JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

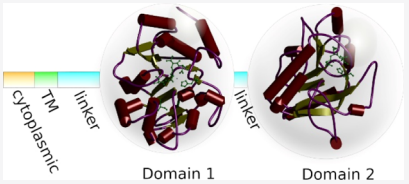
Identification and Modeling of a GT-A Fold in the α-Dystroglycan Glycosylating Enzyme LARGE1

Benedetta Righino, Manuela Bozzi, Davide Pirolli, Francesca Sciandra, Maria Giulia Bigotti,* Andrea Brancaccio, and Maria Cristina De Rosa*

Cite This: J. Chem. Inf. Model. 2020, 60, 3145−3156 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The acetylglucosaminyltransferase-like protein LARGE1 is an enzyme that is responsible for the final steps of the post-translational modifications of dystroglycan (DG), a membrane receptor that links the cytoskeleton with the extracellular matrix in the skeletal muscle and in a variety of other tissues. LARGE1 acts by adding the repeating disaccharide unit [-3Xyl-α1,3GlcAβ1-] to the extracellular portion of the DG complex (α-DG); defects in the *LARGE1* gene result in an aberrant glycosylation of α-DG and consequent impairment of its binding to laminin, eventually affecting the connection between the cell and the extracellular environment. In the skeletal muscle, this leads to degeneration of the muscular tissue and muscular dystrophy. So far, a few missense mutations have been identified within the LARGE1 protein and linked to congenital muscular dystrophy, and because no structural information is available on this enzyme, our understanding of the molecular mechanisms underlying these pathologies is still very limited. Here, we generated a 3D model structure of the two catalytic domains of LARGE1, combining different molecular modeling approaches. Furthermore, by using molecular dynamics simulations, we analyzed the effect on the structure and stability of the first catalytic domain of the pathological missense mutation S331F that gives rise to a severe form of muscle−eye−brain disease.



INTRODUCTION

Glycosylation is one of the most common post-translational modifications of proteins in the cell and plays a key role in physiological and pathological cellular functions.[1−3] Glycosylation patterns are altered in a number of diseases including cancer and neurodegenerative and autoimmune disorders.[4] Dystroglycan (DG) is a highly glycosylated membrane receptor, formed by the two subunits α- and β-dystroglycan (hereinafter α-DG and β-DG), that connects the extracellular matrix with the cytoskeleton in many tissues, such as muscle and brain.[5] α-DG is a peripheral membrane protein formed by two globular domains, the N- and C-terminal domains, separated by an elongated mucin-like region.[6] α-DG binds to a set of extracellular matrix molecules that harbor one or more laminin globular domains (LG-domains),[7] such as laminin,[8] agrin,[9] perlecan,[10] pikachurin,[11] neurexin,[12] and slit.[13] The binding properties of α-DG depend on a complex *O*-mannosyl glycosylation, a multi-step process that involves at least 17 different enzymes.[14] Mutations in any of these enzymes are linked to a number of muscular dystrophies, collectively known as secondary dystroglycanopathies, whose clinical presentations can vary from severe congenital muscular dystrophies to milder limb-girdle muscular dystrophy with manifestation in adulthood.[15] The hallmark of secondary dystroglycanopathies is the absence or the significant reduction of α-DG glycosylation.[16] The α-DG *O*-mannosyl modification starts with the attachment of a trisaccharide molecule GalNAc-β3-GlcNAc-β4-Man to Ser/Thr residues in the α-DG mucin-like region.[17] The mannose is then phosphorylated by POMK to allow further glycan elongation and phosphorylation by enzymes including fukutin, FKRP, TMEM5, B4GAT1, and LARGE1.[18−21] In particular, LARGE1 is a type II transmembrane protein, localized in the Golgi apparatus,[22] characterized by two distinct domains, that can be found within the Golgi lumen, with glycosyltransferase activities: a xylosyltransferase activity (domain 1) and a glucuronyltransferase activity (domain 2)[18] (Figure 1).

Indeed, LARGE1 catalyzes the addition to its substrate of several units of the disaccharide [-3Xyl-α1,3GlcAβ1-], known as matriglycan, which is the functional motif that ensures the

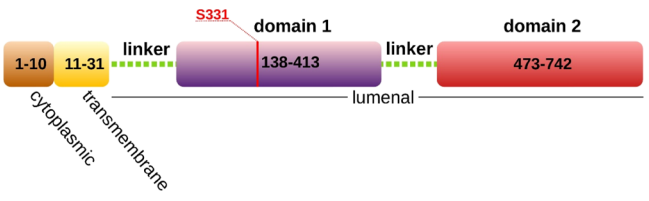


Figure 1. Schematic representation of LARGE1 domains.

Received: March 19, 2020
Published: May 1, 2020

binding of α-DG to the extracellular matrix proteins.[14,23] During DG maturation, LARGE1 binds to the N-terminal domain of α-DG, and this binding is an essential anchor for the enzyme to specifically modify the mucin-like region.[19,24]

Different mutations in *LARGE1* have been identified in patients affected by severe congenital muscular dystrophy with central nervous system involvement.[25−29] These mutations include missense and frameshift mutations in both domain 1 and domain 2, as well as intragenic deletions/insertions. Overexpression of LARGE1 in cells from patients affected by different α-dystroglycanopathies restored α-DG laminin binding, indicating that the modulation of LARGE1 activity may represent a therapeutic strategy for the treatment of secondary dystroglycanopathies.[30,31]

In this context, the effort to reach a fundamental understanding of the molecular details of the glycosylation process is undermined by the lack of high-resolution structural data on LARGE1. In order to start filling this gap, molecular modelling can be used to construct a working three-dimensional model structure of the α-DG glycosylating enzyme. When structural homologs are not available, as in the case of LARGE1, *ab initio* modelling may guide a conformational search based on a designed energy function, but, in spite of recent advances, such an approach has shown low accuracy in prediction.[32] Traditional physics-based potentials for molecular mechanics and conformational searches, first used for organic compounds,[33,34] have not yet been developed to a point where reliable model structures can be generated except for a limited number of proteins.[32] Knowledge-based energy functions were demonstrated to perform better,[35] and among them, the TASSER approach ranked as the top method for automated protein structure prediction in the latest CASP experiments for three-dimensional structure prediction.[36,37] TASSER and its development, I-TASSER gateway, which involves template threading, structural fragment assembly, model refinement, and structure-based protein function annotation, indicate that *ab initio* and template-based modelling may successfully combine.[38] To date, the only *in silico* analysis of LARGE1, confined to the study of LARGE1 domain 2, is that of Bhattacharya et al.[39]

In this study, we present a three-dimensional model for both domain 1 and 2 of LARGE1 and analyze their structural features using the best performing methods for fold recognition and domain boundary prediction.[40,41] The generated three-dimensional structures were refined, and the best reliable models were selected and subjected to molecular dynamics (MD) simulations. MD calculations are useful tools for evaluating the stability of predicted domains as well as the effect of a pathological mutation,[42−44] and we employed them in order to provide insight into the outcome of the S331F mutation found in a patient affected by congenital muscular dystrophy with eye and brain involvement.[26]

MATERIALS AND METHODS

Functional Domains Prediction and Models Building. The 756 amino acid long sequence of LARGE1 from *Mus musculus* was retrieved from the UniProt database (http://www.uniprot.org/)[45] (entry code Q9Z1M7), and the domain analysis was performed using the Conserved Domain Database (CDD) server.[46] Following a successful modelling strategy,[47] two different protein-modelling approaches were employed to build the molecular models of the identified LARGE1 protein domains: HHPRED[48] in combination with MODELLER[49] and the I-TASSER server.[50] The web-service HHpred, available at https://toolkit.tuebingen.mpg.de/tools/hhpred, offers a threading approach that uses the hidden Markov models[51] and was employed to align the sequences and search for suitable template structures. The PDB70 profile database was used for the template search, and the crystallographic structure showing the best HHpred probability score, the largest coverage, and the best resolution was chosen as a template. The resulting alignment was submitted to the program MODELLER v 9.15 as implemented in Discovery Studio 2016 (Dassault Systèmes BIOVIA, Discovery Studio Modelling Environment, Release 2016, San Diego: Dassault Systèmes, 2016) to build the three-dimensional structure of the identified LARGE1 domains. Fifty models with a degree of optimization scored as "high" were generated by the "Build Homology Model" protocol of Discovery Studio and ranked by the PDF total energy. The best-ranked models were then submitted to the "Loop Refinement" protocol of the program generating five models for each loop, with a "high" optimization level. The refined models with the lowest PDF total energy score were chosen as the final models. For comparison, the I-TASSER web-service was also used, a meta-server that automatically employs ten threading algorithms in combination with *ab initio* modelling to build the tertiary structure of a protein as well as replica-exchange Monte Carlo dynamics (REMD) simulations for the atomic-level refinement. PROCHECK[52] and ProSA-Web[53] were employed to evaluate the quality of the model. The structure of the mutant S331F of LARGE1 domain 1 was carried out with the "Build Mutant" protocol of the BIOVIA Discovery Studio suite (Dassault Systèmes BIOVIA, Discovery Studio Modelling Environment, Release 2016, San Diego: Dassault Systèmes, 2016). Fifty models were generated with a "high" optimization level and with no restraints and evaluated based on their PDF total energy and DOPE score.

Molecular Dynamics Simulations of WT and of the S331F Mutant. Molecular dynamics simulations were executed using the version 4.8 of Desmond (D. E. Shaw Research, New York) employing the OPLS2005 force field.[54] An all-hydrogen model of the proteins was first generated using the Protein Preparation Wizard tool[55] available in the Maestro software (Schrödinger, LLC, New York, NY, 2017), for hydrogen insertion and force field atom-types and partial charges assignment. The systems for the simulations were built with the "System Builder" tool of the Maestro suite, drawing a triclinic box around the domain 1 of LARGE1 with a distance buffer of 10 Å between the protein and the side of the box, and filling it with SPC water molecules. A 0.15 mol/L NaCl salt concentration was added, and additional Cl^- ions were added to neutralize the system (Table 1). The resulting systems both consisted of a 71 × 76 × 71 Å-sized box containing a total of about 30000 atoms, among which ∼8800 were water molecules and ∼4590 were protein atoms (Table 1).

The two systems were first minimized in order to relax the molecules into a local energy minimum and to remove the steric clashes, using the default protocol consisting of an initial steepest descent phase followed by a minimization with the LBFGS method, until convergence. The systems were then equilibrated using the Desmond standard NPT relaxation protocol with default parameters. We performed duplicate MD simulations of 500 ns each, using different random seeds for the assignment of the initial velocities, at a constant pressure of 1 atm and 300 K, saving the energies and the trajectories every

https://dx.doi.org/10.1021/acs.jcim.0c00281
J. Chem. Inf. Model. 2020, 60, 3145−3156

Table 1. Details of the Starting Structures for MD Simulations

system property	WT	S331F mutant
number of atoms	30,017	30,963
protein residues	276	276
protein atoms	4585	4594
water molecules	8773	8752
Cl^- ions	31	31
Na^+ ions	24	24
Mn^{2+} ion	1	1
size (Å)	71.2 × 66.4 × 70.5	71.2 × 66.4 × 70.5
ligand atoms	55	55
protein charge	+7	+7
ligand charge	−2	−2

20 ps. The Martyna−Tobias−Klein method was used for pressure coupling, and the Nose−Hoover thermostat was employed for the temperature bath, using the default settings for both, while a cut-off of 9 Å was used for the non-bond interactions. The analysis of trajectories was performed with the "simulation event analysis" tool of Maestro (Maestro-Desmond Interoperability Tools, Schrödinger, New York, NY, 2017), and the solvent accessible surface area (SASA) over the simulation time was calculated with VMD.[56] Discovery Studio (Dassault Systèmes BIOVIA, Discovery Studio, 2018, San Diego: Dassault Systèmes, 2018), Maestro, and VMD were used for the visual inspection of the three-dimensional structures and for the simulations results. The convergence of the MD simulations was assessed by calculating the root mean square inner product (RMSIP) using Bio3D library as implemented in the version 3.5.2 of the R package.[57,58]

The protein motion essential for the biological function of LARGE1 was analyzed by principal component analysis (PCA)-based dimensionality reduction of the atomic fluctuations of the Cα atoms in the simulated trajectory. The Bio3D library,[58] as implemented in the version 3.5.2 of the R package, was employed to perform and graphically represent the PCA of the WT and the S331F mutant trajectories. The conformations from the last 210 ns replica MD simulations were used for the wild-type (WT) and mutant enzymes. All the calculations were carried out employing an NVIDIA M4000 GPU and an Intel Xeon X5660 processor, on a HPZ820 workstation running Linux Centos 7 operating system. FoldX v5.0 was used to estimate the impact of the S331F mutation on protein stability (http://foldxsuite.crg.eu).[59] The algorithm calculates the free energy of folding of the WT and the mutated protein and estimates whether the mutation has a destabilizing ($\Delta\Delta G > 0$) or stabilizing ($\Delta\Delta G < 0$) effect.

Cloning and Expression of LARGE1 Domain 1 in *Escherichia coli*. The nucleotide sequence of LARGE1 domain 1 (residues 138−413) from *M. musculus* was synthesized and optimized for expression in *E. coli* by Invitrogen (GeneArt gene synthesis). The synthetic gene was cloned with extremities 5′BamHI-3′EcoRI downstream the thioredoxin gene in the vector pHisTrx, for expression in *E. coli* as a fusion product; the expression vector thus obtained was transformed into *E. coli* BL21 (DE3). A single colony from the transformants was picked and grown o.n. in LB + Amp 100 mg/L, then diluted 100-fold in fresh LB + Amp 100 mg/mL, and grown at 30 °C shaking at 220 rpm until OD600 ≈ 0.8; the expression was induced by adding IPTG to 0.5 mM final concentration. Protein production was checked on single aliquots collected at specific times after induction by lysing the cell pellets and running them on sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE). The bulk of cells was collected 3 h after induction, and the cell pellet was resuspended and lysed; the soluble proteins were separated from the insoluble ones by high-speed centrifugation and finally checked on SDS-PAGE in order to identify which fraction contained the fusion product. To increase the solubility of LARGE1 domain 1, the same construct was used to transform Arctic Xpress *E. coli* BL21 (DE3) competent cells (Agilent Technologies) following the manufacturer instructions. A single colony from the transformants was picked and grown o.n. in LB containing ampicillin 100 mg/L and gentamicin 20 mg/L, then diluted 200× in fresh LB containing no selection antibiotic, and grown at 30 °C shaking at 220 rpm for 3 h. The expression was then induced by adding IPTG to 1 mM final concentration and growing the cells for additional 24 h at 12 °C shaking at 200 rpm. Cells were harvested, resuspended in binding buffer (20 mM Tris HCl, 0.5 M NaCl, 15 mM imidazole, 2 mM Mn^{2+}), and lysed in a cell disruptor (Constant Systems, model Z plus 1.1 KW). The soluble and insoluble fractions were separated by centrifugation, and the supernatant was loaded on a 5 mL His trap Ni^{2+} column equilibrated in binding buffer, for isolation of the His-tagged fusion product by binding and imidazole elution.

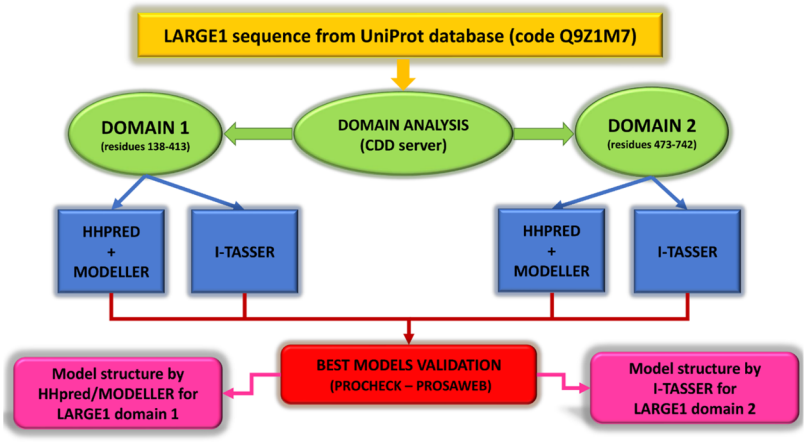


Figure 2. Schematic of the methodology employed for modelling LARGE1 domains 1 and 2.

https://dx.doi.org/10.1021/acs.jcim.0c00281
J. Chem. Inf. Model. 2020, 60, 3145−3156

Expression and purification outcomes were checked on SDS-PAGE.

RESULTS AND DISCUSSION

Molecular Modelling. A schematic summarizing the workflow of our molecular modelling approach is provided in Figure 2.

Molecular Modelling of LARGE1 Domain 1. We decided to use the murine LARGE1 sequence for molecular modelling due to the very high degree of sequence similarity between murine and human variants (100% sequence identity for domain 1 and 98.5 for domain 2, Supporting Information, Figure S1). It is worth noting that we have recently shown that the X-ray structures of murine and human N-terminal of α-dystroglycan (displaying the same degree of sequence similarity found between human and murine LARGE1) are identical.[60] As for the human point mutation of LARGE1 that we have introduced and analyzed in the model and in agreement with the above, all the residues in the region specifically perturbed by the mutation S331F in murine LARGE1 are identical to those in its human counterpart.

The xylosyltransferase and glucuronyltransferase activities of LARGE1 have been attributed to two different domains of the protein.[18] Indeed, the identification of functional domains by the CDD server revealed the presence of two domains, namely, domain 1 (residues 138−413) and domain 2 (residues 473−742), both predicted to belong to two glycosyltransferase families (family 8 and 49, respectively), in agreement with the data reported in the literature.[18,61] The absence of a suitable template structure in the Protein Data Bank for homology modelling of the whole-length protein led us to generate three-dimensional model structures of the two identified domains individually. The HHpred analysis on domain 1 (residues 138−413) identified three potential templates in the Protein Data Bank: the crystal structure of a protein belonging to the glycosyltransferase family 8 from *Anaerococcus prevotii* (PDB code 3TZT[62]), the crystal structure of the galactosyltransferase LgtC from *Neisseria meningitidis* (PDB code 1G9R[63]), and the crystal structure of the xyloside xylosyltransferase 1 from *M. musculus* (PDB code 4WMA[64]). Amongst these, 1G9R (2 Å resolution, 19% sequence identity) was selected as the template structure because it exhibited the lowest number of missing residues. The HHpred alignment was then used by MODELLER for model building. Structural models were generated, and the highest-scoring model (lowest PDF total energy) was selected for further analysis. The quality of the model was evaluated considering the overall stereochemistry by PROCHECK, and the resulting Ramachandran plot showed that 95.7% of residues was located in the most favorable regions (core and allowed), whereas the 3.9% was located within the generously allowed region and only the 0.4%, confined to Lys 386, in the disallowed region. The interaction energy of each residue with the structural surrounding environment, calculated with PROSA-Web, gave a global *Z*-score of −5.6 (276 residues), which falls well within the accepted range for proteins of the same size and indicates an overall good model quality (Supporting Information, Figure S2A). The PROSA-Web local quality analysis of the model, represented in the energy profile plot, shows that all the residues, except those in regions 265−285, have negative scores, where the more negative is the score, the more correctly modeled is the region (Supporting Information, Figure S2B).

For comparison, an alternative approach was used and the model of domain 1 was also generated on the I-TASSER server, which is based on a combination of multiple-threading alignments, iterative template fragment assembly simulations, and *ab initio* modelling algorithms. The template search stage of the I-TASSER procedure identified the crystal structures of the galactosyltransferase LgtC from *N. meningitidis* (PDB entry codes: 1G9R,[63] 1GA8,[63] and 1SS9[65]) of a glycosyltransferase family 8 from *A. prevotii* (PDB code 3TZT[62]) and of the xyloside-α-1,3-xylosyltransferase 1 (XXYLT1) from *Homo sapiens* (PDB code 4WM0[64]) as the best templates. In agreement with the HHpred results, all the templates detected by I-TASSER displayed the typical GT-A fold, consisting of two α/β/α domains with a continuous central β-sheet,[66] all belonging to family 8. Identified templates share sequence identity with the target sequence ranging from 19 to 24% (Supporting Information, Table S1). Five I-TASSER models were obtained, and the best scoring model displayed a TM-score of 0.79 ± 0.09, an estimated root-mean-square deviation (rmsd) of 4.8 ± 3.1 Å and, on a scale of accuracy ranging from −5 (lowest) to 2 (highest), a *C*-score value of 0.57. The model structure obtained from the I-TASSER procedure featured a lower percentage of residues (87%) in the core and allowed regions of the Ramachandran plot and 6.7 and 5.9% residues in the generously allowed and disallowed regions, respectively. The ProSA-web local quality analysis of this model reported negative energy values for all the residues, with the exception of the region 271−296, and the global quality analysis (276 residues) gave a *Z*-score of −6.92, which falls within the range of the experimentally resolved structures of the same size (Supporting Information, Figure S3A,B). The models generated by I-TASSER and HHpred were submitted to the COFACTOR software,[67] which identifies the most structurally similar enzymes in the Protein Data Bank, performing a local and global structure match. This analysis showed that the two generated models share the highest structural similarity with the crystallographic structure of the glycosyltransferase LgtC from *N. meningitidis* (PDB code number 1SS9), displaying a TM-score of 0.89 and 0.92 for the I-TASSER and HHpred model, respectively, as well as a Cα-rmsd lower than 2 and a sequence coverage of 94.6% for both the model structures. Given that a TM-score greater than 0.7 indicates that the structures have a 90% probability of being in the same topology family,[68] the two models of domain 1 can thus confidently be assigned to the GT-A glycosyltransferase family. For comparison, the calculated Cα-rmsd values between the MODELLER and I-TASSER LARGE1 models are 2.3 Å (domain 1) and 4.1 Å (domain 2), indicative of similar models. According to the validation results, the model structure obtained by the combined HHpred/MODELLER approach was selected as the best model for LARGE1 domain 1 for further analysis. The model structure of domain 1 consists of a central seven-stranded mainly parallel β-sheet (Figure 3A), sandwiched between 14 α-helices and a small β-sheet, which is conserved in all the proteins with a GT-A fold (Figure 3A,B).

On one side of the β-sheet, helix-C contacts the β1 and β2 strands, whereas helices G, J, and K contact the rest of the β-sheet. On the other side, a small β-sheet (strands β5 and β9) is located between helices A, B, and N that are in contact with the β1, β2, and β3 strands, whereas helices D, L, M, and E are in contact with the β6 and β8 strands (Figure 3). As found in other GT-A folded proteins, the small β-sheet formed by two antiparallel strands (β5 and β9) is linked to the β4 strand by a

https://dx.doi.org/10.1021/acs.jcim.0c00281
J. Chem. Inf. Model. 2020, 60, 3145−3156

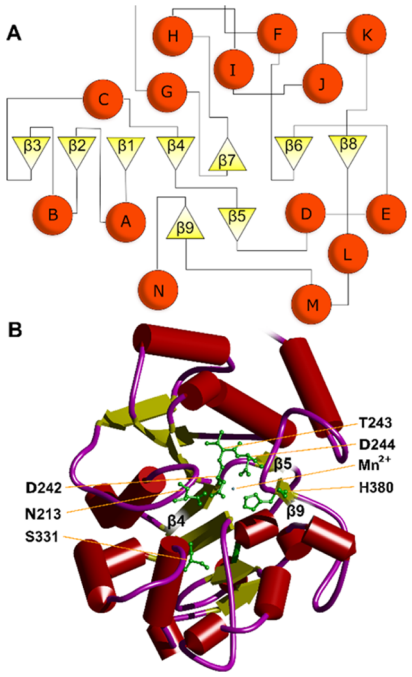


Figure 3. Structure and topology of LARGE1 domain 1. (A) Topology diagram of the domains. Helices are shown as red spheres, and β-strands are shown as yellow triangles. (B) Structural model represented as a cartoon, and the secondary structure is colored (helices in red, β-strands in yellow, and random coils in magenta). Individual side chains are represented in green as balls and sticks.

loop (L4) that contains a conserved DXD motif likely to bind the manganese cation (Mn^{2+})[69] responsible for the catalytic activity of domain 1.[18,22] The Mn^{2+} ion has a crucial role in the catalytic activity of LARGE1, and it is coordinated by the DXD motif residues as well as by the α- and β-phosphate moieties of the substrate.[63] In the predicted model structure of LARGE1 domain 1, the Mn^{2+} ion is coordinated by the NE2 atom of His380 of the β9 strand and by the carboxylate oxygen atoms of Asp242 and Asp244 belonging to the β4−β5 linker loop L4 (Figure 3B). Analogous interactions are established by the Mn^{2+} ion in the neighbor crystallographic structure of LgtC from *N. meningitidis* (1G9R) where the cation is coordinated by His244, Asp103, and Asp105, respectively (ref 63, Figure 4).

Full inspection of the structural alignment reveals that all major secondary structure elements of the model of LARGE1 domain 1 superimpose almost exactly with those of the LgtC glycosyl transferase, with only few exceptions related to the helical content (Figure 4).

Molecular Modelling of LARGE1 Domain 2. An analogous analysis was performed for LARGE1 domain 2. The top-ranking templates identified with a 99.7% probability by the template search methods of HHpred were all *N*-acetylgalactosaminyl transferases, namely, 1XHB,[70] 6E4R,[71] 6H0B,[72] 6IWR,[73] and 2D7I.[74] 1XHB, the crystal structure of the UDP-GalNAc: polypeptide α-*N*-acetylgalactosaminyltransferase-T1 from *M. musculus*, was selected as the best template because it exhibited the highest *P*-value (2.0×10^{-19}), highest probability (99.7%), and a sequence identity of 10% with the target. The stereochemical quality of the model with the lowest PDF value, generated by MODELLER on the basis of the HHpred alignment, was evaluated using PROCHECK. The Ramachandran plot calculation revealed that 91.5% of the residues are found in the most favorable regions (core and

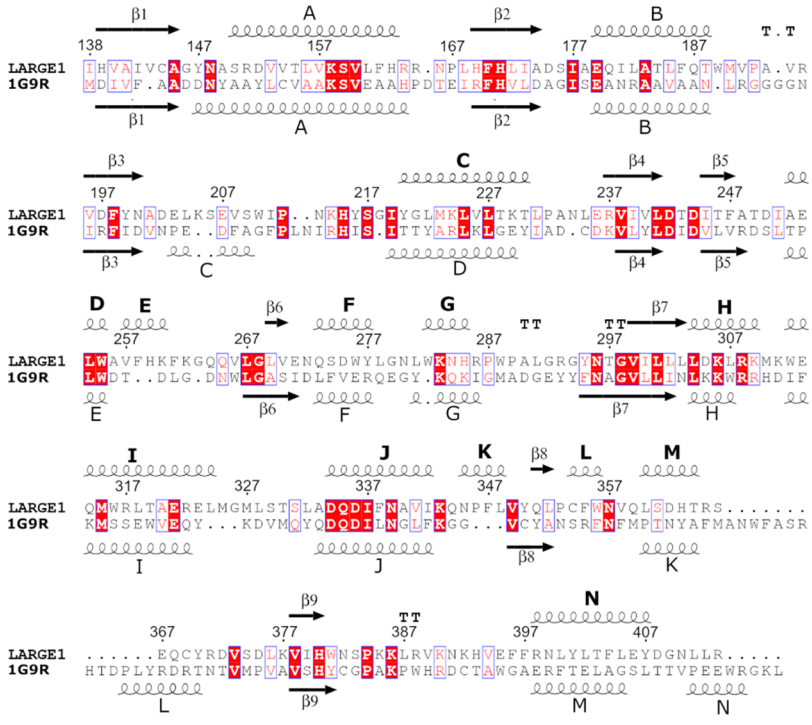


Figure 4. Structure-based sequence alignment of LARGE1 domain 1 and LgtC from *N. meningitidis*. Numbering is relative to the sequence of LARGE1, and the secondary structure elements of LgtC are based on Persson et al.[63] The figure was prepared using the ESPript server (http://espript.ibcp.fr). Identical residues are shown in white on red, and homologous residues are in red letters. Secondary structure details are represented (helices: squiggles, beta strands: arrows, and turns: TT letters).

https://dx.doi.org/10.1021/acs.jcim.0c00281
J. Chem. Inf. Model. 2020, 60, 3145−3156

allowed), 3.7% in the generously allowed region, and the remaining 4.9% in the disallowed region. ProSA-Web analysis of the model (270 residues) revealed a *Z*-score value of the target protein of −2.69, that is, within the range of the experimental protein structures of the same size, which indicates an acceptable overall quality of the model structure. However, the ProSA-Web analysis revealed that about 50% of the residues display positive values of interaction energy, suggesting errors in some regions of the model (Supporting Information, Figure S4A,B). The crystallographic templates found by I-TASSER to be suitable for a threading-*ab initio* model were the crystal structure of chondroitin polymerase from *E. coli* (PDB code 2Z86[75]), the structure of the human UDP-GalNAc:polypeptide *α*-*N*-acetylgalactosaminyltransferase-T2 (PDB codes 2FFU,[76] 5AJO[77] and 4D0T[78]), and the crystal structure of the human pp-GalNAc-T10 (pdb code 2D7I[74]). Notably, the PDB entry 4D0T was also used as a template for LARGE1 domain 2 model building in the work published by Bhattacharya.[39] Sequence identities between identified templates and LARGE1 domain 2 fall in the range 17−21% (see Supporting Information, Table S2, for details). The top scoring model of domain 2 produced by I-TASSER had a C-score value equal to −0.82, an estimated TM-score of 0.61 ± 0.14 Å and an estimated rmsd of 7.8 ± 4.4 Å. The closest structural analogue to the model of domain 2, found by I-TASSER in the Protein Data Bank is 2FFU, with a TM-score of 0.87 (rmsd 1.61 Å), where a TM-score higher than 0.5 generally indicates the same fold in SCOP/CATH. In agreement with the study of Bhattacharya and colleagues,[39] the crystal structure with PDB code 4D0T, that is, the human glycosyltransferase GalNAc-T2 in complex with UDP-GalNAc, the EA2 peptide, and manganese ion, was also found to have a close structural similarity to the model of LARGE1 domain 2 (TM-score of 0.84 and rmsd of 1.66 Å). The analysis of the stereochemical quality of this model, assessed with PROCHECK, showed that 95.5%, 1.6, and 2.8% of the residues are located within the core and allowed, generously allowed, and disallowed regions of the Ramachandran plot, respectively. The overall model quality shows a ProSA-web *Z*-score of −7.2, in agreement with the *Z*-scores calculated for the experimentally resolved structures in the Protein Data Bank (Supporting Information Figure S5A). The ProSA-web server showed that the interaction energy of each residue with the rest of the protein has negative values, apart from the residues within the two small regions 584−590 and 597−601 (Supporting Information, Figure S5B). Comparison of the results obtained using the two modelling approaches led to the selection of the structural model built with I-TASSER as the best model of LARGE1 domain 2 (Figure 5).

Although the I-TASSER server performed better than HHpred in terms of quality of the final model, the results from both methods agree on pointing toward domain 2 to be classified as an enzyme of the GT-A family, with a Rossman-like fold consisting of a two sandwiched *β*-sheets interposed between four *α*-helices on both sides (Figure 5A). Interestingly, although the fold, especially the arrangement of the *β*-sheets, is conserved between the two domains of LARGE1, the three-dimensional model of domain 2 shows a lower number of strands with respect to domain 1 (eight vs nine) (Figures 3A and 5A). Strand *β*8 undergoes a distortion at the level of Pro699 that makes its C-terminal portion parallel to the *β*6 strand and its N-terminal portion antiparallel to the *β*5 strand, thus creating the double *β*-sheet structure shown in Figure 5B.

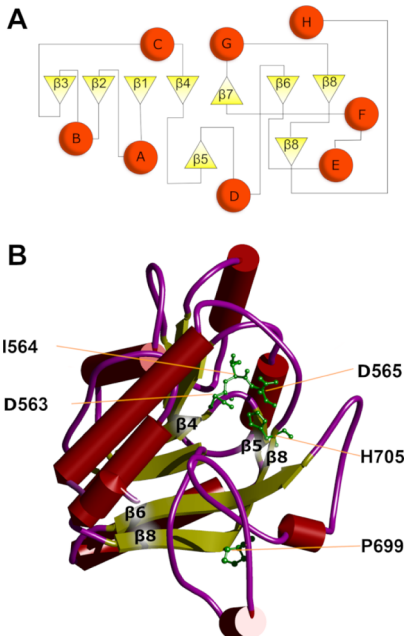


Figure 5. Structure and topology of LARGE1 domain 2. (A) Topology diagram of the domains. Helices are shown as red spheres, and *β*-strands are shown as yellow triangles. (B) Structural model is represented as a cartoon, and the secondary structure is colored (helices in red, *β*-strands in green, and random coils in magenta). Individual side chains are represented as green balls and sticks.

Interestingly, the third DXD motif, that in domain 2 is DID (563−565), essential for the glucuronyltransferase activity of LARGE1,[18,22] is located in a small unstructured turn between the strands *β*4 and *β*5, and it is structurally aligned with the DXH motif of the structural neighbor UDP-GalNAc:polypeptide *α*-*N*-acetylgalactosaminyltransferase-T2 from *H. sapiens*. The turn region is located deep into the binding cleft of the enzyme, with the two aspartates (Asp563 and Asp565) pointing toward the solvent and the isoleucine (Ile564) buried in the protein core, indicating an orientation that could favor the coordination of the carboxyl groups of the aspartates to a metal ion cofactor such as Mn^{2+}[79] (Figure 5B). It is noteworthy that His705, in the *β*8 strand, is conserved in 2FFU (His359, see sequence alignment in Figure 6) and that in the crystal structure, this residue coordinates the metal ion.

His705 in the model of domain 2 is in close contact with the aspartates 563 and 565 (distance shorter than 4 Å), suggesting an involvement of this basic residue in the catalytic cluster (Figure 5B).

Effects of the S331F Mutation on the Structure and Dynamics of LARGE1 Domain 1. The mutation of the LARGE1 residue Ser331 to Phe has been known for some time to cause a congenital muscular dystrophy phenotype characterized by eye malformations and mental retardation.[26] According to our *in silico* generated structure of domain 1, Ser331 is located in a random coil region (residues 325−333) between helices I and J, on the rim of the active-site cleft and close to Asn213, which belongs to the coil region between strand *β*3 and helix C (residues 200−219) (Figure 3B). Molecular dynamics simulations of the WT and of the S331F mutant were carried out to explore the conformational changes that occur upon S331F replacement and to assess its structural and dynamics effects on the enzyme function, with the final aim of understanding the molecular mechanism at the basis of

https://dx.doi.org/10.1021/acs.jcim.0c00281
J. Chem. Inf. Model. 2020, 60, 3145−3156

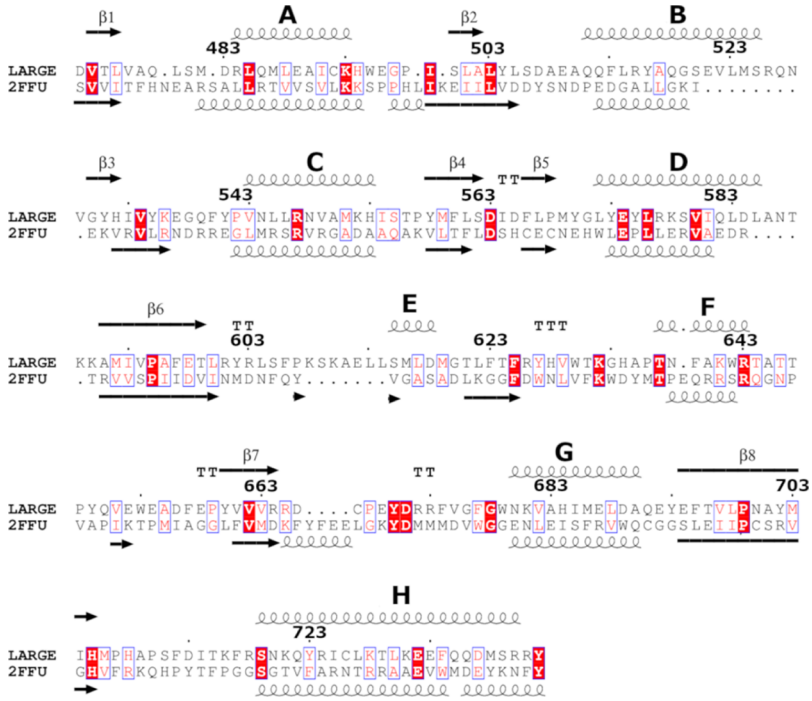


Figure 6. Structure-based sequence alignment of LARGE1 domain 2 and UDP-GalNAc:Polypeptide *α*-*N*-acetylgalactosaminyltransferase-2. The figure was prepared using the ESPript server (http://espript.ibcp.fr). Identical residues are shown in white on red, and homologous residues are in red letters. Secondary structure details are represented (helices: squiggles, beta strands: arrows, and turns: TT letters).

the pathological alterations induced by this mutation. Two different MD replicates were performed for both WT and S331F mutant. The stability of the two systems has been evaluated by measuring the rmsd of the alpha carbons with respect to their positions at time 0, over the simulation time. After an equilibration of 40 ns, the rmsd reached a plateau with a stable value of 4.4 ± 0.2 and of 4.5 ± 0.2 Å for WT and S311F, respectively (Figure 7A), until the end of the simulations. An analysis of the C*α* atom root mean square fluctuation (RMSF) with respect to their average positions calculated over the last 200 ns was then performed, in order to identify the regions of the protein whose movements are mostly affected by the replacement of Ser311 with a Phe residue (Figure 7B). The results indicate that most of the residues share similar fluctuations. The differences in RMSF between the two simulated systems are shown in Figure 7C, which reveals larger fluctuations from the averaged MD conformations in the unstructured regions spanning residues 178−180, 290−292, 324, 367−368, and 410−413.

Although several residues were found to display significant differences in RMSF upon mutation, the important catalytic residues such as the two DXD motifs and His380 were not affected in their atomic fluctuations. An alteration of total, hydrophobic, and hydrophilic SASAs in the mutant protein with respect to the WT was also observed. A representative plot of SASA for all-atom, hydrophobic, and polar residues against time is shown in Figure 8.

Compared to the WT, the S331F mutant showed higher all-atom SASAs, which converged to an average value of 15145 ± 248 Å^2 (14777 ± 327 Å^2 in the WT protein). Notably, the hydrophobic SASA of the mutant increased about 12.7% over the simulation time with respect to the WT protein (5115 ± 265 and 4538 ± 166 Å^2, respectively) (Figure 8), a feature that could significantly affect the interactions at the protein surface.

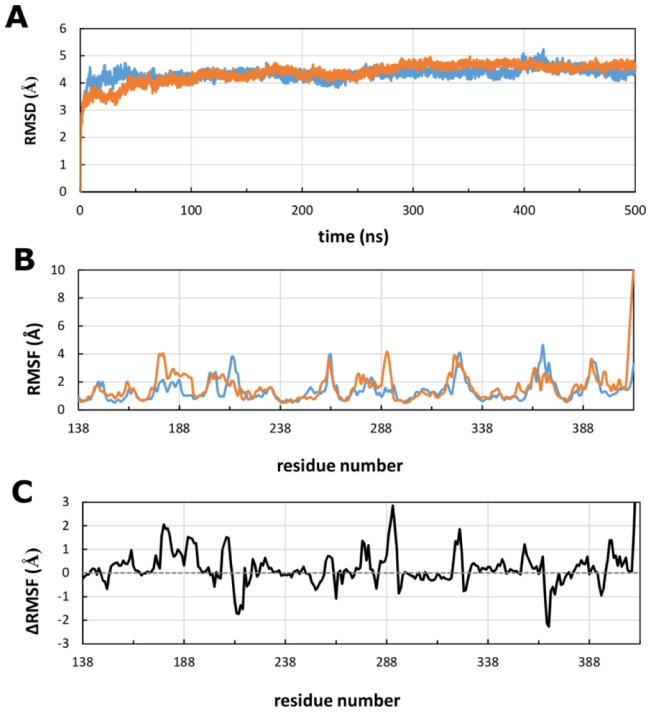


Figure 7. Analysis of MD trajectories. (A) C*α*-rmsd from the starting structure. (B) Residue-based C*α*-RMSF relative to the average structure. (C) ΔRMSF: difference between the RMSF of the mutant S331F and that of the WT for each residue. A positive value indicates larger fluctuations for the mutant, whereas a negative value means larger fluctuations for the WT. Plots relative to the WT system are colored in blue and those relative to the S331F mutant in orange.

In order to check the convergence of the MD simulations, we have performed the RMSIP analysis on two equilibrated

https://dx.doi.org/10.1021/acs.jcim.0c00281
J. Chem. Inf. Model. 2020, 60, 3145−3156

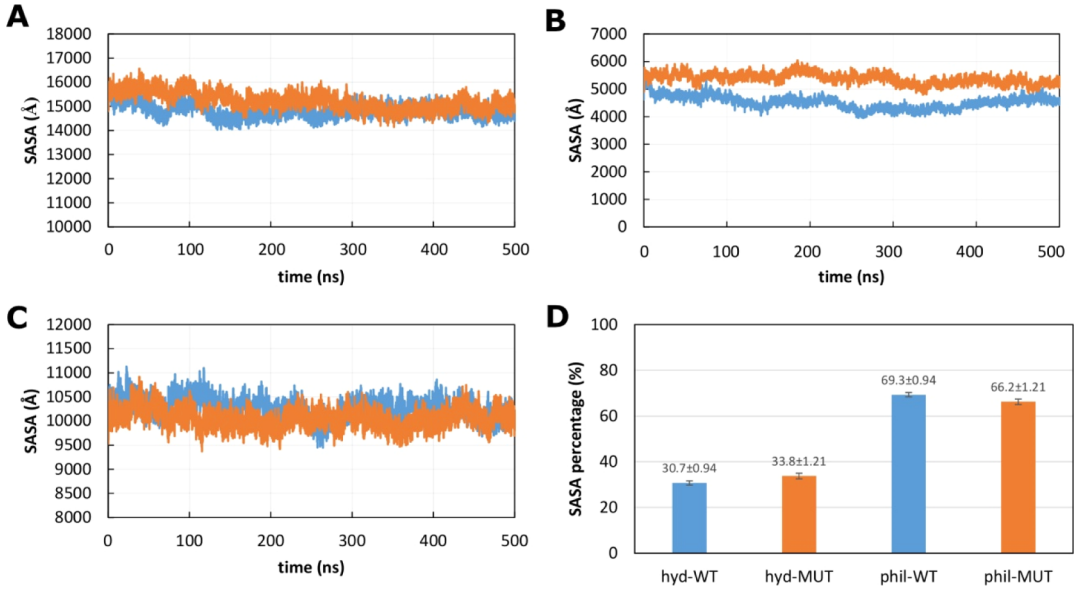


Figure 8. Analysis of MD trajectories. Time series of the SASA of LARGE1 domain 1. The evolution of the total, hydrophobic, and hydrophilic SASA over the simulation time is shown in panels (A–C), respectively. (D) Percentage of hydrophobic SASA (hyd-SASA) and hydrophilic SASA (phil-SASA) with respect to the total SASA. Each time-series refers to both the simulation replicas. For color code, see Figure 7.

regions of the trajectories. The obtained high RMSIP values of 0.80 (replica#1), 0.82 (replica#2) and 0.82 (replica#1), 0.83 (replica#2) for the WT and mutant proteins, respectively, are indicative of an adequate level of convergence.

The local conformational changes that occur upon S331F mutation have been inspected by analyzing the non-bonding interactions established by Ser/Phe331. In the WT simulations, Ser331 is hydrogen bonded, by means of its OG and O backbone oxygen atoms, to Asn213 (O backbone oxygen, ND2 and OD1 atoms), whereas in the mutant, Phe331 moves away from Asn213, thus causing a conformational change in loops 200–219 and 325–333 (Figures 9 and 10). All plots of the duplicate trajectories were similar and are shown in the Supporting Information (Figures S6–S8).

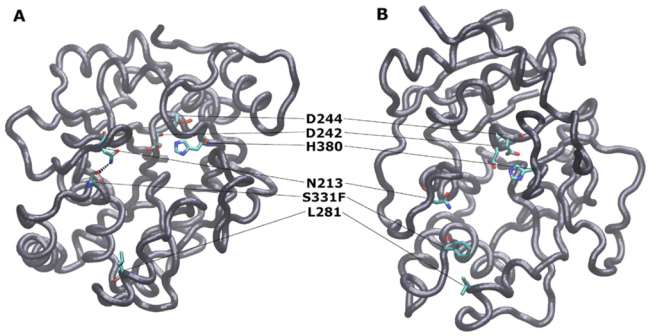


Figure 9. Representative structures of the simulated LARGE1 WT and S331F mutant. The structurally closest frame to the averaged structure from the MD simulations (obtained selecting the frame with the lowest Cα-rmsd with respect to the average Cα-coordinates calculated over the equilibrated trajectory) is reported for each simulated system. Panel (A) shows the average structures from the WT simulation, and panel (B) shows the average structures from the S331F mutant simulation. The backbone is represented as tube colored in grey, the atomic details are represented as sticks colored by atom type, and the black dashed line represents the hydrogen bond interaction.

A weak destabilizing effect of the S331F mutation ($\Delta\Delta G$ = 0.21 kcal/mol) was also calculated by FoldX. This $\Delta\Delta G$ value, which cannot be directly related to a change in enzyme function,[80] suggests that a phenylalanine in position 331 is not fully compatible with the surrounding chemical environment, as already observed in MD. Time evolution of the secondary structure elements along the MD simulation was also analyzed, showing the stability of the β-strand elements during the entire MD simulations and the structural transition of few α-helices into 3_{10}-helices (Supporting Information, Figure S9).

In order to probe the changes in conformational dynamics induced by the S331F mutation, analysis of the MD trajectory by PCA using the Ca atoms was also performed. Results indicate that the first twenty principal components (PCs) account for 90.7% (WT) and 92.8% (mutant) of the motions observed in the last 400 ns of the trajectories, whereas the first three PCs account for the 74.5% (WT) and the 82.3% (mutant) motions (Supporting Information, Figure S10A). The trace values of the diagonalized covariance matrix were found to be 1706.9 Å^2 (WT) and 2145.8 Å^2 (mutant), indicating an increased flexibility of S331F in the collective motion as compared to the WT (Supporting Information, Figure S10A,B). Analysis of the contribution of each residue to the first principal component revealed that the Ser 331 to Phe mutation alters the protein motion in the loop regions surrounding the residue 331. Notably, the peaks corresponding to residues 211–216 and 282–291 are shifted backward (to 201–210 and 276–288, respectively) in the mutated protein (Supporting Information, Figure S10C). This analysis suggests that the S331F mutation introduces local differences in the motion of the active site rim of domain 1, which might affect the interaction with the substrate.

Cloning and Expression of LARGE1 Domain 1 in *E. coli*. A series of first attempts at heterologous expression of domain 1 as revealed by our modelling results were carried out. The DNA sequence coding for the region comprising residues 138–413 of *M. musculus* LARGE1 was custom-synthesized and optimized for the expression in *E. coli*. The synthetic gene thus

https://dx.doi.org/10.1021/acs.jcim.0c00281
J. Chem. Inf. Model. 2020, 60, 3145–3156

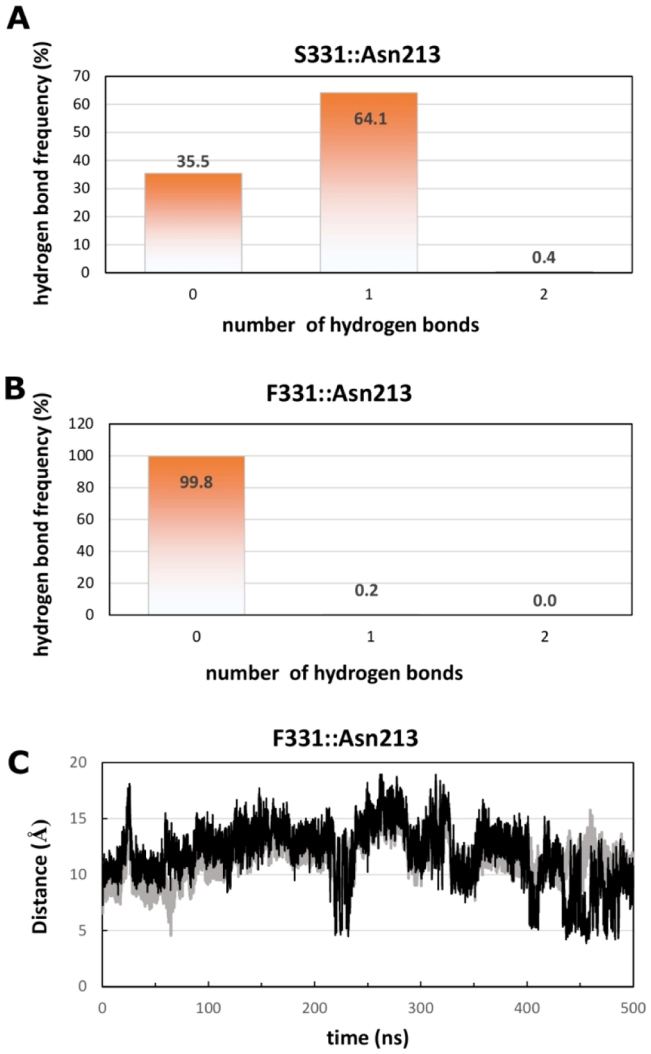


Figure 10. Analysis of MD trajectories. Frequency of the hydrogen bonds (%) between Ser331 and Asn213 in domain 1 WT (A) and S331F (B). Time evolution of the distances between the geometric centroids of the Phe331 and Asn213 side chain atoms (C) and between the O backbone atoms of Phe331 and Asn213 (black and grey, respectively).

obtained was cloned into the pHisTrx vector for expression downstream the fusion partner His_6-tagged thioredoxin and subsequently transformed into *E. coli* strain BL21 (DE3). Although preliminary, the evidence that the protein can be recovered in the soluble form upon heterologous expression in the presence of helper proteins such as the chaperonin GroEL (see Supporting Information, Figures S11−S13) reinforce the notion that the subdomain 1 represents an autonomous folding unit, as suggested by our modelling results.

Further experiments are granted aimed at producing quantitative amounts of domain 1 for future biochemical and structural studies. Recombinant expression of individual LARGE1 domains 1 and 2, as well of the two domains together in a full-length LARGE1 construct, would allow a thorough analysis of the enzymatic proprieties and activity of this glycosyltransferase. The two catalytic domains of LARGE1 are connected by a flexible linker (ranging from amino acid 414 to 472) that due to its nature could not be accurately modeled. A very interesting question is whether the two domains establish long-range interactions and work somehow cooperatively during the sequential synthesis of each added disaccharide block (*i.e.* the final product of every full LARGE1 enzymatic cycle) or rather function in an independent fashion. Whether the binding of each substrate to each catalytic domain induces any allosteric long-range effects influencing the enzymatic behavior of the other catalytic domain remains to be determined at the current stage. Noteworthy, also the N-terminal region of α-dystroglycan[81] harbors two different domains (IG1 and S6), and it is intriguing to foresee a scenario in which the two DG domains would be involved reciprocally in chaperoning the activity of the two catalytic domains of LARGE1.

CONCLUSIONS

Our combined use of molecular modelling approaches allowed for a three-dimensional description of the two domains respectively responsible for the xylosyltransferase and glucuronyltransferase activities of LARGE1. This work characterized the three DXD motifs, each likely to bind a manganese cation (Mn^{2+}), that constitute the active sites for the glycosyltransferase activities of the two domains,[22,69,82] thus offering some initial insight into the possible mechanism of binding of the sugar donor and acceptor. Molecular dynamic simulations of the S331F mutant, carried out in comparison with the WT counterpart, points to a more flexible and less stable enzyme, as suggested by an increased RMSF and all-atom and hydrophobic SASA parameters, compared to the WT protein. Indeed, a reduced stability could be at the basis of the poorer enzymatic activity of this variant of LARGE. This would eventually lead to the hypoglycosylation of α-DG observed in a patient affected by a severe form of congenital muscular dystrophy with anomalies of eye and brain linked to this missense mutation.[26]

Our study demonstrates that in the absence of experimental structural data on LARGE1, template-based three-dimensional modelling of the enzyme can offer some insight into its overall structural features, and the analysis of its dynamic behavior by *in silico* methods can help to elucidate the molecular mechanism underneath the diseases linked to missense mutations of the enzyme. In addition, our modelling study might pave the road for the analysis of potential complexes formed by LARGE1 and the α-dystroglycan N-terminal domain (αDGN), whose structure we have previously solved.[81]

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.0c00281.

Sequence alignment of human and murine LARGE1 proteins; structural analysis of the LARGE1 domain 1 model structure generated by the I-TASSER server; sequence identity between LARGE domain 1 and the templates used by I-TASSER to build the model; structural analysis of the LARGE1 domain 2 model generated by HHPRED/MODELLER approach; structural analysis of the LARGE1 domain 2 model generated by the I-TASSER server; analysis of MD trajectories; time evolution of the secondary structural elements along the MD simulations generated by VMD [78]; projection of the enzyme conformations onto the first two PCs for the wild-type (in blue) and mutant (in

https://dx.doi.org/10.1021/acs.jcim.0c00281
J. Chem. Inf. Model. 2020, 60, 3145−3156

orange) enzymes; and cloning and expression of LARGE1 domain 1 (PDF)

AUTHOR INFORMATION

Corresponding Authors

Maria Giulia Bigotti − *School of Translational Health Sciences, Research Floor Level 7, Bristol Royal Infirmary, BS2 8HW Bristol, U.K.; School of Biochemistry, University Walk, University of Bristol, BS8 1TD Bristol, U.K.;* Email: G.Bigotti@bristol.ac.uk

Maria Cristina De Rosa − *Institute of Chemical Sciences and Technologies "Giulio Natta" (SCITEC)—CNR, 00168 Rome, Italy;* orcid.org/0000-0002-9611-2490; Email: mariacristina.derosa@cnr.it

Authors

Benedetta Righino − *Dipartimento di Scienze Biotecnologiche di Base, Cliniche Intensivologiche e Perioperatorie, Università Cattolica del Sacro Cuore, 00168 Rome, Italy*

Manuela Bozzi − *Dipartimento di Scienze Biotecnologiche di Base, Cliniche Intensivologiche e Perioperatorie, Università Cattolica del Sacro Cuore, 00168 Rome, Italy; Institute of Chemical Sciences and Technologies "Giulio Natta" (SCITEC)—CNR, 00168 Rome, Italy*

Davide Pirolli − *Institute of Chemical Sciences and Technologies "Giulio Natta" (SCITEC)—CNR, 00168 Rome, Italy*

Francesca Sciandra − *Institute of Chemical Sciences and Technologies "Giulio Natta" (SCITEC)—CNR, 00168 Rome, Italy*

Andrea Brancaccio − *Institute of Chemical Sciences and Technologies "Giulio Natta" (SCITEC)—CNR, 00168 Rome, Italy; School of Biochemistry, University Walk, University of Bristol, BS8 1TD Bristol, U.K.;* orcid.org/0000-0003-4690-8826

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.0c00281

Author Contributions

B.R. and M.B. have equally contributed to this work. The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Funding

This work was supported by a Wellcome Trust Career Re-entry Fellowship (097350/Z/11/Z) to M.G.B. and by an AFM (French Telethon) Grant (no. 20009) to A.B.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The funders played no role in the design of the study, in the collection, analysis, and interpretation of data, or in the decision to submit the manuscript for publication. The School of Biochemistry of the University of Bristol is acknowledged for hosting A.B.

ABBREVIATIONS

DG, dystroglycan; GT, glycosyltransferase; MD, molecular dynamics; RMSIP, root mean square inner product; PCA, principal component analysis; WT, wild type; CDD, Conserved Domain Database; PDB, protein data bank; rmsd, root mean square deviation; RMSF, root mean square fluctuation; SASA, solvent accessible surface areas

REFERENCES

(1) Lau, K. S.; Partridge, E. A.; Grigorian, A.; Silvescu, C. I.; Reinhold, V. N.; Demetriou, M.; Dennis, J. W. Complex N-glycan number and degree of branching cooperate to regulate cell proliferation and differentiation. *Cell* **2007**, *129*, 123−134.

(2) Neelamegham, S.; Mahal, L. K. Multi-level regulation of cellular glycosylation: from genes to transcript to enzyme to structure. *Curr. Opin. Struct. Biol.* **2016**, *40*, 145−152.

(3) Spiro, R. G. Protein glycosylation: nature, distribution, enzymatic formation, and disease implications of glycopeptide bonds. *Glycobiology* **2002**, *12*, 43R−56R.

(4) Reily, C.; Stewart, T. J.; Renfrow, M. B.; Novak, J. Glycosylation in health and disease. *Nat. Rev. Nephrol.* **2019**, *15*, 346−366.

(5) Bozzi, M.; Morlacchi, S.; Bigotti, M. G.; Sciandra, F.; Brancaccio, A. Functional diversity of dystroglycan. *Matrix Biol.* **2009**, *28*, 179−187.

(6) Brancaccio, A.; Schulthess, T.; Gesemann, M.; Engel, J. Electron microscopic evidence for a mucin-like region in chick muscle α-dystroglycan. *FEBS Lett.* **1995**, *368*, 139−142.

(7) Dempsey, C. E.; Bigotti, M. G.; Adams, J. C.; Brancaccio, A. Analysis of α-dystroglycan/LG domain binding modes: investigating protein motifs that regulate the affinity of isolated LG domains. *Front. Mol. Biosci.* **2019**, *6*, 18.

(8) Ervasti, J. M.; Campbell, K. A role for the dystrophin-glycoprotein complex as a transmembrane linker between laminin and actin. *J. Cell Biol.* **1993**, *122*, 809−823.

(9) Bowe, M. A.; Deyst, K. A.; Leszyk, J. D.; Fallon, J. R. Identification and purification of an agrin receptor from torpedo postsynaptic membranes: A heteromeric complex related to the dystroglycans. *Neuron* **1994**, *12*, 1173−1180.

(10) Talts, J. F.; Andac, Z.; Göhring, W.; Brancaccio, A.; Timpl, R. Binding of the G domains of laminin α1 and α2 chains and perlecan to heparin, sulfatides, α-dystroglycan and several extracellular matrix proteins. *EMBO J.* **1999**, *18*, 863−870.

(11) Sato, S.; Omori, Y.; Katoh, K.; Kondo, M.; Kanagawa, M.; Miyata, K.; Funabiki, K.; Koyasu, T.; Kajimura, N.; Miyoshi, T.; Sawai, H.; Kobayashi, K.; Tani, A.; Toda, T.; Usukura, J.; Tano, Y.; Fujikado, T.; Furukawa, T. Pikachurin, a dystroglycan ligand, is essential for photoreceptor ribbon synapse formation. *Nat. Neurosci.* **2008**, *11*, 923−931.

(12) Sugita, S.; Saito, F.; Tang, J.; Satz, J.; Campbell, K.; Südhof, T. C. A stoichiometric complex of neurexins and dystroglycan in brain. *J. Cell Biol.* **2001**, *154*, 435−446.

(13) Wright, K. M.; Lyon, K. A.; Leung, H.; Leahy, D. J.; Ma, L.; Ginty, D. D. Dystroglycan organizes axon guidance cue localization and axonal pathfinding. *Neuron* **2012**, *76*, 931−944.

(14) Yoshida-Moriguchi, T.; Campbell, K. P. Matriglycan: a novel polysaccharide that links dystroglycan to the basement membrane. *Glycobiology* **2015**, *25*, 702−713.

(15) Taniguchi-Ikeda, M.; Morioka, I.; Iijima, K.; Toda, T. Mechanistic aspects of the formation of α-dystroglycan and therapeutic research for the treatment of α-dystroglycanopathy: A review. *Mol. Aspect. Med.* **2016**, *51*, 115−124.

(16) Michele, D. E.; Barresi, R.; Kanagawa, M.; Saito, F.; Cohn, R. D.; Satz, J. S.; Dollar, J.; Nishino, I.; Kelley, R. I.; Somer, H.; Straub, V.; Mathews, K. D.; Moore, S. A.; Campbell, K. P. Post-translational disruption of dystroglycan−ligand interactions in congenital muscular dystrophies. *Nature* **2002**, *418*, 417−421.

(17) Yoshida-Moriguchi, T.; Willer, T.; Anderson, M. E.; Venzke, D.; Whyte, T.; Muntoni, F.; Lee, H.; Nelson, S. F.; Yu, L.; Campbell, K. P. SGK196 is a glycosylation-specific O-mannose kinase required for dystroglycan function. *Science* **2013**, *341*, 896−899.

(18) Inamori, K.-i.; Yoshida-Moriguchi, T.; Hara, Y.; Anderson, M. E.; Yu, L.; Campbell, K. P. Dystroglycan function requires xylosyl- and glucuronyltransferase activities of LARGE. *Science* **2012**, *335*, 93−96.

(19) Kanagawa, M.; Kobayashi, K.; Tajiri, M.; Manya, H.; Kuga, A.; Yamaguchi, Y.; Akasaka-Manya, K.; Furukawa, J.-i.; Mizuno, M.; Kawakami, H.; Shinohara, Y.; Wada, Y.; Endo, T.; Toda, T. Identification of a post-translational modification with ribitol-phosphate and its defect in muscular dystrophy. *Cell Rep.* **2016**, *14*, 2209−2223.
(20) Praissman, J. L.; Live, D. H.; Wang, S.; Ramiah, A.; Chinoy, Z. S.; Boons, G.-J.; Moremen, K. W.; Wells, L. B4GAT1 is the priming enzyme for the LARGE-dependent functional glycosylation of α-dystroglycan. *eLife* **2014**, *3*, No. e03943.
(21) Yoshida-Moriguchi, T.; Yu, L.; Stalnaker, S. H.; Davis, S.; Kunz, S.; Madson, M.; Oldstone, M. B. A.; Schachter, H.; Wells, L.; Campbell, K. P. O-mannosyl phosphorylation of α-dystroglycan is required for laminin binding. *Science* **2010**, *327*, 88−92.
(22) Brockington, M.; Torelli, S.; Prandini, P.; Boito, C.; Dolatshad, N. F.; Longman, C.; Brown, S. C.; Muntoni, F. Localization and functional analysis of the LARGE family of glycosyltransferases: significance for muscular dystrophy. *Hum. Mol. Genet.* **2005**, *14*, 657−665.
(23) Briggs, D. C.; Yoshida-Moriguchi, T.; Zheng, T.; Venzke, D.; Anderson, M. E.; Strazzulli, A.; Moracci, M.; Yu, L.; Hohenester, E.; Campbell, K. P. Structural basis of laminin binding to the LARGE glycans on dystroglycan. *Nat. Chem. Biol.* **2016**, *12*, 810.
(24) Hara, Y.; Balci-Hayta, B.; Yoshida-Moriguchi, T.; Kanagawa, M.; Beltrán-Valero de Bernabé, D.; Gündeşli, H.; Willer, T.; Satz, J. S.; Crawford, R. W.; Burden, S. J.; Kunz, S.; Oldstone, M. B. A.; Accardi, A.; Talim, B.; Muntoni, F.; Topaloğlu, H.; Dinçer, P.; Campbell, K. P. A dystroglycan mutation associated with limb-girdle muscular dystrophy. *N. Engl. J. Med.* **2011**, *364*, 939−946.
(25) Clarke, N. F.; Maugenre, S.; Vandebrouck, A.; Urtizberea, J. A.; Willer, T.; Peat, R. A.; Gray, F.; Bouchet, C.; Manya, H.; Vuillaumier-Barrot, S.; Endo, T.; Chouery, E.; Campbell, K. P.; Mégarbané, A.; Guicheney, P. Congenital muscular dystrophy type 1D (MDC1D) due to a large intragenic insertion/deletion, involving intron 10 of the LARGE gene. *Eur. J. Hum. Genet.* **2011**, *19*, 452.
(26) Clement, E.; Mercuri, E.; Godfrey, C.; Smith, J.; Robb, S.; Kinali, M.; Straub, V.; Bushby, K.; Manzur, A.; Talim, B.; Cowan, F.; Quinlivan, R.; Klein, A.; Longman, C.; McWilliam, R.; Topaloglu, H.; Mein, R.; Abbs, S.; North, K.; Barkovich, A. J.; Rutherford, M.; Muntoni, F. Brain involvement in muscular dystrophies with defective dystroglycan glycosylation. *Ann. Neurol.* **2008**, *64*, 573−582.
(27) Godfrey, C.; Clement, E.; Mein, R.; Brockington, M.; Smith, J.; Talim, B.; Straub, V.; Robb, S.; Quinlivan, R.; Feng, L.; Jimenez-Mallebrera, C.; Mercuri, E.; Manzur, A. Y.; Kinali, M.; Torelli, S.; Brown, S. C.; Sewry, C. A.; Bushby, K.; Topaloglu, H.; North, K.; Abbs, S.; Muntoni, F. Refining genotype - phenotype correlations in muscular dystrophies with defective glycosylation of dystroglycan. *Brain* **2007**, *130*, 2725−2735.
(28) Longman, C. Mutations in the human LARGE gene cause MDC1D, a novel form of congenital muscular dystrophy with severe mental retardation and abnormal glycosylation of -dystroglycan. *Hum. Mol. Genet.* **2003**, *12*, 2853−2861.
(29) Mercuri, E.; Messina, S.; Bruno, C.; Mora, M.; Pegoraro, E.; Comi, G. P.; D'Amico, A.; Aiello, C.; Biancheri, R.; Berardinelli, A.; Boffi, P.; Cassandrini, D.; Laverda, A.; Moggio, M.; Morandi, L.; Moroni, I.; Pane, M.; Pezzani, R.; Pichiecchio, A.; Pini, A.; Minetti, C.; Mongini, T.; Mottarelli, E.; Ricci, E.; Ruggieri, A.; Saredi, S.; Scuderi, C.; Tessa, A.; Toscano, A.; Tortorella, G.; Trevisan, C. P.; Uggetti, C.; Vasco, G.; Santorelli, F. M.; Bertini, E. Congenital muscular dystrophies with defective glycosylation of dystroglycan: A population study. *Neurology* **2009**, *72*, 1802−1809.
(30) Barresi, R.; Michele, D. E.; Kanagawa, M.; Harper, H. A.; Dovico, S. A.; Satz, J. S.; Moore, S. A.; Zhang, W.; Schachter, H.; Dumanski, J. P.; Cohn, R. D.; Nishino, I.; Campbell, K. P. LARGE can functionally bypass α-dystroglycan glycosylation defects in distinct congenital muscular dystrophies. *Nat. Med.* **2004**, *10*, 696−703.
(31) Beltrán, D.; Anderson, M. E.; Bharathy, N.; Settelmeyer, T. P.; Svalina, M. N.; Bajwa, Z.; Shern, J. F.; Gultekin, S. H.; Cuellar, M. A.; Yonekawa, T.; Keller, C.; Campbell, K. P. Exogenous expression of the glycosyltransferase LARGE1 restores α-dystroglycan matriglycan and laminin binding in rhabdomyosarcoma. *Skeletal Muscle* **2019**, *9*, 11.
(32) Lee, J.; Freddolino, P. L.; Zhang, Y. Ab Initio Protein Structure Prediction. In *From Protein Structure to Function with Bioinformatics*; Rigden, D., Ed.; Springer Netherlands: Dordrecht, 2017; pp 3−35.
(33) Botta, B.; Delle Monache, G.; De Rosa, M. C.; Carbonetti, E.; Botta, M.; Corelli, F.; Misiti, D.; Misiti, D. Synthesis of C-alkyl calix[4]arenes. 3. Acid-catalyzed rearrangement of 2,6-dimethoxycinnamate prior to tetramerization to calix[4]arenes. *J. Org. Chem.* **1995**, *60*, 3657−3662.
(34) Kollman, P.; van Gunsteren, W. F. Molecular mechanics and dynamics in protein design. *Methods in Enzymology*; Elsevier, 1987; Vol. *154*, pp 430−449.
(35) Summa, C. M.; Levitt, M. Near-native structure refinement using in vacuo energy minimization. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 3177−3182.
(36) Moult, J.; Fidelis, K.; Kryshtafovych, A.; Schwede, T.; Tramontano, A. Critical assessment of methods of protein structure prediction: Progress and new directions in round XI. *Proteins* **2016**, *84*, 4−14.
(37) Zhang, Y.; Skolnick, J. Segment assembly, structure alignment and iterative simulation in protein structure prediction. *BMC Biol.* **2013**, *11*, 44.
(38) Zheng, W.; Zhang, C.; Bell, E. W.; Zhang, Y. I-TASSER gateway: A protein structure and function prediction server powered by XSEDE. *Future Generat. Comput. Syst.* **2019**, *99*, 73−85.
(39) Bhattacharya, S.; Das, A.; Bagchi, A. In-silico structural analysis of E509K mutation in LARGE and T192M mutation in alpha dystroglycan in the inhibition of glycosylation of α dystroglycan by LARGE. *Comput. Biol. Chem.* **2016**, *64*, 313−321.
(40) Söding, J.; Remmert, M. Protein sequence comparison and fold recognition: progress and good-practice benchmarking. *Curr. Opin. Struct. Biol.* **2011**, *21*, 404−411.
(41) Yang, J.; Yan, R.; Roy, A.; Xu, D.; Poisson, J.; Zhang, Y. The I-TASSER Suite: protein structure and function prediction. *Nat. Methods* **2015**, *12*, 7−8.
(42) Adiyaman; McGuffin. McGuffin Methods for the refinement of protein structure 3D models. *Int. J. Mol. Sci.* **2019**, *20*, 2301.
(43) Pirolli, D.; Carelli Alinovi, C.; Capoluongo, E.; Satta, M. A.; Concolino, P.; Giardina, B.; De Rosa, M. C. Insight into a novel p53 single point mutation (G389E) by molecular dynamics simulations. *Int. J. Mol. Sci.* **2010**, *12*, 128−140.
(44) Righino, B.; Minucci, A.; Pirolli, D.; Capoluongo, E.; Conti, G.; De Luca, D.; De Rosa, M. C. In silico investigation of the molecular effects caused by R123H variant in secretory phospholipase A2-IIA associated with ARDS. *J. Mol. Graph. Model.* **2018**, *81*, 68−76.
(45) Bairoch, A.; Apweiler, R.; Wu, C. H.; Barker, W. C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M.; Martin, M. J.; Natale, D. A.; O'Donovan, C.; Redaschi, N.; Yeh, L. S. The Universal Protein Resource (UniProt). *Nucleic Acids Res.* **2005**, *33*, D154−D159.
(46) Marchler-Bauer, A.; Derbyshire, M. K.; Gonzales, N. R.; Lu, S.; Chitsaz, F.; Geer, L. Y.; Geer, R. C.; He, J.; Gwadz, M.; Hurwitz, D. I.; Lanczycki, C. J.; Lu, F.; Marchler, G. H.; Song, J. S.; Thanki, N.; Wang, Z.; Yamashita, R. A.; Zhang, D.; Zheng, C.; Bryant, S. H. NCBI's conserved domain database. *Nucleic Acids Res.* **2015**, *43*, D222−D226.
(47) De Rosa, M. C.; Pirolli, D.; Bozzi, M.; Sciandra, F.; Giardina, B.; Brancaccio, A. A second Ig-like domain identified in dystroglycan by molecular modelling and dynamics. *J. Mol. Graph. Model.* **2011**, *29*, 1015−1024.
(48) Soding, J.; Biegert, A.; Lupas, A. N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* **2005**, *33*, W244−W248.
(49) Sali, A.; Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **1993**, *234*, 779−815.

https://dx.doi.org/10.1021/acs.jcim.0c00281
J. Chem. Inf. Model. 2020, 60, 3145−3156

(50) Roy, A.; Kucukural, A.; Zhang, Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.* **2010**, *5*, 725−738.
(51) Hughey, R.; Krogh, A. Hidden Markov models for sequence analysis: extension and analysis of the basic method. *Bioinformatics* **1996**, *12*, 95−107.
(52) Laskowski, R. A.; Moss, D. S.; Thornton, J. M. Main-chain bond lengths and bond angles in protein structures. *J. Mol. Biol.* **1993**, *231*, 1049−1067.
(53) Wiederstein, M.; Sippl, M. J. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res.* **2007**, *35*, W407−W410.
(54) Shivakumar, D.; Williams, J.; Wu, Y.; Damm, W.; Shelley, J.; Sherman, W. Prediction of absolute solvation free energies using molecular dynamics free energy perturbation and the OPLS force field. *J. Chem. Theory Comput.* **2010**, *6*, 1509−1519.
(55) Madhavi Sastry, G.; Adzhigirey, M.; Day, T.; Annabhimoju, R.; Sherman, W. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *J. Comput. Aided Mol. Des.* **2013**, *27*, 221−234.
(56) Humphrey, W.; Dalke, A.; Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph.* **1996**, *14*, 33−38.
(57) Amadei, A.; Ceruso, M. A.; Di Nola, A. On the convergence of the conformational coordinates basis set obtained by the essential dynamics analysis of proteins' molecular dynamics simulations. *Proteins: Struct., Funct., Genet.* **1999**, *36*, 419−424.
(58) Grant, B. J.; Rodrigues, A. P. C.; ElSawy, K. M.; McCammon, J. A.; Caves, L. S. D. Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics* **2006**, *22*, 2695−2696.
(59) Schymkowitz, J.; Borg, J.; Stricher, F.; Nys, R.; Rousseau, F.; Serrano, L. The FoldX web server: an online force field. *Nucleic Acids Res.* **2005**, *33*, W382−W388.
(60) Covaceuszach, S.; Bozzi, M.; Bigotti, M. G.; Sciandra, F.; Konarev, P. V.; Brancaccio, A.; Cassetta, A. Structural flexibility of human α-dystroglycan. *FEBS Open Bio* **2017**, *7*, 1064−1077.
(61) Willer, T.; Inamori, K.; Venzke, D.; Harvey, C.; Morgensen, G.; Hara, Y.; Beltrán Valero de Bernabé, D.; Yu, L.; Wright, K. M.; Campbell, K. P. The glucuronyltransferase B4GAT1 is required for initiation of LARGE-mediated α-dystroglycan functional glycosylation. *eLife* **2014**, *3*, No. e03941.
(62) Cuff, M. E.; Tesar, C.; Bearden, J.; Joachimiak, A. *The Structure of a Protein in Glycosyl Transferase Family 8 from Anaerococcus Prevotii*, 2011.
(63) Persson, K.; Ly, H. D.; Dieckelmann, M.; Wakarchuk, W. W.; Withers, S. G.; Strynadka, N. C. J. Crystal structure of the retaining galactosyltransferase LgtC from Neisseria meningitidis in complex with donor and acceptor sugar analogs. *Nat. Struct. Biol.* **2001**, *8*, 166−175.
(64) Yu, H.; Takeuchi, M.; LeBarron, J.; Kantharia, J.; London, E.; Bakker, H.; Haltiwanger, R. S.; Li, H.; Takeuchi, H. Notch-modifying xylosyltransferase structures support an SNi-like retaining mechanism. *Nat. Chem. Biol.* **2015**, *11*, 847−854.
(65) Lairson, L. L.; Chiu, C. P. C.; Ly, H. D.; He, S.; Wakarchuk, W. W.; Strynadka, N. C. J.; Withers, S. G. Intermediate trapping on a mutant retaining α-galactosyltransferase identifies an unexpected aspartate residue. *J. Biol. Chem.* **2004**, *279*, 28339−28344.
(66) Ünligil, U. M.; Rini, J. M. Glycosyltransferase structure and mechanism. *Curr. Opin. Struct. Biol.* **2000**, *10*, 510−517.
(67) Zhang, C.; Freddolino, P. L.; Zhang, Y. COFACTOR: improved protein function prediction by combining structure, sequence and protein−protein interaction information. *Nucleic Acids Res.* **2017**, *45*, W291−W299.
(68) Xu, J.; Zhang, Y. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* **2010**, *26*, 889−895.
(69) Chang, A.; Singh, S.; Phillips, G. N.; Thorson, J. S. Glycosyltransferase structural biology and its role in the design of catalysts for glycosylation. *Curr. Opin. Biotechnol.* **2011**, *22*, 800−808.
(70) Fritz, T. A.; Hurley, J. H.; Trinh, L.-B.; Shiloach, J.; Tabak, L. A. The beginnings of mucin biosynthesis: The crystal structure of UDP-GalNAc:polypeptide -N-acetylgalactosaminyltransferase-T1. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 15307−15312.
(71) Ji, S.; Samara, N. L.; Revoredo, L.; Zhang, L.; Tran, D. T.; Muirhead, K.; Tabak, L. A.; Ten Hagen, K. G. A molecular switch orchestrates enzyme specificity and secretory granule morphology. *Nat. Commun.* **2018**, *9*, 3508.
(72) de Las Rivas, M.; Paul Daniel, E. J.; Coelho, H.; Lira-Navarrete, E.; Raich, L.; Compañón, I.; Diniz, A.; Lagartera, L.; Jiménez-Barbero, J.; Clausen, H.; Rovira, C.; Marcelo, F.; Corzana, F.; Gerken, T. A.; Hurtado-Guerrero, R. Structural and mechanistic insights into the catalytic-domain-mediated short-range glycosylation preferences of GalNAc-T4. *ACS Cent. Sci.* **2018**, *4*, 1274−1290.
(73) Yu, C.; Liang, L.; Yin, Y. Structural basis of carbohydrate transfer activity of UDP-GalNAc: Polypeptide N-acetylgalactosaminyltransferase 7. *Biochem. Biophys. Res. Commun.* **2019**, *510*, 266−271.
(74) Kubota, T.; Shiba, T.; Sugioka, S.; Furukawa, S.; Sawaki, H.; Kato, R.; Wakatsuki, S.; Narimatsu, H. Structural basis of carbohydrate transfer activity by human UDP-GalNAc: polypeptide alpha-N-acetylgalactosaminyltransferase (pp-GalNAc-T10). *J. Mol. Biol.* **2006**, *359*, 708−727.
(75) Osawa, T.; Sugiura, N.; Shimada, H.; Hirooka, R.; Tsuji, A.; Shirakawa, T.; Fukuyama, K.; Kimura, M.; Kimata, K.; Kakuta, Y. Crystal structure of chondroitin polymerase from Escherichia coli K4. *Biochem. Biophys. Res. Commun.* **2009**, *378*, 10−14.
(76) Fritz, T. A.; Raman, J.; Tabak, L. A. Dynamic association between the catalytic and lectin domains of human UDP-GalNAc:Polypeptide α- N -Acetylgalactosaminyltransferase-2. *J. Biol. Chem.* **2006**, *281*, 8613−8619.
(77) Lira-Navarrete, E.; de las Rivas, M.; Compañón, I.; Pallarés, M. C.; Kong, Y.; Iglesias-Fernández, J.; Bernardes, G. J. L.; Peregrina, J. M.; Rovira, C.; Bernadó, P.; Bruscolini, P.; Clausen, H.; Lostao, A.; Corzana, F.; Hurtado-Guerrero, R. Dynamic interplay between catalytic and lectin domains of GalNAc-transferases modulates protein O-glycosylation. *Nat. Commun.* **2015**, *6*, 6937.
(78) Lira-Navarrete, E.; Iglesias-Fernández, J.; Zandberg, W. F.; Compañón, I.; Kong, Y.; Corzana, F.; Pinto, B. M.; Clausen, H.; Peregrina, J. M.; Vocadlo, D. J.; Rovira, C.; Hurtado-Guerrero, R. Substrate-guided front-face reaction revealed by combined structural snapshots and metadynamics for the Polypeptide N -Acetylgalactosaminyltransferase 2. *Angew. Chem., Int. Ed.* **2014**, *53*, 8206−8210.
(79) Inamori, K.-i.; Willer, T.; Hara, Y.; Venzke, D.; Anderson, M. E.; Clarke, N. F.; Guicheney, P.; Bönnemann, C. G.; Moore, S. A.; Campbell, K. P. Endogenous glucuronyltransferase activity of LARGE or LARGE2 required for functional modification of α-dystroglycan in cells and tissues. *J. Biol. Chem.* **2014**, *289*, 28138−28148.
(80) Bromberg, Y.; Rost, B. Correlating protein function and stability through the analysis of single amino acid substitutions. *BMC Bioinf.* **2009**, *10*, S8.
(81) Bozic, D.; Sciandra, F.; Lamba, D.; Brancaccio, A. The structure of the N-terminal region of murine skeletal muscle α-dystroglycan discloses a modular architecture. *J. Biol. Chem.* **2004**, *279*, 44812−44816.
(82) Inamori, K.-i.; Yoshida-Moriguchi, T.; Hara, Y.; Anderson, M. E.; Yu, L.; Campbell, K. P. Dystroglycan function requires xylosyl- and glucuronyltransferase activities of LARGE. *Science* **2012**, *335*, 93−96.

https://dx.doi.org/10.1021/acs.jcim.0c00281
J. Chem. Inf. Model. 2020, 60, 3145−3156