scientific reports

OPEN

In silico analysis of the human milk oligosaccharide glycome reveals key enzymes of their biosynthesis

Andrew G. McDonald[1]✉, Julien Mariethoz[2,3], Gavin P. Davey[1] & Frédérique Lisacek[2,3]✉

Human milk oligosaccharides (HMOs) form the third most abundant component of human milk and are known to convey several benefits to the neonate, including protection from viral and bacterial pathogens, training of the immune system, and influencing the gut microbiome. As HMO production during lactation is driven by enzymes that are common to other glycosylation processes, we adapted a model of mucin-type GalNAc-linked glycosylation enzymes to act on free lactose. We identified a subset of 11 enzyme activities that can account for 206 of 226 distinct HMOs isolated from human milk and constructed a biosynthetic reaction network that identifies 5 new core HMO structures. A comparison of monosaccharide compositions demonstrated that the model was able to discriminate between two possible groups of intermediates between major subnetworks, and to assign possible structures to several previously uncharacterised HMOs. The effect of enzyme knockouts is presented, identifying β-1,4-galactosyltransferase and β-1,3-*N*-acetylglucosaminyltransferase as key enzyme activities involved in the generation of the observed HMO glycosylation patterns. The model also provides a synthesis chassis for the most common HMOs found in lactating mothers.

Human milk, aside from its value as a source of food for the new-born infant, has increasingly been shown to have many additional benefits in promoting development and providing protection from diseases. Human milk oligosaccharides (HMOs) are the third most abundant constituent of milk, after lactose, and lipids, with total mean concentrations ranging from 4–30 g/L[1]. The majority of these complex oligosaccharides are based on free lactose. Their compositions and structures have been the subject of several studies dating back to the 1950s, starting with the work of Kuhn[2] and others[3–5], and subsequent characterisations by Kobata, Ginsburg and coworkers[6–15]. Although indigestible by the neonate, HMOs provide several benefits to it, including antimicrobial action[16,17], protection from viral pathogens[18–21] and necrotising enterocolitis[22–24], promotion of a healthy gut microbiota[25,26] and the development of the immune and nervous systems[27–29]. HMOs may also play a role in preventing allergic reactions during childhood[30].

A feature unique to human milk is the heterogeneity of its HMO population, when compared to milk from other species, such as bovine, which are lower in both quantity and structural diversity[31]. Anti-adhesive properties of HMOs rely on competitive inhibition with pathogens for host receptors, as is the case, for example, with the finding that mono- and di-fucosylated oligosaccharides inhibit the binding of cholera toxin to glycosylated receptors of human epithelial cells[32]. By presenting a wide variety of epitopes, the human milk oligosaccharide is able to mask the newborn from such toxins, while simultaneously promoting a colonisation of beneficial gut flora of Bifidobacterial species[33].

In consequence, there has been much interest towards the synthesis of HMOs industrially, for use as probiotics, for which a number of approaches have been used, including direct chemical syntheses, use of recombinant glycosyltransferase (GT) activities[34]; transglucosidase reactions, in which glycohydrolases act in reverse to attach monosaccharides to oligosaccharides[35–37]; and through reconstruction of Leloir-type[38] pathways to synthesise the preferred HMO substrate of bifidobacteria, lacto-*N*-biose I[39]. Although little is known as to the actual enzyme activities expressed in the lactating mammary epithelium[40], some insights can be gleaned from analysis of the more than 200 HMO structures already characterised, many of which have been employed as model substrates of glycosidases and glycosyltransferases (for example,[41]).

Based on the observation that mammary gland is secretory in nature, with properties similar to those of mucosal surfaces[42], a view that has also been supported by the murine case[43], we assumed that some of the

[1]School of Biochemistry and Immunology, Trinity College Dublin, Dublin 2, Ireland. [2]Computer Science Department, University of Geneva, 1227 Geneva, Switzerland. [3]Proteome Informatics Group, SIB Swiss Institute of Bioinformatics, 1211 Geneva, Switzerland. ✉email: amcdonld@tcd.ie; frederique.lisacek@unige.ch

Glycologue single-letter code	SNFG symbol	IUPAC symbol	Definition	Assumed configuration
f	🔺	Fuc	L-fucose	α
G	🔵	Glc	D-glucose	β
L	🟡	Gal	D-galactose	β
S	◆	Neu5Ac	*N*-acetylneuraminate	α
Y	🟦	GlcNAc	*N*-acetyl-α-D-glucosamine	β
a, b	n/a	α, β	anomeric configuration	-

Table 1. Monosaccharide symbols. Definition of monosaccharide units of HMOs represented as single-letter codes in Glycologue, symbols in the Symbol Nomenclature For Glycans (SNFG) notation[58], IUPAC symbols. For example, *N*-acetylgalactosamine (GalNAc) is represented by Glycologue as 'V', and sulfate by the lowercase 's'.

enzymes involved in mucin biosynthesis would also be active in the production of HMOs. Our purpose in this article is to apply an adapted model of the enzymes involved in mucin-type (GalNAc-linked) O-glycosylation to act on free lactose, and to validate it against a large population of experimentally characterised HMOs, and subsequently predict a minimal biosynthetic reaction network. In doing so, we extend the previous work of GlycomeSeq[44], an existing in silico model of HMO biosynthesis developed as an aid in sequencing milk glycans and that was missing sialylated structures. We also complement the metabolic reconstruction approach of Bao et al. (2020)[45] that modelled and scrutinised the activity of ten glycosyltransferases in the synthesis of HMOs detected in cohorts including both secretor and non-secretor healthy mothers. The present article describes the model and the biosynthesis simulations performed on HMOs.

Results

Model description.

We developed a model of the enzymes of O-linked glycosylation[46] to allow their action on free lactose and predict possible human milk oligosaccharide products. The model uses a formal-grammar based approach to apply regular-expression based rules to model the transfer of monosaccharides, represented by a single-letter code, from activated sugars (sugar-nucleotide donors) to an oligosaccharide acceptor. The transformation rules are classified into a number of discrete types: extension, decoration, branching and termination (see Methods).

We compiled a library of 226 lactose-based human milk oligosaccharides from a variety of sources[40,47–55], as shown in Supplementary Table S1, which includes the IUPAC name, GlycoCT[56] structure encoding and a GlyTouCan[57] identifier, where available. The HMOs were found to be composed of the five monosaccharides, L-fucose, D-galactose, D-glucose, *N*-acetyl-D-glucosamine, with a small number with 6-*O*-sulfated residues (4) and a single occurrence of *N*-acetyl-D-galactosamine. A single-letter code was used to denote these monosaccharide units, as shown in Table 1. Free lactose, β-D-Gal*p*-(1,4)-D-Glc*p*-ol, is represented by [L4G]. Branched and decorated oligosaccharides are similarly represented by using bracketed notation, for example the monofucosylated HMOs 2′-FL and 3-FL are represented as [[f2]L4G] and [L4[f3]G].

A table of the most commonly occurring HMOs, ranked according to their frequency of occurrence in the cited literature, is available online at https://glycologue.org/m/sample.php.

Enzymes and reactions.

The enzymes of the model are shown in Table 2, numbered **1**–**11**, ordered by EC number, comprising two galactosyltransferases (enzymes **1** and **6**), three are *N*-acetylglucosaminyltransferases (**4**, **10**, **11**), with three fucosyltransferases (**2**, **3**, **5**) and three siayltransferases (**7**–**9**).

As shown in Fig. 1 (Linkage types), HMOs, in common with mucin-type O-glycans and other glycoconjugates, possess non-reducing termini that are based on Gal-β1,3-GlcNAc and Gal-β1,4-GlcNAc motifs, denoted type 1 and type 2[60,61], respectively, which are formed by the actions of enzymes **6** and **1** on a structure terminating in GlcNAc.

These basic determinants are named lacto-*N*-biose (LNB) and *N*-acetyllactosamine (LacNAc). De-galactosylated HMOs, terminating in GlcNAc, are rare, with only two representatives in Supplementary Table S1: GlcNAcβ1-3Galβ1-4Glc (LNTri II)[62], and one synthesised in vitro by Prudden et al.[48], GlcNAcβ1-3Galβ1-4GlcNAcβ1-3Galβ1-s4GlcNAcβ1-6 (Neu5Acα2-6Galβ1-4GlcNAcβ1-3)Galβ1-4Glc.

The smallest model network employing all of the activities of the enzymes in Table 2, is shown in Fig. 2.

Bifunctional enzymes.

Glycosyltransferases (GTs) are multi-specific, acting on a range of acceptors according to a recognition motif. It is known that many enzymes, including GTs, can display secondary activities, a behaviour sometimes called enzyme promiscuity[63]. Several such enzymes activities are included in the model. For example, in Table 2 the α2FucT enzyme (**3**) combines the activities of the type 1 and type 2 galactoside 2-α-L-fucosyltransferases, which transfer fucose to either lactose, to form 2′-FL, or lacto-*N*-biose or LacNAc termini. The activities of the other fucosyltransferases, α3FucT and α4FucT, are recognised separately; however, since α4FucT bears a 3-α-fucosyltransferase towards the glucose of free lactose[64], this was included as a secondary activity of enzyme **2**.

Enzyme no	EC number	Short name	Accepted name	Rhea Acc. no	Reaction pattern[a]
1	EC 2.4.1.38	β4GalT	β-*N*-acetylglucosaminylglycopeptide β-1,4-galactosyltransferase	22,932	UDP-L + *[Y* = UDP + *[Lb4Y*
2	EC 2.4.1.65	α4FucT	3-galactosyl-*N*-acetylglucosaminide 4-α-L-fucosyltransferase	23,628 62,888	GDP-f + *[Lb3Y* = GDP + *[Lb3[fa4]Y* GDP-f + [L4G] = GDP + [L4[f3]G]
3	EC 2.4.1.69 EC 2.4.1.344	α2FucT	galactoside 2-α-L-fucosyltransferase (type 1 & type 2)	50,664 50,668	GDP-f + *[Lb3Y* = GDP + *[[fa2]Lb3Y* GDP-f + *[Lb4Y* = GDP + *[[fa2]Lb4Y* (also acting on lactose)
4	EC 2.4.1.149	β3GnT (iGnT)	*N*-acetyllactosaminide β-1,3-*N*-acetylglucosaminyltransferase	14,389	UDP-Y + *[Lb4Y* = UDP + *[Yb3Lb4Y* (also acting on lactose)
5	EC 2.4.1.152	α3FucT	4-galactosyl-*N*-acetylglucosaminide 3-α-L-fucosyltransferase	14,257	GDP-f + *[Lb4Y* = GDP + *[Lb4[fa3]Y*
6	EC 2.4.1.86	β3GalT	*N*-acetyl-β-D-glucosaminide β-1,3-galactosyltransferase	53,432	UDP-L + *[Y3* = UDP + *[Lb3Y3*
7	EC 2.4.3.1	ST6Gal	β-galactoside α-2,6-sialyltransferase	52,104	CMP-S + *[Lb4Y* = CMP + *[Sa6Lb4Y* (also acting on lactose)
8	EC 2.4.3.6	ST3Gal	*N*-acetyllactosaminide α-2,3-sialyltransferase	52,317	CMP-S + *[Lb4Y* = CMP + *[Sa3Lb4Y* (can also act on type-1 acceptors)
9	EC 2.4.3.10	ST6GlcNAc	*N*-acetylglucoseaminide α-(2,6)-sialyltransferase	-	CMP-S + [Sa3Lb3Yb* = CMP + [Sa3Lb3[Sa6]Yb* (also acting on [Lb3Yb*)
10	EC 2.4.1.150	cIGnT	*N*-acetyllactosaminide β-1,6-*N*-acetylglucosaminyltransferase	54,821	UDP-Y + *[Lb4Yb3L* = UDP + *[[Yb6][Lb4Yb3]L* (inactive toward 3-FL substrates)
11	EC 2.4.1.386	dIGnT	GlcNAcβ1,3Gal β-1,6-*N*-acetylglucosaminyltransferase (distally acting)	-	UDP-Y + *[Yb3L* = UDP + *[[Yb6][Yb3]L* (inactive toward 3-FL substrates)

Table 2. Enzymes of the model. Proposed enzymes of HMO biosynthesis and the corresponding reaction definition in Rhea[59] (accession number) as well as Glycologue reaction patterns. [a]Asterisks act as a wildcard character, to denote an unspecified portion of the oligosaccharide. Symbols and abbreviations used in reaction patterns are those of Table 1, with the following additions: UDP, uridine 5′-diphosphate; GDP, guanosine. 5′-diphosphate; CMP, cytidine 5′-phosphate; CMP-S, CMP-*N*-acetyl-β-neuraminate; GDP-f , GDP-α-L-fucose; UDP-L, UDP-α-D-galactose;UDP-Y, UDP-*N*-acetyl-β-D-glucosamine.

Atypical sialylation. The ST6GlcNAc gene family is not found in humans[65], nevertheless several studies have demonstrated Neu5Ac 6-linked to the GlcNAc of lacto-*N*-biose, as for example in LST b and DSLNT[49–51,55]. Prudden and co-workers were able to synthesise the latter by means of the ST6GALNAC5[48], thus providing evidence of a potential candidate for the primary activity of EC 2.4.3.10 (**9**) in humans. Although the enzyme from rat liver is able to act on asialylated termini ([L3Y)[66], these are not substrates for ST6GALNAC5[67]. A suitable candidate for the secondary activity of **9** remains to be determined, thus its existence is inferred.

Branched HMOs. The model incorporates two separate I-branching (β-1,6-*N*-acetylglucosaminyltransferase) enzyme activities, one centrally acting and the other distally acting, relative to the reducing end of the oligosaccharide. The central activity, named cIGnT, is that of EC 2.4.1.150, which catalyses the reaction UDP-GlcNAc + ∗[Lb4Yb3L∗ = UDP + ∗[[Yb6][Lb4Yb3]L∗ having a preference for oligosaccharides with type 2 termination[68], i.e., Gal β-1,4-linked to GlcNAc. The second IGnT enzyme, dIGnT acts on the predistal galactose before a terminal GlcNAc, and is a secondary activity of the mucin-type core-4 forming enzyme, C2/4GnT[69,70]. Based on the observation that none of the characterised HMOs that were 3-fucosylated on the base glucose were branched, an additional assumption of the model was that the I-branching enzymes **10** and **11** are inactive towards these substrates.

Performance of the simulator. Starting from lactose, with all 11 enzymes active, 206 of the 226 HMOs in Supplementary Table S1 were predicted in silico, a prediction rate, or "coverage", of 91.1%. When considering the HMOs common to more than one study, 85 out of the 89 structures listed in the online table were obtainable in simulations, for a 96% coverage. Owing to a combinatorial explosion in the number of structures formed[46,71], simulations were limited to a user-defined maximum number of GlcNAc residues incorporated into HMO acceptor-products. The number of structures appearing with each iteration of the enzyme simulator increased logistically with iteration number (Fig. 3), when HMOs were limited to between three and six GlcNAc residues.

Reaction networks eventually closed, with no further products being added at higher iteration numbers (see Methods). The maximum number of observed HMOs found in the simulations, while minimising the total number of theoretical structures, was found to occur at iteration 11, when the maximum number of GlcNAc residues per HMO was limited to four.

Networks and HMO classification. At 11 iterations of the method, and with a limit of four GlcNAc residues incorporated per acceptor, the simulator generated 10,821 unique HMO structures, shown in the reaction network in Fig. 4A, where each node is coloured according to the base core structure of the HMO it represents, as shown in Fig. 1 (Base core patterns).

Chen[47] identified 16 distinct core HMO structures, non-fucosylated and neutrally charged, composed only of alternating galactose and *N*-acetylglucosamine units and the glucose moiety of lactose. Although lactose is

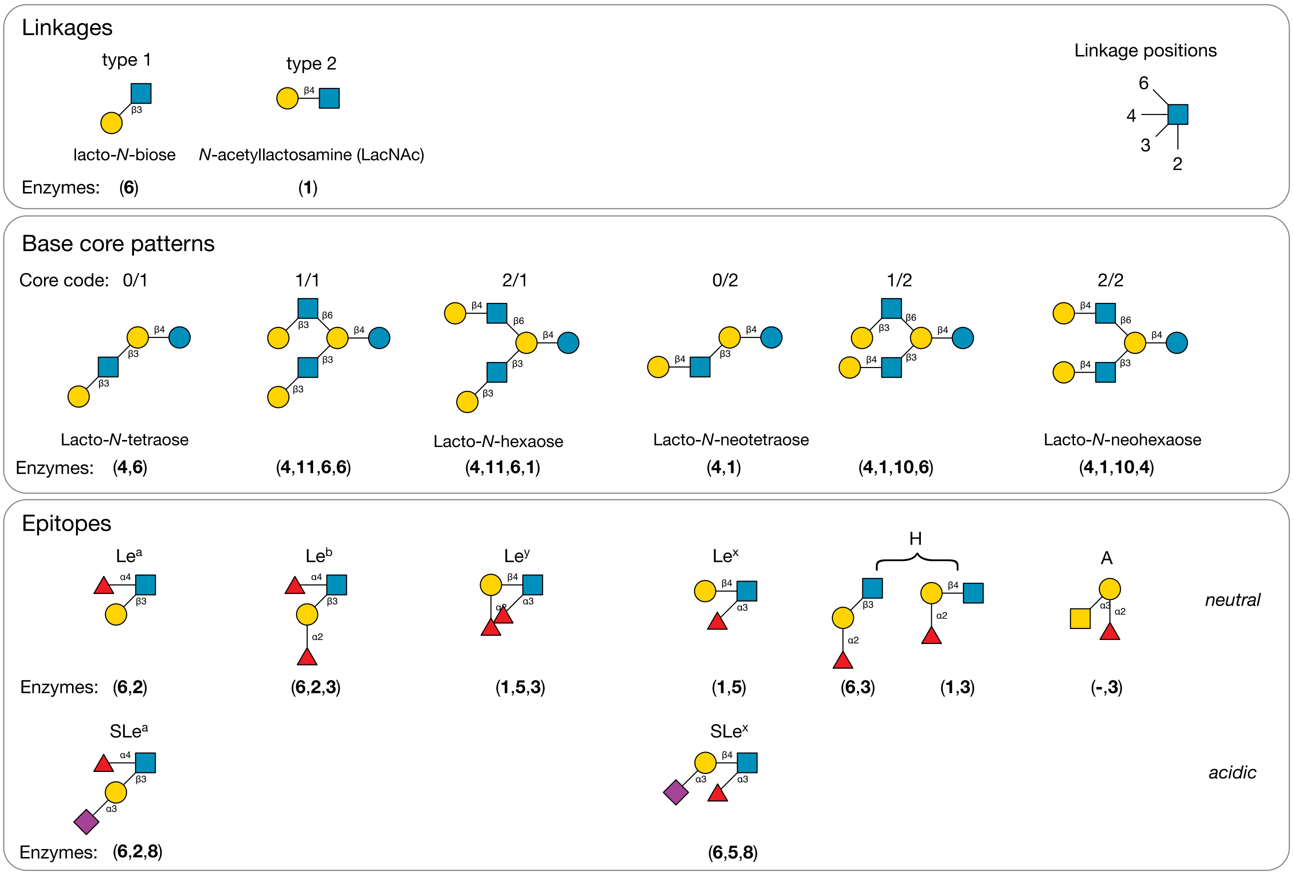

Figure 1. Structural motifs common to HMOs. Core structures are defined as lactose-based oligosaccharides that contain only hexose sugars, Glc and Gal, and the *N*-acetylhexosamine sugar, GlcNAc. Not all cores and epitopes are found within the experimentally determined HMO structures considered by this study. Base core patterns are defined as *a*/*b*, where *a* and *b* refer to the type-1 or -2 linkage of the terminal Gal, either to the 6-linked (*a*) or the 3-linked GlcNAc (*b*) of an I-antigen branch appearing on lactose. Proposed sequences of enzymes (indices of the enzymes in Table 2) involved in the biosynthesis of each motif are displayed beneath its structure. A complete set of core structures is in Supplementary Table S2.

not itself classed as an HMO, under this classification it is Core I. Urashima et al.[40] extended this number of core structures to 19 (I–XIX). Among the HMOs considered in the present study, we proposed an additional five novel cores to those of the latter classification. The full set of HMO core structures found is given in Supplementary Table S2.

Since both earlier classifications provided a large number of distinct sub-populations that were found not to correlate spatially in layouts of the biosynthetic networks (Fig. 4), it made their interpretation difficult. For easier visualisation of the networks, we adopted a simpler classification system based on the initial actions of β3GnT (**4**), followed by those of the two galactosyltransferases, β4GalT (**1**) and β3GalT (**6**), and then the branching *N*-acetylglucosaminyltransferases, cIGnT (**10**) and dIGnT (**11**). This approach divides the oligosaccharides into six classes: two that are linear (types 1 and 2, terminating in β-1,3-Gal and β-1,4-Gal, respectively), and four that combined the actions of cIGnT and dIGnT with subsequent extension by β3GalT and β4GalT. We denoted these four branching possibilities by *a*/*b*, where *a* signifies the type of the 6-linked GlcNAc (upper branch), and *b* the 3-linked GlcNAc (lower branch), leading to four combinations of types 1 and 2, namely, 1/1, 1/2, 2/1 and 2/2. In common biochemical nomenclature, 0/1 is lacto-*N*-tetraose (LNT; Core II), 2/1 is lacto-*N*-hexaose (LNH; Core V), 0/2 is lacto-*N*-neotetraose (LNnT; Core III) and 2/2 is lacto-*N*-neohexaose (LNnH; Core VI).

Figure 4B highlights the observed structures, and their distribution through the network. All regions of the main network have observed counterparts, with the exceptions of the 1/1 (magenta) or 1/2 (green) branching pattern. For the 206 observed HMOs predicted by the model, a minimal biosynthetic network was constructed by reversing the enzymes of glycosylation and compiling a network of all of the reversed-reversed reactions leading from lactose to observed products. The resulting reduced network is shown in Fig. 4C and D, coloured according to the base-core scheme of Fig. 1. The network with 514 HMOs (nodes), including intermediates, and 966 distinct reactions (edges). In Fig. 4D, only the nodes matching the predicted, experimentally observed, HMOs of Supplementary Table S1 are coloured (206 nodes). On account of the multiantennary nature of many HMOs (167 of the 226 HMOs studied bear at least one β-6-GlcNAc residue), multiple routes to the same product are possible, which results in a lattice-like appearance of the networks. This is also seen in other studies of glycosylation reaction networks, including those of HMOs, N-linked and O-linked glycans[45,46,73].

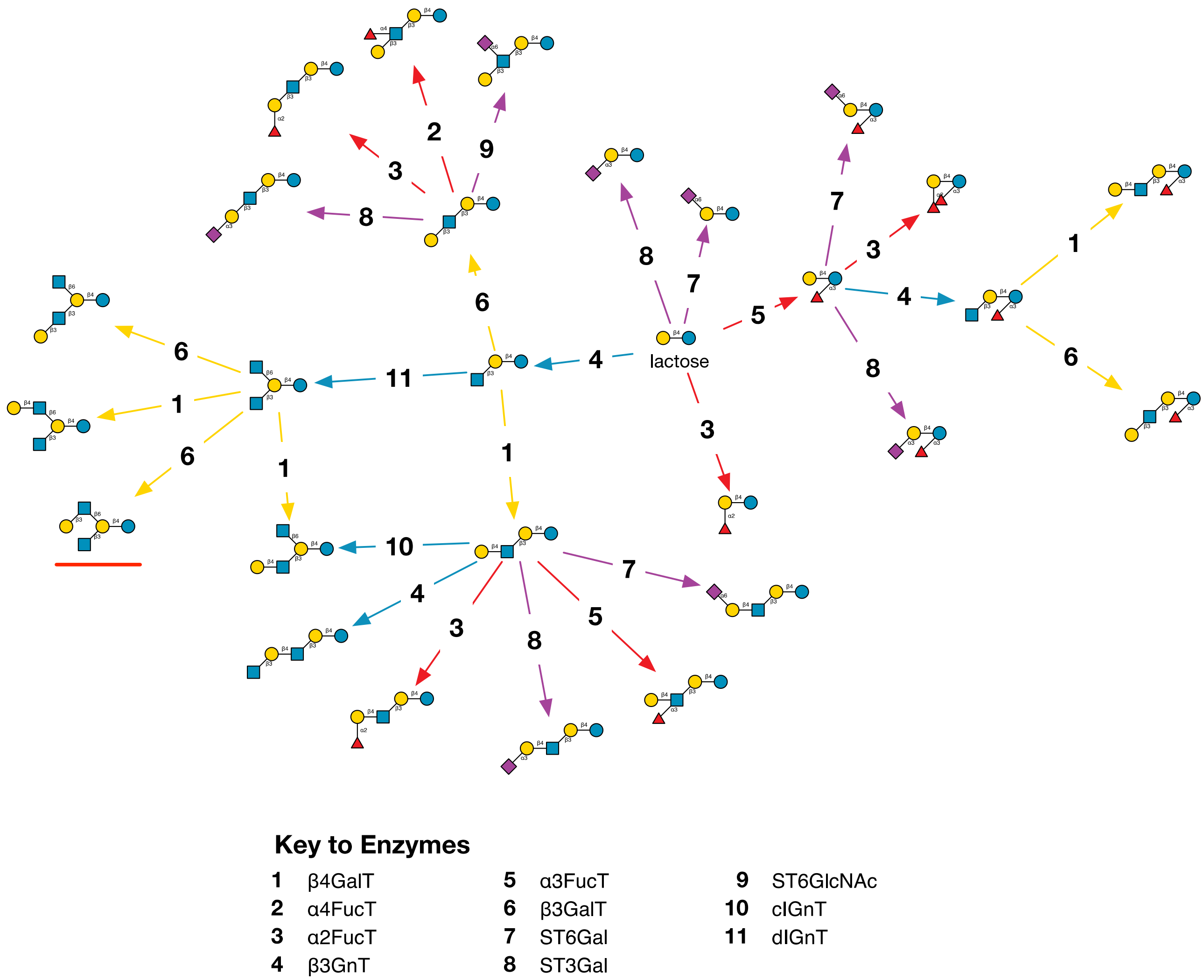

Figure 2. Reactions of the model. A reaction network generated by three iterations of the enzyme simulator, starting from lactose. Enzymes reactions are represented arrows leading from an acceptor substrate to an acceptor product, coloured according to the type of monosaccharide transferred: GlcNAc (blue), Gal (yellow), Fuc (red), Neu5Ac (purple). For simplicity, donor substrates and nucleotide products are omitted. The structure underlined in red was not found among the observational data set (Supplementary Table S1).

A review of HMO concentrations in mature human milk, pooled from 57 studies published between 1966 and 2020, identified the 15 most abundant oligosaccharides[74]. All 15 HMOs were predicted within the model, and a biosynthetic network was constructed (Supplementary Figure S1), in which all of the enzymes except the cIGnT activity (**10**) were used. It was observed that none of the most abundant HMOs were of the base-core 2/2, but were linear 0/1, 0/2, or branched 2/1, according to the classification scheme employed here.

Novel cores/delayed branching.

A possible biosynthetic pathway of eighteen of the core structures I–XIX, along with the structures of the five novel cores, is shown in Fig. 5A.

The only omission from the network is core IV, which was not predicted (see "Discussion", Other enzyme activities). Given that model predicts a subpopulation of 1/1 and 1/2 base-core structures, which are not part of the observational dataset, of interest are the existence of the "delayed branching" structures, *novo*-LNnO[40,76], which is Core XIII[40], and its monofucosylated derivative, F-*novo*-LNnO[40], since they display the 1/1 branching pattern elsewhere in the molecule. An additional core structure, not included in the HMO library, is F-*novo*-LNO, which is type 2 on the upper and type 1 on the lower arm of the branch, and 4-α-fucosylated on the GlcNAc of the terminal lacto-*N*-biose[76]. All three HMOs are predicted by the model, which are defined as 0/2 according to the base-core classification. Their Glycologue structure identifiers (see Supplementary Table S2 for IUPAC and GlycoCT condensed equivalents), trivial names and simulated biosynthetic enzyme sequences are:

[[L4Y6][L4Y3]L4Y3L4G]	*novo*-LNnO	**(4,1,4,1,10,1)**
[[L4Y6][L4[f3]Y3]L4Y3L4G]	F-*novo*-LNnO	**(4,1,4,1,10,1,5)**
[[L4Y6][L3[f4]Y3]L4Y3L4G]	F-*novo*-LNO	**(4,1,4,6,11,2,1)**

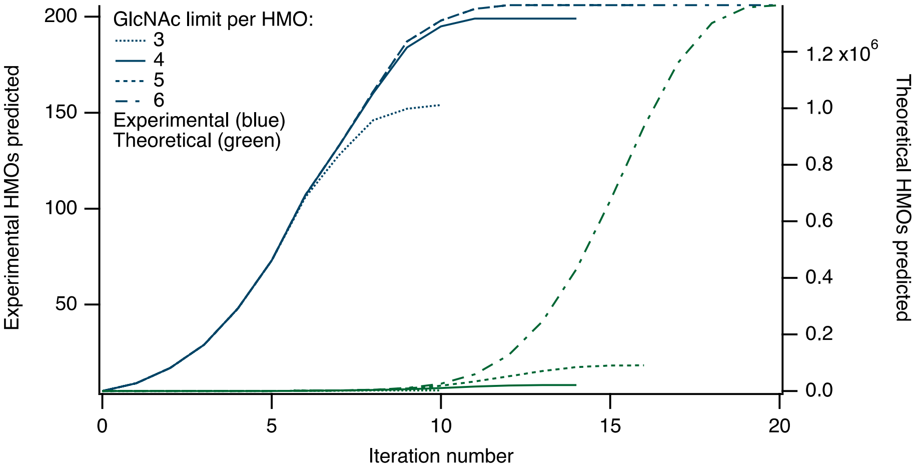

Figure 3. Growth of HMOs predicted by the simulator. Numbers of experimental (left axis) and theoretical (right axis) HMOs predicted, with increasing iteration number, for simulations limited to between 3 and 6 GlcNAc residues per oligosaccharide. At the final point on each curve, reaction networks were closed, with no further structures being added at higher iteration values.

In each case, cIGnT or dIGnT acts on [L3Y3L4Y3L4G] (*para*-Lacto-*N*-neohexaose; Core VII) or [L4Y3L4Y3L4G] (*para*-Lacto-*N*-neohexaose; Core VIII), respectively. The question remains as to the specificity of the 3-β-galactosyltransferase enzyme responsible for extending the 6-linked GlcNAc, whether it is blocked when that residue is immediately connected to lactose, viz., [[Y6][L3Y3]L4G] and [[Y6][L3Y3]L4G].

Biosynthesis of *inverse*-LNnD. A biosynthetic pathway of the difucosylated *inverse*-LNnD[77], with structure identifier [[L4Y6][[L3[f4]Y6][L3[f4]Y3]L4Y3]L4G] (Supplementary Table S1), and composition H6N4, is shown in Fig. 5B. It requires the actions of the two GalT enzymes, and is the product of a set of possible sequences of activities, such as (**4**,**1**,**10**,**1**,**4**,**11**,**6**,**2**,**6**,**2**), when acting on lactose as initial substrate. It is based on lacto-*N*-neohexaose , LnNH, a core-2/2 structure that has been reported in several studies[48,49,53,54], and which is itself derived from the core-0/2 lacto-*N*-neotetraose (LNnT)[48–51,53,54]. Of note is the (**4**,**11**,**6**) subsequence that is indicative of the formation of the 1/1 motif, even though its base core is 2/2 (cf. Fig. 1: Base core patterns). In Fig. 5B the position of the novel core structure, with Glycologue identifier [[L4Y6][[L3Y6][L3Y3]L4Y3]L4G], is marked by an arrow (*cf.* Supplementary Table S2).

HMO compositions vs. structures. As a test of the model, we considered the monosaccharide composition of HMOs as a crude grouping of structures. We entered the compositions of the sub-population defined above into GlyConnect Compozitor[78,79], a tool that roughly simulates the incremental addition of monosaccharides from a nucleotide sugar to an acceptor. The network of compositions is shown in Fig. 6 where we use the condensed notation (hexose = H, hexosamine = N, fucose = F, sialic acid = S and sulfate = s).

In this process, missing intermediates are inferred as virtual nodes depicted in grey. The HMO structures of our library represent 69 compositions stemming from lactose (Hex_2 = H2) (Supplementary Fig. S2) and three from LacNAc (H1N1)) (Supplementary Fig. S3). In these Supplementary figures, the full network is shown with paths highlighted in orange. This colouring is triggered by hovering the mouse on a node to visualise the reachability of and to other nodes from this point (incoming arrows in orange, outgoing arrows in turquoise). The size of the node reflects the number of publications confirming the presence of the corresponding composition. Candidate structures matching a given composition are suggested as part of the library described above that is stored in GlyConnect[80] and accessible in the dedicated HMO section (https://glyconnect.expasy.org/hmo).

We examined the missing compositions as revealed by the virtual grey nodes, especially those with the greatest influence on the connectivity of the graph. In particular, the absence of intermediary structures between H6N4F1 (matching 8 known structures) and H7N5F1 (matching 3 known structures) led Compozitor to consider H7N4F1 or H6N5F1 as potential connectors. Clearly, if not for those virtual nodes the graph would be disrupted. In the same way, missing data connecting H7N5F2 (matching 5 known structures) and H8N6F2 (matching one known structure) are proposed as H7N6F2 or H8N5F2.

The model validated H6N5F1, as 925 generated structures matched this composition, whereas none were associated with H7N4F1. The enzymes of Table 2 can be divided into subsets based on the type of monosaccharide being transferred. Thus, the set of hexosyltransferases are $H_E = \{\mathbf{1},\mathbf{6}\}$, *N*-acetylglucosaminyltransferases are $N_E = \{\mathbf{4},\mathbf{10},\mathbf{11}\}$, fucosyltransferases are $F_E = \{\mathbf{2},\mathbf{3},\mathbf{5}\}$ and sialyltransferases are $S_E = \{\mathbf{7},\mathbf{8},\mathbf{9}\}$. The graph indicates a preference for GlcNAc first and Hexose second, which could be explained by the mutual interplay of subsets H_E and N_E, each of whose members, in our model, act on the products of the other. Starting from lactose (H2), we expect members of N_E to act first, followed by those of H_E, giving H(2 + i)N(i) as the expected composition pattern of cores. By the same reasoning, we would expect a route from H7N5F2 to H8N6F2 to pass through

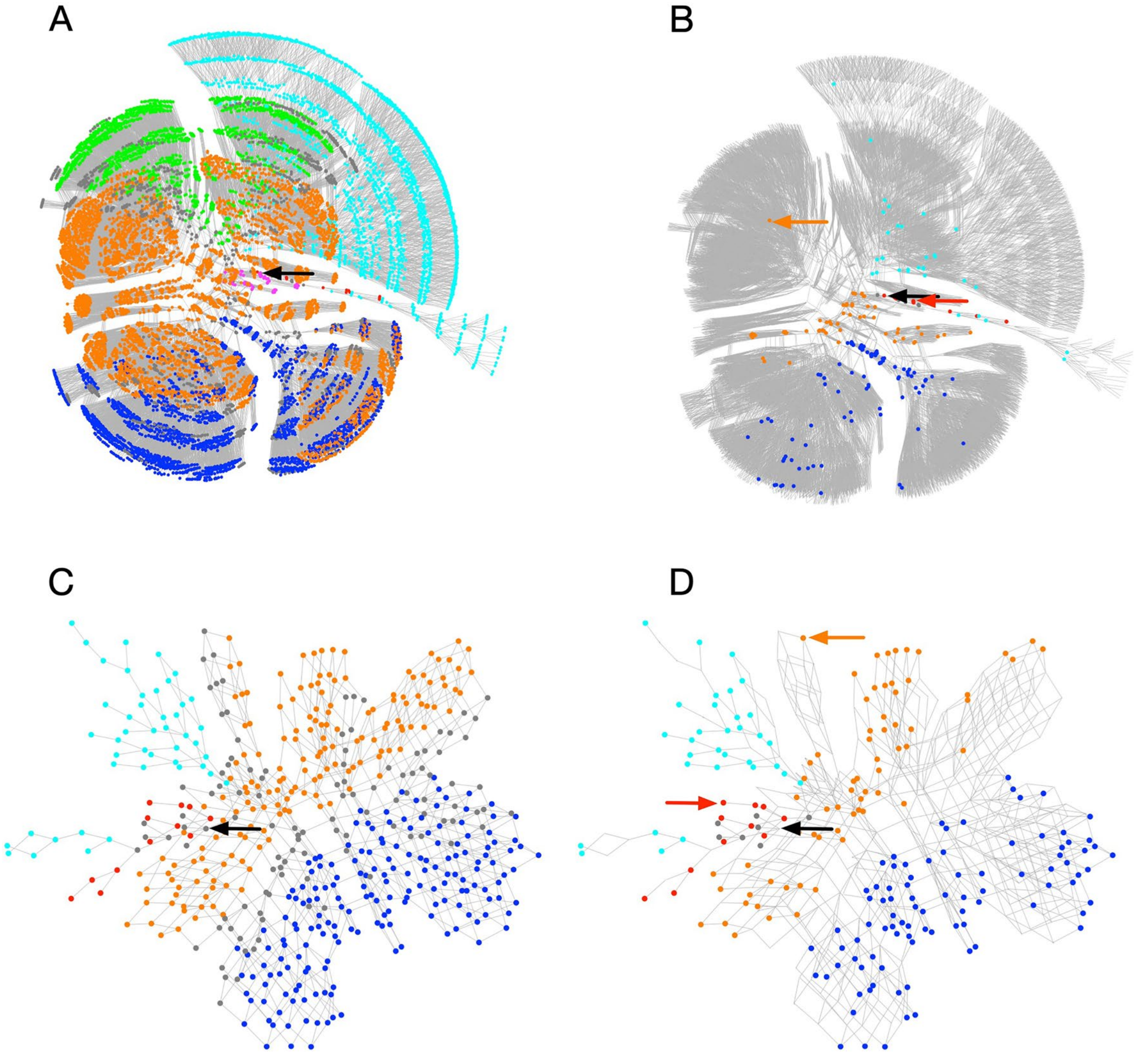

Figure 4. Calculated biosynthetic pathways of human milk oligosaccharides. (**A**) Simulated network of the 11 enzymes of Table 2, limiting to 11 iterations, with each HMO limited to a maximum of 4 GlcNAc residues. (**B**) As A, with 195 experimentally characterised HMOs highlighted and coloured, with theoretical structures shown in grey. (**C**) A minimal reaction network leading to the population of 206 HMOs from the library of observed HMOs (Supplementary Table S1). (**D**) As C, with only the experimentally observed HMOs highlighted. The position of the starting substrate, lactose, is indicated by a black arrow. Nodes are coloured according to the base core configuration given in Fig. 1: magenta (1/1), blue (2/1), green (1/2), orange (2/2), red (0/1), cyan (0/2), grey (other, unclassified). The locations of DSLNT and *inverse*-LNnD, within networks B and D, are indicated by red and orange arrows, respectively. Networks were drawn in Tulip [72] using a stress-minimization layout algorithm.

virtual node H7N6F2 (N + 1, H + 1), in preference to H8N5F2 (H + 1, N + 1). This was verified by simulations, which predicted 12,955 of the former composition, but none of the latter.

Two virtual nodes represent the intermediary linear structures between existing H3, H5 and H7. These are based on preliminary structure assignments from a set of HMOs not included in our validation set, and which are likely to be novel linear chain polygalactosyllactoses (see "Discussion", Other enzyme activities).

The last two virtual nodes are redundant, in the sense that their removal would not disrupt connectivity of the network. However, in their presence, H6N4 (matching 2 known structures) is reachable from 14 nodes in the network and from lactose in particular (blue outgoing arrows in Supplementary Fig. S4). When virtual nodes are not considered then H6N4 is not only reachable through H6N3 or H5N4 and therefore not from the original lactose.

The dataset included in[81] contains 102 HMO compositions some of which representing unexpected extra-large HMOs with a mass greater than 4 KDa. This larger dataset was compared to our library revealing a 70% overlap (Supplementary Fig. S5). Disregarding the extra-large HMOs, 95% of additional compositions were connected to the main network via virtual or existing nodes. Fourteen virtual nodes were required to complete which for most are new in comparison to the six virtual nodes needed in our library.

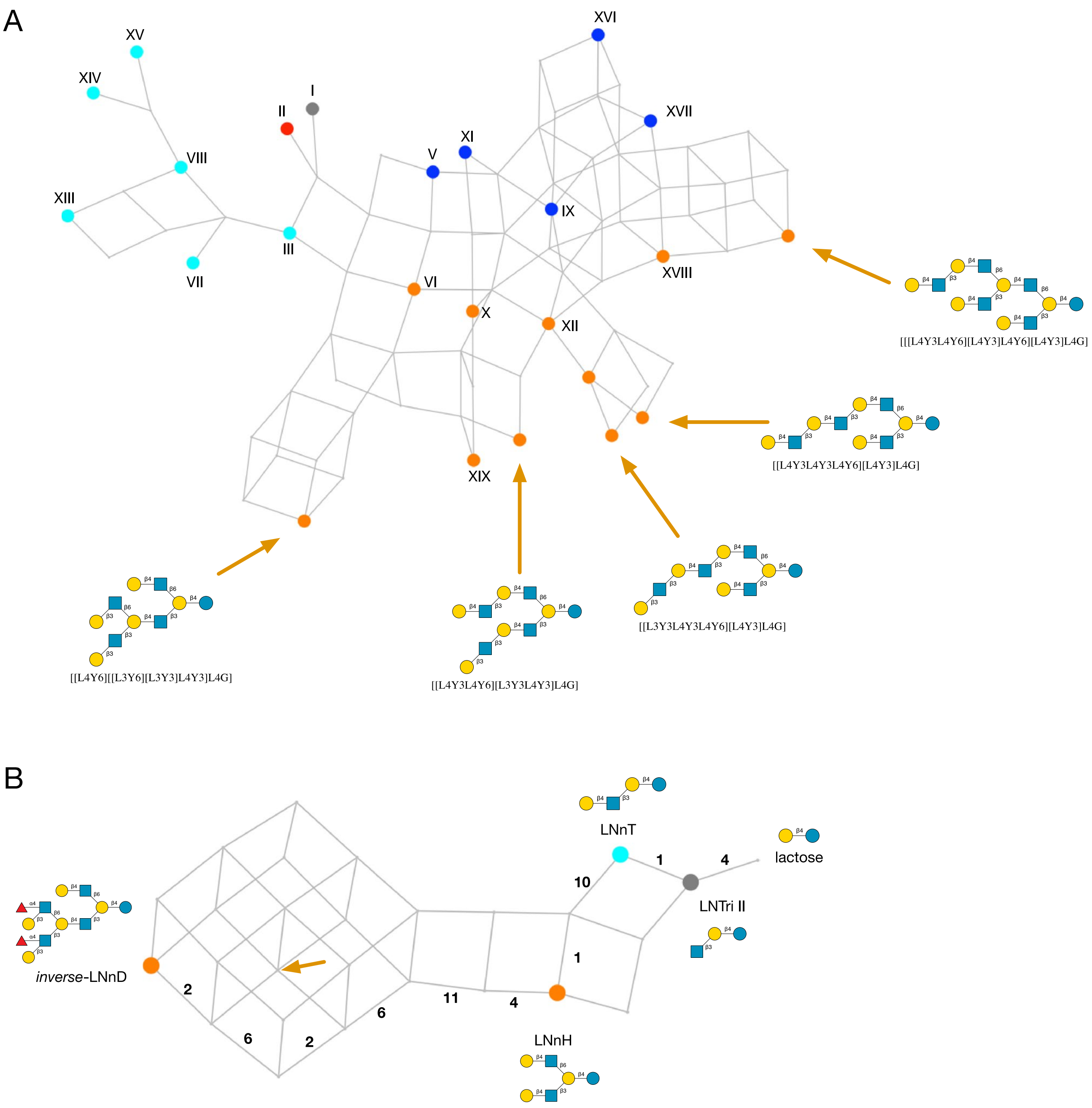
A
XVI
XV
XIV
I
II
XVII
VIII
XI
V
XIII
IX
III
VII
VI
XVIII
X
XII
XIX
[[[L4Y3L4Y6][L4Y3]L4Y6][L4Y3]L4G]
[[L4Y3L4Y3L4Y6][L4Y3]L4G]
[[L3Y3L4Y3L4Y6][L4Y3]L4G]
[[L4Y6][[L3Y6][L3Y3]L4Y3]L4G]
[[L4Y3L4Y6][L3Y3L4Y3]L4G]
B
LNnT
lactose
1
4
10
LNTri II
inverse-LNnD
1
2
11
4
6
6
2
LNnH

Figure 5. Biosynthesis of HMO core structures. (**A**) Network of HMO core structures, labelled I–XIX [40] and with five newly identified cores indicated by arrows, with their Glycologue structural identifiers (see Supplementary Table S2 for IUPAC and GlycoCT condensed formats). (**B**) Proposed biosynthetic network from lactose to *inverse*-LNnD [75], labelled with observed intermediates. Multiple routes to the same product are shown, with the enzyme numbers labelling the edges (reactions) of one possible route. The position of the novel HMO core structure is shown by the orange arrow. Nodes are coloured according to the base-core code of Fig. 4.

Epitopes. HMOs display a wide variety of epitopes, as shown in Fig. 7. The expression of antigenic determinants on HMOs is a function of genetic blood group type and the secretor status of the mother[82–85]. In general, the simulation output matched the percentages of structures with each epitope, except for the Lewis (S)Le^x and Le^y, which are predicted at higher proportions than in the HMO sample library. Since the model lacks an α3GalNAcT enzyme, the only A-antigen bearing HMO, A-hepta[54], was not predicted.

Lewis b occurs less frequently on the lower arm of an I-branched type-2 structure, a motif observed in one HMO by Remoroza et al.[49], although not by Wu et al.[54]. Several examples occur among the HMOs synthesised by Prudden and co-workers[48].

Enzyme knockouts. The effect of knocking out enzyme activities in silico, on coverage of the HMOs in the sample library was investigated. With the null knockout represented by **0**, each activity, **1**–**11** of Table 2 was

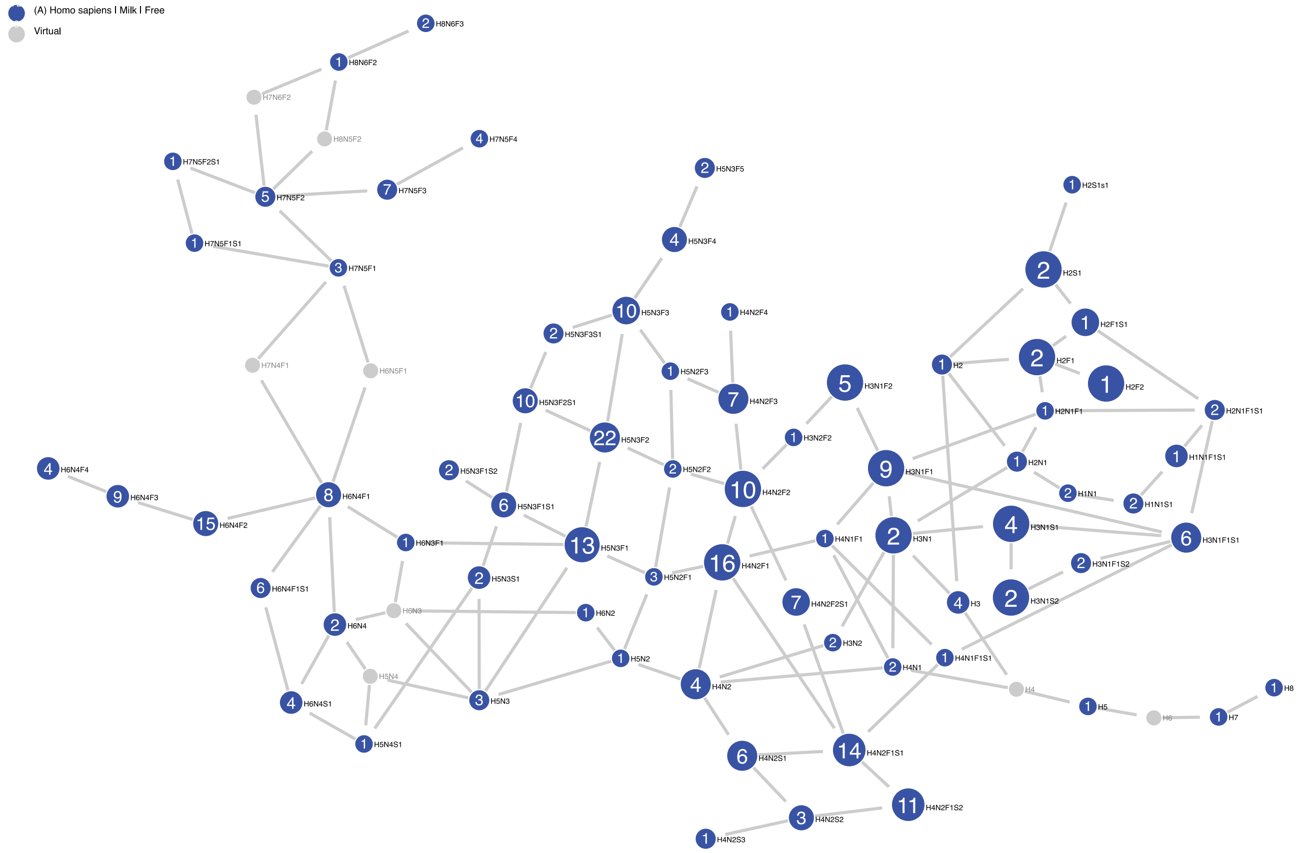

Figure 6. Network of HMO compositions. Nodes represent distinct compositions of hexose, *N*-acetylhexosamine, fucose and sialic acid residues, and differ by one monosaccharide unit from each of their nearest neighbours. Numbers on blue nodes refer to the number of structures in GlyConnect [80] with the composition. Virtual nodes (grey) represent unknown structural intermediates.

disabled in conjunction with another, to form single (***x*/0**) or dual (***x*/*y***) knockouts. The results are summarised in Fig. 8A, in which it is evident that the null knockout, (**0/0**), resulted in the highest coverage of the sample data (91.1%), while the other knockouts decreased the coverage to varying degrees.

Key glycosyltransferases involved in heterogeneity of HMOs are β4GalT (**1**) and β3GnT (**4**), which when knocked out individually resulted in the lowest coverage values consistently. The lowest coverage, of 1.3%, was obtained with the dual knockout of the α4FucT and β3GnT (iGnT) activities. The distal IGnT activity (**11**), having a broader specificity than the central IGnT (**10**), has the greater influence on heterogeneity. The enzymes ranked according to their influence on overall overage of the HMO sample population (the diagonal elements of the square matrix in Fig. 8), are

$$\beta 3\mathrm{GnT} > \beta 4\mathrm{GalT} > \beta 3\mathrm{GalT} > \alpha 3\mathrm{FucT} \ > \ \mathrm{dIGnT} \ > \alpha 2\mathrm{FucT} \ > \alpha 4\mathrm{FucT} \\ > \ \mathrm{ST6Gal} \ > \ \mathrm{ST3Gal} \ > \ \mathrm{ST6GlcNAc} \ > \ \mathrm{null} \ = \ \mathrm{cIGnT}$$

where A > B results in a greater reduction in coverage of the library population when enzyme A is knocked out, compared to the knockout of enzyme B. Maximal coverage was attained with the control (null) knockout, equal to that of the cIGnT knockout.

HMO antigens vary widely between different human populations, but are broadly categorised according to the secretor (Se) and Lewis (Le) status of the mother. Minimal biosynthetic networks corresponding to non-secretor and/or Lewis-negative mothers were constructed by simulating FUT2 and FUT3 genetic knockouts, corresponding to the elimination of 2-α- and (3/4)-α-ʟ-fucosyltransferase activities of Table 2. A single knockout of α2FucT (**3**) removes all H antigen, and Le^b and Le^y epitopes; the resulting network, with all other enzymes available, is shown in Fig. 8B. A FUT3 knockout was modelled through elimination of both the α3FucT and α3FucT activities (enzymes **2** and **5**), with the result shown in Fig. 8C. A still smaller population of HMOs is predicted when all three fucosyltransferase activities were removed (Fig. 8D).

Discussion

We have shown that 11 enzyme activities of the model can account for more than 90% of all HMOs analysed in this study, for which we have computed possible reaction networks. We proposed a biosynthetic pathway leading to *inverse*-LnND, which has revealed a novel HMO core, one of five such cores disclosed by this study. Linking composition data to enzyme activities has enabled us to discriminate between possible routes, where

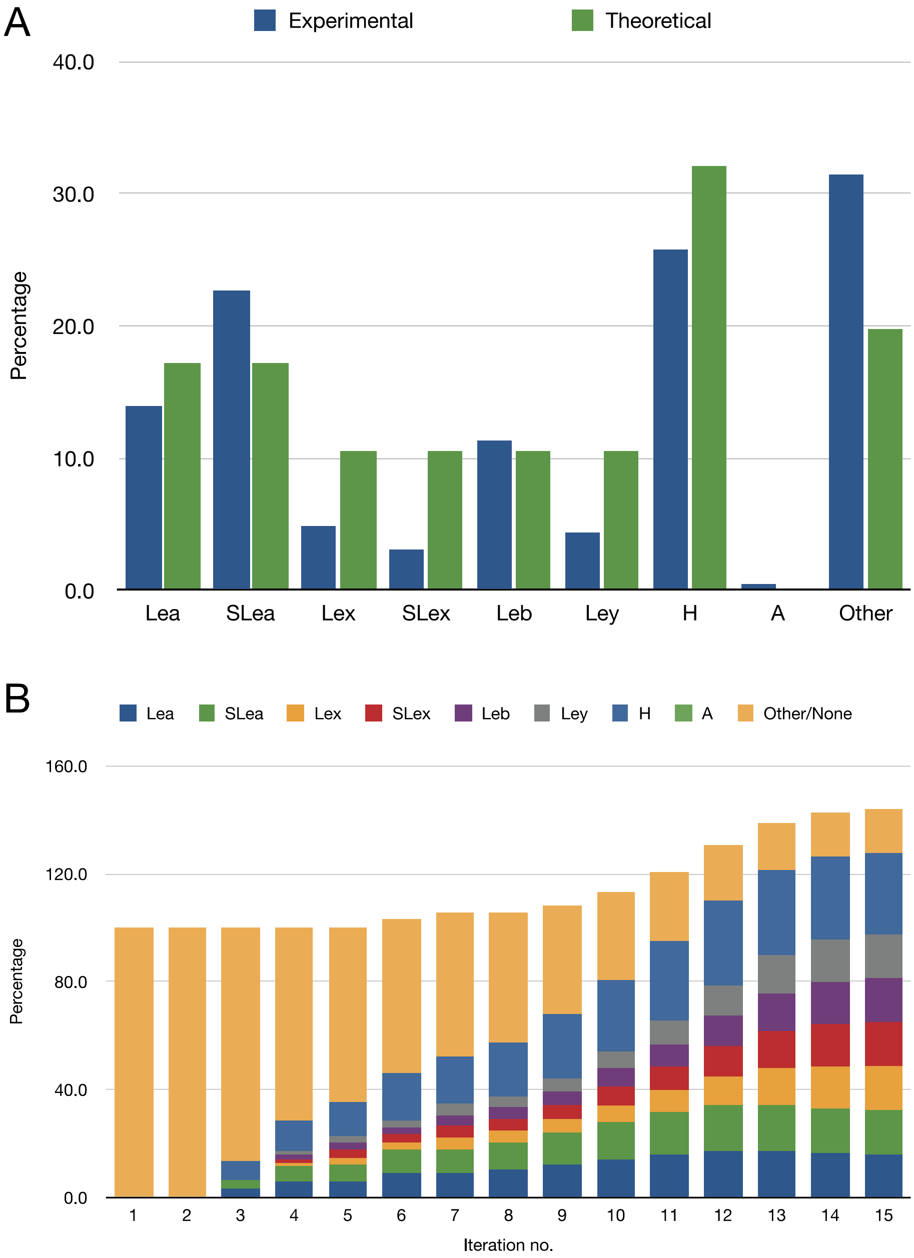

Figure 7. Numbers of epitopes appearing in HMO populations, real and simulated, expressed as a percentage of the total number of structures. (**A**) Percentages of the total numbers of each epitope in the experimental ($N = 226$) and simulated ($N = 51,982$; $i = 13$) HMO populations. As a result of multi-antennarity (branching), more than one epitope can be present on a single HMO. (**B**) Percentages of epitopes appearing in simulations, to which a limit of four GlcNAc residues per HMO was applied, as a function of iteration number (i).

unknown intermediates are inferred to exist. While the HMO-Glycologue simulator can be tailored for 2^{11} possible phenotypes, the single- and dual- knockouts of enzyme activities enabled us to rank the enzymes according to their degree of influence on the observed HMO population. The influence of β3GnT and β4GalT activities on heterogeneity is not unexpected, as both are involved in the extension of oligosaccharides via LacNAc repeats.

Kinetic and genetic regulation of HMO core types. That the majority of branched HMOs fall into the two 2/1 and 2/2 base-core categories could be explained by the higher activity of EC 2.4.1.38 (**1**) towards GlcNAc-β1,6-Gal than to GlcNAc-β1,3-Gal[86]. This would suggest that a kinetic competition exists between the two galactosyltransferases (**1**,**6**) that favours type-2 termination on the 6-linked arm. Kinetic competition might also explain the asymmetry in the existence of LacNAc repeats among the sample population, which appear only on the 6-linked GlcNAc. Of the 226 HMOs in Supplementary Table S1, there are no occurrences of LacNAc-extended (type-2) terminations of "lower" branch and three occurrences of lacto-*N*-biose extended type 1; of

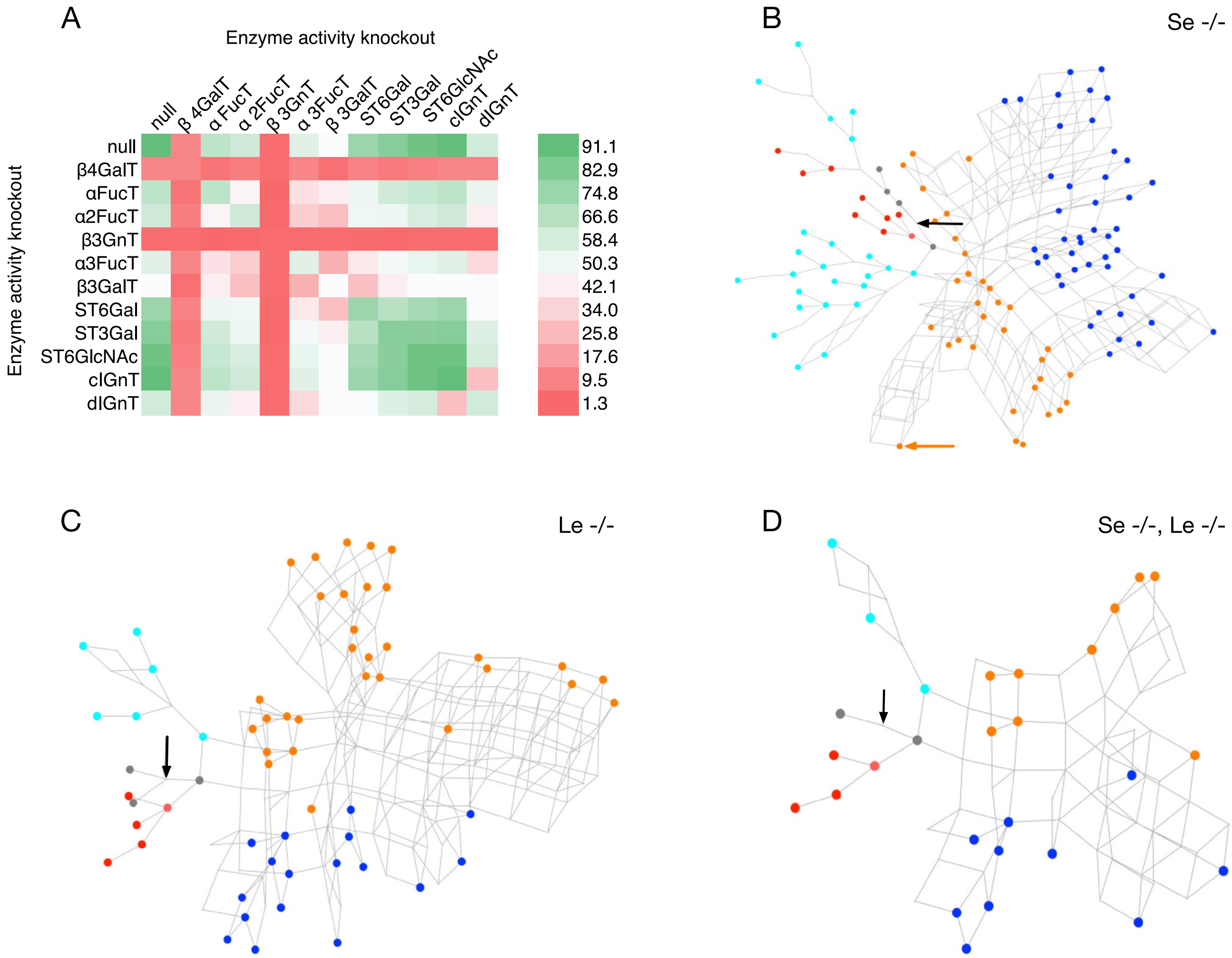

Figure 8. Simulated knockouts of the enzyme activities. (**A**) Effects of dual knockouts of the activities in Table 2 on the percentage coverage of the HMO structure library (Supplementary Table S1). In the null/null knockout all enzymes remain active. Colours range from purple (minimal coverage: 1.3%) to yellow (maximal coverage: 91.1%), corresponding to the null/null knockout (**0**/**0**) and the α4FucT/β3GnT knockout (**2**/**4**), respectively. (**B**–**D**) Predicted HMO biosynthetic networks of (B) non-secretor (Se-/-), (C) Lewis-negative (Le-/-) and (D) non-secretor/Lewis-negative mothers. Observed HMOs are coloured according to their base-core value (Fig. 1): magenta (1/1), blue (2/1), green (1/2), orange (2/2), red (1), cyan (2), grey (other/unclassified). In each network the position of lactose is indicated by a black arrow. In the Se-/- network (B), an orange arrow indicates the position of *inverse*-LNnD.

the "upper" branch, there are 28 extended with lacto-*N*-biose (type 1) and 53 that are LacNAc-extended (type 2). Since changes to HMO concentration are observed to vary over the course of lactation[50,87,88], owing to regulation of the expression of the genes coding for these enzymes, as well as their location in the Golgi, a spatiotemporal separation of the different galactosyltransferase activities **1** and **6** may account for the synthesis of *inverse*-LNnD (Fig. 5).

Disialyl-lacto-*N*-tetraose. The disialylated stucture, disialyllacto-*N*-tetraose (DSLNT) has previously been shown to protect neonatal rats[89] from necrotising enterocolitis (NEC), a serious inflammatory disease of the intestinal wall, to which premature infants especially are prone[90]. Milk from mothers containing DSLNT in elevated quantities has been shown significantly to reduce the likelihood of NEC[23], and its absence from the milk of the mother may be useful as a predictive marker for the disease[22,24]. Our model was able to predict the sequence of enzyme activities leading to this structure, in a four-step process from lactose: (**4**,**6**,**8**,**9**) (*cf.* Table 2), providing an in silico verification of the in vitro synthesis of DSLNT achieved by Prudden et al.[48]. This HMO, which has the composition H3N1S2 (*cf.* Fig. 5), being both terminated and decorated by sialic acid, appears to be unique to human milk: it is not detected, or only in trace amounts, in other species[91]. The mechanism by which milk containing DSLNT acts to protect the infant from NEC is still, to our knowledge, an open question.

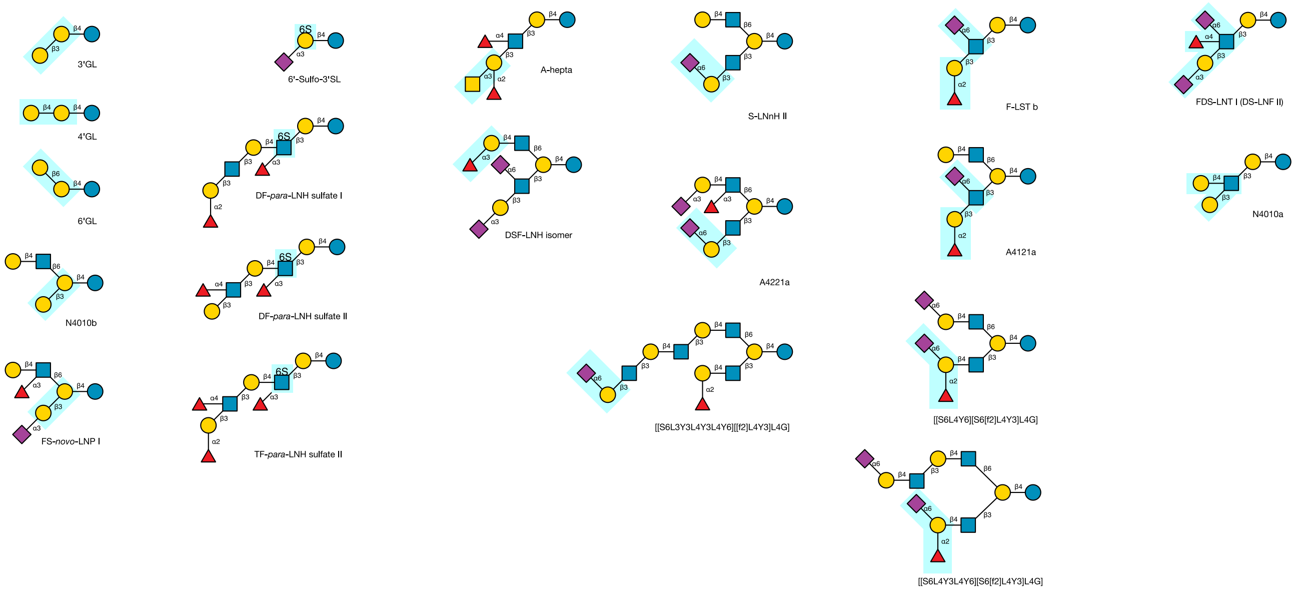

Figure 9. Structures of HMOs that were not predicted by the model. Highlighted motifs (boxed regions) suggest possible additional enzyme activities. See text for details.

Non-predicted structures. The 20 HMOs of Supplementary Table S1 that were not predicted by the model form a set of structures that may point, either to novel enzyme activities unique to milk, or else to alternative functions of known enzymes. We propose possible routes of formation of these unknowns in terms of known, and possibly novel, enzyme activities. The structures and possible enzyme recognition motifs are summarised in Fig. 9.

Galactoside β-galactosyltransferases. Three galactosyllactose structures [L3L4G] (3′-GL), [L4L4G] (4′-GL) and [L6L4G] (6′-GL) were found. The presence of 6′-galactosyllactose in the protein fraction of human milk from non-secretor mothers was first reported by Yamashita and Kobata[10], who also demonstrated that it was the product of a specific galactosyltransferase and not a transgalactosidation reaction of β-galactosidase. Two further galactoside β-galactosyltransferases, acting on positions 3 and 4 of the galactose residue, respectively, would be necessary to explain the other two non-predicted structures in this subset. No examples of extension of galactosyllactose were found in this study. Nevertheless, linear-chain poly-galactosylated HMOs have been discovered in milks from other species, such as lion[92].

GlcNAc β-1,4-galactosyltransferase. The structure of [L3[L4]Y3L4G] (N4010a[49]) is unusual in possessing both type 1 and type 2 termination. While it could be represented within our model in three different ways (see Methods), such as [L4[L3]Y3L4G], no LacNAc extension of LNnT was observed with 3-β-galactosylation of the GlcNAc, therefore we propose that the parent is lacto-*N*-tetraose ([L3Y3L4G], LNT), decorated by an unknown GlcNAc β-1,4-galactosyltransferase. The Glycologue structure identifier thus acts as a record of distinct enzyme activities, even where the structures are indistinguishable.

Alternative sialyltransferase activities. Based on the activities of other sialyltransferases, which act after the fucosyltransferases to cap Lewis-type termini, the ST3Gal and ST6GlcNAc activities of the model could be modified similarly to account for the non-predicted structures, [[f2]L3[S6]Y3L4G] (F-LST b) and [S3L3[S6][f4]Y3L4G] (DS-LNF II). By a similar modification the 6-sialylated H2-antigenic structures, [[S6L4Y6][S6[f2]L4Y3]L4G] and [[S6L4Y3L4Y6][S6[f2]L4Y3]L4G], of Prudden et al.[48] could be modelled.

Although the 6-sialyltransferase activity of EC 2.4.3.1 (**7**) is generic, acting on β-galactosyl termini, we modelled its major activity, which is towards the type-2 LacNAc acceptor[64]. The enzyme from goat and bovine colostrum has a secondary activity towards type 1 , which might explain the three HMOs in which this motif appears, [[L4Y6][S6L3Y3]L4G] (S-LNnH II)[49,55], [[S3L4[f3]Y6][S6L3Y3]L4G] (A4221a)[49] and [[S6L3Y3L4Y3L4Y6][[f2]L4Y3]L4G] (structure B27 synthesised by Prudden et al.[48]).

6-*O*-Sulfotransferases. Only 4 of the 226 HMOs in the library were sulfated, with structure identifiers [S3[s6]L4G], [[f2]L3Y3L4[f3][s6]Y3L4G], [L3[f4]Y3L4[f3][s6]Y3L4G] and [[f2]L3[f4]Y3L4[f3][s6]Y3L4G] (see Supplementary Table S1 for IUPAC and GlycoCT notations). We conclude that there exist two, as-yet uncharacterised, 6-*O*-sulfotransferase enzyme activities of human milk, which are specific for lactose and LNTri II, since no other sulfated compounds were found. All four non-predicted sulfated HMOs have non-sulfated counterparts that were predicted by our model.

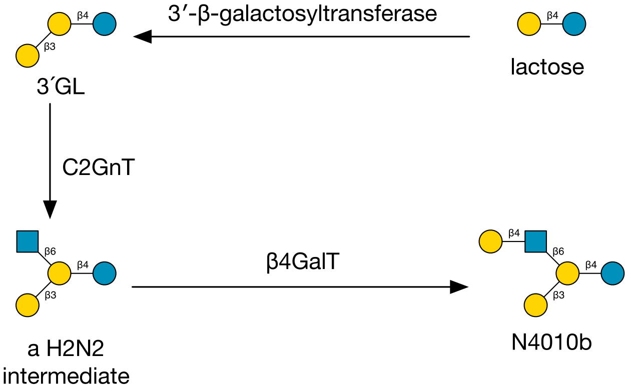

Figure 10. Proposed biosynthetic pathway of lacto-*N*-novopentaose I (N4010b).

Other enzyme activities. The disialomonofucosyllacto-*N*-hexaose structure, [[[f3]L4Y6][S3L3[S6]Y3]L4G], listed by Smith et al.[51], could not be predicted owing to the lack of a 3-α-fucosyltransferase acting on galactose. The corresponding 2-α-fucosylated structure, known as FDS-LNH I[47] and DSF-LNH II[40,49] is predicted within the system described here.

The existence of A-hepta, discovered by Wu et al.[54], is likely to be the product of EC 2.4.1.40, glycoprotein-fucosylgalactoside α-*N*-acetylgalactosaminyltransferase, acting on the H antigen. Since only one occurrence of the antigen was found in the sample dataset, this enzyme activity was excluded from the model.

Following the publication of a mass spectral reference library for HMOs based on the NIST Standard Reference Material (SRM) 1953 dataset[49], the structures of several unknown HMOs were inferred from the original library by means of a bootstrapping approach[92]. Out of 78 of novel HMOs (Table S4 of[92]), 48 (61.5%) were predicted by HMO-Glycologue with the 11 enzymes of the current model active. Owing to preliminary structural assignments and significantly lower coverage, the data was excluded from the library (Supplementary Table S1). Nevertheless, a review of those structures which were not predicted by the model is instructive, since the existence of some of them might be explained by known enzyme activities of human glycosylation. In what follows, R*n* refers to HMO with index number *n* within the cited dataset[92].

HMOs R1 and R5, Galactosyl-FpLNnH (Glycologue structure identifier: [[L4Y3L4[f3]Y6][L3]L4G]) and Galactosyl-TFpLNH ([[[f2]L3[f4]Y3L4[f3]Y6][L3]L4G]), could be formed from one or more of the *O*-glycosyltransferase core enzymes, if these were active against lactose, to give [[L3]L4G] followed by [[Y6][L3]L4G]. The corresponding GalNAc-linked glycan, [L4Y3L4[f3]Y6][L3]VT, is predicted within O-Glycologue. If the two enzymes, C1GalT and C2GnT, were co-expressed in human mammary gland, then such 3-galactosylated “Core-2 HMOs” would be expected to be abundant, but they are not. The [L3]L4 motif is observed in only one other structure, N4010b, from this and a previous study[49], although structure R2, Galactosyl-FpLNnO, might be another candidate. Instead, a possible explanation for N4010b, [[L4Y6][L3]L4G], is that it is the product of C2GnT acting on 3′-GL, formed by the galactoside 3′-β-galactosyltransferase referred to above, followed by β4GalT. The hypothetical pathway is illustrated in Fig. 10. The structure N4010b, which is also designated HMO core IV, or lacto-*N*-novopentaose I[40], and known to occur in milk from marsupial species[93,94], along with further products derived from it[95]. By similar reasoning, Novo-LNP I (R7) could be a “Core-4” HMO. The enzymic origin of these structures therefore remains to be elucidated.

The α-galactosylated structures reported by that study may be the products of two enzymes, fucosylgalactoside 3-α-galactosyltransferase (EC 2.4.1.37) and *N*-acetyllactosaminide 3-α-galactosyltransferase (EC 2.4.1.87). O-Glycologue[96] currently models the activity of the former, EC 2.4.1.37, and thus could be modified to predict these structures. The activity of EC 2.4.1.87 could likewise be added; judging by their absence from this dataset, this enzyme seems not to be active towards fucosylated HMOs.

From our analysis of HMO compositions, our results concluded that certain virtual nodes were predicted within the system, but not others (see "HMO compositions vs. structures"). For instance, composition H8N5F2 that is virtual in our library network (Fig. 6) is real in the Porfirio et al.[81] dataset, thereby supporting the selection of H8N5F2 over H7N6F2 as a valid path connecting H7N5F2 and H8N6F2. If α-galactosyltransferase enzymes are active during lactation, it might also explain the higher ratio of hexose to GlcNAc of such compositions.

The dodecaose series[92] are differentially fucosylated, and all of them would be predicted by the model were it not for the presence of the repeating L3Y3 units (L3Y3L3Y3) on the lower branches of around half of them. This unusual sequence, reported in *O*-glycans of human colonic[97,98] and gastric[99] mucin, are not formed by EC 2.4.1.149, which is instead responsible for the formation of polyLacNAc. Their presence would suggest the activity of an unknown β3GnT enzyme activity that recognises only the terminal [L3Y3*, as acceptor, since the type-1 extended upper arm, L3Y3L3Y6, is not observed. Another sequence that is novel is L4Y3L3Y3, i.e. type-1 structures extended by LacNAc to form type 2, which would infer that β4GalT does not recognise the [Y3L3. We note that the same motif exists in DF-para-LNO I (R58), to which the structure identifier [[f2]L3[f4]Y3L3Y3L4Y3L4G] was assigned.

The novel Iso-LNnO oligosaccharide, which features a terminal Gal 6-linked to a preceding GlcNAc, would require a β6GalT enzyme to form the substructure L6Y. The β(1 → 6)-galactosyltransferase referred to above, which forms 6′-GL, might also be responsible for the addition of Gal to GlcNAc, although, to our knowledge, this functionality has not yet been demonstrated.

The polyGal structures pentagalactosyllactose, heptagalactosyllactose and octagalactosyllactose R71–R73; might be initiated by the enzyme responsible for early heparan/chondroitin biosynthesis, galactosylxylosylprotein 3-β-galactosyltransferase (EC 2.4.1.134), if that enzyme were active towards glucose in addition to xylose. If the di-, tri- and tetra-galactosyllactose precursors of these larger structures exist, they were not reported, nor included in the NIST mass spectral HMO reference library.

Conclusions

As samples produced in various geographical, nutritional and health conditions accumulate, it is very likely that more HMOs will be identified. In light of the results presented here, we envisage that the Glycologue HMO-enzyme simulator will be extended, or adapted, as our knowledge of the enzymes and their substrate specificities improves. The model presented here is a pathway model, which does not account for abundances of individual oligosaccharides, which will require additional knowledge of the cellular environment, and mechanistic details of the enzymes involved, and their localisation within the cell. Such models have already been proposed for the biosynthesis of N-linked[100] and O-linked[101] glycans. The development of a mathematical model based on the core network of Fig. 2, for example, might help to elucidate the relative contributions of the enzyme activities proposed by the current network model, and the overall expected distributions of HMOs given in Fig. 4B and D.

Our analysis of the HMO glycome has raised several questions on individual and combined enzyme activities. Since the model is based on the assumption that O-linked glycosylation enzymes are involved also in milk biosynthesis, our results which will require experimental verification using free-oligosaccharide acceptors. Some questions remain to be answered, such as whether it is possible that the unknown I-branching enzyme, dIGnT, is less specific than cIGnT, and if it can use either type-1 (β3-Gal) or type-2 (β4-Gal) HMOs as acceptors. It is hoped that this analysis will promote further examination of these enzymes, including those yet to be characterised, such as the galactoside β-galactosyltransferases. With its predictive power, the model can also be considered as a guide for experimental synthesis of HMOs, which would potentially enable testing with specialised microarrays.

Methods

Formal language.

The method is based on a formal language, which uses a single-letter code for the monosaccharides, as shown in Table 1, and a set of transformation rules that add one monosaccharide at a time, using a regular-expression based pattern matching to model the enzyme activities. A software implementation of the method, Glycologue, acts iteratively on an initial acceptor-substrate, passing it to each enzyme in turn, and accumulating a set of acceptor-products. The pool of novel acceptor-products become the substrates at the next iteration, until either no new products are formed, or a maximum number of iterations set by the user has been attained. Since extension of oligosaccharides occurs principally by means of LacNAc (Gal-β1,4-GlcNAc), simulations could be limited by the number of GlcNAc residues incorporated. The reaction network was deemed to have *closed* at iteration *i* when no further products were added at the next iteration, $i + 1$. Simulations could also be limited by supplying a target composition value such as H4N3F1S1 (4 Hex, 3 HexNAc, 1 × dHex, 1 × Neu5Ac), such that enzymes would act on a substrate only if its composition did not exceed the prescribed value of any class of monosaccharide. The number of possible oligosaccharide structures matching that composition was then calculated.

Enzymes simulated.

Table 2 lists the biosynthetic enzymes predicted by the model, with an index number, **1**–**11**, the EC number, where available, a short name, a longer accepted name and a reaction pattern. The reactions in Table 2 are based on activities of enzymes already classified within the IUBMB Enzyme List, or from the cited references, wherever an EC number is not available. In reaction patterns, asterisks act as a wildcard character, to denote portions of the molecule of indeterminate length. The glossary in Table 2, footnote b, shows some of the assumptions implicit to the model, such as the anomeric configuration of the donors, from which it can be inferred which enzymes invert or retain the stereochemical configuration of the donor during incorporation into the acceptor. Reactions could, in addition, be limited by Boolean conditional regular expressions. In the case of two I-branching enzymes, **10** and **11**, it was assumed that these enzymes would not be active towards 3-fucolactose (3-FL) substrates.

Glycosyltransferase reaction-pattern classification.

The types of reaction catalysed are classified according to a limited number of transformation patterns. We consider the default mode of action to be the *extension* of a linear oligosaccharide, by

$$\mathrm{Ax} + \mathrm{yB} = \mathrm{xyB} + \mathrm{A} \tag{1}$$

where x and y are monosaccharides, Ax is the nucleotide-sugar donor, and yB the acceptor-substrate.

The formation of a single branch along a linear chain is described as *decoration*, where the pattern is

$$\mathrm{Ax} + \mathrm{yB} = [\mathrm{x}]\mathrm{yB} + \mathrm{A} \tag{2}$$

and we have assumed that [x]y is a shorthand for *[x]y*B, the asterisks acting as a wildcard character. Double branches are used to form symmetric core structures, such as the trimannosyl core of N-glycans, or O-linked glycan cores based on GalNAc:

$$Ax + yB = [x][y]B + A \tag{3}$$

Capping of branches and linearly extended chains is achieved through *termination*, of which sialylation is a typical example. *Modification* of monosaccharides, such as by the actions of sulfotransferases, follows the same pattern as decoration (2). Glycologue structure identifiers order branches by linkage position, writing the branch with the lowest linkage first, reading from right to left. Modifiers are written before sugars units, and multiple modifiers on the same sugar are again ordered by linkage position, from lowest to highest, reading right to left. Boolean conditionals can be applied to enzymes, to prevent action in the presence or absence of a particular recognition motif.

This classification, and the numbering rules, enables the assignment of a unique Glycologue structure identifiers to HMOs with determined structures. Glycosyltransferases **1**, **4** and **6** are involved in extension, and the sialyltransferases in termination, by pattern (1); the fucosyltransferases **2**, **3** and **5**, and ST6GlcNAc (**9**) are involved in decoration according to reaction pattern (2), and GTs **10** and **11** form double branches according to pattern (3).

HMO structure prediction.

The activities of the enzymes could be reversed, and all reaction paths leading from a given HMO to lactose determined. For any set of such initial substrates supplied to the reversed simulator, the complete collection of such paths leading from lactose provided a minimal biosynthetic reaction network. Individual paths are represented by ordered sequences of enzyme activities. For example, the formation of LST b, which has the structure identifier [L3[S6]Y3L4G], has sequence (**4**,**6**,**9**), when starting from lactose. As there can be multiple routes to a given HMO, the networks generated in both the forward and reverse directions possess a lattice-like structure.

Web application.

A web application interface to the HMO-enzyme simulator, the set of experimentally determined HMOs used as validation and the source code of the simulator as a Python 3 script, are available at https://glycologue.org/m. The Glycologue family of simulators[46,96,102] supports import and export of IUPAC short form and GlycoCT condensed formats, along with the native Glycologue structure identifiers, while export as Linear Code is also provided for individual structures. Networks can be exported as SBML, with GlycoCT-XML embedded as annotations of the nodes.

Data availability

The HMO structures analysed in this study are available as CSV files (Supplementary Tables S1 and S2). The enzyme model simulator is available at https://glycologue.org/m/. GlyConnect Compozitor is available at https://glyconnect.expasy.org/compozitor/.

Received: 2 November 2021; Accepted: 3 June 2022
Published online: 27 June 2022

References

1. Hundshammer, C. & Minge, O. In love with shaping you—influential factors on the breast milk content of human milk oligosaccharides and their decisive roles for neonatal development. *Nutrients* **12**, 3568. https://doi.org/10.3390/nu12113568 (2020).
2. Kuhn, R. Vitamine der Milch. *Angew. Chem.* **64**, 493–500. https://doi.org/10.1002/ange.19520641802 (1952).
3. Kuhn, R. *et al.* Aminozucker. *Angew. Chem.* **69**, 23–33. https://doi.org/10.1002/ange.19570690105 (1957).
4. Montreuil, J. Glucides of human milk. *Bull. Soc. Chim. Biol. (Paris)* **39**, 395–411 (1957).
5. Malpress, F. H. & Hytten, F. E. The oligosaccharides of human milk. *Biochem. J.* **68**, 708–717. https://doi.org/10.1042/bj0680708 (1958).
6. Kobata, A., Ginsburg, V. & Tsuda, M. Oligosaccharides of human milk. I. Isolation and characterization. *Arch. Biochem. Biophys.* **130**, 509–513. https://doi.org/10.1016/0003-9861(69)90063-0 (1969).
7. Kobata, A. & Ginsburg, V. Oligosaccharides of human milk. II. Isolation and characterization of a new pentasaccharide, lacto-*N*-fucopentaose 3. *J. Biol. Chem.* **244**, 5496–5502 (1969).
8. Kobata, A. & Ginsburg, V. Oligosaccharides of human milk. 3. Isolation and characterization of a new hexasaccharide, lacto-*N*-hexaose. *J. Biol. Chem.* **247**, 1525–1529 (1972).
9. Kobata, A. & Ginsburg, V. Oligosaccharides of human milk. IV. Isolation and characterization of a new hexasaccharide, lacto-*N*-*neo*hexaose. *Arch. Biochem. Biophys.* **150**, 273–281. https://doi.org/10.1016/0003-9861(72)90036-7 (1972).
10. Yamashita, K. & Kobata, A. Oligosaccharides of human milk: V. Isolation and characterization of a new trisaccharide, 6′-Galactosyllactose. *Arch. Biochem. Biophys.* **161**, 164–170. https://doi.org/10.1016/0003-9861(74)90247-1 (1974).
11. Yamashita, K., Tachibana, Y. & Kobata, A. Oligosaccharides of human milk. Isolation and characterization of three new disialyfucosyl hexasaccharides. *Arch. Biochem. Biophys.* **174**, 582–591. https://doi.org/10.1016/0003-9861(76)90387-8 (1976).
12. Yamashita, K., Tachibana, Y. & Kobata, A. Oligosaccharides of human milk: isolation and characterization of two new nonasaccharides, monofucosyllacto-*N*-octaose and monofucosyllacto-*N*-neooctaose. *Biochemistry* **15**, 3950–3955. https://doi.org/10.1021/bi00663a007 (1976).
13. Ginsburg, V. & Zopf, D. A. Oligosaccharides of human milk. Isolation of a new pentasaccharide, lacto-*N*-fucopentaose V. *Arch. Biochem. Biophys.* **175**, 565–568. https://doi.org/10.1016/0003-9861(76)90546-4 (1976).
14. Yamashita, K., Tachibana, Y. & Kobata, A. Oligosaccharides of human milk: Structures of three lacto-*N*-hexaose derivatives with H-haptenic structure. *Arch. Biochem. Biophys.* **182**, 546–555. https://doi.org/10.1016/0003-9861(77)90536-7 (1977).
15. Yamashita, K., Tachibana, Y. & Kobata, A. Oligosaccharides of human milk. Structural studies of two new octasaccharides, difucosyl derivatives of para-lacto-*N*-hexaose and para-lacto-*N*-*neo*hexaose. *J. Biol. Chem.* **252**, 5408–5411 (1977).
16. Ackerman, D. L. *et al.* Antimicrobial and antibiofilm activity of human milk oligosaccharides against *Streptococcus agalactiae*, *Staphylococcus aureus*, and *Acinetobacter baumannii*. *ACS Infect Dis.* **4**, 315–324. https://doi.org/10.1021/acsinfecdis.7b00183 (2018).
17. Craft, K. M. & Townsend, S. D. The human milk glycome as a defense against infectious diseases: rationale, challenges, and opportunities. *ACS Infect. Dis.* **4**, 77–83. https://doi.org/10.1021/acsinfecdis.7b00209 (2018).
18. Koromyslova, A. *et al.* Human norovirus inhibition by a human milk oligosaccharide. *Virology* **508**, 81–89. https://doi.org/10.1016/j.virol.2017.04.032 (2017).

19. Yu, Y. *et al.* Human milk contains novel glycans that are potential decoy receptors for neonatal rotaviruses. *Mol. Cell. Proteomics* **13**, 2944. https://doi.org/10.1074/mcp.M114.039875 (2014).
20. Morozov, V. *et al.* Human milk oligosaccharides as promising antivirals. *Mol. Nutr. Food Res.* **62**, 1700679. https://doi.org/10.1002/mnfr.201700679 (2018).
21. Moore, R. E., Xu, L. L. & Townsend, S. D. Prospecting human milk oligosaccharides as a defense against viral infections. *ACS Infect. Dis.* **7**, 254–263. https://doi.org/10.1021/acsinfecdis.0c00807 (2021).
22. Autran, C. A. *et al.* Human milk oligosaccharide composition predicts risk of necrotising enterocolitis in preterm infants. *Gut* **67**, 1064–1070. https://doi.org/10.1136/gutjnl-2016-312819 (2018).
23. Hassinger, D. *et al.* Analysis of Disialyllacto-N-Tetraose (DSLNT) content in milk from mothers of preterm infants. *J. Hum. Lact.* **36**, 291–298. https://doi.org/10.1177/0890334420904041 (2020).
24. Masi, A. C. *et al.* Human milk oligosaccharide DSLNT and gut microbiome in preterm infants predicts necrotising enterocolitis. *Gut* **70**, 2273–2282. https://doi.org/10.1136/gutjnl-2020-322771 (2021).
25. Bering, S. B. Human milk oligosaccharides to prevent gut dysfunction and necrotizing enterocolitis in preterm neonates. *Nutrients* https://doi.org/10.3390/nu10101461 (2018).
26. Walsh, C., Lane, J. A., van Sinderen, D. & Hickey, R. M. Human milk oligosaccharides: Shaping the infant gut microbiota and supporting health. *J. Funct. Foods* **72**, 104074. https://doi.org/10.1016/j.jff.2020.104074 (2020).
27. Kulinich, A. & Liu, L. Human milk oligosaccharides: The role in the fine-tuning of innate immune responses. *Carbohydr. Res.* **432**, 62–70. https://doi.org/10.1016/j.carres.2016.07.009 (2016).
28. Ayechu-Muruzabal, V. *et al.* Diversity of human milk oligosaccharides and effects on early life immune development. *Front. Pediatr.* **6**, 239. https://doi.org/10.3389/fped.2018.00239 (2018).
29. Plaza-Díaz, J., Fontana, L. & Gil, A. Human milk oligosaccharides and immune system development. *Nutrients* **10**, 1038. https://doi.org/10.3390/nu10081038 (2018).
30. Zuurveld, M. *et al.* Immunomodulation by human milk oligosaccharides: The potential role in prevention of allergic diseases. *Front. Immunol.* **11**, 801. https://doi.org/10.3389/fimmu.2020.00801 (2020).
31. Urashima, T. *et al.* The predominance of Type I oligosaccharides is a feature specific to human breast milk. *Adv. Nutr.* **3**, 473S-482S. https://doi.org/10.3945/an.111.001412 (2012).
32. Wands, A. M. *et al.* Fucosylated molecules competitively interfere with cholera toxin binding to host cells. *ACS Infect. Dis.* **4**, 758–770. https://doi.org/10.1021/acsinfecdis.7b00085 (2018).
33. Katayama, T. Host-derived glycans serve as selected nutrients for the gut microbe: human milk oligosaccharides and bifidobacteria. *Biosci. Biotechnol. Biochem.* **80**, 621–632. https://doi.org/10.1080/09168451.2015.1132153 (2016).
34. Wang, Y. *et al.* Enzymatic production of HMO mimics by the sialylation of galacto-oligosaccharides. *Food Chem.* **181**, 51–56. https://doi.org/10.1016/j.foodchem.2015.02.064 (2015).
35. Muschiol, J. & Meyer, A. S. A chemo-enzymatic approach for the synthesis of human milk oligosaccharide backbone structures. *Zeitschrift für Naturforschung C* **74**, 85–89. https://doi.org/10.1515/znc-2018-0149 (2019).
36. Zeuner, B. *et al.* Substrate specificity and transfucosylation activity of GH29 α-L-fucosidases for enzymatic production of human milk oligosaccharides. *New Biotechnol.* **41**, 34–45. https://doi.org/10.1016/j.nbt.2017.12.002 (2018).
37. Zeuner, B., Teze, D., Muschiol, J. & Meyer, A. S. Synthesis of human milk oligosaccharides: protein engineering strategies for improved enzymatic transglycosylation. *Molecules* **24**, 2033. https://doi.org/10.3390/molecules24112033 (2019).
38. Leloir, L. F. The enzymatic transformation of uridine diphosphate glucose into a galactose derivative. *Arch. Biochem. Biophys.* **33**, 186–190. https://doi.org/10.1016/0003-9861(51)90096-3 (1951).
39. Nishimoto, M. Large scale production of lacto-*N*-biose I, a building block of type I human milk oligosaccharides, using sugar phosphorylases. *Biosci. Biotechnol. Biochem.* **84**, 17–24. https://doi.org/10.1080/09168451.2019.1670047 (2020).
40. Urashima, T., Hirabayashi, J., Sato, S. & Kobata, A. Human milk oligosaccharides as essential tools for basic and application studies on galectins. *Trends Glycosci. Glycotechnol.* **30**, SE51–SE65. https://doi.org/10.4052/tigg.1734.1SE (2018).
41. Kobata, A. & Ginsburg, V. Uridine diphosphate-*N*-acetyl-D-galactosamine:D-galactose α-3-*N*-acetyl-D-galactosaminyltransferase, a product of the gene that determines blood type A in man. *J. Biol. Chem.* **245**, 1484–1490. https://doi.org/10.1016/S0021-9258(18)63261-2 (1970).
42. Sakwinska, O. & Bosco, N. Host microbe interactions in the lactating mammary gland. *Front Microbiol* **10**, 1863. https://doi.org/10.3389/fmicb.2019.01863 (2019).
43. Betts, C. B. *et al.* Mucosal immunity in the female murine mammary gland. *JI* **201**, 734–746. https://doi.org/10.4049/jimmunol.1800023 (2018).
44. Agravat, S. B. *et al.* Computational approaches to define a human milk metaglycome. *Bioinformatics* **32**, 1471–1478. https://doi.org/10.1093/bioinformatics/btw048 (2016).
45. Bao, B. *et al.* Correcting for sparsity and interdependence in glycomics by accounting for glycan biosynthesis. *Nat. Commun.* **12**, 4988. https://doi.org/10.1038/s41467-021-25183-5 (2021).
46. McDonald, A. G., Tipton, K. F. & Davey, G. P. A knowledge-based system for display and prediction of *O*-glycosylation network behaviour in response to enzyme knockouts. *PLoS Comput. Biol.* **12**, e1004844. https://doi.org/10.1371/journal.pcbi.1004844 (2016).
47. Chen X (2015) Human milk oligosaccharides (HMOS). In: Advances in Carbohydrate Chemistry and Biochemistry. Elsevier, pp 113–190
48. Prudden, A. R. *et al.* Synthesis of asymmetrical multiantennary human milk oligosaccharides. *Proc. Natl. Acad. Sci.* **114**, 6954. https://doi.org/10.1073/pnas.1701785114 (2017).
49. Remoroza, C. A. *et al.* Creating a mass spectral reference library for oligosaccharides in human milk. *Anal. Chem.* **90**, 8977–8988. https://doi.org/10.1021/acs.analchem.8b01176 (2018).
50. Samuel, T. M. *et al.* Impact of maternal characteristics on human milk oligosaccharide composition over the first 4 months of lactation in a cohort of healthy European mothers. *Sci. Rep.* **9**, 11767. https://doi.org/10.1038/s41598-019-48337-4 (2019).
51. Smith, D. F., Zorf, D. A. & Ginsburg, V. Fractionation of sialyl oligosaccharides of human milk by ion-exchange chromatography. *Anal. Biochem.* **85**, 602–608. https://doi.org/10.1016/0003-2697(78)90261-0 (1978).
52. Tachibana, Y., Yamashita, K. & Kobata, A. Oligosaccharides of human milk: Structural studies of Di- and trifucosyl derivatives of lacto-*N*-octaose and lacto-*N*-neooctaose. *Arch. Biochem. Biophys.* **188**, 83–89. https://doi.org/10.1016/0003-9861(78)90359-4 (1978).
53. Thurl, S. *et al.* Systematic review of the concentrations of oligosaccharides in human milk. *Nutr. Rev.* **75**, 920–933. https://doi.org/10.1093/nutrit/nux044 (2017).
54. Wu, S. *et al.* Development of an annotated library of neutral human milk oligosaccharides. *J. Proteome Res.* **9**, 4138–4151. https://doi.org/10.1021/pr100362f (2010).
55. Wu, S., Grimm, R., German, J. B. & Lebrilla, C. B. Annotation and structural analysis of sialylated human milk oligosaccharides. *J. Proteome Res.* **10**, 856–868. https://doi.org/10.1021/pr101006u (2011).
56. Herget, S., Ranzinger, R., Maass, K., Lieth, C. W. & v d,. GlycoCT—a unifying sequence format for carbohydrates. *Carbohydr. Res.* **343**, 2162–2171. https://doi.org/10.1016/j.carres.2008.03.011 (2008).
57. Fujita, A. *et al.* The international glycan repository GlyTouCan version 3.0. *Nucleic Acids Res.* **49**, D1529–D1533. https://doi.org/10.1093/nar/gkaa947 (2021).

58. Neelamegham, S. *et al.* Updates to the symbol nomenclature for glycans guidelines. *Glycobiology* **29**, 620–624. https://doi.org/10.1093/glycob/cwz045 (2019).
59. Bansal, P. *et al.* Rhea, the reaction knowledgebase in 2022. *Nucl. Acids Res.* **50**, D693–D700. https://doi.org/10.1093/nar/gkab1016 (2022).
60. Lloyd, K. O. & Kabat, E. A. Immunochemical studies on blood groups. XLI. Proposed structures for the carbohydrate portions of blood group A, B, H, Lewis-a, and Lewis-b substances. *Proc. Natl. Acad. Sci. USA* **61**, 1470–1477. https://doi.org/10.1073/pnas.61.4.1470 (1968).
61. Kobata, A. & Ginsburg, V. Oligosaccharides of human milk. *J. Biol. Chem.* **244**, 5496–5502. https://doi.org/10.1016/S0021-9258(18)63591-4 (1969).
62. Dabrowski, U., Egge, H. & Dabrowski, J. Proton-nuclear magnetic resonance study of peracetylated derivatives of ten oligosaccharides isolated from human milk. *Arch. Biochem. Biophys.* **224**, 254–260. https://doi.org/10.1016/0003-9861(83)90208-4 (1983).
63. Biswas, A. & Thattai, M. Promiscuity and specificity of eukaryotic glycosyltransferases. *Biochem. Soc. Trans.* **48**, 891–900. https://doi.org/10.1042/BST20190651 (2020).
64. Johnson, P. H., Donald, A. S. R., Feeney, J. & Watkins, W. M. Reassessment of the acceptor specificity and general properties of the Lewis blood-group gene associated α-3/4-fucosyltransferase purified from human mil *Glycoconjugate J.* **9**, 251–264. https://doi.org/10.1007/BF00731137 (1992).
65. Bode, L. Human milk oligosaccharides: Every baby needs a sugar mama. *Glycobiology* **22**, 1147–1162. https://doi.org/10.1093/glycob/cws074 (2012).
66. Paulson, J. C., Weinstein, J. & de Souza-E-Silva, U. Biosynthesis of a disialylated sequence in N-linked oligosaccharides: identification of an *N*-acetylglucosaminide (α2→6)-sialyltransferase in Golgi apparatus from rat liver. *Eur. J. Biochem.* **140**, 523–530. https://doi.org/10.1111/j.1432-1033.1984.tb08133.x (1984).
67. Tsuchida, A. *et al.* Synthesis of disialyl Lewis a (Le^a) structure in colon cancer cell lines by a sialyltransferase, ST6GalNAc VI, responsible for the synthesis of α-series gangliosides. *J. Biol. Chem.* **278**, 22787–22794. https://doi.org/10.1074/jbc.M211034200 (2003).
68. Ujita, M. *et al.* Synthesis of poly-*N*-acetyllactosamine in core 2 branched *O*-glycans. The requirement of novel β-1,4-galactosyltransferase IV and β-1,3-*N*-acetylglucosaminyltransferase. *J. Biol. Chem.* **273**, 34843–34849. https://doi.org/10.1074/jbc.273.52.34843 (1998).
69. Piller, F. *et al.* Biosynthesis of blood group I antigens. Identification of a UDP-GlcNAc:GlcNAcβ1–3Gal(-R)β1–6(GlcNAc to Gal) *N*-acetylglucosaminyltransferase in hog gastric mucosa. *J. Biol. Chem.* **259**, 13385–13390. https://doi.org/10.1016/S0021-9258(18)90706-4
70. Yeh, J.-C., Ong, E. & Fukuda, M. Molecular cloning and expression of a novel β-1,6-*N*-acetylglucosaminyltransferase that forms Core 2, Core 4, and I Branches. *J. Biol. Chem.* **274**, 3215–3221. https://doi.org/10.1074/jbc.274.5.3215 (1999).
71. Cummings, R. D. The repertoire of glycan determinants in the human glycome. *Mol. Biosyst.* **5**, 1087–1104. https://doi.org/10.1039/b907931a (2009).
72. Auber, D. *et al.* Tulip 5. In *Encyclopedia of Social Network Analysis and Mining* (eds Alhajj, R. & Rokne, J.) 1–28 (Springer, 2017).
73. Hossler, P., Goh, L.-T., Lee, M. M. & Hu, W.-S. GlycoVis: visualizing glycan distribution in the protein *N*-glycosylation pathway in mammalian cells. *Biotechnol. Bioeng.* **95**, 946–960. https://doi.org/10.1002/bit.21062 (2006).
74. Soyyılmaz, B. *et al.* The mean of milk: a review of human milk oligosaccharide concentrations throughout lactation. *Nutrients* **13**, 2737. https://doi.org/10.3390/nu13082737 (2021).
75. Blank, D., Dotz, V., Geyer, R. & Kunz, C. Human milk oligosaccharides and Lewis blood group: individual high-throughput sample profiling to enhance conclusions from functional studies. *Adv. Nutr.* **3**, 440S-449S. https://doi.org/10.3945/an.111.001446 (2012).
76. Ashline, D. J. *et al.* Structural characterization by multistage mass spectrometry (MS^n) of human milk glycans recognized by human rotaviruses. *Mol. Cell. Proteomics* **13**, 2961. https://doi.org/10.1074/mcp.M114.039925 (2014).
77. Blank, D. *et al.* Elucidation of a novel lacto-*N*-decaose core structure in human milk using nonlinear analytical technique combinations. *Anal. Biochem.* **421**, 680–690. https://doi.org/10.1016/j.ab.2011.11.030 (2012).
78. Robin, T., Mariethoz, J. & Lisacek, F. Examining and fine-tuning the selection of glycan compositions with GlyConnect Compozitor. *Mol. Cell Proteom.* **19**, 1602–1618. https://doi.org/10.1074/mcp.RA120.002041 (2020).
79. Mariethoz, J., Hayes, C. & Lisacek, F. Glycan compositions with GlyConnect Compozitor to enhance glycopeptide identification. *Methods Mol. Biol.* **2361**, 109–127. https://doi.org/10.1007/978-1-0716-1641-3_7 (2021).
80. Alocci, D. *et al.* GlyConnect: glycoproteomics goes visual, interactive, and analytical. *J. Proteome Res.* **18**, 664–677. https://doi.org/10.1021/acs.jproteome.8b00766 (2019).
81. Porfirio, S. *et al.* New strategies for profiling and characterization of human milk oligosaccharides. *Glycobiology* **30**, 774–786. https://doi.org/10.1093/glycob/cwaa028 (2020).
82. Kunz, C. *et al.* Influence of gestational age, secretor, and Lewis blood group status on the oligosaccharide content of human milk. *J. Pediatr. Gastroenterol. Nutr.* **64**, 789–798. https://doi.org/10.1097/MPG.0000000000001402 (2017).
83. Kunz, C., Rudloff, S. (2017) Compositional analysis and metabolism of human milk oligosaccharides in infants. In: Isolauri, E., Sherman, P.M., Walker, W.A. (eds) *Nestlé Nutrition Institute Workshop Series*. S. Karger AG, pp 137–147
84. Azad, M. B. *et al.* Human milk oligosaccharide concentrations are associated with multiple fixed and modifiable maternal characteristics, environmental factors, and feeding practices. *J. Nutr.* **148**, 1733–1742. https://doi.org/10.1093/jn/nxy175 (2018).
85. Elwakiel, M. *et al.* Human milk oligosaccharides in colostrum and mature milk of Chinese mothers: Lewis positive secretor subgroups. *J. Agric. Food Chem.* **66**, 7036–7043. https://doi.org/10.1021/acs.jafc.8b02021 (2018).
86. Blanken, W. M., Hooghwinkel, G. J. M. & Eijnden, D. H. Biosynthesis of blood-group I and i substances: specificity of bovine colostrum β-*N*-acetyl-D-glucosaminide β1 → 4 galactosyltransferase. *Eur. J. Biochem.* **127**, 547–552. https://doi.org/10.1111/j.1432-1033.1982.tb06906.x (1982).
87. Coppa, G. V. *et al.* Changes in carbohydrate composition in human milk over 4 months of lactation. *Pediatrics* **91**, 637–641 (1993).
88. Austin, S. *et al.* Temporal change of the content of 10 oligosaccharides in the milk of chinese urban mothers. *Nutrients* **8**, 346. https://doi.org/10.3390/nu8060346 (2016).
89. Jantscher-Krenn, E. *et al.* The human milk oligosaccharide disialyllacto-N-tetraose prevents necrotising enterocolitis in neonatal rats. *Gut* **61**, 1417. https://doi.org/10.1136/gutjnl-2011-301404 (2012).
90. Müller, M. J., Paul, T. & Seeliger, S. Necrotizing enterocolitis in premature infants and newborns. *J. Neonatal Perinatal Med.* **9**, 233–242. https://doi.org/10.3233/NPM-16915130 (2016).
91. Monti, L., Cattaneo, T. M. P., Orlandi, M. & Curadi, M. C. Capillary electrophoresis of sialylated oligosaccharides in milk from different species. *J. Chromatogr. A* **1409**, 288–291. https://doi.org/10.1016/j.chroma.2015.07.076 (2015).
92. Remoroza, C. A. *et al.* Increasing the coverage of a mass spectral library of milk oligosaccharides using a hybrid-search-based bootstrapping method and milks from a wide variety of mammals. *Anal. Chem.* **92**, 10316–10326. https://doi.org/10.1021/acs.analchem.0c00342 (2020).
93. Messer, M. & Mossop, G. Milk carbohydrates of marsupials I. Partial separation and characterization of neutral milk oligosaccharides of the eastern grey kangaroo. *Aust. J. Bio. Sci.* **30**, 379. https://doi.org/10.1071/BI9770379 (1977).

94. Messer, M. & Green, B. Milk carbohydrates of marsupials II. Quantitative and qualitative changes in milk carbohydrates during lactation in the Tammar wallaby (*Macropus eugenii*). *Aust. J. Bio. Sci.* **32**, 519. https://doi.org/10.1071/BI9790519 (1979).
95. Urashima, T. *et al.* Evolution of milk oligosaccharides: Origin and selectivity of the ratio of milk oligosaccharides to lactose among mammals. *Biochimica et Biophysica Acta (BBA) General Subjects* **1866**, 130012. https://doi.org/10.1016/j.bbagen.2021.130012 (2022).
96. McDonald, A. G. & Davey, G. P. O-Glycologue: a formal-language-based generator of *O*-glycosylation networks. In *Glycosylation* (ed. Davey, G. P.) 223–236 (Springer, 2022).
97. Podolsky, D. K. Oligosaccharide structures of isolated human colonic mucin species. *J. Biol. Chem.* **260**, 15510–15515. https://doi.org/10.1016/S0021-9258(17)36284-1 (1985).
98. Podolsky, D. K. Oligosaccharide structures of human colonic mucin. *J. Biol. Chem.* **260**, 8262–8271 (1985).
99. Jin, C. *et al.* Structural diversity of human gastric mucin glycans. *Mol. Cell. Proteomics* **16**, 743–758. https://doi.org/10.1074/mcp.M117.067983 (2017).
100. Krambeck, F. J., Bennun, S. V., Andersen, M. R. & Betenbaugh, M. J. Model-based analysis of N-glycosylation in Chinese hamster ovary cells. *PLoS ONE* **12**, e0175376. https://doi.org/10.1371/journal.pone.0175376 (2017).
101. Kouka, T. *et al.* Computational modeling of *O*-linked glycan biosynthesis in CHO cells. *Molecules* **27**, 1766. https://doi.org/10.3390/molecules27061766 (2022).
102. McDonald, A. G. & Davey, G. P. Simulating the enzymes of ganglioside biosynthesis with Glycologue. *Beilstein J. Org. Chem.* **17**, 739–748. https://doi.org/10.3762/bjoc.17.64 (2021).

Acknowledgements

Partly supported by the Swiss Federal Government through the State Secretariat for Education, Research and Innovation (SERI). The Expasy portal is maintained by the web team of the Swiss Institute of Bioinformatics and hosted at the Vital-IT Competency Center. This work was part supported by SNSF grant #31003A_179249 (Frederique Lisacek).

Authors contributions

A.M. and F.L. wrote the main manuscript text, which A.M., F.L. and G.D. edited. Experiments were conceived by A.M., G.D. and F.L. and performed by A.M. and F.L. The software was developed by A.M. (Glycologue) and J.M. (GlyConnect). All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-14260-4.

Correspondence and requests for materials should be addressed to A.G.M. or F.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.