

ProtGlycanDock: Towards Accurate Protein-Glycan Docking with Tailored Dataset, Benchmark and Model

Minghao Xu



Glycans are ubiquitous

Starch



Bread



Rice

Sugars



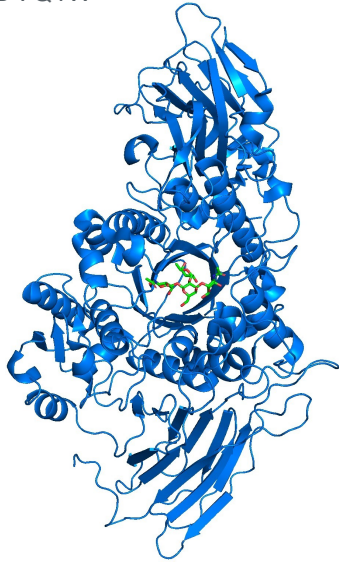
Orange
(fructose)



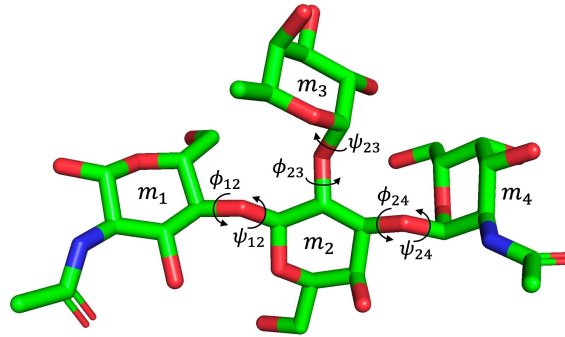
Milk
(lactose)

Protein-glycan interactions determine many biological functions

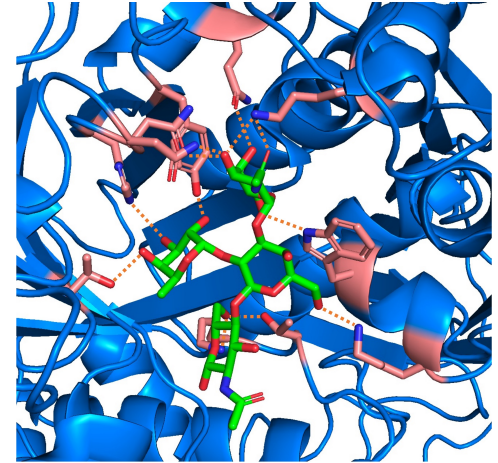
PDB 7Q1W



Protein-glycan complex



Flexible glycosidic bonds



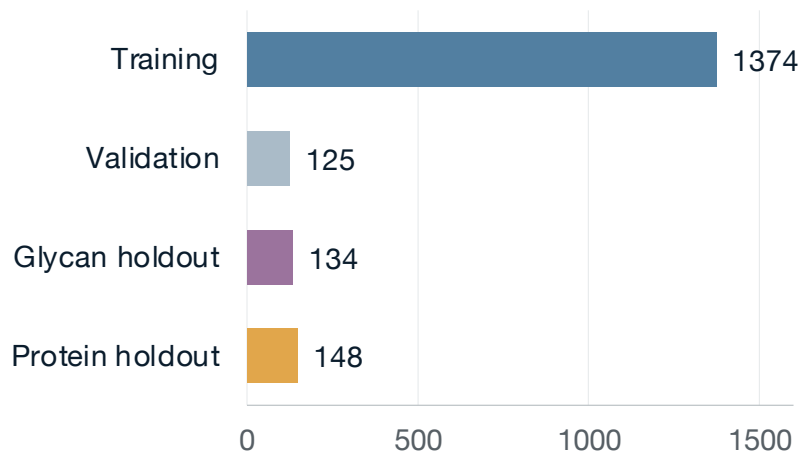
Hydrogen-bond network

Metabolism and immune recognition depend on precise binding

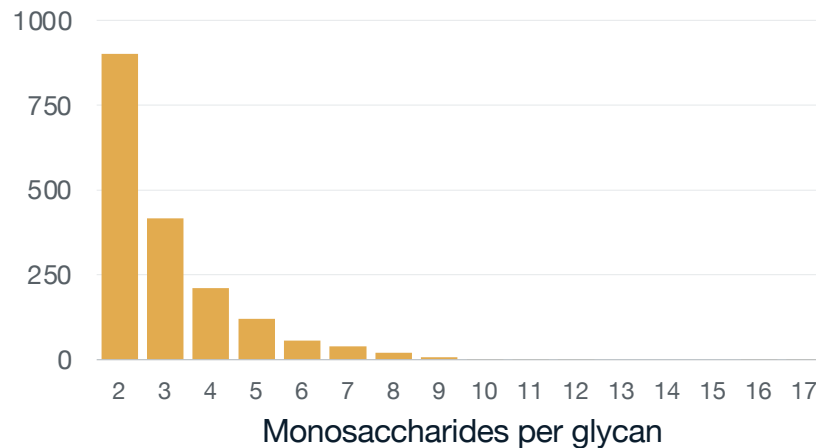
ProtGlycanDock benchmark

1,781 noncovalent complexes and 332 distinct glycans

Training and generalization tests



Glycan size distribution

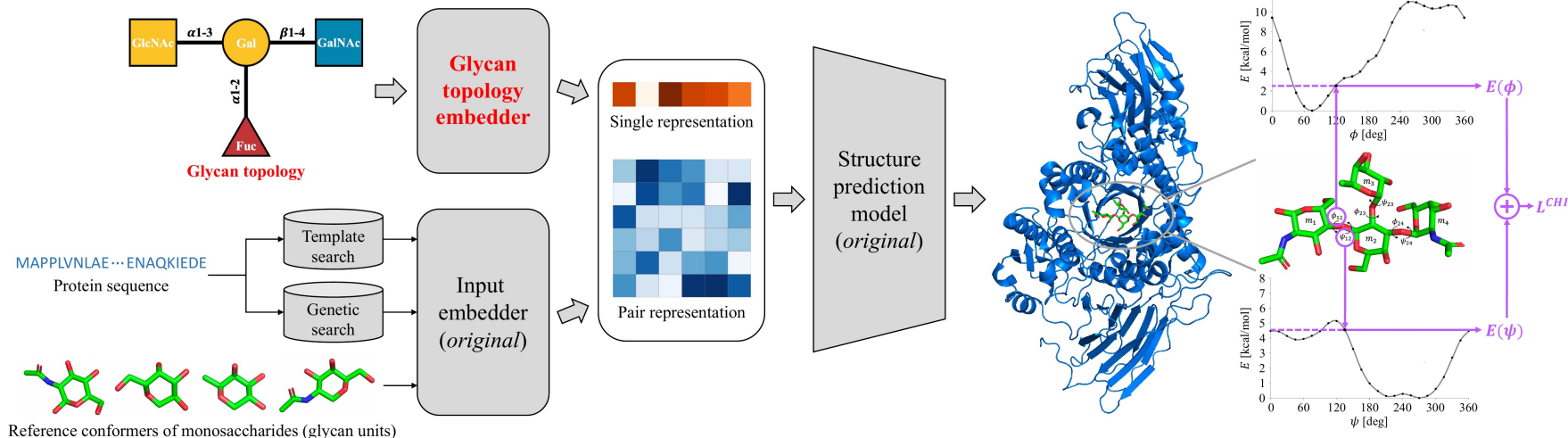


10 baseline models evaluated under blind docking

No ground-truth receptor structure or binding site provided

Glycan-specific model finetuning

Finetuning Protenix and Protenix-constraint



Topology embedding

Monosaccharide and linkage types

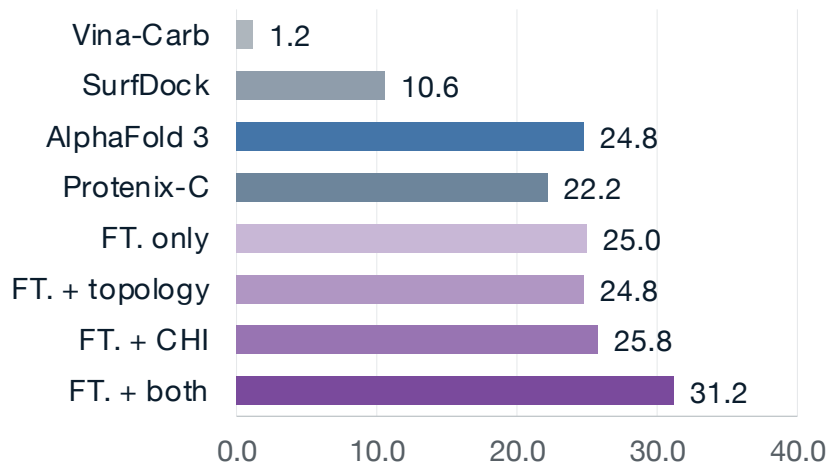
CHI energy loss

Physically plausible glycosidic torsions

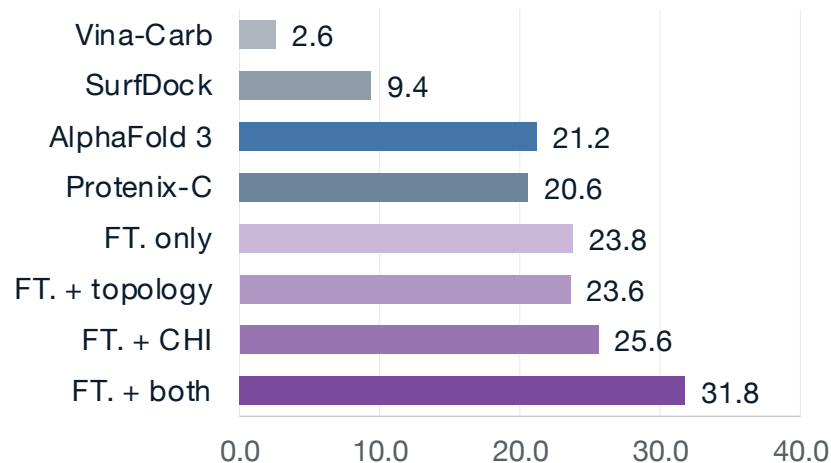
Better docking across both holdouts

Ligand RMSD < 2 Å and PoseBusters-valid (sample count)

Protein holdout (n = 148)



Glycan holdout (n = 134)

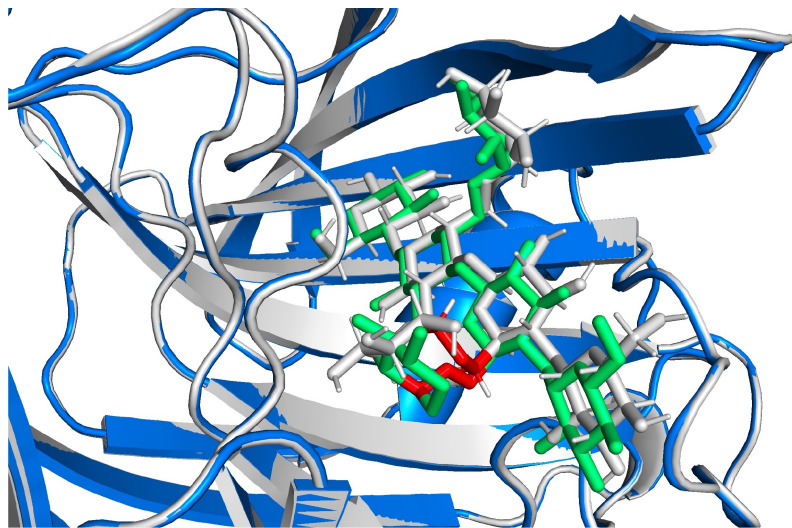


Joint topology + CHI finetuning leads on 11 of 12 metrics

Table 1: mean counts over 5 seeds. All FT. variants use Protenix-constraint (Protenix-C).

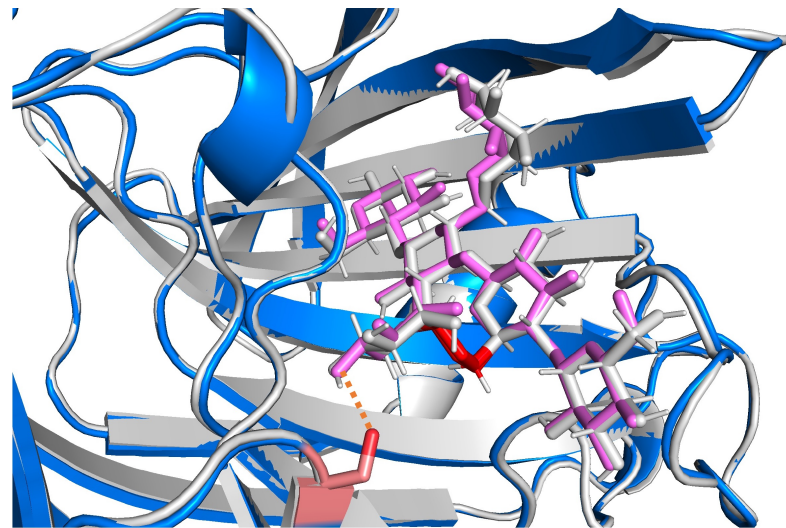
More accurate glycan binding geometry

AlphaFold 3



Ligand RMSD: 4.348 Å

Finetuned Protenix-constraint



Ligand RMSD: 0.699 Å

PDB 2JEN. Gray: ground truth. Red: key glycosidic bond. Orange dashed line: hydrogen bond.

Remaining errors: incorrect binding site or binding depth

Thank you!



Dataset



Code