl resonances ^{1}Hε and ^{13}Cε, located at the end of the side chain, are sampled by the three side-chain torsion angles, making the nuclei very prone to dynamics. For these reasons, even the ^{13}Cε model did not reveal any correlations with the experimental data, although it was possible to explain about 90 % of the DFT shift variance by the χ^3 torsion angle only (using Eureqa model [224]). In addition, the proposed two molecule system might yet be too sparse to sufficiently model the through-space effects for ^{1}H nuclei, solvation effects are not accounted for, and the neutral caps can introduce some additional uncertainty.

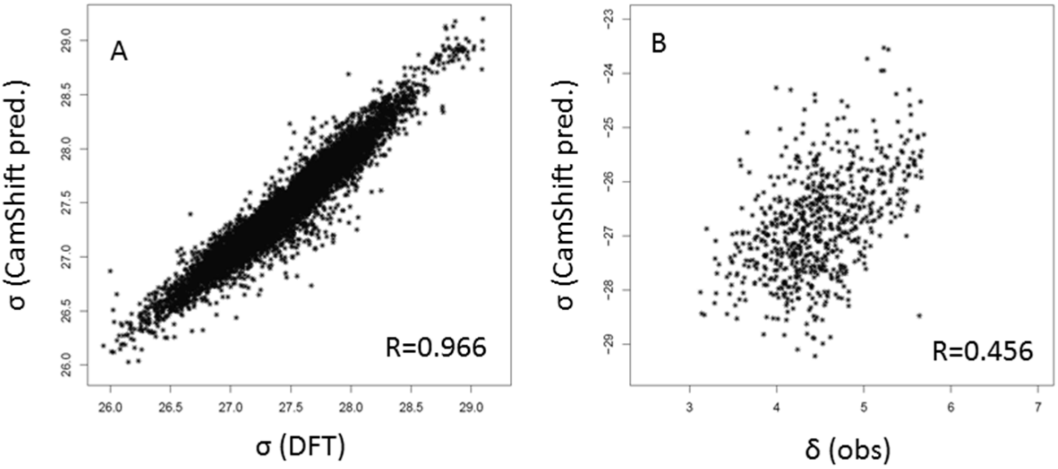


Figure 12. CamShift results of DFT-based chemical shift prediction model for methionine ^{1}Hα. A) vs. DFT calculated shielding constants of model systems B) vs. experimental protein chemical shifts.

Additionally, there was not enough experimental methyl shifts to permit the validation of the proposed model. On the other hand, the study of Christensen [50] has recently shown that although the DFT-based model cannot reproduce the experimental data, it can still be used for structure refinement, an encouraging viewpoint suggesting that this project could be revisited in the future. At the very least, the calculated QM database could be used to develop a method for chemical shift based conformation analysis of the methionine residue in a similar manner as done earlier for other side chains [64,95,96].

5.2.1 QM-derived chemical shift model for phosphorylated amino acids

Another QM-based chemical shift prediction project was targeted for phosphorylated amino acids. Even though phosphorylation is probably the most common post-translational modification [225,226], none of the current predictors is able to predict chemical shifts of phosphorylated amino acids (pAA); this is again due to the shortage of experimental chemical shift data. In an attempt to enable the use of the pAA's in the chemical shift based applications, the correlations between phosphorylated and non-phosphorylated serine, threonine and tyrosine residues were mapped with DFT calculations. This was done in model systems of neutrally capped amino acid residues in several hundreds of different conformations. For each torsional conformation, both ionic forms of the phosphorylated amino acid, as well as the non-phosphorylated version, were built (Fig. 11c). The outcome of this approach would be a set of equations of phosphorylation effect on chemical shifts as a function of torsion angles, although this is not directly comparable with the effect seen in reality, which will always include some conformational changes. Instead, these equations are intended to enable the use of any chemical shift predictor to predict pAA chemical shifts, by predicting the chemical shifts of a corresponding standard amino acid in the same torsional conformation and then by adding the effect of phosphorylation into the results.

The equations for phosphorylation effects were sought using the Eureqa program [224]. Again, the lack of experimental data, which was even more severe than encountered with the methionine project, made the validation of the model almost impossible. Plausible correlations (R from 0.55 to 0.74 for different predictors) against the experimental data of phosphoserine (pSer) ^{1}HN shifts was seen, but on the other hand for many other nuclei, the phosphorylation effect was too large in the DFT calculations. There are several possible reasons for these issues. First, the implicit solvent model used in these calculations may be inadequate, as it has been shown that explicit solvent modelling confers clear benefits on the QM calculations [127,128], and the phosphorus group naturally has many solvent contacts. Furthermore, phosphorus can also make hydrogen bonds to neighboring residues, and the single-residue model is not able to encapsulate these effects. Nonetheless, with the development of a more extensive model system and more experimental data for validation, the proposed approach could still work in the future.

5.3 UNIVERSAL J-COUPLING PREDICTION

In the design of the J-Coupling prediction method Juniper, three main points were emphasized due to its intended use as a part of an automatic spectrum analysis method. These points were i) coverage, ii) maintainability and iii) speed. All these requirements were addressed with the lightweight and general database search method, in which the coupling path information was encoded into hash codes (see chapter 4.2). The outcome was rather successful. First, the main advantage is that the presented method can predict all types of coupling constants within the same framework. This is accomplished by separating the different cases with hash codes and creating the prediction parameters in such a way that coupling paths of all lengths have some general parameters regardless of the elements of the coupling atoms. The ease of maintainability of the method also results from this approach. Should missing or poorly predicted couplings be encountered, they can often be repaired simply by incorporating new data into the database without any extra modifications to the prediction parameters. Ultimately, the method is fast and the prediction is obtained within a few seconds for typical small molecules. This is due to the hash code based data classification that allows the database to be sought with the in-memory binary search algorithm. Moreover, it keeps the number of data points in each class sufficiently low, keeping the kNN algorithm fast.

The prediction accuracy of Juniper was also shown to be satisfactory. In internal tests, using Leave-One-Out validation and non-redundant test set, the total RMS errors were 0.58 and 1.02 Hz for J_{HH} and $J_{PH/FH}$ couplings, respectively (Fig. 13). Generally, errors less than 1 Hz are tolerable in ACA use due to the iterative nature of the method. ^{13}C heterocouplings ($J_{PC/PF}$) can also be predicted but the data is still too sparse to make any realistic evaluations of the accuracy; however, the method works fine as a database lookup for those cases (Fig. 13). Compared with other empirical approaches SPINUS [175] and the commercial ACD/Labs NMR predictor [172], using a test set from the study of Bally and Rablen [11], Juniper was found to be at least as accurate as the other two. However, for the same test set, the quantum mechanical calculations of coupling constants are extremely accurate [11]. Unfortunately, it is still not feasible to undertake the QM calculations for high-throughput spectrum analysis purposes due to the computational cost. Instead, it should be possible to use the QM calculated coupling constants as teaching data for Juniper (in analogy with chapter 5.2), and in this way to enable a rapid retrieval of the accurate QM calculated couplings.

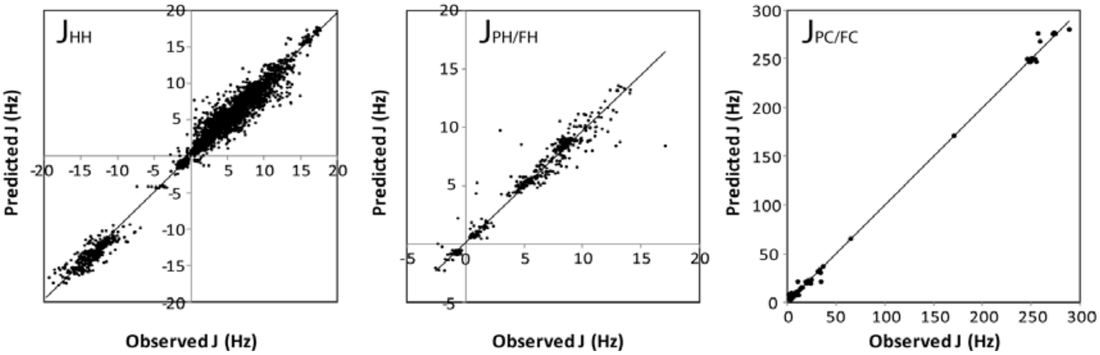


Figure 13. Prediction scatter plots of different J-coupling classes from internal Leave-one-out validation tests of Juniper database.

Juniper has been a part of PERCH NMR Software since the version 2013.1. In practice, it has been shown to improve the throughput of the method, even though in some areas it might not be as accurate as the preceding equation based model. For example, one particular flaw is the prediction of $^3J_{HH}$ couplings over rotatable bonds from a single conformation. Compared with the previous approach that utilized Haasnoot equation [160] averaged over multiple Monte Carlo / MD conformations, this may seem unreasonable. However, when considering ACA throughput, the possible loss of accuracy is clearly compensated by the improved coverage: the iterative approach usually fails if the spectrum contains visible couplings that are not predicted. In addition, the improved prediction accuracy helps by decreasing the number of possible assignments and thus increasing the probability of finding the correct solution.

6 Conclusions and future perspectives

In this dissertation, two different NMR parameter prediction approaches were developed for two different study problems. Both predictors have been shown to successfully model the NMR parameters as a function of the molecular structure. In addition, the two computer programs 4DSPOT and Juniper can be considered as other outcomes of these projects, hopefully serving other scientists working in this field.

The 4DSPOT project was the first study to reveal that protein chemical shifts are better parameterized with ensemble averaged descriptors. At that time, other prediction approaches were using single conformation protein models. Later, the 4DSPOT approach was adopted also in the PPM predictor [106]. The inclusion of dynamics is both a pro and a con, i.e. even though it significantly improves the prediction results and offers a natural solution for accounting for the dynamics nature of proteins, it also adds the uncertainty of the MD simulations (MD contribution to the total variance, S^2_{MD}, is from 0.05 to 0.1 ppm for individual ^1H shifts) to the results. Furthermore, it also complicates the prediction procedure by demanding that dynamic proteins are used as a query, which increases the computational cost. This could be one reason why 4DSPOT has not became a very widely used method in the field of protein NMR even though the non-dynamic model is available in the package too.

Generally, the 4DSPOT project has shown that most of the current protein models are not realistic enough to permit an accurate representation of their chemical shifts. This is reflected by the results of the best proteins in our database, for which the RMS errors are about 50 % smaller than the database averages (0.10, 0.21, 0.56, 0.47, 0.62 and 1.48 ppm for $^1H\alpha$, ^1HN, $^{13}C\alpha$, $^{13}C\beta$, ^{13}CO and backbone ^{15}N chemical shifts, respectively). Furthermore, recent preliminary results have revealed that the mean of coordinate fluctuation of $^{13}C\alpha$ atoms correlates ($R \approx 0.5$) with the mean of prediction error, i.e. the prediction is better for rigid proteins, which again underlines the importance of proper mapping of dynamics.

At present, 4DSPOT is the only predictor in which the teaching data consist solely of NMR derived structures. The success of developing such a predictor, given that the NMR ensembles do not follow Boltzmann statistics, should emphasize the vastly improved quality of NMR structures. Hopefully, dynamically correct ensembles will populate the databases in the future and open new possibilities to improve this presented approach. Overall, the 4DSPOT project can be considered to contribute to the ongoing intense discussion (acknowledged e.g. by the 2013 Nobel Prize in chemistry) about the importance of protein dynamics for understanding how nature works.

So far, reinforcing the structure-based chemical shift prediction with a sequence-based method [55] has been the only successful way to push the prediction accuracy over the above-mentioned resolution barrier. Unfortunately, this gives no extra benefit in most structural studies, in which different conformations of the same protein are compared to each other. On the other hand, although the dynamic approach of 4DSPOT offers some help to overcome the resolution barrier, it is probable that it is also facing a force field barrier. For example, the commonly used protein force fields do not contain parameters for the N-H…O=C hydrogen bond angle (known since the Nobel-awarded studies of Pauling!), evidently contributing to unsatisfactory ^1HN prediction results ($\Delta\delta_{obs}$ vs. $\Delta\delta_{pred}$ R is only 0.72 even for our best NMRE+MD model). More issues can be associated with fixed atomic charges of non-polarizable force fields, incorrect protonation states and improper modelling of aromatic ring stacking. Consequently, dynamic chemical shift prediction approaches possess much room for future improvement via better force fields, whereas the further development of non-dynamic X-ray structure based predictors is discouraged by the fact that the resolution of X-ray models is approaching its limit.

Besides the hunt for better force fields, the future of protein chemical shift prediction is expected to move towards quantum mechanics, which can provide explicit correlations between the shifts and the structure. As the QM calculations link non-averaged observables to exactly known structures, taking account of the dynamics by simply averaging the prediction results of different conformations becomes justified. Another reason to use QM-calculated chemical shifts as teaching data is the lack of experimental data in many cases. For example, this could lead to prediction of ligand binding effects on protein chemical shifts, consequently enabling the use of chemical shift prediction based applications in protein-ligand interaction studies.

The Juniper project established a general framework for J-coupling prediction. The development of Juniper has not stopped at the stage presented in Paper III. First of all, the data flow between Juniper and ACA is an iterative process and thus the Juniper database is being constantly extended. As new spectra are analyzed with ACA, Juniper will gain more teaching data to cover even more cases. Consequently, the throughput of ACA will further improve, and so forth. Moreover, since the framework can deal with any desired coupling constants, it is expected than more coupling classes will be covered in the future. The J_{CH} couplings, required e.g. for HSQC analysis, will be the next class to be added.

In contrast to the 4DSPOT project, which was mainly driven by academic interest, the Juniper project was more focused on building a tool for certain use. Therefore, some features, such as the prediction of couplings over rotatable bonds (see chapter 5.3), are perhaps not as sophisticated as possible. Nonetheless, the prediction accuracy of the data driven approach has been found to be sufficient for ACA use. As expected, most issues associated with Juniper in ACA use were related to the missing predictions for certain couplings. These issues are now readily fixed in Juniper simply by introducing more data.

As with chemical shift prediction, the use of QM calculated data in model teaching is an option also in Juniper. For example, when initiating studies of novel compounds with a certain substructure, one could complement the predictor with QM calculated coupling data of several representative examples of the system. There has also been interest to use Juniper for protein J-coupling prediction, for which the framework should also be suitable via incorporating it to 4DSPOT. In the case of proteins, it might be feasible to use only QM-calculated coupling constants as teaching data, which would enable the results to be averaged over multiple conformations. This should improve the prediction of couplings over rotating bonds, which are the most important ones in protein studies.

The field of NMR parameter prediction exists in two worlds. The first one is the kingdom bordered by the NMR tube, i.e. the world of experimental NMR, which offers accurate observations of the parameters but is limited to the accuracy of the molecular models. The second world is the boundless *in silico* realm, in which the QM-calculated NMR parameters are deterministic and always consistent with the molecular structure, but not necessarily transferrable to the experimental world. As the search for more accurate correlations between structure and NMR parameters continues, uniting these two worlds is a most captivating future goal.

7 References

1. Günther H (1995) NMR Spectroscopy: Basic Principles, Concepts, and Applications in Chemistry, 2nd Edition. Chichester, UK: John Wiley & Sons, Ltd.

2. McConnell HM (1957) Theory of Nuclear Macnetic Shieldin in Molecules. I. Long-Range Dipolar Shielding of Protons. J Chem Phys 27: 226–229.

3. Pople JA (1956) Proton Magnetic Resonance of Hydrocarbons. J Chem Phys 24: 1111.

4. Pople JA (1958) Molecular orbital theory of aromatic ring currents. Mol Phys 1: 175–180.

5. Haigh CW, Mallion RB (1979) Ring current theories in nuclear magnetic resonance. Prog Nucl Magn Reson Spectrosc 13: 303–344.

6. Buckingham AD (1960) Chemical Shifts in the Nuclear Magnetic Resonance Spectra of Molecules Containing Polar Groups. Can J Chem 38: 300–307.

7. Markley JL, Bax A, Arata Y, Hilbers CW, Kaptein R, Sykes BD, Wright PE, Wüthrich K (1998) Recommendations for the Presentation of NMR Structures of Proteins and Nucleic Acids. Pure Appl Chem 70: 117–142.

8. Ramsey NF (1953) Electron Coupled Interactions between Nuclear Spins in Molecules. Phys Rewiev 91: 303–307.

9. Cremer D, Gräfenstein J (2007) Calculation and analysis of NMR spin-spin coupling constants. Phys Chem Chem Phys 9: 2791–2816.

10. Helgaker T, Jaszuński M, Pecul M (2008) The quantum-chemical calculation of NMR indirect spin–spin coupling constants. Prog Nucl Magn Reson Spectrosc 53: 249–268.

11. Bally T, Rablen PR (2011) Quantum-chemical simulation of 1H NMR spectra. 2. Comparison of DFT-based procedures for computing proton-proton coupling constants in organic molecules. J Org Chem 76: 4818–4830.

12. Laatikainen R, Tiainen M, Korhonen S-P, Niemitz M (2011) Computerized Analysis of High-resolution Solution-state Spectra. Encyclopedia of Magnetic Resonance. Chichester, UK: John Wiley & Sons, Ltd.

13. Laatikainen R, Niemitz M, Weber U, Sundelin J, Hassinen T, Vepsäläinen J (1996) General Strategies for Total-Lineshape-Type Spectral Analysis of NMR Spectra Using Integral-Transform Iterator. J Magn Reson Ser A 10: 1–10.

14. McDonald CC, Phillips WD (1967) Manifestations of the tertiary structures of proteins in high-frequency nuclear magnetic resonance. J Am Chem Soc 89: 6332–6341.

15. Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, Livny M, Mading S, Maziuk D, Miller Z, Nakatani E, Schulte CF, Tolmie DE, Wenger RK, Yao H, Markley JL (2008) BioMagResBank. Nucleic Acids Res 36: D402–408.

16. Wishart DS, Watson MS, Boyko RF, Sykes BD (1997) Automated 1H and 13C chemical shift prediction using the BioMagResBank. J Biomol NMR 10: 329–336.

17. Potts BC, Chazin WJ (1998) Chemical shift homology in proteins. J Biomol NMR 11: 45–57.

18. Gronwald W, Willard L, Jellard T, Boyko RF, Rajarathnam K, Wishart DS, Sönnichsen FD, Sykes BD (1998) CAMRA: chemical shift based computer aided protein NMR assignments. J Biomol NMR 12: 395–405.

19. Tjandra N, Bax A (1997) Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium. Science 278: 1111–1114.

20. Pervushin K, Riek R, Wider G, Wüthrich K (1997) Attenuated T2 relaxation by mutual cancellation of dipole-dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. Proc Natl Acad Sci U S A 94: 12366–12371.

21. Ni QZ, Daviso E, Can T V, Markhasin E, Jawla SK, Swager TM, Temkin RJ, Herzfeld J, Griffin RG (2013) High frequency dynamic nuclear polarization. Acc Chem Res 46: 1933–1941.

22. Goldbourt A (2013) Biomolecular magic-angle spinning solid-state NMR: recent methods and applications. Curr Opin Biotechnol 24: 705–715.

23. Ladizhansky V (2014) Recent Advances in Magic-Angle Spinning Solid-State NMR of Proteins. Isr J Chem 54: 86–103.

24. Maslennikov I, Choe S (2013) Advances in NMR structures of integral membrane proteins. Curr Opin Struct Biol 23: 555–562.

25. Murray DT, Das N, Cross TA (2013) Solid State NMR Strategy for Characterizing Native Membrane Protein Structures. Acc Chem Res 46: 2172–2181.

26. Cala O, Guillière F, Krimm I (2014) NMR-based analysis of protein-ligand interactions. Anal Bioanal Chem 406: 943–956.

27. Nováček J, Žídek L, Sklenář V (2014) Toward optimal-resolution NMR of intrinsically disordered proteins. J Magn Reson 241: 41–52.

28. Li C, Liu M (2013) Protein dynamics in living cells studied by in-cell NMR spectroscopy. FEBS Lett 587: 1008–1011.

29. Neira JL (2013) NMR as a tool to identify and characterize protein folding intermediates. Arch Biochem Biophys 531: 90–99.

30. Qu H, Fang X (2013) A brief review on the Human Encyclopedia of DNA Elements (ENCODE) project. Genomics Proteomics Bioinformatics 11: 135–141.

31. Maute RL, Dalla-Favera R, Basso K (2014) RNAs with multiple personalities. Wiley Interdiscip Rev RNA 5: 1–13.

32. Al-Hashimi HM (2013) NMR studies of nucleic acid dynamics. J Magn Reson 237: 191–204.

33. Salmon L, Yang S, Al-Hashimi HM (2014) Advances in the Determination of Nucleic Acid Conformational Ensembles. Annu Rev Phys Chem 65: 293–316.

34. Guerry P, Herrmann T (2011) Advances in automated NMR protein structure determination. Q Rev Biophys 44: 257–309.

35. Nilges M, Clore GM, Gronenborn AM (1988) Determination of three-dimensional structures of proteins from interproton distance data by dynamical simulated annealing from a random array of atoms. Circumventing problems associated with folding. FEBS Lett 239: 129–136.

36. Güntert P, Mumenthaler C, Wüthrich K (1997) Torsion angle dynamics for NMR structure calculation with the new program DYANA. J Mol Biol 273: 283–298.

37. Boomsma W, Frellsen J, Harder T, Bottaro S, Johansson KE, Tian P, Stovgaard K, Andreetta C, Olsson S, Valentin JB, Antonov LD, Christensen AS, Borg M, Jensen JH, Lindorff-Larsen K, Ferkinghoff-Borg J, Hamelryck T (2013) PHAISTOS: a framework for Markov chain Monte Carlo simulation and inference of protein structure. J Comput Chem 34: 1697–1705.

38. Li Z, Scheraga HA (1987) Monte Carlo-minimization approach to the multiple-minima problem in protein folding. Proc Natl Acad Sci U S A 84: 6611–6615.

39. Raman S, Lange OF, Rossi P, Tyka M, Wang X, Aramini J, Liu G, Ramelot TA, Eletsky A, Szyperski T, Kennedy MA, Prestegard J, Montelione GT, Baker D (2010) NMR structure determination for larger proteins using backbone-only data. Science 327: 1014–1018.

40. Schwieters CD, Kuszewski JJ, Tjandra N, Clore GM (2003) The Xplor-NIH NMR molecular structure determination package. J Magn Reson 160: 65–73.

41. Wishart DS, Bigam CG, Holm A, Hodges RS, Sykes BD (1995) 1H, 13C and 15N random coil NMR chemical shifts of the common amino acids. I. Investigations of nearest-neighbor effects. J Biomol NMR 5: 67–81.

42. Bienkiewicz EA, Lumb KJ (1999) Random-coil chemical shifts of phosphorylated amino acids. J Biomol NMR 15: 203–206.

43. Tremblay M-L, Banks AW, Rainey JK (2010) The predictive accuracy of secondary chemical shifts is more affected by protein secondary structure than solvent environment. J Biomol NMR 46: 257–270.

44. Schwarzinger S, Kroon GJ, Foss TR, Chung J, Wright PE, Dyson HJ (2001) Sequence-dependent correction of random coil NMR chemical shifts. J Am Chem Soc 123: 2970–2978.

45. Wishart DS, Case DA (2001) Use of chemical shifts in macromolecular structure determination. Methods Enzymol 338: 3–34.

46. Wishart DS (2011) Interpreting protein chemical shift data. Prog Nucl Magn Reson Spectrosc 58: 62–87.

47. Neal S, Nip AM, Zhang H, Wishart DS (2003) Rapid and accurate calculation of protein 1H, 13C and 15N chemical shifts. J Biomol NMR 26: 215–240.

48. Xu X-P, Case DA (2002) Probing multiple effects on 15N, 13C alpha, 13C beta, and 13C′ chemical shifts in peptides using density functional theory. Biopolymers 65: 408–423.

49. Moon S, Case DA (2007) A new model for chemical shifts of amide hydrogens in proteins. J Biomol NMR 38: 139–150.

50. Christensen AS, Linnet TE, Borg M, Boomsma W, Lindorff-Larsen K, Hamelryck T, Jensen JH (2013) Protein structure validation and refinement using amide proton chemical shifts derived from quantum mechanics. PLoS One 8: e84123.

51. Parker LL, Houk AR, Jensen JH (2006) Cooperative Hydrogen Bonding Effects Are Key Determinants of Backbone Amide Proton Chemical Shifts in Proteins. J Am Chem Soc 128: 9863–9872.

52. Christensen AS, Sauer SPA, Jensen JH (2011) Definitive Benchmark Study of Ring Current Effects on Amide Proton Chemical Shifts: 2078–2084.

53. Dalgarno DC, Levine BA, Williams RJP (1983) Structural information from NMR secondary chemical shifts of peptide alpha C-H protons in proteins. Biosci Rep 3: 443–452.

54. Iwadate M, Asakura T, Williamson MP (1999) Cα and Cβ Carbon-13 Chemical Shifts in Proteins From an Empirical Database. J Biomol NMR 13: 199–211.

55. Han B, Liu Y, Ginzinger SW, Wishart DS (2011) SHIFTX2: significantly improved protein chemical shift prediction. J Biomol NMR 50: 43–57.

56. Wang Y, Jardetzky O (2004) Predicting 15N chemical shifts in proteins using the preceding residue-specific individual shielding surfaces from phi, psi i-1, and chi 1 torsion angles. J Biomol NMR 28: 327–340.

57. Barfield M (2002) Structural dependencies of interresidue scalar coupling (h3)J(NC′) and donor (1)H chemical shifts in the hydrogen bonding regions of proteins. J Am Chem Soc 124: 4158–4168.

58. Kuroki S, Ando S, Ando I, Shoji A, Ozaki T, Webb GA (1990) Hydrogen-bonding Effect on 15N NMR Chemical Shifts of the Glycine Residue of Oligopeptides in the Solid state as Studied by High-Resolution Solid-Stat NMR Spectroscopy. J Mol Struct 240: 19–29.

59. Tsuchiya K, Takahashi A, Takeda N, Asakawa N, Kuroki S, Ando I, Shoji A, Ozaki T (1995) Hydrogen-bonding effect on 13C NMR chemical shifts of amino acid residue carbonyl carbons of some peptides in the crystalline state. J Mol Struct 350: 233–240.

60. Shen Y, Bax A (2010) SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. J Biomol NMR 48: 13–22.

61. Case DA (1995) Calibration of ring-current effects in proteins and nucleic acids. J Biomol NMR 6: 341–346.

62. Sahakyan AB, Vendruscolo M (2013) Analysis of the Contributions of Ring Current and Electric Field Effects to the Chemical Shifts of RNA Bases. J Phys Chem B 117: 1989–1998.

63. Moyna G, Zauhar RJ, Williams HJ, Nachman RJ, Scott a I (1998) Comparison of ring current methods for use in molecular modeling refinement of NMR derived three-dimensional structures. J Chem Inf Comput Sci 38: 702–709.

64. London RE, Wingad BD, Mueller G a (2008) Dependence of amino acid side chain 13C shifts on dihedral angle: application to conformational analysis. J Am Chem Soc 130: 11097–11105.

65. Sahakyan AB, Vranken WF, Cavalli A, Vendruscolo M (2011) Using side-chain aromatic proton chemical shifts for a quantitative analysis of protein structures. Angew Chemie (International ed.in English) 50: 9620–9623.

66. Sahakyan AB, Vranken WF, Cavalli A, Vendruscolo M (2011) Structure-based prediction of methyl chemical shifts in proteins. J Biomol NMR 50: 331–346.

67. Lipari G, Szabo A (1982) Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. Theory and range of validity. J Am Chem Soc 104: 4546–4559.

68. Akke M, Palmer AG (1996) Monitoring Macromolecular Motions on Microsecond to Millisecond Time Scales by R1p - R1 Constant Relaxation Time NMR Spectroscopy. J Am Chem Soc 118: 911–912.

69. Palmer AG, Kroenke CD, Loria JP (2001) Nuclear magnetic resonance methods for quantifying microsecond-to-millisecond motions in biological macromolecules. Methods Enzymol 339: 204–238.

70. Kleckner IR, Foster MP (2011) An introduction to NMR-based approaches for measuring protein dynamics. Biochim Biophys Acta 1814: 942–968.

71. Zeeb M, Balbach J (2004) Protein folding studied by real-time NMR spectroscopy. Methods 34: 65–74.

72. Meiler J, Prompers JJ, Peti W, Griesinger C, Brüschweiler R (2001) Model-free approach to the dynamic interpretation of residual dipolar couplings in globular proteins. J Am Chem Soc 123: 6098–6107.

73. Lakomek NA, Carlomagno T, Becker S, Griesinger C, Meiler J (2006) A thorough dynamic interpretation of residual dipolar couplings in ubiquitin. J Biomol NMR 34: 101–115.

74. Ángyán AF, Gáspári Z (2013) Ensemble-based interpretations of NMR structural data to describe protein internal dynamics. Molecules 18: 10548–10567.

75. Yuji S, Yoko O (1999) Replica exchange molecular dynamics method for protein folding. Chem Phys Lett 314: 141–151.

76. Lindorff-Larsen K, Best RB, Depristo MA, Dobson CM, Vendruscolo M (2005) Simultaneous determination of protein structure and dynamics. Nature 433: 128–132.

77. Richter B, Gsponer J, Várnai 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.

78. Fenwick RB, Esteban-Martín S, Richter B, Lee D, Walter KFA, Milovanovic D, Becker S, Lakomek NA, Griesinger C, Salvatella X (2011) Weak Long-Range Correlated Motions in a Surface Patch of Ubiquitin. J Am Chem Soc 20: 10336–10339.

79. Lange OF, Lakomek N-A, Farès C, Schröder GF, Walter KFA, Becker S, Meiler J, Grubmüller H, Griesinger C, de Groot BL (2008) Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. Science 320: 1471–1475.

80. Best RB, Vendruscolo M (2004) Determination of protein structures consistent with NMR order parameters. J Am Chem Soc 126: 8090–8091.

81. 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.

82. Kukic P, Camilloni C, Cavalli A, Vendruscolo M (2014) Determination of the Individual Roles of the Linker Residues in the Interdomain Motions of Calmodulin Using NMR Chemical Shifts. J Mol Biol 426: 1826–1838.

83. Camilloni C, Robustelli P, De Simone A, Cavalli A, Vendruscolo M (2012) Characterization of the conformational equilibrium between the two major substates of RNase A using NMR chemical shifts. J Am Chem Soc 134: 3968–3971.

84. Kannan A, Camilloni C, Sahakyan AB, Cavalli A, Vendruscolo M (2014) A conformational ensemble derived using NMR methyl chemical shifts reveals a mechanical clamping transition that gates the binding of the HU protein to DNA. J Am Chem Soc 136: 2204–2207.

85. Camilloni C, Cavalli A, Vendruscolo M (2013) Assessment of the use of NMR chemical shifts as replica-averaged structural restraints in molecular dynamics simulations to characterize the dynamics of proteins. J Phys Chem B 117: 1838–1843.

86. Fisher CK, Huang A, Stultz CM (2010) Modeling intrinsically disordered proteins with bayesian statistics. J Am Chem Soc 132: 14919–14927.

87. Jensen MR, Salmon L, Nodet G, Blackledge M (2010) Defining Conformational Ensembles of Intrinsically Disordered and Partially Folded Proteins Directly from Chemical Shifts. J Am Chem Soc 132: 1270–1272.

88. Berlin K, Castañeda CA, Schneidman-Duhovny D, Sali A, Nava-Tudela A, Fushman D (2013) Recovering a representative conformational ensemble from underdetermined macromolecular structural data. J Am Chem Soc 135: 16595–16609.

89. Choy WY, Forman-Kay JD (2001) Calculation of ensembles of structures representing the unfolded state of an SH3 domain. J Mol Biol 308: 1011–1032.

90. Huang J, Grzesiek S (2010) Ensemble calculations of unstructured proteins constrained by RDC and PRE data: a case study of urea-denatured ubiquitin. J Am Chem Soc 132: 694–705.

91. Berjanskii M, Wishart DS (2006) NMR: prediction of protein flexibility. Nat Protoc 1: 683–688.

92. Berjanskii M, Wishart D (2013) A Simple Method to Measure Protein Side-Chain Mobility Using NMR Chemical Shifts. J Am Chem Soc 135: 14536–14539.

93. Robustelli P, Stafford KA, Palmer AG (2012) Interpreting protein structural dynamics from NMR chemical shifts. J Am Chem Soc 134: 6365–6374.

94. Calligari P, Abergel D (2014) Multiple Scale Dynamics in Proteins Probed at Multiple Time Scales through Fluctuations of NMR Chemical Shifts. J Phys Chem B 118: 3823–3831.

95. Hansen DF, Kay LE (2011) Determining valine side-chain rotamer conformations in proteins from methyl 13C chemical shifts: application to the 360 kDa half-proteasome. J Am Chem Soc 133: 8272–8281.

96. Kjaergaard M, Iešmantavičius V, Poulsen FM (2011) The interplay between transient α-helix formation and side chain rotamer distributions in disordered proteins probed by methyl chemical shifts. Protein Sci 20: 2023–2034.

97. Wagner G, Wüthrich K (1982) Sequential Resonance Assignments in Protein 1H Nuclear Magnetic Resonance Spectra: Basic Pancreatic Trypsin Inhibitor. J Mol Biol 155: 347–366.

98. Wagner G, Pardi A, Wüthrich K (1983) Hydrogen bond length and proton NMR chemical shifts in proteins. J Am Chem Soc 105: 5948–5949.

99. Ösapay K, Case DA (1991) A New Analysis of Proton Chemical Shifts in Proteins. J Am Chem Soc 113: 9436–9444.

100. Williamson MP, Asakura T, Nakamura E, Demura M (1992) A method for the calculation of protein α-CH chemical shifts. J Biomol NMR 2: 83–98.

101. Spera S, Bax A (1991) Empirical correlation between protein backbone conformation and C-alpha and C-beta 13C nuclear magnetic resonance chemical shifts. J Am Chem Soc 113: 5490–5492.

102. Le H, Oldfield E (1994) Correlation between 15N NMR chemical shifts in proteins and secondary structure. J Biomol NMR 4: 341–348.

103. 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.

104. Kohlhoff KJ, 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.

105. Seidel K, Etzkorn M, Schneider R, Ader C, Baldus M (2009) Comparative analysis of NMR chemical shift predictions for proteins in the solid phase. Solid State Nucl Magn Reson 35: 235–242.

106. Li D-W, Brüschweiler R (2012) PPM: a side-chain and backbone chemical shift predictor for the assessment of protein conformational ensembles. J Biomol NMR 54: 257–265.

107. Berjanskii M, Liang Y, Zhou J, Tang P, Stothard P, Zhou Y, Cruz J, MacDonell C, Lin G, Lu P, Wishart DS (2010) PROSESS: a protein structure evaluation suite and server. Nucleic Acids Res 38: W633–640.

108. Doreleijers JF, Sousa da Silva AW, Krieger E, Nabuurs SB, Spronk CAEM, Stevens TJ, Vranken WF, Vriend G, Vuister GW (2012) CING: an integrated residue-based structure validation program suite. J Biomol NMR 54: 267–283.

109. Angyán AF, Szappanos B, Perczel A, Gáspári Z (2010) CoNSEnsX: an ensemble view of protein structures and NMR-derived experimental data. BMC Struct Biol 10: 39.

110. Wishart DS, Arndt D, Berjanskii M, Tang P, Zhou J, Lin G (2008) CS23D: a web server for rapid protein structure generation using NMR chemical shifts and sequence data. Nucleic Acids Res 36: W496–502.

111. Shen Y, Lange O, Delaglio F, Rossi P, Aramini JM, Liu G, Eletsky A, Wu Y, Singarapu KK, Lemak A, Ignatchenko A, Arrowsmith CH, Szyperski T, Montelione GT, Baker D, Bax A (2008) Consistent blind protein structure generation from NMR chemical shift data. Proc Natl Acad Sci U S A 105: 4685–4690.

112. 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.

113. Meiler J (2003) PROSHIFT: protein chemical shift prediction using artificial neural networks. J Biomol NMR 26: 25–37.

114. Wang Y (2004) Secondary structural effects on protein NMR chemical shifts. J Biomol NMR 30: 233–244.

115. Nielsen JT, Eghbalnia HR, Nielsen NC (2012) Chemical shift prediction for protein structure calculation and quality assessment using an optimally parameterized force field. Prog Nucl Magn Reson Spectrosc 60: 1–28.

116. Akaike H (1974) A New Look at the Statistical Model Identification. IEEE Trans Automat Contr 19: 716–723.

117. Zeng J, Zhou P, Donald BR (2013) HASH: a program to accurately predict protein Hα shifts from neighboring backbone shifts. J Biomol NMR 55: 105–118.

118. Xu XP, Case DA (2001) Automated prediction of 15N, 13Calpha, 13Cbeta and 13C′ chemical shifts in proteins using a density functional database. J Biomol NMR 21: 321–333.

119. Vila JA, Arnautova YA, Martin OA, Scheraga HA (2009) Quantum-mechanics-derived 13Ca chemical shift server (CheShift) for protein structure validation. Proc Natl Acad Sci U S A 106: 16972–16977.

120. Martin OA, Vila JA, Scheraga HA (2012) CheShift-2: graphic validation of protein structures. Bioinformatics 28: 1538–1539.

121. He X, Wang B, Merz KM (2009) Protein NMR chemical shift calculations based on the automated fragmentation QM/MM approach. J Phys Chem B 113: 10380–10388.

122. Cai L, Fushman D, Kosov DS (2009) Density functional calculations of chemical shielding of backbone 15N in helical residues of protein G. J Biomol NMR 45: 245–253.

123. Frank A, Onila I, Möller HM, Exner TE (2011) Toward the quantum chemical calculation of nuclear magnetic resonance chemical shifts of proteins. Proteins 79: 2189–2202.

124. Frank A, Mo HM, Exner TE (2012) Toward the Quantum Chemical Calculation of NMR Chemical Shifts of Proteins. 2. Level of Theory, Basis Set, and Solvents Model Dependence. J Chem Theory Comput 8: 1480–1492.

125. Zhu T, He X, Zhang JZH (2012) Fragment density functional theory calculation of NMR chemical shifts for proteins with implicit solvation. Phys Chem Chem Phys 14: 7837–7845.

126. Tan H-J, Bettens RPA (2013) Ab initio NMR chemical-shift calculations based on the combined fragmentation method. Phys Chem Chem Phys 15: 7541–7547.

127. Exner TE, Frank A, Onila I, Mo HM (2012) Toward the Quantum Chemical Calculation of NMR Chemical Shifts of Proteins. 3. Conformational Sampling and Explicit Solvents Model. J Chem Theory Comput 8: 4818–4827.

128. Zhu T, Zhang JZH, He X (2013) Automated Fragmentation QM/MM Calculation of Amide Proton Chemical Shifts in Proteins with Explicit Solvent Model. J Chem Theory Comput 9: 2104–2114.

129. Jain R, Bally T, Rablen PR (2009) Calculating accurate proton chemical shifts of organic molecules with density functional methods and modest basis sets. J Org Chem 74: 4017–4023.

130. Lodewyk MW, Siebert MR, Tantillo DJ (2012) Computational prediction of 1H and 13C chemical shifts: a useful tool for natural product, mechanistic, and synthetic organic chemistry. Chem Rev 112: 1839–1862.

131. Dračínský M, Möller HM, Exner TE (2013) Conformational Sampling by Ab Initio Molecular Dynamics Simulations Improves NMR Chemical Shift Predictions. J Chem Theory Comput 9: 3806–3815.

132. Sumowski CV, Hanni M, Schweizer S, Ochsenfeld C (2014) Sensitivity of ab Initio vs Empirical Methods in Computing Structural Effects on NMR Chemical Shifts for the Example of Peptides. J Chem Theory Comput 10: 122–133.

133. Wishart DS, Sykes BD, Richards FM (1992) The chemical shift index: a fast and simple method for the assignment of protein secondary structure through NMR spectroscopy. Biochemistry 31: 1647–1651.

134. Eghbalnia HR, Wang L, Bahrami A, Assadi A, Markley JL (2005) Protein energetic conformational analysis from NMR chemical shifts (PECAN) and its use in determining secondary structural elements. J Biomol NMR 32: 71–81.

135. Wang C-C, Chen J-H, Lai W-C, Chuang W-J (2007) 2DCSi: identification of protein secondary structure and redox state using 2D cluster analysis of NMR chemical shifts. J Biomol NMR 38: 57–63.

136. Mechelke M, Habeck M (2013) A probabilistic model for secondary structure prediction from protein chemical shifts. Proteins 81: 984–993.

137. Camilloni C, De Simone A, Vranken WF, Vendruscolo M (2012) Determination of secondary structure populations in disordered states of proteins using nuclear magnetic resonance chemical shifts. Biochemistry 51: 2224–2231.

138. 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.

139. Shen Y, Delaglio F, Cornilescu G, Bax A (2009) TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. J Biomol NMR 44: 213–223.

140. Berjanskii M V, Neal S, Wishart DS (2006) PREDITOR: a web server for predicting protein torsion angle restraints. Nucleic Acids Res 34: W63–69.

141. Cheung M-S, Maguire ML, Stevens TJ, Broadhurst RW (2010) DANGLE: A Bayesian inferential method for predicting protein backbone dihedral angles and secondary structure. J Magn Reson 202: 223–233.

142. Sahakyan AB, Cavalli A, Vranken WF, Vendruscolo M (2012) Protein Structure Validation Using Side-Chain Chemical Shifts. J Phys Chem B 116: 4754–4759.

143. Krzeminski M, Marsh JA, Neale C, Choy W-Y, Forman-Kay JD (2013) Characterization of disordered proteins with ENSEMBLE. Bioinformatics 29: 398–399.

144. Rieping W, Vranken WF (2010) Validation of archived chemical shifts through atomic coordinates. Proteins 78: 2482–2489.

145. Wang L, Markley JL (2009) Empirical correlation between protein backbone 15N and 13C secondary chemical shifts and its application to nitrogen chemical shift re-referencing. J Biomol NMR 44: 95–99.

146. Wang B, Wang Y, Wishart DS (2010) A probabilistic approach for validating protein NMR chemical shift assignments. J Biomol NMR 47: 85–99.

147. Zhang H, Neal S, Wishart DS (2003) RefDB: A database of uniformly referenced protein chemical shifts. J Biomol NMR 25: 173–195.

148. Hendrickx PMS, Gutmanas A, Kleywegt GJ (2013) Vivaldi: visualization and validation of biomacromolecular NMR structures from the PDB. Proteins 81: 583–591.

149. Montelione GT, Nilges M, Bax A, Güntert P, Herrmann T, Richardson JS, Schwieters CD, Vranken WF, Vuister GW, Wishart DS, Berman HM, Kleywegt GJ, Markley JL (2013) Recommendations of the wwPDB NMR Validation Task Force. Structure 21: 1563–1570.

150. Kuszewski J, Gronenborn AM, Clore GM (1995) The impact of direct refinement against proton chemical shifts on protein structure determination by NMR. J Magn Reson B 107: 293–297.

151. Cavalli A, Salvatella X, Dobson CM, Vendruscolo M (2007) Protein structure determination from NMR chemical shifts. Proc Natl Acad Sci U S A 104: 9615–9620.

152. Rohl CA, Strauss CEM, Misura KMS, Baker D (2004) Protein structure prediction using Rosetta. Methods Enzymol 383: 66–93.

153. Rosato A, Aramini JM, Arrowsmith C, Bagaria A, Baker D, et al. (2012) Blind testing of routine, fully automated determination of protein structures from NMR data. Structure 20: 227–236.

154. Berjanskii M, Tang P, Liang J, Cruz JA, Zhou J, Zhou Y, Bassett E, MacDonell C, Lu P, Lin G, Wishart DS (2009) GeNMR: a web server for rapid NMR-based protein structure determination. Nucleic Acids Res 37: W670–677.

155. Raman S, Huang YJ, Mao B, Rossi P, Aramini JM, Liu G, Montelione GT, Baker D (2010) Accurate automated protein NMR structure determination using unassigned NOESY data. J Am Chem Soc 132: 202–207.

156. Cavanagh J, Fairbrother WJ, III AWH, Skelton NJ, Rance M (2006) Protein NMR Spectroscopy: Principles and Practice. Waltham, MA: Academic Press.

157. Reich HJ (n.d.) Structure Determination Using NMR. http://www.chem.wisc.edu/areas/reich/chem605/. Accessed 22 May 2014.

158. Karplus M (1963) Vicinal Proton Coupling in Nuclear Magnetic Resonance. J Am Chem Soc 85: 2870–2871.

159. Díez E, San-Fabián J, Guilleme J, Altona C, Donders LA (1989) Vicinal proton-proton coupling constants I. Formulation of an equation including interactions between substituents. Mol Phys 68: 49–63.

160. Haasnoot CAG, de Leeuw FAAM, Altona C (1980) The relationship between proton-proton NMR coupling constants and substituent electronegativities—I: An empirical generalization of the karplus equation. Tetrahedron 36: 2783–2792.

161. Constantino MG, Lacerda V, da Silva GV., Tasic L, Rittner R (2001) Principal component analysis of long-range "W" coupling constants of some cyclic compounds. J Mol Struct 597: 129–136.

162. Pretsch E, Bühlmann P, Affolter C (2000) Structure Determination of Organic Compounds: Tables of Spectral Data. 3rd ed. Berlin Heidelberg: Springer.

163. San Fabián J, Guilleme J, Díez E (1998) Vicinal fluorine-proton coupling constants. J Magn Reson 133: 255–265.

164. Lankhorst P, Haasnoot C, Erkelens C, Altona C (1984) Carbon-13 NMR in conformational analysis of nucleic acid fragments. 2. A reparametrization of the Karplus equation for vicinal NMR coupling constants in CCOP and HCOP fragments. J Biomol Struct Dyn 1: 1387–1405.

165. Olaf Kühl (2008) Phosphorus-31 NMR Spectroscopy: A Concise Introduction for the Synthetic Organic and Organometallic Chemist. Berlin Heidelberg: Springer-Verlag.

166. Kwan EE, Huang SG (2008) Structural Elucidation with NMR Spectroscopy: Practical Strategies for Organic Chemists. European J Org Chem 2008: 2671–2688.

167. Bifulco G, Dambruoso P, Gomez-Paloma L, Riccio R (2007) Determination of relative configuration in organic compounds by NMR spectroscopy and computational methods. Chem Rev 107: 3744–3779.

168. Pople JA, Bothner-By AA (1965) Nuclear Spin Coupling Between Geminal Hydrogen Atoms. J Chem Phys 42: 1339–1349.

169. Schaefer T (1962) Correlations of ethylenic proton coupling with electronegativity. Can J Chem 40: 5–8.

170. Cookson RC, Crabb TA, Frankel JJ, Hudec J (1966) Geminal coupling constants in methylene groups. Tetrahedron 22: 355–390.

171. Kraszni M, Szakács Z, Noszál B (2004) Determination of rotamer populations and related parameters from NMR coupling constants: a critical review. Anal Bioanal Chem 378: 1449–1463.

172. ACD/Labs ACD/NMR Predictors http://www.acdlabs.com/products/adh/nmr/nmr_pred/. Accessed 22 May 2014.

173. Mestrelab Mnova NMR Predict Desktop http://mestrelab.com/software/mnova-nmrpredict-desktop/. Accessed 22 May 2014.

174. ChemNMR http://www.upstream.ch/products/chemnmr.html. Accessed 22 May 2014.

175. Binev Y, Marques MMB, Aires-de-Sousa J (2007) Prediction of 1H NMR coupling constants with associative neural networks trained for chemical shifts. J Chem Inf Model 47: 2089–2097.

176. Binev Y, Aires-de-Sousa J (2004) Structure-based predictions of 1H NMR chemical shifts using feed-forward neural networks. J Chem Inf Comput Sci 44: 940–945.

177. Tetko I V (2002) Neural network studies. 4. Introduction to associative neural networks. J Chem Inf Comput Sci 42: 717–728.

178. Di Micco S, Chini MG, Riccio R, Bifulco G (2010) Quantum Mechanical Calculation of NMR Parameters in the Stereostructural Determination of Natural Products. European J Org Chem 2010: 1411–1434.

179. Golotvin SS, Vodopianov E, Lefebvre BA, Williams AJ, Spitzer TD (2006) Automated structure verification based on 1H NMR prediction. Magn Reson Chem 44: 524–538.

180. Golotvin SS, Vodopianov E, Pol R, Lefebvre BA, Williams AJ, Rutkowske RD, Spitzer TD (2007) Automated structure verification based on a combination of 1D 1 H NMR and 2D 1 H – 13 C HSQC spectra. Magn Reson Chem 45: 803–813.

181. Ruan K, Yang S, Van Sant KA, Likos JJ (2009) Application of Hadamard spectroscopy to automated structure verification in high-throughput NMR. Magn Reson Chem 47: 693–700.

182. Plainchont B, Nuzillard J-M (2013) Structure verification through computer-assisted spectral assignment of NMR spectra. Magn Reson Chem 51: 54–59.

183. Elyashberg M, Blinov K, Molodtsov S, Smurnyy Y, Williams AJ, Churanova T (2009) Computer-assisted methods for molecular structure elucidation: realizing a spectroscopist's dream. J Cheminform 1: 3.

184. Plainchont B, de Paulo Emerenciano V, Nuzillard J-M (2013) Recent advances in the structure elucidation of small organic molecules by the LSD software. Magn Reson Chem 51: 447–453.

185. Elyashberg ME, Williams AJ, Martin GE (2008) Computer-assisted structure verification and elucidation tools in NMR-based structure elucidation. Prog Nucl Magn Reson Spectrosc 53: 1–104.

186. Koichi S, Arisaka M, Koshino H, Aoki A, Iwata S, Uno T, Satoh H (2014) Chemical structure elucidation from (13)C NMR chemical shifts: efficient data processing using bipartite matching and maximal clique algorithms. J Chem Inf Model 54: 1027–1035.

187. Elyashberg M, Blinov K, Molodtsov S, Williams AJ (2013) Structure revision of asperjinone using computer-assisted structure elucidation methods. J Nat Prod 76: 113–116.

188. PERCH NMR Software http://www.perchsolutions.com/. Accessed 22 May 2014.

189. Laatikainen R, Hassinen T, Lehtivarjo J, Tiainen M, Jungman J, Tynkkynen T, Korhonen S-P, Niemitz M, Poutiainen P, Jääskeläinen O, Väisänen T, Weisell J, Soininen P, Laatikainen P, Martonen H, Tuppurainen K (2014) Comprehensive Strategy for Proton Chemical Shift Prediction: Linear Prediction with Nonlinear Corrections. J Chem Inf Model 54: 419–430.

190. Santa H, Laatikainen R, Saarela JTA, Tuppurainen K, Peräkylä M (2002) Correlative motions and memory effects in molecular dynamics simulations of molecules: principal components and rescaled range analysis suggest that the motions of native BPTI are more correlated than those of its mutants. Biophys Chem 95: 49–57.

191. Pulkkinen JT, Honkakoski P, Peräkylä M, Berczi I, Laatikainen R (2008) Synthesis and evaluation of estrogen agonism of diaryl 4,5-dihydroisoxazoles, 3-hydroxyketones, 3-methoxyketones, and 1,3-diketones: a compound set forming a 4D molecular library. J Med Chem 51: 3562–3571.

192. Baskaran K, Brunner K, Munte CE, Kalbitzer HR (2010) Mapping of protein structural ensembles by chemical shifts. J Biomol NMR 48: 71–83.

193. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. 28: 235–242.

194. Case DA, Darden TA, Cheatham TEI, Simmerling CL, Wang J, et al. (2012) AMBER 12.

195. 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 65: 712–725.

196. Halgren TA (1996) Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. J Comput Chem 17: 490–519.

197. Indyk P, Motwani R (1998) Approximate nearest neighbors: towards removing the curse of dimensionality. Proceedings of the thirtieth annual ACM symposium on Theory of computing - STOC ’98. New York, New York: ACM Press. pp. 604–613.

198. Hadi AS, Ling RF (1998) Some Cautionary Notes on the Use of Principal Components Regression. Am Stat 52: 15–19.

199. Xie Y-L, Kalivas JH (1997) Evaluation of principal component selection methods to form a global prediction model by principal component regression. Anal Chim Acta 348: 19–27.

200. Depczynski U, Frost VJ, Molt K (2000) Genetic algorithms applied to the selection of factors in principal component regression. Anal Chim Acta 420: 217–227.

201. Breiman L (2001) Random forests. Mach Learn 45: 5–32.

202. Cover T, Hart P (1967) Nearest neighbor pattern classification. IEEE Trans Inf Theory 13: 21–27.

203. Beyer K, Goldstein J, Ramakrishnan R, Shaft U (1999) When Is “Nearest Neighbor” Meaningful? Proceeding ICDT ’99 Proceedings of the 7th International Conference on Database Theory. London: Springer-Verlag. pp. 217–235.

204. Markwick PR, Cervantes CF, Abel BL, Komives EA, Blackledge M, McCammon JA (2010) Enhanced conformational space sampling improves the prediction of chemical shifts in proteins. J Am Chem Soc 132: 1220–1221.

205. Li D-W, Brüschweiler R (2010) Certification of Molecular Dynamics Trajectories with NMR Chemical Shifts. J Phys Chem Lett 1: 246–248.

206. Andrec M, Snyder DA, Zhou Z, Young J, Montelione GT, Levy RM (2007) A large data set comparison of protein structures determined by crystallography and NMR: statistical test for structural differences and the effect of crystal packing. Proteins 69: 449–465.

207. Sikic K, Tomic S, Carugo O (2010) Systematic comparison of crystal and NMR protein structures deposited in the protein data bank. Open Biochem J 4: 83–95.

208. Garbuzynskiy SO, Melnik BS, Lobanov MY, Finkelstein A V, Galzitskaya O V (2005) Comparison of X-ray and NMR structures: is there a systematic difference in residue contacts between X-ray- and NMR-resolved protein structures? Proteins 60: 139–147.

209. Bagaria A, Jaravine V, Güntert P (2013) Estimating structure quality trends in the Protein Data Bank by equivalent resolution. Comput Biol Chem 46: 8–15.

210. Lindorff-Larsen K, Maragakis P, Piana S, Eastwood MP, Dror RO, Shaw DE (2012) Systematic validation of protein force fields against experimental data. PLoS One 7: e32131.

211. Beauchamp KA, Lin Y-S, Das R, Pande VS (2012) Are Protein Force Fields Getting Better? A Systematic Benchmark on 524 Diverse NMR Measurements. J Chem Theory Comput 8: 1409–1414.

212. Li D-W, Brüschweiler R (2010) NMR-based protein potentials. Angew Chem Int Ed Engl 49: 6778–6780.

213. Nerenberg PS, Head-Gordon T (2011) Optimizing Protein - Solvent Force Fields to Reproduce Intrinsic Conformational Preferences of Model Peptides. J Chem Theory Comput 7: 1220–1230.

214. Wang Z-X, Zhang W, Wu C, Lei H, Cieplak P, Duan Y (2006) Strike a Balance: Optimization of Backbone Torsion Parameters of AMBER Polarizable Force Field for Simulations of Proteins and Peptides. J Comput Chem 27: 781–790.

215. Dixon RW, Kollman PA (1997) Advancing beyond the atom-centered model in additive and nonadditive molecular mechanics. J Comput Chem 18: 1632–1646.

216. Ponder JW, Wu C, Pande VS, Chodera JD, Schnieders MJ, Haque I, Mobley DL, Lambrecht DS, Distasio RA, Head-Gordon M, Clark GNI, Johnson ME, Head-Gordon T (2010) Current Status of the AMOEBA Polarizable Force Field. J Phys Chem B 114: 2549–2564.

217. Shi Y, Xia Z, Zhang J, Best R, Wu C, Ponder JW, Ren P (2013) The Polarizable Atomic Multipole-based AMOEBA Force Field for Proteins. J Chem Theory Comput 9: 4046–4063.

218. Ruschak AM, Kay LE (2010) Methyl groups as probes of supra-molecular structure, dynamics and function. J Biomol NMR 46: 75–87.

219. Religa TL, Sprangers R, Kay LE (2010) Dynamic regulation of archaeal proteasome gate opening as studied by TROSY NMR. Science 328: 98–102.

220. Gifford JL, Ishida H, Vogel HJ (2011) Fast methionine-based solution structure determination of calcium-calmodulin complexes. J Biomol NMR 50: 71–81.

221. Bista M, Freund SM, Fersht AR (2012) Domain-domain interactions in full-length p53 and a specific DNA complex probed by methyl NMR spectroscopy. Proc Natl Acad Sci U S A 109: 15752–15756.

222. Stoffregen MC, Schwer MM, Renschler FA, Wiesner S (2012) Methionine scanning as an NMR tool for detecting and analyzing biomolecular interaction surfaces. Structure 20: 573–581.

223. Shapovalov M V, Dunbrack RL (2011) A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. Structure 19: 844–858.

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

225. Lemeer S, Heck AJR (2009) The phosphoproteomics data explosion. Curr Opin Chem Biol 13: 414–420.

226. Dinkel H, Chica C, Via A, Gould CM, Jensen LJ, Gibson TJ, Diella F (2011) Phospho.ELM: a database of phosphorylation sites--update 2011. Nucleic Acids Res 39: D261–267.

APPENDICES I-III:

ORIGINAL PUBLICATIONS

I

4D prediction of protein ^{1}H chemical shifts

Lehtivarjo J, Hassinen T, Korhonen S-P, Peräkylä M and Laatikainen R

Journal of Biomolecular NMR 45: 413-426, 2009

II

Combining NMR ensembles and molecular dynamics simulations provides more realistic models of protein structures in solution and leads to better chemical shift prediction

Lehtivarjo J, Hassinen T, Korhonen S-P, Peräkylä M and Laatikainen R

Journal of Biomolecular NMR 52: 257-267, 2012

III

Universal J-coupling prediction

Lehtivarjo J, Niemitz M and Korhonen S-P

Journal of Chemical Information and Modeling 54: 810-817, 2014.

APPENDIX IV

Use of chemical shift prediction for protein structure evaluation

This chapter was a part of the original manuscript of Paper I, but was removed due to reviewer comments about the excessive length of the manuscript. These results are also published in the following poster: Lehtivarjo J. and Laatikainen R. (**2009**): Use of ^{1}H Chemical Shift Prediction for Protein Structure Evaluation. *The XXXI Finnish NMR Symposium, Kuusamo, Finland. Book of abstracts p. 17.*

Chemical shift prediction in protein structure evaluation

Phosphocarrier protein HPr I14A mutant (PDB 1TXE) was chosen for evaluating differences between ten conformations of its NMR structure ensemble. For all the conformations, chemical shifts were predicted with the 150 ps protocol. For each residue, a combined RMS error, called "*RMS score*" was calculated as an average of backbone chemical shift RMS errors [6]

$$RMS\ score = \frac{RMS_{HN} + (RMS_{H\alpha} * 0.75)}{2} \qquad [6]$$

The Hα shifts were weighted down with the factor of 0.75, as they are only about 75 % dependent of the 3D structure (Wishart and Case 2001). This "*RMS score*" is plotted against the sequence in Fig. 10, showing the problematic areas of the sequence, which in this case seem to be located in the random coil structures. Usually, when such areas are examined, some typical errors are found. This kind of analysis can give hint how the structure is incorrect or could be improved, or if different conformations give different results, the best one may be selected to represent the most correct conformation. In Fig. 11, the structural properties causing the six largest prediction errors of the HPr protein are analyzed.

The shift prediction RMS error as a criterion for selecting the best conformers should be considered, as it is directly connected to observed results, instead of more artificial criteria, such as lowest energies. This was recently proposed also by MINOES approach (Krzeminski et al. 2009), which compares the observed and SHIFTX-predicted (Neal et al. 2003) Hα chemical shifts. In HPr ensemble, the smallest RMS error for HN shifts was 0.44 ppm, compared to the largest one of 0.49 ppm. Although this may sound quite small, it reflects large differences in individual residues, as seen in Fig. 10.

The largest prediction error is found in the D30HN proton. A part of the error can be explained by the nearest backbone torsion angles (Fig. 11A). Still, an over 2 ppm prediction error remains. Because the error of the adjacent residue (S31HN) is also large, it is probable that the whole loop is incorrectly folded. The large observed shift of D30HN is explained by the aromatic ring shielding of F29.

Fig. 11B shows a typical error of β-turns. Only one conformer of the ensemble has a large prediction error in N38HN, and it is the only one with a different type of β-turn. This is not caused by MD, as it is already present in rigid structure ensemble. By means of chemical shift prediction, these kinds of erroneous conformations can easily be ruled out and prevented from entering the final ensemble.

The residues 53-56 form a flexible random coil loop between two β-turns (Fig. 11C). All residues of this region suffer from prediction errors of average size. However, one conformation has been flipped during MD simulation, causing much larger error to V55HN due to hydrogen bond breakup. Typically, if hydrogen bond is missing, due to inaccuracy in original rigid structures or caused by MD simulation, a prediction error of ca. 1 ppm appears.

Fig. 11D presents another missing hydrogen bond: this time the side chain carboxyl serves as hydrogen bond acceptor. Unfortunately, over 1.5 ppm prediction error for G67HN remains even when the hydrogen bond is present. Either the region is incorrectly folded or the hydrogen bonds to side chain carboxyls are not properly described in the prediction protocol.

A typical aromatic ring effect is seen in Fig. 11E, where the aromatic ring of F29 causes strong upfield effect to A73Hα. Although in this example prediction error is mainly caused by aromatic ring distance, an incorrect ring orientation may also cause ±1.5 ppm effects to proximal nuclei. Often MD simulations slightly smooth errors caused by incorrect ring orientations. For example, among rigid structures of HPr ensemble, average error for rigid A73Hα is -1.22 ppm, and for 150 ps structures, it is -1.02 ppm.

Lysine K83 is located as the last residue of an α-helix. Fig. 11F shows how a slight unwinding at the end of the helix is enough to create about 1 ppm prediction error for K83HN. Again, this error mostly arises from the broken hydrogen bond to A79O.

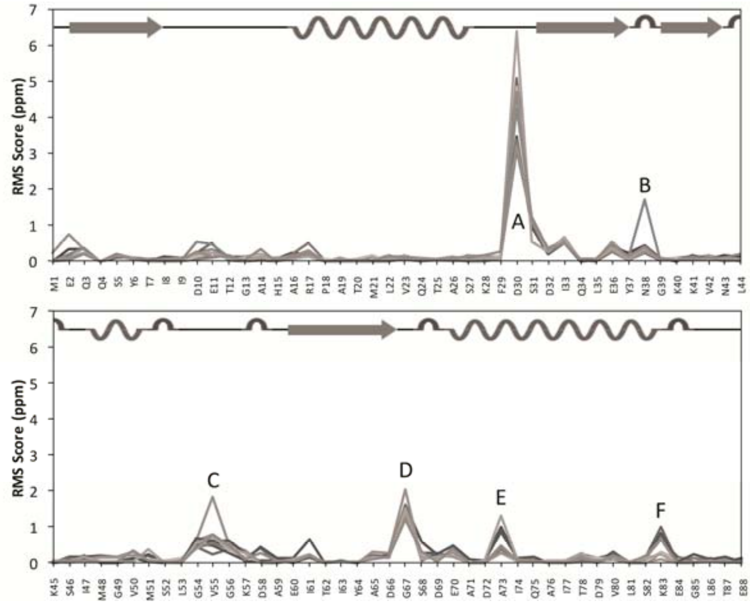
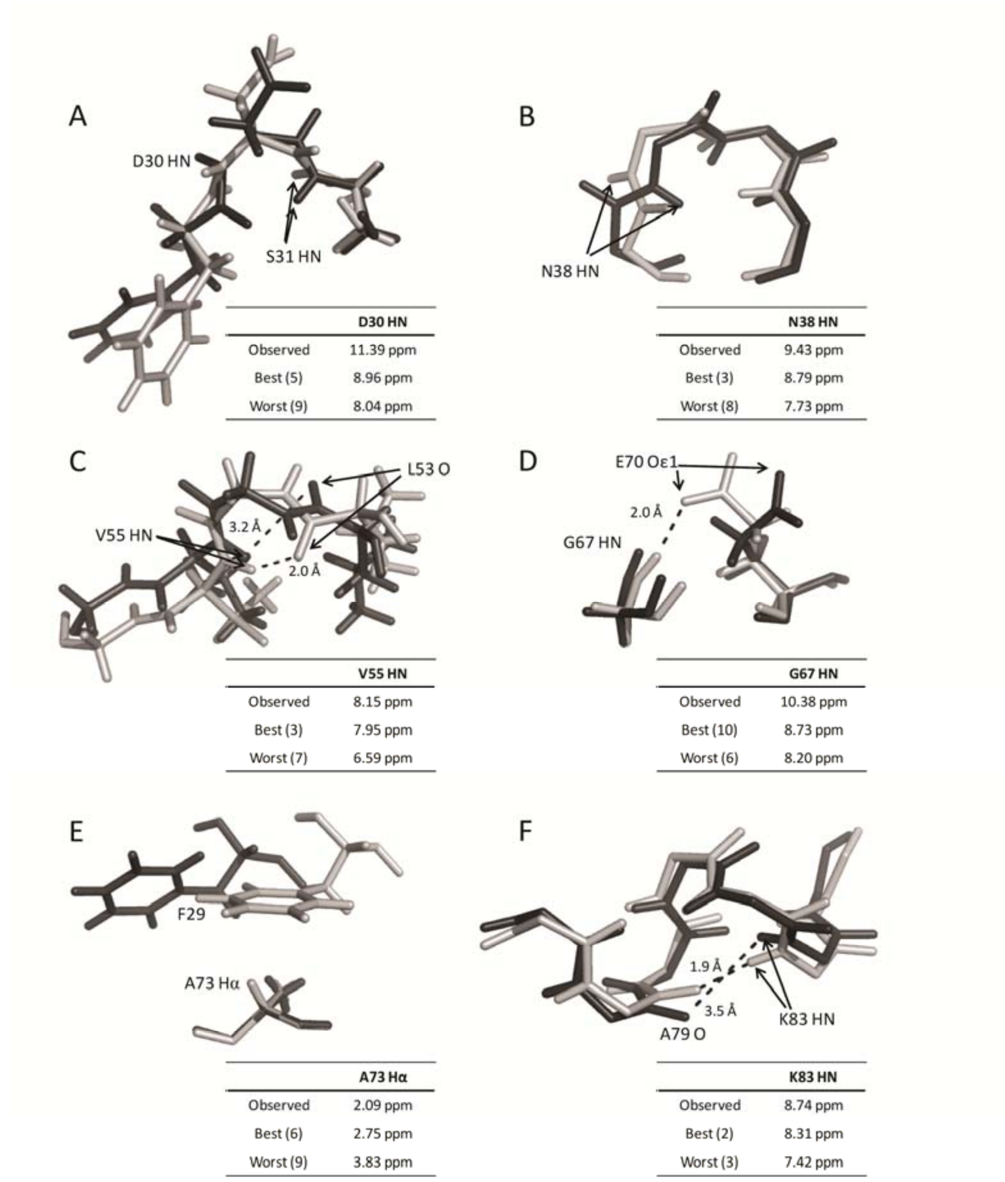


Fig. 10 Prediction RMS-Score vs. sequence for protein Hrp I14A (PDB 1TXE). Each line represents one of the 10 conformers of the NMR structure ensemble. Secondary structure scheme, showing helices, sheets and turns, is the PDB SEQRES sequence, downloaded from the PDB web site http://www.rcsb.org/pdb. The letters from A to F refer to Fig. 11, where the corresponding 3D structures are illustrated.



	D30 HN
Observed	11.39 ppm
Best (5)	8.96 ppm
Worst (9)	8.04 ppm

	N38 HN
Observed	9.43 ppm
Best (3)	8.79 ppm
Worst (8)	7.73 ppm

	V55 HN
Observed	8.15 ppm
Best (3)	7.95 ppm
Worst (7)	6.59 ppm

	G67 HN
Observed	10.38 ppm
Best (10)	8.73 ppm
Worst (6)	8.20 ppm

	A73 Hα
Observed	2.09 ppm
Best (6)	2.75 ppm
Worst (9)	3.83 ppm

	K83 HN
Observed	8.74 ppm
Best (2)	8.31 ppm
Worst (3)	7.42 ppm

Fig. 11 The six largest prediction errors for protein Hrp I14A (PDB 1TXE). The white and black structures represent the best and worst conformers of the NMR ensemble, respectively. In figures B and F, side chain atoms are hidden for clarity. The insert tables present the observed shifts and predicted shifts for the best and worst conformers of the nuclei. The numbers in parenthesis are conformer indices of the ensemble.

Juuso Lehtivarjo

Predicting NMR Parameters from the Molecular Structure

The spectral parameters of the nuclear magnetic resonance (NMR) spectra are dependent on the chemical environment around the nuclei, making NMR spectroscopy a powerful method for studying molecular structure and dynamics at the atomic level. The possibility to predict the spectral parameters from known or proposed molecular structures can be exploited in different applications. This thesis presents two NMR parameter prediction approaches, aimed for protein structure analysis and small molecule structure verification.

UNIVERSITY OF
EASTERN FINLAND

Publications of the University of Eastern Finland
Dissertations in Health Sciences

ISBN 978-952-61-1517-7