Supporting Information

Improved Carbohydrate Structure Generalization Scheme for ^{1}H and ^{13}C NMR simulations

Roman R. Kapaev*,[†], 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 prosp. 47, Moscow 119991, Russia

* Corresponding authors:

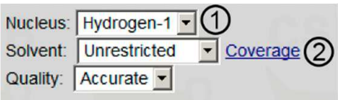
Roman R. Kapaev <kapaev_roman@mail.ru>, Philip V. Toukach <netbox@toukach.ru>

ABSTRACT: The improved Carbohydrate Structure Generalization Scheme has been developed for the simulation of ^{13}C and ^{1}H NMR spectra of oligo- and polysaccharides and their derivatives, including those containing non-carbohydrate constituents found in natural glycans. Besides adding the ^{1}H NMR calculations, we improved the accuracy and performance of prediction and optimized the mathematical model of the precision estimation. This new approach outperformed other methods of chemical shift simulation, including database-driven, neural net based and purely empirical methods and quantum-mechanical calculations at high theory levels. It can process structures with rarely occurring and non-carbohydrate constituents, unsupported by the other methods. The algorithm is transparent to users and allows tracking used reference NMR data to original publications. It was implemented in the Glycan-Optimized Dual Empirical Spectrum Simulation (GODESS) web service, which is freely available on the platform of the Carbohydrate Structure Database (CSDB) project (http://csdb.glycoscience.ru).

Contents of supplementary materials

1. Minor algorithmic improvements and user interface changes

1. In the *fast* mode, the maximal number of generalizations for a single chemical shift is reduced from 10 to 5 to speed up the calculation.
2. In the *accurate* and *extreme* modes, the minimal number of records required for prediction is reduced from 3 to 1 to speed up the calculation.
3. The user interface was modified slightly to support the ^{1}H NMR simulations. From the structure input form (available from the dedicated Web pages http://csdb.glycoscience.ru/bacterial/core/nmrsim.html for predictions based on BCSDB and http://csdb.glycoscience.ru/plant_fungal/core/nmrsim.html for predictions based on PFCSDB), it is now possible to select the nucleus (Figure S-1, **(1)**). In the solvent section, the "Coverage" link displays how many spectra in the database were recorded in each solvent for the selected nucleus (Figure S-1, **(2)**). This information is useful, because the prediction accuracy strongly depends on the database completeness, and there is no solvent generalization provided by the CSGS engine, so all data will be restricted to a certain solvent, if selected. For example, predictions in DMSO using BCSDB will be problematic, because there are only a few NMR spectra recorded in this solvent.



Nucleus: Hydrogen-1

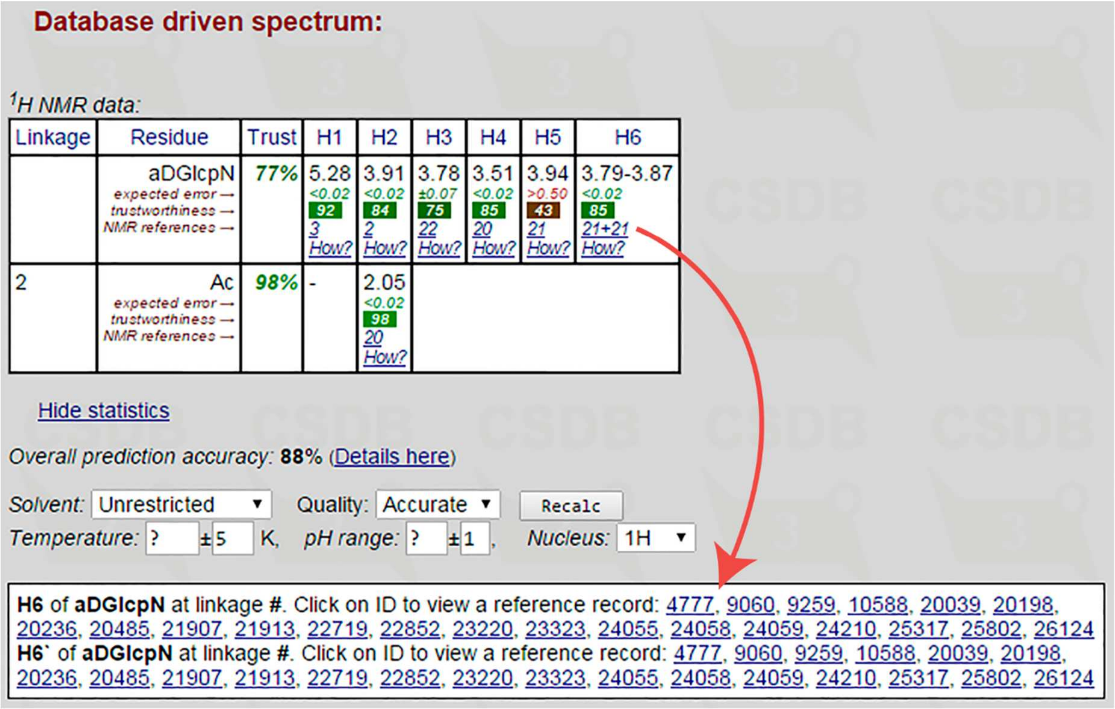
Solvent: Unrestricted Coverage

Quality: Accurate

Figure S-1. New features in the user input interface: **(1)** support of different nucleus selection (^{13}C, ^{1}H), **(2)** solvent coverage viewer.

For both ^{13}C and ^{1}H NMR simulations, the output interface includes assignment tables (see Figure S-2) and sorted NMR spectra. For the ^{13}C NMR prediction, a schematic spectrum is plotted.[1] The assignment table displays chemical shifts for each atom, the trustworthiness level (from 0 to 100%), the expected error (ppm), links to lists of references used for the prediction of each signal and links to generalization reports *("How?")*. If the expected error is very small (< 0.02 ppm for ^{1}H and < 0.2 ppm for ^{13}C NMR simulations) or big (> 0.5 ppm for ^{1}H and > 5 ppm for ^{13}C NMR simulations), the displayed values are "<0.02"/"<0.2" or ">0.5"/">5", respectively. For ^{1}H NMR, if there are more than one atom at the same carbon (e.g., H6 and H6' in glucose), the output has the following features:

a) if the chemical shifts are different, they are hyphenated (see H6 of GlcNAc in Figure S-2);
b) numbers of references used for each atom are displayed separately (e.g., if 10 references are used for H6 and 11 for H6' simulations, the transcription will be *"10+11"*);
c) for each atom, there is a separate reference list (see Figure S-2, below).



Database driven spectrum:

^{1}H NMR data:

Linkage	Residue	Trust	H1	H2	H3	H4	H5	H6
	aDGlcpN expected error → trustworthiness → NMR references →	77%	5.28 <0.02 92 3 How?	3.91 <0.02 84 2 How?	3.78 ±0.07 75 22 How?	3.51 <0.02 85 20 How?	3.94 >0.50 43 21 How?	3.79-3.87 <0.02 85 21+21 How?
2	Ac expected error → trustworthiness → NMR references →	98%	-	2.05 <0.02 98 20 How?				

Hide statistics

Overall prediction accuracy: **88%** (Details here)

Solvent: Unrestricted ▾ Quality: Accurate ▾ Recalc

Temperature: ? ±5 K, pH range: ? ±1 , Nucleus: 1H ▾

H6 of **aDGlcpN** at linkage #. Click on ID to view a reference record: 4777, 9060, 9259, 10588, 20039, 20198, 20236, 20485, 21907, 21913, 22719, 22852, 23220, 23323, 24055, 24058, 24059, 24210, 25317, 25802, 26124

H6` of **aDGlcpN** at linkage #. Click on ID to view a reference record: 4777, 9060, 9259, 10588, 20039, 20198, 20236, 20485, 21907, 21913, 22719, 22852, 23220, 23323, 24055, 24058, 24059, 24210, 25317, 25802, 26124

Figure S-2. Output of the results of ^{1}H NMR simulation of α-D-Glc*p*NAc (using BCSDB, the *accurate* mode; solvent, pH and temperature are unrestricted). The red arrow shows the result of clicking at the reference link for GlcNAc H6.

4. The software was updated to provide detailed reports on generalizations applied during the simulation process. The reports can be displayed by pressing the "*How?*" link in the assignment table. Each generalization is output according to the following syntax:

CENTRAL RESIDUE: *[RED] [TER] [PST] [anomeric][absolute]([basename]* **or** *{[superclass]})[ringsize][qualifiers], [atomic pattern], [stereocode(s)]*

DONOR No.1: [DONOR->(*[goes to]*)CENTRAL] *[anomeric][absolute]([basename]* **or** *{[superclass]})[ringsize][qualifiers], [atomic pattern], [stereocode(s)]*

...

DONOR No.*N*: [DONOR->(*[goes to]*)CENTRAL] *[anomeric][absolute]([basename]* **or** *{[superclass]})[ringsize][qualifiers], [atomic pattern], [stereocode(s)]*

ACCEPTOR: [CENTRAL->(*[goes to]*)ACCEPTOR] *[anomeric][absolute]([basename]* **or** *{[superclass]})[ringsize][qualifiers], [atomic pattern], [stereocode(s)]*

Weight: *[weight]*

[reported if saved as result]

The transcript is the following:

- "CENTRAL RESIDUE": a residue which the predicted atom belongs to.
- *RED* is a flag: if present, the central residue is at the reducing end.
- *TER* is a flag: if present, the central residue is terminal.
- *PST* is a flag: if present, the central residue is not terminal, but may be substituted only at the substituent positions of an input structure. If both *TER* and *PST* flags are absent, the central residue may be substituted at any position.
- *anomeric*: anomeric configuration of a residue (a = α, b = β, or not present).
- *absolute*: absolute configuration of a residue (D, L, R = *(R)*, S = *(S)*, or not present).
- *basename*: residue name (Gal, Glc, Man, etc.).
- if the *basename* property is generalized, the *superclass* property (ald = aldose, ket = ketose, pep = peptide, etc.) is given in braces.
- *ringsize*: ring size of a residue (*p*, *f* or not present).

- *qualifiers*: residue modifiers, such as amino group (N, 3N), carboxyl group (A), etc.
- *atomic pattern*: a string listing the types of carbons in the residue. The following types are used: “o” = >CH−OH (hydroxy); “n” = >CH−NH− (amino); “a” = −COOH (carboxy); “d” = $-CH_2-$ or >CH− (deoxy); “D” = −CH= (deoxy carbon forming double bond); “O” = −C(OH)= (hydroxy sp^2); “N” = −C(NHR)= (amino sp^2) or $-CONH_2$ (amide); “x” = other carbon; “_” = any atom; “%” = any set of atoms.
- *stereocode(s)*: atomic stereo configuration recorded according to the MonosaccharideDB (http://www.monosaccharidedb.org/) stereocode notation: “1” indicates L-configuration of a carbon (OH group pointing to the left in the Fischer projection); “2” indicates D-configuration (OH group pointing to the right in the Fischer projection); “0” is for achiral carbons. If the absolute configuration of a residue is generalized or does not exist, there may be two stereocode variants (one for each configuration): in these cases, both stereocodes are listed and matched to the appropriate configurations.
- “DONOR No.*N*”: enumerated donor of the central residue.
- “ACCEPTOR”: acceptor of the central residue.
- *goes to*: for donors, it is the position in the central residue bonded to the donor; for acceptors, it is the position in the acceptor bonded to the central residue.
- *weight*: weight of the applied generalizations.
- *reported if saved as result*: if the data obtained by generalization are used for simulation, the trustworthiness (trust) and the number of chemical shifts used for the simulation are reported.

Below are some examples of generalization reports for generalized structures:

- searching α-Glc*p*:

  CENTRAL RESIDUE: RED TER {ald}*p*, atomic pattern: oooodo, stereocode: 221220 for D-configuration, 112110 for L-configuration
  Weight: 0
- Searching β-D-Glc*p*-(1→6)-β-D-Gal*p*N, predicting Glc chemical shifts:

  CENTRAL RESIDUE: TER D{ald}*p*, atomic pattern: oooodo, stereocode: 121220
  ACCEPTOR: [CENTRAL->(6)ACCEPTOR]: D{ald}*p*, atomic pattern: onoodo, stereocode: 121120
  Weight: 0
  Saved as result; trust: 3, chemical shifts used: 1
- Searching α-D-RES*p*-(1→6)-β-D-Glc*p*, where RES = Man/Tal/Man4N/Tal4N/…, predicting β-D-Glc*p* C1 (see the first generalization step in Figure 1 of the primary paper)
  CENTRAL RESIDUE: RED PST D{ald}*p*, atomic pattern: oooodo, stereocode: 121220
  DONOR No.1: [DONOR->(6)CENTRAL]: D{ald}*p*, atomic pattern: ooo_do, stereocode: 211_2%
  Weight: 0.04883

5. The weight scale was changed from 0..20 to more intuitive 0..100. Analogously, the trustworthiness scale was changed from 0..4 to 0..100.
6. The SQL queries were optimized to enhance the prediction performance.

2. Weight optimizations

Algorithm S-1 (weight optimization)

00: a) Select a sampling of experimental chemical shifts assigned to certain molecular structures.
b) Predict the chemical shifts for each structure (predictions were carried out using BCSDB in the *accurate* mode with unrestricted solvent, pH, and temperature parameters; during the simulations, the program was prohibited to use database records containing the spectra of the structure currently being predicted to avoid the prediction bias).

c) Exclude those chemical shifts, the prediction of which did not require any generalizations (they would be useless for the weight optimization; see Table S-2 for the resultant sampling characteristics for different weight sets).

```
01: Set the number of chaotic iteration $K = 300$, the population size $SN = 10$, the cluster-specific number of weights under optimization $n$ (depending on residue type), the maximal number of trials for an individual $L = 10$, the number of iterations $G = 100$, $P = 0.7$.
02: for $i = 1$ to $SN$ do
03:     for $j = 1$ to $n$ do
04:         Randomly initialize variable $ch_{0,j} \in (0,1)$
05:         for $k = 1$ to $K$ do
06:             $ch_{k+1,j} = \sin(\pi ch_{k,j})$
07:         end for
08:         Set the $j$-th component (certain weight) of the $i$-th individual (weight set):
            $X_{i,j} = X_{min,j} + ch_{K,j}(X_{max,j} - X_{min,j})$
            where $X_{min,j}$ and $X_{max,j}$ are lower and upper boundaries for $X_{i,j}$. See Table S-3 for boundary values of different descriptor weights. The boundaries were set based on our previous study[1].
09:     end for
10: end for
11: for $i = 1$ to $SN$ do
12:     for $j = 1$ to $n$ do
13:         $OX_{i,j} = X_{max,j} + X_{min,j} - X_{i,j}$
14:     end for
15: end for
16: Select $SN$ fittest individuals from $X$ and $OX$. Save these individuals as the initial generation $\{X_i|i = 1, 2, \ldots, SN\}$, save the fittest individual as $XBest$. Here and below, the fittest individual is the one with the minimal mean absolute deviation ($\Delta(X_i)$) of the predicted chemical shifts from the experimental ones for the sampling.
17: Set the initial number of trials for each individual to zero ($trial_i = 0|i = 1, 2, \ldots, SN$)
18: for $g = 1$ to $G$
19:     for $i = 1$ to $SN$ do
20:         for $j = 1$ to $n$ do
21:             Choose $X_{r1}$, $X_{r2}$ from the current population. $r_1$ and $r_2$ are integers selected randomly from $[1,SN]$ ($r_1 \neq r_2 \neq i$).
22:             Produce a random number $\varphi \in [-1,1]$ (this number is generated anew each time).
23:             Generate the $j$-th component of a new food source $V_i$:
                $V_{i,j} = XBest_j + \varphi(X_{r1,j} - X_{r2,j})$
24:             if ($V_{i,j} \leq 0$) then
25:                 Repeat steps 21-23 (all the weights must be positive)
26:             end if
27:         end for
28:         if ($\Delta(X_i) > \Delta(V_i)$) then
29:             $X_i = V_i$
30:         else then
31:             $trial_i = trial_i + 1$
32:             if ($rand(0,1) < P$) then
33:                 for $j = 1$ to $n$ do
34:                     Select random $k \in \{1, 2, \ldots, SN\}$ ($k \neq i$).
35:                     Produce a random number $\varphi \in [-1,1]$
36:                     Generate the $j$-th component of a new food source $V_i$:
                        $V_{i,j} = X_{i,j} + \varphi(X_{i,j} - X_{k,j})$
37:                     if ($V_{i,j} \leq 0$) then
38:                         Repeat steps 33-35 (all the weights must be positive)
39:                     end if
40:                 end for
41:                 if ($\Delta(X_i) > \Delta(V_i)$) then
42:                     $X_i = V_i$
```

```
43:                 end if
44:             else then
45:                 $trial_i = trial_i + 1$
46:             end if
47:         end if
48:         if ($trial_i > L$) then
49:             Reinitialize the $i$-th individual using steps 03-09
50:             Calculate $\Delta(X_i)$ for the reinitialized individual
51:             $trial_i = 0$
52:         end if
53:         Choose the fittest individual and save it as $XBest$
54:     end for
55: end for
56: Output $XBest$
```

Table S-1. Sampling characteristics used for the weight optimizations.

nucleus	**weight factor set**	**training structures** [a]	**total number of residues**	**total number of atoms**
^{13}C	pyranose	pyranose forms of 6dgulHep, 6dTal, 4dthrHex4enA, Abe, Ara4N, Col, DDmanHep, Fuc, Fuc4N, FucN4N, Gal, GalA, GalN, GalNA, Glc, GlcA, GlcN, GulNA, IdoA, Kdo, Ko, LDmanHep, Man, ManA, ManN, ManNA, Par, Pse, Qui3N, Qui4N, QuiN, Rha, Xyl	100	538
	furanose	furanose forms of 2,7anhKdo, 6dAlt, 6daltHep, Ara, Fru, Fuc, Gal, GalN, Kdo, Par, Rib, Xul, Xyl	88	472
	alditol	1dEry-ol, Ery-ol, Etg, Fuc-ol, Gal-ol, Glc-ol, GlcN-ol, Gro, GroN, Kdo-ol, Man-ol, Qui3N-ol, Rib-ol, Thre-ol	61	226
	other	3HOBut, 4HOBut, Ala, Asp, aThr, Cys, Glu, Gly, GroA, Lac, Me, Orn, SRCetLys, SS3,5HOHex, SSCetLys, Ser, Tyr	78	222
^{1}H	pyranose	pyranose forms of 6dgulHep, 6dmanHep, 6dTal, 4dthrHex4enA, Col, Fuc, Fuc3N, Gal, GalA, GalN, Glc, GlcA, GlcN, GulNA, Kdn, Kdo, LDmanHep, Man, ManNA, Neu, Pse, QuiN, QuiN4N, Rha, RhaN3N, Tyv, Yer	100	566
	furanose	furanose forms of 6dAlt, 6daltHep, Ara, Fru, Fuc, Gal, Par, Rib, Xul	100	575
	alditol	1dEry-ol, Ery-ol, Etg, Fuc-ol, Gal-ol, GalN-ol, Glc-ol, GlcN-ol, Gro, Kdo-ol, Man-ol, Qui3N-ol, Rib-ol, Thre-ol	54	209
	other	3HOBut, 4HOBut, Ala, Asp, aThr, Cys, Gly, GroA, Lac, Me, SRCetLys, SS3,5HOHex, SSCetLys, Ser	67	155

[a] Central residue names are listed. The residues are substituted by pyranoses, furanoses, alditols, amino acids, lipids, phosphoric acid, O- and N-linked acetic acid, methanol, and other residues.

Table S-2. Upper and lower boundaries for different structural descriptors.

Residue type	**Property**	**Property remoteness (bonds)**			
		0	**1**	**2**	**3**
central residue	atom type and stereo configuration	(0,50)	(0,50)	(0,50)	(0,50)
	ring size	(75,100)	(0,75)/(50,75)	-/(0,50)	-
	acceptor presence	(75,100)	(50,75)	(25,50)	(0,25)
	donor(s) presence	(0,100)			
substituent residue	atom type	(25,100)	(10,25)	(0,10)	-
	atom stereo configuration	(20,50)	(1,20)	(0,1)	-
	ring size	(50,100)	(0,50)	-	-
	absolute configuration	(0,100)			
	proximity factor	(1,1)	(0.5,1)	(0.2,0.5)	(0,0.2)

Table S-3. The resultant weight factors of central residues (a residue containing atom under prediction) for different residue types and nuclei (^{1}H, ^{13}C). Property remoteness (R) stands for the number of bonds from the generalized atom(s) to the atom under prediction. Four weight factor sets are provided, corresponding to pyranose, furanose, alditol and common parameter sets. The generalized properties were described and exemplified earlier[1].

nucleus	**property**	**property remoteness (*R*)**	**weight factor**			
			pyranose	**furanose**	**alditol**	**other**
^{13}C	atom stereo configuration and atom type[a)]	≥4	$10^{-R+2}+W_{FAR}$			
		3	$14.6+W_{FAR}$	$92.8+W_{FAR}$	$16.8+W_{FAR}$	$9.9+W_{FAR}$
		2	$80.9+W_{FAR}$	$82.1+W_{FAR}$	$8.9+W_{FAR}$	$47.5+W_{FAR}$
		1	$40.5+W_{FAR}$	$49.0+W_{FAR}$	$57.5+W_{FAR}$	$37+W_{FAR}$
		0	$98.6+W_{FAR}$	$26.5+W_{FAR}$	$85.6+W_{FAR}$	$25.3+W_{FAR}$
	ring size	≥3	10^{-R+2}		inapplicable	inapplicable
		2	1	16.9	inapplicable	inapplicable
		1	38.8	51.6	inapplicable	inapplicable
		0	91.1	74.2	inapplicable	inapplicable
	presence of acceptor (reducing end/inline)	≥4	10^{-R+2}			
		3	15.3	1.4	10.1	10.2
		2	45.8	55.8	29.2	34.8
		1	53.4	87.2	60.7	56.2
		0	71.9	100	79	99.4
	presence of donor(s), for terminal residues	? (any atom)	100	13.1	10	43.9
	presence of donor(s), for non-terminal residues	? (any atom)	71	12.3	9.4	8.3
	Proximity factor (PF)	≥4	10^{-R+2}			
		3	0.1628	0.0035	0.1875	0.1000
		2	0.4782	0.5046	0.2504	0.2028
		1	0.5046	0.6460	0.5213	0.8784
		0	1			
^{1}H	atom stereo configuration and atom type[a)]	≥4	$10^{-R+2}+W_{FAR}$			
		3	$4.3+W_{FAR}$	$54.9+W_{FAR}$	$32.7+W_{FAR}$	$0.4+W_{FAR}$
		2	$86+W_{FAR}$	$16.5+W_{FAR}$	$10+W_{FAR}$	$40.6+W_{FAR}$
		1	$42.2+W_{FAR}$	$12.4+W_{FAR}$	$75.8+W_{FAR}$	$46.2+W_{FAR}$
		0	$95.3+W_{FAR}$	$63.2+W_{FAR}$	$84.9+W_{FAR}$	$34.4+W_{FAR}$
	ring size	≥3	10^{-R+2}		inapplicable	inapplicable
		2	1	29.6	inapplicable	inapplicable
		1	40.4	61.7	inapplicable	inapplicable
		0	76.9	80.6	inapplicable	inapplicable
	presence of acceptor (reducing end/inline)	≥4	10^{-R+2}			
		3	1.7	14.9	14.9	0.2
		2	43.6	35.6	42.6	46.2
		1	51.8	71	63.7	63.7
		0	89.2	100	58.8	89.3
	presence of donor(s), for terminal residues	? (any atom)	27	97.3	55.3	15.6
	presence of donor(s), for non-terminal residues	? (any atom)	20.1	86.9	86	48.4
	Proximity factor (PF)	≥4	10^{-R+2}			
		3	0.0675	0.1528	0.0157	0.1046
		2	0.2094	0.2540	0.4773	0.2351
		1	0.6591	0.5033	0.7872	0.8410

		0	1

[a] Generalizations of stereo configuration and type of atom of a central residue implies generalization of all more distant atoms and full generalization of substituent residues linked to the 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.

Table S-4. The resultant weight factors of substituent residues (donors or acceptors linked to a central residue) for different types of the central residue and nuclei (^1H, ^{13}C). 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.

nucleus	**property**	**property remoteness (*R*)**	**weight factor**			
			pyranose	**furanose**	**alditol**	**other**
^{13}C	atom stereo configuration	≥3	10^{-R+1}×PF			
		2	0.4×PF	0.6×PF	0.01×PF	0.5×PF
		1	13.7×PF	2.2×PF	11.7×PF	14.0×PF
		0	24.2×PF	25.6×PF	31.8×PF	38.5×PF
	atom type	≥3	10^{-R+1}×PF			
		2	0.3×PF	0.1×PF	0.03×PF	0.03×PF
		1	24.8×PF	0.52×PF	21.8×PF	11.6×PF
		0	26.3×PF	11.8×PF	26.4×PF	64.0×PF
	ring size	≥2	10^{-R+1}×PF			
		1	21.2×PF	8.0×PF	20.6×PF	37.6×PF
		0	37.7×PF	83.6×PF	76.8×PF	64.2×PF
	absolute configuration (combination with central residue)	? (any atom)	8.8×PF	25.1×PF	6.36×PF	97.3×PF
^1H	atom stereo configuration	≥3	10^{-R+1}×PF			
		2	0.003×PF	0.1×PF	0.26×PF	0.46×PF
		1	15.6×PF	10.7×PF	17.4×PF	17.3×PF
		0	49.1×PF	45.7×PF	42.5×PF	24.6×PF
	atom type	≥3	10^{-R+1}×PF			
		2	0.2×PF	0.34×PF	0.3×PF	0.4×PF
		1	19.1×PF	23.3×PF	10.3×PF	15.1×PF
		0	30.6×PF	95.0×PF	60.0×PF	75.1×PF
	ring size	≥2	10^{-R+1}×PF			
		1	15.5×PF	29.3×PF	18.5×PF	19.6×PF
		0	93.3×PF	67.0×PF	68.6×PF	95.5×PF
	absolute configuration (combination with central residue)	? (any atom)	69.7×PF	86.9×PF	0.3×PF	75.0×PF

3. Substituent generalizations

Table S-5. Number of substituent generalization steps *n* for different prediction modes. *PF* stands for proximity factor, as described[1].

Mode	***n***
fast	1
accurate	$\lceil 2 \times PF \rceil$
extreme	4

4. Trustworthiness evaluation

Possible values of the trustworthiness are limited: $T \in [0;100]$. This limitation was applied because it was problematic to debug and control the algorithm behavior at very big ($T \rightarrow +\infty$) or small ($T \rightarrow -\infty$) values. Hence, if T calculated with the formula is below zero, it is set to zero.

Generally, increasing the generalization weight (W) and standard deviation in a dataset (σ), and decreasing the number of chemical shifts in the dataset (N) lead to less precise results. This imposes the following restriction:

$$\forall X_1, X_2 \in [0;+\infty)\ X_1 < X_2 : T(X_1) \geq T(X_2) \tag{1}$$

where $T(X)$ is one-parameter dependency of T from a certain parameter X (which is W, $1/N$ or σ):

$$T(X) = 100 - P_X(X) \tag{2}$$

This imposes the following restrictions on $P_X(X)$:

$$\begin{cases} P_X(X) = 100, x_1X + x_2X^2 > 100 \\ P_X(X) = \frac{-x_1^2}{4x_2}, X > -\frac{x_1}{2x_2}, \frac{-x_1^2}{4x_2} \leq 100, x_2 \neq 0 \end{cases} \tag{3}$$

In other cases, $P_X(X)$ is calculated as a polynomial:

$$P_X(X) = x_1X + x_2X^2 \tag{4}$$

In addition, x_1 must be greater than zero to satisfy the condition (1).

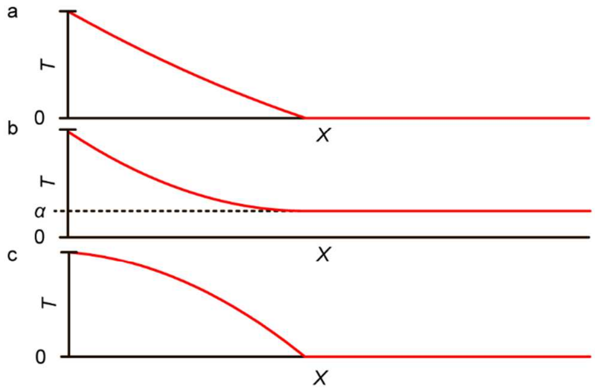


Figure S-3. Three sorts of one parameter dependency of T from X: (a) $x_2 < 0$, $\Delta \geq 0$, (b) $x_2 < 0$, $\Delta < 0$, and (c) $x_2 \geq 0$.

Consequently, there are three cases of the $T(X)$ dependency: when $x_2 < 0$ and the discriminant of $T(X)$ $\Delta \geq 0$ (Figure S-3, a), when $x_2 < 0$ and $\Delta < 0$ (Figure S-3, b), and when $x_2 \geq 0$ (Figure S-3, c). In the second case, $\alpha \in [0;100)$ is a value in the $P_X(X)$ minimum point:

$$\alpha = T\left(-\frac{x_1}{2x_2}\right) = 100 + \frac{x_1^2}{4x_2} \tag{5}$$

Algorithm S-2 (trustworthiness optimization)

```
00: Select a sampling of experimental chemical shifts assigned to certain molecular structures. Predict the chemical shifts for each structure and save the combination of ($W$, $N$, $\sigma$, $\Delta$) for each chemical shift. See Table S-6 for the resultant sampling characteristics.
01: Set the number of chaotic iteration $K = 300$, the population size $SN = 75$, the total number of $x_1$ and $x_2$ under optimization $n = 6$ (two parameters for each $P_X(X)$), the maximal number of trials for an individual $L = 10$, the number of optimization iterations $G = 250$, $P = 0.7$.
02: for $i = 1$ to $SN*2$ do
03:     for $j = 1$ to $n$ do
```

```
04:         Randomly initialize variable $ch_{0,j} \in (0,1)$
05:         **for** $k = 1$ **to** $K$ **do**
06:             $ch_{k+1,j} = \sin(\pi ch_{k,j})$
07:         **end for**
08:         Set the $j$-th component (certain $x_1$ or $x_2$) of the $i$-th individual ($x_1$ and $x_2$ set):
            $X_{i,j} = X_{min,j} + ch_{K,j}(X_{max,j} - X_{min,j})$
            where $X_{min,j}$ and $X_{max,j}$ are lower and upper boundaries for $X_{i,j}$. For each $x_2$, $X_{max,j} = 100$,
            $X_{min,j} = -100$; for each $x_1$, the boundaries are dependent on the corresponding $x_2$ to satisfy
            the above-mentioned limits.
09:     **end for**
10: **end for**
11: Select $SN$ fittest individuals from $X$. Save these individuals as the initial generation $\{X_i|i = 1, 2, \ldots, SN\}$, save the fittest individual as $XBest$. Here and below, the fittest individual is the one with the maximal $f(X_i) = -P_c(X_i)$, where $P_c(X_i)$ is the Pearson correlation coefficient between $\Delta$ and $T$ calculated for the sampling from ($W$, $N$, $\sigma$) using the certain ($x_1$, $x_2$) parameter set.
12: Set the initial number of trials for each individual to zero ($trial_i = 0|i = 1, 2, \ldots, SN$)
13: **for** $g = 1$ **to** $G$
14:     **for** $i = 1$ **to** $SN$ **do**
15:         Select $j$ randomly from $[1,n]$
16:         Choose $X_{r1}$, $X_{r2}$ from the current population. $r_1$ and $r_2$ are integers selected randomly from
            $[1,SN]$ $(r_1 \neq r_2 \neq i)$.
17:         Produce a random number $\varphi \in [-1,1]$ (this number is generated anew each time).
18:         Generate the $j$-th component of a new food source $V_i$:
19:         $V_{i,j} = XBest_j + \varphi(X_{r1,j} - X_{r2,j})$ (other components are set equal to the ones of $X_i$).
20:         **if** ($x_1$ or $x_2$ corresponding to $V_{i,j}$ (which is $x_2$ or $x_1$, respectively) does not satisfy the above-
            mentioned limitations) **then**
21:             Produce a random number $\psi \in [0,1]$
22:             Recalculate the corresponding $x_1$ or $x_2$ (denoted as $V_{i,c}$):
                $V_{i,c} = X_{min,j} + \psi(X_{max,j} - X_{min,j})$
                where $X_{min,j}$ and $X_{max,j}$ are dependent on $V_{i,j}$ to satisfy the limitations
23:         **end if**
24:         **if** ($f(X_i) < f(V_i)$) **then**
25:             $X_i = V_i$
26:         **else then**
27:             $trial_i = trial_i + 1$
28:             **if** ($rand(0,1) < P$) **then**
29:                 Select $j$ randomly from $[1,n]$
30:                 Select random $k \in \{1, 2, \ldots, SN\}$ $(k \neq i)$.
31:                 Produce a random number $\varphi \in [-1,1]$
32:                 Generate the $j$-th component of a new food source $V_i$:
                    $V_{i,j} = X_{i,j} + \varphi(X_{i,j} - X_{k,j})$
33:                 Do steps 20-23
34:                 **if** ($f(X_i) < f(V_i)$) **then**
35:                     $X_i = V_i$
36:                 **end if**
37:             **else then**
38:                 $trial_i = trial_i + 1$
39:             **end if**
40:         **end if**
41:         **if** ($trial_i > L$) **then**
43:             Reinitialize the $i$-th individual using steps 03-09
43:             Calculate $f(X_i)$ for the reinitialized individual
44:             $trial_i = 0$
45:         **end if**
46:         Choose the fittest individual and save it as $XBest$
47:     **end for**
48: **end for**
```

49: Output *XBest*

Table S-6. Sampling characteristics used for the trustworthiness evaluation optimizations, accuracy and performance verification.

nucleus	residue type	structures [a)]	total number of residues	total number of atoms
^{13}C	pyranose	pyranose forms of 6dgulHep, 6dTal, 4dthrHex4enA, Abe, AltNA, Ara4N, Col, DDmanHep, Fuc, Fuc3N, Fuc4N, FucN, FucN4N, Gal, GalA, GalN, GalNA, Glc, Glc3NA, GlcA, GlcN, GlcN3NA, GlcNA, GulNA, IdoA, Kdn, Kdo, Ko, LDmanHep, Man, ManA, ManN, ManN3NA, ManNA, Neu, Par, Pse, Qui3N, Qui4N, QuiN, QuiN4N, Rha, Rha4N, RhaN3N, Xyl	683	4020
	furanose	furanose forms of 2,7anhKdo, 6dAlt, 6daltHep, Ara, Fru, Fuc, Gal, GalN, Kdo, Par, Rib, Xul, Xyl	162	903
	alditol	1dEry-ol, Ery-ol, Etg, Fuc-ol, Gal-ol, GalN-ol, Glc-ol, GlcN-ol, Gro, GroN, Kdo-ol, Man-ol, Qui3N-ol, QuiN-ol, Rib-ol, Thre-ol	118	480
	other	3HOBut, 4HOBut, aThr, Ala, Asp, Cys, Glu, Gly, GroA, Lac, Lys, Me, Orn, SRCetLys, SS3,5HOHex, SSCetLys, Ser, Thr, Tyr	197	483
^{1}H	pyranose	pyranose forms of 6dgulHep, 6dlyxHex-4-ulo, 6dmanHep, 6dTal, 4dthrHex4enA, 6dxylHexN-4-ulo, Abe, Col, DDmanHep, Fuc, Fuc3N, FucN, FucN4N, Gal, GalA, GalN, GalN3NA, GalNA, Glc, Glc3NA, GlcA, GlcN, GlcN3N, GlcN3NA, GlcNA, GulNA, Kdn, Kdo, Kdo8N, LDmanHep, Man, ManA, ManN, ManN3NA, ManNA, Neu, Par, Pse, Qui3N, Qui4N, QuiN, QuiN4N, Rha, Rha4N, RhaN3N, Tyv, Xyl, Yer	626	4054
	furanose	furanose forms of 2,7anhKdo, 6dAlt, 6daltHep, Ara, Fru, Fuc, Gal, GalN, Kdo, Par, Rib, Xul, Xyl	154	999
	alditol	1dEry-ol, Ery-ol, Etg, Fuc-ol, Gal-ol, GalN-ol, Glc-ol, GlcN-ol, Gro, Kdo-ol, Man-ol, Qui3N-ol, QuiN-ol, Rib-ol, Thre-ol	111	650
	other	3HOBut, 4HOBut, Ala, Asp, aThr, Cys, Gc, Glu, Gly, GroA, Lac, Lys, Me, SRCetLys, SS3,5HOHex, SSCetLys, Ser, Thr, Tyr	190	478

[a)] Central residue names are given. The residues are substituted by pyranoses, furanoses, alditols, amino acids, lipids, phosphoric acid, O- and N-linked acetic acid, methanol, and other residues.

Table S-7. The resultant sets of x_1 and x_2 for different types of central residues and nuclei.

nucleus	residue type	parameter	x_1	x_2
^{13}C	pyranose	W	0.3048	-0.0002
		N	35.4382	-2.4013
		σ	27.6909	-2.3835
	furanose	W	22.8231	-1.2843
		N	32.8285	-1.3467
		σ	35.9848	-2.3368
	alditol	W	0.0446	0
		N	33.6428	-2.0476
		σ	36.5949	-3.2725
	other	W	0.0345	0
		N	15.5665	1.0676
		σ	37.6736	-2.2563
^{1}H	pyranose	W	0.5465	-0.0007
		N	8.3225	10.2017
		σ	204.7746	937.7593
	furanose	W	0.4756	0
		N	107.7235	-49.6863
		σ	567.5656	-5.6735
	alditol	W	0.0265	0
		N	8.2974	-0.1696
		σ	150.0168	-245.372
	other	W	0.0039	0
		N	5.0654	-0.0522
		σ	31.3764	10.3615

Table S-8. Resultant parameters A and B for estimation of Δ from T by linear regression ($T = A \times \Delta + B$) for different types of central residue and nuclei.

nucleus	**residue type**	***A***	***B***
^{13}C	pyranose	-5.79	73.35
	furanose	-9.26	57.99
	alditol	-1.53	72.91
	other	-3.25	82.71
^{1}H	pyranose	-67.53	79.39
	furanose	-51.79	38.17
	alditol	-14.31	90.96
	other	-18.32	97.44

5. Accuracy verification

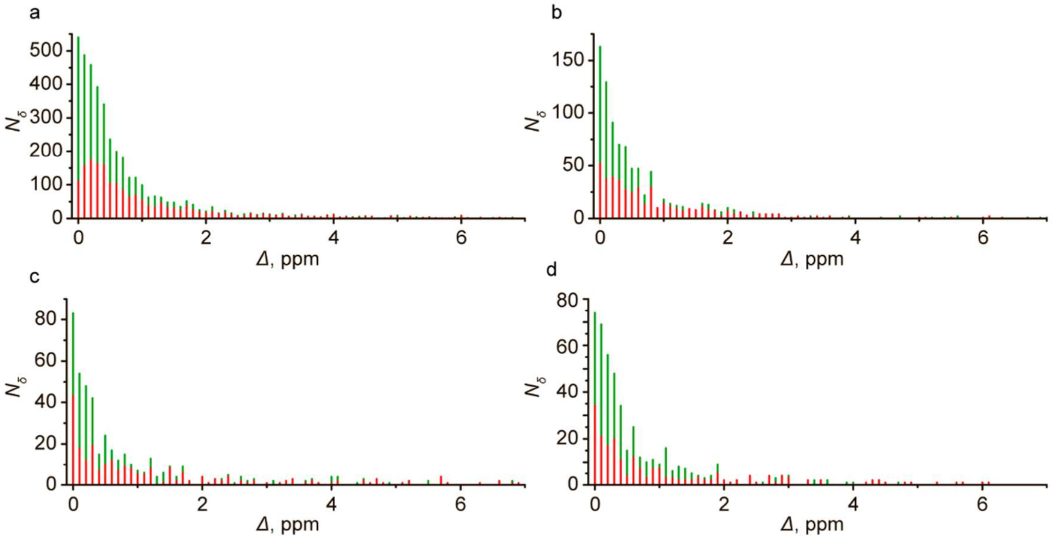


Chart S-1. Number of ^{13}C chemical shifts N_δ depending on the absolute deviation Δ (predicted vs. experimental) for different types of central reside: (a) pyranoses, (b) furanoses, (c) alditols, (d) other residues (see Table S-5 for sampling characteristics). Red bars correspond to chemical shifts prediction of which required generalizations; green bars correspond to other chemical shifts.

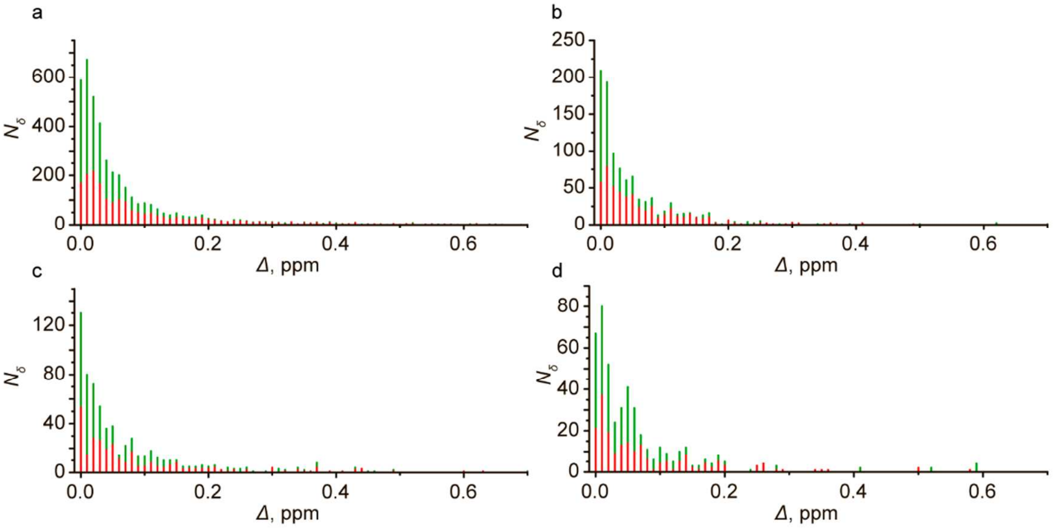


Chart S-2. Number of ^{1}H chemical shifts N_δ depending on the absolute deviation Δ (predicted vs. experimental) for different types of central reside: (a) pyranoses, (b) furanoses, (c) alditols, (d) other residues (see Table S-5 for sampling characteristics). Red bars correspond to chemical shifts prediction of which required generalizations; green bars correspond to other chemical shifts.

6. Trustworthiness verification

Table S-9. Linear correlation coefficients *r* between *T* and observed *Δ* for different types of residues and nuclei (see Table S-5 for samplings). For ^{13}C NMR simulations, the coefficients before optimizations are given.

nuclei	**residue type**	***r* (before optimizations)**	***r* (after optimizations)**
^{13}C	pyranose	0.3270	0.4191
	furanose	0.2791	0.3477
	alditol	0.1183	0.2424
	other	0.3076	0.3614
^{1}H	pyranose	-	0.3455
	furanose	-	0.2543
	alditol	-	0.2998
	other	-	0.6105

7. Abbreviations

Here we explain the residue name abbreviations used in the paper and SI. More detailed explanations are available under the “Monomer Namespace” menu item at the CSDB Web site (http://csdb.glycoscience.ru/database/core/residues.php).

- 1dEry-ol: 1-deoxyerythritol
- 2,7anhKdo: 2,7-anhydro-3-deoxy-D-manno-oct-2-ulosonic acid
- 3HOBut: 3-hydroxybutanoic acid
- 4dthrHex4enA: 4-deoxy-threo-hex-4-enuronic acid
- 4HOBut: 4-hydroxybutanoic acid
- 6dAlt: 6-deoxyaltrose
- 6daltHep: 6-deoxy-altro-heptose
- 6dgulHep: 6-deoxy-gulo-heptose
- 6dlyxHex-4-ulo: 6-deoxy-lyxo-hexos-4-ulose
- 6dmanHep: 6-deoxy-manno-heptose
- 6dTal: 6-deoxytalose
- 6dxylHexN-4-ulo: 2-amino-2,6-dideoxy-xylo-hexos-4-ulose
- Abe: 3,6-dideoxy-D-xylo-hexose (abequose)
- Ac: acetic acid
- Ala: alanine
- AltNA: 2-amino-2-deoxyaltruronic acid
- Ara: arabinose
- Ara4N: 4-amino-4-deoxyarabinose
- Asp: aspartic acid
- aThr: allothreonine
- Col: 3,6-dideoxy-L-xylo-hexose (colitose)
- Cys: cysteine
- DDmanHep: D-glycero-D-manno-heptose
- Ery-ol: erythritol
- Etg: ethylene glycol
- Fru: arabino-hex-2-ulose (fructose)
- Fuc: 6-deoxygalactose (fucose)
- Fuc-ol: 6-deoxy-D-galactitol (D-fucitol), 1-deoxy-D-galactitol (L-fucitol)
- Fuc3N: 3-amino-3,6-dideoxygalactose
- Fuc4N: 4-amino-4,6-dideoxygalactose
- FucN: 2-amino-2,6-dideoxygalactose
- FucN4N: 2,4-diamino-2,4,6-trideoxygalactose
- Gal: galactose
- Gal-ol: galactitol

- GalA: galacturonic acid
- GalN: 2-amino-2-deoxygalactose
- GalN-ol: 2-amino-2-deoxygalactitol
- GalN3NA: 2,3,4-triamino-2,3,4-trideoxygalacturonic acid
- GalNA: 2-amino-2-deoxygalacturonic acid
- Gc: glycolic (2-hydroxyacetic) acid
- Glc: glucose
- Glc-ol: glucitol
- Glc3NA: 3-amino-3-deoxyglucuronic acid
- GlcA: glucuronic acid
- GlcN: 2-amino-2-deoxyglucose
- GlcNAc: 2-acetamido-2-deoxyglucose
- GlcN-ol: 2-amino-2-deoxyglucitol
- GlcN3N: 2,3-diamino-2,3-dideoxyglucose
- GlcN3NA: 2,3-diamino-2,3-dideoxyglucuronic acid
- GlcNA: 2-amino-2-deoxyglucuronic acid
- Glu: glutamic acid
- Gly: glycine
- Gro: glycerol
- GroA: 2,3-dihydroxypropanoic (glyceric) acid
- GroN: 2-amino-2-deoxyglycerol (2-aminopropane-1,3-diol)
- GulNA: 2-amino-2-deoxyguluronic acid
- IdoA: iduronic acid
- Kdn: 3-deoxy-D-glycero-D-galacto-non-2-ulosonic acid (ketodeoxynononic) acid
- Kdo: 3-deoxy-D-manno-oct-2-ulosonic (ketodeoxyoctonic) acid
- Kdo-ol: 3-deoxy-D-manno-oct-2-ulosonic acid alditol
- Kdo8N: 8-amino-3,8-dideoxy-D-manno-oct-2-ulosonic acid
- Ko: D-glycero-D-talo-oct-2-ulosonic (ketooctonic) acid
- Lac: 2-hydroxypropanoic (lactic) acid ether (substituent: 1-carboxyethyl)
- LDmanHep: L-glycero-D-manno-heptose
- Lys: lysine
- Man: mannose
- Man-ol: mannitol
- ManA: mannuronic acid
- ManN: 2-amino-2-deoxymannose
- ManN3NA: 2,3-diamino-2,3-dideoxymannuronic acid
- ManNA: 2-amino-2-deoxymannuronic acid
- Me: methanol
- Neu: 5-amino-3,5-dideoxy-D-glycero-D-galacto-non-2-ulosonic acid (neuraminic) acid
- Orn: ornithine
- Par: 3,6-dideoxy-D-ribo-hexose (paratose)
- Pse: 5,7-diamino-3,5,7,9-tetradeoxy-L-glycero-L-manno-non-2-ulosonic (pseudaminic) acid
- Qui3N: 3-amino-3,6-dideoxyglucose
- Qui3N-ol: 3-amino-3,6-dideoxyglucitol
- Qui4N: 4-amino-4,6-dideoxyglucose
- QuiN: 2-amino-2,6-dideoxyglucose
- QuiN-ol: 2-amino-2,6-dideoxyglucitol
- QuiN4N: 2,4-diamino-2,4,6-trideoxyglucose (bacillosamine)
- Rha: 6-deoxymannose (rhamnose)
- Rha4N: 4-amino-4,6-dideoxymannose
- RhaN3N: 2,3-diamino-2,3,6-trideoxymannose
- Rib: ribose

- Rib-ol: ribitol
- Ser: serine
- SRCetLys: (2S,8R)-N(epsilon)-(1-carboxyethyl)lysine
- SS3,5HOHex: 3S,5S-dihydroxyhexanoic acid
- SSCetLys: (2S,8S)-N(epsilon)-(1-carboxyethyl)lysine
- Thr: threonine (2S,3R)
- Thre-ol: threitol
- Tyr: tyrosine
- Tyv: 3,6-dideoxy-D-arabino-hexose (tyvelose)
- Xul: threo-pent-2-ulose (xylulose)
- Xyl: xylose
- Yer: 3,6-dideoxy-4-C-[(S)-1-hydroxyethyl]-D-xylo-hexose (yersiniose)

References

(1) Kapaev, R. R.; Egorova, K. S.; Toukach, P. V. *J. Chem. Inf. Model.* **2014**, *54*, 2594-2611.