

Supplementary Material

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30

Dataset

The dataset is composed of the following monomers: glucose (Glc), glucosamine (GlcNAc), glucuronic acid (GlcA), fucose (Fuc), mannose (Man), mannosamine (ManNAc), galactose (Gal), galactosamine (GalNAc), galacturonic acid (GalA), neuraminic acid/sialic acid (Neu/Sia), arabinose (Ara), xylose (Xyl), ribose, rhamnose (Rha), abequose (Abe), and fructose (Fru). There are either no training structures or very few for Fru, ManNAc, Abe, Rha, ribose, GalA, and GlcA. Although some have a high average test Dice similarity coefficient, CAPSIF may not accurately predict protein residues that bind those carbohydrate species well. Finally, CAPSIF:Voxel does not perform well on predicting residues that bind Neu and Fuc, likely due to their 9-carbon structure and (*L*) conformation, respectively, as well as GlcNAc.

The supporting Excel file **Supplementary File S1** includes the following information:

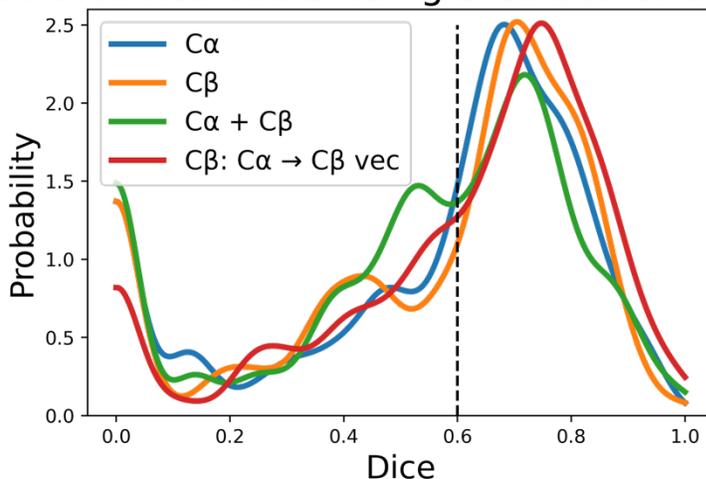
- PDB ID
- Carbohydrate species
- Per-PDB CAPSIF:V Dice coefficient
- Per-carbohydrate species Dice coefficient

Determination of Data Representation

For voxel locations, we compared three representation choices, (1) α carbon ($C\alpha$), (2) β carbon ($C\beta$), or (3) $C\alpha$ and $C\beta$ positions for the location of voxels. We trained and tested each of these models as described in the Methods. We compared the Dice coefficient, sensitivity and positive predictive value to determine which representation performs best (**Figure S1, Table S1**). The $C\beta$ -only representation has an average test Dice coefficient of 0.551, with the $C\alpha$ representation having a test Dice coefficient of 0.545, where when both the $C\alpha$ and $C\beta$ are included together in the representation, the architecture has an average test Dice coefficient of only 0.528.

Finally, we further included orientation information of the residues themselves by concatenating the unit vector of the $C\alpha$ to $C\beta$ bond to the $C\beta$ only representation. This representation had an average test metric of 0.597 ($C\beta$: $C\alpha \rightarrow C\beta$ vec) (**Figure S1, Table S1**). This method performed the best of all three representations, having the largest coverage and highest average test metrics. For these reasons, we chose $C\beta$: $C\alpha \rightarrow C\beta$ as our representation of coordinates and orientation for CAPSIF:V.

Predicted Residue Binding Dice Score: CAPSIF:V



31

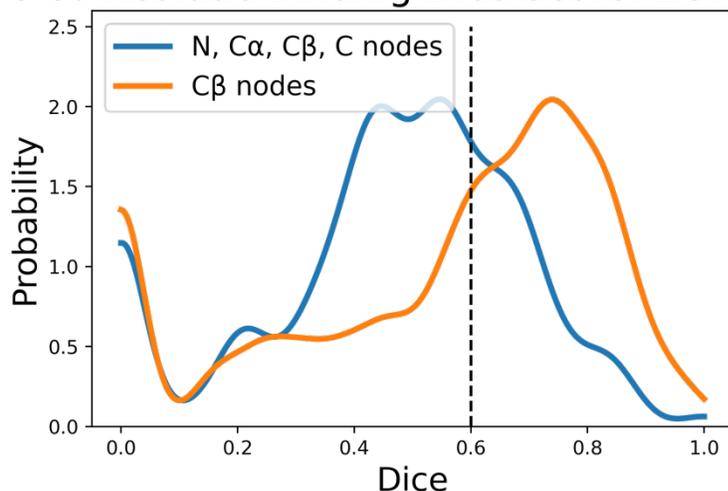
32 **Figure S1: Test Dice coefficient assessment for different representations with CAPSIF:V**
 33 **architectures:** Blue shows a $C\beta$ representation including a normalized vector for alpha carbon
 34 ($C\alpha$) to $C\beta$, orange shows only a $C\beta$ representation, green shows $C\alpha$ representation, and red shows
 35 $C\alpha$ and $C\beta$ representation with all voxels.

36 **Table S1: Performance for each CAPSIF:V model.** Dice coefficient is defined by (Eq 1); PPV
 37 and Sensitivity are same as Table 1.

Voxel Representation	Dice	PPV	Sensitivity
$C\beta$	0.551	0.563	0.583
$C\alpha$	0.545	0.535	0.620
$C\alpha + C\beta$	0.528	0.555	0.554
$C\beta: C\alpha \rightarrow C\beta$	0.597	0.598	0.647

38 Next, we investigated CAPSIF:G node representations, with the architecture described in Methods.
 39 We constructed the following variants: $C\beta$ nodes with φ and ψ angles, $C\beta$ and N, $C\alpha$, and C
 40 backbone nodes (and one-hot encoding for atom type, without φ and ψ angles). The $C\beta$ only node
 41 representation performed the best with a Dice coefficient of 0.543. Further, $C\beta$ takes a fraction of
 42 the time for predictions compared to the backbone due to graph construction time, therefore we
 43 chose the CAPSIF:G to be the $C\beta$ model (**Figure S12, Table S12**).

Predicted Residue Binding Dice Score: EGNN Models



44

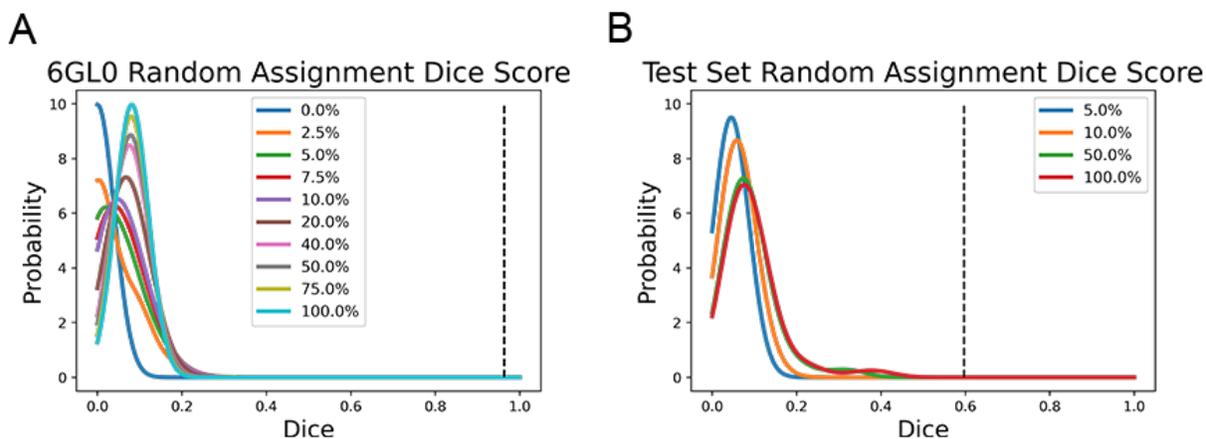
45 **Figure S2: Test Dice coefficient assessment for different representations with EGNN**
46 **architectures.** Blue shows all backbone atoms node representation, orange shows a Cβ node
47 representation.

48 **Table S2: Performance for EGNN model node representation.** Dice coefficient is defined by
49 (Eq 1); PPV and Sensitivity are same as Table 1.

EGNN Representation	Dice	PPV	Sensitivity
Cβ	0.543	0.541	0.590
Backbone	0.458	0.396	0.647

50 **Random Assignment of Carbohydrate Binding Regions**

51 As a control, we compared CAPSIF to a random baseline. For example, for 200 amino acids with
52 a 5.0% positivity rate, we randomly select 10 residues as a true label (sugar binding) and computed
53 the Dice similarity coefficient (Eq 1). Using 1,000 trials for an endoglucanase (6GL0), which has
54 331 total residues with 14 that experimentally bind carbohydrates, we observe a theoretical
55 maximum Dice coefficient at approximately 0.08 when all residues are predicted as carbohydrate
56 binders. At a rate of 5%, we observe a mean Dice coefficient of 0.046, where CAPSIF:V predicts
57 that protein with a Dice coefficient of 0.963 (Fig SI 3A).



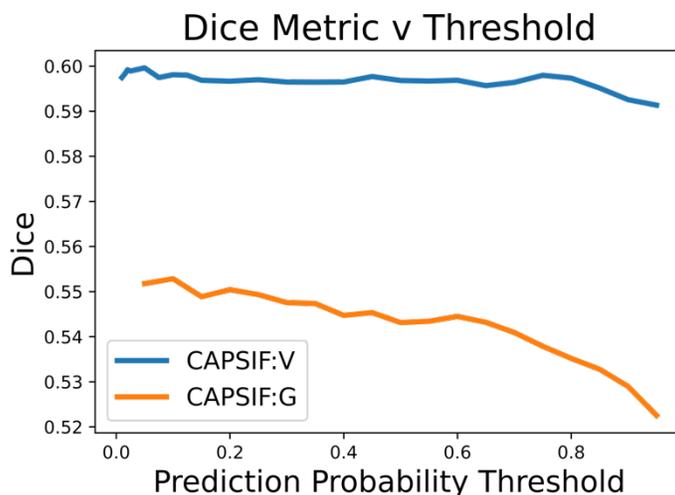
58

59 **Figure S3: Dice coefficient assessment with random assignment smoothed with a kernel**
 60 **density estimate with bandwidth $h = .04$. (A) Dice evaluation of random assignment of an**
 61 **endoglucanase (6GL0). (B) Dice evaluation over entire test set.**

62 The dataset has, on average, 5.16% of protein residues bind carbohydrates. With random
 63 assignment over the entire dataset, random assignment at 5.16% yields an average 0.046 Dice
 64 score, where CAPSIF:V outperforms random assignment by over 12-fold at an average 0.593 Dice
 65 (Fig SI4B).

66 **Determination of CAPSIF probability threshold**

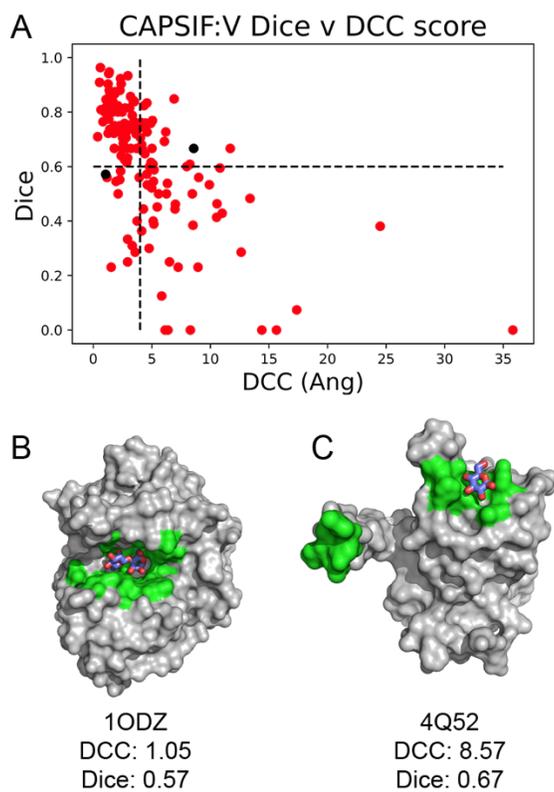
67 To determine the best probability cutoff value for the final activation function, we altered the
 68 threshold on the test dataset (Fig SI5). CAPSIF:V differs minimally for all thresholds while
 69 CAPSIF:G negatively correlates with increasing threshold and drops more sharply after a cutoff
 70 of 0.6. For both architectures we chose a threshold of 0.5.



71

72 **Figure S4: Test Dice coefficient assessment for CAPSIF architectures for various thresholds**
 73 **for the final sigmoid activation function. Blue represents CAPSIF:V, orange represents**
 74 **CAPSIF:G.**

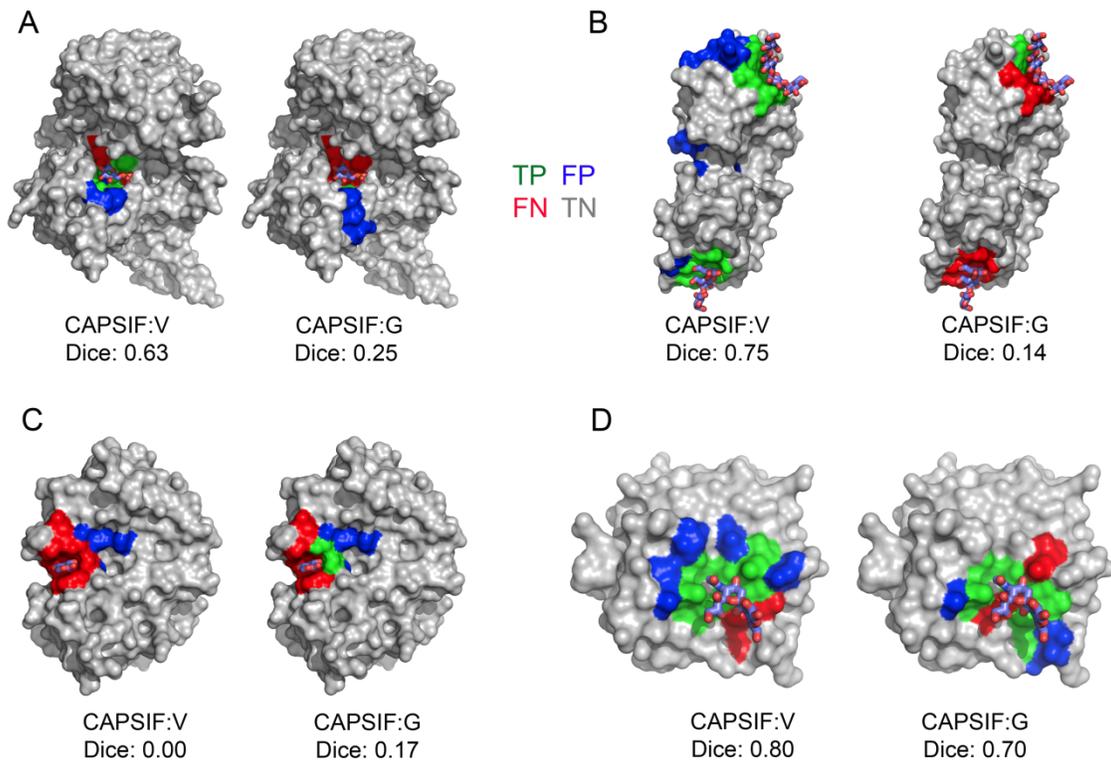
75 Comparison of Dice and DCC metrics



76

77 **Figure S5: Comparison of Dice score and DCC.** (A) Per-target comparison of Dice and DCC
78 for CAPSIF:V predictions on the test set. CAPSIF:V predictions (green) on (B) endo-1,4- β -
79 mannosidase 1ODZ and (C) *C. pinensis* DSM 2588 (4Q52) (gray).

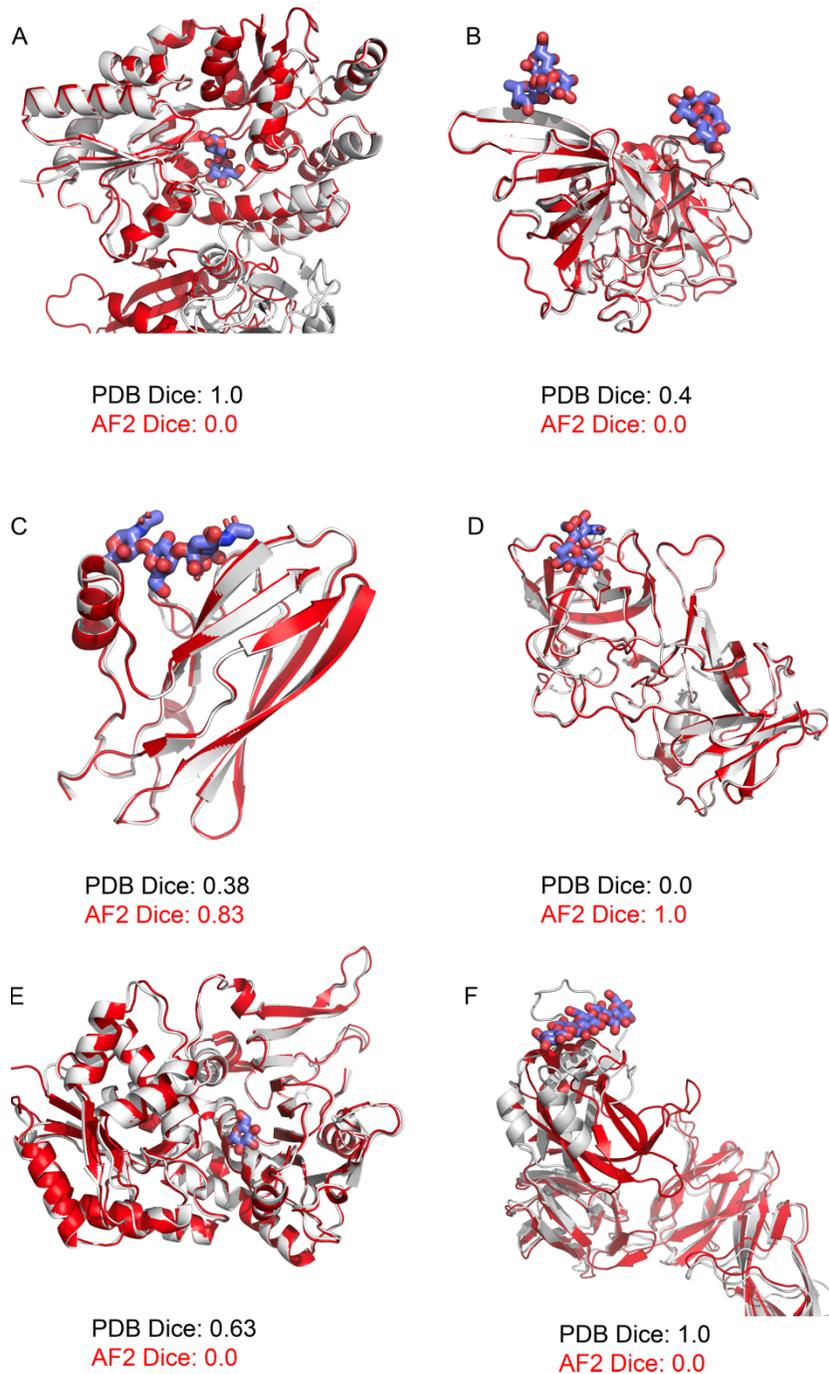
80 **Figures comparing CAPSIF:Voxel and CAPSIF:Graph predictions**



81

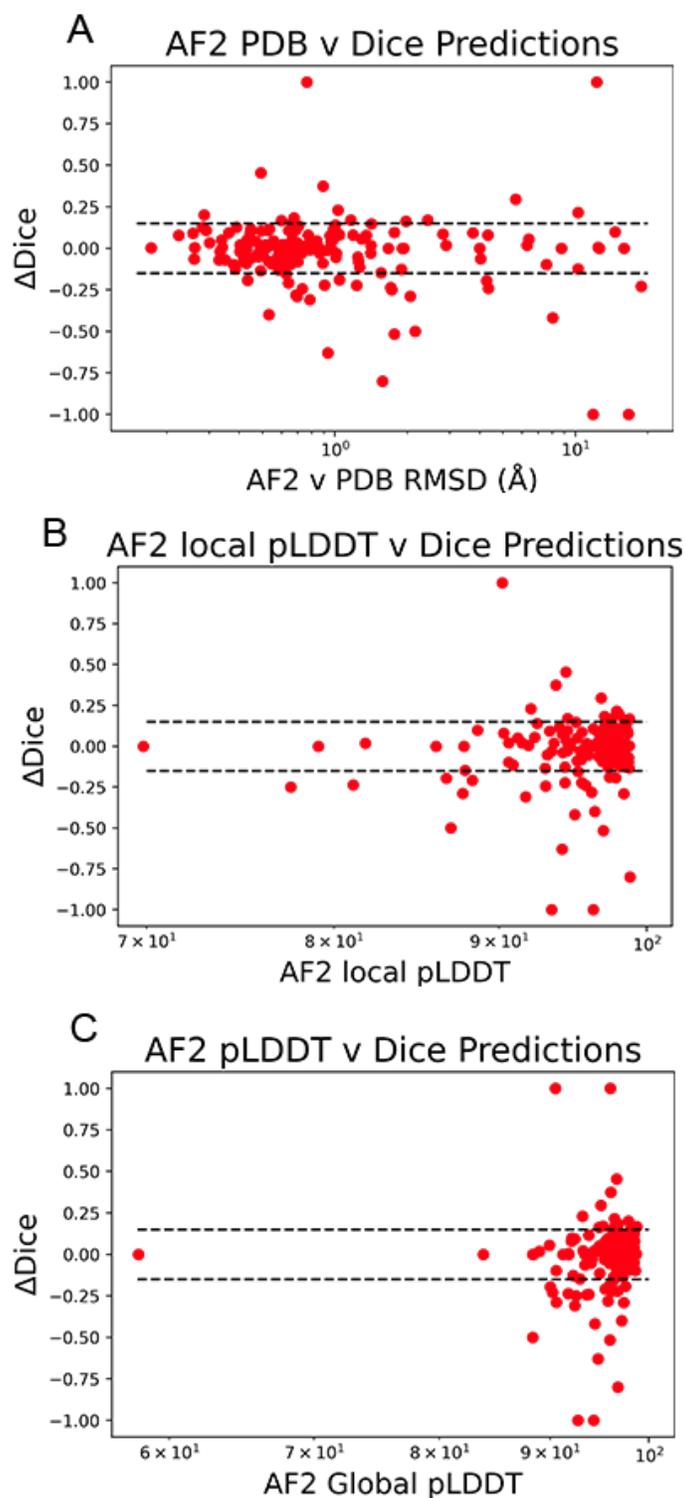
82 **Figure S6: Prediction of carbohydrate binding sites on a protein surface using CAPSIF:V**
83 **and CAPSIF:G. (A)** Glc 6-phosphate dehydrogenase (PDB:5UKW), **(B)** streptococcal virulence
84 factor (PDB:2J44), **(C)** MCR-1 catalytic domain (PDB:5ZJV), and **(D)** CBM40 (PDB:6ER3).
85 Residue labels - green: true positive, blue: false positive, red: false negative, gray: true negative,
86 cyan: bound carbohydrate; Dice coefficient is defined by eq (1) and DCC is distance from center
87 to center of the predicted binding regions.

88 **Comparison of RCSB and AF2 predicted structures**

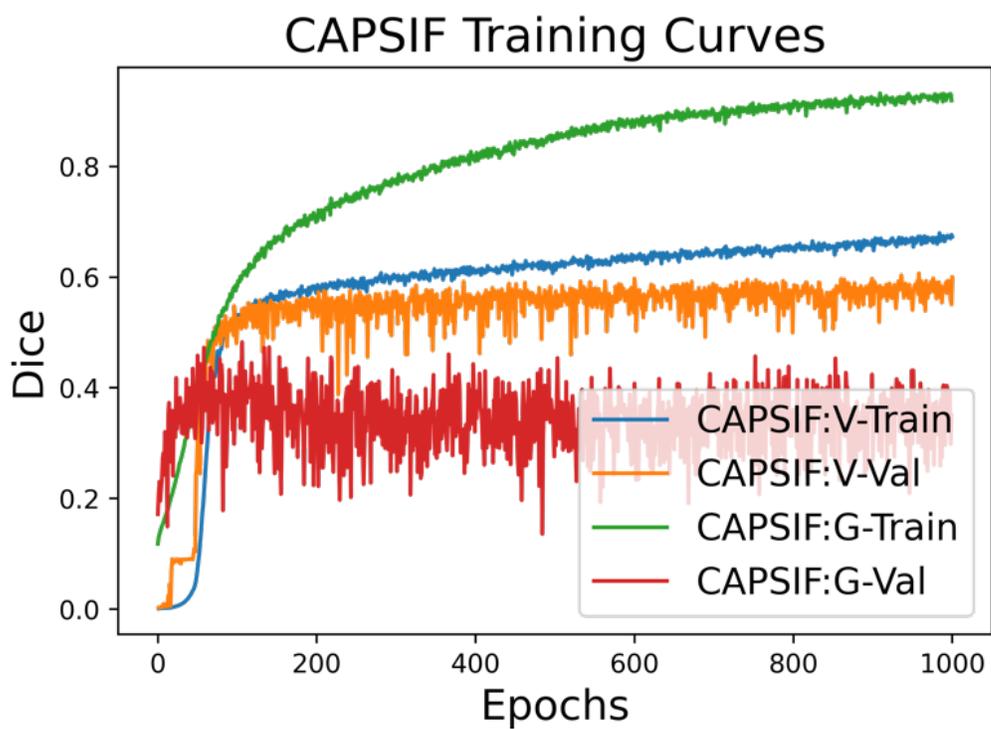


89

90 **Figure S7: AF2 structure prediction (red) of carbohydrate (purple) binding proteins**
91 **compared to experimentally solved structures (white); (A) SUFU (PDB:4BL8) (B) *E. coli***
92 **aminopeptidase N (PDB:4XO5), (C) GspB siglec domain (PDB:5IUC), (D) GII.13 novovirus**
93 **capsid P domain (PDB:5ZVC), (E) Glc 6-phosphate dehydrogenase (PDB:5UKW), and (F)**
94 **surface GBP B (PDB:6E57). Dice coefficient is defined by eq (1).**



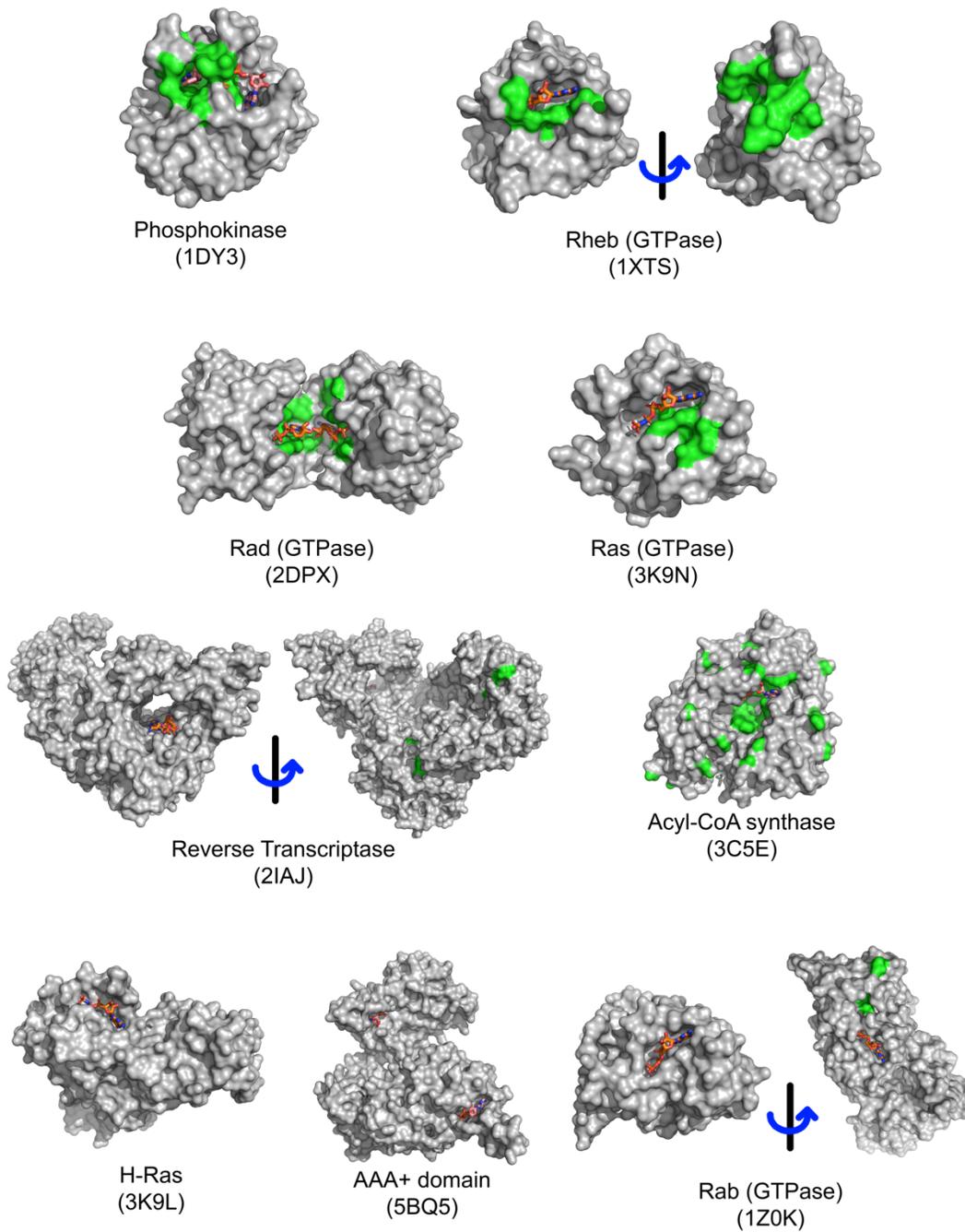
95
 96 **Figure S8: CAPSIF:V accuracy is not correlated with AF2 accuracy or confidence.**
 97 CAPSIF:V predictions on AF2 structure prediction metrics of carbohydrate binding proteins
 98 compared to RCSB structures. Change in Dice metric ($\Delta Dice = AF2\ Dice - RCSB\ Dice$) compared
 99 to (A) the total C α RMSD (log scale), (B) Local average pLDDT score of the carbohydrate binding
 100 region, and (C) total average pLDDT score of the entire structure.



101
102

Figure S9: Training and validation curves of both CAPSIF models.

CAPSIF predicted residues
NTP Ligand



103

104

105

106

Figure S10: Prediction of CAPSIF-V and CAPSIF-G on ATP and GTP-binding proteins. Both CAPSIF models predict similar regions on the ATP/GTP binding proteins, but only qualitatively capture the binding region of the phosphokinase, Acyl-CoA synthase, Rad, and Ras.