***Global MIRAGE specific controlled vocabulary***

In the web form, the user can select predefined glycospecific MIRAGE information. In practise, it mostly relates to specific pretreatment of samples (exoglycosidases, permethylation etc) included in the MIRAGE sample preparation guidelines or in the MS section. A few resources cover this information such as GlycoSuiteDB[22] that is no longer available but now included in GlyConnect (https://glyconnect.expasy.org/) and GlycoDigest (https://glycoproteome.expasy.org/glycodigest/). The treatment list is available in Supplementary Spreadsheet. Being aware that current information about treatments in glycomics is evolving, UniCarb-DR will also accept user-defined treatments as submitted in the spreadsheet. This will expand the controlled vocabulary of specific treatments in glycomics as submission to UniCarb-DR progresses. At some stage, settling on a more rigorous maintenance of the treatment-controlled vocabulary may become necessary.

***Recording of MIRAGE MSn specific metadata***

The MIRAGE guidelines require that MS information for individual structures should be recorded for each structure. By implementing Glycoworkbench as part of a UniCarb-DR submission, the .gwp file format can be used in compliance with MIRAGE. In addition to structural recording and the inclusion of fragment lists with *m/z* (preferentially converted to centroid data) and intensities, Glycoworkbench automatically calculates theoretical masses based on a user-defined charge state, ion mode and derivatization. Glycoworkbench also has modules to calculate and match theoretical fragments with observed ones with a basic score. However, MIRAGE parameters such as “observed parent ion *m/z*”*,* “orthogonal methods” that have been used for identifying individual structures, “scoring” and “validation methods” of fragment data are not recorded in the .gwp file. We propose a model where this information can be included in the ‘Notes’ section in the Glycoworkbench file (Figure 3).

*Orthogonal methods*

In addition to MS, orthogonal methods are classically used in order to fully characterize a glycan structure. To account for this information we propose that the sample preparation methods defined above (Supplementary material) also serve as the controlled vocabulary for orthogonal validation of individual structures. Of course this list also needs to be expanded by input from the community and associated with other glycomic experimental data.

Since the assignment of structures is often based on previous knowledge about the samples, we propose to expand the orthogonal method list with four additional items; this is to capture various aspects of information not necessarily obtained by MS. These are:

1. Residues: Type of monosaccharide that constitutes the structure. MS is usually not sufficient for distinguishing between constituting isomeric monosaccharide units in a structure. A typical question is to establish if previous or biosynthetic knowledge was used in order to assign the monosaccharide composition. If for example, a Mannose is assigned to a certain position rather than the more generic Hexose, is it because of prior knowledge about the sample? This orthogonal method is captured as *Biosynthetic(residue).*
2. Primary Sequence: If the order of monosaccharide units in the structure is assumed based on previous or biosynthetic knowledge, i.e. if the primary sequence of an *N*-linked oligosaccharide core is put down as Hex-(Hex-)Hex-HexNAc-HexNAc, without evidence from MS, the use of this non MS generated additional information should be captured as *Biosynthetic(sequence).*
3. Linkage position: The linkage position in an assigned structure. For example, is Fuc assigned as Fuc**1-2**Gal based on prior or biosynthetic knowledge of blood group H that was shown to be present in the samples? This orthogonal method is captured as *Biosynthetic(linkage).*
4. Linkage configuration: The linkage configuration (usually α and β) in an assigned structure. For example, is Fuc assigned as Fuc**α**1-2Gal based on prior or biosynthetic knowledge of blood group H that was shown to be present in the sample? This orthogonal external information for assigning structures should be recorded as *Biosynthetic(config).*

If only MS is used to assign oligosaccharide structures, we believe that the default should be to include these 4 methods in the MIRAGE file. This is to acknowledge that MS is often not enough for a total characterization of a carbohydrate structure.

*Scoring of MSn fragmentation data*

The first MIRAGE guideline for MS was published in 2013 (23378518) and was based on state of the art glycomic analysis. At the time there were few e-tools used for the interpretation of MS data and scoring of the fragment spectra. Hence, the guidelines only requested the recording of the number of unmatched peaks for each spectrum. This information can be obtained using the peak-matching tool of Glycoworkbench, and could be captured for MIRAGE compliance from this file. However, since the publication of the guidelines, more sophisticated methods for measuring the quality of fragment ions have been developed. We propose to expand on the current guidelines to include this qualitative information. Rather than relying on the number for unmatched peaks, we record the actual scoring. For this we request that the report should include a defined vocabulary for the different types of scoring used in glycomics. Based on our experience in scoring spectra for structural assignment the following 4 items should be included in a MIRAGE report:

1. Scoring method: Answers the question: which method was used? Options would include *manual* interpretation or software aided interpretation such as *de-novo sequencing* methods, *spectral matching or matched/unmatched peaks.* For the scoring method to be relevant there is also a potential need to include:
	* 1. Errors of the mass allowed for parent ion and fragments.
		2. If (and which) database has been used for the scoring
		3. Restrictions i.e. in type of fragments searched, species exclusion or other exclusion from the database
2. Scoring algorithm: Answers the question: Is there a particular algorithm used to perform the scoring? For example, the *normalized dot product* is the most common algorithm for spectral matching*.*
3. Scoring result: Answers the question: what is the value (or values) output by the scoring method?
4. Scoring value format: The experience from proteomics is that a scoring result may not be a single value, so we propose that the format of the result is a string on values (text separated by comma), and that the scoring value format is a controlled vocabulary that defines the layout of the scoring result.

We have for several years defined and used internally a scoring named *UniCarb-DB triplet*. This score is based on the value of the normalized dot product and increased (i) if the matched structure is *identical* to the proposed structure, (ii) if it shares the same *sequence* or if it shares the same *composition.* Information about the rank of the proposed structure in the search result list is also considered. We introduce the triplet notation with an example: “0.99,identical,1” where *0.99* is the dot product score, *identical* indicates 100% similarity between the matched and proposed structures, and *1* indicates the rank of the right answer in the search result list. Other values for the first item can be *no-match*. The scoring value format of *UniCarb-DB triplets* should be defined in the controlled vocabulary for scoring.

*Validation of structures*

The objective of the validation is to give an overview of the structural features that could be determined by MS vs. other information. MS fragmentation is expected to provide primary sequence information. However, we need to use orthogonal methods to determine a full structure and connect it with biological function. The MIRAGE guidelines require information on how a structure was validated. However, the means for how to do so are not defined. Options should cover *manually* or *automatically,* but also *other (*eg false discovery rate). Furthermore, information about the *MSn* level used for validation and their corresponding results are informative. The validation result format should be similar to that of the scoring, i.e. recorded as a string of values separated by commas.

Several features of a structure need to be validated including monosaccharide composition (C), primary sequence (S), linkage position (L), and linkage configuration (C). We suggest the definition of a format notation, and to set the default as the *manual* *CSLC*-format to capture how conclusive the MS and fragment data are for the structure that is proposed. If it is found that the fragment data fully supports each of these 4 items (composition, sequence, linkage and configuration) for a fully assigned structure containing monosaccharide speciation, linkages and configuration, the validation results should be 1,1,1,1. If it is found that nothing is substantiated the results instead should be 0,0,0,0. For easy manual evaluation we propose the following reasoning with a hexasaccharide as an example:

1. Monosaccharide composition (C): The mass of an oligosaccharide provides information about the composition, but is the MS itself conclusive to identify isomeric monosaccharide units? With a manual validation it is always a matter for the researcher to judge, but we can try to provide some guidelines based on our own experience. For a hexasaccharide consisting only of 3 Hexoses and 3 *N*-acetylhexosamines, it is unlikely that only MS and MS2 data will provide information about the type of Hex or HexNAc isomer. Hence the first C value in the validation results should be “0” if the proposed structure suggests specific monosaccharide units for Hex and HexNAc (like Man and GlcNAc). Another example is a hexasaccharide with a composition of Hex2HexNAc2Fuc1NeuAc1. If this is structure was found in previously referenced source, where both fucose and N-acetylneuraminic acid are known to be present, and fragmentation data provides clear evidence that masses corresponding to Fuc and NeuAc residues, one could argue that presence of 2 of the 6 monosaccharides has been validated, because of the lack of isomeric residues in the source. Hence, the validation result should be 2/6 = 0.33 if the proposed structure also contains speciation of Hex (e g Man and/or Gal) and HexNAc (e g GlcNAc) units.
2. Primary sequence (S): How well does the fragmentation data support the proposed sequence? For a hexasaccharide there are 5 linkages that need to be identified. A quick way to validate this is to check if there is any evidence for all glycosidic fragments in the spectra (validation result =1). If one fragment is lacking but still recorded (‘guessed’) in the proposed structure, the primary sequence (S) validation value should be 4/5 = 0.8. In order to perform this manually, we propose the use both single and internal glycosidic fragment assignments. Note that only because all glycosidic linkages are detected, the sequence may not be conclusive and other sequences may also fit the spectra.
3. Linkage position (L): Is there evidence for a specific fragmentation of linkage position? In a hexasaccharide, there are 5 linkage positions that should be determined (assuming the permanence of a link via the anomeric C-1 carbon). If all of the linkages are assigned in the proposed structure but linkage specific fragmentation evidence (usually cross ring fragmentation) is lacking for one of them, the linkage position validation should be 4/5 = 0.8. Note that assignment of cross-ring fragments may be unequivocal.
4. Linkage configuration (C): Usually MS is not the ultimate method to determine α or β configuration, so if these are recorded in all the linkages for a proposed structure the linkage configuration default validation result should be “0”. One could argue that MS may contain this information if for instance the fragmentation (fragment ions and/or intensity) is found to be different for an α or β isomer. This could be the case for instance using MSn methodology [29] or configuration specific fragmentation using ion mobility[40].

It should be pointed out that using this format, orthodox reporting of structures from fragment data provided in the form of numbers of Hex and HexNAc and primary sequence data (all glycosidic fragments) with unknown linkage positions and configurations, are validated with a score of 1,1,1,1. The same structure, recorded instead with Man, Gal, GlcNAc and GalNAc residues and fragments covering all glycosidic linkages, but recorded with linkage position and configuration without MS evidence, will have a validation score of 0,1,0,0. Hence, the validation is not only capturing the quality of the MS data, but also how orthogonal was utilized for interpretation. Other ways of validation of structures for glycomic analysis will inevitably be developed. We assume that our implemented system for MIRAGE recording is flexible enough to incorporate these.