Carbohydrate Structure Generalization Scheme for Database-Driven Simulation of Experimental Observables, Such as NMR Chemical Shifts

ARTICLE *in* JOURNAL OF CHEMICAL INFORMATION AND MODELING · JULY 2014

Impact Factor: 3.74 · DOI: 10.1021/ci500267u · Source: PubMed

3 AUTHORS:

Roman R. Kapaev
Russian Academy of Sciences
4 PUBLICATIONS **4** CITATIONS
SEE PROFILE

Ksenia Egorova
Russian Academy of Sciences
20 PUBLICATIONS **169** CITATIONS
SEE PROFILE

Philip Toukach
Russian Academy of Sciences
75 PUBLICATIONS **877** CITATIONS
SEE PROFILE

JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Carbohydrate Structure Generalization Scheme for Database-Driven Simulation of Experimental Observables, Such as NMR Chemical Shifts

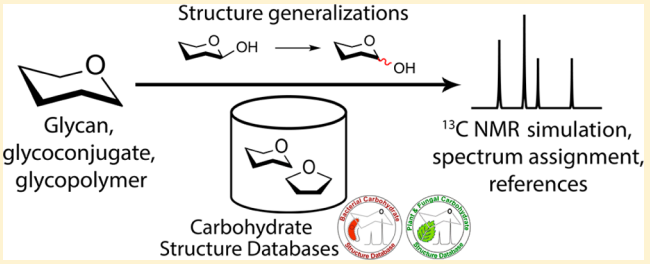
Roman R. Kapaev,[†] Ksenia S. Egorova,*[,‡] and Philip V. Toukach*[,‡]

[†]Higher Chemical College of the Russian Academy of Sciences, Miusskaya sq. 9, Moscow 125047, Russia

[‡]N. D. Zelinsky Institute of Organic Chemistry, Russian Academy of Sciences, Leninsky prospect 47, Moscow 119991, Russia

Supporting Information

ABSTRACT: Carbohydrates play an immense role in different aspects of life. NMR spectroscopy is the most powerful tool for investigation of these compounds. Nowadays, progress in computational procedures has opened up novel opportunities giving an impulse to the development of new instruments intended to make the research simpler and more efficient. In this paper, we present a new approach for simulating ^{13}C NMR chemical shifts of carbohydrates. The approach is suitable for any atomic observables, which could be stored in a database. The method is based on sequential generalization of the chemical surroundings of the atom under prediction and heuristic averaging of database data. Unlike existing applications, the generalization scheme is tuned for carbohydrates, including those containing phosphates, amino acids, alditols, and other non-carbohydrate constituents. It was implemented in the Glycan-Optimized Dual Empirical Spectrum Simulation (GODESS) software, which is freely available on the Internet. In the field of carbohydrates, our approach was shown to outperform all other existing methods of NMR spectrum prediction (including quantum-mechanical calculations) in accuracy. Only this approach supports NMR spectrum simulation for a number of structural features in polymeric structures.



1. INTRODUCTION

Glycoscience has recently exhibited vigorous development and become one of the leading fields in chemical and biological research. Novel design and synthesis of carbohydrates have revealed new opportunities in discovery of medically important compounds.[1−4] Multiple modern drugs and vaccines have active carbohydrate constituents.[5] Carbohydrates compose most of the renewable biomass on Earth;[6] they are the most abundant natural polymers and have acquired industrial significance. However, in spite of the massive development of fascinating applications, carbohydrates remain the least structurally characterized among the major classes of biological molecules.[7]

The major tool for carbohydrate structural studies is NMR spectroscopy, as X-ray crystallography and mass spectrometry were reported to be less efficient because of complications in crystallization and a multitude of stereoisomers, respectively.[8,9] However, further insight into NMR research is limited by difficulties in the interpretation of NMR observables: signal assignment and elucidation of relationships between measured parameters and molecular structure is still a tiresome task.[7] Fortunately, progress in informatics has opened up novel opportunities giving an impulse for the development of new instruments intended to make the research more efficient and simple. Modeling of carbohydrate structure and molecular properties has benefited from a variety of computational methods.[10] Particularly, simulators of NMR spectra are undoubtedly helpful tools, especially in carbon NMR spectroscopy.[7]

Among different methods of simulation and automated assignment of ^{13}C chemical shifts of carbohydrates, empirical approaches realized in such software as ACD/NMR,[11] BIOPSEL,[12] CASPER,[13,14] Modgraph NMRPredict,[15] and others show more accurate results than semiempirical and quantum-chemical calculations.[7]

Methods based on hierarchical organization of spherical environments (HOSE)[16] provide a universal database-driven prediction scheme; however, they possess poorer accuracy than dedicated carbohydrate-optimized methods.[7] We explain this by the limited applicability of structure generalization algorithms common in organic chemistry to carbohydrates with their flexibly connected rigid building blocks and multitude of stereoconfigurations. To our knowledge, HOSE has not been tuned for carbohydrates to date. A few examples of dedicated software have been reported to simulate NMR spectra of glycans, all using an empirical incremental approach and relatively small special databases.[7] However, dedicated NMR databases exist (NMRshiftDB,[17] CSDB,[18] OSCAR,[19,20]

Received: May 5, 2014
Published: July 14, 2014

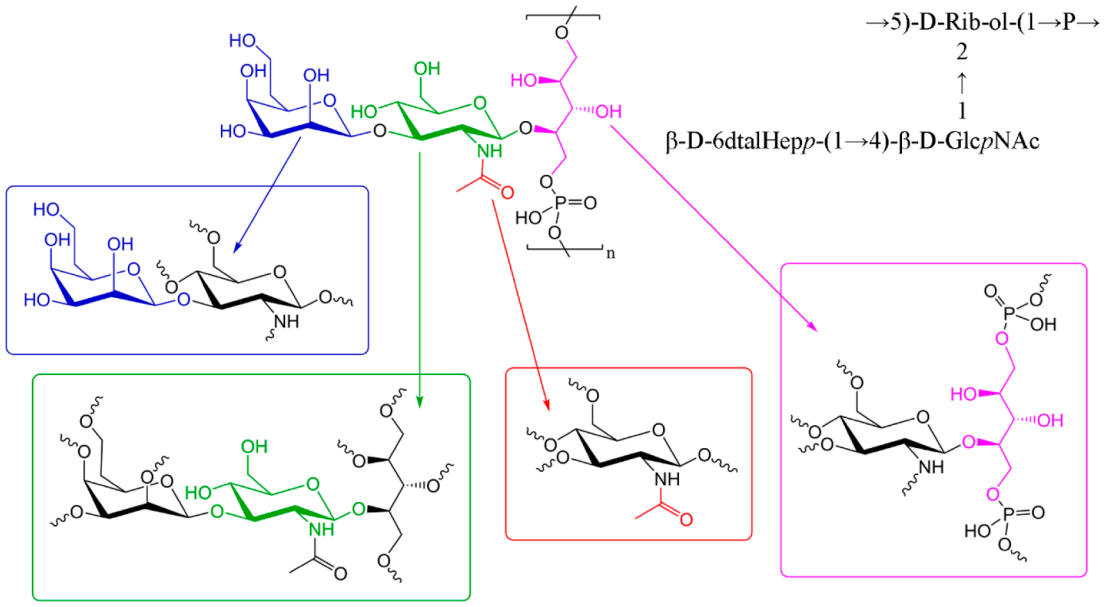


Figure 1. Slicing of a glycan into fragments using as an example the *O*-polysaccharide from *Campylobacter coli* O30.[23] The central residue in each fragment is depicted in color. Wavy bonds indicate any substituent.

and others), and it has been desirable to combine their data power with a glyco-tuned algorithm using a HOSE idea.

Simple averaging and outlier removal are not enough for chemical shift prediction, as they do not distinguish the structural surroundings well enough. Existing databases cannot cover the full scope of chemically possible saccharides; thus, thorough structure generalization is needed. Here we report a carbohydrate structure generalization scheme suitable for simulation of any database-represented atomic parameters. The scheme was verified and approved by simulation of ^{13}C NMR chemical shifts, one of the observables most demanded in structural studies of natural saccharides.

2. SIMULATION SCHEME

The simulation scheme uses a novel structure generalization algorithm and statistical processing of data from the Bacterial Carbohydrate Structure Database (BCSDB)[21] and the Plant & Fungal Carbohydrate Structure Database (PFCSDB).[22] The details are provided in section 4.

2.1. Algorithm of Structure Generalization for ^{13}C NMR Chemical Shift Simulation. Prior to gathering and processing data from a database, the simulation engine divides an input carbohydrate structure into fragments. For each atom, the corresponding fragment includes a "central" residue containing the atom under prediction and all of the adjacent residues that occupy the neighboring topological nodes in the glycan (see Figure 1). Here and below, a residue is a monosaccharide, alditol, amino acid, fatty acid, or any other structural subunit attached to the other part of a molecule by an ester, ether, or amide linkage. The simulation engine processes all fragments to predict properties of all atoms in their central residues. The final predicted NMR spectrum, chemical shift assignment table, and other data are superimposed from independently obtained data for each fragment. The aim of this separation is to make structure alterations less cumbersome and time-consuming.

To predict a carbon chemical shift, the simulator searches the database for a fragment to which the atom belongs. If a sufficient number of instances (depending on the selected quality mode; see section 3) of the fragment is found, the chemical shift values assigned to this atom in all of the fragments are checked for statistical outliers (see section 2.6) and averaged. Otherwise, generalizations of the structural surroundings (from minor to major) are applied until the obtained structural pattern is found in the database (see Figure 2). A generalization act is a modification that changes fragment descriptors so that more potential structures match it. For example, a permutation of α-D-Fuc*p* into D-Fuc*p* is a generalization act, as the former corresponds to a single fully defined residue, while the latter matches two residues, α-D-Fuc*p* and β-D-Fuc*p*.

2.2. Weight Factors. Every structural parameter of a fragment that can be generalized is assigned an empirical weight factor depending on the strength of the expected effect of this parameter on the chemical shift of the predicted atom. The estimation of this dependence empirically accounts for the nature of a parameter, the number of bonds between the permutation and the predicted atom, and the nature of the central residue in the current fragment. The generalization sequence starts from parameters with the lowest weight factors, which are related to the most distal and conformationally flexible groups of atoms. This scheme aims to find generalization acts with a minimal impact on the predicted atom.

The weight factors were obtained by the procedure described in section 2.3. To increase prediction accuracy, we implemented separate sets of weight factors for pyranoses, furanoses, and all of the other residues, including monosaccharaides in the open form. Tables 1 and 2 list the weight factors optimized for three parameter sets. More thorough quantification of structure types and the introduction of dedicated weight factor sets are questions for further studies.

The proximity factor (PF) reflects the number of bonds between the predicted atom and the linkage with a substituent residue. PFs are used to calculate the weights of permutations affecting the substituents of a central residue, as shown in Table 2. The PFs are listed in Table 3.

The weight factors range from 0 to 20, with higher values corresponding to stronger effects. With the exception of special

dx.doi.org/10.1021/ci500267u | J. Chem. Inf. Model. 2014, 54, 2594−2611

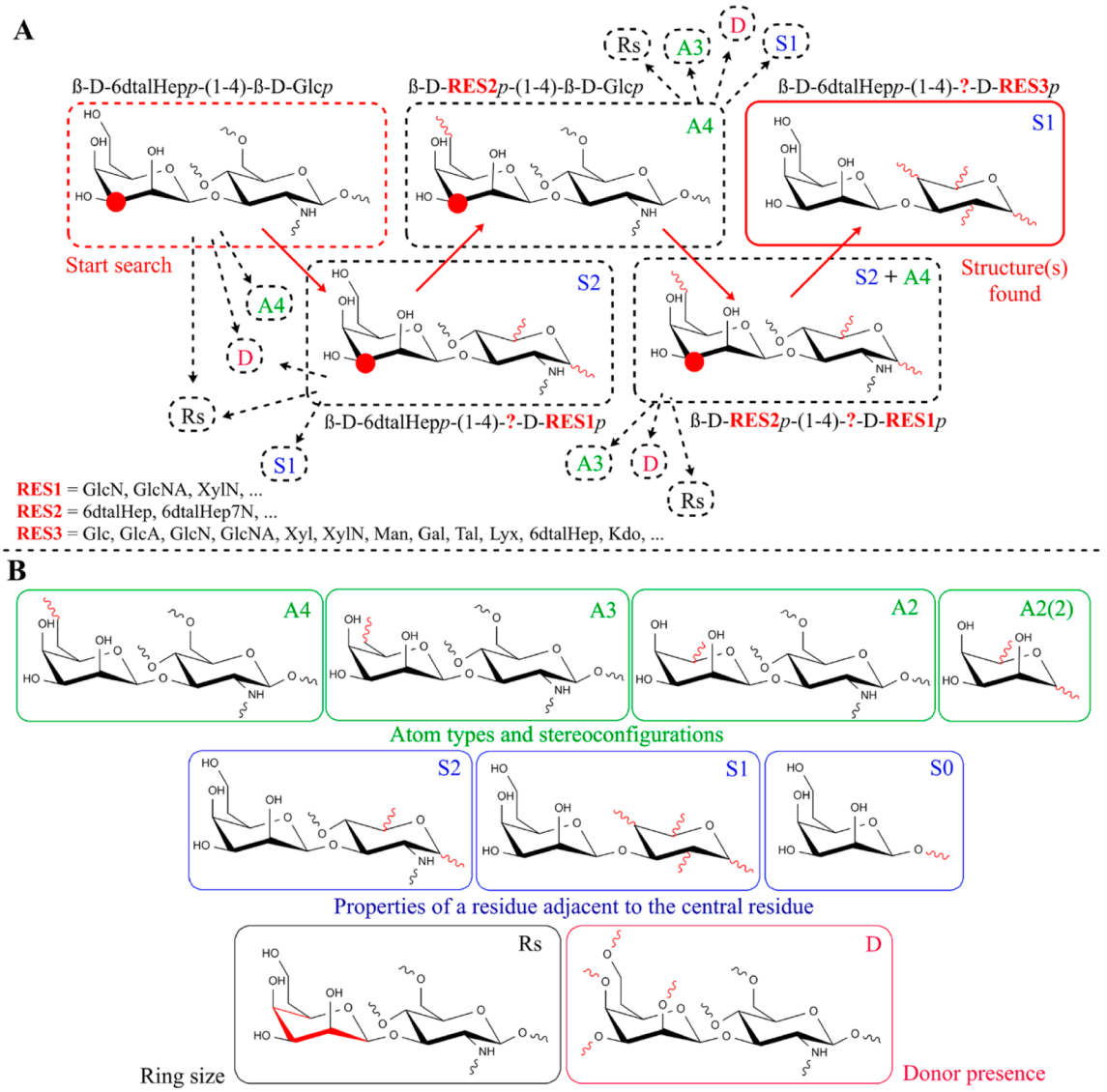


Figure 2. (A) Exemplary generalization scheme: prediction of C3 (red dot) of 6-deoxy-β-D-taloheptose in the blue fragment from Figure 1. The generalization pathway (red arrows) is chosen from among all possible pathways (dashed arrows). (B) Possible elementary generalizations. Letters (A, S, Rs, and D) stand for descriptor types and are deciphered next to each group of pictures. For "A" generalizations, the numbers indicate the remoteness of the nearest generalized atom from the predicted one; for "S" generalizations, the numbers stand for the remoteness of the nearest generalized atom from the atom forming a linkage to the central residue. Wavy bonds indicate atoms forming any possible link; the wavy bond in "S0" indicates any substituent in any stereoconfiguration. Red bonds accentuate certain properties being generalized. For "Rs" generalization, the ring size can be pyranose, furanose, or open-chain.

cases (see section 2.5), each property of a central residue (Table 1) is generalized independently from the other properties and submits only to the principle of total weight minimization. Thus, any allowed combination of listed permutations can appear in the central residue. On the contrary, properties of substituent residues are always generalized conjointly. The generalization pathways for substituents depend on the selected quality mode (see section 3.1).

For atom types and stereoconfigurations of central residues, consecutive generalizations are applied (Table 1). Generalizations of closer atoms are possible if distal atoms have already been generalized. Thus, during processing of C1 of a galactose residue in the exemplary α-D-Glcp-(1→3)-α-D-Galp fragment, generalization of its C6 is applied first (weight W_1 = 7.16). Generalizations of the following atoms are subsequently applied in the next steps: C4 (weight $W_2 = 7.16 + W_{FAR}$; $W_{FAR} = W_1$), C5 ($W_3 = 0.77 + W_{FAR}$; $W_{FAR} = W_2$), C3 ($W_4 = 0.77 + W_{FAR}$; $W_{FAR} = W_3 + W_{SUBST}$, where W_{SUBST} is the weight for the generalization of a glucose residue into any residue), and so on.

The generalization scheme was tuned to achieve the optimal accuracy to performance ratio. Particularly, independent generalization of stereoconfigurations and types of atoms in a central residue would lead to performance loss due to combinatorial growth of the number of possible generalizations. In a hexose residue, there are 4096 mathematically possible variants of orientation and type of atoms (let alone other generalizations), and most of them are inefficient as generalizations. Because of this, the stereo and type properties of atoms are generalized simultaneously, cutting off a priori nonexistent generalizations and decreasing the number of variants ca. 700 times. Similarly, substituent generalizations

dx.doi.org/10.1021/ci500267u | J. Chem. Inf. Model. 2014, 54, 2594−2611

Table 1. Descriptors and Weight Factors of a Residue Containing the Atom under Prediction (the "Central Residue") [Property Remoteness (R) Shows the Number of Bonds from the Permutation to the Atom under Prediction; Three Weight Factor Sets Are Provided, Corresponding to Pyranose, Furanose, and Common Parameter Sets]

		weight factors			generalization example		
changed property	property remoteness (R)	pyranose	furanose	common	atom predicted	structure	structure after generalization
atom stereoconfiguration and atom type[a]	≥4 (distal atom)		$10^{-R+2} + W_{FAR}$		C1	β-D-6dtalHep*p*	β-D-RES1*p* (RES1 = 6dtalHep,6dTal, ...)
						[1111200][b]	[111120*]
						[ooooddo][c]	[oooodd*]
	3	7.16 + W_{FAR}	7.15 + W_{FAR}	0.89 + W_{FAR}		β-D-RES1*p*	β-D-RES2*p* (RES2 = RES1, Tal, TalA, ...)
						[111120*]	[11112*]
						[oooodd*]	[ooood*]
							(also applied for C4 as well as for C6)
	2	0.77 + W_{FAR}	0.71 + W_{FAR}	5.05 + W_{FAR}		β-D-RES3*p* (RES3 = RES2, Tal, Man, Rha, Man4N, ...)	β-D-RES4*p* (RES4 = RES3, Lyx, ...)
						[111?2*]	[111*]
						[ooo?d*]	[ooo*]
							(also applied for C3 as well as for C5)
	1	2.80 + W_{FAR}	3.91 + W_{FAR}	5.69 + W_{FAR}		β-D-RES5*p* (RES5 = RES4, Alt, Ido, Ara, IdoN, ...)	β-D-ANY*p* (ANY = any sugar residue)
						[11*]	[1*]
						[oo*]	[o*]
	0 (predicted atom)	6.62 + W_{FAR}	0.73 + W_{FAR}	1.09 + W_{FAR}		β-D-ANY*p*	D-ANY*p* (α or β)
						[1*]	[*]
						[o*]	[*]
ring size (pyranose/furanose/linear)	≥3 (distal exocyclic atom)		10^{-R+2}	unapplicable	C9	α-Neu*p*	α-Neu (*p* or *f*)
	2 (next to proximal exocyclic atom)	1.00	7.99	unapplicable	C7	β-D-6dtalHep*f*	β-D-6dtalHep (*p* or *f*)
	1 (proximal exocyclic atom)	12.36	14.51	unapplicable	C6	β-D-Fuc*f*	β-D-Fuc (*p* or *f*)
	0 (endocyclic atom)	15.53	19.98	unapplicable	C1	β-D-Fuc*f*	β-D-Fuc (*p* or *f*)
presence of an acceptor (reducing end/inline)	≥4 (atom distant from anomeric)		10^{-R+2}		C7		
	3	3.59	4.03	1.91	C4		
	2 (atom next to neighboring)	9.88	6.73	7.39	C3	β-D-6dtalHep*p*	β-D-6dtalHep*p* or β-D-6dtalHep*p*-(1→
	1 (atom next to anomeric)	12.61	10.42	10.14	C2		
	0 (anomeric atom)	17.13	17.18	17.94	C1		
presence of donor(s), for terminal residues	? (any atom)	6.86	13.76	16.01	any	β-D-6dtalHep*p*-(1→	β-D-6dtalHep*p*-(1→ that may be substituted at C2, C3, C4, C7
presence of donor(s), for nonterminal residues	? (any atom)	6.39	5.25	8.11	any	→4)β-D-6dtalHep*p*-(1→	→4)-β-D-6dtalHep*p*-(1→ that may be substituted at C2, C3, C7

[a]Generalization of stereoconfiguration and type of atom of a central residue implies generalization of all more distant atoms and full generalization of substituent residues linked to generalized atoms. W_{FAR} stands for the sum of the weight factors of all distal atoms and all substituents forming bonds with the altered atom (see the text). Consecutive generalization (from distal to close atoms) of β-D-6dtalHep*p* is given in the example. [b]Here and below, the atomic stereoconfiguration is recorded according to the MonosaccharideDB (http://www.monosaccharidedb.org/) stereocode: "1" indicates carbon with the L configuration (OH group pointing to the left in the Fischer projection); "2" indicates the D configuration (OH group pointing to the right in the Fischer projection); "0" is for achiral carbons. [c]Here and below, the following atom types are used: o = >CH−OH (hydroxy); n = >CH−NH− (amino); a = −COOH (carboxy); d = $-CH_2-$ or >CH− (deoxy); ? = unknown; x = other carbon; * = any set of atoms.

were associated with generalization thresholds (see section 3.1) to bypass inefficient generalization pathways.

To simplify the calculations, we made the weight factors dependent on the number of bonds between the predicted and generalized atom(s) rather than on the interatomic distances. Possible conformational effects on chemical shifts were neglected, as the conformations of sparse flexible bonds existing in glycans depend on connected residues rather than on the whole molecule. In contrast to proteins,[24] glycans do not tend to form tertiary and higher structures that may disturb glycosidic bond torsions; they do not have long-range NOEs,[25] and they display similar 3D structures for fragments with similar primary structures.[26−28]

To corroborate this, we analyzed experimental chemical shifts of interglycosidic atoms for a widespread disaccharide fragment (Gal C1 and Glc C4 in β-D-Gal*p*-(1→4)-β-D-Glc*p*) in 48 molecules deposited in BCSDB (spectra recorded in D_2O). Fragments containing bisubstituted residues (e.g., β-D-Gal*p*-

dx.doi.org/10.1021/ci500267u | J. Chem. Inf. Model. 2014, 54, 2594−2611

Table 2. Weight Factors of Permutations of Substituent Residues[a] [Property Remoteness (R) Stands for the Number of Bonds from the Atom(s) Being Generalized to the Atom Forming a Linkage with the Central Residue; PF Stands for Proximity Factor (See Table 3)]

		weight factor[b]			generalization example	
changed property	property remoteness (R)	pyranose	furanose	common	structure	structure after generalization
atom stereoconfiguration	≥4 (atom distal from linkage)		10^{-R} × PF		α-Leg$p$-(1→RES	β-D-RES1p-(1→RES (RES1 = Leg, 8eLeg, ...) [aodondnod] [0202112?0]
	3	0.004 × PF	0.002 × PF	0.005 × PF		β-D-RES1p-(1→RES (RES1 = Glc, Gal) [121?20] [oooodo]
	2 (atom next to neighboring)	0.16 × PF	0.16 × PF	0.02 × PF	β-D-Glc$p$-(1→RES	β-D-RES1p-(1→RES (RES1 = Glc, All) [12?220] [oooodo]
	1 (atom neighboring to linkage)	1.30 × PF	1.87 × PF	0.24 × PF		β-D-RES1p-(1→RES (RES1 = Glc, Man) [1?1220] [oooodo]
	0 (atom involved in linkage)	6.49 × PF	7.00 × PF	5.30 × PF		D-Glc$p$-(1→RES (α or β) [?21220] [oooodo]
atom type	≥4 (atom distal from linkage)		10^{-R} × PF		β-D-6dtalHep$p$-(1→RES	β-D-RES1p-(1→RES (RES1 = 6dtalHep, ...) [1111200] [oooodd*]
	3	0.008 × PF	0.09 × PF	0.09 × PF		β-D-RES1p-(1→RES (RES1 = Glc, Glc4N, ...) [121220] [ooo?do]
	2 (atom next to neighboring)	0.49 × PF	0.71 × PF	0.74 × PF	β-D-Glc$p$-(1→RES	β-D-RES1p-(1→RES (RES1 = Glc, Glc3N, ...) [121220] [oo?odo]
	1 (atom neighboring to linkage)	2.98 × PF	4.87 × PF	2.41 × PF		β-D-RES1p-(1→RES (RES1 = Glc, GlcN, ...) [121220] [o?oodo]
	0 (atom involved in linkage)	19.03 × PF	7.79 × PF	19.14 × PF		β-D-RES1p-(1→RES (RES1 = Glc, Glc1N, ...) [121220] [?ooodo]
ring size	≥3 (atom distal from endocyclic atoms)		10^{-R} × PF		RES→9)-α-Neu$p$	RES→9)-α-Neu (p or f)
	2	0.93 × PF	0.97 × PF	0.98 × PF	RES→7)-β-D-6dtalHep$f$	RES→7)-β-D-6dtalHep (p or f)
	1 (linked by proximal exocyclic atom)	5.54 × PF	6.79 × PF	8.55 × PF	RES→6)-β-D-Fuc$f$	RES→6)-β-D-Fuc (p or f)
	0 (linked by endocyclic atom)	17.82 × PF	15.71 × PF	12.44 × PF	RES→4)-β-D-Fuc$f$	RES→4)-β-D-Fuc (p or f)
absolute configuration (combination with central residue)[c]	? (always)	8.20 × PF	1.99 × PF	4.02 × PF	α-D-Glcp-(1→4)-α-L-Rha$p$	α-Glcp-(1→4)-α-Rha$p$

[a] In this model, a residue is called a substituent if it is linked to the central residue, regardless of whether it is linked as a donor (by its anomeric center) or as an acceptor (by its non-anomeric center). [b] Three weight factor sets are provided regarding the nature of a central residue, corresponding to pyranose, furanose, and common parameter sets. The residue type of a substituent itself is not taken into account. [c] If after generalization there is only one optically active residue with known absolute configuration, its configuration is made unrestricted. For example, α-Glc$p$-(1→4)-α-L-Rha$p$ is turned into α-Glc$p$-(1→4)-α-Rha$p$.

dx.doi.org/10.1021/ci500267u | J. Chem. Inf. Model. 2014, 54, 2594−2611

Table 3. Proximity Factors for Calculation of Weight Factors in Substituent Residues (Remoteness Represents the Number of Bonds from the Atom under Prediction to the Central Residue Atom Forming a Linkage with the Substituent; *R* Shows the Number of Bonds from the Permutation to the Atom under Prediction)

	PF			example	
remoteness	pyranose	furanose	common	atom predicted[a]	structure
0	0.978	0.949	0.915	Hep C1	
1	0.579	0.719	0.621	Hep C2	
2	0.408	0.277	0.329	Hep C3	β-D-6dtalHepp-(1→4)-β-D-GlcpN
3	0.116	0.158	0.117	Hep C6	
≥4		$2 \times 10^{-R-1}$		Hep C7	

[a]In this example, the carbons of β-D-6dtalHepp are simulated, and the given proximity factors relate to the β-D-GlcpN residue.

(1→4)-[α-D-Galp-(1→3)]-β-D-Glcp) were excluded from the selection. Although these chemical shifts are known to depend on glycosidic bond torsions,[29,30] they showed only minor dispersion (standard deviations of 0.58 ppm for Glc C4 and 0.46 ppm for Gal C1), indicating that all of the molecules had the selected interglycosidic linkage in a similar conformation. Analogously, we chose a trisaccharide fragment with branching at neighboring positions (α-Glc C2 and C3, β-Glc C1, and α-Gal C1 in the β-D-Glcp-(1→3)-[α-D-Galp-(1→2)]-α-D-Glcp fragment in seven structures were analyzed). Standard deviations are presented in Table 4. The deviations were low

Table 4. Standard Deviations for the Observed Chemical Shifts of Interglycosidic Atoms in the β-D-Glcp-(1→3)-[α-D-Galp-(1→2)]-α-D-Glcp Fragment

atom	observed RMSD
α-Glc C2	0.24
α-Glc C3	0.53
β-Glc C1	0.19
α-Gal C1	1.29

(≤0.53 ppm) for all atoms except for α-Gal C1 (1.29 ppm), the latter caused by a number of selected molecules having α-Gal substituted at C2, which strongly affected the C1 chemical shift. Substitution effects of this sort are taken into account within the GODESS approach (see Table 2).

2.3. Weight Factor Optimization. To optimize the weight factors, we used an algorithm inspired by biological evolution, a slightly modified algorithm of self-adaptive evolutionary programming[31] based on optimum search direction (OSDEP).[32]

For each type of weight factor, we used a training set of structures with experimental ^{13}C NMR spectra deposited in CSDB. These sets contained some typical structurally distinct features that are found in natural carbohydrates (Table 5). We implemented the function $\Delta_{avg}(x_1, x_2, ..., x_n)$, where $x_1, x_2, ..., x_n$ stands for the weight factors and proximity factors under optimization. Spectra were simulated for all of the structures in the training set, and Δ_{avg} was returned as the mean absolute deviation between the simulated and experimental data:

$$\Delta_{avg} = \frac{1}{N}\sum_{i=1}^{N} |\delta_i^{exp} - \delta_i^{sim}|$$

where δ^{exp} is an experimental chemical shift, δ^{sim} is a simulated chemical shift, and N is the total number of carbons in all of the structures in the training set. To avoid a prediction bias during training, the program was prohibited from using database records containing the spectra of the structure currently being predicted. Spectra were simulated in the *accurate* mode (described in section 3.1), with unrestricted pH, temperature, and solvent parameters (see section 4.1).

Weight factors were optimized for the widespread atom types. Weight factors of distal atoms that hardly affected the simulation and/or were infrequent (e.g., generalizing C9 of neuraminic acid while predicting its C3) were not optimized; they were calculated according to remoteness-dependent formulas (see Tables 1−3). These formulas were designed to make the weights tend to zero asymptotically with increasing remoteness and imply that there are no properties with zero effect.

Here and below, an *individual* is a set of $(x_i, \eta_i)\ \forall\ i \in \{1, 2, ..., n\}$, where x_i is a certain weight factor, η_i is its standard deviation for Gaussian mutations, and n is the number of factors to optimize. According to our algorithm, each x_i has an interval of allowed values designed by summarizing the influence of structural features on chemical shifts estimated using the BIOPSEL incremental effect database[12] (Table 6). Each η_i is also bounded from zero to $x_i^{max} - x_i^{min}$, where x_i^{max} and x_i^{min} are upper and lower bounds for x_i, respectively.

The optimization procedure included the following steps:

Table 5. Characteristics of Training Sets

weight factor set	training structures	total number of residues	total number of carbons
pyranose	pyranose forms of Glc, GlcN, GlcA, Man, Gal, GalN, GalA, IdoA, FucN4N, QuiN, QuiN4N, Qui3N, Rha, Rha4N, RhaN4N, Abe, Xyl, Neu, Kdo, Ko, pseudaminic acid, 8-epilegionaminic acid, 6dTal, L-gro-D-manHep, AltNA, GulN substituted by other pyranoses, furanoses, alditols, amino acids, lipids, phosphoric acid, O- and N-linked acetic acid, methanol, and other residues	105	637
furanose	furanose forms of Gal, GalN, Ara, Fuc, 6daltHep, 6didoHep, Rib, Fru, Xyl, Xul substituted by pyranoses, other furanoses, lipids, alditols, phosphoric acid, O- and N-linked acetic acid, and other residues	55	328
common	Rib-ol, Man-ol, Qui3N-ol, glycerol, glyceric acid, (*S*,*R*)-carboethoxylysine, alanine, 3-hydroxybutyric acid, O- and N-linked acetic acid, acetimidic acid, malic acid, choline, methanol adjacent to various pyranose and furanose sugars, phosphates, and other residues	81	192

dx.doi.org/10.1021/ci500267u | J. Chem. Inf. Model. 2014, 54, 2594−2611

Table 6. Allowed Intervals for Weight Factor Variation (Minimal, Maximal) [Property and Property Remoteness Have the Same Meanings as in Tables 1−3; "−" Means That the Weight Factor Was Not Optimized]

		property remoteness (bonds)			
residue type	property	0	1	2	3
central residue	atom type and stereoconfiguration	(0, 10)	(0, 10)	(0, 10)	(0, 10)
	ring size	(15, 20)	(0, 15)/(10, 15)[a]	−/(0, 10)[a]	−
	acceptor presence	(15, 20)	(10, 15)	(5, 10)	(0, 5)
	donor(s) presence[b]		(0, 20)		
substituent residue	atom type	(5, 20)	(1, 5)	(0.1, 1)	(0, 0.1)
	atom stereoconfiguration	(4, 10)	(0.2, 4)	(0.01, 0.2)	(0, 0.01)
	ring size	(10, 20)	(1, 10)	(0, 1)	−
	absolute configuration (same/different)		(0, 20)		
	proximity factor	(0.9, 1)	(0.5, 0.9)	(0.2, 0.5)	(0, 0.2)

[a]Intervals for pyranose and furanose sets are presented as (pyranose)/(furanose) if they differ. The common set does not have a ring size generalization (see Table 1). The purpose of pyranose/furanose differentiation is the following: if the pyranose ring size is generalized, its exocyclic atoms never become endocyclic; on the contrary, furanose exocyclic atoms may become endocyclic (e.g., C5 in Gal), which would have a stronger effect on the chemical shifts. [b]For both terminal and nonterminal residues.

1. An initial population of $\mu = 5$ individuals was generated: each x_i was generated at random according to its allowed interval (here and below, all uniformly distributed numbers were generated with Mersene twister[33]); initial η_i values were calculated as $(x_i^{max} - x_i^{min})/2$.
2. For each individual, the value of Δ_{avg} was estimated.
3. The best value in a population was remembered as α.
4. Each individual (parent) gave a single offspring. For $i = 1, 2, \cdots, n$:

$$x_i' = x_i + \eta_i\delta + \eta_i N(0,1)(\Delta_{avg} - \alpha)$$

$$\eta_i' = \eta_i \exp\{\tau' N(0,1) + \tau N(0,1)\}$$

where $N(0,1)$ is a normally distributed one-dimensional random number with mean zero and standard deviation 1 (generated using the Box−Muller transformation[34]) that was generated anew each time during processing, δ is a Cauchy random variable with zero mean and variance 1 that was generated anew for each x_i' estimation, Δ_{avg} is the deviation for the current individual, $\tau' = (2n)^{-1/2}$, and $\tau = (2\sqrt{n})^{-1/2}$. If x_i' or η_i' did not conform to the boundaries, it was recalculated.

5. Δ_{avg} was estimated for the offspring individuals.
6. The best individuals (with minimal average deviations) were selected for the next generation; steps 3−6 were repeated until the variation of Δ_{avg} became less than 0.01 ppm per 10 generations.

2.4. Preparations for Fragment Generalization. *2.4.1. Optimization of Absolute Configurations.* Before the generalization process starts, it is desirable on one hand to maximize the number of database records matching the fragment being searched and on the other hand to minimize the effect of generalization on chemical shifts. NMR observables for enantiomers under achiral conditions are identical; thus, inversion of an absolute configuration of every residue in a fragment may broaden the search scope without negative effects on accuracy. This inversion is performed during a pregeneralization procedure executed for every residue. If a residue has no optically active substituents, its absolute configuration is made unrestricted (Figure 3). If a central residue is optically inactive and has no more than one chiral substituent, the substituent absolute configuration is also made unrestricted.

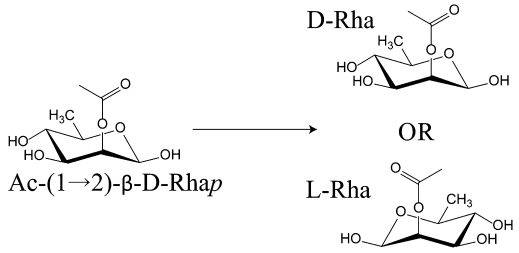


Figure 3. Pregeneralization of β-D-Rha*p*-2OAc. As the L and D isomers have identical NMR spectra, it is desirable to make the absolute configuration unrestricted.

If both the central residue and at least one of the substituent residues are optically active, one of two alternative sets of absolute configurations of the central residue (original and inverted) is selected on the basis of which of them is more populated in the database. If the program selects the inverted configuration, it also inverts all of its chiral substituents (as exemplified in Figure 4). Inversion is carried out between D,L and *R*,*S* pairs.

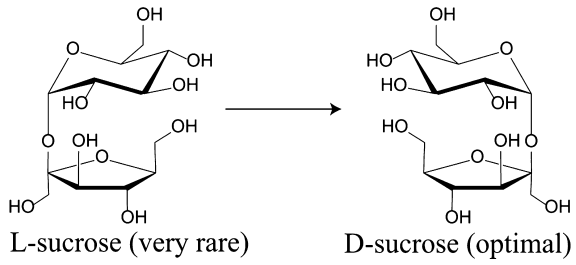


Figure 4. Inversion of the absolute configurations of a fragment (L-sucrose).

2.4.2. Primordial Generalizations. The simulation scheme is able to predict the properties of carbohydrate structures with underdetermined moieties (e.g., D-Glc*p* matching both the α and β forms or α-D-Glc*p*-(1→4)-α-Rha*p* matching both the D and L configurations of rhamnose). In such cases, the weights of these virtual generalizations are precalculated according to the applied weight sets (see section 2.2) and used in subsequent procedures.

2.5. Treatment of Atypical Structures. Because of the structural diversity of carbohydrates and glycoconjugates, it is

dx.doi.org/10.1021/ci500267u | J. Chem. Inf. Model. 2014, 54, 2594−2611

hardly possible to apply an identical generalization algorithm to the whole variety of residue types. Normally, a carbohydrate residue contains from four to nine carbon atoms, and its generalization does not require much time and memory resources. However, biomolecules, including glycoconjugates, often contain residues with more carbon atoms. An increase in the size of the carbon skeleton leads to combinatorial growth of the number of permutations. Therefore, generalization of even a 15-carbon residue may take a few minutes on a PC. This performance becomes crucial when spectrum simulation routines are executed for a multitude of structures, for example in automated NMR-based structure prediction software. On the other hand, larger residues with numerous virtually chemically identical atoms (e.g., such common constituents of bacterial and plant glycans as fatty acids) do not require the full power of the generalization scheme tuned for carbohydrates. We have designed dedicated algorithms and parametrizations for a few such "special cases".

2.5.1. Hydrocarbon Chains. A reduced lipid generalization scheme was developed for residues containing long aliphatic chains (Figure 5). This algorithm is applied for fatty acids, sphingoid structures, and other residues containing long hydrocarbon chains. A residue with an aliphatic chain is separated into three sectors: head (first two atoms), middle, and tail (last three atoms). If a sufficient number of lipid/alkyl-containing fragments is not found in the database, further generalizations of the fragment depend on the sector to which the atom under prediction belongs. For the head atoms, the first generalization step turns a fragment into one containing all of the original atoms from the substituent residue's bond-forming atom to the atom located two bonds away along the hydrocarbon chain. The obtained fragment can undergo further generalizations according to the scheme tuned for carbohydrates. Generalization of the middle and tail atoms retains no information about the substituent residue: the program searches a chain fragment identical to that covering two bonds away from the atom under prediction. During the data search, the algorithm utilizes an atomic pattern distinguishing nine types of carbons on the basis of the attached functional groups and hybridization state.

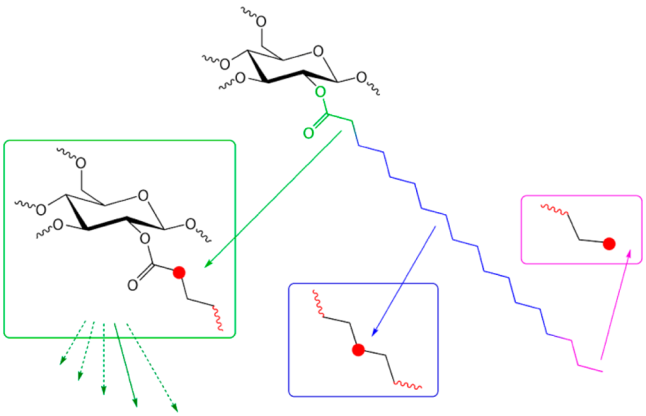


Figure 5. Lipid generalization scheme: prediction of carbon atoms of stearic acid in the fragment Ste-(1→2)-*β*-D−Glc*p*. An aliphatic chain is divided into three sectors by proximity to distinct structural features. Depending on the sector to which an atom under prediction belongs (example atoms are marked as red dots, one for each of the sectors of stearic acid, depicted in green, blue, and magenta), different generalizations are applied. For the head fragment (depicted in green), further generalizations (green arrows) are available.

2.5.2. Other Cases. A problem similar to that discussed above exists for other residue types (e.g., nucleotides). However, because of the relatively small population of such structures in the database, the special treatment of them is postponed. For now, if a large residue (>10 carbons) cannot be treated as a residue with a hydrocarbon chain, only generalizations of its substituents are allowed, while the central residue is never altered.

2.6. Outlier Removal Algorithm. If a database is large enough for accurate predictions, there is always a chance to find erroneous data originating from ambiguous experimental conditions, mistakes in publications, inaccurate annotation processes, or database inconsistencies. To avoid loss of prediction accuracy due to such misfits, a simple data mining algorithm was added to the chemical shift averaging algorithm.

To reject statistical outliers, we used an iterative scheme based on Chauvenet's criterion.[35] For each chemical shift in a data set, a probability value P is calculated according to the normal distribution:

$$P = \frac{N}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

where N is the total number of chemical shifts in the data set, μ is the arithmetic mean, σ is the standard deviation, and x is the value of the chemical shift under testing. If the calculated probability is lower than 0.5, the chemical shift is rejected as an outlier. To detect shielded outliers (values that become noticeable outliers only after the previous outlier has been removed from a data set), the testing procedure is repeated until no more outliers are found. After that, the predicted chemical shift is calculated as the mean value of the data set.

3. EXTRA FEATURES

To expand the simulation abilities, we supplemented the generalization scheme with two extra features: quality mode selection (section 3.1) and trustworthiness evaluation of simulated data (section 3.2). Apart from the statistical approach reported here, the NMR simulator of CSDB uses an incremental approach adopted from the BIOPSEL[12] software. We have developed a "hybrid" NMR spectra simulation tool that benefits from both approaches (section 3.3).

Table 7. Quality Mode Characteristics

mode	maximal number of generalization steps	minimal substituent generalization threshold	minimal number of records required	maximal number of records used	maximal generalization weight
fast	10	0.1	1	20[a]	20
accurate	unlimited	0.01	3	20[a]	20
extreme	unlimited	0.00001	3	unlimited	unlimited

[a]The number increases with the weight of generalizations applied (see the text).

dx.doi.org/10.1021/ci500267u | J. Chem. Inf. Model. 2014, 54, 2594−2611

Table 8. Substituent Generalization Differences between Modes[a]

mode	thresholds used	example: substituent generalization steps	example: step weights
fast	0.1	α-D-RES1*p*-(1→4)-β-D-Gal*p* RES1 atom pattern and stereocode: [oo*][ae*] (any α-D-pyranose with an equatorial hydroxy function at C2: Glc, Gal, Xyl, GalA, Fuc4N, etc.)	0.019 + 0.046 × 2 + 0.057 × 2 + 0.018 × 2 = 0.262
	5	ANY→4)-β-D-Gal*p* (ANY = any substituent)	**0.262** + 0.346 + 0.151 + 2.207 + 0.753 + 0.951 + 2.067 = 6.737
	20	all generalizations for this residue have already been made at the previous stage with a threshold of 5	**6.737**
accurate	0.01	α-D-RES0*p*-(1→4)-β-D-Gal*p* (RES0 = Glc, Gal)[oooodo] [aee?a?]	0.009 × 2 = 0.018
	all of the thresholds for the *fast* mode	same as in the *fast* mode	
extreme	0.00001	same as in the *accurate* mode (this substituent cannot have generalizations with weights less than 0.01)	
	0.0001		
	0.001		
	all of the thresholds for the *accurate* mode		

[a]As an example, generalization pathways of the α-D-Glc*p* substituent in the α-D-Glc*p*-(1→4)-β-D-Gal*p* fragment (predicting C1 of β-D-Gal*p*) are shown, and the weights for each step are provided as a sum of individual descriptor weight factors. Addends derived from previous steps are shown in bold. Atom types and stereoconfigurations are encoded as in Table 2.

3.1. Quality Modes. There are three modes regarding quality versus speed of simulation: *fast*, *accurate*, and *extreme*. Their characteristics and major differences are presented in Table 7. Limitation of structural generalizations prevents deep alterations of a molecule. In the *fast* mode, if the database does not contain any matching structural fragments, after 10 unsuccessful generalization steps the atom is marked as "?" (unpredicted). In the *accurate* mode, the number of generalizations is unlimited, and generalizations are applied until the total weight exceeds 20. Further generalizations (with total weight >20) are often of no use, as their results are hardly credible. However, it is undesirable to leave a chemical shift unpredicted. Therefore, the program searches the database for any fragment containing the central residue in any possible surroundings. If even that fragment is not found, the atom is marked as unpredicted; otherwise, it is marked as predicted with zero trustworthiness. In the *extreme* mode, the allowed generalization number is also unlimited, but in contrast to *accurate* mode, generalizations are executed until all generalization variants have been tried.

The substituent generalization threshold determines which substituent properties are generalized within a certain step. The lower the threshold is, the more gradual the generalization pathway is, which normally leads to more accurate predictions. If the weight of generalization of a certain property is higher than the current threshold, the generalization is not applied. In each step, the current threshold value increases, starting from the mode-specific minimal threshold. The generalization thresholds are given in Table 8. Their values were optimized, and during this process geometric increases (at lower thresholds) and linear increases (at higher thresholds) in different combinations were tried. The results were evaluated according to average accuracy of prediction, execution time, number of steps, and reported trustworthiness for a selection of structures.

In the *fast* mode, if at least one record matches the searched fragment, the program relies on this minimal data set and does not apply further generalizations. In favor of accuracy improvement and minimization of possible negative effects of outlier processing, in the *accurate* and *extreme* modes the minimal number of records is normally set to 3. However, sometimes the algorithm may go deep in generalization in order to find three matching records, but it can find one or two records with minimal generalizations. To prevent accuracy loss, this fact is taken into account: if a data set has fewer than three structures, this intermediate result is stored and further generalization is applied until a larger data set is found. Then the intermediate and final results are compared, and the chemical shift value predicted with better trustworthiness is selected. For trustworthiness estimation, see section 3.2.

The maximal number of structures is limited in the *fast* and *accurate* modes to reduce the calculation time and to simplify further data analysis. Initially, the program uses up to 20 records to predict a chemical shift. However, if deeper generalizations are applied, gathering of more statistics is required to make more precise simulations. Thus, we made the maximal number (N_{max}) dependent on the total weight of generalizations (W_T):

$$N_{max} = 20 + \mathrm{round}(2 \times W_T)$$

where round(x) is an arithmetic rounding function. In the *extreme* mode, the number is unlimited, and for widespread residues, such as acetic acid, the number of used records may run up to hundreds.

We consider the *accurate* mode to have the best quality versus speed ratio. The *fast* mode should be used for approximate evaluations, if the results have to be obtained rapidly. The *extreme* mode may increase the accuracy, but for structures containing rarely occurring residues, the simulation may become very time-consuming (for comparison, see section 5.2).

3.2. Trustworthiness Metrics. The trustworthiness (T) of the simulated data is estimated for every predicted chemical shift. T depends on the extent of generalization (greater total weight leads to lower T), the deviation between chemical shifts found (if the data are contradictory, T decreases), and the number of records found (if the statistics is insufficient, the probability of error increases). According to these facts, we empirically developed the following formula to calculate the trustworthiness:

$$T = 4 - \frac{W_{sum}}{5} - 2 \times \sigma - \frac{1}{2N}$$

dx.doi.org/10.1021/ci500267u | J. Chem. Inf. Model. 2014, 54, 2594−2611

where W_{sum} is the total weight of the generalizations applied, σ is the standard deviation between values in the data set, and N is the number of chemical shifts in the data set. The number is rounded to the first decimal digit. The σ parameter is taken into account only when it exceeds 0.1 so as not to make the T value understated (all values in the data set are identical only in very rare cases). If the calculated value becomes negative, it is set to zero, so the trustworthiness range is from 0 to 4. This formula was obtained by varying different functional dependencies of T on W_{sum}, σ, and N (exponential, square, inverse, linear) with subsequent manual testing on a set of structures with known experimental ^{13}C NMR spectra; the formula is subject to future improvements (see section 6). Table 9 explains the characteristic trustworthiness values in terms of accuracy. These approximate values were obtained according to the experimental trend line (see section 5.3).

Table 9. Approximate Deviation/Trustworthiness Scale

trustworthiness	expected deviation (ppm)
4	0
3	0.5
2	1
1	1.5
0	≥2

3.3. Integration with the Incremental Approach. The dedicated simulation Web service uses two different approaches, empirical (developed earlier[12]) and statistical, so the output includes two NMR spectra. It is not always handy to compare the obtained data: a user has to switch from one assignment table to another and to compare chemical shifts, trustworthiness values, and so on. To minimize this inconvenience, we have implemented a feature that generates a hybrid spectrum that contains the most accurate data combined from the two approaches and can be used as a summary.

To calculate a hybrid chemical shift and to estimate its trustworthiness, the hybridization routine utilizes simulated data from the two approaches, the deviation between them, and trustworthiness values. If the two approaches produce close chemical shifts, their credibility is high, since the result obtained by one approach is proved by the other. To reflect this, if the difference between the simulated chemical shifts is less than 0.2 ppm and at least one of them has a good trustworthiness value (≥2), the hybrid chemical shift is calculated as the mean of the two values and its trustworthiness is computed as $T_H = 4 - |\delta_{stat} - \delta_{emp}|$, where δ_{stat} and δ_{emp} are the simulated chemical shifts obtained using the statistical and empirical approaches, respectively.

If the values are contradictory, the hybrid chemical shift has to be closer to that with the higher trustworthiness. According to this principle, for chemical shifts having an appreciable difference (≥0.2 ppm), the hybrid value δ is given by a linear combination of δ_{stat} and δ_{emp}:

$$\delta = K_{stat}\delta_{stat} + K_{emp}\delta_{emp}$$

where the coefficients are calculated as

$$K_{stat} = \frac{T_{stat}^{\ 3}}{T_{stat}^{\ 3} + T_{emp}^{\ 3}}, \qquad K_{emp} = 1 - K_{stat}$$

in which T_{stat} and T_{emp} are trustworthiness values reported by the statistical and empirical approaches, respectively. If both T_{stat} and T_{emp} are zero, the formula is not applied, and both coefficients are set to 0.5. The trustworthiness of the hybrid shift is calculated analogously, with the only difference that the deviation between the shifts and the confirmation of one by the other are taken into account:

$$T_H = (K_{stat}T_{stat} + K_{emp}T_{emp} - |\delta_{stat} - \delta_{emp}| \times K_{min}) \times (0.9 + 0.2 \times K_{min})$$

where K_{min} is the minimal value of the statistical and empirical coefficients. The deviation part of the equation, $|\delta_{stat} - \delta_{emp}| \times K_{min}$, is needed in order to consider the admixture chemical shift added to the "main" one (with the higher coefficient). This is especially important when $K_{stat} = K_{emp}$. For example, if both approaches report the maximal trustworthiness of 4.0 but the difference between chemical shifts is 2 ppm, there is no way to

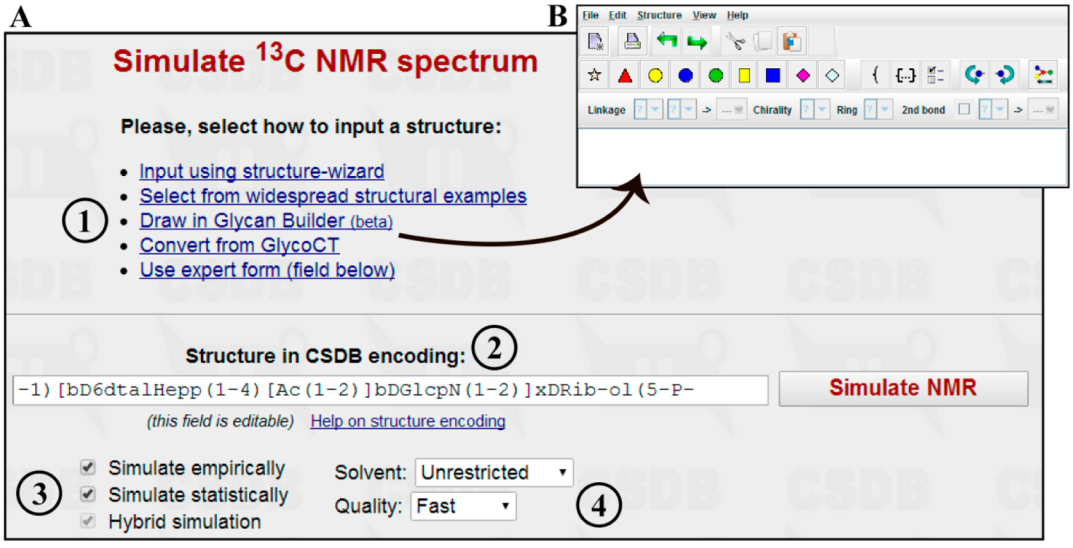


Figure 6. (A) Web interface of the ^{13}C NMR spectra simulator. The structure presented in Figure 1 is input in the CSDB linear format (②). (B) A Glycan Builder[36] window.

dx.doi.org/10.1021/ci500267u | J. Chem. Inf. Model. 2014, 54, 2594−2611

calculate the hybrid precisely because at least one of the approaches is mistaken. Therefore, the hybrid shift is the average, but its trustworthiness T_H is 2.0 instead of 4.0 because of the contradictory data. The outer coefficient, $(0.9 + 0.2 \times K_{min})$, is required in order to consider how the "main" chemical shift is confirmed by the admixture. For example, if one of the values has zero trustworthiness (or is unpredicted), the trustworthiness becomes 90% of the original value, as it cannot be proved by the other approach.

The formulas for the calculation of hybrid chemical shifts were designed in a similar way as described in section 3.2: for K_{stat}, K_{emp}, and T calculations, different functional dependencies on the trustworthiness (square, cubic, exponential, inverse, linear) were tried and then manually tested on a set of structures with known experimental ^{13}C NMR spectra. The formulas are subject to further improvements (see section 6).

4. APPLICATION

The approach is realized as a part of the CSDB module called Glycan-Optimized Dual Empirical Spectrum Simulation (GODESS). The statistical simulation tool is present in the Web interface of both CSDB databases (Bacterial and Plant & Fungal) and selects a database to query accordingly. In the future, we plan to join the two databases to increase accuracy and to reduce the prediction time as a result of less demand for generalizations (for details, see section 6). The feature is available from the dedicated Web pages http://csdb.glycoscience.ru/bacterial/core/nmrsim.html (predictions based on BCSDB) and http://csdb.glycoscience.ru/plant_fungal/core/nmrsim.html (predictions based on PFCSDB).

4.1. Structure and Parameter Input. To input a structure, one of the existing CSDB features is used (Figure 6) (①). They include building in a wizard, drawing in Glycan Builder,[36] conversion from the GlycoCT[37] code, or typing in the CSDB linear format[21] (②). Simulation options (Figure 6) include selection of the approaches (③). When the statistical approach is selected, solvent and quality options are available (④). Unrestricted solvent (default) means no solvent limitations, while selection of a solvent (any water, heavy water explicitly, chloroform, methanol, DMSO, or pyridine) restricts the used reference spectra to those recorded in a particular solvent. For BCSDB, almost all of the spectra were recorded in D_2O (Figure 7A), so it is advisable to leave the solvent unrestricted. For PFCSDB, the predictions are expected to be most accurate for pyridine-d_5 solutions (Figure 7B). In contrast to many other simulation approaches, GODESS considers the solvent effects via a selection of reference data rather than via parametrization. The quality option allows a selection of one of three quality versus speed modes (see section 3.1). The default value is *fast*, implying that the result is generated in less than 15 s.

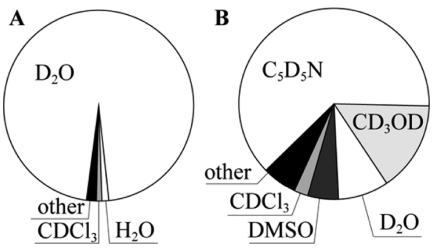


Figure 7. Distributions of ^{13}C NMR spectra recorded in different solvents, as deposited in (A) BCSDB and (B) PFCSDB.

4.2. Output Format. Prediction results are shown as ^{13}C NMR assignment tables and schematic NMR spectra (Figure 8). The assignment table displays the chemical shift of each carbon atom, trustworthiness metrics, and a link to the list of references to the data used for each signal. Every row represents a particular residue in a molecule. Columns are the following:

Linkage: the linkage path to the residue from the oligomer reducing end or the polymer repeating unit rightmost residue.

Residue: residue name, ring size, and configurations.

Trust: residue prediction trustworthiness estimation (averaged over all atoms in the residue).

Other columns: data for every carbon in the residue, including its chemical shift, its trustworthiness, and the number of records used for the chemical shift prediction. A click on this number shows the list of records (Figure 8A, at the bottom), so the user can check the origins of the reference data and track them to the original publications.

The results can be recalculated by pressing the "Recalc" button (Figure 8A). This can be useful if no satisfactory results were obtained (e.g., if the reported trustworthiness values are low, there are unpredicted signals, a solvent was chosen incorrectly, etc.). In addition to the solvent and quality options, sample temperature and pH ranges can be specified. As the temperature dependence of ^{13}C NMR chemical shifts is weak and there are not so many pH data in the database, these features are omitted from the simulation start page for clarity.

The output format for hybrid spectra is similar to that described for the statistical approach. The assignment table indicates the coincidence between approaches ($\Delta = |\delta_{stat} - \delta_{emp}|$) and the schematic spectrum represents data obtained by each method (Figure 9). Small peaks depicted in red and blue stand for empirical and statistical simulations, respectively.

5. VERIFICATION

5.1. Statistical Measurement of Accuracy and Performance. To derive cumulative metrics for the accuracy and performance of the statistical, empirical, and hybrid approaches, we used a test set of structures containing residues of different types (pyranoses, furanoses, alditols, amino acids, lipids, phosphates, and others) with experimentally measured ^{13}C NMR data that were deposited in CSDB and other databases. This experimental data set contained 1322 carbon chemical shifts.

NMR data were simulated for each structure. For the hybrid and statistical approaches, the program was prohibited from using database records containing the spectra of the structure currently being predicted. Simulations were carried out in the *accurate* mode with unrestricted solvent, pH, and temperature parameters.

Absolute deviations between the experimental and simulated chemical shifts were averaged to give the mean deviation $\Delta_{avg} = (1/N)\sum_{i=1}^{N}|\Delta_i|$, where N is the number of chemical shifts in the data set. An average prediction time per atom was measured using the Web server that hosts the CSDB. The results are presented in Table 10. The hybrid approach was shown to be the most precise as benefiting from the two approaches. The statistical approach showed better accuracy than the empirical one, although it took ca. 4000 times longer.

5.2. Representative Simulation Examples. To test the accuracy and performance indicators, we used eight structures

dx.doi.org/10.1021/ci500267u | J. Chem. Inf. Model. 2014, 54, 2594−2611

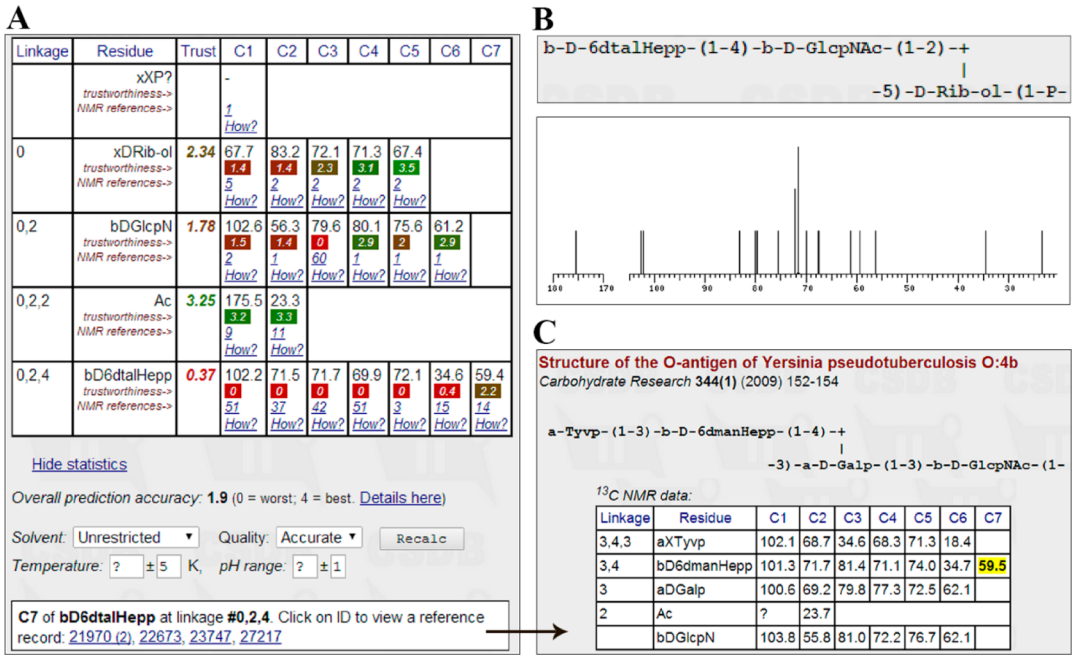


Figure 8. Output of the statistical prediction includes (A) the NMR assignment table and (B, bottom) the schematic spectrum. Clicking on the source ID (A, bottom) opens the corresponding source record (C); the chemical shift used for simulation is highlighted. The predicted structure is shown in (B) at the top. NMR data were simulated in the *accurate* mode with unrestricted solvent, pH, and temperature for the structure represented in Figure 1.

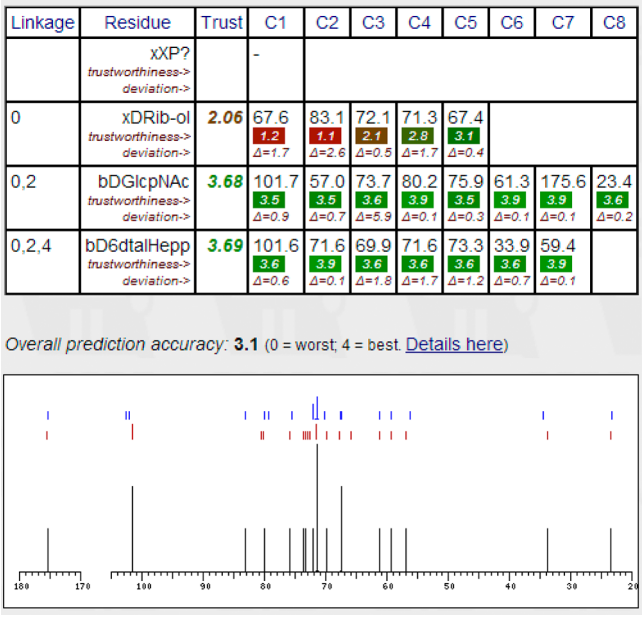


Figure 9. Output of a hybrid spectrum. NMR data were simulated for the structure presented in Figure 1 in the same mode as in Figure 8.

Table 10. Accuracy Comparison of GODESS Approaches

approach	Δ_{avg} (ppm)	prediction time per atom (s)[a]
statistical	0.81	1.1880
empirical	1.02	0.0003
hybrid	0.79	1.1883

[a]Without input/output operations.

covering various structural features of natural carbohydrates and derivatives (see Table 11).

For the statistical approach, data were simulated using BCSDB with unrestricted solvent, temperature, and pH parameters. If structures under prediction were deposited in the database, they were virtually removed before processing to avoid a prediction bias. The approach comparison is summarized in Table 12. For all of the structures except sucrose (**2**), the statistical approach was the most accurate in the chemical shift prediction and strongly outperformed the quantum-mechanical approaches reported as best for carbohydrates[7] both in accuracy and speed. The low root-mean-square deviation (RMSD) for sucrose simulated by the ACD/Labs software was considered a coincidence related to α-D-Glc*p* and β-D-Fru*f*. This assumption was confirmed by changing α-D-Glc*p* to β-D-Gal*p*, which caused only minor changes in the chemical shifts predicted by ACD/Labs even though the differences in C1, C4, and C5 in the experimental spectra were significant. The low sensitivity of ACD/NMR to structural changes is discussed in section 5.3. The mapping of signals predicted by GODESS to the experimental ^{13}C NMR spectrum of sucrose is shown in Figure 10. The largest deviations are observed for atoms forming linkages between residues, which is due to the absence of ketofuranose-(2→1)-aldopyranose fragments in the database. A systematic overestimation of chemical shifts by ca. 1 ppm was observed.

Structure **8**, as well as similar cases, could be predicted only by the statistical approach, as a polymer (polymers are supported only by GODESS, BIOPSEL, and CASPER) containing uncommon residues. Although CASPER was reported to be precise for carbohydrate NMR simulation,[13] none of the structures except **1** and **4** could be predicted using this service because of limitations on allowed residues.

The *extreme* mode often provided more precise simulations than the *accurate* mode; however, accuracy improvements are not guaranteed. For large structures with rarely occurring residues or residue surroundings, such as ones containing high

dx.doi.org/10.1021/ci500267u | J. Chem. Inf. Model. 2014, 54, 2594−2611

Table 11. Structures Used for Accuracy and Performance Testing (References Correspond to the Published Experimental NMR data in D_2O)

structure	peculiarities
(1)[38] α-D-Glc*p*	Simple and common monosaccharide
(2)[7] β-D-Fru*f*-(2→1)-α-D-Glc*p* (sucrose)	Monosaccharaides both in pyranose (populated and well parameterized) and furanose (poorly parameterized) forms
(3)[39] α-D-Glc*p*NAc-(1→P→P→5)-β-D-Rib*f*-(1→N)-uracil (UDP-α-D-Glc*p*NAc)	Rarely occurring and poorly parameterized residues within uridine diphosphate
(4)[40] L-Ala-(2→1)-L-Glu-(2→6)-α-D-Gal*p*NAcA-(1→4)-D-GalNAc-ol	Alditol and amino acid residues
(5)[41] α-D-Rha*p* 1 ↓ 3 →4)-β-D-Xyl*p*-(1→4)-α-L-Fuc*p*-(1→	Polymer repeating unit, bisubstitution at neighboring positions, deoxy sugars, pentose
(6)[42] β-D-Gal*p* 1 ↓ 2 →3)-β-D-Gal*p*-(1→1)-D-Gro-(3→P→	Polymer repeating unit, phosphate group, glycerol, bisubstitution at neighboring positions
(7)[43] →7)-a-Pse*p*5Ac-(2→3)-(*S*)-3HOBut-(1→, where 3HOBut is 3-hydroxybutyric acid and Pse is pseudaminic acid	Polymer repeating unit, aliphatic residue, higher sugar (neuraminic acid) substituted at flexible tail
(8)[44] →8)-a-8eLeg*p*5Ac7Ac-(2→3)-a-L-Fuc*p*NAm-(1→3)-a-D-Qui*p*NAc-(1→, where 8eLeg5Ac7Ac is 5,7-diacetamido-3,5,7,9-tetradeoxy-L-*glycero*-D-*galacto*-nonulosonic (8-epilegionaminic) acid and Am is acetimidic acid residue	Polymer repeating unit, rarely occurring higher sugar (8eLeg5Ac7Ac), uncommon non-carbohydrate constituent (acetimidic acid)

sugars (ketodeoxynononic acid, 8-epilegionaminic acid), predictions in the *extreme* mode can take up to 2−3 min.

The hybrid approach was most precise on average (see section 5.1). Nevertheless, its reported accuracy often lay between those reported for the empirical and statistical approaches. This can account for the low adequacy of empirical trustworthiness metrics (see section 5.3), disallowing correct spectra hybridization (for algorithms, see section 3.3).

5.3. Sensitivity to Structural Permutations. The sensitivity of GODESS and other empirical methods (ACD/NMR, Modgraph, CASPER) to structural permutations was tested on a pool of typical disaccharides with the same monomeric composition but different combinations of anomeric configurations and linkage position, namely, D-Glc*p*-(1→2/3/4/6)-D-Gal*p*. We analyzed chemical shifts of interglycosidic carbons, as those are most affected by permutations of the linkage (Table S1 in the Supporting Information). All three approaches of GODESS, as well as CASPER, could reproduce differences in the predicted chemical shifts, as expected from known glycosylation effects in these structures.[48] ACD/NMR and Modgraph simulations gave identical results for any of four combinations of anomeric configurations of both residues, rendering these two approaches hardly applicable for carbohydrate NMR simulations.

5.4. Trustworthiness Evaluation. To test the trustworthiness metrics against the differences between the experimental and simulated chemical shifts, we selected the same structure set, mathematical assumptions, and prediction modes as described in section 5.1. For each structure, we simulated the ^{13}C NMR spectrum and stored the trustworthiness value (T) and observed deviation (Δ) of each chemical shift. Here and below, a (T, Δ) pair is called a point. All points were grouped by the trustworthiness value (in each group, T = constant). For each group, a mean average (Δ_{avg}) was calculated. The obtained data are plotted in Figure 11.

The adequacy of a trustworthiness valuation (A) can be estimated as $A = -\tan \alpha$, where α is the slope angle of the trend line. If the trend line is more slanted, the program is likely to produce low T values for high Δ and vice versa; if $\tan \alpha \rightarrow 0$, the match between Δ and T is poorer. The A value is the best for the statistical approach and worst for the empirical one. This is a result of the fact that the empirical T is estimated for the whole residue while the statistical approach estimates T separately for each carbon.

6. DRAWBACKS AND FUTURE PROSPECTS

At present, it is undoubtedly possible to improve both the prediction accuracy and performance by further tuning of the algorithm and parameter sets as well as to add a number of useful features. Our list of developmental clauses is the following:

Joining the Bacterial and Plant & Fungal databases into a single database has not been done yet because of our funding policy. The combined database will lead not only to better accuracy and performance but also to better treatment of structures with rarely occurring residues. Structures in PFCSDB

dx.doi.org/10.1021/ci500267u | J. Chem. Inf. Model. 2014, 54, 2594−2611

Table 12. Root-Mean-Square Deviations (Simulated vs Experimental) and Calculation Times (in Parentheses) for Different Approaches Applied to Test Structures 1–8 (see Table 11) [For Statistical and Hybrid Approaches, Results in the *Accurate* Mode Are in Regular Font, Whereas Results in the *Extreme* Mode Are in Italics; X Means That the Simulation Is Not Supported by the Software Used]

method	RMSD in ppm (prediction time)							
	1	2	3[a]	4	5	6	7	8
statistical (accurate, *extreme*)	0.52 (19 s), *0.50 (19 s)*	2.13 (19 s), *1.11 (19 s)*	0.37 (2.4 s), *0.37 (2.4 s)*	2.05 (27 s), *1.06 (48 s)*	0.23 (40 s),[b] *0.23 (48 s)*[b]	0.94 (9.3 s), *0.94 (9.3 s)*	1.93 (16 s), *1.91 (25 s)*	0.87 (56 s), *0.92 (116 s)*
empirical (BIOPSEL)[12]	0.96 (0.1 s)	2.65 (0.2 s)	X	X	1.62 (0.1 s)	0.98 (0.3 s)	2.83 (0.2 s)	X
hybrid (accurate, *extreme*)	0.80 (19.1 s), *0.80 (19.1 s)*	2.62 (19.2 s), *2.62 (19.2 s)*	X	X	0.36 (40 s),[b] *0.36 (48 s)*[b]	0.67 (9.6 s), *0.67 (9.6 s)*	2.32 (16.2 s), *2.08 (25.2 s)*	X
empirical (CASPER)[13]	0.64 (<1 s)	X	X	X	1.32 (<2 s)	X	X	X
HOSE + neural net (ACD/Labs 10)[45]	2.99 (<5 s)	0.65 (<5 s)	5.51 (<5 s)	2.76 (<5 s)	X	X	X	X
Modgraph (ChemBioDraw 13)	2.59 (<2 s)	6.64 (<2 s)	6.77 (<2 s)	3.39 (<2 s)	X	X	X	X
GIAO B3LYP/6-311G++(2d,2p) + COSMO (Gaussian 09)[46]	6.56 (11.8 h[c])	3.9 (67.8 h[c])[7]	8.14 (206 h[c])	X	X	X	X	X
GIAO PBE/TZ2p (PRIRODA 06)[47]	n.d.[d]	5.4 (29 min[c])[7]	n.d.[d]	n.d.[d]	X	X	X	X

[a]Chemical shifts of acetic acid residues were not taken into account because the source publication did not contain those data. [b]The source NMR data[41] are not present in BCSDB; however, the same structure is deposited as BCSDB record ID 27122. During the simulation, the use of data from this record was prohibited. With use of this closely related record enabled, the RMSDs were 0.16 (statistical) and 0.28 (hybrid), and the calculation time decreased to 3 s. [c]Calculations were carried out on a PC (Intel Core 2 X4, 3.0 GHz). [d]PBE/TZ2p simulations for structures **1**, **3**, and **4** were omitted, as this approach was shown to be less accurate than B3LYP-level computations on saccharides.[7]

contain many specific residues missing from BCSDB and vice versa. Nevertheless, the ^{13}C spectroscopic coverage of PFCSDB is ca. 1.5 times smaller than that of BCSDB, so it normally provides less precise predictions. Joining the databases will eliminate this problem.

The weight optimization scheme has potential for further improvement. Particularly, the unconsidered factors are the following: (1) The weight of substituent residue generalization depends not only on the nature of the central residue but also on the nature of the substituent itself. (2) Bonds between atoms in carbohydrates are not equal energetically and stereochemically. Thus, the modified number of bonds between an atom under prediction and atom(s) under generalization may take noninteger values and contribute to the generalization effect. (3) For rarely used parameters (e.g., generalization of a pyranose ring size while predicting a carbon two bonds away from the cycle), the number of matching structures in the training set might be too small. Because of the rarity of such features, it was not possible to find enough experimental data for training.

Other areas of improvement include the following: (1) The development of thorough algorithms for trustworthiness estimation is needed for better spectrum simulation based on credibility metrics. (2) Intuitive formulas for spectral hybridization have not been proved to best match their purpose and are subject to optimization by iterative comparison of simulation results to the experimental data, as has been done for purely statistical prediction. (3) Expansion of special generalization features to special structures other than hydrocarbon chains may improve the accuracy of glycoconjugate and glycoside NMR simulation. (4) Adding temperature, acidity, and other nonstructural parameters to the generalization scheme may help to determine how the solvent or pH affect certain chemical shifts and to develop corresponding mathematical models allowing predictions in solvents or at temperatures unpopulated in the database.

The structure generalization scheme can be used with any atomic parameter, and CSDB contains thousands of ^{1}H NMR spectra. However, proton chemical shifts occupy a much narrower range and are affected by acidity, temperature, solvent, and other factors, limiting usage of proton NMR spectroscopy in automated structural studies of carbohydrates. The adaptation of the reported approach to ^{1}H NMR simulations is a question for the future.

7. CONCLUSIONS

A glyco-tuned approach to simulate any database-represented atomic observables has been developed. Although the approach was optimized for the prediction of ^{13}C NMR data, it is not limited to carbon chemical shifts. The developed algorithm can serve as a proof of concept, which can be adapted to the prediction of other database-populated observables by weight factor optimization only, without algorithmic changes. The weight optimization algorithm used (based on evolutionary programming) is common for any kind of observable, so the required scheme tuning is expected to be simple.

The approach was realized within a service freely available on the Internet. It has shown higher accuracy on a test pool of carbohydrate structures in comparison with modern ^{13}C NMR prediction instruments such as ACD/NMR, Modgraph NMRPredict, BIOPSEL, and others, including quantum-mechanical calculations at high theory levels with large basis sets. It has appreciable speed superiority in comparison with

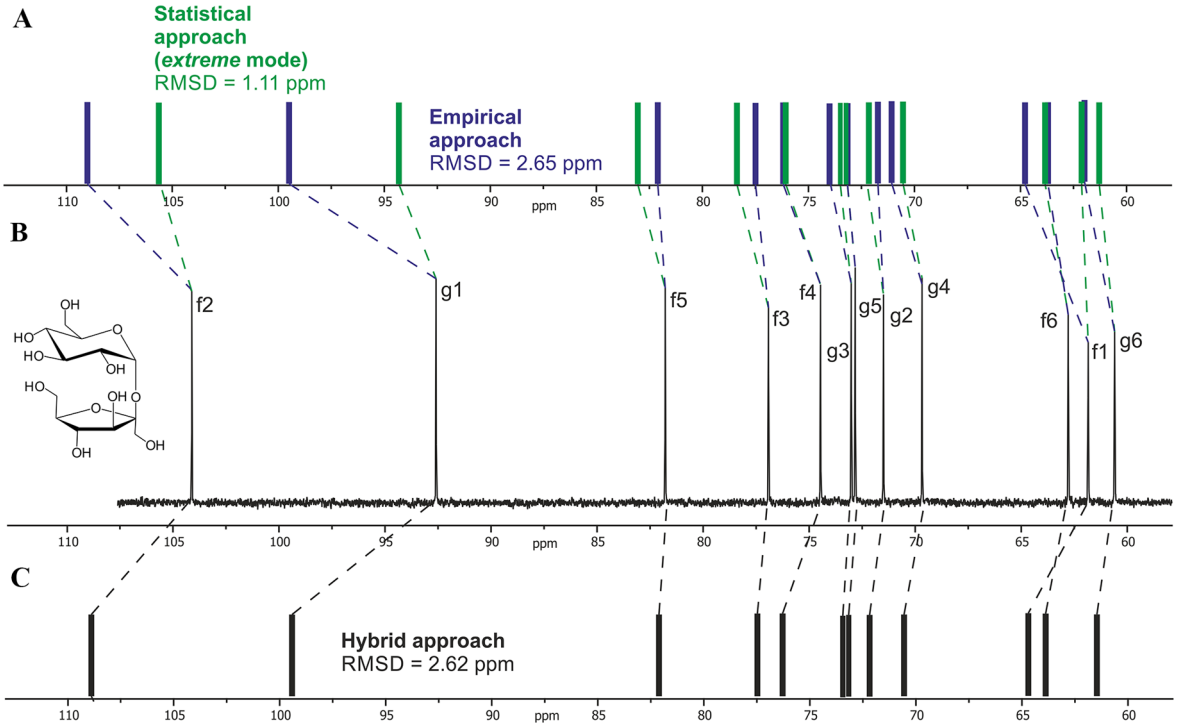


Figure 10. ^{13}C NMR spectra of sucrose simulated by (A) the statistical (green) and empirical (blue) approaches and (C) the hybrid approach, compared with (B) the experimental[7] spectrum recorded at 298 K in D_2O. The assignment of peaks is denoted by "f" (fructose residue) or "g" (glucose residue) and the carbon number. Dashed lines show correlations between the predicted and experimental data. The root-mean-square deviation is given for each approach. Linear correlation values were 0.999 for the statistical approach, 0.993 for the empirical approach, and 0.994 for the hybrid approach.

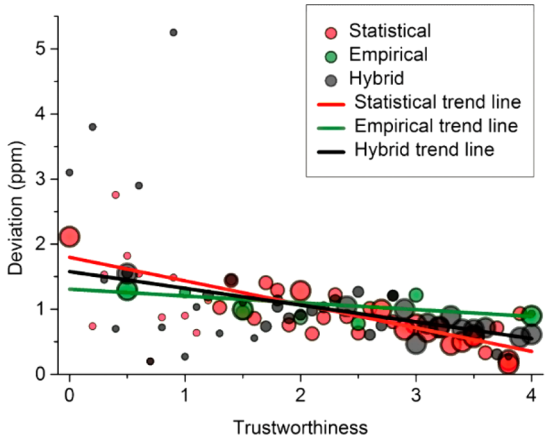


Figure 11. Correspondence between observed deviation (Δ) and trustworthiness. Circle centers are at (T, Δ_{avg}), where Δ_{avg} is the mean Δ for a certain trustworthiness (see the text). Circle sizes reflect the statistical credibility of the calculated mean: big circles contain >30 points, medium circles 10–30 points, and small circles <10 points. Trend lines were built by linear regression of all the points.

quantum-mechanical methods. We have developed a way to estimate the trustworthiness of simulations, indicating the level of precision for every chemical shift.

Using a database makes it possible for the user to find the reference data used in a simulation. In contrast, most of existing instruments cannot allow the origin of the simulated data to be determined, and users often deal with a "black box" and cannot ascertain whether the data used for predictions were credible.

Unlike other approaches, using the generalization scheme makes it possible to simulate atomic observables for under-determined carbohydrate structures, which (with the help of trustworthiness valuation) can be useful for studying the influence of structure descriptors on chemical shift values.

The designed statistical approach was combined with the incremental scheme, which was reported to excel quantum-chemical and unspecific empirical approaches.[7] A combined tool, GODESS, and its hybrid spectrum builder benefit from two simulation schemes that complement and support each other. Unlike the incremental approach,[12] the statistical one does not need a dedicated chemical shift and substitution effect database but uses a regularly updated general purpose database (i.e. the CSDB) and can process structures with rarely occurring and non-carbohydrate constituents.

8. MATERIALS AND METHODS

The chemical shift prediction routine utilizes the Carbohydrate Structure Database (CSDB) powered by the MySQL 5.5.29 relational database engine and PHP 5.3.9 scripts. The Web interface uses DHTML 4, CSS 2, and JavaScript 1.2 and was tested in Google Chrome 31, Mozilla Firefox 25, and Internet Explorer 10. Statistical processing of data was performed in OriginPro 9.0.0 and Microsoft Excel 2013. All services are freely available under the "NMR simulation" menu item at the CSDB Web site (http://csdb.glycoscience.ru).

■ ASSOCIATED CONTENT

Supporting Information

^{13}C chemical shifts of interglycosidic atoms in various glucopyranosyl galactopyranoses predicted by different methods. This material is available free of charge via the Internet at http://pubs.acs.org.

dx.doi.org/10.1021/ci500267u | J. Chem. Inf. Model. 2014, 54, 2594–2611

AUTHOR INFORMATION

Corresponding Authors

*E-mail: netbox@toukach.ru (P.V.T.).
*E-mail: danamad@gmail.com (K.S.E.).

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This study was performed in the framework of development of the Carbohydrate Structure Database funded by the Russian Foundation for Basic Research (Grants 05-07-90099 and 12-04-00324). The authors thank Michael Dolgov for the NMR spectra visualization script. The ^{13}C NMR data obtained using the reported tool can be used free of charge with reference to this paper.

ABBREVIATIONS

Residue Names

3HOBut	3-hydroxybutanoic acid
6dTal	6-deoxytalose
6didoHep	6-deoxyidoheptose
6dtalHep	6-deoxytaloheptose
8eLeg	5,7-diamino-3,5,7,9-tetradeoxy-L-glycero-D-galactonon-2-ulosonic (8-epilegionaminic) acid
Abe	3,6-dideoxy-D-xylohexose (abequose)
Ac	acetic acid
Ala	alanine
All	allose
Alt	altrose
AltNA	2-amino-2-deoxyaltruronic acid
Am	acetimidic acid
Ara	arabinose
Fru	fructose
Fuc	6-deoxygalactose (fucose)
Fuc4N	4-amino-4,6-dideoxygalactose
FucN	2-amino-2,6-dideoxygalactose
FucN4N	2,4-diamino-2,4,6-trideoxygalactose
Gal	galactose
Gal4N	4-aminogalactose
GalA	galacturonic acid
Glc	glucose
Glc1N	1-amino-1-deoxyglucose
Glc3N	3-amino-3-deoxyglucose
Glc4N	4-amino-4-deoxyglucose
GlcA	glucuronic acid
GlcN	2-amino-2-deoxyglucose
Glu	glutamine
Gro	glycerol
GulN	2-amino-2-deoxygulose
Hep	heptose
Ido	idose
IdoA	iduronic acid
IdoN	2-amino-2-deoxyidose
Kdo	3-deoxy-D-mannooct-2-ulosonic (ketodeoxy-octonic) acid
Ko	D-glycero-D-talooct-2-ulosonic (ketooctonic) acid
L-gro-D-manHep	L-glycero-D-mannoheptose
Leg	5,7-diamino-3,5,7,9-tetradeoxy-D-glycero-D-galactonon-2-ulosonic (legionaminic) acid
Lyx	lyxose
Man	mannose
Man-ol	mannitol
Neu	5-amino-3,5-dideoxy-D-glycero-D-galactonon-2-ulosonic (neuraminic) acid
P	phosphoric acid
Pse	5,7-diamino-3,5,7,9-tetradeoxy-L-glycero-L-mannonon-2-ulosonic (pseudaminic) acid
Qui3N-ol	3-amino-3,6-dideoxyglucitol
QuiN	2-amino-2,6-dideoxyglucose
QuiN4N	2,4-diamino-2,4,6-trideoxyglucose (bacillosamine)
Rha	6-deoxymannose (rhamnose)
Rha4N	4-amino-4,6-dideoxymannose
RhaN4N	2,4-diamino-2,4,6-trideoxymannose
Rib	ribose
Rib-ol	ribitol
Ste	stearic acid
Tal	talose
TalA	taluronic acid
UDP	uridine diphosphate
Xul	threopent-2-ulose (xylulose)
Xyl	xylose

Other Abbreviations

B3LYP	Becke three-parameter/Lee–Yang–Parr
BCSDB	Bacterial Carbohydrate Structure Database
BIOPSEL	biopolymer structure elucidation
CASPER	computerized approach to structure determination of polysaccharides
COSMO	conductor-like screening model
CSDB	Carbohydrate Structure Database
CSS	cascading style sheets
DHTML	dynamic hypertext markup language
GIAO	gauge-including atomic orbital
GlycoCT	glyco connection table
GODESS	glycan-optimized dual empirical spectra simulation
HOSE	hierarchical organization of spherical environments
NMR	nuclear magnetic resonance
NOE	nuclear Overhauser effect
OSCAR	oligosaccharide subtree constraint algorithm
OSDEP	optimum search direction evolutionary programming
PBE	Perdew–Burke–Ernzerhof
PF	proximity factor
PFCSDB	Plant & Fungal Carbohydrate Structure Database
PHP	PHP: a hypertext preprocessor
RMSD	root-mean-square deviation

REFERENCES

(1) Gaidzik, N.; Westerlind, U.; Kunz, H. The development of synthetic antitumour vaccines from mucin glycopeptide antigens. *Chem. Soc. Rev.* **2013**, *42* (10), 4421–4442.

(2) Astronomo, R. D.; Burton, D. R. Carbohydrate vaccines: Developing sweet solutions to sticky situations? *Nat. Rev. Drug Discovery* **2010**, *9* (4), 308–324.

(3) Boltje, T. J.; Buskas, T.; Boons, G. J. Opportunities and challenges in synthetic oligosaccharide and glycoconjugate research. *Nat. Chem.* **2009**, *1* (8), 611–622.

(4) Johnson, M. A.; Bundle, D. R. Designing a new antifungal glycoconjugate vaccine. *Chem. Soc. Rev.* **2013**, *42* (10), 4327–4444.

(5) Alper, J. Searching for medicine's sweet spot. *Science* **2001**, *291* (5512), 2338–2343.

(6) Schmidt, L. D.; Dauenhauer, P. J. Chemical engineering: Hybrid routes to biofuels. *Nature* **2007**, *447* (7147), 914–915.

(7) Toukach, F. V.; Ananikov, V. P. Recent advances in computational predictions of NMR parameters for the structure elucidation of carbohydrates: Methods and limitations. *Chem. Soc. Rev.* **2013**, *42* (21), 8376−8415.
(8) Hricovini, M. Structural aspects of carbohydrates and the relation with their biological properties. *Curr. Med. Chem.* **2004**, *11* (19), 2565−2583.
(9) Duus, J.; Gotfredsen, C. H.; Bock, K. Carbohydrate structural determination by NMR spectroscopy: Modern methods and limitations. *Chem. Rev.* **2000**, *100* (12), 4589−4614.
(10) Lutteke, T. The use of glycoinformatics in glycochemistry. *Beilstein J. Org. Chem.* **2012**, *8*, 915−929.
(11) Sasaki, R. R.; Lefebvre, B. A. On the importance of structure stereochemical markers in ^{13}C NMR predictions. Presented at the Small Molecule NMR Conference (SMASH 2006), Burlington, VT, Sept 10−13, 2006.
(12) Toukach, F. V.; Shashkov, A. S. Computer-assisted structural analysis of regular glycopolymers on the basis of ^{13}C NMR data. *Carbohyd. Res.* **2001**, *335* (2), 101−114.
(13) Jansson, P. E.; Stenutz, R.; Widmalm, G. Sequence determination of oligosaccharides and regular polysaccharides using NMR spectroscopy and a novel Web-based version of the computer program CASPER. *Carbohyd. Res.* **2006**, *341* (8), 1003−1010.
(14) Lundborg, M.; Widmalm, G. Structural analysis of glycans by NMR chemical shift prediction. *Anal. Chem.* **2011**, *83* (5), 1514−1517.
(15) Brandolini, A. J. NMRPredict. *J. Am. Chem. Soc.* **2006**, *128* (40), 13313.
(16) Bremser, W. Hose—A novel substructure code. *Anal. Chim. Acta* **1978**, *103* (4), 355−365.
(17) Steinbeck, C.; Krause, S.; Kuhn, S. NMRShiftDB—Constructing a free chemical information system with open-source components. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (6), 1733−1739.
(18) Herget, S.; Toukach, P. V.; Ranzinger, R.; Hull, W. E.; Knirel, Y. A.; von der Lieth, C. W. Statistical analysis of the Bacterial Carbohydrate Structure Data Base (BCSDB): characteristics and diversity of bacterial carbohydrates in comparison with mammalian glycans. *BMC Struct. Biol.* **2008**, *8* (8), No. 35.
(19) Lapadula, A. J.; Hatcher, P. J.; Hanneman, A. J.; Ashline, D. J.; Zhang, H.; Reinhold, V. N. Congruent strategies for carbohydrate sequencing. 3. OSCAR: An algorithm for assigning oligosaccharide topology from MS^n data. *Anal. Chem.* **2005**, *77* (19), 6271−6279.
(20) Zhang, H.; Singh, S.; Reinhold, V. N. Congruent strategies for carbohydrate sequencing. 2. FragLib: An MS^n spectral library. *Anal. Chem.* **2005**, *77* (19), 6263−6270.
(21) Toukach, P. V. Bacterial carbohydrate structure database 3: Principles and realization. *J. Chem. Inf. Model.* **2011**, *51* (1), 159−170.
(22) Egorova, K. S.; Toukach, P. V. Expansion of coverage of Carbohydrate Structure Database (CSDB). *Carbohyd. Res.* **2014**, *389*, 112−114.
(23) Aspinall, G. O.; McDonald, A. G.; Pang, H.; Kurjanczyk, L. A.; Penner, J. L. An antigenic polysaccharide from *Campylobacter coli* serotype O:30. Structure of a teichoic acid-like antigenic polysaccharide associated with the lipopolysaccharide. *J. Biol. Chem.* **1993**, *268* (24), 18321−18329.
(24) Peric-Hassler, L.; Hansen, H. S.; Baron, R.; Hunenberger, P. H. Conformational properties of glucose-based disaccharides investigated using molecular dynamics simulations with local elevation umbrella sampling. *Carbohydr. Res.* **2010**, *345* (12), 1781−1801.
(25) Bubb, W. A. NMR spectroscopy in the study of carbohydrates: Characterizing the structural complexity. *Concepts Magn. Reson.* **2003**, *19A* (1), 1−19.
(26) Martin-Pastor, M.; Bush, C. A. Comparison of the conformation and dynamics of a polysaccharide and of its isolated heptasaccharide repeating unit on the basis of nuclear Overhauser effect, long-range C−C and C−H coupling constants, and NMR relaxation data. *Biopolymers* **2000**, *54* (4), 235−248.
(27) Yuriev, E.; Farrugia, W.; Scott, A. M.; Ramsland, P. A. Three-dimensional structures of carbohydrate determinants of Lewis system antigens: Implications for effective antibody targeting of cancer. *Immunol. Cell Biol.* **2005**, *83* (6), 709−717.
(28) Lutteke, T. Analysis and validation of carbohydrate three-dimensional structures. *Acta Crystallogr., Sect. D* **2009**, *65* (Pt. 2), 156−168.
(29) Grant, D. M.; Cheney, B. V. Carbon-13 magnetic resonance. VII. Steric perturbation of the carbon-13 chemical shift. *J. Am. Chem. Soc.* **1967**, *89* (21), 5315−5318.
(30) Goffin, D.; Bistricky, P.; Shashkov, A.; Lynch, M.; Paquot, M.; Savage, A.; Hanon, E. A Systematic NMR Determination of α-D-Glucooligosaccharides, Effect of Linkage Type, Anomeric Configuration and Combination of Different Linkages Type on ^{13}C Chemical Shifts for the Determination of Unknown Isomaltooligosaccharides. *Bull. Korean Chem. Soc.* **2009**, *30* (11), 2535−2541.
(31) Porto, W. Evolutionary Programming. In *Handbook of Evolutionary Computation*; Bäck, T., Fogel, D. B., Michalewicz, Z., Eds.; IOP Publishing Ltd and Oxford University Press: Bristol, U.K., 1997; Vol. *1*, pp B1.4:1−B1.4:10.
(32) Lin, G.; Lu, X.; Liang, Y.; Kang, L.; Yao, X. A Self-Adaptive Evolutionary Programming Based on Optimum Search Direction. In *Advances in Computation and Intelligence: Third International Symposium on Intelligence Computation and Applications*; Kang, L., Ed.; Springer: Berlin, 2008; pp 9−18.
(33) Matsumoto, M.; Nishimura, T. Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Trans. Model. Comput. Simul.* **1998**, *8* (1), 3−30.
(34) Box, G. E. P.; Muller, M. E. A note on the generation of random normal deviates. *Ann. Math. Stat.* **1958**, *29* (2), 610−611.
(35) Taylor, J. R., Chauvenet's Criterion. In *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*; University Science Books: Sausalito, CA, 1997; pp 166−169.
(36) Damerell, D.; Ceroni, A.; Maass, K.; Ranzinger, R.; Dell, A.; Haslam, S. M. The GlycanBuilder and GlycoWorkbench glycoinformatics tools: Updates and new developments. *Biol. Chem.* **2012**, *393* (11), 1357−1362.
(37) Herget, S.; Ranzinger, R.; Maass, K.; Lieth, C. W. GlycoCT—A unifying sequence format for carbohydrates. *Carbohyd. Res.* **2008**, *343* (12), 2162−2171.
(38) Bagno, A.; Rastrelli, F.; Saielli, G. Prediction of the ^{1}H and ^{13}C NMR spectra of α-D-glucose in water by DFT methods and MD simulations. *J. Org. Chem.* **2007**, *72* (19), 7373−7381.
(39) Schoenhofen, I. C.; McNally, D. J.; Vinogradov, E.; Whitfield, D.; Young, N. M.; Dick, S.; Wakarchuk, W. W.; Brisson, J. R.; Logan, S. M. Functional characterization of dehydratase/aminotransferase pairs from *Helicobacter* and *Campylobacter*: Enzymes distinguishing the pseudaminic acid and bacillosamine biosynthetic pathways. *J. Biol. Chem.* **2006**, *281* (2), 723−732.
(40) Ovchinnikova, O. G.; Arbatsky, N. P.; Chizhov, A. O.; Kocharova, N. A.; Shashkov, A. S.; Rozalski, A.; Knirel, Y. A. Structure of a polysaccharide from *Providencia rustigianii* O11 containing a novel amide of 2-acetamido-2-deoxygalacturonic acid with L-glutamyl-L-alanine. *Carbohydr. Res.* **2012**, *349*, 95−102.
(41) Sigida, E. N.; Fedonenko, Y. P.; Shashkov, A. S.; Zdorovenko, E. L.; Konnova, S. A.; Ignatov, V. V.; Knirel, Y. A. Structural studies of the O-specific polysaccharide(s) from the lipopolysaccharide of *Azospirillum brasilense* type strain Sp7. *Carbohydr. Res.* **2013**, *380*, 76−80.
(42) Shashkov, A. S.; Potekhina, N. V.; Naumova, I. B.; Evtushenko, L. I.; Widmalm, G. Cell wall teichoic acids of *Actinomadura viridis* VKM Ac-1315T. *Eur. J. Biochem.* **1999**, *262* (3), 688−695.
(43) Gil-Serrano, A. M.; Rodriguez-Carvajal, M. A.; Tejero-Mateo, P.; Espartero, J. L.; Menendez, M.; Corzo, J.; Ruiz-Sainz, J. E.; Buendi, A. C. A. M. Structural determination of a 5-acetamido-3,5,7,9-tetradeoxy-7-(3-hydroxybutyramido)-L-glycero-L-mannononulosonic acid-containing homopolysaccharide isolated from *Sinorhizobium fredii* HH103. *Biochem. J.* **1999**, *342* (Pt. 3), 527−535.
(44) King, J. D.; Mulrooney, E. F.; Vinogradov, E.; Kneidinger, B.; Mead, K.; Lam, J. S. *lfnA* from *Pseudomonas aeruginosa* O12 and *wbuX* from *Escherichia coli* O145 encode membrane-associated proteins and

are required for expression of 2,6-dideoxy-2-acetamidino-L-galactose in lipopolysaccharide O antigen. *J. Bacteriol.* **2008**, *190* (5), 1671−1679.
(45) Smurnyy, Y. D.; Blinov, K. A.; Churanova, T. S.; Elyashberg, M. E.; Williams, A. J. Toward more reliable ^{13}C and ^{1}H chemical shift prediction: A systematic comparison of neural-network and least-squares regression based approaches. *J. Chem. Inf. Model.* **2008**, *48* (1), 128−134.
(46) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*; Gaussian, Inc.: Wallingford, CT, 2009.
(47) Laikov, D. N.; Ustynyuk, Y. A. PRIRODA-04: A quantum-chemical program suite. New possibilities in the study of molecular systems with the application of parallel computing. *Russ. Chem. Bull., Int. Ed.* **2005**, *54* (3), 820−826.
(48) Lipkind, G. M.; Shashkov, A. S.; Knirel, Y. A.; Vinogradov, E. V.; Kochetkov, N. K. A computer-assisted structural analysis of regular polysaccharides on the basis of ^{13}C NMR data. *Carbohydr. Res.* **1988**, *175* (1), 59−75.