Predicting Protein NMR Chemical Shifts in the Presence of Ligands and Ions using Force Field-based Features

Anna Katharina Dehof [1,*], Hans-Peter Lenhof [1], Andreas Hildebrandt [2]

[1]Center for Bioinformatics, Saarland University, 66041 Saarbrücken, Germany
[2]Chair for Software-Engineering and Bioinformatics,
Johannes-Gutenberg-University Mainz, 55128 Mainz, Germany

[*]anna.dehof@bioinf.uni-sb.de

Abstract: The computational prediction of Nuclear Magnetic Resonance (NMR) chemical shifts has been studied extensively in the last years. For small molecules, accurate quantum chemical methods can be applied. But for macromolecular systems, such as proteins, the best available techniques that are sufficiently efficient for high-throughput applications use a combination of semi-classical approximations and statistical models. In this work, we study the influence of the presence of ligand molecules or ions on chemical shift prediction of protein atoms. We test a number of ligand-related properties for correlation with shift deviation and describe how a statistical model for the influence of ligand atoms and ions on chemical shifts can be built. The technique can be efficiently applied to large-scale data sets, can be applied on top of an arbitrary protein shift predictor, and can be trained in a fully automated fashion. Our results clearly highlight the importance of including ligand-related features and that even with the currently available small data set and relatively crude feature set, shift prediction in the vicinity of non-protein atoms can be moderately improved.

1 Introduction

Nuclear Magnetic Resonance (NMR) spectroscopy is increasingly becoming one of the cornerstones of structural biology. While the method has its shortcomings, e.g. with respect to the size of systems that can be studied, it offers a number of unique advantages for the study of molecular structures. Most importantly, the experiment can take place in solution, relatively close to physiological conditions. However, solving protein structures by NMR is not a completely straight-forward process. One important step of the NMR workflow is the resonance assignment process, where experimentally measured chemical shifts are mapped to atoms in the molecule under consideration. This then usually leads to a number of distance constraints that can be employed to generate models of the protein structure. Automated chemical shift prediction from a model or candidate structure of a protein can thus help with the structure elucidation process, but it can do even more: chemical shifts are highly sensitive to local characteristics of the molecular structure and hence subsume an enormous amount of structural information. Consequently, all kinds of computational procedures that generate candidate structures as part of their algorithmic process can profit from a comparison of computed chemical shifts or spectra with experimentally observed assigned or unassigned data. This collection includes techniques such as homology modelling and molecular docking, which can use NMR chemical shifts as an experimentally derived scoring function. For the case of protein-protein docking, for instance, previous work has shown that such information can help significantly in scoring candidates [KBM$^+$01, MCS$^+$08, CMV11]. Similarly, chemical shifts have been successfully used as restraints for molecular dynamics simulations [RKCV10].

However, the techniques that are currently available for chemical shift prediction can either cope with small molecules only (in this case, quantum mechanical treatment becomes possible), or only address non-modified proteins. An interaction between proteins and ligands, or proteins and ions for that matter, is out of scope of these techniques. Hence, for an application, e.g., to the scoring of protein-ligand docking results, we are currently lacking the right tools.

In this work, we study the influence of non-protein atoms on protein chemical shift prediction, propose simple features to capture part of this effect, and present a technique for incorporating such features into NMR chemical shift prediction. While the method does not allow to compute the shift for the ligand atoms themselves, the ligand structure is indirectly reflected in the response of the protein atoms close to parts of the ligand. This dependancy can lead to a first step along the way towards a pipeline for including protein-ligand NMR into the scoring of docking results, similar in spirit to [KBM$^+$01], which did the same for the protein-protein case.

In the following, we will present the methodology behind our approach, describe our implementation, and discuss the training and evaluation of a preliminary model.

2 The general idea behind the approach

An NMR experiment records one- or multidimensional spectra $\mathcal{S}$, which are formed by the sum of all responding atoms in the system. The information that is extracted from this recorded spectrum typically comes in the form of so-called chemical shifts, which describe the discrepancy between the NMR response observed for a given atom and its standard response in a simple model system. Thus, the chemical shift value δ of an atom encodes information about its chemical environment (both in molecular topology and space) and how it differs from that in a simple model system. Computing a mapping between the spectral peaks and individual atoms in the molecule corresponds to the so-called resonance assignment problem for the spectrum. The inverse direction, predicting the chemical shift for individual atoms of a molecule, gives the NMR chemical shift prediction problem.

The official repository for protein NMR shift data is the Biological Magnetic Resonance Data Bank (BMRB) [UAD$^+$08], which stores the data in NMRStar format, a file format [HC95] that is derived from CIF [HAB91, BM02]. These data files contain detailed information about the experimental setup and the physico-chemical conditions as well as the recorded chemical shifts for those atoms of the measured molecule for which it could be obtained.

In an idealized NMR chemical shift spectrum $\mathcal{S}$, the peaks are just sharp sticks and the spectrum has the following form

$$\mathcal{S} = \sum_{a \in mS} \delta(a) + \epsilon$$

where ϵ denotes the noise, $\delta(a)$ the recorded chemical shift for atom a, and mS the molecular system.

If the system under consideration is a protein, the spectrum can be further divided into protein and solvent contributions:

$$\mathcal{S} = \sum_{a \in P} \delta^P(a) + \sum_{a \in S} \delta^S(a) + \epsilon$$

where P denotes the set of all protein atoms and S the solvent.

The chemical shifts within a protein–ligand complex thus are:

$$\mathcal{S} = \sum_{a \in P} \delta^P(a) + \sum_{a \in L} \delta^L(a) + \sum_{a \in S} \delta^S(a) + \epsilon$$

where L denotes the set of ligand atoms.

If we take the influence of protein, ligand, and solvent onto each other into account as well, we find:

$$\begin{aligned} \mathcal{S} \quad = \quad & \sum_{a \in P} \delta^P(a) + \delta^{LP}(a) + \delta^{SP}(a) \\ & + \sum_{a \in L} \delta^L(a) + \delta^{PL}(a) + \delta^{SL}(a) \\ & + \sum_{a \in S} \delta^S(a) + \quad \epsilon \end{aligned}$$

where $\delta^{LP}(a)$ ($\delta^{PL}(a)$) denotes the chemical shift contribution induced by the ligand onto a protein (protein onto a ligand) atom a, and $\delta^{SX}(a)$ denotes the influence of the solvent onto the protein ($X = P$) or ligand ($X = L$).

The chemical shift information stored in a BMRB file is restricted to protein atoms only. Thus, our model and the typical content in the BMRB is limited to

$$\mathcal{S} = \sum_{a \in P} \delta^{P}(a) + \delta^{LP}(a) + \delta^{SP}(a) + \epsilon$$

The solvent is assumed to be distributed evenly around the protein and $\delta^{P}(a)$ and $\delta^{SP}(a)$ are usually computed together.

For a protein-ligand shift prediction, we thus have to consider

$$\mathcal{S} = \sum_{a \in P} \delta^{P}(a) + \delta^{LP}(a) + \epsilon$$

where δ^{P} comprises now the influence of solvent and protein on a protein's chemical shift and δ^{LP} the ligands influence.

Given these considerations, we can formulate an additive model for a protein-ligand chemical shift prediction as

$$\mathcal{S} = \sum_{a \in P} \mathcal{M}_{PL}(a) = \sum_{a \in P} \{\mathcal{M}_{P}(a) + \mathcal{M}_{L}(a)\}$$

Thus, for a protein–ligand complex and a given pure protein prediction $\mathcal{M}_P(a)$, we only have to train a model for $\mathcal{M}_L(a)$ that can be simply added.

In pure protein prediction typical models are based on approximations to semi-classical terms, statistical models, or a combination of both. Since to our knowledge, no reliable approximations of classical terms covering the protein-ligand interface are available, we decided for a statistical ligand model

$$\mathcal{M}_{L}(a) = \hat{\delta}(a, \mathrm{lf}_1(a), \ldots, \mathrm{lf}_j(a))$$

where $\hat{\delta}$ is a random forest model, and lf_1,..., lf_j denote ligand related features.

3 Materials and Methods

3.1 The underlying protein model

For the pure protein model, $\mathcal{M}_P$ we employed our recently developed pure protein hybrid approach [DLLH11], a random forest model based on a large collection of features, namely semi-classical approximations for ring current (ring), the electric field (EF), hydrogen bonding (HB), and random coil contributions (coil) as well as sequential, structural, and force field based features:

$$\mathcal{M}_{P} = \mathcal{RF}(\delta^{\text{coil}}, \delta^{\text{ring}}, \delta^{\text{EF}}, \delta^{\text{HB}}, \mathrm{f}_1, \ldots, \mathrm{f}_i)$$

The model was trained on a recent extract of the BMRB, 859 non-homolog NMR resolved PDB files mapped to BMRB files with 100% sequence identity.

In principle, any protein chemical shift prediction method can be used as basic model. However, our pure protein model has the advantage that it was built exclusively on unrereferenced pure protein data (PDB entries with additional ions, ligands, or DNA have been excluded) and focused on NMR resolved PDB structures for high consistency between the PDB data and the underlying NMR experiment (as we are going to use for training the protein-ligand model as well).

3.2 Towards a model of the ligand influence

From our experience on protein-only models, we have reason to expect that force field related features, such as contributions to interaction energies between ligand and protein atoms, will form valuable features for chemical shift prediction in the protein-ligand case as well. However, in contrast to the protein-only case, were the chemistry is uniform and hence force fields are relatively simple, treating protein-ligand interactions correctly is considerably more difficult. Roughly speaking, we will need a force field that is equally suited to protein and ligand atoms. While there are specialized force fields for the treatment of ligands, such as MMFF94 [Hal96], these usually have drawbacks in correctly capturing the protein chemistry. Hence, we decided to base our study on the GAFF force field [WWC$^+$04], which for protein atoms equals the well-known Amber force field. For non-protein (or non-DNA) atoms, GAFF offers a wide range of force field parameters and a simple scheme to extrapolate these to atom types that are not fully covered.

However, fully integrating the GAFF force field into our chemical shift prediction is a non-trivial task. Instead, in this study we aim at first answering whether such an approach is worthwhile at all by addressing two different questions: (a) do chemical shift predictions really fare significantly worse in the presence of ligand atoms and (b) is there a non-trivial correlation between the interactions encoded in GAFF and the deviation of the chemical shifts close to a ligand from their predicted values?

For the first question, we have collected a data set of 581 PDB-BMRB pairs of protein-ligand complexes (c.f. section 3.3). After culling the set to 10% homology using the PISCES package provided by the Dunbrack group [WJ05] and further restricting to 100% sequence similarity between PDB and BMRB protein content, we obtain 151 such pairs.

The data set was generated automatically using the procedure described in [DRB$^+$11]. Out of the 102,427 assignments of shift values to protein atoms that were contained in this set, 39,230 belonged to atoms that were located less than 10Å from a ligand atom.

To answer the second question, we created a set of new features that capture information about the ligand environment. Such information has been intelligently encoded into the GAFF atom types, which are based on the local molecular topology of the ligand. In contrast to other atom typing techniques, GAFF not only relies on pure connectivity information, but also uses the orders of the bonds connecting the atoms to deduce the final atom type. For this to work, however, correct bond order information for the ligand is required.

Unfortunately, the ligand information provided by the Protein Data Bank (PDB) [BWF$^+$00] is often incomplete, hydrogen information is missing and/or bond order information is not provided. In cases where only one of these is missing, the other can be deduced by either filling up free valences with hydrogens or by distributing free valences over bonds. For the latter case, we use BALL's bond order assigner [DRB$^+$11].

In principle, we could now try to proceed by computing GAFF interaction energies between all pairs of ligand- and protein atoms. However, this turns out to be a non-trivial problem: first, the input structures need to be carefully prepared (charges have to be assigned, e.g., with the help of a quantum chemical procedure, force field parameters have to be extrapolated, etc.). Then, we might need to optimize the structures with respect to the GAFF energy in order to relax potential non-optimal configurations due to inaccuracies in the structure or approximations in the force field. The parameters for the minimization will have to be carefully controlled, and quite possibly, structural restraints will have to be employed to keep the structure from deviating from the NMR experiment. Finally, we need to be able to decompose the energies into the intra-protein, intra-ligand and protein-ligand components for each term in the force field's Hamiltonian. Not all of these requirements can be easily satisfied using the commercial implementation of the GAFF force field, but on the other hand, implementing GAFF is a challenging task.

We hence started with a simpler set of features to determine whether such a force field based approach will be likely to succeed: instead of computing GAFF energies, we determine the local chemical neighbourhood from the composition of non-protein GAFF types we find in the vicinity of each atom. As cutoff for neighborhood of protein and ligand atoms we use 10Å.

We then define the following new features for our statistical model: for each protein atom, we use the total number of ligand atoms within a 10Å radius summed over all atom types (*num_het_atoms*), the atomic element (*cl_het_element*), the GAFF type (*cl_gaff_type*), the distance of the closest ligand atom (*cl_het_dist*), and for each GAFF type *i* separately the number and closest distances (*gaff_type_i* and *gaff_type_dist_i*) for all ligand atoms within the 10Å radius as features.

In addition, we added the size of the ligand (*ligand_size*) and an indicator variable for the presence of ions (*has_ion*) to the feature list. For the classification whether a non-protein molecule counts as ion or ligand, we use the PDB RESTful web service (`http://www.rcsb.org/pdb/software/rest.do`).

3.3 Creation of the data set

A number of datasets for NMR protein chemical shift prediction has been employed by former approaches, e.g. the very recent ShiftX2 [HLGW11] training and test set, the PROSHIFT set [Mei03], the TALOS+ set [SDCB09], the RefDB [ZNW03] or the general PDB to BMRB mapping of the BMRB [UAD$^+$08] itself. To the best of our knowledge, none of these specially focuses on protein-ligand complexes. We thus created a new pure protein-ligand data set based on the official BMRB repository.

To this end, we further extended our pipeline for NMR pure protein chemical shift prediction models [DLLH11]. Thus, automated creation of datasets and retraining of the models is easily possible if new data becomes available.

The pipeline starts with the official BMRB mapping, from which we select all entries with either ions or ligands present according to the PDB RESTful web service. The pipeline further restricts the dataset to a subset with solely non-homolog entries and PDB–BMRB pairs with 100% sequence similiarity.

The pipeline then creates an SQLite database storing for each atom the experimenal shift and the features, the ones used for evaluating the pure protein model as well as the new ligand related features. These computations were assisted by the BALL library [HDR$^+$10].

We then apply the pure protein model $\mathcal{M}_P$ and compute for each atom a a residual shift $\delta^{\text{res}}(a)$ by substracting the pure protein prediction from the atom's experimental chemical shift:

$$\delta^{\text{res}}(a) = \delta^{\text{exp}}(a) - \mathcal{M}_P(a, \delta^{\text{coil}}(a), \delta^{\text{ring}}(a), \delta^{\text{EF}}(a), \delta^{\text{HB}}(a), \mathrm{f}_1(a), \ldots, \mathrm{f}_i(a)) \tag{1}$$

For training statistical models, the pipeline uses R [R D11] and its Random Forests package [Bre01]. The pipeline randomly splits the dataset in a ratio of 70:30 into training and test set. The resulting (extended) pipeline is shown in Fig. 1.

3.4 Performance Evaluations of the Model

We evaluated the model on the randomly chosen test sets created by our pipeline. Comparison to state-of-the-art techniques was performed by applying the stand-alone versions of ShiftX2 to our test data sets. The performance of our models can be estimated from the root mean squared error (rmse) and Pearson's Correlation Coefficient (corr) on the test set.

$$\text{rmse} = \sqrt{\frac{\sum_{i=1}^{n} \left(\delta_i^{\text{exp}} - \delta_i^{\text{pred}}\right)^2}{n}} \tag{2}$$

$$\text{corr} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{\delta_i^{\text{exp}} - \hat{\delta}_i^{\text{exp}}}{s_{\delta^{\text{exp}}}} \right) \left(\frac{\delta_i^{\text{pred}} - \hat{\delta}_i^{\text{pred}}}{s_{\delta^{\text{pred}}}} \right) \tag{3}$$

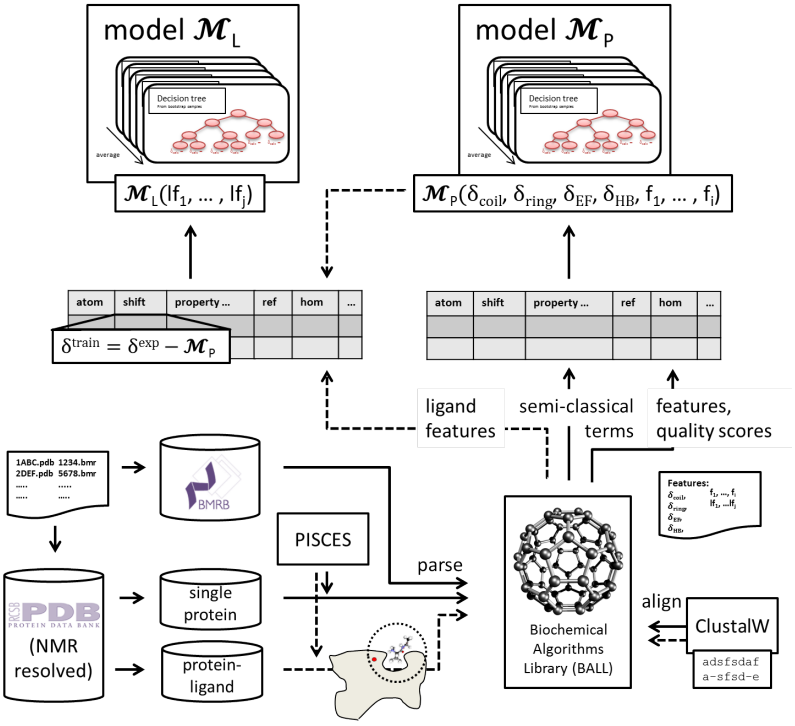

Figure 1: Pipeline for dataset generation and training of protein-ligand NMR chemical shift prediction models.

with δ^{exp} denoting the experimentally measured chemical shift of an atom, δ^{pred} the predicted chemical shift, and n the number of predictions made. $\hat{\delta}^{\text{exp}}(\hat{\delta}^{\text{pred}})$ denotes the standard deviation of the experimentally measured (predicted) shifts and $s_{\delta^{\text{exp}}}(s_{\delta^{\text{pred}}})$ its mean.

4 Results

To answer the first question mentioned above, i.e., to determine whether shift prediction really works significantly worse in the presence of ligand atoms as compared to the bulk, we compared the prediction of an established protein shift prediction package, the recently published ShiftX2 program, on the set of protein atoms in the spatial neighbourhood of a ligand atom. The results can be found in Tab. 1.

As can be seen, the performance of shift prediction close to ligand atoms indeed breaks down considerably. This is in line with our expectations: the ligand atoms are part of the chemical environment, which is what determines the chemical shift values. Since the ligand atoms are not covered by previous shift prediction models, the estimate of the chemical environment does not match that realized in nature and hence, the prediction strongly deviates from the true values. This motivates that additional steps should be taken to model ligand atoms in shift prediction procedures.

To now address the second question, i.e., to decide whether force field based features contain useful information for shift prediction, we started by plotting each feature against the absolute residual shift value, i.e., against the magnitude of the shift not explained by the protein model (of course, this number includes errors in the protein model as well). Not all features show a significant correlation, but some do: from Fig. 2 for the feature *ligand_size*, e.g., we see that larger ligands tend to have a greater influence on shift prediction than smaller ones. The largest ligands in our data set, however, tended to be modified peptides so that the deviation from a pure protein model is not as pronounced as the ligand size might suggest.

Another example of an interesting dependency between residual shift value magnitude and feature can be

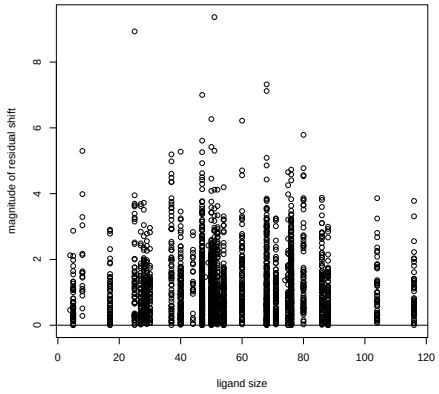

Figure 2: Exemplary plot of the residual shift magnitude as a function of ligand size for carbon atoms.

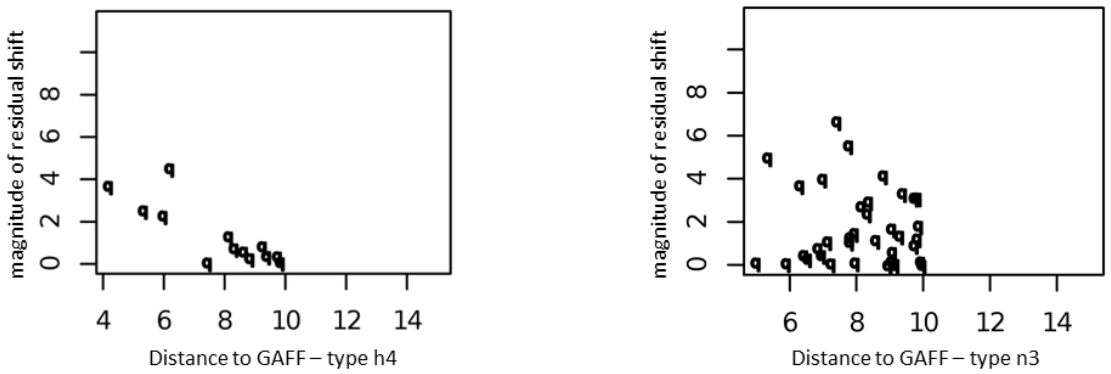

Figure 3: Exemplary plot of the residual shift magnitude as a function of distance to the next atoms of type h4 (left) and n3 (right).

found in Fig. 3, which shows the distances to the closest h4 and n3 atoms. Here, we see that the residual shift tends to decrease with increasing distance from the next h4 or n3 atom, since the interactions that influence the chemical shift are strongly decaying with distance.

In both figures, the results agree with our expectations.

However, plotting the residual itself as opposed to its magnitude as a function of the individual features (c.f. Fig. 4), we found a less clear trend. Roughly speaking, the presence of several GAFF atom types in the vicinity of a protein atom seems to induce a strongly deviating shift, but the direction and strength of the deviation can in most cases not simply be read of from the atomic composition of the neighbourhood.

An intuitive reasoning thus seems to suggest that simple atom type features as described above can indeed help in improving prediction performance, but that their influence will be limited, and more sophisticated features such as GAFF energy components will be needed to further improve the quality. And indeed, this is in line with our results when we proceed with our pipeline and train random forest models to the residual shift values (since we have significantly fewer data points at our disposal as in pure protein shift prediction, we only train three different models, one for C, N, and H atoms each): as can be seen in Tab. 1, which shows the results on the independent, non-homolog test set (30% of the initial data set), shift prediction is assisted by the new ligand model (prediction accuracy increases in all cases), but the increase is moderate in nature.

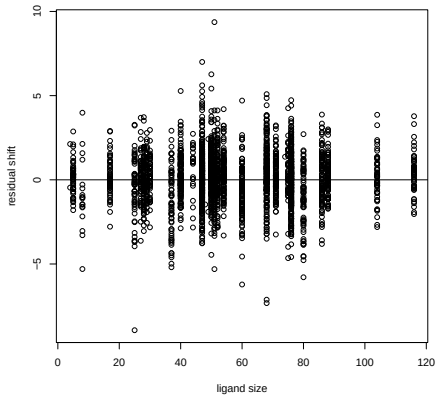

Figure 4: Exemplary plot of the residual shift as a function of ligand size for carbon atoms.

Method	^{15}N correlation (rmse)	^{13}C correlation (rmse)	^{1}H correlation (rmse)
ShiftX2 (reported)	0.9800 (1.1169)	0.9875 (0.497)	0.9729 (0.1471)
ShiftX2	0.72 (3.606)	0.999 (1.804)	0.939 (0.887)
BALL (ligand)	0.811 (2.712)	0.999 (1.592)	0.961 (0.867)
size (training/test)	3873 / 1660	9188 / 3938	14399 / 6172
num features	49	49	50

Table 1: Performance of ShiftX2 as reported, as measured on our new data set of protein atoms close to a ligand, and of our new ligand model on this set. Values shown are Pearson correlation and rmse (in parentheses), respectively, per atom type. The reported values of ShiftX2 have been averaged over all atom types for each element. The last two lines show the size of our data set in terms of shifts and the number of used features in our random forest model.

5 Discussion

As expected from the feature/residual correlations, the inclusion of GAFF atom type features improves the prediction, but only moderately. The limited improvement is probably due to two effects: for some of the atoms, the errors of the protein model are relatively large and these of course do not correlate with any ligand feature. Secondly, and more importantly, the dependency between protein shifts and ligand atoms is complex and cannot be modelled well by just counting atom types. In the terminology of a hybrid shift prediction model, we need semi-classical terms (i.e., models of the underlying physico-chemical processes in contrast to a mere collection of molecular properties) to greatly improve the prediction performance.

Still, considering the complexity of the task (for a similarly complex problem, imagine predicting force field energies using a purely statistical model from the composition of the chemical neighbourhood of a given atom), the slight but consistent increase in model performance strongly indicates that force field based descriptors are a promising ingredient for protein-ligand chemical shift prediction.

Here, we expect to improve matters by using GAFF energies in future work. Even though this might sound like a trivial extension at first glance, it is a highly non-trivial task that involves questions of a technical (implementation, integration, energy decomposition, ...) as well as of a more fundamental (parametrization, treatment of singularities, treatment of missing atoms or non-optimal input structures) nature. However, from the results of this study we are convinced that the effort will be worth its while.

References

[BM02] I. D. Brown and B. McMahon. CIF: The computer language of crystallography. *Acta Crystallographica Section B: Structural Science*, 58(3 PART 1):317–324, 2002.

[Bre01] L. Breiman. Random Forests. *Machine Learning*, (1):5–32, 2001.

[BWF+00] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28:235–242, 2000.

[CMV11] A. Cavalli, R. W. Montalvao, and M. Vendruscolo. Using Chemical Shifts to Determine Structural Changes in Proteins upon Complex Formation. *The journal of physical chemistry. B*, June 2011.

[DLLH11] A.K. Dehof, S.H. Loew, H.P. Lenhof, and A. Hildebrandt. A Pipeline for automated dataset generation and model training for NMR chemical shift prediction. *submitted for publication*, 2011.

[DRB+11] A. K. Dehof, A. Rurainski, Q. A. Bui, S. Böcker, H. P. Lenhof, and A. Hildebrandt. Automated bond order assignment as an optimization problem. *Bioinformatics (Oxford, England)*, 27(5):619–25, March 2011.

[HAB91] S. R. Hall, F. H. Allen, and I. D. Brown. The crystallographic information file (CIF): a new standard archive file for crystallography. *Acta Crystallographica Section A*, 47(6):655–685, Nov 1991.

[Hal96] T.A. Halgren. MMFF VI. MMFF94s option for energy minimization studies. *J Comp Chem*, 17:490–519, 1996.

[HC95] S. R. Hall and A. P. F. Cook. STAR dictionary definition language: Initial specification. *Journal of Chemical Information and Computer Sciences*, 35(5):819–825, 1995.

[HDR+10] A. Hildebrandt, A.K. Dehof, A. Rurainski, A. Bertsch, M. Schumann, N. Toussaint, A. Moll, D. Stockel, S. Nickels, S. Mueller, H.P. Lenhof, and O. Kohlbacher. BALL - Biochemical Algorithms Library 1.3. *BMC Bioinformatics*, 11(1):531, 2010.

[HLGW11] B. Han, Y. Liu, S. Ginzinger, and D. S. Wishart. SHIFTX2: significantly improved protein chemical shift prediction. *Journal of Biomolecular NMR*, pages 1–15, 2011.

[KBM+01] O. Kohlbacher, A. Burchardt, A. Moll, A. Hildebrandt, P. Bayer, and H. P. Lenhof. Structure prediction of protein complexes by an NMR-based protein docking algorithm. *J Biomol NMR*, 20(1):15–21, May 2001.

[MCS+08] R.W. Montalvao, A. Cavalli, X. Salvatella, T.L. Blundell, and M. Vendruscolo. Structure determination of protein-protein complexes using NMR chemical shifts: Case of an endonuclease colicin-immunity protein complex. *Journal of the American Chemical Society*, 130(47):15990–15996, 2008.

[Mei03] J. Meiler. PROSHIFT: protein chemical shift prediction using artificial neural networks. *J Biomol NMR*, 26(1):25–37, May 2003.

[R D11] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.

[RKCV10] P. Robustelli, K. Kohlhoff, A. Cavalli, and M. Vendruscolo. Using NMR Chemical Shifts as Structural Restraints in Molecular Dynamics Simulations of Proteins. *Structure (London, England : 1993)*, 18(8):923–933, August 2010.

[SDCB09] Y. Shen, F. Delaglio, G. Cornilescu, and A. Bax. TALOS+: A hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *J. Biomol. NMR*, (44):213–223, 2009.

[UAD+08] E.L. Ulrich, H. Akutsu, J.F. Doreleijers, Y. Harano, Y.E. Ioannidis, J. Lin, M. Livny, S. Mading, D. Maziuk, Z. Miller, E. Nakatani, C.F. Schulte, D.E. Tolmie, R. Kent Wenger, H. Yao, and J.L. Markley. BioMagResBank. *Nucleic Acids Research*, 36(SUPPL. 1):D402–D408, 2008.

[WJ05] G. Wang and R. L. Dunbrack Jr. PISCES: Recent improvements to a PDB sequence culling server. *Nucleic Acids Research*, 33(SUPPL. 2):W94–W98, 2005.

[WWC+04] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case. Development and testing of a general amber force field. *J Comput Chem*, 25(9):1157–1174, Jul 2004.

[ZNW03] H. Zhang, S. Neal, and D. S. Wishart. RefDB: A database of uniformly referenced protein chemical shifts. *Journal of Biomolecular NMR*, 25:173–195, 2003. 10.1023/A:1022836027055.