BMMC VII

Database-driven simulation of 1D and 2D NMR spectra of carbohydrates

Philip Toukach, Roman Kapaev

GODESS

(Glycan-Optimized Dual Empirical Spectrum Simulation)

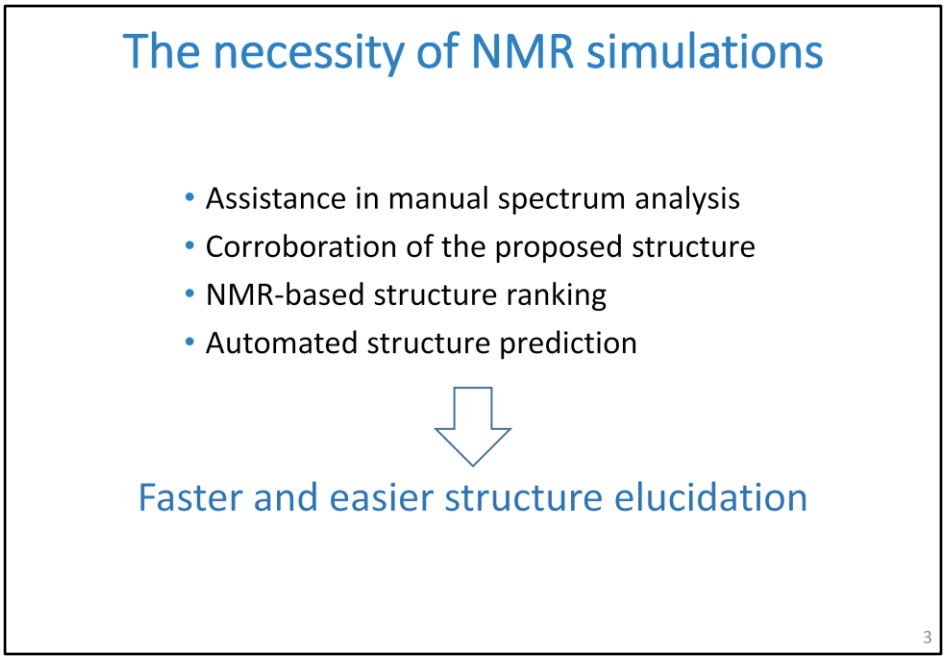
is a web-service for the NMR simulation of glycans and derivatives

This talk:

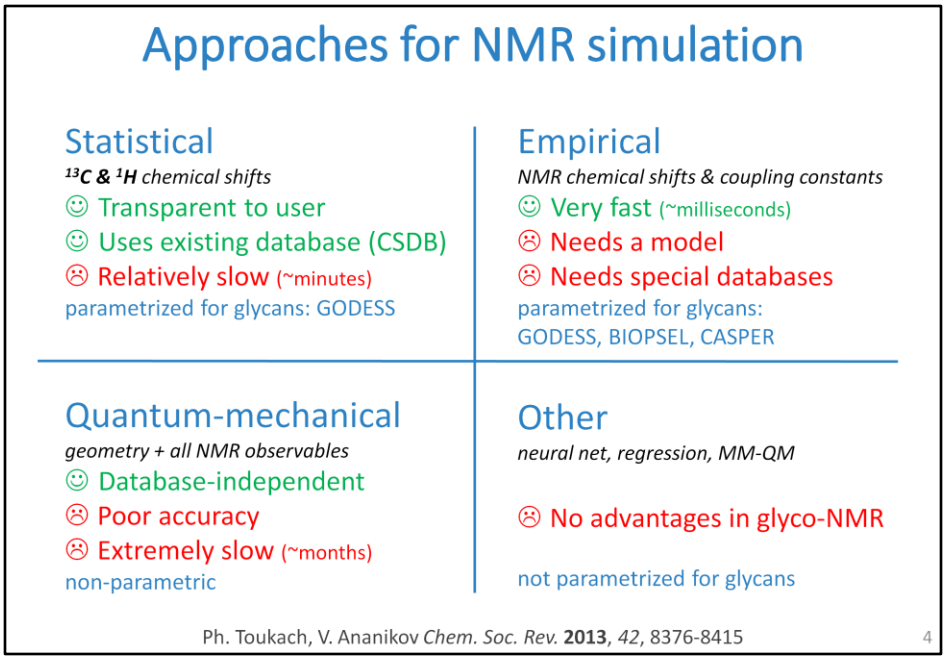
- Existing approaches
- GODESS: principles
- GODESS: how to use

2

I'd like to speak about new features of a web-service named GODESS, which is designed for the NMR simulation of glycopolymers and glycoconjugates.
I will tell you about the existing approaches, the principles lying behind our software, and the interface of the service.



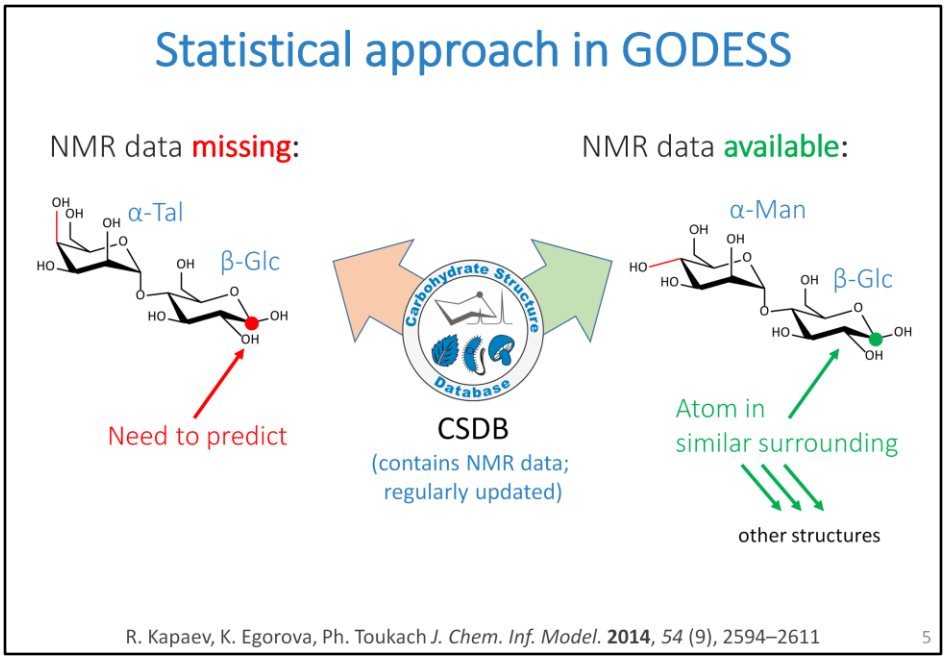
The NMR spectroscopy was proven as one of the most powerful tools for carbohydrate structural studies.
However, the interpretation of the NMR observables is still a tiresome task.
The progress in informatics revealed many opportunities to simplify this research.
In particular, NMR simulators allow the confirmation and ranking of structural hypotheses, and serve as a basis for solving the inverse problem – the software-assisted or even automated elucidation of structure from the experimental NMR data.



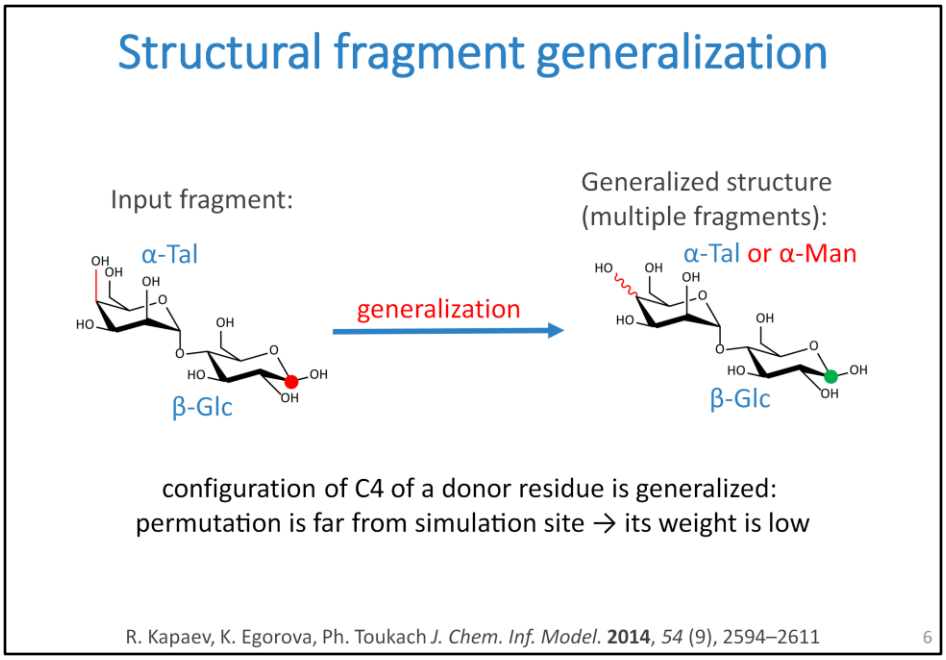
Several approaches exist for the NMR simulation of glycans.
Statistical approach relies on averaging of the database content, and provides most accurate results when the database is good.
The accuracy strongly depends on the algorithm used to find and compare molecular surrounding similar to that predicted.

Empirical approach is based on the incremental chemical shift calculation and detailed parametrization of structural descriptors.
Unlike the others it is very fast, allowing the bulk processing of millions of structures.

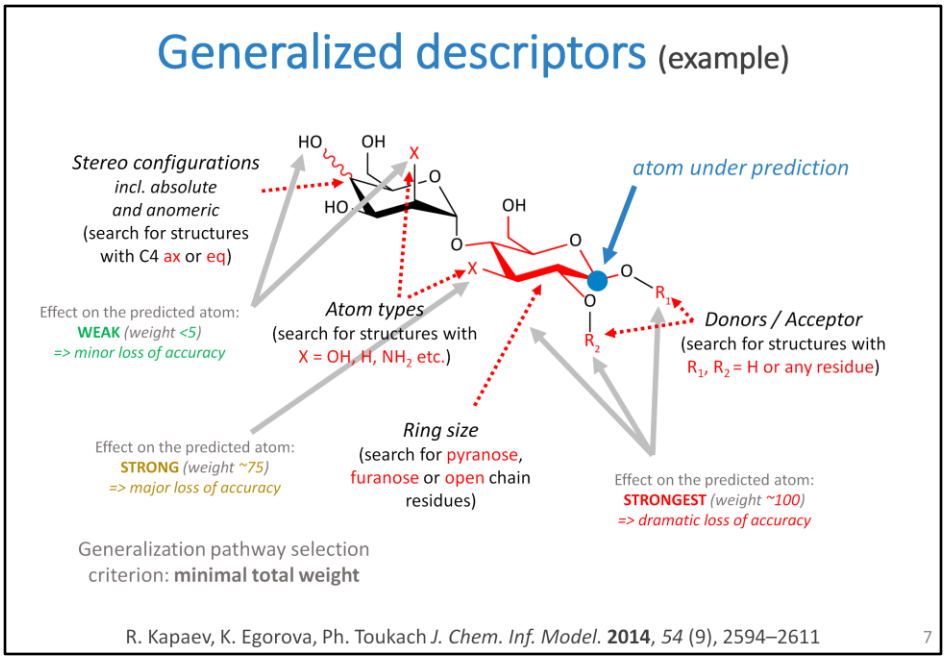
Quantum-chemical calculations don't restrict a compound class. However for glycans, even at high theory levels and in large basis sets, their accuracy, to say nothing of speed, still can not compete with statistical and empirical methods.
Our software combines the statistical and empirical approaches, supporting each other and parametrized for carbohydrate-containing molecules.



The basic principle of the statistical approach is averaging of published chemical shifts deposited in the database.
It uses the curated NMR data from the Carbohydrate Structure Database, currently containing about eight thousand spectra.
These chemical shifts are retrieved for the atoms in structural surrounding close to that in the predicted molecule.
For example, there are no structures with the fragment on the left that have the NMR data in the database.
But a minor structural permutation distant from the predicted atom gives us a popular fragment which we can use instead.

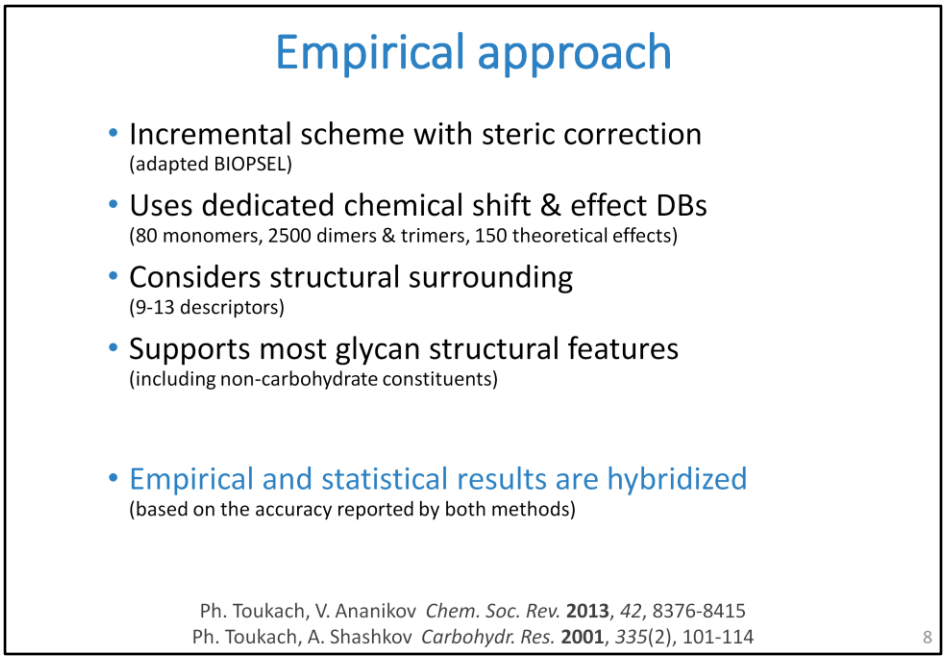


To find such similar fragments, generalizations of a structure are applied.
Every fragment is described by several descriptors.
A generalization is unrestricting a certain descriptor to widen the search scope.
In our example, stereo configuration of a substituent at position 4 of talose is generalized.
For every atom, our software scans the database for the structures containing the generalized fragments, starting from minor permutations, and generalizing more and more descriptors until enough structures are found.

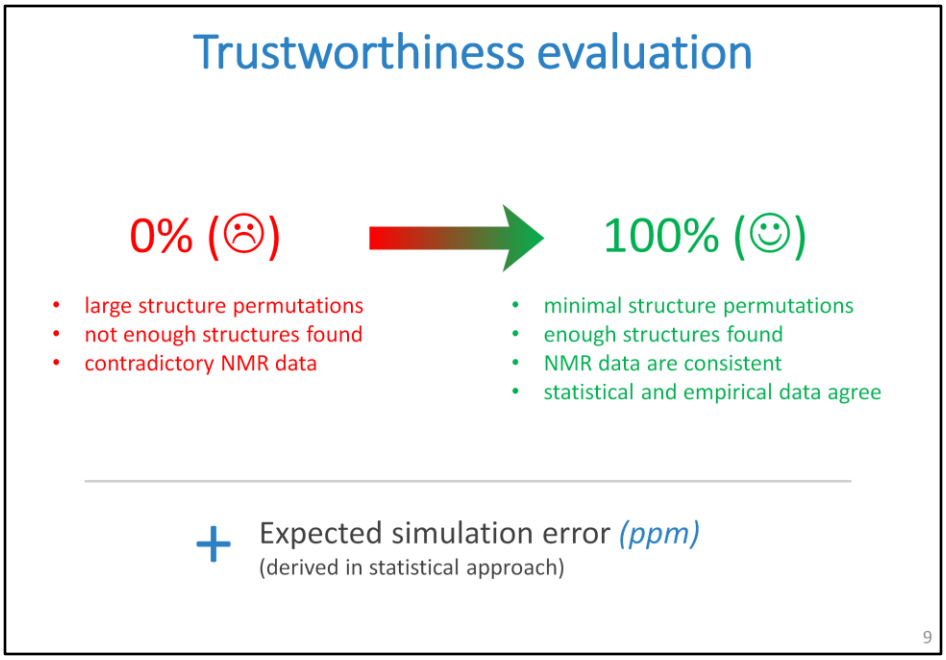


The generalizable descriptors include stereo configurations and substituent types at all atoms, ring sizes of residues, and presence and type of donors and acceptors.
The slide shows a few examples of such elementary generalizations.
A di- or trisaccharide fragment usually has two dozens of descriptors.
All of them are generalized independently, so there is a lot of ways to generalize the whole structure.
So we have to find out, which generalizations should be applied first and which should be applied if previous gave no result.

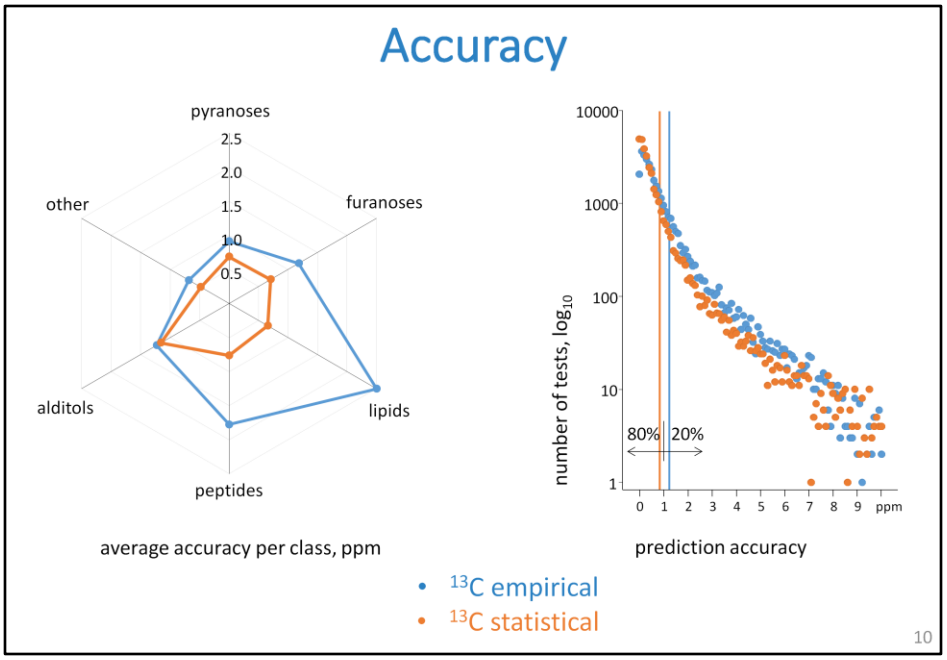
To do this, each generalization is assigned a weight reflecting how strong is its impact on the predicted atom.
These weights were obtained by the iterative fitting procedure.
They depend on the nature of a descriptor and on its distance from the prediction site.
For example, epimerization of a distant carbon has the low weight during the simulation of glucose C1, while changing the anomeric substituent has a dramatic effect.
To find the best-fitting fragments, the software minimizes the total weight of the applied generalizations.



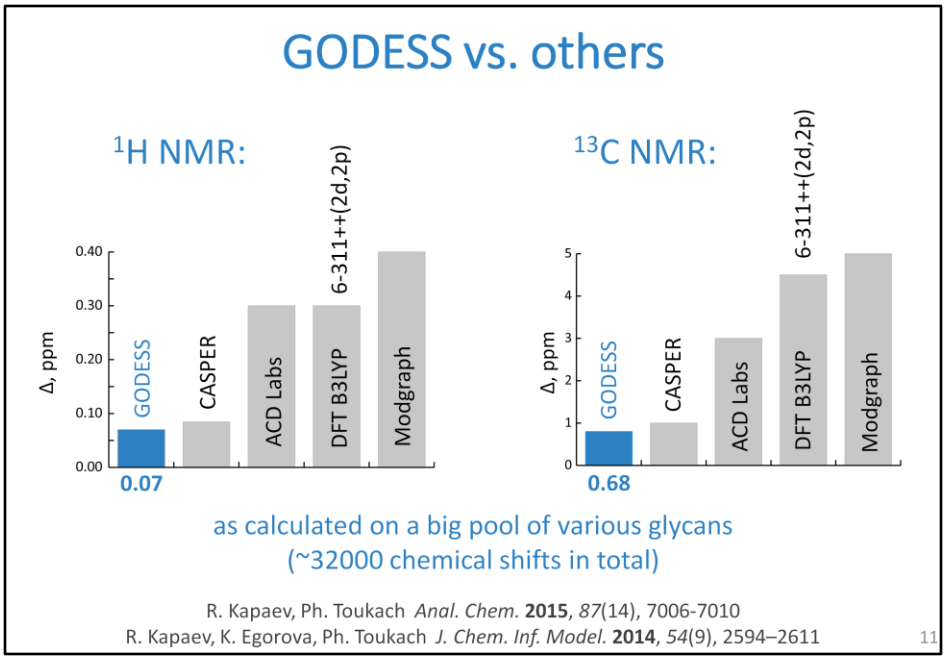
The slide lists the main features of the empirical carbon NMR simulation in GODESS.
It uses the incremental scheme and the dedicated databases of chemical shifts and substitution effects.
To find the best-fitting data, the fragments are compared using several structural descriptors.
This approach has been improving for years, but is not new and I will not focus on it.
Our software yields so called hybrid spectra.
It fuses the statistical and the empirical results basing on how good each atom was simulated by each of them.



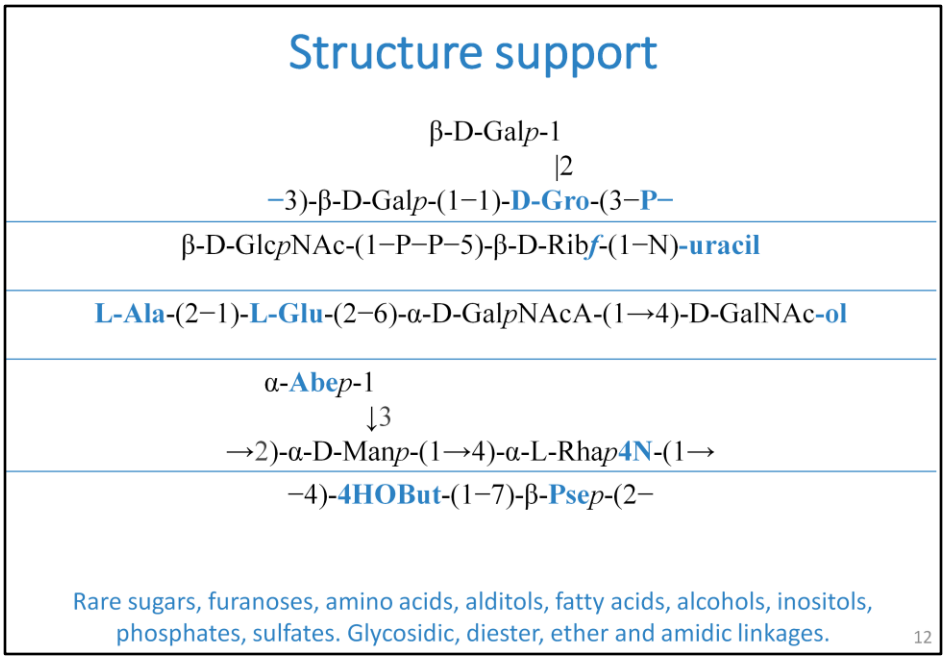
To do this, the trustworthiness is estimated for every prediction.
This metrics is based on how strong were structural permutations affording to find the matching fragments in the database.
It also depends on the amount and consistency of the NMR data found, and on agreement between the statistical and empirical predictions.
The expected simulation error is derived from trustworthiness.
This derivation uses the regression rules obtained from the fitting of predictions on bulk structure samplings.



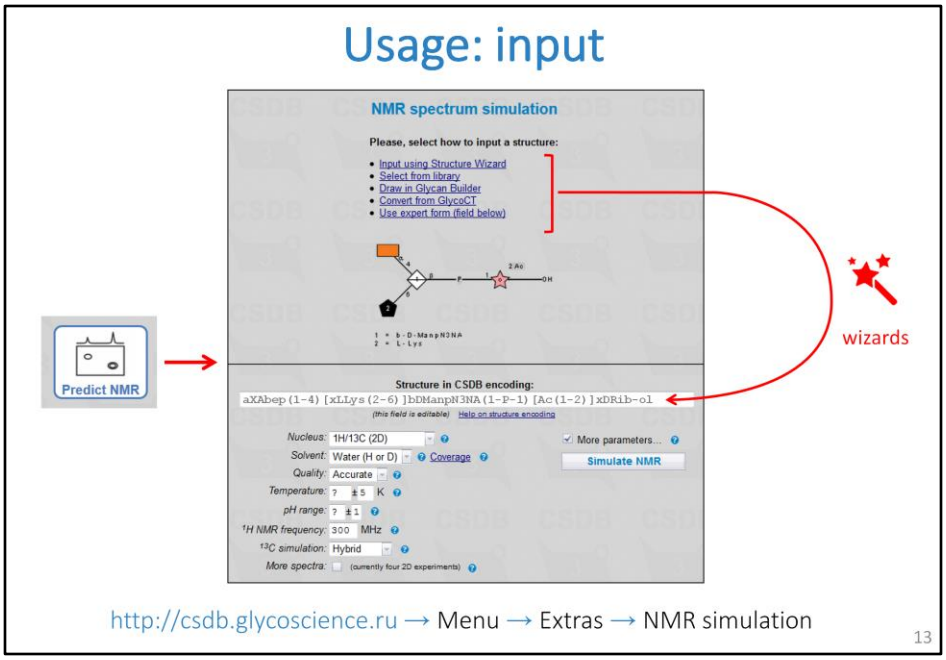
I guess you are interested, what this simulation error is.
On the right is the distribution of the carbon NMR simulation accuracy tested on all database content.
Usage of the data from the predicted structure itself was forbidden during the test.
We ran 30 000 tests and 80% of them resulted better than 1 ppm from the experiment.
On the left is the mean accuracy per residue class.
On average, empirical simulations perform worse, especially for poorly parametrized constituents, such as fatty acids or amino acids.



These graphs show the comparison of accuracy exhibited by various software on a big pool of carbohydrates and derivatives.
Our software outperformed the other methods, including other database-driven, neural net based, and quantum-mechanical calculations.
The speed of GODESS is comparable to that of other empirical approaches.
The accuracy of GODESS is compatible to that of CASPER software, however, the number of structural features supported by GODESS is wider.



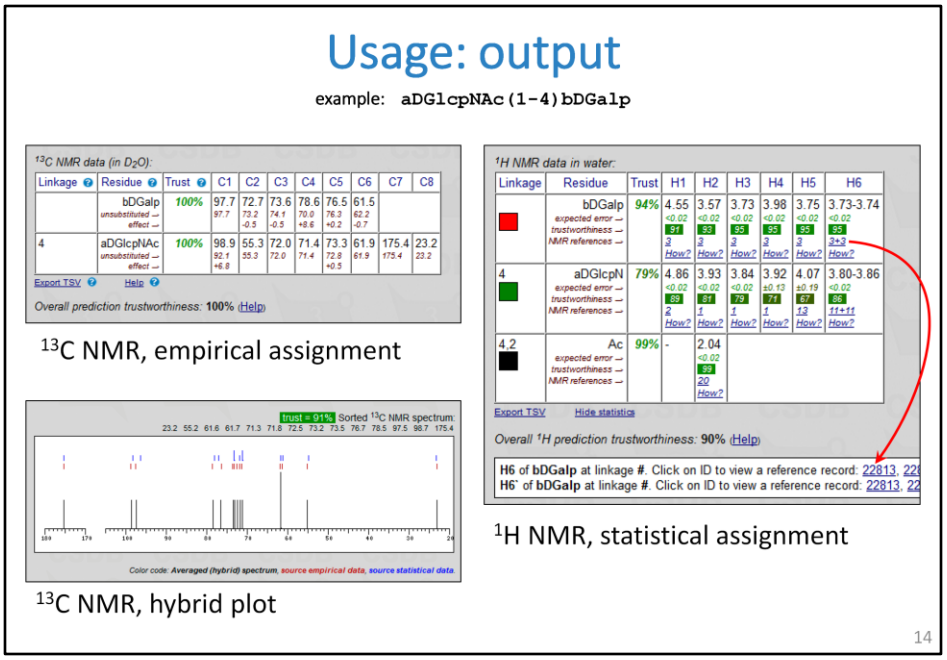
There are a few examples.
Unlike the other software, we can process most of the structural features found in natural carbohydrates, including furanoses, higher sugars, phosphates, alditols, glycopeptides, glycolipids and many other.
The structure scope is limited to oligomers and regular polymers built of residues linked by glycosidic, amidic, ester, or diester bonds.
For many saccharides, especially of bacterial origin, GODESS is the only software that supports their NMR simulation with acceptable accuracy.



Now I am going to present you a web-service built on top of this software.
It is implemented at the platform of the Carbohydrate Structure Database and it is freely available on the Internet.
To access the service press this button on the database main screen.

To run the NMR simulation, you are expected to input a primary glycan structure.
It can be done in several ways, including a wizard to assemble a structure or graphical GlycanBuilder tool.
Alternatively you can write a structure in one of the existing carbohydrate notations.

These optional parameters are related to the NMR simulation. The main of them are nucleus and solvent.
According to the database content, the best predictions are for water and pyridine solutions at room temperature.

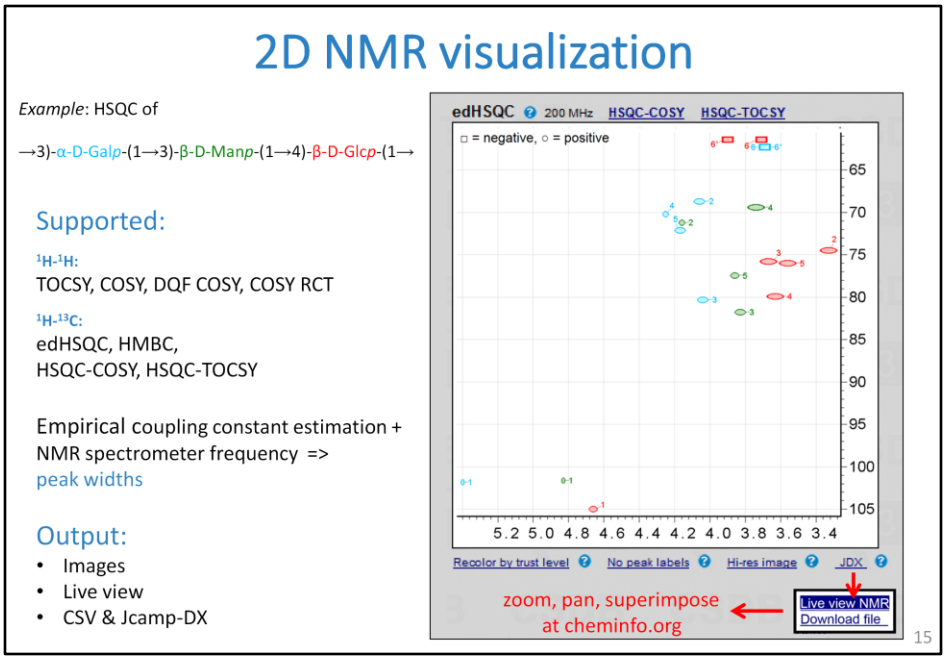


Besides the plots of spectra, results are presented as the signal assignment tables. These tables contain proton and carbon chemical shifts for every residue, trustworthiness values and expected simulation errors.

Clicking on the number of references in statistical assignment shows the list of database records used in the simulation, so you can track the sources of data to original publications.

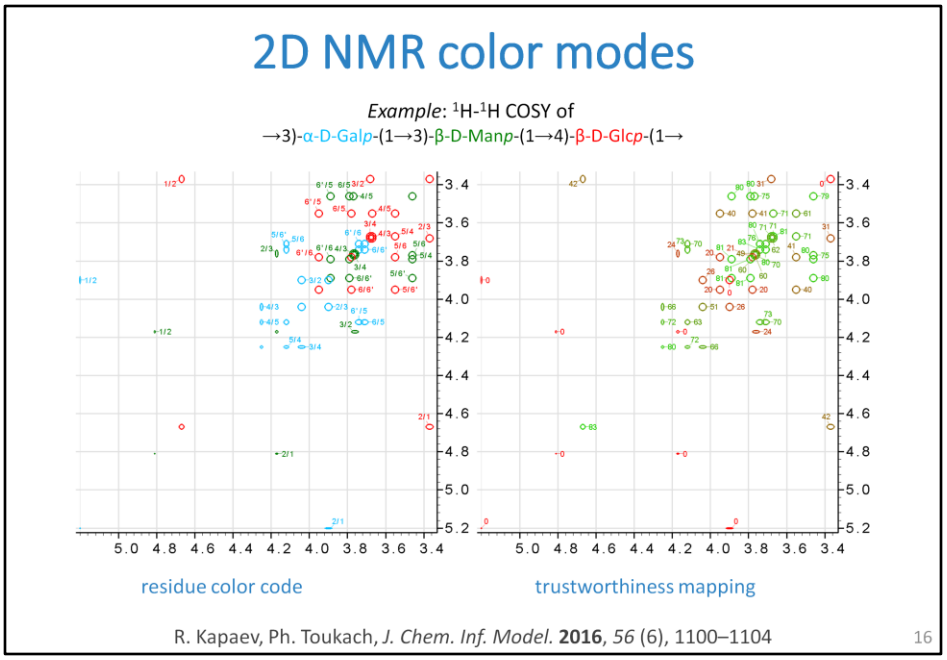
The generalization sequence is available by clicking on “How?”.

The result page allows you to refine the calculation by specifying a solvent, temperature, the NMR frequency and other parameters.



The service provides the plots of the most two-dimensional spin correlations commonly used in glycobiology, except NOESY.
This is the list of supported experiments.
Every peak is colored and labeled according either to its assignment or to its trustworthiness value.
We ask a user for a spectrometer frequency and empirically estimate proton coupling constants.
These values are used to predict the approximate cross-peak shapes.

This is an example of the HSQC spectrum.
It is so called edited experiment, so it has negative and positive signals, which are depicted by rectangles and ellipses, respectively.
The spectra are exportable in a few formats, which you can further process online or in the dedicated NMR software.



This is an example showing the color modes of two-dimensional spectra.

On the left, every peak is colored according to the residue, and labeled by atom numbers.
In COSY, there are only intra-residue correlations, so the labels have a single color, according to the residue color code in the assignment table.
In the spectra, containing inter-residue correlation, the signals can be bicolored.

On the right, colors and labels report the signal simulation trustworthiness, greener is better.

CSDB / GODESS team


Philip Toukach
R&D, CSDB


Roman Kapaev
R&D, GODESS


Ksenia Egorova
CSDB_GT, curation


Nadezhda Kalinchuk
annotation


Yuriy Knirel
curation

http://csdb.glycoscience.ru

soon to come… **2017** // automated structure elucidation

R. Kapaev, Ph. Toukach, *J. Chem. Inf. Model.* **2016**, *56* (6), 1100–1104 // 2D NMR
R. Kapaev, Ph. Toukach, *Anal. Chem.* **2015,** *87*(14), 7006-7010 // Statistical ^{13}C & ^{1}H
R. Kapaev, K. Egorova, Ph. Toukach, *J. Chem. Inf. Model.* **2014**, *54* (9), 2594–2611 // Statistical ^{13}C
Ph. Toukach, V. Ananikov *Chem. Soc. Rev.* **2013**, *42*, 8376-8415 // NMR simulation
Ph. Toukach, A. Shashkov *Carbohydr. Res.* **2001**, *335* (2), 101-114 // Empirical ^{13}C

17

I invite you to play with the NMR prediction using this address.
It hosts a pack of four databases, and three of them have the GODESS service implemented.
These are people involved in the project, and this is further reading.

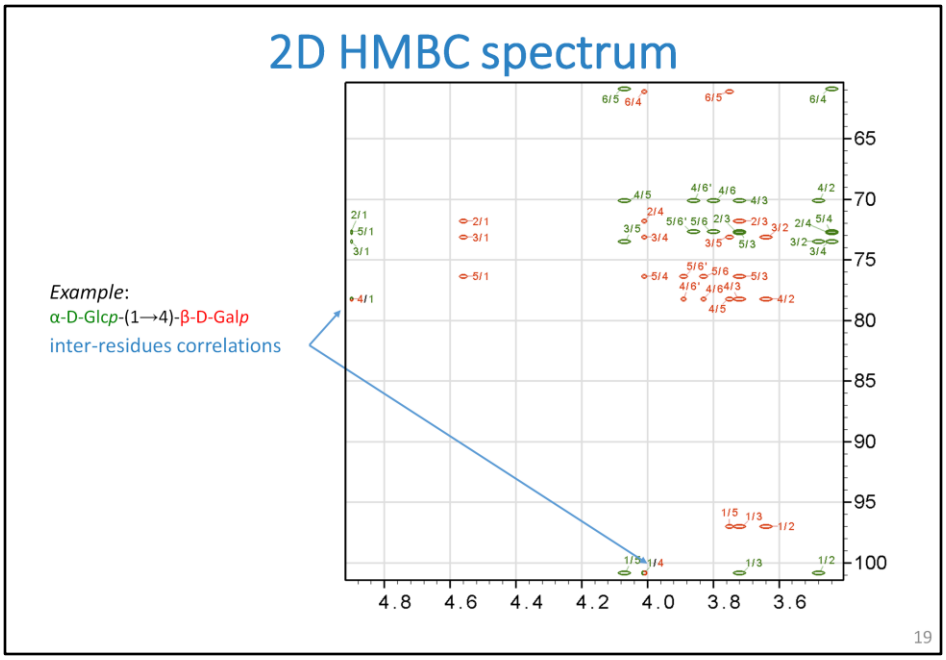
^{13}C NMR spectrum hybridization

example: **aDGlcpNAc(1-4)bDGalp**

- Analyzes trustworthiness reported by both approaches and their agreement
- Combines the most accurate data from both approaches
- Compares chemical shifts from both approaches

18

We have implemented a feature that generates a hybrid spectrum containing the most accurate data combined from the two approaches.
The combining is based on trustworthiness metrics from each method and applied for each atom individually.
It can be used as a summary and for comparison of methods.
The output assignment table contains deviation between the two approaches, and the schematic spectrum has signals from different approaches in different color.



This is an example of the HMBC spectrum.
Inter-residue correlations are shown in two colors.