Supplementary Data

Glycan Reader is Improved to Recognize Most Sugar Types and Chemical Modifications in the Protein Data Bank

Sang-Jun Park, Jumin Lee, Dhilon S. Patel, Hongjing Ma, Hui Sun Lee, Sunhwan Jo, and Wonpil Im

Table S1. Monosaccharides recognized by *Glycan Reader*.

	Abbreviation	Full Name	CHARMM Residue Name[†]
Hex	Glc	D/L-Glucose	GLC
	Man	D/L-Mannose	MAN
	Gal	D/L-Galactose	GAL, GALF
	Gul	D/L-Gulose	GUL
	Alt	D/L-Altrose	ALT
	All	D/L-Allose	ALL
	Tal	D/L-Talose	TAL
	Ido	D/L-Idose	IDO
HexN	GlcN	D-Glucosamine	GLCN
	ManN	D-Mannosamine	MANN
	GalN	D-Galactosamine	GALN
	GulN	D-Gulosamine	GULN
	AltN	D-Altrosamine	ALTN
	AllN	D-Allosamine	ALLN
	TalN	D-Talosamine	TALN
	IdoN	D-Idosamine	IDON
HexNAc	GlcNAc	*N* -acetyl- D/L-Glucosamine	GLCNA
	ManNAc	*N*-acetyl- D/L-Mannosamine	MANNA
	GalNAc	*N* -acetyl- D/L-Galactosamine	GALNA
	GulNAc	*N*-acetyl- D/L-Gulosamine	GULNA
	AltNAc	*N* -acetyl- D/L-Altrosamine	ALTNA
	AllNAc	*N*-acetyl- D-Allosamine	ALLNA
	TalNAc	*N* -acetyl- D/L-Talosamine	TALNA
	IdoNAc	*N*-acetyl-L-Idosamine	IDONA
HexA	GlcA	D-Glucuronic acid	GLCA
	ManA	D-Mannuronic acid	MANA
	GalA	D-Galacturonic acid	GALA
	GulA	D/L-Guluronic acid	GULA
	AltA	D/L-Altruronic acid	ALTA
	AllA	D-Alluronic acid	ALLA
	TalA	D-Taluronic acid	TALA
	IdoA	L-Iduronic acid	IDOA
Deoxy	Qui	D-Quinovose	QUI
	Rha	D/L-Rhamnose	RHM
	Fuc	D/L-Fucose	FUC
	QuiNAc	*N*-acetyl-D-Quinovose	QUINA
	RhaNAc	*N*-acetyl-L-Rhamnose	RHMNA
	FucNAc	*N*-acetyl-L-Fucose	FUCNA
	6dAlt	L-6-deoxy-Altrose	L6ALT

	6dTal	L-6-deoxy-Talose	L6TAL
	Oli	D-Olivose	OLI
	Tyv	D-Tyvelose	TYV
	Abe	D-Abequose	BEQ
	Par	D-Paratose	PAR
	Dig	D-Digitoxose	DIG
	Col	L-Colitose	COL
Pentose	Ara	D/L-Arabinose	ARB, ARBP
	Lyx	D-Lyxose	LYF, LYXP
	Xyl	D/L-Xylose	XYF, XYL, XULF
	Rib	D/L-Ribose	RIB, RIBP, DEO
	Fru	D/L-Fructose	FRU, FRUP
	Tag	D-Tagatose	TAG, TAGP
	Sor	L-Sorbose	SOR, SORP
	Psi	D/L-Psicose	PSI, PSIP
Acidic	Neu5Ac	*N*-acetyl-D-Neuraminic acid	NE5AC
	Neu5Gc	*N*-glycolyl-D-Neuraminic acid	NE5GC
	Neu	D-Neuraminic acid	NEU
	Kdn	Deaminoneuraminic acid	KDN
	Kdo	2-keto-3-deoxy-D-mannooctanoic acid	KDO, KDA
	Dha	3-deoxy-D-lyxo-heptopyran-2-ularic acid	DHA
Etc	LDManHep	L-glycero-D-manno-heptose	HEP
	DDManHep	D-glycero-D-manno-heptose	DHEP
	MurNAc	*N*-acetyl-D-Muramic acid	MU2AC
	MurNGc	*N*-glycolyl-D-Muramic acid	MU2GC
	Mur	D-Muramic acid	MUR
	Bac	D-bacillosamine	BAC

†The α or β anomeric configuration is designated by adding “A” or “B” in front of each CHARMM residue name.

Table S2. Chemical modifications recognized by *Glycan Reader*.

Modification	CHARMM Patch Name
Phosphate	PH, PHF[$]
Pyrophosphate	PPH, PPHF[$]
Sulfate	SUF, SUFF[$]
Sulfonate (Qui)	SFO
Sulfo-amine	NSF
Thiol	SH
Isopropyl thiol	SPRO
Amine	AM
O-methyl	OME, COOME[%]
O-acetyl	OAC
N-acetyl	NAC
Carboxyl	COO
Fluorine	CF, CF2 (for Neu5Ac)
Ethanolamine phosphate	PEA
Ethanolamine diphosphate	ETPP
Glycerol	MGLY
Glyceric acid	GLYA
Glycine amide	NGF
Acyl chain[@]	C8, C10, C12[†]
S-linked acyl chain[@]	SC8, SC10, SC12[†]
Ceramide[@]	CERA, CERB[‡]
Diacyl glycerol[@]	DAGA, DAGB[‡]
Phosphatidyl inositol diacyl glycerol[@]	DPINS, INS2A, INS2B, INS6A, INS6B[#]
2,3-diphytanyl glycerol[@]	DPHGA, DPHGB[†]
Deoxylation	DEO

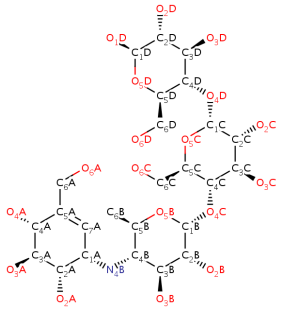
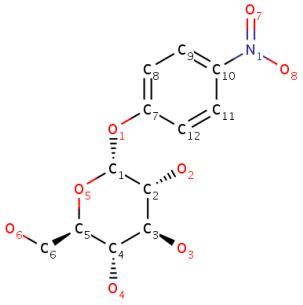
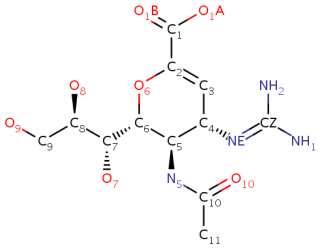
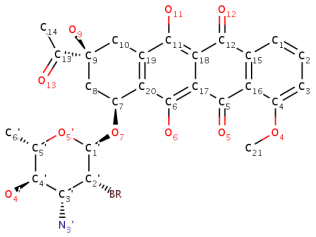
[$]PHF, PPHF, and SUFF are for furanose.
[%]COOME is for uronic acid.
[†]The numbers 8, 10, and 12 represent the number of carbons in the acyl chain.
[‡]"A" or "B" at the end of each patch name represents the α or β anomeric configuration of carbohydrates attached to ceramide.
[#]DPINS is the patch between inositol and diacylglycerol; 2 or 6 is the inositol ring carbon position to which carbohydrates are attached; [‡]"A" or "B" represents the α or β anomeric configuration of carbohydrates attached to inositol.
[@]These 6 templates are used for glycolipid detection; among a total of 25 chemical modifications, there are 19 function groups and 6 glycolipids

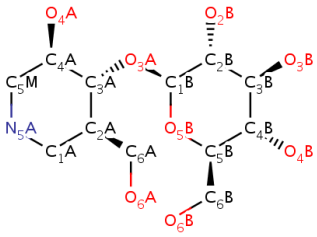
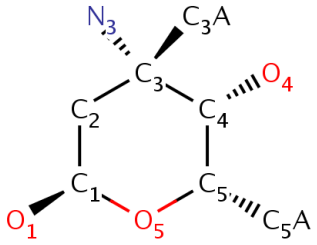
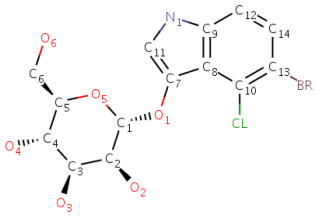
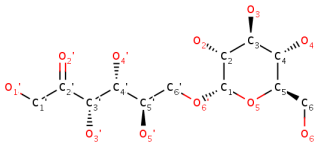
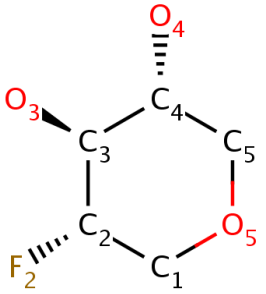
Table S3. Glycosidic linkages recognized by *Glycan Reader*.

Linkage[†]	CHARMM Patch Name
α, β-Pyr-(1→1)-α, β–Pyr	11AA, 11AB, 11BB, 1S1BB
α, β-Pyr-(1→2)-α, β-Pyr	12AA, 12AB, 12BB, 1S2AA, 1S2AB, 1S2BA, 1S2BB
α, β-Pyr-(1→3)-α, β-Pyr	13AA, 13AB, 13BA, 13BB, 1S3AA, 1S3AB, 1S3BA, 1S3BB
α, β-Pyr-(1→4)-α, β-Pyr	14AA, 14AB, 14BA, 14BB, 1S4AA, 1S4AB, 1S4BA, 1S4BB
α, β-Pyr-(1→6)-α, β-Pyr	16AB, 16BB, 1S6AT, 1S6BT
α, β-Pyr-(1→6)-β-Hep	MH16AB, MH16BB
α, β-Pyr-(1→7)-α-Hep	MH17AT, MH17BT
α, β-Pyr-(1→5)-α-Ulo	SA15AA, SA15BA
α-Ulo-(2→4)-α-Ulo	SA24E
α-Ulo-(2→6)-α-Pyr	SA26T
α-Ulo-(2→8)-α-Ulo	SA28T
α-Sia-(2→3)-α, β-Pyr	SA23A
α-Sia-(2→6)-α, β-Pyr	SA26T
α-Sia-(2→8)-α-Sia	SA28E
α-Sia-(2→9)-α-Sia	SA29E
α-Pyr-(1→2)- β-UloF	FP21BA
β-UloF-(2→1)- β-UloF	FF21BB
β-UloF-(2→6)- β-UloF	FF26BB
α-Fur-(1→2)- β-Pyr	FP12AB
α-Fur-(1→3)- β-Pyr	FP13AB
β -Fur-(1→2)-α-Fur	FF12BA
α -Fur-(1→3)-α-Fur	FF13AA
α-Fur-(1→5)-α-Fur	FF15AT

[†]Pyr is for pyranose; Hep for heptopyranose; Ulo for octulosonic acid; Sia for sialic acid; UloF for hex-2-ulofuranose; Fur for pentofuranose.

Table S4. Undefined chemical modifications that appear in more than 10 PDB entries.

# PDB entries	# molecule	Description	Structure / Representative PDB ID
50	97	Acarbose	 PDB ID: 4BQF
47	117	Nitrophenyl-D-glucoside	 PDBID: 1VAL
22	48	Zanamivir	 PDBID: 4CPN
22	28	Daunorubicin	 PDBID: 381D

19	29	Cellobiose-like Isofagomine		PDBID: 3RX8
17	55	Vancomycin		PDBID: 1PNV
14	25	(5-bromo-4-chloro-3-indolyl)-D-mannose		PDBID: 3AX4
13	16	6-O-alpha-D-glucopyranosyl-D-fructose		PDBID: 4HA1
11	15	1,2-deoxy-2-fluoro-xylopyranose		PDBID: 1C5I

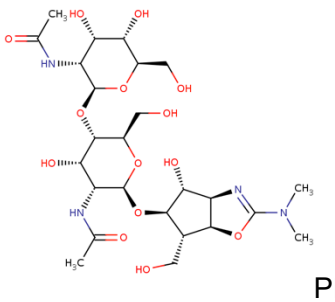
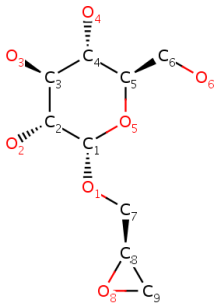
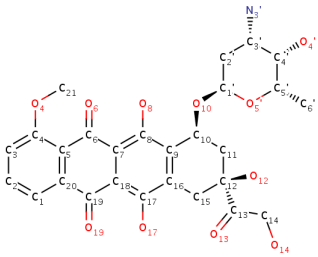
11	19	Allosamidin		PDBID: 1X6N
11	16	2,3- epoxypropyl-alpha-D-glucopyranoside		PDBID: 1J11
10	12	Doxorubicin		PDBID: 4DX7

Table S5. *Glycan Reader*-recognized structural errors among 10,731 PDB glycan chains.

Error type	Number of PDB entries	PDB Example
Missing Oxygen (at anomeric carbon)	285	4MYV
Missing Glycosidic Linkages	339	2Y1K
Surplus Glycosidic Linkages	55	1MQL
Bond Distance > 2.0 Å	85	1QFG
Overlapping Atoms	8	4KQM
	624 (5.81%)	

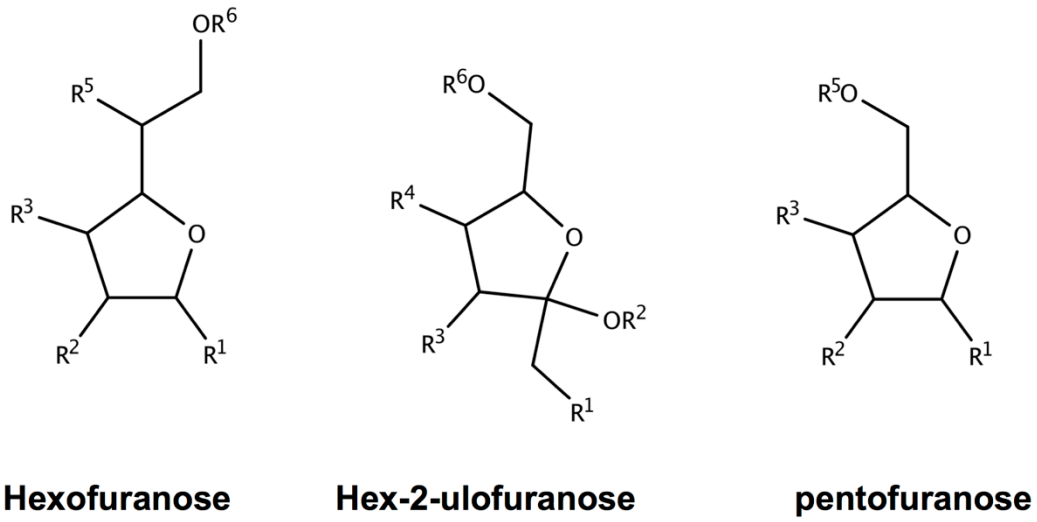


Figure S1. The template graphs of five-membered rings to identify the backbone atoms through VF2 isomorphism. R is the available position of chemical modifications.

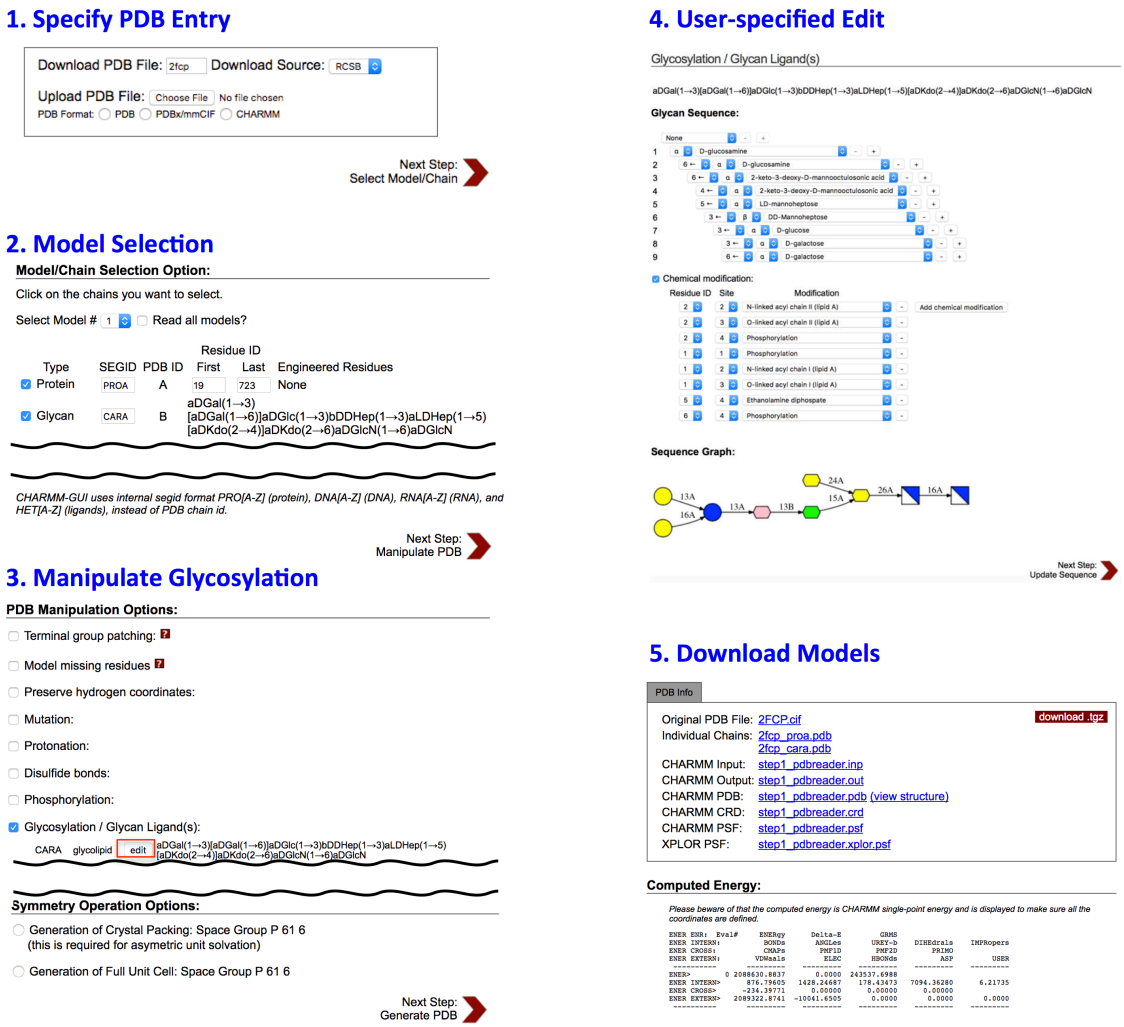


Figure S2. Step-by-step illustration of *Glycan Reader* in CHARMM-GUI: (1) One can upload a PDB file or specify a PDB entry. PDB ID: 2FCP is used as an example. (2) All protein and glycan chains are displayed and selected as a default. (3) Selected glycans are displayed in the PDB Manipulation page, and the edit button (red box) is provided for the glycan sequence manipulation. (4) Upon user's click on the edit button, *Glycosylation / Glycan Ligand(s) Manipulator* is displayed in a new pop-up window, and one can edit the glycosylation and glycan sequence. (5) Finally, one can download structure (PDB and CRD) and data (PSF) files. In the case that a user wants to build a solvated or membrane system containing glycans, the user can use *Quick MD Simulator* or *Membrane Builder* and follow the same PDB reading steps. At the end of these modules, one can obtain the equilibration and production input files for CHARMM, NAMD, GROMACS, AMBER, GENESIS, LAMMPS, Desmond, OpenMM, and CHARMM/ OpenMM.

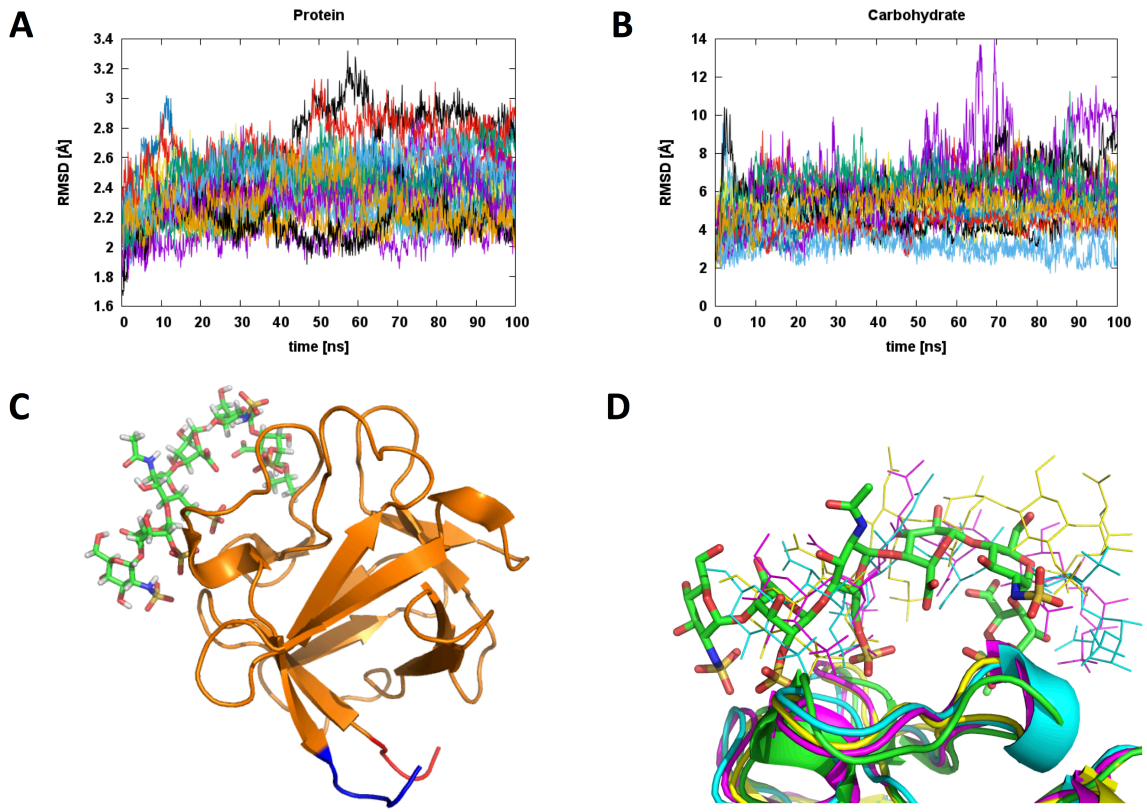


Figure S3. (A and B) RMSD time-series of FGF-1 monomer and heparin analogue. The FGF-1 RMSDs were measured for residue id 26 to 149, excluding (C) the N-terminal (22-25) and C-terminal (150-154) loops due to the high flexibility (gray). Note that the heparin RMSDs were measured after FGF-1 was first superposed to each initial NMR structure. (D) Heparin analogue structures obtained from 10-ns (cyan), 20-ns (magenta), and 50-ns (yellow) snapshots. Green is the initial heparin structure.

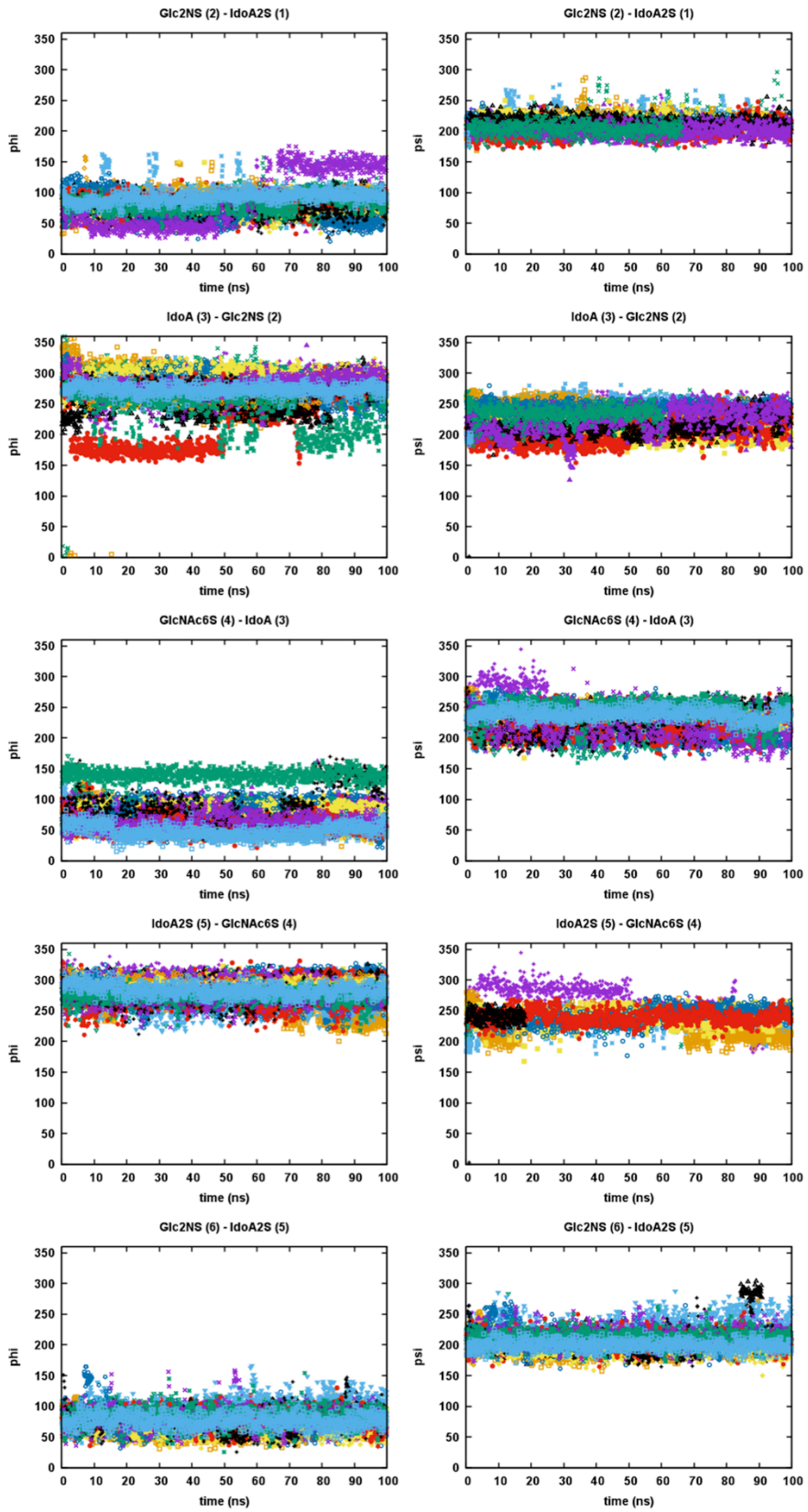


Figure S4. Time-series of each ϕ and ψ glycosidic dihedral angle.

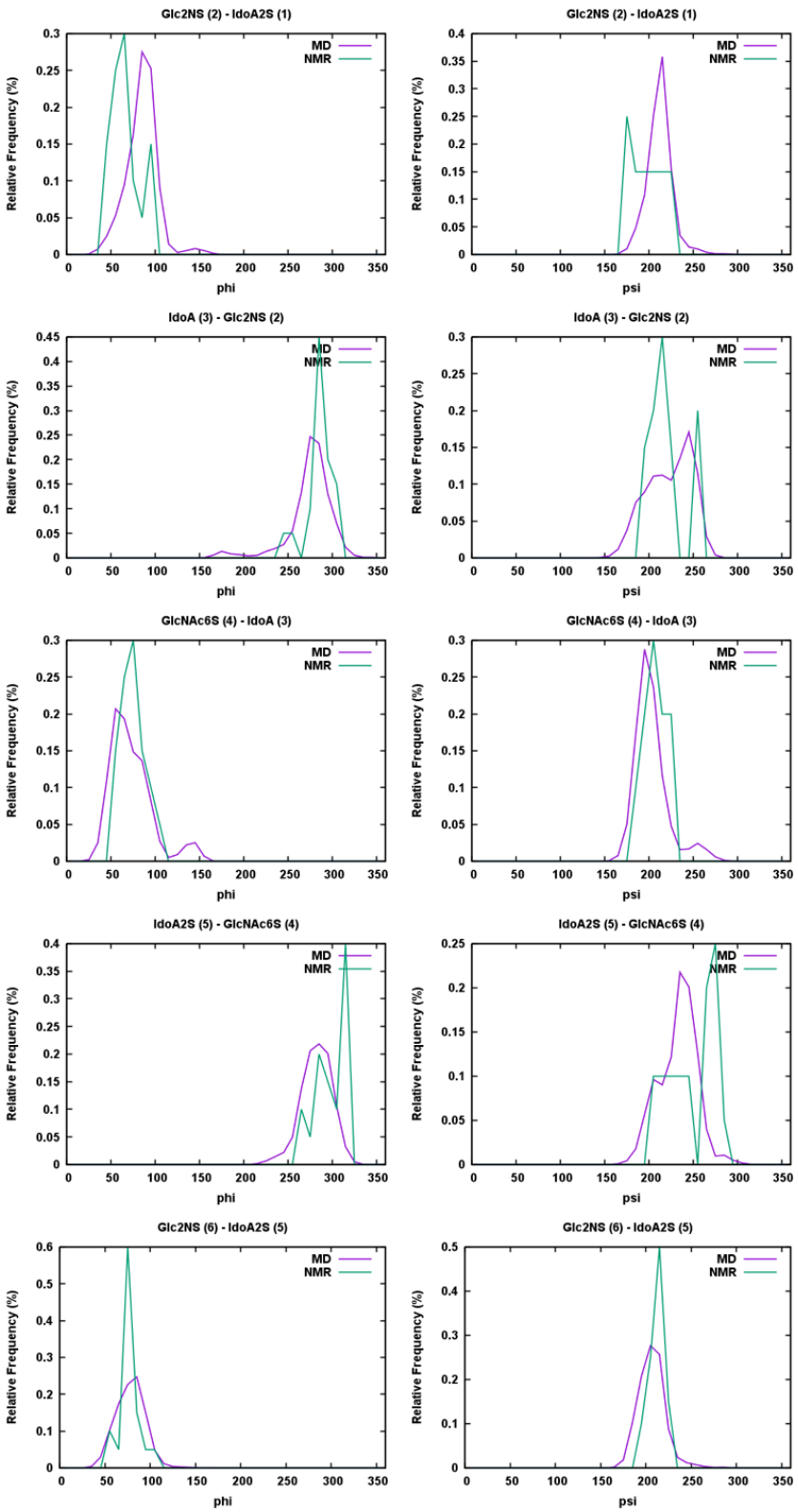


Figure S5. Frequency distribution curves of φ and ψ angles obtained from the MD simulations in comparison with those calculated from the 20 NMR models.