ACS OMEGA

Article

Cite This: ACS Omega 2019, 4, 7323−7329

http://pubs.acs.org/journal/acsodf

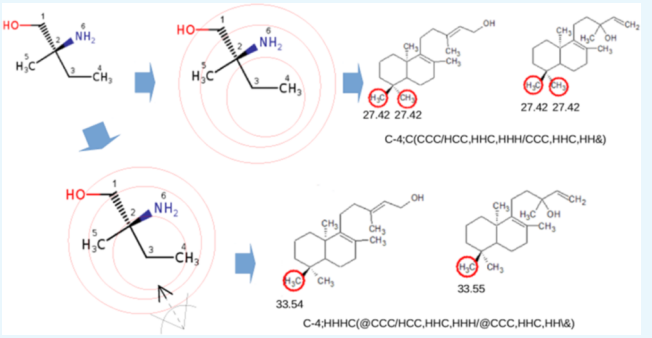
Stereo-Aware Extension of HOSE Codes

Stefan Kuhn*,† and Sean R. Johnson‡,§

†School of Computer Science and Informatics, De Montfort University, The Gateway LE1 9BH, Leicester, U.K.

‡Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, Michigan 48824, United States

ABSTRACT: Descriptions of molecular environments have many applications in chemoinformatics, including chemical shift prediction. Hierarchically ordered spherical environment (HOSE) codes are the most popular such descriptions. We developed a method to extend these with stereochemistry information. It enables distinguishing atoms which would be considered identical in traditional HOSE codes. The use of our method is demonstrated by chemical shift predictions for molecules in the nmrshiftdb2 database. We give a full specification and an implementation.



INTRODUCTION

A long-standing problem in chemoinformatics is the accurate encoding of the environment of an atom in a molecule. This task can be considered the equivalent problem to the accurate encoding of a molecule. Established techniques for encoding molecules include naming conventions like the IUPAC naming[1] or various line notations. The most common of these is the simplified molecular-input line-entry system (SMILES).[2] This has been extended over time, in particular, there is a version containing chirality information. A good notation should contain all informations about a molecule which is found in a typical molecular depiction, such as the atoms and their elements, charges, isotopes, etc. and their connectivity, i.e., the graph structure, including the bond order.

In a similar fashion, an atom environment encoding should contain information about an atom, the atoms in its environment, and how they are connected. Hierarchically ordered spherical environment (HOSE) codes are most commonly used for this purpose.[3] The HOSE codes do not contain stereochemistry information.

Since a stereochemical extension of SMILES has been suggested and successfully applied, we suggest extending the HOSE codes in a similar fashion to encode stereochemistry. This system keeps the advantages of the HOSE code notation and combines it with the established stereochemistry encoding of SMILES. As opposed to other methods, we need no information separate from the HOSE code, and all existing systems based on HOSE codes can use the extended version without any modification in the algorithm. We provide a full specification of the extended HOSE code and also supply a reference implementation.

Examples from the field of chemical shift prediction in nuclear magnetic resonance (NMR) show the strength of our approach.

BACKGROUND

The need for compact line notations of molecules emerged as soon as computers were used to handle chemical information. An early system is represented by the Wiswesser line notation (WLN).[4] This was followed interalia by the SYBYL line notation (SLN)[5] and the simplified molecular-input line-entry system (SMILES). There are also various standards for naming schemes, the most successful of which is probably the IUPAC naming. A recent development is the International Chemical Identifier (InChI).[6] These are to varying degrees, compact, human-readable, machine-processable, canonically specified, and supported by software implementations. A full discussion of their advantages and disadvantages is out of scope here.

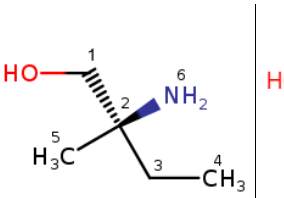
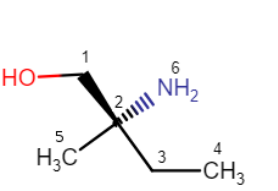
SMILES is of particular interest for us, because it has been extended to contain stereochemistry information.[2] The extension uses local chirality representation (as opposed to absolute chirality) and enables partial specifications. This is relevant because frequently stereochemistry is not fully specified in publications.

SMILES, WLN, SLN, and InChI notations encode the whole of a molecule. In some cases, it is useful to encode the environment of a particular atom in a molecule. This second kind of encoding could be used for database searches,[7,8] drug design,[9,10] or chemical shift prediction. For this purpose, the Hierarchical Organization of Spherical Environments (HOSE) code is a classic method.

The HOSE code encodes the atoms around a center in a sphere-wise manner. The spheres in the code are defined by the distance in bonds from the atom to be described. So, atoms are listed in a hierarchical manner starting with atoms one bond away from the selected center, proceeding to atoms two bonds

Received: February 21, 2019
Accepted: April 2, 2019
Published: April 23, 2019

Table 1. Comparison of the Standard and Stereo HOSE Codes of Two Stereoisomers

			
Atom 1	HOSE code	C-4;CO(CCN,/C,,/)//	
	stereo HOSE code	C-4;CO(@CCN,/C,,/)//	C-4;CO(@CNC,/C,,/)//
Atom 2	HOSE code	C-4;CCCN(C,O,,/,/)//	
	stereo HOSE code	C-4;@CCNC(C,O,,/,/)//	C-4;@CCCN(C,O,,/,/)//
Atom 3	HOSE code	C-4;CC(CCN,/O,,/)//	
	stereo HOSE code	C-4;CC(@CNC,/O,,/)//	C-4;CC(@CCN,/O,,/)//
Atom 5	HOSE code	C-4;C(CCN/C,O,/,)//	
	stereo HOSE code	C-4;C(@CNC/C,,O/,)//	C-4;C(@CCN/C,O,/,)//
Atom 6	HOSE code	N-3;C(CCC/C,O,/,)//	
	stereo HOSE code	N-3;C(@CCC/C,O,/,)//	N-3;C(@CCC/C,,O/,)//

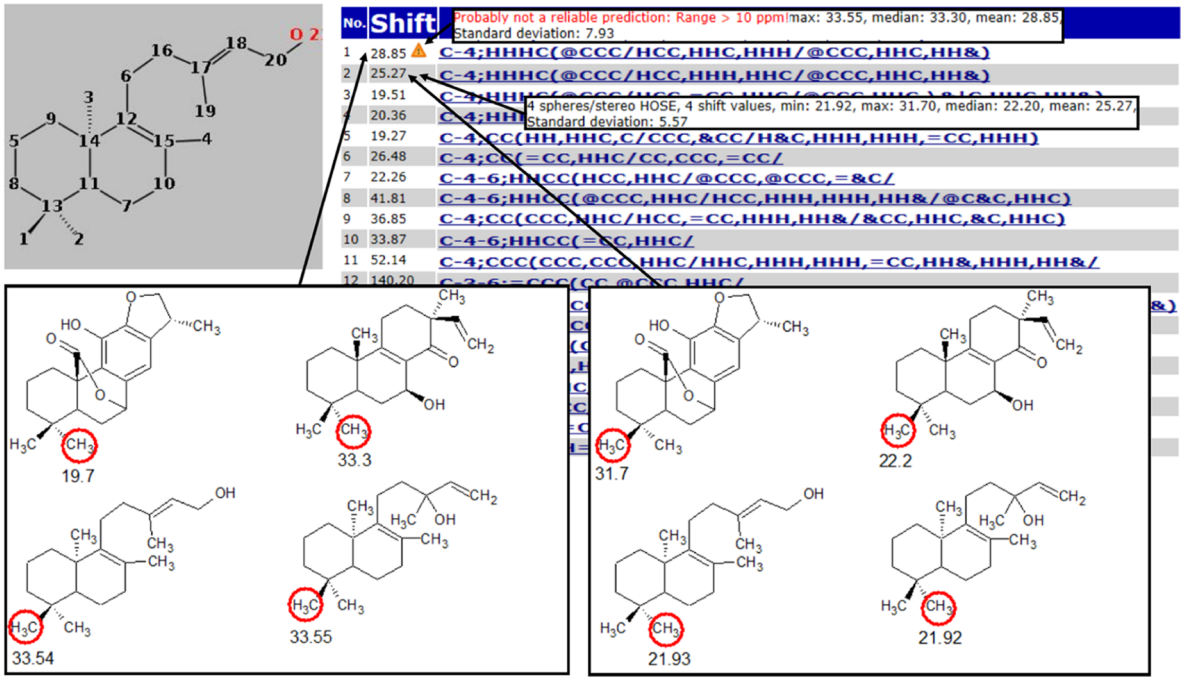


Figure 1. ^{13}C prediction done in nmrshiftdb2 for a compound with several chiral centers, using the stereo HOSE code. The predictions for atoms 1 and 2 are different. For atom 1, nmrshiftdb2 warns about an unusually wide range of possible values, indicating a wrong assignment. The four compounds which are used for the prediction are shown inset. The HOSE code used for atom 1 is *C*−4; *HHHC*(*@CCC/HCC,HHC,HHH/@CCC,HHC,HH*&), for atom 2 it is *C*−4; *HHHC*(*@CCC/HCC,HHH,HHC/@CCC,HHC,HH*&). The atoms marked with the red circles are used for the prediction. They are chemically equivalent, and the value 19.7 on the left and 31.7 on the right is not in line with the other assignments.

away, etc. Bond orders and aromaticity are also included. A graphical explanation of the HOSE code is provided in Kuhn et al.[11]

Chemical shift prediction is done using HOSE codes by generating the HOSE code for an atom in the molecule for which the prediction is done. In a database, all atoms with the same HOSE code are searched. The shift, or, if several are found, the average of the shifts, of the atoms found, gives the predicted shift. Since the HOSE code disregards stereochemistry, for stereochemically different atoms, the same shifts will be predicted.

Several methods for encoding atomic environments have previously been proposed, each with particular disadvantages when compared to stereo-HOSE codes. MNA descriptors[12] encode neighboring atoms in a hierarchical fashion but do not account for stereochemistry. An approach for encoding stereochemical information in the encoding of atomic environments has been suggested by Schütz et al.,[13] but they do not give an actual algorithm. It is unclear from the paper exactly how the stereochemical interactions mentioned are calculated but they depend somehow on an external library of ring skeletons. Spanton and Whittern[14] suggest combining HOSE codes with the InChI descriptor for stereochemistry. Again, it is unclear how this is exactly done, and it requires a separate descriptor. In our approach, the stereo description is part of the HOSE code itself.

DOI: 10.1021/acsomega.9b00488
ACS Omega 2019, 4, 7323−7329

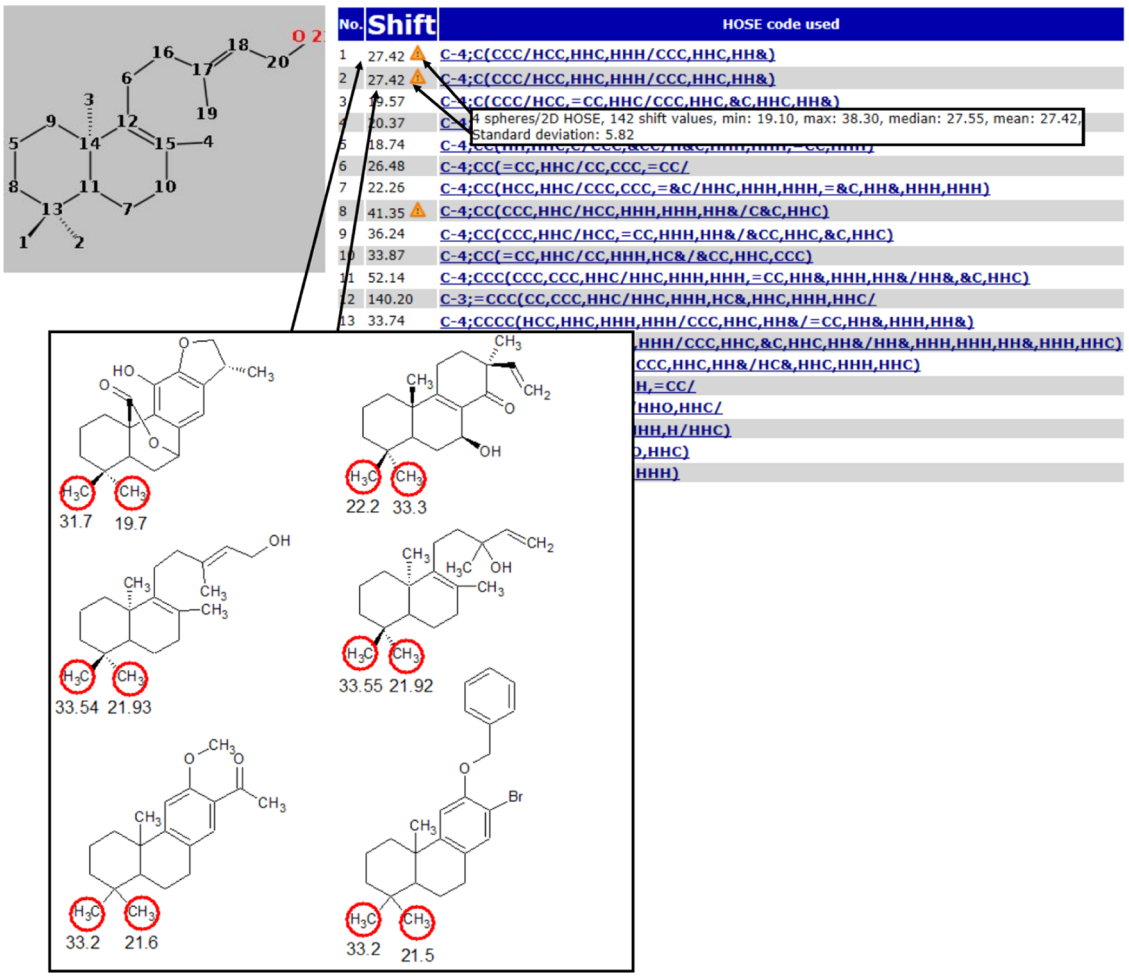


Figure 2. ^{13}C prediction done in nmrshiftdb2 for a compound with several chiral centers, using the standard HOSE code. The predictions for atoms 1 and 2 are identical. For atoms 1 and 2, nmrshiftdb2 warns about an unusually wide range of possible values. This is due to the inclusion of shifts for both positions and does not reveal the wrong assignment. Some of the compounds used for the prediction are shown inset, altogether 76 structures are found. The HOSE code used for the search is *C*−4; *C*(*CCC*/*HCC*,*HHC*,*HHH*/*CCC*,*HHC*,*HH*&). They include the structures from Figure 1, in which two atoms are considered equivalent. More compounds are found, including (third row) compounds with no stereospecification, where both atoms are used for the prediction.

Extended connectivity fingerprints (ECFP)[15] are hashed substructure fingerprints. They can include stereochemical features and could be used in a similar way to our encoding for database searches. Due to the hashing, it is impossible to reconstruct molecules from the fingerprints without a database look up. Also, there is the possibility of collisions, and they are not human-readable. On the other hand, they are very compact and memory-efficient. A major disadvantage is that they cannot handle partial stereo matches or incompletely defined stereochemistry. Finally, Yamashita et al.[16] used atom environments for calculating molecular similarity. These environments provide similarity, not exact matches, and do not take stereochemistry into account.

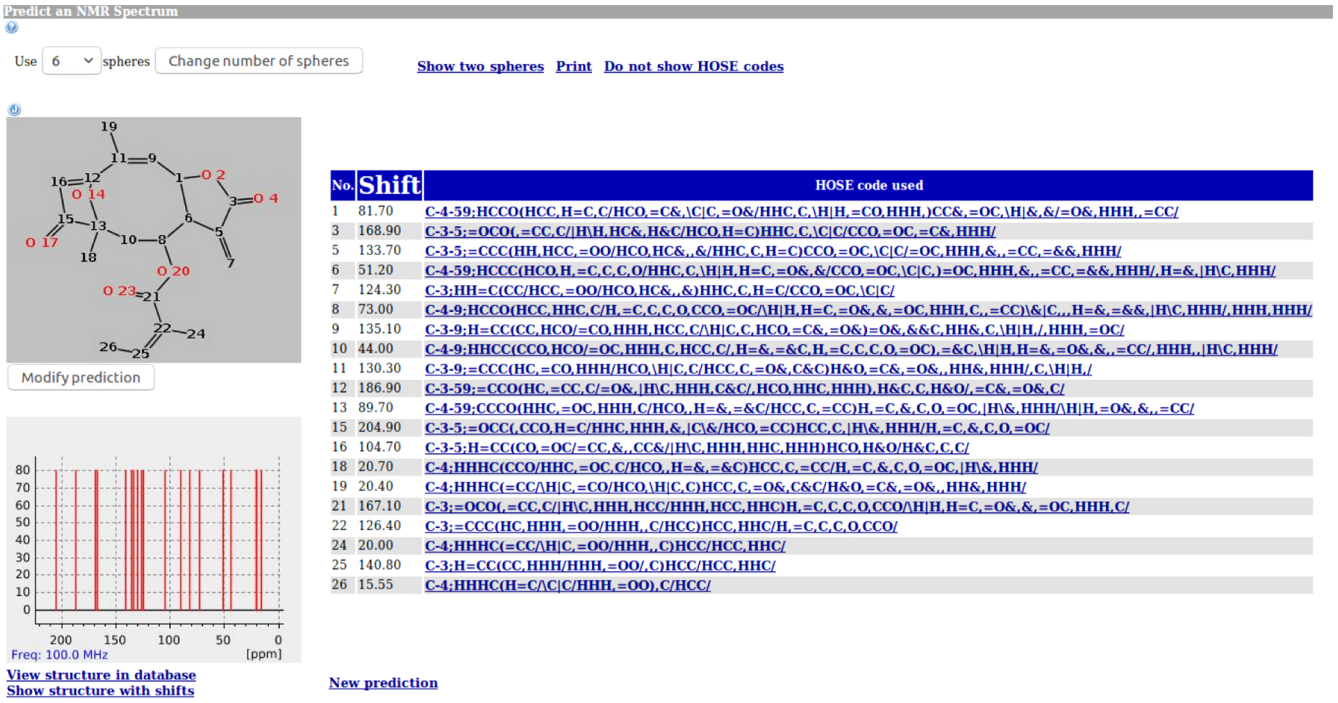
The approach of using artificial intelligence methods, e.g., neural networks, for predicting molecular properties has been widely used.[a] Typically, these methods work on three-dimensional structures, thereby also including stereochemistry. A major drawback of such methods is that three-dimensional coordinates are not known and must be calculated. This supposes a degree of accuracy in the training data, which is typically not there: The actual sample was probably a mixture of conformers.

RESULTS AND DISCUSSION

The stereo HOSE code can show chemical situations not represented by standard HOSE codes. In Table 1, two stereoisomers of the molecule in Figure 5 are shown with the HOSE codes for each atom. It should be noted that in this molecule, atom 1 is a chiral center. The standard HOSE code is identical for equivalent atoms in the two isomers, whereas the stereo HOSE code distinguishes the isomers.

We demonstrate the power of our approach with two examples of chemical shift predictions from nmrshiftdb2.[21] The stereo HOSE code, following the specification given here, has been implemented in nmrshiftdb2. The usefulness of the encoding is demonstrated by Figures 1 and 2. Figure 1 shows a ^{13}C NMR prediction done in nmrshiftdb2 for a compound with a chiral center and diastereotopic methyl groups. Atom 14 is a chiral center; the HOSE codes for its neighbors (e.g., 9) include stereochemistry information and would be different for another isomer. Atoms 1 and 2 are diastereotopic. For them, we get different shift values predicted, which would not be the case with conventional HOSE codes, since they could not distinguish these two atoms. Note that a stereo SMILES would not include stereochemical information here, whereas for the stereo-

DOI: 10.1021/acsomega.9b00488
ACS Omega 2019, 4, 7323−7329



No.	Shift	HOSE code used
1	81.70	C-4-59;HCCO(HCC,H=C,C/HCO,=C&,\C|C,=O&/HHC,C,\H|H,=CO,HHH,)CC&,=OC,\H|&,&/=O&,HHH,,=CC/
3	168.90	C-3-5;=OCO(,=CC,C/|H\H,HC&,H&C/HCO,H=C)HHC,C,\C|C/CCO,=OC,=C&,HHH/
5	133.70	C-3-5;=CCC(HH,HCC,=OO/HCO,HC&,,&/HHC,C,H=C)CCO,=OC,\C|C/=OC,HHH,&,,=CC,=&&,HHH/
6	51.20	C-4-59;HCCC(HCO,H,=C,C,C,O/HHC,C,\H|H,H=C,=O&,&/CCO,=OC,\C|C,)=OC,HHH,&,,=CC,=&&,HHH/,H=&,|H\C,HHH/
7	124.30	C-3;HH=C(CC/HCC,=OO/HCO,HC&,,&)HHC,C,H=C/CCO,=OC,\C|C/
8	73.00	C-4-9;HCCO(HCC,HHC,C/H,=C,C,C,O,CCO,=OC/\H|H,H=C,=O&,&,=OC,HHH,C,,=CC)\&|C,,,H=&,=&&,|H\C,HHH/,HHH,HHH/
9	135.10	C-3-9;H=CC(CC,HCO/=CO,HHH,HCC,C/\H|C,C,HCO,=C&,=O&)=O&,&&C,HH&,C,\H|H,/,HHH,=OC/
10	44.00	C-4-9;HHCC(CCO,HCO/=OC,HHH,C,HCC,C/,H=&,=&C,H,=C,C,C,O,=OC),=&C,\H|H,H=&,=O&,&,,=CC/,HHH,,|H\C,HHH/
11	130.30	C-3-9;=CCC(HC,=CO,HHH/HCO,\H|C,C/HCC,C,=O&,C&C)H&O,=C&,=O&,,HH&,HHH/,C,\H|H,/
12	186.90	C-3-59;=CCO(HC,=CC,C/=O&,\H\C,HHH,C&C/,HCO,HHC,HHH),H&C,C,H&O/,=C&,=O&,C/
13	89.70	C-4-59;CCCO(HHC,=OC,HHH,C/HCO,,H=&,=&C/HCC,C,=CC)H,=C,&,C,O,=OC,|H\&,HHH/\H|H,=O&,&,,=CC/
15	204.90	C-3-5;=OCC(,CCO,H=C/HHC,HHH,&,|C\&/HCO,=CC)HCC,C,|H\&,HHH/H,=C,&,C,O,=OC/
16	104.70	C-3-5;H=CC(CO,=OC/=CC,&,,CC&/|H\C,HHH,HHC,HHH)HCO,H&O/H&C,C,C/
18	20.70	C-4;HHHC(CCO/HHC,=OC,C/HCO,,H=&,=&C)HCC,C,=CC/H,=C,&,C,O,=OC,|H\&,HHH/
19	20.40	C-4;HHHC(=CC/\H|C,=CO/HCO,\H|C,C)HCC,C,=O&,C&C/H&O,=C&,=O&,,HH&,HHH/
21	167.10	C-3;=OCO(,=CC,C/|H\C,HHH,HCC/HHH,HCC,HHC)H,=C,C,C,O,CCO/\H|H,H=C,=O&,&,=OC,HHH,C/
22	126.40	C-3;=CCC(HC,HHH,=OO/HHH,,C/HCC)HCC,HHC/H,=C,C,C,O,CCO/
24	20.00	C-4;HHHC(=CC/\H|C,=OO/HHH,,C)HCC/HCC,HHC/
25	140.80	C-3;H=CC(CC,HHH/HHH,=OO/,C)HCC/HCC,HHC/
26	15.55	C-4;HHHC(H=C/\C|C/HHH,=OO),C/HCC/

Figure 3. Prediction for a compound with an E/Z configuration around bond 22 to 25. The HOSE code for atom 26 contains the | and \ elements in the second sphere. For atom 8, they are found in a higher sphere.

enhanced HOSE code, we use it. In the example, nmrshiftdb2 also informs us that for atom 1 there is an unusually wide range of possible values (from 19.7 to 33.35 ppm). Using the HOSE code, we can look up the values used for the prediction and get the structures in the left inset in Figure 1. Looking at these, it is clear that 19.7 is a suspicious assignment for atom 1. The same way, 31.7 is identified as suspicious for atom 2.

It should be noted that in this example it was specifically the stereo extension, which made the prediction possible. A standard HOSE code (or any other nonstereo aware encoding) would have encoded atoms 1 and 2 in the structure to be predicted identically. It would, therefore, also have performed an identical database search for both and would have predicted both atoms to have indistinguishable shifts. In contrast, we could predict 28.85 and 25.75, respectively. This is due to different encodings. Also due to the different encodings, the search shown in Figure 1 is with the encoding of atom 1 or atom 2 and, therefore, gives only those values assigned to the equivalent atoms in the database. A search with a nonstereo specific encoding would have yielded more results, including eight from the four compounds found with the chiral HOSE code, four of them around 33.5 and four around 19.5. This situation is shown in Figure 2. It would not have been clear that one of them is misassigned, it could only have been concluded that the situation is not properly captured since the values found clearly fall in two groups. To our knowledge, no other encoding system currently used would have captured this situation.

The assignment of 19.7 was most likely either mistranscribed into nmrshiftdb2 or it was misassigned in the original paper.[22] A check of this paper reveals that the assignment there is as currently reported in nmrshiftdb2. A NOESY experiment, which would have revealed the wrong assignment, was not done. A database check not using stereochemistry would also not have revealed the problem. In contrast, the error would have been uncovered by a stereo-aware HOSE code prediction, as provided by nmrshiftdb2.

In Figure 3, we show how the double-bond configuration encoding helps with matching identical structures. Again, we see an nmrshiftdb2 ^{13}C NMR prediction. Looking at the HOSE code for atom 26, we can see the | and \ elements in the second sphere, so only molecules where the configuration is identical will be matched. It also works in higher spheres, for example, the HOSE code for atom 8 shows the symbols in a higher sphere.

To show that there is a systematic improvement and the given examples are not isolated cases, we have used all ^{13}C and ^{1}H chemical shifts contained in nmrshiftdb2 as reference values and predicted a chemical shift for the atom they are assigned to using the stereo HOSE code and the normal HOSE code. For the prediction, we used all of the spectra in nmrshiftdb2 except for the spectrum which contains the reference value (technically, this performs leave-one-out cross-validation). We then calculated the mean error and the root mean squared error for both methods. The values are given in Table 2. There is a clear improvement of the error in all cases. We have only included those predictions where there is a stereochemical situation in the molecule and where a stereochemically matching situation was found for the prediction. If we would have included the other cases, there would have been no difference between the two predictions in this case, and the overall improvement would have been smaller, but still significant.

Table 2. Comparison of the Prediction Results in nmrshiftdb2 Using Standard and Stereo HOSE Codes

		number of examples	mean error	RMSE
^{13}C	standard HOSE code	2703	3.52	0.21
	stereo HOSE code	2703	2.82	0.14
^{1}H	standard HOSE code	1622	0.29	0.03
	stereo HOSE code	1622	0.25	0.02

DOI: 10.1021/acsomega.9b00488
ACS Omega 2019, 4, 7323−7329

Our method also allows for the partial specification of stereochemistry. It follows closely the way stereochemistry is normally depicted by chemists. Therefore, the stereo HOSE code can be generated from structure diagrams alone with no additional information or generation of three-dimensional coordinates needed. We assume that stereochemistry is specified in diagrams following the rules in Brecher,[23] where a drawing is indicating the most specific structure possible, and no additional labels are used. This may in some cases leads to wrong interpretations, but we assume a certain type of encoding here and leave the physical interpretation of drawings to other software or humans.

In a stereo-enhanced HOSE code, the higher spheres are different depending on stereochemistry in lower spheres. This can make it difficult to use interpolation by comparing higher spheres. A possible solution for this would be to store a standard HOSE code in parallel. Furthermore, since NMR shifts are generally more affected by stereochemistry in rigid molecules than in flexible molecules, the advantages of using stereo-HOSE codes will be greater for predicting shifts of rigid molecules.

When applying stereo HOSE codes to chemical shift prediction, we also trace back exactly where our values come from and spot inconsistencies in the database, as demonstrated in the chirality example. This would be difficult to do with machine-learning methods, e.g., neural networks, since with these it is normally not possible to understand how the result was calculated. On the other hand, the method only finds identical fragments, whereas a machine-learning approach can potentially produce good results by finding similarities and implicit rules.

Our approach can also distinguish diastereotopic hydrogens, if they are marked with wedge bonds. Otherwise, a distinction may be made, for example, as part of a shift prediction software, but it is not the part of a structure encoding algorithm, which we discuss here.

In our example, we have shown the utility of stereo-HOSE codes in predicting the NMR shifts of individual atoms within molecules. In addition, stereo-HOSE codes may find a use in similarity searches or property prediction of entire molecules, by converting the stereo-HOSE codes into molecular fingerprints. Although formally specifying a fingerprinting algorithm is beyond the scope of the present work, we can sketch an outline based on previous work including the LINGO[24] algorithm, which counts the frequency of substrings of (nonstereo) SMILES strings, and the ECFP algorithm which is based on connectivity shells centered at each atom in the molecule. Drawing on these algorithms, a (stereo-) HOSE-fingerprint could be as simple as a tabulated count of all HOSE codes from level 1 to *N* in a molecule. In such an algorithm, there would be a substantial difference between fingerprints generated from HOSE codes with stereochemistry assigned fully, partially, or not at all, due to the rearrangement of substrings during the encoding of stereochemistry. The impact of these differences on the quality of search results and structure prediction would depend on the size and quality of the reference database. As mentioned earlier, a simple adaptation would be to store fingerprints from standard and stereo-HOSE codes in parallel.

CONCLUSIONS

HOSE codes are an established tool for describing atom environments and for chemical shift prediction. We have extended the HOSE code to include stereochemical information. The principles are adopted from the well-known stereo-enhanced SMILES notation. Our extended HOSE codes can be generated from the structure without reference to external libraries. They allow specification of partial stereochemistry, similar to how chemical structure diagrams are drawn. Since structure diagrams are still the most common way to communicate compounds in publications, this enables encoding of mainstream chemical information. Using our new stereo-extended HOSE code, we can distinguish atoms which would be encoded identically with traditional HOSE codes. Better chemical shift prediction is a benefit of the approach. The deterministic nature of the algorithms sets it apart from machine-learning approaches. A disadvantage of our approach is that it is affected by a lack of standards in structure drawing and interpretation. The code for generating the extended HOSE codes is available under an open-source licence.

METHOD

The Daylight Theory Manual[2] uses two syntax elements to encode stereochemical information. We use both of these in a similar fashion.

First, configuration around double bonds is specified by the characters / and \. The two compounds in Figure 4 would be

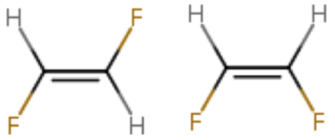


Figure 4. Two possible configurations around a double bond.

represented by the SMILES `F/C=C/F`, respectively, `F/C=C\F`. Alternatively, `F\C=C\F`, respectively, `F\C=C/F` are possible. For the first compound, the slashes are always identical, for the second, they are opposite. A three-sphere stereo HOSE code of the fluorine atom in the left structure in Figure 4 would be `F;C(=C/\F` and for the right structure it would be `F;C(=C/|F`. The rule is to follow the bonds through the molecule (as it is done for HOSE codes) and to notice when we cross a double bond. In front of each of the atoms at the other end of the double bond, we put a \ or | depending on if they are opposite or same side of the double bond seen from where we came from. In the left structure, in Figure 4, the other fluorine is the opposite of where we came from, so we put a \ in front of it. In the right structure, the fluorine is on the same side, so it gets a | (the / character is used in HOSE codes for separating spheres, so we use | instead of / to avoid confusion). Note that in contrast to SMILES, only one version is allowed, so our HOSE code is canonical. If the hydrogen on the carbon would be included in the HOSE code (we do not make assumptions about valencies or explicit hydrogens, just like the original HOSE code), then the codes would be `F;C(=C/|H\F` and `F;C(=C/\H|F`, respectively.

Second, for chiral centers, SMILES uses a specification based on local chirality. For this, the order in which neighbors occur in the SMILES string, seen from a certain atom, is used. When building a HOSE code, a "point of view" is provided, since the encoding starts with the center atom and proceeds outward. Therefore, the principle of stereo SMILES naturally extends to HOSE codes. In Figure 5, there is a tetrahedral configuration around the carbon atom with number 2. If we want to build a stereo HOSE, for example, for atom number 3, we get `C-4;CC(@CNC,/O,,)//` (spheres beyond the third are

DOI: 10.1021/acsomega.9b00488
ACS Omega 2019, 4, 7323−7329

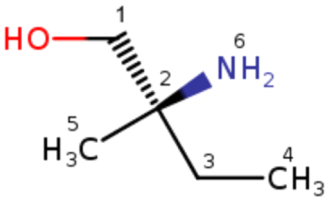


Figure 5. Tetrahedral configuration around a carbon atom.

empty). As in ordinary HOSE codes, we start with the atom for which the HOSE code is built and encode it as C-4. We then get in the first sphere two single carbon atoms (written as CC). The second sphere contains the atoms around the chiral center, therefore we put an @. When specifying chirality, we look along the axis from the atom in the previous sphere to the chiral center (atom 3 to atom 4 in this case), giving `@CNC`. The first atom in the list is the atom which would be first in the standard HOSE code. Notice that in the second sphere, the two carbon atoms are not distinguished, but in the third sphere, one of the carbon atoms has an oxygen atom attached. Consequently, `C-4;CC(@CNC,/,,O/)//` would indicate a different chiral configuration. Note that we do not allow (as opposed to the SMILES specification) the use of `@@` with reverse atom order. This is to keep the HOSE codes canonical. The ring closure symbol & is treated like an atom and put wherever the atom in the ring would be from the point of view. If the chiral center is encountered in a higher sphere, we put the `@` in a higher sphere as well, ordering the atoms according to the same rule. If we want to generate a stereo-HOSE code focused at an atom which is itself a chiral center, the rule is to put the @ at the start of the first sphere. In this case, the first atom following the @ is the atom which would come first in the standard HOSE code and the second is the atom which would come second in the standard HOSE code. The point of view is the axis from the focus atom to the first atom after the @. The other atoms follow in the order they are seen, going counterclockwise starting with the second atom after the @. The encoding of atom 2 in Table 1 is an example for this.

We use the same principle to transfer the encoding of the other types of chirality given in Section 3.3.4 of the Daylight Theory Manual:[2] Only `@` allowed, the point of view is the atom in the previous shell, the first atom to list is given by the HOSE code specification. For square-planar chiralities, only `SP1` is allowed. In all other cases, the default class with a single `@` is used.

If a chemical structure is to be encoded in, e.g., a SMILES string, stereochemistry is only considered around chiral centers. Even if wedge bonds are used somewhere else in a depiction of the structure, they cannot be relevant for the identification of the structure and are, therefore, disregarded. For the stereo-enhanced HOSE codes, we encode also wedge bonds on diastereotopic atoms, since they are chemically different and have to be encoded differently in an atom-centric encoding. We have shown an example of this in section Results and Discussion.

We require all hydrogens to be explicitly specified. This is necessary around chiral centers, but since the space taken up by hydrogen information is not a problem nowadays, we stipulate explicit hydrogens everywhere to generate the stereo HOSE code.

Finally, there are cases where the priority in the sphere is only relevant due to stereochemistry in a higher sphere. In Figure 6, stereochemistry is partially defined. If a standard HOSE code would be built for atom 3, it would read `C-4;CCC-`

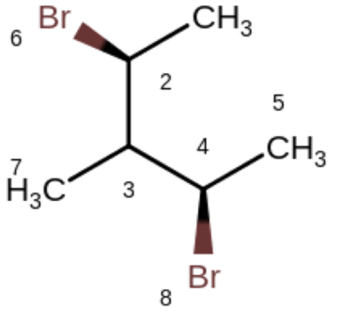


Figure 6. Structure where stereochemistry influences priority.

`(CY,CY,//)` (note that bromine is represented by `Y` in HOSE codes). Here, the first two carbons in the first sphere can represent atom 2 or 4, since they cannot be distinguished in the higher spheres and are, therefore, equivalent. Considering stereochemistry, the higher spheres are different. The HOSE code generated by our method would be `C-4;CCC(@HCY,@HYC,//)` (with hydrogens around the chiral centers). Notice that in the second sphere, the atoms attached are different. We, therefore, need a rule for priority here. For this, we take the atoms attached to the two carbons, for which we need a priority, in the order they appear in the next sphere. The first atom `H` in both cases, the next is `C` respectively `Br`. Since `C` takes priority, the `HCY` group takes priority before the `HYC` group. This example also shows a partial specification of stereochemistry, since there is none given around atom 3. If there would be one, this would give priorities and the HOSE codes would start with an `@` in the first sphere and be different depending on the stereochemistry around atom 3.

AUTHOR INFORMATION

Corresponding Author

*E-mail: stefan.kuhn@dmu.ac.uk.

ORCID

Stefan Kuhn: 0000-0002-5990-4157

Sean R. Johnson: 0000-0001-8261-9015

Present Address

§Conagen Inc., 15 Deangelo Drive, Bedford, MA 01730, United States (S.R.J.).

Notes

The authors declare no competing financial interest.

Java code for generating stereo-aware HOSE codes is available under https://sourceforge.net/p/nmrshiftdb2/code/HEAD/tree/trunk/nmrshiftdb2/src/java/org/openscience/nmrshiftdb/util/ExtendedHOSECodeGenerator.java as part of the nmrshiftdb2 project. It is based on the HOSECodeGenerator class of the Chemistry Development Kit and is licenced under the GNU Affero General Public License.

ACKNOWLEDGMENTS

The authors thank all contributors to the nmrshiftdb2 database and software for their cooperation over many years. Special thanks go to Nils Schlörer and the NMR facility at the Chemistry Department of Universität zu Köln for hosting and advice. Part of the work by S.K. was funded by Deutsche Forschungsgemeinschaft, Grant Numbers: LI 2858/11, SCHL 580/31.

ADDITIONAL NOTE

[a]Ref 17 is a good overview, refs 11, 18−20 are some papers in the area.

DOI: 10.1021/acsomega.9b00488
ACS Omega 2019, 4, 7323−7329

REFERENCES

(1) Favre, H. A.; Powell, W. H. *Nomenclature of Organic Chemistry*; The Royal Society of Chemistry, 2014; pp P001−P1568.
(2) Daylight Theory Manual. http://www.daylight.com/dayhtml/doc/theory/.
(3) Bremser, W. Hose − a novel substructure code. *Anal. Chim. Acta* **1978**, *103*, 355−365.
(4) Wiswesser, W. J. The Wiswesser Line Formula Notation. *Chem. Eng. News* **1952**, *30*, 3523−3526.
(5) Ash, S.; Cline, M. A.; Homer, R. W.; Hurst, T.; Smith, G. B. SYBYL Line Notation (SLN): A Versatile Language for Chemical Structure Representation. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 71−79.
(6) InChI and InChIKeys for chemical structures. https://www.inchi-trust.org/.
(7) Hähnke, V. D.; Bolton, E. E.; Bryant, S. H. PubChem atom environments. *J. Chem. Inf. Comput. Sci.* **2015**, *7*, 1−37.
(8) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Similarity Searching of Chemical Databases Using Atom Environment Descriptors (MOLPRINT 2D): Evaluation of Performance. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1708−1718.
(9) Wasko, M. J.; Pellegrene, K. A.; Madura, J. D.; Surratt, C. K. A Role for Fragment-Based Drug Design in Developing Novel Lead Compounds for Central Nervous System Targets. *Front. Neurol.* **2015**, *6*, No. 197.
(10) Rogers, D.; Brown, R. D.; Hahn, M. Using extended-connectivity fingerprints with Laplacian-modified Bayesian analysis in high-throughput screening follow-up. *J. Biomol. Screening* **2005**, *10*, 682−686.
(11) Kuhn, S.; Egert, B.; Neumann, S.; Steinbeck, C. Building blocks for automated elucidation of metabolites: machine learning methods for NMR prediction. *BMC Bioinf.* **2008**, *9*, 400.
(12) Filimonov, D.; Poroikov, V.; Borodina, Y.; Gloriozova, T. Chemical Similarity Assessment through Multilevel Neighborhoods of Atoms: Definition and Comparison with the Other Descriptors. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 666−670.
(13) Schütz, V.; Purtuc, V.; Felsinger, S.; Robien, W. CSEARCH-STEREO: A new generation of NMR database systems allowing three-dimensional spectrum prediction. *Fresenius' J. Anal. Chem.* **1997**, *359*, 33−41.
(14) Spanton, S. G.; Whittern, D. The development of an NMR chemical shift prediction application with the accuracy necessary to grade proton NMR spectra for identity. *Magn. Reson. Chem.* **2009**, *47*, 1055−1061.
(15) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(16) Yamashita, H.; Higuchi, T.; Yoshida, R. Atom Environment Kernels on Molecules. *J. Chem. Inf. Model.* **2014**, *54*, 1289−1300.
(17) Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design: An Introduction*, 2nd ed.; Wiley: New York, 1999.
(18) Bernazzani, L.; Duce, C.; Micheli, A.; Mollica, V.; Sperduti, A.; Starita, A.; Tiné, M. R. Predicting Physical−Chemical Properties of Compounds from Molecular Structures by Recursive Neural Networks. *J. Chem. Inf. Model.* **2006**, *46*, 2030−2042.
(19) CSEARCH-NMR-Server. https://nmrpredict.orc.univie.ac.at/.
(20) Meiler, J. PROSHIFT: protein chemical shift prediction using artificial neural networks. *J. Biomol. NMR* **2003**, *26*, 25−37.
(21) Kuhn, S.; Schlorer, N. E. Facilitating quality control for spectra assignments of small organic molecules: nmrshiftdb2-a free in-house NMR database with integrated LIMS for academic service laboratories. *Magn. Reson. Chem.* **2015**, *53*, 582−589.
(22) Marrero, J. G.; Andres, L. S.; Luis, J. G. Semisynthesis of rosmanol and its derivatives. Easy access to abietatriene diterpenes isolated from the genus Salvia with biological activities. *J. Nat. Prod.* **2002**, *65*, 986−989.
(23) Brecher, J. Graphical representation of stereochemical configuration (IUPAC Recommendations 2006). *Pure Appl. Chem.* **2006**, *78*, 1897−1970.
(24) Vidal, D.; Thormann, M.; Pons, M. LINGO, an Efficient Holographic Text Based Method To Calculate Biophysical Properties and Intermolecular Similarities. *J. Chem. Inf. Model.* **2005**, *45*, 386−393.

DOI: 10.1021/acsomega.9b00488
ACS Omega 2019, 4, 7323−7329