DISSERTATION | DOCTORAL THESIS

Titel | Title

Bridging Computational and NMR Approaches to Unravel Protein-Ligand Interactions

verfasst von | submitted by

mgr lic. Aleksandra Laura Ptaszek

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Doktorin der Naturwissenschaften (Dr.rer.nat.)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 796 605 419
Dissertationsgebiet lt. Studienblatt | Field of study as it appears on the student record sheet:	Chemie
Betreut von | Supervisor:	Univ.-Prof. Dr. Robert Konrat Dr. Pedro Alejandro Sánchez Murcia

SUPERVISORS:
Univ.-Prof. Dr. Robert Konrat
Assoc.-Prof. Dr. Pedro Alejandro Sánchez Murcia

LOCATION:
Vienna, Austria

TIME FRAME:
November 2021 - April 2025

ACKNOWLEDGMENTS

First, I would like to thank Prof. Robert Konrat for your trust, openness to new ideas and all the inspiring conversations that showed me how fun science can be! Your enthusiasm and passion have kept me motivated throughout this PhD journey.

I am immensely grateful to Prof. Pedro A. Sánchez Murcia, who guided me through my first steps in biomolecular simulations and was a great supporter during my first days abroad and throughout my entire PhD. Thank you for involving me in so many scientific initiatives and projects that have made me a better scientist. It was a joy and an honour to be the first student in your lab and to help build it from the ground up.

Special thanks to Gerald Platzer for our endless discussions about chemical shifts and your enthusiasm for science. I thoroughly enjoyed learning from your experimental perspective over the years, sharing mine with you, and building upon them together. And, of course, for watering our plants:) I am also grateful to the entire MAG-lab, especially Sarah Kratzwald for the protein purification lesson!

My sincere appreciation goes to Moritz Mayer for taking care of our collaboration with Boehringer Ingelheim. I greatly appreciate your input on each project and the fact that I could always count on you. I am also thankful to the whole team, in particular Darryl McConnell, for the always engaging discussions at our JRC meetings, especially those on water molecules.

To all the scientists I have had the pleasure of working with, thank you for your valuable input and efforts.

I would like to acknowledge my first mentor, Prof. Mariusz P. Mitoraj, who supervised me during my bachelor's and master's studies, for sharing with me your fascination with chemical bonds, a passion that resonates with me to this day. Your mentorship gave me a strong foundation in computational chemistry. I am also grateful to Filip Sagan from Jagiellonian University for our great collaboration and for staying in touch over the years.

From UC Berkeley, I would like to thank Prof. Teresa Head-Gordon for being a great inspiration, for the time she dedicated to me, her trust and support. The machine learning era is an incredibly exciting time to be involved in computational chemistry, and I am grateful for the opportunity to be a part of it. Big thanks also to my main collaborators there, Jie Li and Jiashu Liang, it was truly a pleasure doing science together! And to the whole Head-Gordon group for making me feel so welcome.

To my lovely labmates from CAMDgraz lab: Daniel Platero Rochart, Bruno di Geronimo, Špela Mandl, Juan Toledo, Christoph Nusshold, and Ulrika Ferstl as well as all the visiting students and the entire Division of Medicinal Chemistry. Also, to my dear friends from MedUniGraz: Marco, Rina, Alankrita, Ceyhun, Olaya, Barsha, Aneta, Alicja, Tejus, Anastasia and others - thank you for making me feel at home away from home! The atmosphere at MedUniGraz was truly unforgettable.

A special mention to Sonja Knödlstorfer and Giorgia Toscano for all the fun we had together, from science to karaoke and beyond! I would also like to thank the entire Robert's team: Daniel Braun, Andreas Beier, Thomas Schwarz, Mario Migotti, Karin Ledolter, Theresa Höfurthner, Irene Ceccolini, Letizia Piacentini, Georg Kontaxis and all the visiting students. Working in such an interdisciplinary environment has been a fantastic experience, allowing us to learn from each other and combine our strengths. My sincere thanks to Karin Pfeiffer and Iris Gruber for handling the administration and having the answers to all my questions.

To my parents and grandparents, thank you for your unwavering support and for giving me the freedom to make my own choices. To my wonderful friends, I am extremely grateful to have you: Olga, Kabi, Ola, Sonia, JJ, Grzesiu, Adam, Anita, Michał and others.

To my fiancé, Szymon - thank you for moving abroad with me and always being by my side. This journey would not have been possible without your love and support.

PUBLICATIONS

[1] A. L. Ptaszek, S. Kratzwald, F. Sagan, *et al.*, "From weak interactions to strong affinity: Deciphering the streptavidin-biotin interaction through nmr and computational analysis," *bioRxiv*, pp. 2024–12, 2024.

[2] A. L. Ptaszek, J. Li, R. Konrat, G. Platzer, and T. Head-Gordon, "Ucbshift 2.0: Bridging the gap from backbone to side chain protein chemical shift prediction for protein structures," *Journal of the American Chemical Society*, vol. 146, no. 46, pp. 31 733–31 745, 2024.

[3] G. Platzer, A. L. Ptaszek, J. Böttcher, *et al.*, "Ligand 1h nmr chemical shifts as accurate reporters for protein-ligand binding interfaces in solution," *ChemPhysChem*, vol. 25, no. 1, e202300636, 2024.

[4] J. Li, J. Liang, Z. Wang, *et al.*, "Highly accurate prediction of nmr chemical shifts from low-level quantum mechanics calculations using machine learning," *Journal of Chemical Theory and Computation*, vol. 20, no. 5, pp. 2152–2166, 2024.

[5] A. Beier, G. Platzer, T. Höfurthner, *et al.*, "Probing protein–ligand methyl-π interaction geometries through chemical shift measurements of selectively labeled methyl groups," *Journal of Medicinal Chemistry*, vol. 67, no. 15, pp. 13 187–13 196, 2024.

[6] M. Gazorpak, K. M. Hugentobler, D. Paul, *et al.*, "Harnessing protac technology to combat stress hormone receptor activation," *Nature Communications*, vol. 14, no. 1, p. 8177, 2023.

[7] A. L. Ptaszek, F. Sagan, R. Filas, P. Kubisiak, and M. P. Mitoraj, "Theoretical description of hydride-hydride interactions in selected hydrogen storage materials," *ChemPhysChem*, e202400668, 2021.

[8] G. Mahmoudi, M. G. Babashkina, W. Maniukiewicz, *et al.*, "Solvent-induced formation of novel ni (ii) complexes derived from bis-thiosemicarbazone ligand: An insight from experimental and theoretical investigations," *International journal of molecular sciences*, vol. 22, no. 10, p. 5337, 2021.

[9] D. A. Safin, M. G. Babashkina, M. Bolte, *et al.*, "Novel sterically demanding schiff base dyes: An insight from experimental and theoretical calculations," *Journal of Luminescence*, vol. 238, p. 118 264, 2021.

[10] M. P. Mitoraj, F. Sagan, D. W. Szczepanik, *et al.*, "Origin of hydrocarbons stability from a computational perspective: A case study of ortho-xylene isomers," *ChemPhysChem*, vol. 21, no. 6, pp. 494–502, 2020.

ABSTRACT

Nuclear magnetic resonance (NMR) chemical shifts are among the most fundamental observables in chemistry and structural biology. As sensitive probes of the local electronic environment and geometry, they provide valuable insights into molecular structure and interactions. In solution, NMR captures ensemble-averaged information, making it uniquely suited for studying dynamic biological systems. However, mapping experimental observations into structural models remains a bottleneck, requiring robust computational approaches. This thesis explores two computational strategies for calculating chemical shifts in biomolecular systems, including quantum chemistry and data-driven methods.

The first approach shows that protein-bound ligand ^{1}H NMR chemical shifts can serve as precise markers of the immediate chemical environment at protein-ligand interfaces. Cross-validation of chemical shifts calculated using quantum chemistry combined with molecular dynamics against experimental data allows for the refinement of X-ray structures, yielding a representation that better reflects the solution state of the complex. In addition, the origin of perturbations in ^{1}H NMR signals upon protein binding is investigated, extending the study to the role of weak interactions in a model protein-ligand system.

Furthermore, this work introduces UCBShift 2.0, a machine learning-based model for protein chemical shift prediction. By integrating structural features, sequence and structure alignment with curated experimental data, UCBShift 2.0 advances the accuracy of chemical shift prediction for both backbone and side chains, outperforming existing methods. With its enhanced predictive accuracy, UCBShift 2.0 has the potential to automate the NMR signal assignment.

By combining experimental NMR data with advanced computational techniques, this thesis contributes to bridging the gap between static structural data and the dynamic behaviour of biomolecular systems, supporting advances in structural biology and drug discovery.

ZUSAMMENFASSUNG

Die Messungen der chemischen Verschiebungen in der Kernspinresonanzspektroskopie (NMR-Spektroskopie) ist die Grundlage für deren Anwendung in Chemie und Strukturbiologie. Als empfindliche Sonden der lokalen elektronischen Umgebung und Geometrie liefern sie wertvolle Einblicke in die Molekülstruktur und molekulare Wechselwirkungen. In Lösung erfasst die NMR ensemblegemittelte Informationen und eignet sich daher besonders für die Untersuchung dynamischer biologischer Systeme. Die Zuordnung experimenteller Beobachtungen zu strukturellen Modellen bleibt jedoch eine Herausforderung, die robuste rechnergestützte Ansätze erfordert. Die vorliegende Dissertation untersucht zwei rechnergestützte Strategien zur Berechnung chemischer Verschiebungen in biomolekularen Systemen: quantenchemische Methoden und datengesteuerte Ansätze.

Der erste Ansatz zeigt, dass die ^{1}H-NMR-chemischen Verschiebungen von protein-gebundenen Liganden als präzise Marker der unmittelbaren chemischen Umgebung an Protein-Ligand-Grenzflächen dienen können. Die Kreuzvalidierung der chemischen Verschiebungen, die mittels Quantenchemie in Kombination mit Molekulardynamik und experimentellen Daten berechnet wurden, ermöglicht die Verfeinerung von Röntgenstrukturen, wodurch eine verfeinerte Darstellung erhalten wird, die den Lösungsszustand des Komplexes besser widerspiegelt. Darüber hinaus wird der Ursprung der Veränderungen in den ^{1}H-NMR-Signalen nach der Proteinbindung untersucht, wobei die Rolle schwacher Wechselwirkungen in einem Modell-Protein-Ligand-System näher betrachtet wird.

Weiters wird in dieser Arbeit ein neues Programm, UCBShift 2.0, vorgestellt, ein auf maschinellem Lernen basierendes Modell zur Vorhersage chemischer Verschiebungen in Proteinen. Durch die Integration struktureller Merkmale, Sequenz- und Strukturausrichtung mit kuratierten experimentellen Daten verbessert UCBShift 2.0 die Genauigkeit der chemischen Verschiebungsvorhersage sowohl für das Rückgrat als auch für die Seitenketten und übertrifft bestehende Methoden. Mit seiner gesteigerten Vorhersagegenauigkeit hat UCBShift 2.0 das Potenzial, NMR-Signalzuordnungsprogramme zu automatisieren.

Durch die Kombination experimenteller NMR-Daten mit fortgeschrittenen, computergestützten Techniken trägt diese Dissertation dazu bei, die Lücke zwischen statischen Strukturdaten und dem dynamischen Verhalten biomolekularer Systeme zu schließen und damit Fortschritte in der Strukturbiologie und Wirkstoffentwicklung zu unterstützen.

CONTENTS

Part I

INTRODUCTION

1

INTRODUCTION

Nuclear Magnetic Resonance (NMR) spectroscopy stands as one of the most valuable techniques for identifying molecules and understanding their structural and dynamic properties. Central to this methodology is the fact that distinct types of nuclei, such as ^{1}H (proton) and ^{13}C (carbon), resonate at characteristic frequencies in the presence of a strong magnetic field. However, the power of NMR goes beyond simply distinguishing chemical elements. Nuclei of the same isotope, when exposed to different electronic environments, give rise to separate resonance signals. A textbook example is ethanol, with its three chemically distinct proton groups: methyl, methylene, and hydroxyl, each producing a unique peak. This phenomenon, known as a *chemical shift*, arises from variations in the electron density surrounding each nucleus, thereby providing a unique "fingerprint" of the molecular structure. Experimental chemists routinely use this fact to determine the structures of organic and inorganic compounds. More intriguing, however, is how chemical shifts respond to changes in experimental conditions. For instance, while the methyl groups of ethanol exhibit relatively stable chemical shifts across different solvents, the hydroxyl proton can shift dramatically depending on the solvent's polarity [1]. This highlights how NMR captures not only the intrinsic properties of molecules but also their environment, making it a perfect tool for studying intermolecular interactions. Solution-state NMR is now widely used to investigate structures and interactions of biomolecules under near-physiological conditions, extending even to in-cell experiments. In contrast to X-ray crystallography, which provides only static structural information about molecules, NMR in solution reflects an ensemble of conformations that molecules explore dynamically. This capability is particularly valuable in drug discovery, where an accurate representation of the binding interface in solution is critical to understanding how drugs interact with their targets.

The high sensitivity of NMR observables to alterations in the electronic structure, conformation and molecular environment, makes their analysis a highly complicated task. The primary challenge for NMR spectroscopists is mapping the information derived from the spectrum into an accurate representation of the molecular structure. To aid and optimize the resonance assignment process, empirical and *ab initio* methods are routinely used. This thesis presents two computational approaches for the prediction of NMR chemical shifts. The first involves quantum mechanical (QM) calculations, which are ap-

plied to study interactions in protein-ligand systems, as explored in the first two papers, Chapters 3 and 4. In the first article, QM calculations are further integrated with classical and quantum mechanics/molecular mechanics (QM/MM) molecular dynamics (MD) simulations to obtain a more accurate representation of the solution-state system. The second approach employs a machine learning algorithm designed for protein side-chain chemical shift prediction, which is the focus of the third paper, Chapter 5.

In the next chapter, I briefly introduce both *ab initio* and machine learning approaches for chemical shift prediction. I outline the key difficulties in the field and explain how non-covalent interactions contribute to chemical shift variations. With a few exceptions, the discussion is limited to solution-state NMR chemical shift prediction applied to proteins and organic molecules.

2

NMR CHEMICAL SHIFTS PREDICTION

2.1 *ab initio* APPROACH

Nuclei in molecules are surrounded by electrons that are constantly in motion. When an external magnetic field $\mathbf{B_0}$ is applied to the sample, it perturbs the electronic wave function of the system and induces electronic currents. These currents in turn generate a secondary magnetic field at each nuclear site. As a result, the nucleus does not experience the full external field $\mathbf{B_0}$, but rather an effective field given by:

$$\mathbf{B_{nucleus}} = (\mathbf{1} - \boldsymbol{\sigma})\mathbf{B_0} \tag{1}$$

where σ is the *chemical shielding* tensor that includes information about the magnitude and orientation of the shielding effect. Since chemical shielding arises from the electronic structure of molecules, its proper description must account for the molecular Hamiltonian in the presence of an external magnetic field. Because the magnetic effects are minimal compared with electron orbital energies, they can be treated by perturbation theory [2], [3]. Ramsey was the first to implement this approach, deriving the fundamental equations for nuclear magnetic shielding and spin-spin coupling [2], [4]. This thesis focuses solely on the former. Ramsey's expression for chemical shielding separates it into two distinct components:

$$\boldsymbol{\sigma} = \boldsymbol{\sigma}^{dia} + \boldsymbol{\sigma}^{para} \tag{2}$$

where $\boldsymbol{\sigma}^{dia}$ accounts for the diamagnetic and $\boldsymbol{\sigma}^{para}$ for the paramagnetic contribution.

The diamagnetic term comes from the first-order perturbation correction, and its α, β elements are given by the formula:

$$\sigma_{\alpha\beta}^{\mathrm{dia}} = \frac{\mathrm{e}^2\mu_0}{8\pi m_e}\langle\psi_0|\sum_k \frac{r_{kN}^2\delta_{\alpha\beta} - r_{kN\alpha}r_{kN\beta}}{r_{kN}^3}|\psi_0\rangle, \qquad \alpha, \beta = x, y, z \tag{3}$$

where e and m_e are elementary charge and mass of the electron, μ_0 is the vacuum permeability, ψ_0 is a ground-state electronic wave function of the system, $\delta_{\alpha\beta}$ is the Kronecker delta and r_{kN} is the distance between the electron k and the nucleus N. The gauge origin for the magnetic vector potential is at the nucleus site N. The summation runs over all k electrons [4]–[7].

The diamagnetic component was previously derived by Lamb and can be used to calculate the magnetic shielding of atoms [8]. To move

beyond atoms to molecules, the paramagnetic component must also be considered [4]. The paramagnetic term arises from the second-order perturbation theory, and its components are given by the following expression:

$$\sigma_{\alpha\beta}^{\text{para}} = -\frac{e^2\mu_0}{8\pi m_e^2} \sum_n \frac{1}{E_n - E_0} \times \left[\langle \psi_0 | \sum_k \hat{L}_{k\alpha} | \psi_n \rangle \langle \psi_n | \sum_k \frac{\hat{L}_{kN\beta}}{r_{kN}^3} | \psi_0 \rangle + \text{c.c.} \right] \tag{4}$$

Here the sum extends over all excited electronic states and electrons. E_n and E_0 are energies of the excited and ground electronic states, respectively. $\hat{L}_{k\alpha}$ is the α-component angular momentum operator relative to the magnetic field and $\hat{L}_{kN\beta}$ is the β-component angular momentum operator about the nucleus. *c.c.* refers to the complex conjugate [6], [7].

While the diamagnetic term can be computed solely from the ground state, the paramagnetic component requires contributions from excited-state wave functions and their corresponding energies, making it significantly more difficult to calculate. From a computational perspective, Ramsey's theory presents a practical challenge known as the *gauge origin problem*. Within the finite basis set approximation, neither the diamagnetic nor the paramagnetic terms are strictly gauge-invariant, meaning their values depend on the choice of the gauge origin when defining the magnetic vector potential [9].

The most common approaches to address this issue are the gauge-including atomic orbital (GIAO) method [10], [11], which incorporates an explicit field dependence into the basis functions, and the continuous set of gauge transformations (CSGT) approach [12]. GIAO method exhibits faster and smoother convergence with respect to basis set size compared to CSGT [13].

In solution-state NMR, because of the molecular tumbling motions, only the isotropic part of the magnetic shielding tensor $\boldsymbol{\sigma}(\mathbf{r})$ is observed:

$$\sigma(\mathbf{r}) = \frac{\text{Tr}[\boldsymbol{\sigma}(\mathbf{r})]}{3} \tag{5}$$

In practice, NMR experiments do not directly measure chemical shielding but rather the *chemical shift* defined as:

$$\delta = \sigma_{ref}(\mathbf{r}) - \sigma(\mathbf{r}) \tag{6}$$

where $\sigma_{ref}(\mathbf{r})$ represents the chemical shielding of a nucleus in an arbitrarily chosen reference sample. To determine chemical shift, both the shielding of the nucleus of interest and that of the reference must be determined at the same level of theory. Alternatively, when a set of chemical shieldings needs to be converted, a linear correlation between computed chemical shieldings and experimental chemical

shifts across a range of nuclei can be used. When this correlation has a slope close to -1, the intercept corresponds to the shielding of the reference compound, indirectly converting calculated shieldings into chemical shifts [14]–[16].

Overcoming the *gauge origin problem* led to the widespread use of *ab initio* methods for the prediction of NMR parameters. The origins can be traced back to applications of Hartree-Fock theory combined with GIAO [17], [18] and its later extensions to explicitly correlated methods [9], [19]–[21]. Given the limitations of the electron correlation treatment in the Hartree-Fock method and the substantial computational effort of Post-Hartree-Fock approaches, the focus of the community shifted to density functional theory (DFT), which has become a standard in the field [22], [23]. However, post-Hartree-Fock approaches, especially MP2 and CCSD(T), remain essential for DFT functionals benchmarking and building golden-standard quality datasets [24]. Multiple benchmark studies have been conducted for ^{1}H and ^{13}C, as well as less common nuclei, with no consensus on a single best functional or basis set [16], [24]–[32]. While hybrid DFT functionals generally perform well, the optimal choice of theory level depends on the nucleus and molecular system and must be balanced against computational cost and accuracy. In addition, if heavy elements are present in the system under study, relativistic effects may need to be considered [33], [34]. In recent years, composite approaches have gained popularity, aiming to approximate CCSD(T)/large-basis-set results at a reduced computational cost [35]–[37]. One such method, proposed by Liang et al. [37], employs locally dense basis sets, specifically designed for nuclear magnetic shielding calculations [38].

2.1.1 *Multiscale methods for biomolecules*

The methods described above are not directly applicable to biomolecules due to their large size and associated computational cost. To address this challenge, system partitioning methods can be employed. The most common approach is the hybrid quantum mechanics/molecular mechanics (QM/MM) method, originally designed for studying chemical reactions in biomolecules but later adapted for chemical shift predictions. In a QM/MM calculation, the system is divided into two layers treated at different levels of theory. Chemical shifts are computed for a selected region of the biomolecule using a QM approach, while the surrounding environment is modeled using classical molecular mechanics. By iterating region by region, the entire molecule can be considered [39]–[41]. To accurately reproduce fully QM-derived chemical shifts, it has been shown that the QM region must include not only the residue of interest but also nearby functional groups, particularly hydrogen bond acceptors and

aromatic rings. In contrast, more distant residues can be treated classically, providing a good approximation of environmental contributions. Numerous benchmark studies have been conducted on small models, DNA, proteins, and protein-ligand systems, evaluating the impact of QM/MM parameters such as the level of theory in the QM and MM regions, QM region size, structural pre-optimization, solvation effects, and molecular dynamics [42]–[47].

2.1.2 *Conformational flexibility*

The challenge of chemical shift prediction arises not only from the choice of *ab initio* method but also from the molecular geometry used in the calculations. Typically, chemical shift computations are preceded by geometry optimization, which works well for small, rigid molecules. However, under solution-state NMR conditions, molecules sample multiple conformations over the course of an experiment, leading to observed shifts that represent a conformationally averaged value. To accurately reproduce NMR observables under these conditions, it is essential to identify the relevant conformations. A standard approach involves Boltzmann-weighted averaging of chemical shifts, where each conformer's contribution is weighted according to its free energy. The group of Grimme proposed an automated method combining CREST (Conformer-Rotamer Ensemble Sampling Tool) conformational sampling with Boltzmann-weighted shift predictions, enabling full NMR spectra calculations of non-rigid molecules [48], [49].

It was shown that errors in predicted shifts can arise from rovibrational motions, which are often neglected due to their high computational cost [16], [50]. These effects, while usually minor for rigid systems, become important for highly flexible molecules.

For biomolecules, dynamic effects are typically accounted for by averaging chemical shifts over structures sampled from classical MD simulations. These frames are usually selected either at uniform time intervals or randomly to ensure an unbiased representation of conformational dynamics. Various studies have demonstrated that considering ensembles of structures from such sampling methods significantly improves agreement with experimental results [51], [52].

Among different nuclei, ^{13}C chemical shifts are particularly sensitive to conformational changes, making them invaluable for structural elucidation. By cross-validating computed and experimental shifts, it is possible to determine the most populated conformations under experimental conditions. This approach has been widely applied to small molecules and extended to complex natural products, where experimental correlations help to confirm isomeric structures [53], [54]. Moreover, chemical shifts play an essential role in predict-

ing protein secondary structure, as they provide valuable insights into backbone and side-chain conformations [55], [56].

2.1.3 *Solvent effects*

Solvent effects can be incorporated into chemical shift predictions using either implicit or explicit solvation models. Comparisons between gas-phase calculations and continuum solvation models for solution-state NMR predictions generally show a slight improvement when the latter is used. While implicit solvation models can provide reasonable accuracy for non-polar molecules and solvents, they are insufficient for highly polar systems. In particular, explicit hydrogen-bonding interactions with the solvent are necessary to achieve reliable predictions for polar protons [52], [57], [58]. Exner et al. demonstrated that variations in chemical shifts across MD snapshots can reach 6 ppm for hydrogen and 50 ppm for nitrogen, primarily due to intermolecular interactions with solvent molecules [52].

In protein NMR, it has been shown that incorporating first-shell solvent molecules into QM/MM calculations improves predictions for labile protons compared to implicit solvation alone, while having a negligible effect on C-bound protons [46]. Despite these improvements, single-frame QM/MM calculations often perform poorly for N-bound protons. This is because they can form hydrogen bonds with highly dynamic solvent molecules and exchange protons with them, which cannot be captured by a static structural representation.

Zhu et al. also investigated the role of hydrogen bond cooperativity in chemical shift prediction. They found that hydrogen bonds, which do not interact with the proton of interest directly but only influence it through secondary hydrogen bonding networks, can slightly improve ^{1}H shift predictions as well [45]. A computationally efficient approach is to use an explicit solvent model for the solvation shell around the solute while employing an implicit solvation model to approximate the effects of distant solvent molecules [59].

Jin et al. studied solvent effects on chemical shift prediction in RNA using a QM/MM approach. They compared classical and polarizable force fields for MD ensemble averaging and found that the latter provided better accuracy due to its improved treatment of hydrogen bonding [47]. Similar trends were observed for proteins, where the use of polarizable force fields further enhanced chemical shift predictions [45].

A promising framework for predicting chemical shifts in solution was proposed by Atwi et al., integrating classical MD simulations with an automated selection of the most populated conformational clusters for further DFT-based chemical shift calculations. Their approach allows control over the solvation shell size and highlights the limitations of classical force fields as well [60]. A similar approach

has been applied to intrinsically disordered proteins (IDPs), where machine-learning-based clustering is used to extract relevant conformations from MD simulations for chemical shift calculations [61].

Dracinsky et al. demonstrated excellent agreement between experimentally measured chemical shifts for N-methyl acetamide and those calculated using DFT based on frames extracted from Car–Parrinello molecular dynamics (CPMD) simulations [62]. Achieving convergence for amide proton shifts required averaging over more than 600 MD frames, highlighting the necessity of extensive sampling. He further noted that the limitations of classical MD stem from its inaccurate description of hydrogen bond geometries and dynamics compared to *ab initio* MD. Another drawback of classical MD is its neglect of bond breaking and formation, which can influence averaged chemical shifts, particularly for labile protons [63].

So far, *ab initio* MD has been limited to small model systems due to its high computational cost. Future developments in force fields for biomolecules may improve the accuracy of MD-based chemical shift predictions. However, the large number of frames required, especially for flexible biomolecules, poses a computational challenge, preventing its routine application.

2.1.4 *Chemical shift changes upon chemical bond formation*

Quantum chemistry provides a fundamental understanding of chemical shielding, linking it to the electronic structure of molecules. Changes in the shielding tensor are mainly attributed to variations in the paramagnetic component (equation 4), while the diamagnetic component typically shows minimal alteration (equation 3). This distinction arises because diamagnetic shielding is predominantly determined by core atomic orbitals, whereas paramagnetic shielding is influenced by mixing ground and excited state molecular orbitals.

NMR chemical shifts can be rationalised using localised molecular orbitals (LMOs), which provide an intuitive framework for interpreting electronic interactions. LMOs represent atomic cores, lone pairs, and bonding interactions such as σ, π and δ bonds [6], [15], [64]–[67].

To illustrate, Viesser et al. applied this theory to study the effects of electron-donating and -withdrawing substituents on ^{13}C NMR chemical shifts in benzene derivatives. Their findings challenge the traditional organic chemistry explanation, which attributes variations in chemical shifts among benzene derivatives to changes in π electron density. Instead, they demonstrated that the σ-bonding framework plays a dominant role, with the resulting perturbations affecting only the paramagnetic component of the chemical shielding tensor, while the diamagnetic component remains largely unchanged [66].

In the context of drug discovery, ^{1}H NMR chemical shifts are of particular interest as they serve as highly sensitive reporters of the local environment, especially in the presence of hydrogen bond acceptors.

Theoretical studies have demonstrated that hydrogen bonding possesses a complex character, with electrostatic and covalent contributions playing equally significant roles [68]–[73]. One indicator of its covalency is J-coupling between the proton and the hydrogen bond acceptor [74], [75]. The ETS-NOCV (Extended Transition State-Natural Orbital for Chemical Valence) charge and energy decomposition scheme [76] further differentiates charge deformation in classical hydrogen bonds into a dominant σ charge transfer contribution and a significantly weaker resonance-assisted π-polarization component, whose magnitude depends on the specific nature of the system. For clarity, a hydrogen bond is symbolically represented as $X - H \cdots Y$, where X and Y denote N/O hydrogen bond donors and acceptors, respectively. The σ charge transfer component involves electron donation from the lone pair of the acceptor (Y) to the empty σ^*_{X-H} orbital of the donor. In contrast, the π contribution corresponds to electron flow from the π orbital of the donor (X) to the antibonding orbital of the acceptor (Y) [72], [73].

Regarding chemical shifts, hydrogen bond formation typically results in characteristic proton deshielding. A correlation between hydrogen bond strength and chemical shift changes has been documented in the literature [2], [77]. Sutter et al. investigated the impact of classical hydrogen bonds on proton NMR observables [78]. Their detailed study examined how the shielding tensor components vary across different hydrogen-bonded systems. In contrast to previous ^{13}C NMR chemical shift analyses, they observed that in ^{1}H NMR, variations occur in both the paramagnetic and diamagnetic shielding components, with both consistently contributing to greater deshielding as hydrogen bond strength increases. Their findings indicate that deshielding arises primarily from $\sigma(X - H)$, with the extent of deshielding strongly dependent on the elongation of the $X - H$ bond upon hydrogen bond formation. This effect is further amplified by an additional σ contribution from the lone pair of the hydrogen bond acceptor. A π component centred on the hydrogen bond acceptor also plays a role, though it slightly opposes the deshielding effect of the σ components, thereby inducing a minor shielding contribution. Furthermore, orbitals on adjacent atoms can influence the chemical shift, particularly in systems exhibiting resonance-assisted hydrogen bonding. These observations align with the hydrogen bond analysis conducted by Mitoraj et al. [72], [73].

When an aromatic ring serves as a weak hydrogen bond acceptor, the resulting proton NMR signal exhibits shielding rather than the deshielding typically associated with hydrogen bonding. From a drug discovery perspective, $C - H \cdots \pi$ contacts are particularly impor-

tant since aliphatic-aromatic proximities have been identified as the most frequently observed type of contact in protein-ligand complexes across the PDB database [79]. The nature of weak $C-H\cdots\pi$ interactions is primarily governed by London dispersion forces, though electrostatic and charge-transfer contributions also play a role [80], [81], Chapter 4. Due to their weak hydrogen bond character, $C-H\cdots\pi$ interactions induce only minimal deshielding in NMR. However, this effect is largely masked by the shielding contribution from the ring current effect, which is induced by the external magnetic field acting on the aromatic ring [2]. Consequently, a proton within a π system environment experiences net shielding. This phenomenon has been extensively discussed by Scheiner in small model systems involving $O-H\cdots\pi$ interactions [82] and further analysed in protein-ligand complexes in our study, Chapter 4.

2.2 DATA-DRIVEN METHODS

With advances in data science, machine learning methods (ML) have emerged as a powerful alternative to first-principles calculations for molecular properties. For chemical shifts, ML-based approaches have achieved gold-standard accuracy at a significantly reduced computational cost. Their applicability extends beyond small molecules, encompassing larger and more complex systems such as biomolecules and solid-state materials [83]–[87]. When developing predictive models for chemical shifts, or other molecular properties, three key components must be carefully selected: (i) the choice of an experimental or computational database containing the target molecular properties, (ii) the molecular representation, and (iii) the predictive algorithm [88].

A sufficiently large and well-curated chemical shift database is essential for accurate predictions. For small molecules, these are often derived from state-of-the-art CCSD(T) calculations [89], [90]. In contrast, predictive models for biomolecules typically rely on experimental data, given the prohibitive cost of generating high-quality computational datasets and the additional complexity introduced by solvent effects and molecular dynamics. A major challenge in this case is data curation, as datasets may contain errors due to human mistakes, variations in experimental conditions, and inconsistencies in referencing standards [91].

The representation of molecular structures is another crucial factor in chemical shift prediction. Haghighatlari et. al. point out that an ideal descriptor should be universal, extendable to out-of-distribution systems, and computationally efficient [88]. Various molecular representations have been applied, differing in their level of physical detail. Examples range from simple models, such as the Coulomb matrix descriptor, which has been successfully used for chemical shift predic-

tion in organic molecules [92], [93], to more advanced representations, such as atomic environment vectors (AEV) [89], [94]. For solids, models like atom-centered Gaussian density [86] and smooth overlap of atomic positions (SOAPs) [87] have been employed. In proteins, structural features such as dihedral angles, hydrogen bond geometries, and empirical relationships [95]–[97] serve as descriptors in models like SPARTA+, SHIFTX2 or UCBShift [83], [85], [98].

In terms of predictive algorithms, various neural network architectures are widely employed, particularly for small molecules in the gas phase and solution [89], [99]–[101] as well as for proteins [83], [84] and solid-state NMR [86], [87]. Kernel ridge regression has also been successfully applied to chemical shift prediction [102], [103]. Recent studies have found random forest regression to be the most effective approach for protein chemical shift prediction [85], [98].

Hybrid approaches also exist, such as SHIFTX2 and UCBShift (Chapter 5), which combine feature-based machine learning with sequence-based transfer chemical shifts from the database [85], [98]. Another non-standard approach is ΔML, which integrates geometrical descriptors and neural networks with low-cost DFT shielding calculations to bridge the gap between low- and high-level DFT [89]. This method has even been extended to achieve CCSD(T) accuracy [101] by incorporating geometrical, electronic, and magnetic properties as descriptors [90]. Another non-standard model is iShiftML, which learns diamagnetic and paramagnetic contributions separately by utilising a novel tensor environment vector representation. Designed for chemical shift prediction in small organic molecules, iShiftML has demonstrated high transferability, extending its applicability to natural products [104].

2.3 COMPARISON OF MODELS

Several data-driven protein chemical shift predictors, such as SHIFTS, SPARTA+, SHIFTX and SHIFTX2, have been validated against QM/MM and fragment-based ADMA approaches for proteins, generally outperforming quantum methods, particularly for labile protons [44], [46], [105], [106]. However, some studies have identified specific limitations. For instance, Zhu et al. investigated amide proton shifts in two proteins using QM/MM DFT and compared the results with SHIFTX2 and SPARTA+. Both ML methods outperformed QM/MM in terms of RMSE, MUE, and correlation coefficient (R) when compared to experimental data. However, the slope of the regression line was 1 for the QM/MM approach, while SHIFTX2 and SPARTA+ yielded slopes of 0.8 and 0.7, respectively, suggesting potential bias in ML predictions [45]. Similarly, Sumowski et al. studied the sensitivity of *ab initio* and several data-driven models to structural changes and environmental effects. While the *ab initio* chemical shifts of carbon

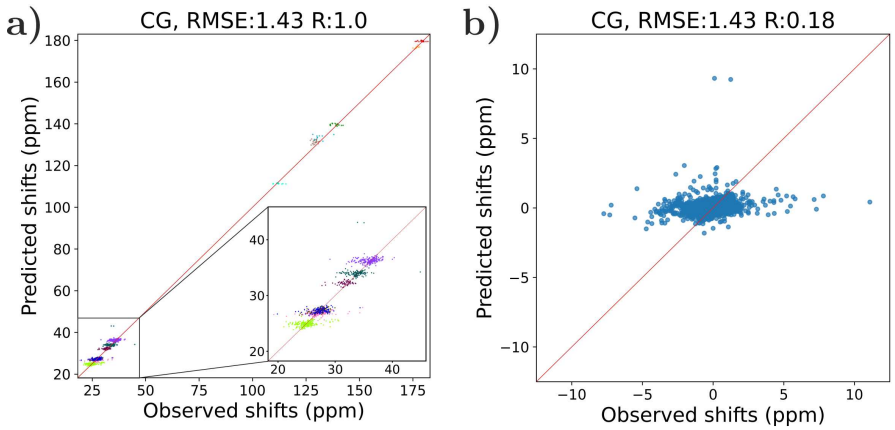


Figure 1: a) Predicted vs. observed C_γ chemical shifts using SHIFTX2 b) Same as (a), but with predicted chemical shifts adjusted by subtracting amino acid-specific random coil values. Panel (b) adapted from [108]. Tested on the dataset described in Chapter 5.

and nitrogen varied significantly upon conformational changes, the results from empirical models remained largely unaffected [107]. This insensitivity can be attributed to the fact that ML predictors rely on X-ray crystal structures, which do not account for protein flexibility in solution-state NMR. As a result, empirical models show reduced sensitivity to subtle structural variations, limiting their reliability for structural validation [107]. Another drawback of data-driven predictors is their lack of transferability to non-standard residues, such as post-translational modifications or protein-ligand systems.

An additional challenge in chemical shift prediction is the potential misinterpretation of accuracy metrics. Protein ML chemical shift predictors are expected to capture variations from random coil values for each atom type, which arise due to conformational changes and interactions with the environment. Papers often report Pearson's correlation coefficient (R) as an indicator of prediction accuracy. However, for SHIFTX2, the reported correlation coefficients are nearly perfect (R = 1.0) for most atom types [98]. This misleading result arises because chemical shifts are primarily determined by amino acid identity, and the reported R values are computed for groups of amino acids rather than for each amino acid type separately. This is a common issue in the scientific literature, known as *spurious correlation* [109]. As demonstrated in Figure 1, SHIFTX2 fails to predict variations within different amino acid types. The model merely reproduces the average random coil value for each amino acid, information that does not necessitate machine learning prediction. Similar statistical misinterpretations have also been observed in other chemical shift prediction studies, including those based on *ab initio* approaches [105], [110], [111].

As highlighted in this introduction, despite advancements in *ab initio* chemical shift prediction, it remains a significant challenge in computational chemistry, particularly for biomolecular systems. The accuracy of such predictions depends on multiple factors, including structural representation, solvent treatment, and the choice of theory level. Encouragingly, QM/MM approaches offer a viable solution for biomolecules, given the local nature of chemical shifts. The choice of approach should be carefully tailored to the specific problem at hand. Additionally, further validation of novel ML-based methods against *ab initio* studies remains essential to ensure their accuracy and reliability.

2.4 SUMMARY OF THE THESIS

This dissertation explores two strategies for chemical shift prediction: *ab initio* methods to gain insights into protein-ligand interactions and a novel data-driven approach for predicting protein side-chain chemical shifts. Chapter 3 focuses on understanding how a drug molecule interacts with its target in solution. By validating computed and experimental ^{1}H chemical shifts in the model system, we identified significant discrepancies between the X-ray structure and solution-state representation. Incorporating classical and QM/MM MD-based averaging allowed us to *refine* the static X-ray structure, yielding a more accurate solution-state model. Building on this approach, in chapter 4, we investigated the factors that perturb ^{1}H chemical shifts of biotin upon streptavidin binding. Additionally, using DFT- and CCSD(T)-based chemical bond descriptors, we quantified and elucidated how specific amino acids in the binding pocket interact with the ligand. Extending beyond *ab initio* methods, Chapter 5 introduces a machine learning-based protein side-chain chemical shift predictor, which integrates a machine learning model trained on engineered features from crystal structures alongside a structure- and sequence-based transfer module.

Part II

SHOWCASE OF PUBLISHED WORK

The core of this cumulative thesis consists of two scientific articles published in peer-reviewed journals and one manuscript currently under revision. Aside from the continuous page numbering, the original layout design should be considered the intellectual property of the respective journals. Supporting Information materials are included in their original format as well. Each chapter corresponds to one article, preceded by its digital object identifier (DOI) and a brief statement outlining my personal contribution.

3

LIGAND ^{1}H NMR SHIFTS AS REPORTERS OF PROTEIN-LIGAND INTERACTIONS

G. Platzer, A. L. Ptaszek, J. Böttcher, *et al.*, "Ligand 1h nmr chemical shifts as accurate reporters for protein-ligand binding interfaces in solution," *ChemPhysChem*, vol. 25, no. 1, e202300636, 2024 [112]

DOI: 10.1101/2021.11.30.470546

PERSONAL CONTRIBUTION

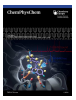
Simulations, data analysis, validation, writing, visualization, shared First Author with G. Platzer

ChemPhysChem

Research Article
doi.org/10.1002/cphc.202300636

Chemistry Europe
European Chemical Societies Publishing

www.chemphyschem.org



Ligand ^{1}H NMR Chemical Shifts as Accurate Reporters for Protein-Ligand Binding Interfaces in Solution**

Gerald Platzer[+],*[a, b] Aleksandra L. Ptaszek[+],[a, c] Jark Böttcher,[d] Julian E. Fuchs,[d] Leonhard Geist,[d] Daniel Braun,[a] Darryl B. McConnell,[d] Robert Konrat,[a] Pedro A. Sánchez-Murcia,*[c] and Moriz Mayer*[d]

The availability of high-resolution 3D structural information is crucial for investigating guest-host systems across a wide range of fields. In the context of drug discovery, the information is routinely used to establish and validate structure-activity relationships, grow initial hits from screening campaigns, and to guide molecular docking. For the generation of protein-ligand complex structural information, X-ray crystallography is the experimental method of choice, however, with limited information on protein flexibility. An experimentally verified structural model of the binding interface in the native solution-state would support medicinal chemists in their molecular design decisions. Here we demonstrate that protein-bound ligand ^{1}H NMR chemical shifts are highly sensitive and accurate probes for the immediate chemical environment of protein-ligand interfaces. By comparing the experimental ligand ^{1}H chemical shift values with those computed from the X-ray structure using quantum mechanics methodology, we identify significant disagreements for parts of the ligand between the two experimental techniques. We show that quantum mechanics/molecular mechanics (QM/MM) molecular dynamics (MD) ensembles can be used to refine initial X-ray co-crystal structures resulting in a better agreement with experimental ^{1}H ligand chemical shift values. Overall, our findings highlight the usefulness of ligand ^{1}H NMR chemical shift information in combination with a QM/MM MD workflow for generating protein-ligand ensembles that accurately reproduce solution structural data.

Introduction

Reversible binding of organic molecules to a target protein is driven by a multitude of non-covalent interactions.[1] Although the contribution of an individual interaction to the overall binding energy can be small, the cumulative effect of multiple interactions can lead to binding strengths approaching those of covalent bonds.[2] A prime example is the binding of streptavidin to biotin/avidin featuring a multitude of weak, non-covalent interactions culminating in the highest observed binding affinity of a small molecule to a protein.[3] Exploiting putative interactions in a binding pocket with maximum efficiency is a major objective in medicinal chemistry to yield highly potent binders. Improving the potency of initial hits relies on gradually altering a starting compound and monitoring the effect on the compound's potency, a process commonly referred to as classical structure-activity relationship. The continuous modification of a compound is monitored in the binding affinity of the altered compound to the target protein, a readout that is of global nature (reflecting the total Gibbs free energy of binding ΔG).[4] One downside of this approach is that information of individual contributions with the protein can be difficult to isolate.[5] However, it is precisely the cumulative effect of many, often-times small additive effects to the overall ΔG of binding that determine if a ligand can subsequently engage with its target protein in a selective manner. Utilizing these weak non-covalent interactions to their fullest potential is one of the objectives of an efficient optimization process. An experimentally validated structural model that allows for an accurate quantification of the individual contributions of non-covalent interactions (NCIs) is an invaluable input for medicinal chemistry, enabling rational design ideas based on the local chemical environment of the protein-drug interface. To achieve this, a variety of structure-based approaches are commonly employed to aid in drug discovery.[6] A reliable structural model

[a] Dr. G. Platzer,[+] A. L. Ptaszek,[+] Dr. D. Braun, Prof. Dr. R. Konrat
Christian Doppler Laboratory for High-Content Structural Biology and Biotechnology, Department of Structural and Computational Biology, Max Perutz Labs, University of Vienna, Campus Vienna Biocenter 5, 1030-Vienna, Austria
E-mail: gerald.platzer@mag-lab.eu

[b] Dr. G. Platzer[+]
MAG-LAB GmbH, Karl-Farkas-Gasse 22, 1030-Vienna, Austria

[c] A. L. Ptaszek,[+] Prof. P. A. Sánchez-Murcia
Laboratory for Computer-Aided Molecular Design, Division of Medicinal Chemistry, Otto Loewi Research Center, Medical University Graz, Neue Stiftingtalstrasse 6/III, 8010-Graz, Austria
E-mail: pedro.murcia@medunigraz.at

[d] Dr. J. Böttcher, Dr. J. E. Fuchs, Dr. L. Geist, Dr. D. B. McConnell, Dr. M. Mayer
Boehringer Ingelheim RCV GmbH & Co. KG, Dr. Boehringer Gasse 5–11, 1121-Vienna, Austria
E-mail: moriz.mayer@boehringer-ingelheim.com

[+] These authors contributed equally to this work.

[**] A previous version of this manuscript has been deposited on a preprint server (DOI: https://doi.org/10.1101/2021.11.30.470546)

Supporting information for this article is available on the WWW under https://doi.org/10.1002/cphc.202300636

ChemPhysChem
Research Article
doi.org/10.1002/cphc.202300636
Chemistry Europe
European Chemical Societies Publishing

is essential for every step in the hit-to-lead optimization process that uses some form of structural information as input.

Several theoretical and experimental methods are commonly employed to obtain protein 3D structural information. Among these, X-ray crystallography is currently the primary experimental method used to derive protein structural information in structural biology and to characterize protein-ligand complexes. However, X-ray crystallography has limitations in that it is a static technique, providing a snapshot of the molecule in a specific conformation, which may not fully capture the dynamic behavior of the molecule or protein.[7] In addition, the presence of a crystal lattice can induce additional rigidity, conformational constraints,[8] and in rare cases even different binding modes.[9] Cryo-electron microscopy (cryo-EM) has emerged as a powerful alternative diffraction-based method that can generate structures of proteins in various conformational states. The application in drug design cycles is continuously evolving with still some limitations regarding throughput and molecular size of the target protein.[10] Theoretical approaches for structure generation include homology modeling[11] (when applicable) and structure generation via deep neural networks, (e.g. AlphaFold, RoseTTAFold)[12] which, do not require experiments for the generation of novel structures. While initially limited to single protein chains, protein-protein complex prediction has become available recently.[13] Given the fast pace the field is evolving it can be expected that the generation of protein-ligand complex structures will become feasible in the foreseeable future.

Nuclear Magnetic Resonance (NMR) has been used for structure determination in the past and is still being used in corner-case scenarios where diffraction methods fail, or dynamic processes are to be analyzed. However, the process is technically challenging and time-consuming and usually not competitive with X-ray and cryo-EM. The main strength of NMR is access to site-specific information about the chemical environment in solution. In the context of drug-discovery, NMR can offer detailed information about the protein-drug interface in solution which can assist in the generation of a structural model that closely resembles the native state.[14] In this work we highlight the use of ligand ^{1}H chemical shift (CS) values as probes for the molecular details of a given protein-ligand interface which has several advantages. Firstly, it nicely supplements structural information derived from diffraction data, which is sensitive to heavy atoms, but "blind" to hydrogens. ^{1}H nuclei are naturally present in organic material and are inherently NMR-active eliminating the need for isotope labeling. In addition, they are highly sensitive probes for intermolecular interactions, most notably the formation of hydrogen bonds and the interaction with the π-electrons of nearby aromatic ring systems.[15] Additionally, the use of ligand proton information does not require time-consuming assignment of protein signals while the assignment of ^{1}H ligand signals is fast, straightforward, and routinely used in analytical chemistry. On top of that, the most abundant ^{1}H reporter protons on the protein-side are methylene protons of beta and gamma sidechain carbons, which often suffer from significant peak broadening due to their unfavorable relaxation properties.[16] Finally, bound ligand protons are near the protein surface where they usually cover a large portion of the total protein-ligand interaction surface. Given the extensive information content, we set out and explored the potential use of ligand ^{1}H shifts for the refinement of the immediate protein-ligand interface.

The relationship between the CS and protein 3D structure is well established but has not been exploited in the context of drug design and is only rarely discussed with respect to protein-ligand interactions in general.[17] Previously, we reported on the interaction of protons of protein sidechains engaged in CH–π interactions with aromatic ring systems of ligands, based solely on ^{1}H CS values of the interacting protein aromatic ring protons. The method can give site-specific information on protein CH-groups and their interaction with ligand π-systems in a protein-detected manner.[18] This study emphasizes the usefulness of ^{1}H CS information in understanding and eventually refining protein-ligand complex structures in solution. We chose the PWWP1 domain of the histone-lysine N-methyltransferase NSD3, a protein implicated in several forms of cancer,[19] as a model system. Potent NSD3 inhibitors binding to the PWWP1 domain have been recently reported.[20] In this work we use the chemical probe molecule BI-9321 (named Ligand 1 here) (Figure 2c) which binds to the PWWP1 domain of NSD3 with a SPR K_d of 170 nM. (Figure 2a/b, PDB ID 6G2O)[20] We demonstrate that quantum mechanics calculations, in combination with protein-ligand ensembles generated by QM/MM MD, can accurately reproduce the ^{1}H CS data obtained via solution NMR. Taken together our work highlights the usefulness of ligand ^{1}H CS information for understanding and validating the interactions of ligands with their target proteins that can significantly benefit early-stage drug discovery and hit-to-lead optimization.

Results

Figure 1 shows the different workflows we applied. Starting points for this approach are structural models that can be derived from various experimental or theoretical sources, most prominently via diffraction-based methods (X-ray, Cryo-EM) or theoretical approaches (e.g. AlphaFold, RoseTTAFold).

A large number of protons cover the interaction surface, both on the side of the protein and the ligand (Figure 2c/d). Backbone NH shifts are often used to characterize ligand binding, however only 2 backbone amide groups are within a 4.0 Å cutoff distance to the ligand as opposed to 37 carbon-bound protons within the same distance (Figure 2c). This highlights the very good surface-coverage of non-exchangeable protons for the analysis of protein/ligand interfaces. However, the use of proton resonances on the protein-side suffers from two major drawbacks. First, the necessity for a signal assignment, and second, most potential ^{1}H reporter protons are methylene protons of beta and gamma sidechain carbons, which suffer from significant peak broadening due to their unfavorable relaxation properties.[16] Bound ligand protons ^{1}H CS offer a promising alternative, and in the case of the PWWP1 domain, non-exchangeable ligand protons can be used to

ChemPhysChem
Research Article
doi.org/10.1002/cphc.202300636
Chemistry Europe
European Chemical Societies Publishing

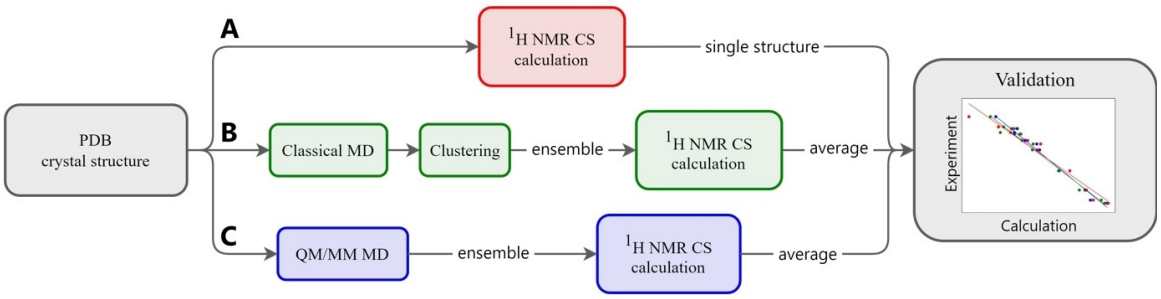


Figure 1. Scheme of the applied methodology for ^{1}H CS calculation: path **A** using the X-ray structure, path **B** using classical forcefield-based MD ensembles, and path **C** using QM/MM MD ensembles.

investigate the immediate chemical environment of the protein-ligand interface in solution (Figure 2a/d).

Starting with the X-ray crystal structure as a reference model, we employed the gauge-independent atomic orbital (GIAO) method combined with density functional theory (DFT)[21] to extract chemical shielding values. We then compared these results with experimentally determined NMR ^{1}H CS data that was obtained using [^{1}H,^{1}H]-NOESY spectroscopy of a 1:1 protein-ligand complex in a fully deuterated protein background (Section S2). Discrepancies, especially for protons involved in hydrogen bonding interactions, were observed. Recognizing that a single structure might not adequately represent the solution state of the protein-ligand interaction, we turned to molecular dynamics (MD).[15,21c, 22] The aim was to create an ensemble of structures that is able to provide a better representation of the native protein state in solution. We generated a classical MD ensemble and validated the computed ensemble-based ^{1}H chemical shifts (CS) by comparing them with NMR solution data. Although classical MD provided structural models that were more consistent with the NMR data, some non-covalent interactions (NCIs) could not be accurately reproduced. While contributions to the CS arising from nearby ring-systems were now better represented (less directionality), hydrogen bond interactions with classical hydrogen bond acceptors (CH···O) were still not fully captured (more directionality). This is due to the simplicity of classical force fields that approximate non-covalent interactions by electrostatic and van der Waals terms, and atomic charge distributions as time-invariant spheres. In a second step, and to ensure the inclusion of local adaptations and molecular intricacies of the protein-ligand interface we turned to a hybrid quantum mechanics/molecular mechanics QM/MM MD approach.[23] The resulting ensemble of structures demonstrated further improvement and was able to address hydrogen bond interactions more precisely than classical MD.

Calculation of ^{1}H-Chemical Shielding

X-Ray Structure

To evaluate the quality of any structural model it is necessary to convert the individual ^{1}H positions in the 3D structure into the physical equivalent of the NMR CS. This is done by calculating the chemical shielding value at each proton position in the 3D model using gauge-independent atomic orbital (GIAO)[21a,24] in conjunction with density functional theory (DFT) at the wB97X-D[25]/def2-TZVP theory level.[26] This allows for comparison of structural models with solution NMR CS data.

The CS values for each of the 12 positions highlighted in Figure 2a are provided in the SI. The experimental data was then compared with calculated shielding data for the same ligand proton positions of the structural model. Figure 3 shows the linear regression of experimental chemical shift vs. calculated chemical shielding for all 12 proton signals using the X-ray structure (PDB ID 6G2O). Frequently ambiguous sidechain amide orientations of Gln and Asn in contact with the ligand were checked for correctness and residues Gln316 and Gln367 were modified (180 flip) accordingly.[27] The three proton signals of the methyl groups are magnetically equivalent and thus converge to a single signal (computationally considered via rotameric averaging).

In principle, there is good overall agreement between the back-calculated bound ^{1}H CS values based on the X-ray structure and the experimental NMR ^{1}H CS values, judging from the R^2. However, some individual positions (1, 2, 3 and 12) show deviations of over 0.5 ppm. The largest outlier is position 1, with a calculated CS overestimated by almost 2 ppm. Interestingly, the three protons with the largest deviations all form hydrogen bonds in the X-ray structure: a CH···π interaction at positions 3 and 12 CH···O interactions at positions 1 and 2. Importantly, the electron densities for both the ligand and the protein are well-defined and can therefore be excluded as the source of the observed deviation (SI Figure 2). These results were consistent regardless of the DFT functional used for the chemical shielding calculation (Figure S5, Section S5).

ChemPhysChem
Research Article
doi.org/10.1002/cphc.202300636
Chemistry Europe
European Chemical Societies Publishing

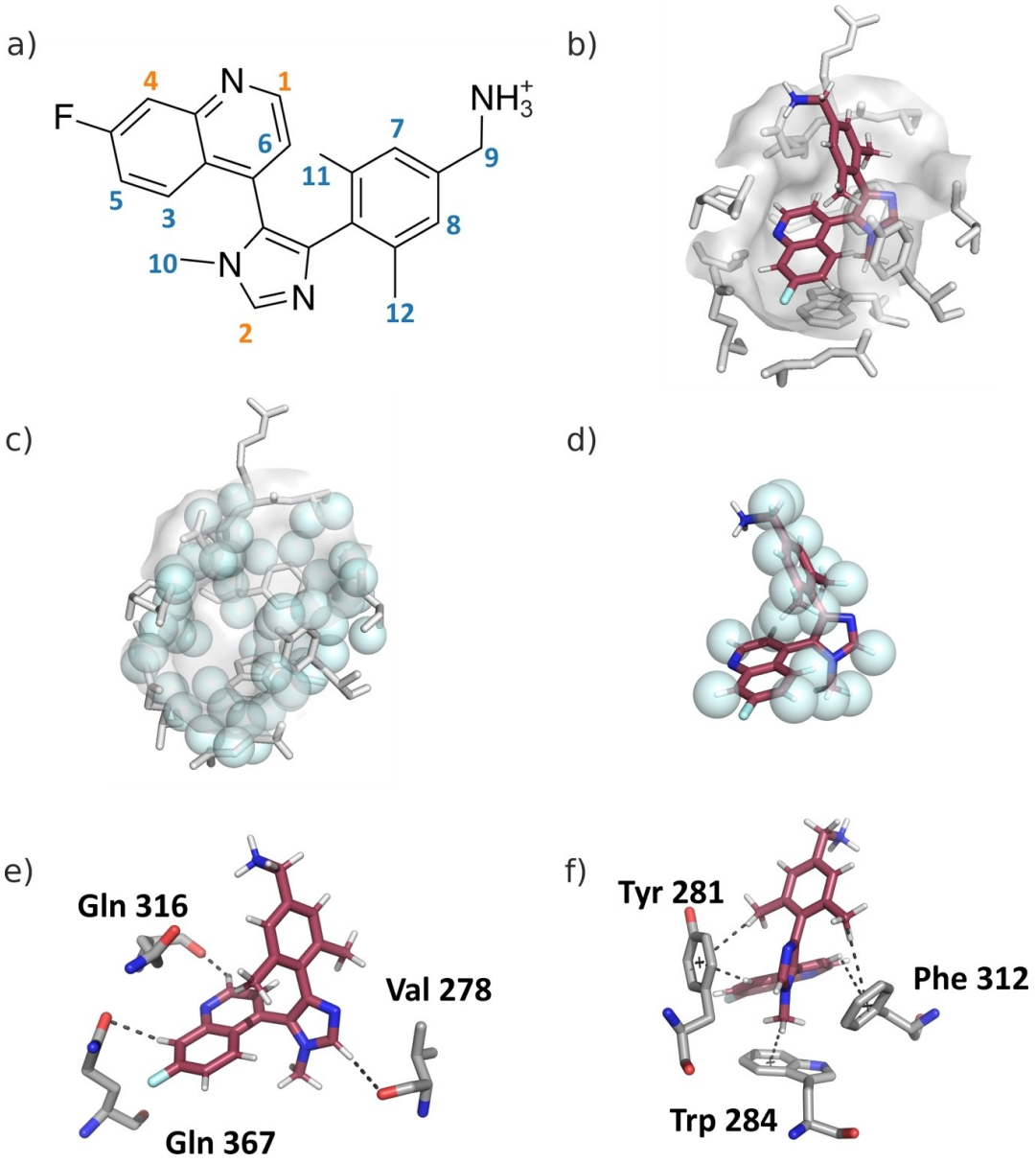


Figure 2. a) Chemical structure and proton signal assignment of ligand 1 (BI-9321).[20] b) Binding mode of ligand 1 to the PWWP1 domain of NSD3 (PDB ID 6G2O). c) Distribution of the 37 non-exchangeable protein protons within 4.0 Å of ligand 1 (light blue spheres). d) Non-exchangeable ligand protons (light blue spheres). e) Non-exchangeable ligand 1 proton interactions with hydrogen-bond acceptors of the protein. f) Ligand 1 protons interactions with aromatic groups.

Classical MD-Derived Structural Ensemble

From the previous subsection, we can see that the crystal structure does not fully represent the behaviour of the protein-ligand system in solution, the condition of our NMR measurements. The protein-ligand structure obtained by soaking potentially biases the conformational flexibility towards a static representation and the binding site is particularly exposed to crystal packing forces, which can distort the structure observed in solution. Following the recent literature, we explore the impact of protein dynamics with the aim of generating structural ensembles that better reproduce experimental NMR solution data.[15,21c, 22]

First, we employ classical MD (cMD) simulations in explicit solvation to highlight the impact of structural dynamics on the ^{1}H CSs (Section S3). We applied an Amber ff19SB[28] force field for the protein and the Generalized Amber Force Field (GAFF)[29] for the ligand, which is a common approach for this type of

ChemPhysChem
Research Article
doi.org/10.1002/cphc.202300636
Chemistry Europe
European Chemical Societies Publishing

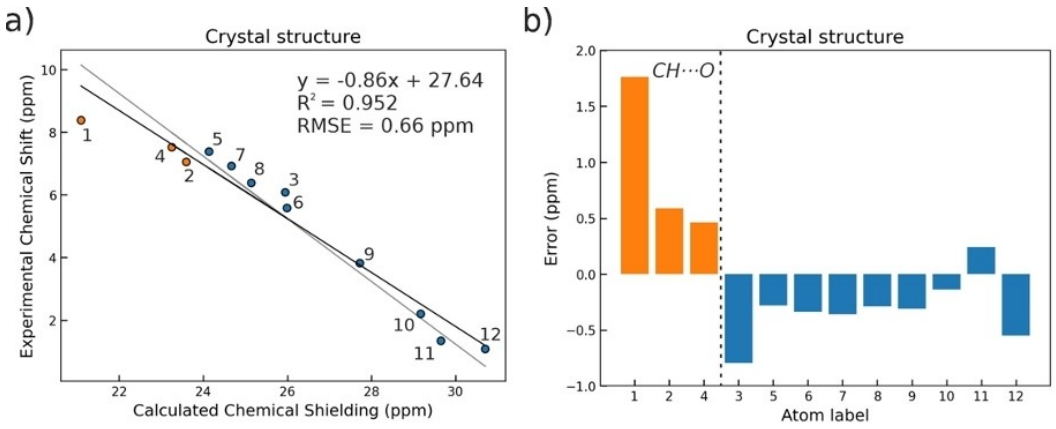


Figure 3. a) Correlation between experimental NMR ^{1}H CSs (ppm) and calculated chemical shielding (ppm) from the crystal structure (PDB ID 6G2O). The grey line corresponds to an ideal regression with slope = −1. b) Absolute error (ppm) of calculation vs. NMR experiment. Non-exchangeable proton interactions with protein hydrogen-bond acceptors are indicated in orange.

system. The use of cMD-ensembles reduces the root mean square error (RMSE) by ~40% to a value of 0.39 ppm (Figure 4). We observe that the errors of protons 1, 3, 4 and 12 are now significantly reduced. Interestingly, proton 4 shows a much broader distance distribution with Gln367 (Section S3). Closer inspection of the MD trajectory revealed alternating hydrogen bonding to the hydroxy group of Ser314 and the amide oxygen on Gln367, hence only marginally affecting the chemical shift of proton 4 (Figure 4, Figure S4). Surprisingly, the absolute error of proton 2 increases to 0.9 ppm. When analysing the cMD trajectories, we found that the distance between proton 2 and the carbonyl oxygen of Val278 was shortened, resulting in a significant deshielding effect on proton 2 (Figure S3 and Figure 4a). This discrepancy is related to the limitations of classical force fields, which approximate hydrogen bonding by a combination of electrostatic and van der Waals components, and use fixed partial charges centred on atoms, neglecting electronic polarisation. Additionally, force fields are parametrized based on a limited training data set (experimental or high-level calculations), so they may not precisely capture all the subtle protein-ligand interactions.[30]

Beyond Classical MD: QM/MM MD-Derived Structural Ensemble

Considering the limitations of the cMD approach, we extend the cMD treatment with short QM/MM MD simulations to generate more accurate geometries for the ^{1}H CSs calculations (Section S4). This method allows for structural refinement according to quantum mechanical energy functions of atoms in the binding mode, while the rest of the system is treated with the force field, details in Section S4. The QM/MM MD ensemble-averaged ^{1}H CSs are compared with the experimental NMR data in Figure 4c. All the statistical parameters are found to be improved compared to the cMD (RMSE is reduced to 0.28 ppm).

Interaction Details: Static Structure versus Dynamic Ensembles

The NMR chemical shift is an ensemble-averaged observable reporting on all occupied protein-ligand conformations populated in solution. MD-averaging offers a way to account for this structural heterogeneity. Geometrical changes as consequence of these dynamics result in changes to the ^{1}H CSs of protons in the affected area. These changes are most pronounced when the orientation of a nearby aromatic ring-system is altered, or a hydrogen bond is formed or broken. To rationalize the effect that the use of cMD and QM/MM-ensembles has on the calculation of ligand ^{1}H CSs, we first analysed the distance distributions of all protons on ligand 1 to the interacting moieties on the protein along the different trajectories (Figure S3) and compared these values to the ones of the X-ray structure. Protons 3, 5 and 6 largely maintain their original interaction distance with the protein, all three interacting with nearby π systems. In addition, cMD and QM/MM seem to cover these CH···π interactions equally well. Protons 10, 11 and 12 all form CH_3···π interactions with the protein, slightly weaker in interaction strength,[31] but similar to CH···π interactions in terms of molecular mechanism (Figure 2f). In all three cases cMD and QM/MM are in good agreement with the X-ray structure and binding site distances relevant for ligand chemical shifts are preserved (Figure S3). It should be noted that even small positional rearrangements in the vicinity of π-systems can lead to substantial changes in the observed ^{1}H CSs of the interacting protons.[18] Protons 1,2 and 4 make up the second interaction type, all three forming CH···O hydrogen-bonds interactions with the protein (Figure 2e,f). It is this group for which cMD and QM/MM methods generate largely different results in terms of interaction distances.

To visualize the underlying molecular details of the discrepancy between the two theoretical approaches, we performed non-covalent interaction (NCI) analysis around proton 1 on some representative frames from the QM/MM MD simulations (Figure 5, Section S6).[32] In NCIPLOT, energetically

ChemPhysChem
Research Article
doi.org/10.1002/cphc.202300636
Chemistry Europe
European Chemical Societies Publishing

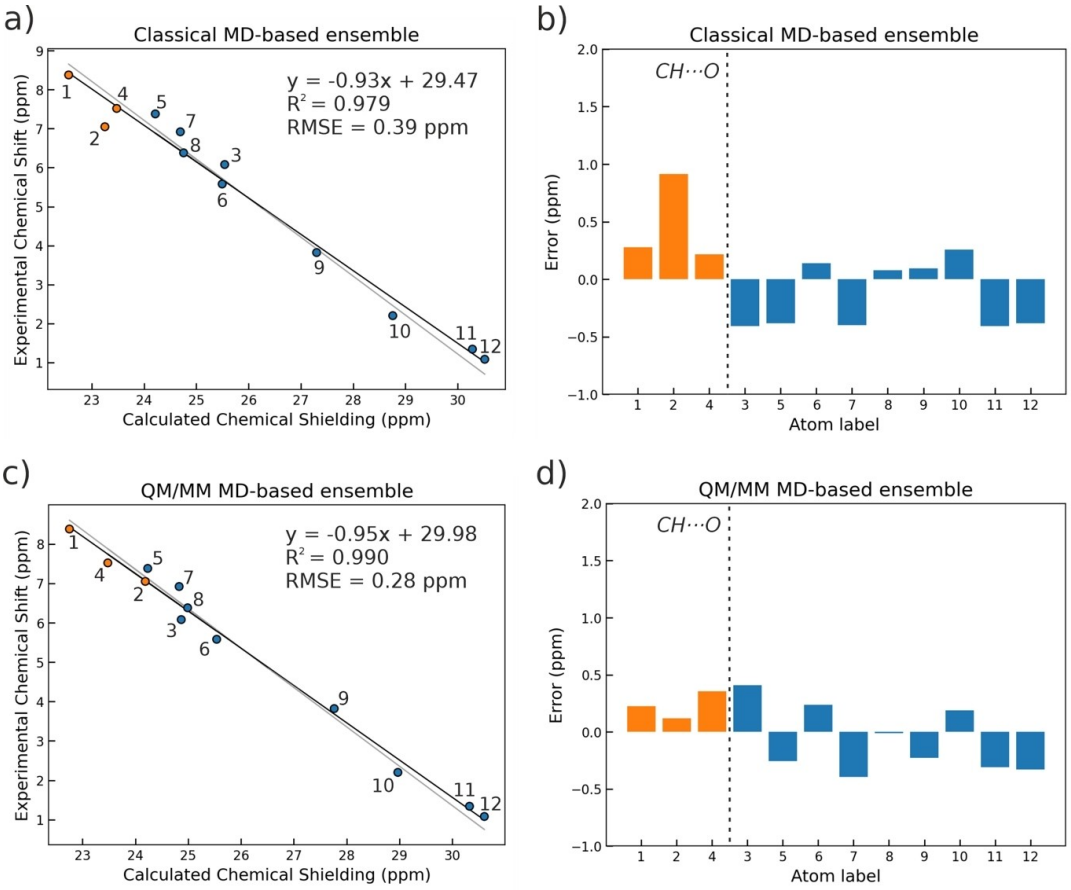


Figure 4. Correlation between experimental ^{1}H CSs (ppm) and calculated chemical shielding (ppm) using a) cMD and c) QM/MM MD-based ensembles. The grey line corresponds to an ideal regression with slope = −1. Absolute error (ppm) of calculation vs. NMR experiment are shown for b) cMD and d) QM/MM MD-based ensembles (average value). Non-exchangeable proton interactions with protein hydrogen-bond acceptors are indicated in orange.

relevant contacts are represented via a color-coded iso-surface between the interacting atoms (Figure 5c/d). NCI analysis on the crystal structure reveals the presence of a weak CH···O hydrogen bond (blue disk, Figure 5c) between proton 1 and the backbone oxygen atom of Gln316 (~2.5 Å). Furthermore, a CH···N hydrogen bond in addition to weak van der Waals interactions between the amino group of the ligand and the amide group of Gln316 (blue and green NCI iso-surfaces, respectively) are identified. In comparison, the former hydrogen bond with Gln316 is lost in the QM/MM ensemble (Figure 5d) and only weak interactions between the ligand and Gln316 are preserved. It is the loss of the CH···O hydrogen bond of proton 1 to the backbone oxygen of Gln316 (Figure 5a, line vs. distribution) which results in an apparent upfield shift (increased shielding and a resulting shift to the higher ppm values) of proton 1 compared to the X-ray structure (Figure 5b, line vs. distribution).

While for proton 1 the different outcomes for cMD and QM/MM can be explained based on dynamic phenomena alone, this is not the case for proton 2. Based on X-ray and cMD, proton 2 establishes a CH···O hydrogen bond with the backbone oxygen of Val278 (H···O distance of ~2.5 Å and light green/blue NCI iso-surface for the interaction). In the QM/MM-ensemble however, this interaction is lost and proton 2 is interacting with the isopropyl group of Val278 via homopolar CH···HC interactions or σ···σ stacking instead (blue/yellow NCI iso-surface). The nature of such interactions is poorly understood, but London dispersion forces are their primary contributor.[33] While the CH···O interaction of proton 2 with Val278 is maintained in the cMD, the QM/MM MD can capture the alternative interaction mode (the formation of an intramolecular hydrogen bond (NH···O) with the amide nitrogen of Tyr281), resulting in a much better agreement with the experimental data (Figure 6).

Discussion

Protein-bound ligand ^{1}H CSs are seldomly looked at even though they are exquisitely sensitive and contain a wealth of information on how ligands interact with the protein. The use of ^{1}H CS information has several benefits for the characterization of protein-ligand interactions. First and foremost is the

ChemPhysChem
Research Article
doi.org/10.1002/cphc.202300636
Chemistry Europe
European Chemical Societies Publishing

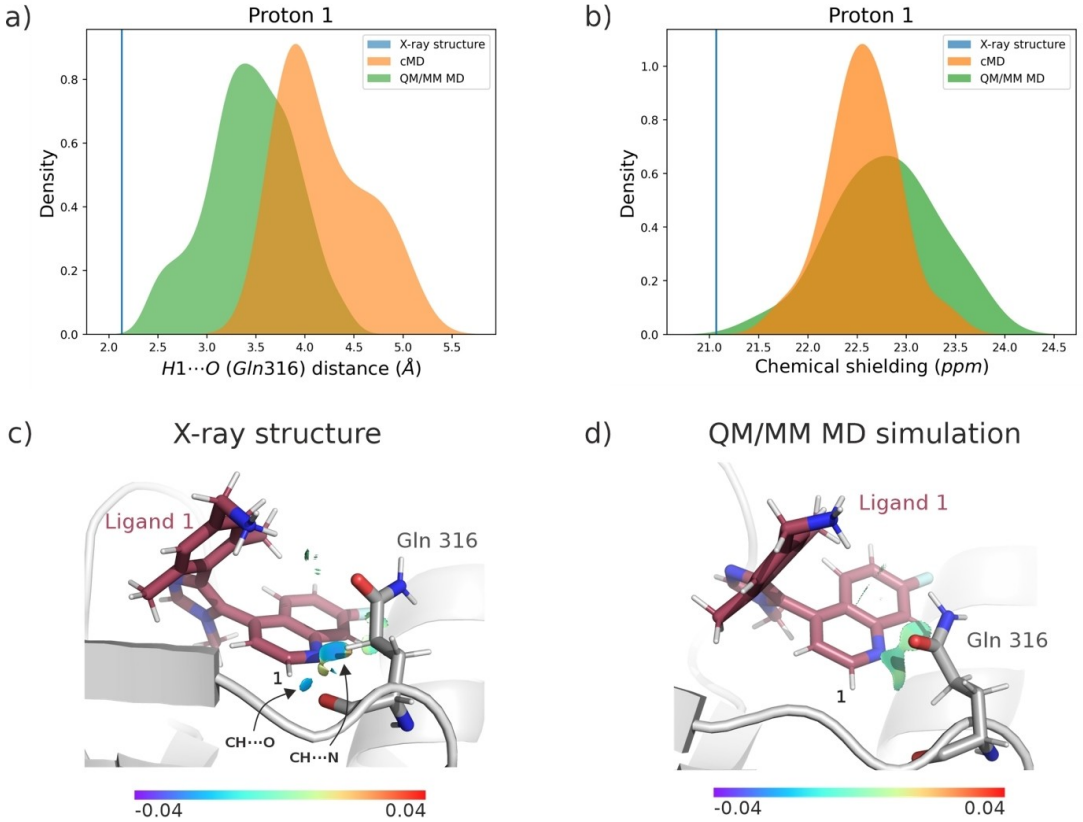


Figure 5. Non-covalent interactions of the ligand proton 1 with residue Gln316. a) Calculated distance distributions (Å) between Gln316 carbonyl oxygen atom (hydrogen bond acceptor) and ligand proton 1 (hydrogen bond donor) from the cMD and QM/MM MD trajectory vs. distance in the crystal structure (blue line). b) Calculated chemical shielding (ppm) from the cMD and QM/MM MD-based ensemble (distribution curves) vs. the crystal structure (blue line) of proton 1. c) Reduced density gradient calculated from the X-ray geometry d) Reduced density gradient calculated based on a representative frame (maximum of the distance and shielding distributions) from QM/MM MD simulation. The iso-surfaces are coloured according to values of sign(λ_2)ρ within the range of −0.04 au (purple) to 0.04 au (red). Purple indicates strong attractive interactions, green weak interactions, and red repulsive interactions. For better visibility, iso-surface representation has been limited to interactions with ligand proton 1.

extensive surface coverage of the protein-ligand interface by protons. As opposed to protein ^{1}H CS information, the assignment of ligand ^{1}H CS information is less time-consuming and does not require sophisticated NMR experiments. Additionally, ligand ^{1}H signals are less likely to be affected by severe line-broadening, for example due to unfavourable motional behaviour, a limitation that often affects CH and CH_2 groups on the side of the protein.

We show that the static structure of the NSD3-Ligand 1 complex derived from X-ray crystallography is not in full agreement with NMR solution data. Discrepancies were found to be largest for ligand protons participating in hydrogen-bonding to protein acceptors. We generated cMD-ensembles with the intent of overcoming the limitation of the static experimental model provided by X-ray crystallography. Consideration of the dynamics based on cMD ensembles allowed us to reduce the RMSE of the ^{1}H CS calculation by ~40%. Since non-covalent interactions in cMD are simplified in classical force field representations, we generated a second ensemble from QM/MM MD simulations with the aim to better reflect the subtle energetic details of the protein-ligand interactions. This resulted in a further improvement of the correlation (additional RMSE reduction of ~10%) with the NMR experiment resulting in a total RMSE reduction of over 50% with respect to the X-ray model. We attribute this improvement to a more complete and accurate view on the protein-ligand complex in solution state as provided by QM/MM MD simulations.

In terms of application, we see numerous areas where ligand ^{1}H CS information can be useful. The analysis of the ^{1}H CS allows to detect and quantify CH···π interactions with the ligand being the H-bond donor instead of the acceptor as shown previously in our "PI by NMR" contribution.[18] This information can be exploited in hit-to-lead as well as in earlier stages of drug discovery programs, where ligands with moderate affinities are optimized towards stronger binding affinities. For example, the (ligand-based) observation of chemical-shift changes induced by aromatic ligand moieties can be utilized to identify compounds which form optimal CH···π interactions with the protein, an information that is helpful in exploring and guiding design ideas. On the contrary, the use of

ChemPhysChem
Research Article
doi.org/10.1002/cphc.202300636
Chemistry Europe
European Chemical Societies Publishing

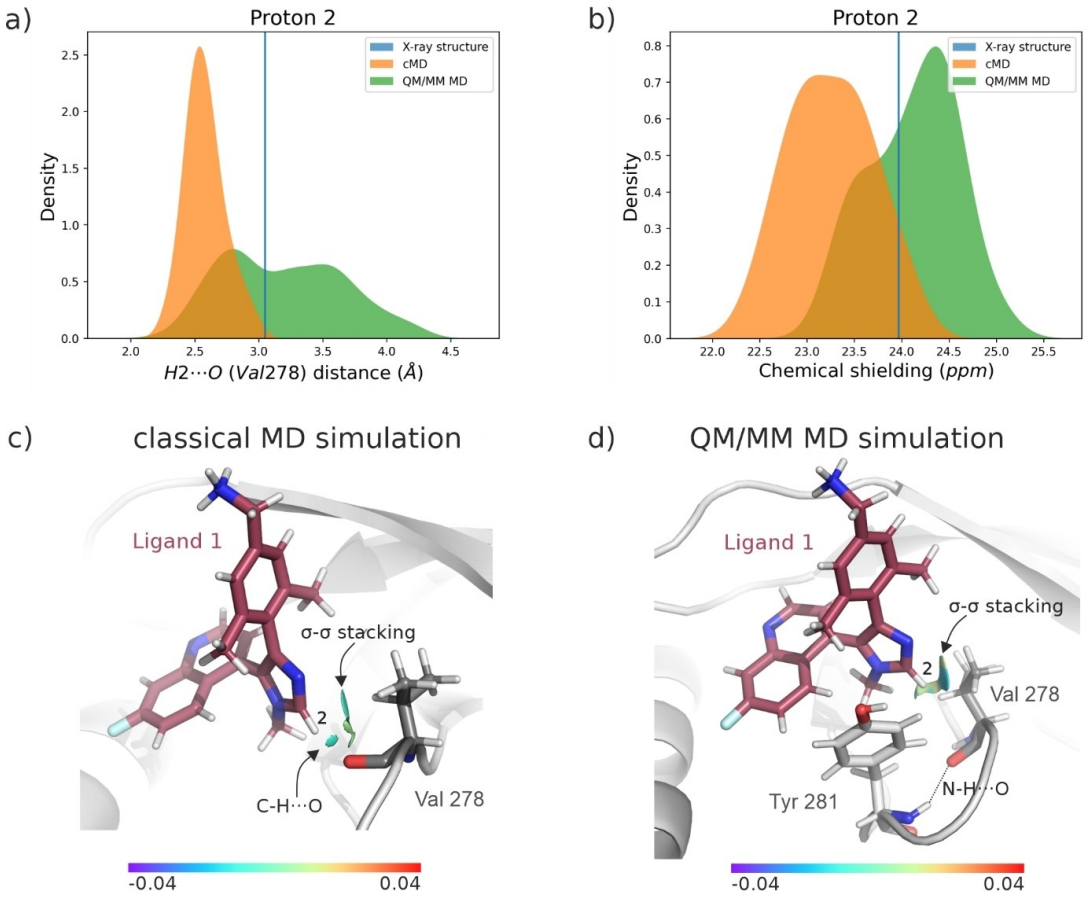


Figure 6. Non-covalent interactions of the ligand proton 2 with residue Val278. a) Calculated distance distributions (Å) between Val278 carbonyl oxygen atom (hydrogen bond acceptor) and ligand proton 1 (hydrogen bond donor) from the cMD and QM/MM MD trajectory vs. distance in the crystal structure (blue line). b) Calculated chemical shielding (ppm) from the cMD and QM/MM MD-based ensemble vs. the crystal structure (blue line) of proton 2. c) Reduced density gradient calculated based on a representative frame (maximum of the distance and shielding distributions) from classical MD simulation. d) Reduced density gradient calculated based on the representative frame (maximum of the distance and shielding distributions) from QM/MM MD simulation. The iso-surfaces are coloured according to values of sign(λ_2)ρ within the range of −0.04 au (purple) to 0.04 au (red). Purple indicates strong attractive interactions, green weak interactions, and red repulsive interactions. For better visibility, Iso-surface representation has been limited to interactions of ligand with proton 2.

ligand ^{1}H CS information for refinement purposes remains largely untapped, with a foreseeable applicability for the verification of docking poses derived from X-ray and cryo-EM models. Most intriguing, however, is the use of ligand ^{1}H CS information in conjunction with ever more present structural models derived from AlphaFold[34] and other machine-learning based approaches to generate accurate and experimentally verified ensembles of protein-ligand complexes that closely resemble the native solution state.

Supporting Information

Additional figures and tables are collected in the Supporting Information together with a detailed description of all experimental and computational methods.

The authors have cited additional references within the Supporting Information.[35–48]

Acknowledgements

G. Platzer and A. L. Ptaszek were funded by the Christian Doppler Laboratory for High-Content Structural Biology and Biotechnology, Austria. The financial support by the Austrian Federal Ministry for Digital and Economic Affairs and the National Foundation for Research, Technology and Development is gratefully acknowledged. A. L. Ptaszek and P. A. Sánchez-Murcia thank the Medical University of Graz for computation time (MedBioNode cluster).

Conflict of Interests

The authors declare no conflict of interest. R. Konrat is a shareholder of MAG-LAB and G. Platzer is an employee of MAG-LAB. M. Mayer, J. Böttcher, J. E. Fuchs, L. Geist and D.B.

ChemPhysChem
Research Article
doi.org/10.1002/cphc.202300636
Chemistry Europe
European Chemical Societies Publishing

McConnell are employees of Boehringer Ingelheim RCV & Co KG.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Keywords: chemical shift · computational chemistry · drug design · NMR spectroscopy · non-covalent interactions

[1] a) C. Bissantz, B. Kuhn, M. Stahl, *J. Med. Chem.* **2010**, *53*, 5061–5084; b) R. Ferreira de Freitas, M. Schapira, *MedChemComm* **2017**, *8*, 1970–1981; c) H. J. Schneider, *Angew. Chem. Int. Ed. Engl.* **2009**, *48*, 3924–3977.
[2] A. S. Mahadevi, G. N. Sastry, *Chem. Rev.* **2016**, *116*, 2775–2825.
[3] D. B. McConnell, *J. Med. Chem.* **2021**, *64*, 16319–16327.
[4] a) G. G. Ferenczy, G. M. Keseru, *J. Chem. Inf. Model.* **2012**, *52*, 1039–1045; b) W. P. Jencks, *Proc. Natl. Acad. Sci. USA* **1981**, *78*, 4046–4050.
[5] a) G. Klebe, *Nat. Rev. Drug Discovery* **2015**, *14*, 95–110; b) T. S. Olsson, M. A. Williams, W. R. Pitt, J. E. Ladbury, *J. Mol. Biol.* **2008**, *384*, 1002–1017.
[6] A. V. Sadybekov, V. Katritch, *Nature* **2023**, *616*, 673–685.
[7] a) B. T. Burnley, P. V. Afonine, P. D. Adams, P. Gros, *eLife* **2012**, *1*, e00311; b) P. T. Lang, H. L. Ng, J. S. Fraser, J. E. Corn, N. Echols, M. Sales, J. M. Holton, T. Alber, *Protein Sci.* **2010**, *19*, 1420–1431.
[8] a) J. J. Chou, S. Li, C. B. Klee, A. Bax, *Nat. Struct. Biol.* **2001**, *8*, 990–997; b) M. A. DePristo, P. I. de Bakker, T. L. Blundell, *Structure* **2004**, *12*, 831–838.
[9] A. Blum, J. Bottcher, S. Dorr, A. Heine, G. Klebe, W. E. Diederich, *J. Mol. Biol.* **2011**, *410*, 745–755.
[10] T. Nakane, A. Kotecha, A. Sente, G. McMullan, S. Masiulis, P. Brown, I. T. Grigoras, L. Malinauskaite, T. Malinauskas, J. Miehling, T. Uchanski, L. Yu, D. Karia, E. V. Pechnikova, E. de Jong, J. Keizer, M. Bischoff, J. McCormack, P. Tiemeijer, S. W. Hardwick, D. Y. Chirgadze, G. Murshudov, A. R. Aricescu, S. H. W. Scheres, *Nature* **2020**, *587*, 152–156.
[11] M. T. Muhammed, E. Aki-Yalcin, *Chem. Biol. Drug Des.* **2019**, *93*, 12–20.
[12] a) J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, *Nature* **2021**, *596*, 583–589; b) M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, C. Millán, H. Park, C. Adams, C. R. Glassman, A. DeGiovanni, J. H. Pereira, A. V. Rodrigues, A. A. van Dijk, A. C. Ebrecht, D. J. Opperman, T. Sagmeister, C. Buhlheller, T. Pavkov-Keller, M. K. Rathinaswamy, U. Dalwadi, C. K. Yip, J. E. Burke, K. C. Garcia, N. V. Grishin, P. D. Adams, R. J. Read, D. Baker, *Science* **2021**, *373*, 871–876.
[13] E. Richard, O. N. Michael, P. Alexander, A. Natasha, S. Andrew, G. Tim, Ž. Augustin, B. Russ, B. Sam, Y. Jason, R. Olaf, B. Sebastian, Z. Michal, B. Alex, P. Anna, C. Andrew, T. Kathryn, J. Rishub, C. Ellen, K. Pushmeet, J. John, H. Demis, *bioRxiv* **2022**, 2021.2010.2004.463034.
[14] a) A. T. Brunger, *Nat. Struct. Biol.* **1997**, *4 Suppl*, 862–865; b) M. Rinaldelli, E. Ravera, V. Calderone, G. Parigi, G. N. Murshudov, C. Luchinat, *Acta Crystallogr. Sect. D* **2014**, *70*, 958–967; c) H. van den Bedem, J. S. Fraser, *Nat. Methods* **2015**, *12*, 307–318; d) N. J. Fowler, M. P. Williamson, *Structure* **2022**, *30*, 925–933 e922.
[15] a) T. E. Exner, A. Frank, I. Onila, H. M. Moller, *J. Chem. Theory Comput.* **2012**, *8*, 4818–4827; b) P. Robustelli, K. A. Stafford, A. G. Palmer, 3rd, *J. Am. Chem. Soc.* **2012**, *134*, 6365–6374.
[16] E. Miclet, D. C. Williams, Jr., G. M. Clore, D. L. Bryce, J. Boisbouvier, A. Bax, *J. Am. Chem. Soc.* **2004**, *126*, 10560–10570.
[17] B. Wang, K. Raha, K. M. Merz, Jr., *J. Am. Chem. Soc.* **2004**, *126*, 11430–11431.
[18] G. Platzer, M. Mayer, A. Beier, S. Bruschweiler, J. E. Fuchs, H. Engelhardt, L. Geist, G. Bader, J. Schorghuber, R. Lichtenecker, B. Wolkerstorfer, D. Kessler, D. B. McConnell, R. Konrat, *Angew. Chem. Int. Ed. Engl.* **2020**, *59*, 14861–14868.
[19] M. W. Sroka, C. R. Vakoc, *Nature* **2021**, *590*, 399–400.
[20] J. Bottcher, D. Dilworth, U. Reiser, R. A. Neumuller, M. Schleicher, M. Petronczki, M. Zeeb, N. Mischerikow, A. Allali-Hassani, M. M. Szewczyk, F. Li, S. Kennedy, M. Vedadi, D. Barsyte-Lovejoy, P. J. Brown, K. V. M. Huber, C. M. Rogers, C. I. Wells, O. Fedorov, K. Rumpel, A. Zoephel, M. Mayer, T. Wunberg, D. Bose, S. Zahn, H. Arnhof, H. Berger, C. Reiser, A. Hormann, T. Krammer, M. Corcokovic, B. Sharps, S. Winkler, D. Haring, X. L. Cockcroft, J. E. Fuchs, B. Mullauer, A. Weiss-Puxbaum, T. Gerstberger, G. Boehmelt, C. R. Vakoc, C. H. Arrowsmith, M. Pearson, D. B. McConnell, *Nat. Chem. Biol.* **2019**, *15*, 822–829.
[21] a) K. Wolinski, J. F. Hinton, P. Pulay, *J. Am. Chem. Soc.* **1990**, *112*, 8251–8260; b) J. Swails, T. Zhu, X. He, D. A. Case, *J. Biomol. NMR* **2015**, *63*, 125–139; c) A. H. Mazurek, L. Szeleszczuk, D. M. Pisklak, *Int. J. Mol. Sci.* **2021**, *22*; d) X. Jin, T. Zhu, J. Z. H. Zhang, X. He, *Front. Chem.* **2018**, *6*, 150.
[22] a) V. Palivec, R. Pohl, J. Kaminský, H. Martinez-Seara, *J. Chem. Theory Comput.* **2022**, *18*, 4373–4386; b) T. Zhu, J. Z. H. Zhang, X. He, *J. Chem. Theory Comput.* **2013**, *9*, 2104–2114; c) M. Dracinsky, H. M. Moller, T. E. Exner, *J. Chem. Theory Comput.* **2013**, *9*, 3806–3815; d) J. Pavlíková Přecechtělová, A. Mládek, V. Zapletal, J. Hritz, *J. Chem. Theory Comput.* **2019**, *15*, 5642–5658; e) M. J. Bakker, A. Mládek, H. Semrád, V. Zapletal, J. Pavlíková Přecechtělová, *Phys. Chem. Chem. Phys.* **2022**, *24*, 27678–27692; f) M. Robinson, P. D. Haynes, *J. Chem. Phys.* **2010**, *133*; g) P. Liu, W. Li, Z. Kan, H. Sun, J. Ma, *J. Phys. Chem. A* **2016**, *120*, 490–502; h) R. Atwi, Y. Chen, K. S. Han, K. T. Mueller, V. Murugesan, N. N. Rajput, *Nat. Comput. Sci.* **2022**, *2*, 112–122.
[23] J. Kollar, V. Frecer, *J. Mol. Model.* **2017**, *24*, 11.
[24] a) F. London, *J. Phys. Radium* **1937**, *8*, 397–409; b) R. Ditchfield, *Mol. Phys.* **1974**, *27*, 789–807; c) R. McWeeny, *Phys. Rev.* **1962**, *126*, 1028–1034; d) J. R. Cheeseman, G. W. Trucks, T. A. Keith, M. J. Frisch, *J. Chem. Phys.* **1996**, *104*, 5497–5509.
[25] J. D. Chai, M. Head-Gordon, *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615–6620.
[26] F. Weigend, R. Ahlrichs, *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297–3305.
[27] J. M. Word, S. C. Lovell, J. S. Richardson, D. C. Richardson, *J. Mol. Biol.* **1999**, *285*, 1735–1747.
[28] C. Tian, K. Kasavajhala, K. A. A. Belfon, L. Raguette, H. Huang, A. N. Migues, J. Bickel, Y. Wang, J. Pincay, Q. Wu, C. Simmerling, *J. Chem. Theory Comput.* **2020**, *16*, 528–552.
[29] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, D. A. Case, *J. Comput. Chem.* **2004**, *25*, 1157–1174.
[30] R. S. Paton, J. M. Goodman, *J. Chem. Inf. Model.* **2009**, *49*, 944–955.
[31] S. Tsuzuki, K. Honda, T. Uchimaru, M. Mikami, K. Tanabe, *J. Am. Chem. Soc.* **2000**, *122*, 3746–3753.
[32] a) J. Contreras-Garcia, E. R. Johnson, S. Keinan, R. Chaudret, J. P. Piquemal, D. N. Beratan, W. Yang, *J. Chem. Theory Comput.* **2011**, *7*, 625–632; b) R. A. Boto, F. Peccati, R. Laplaza, C. Quan, A. Carbone, J. P. Piquemal, Y. Maday, A. J. Contreras-Garci, *J. Chem. Theory Comput.* **2020**, *16*, 4150–4158; c) E. R. Johnson, S. Keinan, P. Mori-Sánchez, J. Contreras-García, A. J. Cohen, W. Yang, *J. Am. Chem. Soc.* **2010**, *132*, 6498–6506.
[33] a) J. P. Wagner, P. R. Schreiner, *Angew. Chem. Int. Ed. Engl.* **2015**, *54*, 12274–12296; b) M. Alonso, T. Woller, F. J. Martin-Martinez, J. Contreras-Garcia, P. Geerlings, F. De Proft, *Chemistry* **2014**, *20*, 4931–4941; c) A. Feng, Y. Zhou, M. A. Y. Al-Shebami, L. Chen, Z. Pan, W. Xu, S. Zhao, B. Zeng, Z. Xiao, Y. Yang, W. Hong, *Nat. Chem.* **2022**, *14*, 1158–1164; d) D. Danovich, S. Shaik, F. Neese, J. Echeverria, G. Aullon, S. Alvarez, *J. Chem. Theory Comput.* **2013**, *9*, 1977–1991.
[34] D. Sala, F. Engelberger, H. S. McHaourab, J. Meiler, *Curr. Opin. Struct. Biol.* **2023**, *81*, 102645.
[35] J. Myers, G. Grothaus, S. Narayanan, A. Onufriev, *Proteins* **2006**, *63*, 928–938.
[36] R. Anandakrishnan, B. Aguilar, A. V. Onufriev, *Nucleic Acids Res.* **2012**, *40*, W537–541.
[37] a) S. Bietz, S. Urbaczek, B. Schulz, M. Rarey, *J. Cheminf.* **2014**, *6*, 12; b) R. Fahrrolfes, S. Bietz, F. Flachsenberg, A. Meyder, E. Nittinger, T. Otto, A. Volkamer, M. Rarey, *Nucleic Acids Res.* **2017**, *45*, W337-W343; c) K. Schoning-Stierand, K. Diedrich, R. Fahrrolfes, F. Flachsenberg, A. Meyder, E. Nittinger, R. Steinegger, M. Rarey, *Nucleic Acids Res.* **2020**, *48*, W48-W53.
[38] Gaussian 16, (Revision C.01), M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, Williams, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G.

ChemPhysChem
Research Article
doi.org/10.1002/cphc.202300636
Chemistry Europe
European Chemical Societies Publishing

Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, D. J. Fox, Gaussian, Wallingford, CT, **2016**.

[39] Amber 21, D. A. Case, K. Belfon, I. Y. Ben-Shalom, J. T. Berryman, S. R. Brozell, D. S. Cerutti, T. E. Cheatham, III, G. A. Cisneros, V. W. D. Cruzeiro, T. A. Darden, R. E. Duke, G. Giambasu, M. K. Gilson, H. Gohlke, A. W. Goetz, R. Harris, S. Izadi, S. A. Izmailov, K. Kasavajhala, M. C. Kaymak, E. King, A. Kovalenko, T. Kurtzman, T. S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, M. Machado, V. Man, M. Manathunga, K. M. Merz, Y. Miao, O. Mikhailovskii, G. Monard, H. Nguyen, K. A. O'Hearn, A. Onufriev, F. Pan, S. Pantano, R. Qi, A. Rahnamoun, D. R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, A. Shajan, J. Shen, C. L. Simmerling, N. R. Skrynnikov, J. Smith, J. Swails, R. C. Walker, J. Wang, J. Wang, H. Wei, R. M. Wolf, X. Wu, Y. Xiong, Y. Xue, D. M. York, S. Zhao, P. A. Kollman, **2021**.

[40] T. Darden, D. York, L. Pedersen, *J. Chem. Phys.* **1993**, *98*, 10089–10092.

[41] S. Miyamoto, P. A. Kollman, *J. Comput. Chem.* **1992**, *13*, 952–962.

[42] a) C. Bannwarth, S. Ehlert, S. Grimme, *J. Chem. Theory Comput.* **2019**, *15*, 1652–1671; b) C. Bannwarth, E. Caldeweyher, S. Ehlert, A. Hansen, P. Pracht, J. Seibert, S. Spicher, S. Grimme, *WIREs Comput. Mol. Sci.* **2021**, *11*, e1493.

[43] F. Neese, *WIREs Comput. Mol. Sci.* **2012**, *2*, 73–78.

[44] D. R. Roe, T. E. Cheatham 3rd, *J. Chem. Theory Comput.* **2013**, *9*, 3084–3095.

[45] M. Ernzerhof, G. E. Scuseria, *J. Chem. Phys.* **1999**, *110*, 5029–5036.

[46] a) A. D. Becke, *J. Chem. Phys.* **1993**, *98*, 5648–5652; b) C. Lee, W. Yang, R. G. Parr, *Phys. Rev. B* **1988**, *37*, 785–789.

[47] Y. Zhao, D. G. Truhlar, *Theor. Chem. Acc.* **2008**, *120*, 215–241.

[48] The PyMOL Molecular Graphics System, Version 1.8, L. Schrödinger, & DeLano, W., **2015**.

Manuscript received: September 6, 2023
Revised manuscript received: October 23, 2023
Version of record online: November 13, 2023

ChemPhysChem

Supporting Information

Ligand ^{1}H NMR Chemical Shifts as Accurate Reporters for Protein-Ligand Binding Interfaces in Solution

Gerald Platzer,* Aleksandra L. Ptaszek, Jark Böttcher, Julian E. Fuchs, Leonhard Geist, Daniel Braun, Darryl B. McConnell, Robert Konrat, Pedro A. Sánchez-Murcia,* and Moriz Mayer*

SECTION S1 SAMPLE PREPARATION, ANALYTICAL DATA AND SYNTHETIC PROCEDURES

Sequence information as well as expression and purification protocols for the PWWP1 domain of NSD3 are given in Böttcher *et al*.[1] In order to obtain deuterated protein, the M9 minimal medium was supplemented with $^{15}NH_4Cl$ (0.5 g/L) and deuterated glucose (2.5 g/L) dissolved in D_2O. After a series of pre-cultures (LB - overnight, 1:1 LB+M9 - 8h, M9 - overnight) containing ampicillin (100 µg/L), M9 (D_2O) medium was inoculated, induced with IPTG (0.25 mM) at an OD_{600} of 0.8 and grown for 18 h at 20 °C. The final buffer for size-exclusion column contained: 50 mM Na-phosphate, 100 mM NaCl, 1 mM TCEP, pH 7.2 – dissolved in D_2O. Compounds were obtained from the central repository of Boehringer Ingelheim. Analytical data and synthesis were described previously.[2] Compound purity of ligand 1 was remeasured to >95% with HPLC, traces shown below.

SECTION S2 PROTEIN NMR SPECTROSCOPY AND [^{1}H-^{1}H]-NOESY SIGNAL ASSIGNMENT

All protein NMR experiments were conducted at 298 K on a Bruker Avance 600 MHz spectrometer equipped with a TCI cryoprobe. [^{1}H-^{1}H]-NOESY experiments were acquired using the pulse sequence *"noesyphpr"* of the Bruker library. NSD3-PWWP1 sample concentrations were 200 µM. Ligand concentrations were 400 µM for Ligand 1 (2:1). The NOESY mixing time was adjusted to 300 ms. Spectra were recorded using 256 (*t1*) x 2048 (*t2*) complex points with acquisition times of 19.3 ms (*t1*) and 155 ms (*t2*). A total of 32 scans were recorded per t1 increment with a recycle delay of 2 sec.

A scheme of the signal assignment strategy is outlined in Figure 1. Resonances of the free ligand can be determined from a simple ^{1}H-1D (Figure S1a / red ^{1}H-1D) while signals of the bound state are observed when the ligand to protein ratio is adjusted to a 1:1 ratio (Figure S1a / blue ^{1}H-1D). The nanomolar affinity of Ligand 1 ensures full ligand saturation. For the analysis of the 2D NOESY spectra a 2:1 ratio of ligand to protein is chosen resulting in the emergence of exchange cross peaks between the free and the bound ligand, making the signal assignment of the bound ligand form via cross peak connectivities a simple task. Signal assignment of the bound state is facilitated by the fact that typically only exchange (EXSY) peaks are observed between the free and bound states (solid black square in Figure 1). As stated before, EXSY cross peaks (solid black) arises from exchange between the free and the bound ligand species while the regular NOESY cross peak (dotted black) originate from spatial proximity. Only in cases where intra-ligand NOEs are already visible in the protein free (apo) form, additional NOESY peaks are observed for the 2:1 mixture. Besides the EXSY cross peak (solid black) an additional exchange-relayed NOESY cross peak connects the free and bound state (dotted black). For Ligand 1 this is the case for neighbouring protons (resonances 1&6 and 3&5). Since all NOE cross peaks observed for the NSD3/Ligand 1 system involving free ligand species have a positive sign, the transfer of magnetization must happen via a relayed NOE effect, where the ligand retains an effective correlation time resembling the bound state. Discrimination between the two types of cross peaks is straightforward shall be discussed with the example of proton 1 of Ligand 1 (Figure 1a). The free signal of Proton 1 is connected to three other signals via cross peaks in the NOESY experiment. Strong signal intensities are found for the EXSY signal connecting 1_{free} to 1_{bound} and the NOESY signal connecting 1_{free} and 6_{free}. The only other signal connecting a free to a bound signal is the relayed NOESY signal 1_{free} to 6_{bound}. Differentiation of the exchange signal vs the relayed NOE signal is possible due to their different signal intensities. The significantly lower signal intensity of the relayed NOE signal vs. the exchange signal stems from the presence of two effects (NOE and exchange)

whereas the exchange signal itself has no NOE contribution. This circumstance makes the identification of the bound signal of each individual proton straightforward. Interestingly and important to note, bound ligand resonances that are close to the detection limit in the ^{1}H-1D can still be detected via NOESY cross peak signals.

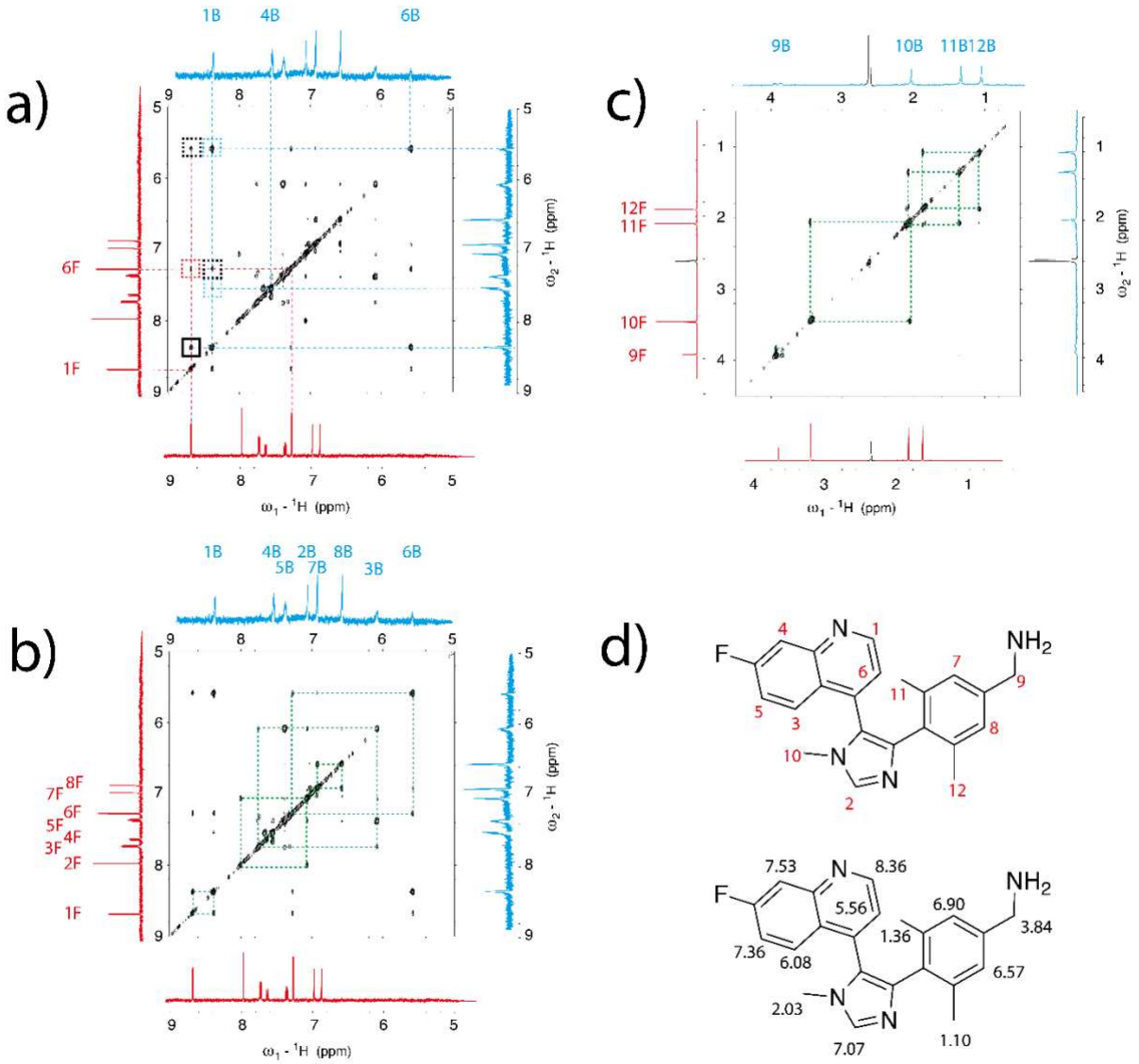


Figure S1. [^{1}H-^{1}H]-NOESY spectra of Ligand 1 (SPR K_D = 170 nM) in complex with the PWWP1 domain of NSD3. 2D-NOESY spectra of ligand 1 were recorded with a 2-fold excess over the protein. 1D traces represent ligand signals in the free state (red) and the bound state (blue) and were derived from separate ^{1}H-1D measurements. For the 1D, the free state (red) was measured in the absence of protein and the bound state (blue) with a protein to ligand ration of 1:1 for ligand 1. 1D peak-suffixes F and B refer to the free and bound species respectively. a) Scheme for the assignment strategy of NOE cross peaks showing exchange peaks (solid black), exchange relayed NOE cross peaks (dotted black), NOEs between unbound signals (dotted red) and NOEs between bound signals (dotted blue). b) NOE exchange peaks correlate free and bound signals (dotted green) for aromatic protons, and aliphatic protons c). d) Proton signal numbering (top) and ^{1}H chemical shifts in the protein bound state (bottom).

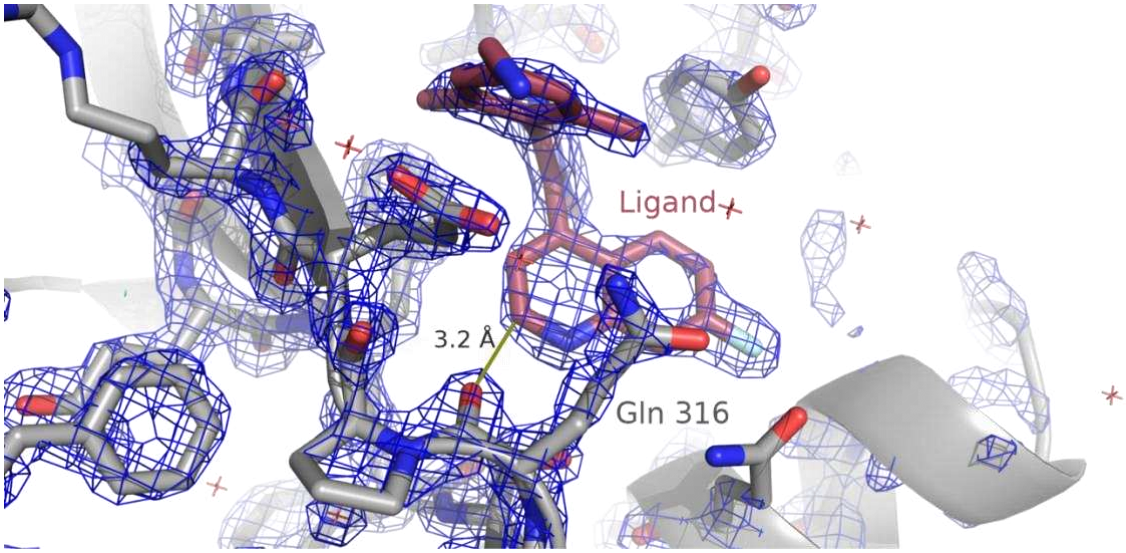


Figure S2. X-ray electron density map of ligand 1 bound to the PWWP1 domain of human NSD3 (PDB ID 6G2O).[1]

SECTION S3 CLASSICAL MOLECULAR DYNAMICS (MD) SIMULATION

The X-ray structure of the PWWP1 domain of NSD3 co-crystallized in complex with the compound BI-9321 (ligand 1) (136 residues, PDB ID 6G2O, resolution 1.81 Å)[1] was used as initial Cartesian coordinates of the system. The missing loop region containing residues 343-353 was modelled manually based on the Q9BZ95 AlphaFold model. After that, the loop was minimized over 300 steps using the steepest descent algorithm in gas phase, applying restrains in all atoms except on the residues of the modelled loop (res. 339-361). The protonation state of the protein residues at pH 8.0 was estimated using the software H++ v. 4.0.[3] Ligand 1 was protonated using the ProteinsPlus server[4] and its geometry was optimized at the B3LYP-D3/6-311G(d,p) level of theory with Gaussian16.[5] Using the optimized structure, the RESP point charges were derived from the computed electrostatic potential at the HF/6-31G(d) level of theory. Amber ff19SB force field[6] was used for the amino acid residues and the GAFF force field[7] was used to describe the atom types of ligand 1. The system was neutralized with chloride ions and soaked in an octahedral box with TIP3P water molecules that extended 10 Å away from the solute using the LEaP program (AmberTools21).[8] The MD simulations were performed under periodic boundary conditions. For all non-bonded interactions, a 10 Å cut-off distance was applied, whereas long range electrostatic interactions were treated using the smooth particle mesh Ewald (PME) method.[9] The SHAKE algorithm was used to constrain bonds involving hydrogen atoms.[10] The system was minimized in three consecutive steps using for the first 10,000 cycles the steepest descent algorithm, and the conjugate gradient method for the last 10,000 cycles. In the first two steps, the hydrogen atoms and the solvent molecules were sequentially minimized, and in the last step, the whole system was allowed to relax. Afterwards, the solvated system was heated from 100.00 to 298.15 K during 20 ps with a time step of 0.2 fs in an NVT ensemble using the Langevin thermostat (1 ps^{-1} collision frequency). The position of all solute heavy atoms was restrained with a weight of 40 kcal mol^{-1} $Å^{-2}$. Then, the system was equilibrated in 5 steps of 20 ps each (100 ps in total, 2 fs timestep). The first 4

equilibration steps were run with a NVT ensemble and in the last step, the system was switched to an isothermal-isobaric ensemble (NPT) with a pressure of 1 bar. In parallel, the harmonic constraints were gradually removed. The system was further simulated for 0.5 µs at 298.15 K. All frames from the MD trajectory were collected into **cMD - full set** (Figure S3). The MD trajectory was clustered with the *k-means* algorithm (cpptraj) into 3 conformational clusters based on RMSD of the heavy atoms of the residues Val278, Tyr281, Trp284, Phe312, Gln316, and Gln367 and ligand 1. From the first cluster, that covers 50% of the frames of the entire trajectory, 20 random frames were selected for further ^{1}H NMR calculation (**cMD - 20 frames set**). Figure S3 shows the distance distributions of these MD ensembles (green and red plots) between relevant atoms on ligand 1 and their interacting partners in the protein. **cMD – 20 frames set** with 20 frames from the major cluster represents maxima of the distance distributions of the **cMD - full set**. It can therefore be concluded that the choice of 20 frames is an accurate approximation of the 0.5 µs trajectory, thus reducing the computation time of CSs calculation.

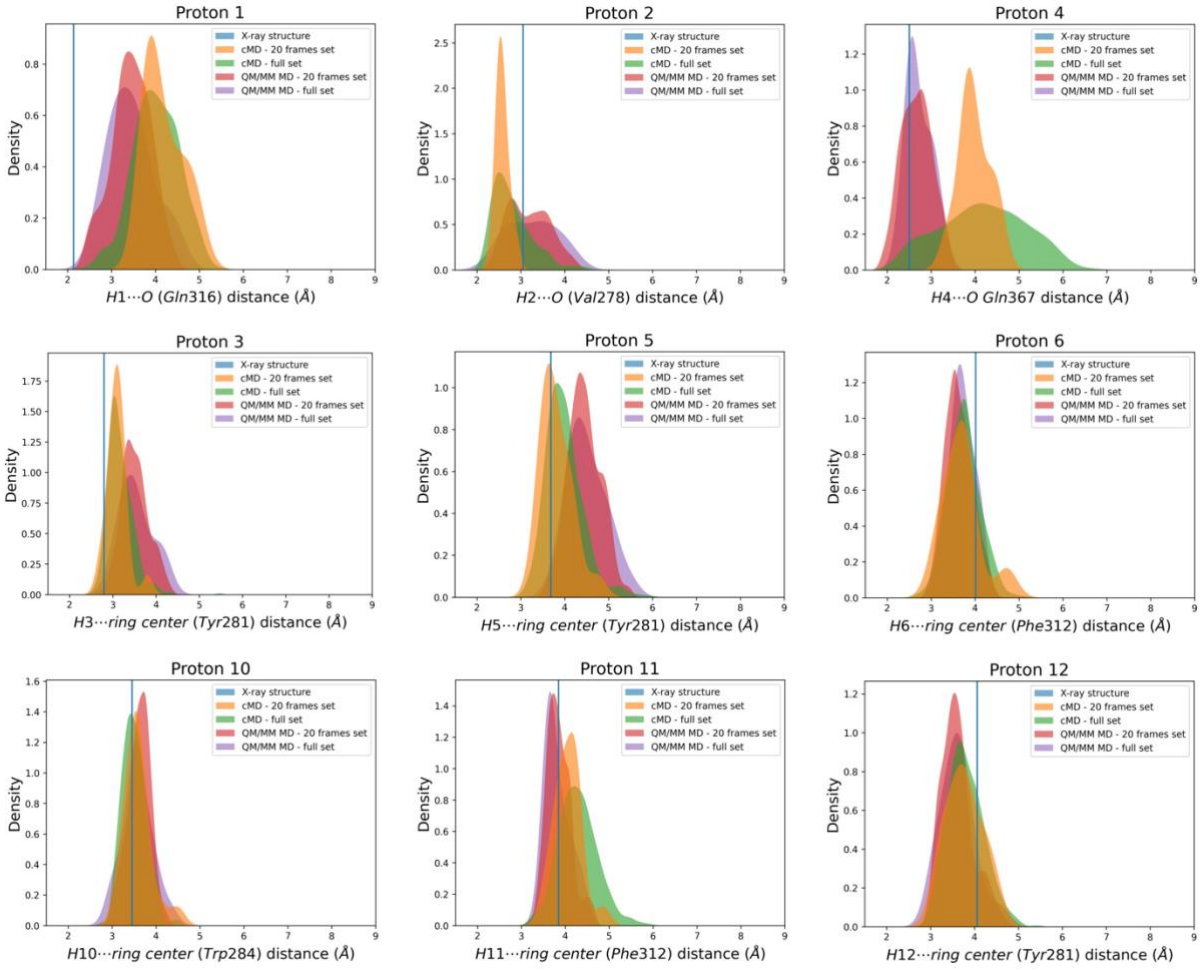


Figure S3. Distance distributions (Å) between ligand protons and interacting residues from **X-ray structure** (reference, blue), **cMD - 20 frames set** (orange), **cMD - full set** (0.5 µs, green), **QM/MM MD - 20 frames set** (red), and **QM/MM MD - full set** (purple).

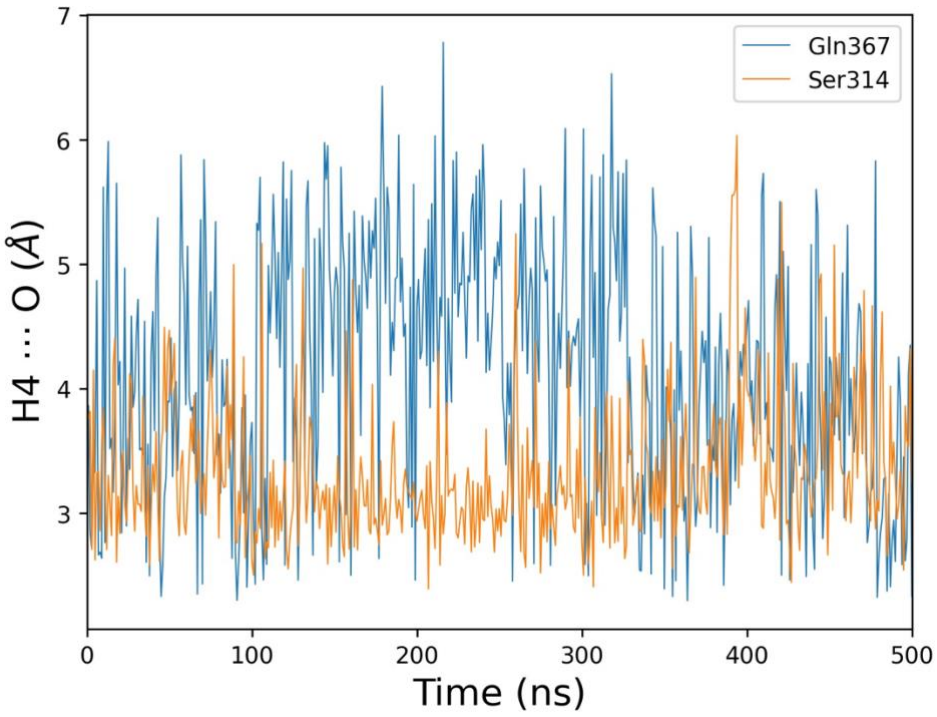


Figure S4. Fluctuations of hydrogen bond distance between ligand proton 4 and interacting residues over the classical MD trajectory (**cMD - full set**).

SECTION S4 QM/MM MD SIMULATIONS

The equilibrated solvated system before MD production was taken as initial geometry for the QM/MM MD simulations. The definition of the QM region included the side chains from the C_β of the interacting residues on the protein (Val278, Tyr281, Trp284, Phe312, Gln316, Gln367; plus, the backbone atoms of Val278 and Gln316) and atoms of ligand 1 (146 total number of atoms). Importantly, the N- and C-terminus of the latter two residues were capped as amides with the previous carbonyl carbon and with the following amino group, respectively. The QM region was treated quantum mechanically using the semi-empirical tight-binding GFN2-xTB method.[11] It was shown that GFN2-xTB outperforms other semiempirical methods in its ability to accurately reproduce the electronic density.[11] A trajectory of 10 ps length was produced with a time step of 0.2 fs. The trajectory was saved every 10 fs as **QM/MM MD - full set**. 20 frames were selected in uniform time intervals for further chemical shifts calculations (**QM/MM MD - 20 frames set**). SHAKE was activated only for the MM region and all simulations were run under an NPT ensemble at 298.15 K with a Langevin thermostat. The QM/MM MD simulation was performed within Amber20 with sander.MPI[8] interfaced to Orca v. 5.0.2.[12] The trajectory was analysed using cpptraj v5.1.0.[13]

The use of the geometries from 10 ps QM/MM MD simulation gave a sufficient prediction of the ^{1}H-NMR chemical shifts, absolute error below 0.5 pm for all atoms (Figure 4d). Importantly, this methodology allowed us to reduce and understand the significant error on proton 2 that we identified in the case of the cMD approach (Figure 4b, Figure 6). To verify if the improvement of the proton 2 CS does not arise from the short simulation time of QM/MM MD simulation, we performed three 10 ps QM/MM MD replicas starting from cMD snapshots

from **cMD – 20 frames set**. Figure S5 shows the distribution of the proton 2 CH···O hydrogen bond distances based on the three 10 ps trajectories. We observe the same behaviour of the hydrogen bond as from the single trajectory chosen for the CS calculations (Figure S3). Longer QM/MM MD simulations could lead to artefacts due to the partition between QM and MM regions and are not feasible in reasonable computational time.

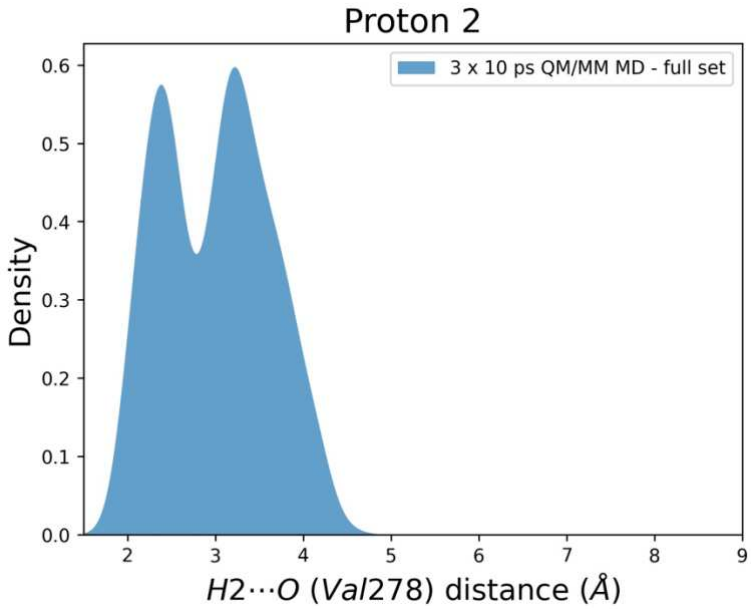


Figure S5. Distance distribution (Å) of proton 2 CH···O hydrogen bond from three 10 ps QM/MM MD simulation replicas.

SECTION S5 LIGAND ^{1}H-NMR CHEMICAL SHIFTS CALCULATION

For the prediction of ^{1}H-CS of ligand 1, three different paths were used to obtain the input geometries (Figure 1 main text): (i) the X-ray structure of ligand 1 in complex with NSD3-PWWD1 (path **A**), (ii) cMD-ensemble (path **B**), and (iii) QM/MM MD-ensemble (path **C**).

In all cases, the isotropic chemical shielding values were calculated using the gauge-independent atomic orbital GIAO method,[14] with the wB97XD[15] DFT functional, and the def2-TZVP[16] basis set as implemented in Gaussian16 package. The ^{1}H-NMR chemical shielding values were calculated for the X-ray crystal structure, cMD - 20 frames set, and the QM/MM MD – 20 frames set (see sections S3 and S4). Regarding the X-ray, all hydrogen atoms of the structure, including crystal water molecules, were firstly minimized with the Amber ff19SB/GAFF force fields according to the first step of the minimization protocol (see section S3). In all cases, atoms defined as the QM region (see section S4) were treated explicitly, and the rest of the environment was included as RESP point charges (electrostatic embedding). In

order to compare the results with the experiment, ^{1}H-NMR chemical shielding values ($\boldsymbol{\sigma}$) need to be converted to chemical shifts ($\boldsymbol{\delta}$) using a standard according to the equation:

$$\delta = \sigma_{standard} - \sigma$$

Chemical shielding of the standard (e.g., TMS, H_2O) is usually calculated at the same level of theory as the system of interest in gas phase or with implicit solvation. Nevertheless, this method neglects the dynamics of the system and intermolecular interactions, such as hydrogen bonds. Therefore, we decided for an alternate way to calculate the chemical shielding of the standard ($\boldsymbol{\sigma_{standard}}$), water in our case, using a linear model with fixed slope of -1 according to the literature.[17] The standard chemical shielding was re-calculated for each of the paths taken for ^{1}H-CS calculation (Figure 1).

To verify that the X-ray structure-based NMR CS prediction is not an artefact of the chosen DFT-functional, we performed a benchmarking study (Figure S6). We selected two hybrid DFT-functionals: PBE0[18] and B3LYP[19], as well as one hybrid meta-GGA functional, M06-2X[20] covering "medium-range" correlation energy. According to the literature, all of them yield satisfactory CS outcomes.[21] Figure S6 show that the results differ very little when other density functions are used.

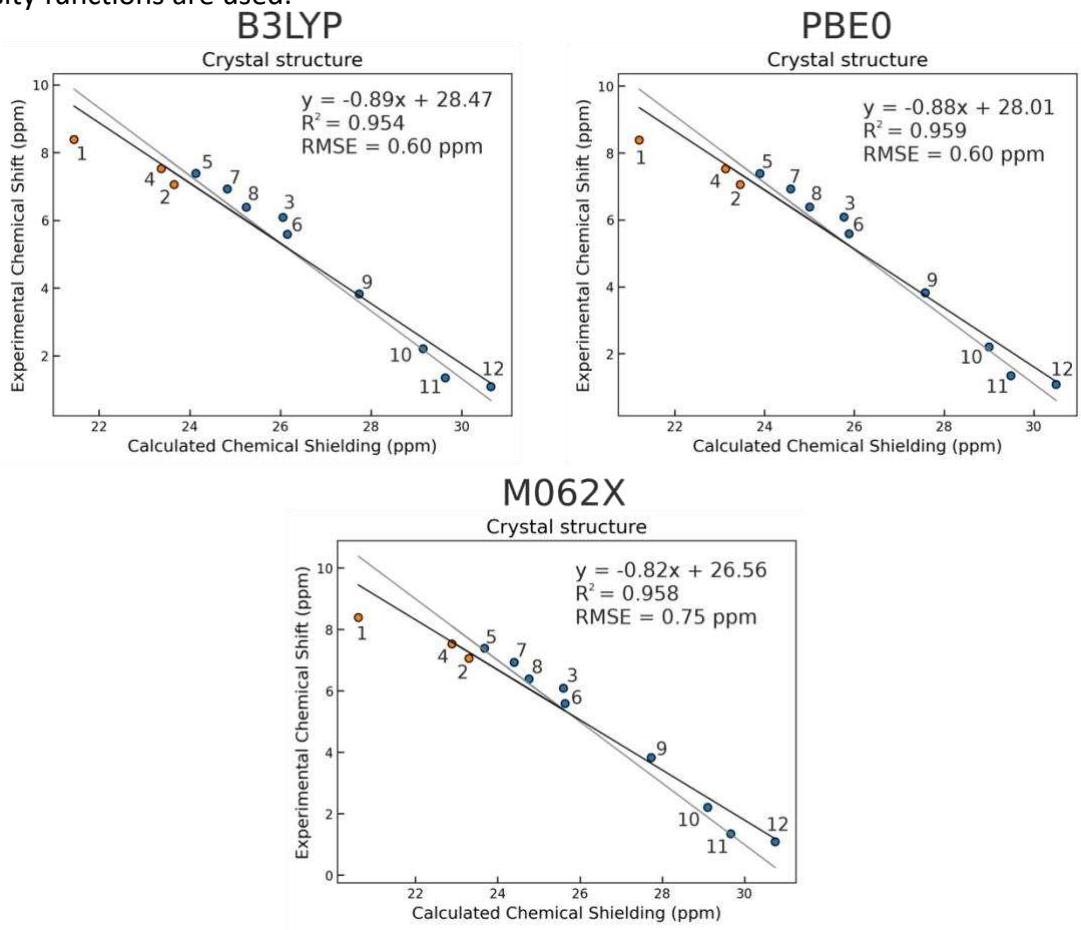


Figure S6. Correlation between experimental ^{1}H chemical shifts (ppm) and calculated chemical shielding (ppm) using different DFT functionals with the Def2-TZVP basis set.

SECTION S6 NON-COVALENT INTERACTIONS INDEX (NCI) ANALYSIS

The non-covalent interactions (NCI) index is applied to identify the intermolecular interactions based on the electronic density ($\rho(\vec{r})$) and its derivatives. Plot of reduced density gradient vs the electronic density

$$s = \frac{1}{2(3\pi^2)^{\frac{1}{3}}} \frac{|\nabla\rho|}{\rho^{\frac{4}{3}}}$$

reveals characteristic picks in the range of low $\rho(\vec{r})$ and $s(\vec{r})$ corresponding to weak interactions. The sign of the second eigenvalue of the electron density Hessian ($\nabla\rho^2 = \lambda_1 + \lambda_2 + \lambda_3$), indicates whether interactions are bonding ($\lambda_2 < 0$) or non-bonding ($\lambda_2 > 0$). Consequently, the $sign(\lambda_2)\rho$ as a function of $s(\vec{r})$ is plotted to visualize interaction types: (a) attractive interactions, e.g., hydrogen bonds with the $sign(\lambda_2)\rho < 0$, (b) van der Waals interactions with the $sign(\lambda_2)\rho \approx 0$, (c) non-bonding contacts (steric clashes) when $sign(\lambda_2)\rho > 0$. The electronic density gives insight into how strong those interactions are. In this work, the contours of $s(\vec{r})$ in the colours of $sign(\lambda_2)\rho$ values are presented. Purple indicates strong attractive interactions, green weak interactions, and red repulsive interactions. Reduced density gradient for the protein-ligand interactions was calculated based on the pro-molecular densities using the NCIPLOT4 following the protocol dedicated for protein-ligand interactions.[22] Contours were visualized with Pymol.[23]

SECTION S7 COMPUTATIONAL TIMINGS

Table 1 Timings for classical MD, QM/MM MD and GIAO/DFT calculations

Calculation	Computational time for 1 time step	Total time to complete the task	Resources
Classical MD	0.87 ms	60 hrs 25 min	RTX 3090 GPU
QM/MM MD	6.36 s	88 hrs 22 min	16 CPU-cores
GIAO/DFT	-	9 hrs 11 min	8 CPU-cores

SECTION S8 EXPERIMENTAL AND COMPUTATIONAL CONSIDERATIONS

With respect to protein size, the NMR-based detection of the ligand-bound state is limited by the correlation-time, setting an upper limit of about 50kDa for the protein-system under investigation. Expression of the protein in question must give sufficient yields in the order to reach concentrations of about 100 µM necessary for the detection of the ligand-bound state in the [^{1}H-^{1}H]-NOESY experiment. High affinity ligands in the slow exchange regime have to be available in similar concentrations for 1:1 and 2:1 ligand to protein mixtures. For ligands

in the intermediate exchange regime, often higher concentrations are necessary to be able to observe the free and bound exchange peaks in the NOESY. This can sometimes pose challenges if the ligand solubility is limited. Additionally, detection of the ligand-bound state is dependent on the kinetics of the protein-ligand binding event. In order to observe the exchange cross-peaks in the NOESY two conditions must be fulfilled. First, the residence lifetime must be on the order of the NOESY mixing time (~ 100 ms). Secondly, the exchange rate between free and bound state must be smaller than the frequency difference of the two states (to ensure that the slow to intermediate exchange limit prevails). Specifically, data obtained on a weak binder (K_D ~ 100 µM; data not shown) demonstrated the existence of a fast exchange regime (coalescence of signals) resulting in a limitation with to respect to weaker binding ligands.

The computation time required for a particular system will depend on the initial coordinates of the complex, the nature of the binding, and the level of theory (Table S1, Section S7). A more flexible system may need longer MD-based refinement steps, as for the case of low-affinity binders. The QM/MM MD calculation time depends on the level of theory and the size and chemical nature of the QM region. Additionally, QM/MM requires careful selection of the atoms belonging to the QM region.

References

[1] J. Bottcher, D. Dilworth, U. Reiser, R. A. Neumuller, M. Schleicher, M. Petronczki, M. Zeeb, N. Mischerikow, A. Allali-Hassani, M. M. Szewczyk, F. Li, S. Kennedy, M. Vedadi, D. Barsyte-Lovejoy, P. J. Brown, K. V. M. Huber, C. M. Rogers, C. I. Wells, O. Fedorov, K. Rumpel, A. Zoephel, M. Mayer, T. Wunberg, D. Bose, S. Zahn, H. Arnhof, H. Berger, C. Reiser, A. Hormann, T. Krammer, M. Corcokovic, B. Sharps, S. Winkler, D. Haring, X. L. Cockcroft, J. E. Fuchs, B. Mullauer, A. Weiss-Puxbaum, T. Gerstberger, G. Boehmelt, C. R. Vakoc, C. H. Arrowsmith, M. Pearson, D. B. McConnell, *Nat Chem Biol* **2019**, *15*, 822-829.

[2] J. Myers, G. Grothaus, S. Narayanan, A. Onufriev, *Proteins* **2006**, *63*, 928-938.

[3] R. Anandakrishnan, B. Aguilar, A. V. Onufriev, *Nucleic Acids Res* **2012**, *40*, W537-541.

[4] aS. Bietz, S. Urbaczek, B. Schulz, M. Rarey, *J Cheminform* **2014**, *6*, 12; bR. Fahrrolfes, S. Bietz, F. Flachsenberg, A. Meyder, E. Nittinger, T. Otto, A. Volkamer, M. Rarey, *Nucleic Acids Res* **2017**, *45*, W337-W343; cK. Schoning-Stierand, K. Diedrich, R. Fahrrolfes, F. Flachsenberg, A. Meyder, E. Nittinger, R. Steinegger, M. Rarey, *Nucleic Acids Res* **2020**, *48*, W48-W53.

[5] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, Williams, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, D. J. Fox, Wallingford, CT, **2016**.

[6] C. Tian, K. Kasavajhala, K. A. A. Belfon, L. Raguette, H. Huang, A. N. Migues, J. Bickel, Y. Wang, J. Pincay, Q. Wu, C. Simmerling, *J Chem Theory Comput* **2020**, *16*, 528-552.

[7] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, D. A. Case, *J Comput Chem* **2004**, *25*, 1157-1174.

[8] D.A. Case, K. Belfon, I.Y. Ben-Shalom, J.T. Berryman, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, G.A. Cisneros, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, G. Giambasu, M.K. Gilson, H. Gohlke, A.W. Goetz, R. Harris, S. Izadi, S.A. Izmailov, K. Kasavajhala, M.C. Kaymak, E. King, A. Kovalenko, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, M. Machado, V. Man, M. Manathunga, K.M. Merz, Y. Miao, O. Mikhailovskii, G. Monard, H. Nguyen, K.A. O'Hearn, A. Onufriev, F. Pan, S. Pantano, R. Qi, A. Rahnamoun, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, A. Shajan, J. Shen, C.L. Simmerling, N.R. Skrynnikov, J. Smith, J. Swails, R.C. Walker, J. Wang, J. Wang, H. Wei, R.M. Wolf, X. Wu, Y. Xiong, Y. Xue, D.M. York, S. Zhao, and P.A. Kollman, **2022**.

[9] T. Darden, D. York, L. Pedersen, *The Journal of Chemical Physics* **1993**, *98*, 10089-10092.

[10] S. Miyamoto, P. A. Kollman, *Journal of Computational Chemistry* **1992**, *13*, 952-962.

[11] aC. Bannwarth, S. Ehlert, S. Grimme, *J Chem Theory Comput* **2019**, *15*, 1652-1671; bC. Bannwarth, E. Caldeweyher, S. Ehlert, A. Hansen, P. Pracht, J. Seibert, S. Spicher, S. Grimme, *WIREs Computational Molecular Science* **2021**, *11*, e1493.
[12] F. Neese, *WIREs Computational Molecular Science* **2012**, *2*, 73-78.
[13] D. R. Roe, T. E. Cheatham, 3rd, *J Chem Theory Comput* **2013**, *9*, 3084-3095.
[14] aF. London, *J. Phys. Radium* **1937**, *8*, 397-409; bR. McWeeny, *Physical Review* **1962**, *126*, 1028-1034; cR. Ditchfield, *Molecular Physics* **1974**, *27*, 789-807; dK. Wolinski, J. F. Hinton, P. Pulay, *Journal of the American Chemical Society* **1990**, *112*, 8251-8260; eJ. R. Cheeseman, G. W. Trucks, T. A. Keith, M. J. Frisch, *The Journal of Chemical Physics* **1996**, *104*, 5497-5509.
[15] J. D. Chai, M. Head-Gordon, *Phys Chem Chem Phys* **2008**, *10*, 6615-6620.
[16] F. Weigend, R. Ahlrichs, *Phys Chem Chem Phys* **2005**, *7*, 3297-3305.
[17] M. Dracinsky, H. M. Moller, T. E. Exner, *J Chem Theory Comput* **2013**, *9*, 3806-3815.
[18] M. Ernzerhof, G. E. Scuseria, *The Journal of Chemical Physics* **1999**, *110*, 5029-5036.
[19] aA. D. Becke, *The Journal of Chemical Physics* **1993**, *98*, 5648-5652; bC. Lee, W. Yang, R. G. Parr, *Physical Review B* **1988**, *37*, 785-789.
[20] Y. Zhao, D. G. Truhlar, *Theoretical Chemistry Accounts* **2008**, *120*, 215-241.
[21] aT. Zhu, J. Z. H. Zhang, X. He, *Journal of Chemical Theory and Computation* **2013**, *9*, 2104-2114; bA. H. Mazurek, L. Szeleszczuk, D. M. Pisklak, *Int J Mol Sci* **2021**, *22*; cR. Atwi, Y. Chen, K. S. Han, K. T. Mueller, V. Murugesan, N. N. Rajput, *Nature Computational Science* **2022**, *2*, 112-122.
[22] aE. R. Johnson, S. Keinan, P. Mori-Sánchez, J. Contreras-García, A. J. Cohen, W. Yang, *Journal of the American Chemical Society* **2010**, *132*, 6498-6506; bJ. Contreras-Garcia, E. R. Johnson, S. Keinan, R. Chaudret, J. P. Piquemal, D. N. Beratan, W. Yang, *J Chem Theory Comput* **2011**, *7*, 625-632; cR. A. Boto, F. Peccati, R. Laplaza, C. Quan, A. Carbone, J. P. Piquemal, Y. Maday, A. J. Contreras-Garci, *J Chem Theory Comput* **2020**, *16*, 4150-4158.
[23] L. Schrödinger, & DeLano, W., **2020**.

4

NMR AND COMPUTATIONAL STUDY OF STREPTAVIDIN–BIOTIN BINDING

A. L. Ptaszek, S. Kratzwald, F. Sagan, *et al.*, "From weak interactions to strong affinity: Deciphering the streptavidin-biotin interaction through nmr and computational analysis," *bioRxiv*, pp. 2024–12, 2024 [113]

DOI: 10.1101/2024.12.19.629369
The manuscript is currently under the second revision in The Journal of Physical Chemistry B.

PERSONAL CONTRIBUTION

Conceptualization, simulations, data analysis, validation, writing, visualization, shared First Author with S. Kratzwald

From Weak Interactions to Strong Affinity: Deciphering the Streptavidin-Biotin Interaction through NMR and Computational Analysis

Aleksandra L. Ptaszek,[†,‡,¶,@] Sarah Kratzwald,[§,¶,‖,@] Filip Sagan,[⊥] Mario Migotti,[¶] Pedro A. Sánchez-Murcia,[‡,#] Robert Konrat,[†,¶,§] and Gerald Platzer[*,§,†]

†*Christian Doppler Laboratory for High-Content Structural Biology and Biotechnology, Department of Structural and Computational Biology, Max Perutz Labs, University of Vienna, Campus Vienna Biocenter 5, 1030-Vienna, Austria*

‡*Laboratory of Computer-Aided Molecular Design, Division of Medicinal Chemistry, Otto-Loewi Research Center, Medical University of Graz, Neue Stiftintalstr. 6/III, A-8010 Graz, Austria*

¶*Department of Structural and Computational Biology, Max Perutz Labs, University of Vienna, Campus Vienna Biocenter 5, 1030-Vienna, Austria*

§*MAG-LAB GmbH, Karl-Farkas-Gasse 22, 1030- Vienna, Austria*

‖*Institute of Organic Chemistry, Faculty of Chemistry, University of Vienna, Währingerstr. 38, Vienna 1090, Austria*

⊥*Faculty of Chemistry, Jagiellonian University, Gronostajowa 2, 30-387 Cracow, Poland*

#*BioTechMed-Graz, Mozartgasse 12/II, A-8010 Graz, Austria*

@*Equal contribution*

E-mail: gerald.platzer@mag-lab.eu

Abstract

Understanding weak interactions in protein-ligand complexes is essential for advancing drug design. Here, we combine experimental and quantum mechanical approaches to study the streptavidin-biotin complex, one of the strongest interacting protein-ligand systems. Using a monomeric streptavidin mutant, we analyze ^{1}H NMR chemical shift perturbations (CSPs) of biotin upon binding, identifying remarkable upfield shifts of up to -3.2 ppm. Quantum chemical calculations attribute these shifts primarily to aromatic ring currents, with additional contributions from charge transfer effects linked to weak interactions. The agreement between experimental and computed chemical shifts validated the X-ray structure as a reliable basis for detailed computational analyses. Energy decomposition analysis reveals that electrostatics dominate the biotin-streptavidin interaction, complemented by significant orbital and dispersion contributions. Notably, weak non-covalent interactions, such as CH$\cdots$S, CH$\cdots\pi$, and CH$\cdots$HC contacts, driven by London dispersion forces, contribute $\sim$44% to the complex's stability.

Introduction

Nuclear Magnetic Resonance (NMR) spectroscopy is an indispensable tool in chemistry and structural biology, providing atomic-level insights into molecular dynamics, conformational changes, and intricate intermolecular interactions. Its unique ability to examine molecules under near-physiological conditions makes NMR particularly valuable for studying complex systems such as proteins and protein-ligand interactions.[1]

Traditionally, scalar couplings (J-couplings) have been extensively employed to probe bonding relationships and conformational states, particularly in small molecules. The presence of scalar couplings, which were initially observed in covalent bonds, is now known to extend to hydrogen bonds, demonstrating their complex nature driven by both electrostatic forces and electronic density transfer.[2] Over recent years, *through-space* scalar couplings have been identified in interactions at the very edge of hydrogen bonding, including CH$\cdots\pi$ contacts, where a soft acidic CH bond interacts with a soft basic π-system.[3,4] Even more intriguingly, such couplings have been detected between homopolar CH$\cdots$HC contacts, suggesting that these interactions, though weak, exhibit characteristics of true bonding.[5–8] While J-couplings provide rich structural information, their application is challenging in large protein systems. The spectral crowding and faster relaxation times in these systems reduce the practical utility of J-couplings, particularly for routine applications like drug discovery.

Another valuable NMR observable that provides extensive insight into the local chemical environment is the chemical shift. For example, a proton chemical shift serves as an effective indicator of hydrogen bonding, where stronger hydrogen bonds lead to increased deshielding effects. Additionally, the proximity of aromatic rings induces shielding effects due to ring currents, perturbing chemical shifts.[9] Upon protein binding, ligand proton chemical shifts can change due to a combination of factors, including loss of hydrogen bonds with solvent and formation of new hydrogen bonds within the binding pocket. Moreover, binding modes frequently involve aromatic rings, leading to additional shielding effects from ring currents. These combined factors provide a rich source of structural information that can be harnessed

to study protein-ligand interactions in detail.

In this study, we explore the interaction between streptavidin and biotin, which is recognized as one of the most robust non-covalent interactions found in nature, and has been a target of numerous studies in the past, Figure 1.[10–14] The high affinity of this interaction is exploited within a wide range of biotechnological applications (e.g. *Strep-tags*).[15] Despite long-lasting scientific interest in the streptavidin-biotin complex, the origin of the high binding strength remains elusive.

When binding to wt streptavidin, biotin engages in a strongly enthalpic interaction (ΔH = 24.5 kcal/mol), along with an entropic penalty (TΔS = -6.4 kcal/mol).[16] Previous studies that focused on the wt streptavidin have investigated the contribution of the entropic term for biotin´s high binding affinity. In calculations, a highly structured five-ring of water molecules was identified in the polar region of the wt streptavidin´s biotin binding pocket.[17] Further calculations have demonstrated that biotin needs to compensate for an enthalpic binding energy of -6.85 kcal/mol, through entropic compensation, to optimize its interaction in the wt streptavidin.[10,18] In this work, however, we solely focus on the enthalpic contribution of the binding energy of biotin to a monomeric mutant. Hence, the entropic contribution will not be further discussed.

Besides a high shape complementarity, a large number of classical hydrogen bonds involving the biotin head-piece nitrogens and oxygen as well as the carboxy terminus of the aliphatic tail contribute significantly to the high binding enthalpy of the system. It is the "hydrophobic tail" however, that adds several non-classical hydrogen bonding interactions that are rarely discussed in a protein-ligand binding context which have been shown to majorly contribute to the total free binding energy of the system. The term "hydrophobic interaction" is often misused in the context of CH, CH_2 and CH_3 groups interacting with their environment. In addition to hydrogen bonds to classical acceptors (O,N,S) and the distinct case of Fluorine,[19,20] these groups also form CH$\cdots\pi$[3,9] interactions with nearby aromatic and homopolar CH$\cdots$HC interactions[21–23] to nearby aliphatic residues. Interactions

of carbon-bound protons are of special interest given their high frequency in protein-ligand complexes accounting for over 50% of all non-covalent interactions present in guest-host systems.[24] In addition, their presence can be detected straightforwardly via simple 2D NMR experiments.[9] Since native Streptavidin (PDB ID: 3RY2) forms a tetramer in solution, we opted to use a slightly modified, monomeric form of Streptavidin (PDB ID: 4JNJ) to make the protein construct amenable for NMR measurements.[25,26] In the streptavidin-biotin complex under study (PDB ID: 4JNJ), only two out of fifteen hydrogen atoms form classical NH$\cdots$O hydrogen bonds with Asn45 (Ser45 in PDB ID: 3RY2) and Asp128, Figure 1 c). Beyond that, the X-ray structure reveals weaker CH$\cdots$O interactions with Thr48 (Val47 in PDB ID: 3RY2) and Gly49, and a CH$\cdots$S contact with Cys59, Figure 1 c,d). Further, biotin's binding pocket showcases multiple CH$\cdots\pi$ interactions with aromatic rings of Trp79 and Trp108 residues. Additionally, homopolar CH$\cdots$HC interactions with Leu110 highlight the complex's reliance on non-classical, dispersion-driven forces that collectively support biotin's snug fit within streptavidin. Amid the ongoing scientific debate, considerable research has been devoted to deciphering the nature of CH$\cdots$HC interactions. Until recently, they were mostly perceived with the notion of "steric crowding", but more and more works emphasize the stabilization provided by such contacts, stemming jointly from London dispersion and orbital interaction.[27–41]

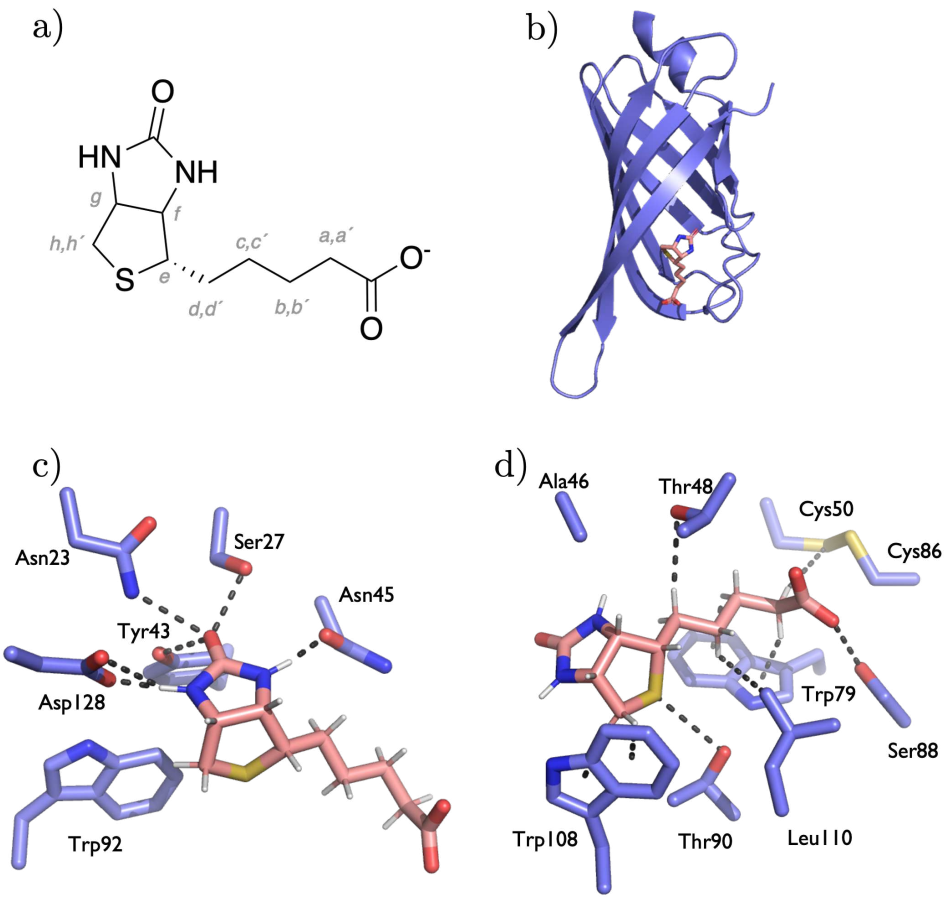


Figure 1: (a) Chemical structure of biotin at pH 7.4 with H-labeling used throughout this work; (b) X-ray structure of mSA2 and biotin (PDB ID 4JNJ);[25] (c) Interactions of biotin's ureido ring with streptavidin pocket; (d) Interactions of biotin´s tetrahydrothiophene ring and valeric acid side chain with surrounding streptavidin residues.

Previously, we reported the PI by NMR methodology which allows the quantification of the strength of these CH$\cdots\pi$ interactions.[9] Recently, the approach has been developed further. Thereby, protein deuteration and ligand-based ^{1}H NMR experiments were used to probe CH^{ligand}-$\pi^{protein}$ interactions by ^{1}H NMR experiments.[42] We showed that ligand proton chemical shifts, when combined with computational shift predictions and molecular dynamics simulations, can refine protein-ligand interfaces, providing valuable insights into solution structures. This study employs a ^{1}H-^{1}H NOESY experiment to quantify chemical shift perturbation (CSP) for the carbon-bound protons of biotin in both its free and bound

forms. To validate the solution structure against crystallographic data, we compare the experimentally obtained shifts with quantum mechanically (QM) predicted shifts based on the X-ray structure of the streptavidin-biotin complex. Following this validation, we use modern chemical bonding descriptors such as ETS-NOCV[43] and LED-DLPNO-CCSD(T)[44–51] to explore the interactions of biotin with the binding pocket, focusing on the non-classical hydrogen bonds. We also decompose the carbon-bound proton chemical shifts to separate the shielding effects of environmental ring currents from deshielding due to hydrogen bond formation. Our findings show that the large CSPs are primarily due to ring current effects, with hydrogen bonding playing a secondary role. In conclusion, by integrating NMR and computational methods, we offer a comprehensive understanding of the forces that govern biotin-streptavidin molecular interactions.

Methods

Plasmid Design. DNA sequences encoding the fusion construct H6-MBP-3C-mSA2 were synthesized by GeneScript and cloned into the pETM44 vector via Sequence and Ligation Independent Cloning (SLIC). This construct integrates a hexahistidine (H6) tag and maltose-binding protein (MBP) for purification, linked to the mSA2 protein through a 3C protease cleavage site to allow for H6-MBP post-purification. The integrity and sequence accuracy of the insert were confirmed by sequencing.

Protein Expression and Purification. Recombinant H6-MBP-3C- andmSA2 (monomeric streptavidin 2) was expressed in *E. coli BL21 (DE3)* and contains N-terminal 3C protease cleavable H6-MBP tags. ^{15}N labeled H6-MBP-3C-mSA2 was expressed in M9 medium containing $^{15}NH_4Cl$ (1 g/l). Cells were grown to an OD_{600} (Optical Density at 600 nm) of 0.9 and induced with 0.4 mM IPTG (isopropyl-β-D-1-thiogalactopyranosid). The expression medium was additionally supplemented with biotin (1 mg/l) for the mSA2-biotin construct.

Cells were harvested after 16 h of expression at 20°C. Perdeuterated H6-MBP-3C-mSA2 was expressed by transferring 100 μl of transformed *E. coli BL21(DE3)* cells from M9 medium (no isotope-enriched material, 100% H_2O) to $M9^{50\%D2O}$ (no isotope-enriched material, 50% D_2O, 50% H_2O) until an OD_{600} of 0.7 was reached. A small amount of *E. coli* cells were transferred to 50 ml $M9^{100\%D_2O}$ (no isotope-enriched material, 100% D_2O) preculture. The next day, a small aliquot of the cells was pelleted and transferred to the $M9^{100\%D_2O}$ expression medium (D-glucose-d7, 100% D_2O). After reaching an OD_{600} of 0.9, induction was performed using 0.4 mM IPTG. Cells were harvested after 20 h at 20°C, and the pellet was resuspended in 30 ml of lysis buffer (20 mM TrisHCl pH 7.4, 300 mM NaCl, 10 mM Imidazole, 1 mM TCEP, 0.5% (v/v) CHAPS, 1 uM PMSF, 1 mM Biotin, Protease Inhibitor Cocktail). Bacteria were lysed by sonication (4 x 3 min, 50% amplitude, 01 s on 01 s off). Proteins were purified by Ni-NTA (2 x HiTrap Chelating HP, 5 mL, GE Healthcare), buffer exchanged to cleavage buffer (20 mM TrisHCl pH 7.4, 100 mM NaCl, 1 mM TCEP), and cleaved with an in-house 3C protease (H10-GST-3C) overnight at 4°C. The cleaved protein was again loaded to a Ni-NTA to bind the cleaved H6-MBP and the H10-GST-3C. The flow-through containing mSA2 was collected. The mSA2-containing fractions were pooled, concentrated and loaded onto a gel filtration column (HiLoad 16/600 Superdex 75 pg) equilibrated with phosphate buffered saline (PBS) buffer pH 7.4 (containing 0.5 mM TCEP). The mSA2-containing fractions were concentrated in an Amicon Ultra-15 centrifugal filter device (3 kDa MW cutoff) and either directly measured by NMR or buffer exchanged ($PBS^{100\%D2O}$) using the Amicon Ultra-15 centrifugal filter device. Protein concentration was measured using a NanoDrop. For apo-mSA2, no biotin was added to the expression medium, and the sample was used immediately after the second Ni-NTA column after the cleavage of the H6-MBP tags due to the limited stability of apo-mSA2 in solution.

NMR Spectra Acquisitions. 1D and NOESY experiments were carried out at 298 K on a Bruker 600 MHz spectrometer equipped with a TCI cryoprobe. The initial sample

concentration was 860 μM. For 1D experiments, the number of scans was set to 512. For the NOESY experiment, the number of scans was set to 32, the size of fid was set to 2048, 512 (F2, F1) and the mixing time was set to 100 ms.

Ligand ^1H-NMR chemical shifts computation. The X-ray crystal structure of biotin bound to the mutant streptavidin mSA2 (PDB ID: 4JNJ)[25] was used as the starting model for the computational study. The protein was protonated at pH 7.4 using the H++ web server (version 4.0),[52] while the ligand was protonated using the ProteinsPlus server.[53–56] The restrained electrostatic potential (RESP) point charges for biotin were computed by means of HF/6-31G(d) on the optimized geometry of the ligand at the B3LYP-D3/6-311G(d,p) level of theory using Gaussian16.[57] The Amber ff19SB force field[58] was applied to the protein, while the General Amber Force Field (GAFF)[59] was used for the ligand. The system was neutralized by adding chloride ions and solvated in an octahedral box of TIP3P water molecules,[60] extending 10 Å from the solute, using the LEaP module of AmberTools21.[61] Water molecules present in the crystal structure were retained. Energy minimization was conducted under periodic boundary conditions using the Amber20 software.[61] Non-bonded interactions were calculated with a 10 Å cutoff, while long-range electrostatic interactions were treated using the smooth Particle Mesh Ewald (PME) method.[62] The SHAKE algorithm was applied to constrain all bonds involving hydrogen atoms.[63] Minimization was performed in two stages: initially, 10,000 cycles using the steepest descent algorithm, followed by 10,000 cycles with the conjugate gradient method. In the first stage, only hydrogen atoms were minimized, followed by the minimization of solvent molecules in the second stage.

The ligand, accompanied by two water molecules H-bonded to its carboxylate group, was chosen for further QM calculations, alongside interacting amino acids: Asn45, Ala47, Trp79, Cys86, Ser88, Thr90, Trp108, Leu110, Asn23, Ser27, Tyr43, Trp92, Asp128, Thr48, Gly49 and Cys50. The side chains up to the C_α atoms of Asn45, Ala47, Trp79, Cys86, Ser88, Thr90, Trp108, and Leu110 were included, with the C_α-N and C_α-$C_{Carbonyl}$ bonds

being cleaved and filled with hydrogen atoms. For Asn23, Ser27, Tyr43, Trp92, and Asp128, the bonds between C_β and C_α were cut and filled with hydrogen atoms. Additionally, the sequence Thr48-Gly49-Cys50 underwent cleavage at the C_α-N bond of Thr48 and the C_α-$C_{Carbonyl}$ bond of Cys50, followed by filling with hydrogen atoms.

The isotropic chemical shielding values were determined using the Gauge-Independent Atomic Orbital (GIAO) approach[64–68] and the wB97XD functional[69] in combination with the def2-TZVP basis set, as implemented in Gaussian16.[57] The choice of theory level is based on the benchmark carried out in our previous study.[42]

To compare the prediction with experimental data, the calculated ^{1}H NMR chemical shielding values (σ) need to be converted to chemical shifts (δ) using a standard reference:

$$\delta = \sigma_{standard} - \sigma \tag{1}$$

In this study, we used a linear model with a fixed slope of -1 to calculate the chemical shielding of the reference standard, specifically water.[42]

^1H-NMR chemical shifts decomposition. Building on the approach outlined by Scheiner,[70] we decompose the chemical shift of protons in the bound form into two physically meaningful components: the "Environmental shielding effect" and the "Deshielding due to H-bond formation," as referenced further in the text. Following the setup of the QM region, we computed the "Environmental shielding effect" by removing the ligand and replacing its C-bound protons with "dummy atoms" positioned at the original proton locations. Shielding was then calculated at these precise positions. This effect arises from the electron density of hydrogen-accepting groups (here, interacting residues), which can respond to external magnetic fields even without forming a hydrogen bond (HB) or having a proton donor nearby. Therefore, this positional effect is independent of any hydrogen bond formation.

If a hydrogen bond does form, there are additional contributions from charge transfer

between the donor and acceptor, along with internal polarization changes within each group. This "Deshielding due to H-bond formation" was calculated by subtracting two components from the chemical shifts computed for the entire ligand-binding pocket system: (1) the chemical shifts of the ligand alone and (2) “Environmental shielding effect”. As demonstrated in Scheiner's work,[70] this component serves as an indicator of the presence and strength of a hydrogen bond, as it has shown correlation with binding energy in model systems.

Extended Transition State Natural Orbitals for Chemical Valence (ETS-NOCV) charge and energy decomposition method. Extended Transition State (ETS)[71] is applied to partition the total binding energy between interacting fragments into chemically meaningful contributions:

$$\Delta E_{total} = \Delta E_{elstat} + \Delta E_{disp} + \Delta E_{Pauli} + \Delta E_{orb} \tag{2}$$

The first term, ΔE_{elstat}, in the equation represents the classical electrostatic interactions between selected fragments. The ΔE_{disp} concerns semiempirical Grimme dispersion correction (D3).[72] The third component, ΔE_{Pauli}, accounts for Pauli repulsion between occupied orbitals of fragments. The last, orbital interactions contribution, ΔE_{orb}, is stabilizing and covers the interaction energy between occupied orbitals of the first fragment with unoccupied orbitals of the second one and vice versa, as well as the polarization effects. Natural Orbitals for Chemical Valence (NOCV) theory[73] enables the decomposition of the overall deformation density, $\Delta\rho_{orb} = \rho - \rho_0$ (where ρ is a density of the molecule and ρ_0 is a density of non-interacting fragments), into the distinct bonding channels such as σ, π, δ, etc: $\Delta\rho_{orb} = \sum_i \Delta\rho_{orb}(i)$.

Finally, combining the ETS method with Natural Orbitals for Chemical Valence (NOCV) theory,[43] allows us to quantify the energy values of the ΔE_{orb}, corresponding to $\Delta\rho_{orb}(i)$ contributors: $\Delta E_{orb} = \sum_i \Delta E_{orb}(i)$. Overall, the ETS-NOCV method provides a qualitative and quantitative, picture of a chemical bonding.

ETS-NOCV analyses were performed using the Amsterdam Density Functional (ADF) software, version 2023.104.[74,75] The BLYP functional, coupled with Grimme's D3 dispersion correction with the Becke-Johnson damping,[72] was applied for bonding analysis due to its recognized reliability in describing non-covalent interactions.[76,77] All calculations utilized a triple-ζ STO basis set with polarization functions (TZP), and were conducted without imposing symmetry constraints, using an all-electron basis set.

The Domain-based Localized Pair-Natural Orbital Singles and Doubles Coupled Cluster with perturbative Triples (DLPNO-CCSD(T)) and Local Energy Decomposition. DLPNO-CCSD(T)[44–48] Local Energy Decomposition (LED)[49–51] calculations were performed in ORCA 5.0.4 program,[78] using def2-TZVP basis,[79,80] RIJCOSX approximation[81] and TightPNO settings for PNO convergence.

The Domain-based Localized Pair-Natural Orbital Singles and Doubles Coupled Cluster with perturbative Triples (DLPNO-CCSD(T)) is a local approximation of the Coupled Clusters method, which itself is often referred to as a 'golden standard' for quantum-chemical calculations. Its main feature is the localization of the orbitals and generation of the Projected Atomic Orbitals (PAOs) from occupied localized orbitals, followed by creation of Pair Natural Orbitals from selected PAOs.

Local Energy Decomposition[49–51] is defined within the DLPNO-framework, and decomposes the interaction energy into the Hartree-Fock and correlation energies:

$$\Delta E_{int} = \Delta E_{int}^{HF} + \Delta E_{int}^{C} + \Delta E_{el-prep} \tag{3}$$

Electronic preparation terms of the distorted fragments ($\Delta E_{el-prep}^{C}$ and $\Delta E_{el-prep}^{HF}$) were added for clarity. Hartree-Fock interaction energy consists of the electrostatic and exchange terms:

$$\Delta E_{int}^{HF} = \Delta E_{elstat} + \Delta E_{exch} \tag{4}$$

whereas the correlation interaction consists of the following terms:

$$\Delta E^{C}_{int} = \Delta E^{C}_{CT1\rightarrow 2} + \Delta E^{C}_{CT1\rightarrow 2} + \Delta E_{disp} + \Delta E^{C}_{rest} \tag{5}$$

where ΔE^{C}_{CT} terms correspond to the instantaneous strong-pair charge transfer terms, ΔE_{disp} is a sum of dispersion stemming from strong and weak pairs, and ΔE^{C}_{rest} is a summation of the non-dispersive part of weak-pair interaction energy and the triples correction.

Results and Discussion

Detection of Chemical Shift Perturbations of biotin's H-Atoms

In this work we study a monomeric mutant of the streptavidin protein (mSA2)[25,26] which, due to its more favorable relaxation properties is more suitable for NMR applications compared to the homotetrameric wild-type (WT). In the tetrameric WT streptavidin, a tryptophan residue (Trp120) of an adjacent subunit forms additional stacking interactions with biotin which is associated with the extraordinarily high wild-type affinity towards biotin.[82] The mSA2 mutant used in this study has a binding affinity in the low nanomolar range ($K_D = 1.9 \pm 0.4 \times M^{-9}$),[26] whereas the WT streptavidin has a binding affinity in the picomolar range ($K_D = 4 \times 10^{-14}$ M).[83] The mSA2 mutant is a hybrid of rhizavidin and streptavidin (Figure S1). Unlike streptavidin, rhizavidin binds to biotin with a single subunit. In creating mSA2, amino acids in the biotin binding pocket of streptavidin were replaced with those from rhizavidin. Notably, many of these residues are common to both proteins (see Figure S2).[84]

To determine chemical shift perturbations (CSP) of biotin upon binding to streptavidin, we perdeuterated mSA2 to render all protein ^{1}H signals NMR inactive. To evaluate CSPs for each of biotin's non-exchanging hydrogen atoms, ^{1}H peak assignment of biotin's protons was performed as described previously[42] via a ^{1}H-^{1}H NOESY experiment of a perdeuterated 1:2

streptavidin-biotin complex. This allows for an unambiguous correlation of biotin's unbound H-signals with biotin's H-signals bound to streptavidin (see Figure S4).

Figure 2 shows the ^{1}H NMR spectra of free biotin (top) and biotin bound to deuterated streptavidin (bottom). Remarkably, CSPs of -3.2, -2.9, -2.1 and -1.9 ppm for biotin's H-atoms *h, b, h´* and *a* could be observed. Large ^{1}H upfield shifts indicate shielding for affected H-atoms which is induced by the π cloud of aromatic ring systems when located above/below of the respective H-atom thereby forming favorable CH$\cdots\pi$ interactions with streptavidin's aromatic indole moieties of Trp79 (for *a* and *b*) and Trp108 (for *h* and *h´*) respectively (see Figure 2 b). The interaction can be considered as a special case of hydrogen bonding where the C-H group serves as the H-bond donor and the aromatic ring as the H-bond acceptor.[85]

All H-atoms of biotin's aliphatic tail (*a, a´, b, b´, c, c´, d, d´ and e*) undergo an upfield shift upon binding (min - 0.49 ppm; max - 2.89 ppm). The general observation that carbon-bound ligand protons shift to lower ppm values upon binding can be attributed to two factors; 1) Protons moving from the solvated free state to a solvent-inaccessible bound state lose hydrogen-bonding to water resulting in an apparent upfield shift (hydrogen-bonding itself has a deshielding character) and 2) ring systems exert their effect up to a distance of 4Å away and can influence protons that are not directly interacting with them. (*b´,c, c´, d, d´, e*). The last section of the results discusses the detailed decomposition of the ^{1}H chemical shifts of biotin in the bound form into environmental shielding and deshielding due to the hydrogen bond.

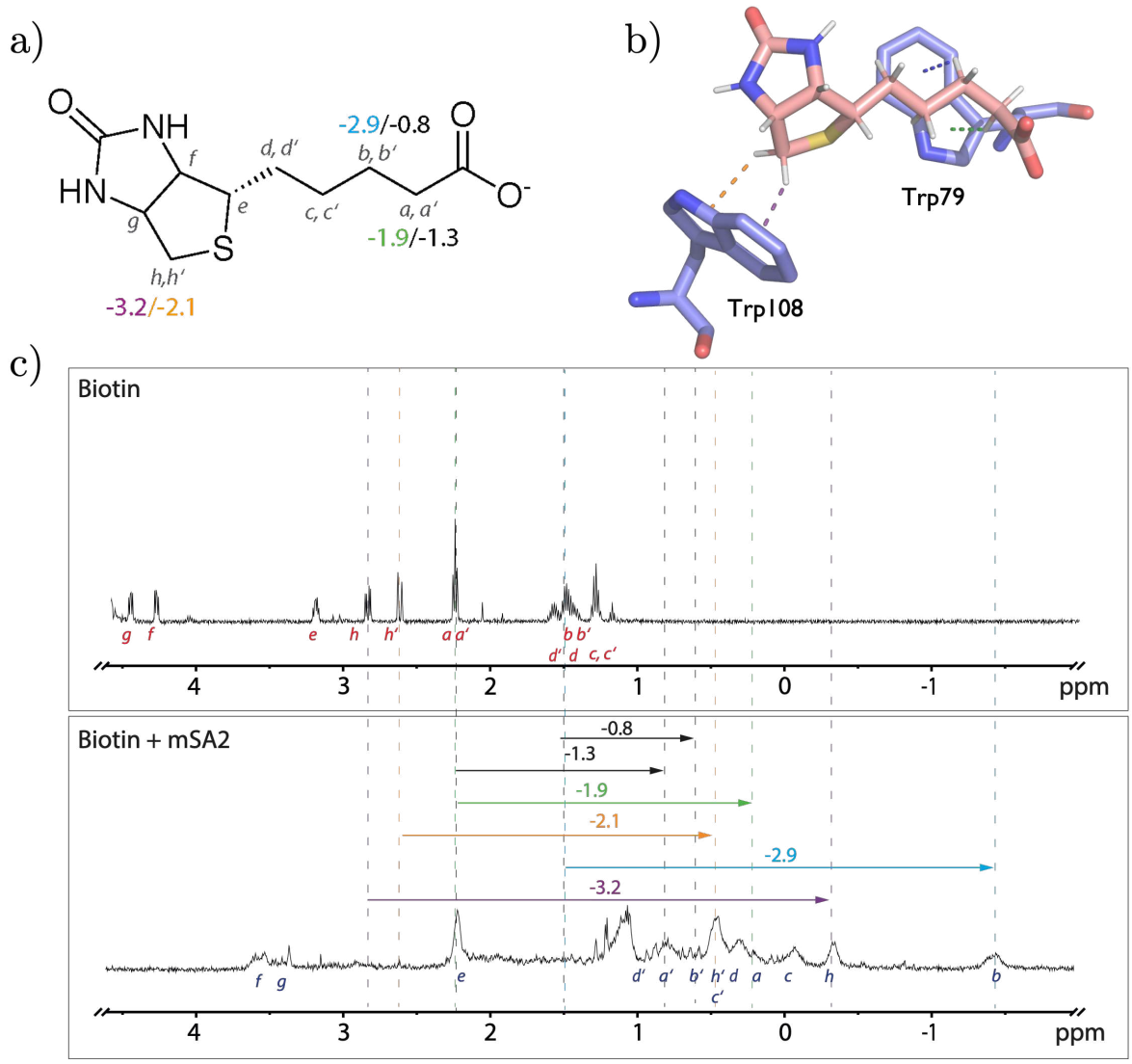


Figure 2: (a) Chemical structure of biotin highlighting the protons undergoing the largest CSP. (b) Crystal structure of biotin-mSA2. Protons a, b, h and h' are placed right in top of the indole moiety of Trp79 (proton a and b) and Trp108 (proton h and h'). (c) *top:* ^{1}H-NMR of biotin in PBS buffer, *bottom:* ^{1}H-NMR of biotin bound to deuterated mSA2. The dashed lines depict the ^{1}H CSPs of biotin's free to bound state for proton a, b, h and h'. Detailed results provided in Supplementary Table S1.

^{1}H NMR Chemical Shifts Prediction

In a previous study, we investigated the ^{1}H chemical shifts (CSs) of the BI-9321 molecule bound to the PWWP1 domain of NSD3.[42] By validating calculated CSs against experimen-

tal ligand ^{1}H CSs in the bound form, we identified significant discrepancies for certain protons. To address these, we applied a quantum mechanics/molecular mechanics (QM/MM) molecular dynamics (MD) ensemble approach to refine the initial X-ray structure, which significantly improved agreement with the experimental ^{1}H CSs in solution, achieving an absolute error (RMSE) of less than 0.5 ppm for each proton and a root mean square error (RMSE) of 0.28 ppm.

In this study, we used the same theoretical approach to compute chemical shieldings for the biotin-mSA2 X-ray structure (PDB ID: 4JNJ)[25] and validated the results with experimental ^{1}H CSs. Figure 3 a) demonstrates a strong correlation between calculated and experimental values, with an RMSE of 0.24 ppm. Furthermore, absolute errors for all protons remained below 0.5 ppm (Figure 3 b). These results confirm that the X-ray structure of biotin-mSA2 provides an accurate representation of the complex in solution. Notably, unlike our previous study where QM/MM MD refinement was required to achieve agreement with experimental data,[42] here the chemical shifts calculated directly from the static X-ray structure already yielded higher accuracy, so that conformational sampling through classical or QM/MM MD simulations was not necessary.

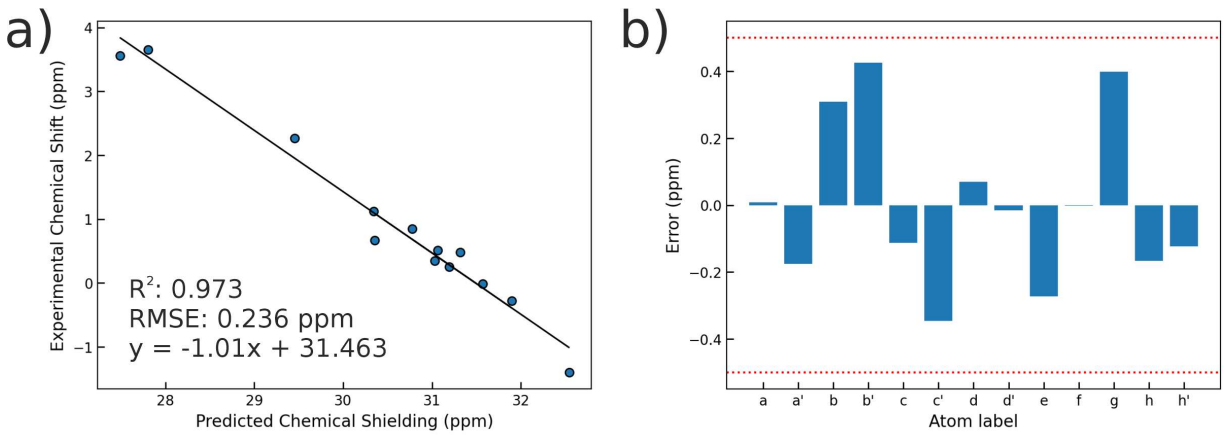


Figure 3: a) Correlation between experimental NMR ^{1}H CSs (ppm) and calculated chemical shielding (ppm) from the X-ray crystal structure (PDB ID: 4JNJ). b) Differences between calculated and experimental chemical shifts (ppm). GIAO/wB97X-D/def2-TZVP level of theory was applied.[64–69]

Biotin-mSA2 Interaction Energy Analysis

To gain a comprehensive understanding for the physico-chemical properties driving Biotin-mSA2 complex formation, we used the ETS-NOCV charge and energy decomposition scheme[43] at the DFT level as well as the LED energy decomposition based on the DLPNO approximation of the *gold-standard* CCSD(T) method.

In the first step, we consider the interaction of the biotin molecule (*fragment 1*) with the binding site and water molecules in direct contact with biotin (*fragment 2*) using the geometric cluster previously defined for the ^{1}H CS calculations, as described in the Methods section, Figure 4 a). The overall interaction energy reaches -86.67 kcal/mol and is dominated by electrostatic interactions (40%) followed by orbital (36%) and dispersion (24%) components, Table 1 a). The deformation density reveals numerous charge-flow channels associated with conventional $N/O - H \cdots O$ hydrogen bond formation in the ureido ring and carboxylate regions, as well as weaker $C - H \cdots O$ and $O - H \cdots S$ hydrogen bonds, Figure 4 a). In order to exclude the strong hydrogen bonds with water molecules from the interaction energy, we include water molecules in *fragment 1* getting the total interaction energy equal to -83.36 kcal/mol (Table 1 b), Figure 4 b)) which is in an excellent agreement with the *state-of-the-art* DLPNO-CCSD(T) interaction energy equal to -83.23 kcal/mol. In comparison to the initial system (Table 1 a)), the absence of two prominent hydrogen bond donors results in a rise of approximately 20 kcal/mol in both electrostatic and orbital energy terms, which is counterbalanced by the reduction in Pauli repulsion.

Delving deeper into the contributions of individual regions within the binding pocket to the overall stability, we analyse two distinct cases: (1) the interactions of biotin with only those residues from the binding pocket that interact with its ureido ring, referred to in the text as the *ureido subpocket* (Figure 1 c)), and (2) interactions of biotin with the remaining residues of the binding pocket, referred to as the *tail subpocket* (Figure 1 d)). Total interactions between the biotin and the ureido subpocket amount to approximately 28% of biotin-mSA2 stability, Table 1 c). Analysis of the deformation density uncovers

charge-flow channels related to the formation of five classical hydrogen bonds (Figure 4 c)), yielding an energy of -41.40 kcal/mol, which accounts for approximately 43% of the overall attractive forces in this system. Additionally, electrostatic interactions contribute around 36% of the overall attraction. The dominance of orbital and electrostatic factors aligns with prior analyses of classical hydrogen bonds.[86]

Focusing now on the interactions of biotin with the tail subpocket, electrostatic forces emerge as the primary contributor, constituting 38% of the interactions, likely influenced by the negatively charged carboxylic acid moiety at physiological pH (pK_a ∼4.5) (Table 1 d). Interestingly, calculations indicate that approximately 35% of the stabilizing interactions are attributed to dispersion forces. To validate this empirical estimation of London dispersion, we cross-referenced it with results obtained using the LED-DLPNO-CCSD(T) level of theory, and found the DFT overestimating dispersion (-51.6 kcal/mol vs -42.8 kcal/mol) in this fragmentation. Nevertheless, the dispersion contribution within the biotin - tail subpocket model remains twice as high as for the biotin - ureido subpocket, owing to much higher degree of packing around the hydrophobic tail. As evidenced by the Dispersion Interaction Density (Figure S6 b), the entirety of tail section is engaged in dispersive stabilization, whereas in the ureido subpocket case, only hydrogen bond donors and acceptors are stabilized in this way (Figure S6 a). Furthermore, orbital part of the energy is on par with the ureido subpocket fragment, and constitues 27% of total stabilization within the tail subsystem (Table 1 d), Figure 4 d). Figure S5 illustrates the NOCV-based deformation density channels, highlighting both σ and notably weaker (∼-1 kcal/mol) π components of the $N - H \cdots O$ hydrogen bonds. Additionally, the analysis reveals the involvement of sulfur in hydrogen bonding interactions, not only in the $O - H \cdots S$ bond with Thr90, but also in weaker (∼ -1.6 kcal/mol) $C - H \cdots S$ bonds with Trp79 and the methyl group of Thr90. Moreover, the NOCV channels corresponding to the $C - H \cdots \pi$ interactions are identified, Figure S5.

Table 1: ETS energy decomposition results describing the interaction of biotin with the binding site: a) biotin molecule interacting with two water molecules and binding pocket, b) biotin and two water molecules interacting with the binding pocket, c) biotin and two water molecules interacting with the ureido subpocket (Asn23, Ser27, Tyr43, Asn45, Trp92, Asp128), d) biotin and two water molecules interacting with the tail subpocket (Ala47, Thr48, Gly49, Cys50, Cys86, Trp79, Ser88, Thr90, Trp108, Leu110). The percentages represent the proportionate contribution to the overall attractive interactions ($E_{elstat} + \Delta E_{orb} + \Delta E_{disp}$). All energies in kcal/mol.

Energy term	**a) Biotin $\cdots$ pocket+H_2O**	**b) Biotin+H_2O $\cdots$ pocket**	**c) Biotin+H_2O $\cdots$ ureido subpocket**	**d)Biotin+H_2O $\cdots$ tail subpocket**
ΔE_{total}	-86.67	-83.36	-23.49	-61.27
ΔE_{elstat}	-113.16 (40%)	-89.17 (37%)	-34.85 (36%)	-55.57 (38%)
ΔE_{orb}	-101.60 (36%)	-79.46 (33%)	-41.40 (43%)	-39.55 (27%)
ΔE_{disp}	-68.72 (24%)	-71.93 (30%)	-20.35 (21%)	-51.59 (35%)
ΔE_{Pauli}	196.82	157.20	73.11	85.45

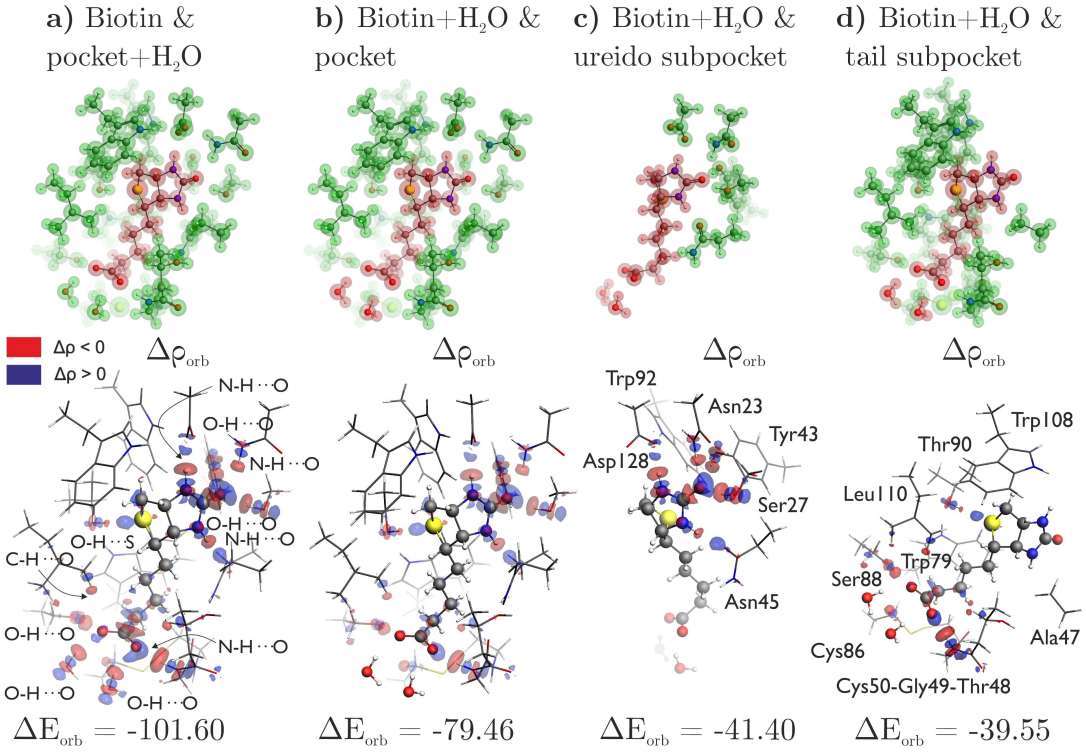


Figure 4: The initial row illustrates the ETS-NOCV fragmentation using red and blue, as described in Table 2. Below, the overall deformation density $\Delta\rho_{orb}$ contours with the corresponding ΔE_{orb} for 4 models. The biotin molecule is represented by balls and sticks, and the binding site is represented by sticks. The inflow of electronic density in blue and the outflow in red. Energy in kcal/mol.

We also analysed how biotin interacts with individual residues from the tail subpocket to elucidate their role in overall stability, Table 2, Figure S7. The results show that fragments interacting with biotin mainly through $N/O - H \cdots O$ classical hydrogen bonds, such as Gly49 and Ser88, are predominantly governed by electrostatic interactions. Collectively, these fragments provide a stabilization comparable to 29% of the mSA2 $\cdots$ biotin stability. Deformation density analysis additionally reveals charge transfer channels associated with the weaker $C-H \cdots O/S$ hydrogen bonds with Thr48 and Cys50-Cys86, Figure S7. The $O-H \cdots S$ type hydrogen bond between Thr90 and biotin is equally dominated by electrostatic and orbital components and compares to 5% of the global binding. The residues Trp79, Trp108, and Ala47 compare to 15% of the total binding energy, predominantly contributing through $C-H \cdots \pi$ contacts. An additional 6% of the equivalent of binding energy is derived from $C-H \cdots H-C$ interactions with Leu110. Both of these interaction types are primarily governed by London dispersion forces.

Table 2: ETS energy decomposition results describing the interaction of biotin (**fragment 1**) with individual residues of the tail subpocket. Within parentheses, the contribution to the overall biotin+H_2O $\cdots$ pocket stability is outlined. Energies in kcal/mol.

Interacting fragment 2	ΔE_{total}	ΔE_{elstat}	ΔE_{orb}	ΔE_{disp}	ΔE_{Pauli}
Gly49	-14.33 (16%)	-21.61	-14.12	-6.84	28.24
Ser88	-11.58 (13%)	-9.11	-5.34	-2.87	5.73
Cys50-Cys86	-8.16 (9%)	-5.18	-4.41	-7.30	8.73
Trp79	-7.44 (9%)	-4.71	-3.14	-10.54	10.95
Thr48	-6.62 (8%)	-3.74	-3.15	-5.13	5.40
Trp108	-5.47 (6%)	-2.99	-1.70	-7.38	6.59
Leu110	-5.07 (6%)	-2.61	-2.63	-5.33	5.50
Thr90	-3.96 (5%)	-5.66	-5.54	-4.23	11.48
Ala47	-1.03 (1%)	-1.80	-0.84	-3.10	4.71

Probing Hydrogen Bonding Effects via ^{1}H NMR Shifts

In this paragraph, we aim to elucidate the physical factors behind the chemical shifts observed in biotin carbon-bound protons upon protein binding. By following the approach proposed

by Scheiner,[70] we identify two primary components affecting the chemical shift of interest: 1) an environmental shielding effect and 2) deshielding due to hydrogen bond formation (details are given in the Methods section), Table 3.

The environmental shielding component accounts for the magnetic shielding effects arising from the surrounding molecular environment. These include ring current effects caused by aromatic residues, as well as positional shielding from nearby atoms or groups. Computationally, this effect was isolated by replacing the biotin protons with dummy atoms while retaining the protein binding pocket. The highest shielding values were observed for protons **h**, **b**, **h'**, and **a**, which explains the significant experimental CSPs noted in these regions, Table 3. The shielding effects extend beyond direct hydrogen bond lengths, demonstrating the long-range influence of aromatic rings, with observed values ≥ 0.3, even for protons significantly distant from the aromatic ring. These findings align with our empirical estimate of the ring current effect as predicted by the classical Pople model (Table S3).

The deshielding component arises from hydrogen bond formation between biotin protons and hydrogen bond acceptors within the protein binding pocket. This interaction introduces localized charge transfer and polarization effects, leading to reduced electron density around the protons and thus a deshielding effect. The analysis of various types of hydrogen bonds presented in Schreiner's work[70] suggests that the "deshielding due to H-bond" correlates with the bonding strength. The largest "deshielding effect due to the H-bond" is observed for protons **a** and **h'**, both of which are located directly above the pyrrole rings of tryptophan amino acids (Table 3, Figure 5). In the paper of Scheiner,[70] this position was identified as the most energetically favourable for the water $\cdots$ indole dimer. Moreover, the hydrogen bond character of protons **a**, **b**, and **d** is further enhanced by weak CH $\cdots$ O and CH $\cdots$ N bonds, as evidenced by the deformation density channels visualized in Figure 5. Interestingly, the lowest, but nonzero deshielding due to H-bond formation is identified for protons **f** and **e** that exhibit only homopolar CH$\cdots$HC interactions, corresponding charge-flow channels are presented in Figure 5.

Surprisingly, the proton **c'** does not follow the expected trend (positive sign) for "deshielding due to H-bond formation", Table 3. For this specific proton, the deshielding effect due to hydrogen bonding in the bound state is masked by a stronger deshielding effect in the unbound state. In the absence of a binding pocket, the carboxylate group of biotin carries a negative charge, causing it to polarize the proton **c'** leading to the formation of an intramolecular hydrogen bond in the free form. In the bound form, this interaction is relieved and the **c'** proton interaction is replaced with weaker H-bond acceptors in the binding pocket. This leads to an apparent shielding as reflected by the negative value (-0.27) for proton **c'** in Column 3 of Table 3. This phenomenon is observed in the chemical shift calculation of the free form because in our model the ligand is maintained in the same conformation as in the bound form, thus preventing the rotation of the biotin tail.

Table 3: Decomposition of the factors contributing to chemical shifts of individual biotin C-bounded protons along with the lists of their H-bond acceptors and hydrogen bond lengths. In the case of C-H$\cdots\pi$ interactions, the distance from the proton to the center of the ring is considered.

Proton label	Environmental shielding effect	Deshielding due to H-bond formation	Exp. CSP	H-bond acceptor	H-bond length (Å)
a	-1.93	0.64	-1.88	π(Trp79)	2.9
				O(Ser88)	3.0
a'	-0.93	0.61	-1.28	S(Cys50)	2.9
				H-C(Cys50)	2.2
				π(Trp79)	4.1
b	-2.32	0.52	-2.89	π(Trp79)	3.1
				N(Asn45)	2.9
b'	-0.82	0.49	-0.82	O(Gly49)	2.4
				N(Asn45)	2.8
				π(Trp79)	4.8
c	-1.42	0.46	-1.31	π(Trp79)	3.1
c'	-0.89	-0.27	-0.82	H-C(Leu110)	2.3
				O(biotin CO_2^-)	3.2
				π(Trp79)	4.6
d	-0.84	0.62	-1.17	π(Trp79)	3.9
				O(Asn45)	2.9
d'	-0.32	0.38	-0.49	O(Thr48)	2.7
				N(Asn45)	3.0
				π(Trp79)	5.3
e	-0.72	0.06	-0.98	π(Trp108)	4.9
				H-C(Leu110)	2.6
f	-0.31	0.12	-0.70	π(Trp108)	5.4
				H-C(Thr48)	2.5
g	-0.63	0.28	-0.94	π(Trp108)	3.6
h	-3.35	0.25	-3.17	π(Trp108)	2.7
h'	-1.99	0.66	-2.15	π(Trp108)	3.0

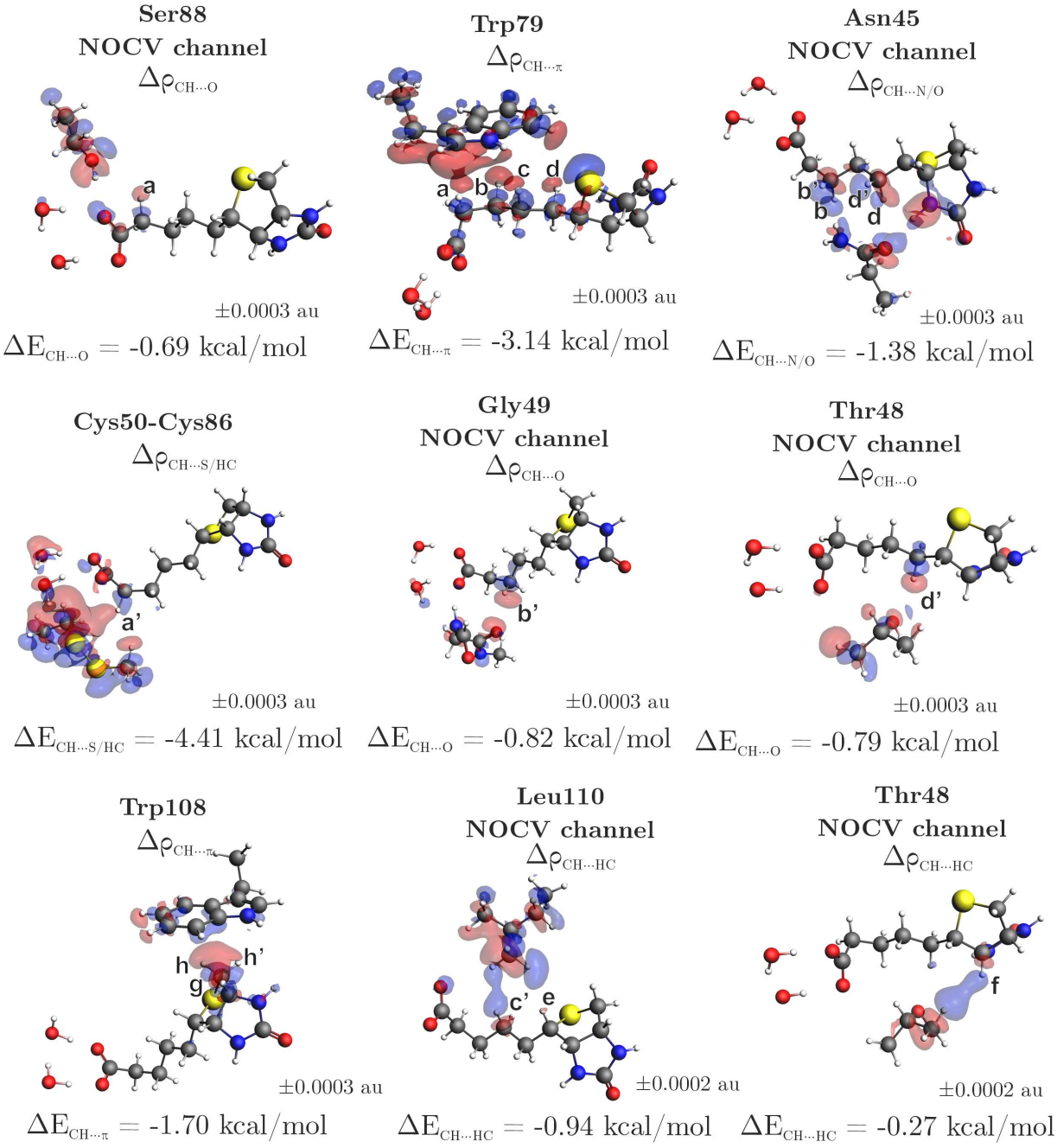


Figure 5: The contours of deformation density channels revealing the charge transfer of carbon-bounded biotin protons caused by hydrogen bonds formation. The inflow of electronic density is shown in blue and the outflow in red. Carbon-bounded protons of interest are labeled.

Conclusions

In this study, we aimed to provide a detailed understanding of the molecular interactions governing the binding of biotin to the monomeric streptavidin mutant (mSA2), a system known for its exceptionally strong non-covalent protein-ligand interaction. Through ^{1}H NMR chemical shift analysis and quantum chemical calculations, we uncovered important insights into the nature of these interactions.

Upon binding to mSA2, all carbon-bound protons in biotin shifted to significantly lower ppm values. Chemical shifts were accurately predicted using quantum chemical methods based on the X-ray structure. The strong agreement between the predicted and experimental shifts confirmed that the X-ray structure is a reliable representation of the streptavidin-biotin complex in solution, allowing for further computational studies of the binding mechanism.

The decomposition of proton chemical shifts in the bound form revealed that the strongest chemical shift perturbations are primarily due to the ring current effect from nearby aromatic residues. Moreover, the chemical shifts of all carbon-bound protons participating in weak interactions, like CH$\cdots$O, CH$\cdots$S, CH$\cdots\pi$, and even homopolar CH$\cdots$HC, are impacted by the contributions arising from charge transfer as well.

Energy decomposition analysis revealed that the biotin-mSA2 interaction is primarily driven by electrostatic forces, with additional contributions from orbital and dispersion interactions. These findings were corroborated by LED-DLPNO-CCSD(T) calculations, which showed strong agreement with the density functional theory results. Further NOCV analysis and fragmentation of the system allowed us to pinpoint weaker non-covalent interactions, such as CH$\cdots$S, CH$\cdots\pi$, and homopolar CH$\cdots$HC, alongside classical hydrogen bonds. Our analysis quantified that non-classical hydrogen bonds, including OH$\cdots$S, CH$\cdots$S, CH$\cdots\pi$, and CH$\cdots$HC, contribute approximately 44% to the overall stability of the complex. We demonstrated that these interactions are predominantly governed by London dispersion forces, with smaller contributions from electrostatics and orbital interactions. This contrasts with classical hydrogen bonds, which typically rely on electrostatic and orbital components.

These findings highlight the nuanced role of weak non-covalent interactions in stabilizing protein-ligand complexes. Such a detailed understanding of binding mechanisms and the nature of non-covalent interactions is crucial for drug discovery, as it enables the design of molecules that leverage these weak forces to achieve high-affinity binding.

Acknowledgement

The authors thank Mariusz P. Mitoraj and Moriz Mayer for their scientific input and critical revision of the manuscript. A. L. Ptaszek was funded by the Christian Doppler Laboratory for High-Content Structural Biology and Biotechnology, Austria. The financial support by the Austrian Federal Ministry of Labour and Economy, the National Foundation for Research, Technology and Development and the Christian Doppler Research Association is gratefully acknowledged. A. L. Ptaszek and Pedro A. Sánchez-Murcia thank the Medical University of Graz for computation time at the MedBioNode cluster. Parts of the calculations were carried out with the equipment purchased thanks to the financial support of the European Regional Development Fund in the framework of the Polish Innovation Economy Operational Program (Contract No. POIG.02.01.00-12-023/08). The authors gratefully acknowledge the Polish high-performance computing infrastructure PLGrid (HPC Centers: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2024/016976.

Supporting Information

Contents include details of mSA2 mutant similarity to WT streptavidin, NMR spectra and measured chemical shifts, LED/DLPNO-CCSD(T) results, DID interfragment dispersion visualisation, more detailed NOCV analysis and CSP predictions using the Pople model.

References

(1) Ziarek, J. J.; Baptista, D.; Wagner, G. Recent developments in solution nuclear magnetic resonance (NMR)-based molecular biology. *Journal of Molecular Medicine* **2018**, *96*, 1–8.

(2) Dingley, A. J.; Cordier, F.; Grzesiek, S. An introduction to hydrogen bond scalar couplings. *Concepts in Magnetic Resonance: An Educational Journal* **2001**, *13*, 103–127.

(3) Plevin, M. J.; Bryce, D. L.; Boisbouvier, J. Direct detection of CH/π interactions in proteins. *Nature chemistry* **2010**, *2*, 466–471.

(4) Hierso, J.-C. Indirect nonbonded nuclear spin–spin coupling: a guide for the recognition and understanding of "through-space" NMR J constants in small organic, organometallic, and coordination compounds. *Chemical reviews* **2014**, *114*, 4838–4867.

(5) Dračínskỳ, M.; Buchta, M.; Buděšínskỳ, M.; Vacek-Chocholoušová, J.; Stará, I. G.; Starỳ, I.; Malkina, O. L. Dihydrogen contacts observed by through-space indirect NMR coupling. *Chemical Science* **2018**, *9*, 7437–7446.

(6) Fukuda, H.; Tsurumaki, E.; Wakamatsu, K.; Toyota, S. Unusually short H··· H contacts in intramolecularly cyclized helically fused anthracenes. *Chemistry–A European Journal* **2024**, e202401627.

(7) Li, J.; Wang, Y.; An, L.; Chen, J.; Yao, L. Direct Observation of CH/CH van der Waals Interactions in Proteins by NMR. *Journal of the American Chemical Society* **2018**, *140*, 3194–3197.

(8) Dračínskỳ, M.; Jansa, P.; Bouř, P. Computational and Experimental Evidence of Through-Space NMR Spectroscopic J Coupling of Hydrogen Atoms. *Chemistry–A European Journal* **2012**, *18*, 981–986.

(9) Platzer, G.; Mayer, M.; Beier, A.; Brüschweiler, S.; Fuchs, J. E.; Engelhardt, H.; Geist, L.; Bader, G.; Schörghuber, J.; Lichtenecker, R.; others PI by NMR: probing CH–π interactions in protein–ligand complexes by NMR spectroscopy. *Angewandte Chemie* **2020**, *132*, 14971–14978.

(10) McConnell, D. B. Biotin's lessons in drug design. *Journal of medicinal chemistry* **2021**, *64*, 16319–16327.

(11) Weber, P. C.; Ohlendorf, D. H.; Wendoloski, J.; Salemme, F. Structural origins of high-affinity biotin binding to streptavidin. *Science* **1989**, *243*, 85–88.

(12) Grubmüller, H.; Heymann, B.; Tavan, P. Ligand binding: molecular mechanics calculation of the streptavidin-biotin rupture force. *Science* **1996**, *271*, 997–999.

(13) Yuan, C.; Chen, A.; Kolb, P.; Moy, V. T. Energy landscape of streptavidin- biotin complexes measured by atomic force microscopy. *Biochemistry* **2000**, *39*, 10219–10223.

(14) Ozawa, M.; Ozawa, T.; Nishio, M.; Ueda, K. The role of CH/π interactions in the high affinity binding of streptavidin and biotin. *Journal of Molecular Graphics and Modelling* **2017**, *75*, 117–124.

(15) Dundas, C. M.; Demonte, D.; Park, S. Streptavidin–biotin technology: improvements and innovations in chemical and biological applications. *Applied microbiology and biotechnology* **2013**, *97*, 9343–9353.

(16) Hyre, D. E.; Le Trong, I.; Merritt, E. A.; Eccleston, J. F.; Green, N. M.; Stenkamp, R. E.; Stayton, P. S. Cooperative hydrogen bond interactions in the streptavidin–biotin system. *Protein Science* **2006**, *15*, 459–467.

(17) Young, T.; Abel, R.; Kim, B.; Berne, B. J.; Friesner, R. A. Motifs for molecular recognition exploiting hydrophobic enclosure in protein–ligand binding. *Proceedings of the National Academy of Sciences* **2007**, *104*, 808–813.

(18) DeChancie, J.; Houk, K. The origins of femtomolar protein- ligand binding: hydrogen-bond cooperativity and desolvation energetics in the biotin-(strept) avidin binding site. *Journal of the American Chemical Society* **2007**, *129*, 5419–5429.

(19) Kryachko, E.; Scheiner, S. CH... F Hydrogen Bonds. Dimers of Fluoromethanes. *The Journal of Physical Chemistry A* **2004**, *108*, 2527–2535.

(20) Dalvit, C.; Invernizzi, C.; Vulpetti, A. Fluorine as a hydrogen-bond acceptor: Experimental evidence and computational calculations. *Chemistry–A European Journal* **2014**, *20*, 11058–11068.

(21) Alonso, M.; Woller, T.; Martín-Martínez, F. J.; Contreras-García, J.; Geerlings, P.; De Proft, F. Understanding the Fundamental Role of π/π, σ/σ, and σ/π Dispersion Interactions in Shaping Carbon-Based Materials. *Chemistry–A European Journal* **2014**, *20*, 4931–4941.

(22) Danovich, D.; Shaik, S.; Neese, F.; Echeverria, J.; Aullon, G.; Alvarez, S. Understanding the Nature of the CH··· HC Interactions in Alkanes. *Journal of Chemical Theory and Computation* **2013**, *9*, 1977–1991.

(23) Echeverría, J.; Aullón, G.; Danovich, D.; Shaik, S.; Alvarez, S. Dihydrogen contacts in alkanes are subtle but not faint. *Nature chemistry* **2011**, *3*, 323–330.

(24) de Freitas, R. F.; Schapira, M. A systematic analysis of atomic protein–ligand interactions in the PDB. *Medchemcomm* **2017**, *8*, 1970–1981.

(25) DeMonte, D.; Drake, E. J.; Lim, K. H.; Gulick, A. M.; Park, S. Structure-based engineering of streptavidin monomer with a reduced biotin dissociation rate. *Proteins: Structure, Function, and Bioinformatics* **2013**, *81*, 1621–1633.

(26) Demonte, D.; Dundas, C. M.; Park, S. Expression and purification of soluble monomeric

streptavidin in Escherichia coli. *Applied microbiology and biotechnology* **2014**, *98*, 6285–6295.

(27) Mandal, N.; Pratik, S. M.; Datta, A. Exploring Ultrashort Hydrogen–Hydrogen Nonbonded Contacts in Constrained Molecular Cavities. *The Journal of Physical Chemistry B* **2017**, *121*, 825–834.

(28) Liptrot, D. J.; Power, P. P. London dispersion forces in sterically crowded inorganic and organometallic molecules. *Nature Reviews Chemistry* **2017**, *1*, 1–12.

(29) Cukrowski, I.; Govender, K. K.; Mitoraj, M. P.; Srebro, M. QTAIM and ETS-NOCV Analyses of Intramolecular CH···HC Interactions in Metal Complexes. *The Journal of Physical Chemistry A* **2011**, *115*, 12746–12757.

(30) Wolstenholme, D. J.; Flogeras, J.; Che, F. N.; Decken, A.; McGrady, G. S. Homopolar Dihydrogen Bonding in Alkali Metal Amidoboranes: Crystal Engineering of Low-Dimensional Molecular Materials. *Journal of the American Chemical Society* **2013**, *135*, 2439–2442.

(31) Wolstenholme, D. J.; Dobson, J. L.; McGrady, G. S. Homopolar dihydrogen bonding in main group hydrides: discovery, consequences, and applications. *Dalton Transactions* **2015**, *44*, 9718–9731.

(32) Poater, J.; Solà, M.; Bickelhaupt, F. M. Hydrogen–Hydrogen Bonding in Planar Biphenyl, Predicted by Atoms-In-Molecules Theory, Does Not Exist. *Chemistry – A European Journal* **2006**, *12*, 2889–2895.

(33) Weinhold, F.; Schleyer, P. v. R.; McKee, W. C. Bay-type H···H "bonding" in cis-2-butene and related species: QTAIM versus NBO description. *Journal of Computational Chemistry* **2014**, *35*, 1499–1508.

(34) Dillen, J. Congested molecules. Where is the steric repulsion? An analysis of the electron density by the method of interacting quantum atoms. *International Journal of Quantum Chemistry* **2013**, *113*, 2143–2153.

(35) Echeverría, J.; Aullón, G.; Alvarez, S. Dihydrogen intermolecular contacts in group 13 compounds: H$\cdots$H or E$\cdots$H (E = B, Al, Ga) interactions? *Dalton Trans.* **2017**, *46*, 2844–2854.

(36) Belkova, N. V.; Epstein, L. M.; Filippov, O. A.; Shubina, E. S. Hydrogen and Dihydrogen Bonds in the Reactions of Metal Hydrides. *Chemical Reviews* **2016**, *116*, 8545–8587.

(37) Eskandari, K.; Van Alsenoy, C. Hydrogen–hydrogen interaction in planar biphenyl: A theoretical study based on the interacting quantum atoms and Hirshfeld atomic energy partitioning methods. *Journal of Computational Chemistry* **2014**, *35*, 1883–1889.

(38) Grabowski, S. J.; Sokalski, W. A.; Leszczynski, J. Nature of X-H$^{+\delta}\cdots{}^{-\delta}$H-Y Dihydrogen Bonds and X-H$\cdots\sigma$ Interactions. *The Journal of Physical Chemistry A* **2004**, *108*, 5823–5830.

(39) Pinter, B.; Fievez, T.; Bickelhaupt, F. M.; Geerlings, P.; De Proft, F. On the origin of the steric effect. *Phys. Chem. Chem. Phys.* **2012**, *14*, 9846–9854.

(40) Bates, T. G.; de Lange, J. H.; Cukrowski, I. The CH$\cdots$HC interaction in biphenyl is a delocalized, molecular-wide and entirely non-classical interaction: Results from FALDI analysis. *Journal of Computational Chemistry* **2021**, *42*, 706–718.

(41) Mitoraj, M. P.; Sagan, F.; Szczepanik, D. W.; de Lange, J. H.; Ptaszek, A. L.; van Niekerk, D. M.; Cukrowski, I. Origin of hydrocarbons stability from a computational perspective: a case study of ortho-xylene isomers. *ChemPhysChem* **2020**, *21*, 494–502.

(42) Platzer, G.; Ptaszek, A. L.; Böttcher, J.; Fuchs, J. E.; Geist, L.; Braun, D.; McConnell, D. B.; Konrat, R.; Sánchez-Murcia, P. A.; Mayer, M. Ligand 1H NMR Chemical Shifts as Accurate Reporters for Protein-Ligand Binding Interfaces in Solution. *ChemPhysChem* **2024**, *25*, e202300636.

(43) Mitoraj, M. P.; Michalak, A.; Ziegler, T. A combined charge and energy decomposition scheme for bond analysis. *Journal of chemical theory and computation* **2009**, *5*, 962–975.

(44) Neese, F.; Wennmohs, F.; Hansen, A. Efficient and accurate local approximations to coupled-electron pair approaches: An attempt to revive the pair natural orbital method. *The Journal of Chemical Physics 130*, 114108.

(45) Neese, F.; Hansen, A.; Liakos, D. G. Efficient and accurate approximations to the local coupled cluster singles doubles method using a truncated pair natural orbital basis. *The Journal of Chemical Physics 131*, 064103.

(46) Riplinger, C.; Sandhoefer, B.; Hansen, A.; Neese, F. Natural triple excitations in local coupled cluster calculations with pair natural orbitals. *The Journal of Chemical Physics 139*, 134101.

(47) Guo, Y.; Riplinger, C.; Becker, U.; Liakos, D. G.; Minenkov, Y.; Cavallo, L.; Neese, F. Communication: An improved linear scaling perturbative triples correction for the domain based local pair-natural orbital based singles and doubles coupled cluster method [DLPNO-CCSD(T)]. *The Journal of Chemical Physics 148*, 011101.

(48) Saebo, S.; Pulay, P. Local Treatment of Electron Correlation. *Annual Review of Physical Chemistry 44*, 213–236.

(49) Altun, A.; Saitow, M.; Neese, F.; Bistoni, G. Local Energy Decomposition of Open-Shell Molecular Systems in the Domain-Based Local Pair Natural Orbital Coupled Cluster Framework. *Journal of Chemical Theory and Computation 15*, 1616–1632.

(50) Altun, A.; Izsák, R.; Bistoni, G. Local energy decomposition of coupled-cluster interaction energies: Interpretation, benchmarks, and comparison with symmetry-adapted perturbation theory. *International Journal of Quantum Chemistry 121*, e26339.

(51) Schneider, W. B.; Bistoni, G.; Sparta, M.; Saitow, M.; Riplinger, C.; Auer, A. A.; Neese, F. Decomposition of Intermolecular Interaction Energies within the Local Pair Natural Orbital Coupled Cluster Framework. *Journal of Chemical Theory and Computation 12*, 4778–4792.

(52) Anandakrishnan, R.; Aguilar, B.; Onufriev, A. V. H++ 3.0: automating p K prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic acids research* **2012**, *40*, W537–W541.

(53) `https://proteins.plus`.

(54) Schöning-Stierand, K.; Diedrich, K.; Ehrt, C.; Flachsenberg, F.; Graef, J.; Sieg, J.; Penner, P.; Poppinga, M.; Ungethüm, A.; Rarey, M. Proteins Plus: A comprehensive collection of web-based molecular modeling tools. *Nucleic Acids Research* **2022**, *50*, W611–W615.

(55) Fährrolfes, R.; Bietz, S.; Flachsenberg, F.; Meyder, A.; Nittinger, E.; Otto, T.; Volkamer, A.; Rarey, M. Proteins Plus: a web portal for structure analysis of macromolecules. *Nucleic acids research* **2017**, *45*, W337–W343.

(56) Schöning-Stierand, K.; Diedrich, K.; Fährrolfes, R.; Flachsenberg, F.; Meyder, A.; Nittinger, E.; Steinegger, R.; Rarey, M. Proteins Plus: interactive analysis of protein–ligand binding interfaces. *Nucleic acids research* **2020**, *48*, W48–W53.

(57) Frisch, M. J. et al. Gaussian˜16 Revision C.01. 2016; Gaussian Inc. Wallingford CT.

(58) Tian, C.; Kasavajhala, K.; Belfon, K. A.; Raguette, L.; Huang, H.; Migues, A. N.; Bickel, J.; Wang, Y.; Pincay, J.; Wu, Q.; others ff19SB: amino-acid-specific protein

backbone parameters trained against quantum mechanics energy surfaces in solution. *Journal of chemical theory and computation* **2019**, *16*, 528–552.

(59) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *Journal of computational chemistry* **2004**, *25*, 1157–1174.

(60) Mark, P.; Nilsson, L. Structure and dynamics of the TIP3P, SPC, and SPC/E water models at 298 K. *The Journal of Physical Chemistry A* **2001**, *105*, 9954–9960.

(61) Case, D. A.; Aktulga, H. M.; Belfon, K.; Ben-Shalom, I.; Brozell, S. R.; Cerutti, D. S.; Cheatham III, T. E.; Cruzeiro, V. W. D.; Darden, T. A.; Duke, R. E.; others *Amber 2021*; University of California, San Francisco, 2021.

(62) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An N · log N method for Ewald sums in large systems. *The Journal of chemical physics* **1993**, *98*, 10089–10092.

(63) Miyamoto, S.; Kollman, P. A. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *Journal of computational chemistry* **1992**, *13*, 952–962.

(64) London, F. Théorie quantique des courants interatomiques dans les combinaisons aromatiques. *J. phys. radium* **1937**, *8*, 397–409.

(65) McWeeny, R. Perturbation theory for the Fock-Dirac density matrix. *Physical Review* **1962**, *126*, 1028.

(66) Cheeseman, J. R.; Trucks, G. W.; Keith, T. A.; Frisch, M. J. A comparison of models for calculating nuclear magnetic resonance shielding tensors. *The Journal of chemical physics* **1996**, *104*, 5497–5509.

(67) Wolinski, K.; Hinton, J. F.; Pulay, P. Efficient implementation of the gauge-independent

atomic orbital method for NMR chemical shift calculations. *Journal of the American Chemical Society* **1990**, *112*, 8251–8260.

(68) Ditchfield, R. Self-consistent perturbation theory of diamagnetism: I. A gauge-invariant LCAO method for NMR chemical shifts. *Molecular Physics* **1974**, *27*, 789–807.

(69) Chai, J.-D.; Head-Gordon, M. Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections. *Physical Chemistry Chemical Physics* **2008**, *10*, 6615–6620.

(70) Scheiner, S. Assessment of the presence and strength of H-bonds by means of corrected NMR. *Molecules* **2016**, *21*, 1426.

(71) Ziegler, T.; Rauk, A. On the calculation of bonding energies by the Hartree Fock Slater method: I. The transition state method. *Theoretica chimica acta* **1977**, *46*, 1–10.

(72) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *Journal of computational chemistry* **2011**, *32*, 1456–1465.

(73) Mitoraj, M.; Michalak, A. Natural orbitals for chemical valence as descriptors of chemical bonding in transition metal complexes. *Journal of molecular modeling* **2007**, *13*, 347–355.

(74) Te Velde, G. t.; Bickelhaupt, F. M.; Baerends, E. J.; Fonseca Guerra, C.; van Gisbergen, S. J.; Snijders, J. G.; Ziegler, T. Chemistry with ADF. *Journal of Computational Chemistry* **2001**, *22*, 931–967.

(75) ADF 2024.1, SCM, Theoretical Chemistry, Vrije Universiteit, Amsterdam, The Netherlands, `http://www.scm.com`.

(76) Stasyuk, O. A.; Sedlak, R.; Guerra, C. F.; Hobza, P. Comparison of the DFT-SAPT and

canonical EDA schemes for the energy decomposition of various types of noncovalent interactions. *Journal of Chemical Theory and Computation* **2018**, *14*, 3440–3450.

(77) Gao, W.; Feng, H.; Xuan, X.; Chen, L. The assessment and application of an approach to noncovalent interactions: the energy decomposition analysis (EDA) in combination with DFT of revised dispersion correction (DFT-D3) with Slater-type orbital (STO) basis set. *Journal of molecular modeling* **2012**, *18*, 4577–4589.

(78) Neese, F.; Wennmohs, F.; Becker, U.; Riplinger, C. The ORCA quantum chemistry program package. *The Journal of Chemical Physics 152*, 224108.

(79) Weigend, F. Accurate Coulomb-fitting basis sets for H to Rn. *Physical Chemistry Chemical Physics 8*, 1057–1065.

(80) Weigend, F.; Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Physical Chemistry Chemical Physics 7*, 3297–3305.

(81) Helmich-Paris, B.; de Souza, B.; Neese, F.; Izsák, R. An improved chain of spheres for exchange algorithm. *The Journal of Chemical Physics 155*, 104109.

(82) Sano, T.; Cantor, C. R. Intersubunit contacts made by tryptophan 120 with biotin are essential for both strong biotin binding and biotin-induced tighter subunit association of streptavidin. *Proceedings of the National Academy of Sciences* **1995**, *92*, 3180–3184.

(83) Green, N. M. *Methods in enzymology*; Elsevier, 1990; Vol. 184; pp 51–67.

(84) Lim, K. H.; Huang, H.; Pralle, A.; Park, S. Stable, high-affinity streptavidin monomer for protein labeling and monovalent biotin detection. *Biotechnology and bioengineering* **2013**, *110*, 57–67.

(85) Panigrahi, S. K.; Desiraju, G. R. Strong and weak hydrogen bonds in the protein–ligand interface. *Proteins: Structure, Function, and Bioinformatics* **2007**, *67*, 128–141.

(86) Mitoraj, M. P.; Kurczab, R.; Boczar, M.; Michalak, A. Theoretical description of hydrogen bonding in oxalic acid dimer and trimer based on the combined extended-transition-state energy decomposition analysis and natural orbitals for chemical valence (ETS-NOCV). *Journal of molecular modeling* **2010**, *16*, 1789–1795.

CSs chemical shifts

CSP chemical shift perturbations

NMR Nuclear Magnetic Resonance

WT wild-type

Supporting Information: From Weak Interactions to Strong Affinity: Deciphering the Biotin-Streptavidin Interaction through NMR and Computational Analysis

Aleksandra L. Ptaszek,[†,‡,¶,#] Sarah Kratzwald,[§,¶,#] Filip Sagan,[‖] Mario Migotti,[¶] Pedro A. Sánchez-Murcia,[‡,⊥] Robert Konrat,[†,¶] and Gerald Platzer[*,§,†]

†*Christian Doppler Laboratory for High-Content Structural Biology and Biotechnology, Department of Structural and Computational Biology, Max Perutz Labs, University of Vienna, Campus Vienna Biocenter 5, 1030-Vienna, Austria*

‡*Laboratory of Computer-Aided Molecular Design, Division of Medicinal Chemistry, Otto-Loewi Research Center, Medical University of Graz, Neue Stiftintalstr. 6/III, A-8010 Graz, Austria*

¶*Department of Structural and Computational Biology, Max Perutz Labs, University of Vienna, Campus Vienna Biocenter 5, 1030-Vienna, Austria*

§*MAG-LAB GmbH, Karl-Farkas-Gasse 22, 1030- Vienna, Austria*

‖*Faculty of Chemistry, Jagiellonian University, Gronostajowa 2, 30-387 Cracow, Poland*

⊥*BioTechMed-Graz, Mozartgasse 12/II, A-8010 Graz, Austria*

#*Equal contribution*

E-mail: gerald.platzer@mag-lab.eu

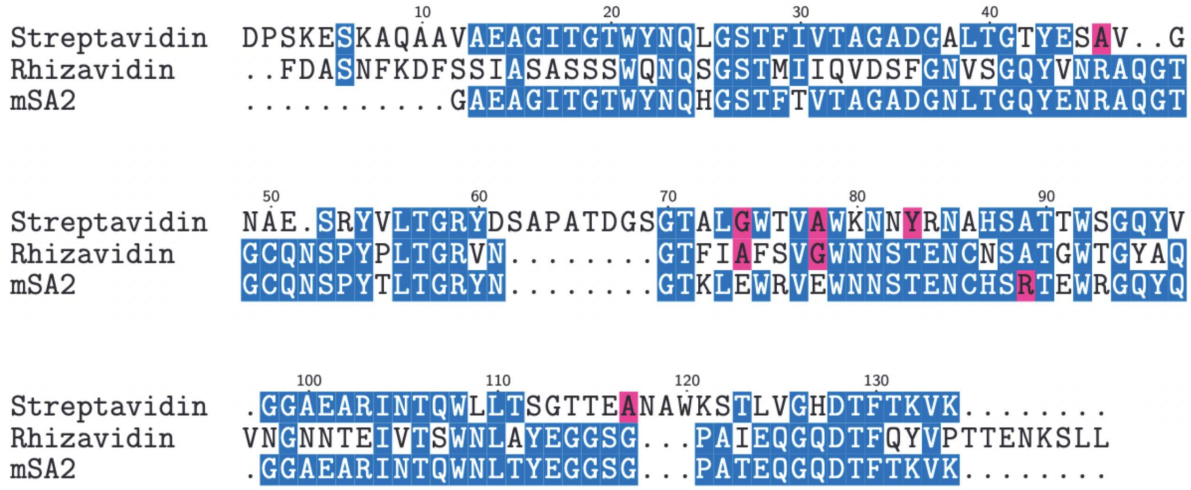


Figure S1: Sequence similarities of WT streptavidin, WT rhizavidin and mSA2.[S1]

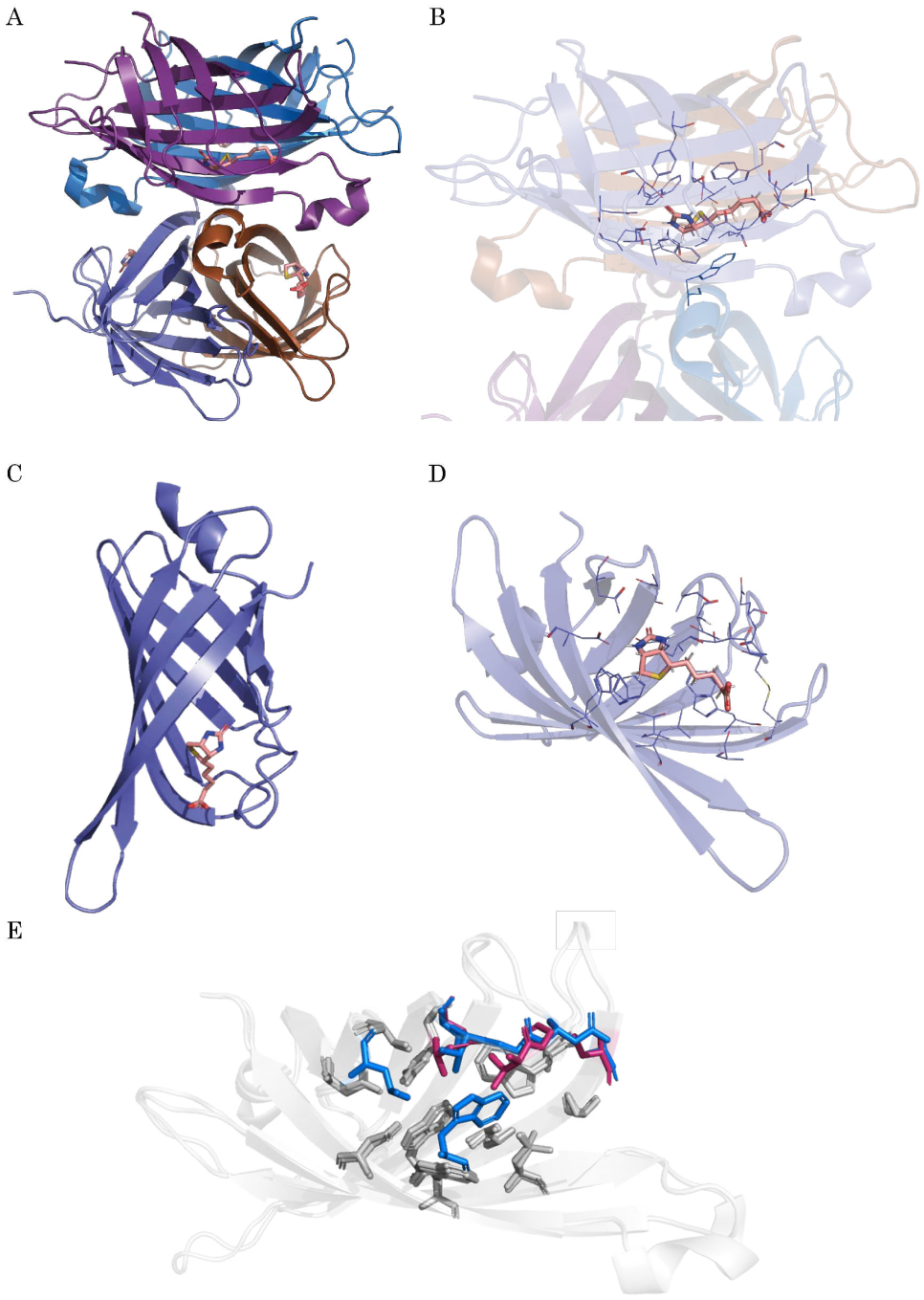


Figure S2: (A) The crystal structure of WT streptavidin is shown as a homotetrameric protein with (B) the binding pocket of biotin bound to the WT streptavidin. (PDB ID: 3RY2) (C) The crystal structure of the mSA2 mutant is shown, with (D) the binding pocket of biotin. (PDB ID: 4JNJ) (E) The aligned structures of the WT and mutant are illustrated, where the differences in amino acids in the biotin-binding pocket are highlighted in blue for the WT and magenta for the mSA2 mutant.

HSQC

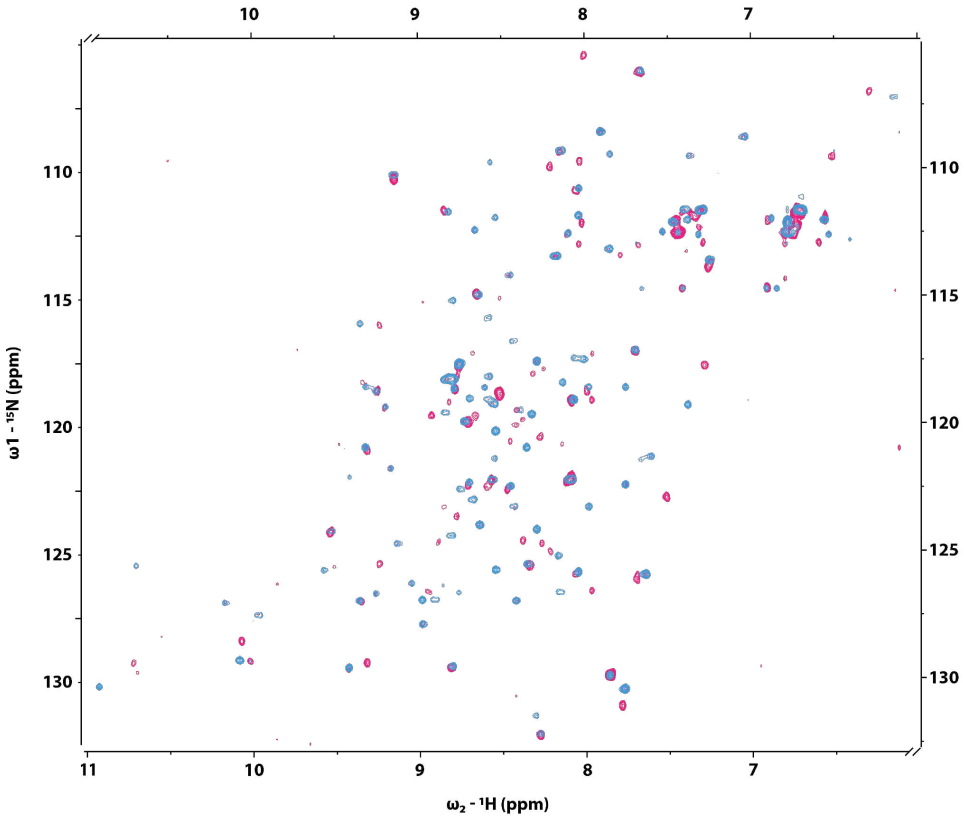


Figure S3: ^{1}H-^{15}N HSQC of mSA2-apo (pink) and mSA2-biotin (blue).

NOESY

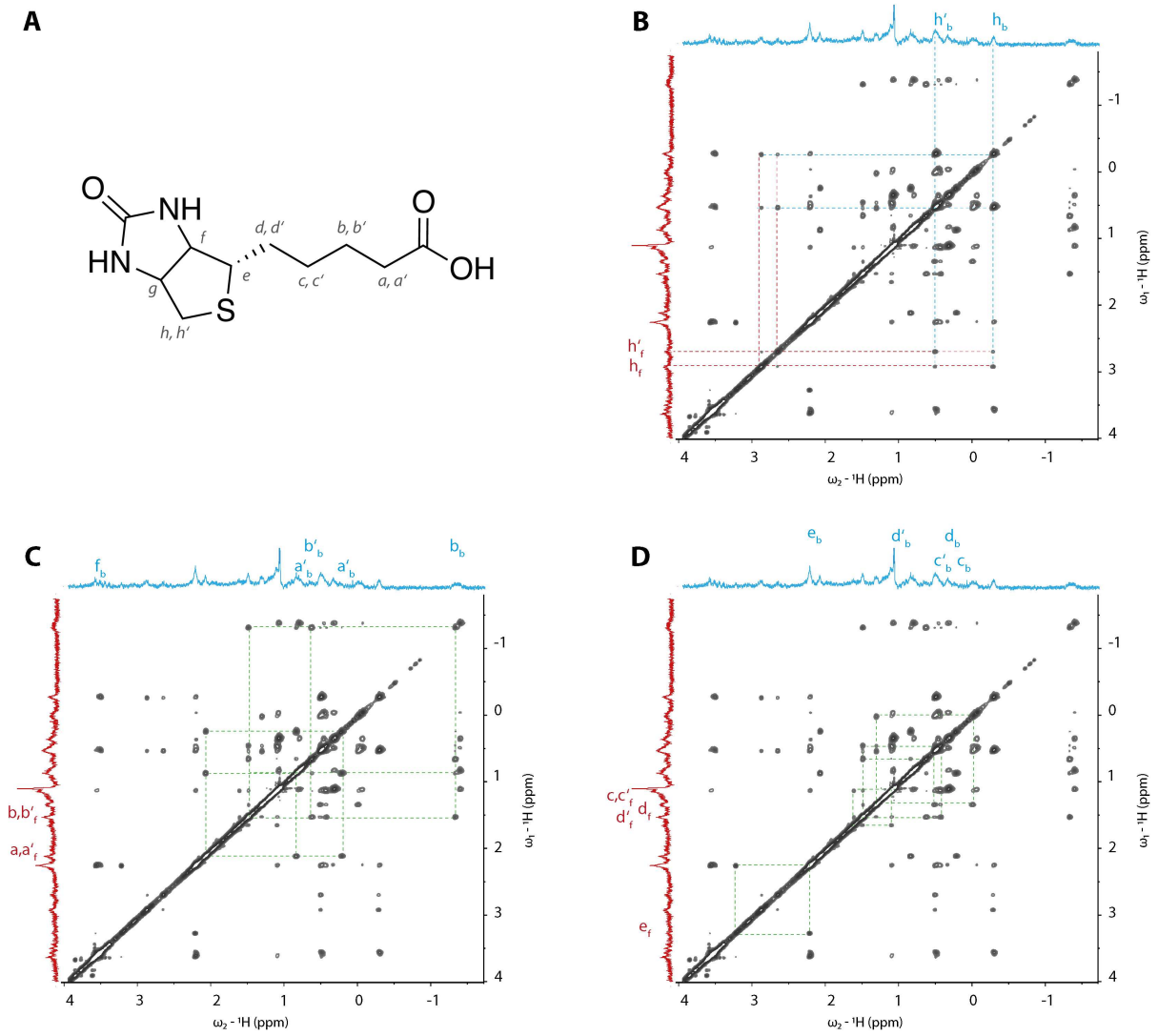


Figure S4: ^{1}H-^{1}H NOESY spectra of biotin in complex with mSA2. Spectra were recorded as mSA2-biotin 1:1 complex, due to partial precipitation of mSA2, biotin is in 1.5-fold excess. (A) Chemical structure of biotin: protons are numbered with letters a - h'. (B) Scheme for the assignment strategy of NOE cross peaks showing exchange peaks between the free signals (red) and the bound signals (blue), NOE between two bound signals, and NOE between free to relay bound signals. (C and D) CSPs between free and bound signals are shown in dotted green lines.

Table S1: Experimentally determined CS and CSP for biotin's carbon-bound H-atoms.

Proton	δ Biotin (ppm)	$CSP_{Experimental}$ (ppm)	δ Biotin:mSA2 (ppm)
a	2.13	-1.88	0.25
a´	2.13	-1.28	0.85
b	1.49	-2.89	-1.41
b´	1.49	-0.82	0.67
c	1.30	-1.31	-0.01
c´	1.30	-0.82	0.48
d	1.52	-1.17	0.35
d´	1.61	-0.49	1.12
e	3.25	-0.98	2.27
f	4.35	-0.70	3.65
g	4.50	-0.94	3.56
h	2.89	-3.17	-0.28
h´	2.66	-2.15	0.51

Table S2: LED/DLPNO-CCSD(T) energy decomposition describing the interaction between the fragments analogous to ones described in main manuscript: biotin together with nearby water molecules (fragment 1 in all fragmentations) and the binding pocket (fragment 2); biotin with waters and the ureido subpocket (Asn23, Ser27, Tyr43, Asn45, Trp92, Asp12); biotin with waters and the tail subpocket (Ala47, Thr48, Gly49, Cys50, Cys86, Trp79, Ser88, Thr90, Trp108, Leu110). Energies listed in kcal/mol.

Energy term	**Biotin+H_2O $\cdots$ pocket**	**Biotin+H_2O $\cdots$ ureido subpocket**	**Biotin+H_2O $\cdots$ tail subpocket**
ΔE_{int}	-83.23	-22.18	-58.71
ΔE^{HF}_{elst}	-348.78	-170.09	-182.48
ΔE^{HF}_{exch}	-71.28	-31.95	-39.95
ΔE_{disp}	-62.53	-20.14	-42.82
$\Delta E^{C}_{CT1\rightarrow2}$	-41.42	-16.74	-24.71
$\Delta E^{C}_{CT2\rightarrow1}$	-31.85	-16.30	-14.83
ΔE^{C}_{rest}	-38.13	-12.62	-21.76
$\sum \Delta E_{el-prep}$	510.76	245.67	268.85

LED analysis

DLPNO-CCSD(T)-based Local Energy Decomposition is a powerful tool that allows for an unprecedented level of detail in describing London dispersion.[S2–S9] Although not strictly additive, the contributions from biotin-ureido subpocket and biotin-tail subpocket models sum

up almost perfectly to their counterparts in the whole model, allowing us to draw conclusions on their respective magnitudes, Table S2. Immediately noticeable is the greater stabilization stemming from the tail part of binding site. Almost all energy terms are more stabilizing for the tail part, owing to greater number of contacts in this region, showing clearly, that many weaker interactions can dominate over the regular hydrogen bonds. Particularly, the dispersion energy is twice as high as for the biotin-ureido subpocket, followed by significantly higher exchange energy term, jointly indicating non-classical nature of the said stabilization. Another interesting difference between the fragmentations is the prevalence of biotin→binding site correlation charge-transfer term (which can be interpreted as the instantaneous ion pair formation), which is consistent with the direction of typical charge-transfer.[S9] It may be due to the negative charge on biotin, which probably increases charge-transfer in this direction.

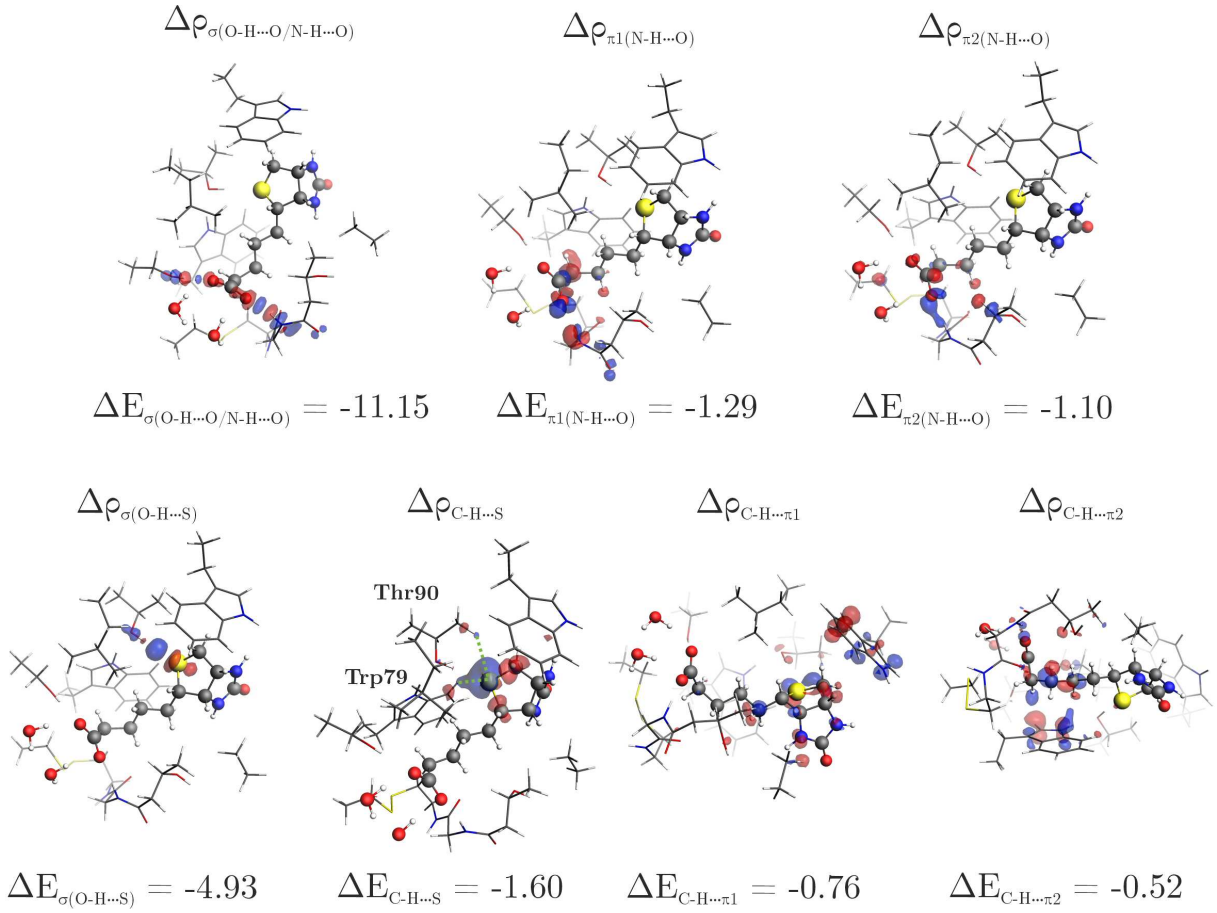


Figure S5: Selected NOCV deformation density channels along with the energy contributions of biotin molecule and two water molecules (fragment 1) interacting with the protein tail subpocket (fragment 2). The biotin and water molecules (fragment 1) are represented by balls and sticks, and the residues from the binding site (fragment 2) are represented by sticks.

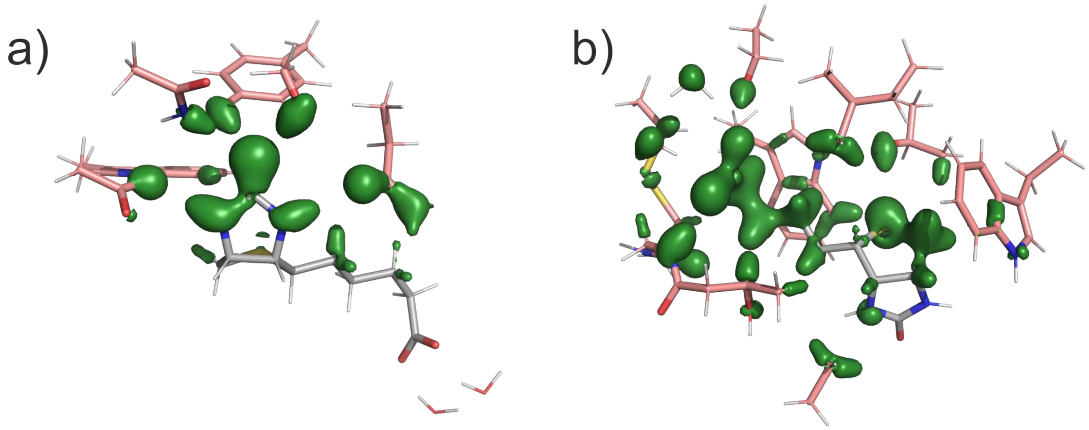


Figure S6: DID interfragment dispersion visualization of biotin and water molecules interaction with a) the ureido subpocket and b) the tail subpocket. The biotin and water molecules (fragment 1) are coloured gray, and the residues from the binding site (fragment 2) are coloured pink. LED-DLPNO-CCSD(T)/def2-tzvp.

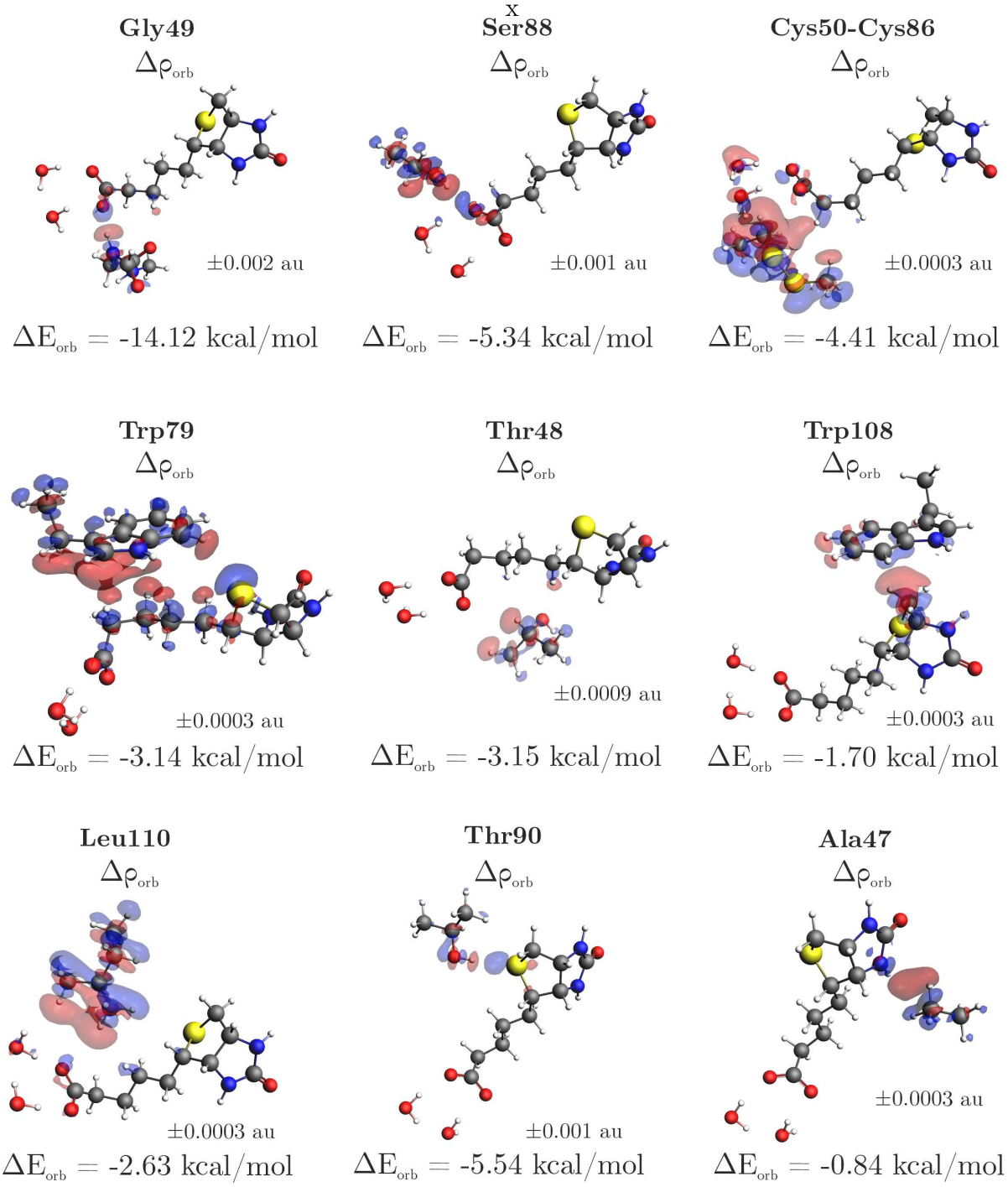


Figure S7: The overall deformation density $\Delta\rho_{orb}$ with the corresponding ΔE_{orb} of biotin molecule and two water molecules (fragment 1) interacting with individual residues from the tail subpocket (fragment 2).

Pople's Model

In Pople's equation, the center of the aromatic ring induces a magnetic field according to Equation 1:[S10]

$$\Delta\sigma = 10^6 \times \frac{ne^2a^2}{4\pi mc^2} \times \frac{3cos^2\theta - 1}{r^3} \quad (1)$$

In Equation 1, $\Delta\sigma$ represents the change in the chemical shift in ppm, n equals the number of π-electrons (6 for each of the aromatic rings of Trp), e equals the elementary charge in Franklin (1 e = 4.803× 10^{10} Franklin), a equals the radius of the aromatic rings in Trp (6-ring: 1.39×10^{-8} cm; 5-ring: 1.22×10^{-8} cm), m equals the electron mass in gram (9.109×10^{-28} g), c is the speed of light (2.998×10^{10} m×s^{-1}), θ represents the angle between the normal vector of the plane of one of the aromatic rings of Trp and the proton to ring center vector in rad, and r represents the distance from the proton to either the 5-ring or 6-ring center of the aromatic ring. Values (r, θ and a) were extracted from the crystal structure (PDB ID: 4JNJ).

Table S3: CSP prediction from the Pople model.

Proton	$CSP_{Experimental}$ (ppm)	CSP_{Pople} (ppm)
a	-1.88	-2.40
a´	-1.28	-0.84
b	-2.89	-2.12
b´	-0.82	-0.70
c	-1.31	-0.86
c´	-0.82	-0.48
d	-1.17	-0.51
d´	-0.49	-0.40
e	-0.98	-0.56
f	-0.70	-0.27
g	-0.94	-0.64
h	-3.17	-3.94
h´	-2.15	-1.86

References

(S1) DeMonte, D.; Drake, E. J.; Lim, K. H.; Gulick, A. M.; Park, S. Structure-based engineering of streptavidin monomer with a reduced biotin dissociation rate. *Proteins: Structure, Function, and Bioinformatics* **2013**, *81*, 1621–1633.

(S2) Neese, F.; Wennmohs, F.; Hansen, A. Efficient and accurate local approximations to coupled-electron pair approaches: An attempt to revive the pair natural orbital method. *The Journal of Chemical Physics 130*, 114108.

(S3) Neese, F.; Hansen, A.; Liakos, D. G. Efficient and accurate approximations to the local coupled cluster singles doubles method using a truncated pair natural orbital basis. *The Journal of Chemical Physics 131*, 064103.

(S4) Riplinger, C.; Sandhoefer, B.; Hansen, A.; Neese, F. Natural triple excitations in local coupled cluster calculations with pair natural orbitals. *The Journal of Chemical Physics 139*, 134101.

(S5) Guo, Y.; Riplinger, C.; Becker, U.; Liakos, D. G.; Minenkov, Y.; Cavallo, L.; Neese, F. Communication: An improved linear scaling perturbative triples correction for the domain based local pair-natural orbital based singles and doubles coupled cluster method [DLPNO-CCSD(T)]. *The Journal of Chemical Physics 148*, 011101.

(S6) Saebo, S.; Pulay, P. Local Treatment of Electron Correlation. *Annual Review of Physical Chemistry 44*, 213–236.

(S7) Altun, A.; Saitow, M.; Neese, F.; Bistoni, G. Local Energy Decomposition of Open-Shell Molecular Systems in the Domain-Based Local Pair Natural Orbital Coupled Cluster Framework. *Journal of Chemical Theory and Computation 15*, 1616–1632.

(S8) Altun, A.; Izsák, R.; Bistoni, G. Local energy decomposition of coupled-cluster inter-

action energies: Interpretation, benchmarks, and comparison with symmetry-adapted perturbation theory. *International Journal of Quantum Chemistry 121*, e26339.

(S9) Schneider, W. B.; Bistoni, G.; Sparta, M.; Saitow, M.; Riplinger, C.; Auer, A. A.; Neese, F. Decomposition of Intermolecular Interaction Energies within the Local Pair Natural Orbital Coupled Cluster Framework. *Journal of Chemical Theory and Computation 12*, 4778–4792.

(S10) Pople, J. Proton magnetic resonance of hydrocarbons. *The Journal of Chemical Physics* **1956**, *24*, 1111–1111.

5

UCBSHIFT 2.0: SIDE CHAIN PROTEIN CHEMICAL SHIFT PREDICTION

Reprinted with permission from
A. L. Ptaszek, J. Li, R. Konrat, G. Platzer, and T. Head-Gordon, "Ucbshift 2.0: Bridging the gap from backbone to side chain protein chemical shift prediction for protein structures," *Journal of the American Chemical Society*, vol. 146, no. 46, pp. 31 733–31 745, 2024 [108]

DOI: 10.1021/jacs.4c10474

PERSONAL CONTRIBUTION

Implementation, data analysis, writing, visualization, conceptualization, shared First Author with J. Li

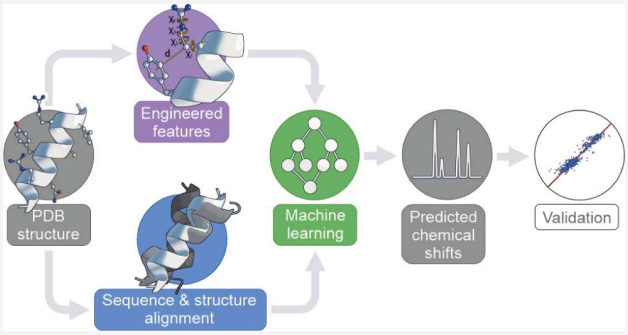
JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY
pubs.acs.org/JACS
Article

UCBShift 2.0: Bridging the Gap from Backbone to Side Chain Protein Chemical Shift Prediction for Protein Structures

Aleksandra L. Ptaszek,[#] Jie Li,[#] Robert Konrat, Gerald Platzer, and Teresa Head-Gordon*

Cite This: J. Am. Chem. Soc. 2024, 146, 31733−31745
Read Online
ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Chemical shifts are a readily obtainable NMR observable that can be measured with high accuracy, and because they are sensitive to conformational averages and the local molecular environment, they yield detailed information about protein structure in solution. To predict chemical shifts of protein structures, we introduced the UCBShift method that uniquely fuses a transfer prediction module, which employs sequence and structure alignments to select reference chemical shifts from an experimental database, with a machine learning model that uses carefully curated and physics-inspired features derived from X-ray crystal structures to predict backbone chemical shifts for proteins. In this work, we extend the UCBShift 1.0 method to side chain chemical shift prediction to perform whole protein analysis, which, when validated against well-defined test data shows higher accuracy and better reliability compared to the popular SHIFTX2 method. With the greater abundance of cleaned protein shift-structure data and the modularity of the general UCBShift algorithms, users can gain insight into different features important for residue-specific stabilizing interactions for protein backbone and side chain chemical shift prediction. We suggest several backward and forward applications of UCBShift 2.0 that can help validate AlphaFold structures and probe protein dynamics.

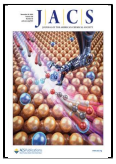


INTRODUCTION

Nuclear magnetic resonance (NMR) spectroscopy is a primary experimental tool for characterizing dynamics and the solution structure of biomolecules. NMR chemical shifts for organic systems containing ^{1}H, ^{13}C and ^{15}N nuclei can provide detailed descriptions of the structure of drug molecules,[1,2] proteins and their complexes,[3−5] and disordered protein states.[6−8] NMR chemical shifts in particular are sensitive not only to changes in local structure, but particularly to conformational changes that depend on sequence context and peptide length, solvent exposure or protein hydrogen-bonding environments, and even vibrational averaging. While quantum chemical calculations of magnetic properties are often powerful, the computational requirements associated with high level quantum mechanical (QM) chemical shift predictions are far too demanding for "on the fly" evaluation,[9] although QM calculated chemical shifts have been used in machine learning training for chemical shift predictions.[1,2,10,11]

Therefore, it is common practice to develop a database-trained expert system that can produce "predicted" chemical shifts by eliminating QM calculations altogether and instead predict the experimental observable directly. These heuristic calculators have been primarily focused on backbone chemical shifts, and must be trained to account for how NMR observables depend not only on backbone dihedral angles, but other features such as bond angles, deviations from planarity of the peptide bond, and hydrogen-bonding with the surrounding protein or solvent. Meiler and coworkers[12] and later Shen and Bax[13] showed that Artificial Neural Networks (ANNs) are well suited to utilize such features for protein backbone chemical shift prediction. The single-layer feed-forward network developed and packaged as SPARTA+[13] is a popular such example, with other feature-based methods including SHIFTCALC,[14] SHIFTX,[15] PROSHIFT,[12] CamShift,[16] and PPM[17,18] also showing comparable quality predictions. It is also worth distinguishing the SHIFTX+ component of SHIFTX2[19] as it uses not only backbone geometric features but also residue biological similarity properties like block substitution matrix (BLOSUM) numbers to predict chemical shifts using either ANNs or Bagging and Boosting ensemble models. Even so, these models still suffer from inaccuracies—presumably because the features drawn from X-ray crystal structures are incomplete or unrepresentative- such that a pragmatic solution is to substitute known experimental chemical shift values of another protein

Received: July 31, 2024
Revised: October 17, 2024
Accepted: October 18, 2024
Published: November 12, 2024



https://doi.org/10.1021/jacs.4c10474
J. Am. Chem. Soc. 2024, 146, 31733−31745

Journal of the American Chemical Society pubs.acs.org/JACS Article

that has high sequence homology to the target protein of interest. For example, SHIFTX2 also takes advantage of existing databases through the SHIFTY+ component that introduces an alignment-and-transfer technique to fully exploit sequence homology in order to make more accurate predictions.

We recently introduced the UCBShift 1.0 method[20] that offers several advancements on these early foundational models for chemical shift prediction of backbone atoms. The first improvement is to modernize the machine learning to utilize a random forest regression model with a greatly expanded X-ray data set and feature extraction and transformations, and is referred to as the UCBShift-X predictor. We also qualitatively change the nature of homology to not only include high sequence homology but also to introduce proteins with low sequence but high structural homology, that comprises the UCBShift-Y module. A final random forest regression step combines the two modules to make chemical shift predictions if homology to the target protein is available, otherwise shift predictions are made with only the UCBShift-X module. The mean absolute error (MAE) of UCBShift 1.0 for backbone atoms of proteins is 0.31 ppm for amide hydrogens, 0.19 ppm for H_α, 0.84 ppm for C', 0.81 ppm for C_α, 1.00 ppm for C_β, and 1.81 ppm for N.

In this work we seek to improve accuracy and robustness, as well as to gain insight, for chemical shift calculations for the carbon, hydrogen, and nitrogen atoms of side chains of aqueous proteins by retraining the UCBShift 1.0 model over an expanded data set. Most existing chemical shift predictors like SPARTA+ are restricted to backbone atoms but the SHIFTX2 algorithm also considers side chains and is a comparative method we will consider in this work. Because both models extract engineered features from high quality protein X-ray crystal structures that are insensitive to alternate conformations of a protein in the thermalized ensemble, we test under what circumstances the −X modules provide good discriminative power for side chain C, H, and N chemical shifts in different atomic environments. The primary importance of sequence and structural homology based alignment in the −Y component is designed to overcome the limitations of extracting static features from crystal structures that undoubtedly miss the underlying dynamics that is inherent in the substitution with an experimental chemical shift from solution-based NMR.

The resulting UCBShift-2.0 algorithm achieves significantly lower mean absolute error (MAE) and lower root-mean-square-error (RMSE) compared to SHIFTX2 (which does not predict on some of the atom types at all) when evaluated on an independently generated test data set for side chains, with average MAEs of 0.80 ppm for C, 0.16 ppm for H, and 0.99 ppm for N chemical shifts. Overall the performance enhancement of UCBShift-2.0 arises from having more training data relative to the original SHIFTX2 study,[19] and more capacity and better exploitation of features and homology in the overall machine learning model. More specifically, UCBShift-2.0 can better predict the observed variations in chemical shift values for each amino acid, whereas SHIFTX2 mostly classifies shift predictions by residue type with little variation from random coil values. Furthermore, we find that while the Y-module is responsible for ∼50−60% of the performance enhancement of UCBShift-2.0 over SHIFTX2, side chain features such as ring currents and χ_2 rotamer states found through the X-module are also critically important for side chain H and C chemical shifts, which are dominated by C−H groups. By providing both source code and cleaned data sets available from an open source github, we recommend that UCBShift-1.0 and UCBShift-2.0, and their UCBShift-X and UCBShift-Y modules, can each offer a current state-of-the-art side chain chemical shift prediction algorithm that can be tailored to desired applications for protein backbone and side chain chemical shift prediction.

METHODS AND MODELS

Preparation of Enhanced Data Sets. The training set of the original UCBShift-1.0[20] was constructed by consolidating the training and testing sets SPARTA+ and SHIFTX+ into one comprehensive data set. For predicting protein side chain chemical shifts, we enriched this data set with chemical shifts available in the BMRB.[21] Chemical shift data obtained from the BMRB were rereferenced to high-resolution X-ray structures using Rereferenced Protein Chemical shift Database (RefDB).[22] The final number of experimental chemical shifts for training and testing are provided in Table 1, and broken down further in Tables S1 and S2. The PDB IDs with corresponding BMRBs of the 851 protein structures are provided in Table S3. The test set consists of 200 structures that were prepared in the previous release of the UCBShift-1.0, with details provided in Table S4.[20]

The structures from training and test sets were downloaded from RCSB and protonated using the PDB2PQR software.[23,24] We chose this algorithm because it allows for Propka-based protonation of ionizable residues based on the pH value. The PDB2PQR protonation was preceded by the optimization of adjustable protons (OH, SH, NH_3^+, Met-CH_3) and ambiguous Asn, Gln and His side chain orientation with the REDUCE software.[25] Zhang et al. point out the existence of assignment errors present in the BMRB database.[22] The training and test sets filtered out outliers that were 6 and 12 ppm bigger or smaller than the average value for the particular residue for hydrogens and carbon atoms, respectively. This high threshold allowed us to filter out only missassigned shifts following the original UCBShift-1.0 idea of the "real world" data.[20,26]

Harsch et al. have noted substantial misassignments of asparagine and glutamine side chain amide hydrogens within the BMRB.[27,28] According to their analysis, when the chemical shift differences between the two hydrogens are ⩾0.40 ppm for Asn and ⩾0.42 ppm for Gln, the resonance signal at the higher chemical shift (referred to as "downfield shifted") should be assigned to HD21 for Asn and HE21 for Gln, while the signal at the lower chemical shift ("upfield shifted") should be assigned to HD22 and HE22, respectively. We corrected the assignments in our training and test sets based on this rule. Specifically, we swapped the assignments of HD21/HD22 or HE21/HE22 when δ(HD21) < δ(HD22) or δ(HE21) < δ(HE22), and the difference between them was ⩾0.40 ppm for Asn or ⩾0.42 ppm for Gln. The resulting distributions of chemical shifts before and after correction are shown in Figure S1.

Machine Learning Workflow. The overall design of the UCBShift-1.0 and -2.0 chemical shift prediction algorithm is based on a random forest machine learning model, and consists of two modules, UCBShift-Y and UCBShift-X as shown in Figure 1. We used an Extra Tree Regressor and Random Forest Regressor as implemented in the scikit-learn package.[29] The hyperparameters were initially optimized with TPOT[30] using 3-fold cross-validation on the training set, then fine-tuned with a temporal validation set of 50 randomly selected structures. From here on out we refer to UCBShift-2.0 as UCBShift unless otherwise noted.

https://doi.org/10.1021/jacs.4c10474
J. Am. Chem. Soc. 2024, 146, 31733−31745

Journal of the American Chemical Society pubs.acs.org/JACS Article

Table 1. Number of Training and Testing Set Shift-Structure Examples for Every Type of Atom

Atom	Train	Test
CG	14 995	1289
CD1	4590	621
CG2	4300	731
CD	3467	515
CG1	2919	472
CD2	2735	367
CE	1649	256
CE1	1090	114
CE2	600	49
CZ	428	55
CZ2	176	16
CH2	167	13
CE3	156	17
CZ3	151	15
HB2	17 423	4554
HB3	16 259	4318
HG2	9667	2684
HB	6442	1820
HD2	6104	1705
HG3	5901	1416
HD1	4203	1234
HD3	2911	758
HE2	2612	645
HG	2076	450
HE1	2045	481
HE3	1409	345
HG12	1373	290
HG13	1239	274
HG1	1234	493
HD21	722	184
HD22	720	185
HE	604	162
HE21	595	164
HE22	594	162
HZ	558	164
HZ2	242	61
HH2	222	59
HZ3	211	56
ND2	672	120
NE2	562	126
NE1	287	47

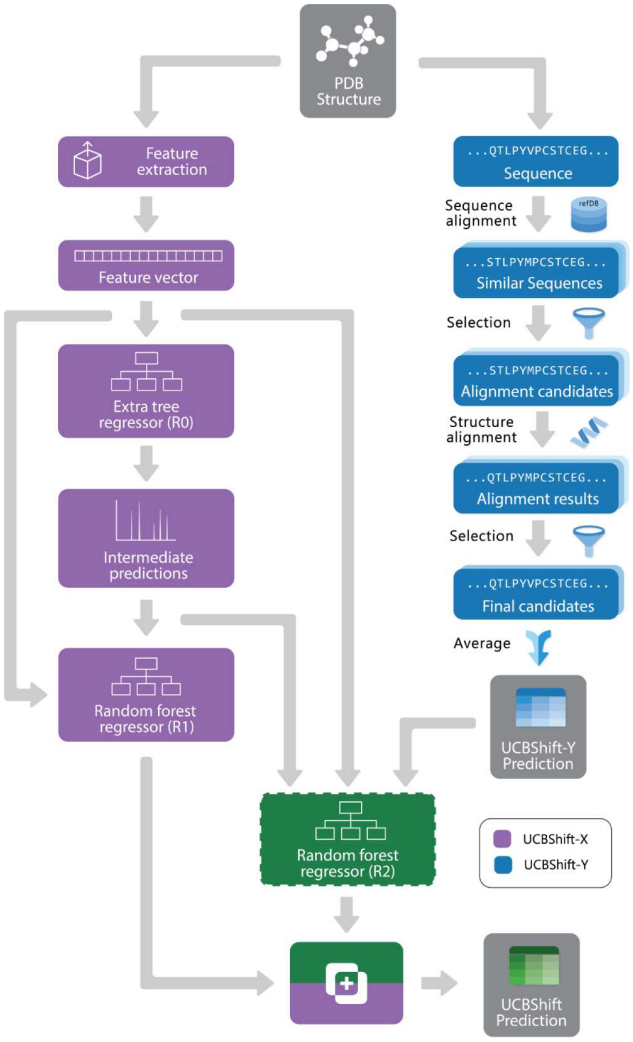


Figure 1. The overall design of the original UCBShift-1.0 chemical shift prediction algorithm used for UCBShift-2.0. The general UCBShift algorithm combines both a transfer prediction module that relies on both sequence and structural alignments, and a machine learning module that trains a tree regression model on augmented feature extracted data. The feature vector has been augmented for side chains as explained in the text. Reproduced from ref.[20] with permission from the Royal Society of Chemistry.

The UCBShift-Y component (blue path in Figure 1) transfers experimental chemical shifts from a reference database to a query protein based on sequence and structural similarity. The idea is similar to the SHIFTY+ module of SHIFTX2 predictor that utilizes the transfer of experimental chemical shifts from the protein database if the sequence is identical or closely matches the sequence of the query protein. UCBShift-Y also takes advantage of structural similarity in that it filters out mismatching chemical shifts of proteins with high sequence similarity but significantly different structure or when there is poor sequence alignment but significant structural similarity.

The UCBShift-X prediction algorithm (purple path in Figure 1) uses a feature vector and employs an extra tree regressor (R0) followed by a random forest regressor (R1).[31,32] The second Random Forest regressor (R2) incorporates the feature vector, the R1 regressor, and the secondary shift output from UCBShift-Y, along with additional scores and coverage metrics that indicate the quality of the alignments. The final chemical shift prediction is generated by either R1 (if UCBShift-Y predictions are unavailable) or R2 (if UCBShift-Y predictions are available). In the last step of the chemical shifts prediction, the random coil values are added back to the secondary shift predictions. Every type of atom chemical shift is trained individually. Some important details of the two components are described here for the readers benefit, but additional detail can be found in reference.[20]

UCBShift-Y Module. The algorithm begins by aligning the target sequence with all sequences available in the RefDB database[22] and selecting significant matches using the BLAST algorithm.[33] Subsequently, the PDB files of identified sequences undergo structural alignment with the query protein via the mTM-align algorithm.[34] Only the alignments with a TM score exceeding 0.8 and an RMSD with a target below 1.75 Å are retained. Afterward, the Needleman−Wunsch algorithm is used to determine the optimal alignment of each of the PDB

https://doi.org/10.1021/jacs.4c10474
J. Am. Chem. Soc. 2024, 146, 31733−31745

Journal of the American Chemical Society pubs.acs.org/JACS Article

Table 2. Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Pearson's Correlation Coefficients (R) for UCBShift and SHIFTX2[a,b,c]

	UCBShift			SHIFTX2			Improvement factors	
Atom	RMSE	MAE	R	RMSE	MAE	R	ΔRMSE	ΔMAE
CG	1.028(2)	0.635(1)	0.68	1.429(3)	0.909(2)	0.18	0.401	0.274
CD1	1.1253(3)	0.719(2)	0.74	1.530(4)	1.087(2)	0.44	0.405	0.368
CG2	1.0037(2)	0.671(1)	0.69	1.311(2)	0.916(1)	0.35	0.307	0.245
CD	1.2102(4)	0.745(2)	0.60	1.445(4)	0.901(3)	0.29	0.235	0.156
CG1	1.0935(4)	0.715(2)	0.69	1.373(4)	0.954(3)	0.44	0.280	0.239
CD2	1.2031(2)	0.846(2)	0.77	1.697(5)	1.252(3)	0.40	0.494	0.406
CE	0.9221(4)	0.584(2)	0.47	1.026(3)	0.671(2)	0.26	0.104	0.087
CE1	0.935(4)	0.637(3)	0.66	1.32(1)	0.889(5)	0.06	0.385	0.252
CE2	0.9975(4)	0.760(5)	0.53	1.128(1)	0.878(4)	0.23	0.131	0.118
CZ	1.5504(2)	0.97(1)	0.64	2.43(2)	1.64(1)	−0.25	0.88	0.67
CZ2	0.9999(8)	0.73(1)	0.60	-	-	-	-	-
CH2	1.5763(2)	1.02(2)	−0.09	-	-	-	-	-
CE3	1.3219(1)	1.08(1)	0.52	-	-	-	-	-
CZ3	1.9102(4)	1.11(2)	0.55	-	-	-	-	-
HB2	0.2257(3)	0.1228(1)	0.77	0.2632(3)	0.1479(1)	0.68	0.0375	0.0251
HB3	0.2278(3)	0.1265(1)	0.77	0.2882(4)	0.1720(2)	0.61	0.0604	0.0455
HG2	0.182(3)	0.0964(1)	0.80	0.2169(3)	0.1266(1)	0.68	0.0349	0.0302
HB	0.1696(5)	0.0906(2)	0.86	0.2115(4)	0.1156(2)	0.78	0.0419	0.0250
HD2	0.2181(1)	0.1116(3)	0.77	0.264(1)	0.1436(2)	0.64	0.046	0.0320
HG3	0.2069(5)	0.1153(2)	0.76	0.2457(5)	0.1394(2)	0.64	0.0388	0.0241
HD1[d]	0.2095(1)	0.1036(3)	0.82	0.247(1)	0.1385(3)	0.64	0.049	0.0374
HD3	0.2842(1)	0.1421(4)	0.69	0.315(1)	0.1577(4)	0.60	0.031	0.0156
HE2[d]	0.1668(1)	0.0931(3)	0.82	0.226(1)	0.1269(3)	0.64	0.060	0.0343
HG[d]	0.2504(1)	0.1430(4)	0.71	0.262(1)	0.1514(5)	0.66	0.015	0.0107
HE1	0.2688(1)	0.1416(5)	0.77	0.322(1)	0.1595(5)	0.66	0.053	0.0179
HE3	0.1698(1)	0.1016(3)	0.82	0.252(1)	0.1293(5)	0.55	0.082	0.0277
HG12	0.2782(1)	0.1965(1)	0.72	0.327(1)	0.228(1)	0.59	0.049	0.032
HG13	0.2932(1)	0.2048(1)	0.76	0.357(1)	0.250(1)	0.61	0.064	0.045
HG1[d]	0.2911(2)	0.1254(5)	0.58	0.215(1)	0.1221(3)	0.69	0.029	0.0207
HD21	0.402(2)	0.255(1)	0.53	0.498(2)	0.301(1)	0.24	0.096	0.046
HD22	0.317(1)	0.210(1)	0.44	0.361(1)	0.237(1)	0.23	0.044	0.027
HE	0.4575(4)	0.215(1)	0.77	0.526(4)	0.246(1)	0.69	0.069	0.031
HE21	0.412(3)	0.231(1)	0.35	0.428(4)	0.215(1)	0.41	0.016	−0.016
HE22	0.255(1)	0.1664(5)	0.47	0.291(1)	0.180(1)	0.35	0.036	0.014
HZ	0.3541(4)	0.172(1)	0.72	0.40(1)	0.179(1)	0.62	0.05	0.007
HZ2	0.2393(1)	0.161(1)	0.67	-	-	-	-	-
HH2	0.2216(1)	0.169(1)	0.62	-	-	-	-	-
HZ3	0.3701(3)	0.251(2)	0.53	-	-	-	-	-
ND2	1.7795(1)	1.02(1)	0.79	-	-	-	-	-
NE2	1.71(2)	0.88(1)	0.73	-	-	-	-	-
NE1	1.4729(1)	1.07(1)	0.77	-	-	-	-	-

[a]We found that 25% and 3.5% of the testing structures are only predicted by the UCBShift-X and SHIFTX+ algorithms, respectively, in which no alignment is possible. [b]For R coefficient calculations, chemical shifts were subtracted from the random coil references. [c]Uncertainties were computed based on 50 random samples from 75% of the test data. All units in ppm. [d]RMSE and MAE excluding atom types which SHIFTX2 does not predict. RMSE for HD1, HE2, HG, HG1: 0.198(1), 0.166(1), 0.247(1), 0.186(1), respectively. MAE for HD1, HE2, HG, HG1: 0.1011(3), 0.0926(3), 0.1407(4), 0.1014(3), respectively.

sequences with the RefDB sequence.[35] Figure S2 shows the calculated TM scores between each test sequence and the best-aligned sequence in the UCBShift and SHIFTX2 training sets. Even for cases exhibiting full sequence homology with the training set, the chemical shifts sourced from the BMRB database do not coincide between the training and test sets, since they correspond to different types of atoms.

In case of identical residues, the chemical shifts from RefDB are directly assigned to the query protein. Otherwise, the target shift for atom A in residue I is calculated from the matching residue J according to the equation:

$$\delta_{I,A} = \delta_{rc,I,A} + (\delta_{J,A} - \delta_{rc,J,A}) \quad (1)$$

where $\delta_{J,A}$ is the chemical shift of the matching residue, and δ_{rc}, and $\delta_{rc,I/J,A}$ are the random coil chemical shifts in residues I and J for atom type A. When there is more than one aligned structure the chemical shift is calculated as the weighted average with weights w_I given by

$$w_I = \begin{cases} e^{5(S_{NA}\times S_{TM})+B_{IJ}}, & B_{IJ} \geq 0 \\ 0, & B_{IJ} < 0 \end{cases} \quad (2)$$

https://doi.org/10.1021/jacs.4c10474
J. Am. Chem. Soc. 2024, 146, 31733−31745

Journal of the American Chemical Society pubs.acs.org/JACS Article

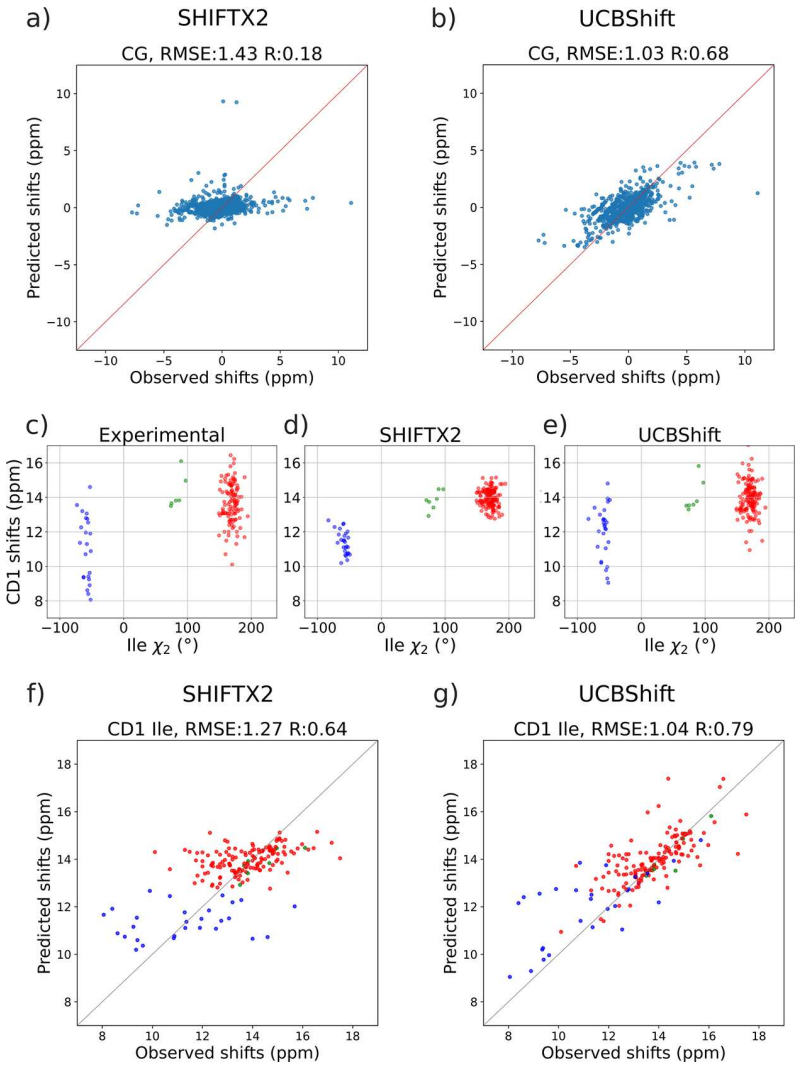


Figure 2. Predicted CG and CD1 chemical shifts by UCBShift and SHIFTX2 compared to experiment. Correlation between experimental and predicted CG chemical shifts subtracted from the random coil references for (a) SHIFTX2 and (b) UCBShift. (c−e) CD1 chemical shifts of isoleucine plotted against the χ_2 dihedral angle. Correlation between experimental and predicted CD1 chemical shifts of isoleucine for (f) SHIFTX2 and (g) UCBShift. In (c−g), three conformational clusters are marked with different colors.

where $S_{NA} = S_{blast}/\max(S_{blast})$ is the sequence alignment blast score normalized by the maximum blast score; S_{TM} is the structure alignment TM score; B_{IJ} represents the substitution score between the amino acid at position I in the target sequence and the amino acid at position J in the matching sequence, based on the BLOSUM62 matrix.[36] Weights are set to zero when the substitution scores are negative, usually indicating the substitutions in the target sequence are dissimilar residues.

UCBShift-X Module. UCBShift-X is a decision tree-based machine learning module, which extracts structural features from a PDB file or property calculations that depend on the coordinates for each atom type. These residue- and atom type-specific features extracted from the protonated PDB structures include the following:

- BLOSUM62 numbers: Substitution scores derived from the BLOSUM62 matrix, indicating the probability of replacing the query residue with any other amino acid.[36]
- Backbone dihedral Angles: Sine and cosine values of the ϕ and ψ torsion angles of the query residue, as well as for the preceding and following residues.
- Side-Chain Dihedral Angles: Binary indicators for the existence of $\chi_1, \chi_2, \chi_3, \chi_4$ and χ_5 side-chain dihedral angles, and their corresponding sine and cosine values. They are considered for the query residue and the adjacent residues.
- Hydrogen Bonds: For every side chain hydrogen, hydrogen bond's features are described by five numbers: existence (Boolean), distance between donor−acceptor pairs, cosine values of the angles at the donor hydrogen and acceptor atom, and hydrogen bond energy calculated via the DSSP model.[37] A hydrogen bond acceptor can be any oxygen or nitrogen atom in the protein.

  In addition, for each atom type, the backbone hydrogen bond descriptors related to the query residue are considered. These include the hydrogen bond between the amide hydrogen and carboxyl oxygen, and between the Cα hydrogen and a carboxyl group. For the query residue all five hydrogen bond descriptors are included for: (1) amide hydrogen (2) carboxyl oxygen, (3) α hydrogen and additionally (4) carboxyl oxygen features for the previous residue and (5) the amide hydrogen

https://doi.org/10.1021/jacs.4c10474
J. Am. Chem. Soc. 2024, 146, 31733−31745

Journal of the American Chemical Society pubs.acs.org/JACS Article

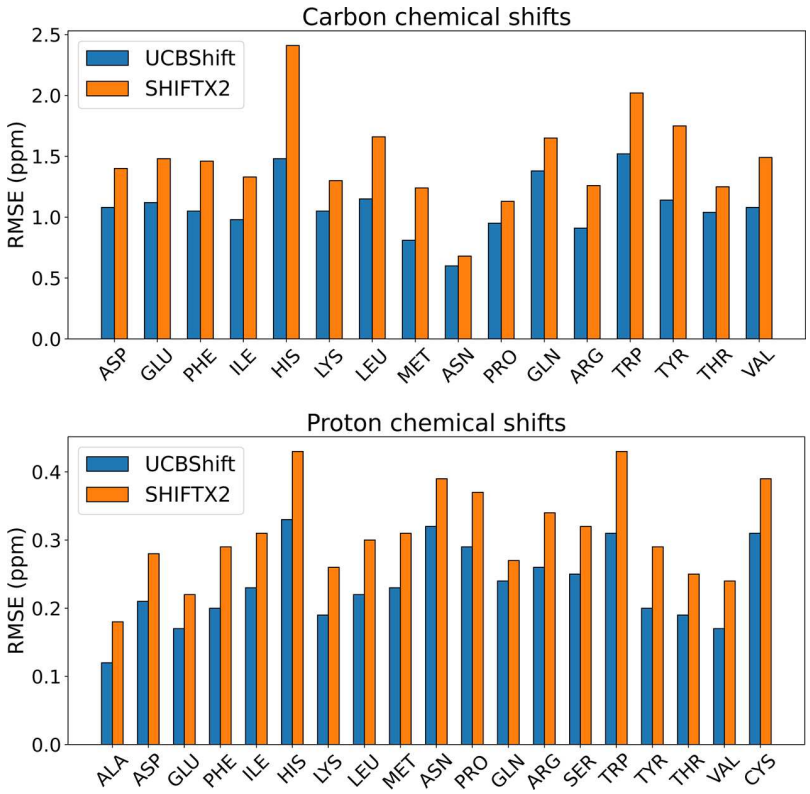


Figure 3. Root-mean-square error (RMSE) of carbon and hydrogen chemical shifts for every amino acid predicted by UCBShift and SHIFTX2 compared to experiment.

features for the next residue. This gives 25 backbone hydrogen bonding features for each atom type.

- Polynomial Transformations: Squared and inverse polynomial transformations of hydrogen bond distances and squares of dihedral angle cosine values, which are included in several empirical chemical shift calculation formulas.[38,39]
- S^2 order Parameters: NMR S^2 order parameters of N−H bond calculated using the contact model.[40]
- Accessible Surface Area (ASA): Absolute and relative accessible surface areas determined by the DSSP program.
- Secondary Structure: One-hot encoded secondary structure representation in 8 categories from the DSSP program.
- B Factor: Average B factor of the residue, extracted from the PDB file.
- Half Surface Exposure (HSE): A measure of the residue's exposure in the protein structure.[41]
- Hydrophobicity: Hydrophobicity values from the Wimley−White whole residue hydrophobicity scales.[42]
- Ring Current Effect: Calculated using the Haigh−Mallion model, including the ring current for the specific atom type in the training model.[43]
- pH value of the NMR experiment.
- Electric Field: Electric field effect calculated using the formula:

$$\delta_{EF} = \sum_i \frac{q_i \varepsilon \cos \theta_i}{d_i^2} \tag{3}$$

for hydrogen and nitrogen atoms. In the equation, q_i is the charge of the interacting atom i, θ is the i-H-X angle (where X is a heavy atom bound to the target hydrogen) and d_i is the distance between the hydrogen i and the target hydrogen. Interacting atom charges are derived from the Amber force field parameters.[44]

- Protonation Indicator: This binary feature indicates the presence (1) or absence (0) of atoms corresponding to ionizable residues, as determined by protonation state prediction software. This indicator relies on whether specific atoms, associated with protonation states, are present in the protonated PDB structure of the protein, reflecting significant shifts due to protonation changes.

RESULTS

Table 2 displays the MAE, root-mean-square error (RMSE), Pearson's correlation coefficients (R), and improvement factors computed using the UCBShift and SHIFTX2 models for chemical shift prediction by side chain atom type. The data reveals that UCBShift consistently outperforms SHIFTX2 across all side chain atom types, on average by ∼0.28 ppm and ∼0.03 ppm for carbons and hydrogens, respectively. Nitrogen chemical shifts are predicted by UCBShift, unlike SHIFTX2, with an average MAE of 0.99 ppm. The more detailed results, which distinguish between individual amino acids, also demonstrate improvement over SHIFTX2 in all types of nuclei (Table S5). It is also important to note that we utilized a "test mode" criteria for UCBShift predictions that excludes sequences with more than 99% similarity to the query sequence, while SHIFTX2 testing data may include 100% similarity cases. Overall UCBShift is a more consistent performing model for side chain chemical shift prediction despite this small handicap.

https://doi.org/10.1021/jacs.4c10474
J. Am. Chem. Soc. 2024, 146, 31733−31745

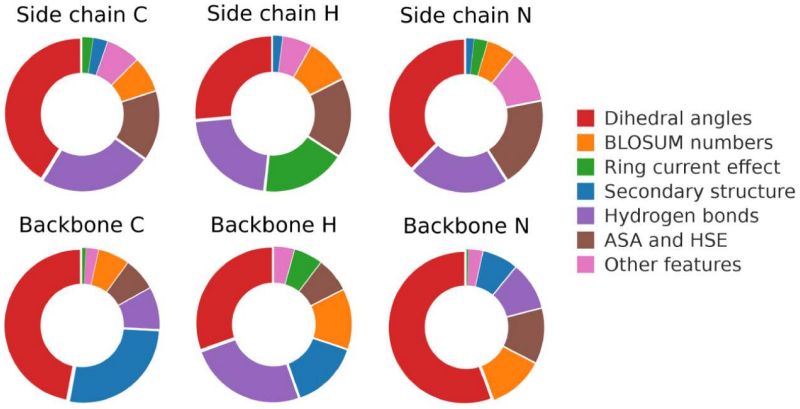


Figure 4. Relative importance of all the input features analyzed from the UCBShift model for side chain and backbone C, H, and N atom types. Importance values from the extra tree regressor (R0) were scaled in proportion to the "*Prediction from R0*" contribution in the random forest regressor (R1), Tables S9−S11. These scaled values were then added to the corresponding importance values from the R1 regressor, giving the overall R0/R1 importance of features. Feature importance was calculated as the mean decrease in impurity across all trees, using the built-in feature importance method from Scikit-learn.[29]

Protein side chains exhibit greater flexibility compared to the backbone dihedral angles, which are usually determined by the secondary structure. This higher flexibility leads to the presence of multiple rotameric states that makes accurate prediction of side chain chemical shifts more challenging.[45−47] In Figure 2 we provide a more detailed comparison of chemical shift accuracy using UCBShift versus SHIFTX2 illustrated with the CG and CD1 side chain atoms, which possess the largest amount of testing data as indicated in Table 1 and is present in many of the amino acid side chains. As seen in Figure 2a, the SHIFTX2 predictions exhibit a tendency to cluster around the amino acid random coil values, rather than accurately representing the experimental spread of chemical shift values that vary by each type of amino acid and ranges of environment. The UCBShift model, however, appears to better account for the shift variations within each amino acid type, providing predictions that are more finely tuned to the actual spread of observed shifts of CG as seen in Figure 2b.

As another example, we identify three conformational clusters of the experimental CD1 chemical shift for the isoleucine residue based on the χ_2 dihedral angle as seen in Figure 2c−e, in which the conformation with $\chi_2 = -60°$ has been reported to correspond to the chemical shift value of ∼11.4 ppm.[46] However, the broad experimental CD1 chemical shift distribution in each of the $\chi_2 = -60°$ as well as $\chi_2 = 180°$ clusters suggests that many conformational states can be present in solution, Figure 2c. We observe that the UCBShift predictor better captures the large spread of chemical shifts within each geometrical cluster, while SHIFTX2 does not predict the same spread of CD1 shifts within clusters observed in the experimental data. UCBShift additionally reveals a more favorable correlation between the predicted and experimental chemical shifts for the CD1 shifts of isoleucine, Figure 2f,g. Similar trends are observed for other atom types such as CG2 and including nitrogen which is only predicted by UCBShift, as summarized in Figures S3−S10.

The bar plots depicted in Figure 3 summarize the enhancement of UCBShift shift prediction for carbon and hydrogen side chain atoms in comparison to SHIFTX2 across various amino acid types. The overall RMSE using UCBShift systematically improves over SHIFTX2 for all amino acids. Histidine demonstrates a particularly pronounced enhancement in carbon chemical shifts prediction, which we attribute to the effective handling of protonation states by UCBShift.[26] Tyrosine also displays a notable decrease in errors. Looking at the details in Table S5, we identify the most significant discrepancies for CG and CZ. Reported experimental data for these nuclei is very limited (Table S1), for reasons discussed in later section. The improved performance of UCBShift for cases with little training data can be attributed to the extensive training set employed during its development, consisting of 851 structures compared to the 197 structures incorporated in SHIFTX2.

Next we consider why there are performance enhancements for UCBShift relative to SHIFTX2. Table S6 makes clear that the X-module of UCBShift shows small performance enhancements relative to SHIFTX2 over most of the C and H side chain chemical shift test data, although there are some exceptions as well. However, the more notable component of success for UCBShift appears to be in the Y-module predictions as reported in Table S6. Hence when any type of homology is available, along with an assigned experimental chemical shift, it improves average chemical shift MAEs by 0.5 ppm for side chain carbons and 0.2 ppm for side chain hydrogens compared to the standalone UCBShift-X component.

We believe there are several contributing factors to this overall improvement through the Y-module. First is that we have more training data than the much earlier SHIFTX2 study, and in particular more sequence/structural homology data, that can be exploited better by the more sophisticated R2 regressor (see Figure 1). But to test this possibility, we also retrained the UCBShift model on the original SHIFTX2 data set as seen in Table S7. This demonstrates that the UCBShift algorithm is not solely benefiting from an expanded data set but a better algorithm as well. In addition, it appears that, on average, feature extraction using static crystal structures cannot represent the solution NMR experiment, in which the substituted experimental chemical shift better represents the thermal fluctuations and time averaging over variable chemical environments not available in the X-module. Finally, given that ∼75% of the data has available homology, UCBShift is trained to rely more heavily on this component of the machine learning algorithm.

This is not to say that feature extraction data is not important, since the R2 regressor takes into account not only the alignment, but direct feature data as well as the extra tree regressor R0 (see

https://doi.org/10.1021/jacs.4c10474
J. Am. Chem. Soc. 2024, 146, 31733−31745

Journal of the American Chemical Society pubs.acs.org/JACS Article

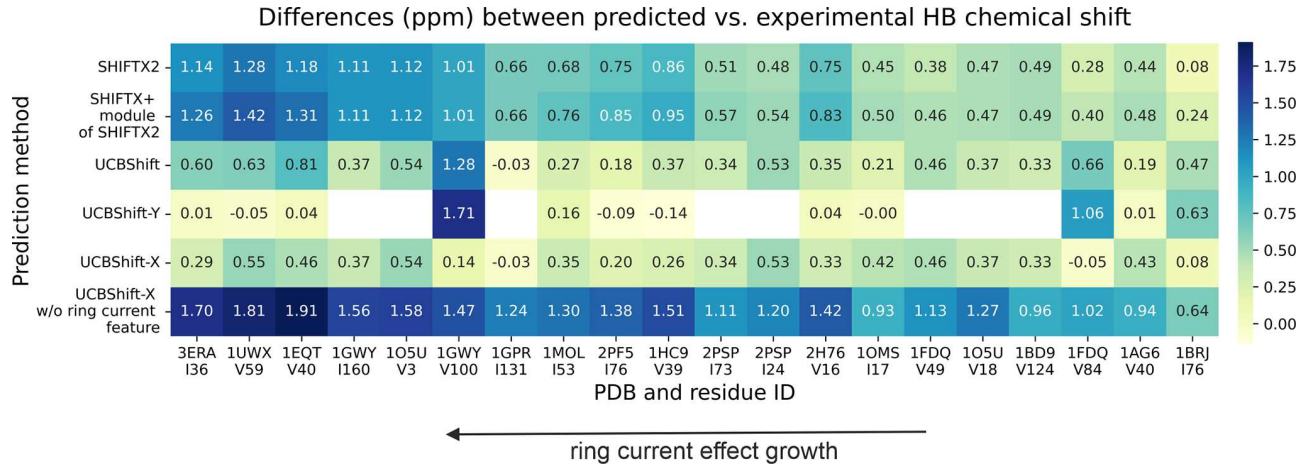


Figure 5. Heatmap presenting differences between predicted and experimental HB chemical shifts of Ile and Val from the subset including hydrogens involved in CH···π interactions. UCBShift-X with turned off ring current effect feature, UCBShift-X, UCBShift-Y, UCBShift, SHIFTX+ module of SHIFTX2 and SHIFTX2 prediction models are considered. Chemical shifts are in ascending order from left to right (from the strongest to the weakest ring current effect). The average absolute error (ppm) across this data set is 0.78 for SHIFTX2, 0.84 for SHIFTX+, 0.52 for UCBShift, 0.61 for UCBShift-Y, and 0.36 for UCBShift-X.

Figure 1). In Table S6 we see that the R2 regressor performs better than UCBShift-Y in regards MAE for CG, HB, HB2/HB3, HD21/22, HE21/22 atom types, and often eliminating outliers for many atoms by reducing the RMSE. The enhancement of the R2 regressor for hydrogens HD21/22, HE21/22 is especially noteworthy, as the underperformance of module Y may stem from the previously discussed misassignments of these hydrogens in asparagine and glutamine. We also constructed a low-homology test set of 59 proteins with a sequence identity of 50% or less to those in the training set. Table S8 shows that UCBShift outperforms SHIFTX2 on this more restrictive low homology test set that also demonstrates that the X-module is critically important as well.

Finally, we examine the various features used in the UCBShift model for predicting chemical shifts for each atom type in backbone and side chains, summarized in Tables S9−S11. Unsurprisingly, the UCBShift-Y prediction dominates the R2 regressor, with influence of features through the unoptimized node splits of the R0 regressor, for reasons explained above. Hence we analyze the test examples for the features that are important for the unoptimized node splits of the R0 regressor, and then corrected by the optimized R1 random forest regressor and displayed graphically in Figure 4.

Our previous investigation using UCBShift-1.0 for backbone atoms revealed that for carbon and nitrogen chemical shifts the backbone dihedral angle features are most important (Tables S9 and S11.). Additionally for backbone carbons the secondary structure classification dominates the R0/R1 regressor as seen in Figure 4 and Table S9. For backbone hydrogen atoms the same geometric features are important as well, but with the added component of untransformed and transformed hydrogen-bonding together with the stability and conservation of those structural features as determined by the BLOSUM number (Figure 4 and Table S10). These dominant features are perhaps unsurprising but reassuring, and support the essential role that backbone chemical shifts provide in protein structure determination.[48]

Moreover, various studies have explored the utility of side chain chemical shifts in elucidating protein conformation and dynamics.[46,47,49] It is notable that analysis of R0/R1 reveals the predominant relevance of various forms of side-chain dihedral angles for carbon, nitrogen, and hydrogen chemical shifts, BLOSUM numbers, accessible surfaces, and various forms of hydrogen bond features as shown in Figure 4. Interestingly, B factors, S2 order parameters, hydrophobicity, pH values, and electric fields play more negligible roles for chemical shift predictions for any side chain C, H, or N atom.

Perhaps the most distinguishing feature of side chain hydrogen chemical shift predictions with UCBShift is the importance of ring currents as seen in Figure 4. When a hydrogen is surrounded by an aromatic ring, it undergoes a significant shielding due to the electric current induced by an external magnetic field. Hence hydrogen chemical shifts are highly sensitive reporters of weak CH···π interactions that play important role in the stabilization of protein structures[50−53] and protein−ligand complexes.[51,54,55] For most of the hydrogen types including carbon bound hydrogens HB, HB2/3, HG2/3/12/13, HD2/3, and HE21 the ring current effect is a dominating feature with importance greater than 20% (Table S10). Interestingly, the carbon-bonded HA hydrogen reveals a significantly lower, 5.3%, ring current feature importance. The observation can be attributed to the analysis of protein PDB structures, which indicates that HA atoms, compared to the side-chain hydrogens, exhibit the least participation in CH···π interactions.[56]

Thus, we further analyze UCBShift on a test subset containing hydrogen atoms that reveal notable shielding as a result of the ring current impact stemming from CH···π interactions in Figure 5. From the testing set, we select 20 chemical shifts below 1 ppm of HB atoms in Ile and Val, where the distance from the hydrogen to the center of the aromatic ring $\leqslant$ 3.7, Table S12. Figure 5 demonstrates that UCBShift surpasses SHIFTX2 in handling these extreme examples, and among all the submodels, UCBShift-X exhibits the highest level of performance. In order to assess the impact of the ring current feature on the prediction outcome, we conduct a prediction with the UCBShift-X model while deactivating the ring current feature. Analysis of the last two rows of the heatmap in Figure 5 shows that the ring current feature decreases chemical shift values throughout the data set. Additionally, it is observed that SHIFTX+ and SHIFTX2 exhibit

https://doi.org/10.1021/jacs.4c10474
J. Am. Chem. Soc. 2024, 146, 31733−31745

a lower accuracy in predicting the chemical shifts of hydrogens located in the most shielded positions on the left side. In contrast, the consistency of UCBShift-X and UCBShift performance remains relatively stable throughout the data set.

Turning to the hydrogen types including N-bound hydrogens, e.g., backbone amide hydrogen, HD21/22 (Asn), HE1 (Trp), HE (Arg) we observe that H-bond features may overcome the ring current effect, Table S10. This is caused by the ability of N-bound hydrogens to act as effective hydrogen bond donors. Strong hydrogen bonds in the form of N−H···N/O lead to a pronounced deshielding effect resulting in a large downfield shift of the interacting hydrogen.[52] Due to the large difference in chemical shift between strong, weak and non-hydrogen-bonded NH hydrogens the chemical shift range of these nuclei is broad and hence more difficult to properly assign.[27,28] For those hydrogens the prediction is usually worse, RMSE ∼ 0.5 pmm, than for carbon-bound hydrogens with only marginal improvement relative to SHIFTX2 as seen in Table S5. Additionally, there is highly limited experimental chemical shifts data for N-bound side chain hydrogens (Tables S1 and S2). Interestingly, the relevance of hydrogen bond features extends to the chemical shifts of carbon-bound hydrogens, which can be attributed to the occurrence of weak CH···O/N hydrogen bonds in protein structures.[57]

Among ionizable residues, histidine is particularly noteworthy because its protonation state can vary within physiological pH ranges. Our analysis indicates that the protonation feature impacts the R0 regression for CD2 and HE1 atom types by 3.1% and 2.2% respectively, Tables S9 and S10. This is in line with literature revealing a significant variation in histidine CD2 and HE1 chemical shifts upon ionization.[26] Overall, both, various protonation states of the histidine as well as its ability to form hydrogen bonds lead to a broader distribution of reported chemical shifts, resulting in a higher prediction error for this residue type (Figure 3 and Table S5).[58]

The UCBShift-2.0 algorithm covers all side chain carbon atom types, but there is still a shortage of experimentally determined chemical shifts for 5 types of hydrogen atoms, specifically the hydroxyl group hydrogen (HH) of tyrosine and the terminal NH hydrogen atoms (HH11/12/21/22) of the guanidinium group in arginine. There is also insufficient experimental data for five nitrogen atom types, namely NH1/2 and NE of the guanidinium group in arginine, NZ of lysine, and ND1 of histidine. In addition, there is a limited amount of experimental chemical shift data and therefore relatively poor prediction performance for O/S- and some of the N-bound hydrogens (HG of Ser and Cys, HG1 of Thr, HZ of Lys, and HD1 and HE2 of His), Tables S1 and S5. Acidic hydrogens are known to be particularly challenging to measure due to their rapid exchange with water.[26,59−62] Direct observation of these nuclei is limited to situations where they are protected from rapid solvent exchange, for example when buried within a folded protein structure (due to hydrogen bonds or steric hindrance) and/or under conditions of low temperature or low pH. The HH11/12/21/22 hydrogen peaks of arginine are even more difficult to detect as they not only undergo proton exchange but are also affected by the NE-CZ bond rotation resulting in a single broad signal for the four hydrogens.[63] Additionally, the detection of O- and S-bound hydrogens (HG of Ser and Cys, HG1 of Thr, HH of Tyr), even under the rare slow exchange conditions, is hampered by the inapplicability of conventional 2D NMR methods.[64−67] In their discussion, Takeda and coauthors pointed out that NMR signals of labile protons can often be misinterpreted due to the above-mentioned challenges.[64]

When it comes to carbon side chain chemical shifts, there is a noticeable shortage of experimental chemical shifts available for quaternary carbon atoms in the data set (CG chemical shifts of Phe, His, Trp, Tyr, CZ of Arg, Tyr, and CD2, CE2 of Trp), Table S1. The assignment of those nuclei is experimentally more challenging due to a lack of a directly attached hydrogen and requires more complex long-range NMR experiments explaining the sparse data in the BMRB.[68] This is for most of the quaternary carbon atom types reflected in the somewhat subpar predictive performance shown in Table S5.

■ DISCUSSION AND CONCLUSIONS

In this work we build upon the UCBShift-1.0 model for backbone atoms by extending it to side chain chemical shift predictions. We have used numerical and categorical features derived from high quality but static crystal structures that are greatly expanded in terms of training and test data, and we have redesigned a structure based alignment module to directly transfer the chemical shifts from the experimental database to the query protein when not only sequence but when the structure homology is sufficiently high, all incorporated in a random forest regression model. Overall UCBShift-1.0 and -2.0 stands as a modernized and state-of-the-art predictor of both backbone and side chain carbon, hydrogen, and nitrogen chemical shifts.

UCBShift yields a far superior side chain chemical shift prediction than SHIFTX2 through small improvements to its X-module, but mostly through its Y-module which is largely responsible for raising overall UCBShift performance within and across all amino acids and their various atom types. This is because as NMR data accumulates over time, 75% of folded proteins have a substitutable chemical shift from another protein with high sequence and/or structural homology, and the experimental chemical shifts contain dynamical and ensemble averaging over environments that is not easily extracted from static crystal structures. But UCBShift-Y module still does not outperform the R2 regressor in all cases, because the engineered feature set also help predict the environmental variations of observed chemical shift values within each amino acid type. In particular, the engineered features of ring currents are especially important for hydrogen chemical shifts of nearly all side chains, especially for CH···π interactions, and the greater environmental variations of observed C chemical shift values for each amino acid depend on a dominant χ_2 rotamer. For extreme cases and/or underrepresented examples such as low-frequency CH···π interactions, the UCBShift-X module outperforms UCBShift itself, due to interference by the UCBShift-Y where the transferred shift is unhelpful. Hence while UCBShift-1.0 and -2.0 can be used as a black-box algorithm for chemical shift prediction, we encourage users to take advantage of all modular UCBShift-X and UCBShift-Y components that can be tailored to desired applications for protein backbone and side chain chemical shift prediction.

Encouraged by the performance of our prediction tool we anticipate several valuable applications with significant impact on biomolecular NMR spectroscopy. Specifically, we suggest backward and forward applications of UCBShift 2.0. The most obvious backward application would be a 3D structure-based rigorous validation of experimental NMR signal assignments either obtained via automated or manual approaches. The example of Asn and Gln side-chain signals described in our study

https://doi.org/10.1021/jacs.4c10474
J. Am. Chem. Soc. 2024, 146, 31733−31745

Journal of the American Chemical Society pubs.acs.org/JACS Article

suggests valuable applications for these notoriously difficult amino acids. It is interesting to note that a related correction tool was developed in the past for correcting erroneously annotated Asn/Gln rotamers in experimental X-ray structures.[69]

The introduction of AlphaFold2 (AF2) has clearly revolutionized structural biology.[70] Despite the high level of accuracy of AF2 models there is still the need for independent model evaluation and several experimental methods (SAXS, X-ray crystallography, cryo-EM or NMR) have been proposed for this task.[71] The availability of side-chain information in UCBShift-2.0 will clearly boost the relevance of NMR spectroscopy in this endeavor. The necessary prerequisite for NMR-based evaluation of AF2 models is the availability of at least a backbone resonance assignment. With the advancement of machine learning techniques in biomolecular NMR spectroscopy this bottleneck no longer exists. For example, the ARTificial Intelligence for NMR Applications method (ARTINA)[72] and the NMRtist Web server[73] were introduced for visual spectrum analysis and automated signal assignment starting from raw experimental NMR data. More recently, a novel time-optimized deep learning approach for protein NMR assignment was introduced that employs AlphaFold and chemical shift prediction.[74] Importantly, in this approach the previous version of UCBShift was employed to provide backbone chemical shift data. Clearly, the now available side-chain information will lead to a further improvement of this powerful assignment approach.

While applications to automated NMR signal assignment and validation of AF2 predicted structural models are clearly important and highly relevant, we strongly believe that potentially transformative forward applications will be in the study of protein conformational dynamics. It is common and undisputed knowledge that NMR provides unique insight into the structural dynamics of proteins in solution and numerous NMR spin relaxation experiments have been designed over the years.[75] Importantly, the observed solution chemical shifts in proteins are the result of subtle conformational averaging process and therefore probe the entirety of the proteins' dynamics. An interesting exploitation of this intricate relationship was recently introduced with an approach to assess structural dynamics along the protein backbone via a method called Accuracy of NMR Structures Using the so-called NMR random coil index (RCI)[76] and Rigidity (ANSURR).[77] It calculates the local rigidity of a protein structure[78] and compares it with the local rigidity as measured using a version of the RCI based on backbone NMR chemical shifts. While this method was merely introduced to assess the accuracy of solution structures an extension of this conceptual thinking to side-chain chemical shift data could radically change the way protein dynamics are probed by NMR spectroscopy.

Specifically, there is a well established dependence of ^{13}C side-chain chemical shifts on dihedral angles as a tool for conformational analysis in proteins,[46] and simple relations exist to relate Ile ($^{13}C^{\delta_1}$)[47] and Leu ($^{13}C^{\delta_1,\delta_2}$)[49,79] methyl ^{13}C chemical shifts into χ_2 rotamer populations for these residues. Analogous relationships and procedures exist for obtaining χ_1 torsion angle distributions for Val residues in proteins on the basis of measured $^{13}C^{\gamma_1,\gamma_2}$ shifts exclusively. Importantly, such chemical shifts, when available through relaxation dispersion NMR measurements, can even be used to infer "invisible" excited protein states,[80−82] that are only sparsely and transiently populated. Analogous extensions to other amino acids can be envisaged.

Moreover, NMR spin relaxation measurements obtained for methyl group carbons revealed that, for example, extracted order parameters correlate with conformational averaging along the amino acid side-chain and thus again providing support for the expected correlation between chemical shift and conformational dynamics. While ^{13}C chemical shifts primarily report on side-chain conformational averaging, ^{1}H chemical shifts will be exquisitely sensitive to subtle structural arrangements of aromatic ring systems within a protein. We have recently demonstrated how this information can be used to extract binding poses in protein−ligand complexes with significant ramifications in structure-based drug design programs.[83,84] Most importantly, QM calculations of ligand ^{1}H chemical shifts for the protein-bound state allowed for refinement of the solution structures of protein−ligand (drug) complexes.[83] Developments (exploiting machine learning techniques) are ongoing to improve QM calculations and make them applicable also to larger biomolecular systems.

To conclude, we anticipate that the improved chemical shift prediction tool presented here, and the explosive growth of machine learning to NMR structural biology[2,11,85−87] to generate protein conformational ensembles will offer exciting possibilities that cannot be analyzed using more conventional tools of structural biology. A better quantitative understanding of the intricate relationship between protein structure and NMR chemical shift will be highly valuable in this endeavor.

■ ASSOCIATED CONTENT

Data Availability Statement

The code is provided through a GitHub repository link: https://github.com/THGLab/CSpred/tree/SideChain.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/jacs.4c10474.

Details of proteins and chemical shifts used in training and test sets; more analysis of UCBShift and SHIFTX2 performance on various benchmarks (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Teresa Head-Gordon − *Pitzer Center for Theoretical Chemistry, Department of Chemistry and Departments of Bioengineering and Chemical and Biomolecular Engineering, University of California, Berkeley, California 94720, United States;* orcid.org/0000-0003-0025-8987; Email: thg@berkeley.edu

Authors

Aleksandra L. Ptaszek − *Christian Doppler Laboratory for High-Content Structural Biology and Biotechnology, Department of Structural and Computational Biology, Max Perutz Laboratories, University of Vienna, Vienna 1030, Austria; Laboratory of Computer-Aided Molecular Design, Division of Medicinal Chemistry, Otto-Loewi Research Center, Medical University of Graz, A-8010 Graz, Austria*

Jie Li − *Pitzer Center for Theoretical Chemistry, Department of Chemistry, University of California, Berkeley, California 94720, United States;* orcid.org/0000-0002-4727-1786

Robert Konrat − *Christian Doppler Laboratory for High-Content Structural Biology and Biotechnology, Department of Structural and Computational Biology, Max Perutz Laboratories, University of Vienna, Vienna 1030, Austria*

https://doi.org/10.1021/jacs.4c10474
J. Am. Chem. Soc. 2024, 146, 31733−31745

Gerald Platzer − *MAG-LAB GmbH, Vienna 1030, Austria*

Complete contact information is available at: https://pubs.acs.org/10.1021/jacs.4c10474

Author Contributions

#A.L.P. and J.L. contributed equally to this paper.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

A.L.P. was funded by the Christian Doppler Laboratory for High-Content Structural Biology and Biotechnology, Austria, and received further support from the DosChem doctoral school program Faculty of Chemistry, University of Vienna. UC Berkeley investigators thank the National Institutes of Health (2R01GM127627-05) for funding this work.

■ REFERENCES

(1) Paruzzo, F. M.; Hofstetter, A.; Musil, F.; De, S.; Ceriotti, M.; Emsley, L. Chemical shifts in molecular solids by machine learning. *Nat. Commun.* **2018**, *9*, 4501.
(2) Liu, S.; Li, J.; Bennett, K. C.; Ganoe, B.; Stauch, T.; Head-Gordon, M.; Hexemer, A.; Ushizima, D.; Head-Gordon, T. Multiresolution 3D-DenseNet for Chemical Shift Prediction in NMR Crystallography. *J. Phys. Chem. Lett.* **2019**, *10*, 4558−4565.
(3) Case, D. A. Molecular dynamics and NMR spin relaxation in proteins. *Acc. Chem. Res.* **2002**, *35*, 325−331.
(4) Case, D. A. Chemical shifts in biomolecules. *Curr. Opin. Struct. Biol.* **2013**, *23*, 172−176.
(5) Christensen, A.; Linnet, T.; Borg, M.; Boomsma, W.; Lindorff-Larsen, K.; Hamelryck, T.; Jensen, J. Protein Structure Validation and Refinement Using Amide Proton Chemical Shifts Derived from Quantum Mechanics. *PLoS One* **2013**, *8*, No. e84123.
(6) Dyson, H. J.; Wright, P. E. Unfolded Proteins and Protein Folding Studied by NMR. *Chem. Rev.* **2004**, *104*, 3607−3622.
(7) Ball, K. A.; Wemmer, D. E.; Head-Gordon, T. Comparison of structure determination methods for intrinsically disordered amyloid-beta peptides. *J. Phys. Chem. B* **2014**, *118*, 6405−6416.
(8) Bhowmick, A.; Brookes, D. H.; Yost, S. R.; Dyson, H. J.; Forman-Kay, J. D.; Gunter, D.; Head-Gordon, M.; Hura, G. L.; Pande, V. S.; Wemmer, D. E.; Wright, P. E.; Head-Gordon, T. Finding Our Way in the Dark Proteome. *J. Am. Chem. Soc.* **2016**, *138*, 9730−9742.
(9) Swails, J.; Zhu, T.; He, X.; Case, D. A. AFNMR: Automated fragmentation quantum mechanical calculation of NMR chemical shifts for biomolecules. *J. Biomol. NMR* **2015**, *63*, 125−139.
(10) Haghighatlari, M.; Li, J.; Heidar-Zadeh, F.; Liu, Y.; Guan, X.; Head-Gordon, T. Learning to Make Chemical Predictions: The Interplay of Feature Representation, Data, and Machine Learning Methods. *Chem* **2020**, *6*, 1527−1542.
(11) Li, J.; Liang, J.; Wang, Z.; Ptaszek, A. L.; Liu, X.; Ganoe, B.; Head-Gordon, M.; Head-Gordon, T. Highly Accurate Prediction of NMR Chemical Shifts from Low-Level Quantum Mechanics Calculations Using Machine Learning. *J. Chem. Theory Comput.* **2024**, *20*, 2152−2166.
(12) Meiler, J.; Baker, D. Coupled prediction of protein secondary and tertiary structure. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100*, 12105−12110.
(13) Shen, Y.; Bax, A. SPARTA+: A modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. *J. Biomol. NMR* **2010**, *48*, 13−22.
(14) Iwadate, M.; Asakura, T.; Williamson, M. P. Cα and Cβ carbon-13 chemical shifts in proteins from an empirical database. *J. Biomol. NMR* **1999**, *13*, 199−211.
(15) Neal, S.; Nip, A. M.; Zhang, H.; Wishart, D. S. Rapid and accurate calculation of protein 1 H, 13 C and 15 N chemical shifts. *J. Biomol. NMR* **2003**, *26*, 215−240.
(16) Kohlhoff, K. J.; Robustelli, P.; Cavalli, A.; Salvatella, X.; Vendruscolo, M. Fast and accurate predictions of protein NMR chemical shifts from interatomic distances. *J. Am. Chem. Soc.* **2009**, *131*, 13894−13895.
(17) Li, D.-W.; Brüschweiler, R. PPM: A side-chain and backbone chemical shift predictor for the assessment of protein conformational ensembles. *J. Biomol. NMR* **2012**, *54*, 257−265.
(18) Li, D.; Brüschweiler, R. PPM_One: A static protein structure based chemical shift predictor. *J. Biomol. NMR* **2015**, *62*, 403−409.
(19) Han, B.; Liu, Y.; Ginzinger, S. W.; Wishart, D. S. SHIFTX2: Significantly improved protein chemical shift prediction. *J. Biomol. NMR* **2011**, *50*, 43−57.
(20) Li, J.; Bennett, K. C.; Liu, Y.; Martin, M. V.; Head-Gordon, T. Accurate prediction of chemical shifts for aqueous protein structure on "Real World" data. *Chem. Sci.* **2020**, *11*, 3180−3191.
(21) Ulrich, E. L.; et al. BioMagResBank. *Nucleic Acids Res.* **2007**, *36*, D402−D408.
(22) Zhang, H.; Neal, S.; Wishart, D. S. RefDB: A database of uniformly referenced protein chemical shifts. *J. Biomol. NMR* **2003**, *25*, 173−195.
(23) Dolinsky, T. J.; Nielsen, J. E.; McCammon, J. A.; Baker, N. A. PDB2PQR: An automated pipeline for the setup of Poisson−Boltzmann electrostatics calculations. *Nucleic Acids Res.* **2004**, *32*, W665−W667.
(24) Dolinsky, T. J.; Czodrowski, P.; Li, H.; Nielsen, J. E.; Jensen, J. H.; Klebe, G.; Baker, N. A. PDB2PQR: Expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Res.* **2007**, *35*, W522−W525.
(25) Word, J. M.; Lovell, S. C.; Richardson, J. S.; Richardson, D. C. Asparagine and glutamine: Using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.* **1999**, *285*, 1735−1747.
(26) Platzer, G.; Okon, M.; McIntosh, L. P. pH-dependent random coil 1 H, 13 C, and 15 N chemical shifts of the ionizable amino acids: A guide for protein p K a measurements. *J. Biomol. NMR* **2014**, *60*, 109−129.
(27) Harsch, T.; Dasch, C.; Donaubauer, H.; Baskaran, K.; Kremer, W.; Kalbitzer, H. Stereospecific assignment of the asparagine and glutamine side chain amide protons in random-coil peptides by combination of molecular dynamic simulations with relaxation matrix calculations. *Appl. Magn. Reson* **2013**, *44*, 319−331.
(28) Harsch, T.; Schneider, P.; Kieninger, B.; Donaubauer, H.; Kalbitzer, H. R. Stereospecific assignment of the asparagine and glutamine sidechain amide protons in proteins from chemical shift analysis. *J. Biomol. NMR* **2017**, *67*, 157−164.
(29) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(30) Olson, R. S.; Urbanowicz, R. J.; Andrews, P. C.; Lavender, N. A.; Kidd, L. C.; Moore, J. H. Automating biomedical data science through tree-based pipeline optimization. In *Applications of Evolutionary Computation: 19th European Conference, EvoApplications 2016*, Springer: 2016; pp 123−137.
(31) Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **2006**, *63*, 3−42.
(32) Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5−32.
(33) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403−410.
(34) Dong, R.; Peng, Z.; Zhang, Y.; Yang, J. mTM-align: An algorithm for fast and accurate multiple protein structure alignment. *Bioinformatics* **2018**, *34*, 1719−1725.
(35) Needleman, S. B.; Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **1970**, *48*, 443−453.
(36) Henikoff, S.; Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U. S. A.* **1992**, *89*, 10915−10919.

(37) Kabsch, W.; Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577−2637.
(38) Wishart, D.; Sykes, B.; Richards, F. Relationship between nuclear magnetic resonance chemical shift and protein secondary structure. *J. Mol. Biol.* **1991**, *222*, 311−333.
(39) Wagner, G.; Pardi, A.; Wuethrich, K. Hydrogen bond length and proton NMR chemical shifts in proteins. *J. Am. Chem. Soc.* **1983**, *105*, 5948−5949.
(40) Zhang, F.; Bruschweiler, R. Contact Model for the Prediction of NMR N-H Order Parameters in Globular Proteins. *J. Am. Chem. Soc.* **2002**, *124*, 12654−12655.
(41) Hamelryck, T. An amino acid has two sides: A new 2D measure provides a different view of solvent exposure. *Proteins: Struct., Funct., Bioinf.* **2005**, *59*, 38−48.
(42) Wimley, W. C.; White, S. H. Experimentally determined hydrophobicity scale for proteins at membrane interfaces. *Nat. Struct. Biol.* **1996**, *3*, 842−848.
(43) Haigh, C.; Mallion, R. Ring current theories in nuclear magnetic resonance. *Prog. Nucl. Magn. Reson. Spectrosc.* **1979**, *13*, 303−344.
(44) Case, D. A.; Aktulga, H. M.; Belfon, K.; Ben-Shalom, I.; Brozell, S. R.; Cerutti, D. S.; Cheatham, T. E., III; Cruzeiro, V. W. D.; Darden, T. A.; Duke, R. E., et al. *Amber 2021*; University of California: San Francisco, 2021.
(45) MacArthur, M. W.; Thornton, J. M. Protein side-chain conformation: A systematic variation of $\chi 1$ mean values with resolution−a consequence of multiple rotameric states? *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **1999**, *55*, 994−1004.
(46) London, R. E.; Wingad, B. D.; Mueller, G. A. Dependence of amino acid side chain 13C shifts on dihedral angle: Application to conformational analysis. *J. Am. Chem. Soc.* **2008**, *130*, 11097−11105.
(47) Hansen, D. F.; Neudecker, P.; Kay, L. E. Determination of isoleucine side-chain conformations in ground and excited states of proteins from chemical shifts. *J. Am. Chem. Soc.* **2010**, *132*, 7589−7591.
(48) Cavalli, A.; Salvatella, X.; Dobson, C. M.; Vendruscolo, M. Protein structure determination from NMR chemical shifts. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 9615−9620.
(49) Mulder, F. A. Leucine Side-Chain Conformation and Dynamics in Proteins from 13C NMR Chemical Shifts. *ChemBiochem* **2009**, *10*, 1477−1479.
(50) Brandl, M.; Weiss, M. S.; Jabs, A.; Suhnel, J.; Hilgenfeld, R. CH-π-interactions in proteins. *J. Mol. Biol.* **2001**, *307*, 357−377.
(51) Platzer, G.; Mayer, M.; Beier, A.; Brüschweiler, S.; Fuchs, J. E.; Engelhardt, H.; Geist, L.; Bader, G.; Schörghuber, J.; Lichtenecker, R. PI by NMR: Probing CH−π interactions in protein−ligand complexes by NMR spectroscopy. *Angew. Chem.* **2020**, *132*, 14971−14978.
(52) Scheiner, S. Assessment of the presence and strength of H-bonds by means of corrected NMR. *Molecules* **2016**, *21*, 1426.
(53) Carter-Fenk, K.; Liu, M.; Pujal, L.; Loipersberger, M.; Tsanai, M.; Vernon, R. M.; Forman-Kay, J. D.; Head-Gordon, M.; Heidar-Zadeh, F.; Head-Gordon, T. The Energetic Origins of Pi-Pi Contacts in Proteins. *J. Am. Chem. Soc.* **2023**, *145*, 24836−24851.
(54) Höfurthner, T.; Toscano, G.; Kontaxis, G.; Beier, A.; Mayer, M.; Geist, L.; McConnell, D. B.; Weinstabl, H.; Lichtenecker, R.; Konrat, R. Synthesis of a 13C-methylene-labeled isoleucine precursor as a useful tool for studying protein side-chain interactions and dynamics. *J. Biomol. NMR* **2024**, *78*, 1−8.
(55) Toscano, G.; Höfurthner, T.; Nagl, B.; Beier, A.; Mayer, M.; Geist, L.; McConnell, D. B.; Weinstabl, H.; Konrat, R.; Lichtenecker, R. J. $^{13}C\beta$-Valine and $^{13}C\gamma$-Leucine Methine Labeling To Probe Protein Ligand Interaction. *ChemBiochem* **2024**, *25*, No. e202300762.
(56) Kumar, M.; Balaji, P. V. C-H...pi interactions in proteins: Prevalence, pattern of occurrence, residue propensities, location, and contribution to protein stability. *J. Mol. Model.* **2014**, *20*, 2136.
(57) Derewenda, Z. S. CH groups as donors in hydrogen bonds: A historical overview and occurrence in proteins and nucleic acids. *Int. J. Mol. Sci.* **2023**, *24*, 13165.
(58) Hass, M. A.; Hansen, D. F.; Christensen, H. E.; Led, J. J.; Kay, L. E. Characterization of conformational exchange of a histidine side chain: Protonation, rotamerization, and tautomerization of His61 in plastocyanin from Anabaena variabilis. *J. Am. Chem. Soc.* **2008**, *130*, 8460−8470.
(59) Liepinsh, E.; Otting, G. Proton exchange rates from amino acid side chains−implications for image contrast. *Magn. Reson. Med.* **1996**, *35*, 30−42.
(60) Kougentakis, C. M.; Grasso, E. M.; Robinson, A. C.; Caro, J. A.; Schlessman, J. L.; Majumdar, A.; García-Moreno, E. B. Anomalous properties of Lys residues buried in the hydrophobic interior of a protein revealed with 15N-detect NMR spectroscopy. *J. Phys. Chem. Lett.* **2018**, *9*, 383−387.
(61) Iwahara, J.; Jung, Y.-S.; Clore, G. M. Heteronuclear NMR spectroscopy for lysine NH3 groups in proteins: Unique effect of water exchange on 15N transverse relaxation. *J. Am. Chem. Soc.* **2007**, *129*, 2971−2980.
(62) Nguyen, D.; Chen, C.; Pettitt, B. M.; Iwahara, J. NMR methods for characterizing the basic side chains of proteins: Electrostatic interactions, hydrogen bonds, and conformational dynamics. *Methods Enzymol.* **2019**, *615*, 285−332.
(63) Henry, G. D.; Sykes, B. D. Determination of the rotational dynamics and pH dependence of the hydrogen exchange rates of the arginine guanidino group using NMR spectroscopy. *J. Biomol. NMR* **1995**, *6*, 59−66.
(64) Takeda, M.; Jee, J.; Ono, A. M.; Terauchi, T.; Kainosho, M. Hydrogen exchange study on the hydroxyl groups of serine and threonine residues in proteins and structure refinement using NOE restraints with polar side-chain groups. *J. Am. Chem. Soc.* **2011**, *133*, 17420−17427.
(65) Chen, J.; Yadav, N. N.; Stait-Gardner, T.; Gupta, A.; Price, W. S.; Zheng, G. Thiol-water proton exchange of glutathione, cysteine, and N-acetylcysteine: Implications for CEST MRI. *Nmr Biomed.* **2020**, *33*, No. e4188.
(66) Takeda, M.; Jee, J.; Terauchi, T.; Kainosho, M. Detection of the sulfhydryl groups in proteins with slow hydrogen exchange rates and determination of their proton/deuteron fractionation factors using the deuterium-induced effects on the 13Cβ NMR signals. *J. Am. Chem. Soc.* **2010**, *132*, 6254−6260.
(67) Nordstrand, K.; Åslund, F.; Meunier, S.; Holmgren, A.; Otting, G.; Berndt, K. D. Direct NMR observation of the Cys-14 thiol proton of reduced *Escherichia coli* glutaredoxin-3 supports the presence of an active site thiol-thiolate hydrogen bond. *FEBS Lett.* **1999**, *449*, 196−200.
(68) Davis, D. G. Proton NMR detection of long-range heteronuclear multiquantum coherences in proteins: The complete assignment of the quaternary aromatic carbon-13 chemical shifts in lysozyme. *J. Am. Chem. Soc.* **1989**, *111*, 5466−5468.
(69) Weichenberger, C. X.; Sippl, M. J. NQ-Flipper: Recognition and correction of erroneous asparagine and glutamine side-chain rotamers in protein structures. *Nucleic Acids Res.* **2007**, *35*, W403−W406.
(70) Jumper, J.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583−589.
(71) Agarwal, V.; McShan, A. C. The power and pitfalls of AlphaFold2 for structure prediction beyond rigid globular proteins. *Nat. Chem. Biol.* **2024**, *20*, 950−959.
(72) Klukowski, P.; Riek, R.; Güntert, P. Rapid protein assignments and structures from raw NMR spectra with the deep learning technique ARTINA. *Nat. Commun.* **2022**, *13*, 6151.
(73) Klukowski, P.; Riek, R.; Güntert, P. NMRtist: An online platform for automated biomolecular NMR spectra analysis. *Bioinformatics* **2023**, *39*, btad066.
(74) Klukowski, P.; Riek, R.; Güntert, P. Time-optimized protein NMR assignment with an integrative deep learning approach using AlphaFold and chemical shift prediction. *Sci. Adv.* **2023**, *9*, No. eadi9323.
(75) Sekhar, A.; Kay, L. E. An NMR view of protein dynamics in health and disease. *Annu. Rev. Biophys.* **2019**, *48*, 297−319.
(76) Berjanskii, M. V.; Wishart, D. S. Application of the random coil index to studying protein flexibility. *J. Biomol. NMR* **2008**, *40*, 31−48.

https://doi.org/10.1021/jacs.4c10474
J. Am. Chem. Soc. 2024, 146, 31733−31745

(77) Fowler, N. J.; Sljoka, A.; Williamson, M. P. A method for validating the accuracy of NMR protein structures. *Nat. Commun.* **2020**, *11*, 6321.
(78) Jacobs, D. J.; Rader, A. J.; Kuhn, L. A.; Thorpe, M. F. Protein flexibility predictions using graph theory. *Proteins: Struct., Funct., Bioinf.* **2001**, *44*, 150−165.
(79) Hansen, D. F.; Neudecker, P.; Vallurupalli, P.; Mulder, F. A.; Kay, L. E. Determination of Leu side-chain conformations in excited protein states by NMR relaxation dispersion. *J. Am. Chem. Soc.* **2010**, *132*, 42−43.
(80) Hansen, D. F.; Vallurupalli, P.; Kay, L. E. Using relaxation dispersion NMR spectroscopy to determine structures of excited, invisible protein states. *J. Biomol. NMR* **2008**, *41*, 113−120.
(81) Vallurupalli, P.; Hansen, D. F.; Kay, L. E. Structures of invisible, excited protein states by relaxation dispersion NMR spectroscopy. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 11766−11771.
(82) Korzhnev, D. M.; Religa, T. L.; Banachewicz, W.; Fersht, A. R.; Kay, L. E. A transient and low-populated protein-folding intermediate at atomic resolution. *Science* **2010**, *329*, 1312−1316.
(83) Platzer, G.; Ptaszek, A. L.; Böttcher, J.; Fuchs, J. E.; Geist, L.; Braun, D.; McConnell, D. B.; Konrat, R.; Sánchez-Murcia, P. A.; Mayer, M. Ligand 1H NMR Chemical Shifts as Accurate Reporters for Protein-Ligand Binding Interfaces in Solution. *ChemPhyschem* **2024**, *25*, No. e202300636.
(84) Beier, A.; Platzer, G.; Höfurthner, T.; Ptaszek, A. L.; Lichtenecker, R. J.; Geist, L.; Fuchs, J. E.; McConnell, D. B.; Mayer, M.; Konrat, R. Probing Protein−Ligand Methyl- π Interaction Geometries through Chemical Shift Measurements of Selectively Labeled Methyl Groups. *J. Med. Chem.* **2024**, *67*, 13187−13196.
(85) Kuhn, S. Applications of machine learning and artificial intelligence in NMR. *Magn. Reson. Chem.* **2022**, *60*, 1019−1020.
(86) Zhang, O.; Haghighatlari, M.; Li, J.; Liu, Z. H.; Namini, A.; Teixeira, J. M. C.; Forman-Kay, J. D.; Head-Gordon, T. Learning to evolve structural ensembles of unfolded and disordered proteins using experimental solution data. *J. Chem. Phys.* **2023**, *158*, 174113.
(87) Cortés, I.; Cuadrado, C.; Hernández Daranas, A.; Sarotti, A. M. Machine learning in computational NMR-aided structural elucidation. *Front. Nat. Prod.* **2023**, *2*, 1122426.

https://doi.org/10.1021/jacs.4c10474
J. Am. Chem. Soc. 2024, 146, 31733−31745

Supporting Information: UCBShift 2.0: Bridging the Gap from Backbone to Side Chain Protein Chemical Shift Prediction for Protein Structures

Aleksandra L. Ptaszek,[†,‡,¶] Jie Li,[§,¶] Robert Konrat,[†] Gerald Platzer,[‖,†] and Teresa Head-Gordon[*,§,⊥]

†*Christian Doppler Laboratory for High-Content Structural Biology and Biotechnology, Department of Structural and Computational Biology, Max Perutz Labs, University of Vienna, Campus Vienna Biocenter 5, 1030-Vienna, Austria*

‡*Laboratory of Computer-Aided Molecular Design, Division of Medicinal Chemistry, Otto-Loewi Research Center, Medical University of Graz, Neue Stiftintalstr. 6/III, A-8010 Graz, Austria*

¶*A.L.P. and J.L. contributed equally to this paper*

§ *Pitzer Center for Theoretical Chemistry, Department of Chemistry, University of California, Berkeley CA 94720, USA*

‖*MAG-LAB GmbH, Karl-Farkas-Gasse 22, 1030- Vienna, Austria*

⊥*Departments of Bioengineering and Chemical and Biomolecular Engineering, University of California, Berkeley, CA, 94720, USA*

E-mail: thg@berkeley.edu

Table S1: *The total number of chemical shifts in the training set for each type of atom and amino acid.*

Atom \ Amino acid	Ala	Asp	Glu	Phe	Ile	His	Lys	Leu	Met	Asn	Pro	Gln	Arg	Ser	Trp	Tyr	Thr	Val	Cys
CG	-	208	2875	41	-	32	2708	3430	665	316	1508	1516	1622	-	32	42	-	-	-
CD1	-	-	-	533	1391	-	-	2034	-	-	-	-	-	-	191	441	-	-	-
CG2	-	-	-	-	1391	-	-	-	-	-	-	-	-	-	-	-	1284	1625	-
CD	-	-	171	-	-	-	1468	-	-	-	765	150	913	-	-	-	-	-	-
CG1	-	-	-	-	1262	-	-	-	-	-	-	-	-	-	-	-	-	1657	-
CD2	-	-	-	320	-	226	-	1899	-	-	-	-	-	-	13	277	-	-	-
CE	-	-	-	-	-	-	1405	-	244	-	-	-	-	-	-	-	-	-	-
CE1	-	-	-	472	-	184	-	-	-	-	-	-	-	-	-	434	-	-	-
CE2	-	-	-	292	-	-	-	-	-	-	-	-	-	-	28	280	-	-	-
CZ	-	-	-	349	-	-	-	-	-	-	-	-	55	-	-	24	-	-	-
CZ2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	176	-	-	-	-
CH2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	167	-	-	-	-
CE3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	156	-	-	-	-
CZ3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	151	-	-	-	-
HB2	-	1772	2015	1141	-	510	1922	2252	491	1183	969	937	1190	1582	348	891	-	-	220
HB3	-	1672	1802	1090	-	492	1777	2159	446	1124	892	856	1092	1425	338	882	-	-	212
HG2	-	-	1837	-	947	-	1645	-	427	-	826	864	1012	-	-	-	956	1153	-
HB	1290	-	-	-	1574	-	-	-	-	-	-	-	-	-	-	-	1584	1997	-
HD2	-	-	-	622	-	336	1445	1345	-	-	834	-	981	-	-	541	-	-	-
HG3	-	-	1609	-	-	-	1479	-	397	-	752	759	905	-	-	-	-	-	-
HD1	-	-	-	849	939	35	-	1403	-	-	-	-	-	-	274	703	-	-	-
HD3	-	-	-	-	-	-	1226	-	-	-	813	-	872	-	-	-	-	-	-
HE2	-	-	-	591	-	23	1483	-	-	-	-	-	-	-	-	515	-	-	-
HG	-	-	-	-	-	-	-	2030	-	-	-	-	-	31	-	-	-	-	15
HE1	-	-	-	780	-	276	-	-	-	-	-	-	-	-	317	672	-	-	-
HE3	-	-	-	-	-	-	1188	-	-	-	-	-	-	-	221	-	-	-	-
HG12	-	-	-	-	1373	-	-	-	-	-	-	-	-	-	-	-	-	-	-
HG13	-	-	-	-	1239	-	-	-	-	-	-	-	-	-	-	-	-	-	-
HG1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	58	1176	-
HD21	-	-	-	-	-	-	-	-	-	722	-	-	-	-	-	-	-	-	-
HD22	-	-	-	-	-	-	-	-	-	720	-	-	-	-	-	-	-	-	-
HE	-	-	-	-	-	-	-	-	275	-	-	-	329	-	-	-	-	-	-
HE21	-	-	-	-	-	-	-	-	-	-	-	595	-	-	-	-	-	-	-
HE22	-	-	-	-	-	-	-	-	-	-	-	594	-	-	-	-	-	-	-
HZ	-	-	-	543	-	-	15	-	-	-	-	-	-	-	-	-	-	-	-
HZ2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	242	-	-	-	-
HH2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	222	-	-	-	-
HZ3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	211	-	-	-	-
ND2	-	-	-	-	-	-	-	-	-	672	-	-	-	-	-	-	-	-	-
NE2	-	-	-	-	-	-	-	-	-	-	-	562	-	-	-	-	-	-	-
NE1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	287	-	-	-	-

Table S2: *The total number of chemical shifts in the testing set for each type of atom and amino acid.*

Atom \ Amino acid	Ala	Asp	Glu	Phe	Ile	His	Lys	Leu	Met	Asn	Pro	Gln	Arg	Ser	Trp	Tyr	Thr	Val	Cys
CG	-	21	226	17	-	12	249	272	61	12	119	140	125	-	13	22	-	-	-
CD1	-	-	-	45	191	-	-	304	-	-	-	-	-	-	23	58	-	-	-
CG2	-	-	-	-	204	-	-	-	-	-	-	-	-	-	-	-	247	280	-
CD	-	-	17	-	-	-	225	-	-	-	122	27	124	-	-	-	-	-	-
CG1	-	-	-	-	167	-	-	-	-	-	-	-	-	-	-	-	-	305	-
CD2	-	-	-	20	-	31	-	280	-	-	-	-	-	-	6	30	-	-	-
CE	-	-	-	-	-	-	216	-	40	-	-	-	-	-	-	-	-	-	-
CE1	-	-	-	38	-	19	-	-	-	-	-	-	-	-	-	57	-	-	-
CE2	-	-	-	12	-	-	-	-	-	-	-	-	-	-	7	30	-	-	-
CZ	-	-	-	32	-	-	-	-	-	-	-	-	7	-	-	16	-	-	-
CZ2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	16	-	-	-	-
CH2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	13	-	-	-	-
CE3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	17	-	-	-	-
CZ3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	15	-	-	-	-
HB2	-	431	472	257	-	115	529	553	98	314	257	264	286	376	82	266	-	-	254
HB3	-	415	439	248	-	108	496	523	92	301	244	247	266	353	81	258	-	-	247
HG2	-	-	379	-	305	-	400	-	78	-	221	217	221	-	-	-	392	471	-
HB	536	-	-	-	357	-	-	-	-	-	-	-	-	-	-	-	419	508	-
HD2	-	-	-	164	-	75	350	469	-	-	246	-	220	-	-	181	-	-	-
HG3	-	-	361	-	-	-	380	-	73	-	205	195	202	-	-	-	-	-	-
HD1	-	-	-	186	286	3	-	488	-	-	-	-	-	-	66	205	-	-	-
HD3	-	-	-	-	-	-	317	-	-	-	243	-	198	-	-	-	-	-	-
HE2	-	-	-	152	-	1	317	-	-	-	-	-	-	-	-	175	-	-	-
HG	-	-	-	-	-	-	-	441	-	-	-	-	-	7	-	-	-	-	2
HE1	-	-	-	165	-	58	-	-	-	-	-	-	-	-	67	191	-	-	-
HE3	-	-	-	-	-	-	286	-	-	-	-	-	-	-	59	-	-	-	-
HG12	-	-	-	-	290	-	-	-	-	-	-	-	-	-	-	-	-	-	-
HG13	-	-	-	-	274	-	-	-	-	-	-	-	-	-	-	-	-	-	-
HG1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	15	478	-
HD21	-	-	-	-	-	-	-	-	-	184	-	-	-	-	-	-	-	-	-
HD22	-	-	-	-	-	-	-	-	-	185	-	-	-	-	-	-	-	-	-
HE	-	-	-	-	-	-	-	-	56	-	-	-	106	-	-	-	-	-	-
HE21	-	-	-	-	-	-	-	-	-	-	-	164	-	-	-	-	-	-	-
HE22	-	-	-	-	-	-	-	-	-	-	-	162	-	-	-	-	-	-	-
HZ	-	-	-	135	-	-	29	-	-	-	-	-	-	-	-	-	-	-	-
HZ2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	61	-	-	-	-
HH2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	59	-	-	-	-
HZ3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	56	-	-	-	-
ND2	-	-	-	-	-	-	-	-	-	120	-	-	-	-	-	-	-	-	-
NE2	-	-	-	-	-	-	-	-	-	-	-	114	-	-	-	-	-	-	-
NE1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	47	-	-	-	-

Table S3: *The PDB IDs and corresponding BMRB entries for 851 protein structures in the training set.* Chemical shift IDs in italics were obtained from the SPARTA+ and SHIFTX+ datasets, which include data from TALOS (Cornilescu et al. 1999) and other databases.

PDB ID	BMRB	PDB ID	BMRB	PDB ID	BMRB	PDB ID	BMRB	PDB ID	BMRB
1QL3	4471	1P9Y	15813	2FD6	4049	2OG3	6372	1BYQ	7003
1FGY	15669	3EGV	4965	1RW6	6236	2NYZ	5042	2VYI	5709
1Z2M	5658	1CKT	4079	2OTA	15317	1G49	15396	1XTQ	15048
2Z3Y	10011	2GRR	6305	2O2V	15332	1AGR	4386	1OK3	5506
1APM	6183	1M48	6621	2AL4	15496	1C4Z	15498	1MSP	4242
1CMX	4983	1E50	4092	1OPK	4251	2OZB	7249	1MU4	4972
1QGT	15969	2GOL	7250	1BBU	4134	2EKT	4061	1ZGY	15518
1JHE	6373	2R6G	15911	2Z6O	6546	2D2A	6224	2V6Q	5897
3CMM	4769	3ERR	11013	2P3K	5275	2W1T	11034	2RIQ	15602
2AXI	15945	1WRK	15388	1CDC	4109	2HIQ	6868	1YJM	10104
3C1V	4892	2NVH	1062	2AUG	5314	1ZYJ	6468	1RLW	4188
2FCW	6950	2F4M	5868	1JF4	4038	3DZD	6611	1M9F	5258
1QZM	5107	1XTM	6073	2CKX	15444	1KYF	6034	2OJ4	15178
1EY4	4052	1DN1	15646	1ZOQ	6874	1MH1	6970	1DTL	15388
1PJX	5618	3CX2	6015	2QFA	6342	2P4T	5237	1HY7	15120
1MPD	4987	1DFJ	16011	2BQQ	5482	2CYX	6711	1L8B	7115
1B79	4297	3BPY	4675	1XW3	6590	1J1D	15427	3BPO	4843
2H7W	6876	1ZCK	6080	1XF9	16393	3B5G	15276	1NMN	5270
2GMF	15531	2QQ9	4183	2OG0	5539	1Z40	6510	2HZ5	6210
2FUN	15932	2O6P	15913	1KCQ	4794	1COL	16523	1PRX	6148
2AXW	5946	1I27	5685	2BWF	15769	3CWI	15844	1I5Z	4388
1DIV	15884	2FD9	6455	1OMR	4030	2OOB	15111	1F5W	5516
1BYL	4786	2NMZ	5967	3BJ5	15303	3BR8	6884	2PT5	4722
1CZP	447	1QYO	5666	3BY8	4821	1GP0	7131	2Z3Q	6882
1FBV	15796	1I8D	4954	3DWY	6960	3BA0	15578	2NZ7	6843
3F17	6391	1R0X	16367	2FMP	4326	2J0E	5468	1AY9	5022
1LE8	4637	1NM5	4236	1N8L	15449	2ESK	6277	1O8X	5222
2PEH	15882	2F8K	6956	2I04	6911	1R8J	5031	2H3L	5631
1KNL	5772	1LCI	6646	1YRC	5759	1BEB	10010	1STF	15548
1W6N	15800	3BKZ	6685	2R0O	15808	1U8B	6605	1DXS	4413
2QWO	4716	2BF2	4560	2QYP	5465	1P1V	6821	2A1I	15240
2NSZ	6900	2F0R	5532	2BX6	6744	1KWI	5688	1Z6S	16010
1RZX	6126	2VZC	15760	1X10	10052	2GIB	15511	1IWL	10078
2BCQ	5766	1XAK	6824	3DLG	5347	1J40	5856	1LKK	5794
1TIP	5935	1ENH	15520	1YTC	5003	3DEO	6589	2R2O	6346
1BRF	5601	1CRC	5827	1NAY	16090	1JYR	11055	2EA9	15088
1WVH	6986	1SK7	6983	1JET	10053	3DXB	15885	2GIM	5505
1TSF	5805	1ZYM	4106	1NCO	5343	2IG0	5878	1DU0	15536
1I4M	4379	2DS5	5665	2A1J	6502	2MB5	16500	2P1T	15219
1VM9	4992	1WZD	15070	1M9Z	4779	3ERJ	6058	1ZLI	15730
2I0L	15209	1TQG	7132	2PE8	15882	2F1C	15426	1T6O	6567
3BEG	7301	1CMI	4305	2VBU	15530	1QAU	4304	1R6Z	5999
1L6M	4267	1Q6H	5863	2VNF	7210	2FKW	6348	1QMY	15278
2DCY	4704	1GRN	15424	2Q79	5952	1F9F	4998	1PUF	5349
2BOP	4087	1DT9	6116	1SW0	15064	1M8N	6111	1NMU	4339
1WM3	6801	3DWV	15597	2CUA	5819	2DPK	15764	2Q00	15265
2BKB	4315	2VQE	4369	2IQY	6783	1E87	15703	2EGD	6531
2OUO	15057	1DKI	5547	1LCK	4860	1QCF	4122	1SRR	15008
2DTM	6785	2V8F	15452	1N18	4202	1P42	5627	1IFR	5224

Table S3 – continued from previous page

PDB ID	BMRB	PDB ID	BMRB	PDB ID	BMRB	PDB ID	BMRB	PDB ID	BMRB
2FOJ	6939	1XU1	6384	2PQH	15013	1JBE	4083	1ERQ	6024
3D48	5599	1A1V	4885	1EV2	5943	1CEI	7188	1MIX	7061
1FPO	15541	2ZEQ	5496	1RA9	4554	3BRI	15077	1USE	5955
2HW4	6625	2FY6	15280	1G6L	4356	2V8L	7083	1P6O	6223
2OZN	7270	1LOP	4765	2A9I	10092	1JMC	5821	1V7P	5752
1BM8	4254	1S0P	5393	1WPG	5161	2PQI	15369	2E7P	6410
2O0Q	15281	2ZJD	15877	1R8S	5368	1ZLH	15372	2A4D	7219
1OR3	15744	1Y9L	15504	1RKE	15653	2E3W	4032	2D1Z	5679
2FWE	5497	1HYU	15264	2O5G	4056	3BWU	15032	1MD0	5991
1MK7	15792	1S69	5269	1MEE	15706	3DAW	7172	1O08	7234
2AO3	6226	1N26	5940	2GKR	5472	8PAZ	6043	2VN1	15038
3F6Q	4884	1ATZ	5456	2CWR	6829	1KS9	15643	2YXJ	6578
1OVL	16541	2OFP	15643	2ES7	7178	2QJL	7165	1ZYN	16265
1YTY	15727	1CSG	15531	1IEE	4562	1YTB	6700	1THF	15741
1TOL	4771	1BR9	4214	2RGT	6658	2ZD1	4097	1GNS	15250
1GY6	5887	2DTQ	15514	1PLQ	7240	1ZUD	4470	1EM9	4384
1A17	6696	1KF3	4032	1TP9	6132	2IN0	15560	1W41	5485
1GZI	15817	4FGF	4091	1SN8	6122	2A38	5760	1U8T	4472
2ITL	4127	1UJ8	6776	1T8L	5358	2FCL	7086	1YSB	6223
2OYN	15530	1EXP	7264	1G4C	4299	2DFK	15424	1EPF	4162
1ZX8	16007	2CJR	15511	1UB4	6833	2C5L	6635	1SNC	4052
1F46	4717	1DYT	15757	1DQE	6313	1NQD	15722	1MV9	15219
1T15	6114	2CDN	4840	1TP5	6193	1Y2G	4717	1KJL	4909
1L2H	1062	1JTG	6357	1O4D	6604	2DYI	10139	2NNR	6876
2F1Y	6939	2GFE	5182	1SNM	4053	2GRO	4132	2O0P	15281
2BE6	4174	1B1H	10053	1RUW	6197	4ICB	6699	1FH9	7264
1IWT	5142	1RRO	7322	1RWY	15517	1SKM	16395	1HQ2	4300
1BT5	6024	1H4G	5352	1IWM	10096	1RMM	5666	2EWR	7086
2TRX	1812	1RGE	4259	1UDR	4083	2CIA	6575	2AOJ	5967
1UV0	6231	1TVG	6344	1P7T	5471	1A2P	4964	1CEX	4101
1LM4	4834	1VFQ	6283	2ESP	6277	1YZ1	15243	1YKT	5359
1UBQ	5387	1C7F	5571	1CLL	6023	1FE4	6266	1T2W	4879
1NW2	5241	1VP6	15249	1MXE	1634	1MN8	5623	2ADF	5456
1U7B	15501	1H4A	16173	1CWC	2208	1N2D	6332	2F3Y	15852
1BKF	4077	1SYD	15232	2FKE	4077	1HL5	4202	1HFC	4064
1HUU	4047	1HPC	4336	2HZE	4113	1TW4	15854	2H30	6709
2B02	6597	1O13	6198	1KD	6250	256B	6560	1RX2	5740
1VDQ	4562	1W80	6034	2F5M	16171	1TOP	4401	1YJ7	6252
2A0N	15741	3EZM	cvn	1ZW9	5355	1ICM	7356	1EB0	5826
1Y93	6391	2Z2I	7055	1HCB	4022	1IIB	4955	1OSP	4076
1SMX	6122	1UCK	4259	1BRI	4964	1JR2	7242	1SCJ	15706
1A6K	4061	2ES9	15089	1FF3	6786	1XMT	6338	2BF5	4560
1V9T	4037	1I1J	5220	1F94	5097	1CM2	2371	2GAB	5508
1OQR	4149	1CYN	15505	1ODV	6321	1SGZ	6016	2B8	4094
1FGZ	15669	1LAX	4354	1ZE3	6779	2ALE	15412	1B2V	5081
2D3G	6457	1F5R	4968	1KQR	5275	3LZT	4562	1TVQ	16310
1YLA	17362	1T3Y	6032	1JL3	6075	1YKY	4831	1ZJL	7114
1YP7	4340	2GBT	15713	1JIW	6292	1IV7	5529	5PTI	bpti
1HKA	4299	1FIL	4082	1L3K	4084	1BFG	4091	1A7T	4102
1EMV	4115	1U9B	4132	1JV4	4340	1FQA	4354	1G8I	4378
1GXQ	4421	1C44	4438	1XUO	4553	1FZY	4567	1M5E	5182
1TJM	5194	1H7M	5485	1F4P	5571	1IPB	5712	1N0S	5756

Table S3 – continued from previous page

PDB ID	BMRB	PDB ID	BMRB	PDB ID	BMRB	PDB ID	BMRB	PDB ID	BMRB
1IU1	5761	1EW4	5792	1J97	5799	1Q4R	5843	1UOH	5898
1VC1	5921	1H70	6074	1J54	6184	1M45	6332	1U07	6375
1ZLQ	6416	1DBF	6494	1F2F	6503	1QAV	6754	2D3D	6922
2AWG	6923	1KBL	6932	2D58	6980	1NEY	7216	3RN3	4031
1DDW	4766	1AYF	4566	1CDL	1634	1CBS	4186	1MJC	4296
2END	5244	1EKG	4342	1GNU	5058	2A0B	4857	1AM1	5355
1ZNB	4102	1AIL	4317	1PLC	4019	1MHO	5206	1D4T	5211
1EZ3	4198	1H4H	5352	8ABP	6136	1BDO	4425	1NCX	5738
1RSY	4039	1RUV	4031	1LZ1	5142	1UP1	4084	1CNR	6504
2VFX	5299	1Y9T	15503	1YPC	4974	1ONC	4371	1BCX	4705
1HST	30	1CHO	5469	2PLH	55	1MOX	246	1B0C	264
1WEJ	274	2HXK	283	2IFQ	284	1EHB	294	3ICB	15594
1IGV	327	1TPK	349	1RPG	385	1KF8	443	1STA	496
1FKH	16931	1NOA	1766	1P6Z	17013	2BUO	16555	1DPF	16669
1CYO	1208	1IVO	1377	1JL9	1378	1F93	1414	2ETL	17260
1D66	1517	1ET1	3427	5CRO	1743	2SNI	1869	1VCX	1991
1IAR	2031	1JS2	2219	1XLQ	2278	1Y5K	16898	3SXL	4028
1A7G	4035	2I9A	4049	1JRL	4060	1PPF	5519	2GHY	4072
1VAP	4078	1BI7	4086	1IO4	4092	1HRH	5931	1LFO	15434
4AKE	5720	2IRF	4161	1G05	4173	2CBS	4186	1E4V	5746
1BQV	16426	1W0T	4210	1C76	4215	1U2G	4236	2BVU	4242
1X8U	4267	1SFC	4272	1I5K	4276	1EGJ	4308	1DK0	4309
1YCE	4316	1QST	4321	1GGQ	4323	1UEA	5099	1T0K	4339
1AKE	4350	4MBP	4987	1B72	4358	2CM6	4360	1SLM	15120
1XMQ	4369	1EDH	4380	1ZRF	4388	1DFU	4395	1F2L	4397
1QMR	4417	1OB1	4437	1QA9	4455	1JVO	4465	1ADW	4466
1XDG	4553	1K46	4961	1QIB	4565	1DHN	4573	1QD7	4577
1JOC	4579	1QOJ	4622	1F3V	4636	1AKH	4637	1XCT	4669
2J03	4688	2EB9	4695	1BM9	4718	2UUB	4727	1F35	4735
1GO4	4775	1BYF	4782	1AEP	4814	2C6Y	4829	1MVP	4839
1X8R	4848	2FJY	4849	1QQY	4876	1YOB	4886	2CWI	4887
1J8R	4897	1QJ8	4936	1R5R	4940	1BJA	4957	1H0A	4959
2NXN	7315	1MO1	4972	2CH4	4984	1F80	4989	1SCE	5008
1UMU	5022	1NF0	16565	2HTX	6415	2B5Z	5069	1B56	7107
1KJT	5128	1S3S	5155	2VPF	5186	1DD5	5191	1ZQW	5208
1JN3	5209	2IIM	5223	2O3B	5261	5PNT	5350	1HUR	5368
2HDZ	5392	1TJF	5393	1XEO	6142	1EH1	5458	1G6H	5462
2GTG	6158	2OVO	5470	1G7F	5474	1RGH	5492	2B9A	5507
1E96	6970	2J1K	5516	1CRB	5579	2BAY	5594	1N9D	6643
1XM0	17008	1VSA	5650	1MVK	5654	2D5V	5678	1GRI	5693
1ND4	5721	1RX4	5741	1XZO	5742	2BX9	16492	1A6J	5789
1X27	5794	2DQE	5803	1V2Z	5825	2FFM	5844	1Q6U	5863
1SKY	5886	1R7J	5891	2BKY	5892	1PEY	5899	1BF4	5910
1RIL	5918	1K6M	5935	1E0R	5936	1BY9	6877	1UGM	5958
1ZND	5996	1TN3	6008	1MMQ	6014	1ZH5	6044	1NG2	6057
1XM2	15949	1UUH	6093	1UW2	6097	2IDO	6127	1VYF	6150
1OC0	6160	2CLY	6212	2HGU	6255	2J4T	6279	1T3W	6284
2I32	6298	1MFT	6302	2IO2	6306	1UTX	6317	3PYP	6322
1FD9	15507	1E31	6342	1TTZ	6363	1TPX	6403	1HH8	6399
1SZ9	6404	1B88	6406	1UHI	6418	1MZN	6429	1SYR	6442
1K3E	6451	2PKT	6452	2FSO	6468	1COM	6494	1QZ8	6501
1Y62	6506	2AHM	16981	2H2K	6532	2IN1	6546	1ZEO	6549

Table S3 – continued from previous page

PDB ID	BMRB	PDB ID	BMRB	PDB ID	BMRB	PDB ID	BMRB	PDB ID	BMRB
1VJH	6585	2A01	6588	1PEX	6617	2NMM	6625	2F69	6628
1VDR	15253	1GAW	6695	1U2P	6722	2BP3	6730	2H2R	6735
2F1N	6758	4CRO	6771	2CHD	6787	2D07	6801	1EZ9	6807
2ASW	6822	2FX5	6832	1G6A	6838	1FO1	6853	2ION	6900
2FA4	6913	2AV5	6917	1M1F	6925	2FJJ	6930	1IHO	6940
1EJF	6973	1U7E	6984	2UWD	7003	2GHP	7070	2PST	7075
108M	7076	2EB8	16501	1PF9	7091	1IPC	7115	1E0C	7130
2LZH	7136	1QFJ	7138	2F30	7159	1KMV	7195	2JMQ	7210
1SAI	7251	2QGV	7256	2IHB	7272	1MMS	7308	1TV9	7319
1IFC	15082	1BED	7360	1R1U	15177	1NIW	7424	2F3Z	7425
2CZT	10137	2CZU	11062	1U4E	11067	1IW0	15073	2PG1	16847
1J7D	15092	2OOA	15111	1XD3	15121	1A43	15137	1FCC	15156
2A5D	15162	2C7P	15170	1L6X	15204	1D8C	*malate*	1Q5P	*maxacal*
1REC	*recoverin*	1FHM	*rubredoxin*	1HIW	*gag*	1SVN	*maxacal*	1PUC	*suc1*
11REC	*recoverin*	1FHM	*rubredoxin*	1HIW	*gag*	1SVN	*maxacal*	1PUC	*suc1*
1RCB	*interleukin4*	1KZN	*gyraseB*	1ACF	*profilin*	2CPL	*cyclophilin*	1BPB	*dna polym B*
1QA7	*hav*	1G6S	*epsps complexed*	2ALP	*alpha LP*	1MKA	*dehydrase*	1ERT	*thioredoxin red*
2RNT	*Ribonouclease T1*	1F3G	*IIIglc*	1HUL	*interleukin5*	1HUF	*YopH*	2CI2	*ci2*
1HTP	*Hmet*	1QRX	*alpha LP*	1A30	*HIVprotease*	1BNJ	*barnase*	1AVS	*sNTnC*
1EYH	*epsin NTH*	4I1B	*interleukin 1beta*	1A3K	*crd galectin 3*	1JF8	*arsenate reduct*	1DMB	*mbp*

Table S4: The PDB IDs and corresponding BMRB entries for 200 protein structures in the testing set.

PDB ID	BMRB	PDB ID	BMRB	PDB ID	BMRB	PDB ID	BMRB	PDB ID	BMRB
1GN0	17136	1T85	17415	2AFG	15783	2Q2T	16059	1QG7	16142
2NWG	16143	1GWY	16362	1YU8	15245	1TPH	15064	2HWN	4473
1PGX	2575	2UYZ	4132	1JPS	16838	2IN8	16634	2PSP	2384
1QE6	280	2IGD	15283	2GIT	5784	1CY5	4661	1W7Z	397
1DSB	16327	1JHF	6373	451C	10133	1GOA	4012	5NUC	5536
1KBA	1675	2ALG	16294	2SEM	5729	1SNO	1878	1TXX	1813
2HPW	16600	1AR0	5888	2ASD	16869	1VYK	17357	1AM7	16664
2HPR	932	2GM5	4269	1O82	4112	1OVH	915	7RXN	15374
1QVE	4918	1C8C	4570	1R1T	4306	1CKU	2999	2BDY	16940
1XIO	17064	2H9H	4836	1WYW	6304	1GZZ	4204	1DS3	1374
1OKH	4587	2GOO	15956	2DGC	1397	1EZG	5323	1IOZ	10051
1YNR	10135	5HPG	16466	2ILN	5617	2GYK	4115	1NAQ	15094
1LP1	1120	1TUK	4977	1L0S	5573	1MID	15143	1SKO	6181
2OW9	4679	3ERA	7211	7TIM	16566	2BJD	6398	1VKB	16380
1ASS	5930	2A3G	1442	2F91	5274	1OS3	1633	1CLV	4490
1V6P	1381	1OMY	7330	2HSH	257	1HSL	16205	1P2O	45
1M8A	15596	1UWX	1639	1J3F	4568	1Z7X	4370	2IDU	16741
1BNE	16170	2FFG	15529	2GSV	15350	1AG6	79	1BD9	17382
1OMS	15813	2GGM	6687	1HFZ	4811	1FU0	4774	1L1D	6051
2H61	4105	1KTZ	4411	1LU0	4246	1F2M	1875	2FI4	4877
1CM9	4852	1YU7	4428	1RPJ	16984	2OCT	15548	1RDG	5163
2HZI	15488	1PZ4	16662	1WTQ	5908	1MYW	15826	1Q2U	17507
1EK8	5190	2GT8	17251	1SPQ	15066	109M	5158	2GCN	16668
1EQT	16803	1QKR	15653	1DCD	5249	1NBP	6621	1BRJ	975
1H0J	4966	2GJ2	7099	1FD3	4642	2IM8	16113	2BIU	6141
1YW5	16690	1LJU	4944	1VB0	16060	1HC9	15130	2UV0	6271
1R69	2539	2EYO	16585	2B59	5267	1D03	1673	2PF5	6392
1DFN	16254	2SGD	1375	1MOL	4633	1RB4	371	1K82	5219
1AE3	2039	1BNZ	6050	1OBO	5011	1AJ6	5218	2BEM	17160
2IE2	17162	1IP2	5125	1KDJ	7370	1XRK	4786	1IKO	7220
2GP0	15680	1ZNI	1585	1FDQ	16049	1U06	7306	1LW6	4974
2AAO	5324	1BWO	4383	1HB8	2049	1RN1	1658	1MI4	4854
1YCQ	6248	2H76	62	1ICF	5583	1BGF	5997	1J1V	5200
2IDS	16740	1TCF	1553	1PCS	5475	1RCF	16593	2H9E	4396
1IV9	4222	1GPR	1663	1MR3	15626	1QOG	447	1WKX	6123
1B68	6524	1ENF	16146	1PHT	16448	1P7J	7386	1ZDN	17437
1KX9	5094	1WZV	16321	1N3Z	4980	1CQM	16344	3WRP	17010
1PPE	2527	1BAZ	395	2CA5	15214	2CG7	15756	1THQ	6234
1POH	29	1OKS	6569	1YJF	5514	1PHP	16464	2BC5	1672
1OFF	16024	1O5U	16006	3DFR	7200				

Table S5: *Root mean square errors (RMSE) and mean absolute errors (MAE) of H, C and N chemical shifts prediction using UCBShift and ShiftX2 distinguishing between the individual amino acids.* In addition, the standard deviations of the test data (SD) are reported.

Atom type	Resid. name	UCBShift		ShiftX2		
		RMSE	**MAE**	**RMSE**	**MAE**	**SD**
CG	Asp	1.08	0.8	1.46	1.06	1.38
	Glu	0.99	0.61	1.35	0.9	1.36
	Phe	1.73	1.38	2.12	1.7	1.86
	His	2.41	1.97	3.47	3.21	2.67
	Lys	0.97	0.62	1.27	0.81	1.35
	Leu	0.94	0.57	1.32	0.81	1.35
	Met	0.79	0.48	1.26	0.88	1.17
	Asn	0.6	0.45	0.64	0.45	0.84
	Pro	0.79	0.54	0.95	0.7	1.08
	Gln	1.26	0.67	1.83	1.03	1.51
	Arg	0.94	0.59	1.37	0.88	1.46
	Trp	1.57	0.96	1.81	1.33	1.66
	Tyr	1.4	1.25	2.42	2.05	1.18
CD1	Phe	0.96	0.56	1.34	0.86	1.56
	Ile	1.0	0.66	1.26	0.91	1.62
	Leu	1.21	0.79	1.67	1.24	1.72
	Trp	1.77	0.99	2.45	1.54	2.7
	Tyr	0.82	0.52	1.23	0.82	1.0
CG2	Ile	0.82	0.54	1.16	0.81	1.23
	Thr	1.04	0.68	1.23	0.84	1.26
	Val	1.09	0.76	1.47	1.06	1.56
CD	Glu	2.22	1.37	2.54	1.5	2.64
	Lys	1.23	0.75	1.5	0.88	1.5
	Pro	1.08	0.75	1.28	0.9	1.26
	Gln	1.92	1.26	2.21	1.6	2.21
	Arg	0.83	0.53	1.01	0.69	1.06
CG1	Ile	1.14	0.69	1.32	0.87	1.46
	Val	1.07	0.73	1.4	1.0	1.55
CD2	Phe	1.11	0.8	1.83	1.2	1.89
	His	0.96	0.67	2.11	1.34	2.15
	Leu	1.26	0.88	1.66	1.26	1.81
	Trp	0.84	0.65	1.4	1.2	1.05
	Tyr	1.02	0.76	1.52	1.09	1.17
CE	Lys	0.94	0.58	1.01	0.65	0.99
	Met	0.83	0.59	1.13	0.79	1.11
CE1	Phe	0.8	0.54	1.21	0.81	1.06
	His	1.41	1.01	2.0	1.45	1.87
	Tyr	0.81	0.58	1.09	0.76	0.91
CE2	Phe	1.1	0.93	1.2	0.93	0.86
	Trp	0.75	0.54	1.2	0.93	0.58
	Tyr	1.0	0.74	1.16	0.89	0.88

Continued on next page

Atom type	Resid. name	UCBShift RMSE	UCBShift MAE	ShiftX2 RMSE	ShiftX2 MAE	SD
CZ	Phe	0.86	0.58	1.17	0.81	1.16
	Arg	1.46	0.99	1.36	0.87	0.95
	Tyr	2.42	1.74	4.1	3.65	1.95
CZ2	Trp	1.0	0.73	-	-	1.28
CH2	Trp	1.58	1.02	-	-	1.16
CE3	Trp	1.32	1.08	-	-	1.46
CZ3	Trp	1.91	1.11	-	-	1.83
HB2	Cys	0.25	0.14	0.34	0.25	0.44
	Asp	0.2	0.12	0.25	0.15	0.28
	Glu	0.17	0.09	0.18	0.1	0.23
	Phe	0.24	0.14	0.31	0.2	0.41
	His	0.3	0.15	0.3	0.18	0.46
	Lys	0.16	0.09	0.21	0.1	0.28
	Leu	0.22	0.13	0.28	0.17	0.39
	Met	0.22	0.13	0.25	0.16	0.32
	Asn	0.29	0.15	0.31	0.18	0.39
	Pro	0.27	0.17	0.29	0.19	0.37
	Gln	0.17	0.1	0.2	0.1	0.29
	Arg	0.23	0.11	0.27	0.12	0.36
	Ser	0.26	0.12	0.3	0.14	0.36
	Trp	0.23	0.13	0.26	0.15	0.33
	Tyr	0.26	0.15	0.32	0.19	0.39
HB3	Cys	0.3	0.2	0.39	0.29	0.42
	Asp	0.2	0.11	0.26	0.15	0.3
	Glu	0.17	0.09	0.2	0.1	0.24
	Phe	0.22	0.13	0.3	0.21	0.37
	His	0.27	0.17	0.35	0.25	0.36
	Lys	0.14	0.09	0.18	0.1	0.26
	Leu	0.24	0.15	0.33	0.22	0.39
	Met	0.21	0.13	0.25	0.16	0.34
	Asn	0.28	0.14	0.38	0.2	0.43
	Pro	0.3	0.18	0.33	0.22	0.41
	Gln	0.19	0.11	0.24	0.14	0.32
	Arg	0.22	0.12	0.26	0.15	0.34
	Ser	0.24	0.12	0.28	0.15	0.35
	Trp	0.32	0.17	0.4	0.27	0.47
	Tyr	0.28	0.16	0.35	0.23	0.41
HG2	Glu	0.16	0.09	0.2	0.12	0.25
	Ile	0.16	0.09	0.21	0.12	0.3
	Lys	0.18	0.1	0.23	0.13	0.33
	Met	0.25	0.15	0.29	0.2	0.38
	Pro	0.26	0.14	0.31	0.17	0.39
	Gln	0.18	0.09	0.2	0.11	0.27
	Arg	0.18	0.1	0.22	0.14	0.27
	Thr	0.16	0.07	0.2	0.12	0.27
	Val	0.17	0.09	0.2	0.12	0.29

Continued on next page

Atom type	Resid. name	UCBShift RMSE	UCBShift MAE	ShiftX2 RMSE	ShiftX2 MAE	SD
HB	Ala	0.12	0.07	0.16	0.08	0.29
	Ile	0.18	0.1	0.24	0.13	0.37
	Thr	0.18	0.1	0.23	0.14	0.33
	Val	0.19	0.1	0.23	0.12	0.36
HD2	Phe	0.15	0.09	0.21	0.13	0.31
	His	0.38	0.2	0.49	0.31	0.54
	Lys	0.22	0.11	0.22	0.11	0.3
	Leu	0.18	0.09	0.2	0.11	0.3
	Pro	0.26	0.16	0.33	0.19	0.41
	Arg	0.24	0.1	0.29	0.13	0.34
	Tyr	0.17	0.09	0.26	0.19	0.29
HG3	Glu	0.18	0.1	0.21	0.12	0.27
	Lys	0.17	0.11	0.22	0.13	0.31
	Met	0.24	0.16	0.31	0.2	0.42
	Pro	0.27	0.15	0.34	0.19	0.41
	Gln	0.22	0.1	0.23	0.12	0.31
	Arg	0.21	0.12	0.23	0.14	0.2
HD1	Phe	0.15	0.09	0.23	0.15	0.3
	His	1.39	1.14	-	-	0.57
	Ile	0.23	0.1	0.25	0.12	0.36
	Leu	0.19	0.1	0.24	0.12	0.31
	Trp	0.17	0.11	0.34	0.25	0.37
	Tyr	0.21	0.11	0.23	0.15	0.29
HD3	Lys	0.22	0.12	0.23	0.12	0.3
	Pro	0.35	0.2	0.4	0.23	0.48
	Arg	0.28	0.11	0.32	0.13	0.37
HE2	Phe	0.19	0.11	0.23	0.15	0.31
	His	0.41	0.41	-	-	-
	Lys	0.18	0.09	0.24	0.11	0.3
	Tyr	0.12	0.08	0.18	0.14	0.2
HG	Cys	0.6	0.5	-	-	0.73
	Leu	0.25	0.14	0.26	0.15	0.35
	Ser	0.27	0.18	-	-	0.38
HE1	Phe	0.18	0.11	0.21	0.12	0.33
	His	0.41	0.27	0.5	0.32	0.62
	Trp	0.49	0.27	0.58	0.35	0.65
	Tyr	0.13	0.08	0.17	0.08	0.23
HE3	Lys	0.15	0.09	0.22	0.1	0.27
	Trp	0.24	0.16	0.37	0.27	0.41
HG12	Ile	0.28	0.2	0.33	0.23	0.4
HG13	Ile	0.29	0.2	0.36	0.25	0.44
HG1	Thr	1.3	0.89	-	-	0.95
	Val	0.19	0.1	0.21	0.12	0.29
HD21	Asn	0.40	0.26	0.50	0.30	0.48
HD22	Asn	0.32	0.21	0.36	0.24	0.34

Continued on next page

Atom type	Resid. name	UCBShift RMSE	UCBShift MAE	ShiftX2 RMSE	ShiftX2 MAE	SD
HE	Met	0.23	0.13	0.27	0.17	0.33
	Arg	0.54	0.26	0.62	0.29	0.74
HE21	Gln	0.41	0.23	0.43	0.22	0.44
HE22	Gln	0.26	0.17	0.29	0.18	0.29
HZ	Phe	0.21	0.14	0.25	0.15	0.36
	Lys	0.7	0.3	0.8	0.32	0.9
HZ2	Trp	0.24	0.16	-	-	0.32
HH2	Trp	0.22	0.17	-	-	0.28
HZ3	Trp	0.37	0.25	-	-	0.44
ND2	Asn	1.78	1.02	-	-	2.8
NE2	Gln	1.71	0.88	-	-	2.43
NE1	Trp	1.47	1.07	-	-	2.21

Table S6: *RMSE and MAE for ShiftX+ module of ShiftX2, UCBShift-X, UCBShift-Y and UCBShift R2 random forest regressor.* Numbers in brackets show the ± uncertainty on the last digit. ShiftX+ and UCBShift-X are evaluated with a subset of the test data for which there is no homology. UCBShift-Y and R2 Regressor are evaluated with a subset of the test data for which there is only proteins that have homology.

	ShiftX+		**UCBShift-X (R1)**		**UCBShift-Y**		**R2 Regressor**	
Atom	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
CG	1.25(1)	0.830(5)	1.17(1)	0.78(1)	2.00(4)	0.605(4)	0.998(3)	0.604(1)
CD1	1.62(2)	1.19(1)	1.63(1)	1.21(1)	0.842(5)	0.467(2)	0.991(4)	0.618(2)
CG2	1.18(1)	0.915(5)	1.30(1)	0.99(1)	0.915(3)	0.546(2)	0.930(2)	0.606(2)
CD	1.52(4)	0.94(1)	1.48(3)	1.03(1)	0.950(2)	0.571(2)	1.139(3)	0.680(3)
CG1	1.40(2)	1.05(1)	1.44(2)	1.08(1)	1.004(4)	0.546(2)	1.011(3)	0.641(2)
CD2	1.61(1)	1.27(1)	1.60(2)	1.31(1)	1.056(5)	0.641(3)	1.112(3)	0.760(3)
CE	0.76(1)	0.60(1)	0.77(1)	0.65(1)	0.893(4)	0.505(3)	0.953(4)	0.566(2)
CE1	1.17(4)	0.81(3)	0.66(1)	0.52(1)	0.879(6)	0.520(4)	0.95(1)	0.646(3)
CE2	-	-	-	-	0.988(5)	0.736(5)	0.998(4)	0.760(5)
CZ	0.36(1)	0.28(1)	0.51(1)	0.45(1)	0.977(4)	0.693(5)	1.59(2)	1.00(1)
CZ2	-	-	1.36(4)	1.07(5)	0.44(1)	0.224(4)	0.85(1)	0.62(1)
CH2	-	-	2.5(1)	1.7(1)	0.50(1)	0.298(5)	0.91(1)	0.74(1)
CE3	-	-	1.06(4)	0.84(4)	0.44(1)	0.234(4)	1.37(2)	1.13(1)
CZ3	-	-	1.14(4)	0.95(3)	0.44(1)	0.225(3)	2.06(4)	1.15(3)
HB2	0.331(1)	0.233(1)	0.315(1)	0.215(1)	5.5(1)*	0.266(5)	0.194(1)	0.0971(1)
HB3	0.324(1)	0.226(1)	0.310(1)	0.218(1)	5.0(1)*	0.254(4)	0.1975(4)	0.0998(1)
HG2	0.255(2)	0.17(1)	0.244(1)	0.160(1)	0.221(2)	0.0661(3)	0.1569(4)	0.0758(2)
HB	0.265(1)	0.179(1)	0.242(2)	0.162(1)	0.81(2)*	0.083(1)	0.1420(5)	0.0702(2)
HD2	0.300(3)	0.184(1)	0.303(3)	0.185(1)	0.174(1)	0.0668(3)	0.185(1)	0.0891(3)
HG3	0.275(2)	0.183(1)	0.278(2)	0.189(1)	0.162(1)	0.0661(3)	0.1767(5)	0.0907(2)
HD1	0.246(2)	0.165(1)	0.245(2)	0.167(1)	0.209(3)	0.0627(5)	0.198(1)	0.0851(3)
HD3	0.431(5)	0.251(2)	0.423(4)	0.251(3)	0.192(2)	0.0700(4)	0.219(1)	0.1053(4)
HE2	0.25(1)	0.148(2)	0.236(3)	0.148(1)	0.118(1)	0.0492(2)	0.1412(5)	0.0776(3)
HG	0.320(4)	0.220(2)	0.344(4)	0.236(2)	0.185(1)	0.0744(5)	0.214(1)	0.115(1)
HE1	0.46(1)	0.274(3)	0.45(1)	0.273(3)	0.260(3)	0.081(1)	0.190(1)	0.1061(4)
HE3	0.224(3)	0.157(1)	0.192(2)	0.144(1)	0.121(1)	0.0458(3)	0.162(1)	0.0885(4)
HG12	0.422(4)	0.333(3)	0.403(4)	0.328(3)	0.166(1)	0.079(1)	0.221(1)	0.152(1)
HG13	0.456(4)	0.359(3)	0.424(5)	0.335(4)	0.171(1)	0.085(1)	0.235(1)	0.162(1)
HG1	0.277(4)	0.186(2)	0.37(1)	0.216(3)	0.37(1)	0.069(1)	0.261(3)	0.098(1)
HD21	0.49(1)	0.324(4)	0.48(1)	0.283(5)	0.491(2)	0.301(2)	0.378(2)	0.248(1)
HD22	0.419(4)	0.319(4)	0.385(5)	0.291(4)	0.528(3)	0.303(2)	0.295(1)	0.187(1)
HE	0.31(1)	0.207(3)	0.33(1)	0.214(4)	0.162(2)	0.063(1)	0.478(5)	0.215(2)
HE21	0.310(5)	0.220(3)	0.305(3)	0.233(2)	0.521(3)	0.269(2)	0.446(5)	0.230(2)
HE22	0.278(4)	0.210(2)	0.271(5)	0.186(3)	0.476(2)	0.249(2)	0.249(1)	0.159(1)
HZ	0.265(3)	0.212(3)	0.268(3)	0.198(2)	0.66(2)	0.123(3)	0.37(1)	0.165(1)
HZ2	-	-	0.249(5)	0.184(4)	0.172(2)	0.078(1)	0.236(2)	0.152(1)
HH2	-	-	0.273(5)	0.229(4)	0.201(5)	0.061(2)	0.199(1)	0.147(1)
HZ3	-	-	0.307(3)	0.279(3)	0.294(5)	0.096(3)	0.387(4)	0.243(3)
ND2	-	-	2.61(5)	2.04(4)	1.45(2)	0.48(1)	1.58(2)	0.83(1)
NE2	-	-	1.94(4)	1.45(2)	0.79(1)	0.228(4)	2.04(2)	0.99(1)
NE1	-	-	1.72(3)	1.43(3)	0.455(5)	0.219(4)	1.41(1)	0.99(1)

* After removing the enormous outlayer, the RMSE values for HB2, HB3 and HB are reduced to 0.190(1), 0.180(1) and 0.130(1), respectively.

Table S7: *RMSE and MAE for ShiftX2 and UCBShift trained and tested on ShiftX2 data sets.* The UCBShift algorithm was trained and tested using the ShiftX2 datasets as reported in the original ShiftX2 article. In "test mode", sequences demonstrating greater than 99% identity with the query sequence are excluded from the prediction process. Numbers in brackets show the ± uncertainty on the last digit.

	ShiftX2		**UCBShift**		**UCBShift without the "testing mode"**	
Atom	RMSE	MAE	RMSE	MAE	RMSE	MAE
CG	0.828(3)	0.474(1)	0.650(3)	0.3395(9)	0.426(2)	0.194(1)
CD1	1.072(3)	0.675(2)	0.967(3)	0.589(2)	0.567(3)	0.321(1)
CG2	0.985(3)	0.612(2)	0.819(4)	0.501(2)	0.527(3)	0.307(1)
CD	0.714(5)	0.380(2)	0.635(5)	0.311(2)	0.50(1)	0.207(1)
CG1	0.99(1)	0.560(3)	0.855(4)	0.509(2)	0.65(1)	0.319(2)
CD2	1.38(1)	0.839(4)	1.118(5)	0.693(2)	0.82(1)	0.461(2)
CE	0.51(1)	0.272(2)	0.49(1)	0.239(2)	0.44(1)	0.187(1)
CE1	1.16(1)	0.595(5)	1.02(1)	0.499(4)	0.93(1)	0.406(3)
CE2	0.77(1)	0.466(3)	0.78(1)	0.497(4)	0.75(1)	0.458(4)
CZ	1.35(1)	0.84(1)	1.33(1)	0.90(1)	1.19(1)	0.79(1)
CZ2	-	-	0.77(1)	0.60(1)	0.68(1)	0.53(1)
CH2	-	-	1.03(1)	0.74(1)	0.95(1)	0.67(1)
CE3	-	-	1.13(1)	0.93(1)	1.05(1)	0.81(1)
CZ3	-	-	1.27(1)	0.95(1)	1.29(1)	0.98(1)
HB2	0.220(1)	0.1126(3)	0.2048(5)	0.1205(2)	0.1378(4)	0.0753(1)
HB3	0.252(1)	0.1385(3)	0.218(1)	0.1216(2)	0.144(1)	0.0728(2)
HG2	0.1789(5)	0.0962(2)	0.1465(4)	0.0861(2)	0.1064(4)	0.0553(2)
HB	0.202(3)	0.0836(4)	0.201(3)	0.0896(5)	0.159(4)	0.0574(3)
HD2	0.243(1)	0.1185(4)	0.200(1)	0.1088(3)	0.150(1)	0.0718(3)
HG3	0.201(1)	0.1052(4)	0.171(1)	0.0981(3)	0.125(1)	0.0664(3)
HD1	0.211(1)	0.1131(4)	0.223(2)*	0.1105(4)*	0.169(2)**	0.0672(3)**
HD3	0.225(1)	0.111(1)	0.198(1)	0.1188(5)	0.157(1)	0.0844(4)
HE2	0.195(1)	0.108(1)	0.256(5)*	0.105(1)*	0.22(1)**	0.079(1)**
HG	0.234(1)	0.122(1)	0.298(2)*	0.155(1)*	0.177(2)**	0.0972(4)**
HE1	0.275(1)	0.134(1)	0.258(1)	0.153(1)	0.181(1)	0.106(1)
HE3	0.200(2)	0.087(1)	0.177(2)	0.089(1)	0.128(1)	0.0677(5)
HG12	0.312(1)	0.202(1)	0.305(1)	0.210(1)	0.222(1)	0.150(1)
HG13	0.308(1)	0.212(1)	0.255(1)	0.191(1)	0.184(1)	0.133(1)
HG1	0.164(1)	0.082(1)	0.61(1)*	0.194(2)*	0.65(1)**	0.168(2)**
HD21	0.380(4)	0.215(2)	0.354(3)	0.241(2)	0.335(3)	0.224(2)
HD22	0.316(5)	0.171(2)	0.297(5)	0.16(2)	0.30(1)	0.152(2)
HE	0.48(1)	0.225(4)	0.42(1)	0.192(2)	0.39(1)	0.158(3)
HE21	0.303(2)	0.191(2)	0.294(1)	0.220(1)	0.286(1)	0.216(1)
HE22	0.286(2)	0.156(1)	0.219(2)	0.143(1)	0.220(2)	0.145(1)
HZ	0.50(1)	0.210(3)	0.45(1)	0.242(2)	0.42(1)	0.200(2)
HZ2	-	-	0.28(1)	0.172(3)	0.260(5)	0.147(3)
HH2	-	-	0.195(2)	0.148(1)	0.178(1)	0.138(1)
HZ3	-	-	0.254(2)	0.173(2)	0.204(2)	0.146(2)
ND2	-	-	1.56(1)	1.04(1)	1.20(1)	0.82(1)
NE2	-	-	1.17(1)	0.75(1)	1.04(1)	0.65(1)
NE1	-	-	1.46(1)	1.05(1)	1.03(1)	0.77(1)

* RMSE and MAE excluding atom types which ShiftX2 does not predict. RMSE for HD1, HE2, HG, HG1: 0.180(1),0.174(2), 0.231(1), 0.157(1), respectively. MAE for HD1, HE2, HG, HG1: 0.1036(3), 0.0959(5), 0.14(1), 0.099(1).

** RMSE and MAE excluding atom types which ShiftX2 does not predict. RMSE for HD1, HE2, HG, HG1: 0.109(1), 0.117(1), 0.146(1), 0.122(1), respectively. MAE for HD1, HE2, HG, HG1: 0.0607(2), 0.0701(4), 0.0882(5), 0.0698(5).

Table S8: *Root mean square error (RMSE) and mean absolute error (MAE) for UCBShift and ShiftX2 tested only on low homology subset.* The algorithm was tested using a subset of 59 proteins from the entire test set, exhibiting sequence homology of less than 50% compared to the training set. Uncertainties were computed based on 50 random samples from 75% of the test data. All units in ppm.

	UCBShift		**ShiftX2**	
Atom	RMSE	MAE	RMSE	MAE
CG	1.04(1)	0.608(4)	1.36(1)	0.826(5)
CD1	0.94(1)	0.554(4)	1.38(1)	0.90(1)
CG2	0.98(1)	0.611(4)	1.31(1)	0.814(5)
CD	1.36(1)	0.82(1)	1.75(2)	1.02(1)
CG1	1.25(1)	0.74(1)	1.44(2)	0.87(1)
CD2	0.84(1)	0.54(1)	1.41(3)	0.89(1)
CE	1.22(1)	0.77(1)	1.27(1)	0.81(1)
CE1	0.54(1)	0.35(1)	0.94(1)	0.56(1)
CE2	0.236(4)	0.191(4)	0.37(1)	0.27(1)
CZ	0.86(3)	0.49(2)	1.76(3)	1.30(3)
CZ2	1.29(2)	1.20(2)	-	-
CH2	1.39(1)	1.38(1)	-	-
CE3	1.62(5)	1.24(4)	-	-
CZ3	1.00(2)	0.88(2)	-	-
HB2	0.232(1)	0.1265(4)	0.260(1)	0.1370(5)
HB3	0.209(1)	0.1245(3)	0.271(1)	0.1571(5)
HG2	0.191(1)	0.0945(4)	0.212(1)	0.1109(5)
HB	0.189(2)	0.102(1)	0.235(2)	0.120(1)
HD2	0.198(2)	0.104(1)	0.238(3)	0.117(1)
HG3	0.222(2)	0.115(1)	0.239(2)	0.118(1)
HD1	0.208(3)	0.100(1)	0.224(1)	0.119(1)
HD3	0.238(2)	0.118(1)	0.247(2)	0.118(1)
HE2	0.209(3)	0.104(1)	0.255(4)	0.128(2)
HG	0.342(3)	0.190(2)	0.318(3)	0.169(2)
HE1	0.234(2)	0.143(1)	0.267(4)	0.126(2)
HE3	0.187(2)	0.104(1)	0.280(4)	0.139(2)
HG12	0.263(3)	0.178(1)	0.288(2)	0.178(2)
HG13	0.248(2)	0.172(1)	0.308(2)	0.197(3)
HG1	0.27(1)	0.120(2)	0.192(2)	0.102(1)
HD21	0.394(4)	0.265(3)	0.51(1)	0.317(4)
HD22	0.405(4)	0.272(3)	0.455(4)	0.320(4)
HE	0.35(1)	0.200(3)	0.41(1)	0.204(4)
HE21	0.41(1)	0.255(3)	0.427(5)	0.233(3)
HE22	0.303(5)	0.194(3)	0.335(5)	0.193(3)
HZ	0.231(2)	0.157(2)	0.193(3)	0.107(3)
HZ2	0.253(5)	0.163(3)	-	-
HH2	0.205(2)	0.180(2)	-	-
HZ3	0.264(2)	0.217(3)	-	-
ND2	2.1(1)	1.17(2)	-	-
NE2	2.09(5)	1.05(2)	-	-
NE1	1.65(2)	1.22(2)	-	-

Table S9: Carbon atom features.

	C			**CA**			**CB**			**CG**		
Feature	R0	R1	R2	R0	R1	R2	R0	R1	R2	R0	R1	R2
Backbone dihedral angles	23.2	17.3	5.8	32.3	27.3	13.0	34.3	21.3	7.7	6.3	7.2	1.2
Sidechain dihedral angles	6.2	3.2	0.5	5.4	2.1	0.5	7.6	3.5	0.7	21.7	13.0	2.6
Transformed dihedral angles	4.8	2.7	0.5	4.0	1.6	0.3	8.9	3.2	0.9	10.3	8.0	1.6
Secondary structure	37.9	9.4	6.7	39.7	4.7	2.5	21.7	3.6	1.3	7.1	0.3	0.0
Atomic surface area	1.7	1.2	0.2	0.8	0.6	0.1	1.2	1.1	0.2	3.7	3.5	0.7
Half surface exposure	4.3	3.0	0.5	3.5	2.7	1.1	3.9	3.2	0.8	8.1	6.7	1.4
BLOSUM numbers	10.0	2.3	0.4	7.5	1.7	0.4	11.3	2.3	0.6	18.0	4.4	1.1
Hydrogen bond	4.0	3.1	0.5	2.3	1.5	0.3	3.8	2.6	0.4	7.5	7.1	1.4
Transformed hydrogen bond	5.5	3.8	0.7	3.0	1.8	0.3	4.7	3.1	0.5	10.1	8.8	1.8
Ring current	0.5	0.7	0.1	0.2	0.4	0.1	0.7	0.8	0.1	2.4	2.5	0.5
B Factor	0.4	0.9	0.2	0.3	0.5	0.1	0.4	0.8	0.2	1.4	3.7	2.6
S^2 order parameters	0.4	0.8	0.1	0.4	0.5	0.1	0.5	0.8	0.1	1.1	1.9	0.6
Hydrophobicity	0.2	0.1	0.0	0.1	0.1	0.0	0.2	0.1	0.0	0.4	0.3	0.0
pH values	0.2	0.4	0.2	0.1	0.2	0.1	0.2	0.4	0.1	0.5	0.8	0.3
Protonation	0.6	0.0	0.0	0.5	0.0	0.0	0.7	0.0	0.0	1.4	0.1	0.1
Prediction from R0	-	51.2	23.3	-	54.2	25.0	-	53.1	24.2	-	31.6	9.5
UCBShift-Y prediction	-	-	58.9	-	-	55.3	-	-	60.8	-	-	72.1
UCBShift-Y metrics	-	-	1.3	-	-	0.8	-	-	1.4	-	-	2.4
	CD1			**CG2**			**CD**			**CG1**		
Feature	R0	R1	R2	R0	R1	R2	R0	R1	R2	R0	R1	R2
Backbone dihedral angles	5.7	8.0	1.4	8.0	8.1	1.9	6.0	7.7	1.4	8.7	8.1	2.1
Sidechain dihedral angles	23.1	12.8	4.1	20.6	12.4	3.9	17.9	14.4	4.8	17.9	10.6	2.8
Transformed dihedral angles	10.3	9.5	1.9	12.4	9.0	1.7	9.7	8.9	2.1	9.8	7.5	1.5
Secondary structure	5.4	0.3	0.1	7.3	0.3	0.1	7.7	0.3	0.1	17.4	6.1	2.9
Atomic surface area	3.3	3.4	1.0	2.9	2.8	0.5	5.9	6.2	1.8	2.9	3.4	0.4
Half surface exposure	8.4	7.5	1.5	9.1	7.7	1.5	12.0	8.6	1.3	7.6	6.6	1.2
BLOSUM numbers	18.2	4.8	2.1	17.4	3.5	0.8	16.9	3.8	0.9	13.2	3.0	0.6
Hydrogen bond	7.6	8.8	1.9	7.2	7.5	1.2	7.0	6.5	1.3	7.3	7.7	1.3
Transformed hydrogen bond	11.2	10.4	2.5	9.7	9.5	1.5	9.6	8.4	2.4	10.1	9.4	1.6
Ring current	3.2	3.2	0.3	2.6	2.9	0.1	2.7	2.4	0.2	2.2	2.3	0.5
B Factor	1.2	2.9	1.1	1.2	2.8	1.0	2.2	3.6	0.7	1.3	2.7	1.1
S^2 order parameters	0.7	2.2	0.6	0.8	2.6	0.3	0.8	2.3	0.3	0.8	1.9	0.3
Hydrophobicity	0.3	0.1	0.1	0.2	0.0	0.0	0.3	0.0	0.0	0.2	0.0	0.0
pH values	0.5	0.8	0.2	0.5	0.6	0.1	0.6	1.1	0.1	0.6	0.6	0.1
Protonation	0.9	0.1	0.1	0.0	0.0	0.0	1.0	0.1	0.0	0.0	0.0	0.0
Prediction from R0	-	25.3	11.3	-	30.4	11.0	-	25.6	5.9	-	30.1	9.5
UCBShift-Y prediction	-	-	66.1	-	-	72.3	-	-	74.9	-	-	72.6
UCBShift-Y metrics	-	-	3.7	-	-	1.9	-	-	1.7	-	-	1.5

Continued on next page

	CD2			**CE**			**CE1**			**CE2**		
Feature	R0	R1	R2	R0	R1	R2	R0	R1	R2	R0	R1	R2
Backbone dihedral angles	5.5	8.2	1.5	4.7	7.9	1.6	6.8	9.1	5.1	6.2	11.1	1.9
Sidechain dihedral angles	18.5	12.1	2.6	22.5	15.1	3.1	16.5	14.9	2.5	18.1	16.3	1.6
Transformed dihedral angles	13.0	9.7	2.8	11.0	10.0	2.8	10.3	9.3	2.4	10.8	8.4	3.4
Secondary structure	7.4	0.3	0.1	9.8	1.3	0.1	5.5	0.2	0.1	6.6	0.3	0.0
Atomic surface area	3.6	3.9	0.6	3.3	4.6	2.4	4.6	7.0	0.7	3.8	6.2	1.2
Half surface exposure	7.9	9.1	1.2	6.5	6.4	2.1	8.2	7.7	1.8	8.1	8.7	0.9
BLOSUM numbers	17.2	5.3	4.5	16.9	5.5	1.2	23.3	8.5	1.8	18.1	4.1	0.4
Hydrogen bond	8.2	8.2	1.2	7.8	9.3	2.5	9.1	11.2	3.5	9.7	8.9	1.7
Transformed hydrogen bond	11.3	9.6	1.7	10.5	11.4	3.3	12.2	12.3	2.8	13.6	9.5	1.9
Ring current	1.7	3.0	0.3	1.8	2.9	0.6	0.4	0.6	0.1	0.6	1.7	0.1
B Factor	1.2	4.5	1.4	3.3	5.6	1.0	0.8	2.2	2.8	0.9	3.0	0.3
S^2 order parameters	0.6	2.1	0.5	0.7	1.9	1.0	0.8	1.9	0.5	1.8	3.5	0.2
Hydrophobicity	0.1	0.1	0.1	0.3	0.2	0.0	0.1	0.1	0.0	0.2	0.1	0.0
pH values	0.7	1.1	1.1	0.4	2.3	0.3	0.8	3.5	1.0	1.1	8.8	6.0
Protonation	3.1	0.4	0.1	0.6	0.0	0.0	0.5	0.2	1.0	0.1	0.0	0.0
Prediction from R0	-	22.4	5.8	-	15.7	5.0	-	11.4	3.3	-	9.4	0.5
UCBShift-Y prediction	-	-	72.5	-	-	70.6	-	-	67.3	-	-	75.3
UCBShift-Y metrics	-	-	2.0	-	-	2.4	-	-	3.3	-	-	4.6
	CZ			**CZ2**			**CH2**			**CE3**		
Feature	R0	R1	R2	R0	R1	R2	R0	R1	R2	R0	R1	R2
Backbone dihedral angles	4.7	6.0	1.3	8.9	16.3	3.3	8.2	12.0	4.1	4.4	5.7	2.4
Sidechain dihedral angles	15.4	21.0	2.8	17.8	9.1	1.8	15.9	9.5	2.4	19.1	19.0	14.2
Transformed dihedral angles	10.8	10.3	1.8	11.5	16.5	6.0	10.3	19.6	0.2	11.8	9.1	1.6
Secondary structure	7.0	0.2	0.0	5.6	0.2	0.0	6.7	0.2	0.0	11.1	0.4	0.4
Atomic surface area	6.7	4.7	0.9	4.6	4.9	0.7	3.0	3.8	5.1	3.2	4.7	0.3
Half surface exposure	8.3	5.1	0.9	14.0	16.0	3.1	8.7	8.8	13.9	7.7	12.1	1.5
BLOSUM numbers	22.7	6.2	1.9	14.3	4.6	1.2	16.5	3.8	0.6	16.8	3.7	0.4
Hydrogen bond	8.4	12.0	5.4	9.3	7.5	2.4	11.3	11.1	2.6	10.2	17.8	4.6
Transformed hydrogen bond	12.4	12.7	4.6	11.5	8.5	12.9	17.3	20.3	3.8	13.7	20.3	5.8
Ring current	0.2	0.5	0.0	0.1	0.7	0.1	0.2	5.0	0.0	0.4	0.7	0.0
B Factor	2.0	3.0	9.3	1.4	5.2	0.1	1.0	2.8	0.3	1.1	0.6	0.2
S^2 order parameters	0.7	1.7	0.7	0.5	1.6	0.7	0.5	1.4	0.1	0.4	4.2	0.8
Hydrophobicity	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
pH values	0.4	0.6	0.0	0.5	0.9	0.0	0.3	0.8	0.0	0.2	0.3	0.0
Protonation	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Prediction from R0	-	16.0	3.0	-	8.1	0.0	-	0.8	0.0	-	1.5	4.9
UCBShift-Y prediction	-	-	67.0	-	-	56.5	-	-	67.0	-	-	62.9
UCBShift-Y metrics	-	-	0.4	-	-	11.1	-	-	0.0	-	-	0.0

Continued on next page

	CZ3					
Feature	R0	R1	R2			
Backbone dihedral angles	5.7	11.8	9.7			
Sidechain dihedral angles	23.0	17.1	18.3			
Transformed dihedral angles	11.6	16.6	4.5			
Secondary structure	4.8	0.1	0.0			
Atomic surface area	2.4	6.2	0.5			
Half surface exposure	8.1	6.6	2.6			
BLOSUM numbers	14.8	4.5	3.0			
Hydrogen bond	12.2	13.4	3.2			
Transformed hydrogen bond	16.1	17.9	5.2			
Ring current	0.3	0.5	2.0			
B Factor	0.5	1.3	0.2			
S^2 order parameters	0.4	1.3	0.4			
Hydrophobicity	0.0	0.0	0.0			
pH values	0.1	0.4	0.0			
Protonation	0.0	0.0	0.0			
Prediction from R0	-	2.4	0.2			
UCBShift-Y prediction	-	-	50.2			
UCBShift-Y metrics	-	-	0.1			

Table S10: Hydrogen atom features.

	H			**HA**			**HB2**			**HB3**		
Feature	R0	R1	R2	R0	R1	R2	R0	R1	R2	R0	R1	R2
Backbone dihedral angles	20.8	12.4	4.0	17.3	10.1	5.6	6.9	6.4	1.8	6.2	6.0	1.6
Sidechain dihedral angles	6.0	3.2	0.5	5.4	1.6	0.4	10.7	6.7	1.4	11.2	6.6	1.5
Transformed dihedral angles	4.2	2.7	0.4	2.7	1.4	0.4	6.2	5.4	1.2	6.1	5.1	1.2
Secondary structure	15.5	0.4	0.1	22.5	3.3	1.7	7.1	0.4	0.1	7.9	1.1	0.3
Atomic surface area	1.5	1.3	0.2	0.9	0.5	0.2	2.3	2.6	0.7	2.8	2.6	0.6
Half surface exposure	5.4	3.3	0.4	2.9	3.3	0.9	5.9	5.5	1.4	6.8	5.5	1.2
BLOSUM numbers	10.4	2.0	0.4	22.2	3.4	0.9	18.2	3.8	1.1	17.5	4.2	1.0
Hydrogen bond	10.6	4.7	1.7	7.3	2.6	0.9	7.2	5.8	1.4	6.5	5.7	1.2
Transformed hydrogen bond	16.9	10.8	3.3	10.3	5.8	1.4	9.2	7.1	1.8	8.3	6.7	1.5
Ring current	4.9	2.5	0.7	5.3	3.3	1.3	22.7	12.5	5.6	23.3	12.1	5.4
B Factor	0.7	1.0	0.2	0.2	0.4	0.1	0.5	1.4	0.6	0.6	1.5	0.5
S^2 order parameters	0.6	0.9	0.1	0.2	0.4	0.1	0.5	1.5	0.3	0.5	1.4	0.3
Hydrophobicity	0.2	0.1	0.0	0.9	0.2	0.1	0.3	0.1	0.0	0.3	0.1	0.0
pH values	0.3	0.2	0.1	0.1	0.1	0.1	0.5	0.5	0.9	0.5	0.9	1.3
Electric field	1.7	0.8	0.2	0.3	0.2	0.0	0.3	0.5	0.1	0.3	0.6	0.1
Protonation	0.5	0.0	0.0	1.7	0.0	0.0	1.5	0.2	0.0	1.0	0.1	0.0
Prediction from R0	-	53.7	19.9	-	63.3	26.4	-	39.5	15.8	-	39.8	13.6
UCBShift-Y prediction	-	-	66.6	-	-	58.6	-	-	62.7	-	-	66.0
UCBShift-Y metrics	-	-	1.1	-	-	1.0	-	-	3.0	-	-	2.5
	HG2			**HB**			**HD2**			**HG3**		
Feature	R0	R1	R2	R0	R1	R2	R0	R1	R2	R0	R1	R2
Backbone dihedral angles	4.4	4.5	1.1	9.3	6.1	2.0	4.5	4.9	1.0	4.9	5.7	2.1
Sidechain dihedral angles	9.7	6.1	1.3	9.3	4.7	1.5	11.7	7.7	1.3	10.2	6.7	1.3
Transformed dihedral angles	5.4	4.4	1.1	6.5	3.7	1.0	6.3	5.3	1.0	5.6	5.4	1.2
Secondary structure	6.3	0.4	0.1	9.3	0.2	0.1	5.3	0.2	0.1	7.7	0.6	0.4
Atomic surface area	3.6	2.8	0.8	1.8	1.6	0.3	3.7	2.1	0.6	3.6	3.1	1.3
Half surface exposure	5.7	3.8	1.3	4.8	3.6	0.7	5.9	4.4	1.4	6.3	4.9	1.4
BLOSUM numbers	13.2	2.6	0.7	11.3	1.9	0.5	18.1	4.3	1.5	14.8	2.6	1.3
Hydrogen bond	5.3	4.3	1.0	7.4	4.3	0.9	9.0	5.7	1.2	5.8	4.8	1.7
Transformed hydrogen bond	7.0	5.4	1.2	9.2	5.1	1.3	10.0	6.8	1.9	7.8	6.5	1.6
Ring current	36.8	17.2	8.3	29.5	15.2	8.2	22.3	14.3	5.1	30.7	17.4	9.8
B Factor	0.6	1.1	0.3	0.5	0.9	0.2	0.6	1.1	0.6	0.5	1.5	0.5
S^2 order parameters	0.4	1.2	0.4	0.4	0.9	0.1	0.7	1.2	0.2	0.5	1.6	0.5
Hydrophobicity	0.2	0.0	0.0	0.1	0.0	0.0	0.2	0.0	0.0	0.2	0.0	0.0
pH values	0.4	0.5	0.3	0.2	0.3	0.1	0.5	0.4	0.2	0.5	0.7	0.5
Electric field	0.3	0.8	0.1	0.4	0.4	0.1	0.4	1.0	0.2	0.2	0.5	0.2
Protonation	0.5	0.1	0.0	0.0	0.0	0.0	0.8	0.1	0.0	0.5	0.1	0.0
Prediction from R0	-	44.7	19.6	-	51.1	22.6	-	40.7	14.0	-	37.9	13.3
UCBShift-Y prediction	-	-	60.1	-	-	59.8	-	-	67.4	-	-	60.9
UCBShift-Y metrics	-	-	2.4	-	-	0.6	-	-	2.1	-	-	2.1

Continued on next page

	HD1			**HD3**			**HE2**			**HG**		
Feature	R0	R1	R2	R0	R1	R2	R0	R1	R2	R0	R1	R2
Backbone dihedral angles	3.0	3.7	0.7	4.3	4.5	1.2	2.3	4.2	1.4	5.8	6.7	1.2
Sidechain dihedral angles	7.3	4.4	1.0	11.7	7.6	2.6	9.5	6.8	1.6	10.0	5.7	2.1
Transformed dihedral angles	4.3	4.7	0.8	5.9	5.8	2.0	7.1	5.8	1.7	6.1	5.1	1.3
Secondary structure	3.3	0.2	0.0	6.0	0.3	0.2	2.7	0.3	0.1	8.1	0.5	0.0
Atomic surface area	4.0	3.8	1.1	3.9	3.7	1.7	13.8	16.3	1.6	2.4	1.7	0.3
Half surface exposure	5.2	3.5	1.0	6.2	5.7	2.1	6.3	5.5	0.7	6.7	5.6	1.0
BLOSUM numbers	34.4	8.4	1.1	16.3	2.9	1.8	28.6	11.4	5.5	20.2	4.0	0.6
Hydrogen bond	5.6	4.8	1.1	6.7	5.5	1.4	11.4	4.6	2.0	8.9	9.3	1.6
Transformed hydrogen bond	8.4	6.9	1.4	8.7	6.5	2.0	9.8	5.1	4.3	11.7	11.4	3.3
Ring current	20.7	6.2	5.7	28.2	17.2	10.5	4.3	3.2	4.1	17.4	12.5	8.8
B Factor	0.8	1.9	0.4	0.5	1.2	0.5	0.5	0.9	0.3	0.7	0.7	0.6
S^2 order parameters	0.3	1.0	0.3	0.7	1.6	0.4	1.3	1.0	1.6	0.5	1.1	0.3
Hydrophobicity	0.6	0.3	0.1	0.1	0.0	0.1	0.8	1.4	0.2	0.3	0.2	0.0
pH values	0.4	0.2	0.2	0.5	0.5	0.4	1.2	0.9	0.2	0.5	0.6	0.3
Electric field	0.8	8.9	0.4	0.2	0.6	0.1	0.4	1.0	0.1	0.7	5.2	0.1
Protonation	0.7	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0
Prediction from R0	-	41.0	16.1	-	36.4	10.5	-	31.5	10.6	-	29.7	14.5
UCBShift-Y prediction	-	-	66.7	-	-	60.1	-	-	63.2	-	-	62.2
UCBShift-Y metrics	-	-	1.7	-	-	2.6	-	-	0.8	-	-	1.7
	HE1			**HE3**			**HG12**			**HG13**		
Feature	R0	R1	R2	R0	R1	R2	R0	R1	R2	R0	R1	R2
Backbone dihedral angles	3.6	5.2	0.9	3.4	4.5	0.9	5.4	8.2	2.4	4.0	7.9	2.6
Sidechain dihedral angles	12.0	9.1	1.0	16.4	11.1	2.5	11.9	7.4	2.9	14.1	11.4	0.9
Transformed dihedral angles	6.7	7.5	1.2	7.0	10.3	1.4	7.5	5.9	2.0	7.7	6.4	1.1
Secondary structure	4.3	0.2	0.0	5.1	0.3	0.1	4.8	0.2	0.1	3.6	0.1	0.2
Atomic surface area	5.3	5.0	2.1	6.1	5.7	3.0	4.0	2.9	1.7	4.0	2.9	0.3
Half surface exposure	6.7	5.1	1.4	10.0	7.5	1.7	8.5	8.5	1.6	7.9	6.1	2.6
BLOSUM numbers	29.8	11.8	4.2	16.2	4.6	2.3	12.6	3.3	0.8	15.3	3.3	0.8
Hydrogen bond	8.9	7.3	1.9	7.9	6.5	1.5	7.7	7.3	1.7	9.6	10.2	2.0
Transformed hydrogen bond	12.8	8.7	2.7	10.1	7.3	1.5	10.9	8.9	2.4	12.5	11.9	2.1
Ring current	1.6	2.3	0.2	15.0	14.7	5.3	24.4	16.4	9.7	18.5	21.4	10.8
B Factor	0.7	1.7	0.7	0.6	1.1	0.6	0.7	1.8	0.3	0.9	1.7	0.4
S^2 order parameters	0.4	1.4	0.2	1.0	1.6	0.4	0.5	3.1	0.4	0.6	2.4	1.1
Hydrophobicity	2.7	0.5	1.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
pH values	1.3	1.2	0.1	0.5	1.1	0.5	0.7	1.4	0.4	1.0	0.9	0.3
Electric field	1.0	3.1	0.7	0.3	1.4	0.2	0.4	0.8	0.2	0.5	0.6	0.1
Protonation	2.2	0.3	0.0	0.4	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0
Prediction from R0	-	29.7	9.5	-	22.3	14.4	-	23.9	11.1	-	12.9	3.9
UCBShift-Y prediction	-	-	71.3	-	-	63.3	-	-	61.6	-	-	68.6
UCBShift-Y metrics	-	-	0.8	-	-	0.4	-	-	0.7	-	-	2.2

Continued on next page

	HG1			**HD21**			**HD22**			**HE**		
Feature	R0	R1	R2	R0	R1	R2	R0	R1	R2	R0	R1	R2
Backbone dihedral angles	3.3	6.7	1.6	7.8	10.4	3.5	6.3	9.0	2.1	3.4	3.7	0.3
Sidechain dihedral angles	6.6	1.4	0.9	12.5	7.9	3.3	13.1	7.2	4.5	14.2	6.3	1.7
Transformed dihedral angles	7.4	4.8	1.1	6.9	6.2	3.1	7.6	6.8	3.2	7.5	6.9	1.0
Secondary structure	4.8	0.0	0.1	12.3	1.2	3.4	9.9	1.3	0.1	9.6	0.2	0.0
Atomic surface area	2.0	7.0	0.8	5.5	6.6	7.2	5.1	5.2	5.4	3.0	7.0	1.0
Half surface exposure	4.7	2.0	0.8	13.0	11.8	6.3	13.6	16.1	12.8	5.6	3.7	0.3
BLOSUM numbers	29.9	3.1	0.5	14.6	4.2	3.2	13.1	3.1	1.0	22.0	3.4	0.4
Hydrogen bond	12.7	18.6	1.6	9.1	7.1	3.8	9.9	7.4	3.4	8.4	8.1	3.5
Transformed hydrogen bond	17.7	14.6	2.0	10.6	8.0	4.2	12.4	9.8	4.4	9.2	6.5	3.6
Ring current	6.4	5.8	16.1	4.1	9.0	1.6	6.3	13.7	1.9	8.6	14.5	11.7
B Factor	1.7	5.2	0.2	0.8	1.8	0.5	0.9	1.3	1.1	0.5	1.8	0.1
S^2 order parameters	0.5	0.2	0.2	0.5	1.3	0.5	0.7	1.9	0.9	0.5	0.8	0.4
Hydrophobicity	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.2	0.2	0.0
pH values	0.4	1.5	0.1	1.3	1.2	0.4	0.6	0.6	0.4	3.0	0.3	0.0
Electric field	1.7	0.6	0.0	1.1	3.2	0.2	0.5	0.4	0.2	1.7	3.6	0.1
Protonation	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.5	0.1	0.0
Prediction from R0	-	28.3	10.9	-	20.1	14.5	-	16.2	7.0	-	32.8	8.8
UCBShift-Y prediction	-	-	62.2	-	-	43.7	-	-	50.7	-	-	67.0
UCBShift-Y metrics	-	-	0.8	-	-	0.7	-	-	1.0	-	-	0.1
	HE21			**HE22**			**HZ**			**HZ2**		
Feature	R0	R1	R2	R0	R1	R2	R0	R1	R2	R0	R1	R2
Backbone dihedral angles	5.6	8.9	2.9	4.0	5.0	4.1	3.2	5.6	2.7	3.6	4.1	4.1
Sidechain dihedral angles	16.5	9.9	7.4	13.0	5.3	4.5	14.6	12.9	3.7	17.3	13.9	1.8
Transformed dihedral angles	9.9	7.4	3.7	10.3	5.0	3.5	8.2	9.7	3.5	8.2	8.6	13.4
Secondary structure	5.2	0.1	0.1	3.0	0.1	0.0	3.2	0.1	0.0	3.2	0.4	0.0
Atomic surface area	3.8	4.5	0.9	3.7	4.2	1.0	10.4	11.7	2.7	10.7	17.2	1.2
Half surface exposure	10.7	13.0	6.5	9.4	7.1	6.7	6.3	6.9	1.6	10.4	3.8	1.4
BLOSUM numbers	14.1	2.4	1.3	11.3	2.0	0.5	15.2	3.3	1.9	13.1	4.6	1.7
Hydrogen bond	9.3	8.0	5.3	7.2	9.1	4.4	15.4	13.9	2.2	15.6	16.5	1.2
Transformed hydrogen bond	12.1	8.9	5.1	10.8	7.3	4.9	19.7	15.5	5.8	13.9	16.8	1.5
Ring current	10.8	23.0	9.9	25.6	37.4	5.9	0.4	1.8	0.2	1.9	1.7	0.0
B Factor	0.8	1.0	1.2	0.7	0.8	3.9	1.0	1.9	0.2	0.8	1.1	0.1
S^2 order parameters	0.4	1.6	1.0	0.4	0.9	1.4	0.6	1.8	0.7	0.7	2.3	0.4
Hydrophobicity	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0
pH values	0.2	0.4	0.1	0.2	1.7	0.1	0.4	0.7	0.0	0.1	0.1	0.0
Electric field	0.5	0.5	0.9	0.3	0.4	0.1	0.9	4.0	0.1	0.4	0.5	0.2
Protonation	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0
Prediction from R0	-	10.3	7.4	-	13.6	3.8	0.0	10.1	5.2	0.0	8.3	0.6
UCBShift-Y prediction	-	-	44.3	-	-	54.9	-	-	68.7	-	-	68.3
UCBShift-Y metrics	-	-	2.0	-	-	0.2	-	-	0.8	-	-	3.9

Continued on next page

	HH2			**HZ3**				
Feature	R0	R1	R2	R0	R1	R2		
Backbone dihedral angles	5.0	5.4	1.9	3.9	6.9	1.2		
Sidechain dihedral angles	20.9	13.9	0.4	14.2	8.4	1.9		
Transformed dihedral angles	8.3	7.8	0.9	10.1	20.4	9.6		
Secondary structure	4.9	0.0	0.0	4.6	0.2	0.3		
Atomic surface area	15.7	26.3	16.1	10.3	14.6	11.4		
Half surface exposure	9.5	11.5	3.3	15.4	19.0	9.5		
BLOSUM numbers	14.8	2.5	0.4	16.3	6.0	1.1		
Hydrogen bond	7.5	7.8	2.2	9.8	8.3	1.7		
Transformed hydrogen bond	10.3	17.8	8.6	12.3	10.3	3.5		
Ring current	0.6	3.9	0.0	0.5	1.6	0.2		
B Factor	1.3	0.9	0.8	0.9	1.0	0.3		
S^2 order parameters	0.7	1.2	0.0	1.2	1.8	0.1		
Hydrophobicity	0.0	0.0	0.0	0.0	0.0	0.0		
pH values	0.4	0.1	0.5	0.1	0.1	0.1		
Electric field	0.3	0.2	0.1	0.5	0.3	0.0		
Protonation	0.0	0.0	0.0	0.0	0.0	0.0		
Prediction from R0	-	0.7	0.2	-	1.0	2.1		
UCBShift-Y prediction	-	-	57.9	-	-	56.9		
UCBShift-Y metrics	-	-	6.8	-	-	0.0		

Table S11: Nitrogen atom's features.

	N			**ND2**			**NE2**			**NE1**		
Feature	R0	R1	R2	R0	R1	R2	R0	R1	R2	R0	R1	R2
Backbone dihedral angles	27.9	14.2	5.9	5.3	6.9	1.7	4.3	6.4	2.1	5.8	6.6	4.3
Sidechain dihedral angles	12.3	3.2	0.6	15.9	12.6	3.2	17.8	10.3	2.1	23.1	18.0	1.1
Transformed dihedral angles	15.7	4.1	1.2	14.7	10.4	4.1	13.6	9.3	2.1	14.2	7.2	1.7
Secondary structure	11.6	0.7	0.1	8.9	0.8	0.1	3.3	0.1	0.0	8.1	0.1	0.0
Atomic surface area	1.3	1.0	0.2	4.7	5.1	0.9	7.3	8.1	2.8	3.1	3.9	0.2
Half surface exposure	6.3	5.7	2.5	9.0	7.4	3.8	13.2	14.6	3.4	6.5	9.1	2.5
BLOSUM numbers	15.5	2.4	0.5	12.9	3.9	0.7	13.4	2.6	0.8	14.9	3.8	1.2
Hydrogen bond	3.1	2.5	0.4	9.2	8.1	2.4	6.0	5.5	1.7	10.6	9.8	6.7
Transformed hydrogen bond	4.3	3.1	0.6	12.4	9.1	2.7	9.1	6.5	1.9	11.2	12.7	7.2
Ring current	0.2	0.3	0.1	0.3	2.8	0.1	4.7	4.7	1.4	0.2	1.1	0.0
B Factor	0.4	0.7	0.2	1.6	3.5	0.3	0.8	1.8	0.5	0.9	5.1	2.0
S^2 order parameters	0.4	0.7	0.1	0.8	1.9	0.5	0.4	3.1	0.2	0.5	2.5	1.2
Hydrophobicity	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
pH values	0.1	0.2	0.1	3.6	0.9	0.0	4.4	3.4	0.4	0.8	0.6	0.0
Electric field	0.1	0.2	0.0	0.8	2.1	0.2	1.6	2.1	0.2	0.3	1.3	0.0
Protonation	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Prediction from R0	-	60.9	26.2	-	24.5	10.2	-	21.6	6.6	-	18.2	9.7
UCBShift-Y prediction	-	-	60.0	-	-	68.5	-	-	72.6	-	-	61.6
UCBShift-Y metrics	-	-	1.4	-	-	0.7	-	-	1.1	-	-	0.5

Table S12: PDB IDs, amino acids, chemical shifts and the distance from the proton to the center of the ring for isoleucin and valine HB protons under study.

PDB ID	Amino acid	HB chemical shift (ppm)	CH $\cdots$ π distance ($\AA$)
3ERA	Ile36	0.040	3.0
1UWX	Val59	0.050	2.7
1EQT	Val40	0.053	3.0
1GWY	Ile160	0.061	2.8
1O5U	Val3	0.310	2.9
1GWY	Val100	0.354	3.1
1GPR	Ile131	0.440	3.3
1MOL	Ile53	0.500	3.0
2PF5	Ile76	0.500	3.2
1HC9	Val39	0.520	3.1
2PSP	Ile73	0.670	2.9
2PSP	Ile24	0.680	2.9
2H76	Val16	0.700	3.3
1OMS	Ile17	0.742	2.7
1FDQ	Val49	0.757	3.1
1O5U	Val18	0.874	3.4
1BD9	Val124	0.892	3.2
1FDQ	Val84	0.940	2.9
1AG6	Val40	0.940	3.5
1BRJ	Ile76	0.950	3.7

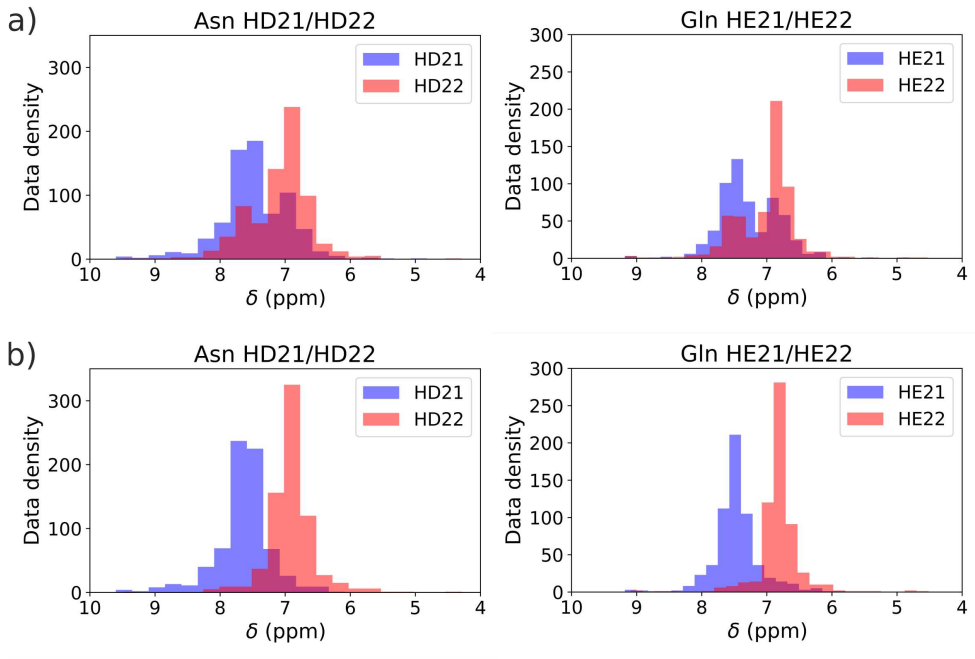


Figure S1: *Chemical shift distributions of HD21/22 and HE21/22 chemical shifts before (a) and after (b) data correction.*

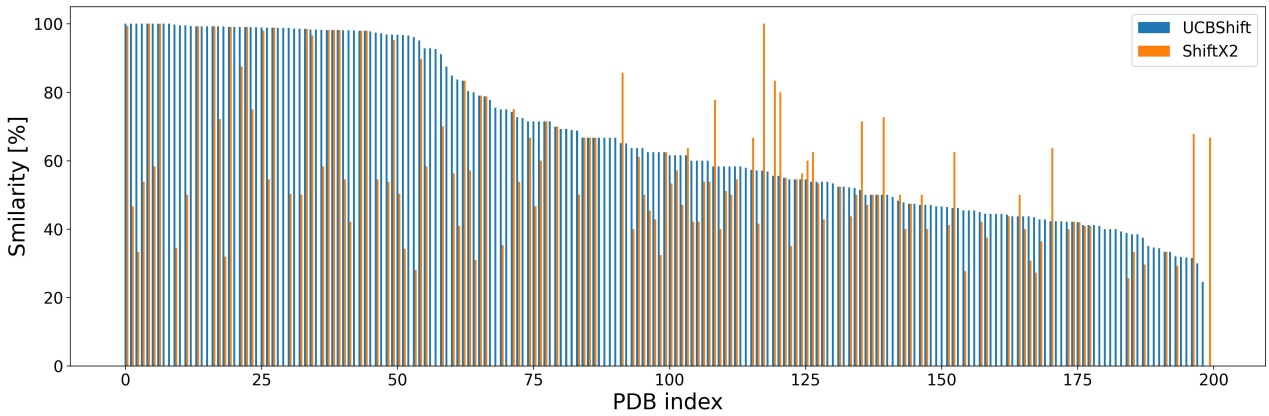


Figure S2: *TM scores calculated between each test sequence and the best-aligned sequence in the UCBShift (blue) and ShiftX2 (orange) training sets.* For cases exhibiting full sequence homology with the training set, the chemical shifts sourced from the BMRB database do not coincide between the training and test sets, since they correspond to different types of atoms.

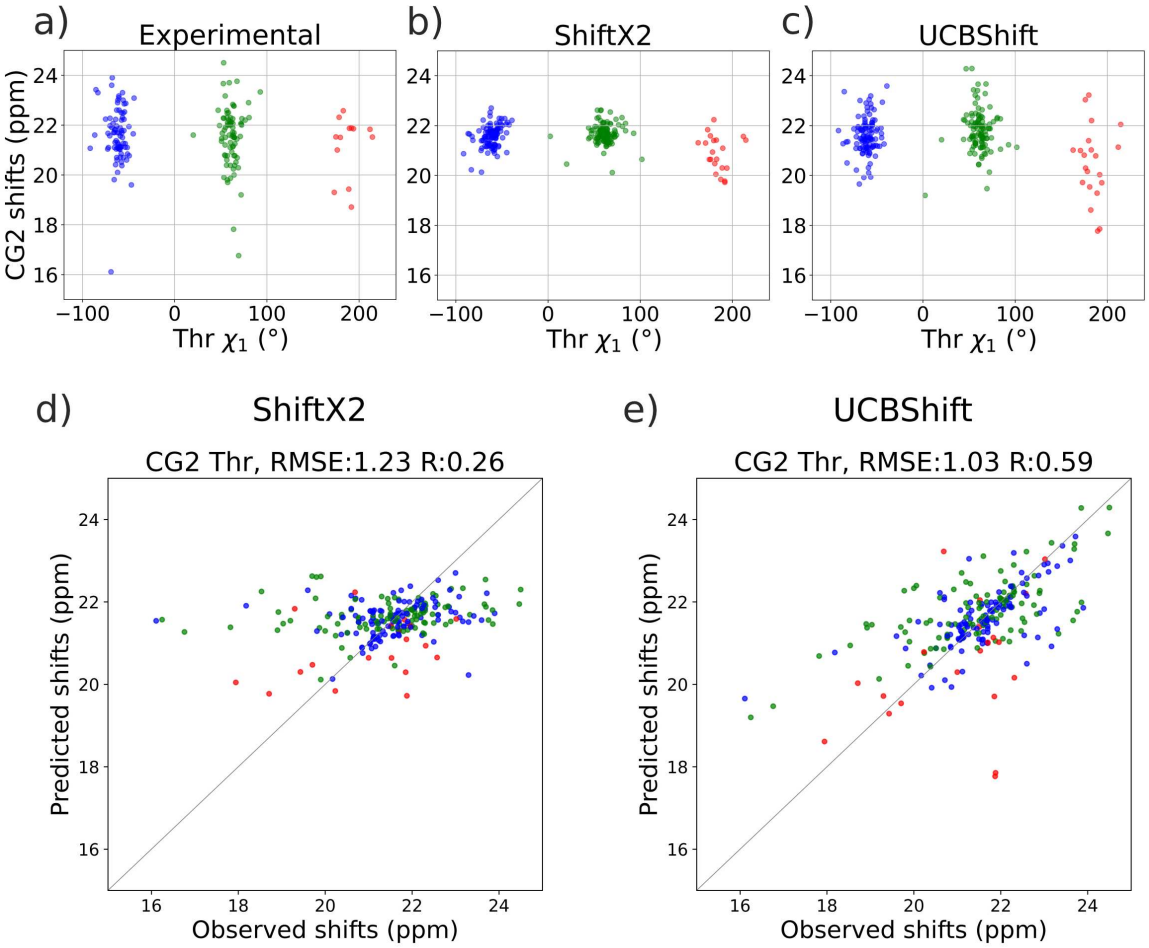


Figure S3: *Predicted CG2 chemical shifts of threonine by UCBShift and SHIFTX2 compared to experiment.* (a-c) CG2 chemical shifts of threonine plotted against the χ_1 dihedral angle. Correlation between experimental and predicted CG2 chemical shifts of threonine for (d) SHIFTX2 and (e) UCBShift. Three conformational clusters are marked with different colours.

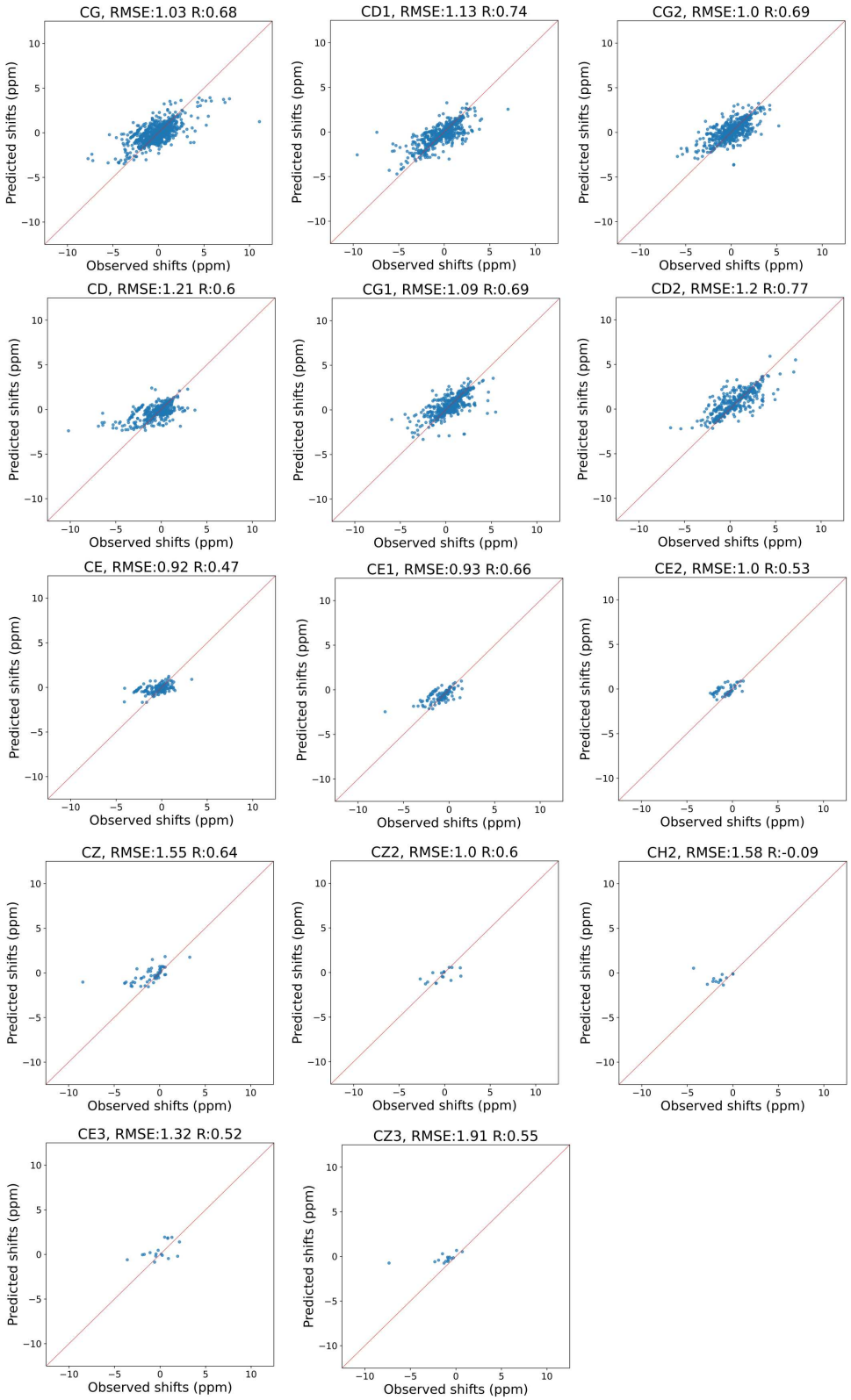


Figure S4: *Correlation plot of carbon chemical shifts with UCBShift compared to experiment. Chemical shifts are subtracted from the random coil reference.*

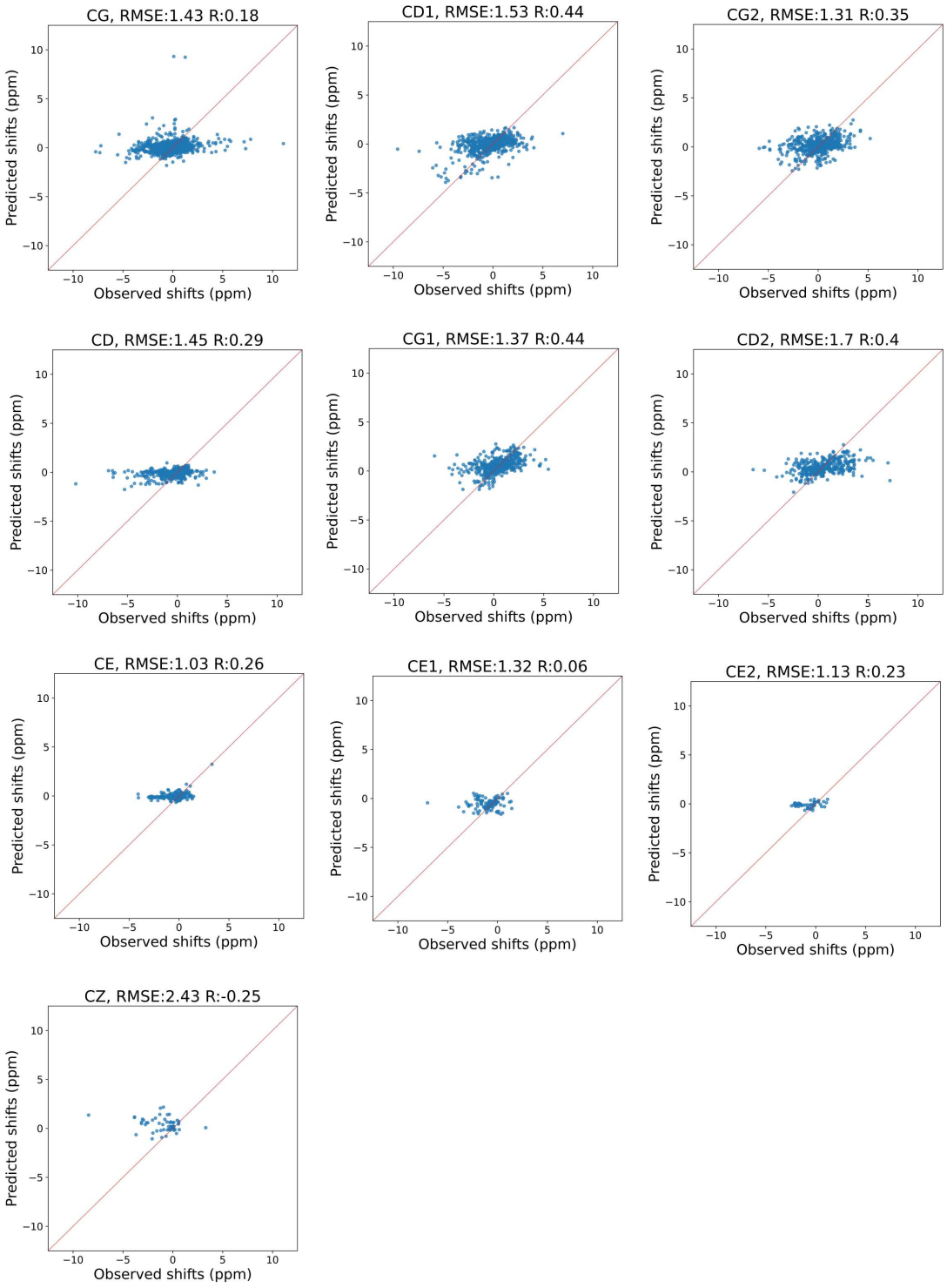


Figure S5: *Correlation plot of carbon chemical shifts with ShiftX2 compared to experiment. Chemical shifts are subtracted from the random coil reference.*

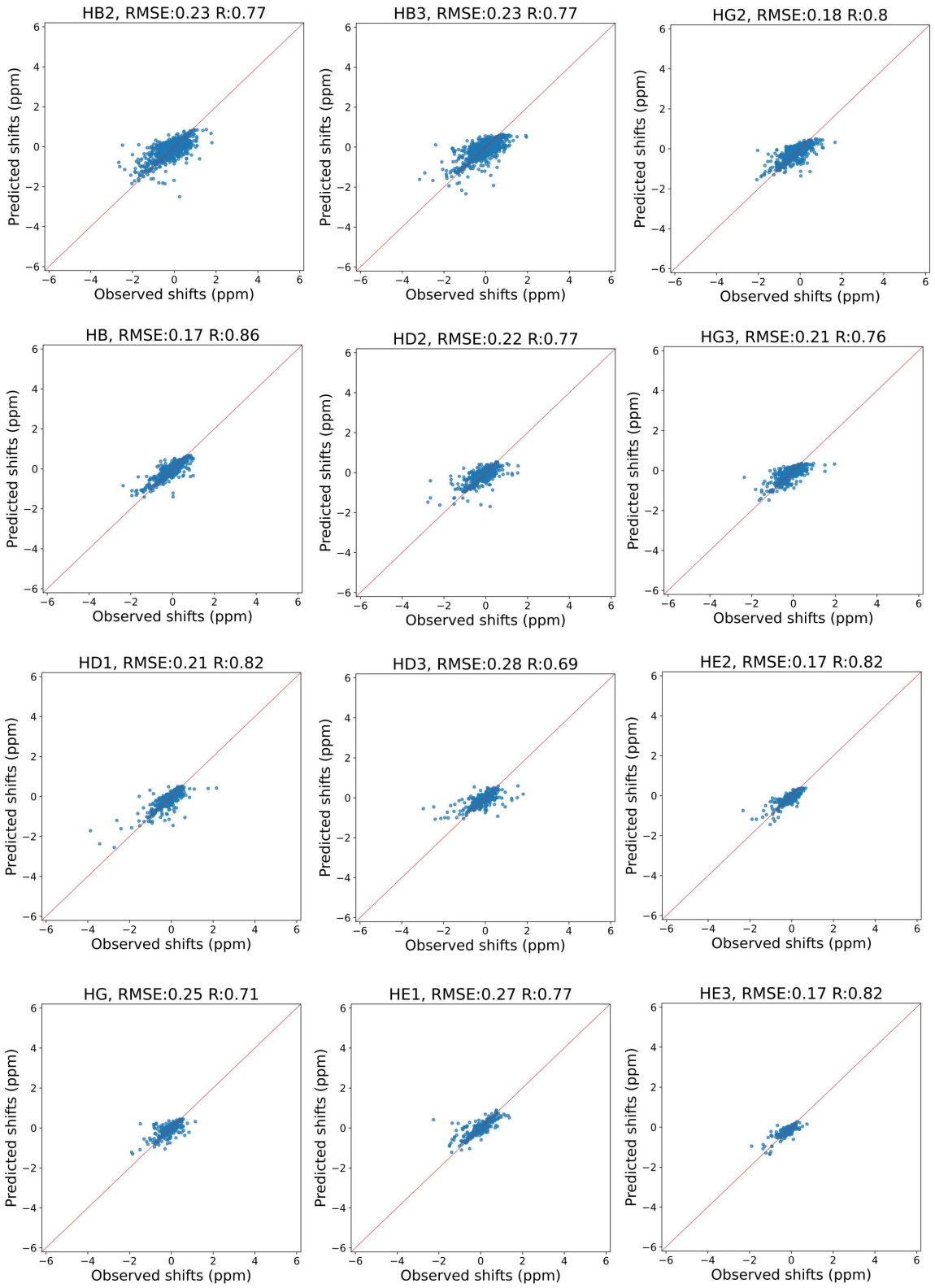


Figure S6: *Correlation plot of hydrogen chemical shifts with UCBShift compared to experiment. Chemical shifts are subtracted from the random coil reference.*

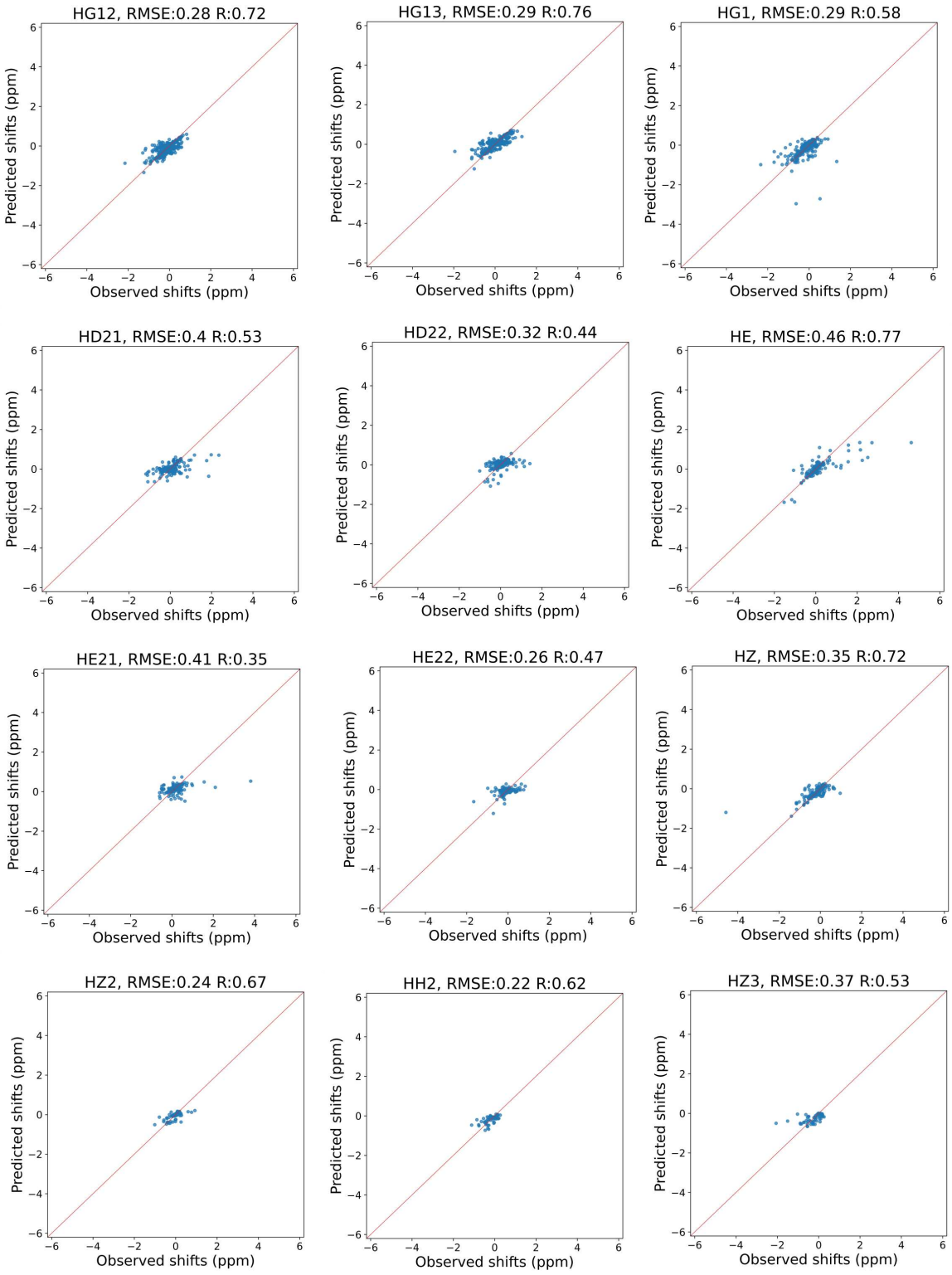


Figure S7: *Correlation plot of hydrogen chemical shifts with UCBShift compared to experiment (continued). Chemical shifts are subtracted from the random coil reference.*

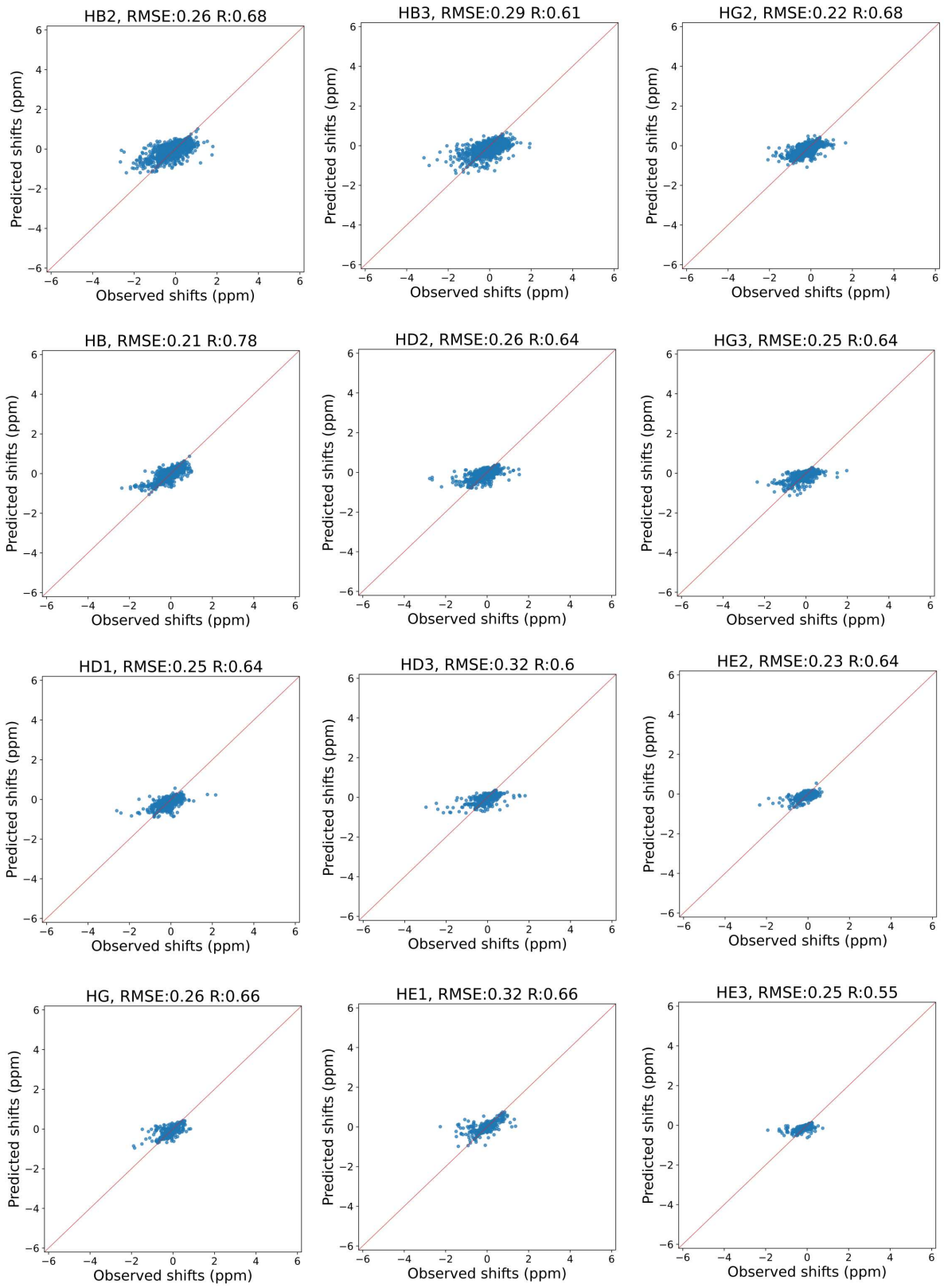


Figure S8: *Correlation plot of hydrogen chemical shifts with ShiftX2 compared to experiment. Chemical shifts are subtracted from the random coil reference.*

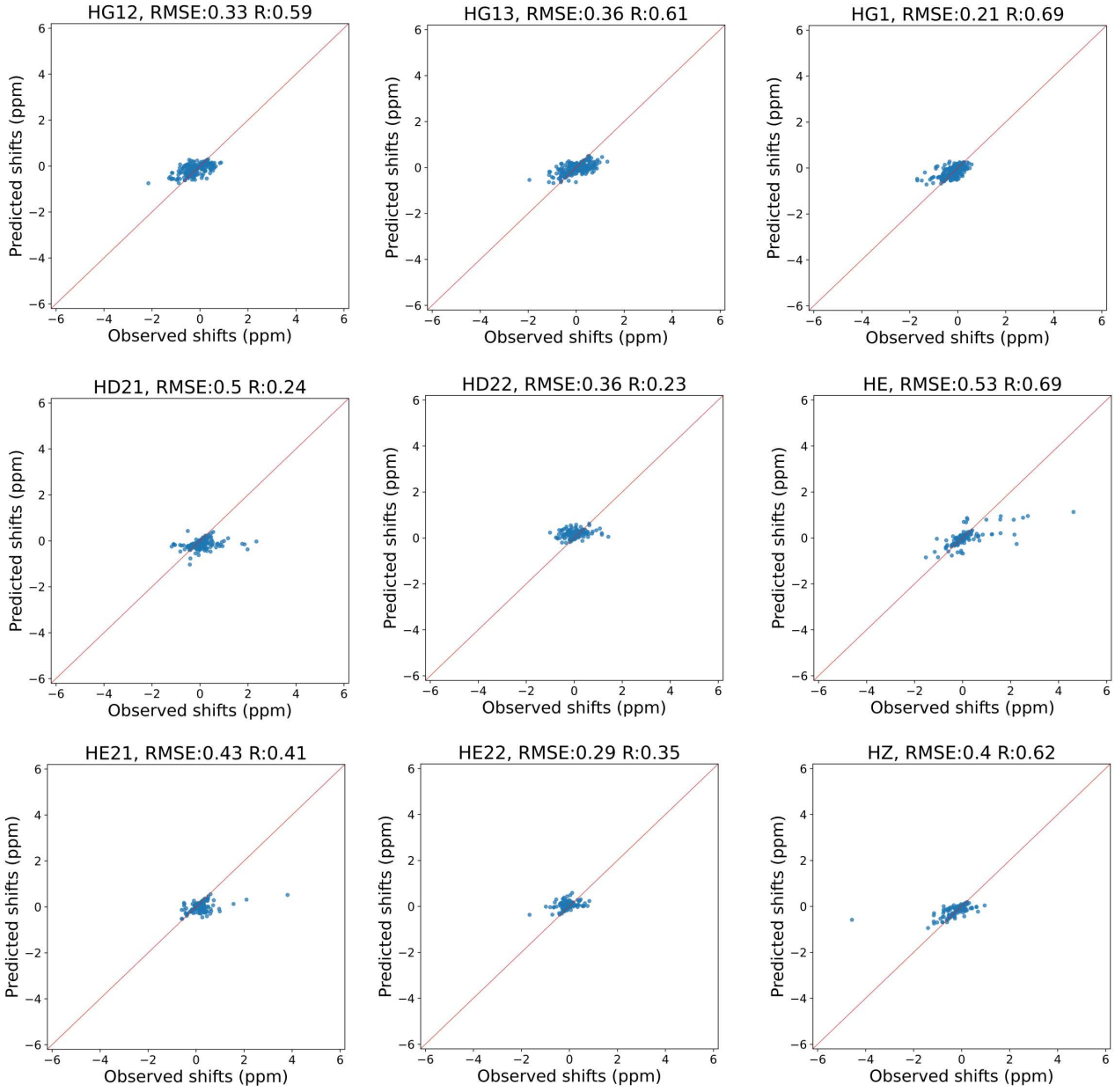


Figure S9: *Correlation plot of hydrogen chemical shifts with ShiftX2 compared to experiment (continued). Chemical shifts are subtracted from the random coil reference.*

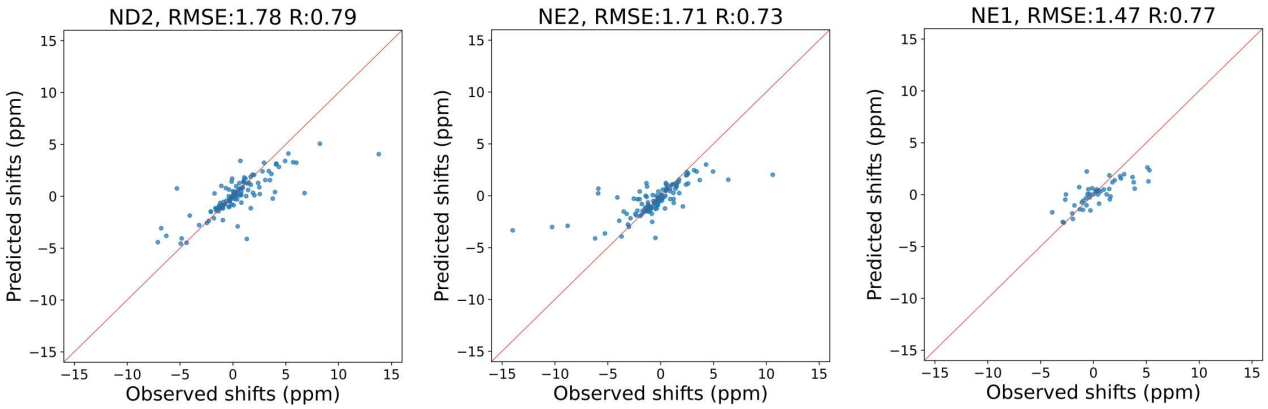


Figure S10: *Correlation plot of nitrogen chemical shifts with UCBShift compared to experiment. Chemical shifts are subtracted from the random coil reference.*

Part III

DISCUSSION AND OUTLOOK

6

DISCUSSION AND OUTLOOK

This thesis contributes to the broader interdisciplinary effort of developing computational and NMR tools to enhance and extend the capabilities of structure-based drug design. A crucial step in this process is to obtain structural information on the protein and its protein-ligand complexes. While X-ray crystallography remains the gold standard for high-resolution structures, its application is often limited by the need for crystalline samples, making it impractical for highly flexible biomolecules. NMR spectroscopy offers an alternative, providing structural information on proteins in solution, as well as insights into their dynamics and molecular interactions. However, a major challenge in NMR analysis is the labour-intensive and technically demanding process of assigning observed chemical shifts to their corresponding nuclei. To address this, accurate chemical shift predictors are essential, as they can significantly facilitate and accelerate the peak assignment process, making NMR-based structural studies more efficient and accessible [114]–[116].

The thesis presents UCBShift 2.0, an extension of UCBShift 1.0, which expands chemical shift prediction beyond the backbone to the whole protein [85]. The model integrates a feature-based machine learning algorithm with a chemical shift transfer module sourced from an experimental database. In benchmark tests, it outperforms SHIFTX2, the only other existing whole-protein chemical shift predictor. Notably, recent studies have demonstrated that integrating UCBShift 1.0 with AlphaFold-derived structures enhances the accuracy of backbone NMR assignment [116]. Despite these advancements, sidechain assignment remains a challenge due to the high conformational flexibility of side chains in solution and the influence of environmental effects. Our current work focuses on applying UCBShift 2.0 to the assignment of Valine and Leucine methyl groups using isotope labelling strategies. Given that ring current effects, one of the primary factors influencing methyl chemical shifts, are well captured by UCBShift 2.0, particularly through its UCBShift-X module, the model has strong potential for methyl group assignment without requiring structural similarity corrections. Moreover, since UCBShift 2.0 predicts chemical shifts for nearly all atom types, its application could be extended to other amino acid side chains, further broadening its utility in NMR-based protein studies.

Another crucial aspect of drug discovery is understanding how a drug interacts with its target in solution. While high-quality X-ray crystal structures provide the most reliable structural models of

biomolecules [117], [118], they represent a static snapshot that does not fully capture the conformational flexibility of proteins and ligands in solution. To incorporate dynamic information, we utilised ^{1}H chemical shifts of the ligand in the protein-bound state, which are straightforward to obtain and assign. Validating quantum mechanics/molecular mechanics (QM/MM)-calculated chemical shifts against experimental data allowed us to identify interactions that differ between the crystalline and solution states. To account for solution-state dynamics, we *refined* X-ray structures with QM/MM molecular dynamics (MD), successfully predicting chemical shifts from the ensemble of conformations. This approach confirms that while X-ray structures serve as a strong starting point, they often require further computational refinement to accurately represent solution-state interactions. This methodology could be extended to AlphaFold-predicted structures, incorporating ligand docking to gain insights into binding interfaces without relying on experimental structures. There are initial studies demonstrating the potential of QM/MM chemical shift predictions for molecular docking validation [119], [120], but experimental NMR data for protein-ligand systems remain limited, making large-scale validation of this approach challenging. Additionally, data-driven methods could simplify chemical shift predictions in protein-ligand systems, offering a more accessible alternative to purely quantum-mechanical approaches. Furthermore, AI-driven machine learning force fields, a rapidly growing area of research, may enable the generation of more reliable structural ensembles with improved solvent representation, overcoming the limitations of classical force fields [121].

Beyond structural determination, drug design could also benefit from a detailed understanding of the forces driving protein-ligand interactions. Traditionally, these interactions are described in terms of hydrogen bonding, hydrophobic contacts and π-stacking. However, modern chemical bonding and energy decomposition descriptors offer a more detailed perspective, allowing the identification of weaker hydrogen bonds and quantification of their role in the overall interaction network. We considered the streptavidin-biotin system as one of the most strongly bound protein-ligand complexes. After validating the structure by ^{1}H chemical shifts, we investigated the interactions that stabilise the complex, unravelling the important role of weak interactions like CH$\cdots\pi$, CH$\cdots$HC and CH$\cdots$S. Analysis of interactions with individual amino acids can offer valuable guidance on how to modify a ligand to fine-tune its binding affinity and specificity [122]. Energy decomposition databases for ligand fragments interacting with amino acids could be of great benefit to medicinal chemists.

BIBLIOGRAPHY

[1] G. R. Fulmer, A. J. Miller, N. H. Sherden, *et al.*, "Nmr chemical shifts of trace impurities: Common laboratory solvents, organics, and gases in deuterated solvents relevant to the organometallic chemist," *Organometallics*, vol. 29, no. 9, pp. 2176–2179, 2010.

[2] E. D. Becker, *High resolution NMR: theory and chemical applications*. Elsevier, 1999.

[3] A. C. de Dios, "Ab initio calculations of the nmr chemical shift," *Progress in Nuclear Magnetic Resonance Spectroscopy*, vol. 29, no. 3-4, pp. 229–278, 1996.

[4] N. F. Ramsey, "Magnetic shielding of nuclei in molecules," *Physical Review*, vol. 78, no. 6, p. 699, 1950.

[5] C. J. Jameson, "Understanding nmr chemical shifts," *Annual review of physical chemistry*, vol. 47, no. 1, pp. 135–169, 1996.

[6] C. M. Widdifield and R. W. Schurko, "Understanding chemical shielding tensors using group theory, mo analysis, and modern density-functional theory," *Concepts in Magnetic Resonance Part A: An Educational Journal*, vol. 34, no. 2, pp. 91–123, 2009.

[7] C. J. Jameson and H. Gutowsky, "Calculation of chemical shifts. i. general formulation and the z dependence," *The Journal of Chemical Physics*, vol. 40, no. 6, pp. 1714–1724, 1964.

[8] W. Lamb Jr, "Internal diamagnetic fields," *Physical Review*, vol. 60, no. 11, p. 817, 1941.

[9] T. Helgaker, M. Jaszunski, and K. Ruud, "Ab initio methods for the calculation of nmr shielding and indirect spin-spin coupling constants," *Chemical Reviews*, vol. 99, pp. 293–352, 1999.

[10] F. London, "Théorie quantique des courants interatomiques dans les combinaisons aromatiques," *J. phys. radium*, vol. 8, no. 10, pp. 397–409, 1937.

[11] R. Ditchfield, "Self-consistent perturbation theory of diamagnetism: I. a gauge-invariant lcao method for nmr chemical shifts," *Molecular Physics*, vol. 27, no. 4, pp. 789–807, 1974.

[12] T. A. Keith and R. F. Bader, "Calculation of magnetic response properties using a continuous set of gauge transformations," *Chemical physics letters*, vol. 210, no. 1-3, pp. 223–231, 1993.

[13] J. R. Cheeseman, G. W. Trucks, T. A. Keith, and M. J. Frisch, "A comparison of models for calculating nuclear magnetic resonance shielding tensors," *The Journal of chemical physics*, vol. 104, no. 14, pp. 5497–5509, 1996.

[14] R. K. Harris, P. Hodgkinson, C. J. Pickard, J. R. Yates, and V. Zorin, "Chemical shift computations on a crystallographic basis: Some reflections and comments," *Magnetic Resonance in Chemistry*, vol. 45, no. S1, S174–S186, 2007.

[15] W. Kutzelnigg, U. Fleischer, and M. Schindler, "The iglo-method: Ab initio calculation and interpretation of nmr chemical shifts and magnetic susceptibilities," *NMR basic principles and progress*, vol. 23, pp. 165–262, 1990.

[16] M. W. Lodewyk, M. R. Siebert, and D. J. Tantillo, "Computational prediction of 1h and 13c chemical shifts: A useful tool for natural product, mechanistic, and synthetic organic chemistry," *Chemical Reviews*, vol. 112, no. 3, pp. 1839–1862, 2012.

[17] D. Chesnut, "Some recent ab initio calculations of the nmr chemical shift," in *Annual reports on NMR spectroscopy*, vol. 21, Elsevier, 1989, pp. 51–97.

[18] K. Wolinski, J. F. Hinton, and P. Pulay, "Efficient implementation of the gauge-independent atomic orbital method for nmr chemical shift calculations," *Journal of the American Chemical Society*, vol. 112, no. 23, pp. 8251–8260, 1990.

[19] J. Gauss and J. F. Stanton, "Coupled-cluster calculations of nuclear magnetic resonance chemical shifts," *The Journal of chemical physics*, vol. 103, no. 9, pp. 3561–3577, 1995.

[20] J. Gauss and J. F. Stanton, "Perturbative treatment of triple excitations in coupled-cluster calculations of nuclear magnetic shielding constants," *The Journal of chemical physics*, vol. 104, no. 7, pp. 2574–2583, 1996.

[21] T. Helgaker, S. Coriani, P. Jørgensen, *et al.*, "Recent advances in wave function-based methods of molecular-property calculations," *Chemical reviews*, vol. 112, no. 1, pp. 543–631, 2012.

[22] G. Schreckenbach and T. Ziegler, "Calculation of nmr shielding tensors using gauge-including atomic orbitals and modern density functional theory," *The Journal of Physical Chemistry*, vol. 99, no. 2, pp. 606–611, 1995.

[23] S. Wolff and T Ziegler, "Calculation of dft-giao nmr shifts with the inclusion of spin-orbit coupling," *The Journal of chemical physics*, vol. 109, no. 3, pp. 895–905, 1998.

[24] C. J. Schattenberg and M. Kaupp, "Extended benchmark set of main-group nuclear shielding constants and nmr chemical shifts and its use to evaluate modern dft methods," *Journal of Chemical Theory and Computation*, vol. 17, no. 12, pp. 7602–7621, 2021.

[25] D. Flaig, M. Maurer, M. Hanni, *et al.*, "Benchmarking hydrogen and carbon nmr chemical shifts at hf, dft, and mp2 levels," *Journal of Chemical Theory and Computation*, vol. 10, no. 2, pp. 572–578, 2014.

[26] E. Benassi, "Benchmarking of density functionals for a soft but accurate prediction and assignment of 1h and 13c nmr chemical shifts in organic and biological molecules," *Journal of computational chemistry*, vol. 38, no. 2, pp. 87–92, 2017.

[27] T. H. Sefzik, D. Turco, R. J. Iuliucci, and J. C. Facelli, "Modeling nmr chemical shift: A survey of density functional theory approaches for calculating tensor properties," *The Journal of Physical Chemistry A*, vol. 109, no. 6, pp. 1180–1187, 2005.

[28] J. B. Stückrath, T. Gasevic, M. Bursch, and S. Grimme, "Benchmark study on the calculation of 119sn nmr chemical shifts," *Inorganic Chemistry*, vol. 61, no. 9, pp. 3903–3917, 2022.

[29] M. Bühl and T. van Mourik, "Nmr spectroscopy: Quantum-chemical calculations," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 1, no. 4, pp. 634–647, 2011.

[30] M. E. Harding, J. Gauss, and P. v. R. Schleyer, "Why benchmark-quality computations are needed to reproduce 1-adamantyl cation nmr chemical shifts accurately," *The Journal of Physical Chemistry A*, vol. 115, no. 11, pp. 2340–2344, 2011.

[31] Y. Zhang, A. Wu, X. Xu, and Y. Yan, "Opbe: A promising density functional for the calculation of nuclear shielding constants," *Chemical Physics Letters*, vol. 421, no. 4-6, pp. 383–388, 2006.

[32] M. Bühl, "Dft computations of transition-metal chemical shifts," *Annual Reports on NMR Spectroscopy*, vol. 64, pp. 77–126, 2008.

[33] J Autschbach and S Zheng, "Relativistic computations of nmr parameters from first principles: Theory and applications," *Annual reports on NMR spectroscopy*, vol. 67, pp. 1–95, 2009.

[34] J. Vıcha, J. Novotny, S. Komorovsky, *et al.*, "Relativistic heavy-neighbor-atom effects on nmr shifts: Concepts and trends across the periodic table," *Chemical reviews*, vol. 120, no. 15, pp. 7065–7103, 2020.

[35] S. Grimme, A. Hansen, S. Ehlert, and J.-M. Mewes, "R2scan-3c: A "swiss army knife" composite electronic-structure method," *The Journal of Chemical Physics*, vol. 154, no. 6, 2021.

[36] M. Müller, A. Hansen, and S. Grimme, "ωB97x-3c: A composite range-separated hybrid dft method with a molecule-optimized polarized valence double-ζ basis set," *The Journal of Chemical Physics*, vol. 158, no. 1, 2023.

[37] J. Liang, Z. Wang, J. Li, *et al.*, "Efficient calculation of nmr shielding constants using composite method approximations and locally dense basis sets," *Journal of chemical theory and computation*, vol. 19, no. 2, pp. 514–523, 2023.

[38] F. Jensen, "Segmented contracted basis sets optimized for nuclear magnetic shielding," *Journal of Chemical Theory and Computation*, vol. 11, no. 1, pp. 132–138, 2015.

[39] A. Warshel and M. Levitt, "Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme," *Journal of molecular biology*, vol. 103, no. 2, pp. 227–249, 1976.

[40] H. M. Senn and W. Thiel, "Qm/mm methods for biomolecular systems," *Angewandte Chemie International Edition*, vol. 48, no. 7, pp. 1198–1229, 2009.

[41] Q. Cui, T. Pal, and L. Xie, "Biomolecular qm/mm simulations: What are some of the "burning issues"?" *The Journal of Physical Chemistry B*, vol. 125, no. 3, pp. 689–702, 2021.

[42] Q. Cui and M. Karplus, "Molecular properties from combined qm/mm methods. 2. chemical shifts in large molecules," *The Journal of Physical Chemistry B*, vol. 104, no. 15, pp. 3721–3743, 2000.

[43] B. Wang and K. M. Merz, "A fast qm/mm (quantum mechanical/molecular mechanical) approach to calculate nuclear magnetic resonance chemical shifts for macromolecules," *Journal of chemical theory and computation*, vol. 2, no. 1, pp. 209–215, 2006.

[44] X. He, B. Wang, and K. M. Merz Jr, "Protein nmr chemical shift calculations based on the automated fragmentation qm/mm approach," *The Journal of Physical Chemistry B*, vol. 113, no. 30, pp. 10 380–10 388, 2009.

[45] T. Zhu, J. Z. Zhang, and X. He, "Automated fragmentation qm/mm calculation of amide proton chemical shifts in proteins with explicit solvent model," *Journal of chemical theory and computation*, vol. 9, no. 4, pp. 2104–2114, 2013.

[46] J. Swails, T. Zhu, X. He, and D. A. Case, "Afnmr: Automated fragmentation quantum mechanical calculation of nmr chemical shifts for biomolecules," *Journal of biomolecular NMR*, vol. 63, pp. 125–139, 2015.

[47] X. Jin, T. Zhu, J. Z. Zhang, and X. He, "A systematic study on rna nmr chemical shift calculation based on the automated fragmentation qm/mm approach," *RSC advances*, vol. 6, no. 110, pp. 108 590–108 602, 2016.

[48] S. Grimme, C. Bannwarth, S. Dohm, *et al.*, "Fully automated quantum-chemistry-based computation of spin–spin-coupled nuclear magnetic resonance spectra," *Angewandte Chemie International Edition*, vol. 56, no. 46, pp. 14 763–14 769, 2017.

[49] P. Pracht, F. Bohle, and S. Grimme, "Automated exploration of the low-energy chemical space with fast quantum chemical methods," *Physical Chemistry Chemical Physics*, vol. 22, no. 14, pp. 7169–7192, 2020.

[50] A. A. Auer, J. Gauss, and J. F. Stanton, "Quantitative prediction of gas-phase c13 nuclear magnetic shielding constants," *The Journal of chemical physics*, vol. 118, no. 23, pp. 10 407–10 417, 2003.

[51] P. Robustelli, K. A. Stafford, and A. G. Palmer III, "Interpreting protein structural dynamics from nmr chemical shifts," *Journal of the American Chemical Society*, vol. 134, no. 14, pp. 6365–6374, 2012.

[52] T. E. Exner, A. Frank, I. Onila, and H. M. Möller, "Toward the quantum chemical calculation of nmr chemical shifts of proteins. 3. conformational sampling and explicit solvents model," *Journal of Chemical Theory and Computation*, vol. 8, no. 11, pp. 4818–4827, 2012.

[53] J. Meiler and M. Will, "Genius: A genetic algorithm for automated structure elucidation from 13c nmr spectra," *Journal of the American Chemical Society*, vol. 124, no. 9, pp. 1868–1870, 2002.

[54] M. Elyashberg, A. J. Williams, and K. Blinov, "Structural revisions of natural products by computer-assisted structure elucidation (case) systems," *Natural Product Reports*, vol. 27, no. 9, pp. 1296–1328, 2010.

[55] A. Cavalli, X. Salvatella, C. M. Dobson, and M. Vendruscolo, "Protein structure determination from nmr chemical shifts," *Proceedings of the National Academy of Sciences*, vol. 104, no. 23, pp. 9615–9620, 2007.

[56] S. P. Mielke and V. V. Krishnan, "Characterization of protein secondary structure from nmr chemical shifts," *Progress in nuclear magnetic resonance spectroscopy*, vol. 54, no. 3-4, pp. 141–165, 2009.

[57] M. Dracinsky and P. Bour, "Computational analysis of solvent effects in nmr spectroscopy," *Journal of Chemical Theory and Computation*, vol. 6, no. 1, pp. 288–299, 2010.

[58] V. Palivec, R. Pohl, J. Kaminský, and H. Martinez-Seara, "Efficiently computing nmr 1h and 13c chemical shifts of saccharides in aqueous environment," *Journal of Chemical Theory and Computation*, vol. 18, no. 7, pp. 4373–4386, 2022.

[59] S. K. Chandy, B. Thapa, and K. Raghavachari, "Accurate and cost-effective nmr chemical shift predictions for proteins using a molecules-in-molecules fragmentation-based method," *Physical Chemistry Chemical Physics*, vol. 22, no. 47, pp. 27 781–27 799, 2020.

[60] R. Atwi, Y. Chen, K. S. Han, *et al.*, "An automated framework for high-throughput predictions of nmr chemical shifts within liquid solutions," *Nature Computational Science*, vol. 2, no. 2, pp. 112–122, 2022.

[61] M. J. Bakker, A. Mládek, H. Semrád, V. Zapletal, and J. P. Přecechtělová, "Improving idp theoretical chemical shift accuracy and efficiency through a combined md/adma/dft and machine learning approach," *Physical Chemistry Chemical Physics*, vol. 24, no. 45, pp. 27 678–27 692, 2022.

[62] M. Dracinsky, H. M. Möller, and T. E. Exner, "Conformational sampling by ab initio molecular dynamics simulations improves nmr chemical shift predictions," *Journal of chemical theory and computation*, vol. 9, no. 8, pp. 3806–3815, 2013.

[63] A. H. Mazurek, Ł. Szeleszczuk, and D. M. Pisklak, "A review on combination of ab initio molecular dynamics and nmr parameters calculations," *International Journal of Molecular Sciences*, vol. 22, no. 9, p. 4378, 2021.

[64] M. Schindler and W. Kutzelnigg, "Theory of magnetic susceptibilities and nmr chemical shifts in terms of localized quantities. ii. application to some simple molecules," *The Journal of Chemical Physics*, vol. 76, no. 4, pp. 1919–1933, 1982.

[65] C. P. Gordon and C. Copéret, "Chemical shift tensors–why should we care?" *Chimia*, vol. 73, no. 4, pp. 252–252, 2019.

[66] R. V. Viesser, L. C. Ducati, C. F. Tormena, and J. Autschbach, "The unexpected roles of σ and π orbitals in electron donor and acceptor group effects on the 13 c nmr chemical shifts in substituted benzenes," *Chemical science*, vol. 8, no. 9, pp. 6570–6576, 2017.

[67] R. V. Viesser, L. C. Ducati, C. F. Tormena, and J. Autschbach, "The halogen effect on the 13 c nmr chemical shift in substituted benzenes," *Physical Chemistry Chemical Physics*, vol. 20, no. 16, pp. 11 247–11 259, 2018.

[68] P. Gilli, V. Bertolasi, V. Ferretti, G. Gilli, *et al.*, "Covalent nature of the strong homonuclear hydrogen bond. study of the oh–o system by crystal structure correlation methods," *Journal of the American Chemical Society*, vol. 116, no. 3, pp. 909–915, 1994.

[69] E. Isaacs, A Shukla, P. Platzman, *et al.*, "Covalency of the hydrogen bond in ice: A direct x-ray measurement," *Physical Review Letters*, vol. 82, no. 3, p. 600, 1999.

[70] C. Fonseca Guerra, F. M. Bickelhaupt, J. G. Snijders, and E. J. Baerends, "The nature of the hydrogen bond in dna base pairs: The role of charge transfer and resonance assistance," *Chemistry–A European Journal*, vol. 5, no. 12, pp. 3581–3594, 1999.

[71] G. Gilli and P. Gilli, "Towards an unified hydrogen-bond theory," *Journal of Molecular Structure*, vol. 552, no. 1-3, pp. 1–15, 2000.

[72] R. Kurczab, M. P. Mitoraj, A. Michalak, and T. Ziegler, "Theoretical analysis of the resonance assisted hydrogen bond based on the combined extended transition state method and natural orbitals for chemical valence scheme," *The Journal of Physical Chemistry A*, vol. 114, no. 33, pp. 8581–8590, 2010.

[73] M. P. Mitoraj, R. Kurczab, M. Boczar, and A. Michalak, "Theoretical description of hydrogen bonding in oxalic acid dimer and trimer based on the combined extended-transition-state energy decomposition analysis and natural orbitals for chemical valence (ets-nocv)," *Journal of molecular modeling*, vol. 16, pp. 1789–1795, 2010.

[74] M. Pecul, J. Leszczynski, and J. Sadlej, "Comprehensive ab initio studies of nuclear magnetic resonance shielding and coupling constants in xh$\cdots$ o hydrogen-bonded complexes of simple organic molecules," *The Journal of Chemical Physics*, vol. 112, no. 18, pp. 7930–7938, 2000.

[75] T. Tuttle, J. Gräfenstein, A. Wu, E. Kraka, and D. Cremer, "Analysis of the nmr spin- spin coupling mechanism across a h- bond: Nature of the h-bond in proteins," *The Journal of Physical Chemistry B*, vol. 108, no. 3, pp. 1115–1129, 2004.

[76] M. P. Mitoraj, A. Michalak, and T. Ziegler, "A combined charge and energy decomposition scheme for bond analysis," *Journal of chemical theory and computation*, vol. 5, no. 4, pp. 962–975, 2009.

[77] G. A. Kumar and M. A. McAllister, "Theoretical investigation of the relationship between proton nmr chemical shift and hydrogen bond strength," *The Journal of Organic Chemistry*, vol. 63, no. 20, pp. 6968–6972, 1998.

[78] K. Sutter, G. A. Aucar, and J. Autschbach, "Analysis of proton nmr in hydrogen bonds in terms of lone-pair and bond orbital contributions," *Chemistry–A European Journal*, vol. 21, no. 50, pp. 18 138–18 155, 2015.

[79] R. F. de Freitas and M. Schapira, "A systematic analysis of atomic protein–ligand interactions in the pdb," *Medchemcomm*, vol. 8, no. 10, pp. 1970–1981, 2017.

[80] M. Nishio, "The ch/π hydrogen bond in chemistry. conformation, supramolecules, optical resolution and interactions involving carbohydrates," *Physical Chemistry Chemical Physics*, vol. 13, no. 31, pp. 13 873–13 900, 2011.

[81] M. J. Plevin, D. L. Bryce, and J. Boisbouvier, "Direct detection of ch/π interactions in proteins," *Nature chemistry*, vol. 2, no. 6, pp. 466–471, 2010.

[82] S. Scheiner, "Assessment of the presence and strength of h-bonds by means of corrected nmr," *Molecules*, vol. 21, no. 11, p. 1426, 2016.

[83] Y. Shen and A. Bax, "Sparta+: A modest improvement in empirical nmr chemical shift prediction by means of an artificial neural network," *Journal of biomolecular NMR*, vol. 48, pp. 13–22, 2010.

[84] M. Darrows, D. Kodituwakku, J. Xue, *et al.*, "Legolas: A machine learning method for rapid and accurate predictions of protein nmr chemical shifts.," 2025.

[85] J. Li, K. C. Bennett, Y. Liu, M. V. Martin, and T. Head-Gordon, "Accurate prediction of chemical shifts for aqueous protein structure on "real world" data," *Chemical science*, vol. 11, no. 12, pp. 3180–3191, 2020.

[86] S. Liu, J. Li, K. C. Bennett, *et al.*, "Multiresolution 3d-densenet for chemical shift prediction in nmr crystallography," *The journal of physical chemistry letters*, vol. 10, no. 16, pp. 4558–4565, 2019.

[87] Y. Guan, S. S. Sowndarya, L. C. Gallegos, P. C. S. John, and R. S. Paton, "Real-time prediction of 1 h and 13 c chemical shifts with dft accuracy using a 3d graph neural network," *Chemical Science*, vol. 12, no. 36, pp. 12 012–12 026, 2021.

[88] M. Haghighatlari, J. Li, F. Heidar-Zadeh, *et al.*, "Learning to make chemical predictions: The interplay of feature representation, data, and machine learning methods," *Chem*, vol. 6, no. 7, pp. 1527–1542, 2020.

[89] P. A. Unzueta, C. S. Greenwell, and G. J. Beran, "Predicting density functional theory-quality nuclear magnetic resonance chemical shifts via δ-machine learning," *Journal of Chemical Theory and Computation*, vol. 17, no. 2, pp. 826–840, 2021.

[90] J. B. Kleine Bu"ning and S. Grimme, "Computation of ccsd (t)-quality nmr chemical shifts via Δ-machine learning from dft," *Journal of Chemical Theory and Computation*, vol. 19, no. 12, pp. 3601–3615, 2023.

[91] H. Zhang, S. Neal, and D. S. Wishart, "Refdb: A database of uniformly referenced protein chemical shifts," *Journal of biomolecular NMR*, vol. 25, pp. 173–195, 2003.

[92] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. Von Lilienfeld, "Fast and accurate modeling of molecular atomization energies with machine learning," *Physical review letters*, vol. 108, no. 5, p. 058 301, 2012.

[93] M. Rupp, R. Ramakrishnan, and O. A. Von Lilienfeld, "Machine learning for quantum mechanical properties of atoms in molecules," *The Journal of Physical Chemistry Letters*, vol. 6, no. 16, pp. 3309–3313, 2015.

[94] J. S. Smith, O. Isayev, and A. E. Roitberg, "Ani-1: An extensible neural network potential with dft accuracy at force field computational cost," *Chemical science*, vol. 8, no. 4, pp. 3192–3203, 2017.

[95] D. S. Wishart, B. D. Sykes, and F. M. Richards, "Relationship between nuclear magnetic resonance chemical shift and protein secondary structure," *Journal of molecular biology*, vol. 222, no. 2, pp. 311–333, 1991.

[96] G. Wagner, A. Pardi, and K. Wuethrich, "Hydrogen bond length and proton nmr chemical shifts in proteins," *Journal of the American Chemical Society*, vol. 105, no. 18, pp. 5948–5949, 1983.

[97] C. Haigh and R. Mallion, "Ring current theories in nuclear magnetic resonance," *Progress in nuclear magnetic resonance spectroscopy*, vol. 13, no. 4, pp. 303–344, 1979.

[98] B. Han, Y. Liu, S. W. Ginzinger, and D. S. Wishart, "Shiftx2: Significantly improved protein chemical shift prediction," *Journal of biomolecular NMR*, vol. 50, pp. 43–57, 2011.

[99] C. Han, D. Zhang, S. Xia, and Y. Zhang, "Accurate prediction of nmr chemical shifts: Integrating dft calculations with three-dimensional graph neural networks," *Journal of Chemical Theory and Computation*, 2024.

[100] P. Gao, J. Zhang, Q. Peng, J. Zhang, and V.-A. Glezakou, "General protocol for the accurate prediction of molecular 13c/1h nmr chemical shifts via machine learning augmented dft," *Journal of Chemical Information and Modeling*, vol. 60, no. 8, pp. 3746–3754, 2020.

[101] J. Gauss, "Analytic second derivatives for the full coupled-cluster singles, doubles, and triples model: Nuclear magnetic shielding constants for bh, hf, co, n 2, n 2 o, and o 3," *The Journal of chemical physics*, vol. 116, no. 12, pp. 4773–4776, 2002.

[102] F. M. Paruzzo, A. Hofstetter, F. Musil, *et al.*, "Chemical shifts in molecular solids by machine learning," *Nature communications*, vol. 9, no. 1, p. 4501, 2018.

[103] W. Gerrard, L. A. Bratholm, M. J. Packer, *et al.*, "Impression–prediction of nmr parameters for 3-dimensional chemical structures using machine learning with near quantum chemical accuracy," *Chemical science*, vol. 11, no. 2, pp. 508–515, 2020.

[104] J. Li, J. Liang, Z. Wang, *et al.*, "Highly accurate prediction of nmr chemical shifts from low-level quantum mechanics calculations using machine learning," *Journal of Chemical Theory and Computation*, vol. 20, no. 5, pp. 2152–2166, 2024.

[105] A. Frank, I. Onila, H. M. Möller, and T. E. Exner, "Toward the quantum chemical calculation of nuclear magnetic resonance chemical shifts of proteins," *Proteins: Structure, Function, and Bioinformatics*, vol. 79, no. 7, pp. 2189–2202, 2011.

[106] A. Frank, H. M. Moöller, and T. E. Exner, "Toward the quantum chemical calculation of nmr chemical shifts of proteins. 2. level of theory, basis set, and solvents model dependence," *Journal of Chemical Theory and Computation*, vol. 8, no. 4, pp. 1480–1492, 2012.

[107] C. V. Sumowski, M. Hanni, S. Schweizer, and C. Ochsenfeld, "Sensitivity of ab initio vs empirical methods in computing structural effects on nmr chemical shifts for the example of peptides," *Journal of Chemical Theory and Computation*, vol. 10, no. 1, pp. 122–133, 2014.

[108] A. L. Ptaszek, J. Li, R. Konrat, G. Platzer, and T. Head-Gordon, "Ucbshift 2.0: Bridging the gap from backbone to side chain protein chemical shift prediction for protein structures," *Journal of the American Chemical Society*, vol. 146, no. 46, pp. 31 733–31 745, 2024.

[109] T. R. Makin and J.-J. Orban de Xivry, "Ten common statistical mistakes to watch out for when writing or reviewing a manuscript," *Elife*, vol. 8, e48175, 2019.

[110] S. Neal, A. M. Nip, H. Zhang, and D. S. Wishart, "Rapid and accurate calculation of protein 1 h, 13 c and 15 n chemical shifts," *Journal of biomolecular NMR*, vol. 26, pp. 215–240, 2003.

[111] D. A. Case, "Using quantum chemistry to estimate chemical shifts in biomolecules," *Biophysical chemistry*, vol. 267, p. 106 476, 2020.

[112] G. Platzer, A. L. Ptaszek, J. Böttcher, *et al.*, "Ligand 1h nmr chemical shifts as accurate reporters for protein-ligand binding interfaces in solution," *ChemPhysChem*, vol. 25, no. 1, e202300636, 2024.

[113] A. L. Ptaszek, S. Kratzwald, F. Sagan, *et al.*, "From weak interactions to strong affinity: Deciphering the streptavidin-biotin interaction through nmr and computational analysis," *bioRxiv*, pp. 2024–12, 2024.

[114] P. Klukowski, R. Riek, and P. Güntert, "Rapid protein assignments and structures from raw nmr spectra with the deep learning technique artina," *Nature Communications*, vol. 13, no. 1, p. 6151, 2022.

[115] H. Wetton, P. Klukowski, R. Riek, and P. Güntert, "Chemical shift transfer: An effective strategy for protein nmr assignment with artina," *Frontiers in Molecular Biosciences*, vol. 10, p. 1 244 029, 2023.

[116] P. Klukowski, R. Riek, and P. Güntert, "Time-optimized protein nmr assignment with an integrative deep learning approach using alphafold and chemical shift prediction," *Science Advances*, vol. 9, no. 47, eadi9323, 2023.

[117] M. Schneider, X. Fu, and A. E. Keating, "X-ray vs. nmr structures as templates for computational protein design," *Proteins: Structure, Function, and Bioinformatics*, vol. 77, no. 1, pp. 97–110, 2009.

[118] M. Berjanskii, Y. Liang, J. Zhou, *et al.*, "Prosess: A protein structure evaluation suite and server," *Nucleic acids research*, vol. 38, no. suppl_2, W633–W640, 2010.

[119] X. Jin, T. Zhu, J. Z. Zhang, and X. He, "Automated fragmentation qm/mm calculation of nmr chemical shifts for protein-ligand complexes," *Frontiers in Chemistry*, vol. 6, p. 150, 2018.

[120] Z. Yu, P. Li, and K. M. J. Merz, "Using ligand-induced protein chemical shift perturbations to determine protein–ligand structures," *Biochemistry*, vol. 56, no. 18, pp. 2349–2362, 2017.

[121] O. T. Unke, S. Chmiela, H. E. Sauceda, *et al.*, "Machine learning force fields," *Chemical Reviews*, vol. 121, no. 16, pp. 10 142–10 186, 2021.

[122] M. J. S. Phipps, T. Fox, C. S. Tautermann, and C.-K. Skylaris, "Energy decomposition analysis approaches and their evaluation on prototypical protein–drug interaction patterns," *Chem. Soc. Rev.*, vol. 44, pp. 3177–3211, 10 2015.