



^{19}F NMR chemical shift prediction with fluorine fingerprint descriptor

Anna Vulpetti*, Gregory Landrum, Simon Rüdissler, Paulus Erbel, Claudio Dalvit

Novartis Institute for Biomedical Research, Novartis Pharma AG, CH-4002 Basel, Switzerland

ARTICLE INFO

Article history:

Received 1 October 2009

Received in revised form 28 December 2009

Accepted 31 December 2009

Available online 7 January 2010

Keywords:

Fluorine

Chemical shift

NMR

Fragment screening

ABSTRACT

A novel strategy for ^{19}F chemical shift prediction is described. The approach is based on a new fluorine fingerprint descriptor and a distance-weighted k -nearest neighbors algorithm applied on a training set of known chemical shifts measured for different fluorine local chemical environments. It is simple, fast, accurate and interpretable, as it allows the user to compare the predicted chemical shift with the experimental chemical shifts of the neighbor structures, analyse the variability in their chemical shifts, and based on that have a knowledge-based assessment of the reliability of the prediction. Possible applications of this approach in combination with ^{19}F NMR-based screening in drug-discovery projects are discussed.

© 2010 Elsevier B.V. All rights reserved.

1. Introduction

Identification of bioactive compounds from chemical libraries by use of ^{19}F NMR spectroscopy has gained increasing importance in the past years due to the high sensitivity of the methodology [1–5]. Screening approaches based on ^{19}F NMR spectroscopy have been thoroughly described in recent publications [6–8]. One of the approaches consists first screening a fluorinated library of compounds to identify binders; these are then used as reporters for the FAXS (Fluorine chemical shift Anisotropy and eXchange for Screening) experiments [1,2] for the subsequent screening of compounds containing (or not) fluorine atom(s) and for measuring the binding constant of the identified actives. Molecules of the fluorinated library binding to a protein can be easily identified by inspecting the ^{19}F NMR spectra recorded in the absence and in the presence of the protein. The binding molecules are detected by the reduction in signal intensity or even disappearance of the ^{19}F NMR resonance in the spectrum recorded in the presence of the protein. This is due to an increase in the ^{19}F signal line width of the molecule interacting with the receptor. Therefore ^{19}F NMR spectroscopy is a straightforward and efficient way to monitor ligand–protein interactions, and offers various advantages: the ^{19}F nucleus, the active NMR isotope, has (a) 100% natural abundance, (b) good sensitivity (0.83 compared to that of ^1H), (c) large dispersion in chemical shifts, thus allowing the screening of carefully prepared large mixtures without severe problems of

chemical shift overlap, (d) although the presence of fluorine is rare in natural products (about a dozen natural metabolites are known) it is quite frequent in commercially available compounds (~15% of ACD [9] molecules contain at least one fluorine atom). (e) The presence of buffer, solvent, detergents used for the screening do not interfere with the ^{19}F NMR detection of the low concentrated fluorinated molecules. (f) Finally, the large chemical shift anisotropy contribution of the bound state and the large exchange contribution deriving from the significant chemical shift difference between free and bound state to the observed transverse relaxation rate result in a very sensitive screening method as theoretically demonstrated [6,10].

^{19}F NMR-based screening is often performed with mixtures to enhance the throughput screening capability. The size of the mixture can range from few molecules to many molecules as described previously [11]. The ability to predict the ^{19}F chemical shifts of small molecules can be used in the generation of large mixtures by reducing the likelihood of spectral overlap and for the virtual deconvolution of the identified active mixtures.

In addition, it can find useful application in the analysis by ^{19}F NMR of enzymatic or chemical reactions involving fluorinated molecules, for predicting the chemical shift of the formed product(s) [6–8].

The ^{19}F isotropic chemical shifts have been studied since high resolution NMR has been applied to small molecules, due to the simplicity in the acquisition of the fluorine spectra with the low-field spectrometers available in those days and the large dispersion in chemical shift [12–14]. Extensive and comprehensive monographies on ^{19}F NMR were already published in the 1960s and early 1970s [15,16]. Over the last few years there has been an augmented interest on this topic due to the steady increase of

* Corresponding author. Tel.: +41 061 32 41016.

E-mail addresses: annavulpetti@hotmail.com, anna.vulpetti@novartis.com (A. Vulpetti).

synthesized fluorine-containing molecules with biological activity. An updated in-depth book on this subject was recently published [17]. Several approaches have been proposed in the literature for predicting the isotropic fluorine chemical shift ranging from *ab initio* calculations based on Hartree–Fock GIAO methods ([18–22] and references cited therein) to empirical-rules based on experimentally measured substituent effects (see for example [23–25] and references cited therein). Here we present a novel method for predicting the fluorine chemical shift that is based on the recently introduced fluorine fingerprint descriptor [11].

2. Results and discussion

2.1. Local fluorine chemical environment fingerprints

The ^{19}F NMR resonance frequencies are influenced by the chemical environments in which the fluorine atom(s) is embedded. To characterize the local chemical environment around the F atoms and the CF_3 groups, a new type of fingerprint was introduced [11]. The fingerprint generation process is identical for the CF and CF_3 -containing molecules as the CF_3 is treated as a dummy atom. Fingerprints are generated for each F atom or CF_3 moiety, as follows: the set of all paths of length one to L bonds rooted at the fluorine atom or at the CF_3 moiety are enumerated. Atom types are generated for the atoms in the path, and these atom-typed paths are then hashed to generate integer bit ids. An atom's type is determined by its atomic number, number of π electrons, and number of heavy-atom neighbors (not counting those in the path). In this paper all the paths of length $L = 5, 6, 7$ were explored to analyze the influence of the parameter L on the chemical shift prediction performance. Depending on the value of L three sets of fingerprints (FP) are generated, namely F-FP-5, F-FP-6 and F-FP-7. Although the two CF and CF_3 chemical shift datasets were treated separately, a single fingerprint acronym (F-FP- L) is used in this paper. In fact identical fingerprints are obtained for CF and CF_3 moieties embedded in identical chemical environments. These fluorine-environment fingerprints were inspired by the topological torsion descriptors published by Nilakantan et al. [26] Our fluorine-environment fingerprints differ from standard topological torsions (where L is kept fixed to four) in that we include all paths between one to five, six or seven bonds and only paths that start from the fluorine atom or the CF_3 moiety. This descriptor is distantly related to the HOSE descriptor [27] that has been used for ^1H and ^{13}C chemical shift predictions [28]. The fluorine-environment fingerprints have previously been utilized for the design of a diverse fluorinated fragment library containing different local environment of fluorine (LEF) [11]. This library is used in combination with ^{19}F NMR-based screening for identifying molecules that interact with the biomolecular target and for probing the fluorophilic protein environment.

2.2. Training sets: description and analysis

In this paper we explore about the efficacy of the developed descriptor for ^{19}F NMR chemical shift prediction. For this purpose two datasets of 640 CF and 550 CF_3 chemical shifts were used (the training sets). All the ^{19}F NMR spectra were recorded in aqueous solution as described in Section 4. Fig. 1 shows the distribution of chemical shifts of the two datasets. The dispersion in chemical shift is ca. 25 ppm for the CF_3 signals and ca. 40 ppm for the CF ones. However there are two subsets of CF-containing molecules that are located outside the 40 ppm range as indicated by the black bars in Fig. 1. These are the 6 member heteroaromatic rings where the heteroatoms are nitrogens. The ^{19}F resonances in the range -150 to -166 in Fig. 1b originate from molecules with two nitrogens in meta positions with respect to the fluorine atom whereas the

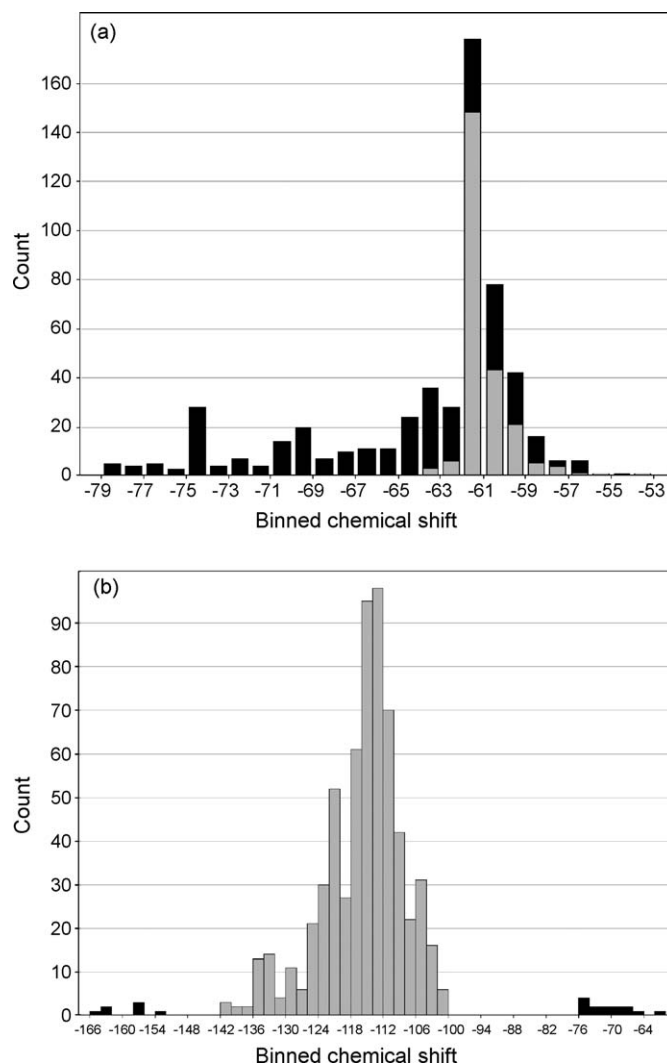


Fig. 1. Distribution of chemical shifts for the (a) CF_3 and (b) CF training sets. The bars in grey correspond to F and CF_3 substituted benzenes.

resonances in the range -60 to -76 originate from molecules with one or two nitrogens in ortho position with respect to the fluorine atom. Fig. 1 reports in grey the F and CF_3 substituted benzenes. It is worthwhile noticing that for the subset of CF and CF_3 -containing phenyl molecules the chemical shift range is 4-fold larger when compared to the CF_3 signals.

To assess if compounds with similar local fluorine-environment have similar chemical shifts, the training set compounds were analyzed in a pairwise fashion to identify how structural differences correlate with changes in chemical shifts. To each compound pair a score that combines their pairwise similarity and the difference between their chemical shifts was calculated as follows:

$$\text{Disparity Score}_{ij} = \frac{|\delta_i - \delta_j|}{1 - \text{sim}_{ij}} \quad (1)$$

where δ_i and δ_j are the isotropic chemical shift of the i th and the j th molecules, and sim_{ij} , ranging from zero to one, is the similarity coefficient between the two molecules using the F-FP- L description. Similarity between two fluorine fingerprints was calculated using the Dice metric, as recently described [11]. This approach of analyzing data has also been applied to the analysis of structure–activity relationships [29–32]. According to Eq. (1), the highest

Table 1
Total fraction of pairs and percentage of pairs with a difference in chemical shift smaller or equal to X ppm having a specific pairwise Dice similarity greater than a similarity threshold for (a) the CF₃ set and (b) the CF set.

(a)								
Similarity threshold	F-FP-5		F-FP-6		F-FP-7			
	% pairs	% of pairs with ≤ 0.5 ppm	% pairs	% of pairs with ≤ 0.5 ppm	% pairs	% of pairs with ≤ 0.5 ppm		
0.9	0.67	88.44	0.22	85.21	0.09	82.01		
0.8	1.50	73.39	0.60	88.61	0.32	84.38		
0.7	3.90	78.22	1.00	81.93	0.53	85.95		
0.6	4.92	68.22	3.69	79.92	1.09	80.99		
0.5	6.91	60.34	4.71	69.77	3.72	78.63		
0.4	14.70	47.32	5.74	63.97	4.71	69.59		
0.3	25.64	37.81	10.62	48.93	6.32	60.24		
0.2	40.48	29.63	27.70	35.76	17.96	41.42		
0.1	55.35	23.88	49.85	26.07	43.13	28.37		
0	100.00	14.91	100.00	14.91	100.00	14.91		
(b)								
Similarity threshold	F-FP-5	F-FP-6	F-FP-6	F-FP-6	F-FP-7	F-FP-7	F-FP-7	
	% pairs	% of pairs with ≤ 2 ppm	% pairs	% of pairs with ≤ 2 ppm	% pairs	% of pairs with ≤ 0.5 ppm	% of pairs with ≤ 1 ppm	
0.9	2.66	40.66	0.46	70.11	0.28	41.12	61.86	75.22
0.8	13.00	25.20	1.79	51.13	0.60	30.34	47.65	62.86
0.7	14.69	28.23	4.28	39.59	1.59	21.16	35.66	52.28
0.6	18.87	25.07	15.92	27.10	5.76	11.17	20.02	32.47
0.5	22.31	23.75	18.66	25.23	16.02	8.50	15.43	26.13
0.4	35.94	20.35	20.29	24.52	18.44	8.20	14.84	25.35
0.3	58.58	19.13	28.83	21.81	20.97	7.77	14.07	24.24
0.2	88.57	18.36	61.29	19.02	42.55	5.71	10.66	19.57
0.1	98.59	17.61	97.35	17.70	89.49	4.88	9.48	18.32
0	100.00	17.47	100.00	17.47	100.00	4.63	9.03	17.47

The bold values highlight those cases for which the percentage of pairs of molecules with a difference in chemical shifts smaller or equal to (a) 0.5 or (b) 2 ppm is (a) >80% or (b) 70%. This also highlights in bold that (a) 14.91% of CF₃ molecules and (b) 17.47% of CF molecules have similar chemical shifts despite having very different chemical environments.

Disparity Score values are obtained for compound pairs that have a high degree of local fluorine-environment similarity, but large differences in chemical shifts. The dataset of 550 CF₃ chemical shifts produces 150,975 pairs [equal to $n(n-1)/2$ pairs with n , the number of CF₃ molecules, equal to 550], while the 640 CF training set produces 204,480 pairs. A simple and intuitive graphical representation of the datasets consists in plotting, for each pair, the difference in chemical shift, $|\delta_i - \delta_j|$, against the $(1 - \text{sim}_{ij})$ value, and color coding the points (each pairs) by different Disparity Score intervals. An interactive navigation in this plot with simultaneous

visualization of chemical structure enables a detailed analysis of relevant information in the dataset.

Table 1 reports, for the three fingerprints (F-FP-5, F-FP-6 and F-FP-7) and the two datasets, the fraction of the pairs having a specific pairwise similarity greater than a similarity threshold. The percentages of pairs of molecules with a difference of chemical shift ≤ 0.5 , 1 and 2 ppm presented in Table 1 are also shown in Fig. 2.

For both the datasets and all three fingerprints, the trend is to have a high percentage of pairs of compounds with similar chemical shifts (i.e., $|\delta_i - \delta_j| \leq X$ ppm) as the similarity between the

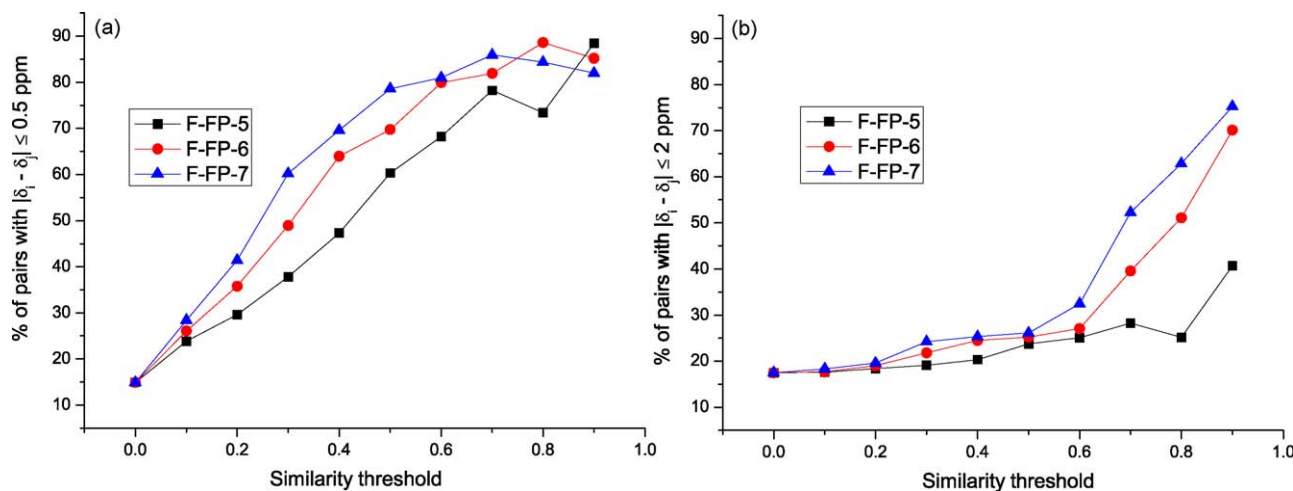


Fig. 2. Plot of percentage of pairs with a difference in chemical shift smaller or equal to X ppm (i.e., $|\delta_i - \delta_j| \leq X$ ppm) vs. Dice similarity threshold for (a) CF₃ and (b) CF training sets using F-FP-5 (black squares), F-FP-6 (red circles), F-FP-7 (blue triangles). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

Table 2

Percentage of chemical shifts predicted with errors ≤ 0.5 , ≤ 1 , ≤ 2 , ≤ 3 ppm or ≤ 10 ppm as a function of different similarity intervals (*SimBin*) and descriptor types (F-FP-L) for the (a) CF₃ and the (b) CF dataset.

(a)		RMSE	% Predicted	% Correct prediction				
F-FP-L	<i>SimBin</i>			≤ 0.5 ppm	≤ 1 ppm	≤ 2 ppm	≤ 3 ppm	≤ 10 ppm
F-FP-6_A	Class I	0.41	30.36	91.62	98.20	98.20	98.80	100.00
F-FP-7_A		0.62	19.82	88.07	97.25	97.25	98.17	100.00
F-FP-7_B		0.52	39.27	82.24	92.52	98.13	100.00	100.00
F-FP-5_A		0.84	52.73	81.03	88.97	93.10	97.93	100.00
F-FP-6_B		0.89	56.72	69.66	82.07	93.10	97.24	100.00
F-FP-7_C	Class II	1.8	59.09	66.97	77.06	87.16	92.66	100.00
F-FP-5_B		1.47	72.55	59.63	77.98	90.83	96.33	100.00
F-FP-7_D		1.95	74.91	58.62	73.56	89.66	91.95	100.00
F-FP-6_C		1.57	75.08	58.42	71.29	93.07	94.06	100.00
F-FP-5_C	Class III	2.03	84.91	50.00	58.82	80.88	86.76	100.00
F-FP-7_E		4.66	84	38.00	58.00	80.00	86.00	100.00
F-FP-6_E		2.71	88.9	36.00	52.00	64.00	72.00	100.00
F-FP-6_D		2.48	84.35	35.29	49.02	70.59	84.31	100.00
F-FP-5_D		3.03	89.64	26.92	46.15	61.54	65.38	100.00
F-FP-5_E		4.47	94.73	17.86	39.29	50.00	60.71	92.86
(b)		RMSE	% Predicted	% Correct prediction				
F-FP-L	<i>SimBin</i>			≤ 0.5 ppm	≤ 1 ppm	≤ 2 ppm	≤ 3 ppm	≤ 10 ppm
F-FP-6_A	Class I	3.45	56.25	43.89	63.61	81.11	87.50	97.22
F-FP-7_A		2.01	39.84	50.98	68.63	80.78	89.80	100.00
F-FP-7_B		5.4	61.56	35.79	54.57	72.59	81.47	95.94
F-FP-7_C	Class II	3.85	81.56	29.89	46.55	65.52	76.05	96.36
F-FP-5_A		3.35	72.97	31.69	49.46	63.38	75.59	97.43
F-FP-6_B		4.2	76.72	26.68	42.36	61.10	73.73	93.69
F-FP-7_D		3.75	92.97	20.50	34.96	53.28	66.72	94.96
F-FP-6_C		4.18	90.94	19.42	34.02	51.89	65.46	95.53
F-FP-5_B	Class III	4.73	90.16	22.36	35.70	49.91	62.56	91.33
F-FP-7_E		4.93	96.88	16.29	29.68	44.68	54.52	91.77
F-FP-5_C		5.01	96.25	17.21	29.71	44.32	53.57	89.94
F-FP-6_D		4.66	97.03	14.49	27.70	40.10	52.82	92.43
F-FP-5_D		5.44	98.59	14.90	24.72	39.78	52.14	88.43
F-FP-5_E		5.53	99.22	14.49	24.09	39.06	51.18	87.24
F-FP-6_E		5.23	98.44	13.49	25.24	37.14	50.00	90.00

pairs increases (right side of the Fig. 2 plots). The definition of X is dataset dependent. Here we consider as an appropriate value $X = 0.5$ ppm for CF₃ training set and $X = 2$ ppm for CF training set. This is due to the wider spread of CF vs. the CF₃ chemical shifts: e.g., for the substituted CF/CF₃ benzenes, see Fig. 1, is of about 4-fold. For comparison a X value of 0.5 and 1 ppm was also tabulated for the application of F-FP-7 to the CF training set to monitor the trend in predictivity as a function of X (Table 1b).

Fig. 2a and Table 1a show that in order to have the chemical shifts of more than 80% of CF₃ pairs differ by ≤ 0.5 ppm, a similarity threshold of ≥ 0.9 is required with F-FP-5. For F-FP-6 and F-FP-7, the similarity threshold can be reduced to 0.7 and 0.6 respectively. For the CF dataset (Fig. 2b and Table 1b) a F-FP-7 pairwise similarity threshold ≥ 0.9 is required to have the chemical shifts of about 75% of the pairs ≤ 2 ppm.

Table 1 and Fig. 2 also indicate that about 15% of pairs of CF₃ molecules and 17% of CF molecules have similar chemical shifts despite having very different fluorine chemical environments (i.e., $\text{sim}_{ij} < 0.1$). This effect, which was already observed and discussed in Ref. [11], does not represent a concern for an approach aiming to predict ¹⁹F chemical shifts based on fluorine chemical environments. If the goal were to elucidate possible structural topology around the fluorine atom(s) having particular experimental chemical shifts, as it is common for ¹³C and ¹H NMR prediction, the presence of pairs with different environments but with similar chemical shifts could represent a problem.

2.3. Algorithm

Encouraged by these results, we applied a distance-weighted k -nearest neighbors algorithm (dw-KNN) to predict the isotropic chemical shifts of new query molecules, δ_q , based on the weighted average of the isotropic chemical shifts δ_i , of the k closest molecules of known chemical shift (the *training set*) [33]. The chemical shift of molecule q is calculated as follows:

$$\delta_q = \frac{\sum_{i=1}^k w_{iq} \delta_i}{\sum_{i=1}^k w_{iq}} \quad (2)$$

where w_{iq} is defined by the relative distance of each neighbor from the query molecule:

$$w_{iq} = \frac{1}{(1 - \text{sim}_{iq})^2} \quad (3)$$

The quadratic term in the denominator of Eq. (3), typically used in the machine-learning research field [33], serves to emphasize the contribution of closer neighbors to the prediction. Thus, because the distance weighting lowers the impact of the choice of k on predictive accuracy, we arbitrarily set k to 50. Although dw-KNN is generally quite robust, it can break down when presented with a new example to predict that is dissimilar to examples in the training set. In this situation, the normalized distance weighting

does not help because all neighbors are distant from the new example.

Leave-one-out cross-validation was applied to identify the minimum similarity required for good predictive accuracy and to verify that the choice of k has little effect on the accuracy of our dw-KNN model. Similarity values were binned in 0.1 unit intervals, and the first five bins of higher similarity (*SimBin*) considered (i.e., A: [1–0.9], B: [0.9–0.8], C: [0.8–0.7], D: [0.7–0.6], E: [0.6–0.5]). Then each molecule in the training set was removed in turn and its chemical shift was predicted considering only the neighbors of the training set that lay inside the five intervals of similarity.

Table 2 contains the percentage of chemical shifts predicted with errors ≤ 0.5 , ≤ 1 , ≤ 2 , ≤ 3 or ≤ 10 ppm for the different similarity intervals and descriptor types (F-FP- L _SimBin).

Additional experiments (not presented here) showed that varying k , the number of neighbors, from 50 to 5 does not have a large impact on prediction accuracy.

In cases where compounds have no neighbors within the specified similarity threshold, the dw-KNN model does not give a prediction. The percentage of compounds for which a prediction was generated are indicated in the column labeled '% predicted' in Table 2. For the same descriptor, predictions can be generated for a larger percentage of molecules, at the expenses of somewhat lower accuracy, by reducing the similarity threshold. To help the data analysis, Table 2a and b is sorted by the column '% of correct prediction with an error ≤ 0.5 ppm' and ' ≤ 2 ppm', respectively. An accuracy ≤ 1 ppm for more than 80% of the CF₃ cases is obtained with $L = 6$ and 7 (F-FP-6_SimBin, F-FP-7_SimBin) and similarity bins (*SimBin*) A and B and in the case $L = 5$ with similarity bin A. These combinations, which provide accurate predictions, are labelled Class I. The percentage of CF₃ shifts predicted with an accuracy ≤ 1 ppm varies from 70% to 78% with F-FP-7_C and F-FP-7_D and in the case of F-FP-6_C, F-FP-5_B (we labelled these Class II). All the other F-FP- L _SimBin combinations give poor prediction (Class III). For the CF molecules the percentage of shifts predicted with errors ≤ 2 ppm is greater than 70% with F-FP-7_A, F-FP-7_B and F-FP-6_A (Class I). The percentage of shifts predicted with errors ≤ 3 ppm varies from 65% to 76% using the five F-FP- L _SimBin combinations reported in Table 2b (Class II).

These results indicate that similarity can be used as a metric of the deviation from the experimental data and that the minimum similarity required to have accurate prediction depends on the type of descriptor.

The absence of F-FP-5 descriptor in more accurate prediction (Class I) for the CF-containing molecules should not surprise. This can be simply ascribed to the fact that the F-FP-5 descriptor starts to count the bond from the carbon of the CF₃ in the CF₃-containing molecules and from the fluorine atom directly for the CF-containing molecules. Moreover, the fluorine chemical shift in aromatic systems is known to be sensitive to substituents in ortho and para position particularly, and less at meta position. To be able to properly describe the type of substituent in para position a F-FP descriptor with $L > 5$ is needed.

2.4. Comparison with a commercially available tool

Our dw-KNN results were compared with those obtained using the ACD/FNMR Predictor software [34] for the CF₃ dataset. The associated error of the predicted chemical shift provided by ACD/FNMR software is of 5 ppm for most of the molecules. This is consistent with what found with our dataset: 93% of the chemical shifts are correctly predicted within an accuracy ≤ 5 ppm, while about 17% of the dataset are predicted with an accuracy ≤ 1 ppm. Molecules predicted by ACD/FNMR with a reported error greater than 8 ppm (52 molecules) are not considered in our analysis. Fig. 3a shows the scatter plot of ACD/FNMR predicted vs. experimental isotropic chemical shifts for the CF₃ molecules (RMSE = 2.77, slope = 0.81, intercept = -13.86). The subset of molecules (392 compounds) for which a prediction using either Class I or Class II descriptors (Table 2) can also be obtained is shown in red circles, whereas the remaining set is shown in black squares. The RMSE value for the ACD/FNMR prediction for the 392 compounds (in red in Fig. 3a) is 2.60.

The dw-KNN prediction for the 392 compounds using Class I (green circles) and Class II (blue triangles) descriptors is shown in Fig. 3b. The RMSE values for Class I prediction (301 compounds) and for all 392 molecules (predicted by either Class I or Class II) are 0.81 and 1.28, respectively. Considering that the two methods use different training sets, the compounds in red (Fig. 3a) are not necessarily those predicted best by the ACDLab tool.

An offset in the reference chemical shift would simply change the Y intercept of the fitting lines, but has no effect on the observed scatter. An explanation for the lower accuracy of the ACD/NMR in the prediction of the 392 molecules chemical shift is probably due to the solvent effect. The spectra of our training set were acquired in aqueous solution whereas the training set used by ACD/NMR

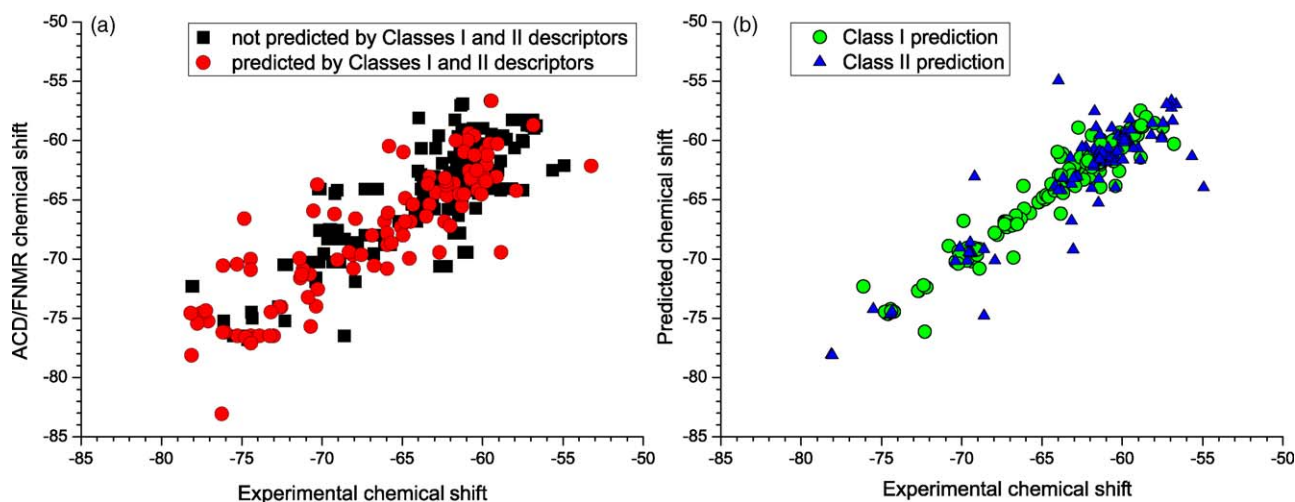
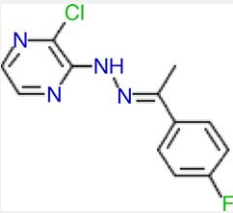
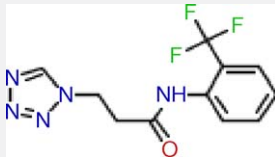


Fig. 3. (a) ACD/FNMR predicted vs. experimental chemical shift values for the CF₃ training set; (b) Predicted chemical shifts vs. experimental chemical shift values for 392 CF₃ training set molecules. The 392 compounds shown in the plot (b), correspond to the points shown in red circles in the plot (a). The green circles in the plot (b) report the chemical shifts predicted only by Class I descriptors (301 compounds), where the blue triangles correspond to the chemical shifts predicted by Class II (91 molecules, with the exclusion of those predicted by Class I). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

Table 3

The ^{19}F chemical shift values are reported for two molecules: (a) containing CF and (b) containing CF_3 .

(a)				(b)			
							
Molecule 23				Molecule 48			
F-FP- <i>L-SimBin</i>		δ	N	F-FP- <i>L-SimBin</i>		δ	N
F-FP-6_A	Class I		0	F-FP-6_A	Class I		0
F-FP-7_A			0	F-FP-7_A			0
F-FP-7_B		-110.30	2	F-FP-7_B			0
F-FP-7_C	Class II	-109.67	21	F-FP-5_A			0
F-FP-5_A		-108.53	50	F-FP-6_B			0
F-FP-6_B		-109.67	20	F-FP-5_B	Class II	-60.20	2
F-FP-7_D		-108.14	50	F-FP-7_C		-60.21	3
F-FP-6_C		-109.38	50	F-FP-7_D		-59.82	4
STD		0.80		F-FP-6_C		-59.85	4
Predicted		-110.30		STD		0.21	
Experimental		-110.76		Predicted		-60.20	
Δ		0.47		Experimental		-60.51	
				Δ		0.31	

The values reported in bold correspond to the predicted chemical shifts.

software utilizes most likely the ^{19}F chemical shifts of the molecules dissolved in organic solvents. It has been demonstrated that the ^{19}F chemical shift is very much dependent on the solvent [35,36] and consequently this could account in part for the scatter observed in Fig. 3a.

2.5. Application to external datasets

The quality of the dw-KNN model to make prediction for completely new molecules was then tested. The ^{19}F chemical shifts of 59 small molecules (43 molecules containing CF_3 and 16 containing fluorine substituted benzenes, the *test sets*) were calculated using the dw-KNN procedure described above. Multiple chemical shift values are calculated for each molecule based on the different combinations of *L* and similarity bins in Classes I and II as described in Table 2. The F-FP-*L-SimBin* combinations are used in sequence as reported in Table 2. For example, the predicted ^{19}F chemical shift values for two molecules (one containing CF_3 and

one containing CF) are reported in Table 3. For molecule 23 two neighbor compounds (column N of Table 3a) are found in the training set using the F-FP-7_B combination (Class I). For molecule 48 (Table 3b) no neighbor compounds are found in the training set using the five F-FP-*L-SimBin* combinations associated with Class I prediction. Two neighbor molecules are found using the Class II F-FP-5-B combination. The chemical shift value is calculated by weighted averaging the chemical shift values of these two molecules with similar fluorine local chemical environment.

Fig. 4 shows the scatter plot of predicted chemical shifts (Classes I and II) vs. the experimental values for the CF_3 test set (Fig. 4a, RMSE = 1.08) and the CF test set (Fig. 4b, RMSE = 2.37). The procedure provides a calculated value for 37 out of 43 CF_3 molecules (27 predictions of Class I, and 10 predictions of Class II) and for 15 out of 16 CF molecules (10 predictions of Class I, and 5 predictions of Class II). Despite the larger size, the approach can be also used for the assignment of ^{19}F chemical shifts in molecules containing multiple fluorine atoms, as shown in Fig. 5. The two

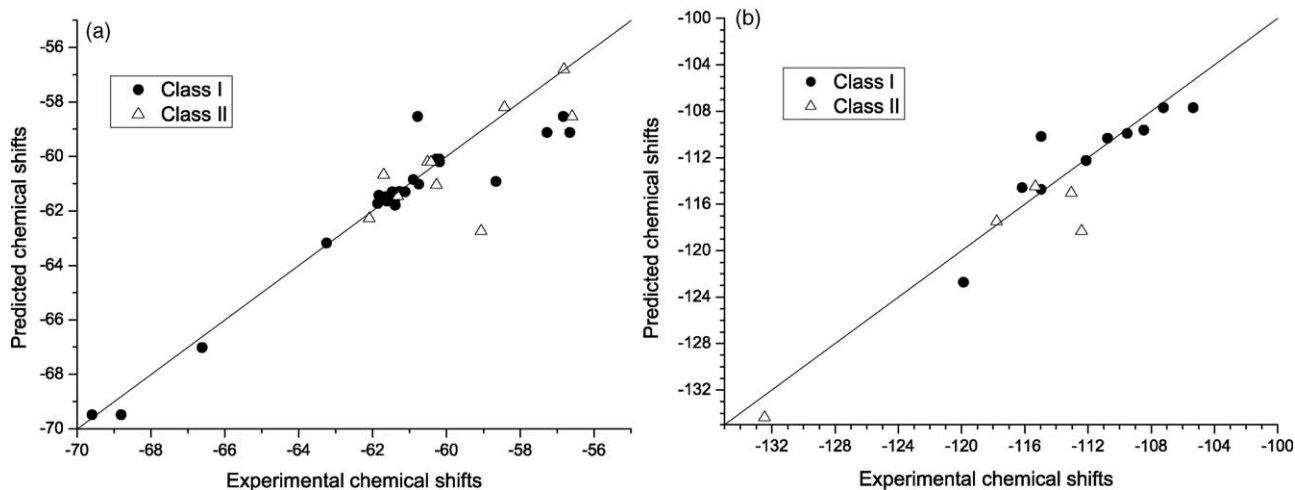


Fig. 4. Predicted vs. experimental chemical shift values for the (a) CF_3 test set and (b) the CF test set. Plot (a) shows the predicted chemical shifts values of Class I in black circles, and the predicted chemical shifts values of Class II in white triangles, respectively for (a) 37 CF_3 molecules and (b) 15 CF molecules.

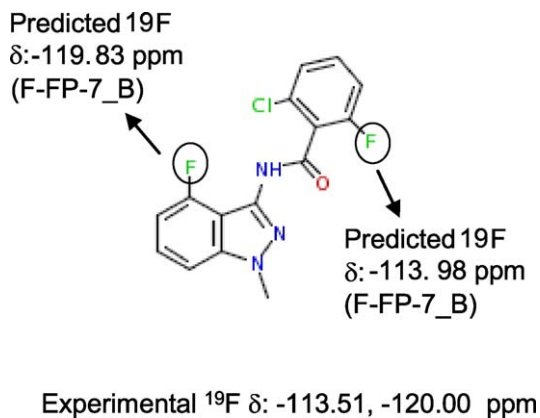


Fig. 5. ^{19}F chemical shift values assignment.

fluorine atoms are eight bonds apart and their chemical shifts are predicted quite accurately by using the F-FP-7_B. It is worth noting that the MW of this molecule (MW = 321.7) and other molecules of Fig. 4 is larger than 300, i.e. the upper MW limit of the training set molecules. Despite the larger size their chemical shifts are well predicted suggesting that the description of the fluorine local environment with a two-dimensional fingerprint descriptor was sufficient for these molecules.

The dw-KNN algorithm also allows us to visualize the experimental chemical shift values and chemical structures of all the neighboring molecules for each prediction. This display helps the further understanding of the applicability domain of the dw-KNN algorithm. In particular we found that the visual inspection of the neighbor structures and the variability in their chemical shifts provides the NMR experts with the possibility to assess, based on their knowledge, the reliability of the prediction beyond the simple numerical accuracy expected from cross-validation.

2.6. Outlook

The quality of the prediction depends on both the training set size and its diversity. The current training datasets have been assembled considering CF and CF₃ molecules that are able to cover as much as possible the different fluorine local chemical environments present in in-house fragments [11]. The expansion of our current chemical shift training set is clearly an area of future investigation for covering the local fluorine-environment fingerprints that currently are not yet represented. The histogram of Fig. 1 will be used to define the most relevant molecules either to be searched for and/or synthesized in order to increase the coverage for chemical shift prediction. In addition molecules that contain multiple fluorine atoms in close proximity and chemically equivalent multiple fluorine atoms will be included.

A useful feature of the presented approach resides in its ability to be retrained with a specific data set of compounds if improved predictions for related structures are required.

3. Conclusion

In summary, we have presented a novel strategy, based on a fluorine fingerprint descriptor, for the ^{19}F chemical shift prediction that takes into account the local environment of fluorine. The approach is simple, fast, accurate and interpretable. It relies on a representative set of known chemical shifts measured for different fluorine local chemical environments.

It is anticipated that this approach could find many useful applications: (a) ^{19}F NMR-based screening is efficiently used to

identify binders within a library of fluorinated molecules assembled in mixtures. The size of the mixtures can range from few molecules to many molecules as described previously [11]. This approach supports the generation of large mixtures by reducing the likelihood of spectral overlap and enables the virtual deconvolution of the identified active mixtures. These two features allow for high throughput NMR screening and rapid identification of the active ingredients. Moreover, (b) it could find useful applications in the ^{19}F NMR functional screening experiments [6] and in the analysis of ^{19}F NMR spectra of chemical reactions involving fluorinated molecules, for predicting the chemical shift of the formed product(s).

4. Experimental

4.1. NMR chemical shift determination

The NMR samples were in 50 mM Tris, 100 mM NaCl, pH 7.5 and contained 10% D₂O for the lock signal. The small molecules were prepared in concentrated stock solutions in deuterated DMSO and stored at 4 °C. All the ^{19}F NMR experiments were recorded at 23 °C with a Bruker DRX-600 spectrometer operating at a ^{19}F Larmor frequency of 564 MHz and equipped with a SampleJet robot for sample tube automation. The spectra were acquired with proton decoupling using the Waltz-16 composite pulse sequence with a 90° pulse of 150 μs. The data were collected with a spectral width of 42.17 and 29.92 ppm for the CF and CF₃ mixtures, respectively. The acquisition and repetition times were 0.8 and 3.8 s, respectively. The data were multiplied with a squared cosine window function prior to Fourier transformation. Typically 32 scans were recorded for each spectrum. Chemical shifts are referenced to the CFCl₃ signal in water.

4.2. Modelling

Molecule preparation, generation of the F-FP-L fingerprints, dw-KNN approach were carried out using the open-source cheminformatics toolkit RDKit [37]. Python scripts for the tasks described in this work are included in the Supplementary material. The developed approach outputs a text file which is visualized in Spotfire [38], a commercial program which enables data/structure visualization.

The figures and tables were performed using the Origin 7.0 and Microsoft Excel software packages.

References

- [1] C. Dalvit, M. Flocco, M. Veronesi, B.J. Stockman, *Comb. Chem. High Throughput Screening* 5 (2002) 605–611.
- [2] C. Dalvit, P.E. Fagerness, D.T.A. Hadden, R.W. Sarver, B.J. Stockman, *J. Am. Chem. Soc.* 125 (2003) 7696–7703.
- [3] T. Tengel, T. Fex, H. Emtenas, F. Almqvist, I. Sethson, J. Kihlberg, *Org. Biomol. Chem.* 2 (2004) 725–731.
- [4] J. Klages, M. Coles, H. Kessler, in: P.A. Bartlett, M. Etzeroth (Eds.), *Exploiting Chemical Diversity for Drug Discovery*, RCS Publishing, 2006, pp. 263–290.
- [5] L. Poppe, T.S. Harvey, C. Mohr, J. Zondlo, C.N. Tegley, O. Nuanmanee, J. Cheetham, *J. Biomol. Screening* 12 (2007) 301–311.
- [6] C. Dalvit, *Prog. NMR Spectrosc.* 51 (2007) 243–271.
- [7] D.B. Berkowitz, K.R. Karukurichi, R. de la Salud-Bea, D.L. Nelson, C.D. McCune, *J. Fluorine Chem.* 129 (2008) 731–742.
- [8] C. Kreutz, R. Micura, in: P. Herdewijn (Ed.), *Modified Nucleosides in Biochemistry, Biotechnology and Medicine*, Wiley-VCH, Weinheim, 2008, pp. 1–27.
- [9] ACD (Available Chemicals Directory) data set is available from Symyx Technologies, Inc.
- [10] C. Dalvit, *Concepts Magnetic Reson. Part A* 32A (2008) 341–372.
- [11] A. Vulpetti, U. Hommel, G. Landrum, R. Lewis, C. Dalvit, *J. Am. Chem. Soc.* 131 (2009) 12949–12959.
- [12] H.S. Gutowsky, D.W. McCall, B.R. McGarvey, L.H. Meyer, *J. Am. Chem. Soc.* 74 (1952) 4809–4817.
- [13] L.H. Meyer, H.S. Gutowsky, *J. Phys. Chem.* 57 (1953) 481–486.
- [14] R.W. Taft, *J. Am. Chem. Soc.* 79 (1957) 1045–1049.
- [15] J.W. Emsley, J. Feeney, L.H. Sutcliffe, *Prog. Nucl. Magn. Reson.* 1 (1966) 871–968.

- [16] J.W. Emsley, L. Phillips, *Prog. Nucl. Magn. Reson.* 7 (1971) 1–520.
- [17] W.R. Dolbier Jr., *Guide to Fluorine NMR for Organic Chemists*, John Wiley and Sons, 2009.
- [18] A.C. de Dios, E. Oldfield, *J. Am. Chem. Soc.* 116 (1994) 743–7454.
- [19] T. Tanuma, J. Irisawa, *J. Fluorine Chem.* 99 (1999) 157–160.
- [20] E. Oldfield, *Annu. Rev. Phys. Chem.* 53 (2002) 349–378.
- [21] E. Oldfield, *Phil. Trans. R. Soc. B* 360 (2005) 1347–1361.
- [22] D.E. Williams, M.B. Peters, B. Wang, K.M. Merz Jr., *J. Phys. Chem. A* 112 (2008) 8829–8838.
- [23] M.J. Fifolt, S.A. Sojka, R.A. Wolfe, D.S. Hojnicky, J.F. Bieron, F.J. Dinan, *J. Org. Chem.* 54 (1989) 3019–3023.
- [24] R.E. Jetton, J.R. Nanney, C.A.L. Mahaffy, *J. Fluorine Chem.* 72 (1995) 121–133.
- [25] G. Bauduin, B. Boutevin, Y. Pietrasanta, *J. Fluorine Chem.* 71 (1995) 39–42.
- [26] R. Nilakantan, N. Bauman, J.S. Dixon, R. Venkataraghavan, *J. Chem. Inf. Comput. Sci.* 27 (1987) 82–85.
- [27] W. Bremser, *Analytica Chim. Acta* 108 (1978) 355–365.
- [28] S.G. Spanton, D. Whittern, *Magn. Reson. Chem.* 47 (2009) 1055–1061.
- [29] J. Bajorath, L. Peltason, M. Wawer, R. Guha, M.S. Lajiness, J.H. Van Drie, *Drug Discov Today* 14 (2009) 698–705.
- [30] G.M. Maggiora, *J. Chem. Inf. Model* 46 (2006) 1535.
- [31] L. Peltason, J. Bajorath, *J. Chem. Inf. Model* 50 (2007) 5571–5578.
- [32] R. Guha, J.H. Van Drie, *J. Chem. Inf. Model* 48 (2008) 646–658.
- [33] T.M. Mitchell, *Machine Learning*, McGraw-Hill International Edit, 1997.
- [34] <http://www.nmrsoftware.com/>.
- [35] E.Y. Lau, J.T. Gerig, *J. Am. Chem. Soc.* 118 (1996) 1194–1200.
- [36] V.N. Plakhotnyk, R. Schmutzler, L. Ernst, Y.V. Kovtun, A.V. Plakhotnyk, *J. Fluorine Chem.* 116 (2002) 41–44.
- [37] “RDKit: Open-source cheminformatics”, <http://www.rdkit.org>, last accessed: May 2009.
- [38] <http://spotfire.tibco.com/>.