JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Universal *J*-Coupling Prediction

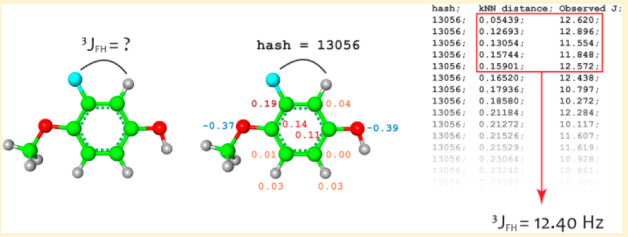
Juuso Lehtivarjo,*[,†,‡] Matthias Niemitz,[‡] and Samuli-Petrus Korhonen[‡]

[†]School of Pharmacy, University of Eastern Finland, P.O. Box 1627, 70211 Kuopio, Finland

[‡]PERCH Solutions Ltd., Puijonkatu 24 B 5, 70110 Kuopio, Finland

Supporting Information

ABSTRACT: A data driven approach for small molecule *J*-coupling prediction is presented. The method is targeted for use as part of an automatic spectrum analysis, therefore emphasizing prediction coverage, maintainability, and speed in the design. The database search involves encoding the coupling path atom types into hash codes, which are used to retrieve the matching coupling constant entries from the database. The fast hash dictionary search is followed by a *k* Nearest Neighbors regression to resolve the substituent and conformational dependencies, parametrized with atomic charges, torsion angles, and steric bulk.



INTRODUCTION

Automatic NMR spectrum analysis methods require *J*-coupling prediction probing the consistency between the proposed structure and the extracted NMR-parameters. Accurate prediction is naturally desirable, but, due to the iterative nature of the automatic spectrum analysis methods, small prediction errors are usually tolerated if the magnitude is correct. Instead, the prediction coverage is essential for a successful method. If the spectrum contains visible coupling constants that are not predicted at all, the automatic spectrum analysis will often remain ambiguous or fail completely.

The different small molecule *J*-couplings can be predicted with different models and equations, the first ones dating back to the 1960s. For example, ^3J couplings can be estimated quite well using the descendants of the Karplus equation,[1] such as the equations for $^3J_{HH}$, $^3J_{FH}$, and $^3J_{FF}$ coupling constants.[2−5] The physics behind $^2J_{HH}$ couplings have also been long known.[6−8] However, there are only a few general prediction programs available. The SPINUS program contains J_{HH} prediction based on a neural network trained for chemical shifts.[9] Some commercial programs also exist.[10−13] Also, quantum mechanical calculation of coupling constants is possible for all coupling paths and proven to be very accurate,[14] at least when heavy elements are not involved. However, even the lightest feasible QM methods require considerable computational resources for rather small molecules, for example 9, 48, and 173 min for chloromethane, benzene, and toluene, respectively.[14] In comparison, data driven methods are near instantaneous and do not require special hardware. Considering the computational cost, as well as the financial cost of the QM software themselves, the use of QM methods is not feasible in high-throughput automatic NMR spectrum analysis. Thus, there is an indisputable need for a lightweight but universal prediction tool.

Automatic spectrum analysis methods can rapidly create a great quantity of reliable coupling constant data. As couplings are naturally divisible into groups by the coupling pair and path atoms, a database search method is applicable to exploit the wealth of this constantly growing data. Compared with chemical shifts, scalar coupling constants are rather prone to through-space effects, which also promotes the use of a data driven method.

The presented method combines the hash dictionary search with *k* Nearest Neighbor (kNN) regression. The hash dictionary search is used to retrieve similar coupling paths from the coupling constant database currently containing over 40 000 data points. For the matching data, kNN regression is applied to resolve the substituent and conformational dependencies and to calculate the coupling values. With this approach, coupling constants of paths of all kind, including heteroatom couplings, are predictable with a single method. The method is primarily designed for use in automatic spectrum analysis, thus emphasizing coverage, maintainability, and speed. The developed method, called Juniper, is freely available as a Web server at www.perchsolutions.com/juniper.html. It is also implemented in the PERCH NMR Software[10,15] (free trial available).

METHODS

An outline of the procedure is presented in Figure 1 for the $^3J_{FH}$ coupling of 3-fluoro-4-methoxyphenol used as an example query. The procedure begins by calculating the hash code (steps 1 and 2) and prediction parameters (step 3) for the query coupling. The hash dictionary search is performed in step 4 and the *k* Nearest Neighbor regression in steps 5 and 6. The detailed description of each step can be found in the following chapters.

Received: January 27, 2014
Published: March 4, 2014

   dx.doi.org/10.1021/ci500057f | *J. Chem. Inf. Model.* 2014, 54, 810−817

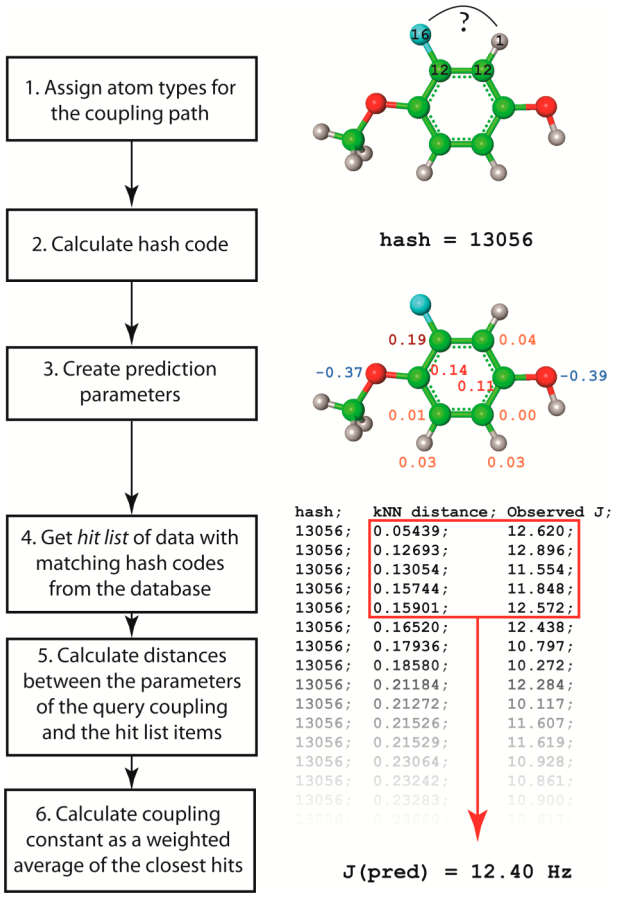


Figure 1. Prediction workflow. In the figure, the $^3J_{FH}$ coupling in 3-fluoro-4-methoxyphenol is used as an example query. In the case of aromatic ring couplings, the prediction parameters (step 3) are the partial charges of the ring and substituent atoms. The observed $^3J_{FH}$ coupling constant is 12.15 Hz.

Atom Classification. To facilitate the database search, the presented method groups the coupling constant data depending on the chemistry of the coupling path atoms. This is done by assigning a certain atom type to each atom of the coupling path and calculating a hash code based on the types. The atoms are classified to different types considering the following points. First, the most important and common couplings in small organic molecules are obviously carried over carbon atoms. Therefore, most of the atom types are for carbons (11 in total). The atom types are selected in a way that the most important coupling groups are already separated just by the the hash code. For example, for the $^3J_{HH}$ couplings the hash code is different for aliphatic chain paths, aliphatic ring paths of different ring sizes, aromatic ring paths, and sp^2-sp^2 paths. On the other hand, there is no need to separate hydrogen atoms of different types, since this information is already carried in the heavy atom to which they are connected. To gain enough data points for the prediction of the rare coupling paths, we also decided to keep the heteroatom classification as simple as possible. Following these considerations, 20 atom types (Table 1) were set up.

In this method, the coupling path is defined as the shortest possible route between two atoms by means of number of bonds. If the path is not symmetric with respect to the atom type indices, the path is read from the direction with the largest atom type index. In the case of two or more paths of equal length, the path with more sp^2 atoms than sp^3 atoms is selected.

Table 1. Atom Classification

atom type index	element and description	atom type index	element and description
1	H	11	C, in aromatic 5-ring
2	C, sp^3 in chain	12	C, in aromatic 6-ring
3	C, sp^3 in 3-membered ring	13	N, sp^3
4	C, sp^3 in 4-membered ring	14	N, sp^2
5	C, sp^3 in 5-membered ring	15	O
6	C, sp^3 in 6-membered ring	16	F
7	C, sp^3 in >6-membered ring	17	Si
8	C, sp^2 in chain	18	P
9	C, sp^2 in ring	19	S
10	C, sp	20	all other elements

In aromatic rings the route with the most heteroatoms is favored; in aliphatic rings *vice versa*. In fully symmetric paths, e.g. the para coupling in a phenyl ring, the path is chosen by chance. In these cases, the prediction parameters are symmetric in order to get the same set of parameters regardless of which path is chosen.

Hash Dictionary Search. The hash code used in the search algorithm is calculated using eq 1 as follows. The atoms of the coupling path (including the coupling pair atoms) are given indices from 1 to n. Every atom type (Table 1) has an index number T.

$$\text{hash} = \sum_{i=1}^{n} T_i * 20^{i-1} \quad (1)$$

With this kind of equation, the hash code is collision-free and reversible. The hash code is stored in a 64-bit unsigned integer in the data files, having space for 137 different atom types if eight bonds are considered as the longest predictable coupling path length. Therefore, it is possible to extend the atom classification in the future if necessary. Anyway, coupling paths longer than six bonds are very seldom visible.

Secondary hash codes complement the primary hash code in certain cases. There are two main reasons to apply a secondary hash code. First, in some cases the data within a certain primary hash code has clearly separate geometry and coupling constant values, thus promoting further separation. This allows the data to be split to highly homogeneous subsets, allowing the kNN regression to operate with the least amount of variables.[16] These cases include, for example, the *trans* and *gauche* $^3J_{HH}$ couplings of aliphatic six-membered ring systems and the *trans* and *cis* $^3J_{HH}$ couplings of alkenes. Second, since the kNN regression is fast as long as the number of data points is relatively small, a secondary hash code can be used to improve the performance. For example, aromatic J_{HH} couplings are roughly grouped by their substituent status: an ortho coupling with a substituent in the adjacent carbon sees only the data with a substituent in the same position, thus decreasing the hit list size and consequently speeding up the prediction. A full list of the secondary hash codes is presented in the Supporting Information Table S1.

A binary search algorithm[17] is applied to retrieve the coupling constant data with matching hash codes from the database. The result of this search is a *hit list* containing the

dx.doi.org/10.1021/ci500057f | J. Chem. Inf. Model. 2014, 54, 810−817

observed coupling constants and the prediction parameters, which are subsequently passed to the kNN analysis. In practice, the coupling constant data is written to two files: hash directory files and prediction parameter files. The hash directory files contain only the primary and secondary hash codes for the data point with an offset value as a link to the binary files. The hash directory is sorted by the primary and secondary hash values, enabling the use of a binary search algorithm. The prediction parameter files contain the observed coupling constant values and the actual prediction parameters. The data points in the prediction parameter files are in the same order than in the hash directory, so the matching data can be read in sequentially. For better performance, the coupling data of different path lengths and types is separated into different files, as it is unnecessary to read in all possible data if there are no such coupling paths in the query molecule.

Prediction Parameters. Prediction parameters are the contributions to the coupling constant used in the kNN analysis. Different sets of prediction parameters are built for different couplings path groups, still using as general code as possible. As too high dimensionality is a known problem in kNN,[16] the number of parameters was kept as low as possible. All prediction parameters are scaled to have values from 0 to 1 to equalize their contribution to the kNN distances, unless otherwise mentioned. Generally, additional scaling of the parameters was avoided to reduce the number of adjustable factors of the prediction algorithm. This ensures the maximal statistical robustness and parsimony of the kNN fitting.[18] The full list of prediction parameters is shown in the Supporting Information Table S2. Four main classes of parameters are used:

1) Charge Parameters. The relation between coupling constants and electronegativities is long-known and widely studied.[2,3,6,7] Yet a better approach should be the use atomic charges that take also the neighbor effects into account. Thus, the Gasteiger-Marsili partial charges[19] of the atoms of the coupling path or the closest substituents are used as prediction parameters for most couplings. The Gasteiger-Marsili partial charges have been previously used successfully for chemical shift prediction.[20,21]

2) Torsion angle parameters. As known since the Karplus equation,[1] torsion angles are the key contributors for the aliphatic $^3J_{HH}$ and $^4J_{HH}$ paths, despite a rough separation already being done for some $^3J_{HH}$ in the secondary hash codes. Cosine squares are used as the parameter for the ^3J couplings. For longer paths the sum of the cosines, or cosine squares, is used.

3) Steric effect parameters. In order to describe the steric bulk in the coupling path substituents the sum of atomic van der Waals radii of the substituents are calculated. The steric bulk affects the coupling constants by altering the rotational freedom around the bonds. In aliphatic ring systems, a parameter derived from steric bulk is used to estimate the flexibility of the ring.

4) Solvent. The solvent polarity index[22] is used as a parameter for all data points. However, the effect of the solvent to the coupling values was found to be rather small. Thus, its effect in the kNN regression has been reduced by effectively weighting the solvent parameters down by a factor of 10.

Additionally, there are some more special parameters to fix certain deficiencies. For example, the above charge parameters do not sufficiently carry the information about neighboring π-systems for the $^2J_{HH}$ couplings. Therefore, a neighboring sp^2 and sp system count is used as a parameter.

***k* Nearest Neighbors Regression.** The actual coupling constant prediction is done using the *k* Nearest Neighbors algorithm[23] from the data in the hit list. In contrast to SPINUS,[9] where kNN works as a part of the Associative Neural Network (ASNN) approach[24] to correct the remaining bias left from neural network, we used kNN to coupling constant prediction as such. Briefly, the Euclidian distances in the prediction parameter space between the query coupling and the hit list couplings are calculated as shown in eq 2

$$r = \sqrt{\frac{\sum_{i=1}^{n} (X_i - P_i)^2}{n}} + 0.01 \tag{2}$$

where X are the parameter values of the query coupling, and P are the corresponding values of the current hit list coupling. To prevent overweighting of exact hits, a fixed value of 0.01 is added to the distances. The predicted coupling is then calculated as a weighted average of the k hits with smallest distance, effectively the k nearest neighbors in prediction parameter space (eq 3).

$$J_{pred} = \sum_{i=1}^{k} \frac{1}{r_i} J_i \Big/ \sum_{i=1}^{k} \frac{1}{r_i} \tag{3}$$

In eq 3, the inverse value of the distances r_i is used for weighting, and J_i are the observed coupling constants of the hits. The value of k was optimized among small odd numbers (3, 5, 7, 9, and 11), from which $k = 5$ was found to give the best prediction results for the internal data. Finally, the predicted couplings smaller than the cutoff value (normally 0.25 Hz) are removed, as they are not visible in standard spectra anyway.

Data. The coupling constant data in the current teaching database originates from two sources. First, there are about 1300 coupling constants from the literature.[25−28] This data forms the basic set for the prediction, also containing a number of more uncommon couplings to ensure that at least one value is present for as many coupling paths as possible.

The majority of the coupling constant data is derived from automatic spectrum analysis performed in-house using the Automated Consistency Analysis (ACA) program of PERCH NMR Software.[15] ACA performs a complete spectral analysis extracting all relevant NMR parameters, chemical shifts, coupling constants, and line widths from the experimental data, and evaluates the consistency between predicted and found NMR parameters for a given structure resulting in a match index. This section of the database is expanded constantly. Additionally, in the PERCH NMR Software implementation of the prediction method, the users can add their own data to the prediction model using a tool provided. This is relevant for users focused on certain chemical environments or substructures, possibly unpatented and thus not represented in the database well enough, which is often the case e.g. in drug development studies.

Before creating the hash codes and prediction parameters, the molecular structures were geometry optimized in the Molecular Modeling System program of the PERCH NMR Software[10] using an extended version of the MMFF94 force field.[29] The manually built literature-derived data was written to the prediction database as such. Prior to adding the automatically analyzed data to the database the data is processed using a preparation algorithm identifying possible errors left from the automatic analysis. The data preparation module has several functions: 1) to remove couplings likely to

Table 2. Amount of Coupling Constant Data for Different Path Types[a]

	J_{HH}	J_{HF} and J_{HP}	J_{CF} and J_{CP}	total
^1J	-	9	69	**78**
^2J	499 (945)	41 (65)	57 (88)	**597 (1098)**
^3J (aliphatic)	3973 (9057)	94 (182)	37 (60)	**4104 (9299)**
^3J (aromatic)	3733 (11503)	178 (408)	22 (29)	**3933 (11940)**
^4J (aliphatic)	1129 (2078)	186 (365)	-	**1315 (2443)**
^4J (aromatic)	4178 (10324)	169 (426)	15 (16)	**4362 (10766)**
^5J (aliphatic)	98 (98)	115 (238)	-	**213 (336)**
^5J (aromatic)	1635 (4772)	40 (81)	1 (1)	**1676 (4854)**
^6J	188 (431)	53(103)	-	**241 (534)**
total	**15433 (39209)**	**885 (1877)**	**201 (263)**	**16519 (41348)**

[a]The values in parentheses are the numbers before the packing algorithm. A coupling path is considered "aromatic" when all coupling path atoms belong to aromatic rings: all the rest are assigned to the "aliphatic" group.

be unreliable due to line width, 2) to remove couplings likely to be unreliable due to symmetry, and 3) to handle AA′BB′ cases that are often ambiguously assigned in the analysis results. A more detailed explanation of the operations of the module can be found in the Supporting Information text. Finally, ambiguous cases were manually verified.

In principle, only the couplings constants which are visible in standard spectra are added to the database. For all other paths, there will be no hash code hits, and the prediction returns a zero. However, there are some cases where couplings with the same hash code can be visible or not, depending on the molecular structure. For example, the aliphatic $^4J_{HH}$ paths are usually not visible unless W-shaped, and the $^2J_{HH}$ couplings in five membered rings can be close to zero when substituted with oxygen. For these paths, some data points with zero couplings have to be added: otherwise the kNN regression will use the data from the visible couplings to predict also the invisible ones, leading to spurious couplings.

Using automatically analyzed data means that there will be multiple similar data points in the database, arising from different molecules with similar substructures or molecules with internal symmetry. Obviously, this is not wanted as it will just enlarge the database size without gaining any predictive value. Furthermore, it can flaw the kNN regression results if there is nothing but copies of the same data as nearest neighbors, effectively collapsing the kNN prediction to simple data retrieval based on a single data point. The problem is increasingly relevant when small k values, such as $k = 5$ of this study, are used. As previously suggested by Binev et al.,[9] a data packing algorithm based on a Kohonen self-organizing map was applied after the database was built. The self-organizing map was built with a number of neurons of 0.9 times the number of data points to ensure that clustering is not done too tightly. From each neuron having multiple data points, only one was retained in the data. The packing algorithm was written in R programming language[30] using the package "kohonen"[31] to build the self-organizing maps.

Currently, there are 41348 coupling constant data points, from 4824 molecules, in the database. After the above packing protocol is applied, 16519 data points remain. The packing protocol was not applied to the ^1J couplings due to data sparsity. The amount of data for different paths is listed in Table 2. The database size is expected to be growing constantly as more data are analyzed by ACA.

Prediction Error Model. An error estimate, "range", is calculated for the predicted couplings. The reliable error estimation is crucial to automatic spectrum analysis methods, especially for ACA since the coupling similarity (deviation between the observed and predicted couplings) is one of the parameters when calculating the match index,[15] and a penalty is given if the fitted parameter lies outside the given range. Besides, trustworthy prediction ranges will speed up the analysis by reducing the number of possible assignments. The range calculation model was based on three parameters: 1) the kNN distance of the best hit, 2) the standard deviation of the observed coupling values of the kNN hits, and 3) the inverse value of the hit list size. The Eureqa program[32] was used to resolve the correlation of the prediction error against the above parameters. The model was taught against a set of ACA derived coupling data of the teaching database predicted with a leave-one-out protocol. All coupling data from different coupling path groups was used in the same error model by using the relative error (prediction error divided by the predicted coupling) as the dependent variable. Finally, the range values given by the Eureqa-derived equation were multiplied by two to make the majority (>95%) of the prediction errors fall inside the given range. The prediction range correlated with the absolute prediction error with a Pearson correlation coefficient R of 0.52.

RESULTS AND DISCUSSION

Prediction Accuracy. For the evaluation of the prediction accuracy, the coupling constants of the teaching database were predicted using a leave-one-out (LOO) validation method. In the LOO method, the molecule for which the coupling constants are currently predicted is excluded from the teaching data in order to assess the external predictivity of the model. To achieve a more realistic image of the prediction accuracy for typical small molecules, the literature-derived section of the teaching data, containing a large number of more uncommon coupling paths, was omitted. Instead, the remaining ACA-derived section, containing a divergent set of industrially and pharmacologically relevant small molecules, was used in the test. Furthermore, as the duplicate data points are removed with the above-mentioned protocol, the tested coupling constants are nonredundant. This prevents the most common coupling paths, likely to get good predictions, to dominate the results. In the current database, some hash codes have only one coupling constant entry. In the LOO evaluation, these cases do not get any prediction and are thus omitted from the results. There were 101 such cases among the 15750 predicted couplings.

dx.doi.org/10.1021/ci500057f | J. Chem. Inf. Model. 2014, 54, 810−817

The RMS errors and the R correlation coefficients of the evaluation are presented in Table 3. In the table, the most evident trend is prediction errors of aromatic couplings being much smaller than aliphatic couplings of same path length. This is due to the fixed geometry of aromatic rings, which allows couplings to be parametrized solely through atomic charges. In aliphatic couplings the path geometry dominates the coupling value, but the optimal geometry is not always known or correct, causing uncertainty in the teaching data. Overall, the R correlations coefficients are better than 0.9 for all J_{HH} couplings except for long paths (^{5-6}J) where the observed shifts are of small magnitude. While it was possible to build adequate prediction for ^{5}J couplings, visible ^{6}J couplings are very rare, and in those cases the prediction was done without any prediction parameters, thus functioning as a plain database search. Anyway, in these cases the prediction will be sufficient for the use in iterative spectrum analysis just by telling if the coupling is visible or not. In addition, it is to be noted that the results for the J_{HF} and J_{HP} couplings are not much worse than for the J_{HH} couplings, despite the much smaller amount of data. For the ^{13}C couplings, the current data is too sparse to make any assumptions about the predictivity. However, the method works fine as a database lookup for those systems.

Table 3. Leave-One-out Validated Prediction RMS Errors (Hz, in Bold) and Pearson R Correlation Coefficients (in Italic) for the Experimentally Derived Part of the Teaching Data[a]

	J_{HH}	J_{HF} and J_{HP}	J_{CF} and J_{CP}
^{1}J	-	-	**5.10** *0.997* (30)
^{2}J	**1.03** *0.982* (436)	**6.23** *0.963* (14)	**4.25** *0.894* (21)
^{3}J (aliphatic)	**0.99** *0.929* (3742)	**0.99** *0.885* (43)	**0.65** *0.916* (16)
^{3}J (aromatic)	**0.34** *0.950* (3638)	**0.74** *0.932* (173)	**2.51** *0.764* (15)
^{4}J (aliphatic)	**0.43** *0.924* (974)	**0.49** *0.826* (162)	-
^{4}J (aromatic)	**0.20** *0.937* (4162)	**0.57** *0.940* (167)	**0.77** *0.463* (6)
^{5}J (aliphatic)	**0.22** *0.946* (74)	**0.11** *0.987* (113)	-
^{5}J (aromatic)	**0.11** *0.752* (1593)	**0.33** *0.947* (36)	-
^{6}J	**0.16** *0.321* (183)	**0.13** *0.093* (50)	-
all	**0.58** *0.992* (14802)	**1.02** *0.993* (758)	**3.79** *0.999* (88)

[a]The values in parentheses are the amounts of data points in the coupling group.

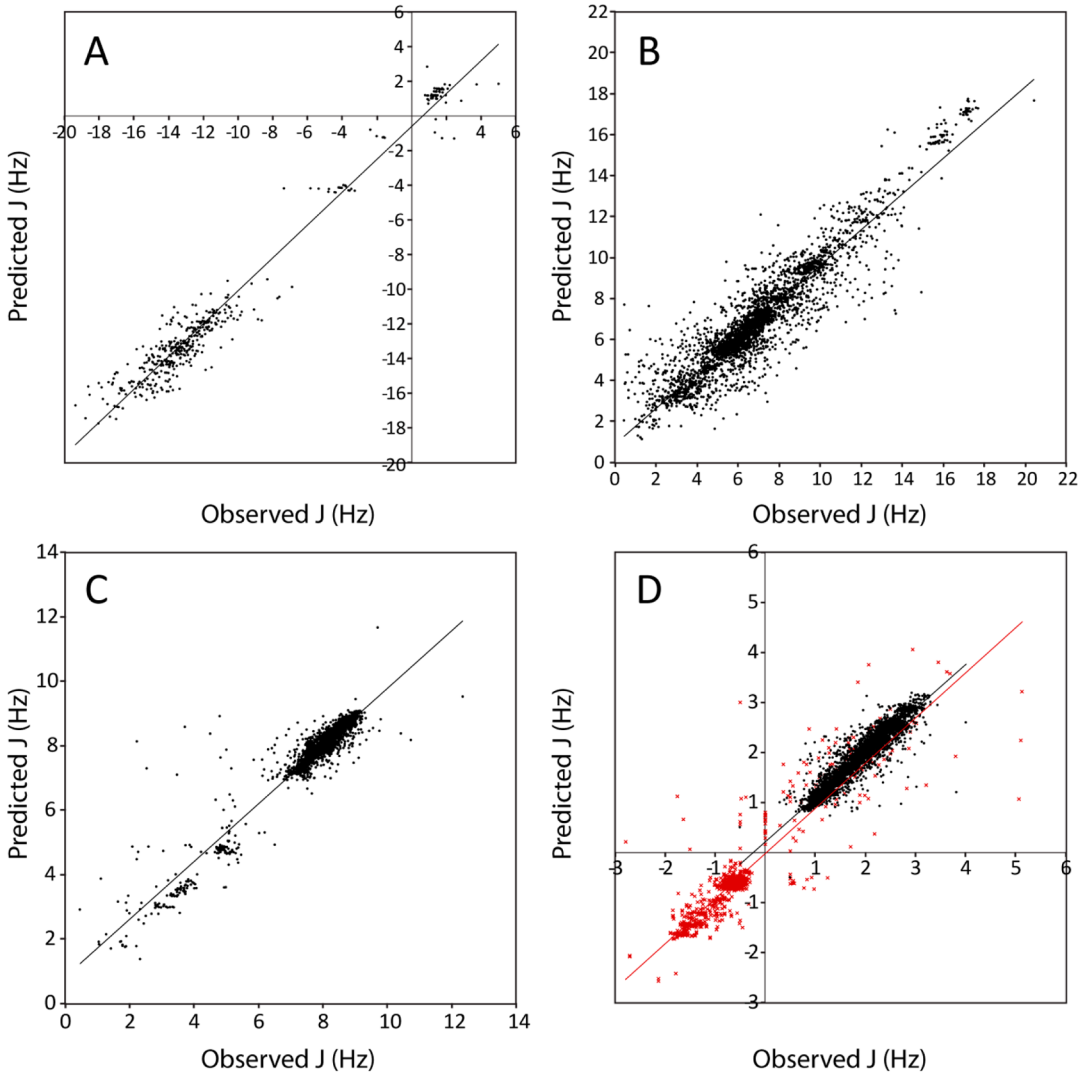


Figure 2. Prediction scatter plots for A) $^{2}J_{HH}$ coupling constants, B) aliphatic $^{3}J_{HH}$ coupling constants, C) aromatic $^{3}J_{HH}$ coupling constants, and D) $^{4}J_{HH}$ coupling constants. In plot D), the black dots and trend line are for the aromatic $^{4}J_{HH}$ couplings, and the red crosses and trend line are for the aliphatic $^{4}J_{HH}$ couplings.

dx.doi.org/10.1021/ci500057f | J. Chem. Inf. Model. 2014, 54, 810−817

Figure 2 shows the prediction scatter plots for certain J_{HH} coupling path groups. The rest of the plots are shown in Supporting Information Figure S1. Even though the RMS errors for these groups are acceptable, the figure shows a large number of outliers. This is unfortunately unavoidable when most of the data are automatically analyzed, even with the preparator module and manual investigation. However, not all outliers are only falsely interpreted or assigned coupling constants, but many arise also from incorrect geometry of the molecule in question. There are also some errors where the parametrization is not yet sufficient to reproduce the coupling constants. For example, one specially challenging group is the polycyclic aliphatic compounds, where the torsion angles are often locked to untypical values. Being uncommon among the teaching data, these compounds are often predicted with poor results. The problem could be solved by adding more suitable data or by adding more prediction parameters to separate these cases from the others. In addition, Figures 2A and 2D show some predictions with false sign. The sign determination is often impossible from standard NMR spectra and inconsistency may be present in the data. Many such cases are already manually corrected.

Using LOO and a nonredundant test set makes this evaluation to reflect the situation where the predictor does not find any exact matches and needs to interpolate between the data points. Without LOO validation, the RMS errors are much smaller (e.g., 0.40 and 0.17 Hz for the aliphatic and aromatic $^3J_{HH}$ couplings, respectively), which confirms the predictor to be able to successfully retrieve the database hits when available. In reality, the average prediction accuracy lies between these values.

The original teaching data contains a large number of duplicates, and about 60% of the data is removed in the packing algorithm. If the duplicate removal is not applied to the database in the model building and testing phases, the RMS errors are 0.76 and 0.25 for the aliphatic and aromatic $^3J_{HH}$ couplings, respectively. This suggests that most of the couplings in the prediction database are indeed clustered and that the prediction for common structures is better.

Comparison with Other Approaches. The accuracy of the presented prediction method Juniper was compared with other approaches. A test set of 99 molecules containing 255 coupling constants was taken from the study of Bally and Rablen,[14] in which the authors compared different QM methods for coupling constant prediction. Here, their "test set", used for method testing and parametrization of linear scaling, and "probe set", used for method evaluation purposes only, are combined to a single test set. In their paper, Bally and Rablen found one QM method (B3LYP/6-31G(d,p)u+1s) being the best by means of accuracy and speed. The results for this method, presented in their paper, are used in this comparison as such. Other tested coupling constant prediction programs were SPINUS[9] and the predictor from ACD/Labs (ACD/Labs NMR Predictors 2012 version 14.00). SPINUS was accessed via a Web server.[33] The predictor in MestReNova software[12] (Modgraph NMRPredict version 4.925) and ChemNMR[11] (in ChemBioDraw version 11.0.1) were also evaluated, but due to their poor coverage and accuracy they were not included in the comparison.

Only the couplings covered by all the programs are included in the comparison; the rest of the predictions can be seen in raw data in the Supporting Information. Formaldehyde and ketene were excluded due to their aberrant coupling constant values, causing large outliers for all three empirical predictors. SPINUS does only predict the absolute values of the couplings. ACD/Labs predictor gives the coupling signs, but many negative signs were obviously missing. To enable a fair comparison of the RMS errors, the SPINUS and ACD/Labs predictions were given a negative sign if the corresponding experimental coupling constant was negative. In addition, the SPINUS predictions for aromatic $^5J_{HH}$ couplings were way off. Compared with the SPINUS performance for other coupling paths, this seems to be a bug. Thus, $^5J_{HH}$ couplings were removed from the comparison.

The results of the comparison are shown as scatter plots in Figure 3 and separately for different coupling path classes in the

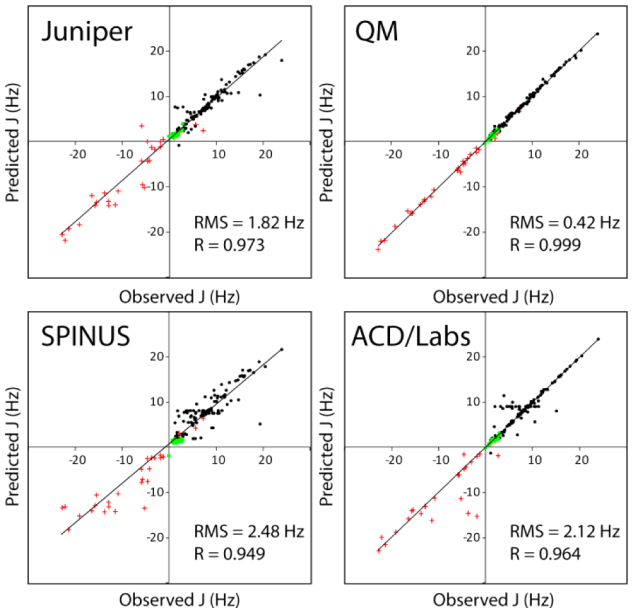


Figure 3. Prediction scatter plots for different prediction methods using the test set from the paper of Bally and Rablen (2011). From that paper, the prediction results with the suggested QM method are shown in the "QM" panel. $^2J_{HH}$, $^3J_{HH}$, and $^4J_{HH}$ coupling constants are shown as red crosses, black dots, and green squares, respectively.

Supporting Information Table S3. The results show superior accuracy for the QM approach. Considering the empirical methods, Juniper was somewhat better than SPINUS and ACD/Labs by means of RMS errors and R correlation coefficients. The experimental methods show rather different prediction profiles, seen in Figure 3. Whereas Juniper and SPINUS have a quite smooth distribution of the prediction errors, ACD/Labs apparently relies on direct search from a large database, shown as a large number of exactly matching predictions. However, when the method failed, the errors were larger than with Juniper, leading to the larger total RMS error.

The test set contains a number of basic molecules that are present in our teaching database. Most probably many of those are included also in SPINUS and ACD/Labs databases: considering the latter, this is evident since for many couplings the predicted values are exactly the same than the experimental ones. The results with the Juniper database hits omitted are shown in Supporting Information Table S4. Anyway, the outcome of the evaluation is the same, with the presented method still being at least as accurate as the other two empirical predictors.

dx.doi.org/10.1021/ci500057f | J. Chem. Inf. Model. 2014, 54, 810−817

The prediction coverage, emphasized in the Juniper design (see the Introduction), was also evaluated. The prediction coverage for the Bally and Rablen set for the tested predictors were 96%, 83%, and 99% for Juniper, SPINUS, and ACD/Labs, respectively. It is to be noted that the Bally and Rablen set contains a number of symmetric molecules (for example cyclohexane and benzene), in which the coupling constants of AA′ spin systems are not visible in standard spectra. SPINUS did not give predictions for these systems which decreases its coverage percentage.

It may be questioned if the coupling signs and couplings with small absolute value are really necessary for the automatic spectral analysis and structure verification purposes. Indeed, they are not often shown in production quality spectra if the spectral simulation is done considering only the first order effects. However, the spectral simulation of ACA takes the second order effects into account, and thus e.g. aromatic para couplings are required to successfully simulate the complex splitting patterns of aromatic regions. Even the coupling signs can sometimes be relevant.[34]

CONCLUSIONS

The development of the presented program Juniper began because a lightweight, general, and easily maintainable coupling constant predictor was needed in the automatic spectrum analysis program ACA[10,15] workflow. All these goals were achieved. Juniper can be taught to predict any type of *J*-couplings, including ^{13}C heterocouplings relevant due to the growing interest for analyzing ^{13}C spectra. For typical small molecules, the prediction results are returned almost instantly. For maintainability, the missing or poorly predicted couplings can just be added to the database and immediately used in the prediction model. Even though all paths are not specifically parametrized, the set of general parameters takes care that predictions are returned also for more uncommon paths.

The prediction accuracy of Juniper showed also to be competitive against other tested data driven predictors. The direct database search approaches, apparently used in e.g. the ACD/Labs predictor,[13] gives very accurate results when the query coupling is included in the database. However, when the query structure is lacking, it may lead to larger errors compared with the more predictive methods. Considering the line width (about 1.2 to 1.6 Hz) of the standard production quality spectra, small errors are often irrelevant for automatic spectrum analysis. Therefore, a consistent level of moderate prediction error is preferable against a mixture of mostly very small errors with some large outliers: actually, large prediction errors usually cause more problems than the really accurate predictions may give benefit.

The prediction accuracy is to be improved in the future when new data is analyzed and added to the model. Still, it might never reach the superior accuracy of quantum mechanical calculation of coupling constants. Instead, it is possible to exploit the accuracy of QM methods by using QM calculated coupling constants as teaching data of Juniper. This way the presented program could work as a wrapper method to retrieve the QM calculated couplings, and the computational cost of QM calculations could be amortized over multiple predictions. Moreover, this could also be a way to complement and improve the current database. For example, before the synthesis of novel compounds of a certain substructure, it could be beneficial to add several representative QM-calculated examples of the substructure to the Juniper database to ensure that the couplings are correctly predicted in the structure verification phase.

ASSOCIATED CONTENT

Supporting Information

Chapter on data preparation module; Tables S1 to S4; Figure S1; high-resolution versions of Figures 2 and 3; raw data of the prediction method comparison as an Excel spreadsheet. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: juuso.lehtivarjo@uef.fi.

Notes

The authors declare the following competing financial interest(s): The presented method is freely available for academic use, but also a part of commercial software (PERCH NMR software). JL is a part-time employee of PERCH Solutions Ltd. MN and SPK are employees of PERCH Solutions Ltd.

ACKNOWLEDGMENTS

J.L. is grateful for funding from the National Doctoral Programme of Information and Structural Biology (ISB).

REFERENCES

(1) Karplus, M. Vicinal Proton Coupling in Nuclear Magnetic Resonance. *J. Am. Chem. Soc.* **1963**, *85*, 2870−2871.

(2) Haasnoot, C. A. G.; de Leeuw, F. A. A. M.; Altona, C. The Relationship between Proton-Proton NMR Coupling Constants and Substituent Electronegativities—I: An Empirical Generalization of the Karplus Equation. *Tetrahedron* **1980**, *36*, 2783−2792.

(3) Díez, E.; San-Fabián, J.; Guilleme, J.; Altona, C.; Donders, L. A. Vicinal Proton-Proton Coupling Constants I. Formulation of an Equation Including Interactions between Substituents. *Mol. Phys.* **1989**, *68*, 49−63.

(4) San Fabián, J.; Guilleme, J.; Díez, E. Vicinal Fluorine-Proton Coupling Constants. *J. Magn. Reson.* **1998**, *133*, 255−265.

(5) San Fabián, J.; Westra Hoekzema, A. J. A. Vicinal Fluorine-Fluorine Coupling Constants: Fourier Analysis. *J. Chem. Phys.* **2004**, *121*, 6268−6276.

(6) Schaefer, T. Correlations of Ethylenic Proton Coupling with Electronegativity. *Can. J. Chem.* **1962**, *40*, 5−8.

(7) Pople, J. A.; Bothner-By, A. A. Nuclear Spin Coupling Between Geminal Hydrogen Atoms. *J. Chem. Phys.* **1965**, *42*, 1339−1349.

(8) Cookson, R. C.; Crabb, T. A.; Frankel, J. J.; Hudec, J. Geminal Coupling Constants in Methylene Groups. *Tetrahedron* **1966**, *22*, 355−390.

(9) Binev, Y.; Marques, M. M. B.; Aires-de-Sousa, J. Prediction of 1H NMR Coupling Constants with Associative Neural Networks Trained for Chemical Shifts. *J. Chem. Inf. Model.* **2007**, *47*, 2089−2097.

(10) PERCH NMR Software, http://www.perchsolutions.com/ (accessed Feb 24, 2014).

(11) ChemNMR, http://www.upstream.ch/products/chemnmr.html (accessed Feb 24, 2014).

(12) Mestrelab Mnova NMR Predict Desktop, http://mestrelab.com/software/mnova-nmrpredict-desktop/ (accessed Feb 24, 2014).

(13) ACD/Labs ACD/NMR Predictors, http://www.acdlabs.com/products/adh/nmr/nmr_pred/ (accessed Feb 24, 2014).

(14) Bally, T.; Rablen, P. R. Quantum-Chemical Simulation of 1H NMR Spectra. 2. Comparison of DFT-Based Procedures for Computing Proton-Proton Coupling Constants in Organic Molecules. *J. Org. Chem.* **2011**, *76*, 4818−4830.

(15) Laatikainen, R.; Tiainen, M.; Korhonen, S.-P.; Niemitz, M. Computerized Analysis of High-Resolution Solution-State Spectra. In

Encyclopedia of Magnetic Resonance; John Wiley & Sons, Ltd.: Chichester, UK, 2011.
(16) Indyk, P.; Motwani, R. Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing - STOC '98*; ACM Press: New York, NY, 1998; pp 604−613.
(17) Knuth, D. E. *Art of Computer Programming, Vol. 3: Sorting and Searching*, 2nd ed.; Addison-Wesley Professional: Reading, MA, 1998; pp 409−417.
(18) Golbraikh, A.; Tropsha, A. Beware of q2! *J. Mol. Graphics Modell.* **2002**, *20*, 269−276.
(19) Gasteiger, J.; Marsili, M. Iterative Partial Equalization of Orbital Electronegativity—A Rapid Access to Atomic Charges. *Tetrahedron* **1980**, *36*, 3219−3228.
(20) Binev, Y.; Aires-de-Sousa, J. Structure-Based Predictions of 1H NMR Chemical Shifts Using Feed-Forward Neural Networks. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 940−945.
(21) Laatikainen, R.; Hassinen, T.; Lehtivarjo, J.; Tiainen, M.; Jungman, J.; Tynkkynen, T.; Korhonen, S.-P.; Niemitz, M.; Poutiainen, P.; Jääskeläinen, O.; Väisänen, T.; Weisell, J.; Soininen, P.; Laatikainen, P.; Martonen, H.; Tuppurainen, K. Comprehensive Strategy for Proton Chemical Shift Prediction: Linear Prediction with Nonlinear Corrections. *J. Chem. Inf. Model.* **2014**, *54*, 419−430.
(22) Reichardt, C. Solvatochromic Dyes as Solvent Polarity Indicators. *Chem. Rev.* **1994**, *94*, 2319−2358.
(23) Cover, T.; Hart, P. Nearest Neighbor Pattern Classification. *IEEE Trans. Inf. Theory* **1967**, *13*, 21−27.
(24) Tetko, I. V. Neural Network Studies. 4. Introduction to Associative Neural Networks. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 717−728.
(25) Pretsch, E.; Bühlmann, P.; Affolter, C. *Structure Determination of Organic Compounds: Tables of Spectral Data*; 3rd ed.; Springer: Berlin, Heidelberg, 2000.
(26) Emsley, J. W.; Phillips, L.; Wray, V. *Fluorine Coupling Constants*; Pergamon Press: Oxford, UK, 1977.
(27) Reich, H. J. Proton-proton coupling, http://www.chem.wisc.edu/areas/reich/handouts/nmr-h/h-coupling.htm (accessed Feb 24, 2014).
(28) Constantino, M. G.; Lacerda, V.; da Silva, G. V.; Tasic, L.; Rittner, R. Principal Component Analysis of Long-Range "W" Coupling Constants of Some Cyclic Compounds. *J. Mol. Struct.* **2001**, *597*, 129−136.
(29) Halgren, T. A. Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization, and Performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490−519.
(30) R Core Team. R: A Language and Environment for Statistical Computing, 2013.
(31) Wehrens, R.; Buydens, L. M. C. Self- and Super-Organizing Maps in R: The Kohonen Package. *J. Stat. Softw.* **2007**, *21*, 1−19.
(32) Schmidt, M.; Lipson, H. Distilling Free-Form Natural Laws from Experimental Data. *Science* **2009**, *324*, 81−85.
(33) SPINUS Web server, http://www2.dq.fct.unl.pt/spinus/ (accessed Feb 24, 2014).
(34) Laatikainen, R.; Niemitz, M.; Weber, U.; Sundelin, J.; Hassinen, T.; Vepsäläinen, J. General Strategies for Total-Lineshape-Type Spectral Analysis of NMR Spectra Using Integral-Transform Iterator. *J. Magn. Reson., Ser. A* **1996**, *10*, 1−10.