



HAL
open science

CBMcarb-DB: Interface of the Three-Dimensional Landscape of Carbohydrate-Binding Modules

D O Ribeiro, F Bonnardel, A S Palma, A L M Carvalho, Serge Perez

► **To cite this version:**

D O Ribeiro, F Bonnardel, A S Palma, A L M Carvalho, Serge Perez. CBMcarb-DB: Interface of the Three-Dimensional Landscape of Carbohydrate-Binding Modules. The Royal Society of Chemistry. Carbohydr. Chemistry, 46, pp.1-22, 2024, 978-1-83767-217-2. hal-04807891

HAL Id: hal-04807891

<https://hal.science/hal-04807891v1>

Submitted on 27 Nov 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 **Chapter X**

2 **CBMcarb-DB: Interface of the Three-Dimensional Landscape of Carbohydrate-** 3 **Binding Modules**

4
5 D.O. Ribeiro^{a,b}, F. Bonnardel^c, A.S. Palma^{a,b,d}, A.L.M. Carvalho^{a,b} and S. Perez^{c*}.

6 ^aAssociate Laboratory i4HB – Institute for Health and Bioeconomy, School of Science
7 and Technology, Universidade NOVA de Lisboa, 2829-516 Caparica, Portugal,

8 ^bUCIBIO, Department of Chemistry, School of Science and Technology, Universidade
9 NOVA de Lisboa, 2829-516 Caparica, Portugal, ^cCentre de Recherche sur les
10 Macromolécules Végétales (CERMAV), CNRS, Université Grenoble-Alpes, Grenoble,
11 France.

12 *Corresponding email address: spsergeperez@gmail.com;
13 serge.perez@cermav.cnrs.fr

14 15 **ABSTRACT**

16 Carbohydrate-Binding-Modules (CBMs) are discrete auxiliary protein modules with a
17 non-catalytic carbohydrate-binding function and that exhibit a great diversity of binding
18 specificities. CBMcarb-DB is a curated database that classifies the three-dimensional
19 structures of CBM-carbohydrate complexes determined by single-crystal X-ray
20 diffraction methods and solution NMR spectroscopy. We designed the database
21 architecture and the navigation tools to query the database with the Protein Data Bank
22 (PDB), UniProtKB, and GlyTouCan (universal glycan repository) identifiers. Special
23 attention was devoted to describing the bound glycans using simple graphical

24 representation and numerical format for cross-referencing to other glycosciences and
25 functional data databases. CBMcarb-DB provides detailed information on CBMs and
26 their bound oligosaccharides and features their interactions using several open-
27 access applications. We also describe how the curated information provided by
28 CBMcarb-DB can be integrated with AI algorithms of 3D structure prediction,
29 facilitating structure-function studies. Also in this chapter, we open to the exciting
30 convergence of CBMcarb-DB with the glycan array repositories, which serve as
31 valuable resources for investigating the specific binding interactions between glycans
32 and various biomolecular targets. The interaction of the two fields represents a
33 significant milestone in glycosciences. CBMcarb-DB is freely available at
34 <https://cbmdb.glycopedia.eu/> and <https://cbmcarb.webhost.fct.unl.pt>.

35 1. Introduction

36 Carbohydrate-Binding Modules (CBMs) are a class of carbohydrate-binding proteins,
37 defined as non-catalytic protein domains with amino acid sequences ranging from 30
38 to 200 amino acids^{1,2}. Currently, the amino acid sequence similarity dictates the
39 classification of CBMs into different families³. The number of newly identified CBM
40 sequences with putative carbohydrate-binding function is growing fast due to the
41 exponential increase of sequence information derived from microbial genomics,
42 metagenomics and transcriptomics data^{4,5}. CBMs are typically associated with
43 modular enzymes but can also be isolated from a carbohydrate-active enzyme
44 polypeptide chain or associated with other non-enzymatic proteins^{4,6,7}. Several CBMs
45 are also classified as lectins, such as *Ricinus communis* ricin toxin B chain in CBM
46 family 13 and the human lectin malectin in family CBM57⁸. Members from CBM family
47 50 are also known as LysM domains, found in peptidoglycan- and chitin-binding
48 proteins⁹. CBMs have been described to exert critical functions in enhancing enzyme
49 catalytic efficiency (e.g. substrate targeting), as carbohydrate-sensing domains or as
50 adhesion molecules (e.g. mediating bacterial adhesion to the host cell surface). These
51 modules show diverse carbohydrate-binding specificities between families and even
52 within the same family. Several characterised CBMs are associated with bacterial and
53 fungal systems that recognise non-crystalline cellulose and chitin, insoluble storage
54 polysaccharides, such as starch and glycogen, and different hemicellulosic substrates,
55 such as xylans, mannans, galactans, soluble α - and β -glucans³. More recently,
56 attention is also drawn to CBMs of systems from bacterial pathogens and commensal
57 bacteria of the human microbiome, which can recognise the complex glycosylation of
58 mucin glycoproteins covering the epithelial gastrointestinal layer¹⁰ (see Chapter XX of
59 this volume for a discussion on CBMs from family 32 and binding to mammalian-type

84 or pocket-shaped surface that matches the shape and size of the ligand. Some CBMs
85 can also disrupt the ordered structure of the polysaccharide by inserting tryptophan
86 residues into the substrate, creating additional binding sites for the catalytic domain.
87 The specificity and affinity of CBMs for their ligands depend on the number, type, and
88 arrangement of the amino acids involved in the binding site. Different CBMs may have
89 different mechanisms of ligand recognition depending on their structure, function, and
90 evolution.

91 Effective bioinformatic tools are nowadays contributing to an increasing number
92 of annotated CBMs which, together with the increasing assignment of carbohydrate-
93 binding specificities, makes CBMs an excellent model for studying the protein-
94 carbohydrate recognition event. CBMs comprise also valuable candidates for various
95 biotechnological applications, as proven by the generation of engineered CBMs with
96 new and diverse functional properties¹⁷.

97 As the development of databases is providing valuable information on the proteins
98 involved in carbohydrate recognition, there is a need to establish an integrated web-
99 based platform that brings together structural and functional knowledge on CBMs and
100 their carbohydrate ligands. CBMcarb-DB (<https://cbmdb.glycopedia.eu> or
101 <https://cbmcarb.webhost.fct.unl.pt/>) arises as a novel database dedicated to displaying
102 and analysing the 3D structures of CBM-carbohydrate complexes, providing curated
103 structural data about CBM-carbohydrate binding interactions, carbohydrate
104 conformation, and functional information on binding specificity.

105

106 2. The CBMcarb-DB Database Construction and Utilisation

107 2.1. Database Construction and Data Curation

108 To populate CBMcarb-DB, an initial list of 638 entries containing the PDB¹⁸ IDs of
109 CBM-carbohydrate structures was extracted from the CAZy database³ for all families
110 assigned, using a web scraper program written specifically for this purpose. This
111 scraper ran through the publicly available pages from the CAZy website and extracted
112 all the relevant information for our purposes directly from the page's HTML. This list
113 was checked and filtered for possible duplicated IDs, resulting in 520 unique PDB
114 entries. Each structure was then manually inspected using Mol* viewer¹⁹, through
115 RCSB PDB²⁰ 3D View, to confirm the presence of the CBM-carbohydrate complex and
116 the relevance of the protein-ligand interaction. In cases where multiple proteins were
117 present in the same structure, correctly identifying the CBM module and its respective
118 ligand was done by checking the original publication and/or running an alignment
119 search using InterPro²¹. When distinct CBMs in complex with ligands from the same
120 or different families occurred in the same structure, a separate entry was created for
121 each CBM-carbohydrate complex. The corresponding information was annotated for
122 each entry manually, cross-referencing with CAZy, PDBsum²² and Yorodumi²³. This
123 analysis resulted in a final list of 362 curated entries that compose the original
124 CBMcarb-DB. The upkeep of CBMcarb-DB will be continuous over time, as the
125 information sources will be searched for new data every two weeks. Currently,
126 information associated with each database entity is added manually. It allows for
127 proper curation and annotation at the expense of a time lag between the date of
128 deposition and the date of release in the database. Later, a more automated strategy
129 for the data input will be explored.

154 from family 20 (37), 35 (25), 32 (23), 6 (18), 40 (16), 48 (16), and 50 (10). Glucose
155 (134) and galactose (76) are the monosaccharide type mostly found in these
156 structures, followed by *N*-acetyl-glucosamine (31), xylose (28), neuraminic acid (19),
157 mannose (16) and *N*-acetyl-galactosamine (14). The size distribution of
158 oligosaccharides complexed with a CBM varies from monosaccharides to
159 octasaccharides, with a higher representation of those with a degree of polymerisation
160 (DP) of DP2 (99) and then DP1(89), decreasing then from DP3 to DP8 (DP3 (58), DP4
161 (45), DP5 (28), DP6 (16), DP7 (14) and DP8 (3)). These numbers may reflect the
162 biological and biotechnological interest of the community in investigating these CBMs,
163 as well as the difficulty in obtaining crystals and high-resolution structures of proteins
164 in complex with larger ligands.

165 

166 **2.3. Description of the Search Interface**

167 The database can be searched and explored with an advanced search tool handling
168 a range of criteria.

169 *Searching by CBM*

170 The search can be performed by the protein types included in each CBM family by
171 clicking on the desired panels to expand and access each corresponding entry or
172 clicking on the blue search button to explore all the associated entries (see Figure 4).

173 

174 *Searching by fields*

175 The database can be searched and explored with an advanced search tool handling
176 a range of criteria:

- 177 • *pdb*: PDB ID.
- 178 • *cbm family*: Family classification as annotated in the CAZy database.
- 179 • *protein name*: Attributed name of the protein that includes the CBM module.
- 180 • *organism*: Origin of the protein.
- 181 • *resolution*: Value in Angstrom of the data resolution.
- 182 • *carb pdb*: Type of monosaccharide (3-character PDB codes).
- 183 • *carb iupac*: Oligosaccharide nomenclature according to the IUPAC condensed
- 184 form.
- 185 • *carb length*: Degree of polymerisation of the oligosaccharide.

186 This will give access to the search by fields (see Figure 5).

187 

188 Clicking on any of the panels will display the existing content. The user is invited
189 to select the entry of his/her choice or to search a desired content by typing on the
190 corresponding panel, for example, typing a PDB ID in the *pdb* panel, a type of
191 carbohydrate monomer such as "Glc" in the *carb iupac* panel, or the degree of
192 polymerisation in the *carb length* panel. It is a multi-field selection process aiming to
193 search the structure and direct the user to access the desired target.

194 The selected criteria will lead to the corresponding entries (see Figure 6).

195 

196 Clicking on 'visualise the 3D structure' yields the detailed page (see Figure 7).

197 

198 On this page, an interactive view of the CBM-oligosaccharide three-
199 dimensional structure is displayed. The LigPlot+²⁸ diagram, oligosaccharide sequence

200 according to the SNFG symbol representation, and the oligosaccharide and the CBM
201 structures are also shown, and the user can download the respective images by
202 clicking on them. For CBM complexes with ligands over three monosaccharides, an
203 interactive view of the oligosaccharide is also displayed, and the user can download
204 the corresponding PDB file by clicking on the *Oligo PDB file* link.

205 Links are also available to access other related information and resources:
206 information about the original publication through PubMed or DOI; the protein
207 sequence through UniProt²⁹; protein and carbohydrate structure and interaction using
208 the various PDB databases sites (RCSG PDB, PDBe PDB³⁰, and PDBsum,
209 SwissModel³¹, PLIP³² and PISA^{33,34}).

210 CBMcarb-DB cross-references to several other databases that rely on various
211 strategies for visualising the interaction between carbohydrate ligands and their
212 protein environment (see Figure 7).

213 Additional information on the oligosaccharide and protein interactions can be
214 obtained from the ligand-protein interaction profiler (PLIP) server³². The NGL viewer³⁵
215 adapted to SwissModel, displays the interactions identified by the PLIP application
216 that calculates and displays atomic-level interactions (hydrogen bonds, hydrophobic,
217 water bridge, etc.) occurring between carbohydrates and proteins. The specific
218 features of the glycans interacting with the surrounding amino acid residues and
219 possible metal ions are shown in 3D. The SwissModel application provides direct
220 access to the PDBsum deployed by the EMBL-EBI, CATH³⁶, and PLIP.

221 A cross-link to the PISA application enables the exploration of quaternary structure
222 formation and stability. The potential contribution of the carbohydrate to the formation
223 of quaternary macromolecular complexes requires the evaluation of energetic stability.

224 The structural information relates to the interfaces between the macromolecular
225 entities, the individual monomers, and the resulting assemblies, from which complex
226 stability can be assessed or predicted.

227 **3. Utilisation of CBMcarb-DB for Analysis and Predictions of CBMs and** 228 **Oligosaccharide structures**

229 The interest in carbohydrate-binding modules materialises through several databases,
230 such as CAZy, dbCAN3, and CBMDB. The CAZy database covers carbohydrate-
231 active enzymes' biochemical knowledge, including CBMs. CAZy revolves around
232 amino acid sequence annotation and has grown closely related to NCBI genomes,
233 Swiss-Prot, and UniProt, allowing the unambiguous characterisation of CAZymes via
234 sequence accession numbers⁸. dbCAN3 provides search tools and automated
235 CAZyme annotations for newly sequenced genomes³⁷.

236 The reported CBMDB integrates more direct data related to CBMs, including
237 sequence similarity searches, pairwise alignment, multiple sequence alignment,
238 structure similarity searches, and phylogenetic visualisation³⁸. Regarding the
239 prediction of binding sites, CBMDB classifies sequence-based methods to use the
240 information derived from the amino acid sequence of a protein, such as conservation,
241 similarity, or motifs.

242 While being more modest regarding the number of assembled information,
243 CBMcarb-DB offers an insight inspection of detailed three-dimensional features
244 relevant to the CBMs and interacting carbohydrates. Besides, it provides a structure-
245 based method, where the information derived from the 3D structure of a protein, such
246 as a shape, surface, or energy, is used to predict the binding sites. CBMcarb-DB
247 provides a unique platform to analyse, predict, and interrogate critical structural
248 features of carbohydrate-protein complexes.

249 Within the overview offered by the CBMcarb-DB, the absence of any X-ray
250 crystallography complexes displaying the type A mode of interaction is a striking
251 feature. Such a mode of interaction involves a planar hydrophobic surface of the CBM,
252 decorated by aromatic residues that interact with flat crystalline polysaccharides, such
253 as chitin or cellulose (see Figure 1). It must be rationalised that the complexity of such
254 interactions is not favourable for conventional X-ray crystallography investigation, both
255 from the standpoint of co-crystallisation and 3D structure solution. Therefore, the
256 following elements of analysis and discussion will be restricted to the type B and C
257 modes of interaction, with a particular emphasis on type B.

258 **3.1. *Deciphering the anatomy of the amino acid-carbohydrate interactions***

259 CBMcarb-DB provides information on the specific carbohydrate ligands bound by
260 determinant amino acid residues of CBMs. The database enables researchers to
261 analyse the interactions between CBMs and carbohydrates, including hydrogen
262 bonding (water-mediated), van der Waals contacts, and hydrophobic interactions.
263 Understanding the binding interactions is crucial for elucidating the molecular
264 recognition and specificity of CBMs. The most critical characteristic driving force
265 mediating protein-carbohydrate interactions is the position and orientation of aromatic
266 amino acid residues (Trp, Tyr and sometimes Phe) within the binding site. These
267 essential planar residues provide a hydrophobic platform for the planar face of sugar
268 rings, an interaction resembling hydrophobic stacking interactions.

269 **3.2. *Integrating AI 3D structure predictions and curated information from*** 270 ***CBMcarb-DB in structure-function studies***

271 The construction of CBMcarb-DB follows the generation of findable, accessible,
272 interoperable and reusable (FAIR) biological data. This is an indispensable
273 organisation to feed and train Machine Learning-based applications to predict different

274 levels of structural organisation and characterise the unique features of the recognition
275 and binding of carbohydrate structures by dedicated proteins, such as CBMs. They
276 play an essential role in enhancing the catalytic efficiency of polysaccharide-degrading
277 enzymes by promoting their proximity and affinity to the substrate and facilitating the
278 enzyme activity. This is particularly notorious for insoluble substrates.

279 CBMcarb-DB facilitates studying conformational changes that occur in CBMs
280 upon binding to carbohydrates. The comparison of the structures of apo (ligand-free)
281 and holo (ligand-bound) CBM offers a way to identify structural rearrangements,
282 including changes in secondary structure elements, loop movements, or global domain
283 motions. Such analyses provide insights into the dynamic nature of CBMs and their
284 adaptation to carbohydrate recognition. Integrating the curated data in the database
285 with other bioinformatics tools provides a thorough understanding of CBM structure-
286 function relationships and the molecular mechanisms underlying carbohydrate
287 recognition. Recent developments in innovative Artificial Intelligence (AI) algorithms,
288 such as AlphaFold³⁹, RoseTTAFold⁴⁰, RaptorX⁴¹ and others, open the route to
289 improving the accuracy and efficiency of protein structure prediction.

290 Comparing known protein structures with AI-predicted models offers several
291 advantages, as illustrated through the success of AlphaFold in the CASP (Critical
292 Assessment of Structure Prediction) competitions⁴². AlphaFold can predict protein
293 structures relatively quickly and for a wide range of proteins across diverse proteomes,
294 making it helpful in analysing many proteins in a short period, including those that are
295 difficult to study experimentally. Such progresses cover the protein-carbohydrate
296 complexes for which the definition of true ligand specificity is challenging. Even when
297 the true (natural) ligand is known, these complexes are often difficult to crystallise. AI
298 structure-prediction tools have shown impressive accuracy in predicting protein 3D

299 structures, particularly for proteins with no similar structures in the PDB. Among many
300 other advantages, comparing AlphaFold predictions with known structures can
301 highlight conformational differences, such as loop movements or side-chain
302 orientations. These variations can provide valuable insights not only into protein
303 flexibility and dynamics but also into the mechanisms of ligand recognition.

304 While considering the direct interaction between CBMcarb-DB curated
305 information and AI algorithms, several approaches are available:

306 i) AI structure predictions can provide high-quality 3D structures for CBMs that
307 might not have been experimentally determined. They help visualise the CBMs'
308 binding sites and critical residues involved in carbohydrate recognition, enhancing
309 structural insight into the particular CBM.

310 ii) The predicted 3D structures can be superposed with known carbohydrate-
311 bound CBM structures within the same family in CBMcarb-DB (e.g. the structural
312 comparison of the AI-predicted structure of CBM11 from *Microbacterium*
313 *hydrocarbonoxydans* with the CBM11 crystal structure from *Clostridium thermocellum*
314 illustrated in Figure 8 and described below). It allows the mapping of ligand-binding
315 sites and provides insights into the specific interactions between CBMs and
316 carbohydrates.

317 iii) By comparing AI-predicted structures with known CBM-carbohydrate
318 complexes in CBMcarb-DB, researchers can analyse the structural features that
319 dictate ligand specificity and affinity. This information is valuable for understanding
320 how CBMs recognise different carbohydrate ligands.

321 iv) AI predictions of apo and holo CBMs (ligand-bound) can be compared to
322 identify conformational changes upon carbohydrate binding. Integrating this

323 information with CBMcarb-DB data can shed light on the dynamic behaviour of CBMs
324 during ligand recognition.

325 v) AI predictions can be cross-validated with the curated CBM structures in
326 CBMcarb-DB to assess the accuracy and reliability of the predictions. Any
327 discrepancies can be addressed, and the predicted models can be refined using
328 experimental data from CBMcarb-DB.

329 vi) CBMcarb-DB may provide additional information about the presence of
330 catalytic domains or other functional modules within CBMs (modular organisation of
331 the genome). Integrating this information with AI predictions aids in understanding the
332 structural basis of CBMs' multifunctional roles.

333 vii) AI predictions can be used to probe the binding of CBMs with carbohydrates
334 that have not been previously characterised. It helps identify potential novel
335 interactions and expands the knowledge of CBM-carbohydrate recognition.

336 viii) Combining AI predictions with curated CBM structures can guide the
337 rational design of mutations in CBMs to investigate the impact of specific amino acids
338 on ligand binding or functional properties.

339 The case study of family 11 CBM highlights the insights gained from such
340 analyses and their implications for understanding CBM function and carbohydrate
341 recognition. CBMcarb-DB reports the two crystal structures of the CBM11 from
342 *Clostridium thermocellum* (CtCBM11) determined in complex with mixed-linked
343 oligosaccharides featuring a β 1,3-linkage at the reducing end (tetrasaccharide
344 Glc β 1,4Glc β 1,4Glc β 1,3Glc) and both at the reducing end and at an internal position
345 (hexasaccharide Glc β 1,4Glc β 1,3Glc β 1,4Glc β 1,4Glc β 1,3Glc), as informed by results
346 from glycan microarrays¹³.

347 The following illustrates the usage of the CBMcarb-DB throughout the
348 comparison of CtCBM11 structures in complex with the mixed-linked ligand with the
349 AlphaFold-predicted structure of a GH26-associated CBM11 from *Microbacterium*
350 *hydrocarbonoxydans* (*MhCBM11*). The bacteria display hydrocarbon-degrading
351 capabilities, which are of interest for bioremediation applications, especially in the
352 cleanup of oil spills and hydrocarbon-contaminated environments⁴³.

353 Insert Figure 8

354 The reason why hydrocarbon-degrading bacteria may possess family 26 glycoside
355 hydrolases could be related to the presence of complex carbohydrates in the
356 environments where they thrive or, as recently found, some members of family 26
357 glycoside hydrolases exhibit broader substrate specificity and can act on other
358 molecules beyond galactomannan^{44,45}. The amino acid sequence alignment of the two
359 CBM11 modules indicates a 20% of primary sequence identity. Figure X.8 shows the
360 superposition of the CtCBM11-hexasaccharide complex (PDB ID: 6r31) with the AF-
361 predicted structure of *MhCBM11* (A0A0M2HNJ1 in InterPro), producing a rmsd of
362 0.858 Å for 69 C-alpha atoms (and 5.830 Å across all 142 C-alpha pairs). Despite the
363 uncertainty on the binding specificity of *MhCBM11*, the superposition analysis reveals
364 the protein stretches (loops) that may accommodate a carbohydrate ligand. It suggests
365 that residues Tyr651, Tyr561, Gln567, Trp628 and Tyr651 from the CBM11 from *M.*
366 *hydrocarbonoxydans* may constitute harbouring sub-sites in the protein, having a
367 putative role in ligand recognition. Predicting the aminoacid residues responsible for
368 binding in a yet unsolved family member can significantly advance the rational design
369 of engineered CBM11s produced by recombinant methods.

370 **3.3. A further prediction of the topology of the CBM binding site from its**
371 **sequence**

372 One can predict the topology of the CBM binding site from its sequence using various
373 computational methods. These methods typically use sequence- or structure-based
374 features, or a combination of both, to identify and classify the binding sites. Some
375 examples of these methods are:

376 • **ConCavity**: This method combines evolutionary sequence conservation estimates
377 with structure-based methods for identifying protein surface cavities.

378 <https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000585>.

379 It can accurately predict 3D ligand binding pockets and individual ligand binding
380 residues.

381 • **PUResNet**: This method uses a deep residual neural network to predict protein-
382 ligand binding sites based on amino acids' physicochemical properties and spatial
383 arrangement.

384 <https://jcheminf.biomedcentral.com/articles/10.1186/s13321-021-00547-7>. It can
385 handle large and complex protein structures and outperforms existing methods on
386 several metrics.

387 • **LIGSITE**: This method uses a regular Cartesian grid to detect pockets and cavities
388 on the protein surface based on solvent accessibility and geometric criteria

389 <https://bip.weizmann.ac.il/toolbox/structure/binding.htm>.

390 It can identify potential ligand binding sites in protein structures.

391 • **EASYMIFs and SITEHOUND**: These methods use molecular interaction fields
392 (MIFs) to identify probable binding sites based on the interaction energy between the
393 protein and a probe molecule. They can filter and cluster the MIFs to locate the most
394 favourable binding sites.

395 **3.4. Are the long oligosaccharides retaining their low-energy conformation**
396 **while interacting with Carbohydrate-Binding Modules?**

397 The dimensions of the carbohydrates, as assessed by the number of constituting
398 monosaccharides, varies from simple mono; disaccharides to octa- and
399 nona-saccharides; 62 oligosaccharides range between five to nine constitutive
400 monomers. The structural diversity is somehow limited, as about 80% of the
401 oligosaccharides are polysaccharide fragments made up of 1-4 β -linked hexoses (Glc,
402 GlcNAc, Xyl, Man), with a rare 1-3 β -linkage occurrence. Branched structures occur in
403 fragments of Xyloglucans. The remaining 20% are 1-4 α -linked Glc residues, either as
404 small amylose linear fragments or cyclodextrins.

405 The limited number of disaccharide segments found when complexed to CBM
406 facilitates the analysis of their bound conformations. The distribution of the
407 experimentally observed conformations of such disaccharide segments can be
408 schematically represented when plotted on computed potential energy surfaces as a
409 function of the glycosidic torsion angles Φ and Ψ . Thanks to the similarity of their
410 configuration of the β -D-Glc-1-4- β -D-Glc, the β -D-GlcNAc-1-4- β -D-GlcNAc, the β -D-
411 Xyl-1-4- β -D-Xyl, the β -D-Man-1-4- β -D-Man, all the observed conformations can be
412 displayed on the same potential energy surface. Other potential energy surfaces,
413 computed for β -D-Glc-1-3- β -D-Glc and α -D-Glc-1-4- α -D-Glc, complete the analysis.
414 As expected, the experimentally observed glycosidic torsion angles displayed some
415 scattering when plotted on the corresponding potential energy surfaces, but they are
416 all located on the lowest energy basins (see Figure 9).

417

Insert Figure 9

418 A detailed inspection of the conformation of 62 hetero-pentasaccharides and
419 higher confirm the conclusions drawn above. The finding that the type B binding mode
420 does not induce any major/significant conformational change in the bound
421 carbohydrate is vital in a predictive context. From the knowledge of a
422 carbohydrate/polysaccharide sequence, several computational tools can provide
423 reliable low-energy 3D structures (glycan-builder.cermav.cnrs.fr⁴⁶;
424 bitbucket.org/mkuttel/carbbuilderii⁴⁷, glycam.org⁴⁸), which can serve as starting points
425 in docking simulations⁴⁹.

426 The conclusions derived from the analysis of combining sites of the protein on one
427 side and the conformations of the bound carbohydrate on the other side point towards
428 the absence of any induced-fit effect and indicate favourable conditions to perform
429 molecular docking.

430

431 **3.5. Integration with glycan microarray data to derive specificity of binding**

432 As mentioned above, a challenge to recognising functional CBM-carbohydrate
433 complexes is the identification of the natural carbohydrate ligands for CBMs and the
434 experimental assignment of the binding specificity at the oligosaccharide level. Over
435 the last two decades, carbohydrate (or glycan) microarrays have become instrumental
436 tools for assigning carbohydrate-protein interactions. They provide targets for
437 structural biology approaches and contribute to the elucidation of the function of
438 diverse endogenous and microbial carbohydrate-recognition systems^{13,50–55}.
439 Examples of glycan microarray analysis of bacterial CBMs are given in Figure 10,
440 highlighting different oligosaccharide-binding specificities.

441

Insert Figure 10

442 As observed in the microarray analysis, the carbohydrate chain length
443 requirements correlate with the different modes of interaction of the CBMs with the
444 ligand in solution using STD-NMR and their assigned functional type B and C⁵⁷. Such
445 findings and other curated glycan microarray data on CBMs can be visualised through
446 the Imperial College Glycosciences Lab web portal under category G
447 (<https://glycosciences.med.ic.ac.uk/data.html>).

448 In the foreseeable future, CBMcarb-DB will be connected to the experimental
449 glycan microarray data. Integrating CBMcarb-DB with curated glycan microarray data
450 will allow for atomic-level insight into the glycan-binding preferences and natural
451 ligands for CBMs and correlate glycan sequence recognition with molecular
452 determinants at the protein binding site.

453 In parallel with the implementation of glycan microarrays, there has been the
454 much-needed development of bioinformatic tools and web resources for handling and
455 using glycan microarray data (please see reference Bojar et al, 2022⁵⁸ and references
456 therein for a comprehensive review of artificial intelligence methods and
457 glycobioinformatics). Examples include the Carbohydrate Micro-Array Analysis and
458 Reporting Tool (CarbArrayART), which is a distributable software tool that
459 accommodates glycan microarray data and metadata storage with retrieval,
460 comparison, mining, and sharing of generated data; the glycan Array Dashboard
461 (GLAD)⁵⁹, which displays glycan microarray results and searchable glycan-binding
462 motifs; and DAGR⁶⁰, MCAW-DB⁶¹, GlyMDB⁶², and CarboGrove⁶³, which are
463 databases for the interpretation of glycan microarray data that store glycan
464 determinants for diverse recognition systems.

465 Another significant development has been the creation of an international
466 glycan array repository (<https://glygen.ccruc.uga.edu/ggarray/>), established under the
467 GlyGen project⁶⁴. There is a provision to release curated glycan microarray data on
468 CBMs to this repository (only two entries are recorded for members of family 40). The
469 data can be displayed through the CarbArrayART software interface that handles
470 metadata compliant with the glycan microarray guidelines for Minimum Information
471 Required for A Glycomics Experiment⁶⁵ for FAIR data.

472 The accessibility to glycan microarray data also provides the means for
473 developing powerful tools to predict glycan binding. These include MotifFinder, a
474 software tool for predicting glycan-binding motifs⁶⁶, and LectinOracle, which combines
475 transformer-based representations for proteins and graph convolutional neural
476 networks for glycans to predict lectin-glycan binding specificities⁶⁷.

477 Bridging CBMcarb-DB with related artificial intelligence and glycobioinformatic
478 tools will provide a structure-based informed rationale to add to predictors of glycan-
479 binding specificities for newly identified CBM sequences from available genomics
480 data. This will advance knowledge of glycan function and CBM engineering for
481 biotechnological and biomedical applications. Importantly, these approaches will fine-
482 tune the characterisation of multidomain enzymatic systems targeting complex
483 carbohydrate substrates, e.g. polysaccharides or mucin glycoproteins.

484

485 **4. Conclusion**

486 In the ever-expanding field of glycobiology, the study of glycan structures and their
487 interactions with biomolecules is paramount. The binding of carbohydrate ligands
488 to CBMs triggers structural conformational changes that optimise the recognition and

489 catalytic activity of CAZymes. These changes range from localised adjustments to
490 global domain motions, and their dynamic nature allows for efficient substrate binding
491 and enzymatic activity. Understanding the structural dynamics of CBMs and their
492 conformational changes upon ligand binding is crucial for elucidating the
493 molecular mechanisms underlying carbohydrate recognition and catalysis. Structural
494 biology databases provide unique information about the three-dimensional structures
495 of proteins, nucleic acids, and other biomolecules.

496 CBMcarb-DB is a valuable resource for researchers investigating CBM
497 structures and their interactions with carbohydrates, facilitating the analysis of 3D
498 structures, the exploration of carbohydrate-binding interactions, and the examination
499 of conformational changes upon ligand binding. By using CBMcarb-DB and integrating
500 its data with other bioinformatics tools researchers can better understand CBM
501 structure-function relationships and the molecular mechanisms
502 underlying carbohydrate recognition. The database can be utilised in conjunction with
503 other bioinformatics resources and tools, as, for instance, by cross-referencing the
504 CBM structures with databases such as CAZy and obtain additional information on
505 the associated catalytic domains and substrate specificities. Furthermore,
506 integrating structural analysis with sequence analysis tools enables a
507 comprehensive understanding of CBM structure-function relationships.
508 In addition to 3D structures, CBMcarb-DB provides valuable annotations and
509 information related to CBMs. Further research combining experimental techniques
510 and computational simulations will provide deeper insights into these fascinating
511 processes. The representation of CBM-oligosaccharide structures covered by
512 CBMcarb-DB so far may reflect the biological and biotechnological interest of the
513 community in investigating these CBMs, as well as the difficulty in obtaining crystals

514 and high-resolution structures of proteins in complex with larger ligands. CBMcarb-DB
515 curated information will promote an understanding of the structural basis of CBMs
516 ligand specificity, which may generate more structural and functional data to supply
517 the database.

518 In a different approach, glycan array repositories serve as valuable resources
519 for investigating the specific binding interactions between glycans and various
520 biomolecular targets. Such repositories house collections of diverse glycans
521 immobilised on different surfaces, enabling high-throughput screening of glycan-
522 protein interactions. These arrays enable researchers to explore the recognition
523 patterns and binding specificities of various proteins, antibodies, and other
524 biomolecules towards different glycan structures. By systematically probing these
525 interactions, scientists can decipher the "glycan code" and unravel the
526 intricate language of glycan recognition. In this chapter, we have explored the
527 integration of structural biology databases and glycan array repositories, offering
528 new opportunities for understanding glycan-mediated recognition events.

529

530

531 **ACKNOWLEDGEMENTS**

532 The authors would like to thank Doctor Luis Gomes (Department of Informatics,
533 Faculty of Sciences, University of Lisbon) for his kind assistance in extracting the
534 publicly available PDB IDs of CBM carbohydrate structures and Doctor José Braga
535 (UCIBIO, FCT-NOVA) for implementing CBMcarb-DB in the FCT-NOVA web server.
536 Funding by the Portuguese Foundation for Science and Technology (FCT-MCTES)
537 for the project grants PTDC/BIA-MIB/31730/2017 and 2022.06104.PTDC; the Applied

538 Molecular Biosciences Unit - UCIBIO (UIDP/04378/2020 and UIDB/04378/2020) and
539 the Associate Laboratory Institute for Health and Bioeconomy - i4HB
540 (LA/P/0140/2020), financed by FCT. We also acknowledge the project HORIZON-
541 WIDERA-2021-101079417-GLYCOTwinning, financed by European funds and the
542 COST Action Innogly CA18103 for the E-COST Virtual Mobility Grant to DR (E-COST-
543 GRANT-CA18103-02b37039). The CrossDisciplinary Program Glyco@Alps
544 supported this work within the framework 'Investissement d'Avenir' program [ANR-
545 15IDEX-02].

546

547 **References**

- 548 1 A. B. Boraston, D. N. Bolam, H. J. Gilbert and G. J. Davies, *Biochemical Journal*,
549 2004, **382**, 769–781.
- 550 2 H. J. Gilbert, J. P. Knox and A. B. Boraston, *Curr Opin Struct Biol*, 2013, **23**,
551 669–677.
- 552 3 E. Drula, M.-L. Garron, S. Dogan, V. Lombard, B. Henrissat and N. Terrapon,
553 *Nucleic Acids Res*, 2022, **50**, D571–D577.
- 554 4 H. J. Gilbert, *Plant Physiol*, 2010, **153**, 444–455.
- 555 5 M. E. Berg Miller, D. A. Antonopoulos, M. T. Rincon, M. Band, A. Bari, T. Akraiko,
556 A. Hernandez, J. Thimmapuram, B. Henrissat, P. M. Coutinho, I. Borovok, S.
557 Jindou, R. Lamed, H. J. Flint, E. A. Bayer and B. A. White, *PLoS One*,
558 DOI:10.1371/journal.pone.0006650.
- 559 6 O. Yaniv, G. Fichman, I. Borovok, Y. Shoham, E. A. Bayer, R. Lamed, L. J. W.
560 Shimon and F. Frolow, *Acta Crystallogr D Biol Crystallogr*, 2014, **70**, 522–534.
- 561 7 G. Buist, A. Steen, J. Kok and O. P. Kuipers, *Mol Microbiol*, 2008, **68**, 838–847.
- 562 8 Alicia Lammerts van Bueren and Elizabeth Ficko-Blean, Carbohydrate-binding
563 modules, [https://www.cazypedia.org/index.php/Carbohydrate-](https://www.cazypedia.org/index.php/Carbohydrate-binding_modules)
564 [binding_modules](https://www.cazypedia.org/index.php/Carbohydrate-binding_modules), (accessed 24 July 2023).

- 565 9 T. Ohnuma and T. Taira, Carbohydrate Binding Module Family 50,
566 [https://www.cazypedia.org/index.php/Carbohydrate_Binding_Module_Family_](https://www.cazypedia.org/index.php/Carbohydrate_Binding_Module_Family_50)
567 50, (accessed 24 July 2023).
- 568 10 S. Etzold and N. Juge, *Curr Opin Struct Biol*, 2014, **28**, 23–31.
- 569 11 F. Bonnardel, S. M. Haslam, A. Dell, T. Feizi, Y. Liu, V. Tajadura-Ortega, Y.
570 Akune, L. Sykes, P. R. Bennett, D. A. MacIntyre, F. Lisacek and A. Imberty, *NPJ*
571 *Biofilms Microbiomes*, 2021, **7**, 49.
- 572 12 T. Nakamura, S. Mine, Y. Hagihara, K. Ishikawa, T. Ikegami and K. Uegaki, *J*
573 *Mol Biol*, 2008, **381**, 670–680.
- 574 13 D. O. Ribeiro, A. Viegas, V. M. R. Pires, J. Medeiros-Silva, P. Bule, W. Chai, F.
575 Marcelo, C. M. G. A. Fontes, E. J. Cabrita, A. S. Palma and A. L. Carvalho,
576 *FEBS J*, 2020, **287**, 2723–2743.
- 577 14 V. Notenboom, A. B. Boraston, D. G. Kilburn and D. R. Rose, *Biochemistry*,
578 2001, **40**, 6248–6256.
- 579 15 E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E.
580 C. Meng and T. E. Ferrin, *J Comput Chem*, 2004, **25**, 1605–1612.
- 581 16 S. M. Tschampel and R. J. Woods, *J Phys Chem A*, 2003, **107**, 9175–9181.
- 582 17 S. Armenta, S. Moreno-Mendieta, Z. Sánchez-Cuapio, S. Sánchez and R.
583 Rodríguez-Sanoja, *Proteins: Structure, Function, and Bioinformatics*, 2017, **85**,
584 1602–1617.
- 585 18 S. Mir, Y. Alhroub, S. Anyango, D. R. Armstrong, J. M. Berrisford, A. R. Clark,
586 M. J. Conroy, J. M. Dana, M. Deshpande, D. Gupta, A. Gutmanas, P. Haslam,
587 L. Mak, A. Mukhopadhyay, N. Nadzirin, T. Paysan-Lafosse, D. Sehnal, S. Sen,
588 O. S. Smart, M. Varadi, G. J. Kleywegt and S. Velankar, *Nucleic Acids Res*,
589 2018, **46**, D486–D492.
- 590 19 D. Sehnal, A. Rose, J. Koca, S. Burley and S. Velankar, The Eurographics
591 Association, 2018, pp. 29–33.
- 592 20 H. M. Berman, *Nucleic Acids Res*, 2000, **28**, 235–242.

- 593 21 M. Blum, H.-Y. Chang, S. Chuguransky, T. Grego, S. Kandasamy, A. Mitchell,
594 G. Nuka, T. Paysan-Lafosse, M. Qureshi, S. Raj, L. Richardson, G. A. Salazar,
595 L. Williams, P. Bork, A. Bridge, J. Gough, D. H. Haft, I. Letunic, A. Marchler-
596 Bauer, H. Mi, D. A. Natale, M. Necci, C. A. Orengo, A. P. Pandurangan, C.
597 Rivoire, C. J. A. Sigrist, I. Sillitoe, N. Thanki, P. D. Thomas, S. C. E. Tosatto, C.
598 H. Wu, A. Bateman and R. D. Finn, *Nucleic Acids Res*, 2021, **49**, D344–D354.
- 599 22 R. A. Laskowski, J. Jabłońska, L. Pravda, R. S. Vařeková and J. M. Thornton,
600 *Protein Science*, 2018, **27**, 129–134.
- 601 23 A. R. Kinjo, G.-J. Bekker, H. Suzuki, Y. Tsuchiya, T. Kawabata, Y. Ikegawa and
602 H. Nakamura, *Nucleic Acids Res*, 2017, **45**, D282–D288.
- 603 24 A. D. McNaught, *Adv Carbohydr Chem Biochem*, 1997, **52**, 43–177.
- 604 25 S. Herget, R. Ranzinger, K. Maass and C.-W. v. d. Lieth, *Carbohydr Res*, 2008,
605 **343**, 2162–2171.
- 606 26 T. Lütteke, A. Bohne-Lang, A. Loss, T. Goetz, M. Frank and C.-W. von der Lieth,
607 *Glycobiology*, 2006, **16**, 71R-81R.
- 608 27 S. Neelamegham, K. Aoki-Kinoshita, E. Bolton, M. Frank, F. Lisacek, T. Lütteke,
609 N. O'Boyle, N. Packer, H. P. Stanley, P. Toukach, A. Varki, R. J. Woods, A.
610 Darvill, A. Dell, B. Henrissat, C. Bertozzi, G. Hart, H. Narimatsu, H. Freeze, I.
611 Yamada, J. Paulson, J. Prestegard, J. Marth, J. F. G. Vliegenthart, M. Etzler, M.
612 Aebi, M. Kanehisa, N. Taniguchi, N. Edwards, P. Rudd, P. Seeberger, R.
613 Mazumder, R. Ranzinger, R. Cummings, R. Schnaar, S. Perez, S. Kornfeld, T.
614 Kinoshita, W. York and Y. Knirel, *Glycobiology*, 2019, **29**, 620–624.
- 615 28 R. A. Laskowski and M. B. Swindells, *J Chem Inf Model*, 2011, **51**, 2778–2786.
- 616 29 *Nucleic Acids Res*, 2019, **47**, D506–D515.
- 617 30 A. R. Kinjo, G.-J. Bekker, H. Wako, S. Endo, Y. Tsuchiya, H. Sato, H. Nishi, K.
618 Kinoshita, H. Suzuki, T. Kawabata, M. Yokochi, T. Iwata, N. Kobayashi, T.
619 Fujiwara, G. Kurisu and H. Nakamura, *Protein Science*, 2018, **27**, 95–102.

- 620 31 A. Waterhouse, M. Bertoni, S. Bienert, G. Studer, G. Tauriello, R. Gumienny, F.
621 T. Heer, T. A. P. de Beer, C. Rempfer, L. Bordoli, R. Lepore and T. Schwede,
622 *Nucleic Acids Res*, 2018, **46**, W296–W303.
- 623 32 S. Salentin, S. Schreiber, V. J. Haupt, M. F. Adasme and M. Schroeder, *Nucleic*
624 *Acids Res*, 2015, **43**, W443–W447.
- 625 33 E. Krissinel, *J Comput Chem*, 2010, **31**, 133–143.
- 626 34 E. Krissinel and K. Henrick, *J Mol Biol*, 2007, **372**, 774–797.
- 627 35 A. S. Rose, A. R. Bradley, Y. Valasatava, J. M. Duarte, A. Prlić and P. W. Rose,
628 *Bioinformatics*, 2018, **34**, 3755–3758.
- 629 36 I. Sillitoe, N. Dawson, T. E. Lewis, S. Das, J. G. Lees, P. Ashford, A. Tolulope,
630 H. M. Scholes, I. Senatorov, A. Bujan, F. Ceballos Rodriguez-Conde, B.
631 Dowling, J. Thornton and C. A. Orengo, *Nucleic Acids Res*, 2019, **47**, D280–
632 D284.
- 633 37 J. Zheng, Q. Ge, Y. Yan, X. Zhang, L. Huang and Y. Yin, *Nucleic Acids Res*,
634 2023, **51**, W115–W121.
- 635 38 X. Lin, X. Xie, X. Wang, Z. Yu, X. Chen and F. Yang, *Applied Sciences*, 2022,
636 **12**, 7842.
- 637 39 J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K.
638 Tunyasuvunakool, R. Bates, A. Židek, A. Potapenko, A. Bridgland, C. Meyer, S.
639 A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J.
640 Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger,
641 M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W.
642 Senior, K. Kavukcuoglu, P. Kohli and D. Hassabis, *Nature*, 2021, **596**, 583–589.
- 643 40 M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J.
644 Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, C. Millán, H. Park, C. Adams, C.
645 R. Glassman, A. DeGiovanni, J. H. Pereira, A. V. Rodrigues, A. A. van Dijk, A.
646 C. Ebrecht, D. J. Opperman, T. Sagmeister, C. Buhlheller, T. Pavkov-Keller, M.
647 K. Rathinaswamy, U. Dalwadi, C. K. Yip, J. E. Burke, K. C. Garcia, N. V. Grishin,
648 P. D. Adams, R. J. Read and D. Baker, *Science (1979)*, 2021, **373**, 871–876.

- 649 41 J. Xu, *Proceedings of the National Academy of Sciences*, 2019, **116**, 16856–
650 16865.
- 651 42 J. Pereira, A. J. Simpkin, M. D. Hartmann, D. J. Rigden, R. M. Keegan and A.
652 N. Lupas, *Proteins: Structure, Function, and Bioinformatics*, 2021, **89**, 1687–
653 1699.
- 654 43 A. Schippers, K. Bosecker, C. Spröer and P. Schumann, *Int J Syst Evol*
655 *Microbiol*, 2005, **55**, 655–660.
- 656 44 A. B. Patel, A. K. Patel, M. P. Shah, I. K. Parikh and C. G. Joshi, *Biotechnol Appl*
657 *Biochem*, 2016, **63**, 257–265.
- 658 45 E. M. Glasgow, K. A. Vander Meulen, T. E. Takasuka, C. M. Bianchetti, L. F.
659 Bergeman, S. Deutsch and B. G. Fox, *J Mol Biol*, 2019, **431**, 1217–1233.
- 660 46 S. B. Engelsens, P. I. Hansen and S. Pérez, *Biopolymers*, 2014, **101**, 733–743.
- 661 47 M. M. Kuttel, J. Stähle and G. Widmalm, *J Comput Chem*, 2016, **37**, 2098–2105.
- 662 48 K. N. Kirschner, A. B. Yongye, S. M. Tschampel, J. González-Outeiriño, C. R.
663 Daniels, B. L. Foley and R. J. Woods, *J Comput Chem*, 2008, **29**, 622–655.
- 664 49 R. Marchetti, S. Perez, A. Arda, A. Imberty, J. Jimenez-Barbero, A. Silipo and
665 A. Molinaro, *ChemistryOpen*, 2016, **5**, 274–296.
- 666 50 H. L. Pedersen, J. U. Fangel, B. McCleary, C. Ruzanski, M. G. Rydahl, M.-C.
667 Ralet, V. Farkas, L. von Schantz, S. E. Marcus, M. C. F. F. Andersen, R. Field,
668 M. Ohlin, J. P. Knox, M. H. Clausen and W. G. T. T. Willats, *Journal of Biological*
669 *Chemistry*, 2012, **287**, 39429–39438.
- 670 51 C. Ruprecht, M. P. Bartetzko, D. Senf, P. Dallabernadina, I. Boos, M. C. F.
671 Andersen, T. Kotake, J. P. Knox, M. G. Hahn, M. H. Clausen and F. Pfrengle,
672 *Plant Physiol*, 2017, **175**, 1094–1104.
- 673 52 C. D. Rillahan and J. C. Paulson, *Annu Rev Biochem*, 2011, **80**, 797–823.
- 674 53 A. S. Palma, T. Feizi, R. A. Childs, W. Chai and Y. Liu, *Curr Opin Chem Biol*,
675 2014, **18**, 87–94.

- 676 54 C. Ruprecht, A. Geissner, P. H. Seeberger and F. Pfrenkle, *Carbohydr Res*,
677 2019, **481**, 31–35.
- 678 55 V. G. Correia, F. Trovão, B. A. Pinheiro, J. L. A. Brás, L. M. Silva, C. Nunes, M.
679 A. Coimbra, Y. Liu, T. Feizi, C. M. G. A. Fontes, B. Mulloy, W. Chai, A. L.
680 Carvalho and A. S. Palma, *Microbiol Spectr*, , DOI:10.1128/Spectrum.01826-21.
- 681 56 D. O. Ribeiro, B. A. Pinheiro, A. L. Carvalho and A. S. Palma, in *Carbohydrate*
682 *Chemistry: Chemical and biological approaches*, eds. A. P. Rauter, T. Lindhorst
683 and Y. Queneau, Royal Society of Chemistry, 2017, pp. 159–176.
- 684 57 A. S. Palma, Y. Liu, H. Zhang, Y. Zhang, B. V. McCleary, G. Yu, Q. Huang, L.
685 S. Guidolin, A. E. Ciocchini, A. Torosantucci, D. Wang, A. L. Carvalho, C. M. G.
686 A. Fontes, B. Mulloy, R. A. Childs, T. Feizi and W. Chai, *Molecular & Cellular*
687 *Proteomics*, 2015, **14**, 974–988.
- 688 58 D. Bojar and F. Lisacek, *Chem Rev*, 2022, **122**, 15971–15988.
- 689 59 A. Y. Mehta and R. D. Cummings, *Bioinformatics*, 2019, **35**, 3536–3537.
- 690 60 E. Sterner, N. Flanagan and J. C. Gildersleeve, *ACS Chem Biol*, 2016, **11**,
691 1773–1783.
- 692 61 M. Hosoda, Y. Takahashi, M. Shiota, D. Shinmachi, R. Inomoto, S. Higashimoto
693 and K. F. Aoki-Kinoshita, *Carbohydr Res*, 2018, **464**, 44–56.
- 694 62 Y. Cao, S.-J. Park, A. Y. Mehta, R. D. Cummings and W. Im, *Bioinformatics*,
695 2020, **36**, 2438–2442.
- 696 63 Z. L. Klamer, C. M. Harris, J. M. Beirne, J. E. Kelly, J. Zhang and B. B. Haab,
697 *Glycobiology*, 2022, **32**, 679–690.
- 698 64 W. S. York, R. Mazumder, R. Ranzinger, N. Edwards, R. Kahsay, K. F. Aoki-
699 Kinoshita, M. P. Campbell, R. D. Cummings, T. Feizi, M. Martin, D. A. Natale,
700 N. H. Packer, R. J. Woods, G. Agarwal, S. Arpinar, S. Bhat, J. Blake, L. J. G.
701 Castro, B. Fochtman, J. Gildersleeve, R. Goldman, X. Holmes, V. Jain, S.
702 Kulkarni, R. Mahadik, A. Mehta, R. Mousavi, S. Nakarakommula, R. Navelkar,
703 N. Pattabiraman, M. J. Pierce, K. Ross, P. Vasudev, J. Vora, T. Williamson and
704 W. Zhang, *Glycobiology*, 2020, **30**, 72–73.

705 65 Y. Liu, R. McBride, M. Stoll, A. S. Palma, L. Silva and S. Agravat, *Glycobiology*,
706 2016, 1–6.

707 66 Z. Klamer and B. Haab, *Anal Chem*, 2021, **93**, 10925–10933.

708 67 J. Lundstrøm, E. Korhonen, F. Lisacek and D. Bojar, *Advanced Science*, 2022,
709 **9**, 2103807.

710

711

712

713 **FIGURE AND TABLE CAPTIONS**

714 **Figure 1.** Schematic representation of the three modes of CBM recognition of
 715 carbohydrate substrates. The "planar" approach (classified as Type A) to crystalline
 716 polysaccharides is represented by the chitin-binding family 2 CBM from *Pyrococcus*
 717 *furiosus* (PDB ID 2CWR)¹²; the "endo" approach (classified as Type B) is represented
 718 by the mixed-linked beta-glucan-binding family 11 CBM from *Clostridium*
 719 *thermocellum* bound to Glc β 1,4Glc β 1,3Glc β 1,4Glc β 1,4Glc β 1,3Glc (PDB ID 6R31)¹³
 720 and exhibiting the typically extended groove (or cleft) with binding subsites capable of
 721 accommodating isolated carbohydrate chains with degrees of polymerisation longer
 722 than 4; the "exo" approach (classified as Type C) is represented by the family 9
 723 carbohydrate-binding module from *Thermotoga maritima* bound to cellobiose (PDB ID
 724 1I82)¹⁴ and exhibiting the typical pocket recognising the termini of glycans containing
 725 one to three monosaccharide units. Polypeptide chains are represented in a blue
 726 ribbon, while carbohydrate chains are represented in stick model. The picture was
 727 prepared with program UCSF Chimera¹⁵.

728 **Figure 2.** Overview of the content of CBMcarb-DB. Screenshot image taken from the
 729 front page of CBMcarb-DB website, depicting the search fields and two examples of
 730 results.

731 **Figure 3.** Screenshot image listing the CBM families represented in CBMcarb-DB.

732 **Figure 4.** Screenshot image depicting an example search within family 16 CBMs on
 733 the main page of CBMcarb-DB.

734 **Figure 5.** View of the webpage depicting the multiple criteria of the advanced search
 735 in CBMcarb-DB.

736 **Figure 6.** Summary view of the field search results using PDB ID 6R31 in the *pdb*
737 search field.

738 **Figure 7.** Full results from the search on CBMcarb-DB, using PDB ID 6R31 in the *pdb*
739 search field. More information can be obtained by clicking on the green buttons.

740 **Figure 8.** Ribbon representation of the crystal structure of the Family 11 CBM from
741 *Clostridium thermocellum* (CtCBM11) in complex with
742 Glc β 1,4Glc β 1,3Glc β 1,4Glc β 1,4Glc β 1,3Glc (PDB 6R31), in beige, superposed with the
743 AlphaFold-predicted structure of a GH26-associated CBM11 from *Microbacterium*
744 *hydrocarbonoxydans* (MhCBM11; A0A0M2HNJ1 in InterPro), in light blue. The
745 alignment was performed using the MatchMaker tool from UCSF Chimera¹⁵, and the
746 rmsd of the superposition is 0.858 Å for 69 matching C-alpha atoms (and 5.830 Å
747 across all 142 C-alpha pairs). The concave side of both CBM11 forms the binding cleft,
748 where ligands are accommodated. The carbohydrate atoms and the side chains of the
749 amino acid residues inside the binding cleft of CtCBM11 that interact with the ligand
750 are shown as stick models and labelled in beige characters. Amino acid residues
751 Tyr651, Tyr561, Gln567, Trp628 and Tyr651 from the MhCBM11, also in stick and
752 labelled in blue characters, constitute the potential residues involved in ligand
753 recognition. Calcium atoms (in green spheres) are surrounded by their coordinating
754 residues, shown as sticks. The picture was prepared with the program UCSF
755 Chimera¹⁵.

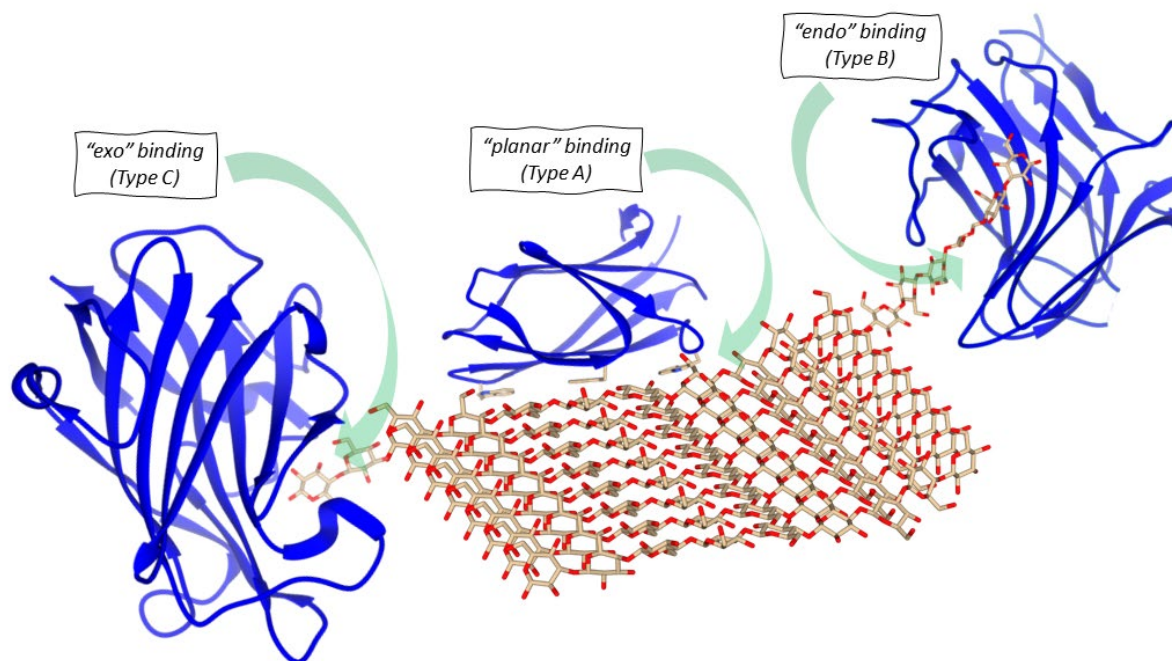
756 **Figure 9.** Φ and Ψ measured in the 3D structures of oligosaccharides of CBM
757 complexes reported on potential energy surfaces. (a) β -D-Glc-1-4- β -D-Glc, the β -D-
758 GlcNAc-1-4- β -D-GlcNAc, the β -D-Xyl-1-4- β -D-Xyl, the β -D-Man-1-4- β -D-Man. (b) α -
759 D-Glc-1-4- α -D-Glc. (c) β -D-Glc-1-3- β -D-Glc.

760 **Figure 10.** Differing specificities and chain length requirements obtained in the
761 microarray analysis of glucan-binding CBMs from different CAZy families: family 41
762 and family 4 CBMs from the marine hyperthermophile *Thermotoga maritima*
763 (*TmCBM41* and *TmCBM4-2*, respectively); family 11 CBM from *Clostridium*
764 *thermocellum* (*CtCBM11*); and family 6 from the aerobic soil bacterium *Cellvibrio*
765 *mixtus* (*CmCBM6-2*); the inset shows the binding epitopes as determined by STD
766 NMR of β 1–3-linked D-glucose trisaccharide in the presence of *TmCBM4-2* and
767 *CmCBM6-2* (dark, medium and light grey circles indicate strong, medium, and weak
768 STD effects, respectively); NGL- lipid-linked (neoglycolipid) probe. Depiction of the
769 glucan-oligosaccharide sequence diversity in the microarray is on the top of the panel.
770 DP: degree of polymerisation. Reproduced from Ribeiro et al. 2018⁵⁶ with permission
771 from The Royal Society of Chemistry.

772

773

774 **Figures**



775

776

777

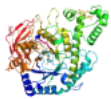
Figure 1

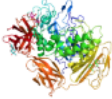
CBMcarb-DB field search

pdb	<input type="text"/>	cbm family	<input type="text"/>
protein name	<input type="text"/>	organism	<input type="text"/>
resolution	<input type="text" value="0"/>	carb pdb	<input type="text"/>
carb iupac	<input type="text"/>	carb length	<input type="text"/>

Scrolling through the database

354 CBM RESULTS

1B9Z-CBM20		Visualize the 3D structure	
pdb	1b9z	cbm_family	CBM20
protein_function	Hydrolase	protein_name	b-amylase (Spoll)
pdb_title	Bacillus cereus beta-amylase cc	organism	Bacillus cereus VAR. MYCOIE
domain	Bacteria	carb_pdb	GLC-GLC
carb_iupac	Glc(a1-4)Glc	linucs	[[[b-D-Glcp]](4+1)][a-D-Glcp]]
carb_mass	342,297	wurcs	WURCS=2.0/1,2,1/[a2122h-1a_
glytoucan		carb_length	2
		glycostructure	a-D-Glcp-(1-4)-a-D-Glcp
		comments2	
			

1CDG-CBM20		Visualize the 3D structure	
pdb	1cdg	cbm_family	CBM20
protein_function	Transferase	protein_name	b-cyclodextrin glucanotrans
pdb_title	Nucleotide sequence and x-ray	organism	Bacillus circulans 251
domain	Bacteria	carb_pdb	GLC-GLC
carb_iupac	Glc(a1-4)Glc	linucs	[[[a-D-Glcp]](4+1)][a-D-Glcp]]
carb_mass	342,297	wurcs	WURCS=2.0/1,2,1/[a2122h-1a_
glytoucan		carb_length	2
		glycostructure	a-D-Glcp-(1-4)-a-D-Glcp
		comments2	
			

778

779

Figure 2

Browse by CBM Family > Protein Name > PDB

(Click on a panel to expand it and on the blue search button to explore)

<input type="text" value="CBM4"/> <input type="button" value="Q"/>	<input type="text" value="CBM25"/> <input type="button" value="Q"/>	<input type="text" value="CBM50"/> <input type="button" value="Q"/>
<input type="text" value="CBM6"/> <input type="button" value="Q"/>	<input type="text" value="CBM27"/> <input type="button" value="Q"/>	<input type="text" value="CBM51"/> <input type="button" value="Q"/>
<input type="text" value="CBM9"/> <input type="button" value="Q"/>	<input type="text" value="CBM28"/> <input type="button" value="Q"/>	<input type="text" value="CBM57"/> <input type="button" value="Q"/>
<input type="text" value="CBM11"/> <input type="button" value="Q"/>	<input type="text" value="CNM29"/> <input type="button" value="Q"/>	<input type="text" value="CBM58"/> <input type="button" value="Q"/>
<input type="text" value="CBM12"/> <input type="button" value="Q"/>	<input type="text" value="CBM32"/> <input type="button" value="Q"/>	<input type="text" value="CBM60"/> <input type="button" value="Q"/>
<input type="text" value="CBM13"/> <input type="button" value="Q"/>	<input type="text" value="CBM33"/> <input type="button" value="Q"/>	<input type="text" value="CBM61"/> <input type="button" value="Q"/>
<input type="text" value="CBM14"/> <input type="button" value="Q"/>	<input type="text" value="CBM34"/> <input type="button" value="Q"/>	<input type="text" value="CBM62"/> <input type="button" value="Q"/>
<input type="text" value="CBM15"/> <input type="button" value="Q"/>	<input type="text" value="CBM35"/> <input type="button" value="Q"/>	<input type="text" value="CBM63"/> <input type="button" value="Q"/>
<input type="text" value="CBM16"/> <input type="button" value="Q"/>	<input type="text" value="CBM36"/> <input type="button" value="Q"/>	<input type="text" value="CBM65"/> <input type="button" value="Q"/>
<input type="text" value="CBM17"/> <input type="button" value="Q"/>	<input type="text" value="CBM39"/> <input type="button" value="Q"/>	<input type="text" value="CBM66"/> <input type="button" value="Q"/>
<input type="text" value="CBM18"/> <input type="button" value="Q"/>	<input type="text" value="CBM40"/> <input type="button" value="Q"/>	<input type="text" value="CBM67"/> <input type="button" value="Q"/>
<input type="text" value="CBM19"/> <input type="button" value="Q"/>	<input type="text" value="CBM41"/> <input type="button" value="Q"/>	<input type="text" value="CBM80"/> <input type="button" value="Q"/>
<input type="text" value="CBM20"/> <input type="button" value="Q"/>	<input type="text" value="CBM42"/> <input type="button" value="Q"/>	<input type="text" value="CBM81"/> <input type="button" value="Q"/>
<input type="text" value="CBM21"/> <input type="button" value="Q"/>	<input type="text" value="CBM47"/> <input type="button" value="Q"/>	<input type="text" value="CBM86"/> <input type="button" value="Q"/>
<input type="text" value="CBM22"/> <input type="button" value="Q"/>	<input type="text" value="CBM48"/> <input type="button" value="Q"/>	

780

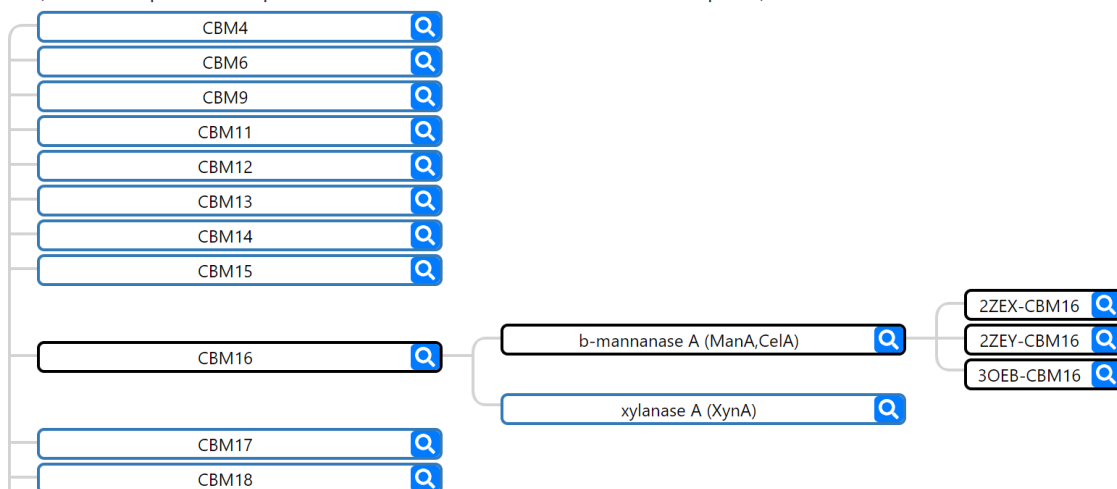
781

Figure 3

782

Browse by Type > Name > PDB

(Click on a panel to expand it and on the blue search button to explore)



783

784

785

Figure 4

CBMcarb-DB field search

pdb	<input type="text"/>	cbm family	<input type="text"/>
protein name	<input type="text"/>	organism	<input type="text"/>
resolution	<input type="text" value="0"/>	carb pdb	<input type="text"/>
carb iupac	<input type="text"/>	carb length	<input type="text"/>

786

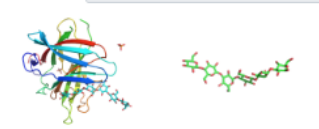
787

788

Figure 5

1 CBM RESULTS

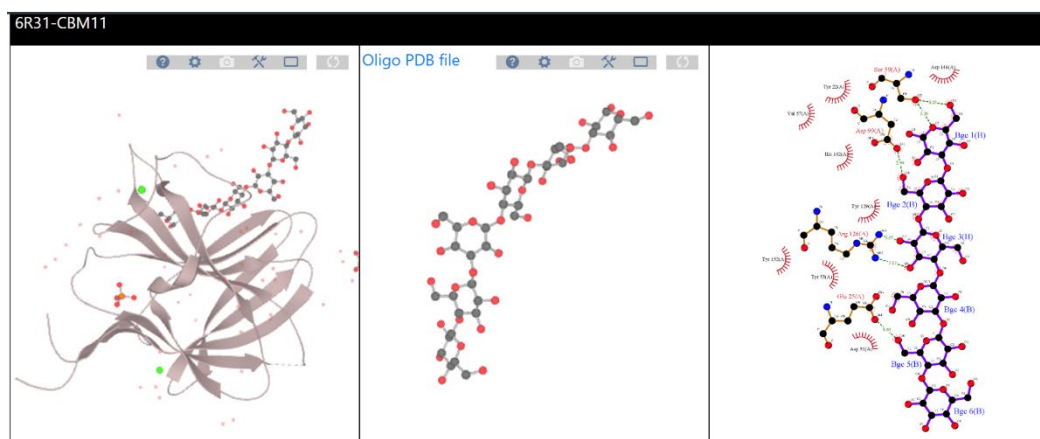
6R31-CBM11		Visualize the 3D structure	
pdb	6r31	cbm_family	CBM11
protein_name	Family 11 carbohydrate-binding	protein_name	bifunctional endo-b-1,4-glu
domain	Bacteria	resolution	2.6
carb_iupac	Glc(b1-4)Glc(b1-4)Glc(b1-3)	wurcs	WURCS=2.0/1,6,5/[a2122h-1b_
carb_mass	990.86	carb_length	6
glytoucan		comments	
protein_function	Sugar binding protei		
organism	Hungateiclostridium thermo		
carb_pdb	BGC-BGC-BGC-BGC-BGC-BC		
linucs	[[[b-D-Glcp](((3+1))][b-D-Glcp][[
glycostructure	b-D-Glcp-(1-4)-b-D-Glcp-(1-3)-		
comments2			



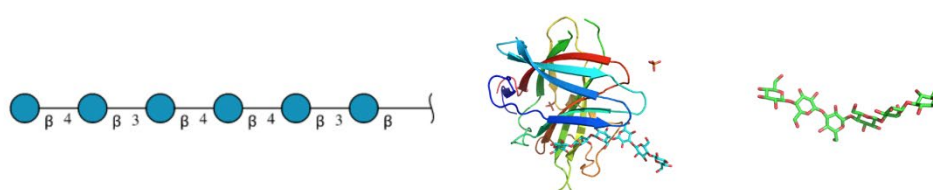
789

790

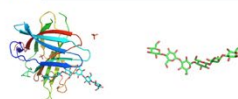
Figure 6



Glyco structure



RCSB.ORG	PDBe	PDBj	PDBsum
CAZy CBM family	PDBe ligands	SwissModel (ligand interactions)	PLIP
PISA	PubMed	DOI CR	UniProt
pdb	6r31	cbm_family	CBM11
pdb_title	Family 11 carbohydrate-binding	protein_name	bifunctional endo-b-1,4-glucan
domain	Bacteria	resolution	2.6
carb_iupac	Glc(b1-4)Glc(b1-4)Glc(b1-3)Glc	wurcs	WURCS=2.0/1,6,5/[a2122h-1b_
carb_mass	990.86	carb_length	6
glytoucan		comments	
		protein_function	Sugar binding protei
		organism	Hungateiclostridium thermocel
		carb_pdb	BGC-BGC-BGC-BGC-BGC-BGC
		linucs	[[[b-D-Glcp]]{((3+1))][b-D-Glcp]]{
		glycostructure	b-D-Glcp-(1-4)-b-D-Glcp-(1-3)-
		comments2	

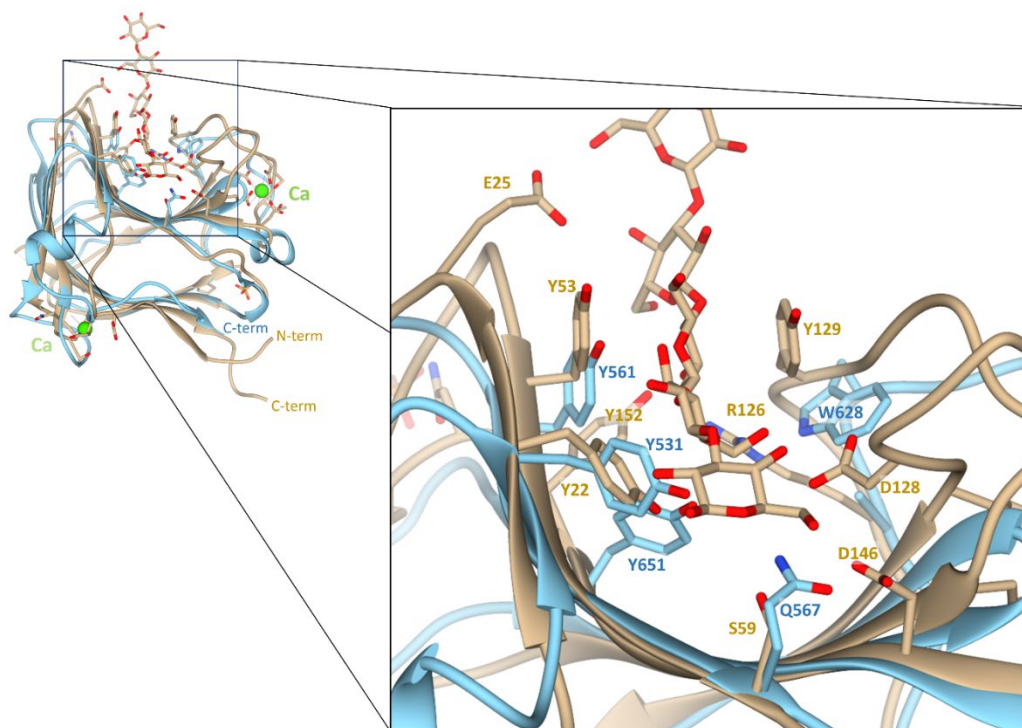


791

792

793

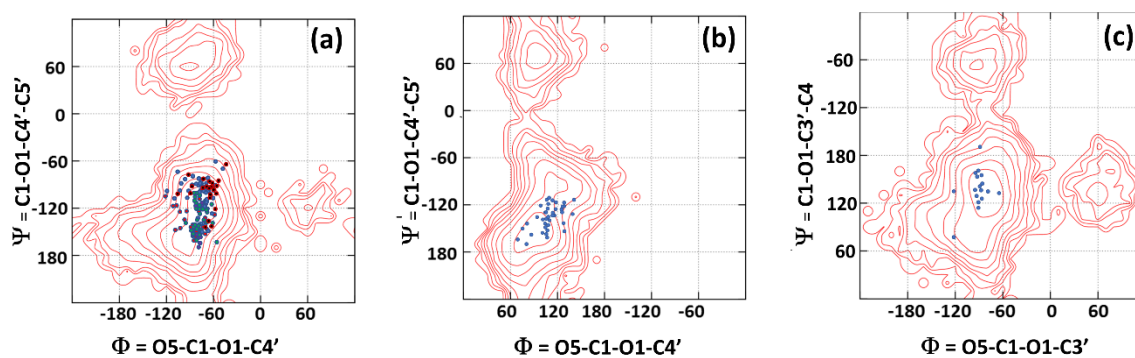
Figure 7



794

795

Figure 8

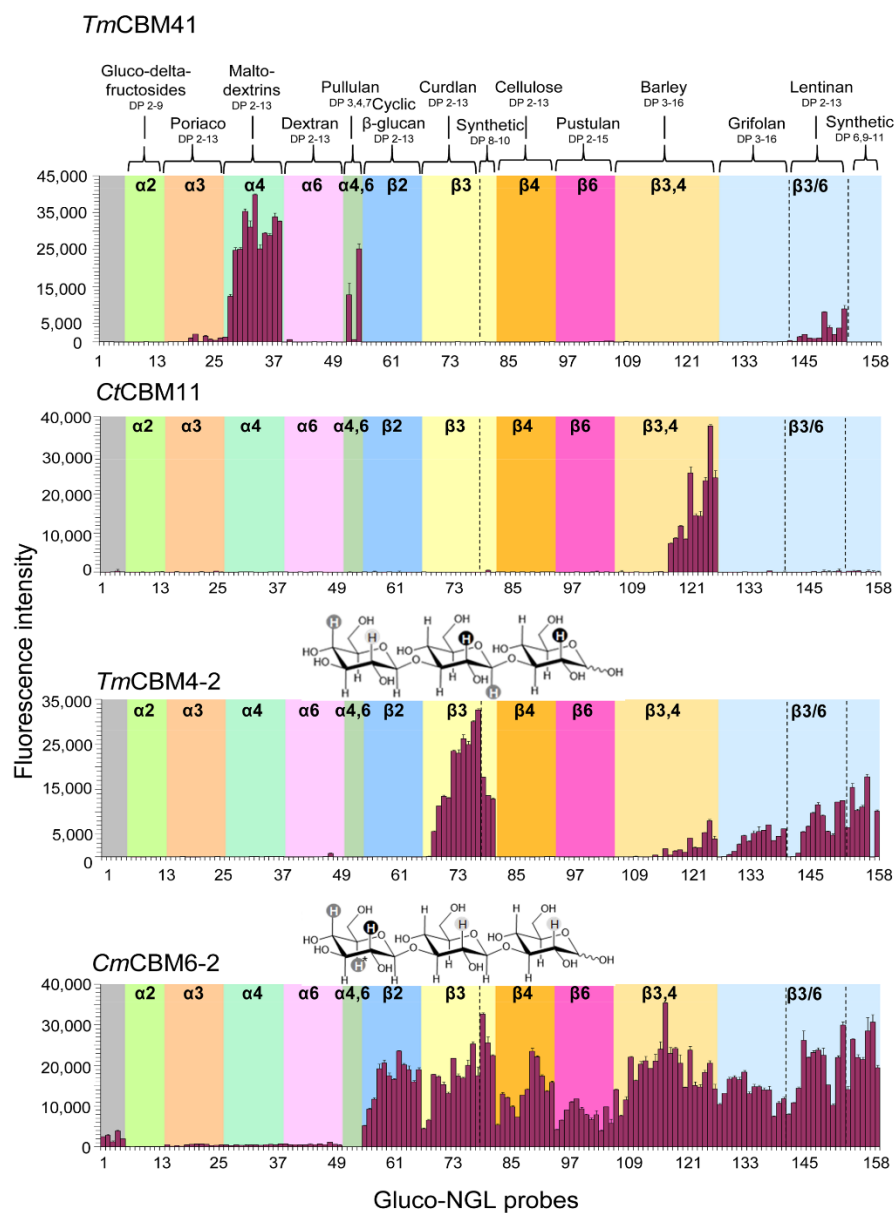


796

797

798

Figure 9



799

800

801

Gluco-NGL probes

Figure 10

802 The authors

803



Diana O. Ribeiro

Diana Ribeiro completed a PhD in Biochemistry in 2020 and has since worked as Junior Researcher at UCIBIO-NOVA. Her research interests focus on developing platforms that use carbohydrate microarray technology to allow the detailed characterisation of protein carbohydrate interactions in bacterial systems, such as in the human gut microbiome and host-pathogen interactions. She is also committed to creating web-based resources providing

813 curated structural and functional data, particularly on carbohydrate-binding modules
814 ligand interactions and bacterial polysaccharides.

815

816 **François Bonnardel**



François Bonnardel holds a PhD in Bioinformatics and Structural Biology delivered jointly from the University of Grenoble Alpes, France, and the University of Geneva, Switzerland, in 2021. During his PhD, he gained data science, programming, Development, Computational, Structural and System Biology competencies. He conceived and developed several databases, including the UniLectin portal, a worldwide recognised contribution for the exploration, classification and prediction of lectin embedding

828 UniLectin3D, PropLec and LectomeXplore databases. Since 2021, he has operated
829 as a Bioinformatics Pipeline developer at Inivata, NeoGenomics, in the UK.

830

831

832 **Ana Luísa M. Carvalho**



833 Ana Luísa completed a PhD in Structural Biology in 2002 and holds a position as Senior Researcher at UCIBIO-NOVA, starting in 2003. Her research interests focus on searching for uncharacterised biological protein-carbohydrate interactions with potential biotechnological use in nutrition, in pharma and the biofuel industry. For this, among other structural biology and biophysics methods, she uses the powerful information provided by X-ray Crystallography. Recently, she is committed to implementing Cryo-Electron Microscopy methods in UCIBIO and for this she coordinates the FCT-NOVA node

844 of the CryoEM-PT National Roadmap Infrastructure.

845 .

846 **Angelina S. Palma**



Angelina S. Palma completed PhD in Biochemistry in 2007, and presently is an Assistant Professor and leader of the Functional Glycobiology Lab at UCIBIO, FCT-NOVA. She developed pioneering work in applying glycan microarrays to study endogenous recognition systems and virus-host interactions. In 2013, she was awarded a 5-year FCT Investigator Grant to apply glycan-microarrays and structural biology to study glycan recognition by microbes. Her research fosters

856 knowledge in Glycobiology and glyco-technologies to study human microbiome
857 interactions and cancer. Currently, she co-coordinates the GLYCOTwinning
858 international network in Glycoscience, funded by Horizon Europe.

859

860

861

862

863

864 **Serge Perez**

Serge Perez holds a Doctorate es Sciences from the University of Grenoble, France. He had international exposure throughout several academic and industry positions in research laboratories in the U.S.A., Canada, and France (Centre de Recherches sur les Macromolécules Végétales, CNRS, Grenoble, as the chairperson and as Director of Research at the European Synchrotron Radiation Facility). His research interests span structural glycoscience and e-learning, for which he created

874 www.glycopedia.eu. He is involved in scientific societies as president and past
875 president of the European Carbohydrate Organisation. He is the author of more than
876 330 research publications and three books on the bioeconomy of natural resources.

877

878

879

880

881

882