Electronic Supplementary Information for

The Energetic Landscape of CH-π Interactions in Protein-Carbohydrate Binding

Allison M. Keys[1,2,3], David W. Kastner[2,3,4], Laura L. Kiessling[3,5,6,*], and Heather J. Kulik[2,3,5]*
[1]*Computational and Systems Biology Program, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*
[2]*Department of Chemical Engineering, MIT, Cambridge, MA 02139, USA*
[3]*Department of Chemistry, MIT, Cambridge, MA, USA 02139, USA*
[4]*Department of Biological Engineering, MIT, Cambridge, MA, USA 02139, USA*
[5]*The Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA*
[6]*Koch Institute for Integrative Cancer Research, MIT, Cambridge, MA 02142, USA*
*co-corresponding authors: kiessling@mit.edu, hjkulik@mit.edu

Contents

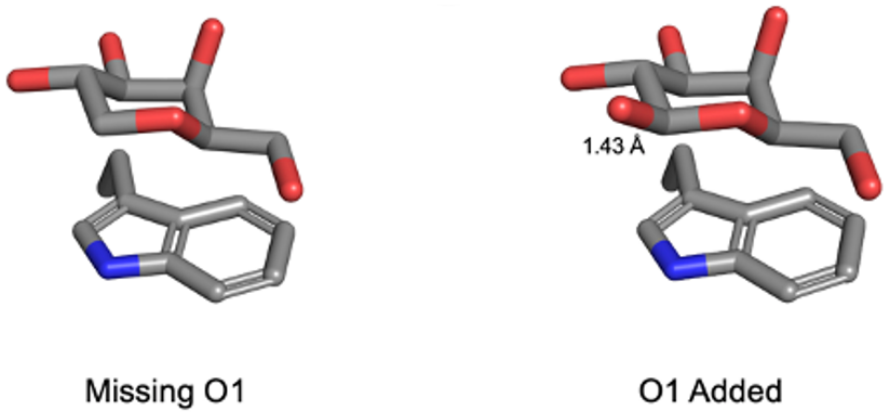

Figure S1. Example of adding a missing anomeric substituent to the C1 on the β-D-galactose ring. The missing atom is added at a bond distance of 1.43 Å. Atoms are colored as follows: carbon in gray, oxygen in red, and nitrogen in blue.

Table S1. Interaction totals for the full dataset of native close contacts grouped by amino acid interacting with β-D-galactose and by $EDIA_m$ scores evaluated for each molecular fragment (i.e. the amino acid or β-D-galactose). Close contacts are defined as aromatic amino acid– β-D-galactose pairs in which the centroids of the two species are within 7 Å of one another.

	All Close Contacts	Both $EDIA_m > 0.8$	At least one $EDIA_m < 0.8$
Tryptophan	524	351	173
Tyrosine	228	154	74
Phenylalanine	82	45	37
Total	834	550	284

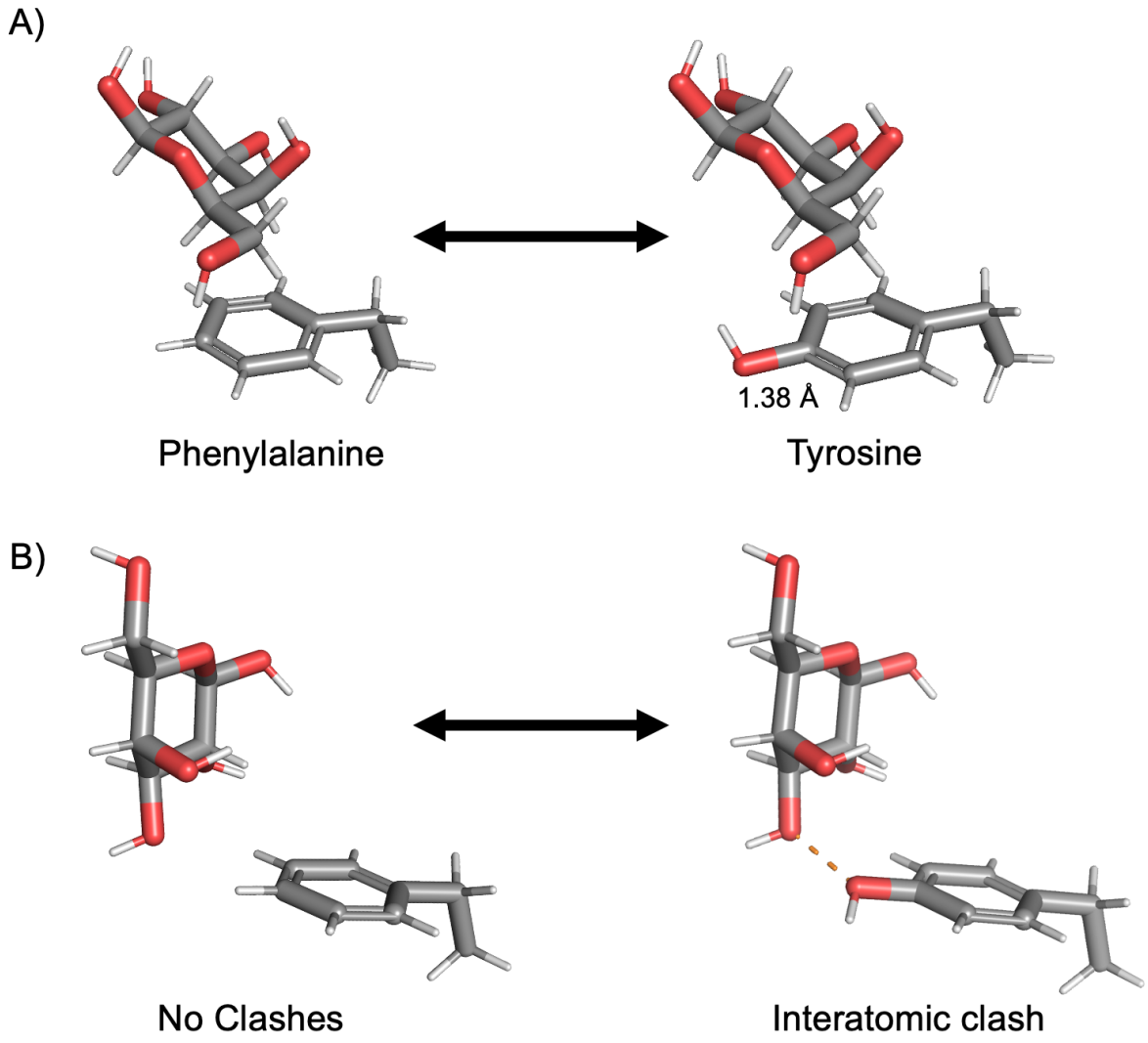

Figure S2. A) Example of adjustment of structures to convert a phenylalanine close contact into one with tyrosine and vice versa. For the conversion of Phe to Tyr, the C–O distance is set to 1.38 Å along the *para* C–H bond vector of phenylalanine. For the conversion of Tyr to Phe, an H is placed along the C–O bond vector and the original O and H are deleted. For both conversions, hydrogens are added by PyMOL v. 2.5.2 and optimized using B3LYP-D3/aug-cc-pVDZ. B) Visualization of an atomic clash formed by the conversion of phenylalanine (left) to tyrosine (right). Clashes are defined as having a distance relative to the sum of van der Waals radii of < 0.75 for any pair of atoms. Atoms are colored as follows: carbon in gray, oxygen in red, and hydrogen in white.

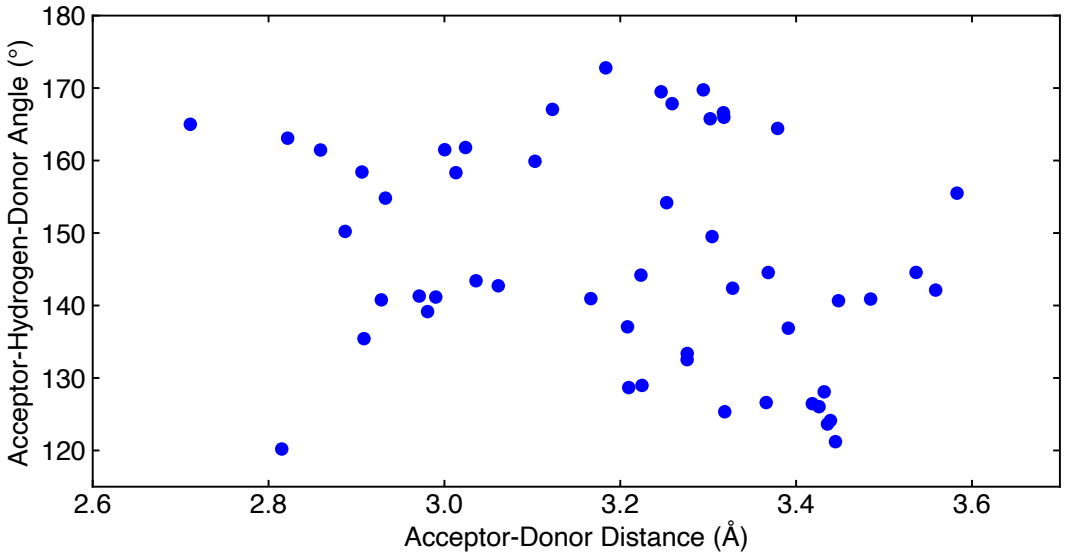

Figure S3. Scatter plot of intermolecular hydrogen bonds observed in the dataset: heavy atom acceptor-donor distance (in Å) versus acceptor-hydrogen-donor angle (in °, i.e., a number close to 180 corresponds to a more linear hydrogen bond). Hydrogen positions were optimized using DFT. Each hydrogen bond is plotted separately when close contacts form multiple intermolecular hydrogen bonds.

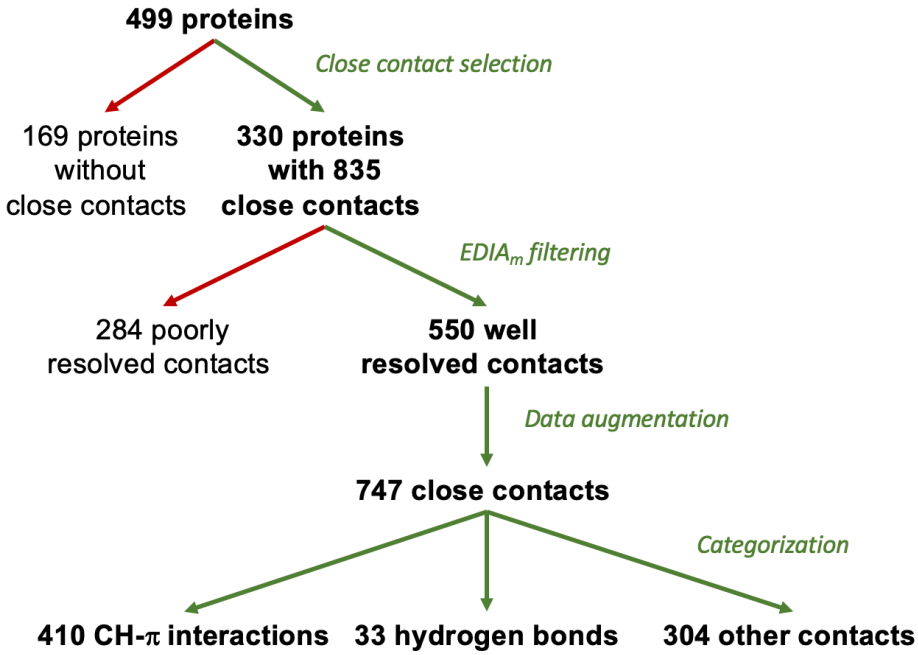

Figure S4. Schematic depicting the four data filtering steps performed to curate the full dataset. All stated numbers include all interactions formed between β-D-galactose and the three amino acids, tryptophan, tyrosine, and phenylalanine.

Table S2. Interaction totals for the full close contacts dataset grouped by amino acid interacting with galactose and by interaction type. Each close contact is assigned to a single type of interaction.

		Well-resolved Close Contacts	Stacking Interactions	Hydrogen Bonds	Other
Tryptophan	All (Native)	351	272	29	50
Tyrosine	Native	154	51	4	99
	Non-Native	43	18	0	25
	All	197	69	4	124
Phenylalanine	Native	45	18	0	27
	Non-Native	154	51	0	103
	All	199	69	0	130
Total	All	747	410	33	304

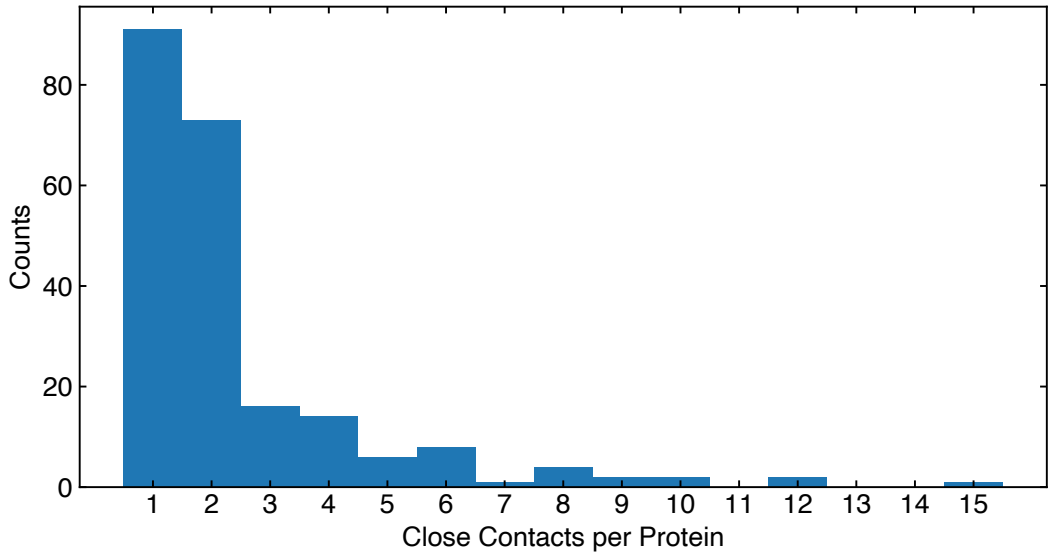

Figure S5. Histogram of the 550 close contacts formed between galactose and tryptophan, tyrosine, or phenylalanine that had an $EDIA_m > 0.8$ for both species, binned by the number of close contacts obtained from each protein.

Table S3. Characterization of the 550 close contacts (CCs) observed in the dataset. For contacts that are duplicates found in repeating subunits of a multimeric protein, original instances are not counted as duplicates while all other duplicates are (e.g., a protein with 4 repeating subunits would contribute 3 duplicates and 1 original, non-duplicate CC). For contacts sharing a galactose with 1+ other residues, all CCs that share a galactose with another CC are included. Since some CCs are duplicates from a repeating subunit and share a galactose with another CC, contacts for which neither is true are specified in the third row.

Close Contact (CC) Origins	**# of CCs**	**# of CCs that are not**
Contacts that are duplicates found in repeating subunits	192	358
Contacts sharing a galactose with 1+ other residues	193	357
Contacts found in proteins with multiple close contacts that are not duplicated and don't share a galactose	56	494
Contacts found in proteins with multiple close contacts	459	91

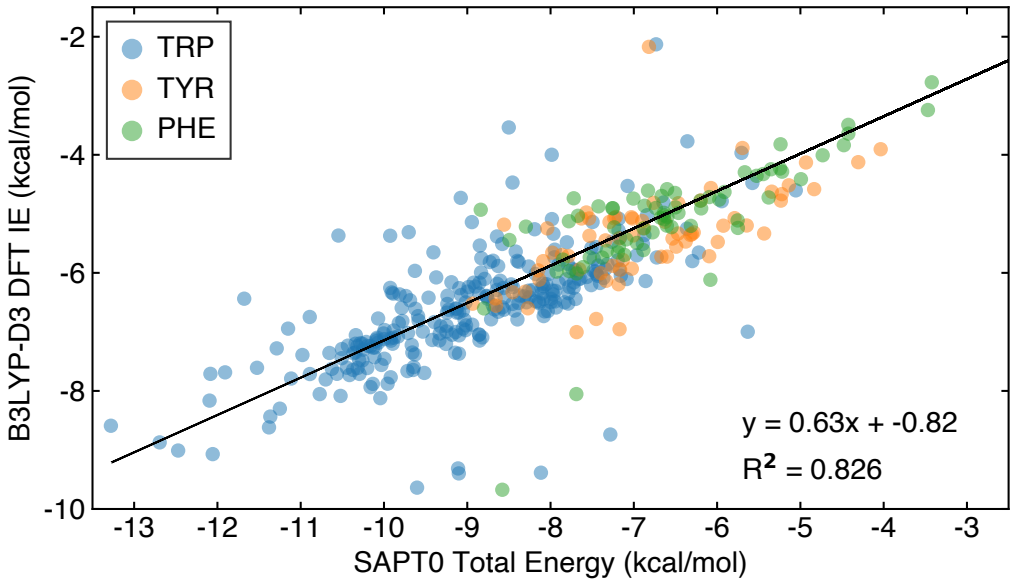

Figure S6. Comparison of B3LYP-D3 DFT IEs relative to SAPT0 total interaction energies of 410 CH–π interactions (all in kcal/mol). Both were evaluated using the aug-cc-pVDZ basis set, and DFT calculations were computed using implicit solvent corrections with the conductor-like polarizable continuum model (C-PCM) with ε=10.

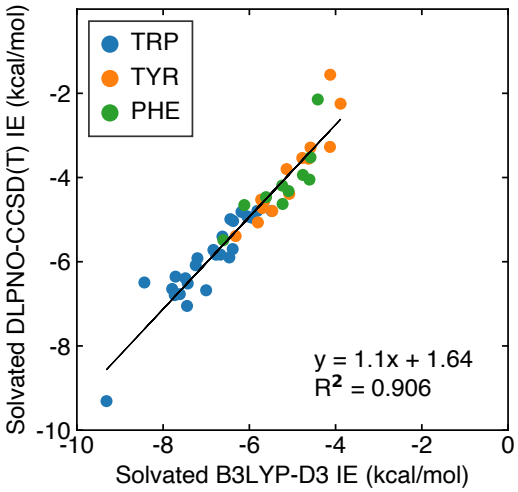

Figure S7. Comparison of interaction energies (IEs) obtained using solvent-corrected B3LYP-D3 DFT against those obtained using solvent-corrected DLPNO-CCSD(T) on the benchmarking dataset of 50 CH–π interactions (all in kcal/mol). DFT IEs were evaluated using the aug-cc-pVDZ basis set and were solvated using C-PCM with ε=10. DLPNO-CCSD(T) IEs were evaluated at the aug-CBS, using the two-point extrapolation formula and the aug-cc-pVDZ and aug-cc-pVTZ basis sets, used the DLPNO approximation and a two-point extrapolation to the complete pair natural orbital space (CPS), and included MP2-derived solvent correction.

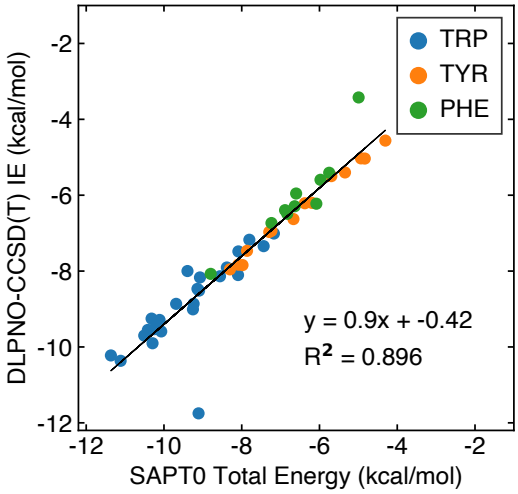

Figure S8. Comparison of gas-phase SAPT0 total energies against gas-phase DLPNO-CCSD(T) interaction energies (IEs) on the benchmarking dataset of 50 CH– π interactions (all in kcal/mol) between galactose and Trp (blue circles), Tyr (orange circles), and Phe (green circles). A best-fit line (black line) with R^2 value is annotated in inset. SAPT0 energies were evaluated using the aug-cc-pVDZ basis set. DLPNO-CCSD(T) IEs were extrapolated to the complete basis set limit (aug-CBS) using the two-point extrapolation formula and the aug-cc-pVDZ and aug-cc-pVTZ basis sets. The DLPNO-CCSD(T) energies also used a two-point extrapolation to the complete pair natural orbital space (CPS). The single tryptophan outlier is likely caused by optimization of the galactose monomer into a less favorable local minima, leading to a stronger interaction energy by DLPNO-CCSD(T) calculation than SAPT0, which computes the total energy using only the dimer.

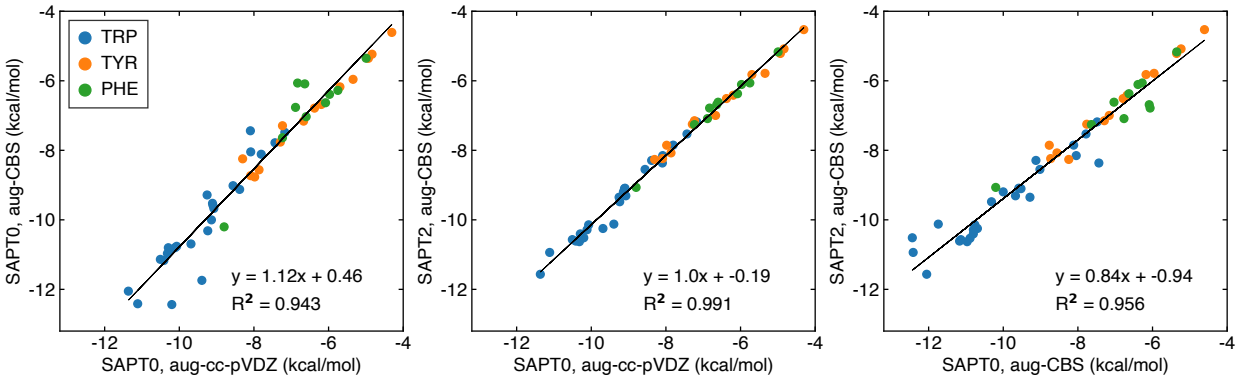

Figure S9. Comparison of total energies obtained using SAPT0 and SAPT2 on the benchmarking dataset of 50 CH–π interactions. All energies were obtained with either the aug-cc-pVDZ basis set or the two-point extrapolation to the augmented complete basis set limit (aug-CBS) using the aug-cc-pVDZ and aug-cc-pVTZ basis sets, as labeled.

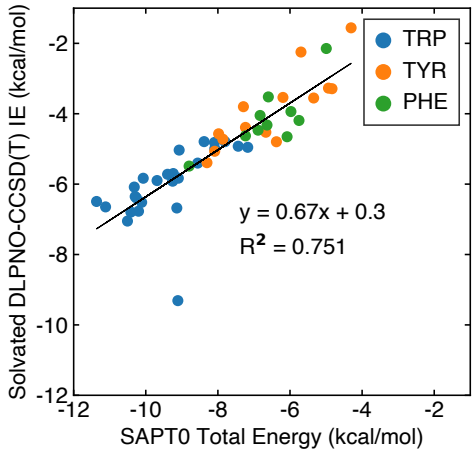

Figure S10. Comparison of gas-phase SAPT0 total energies against solvent-corrected DLPNO-CCSD(T) interaction energies (IEs) on the benchmarking dataset of 50 CH–π interactions (all in kcal/mol). SAPT0 energies were evaluated using the aug-cc-pVDZ basis set. DLPNO-CCSD(T) IEs were evaluated at the aug-CBS, using the two-point extrapolation formula and the aug-cc-pVDZ and aug-cc-pVTZ basis sets, the DLPNO approximation with a two-point extrapolation to the complete pair natural orbital space (CPS), and an MP2 derived solvent correction. The single tryptophan outlier is likely caused by optimization of the galactose monomer into a less favorable local minima, leading to a stronger interaction energy by DLPNO-CCSD(T) calculation than SAPT0, which computes the total energy using only the dimer.

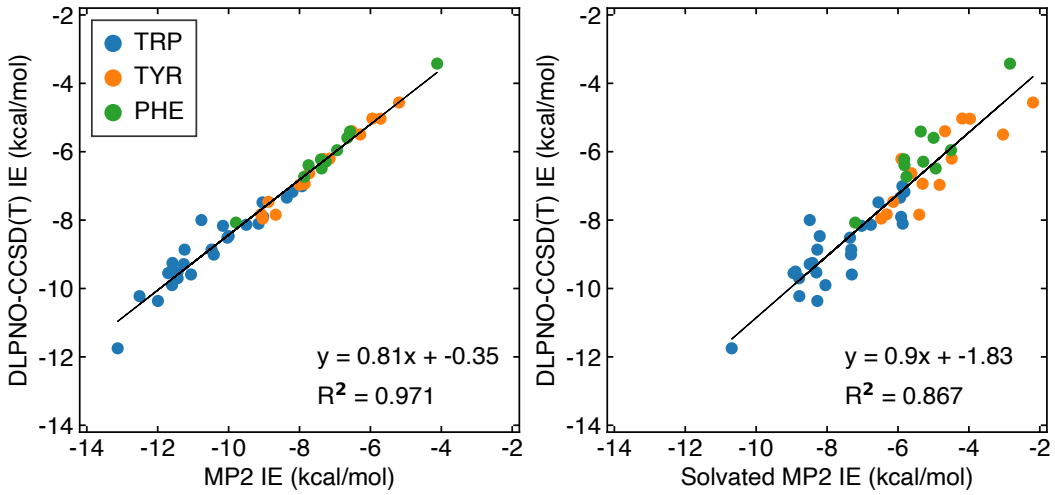

Figure S11. Evaluation of interaction energies (IEs) obtained using RI-MP2 with and without implicit solvent compared against DLPNO-CCSD(T) IEs on the benchmarking dataset of 50 CH–π interactions (all in kcal/mol). All energies were obtained at the aug-CBS, using the two-point extrapolation formula and the aug-cc-pVDZ and aug-cc-pVTZ basis sets. Solvated MP2 IEs were obtained by using the conductor-like polarizable continuum model (C-PCM) with an ε set to 10. DLPNO-CCSD(T) calculations were done using the DLPNO approximation and a two-point extrapolation to the complete pair natural orbital space (CPS).

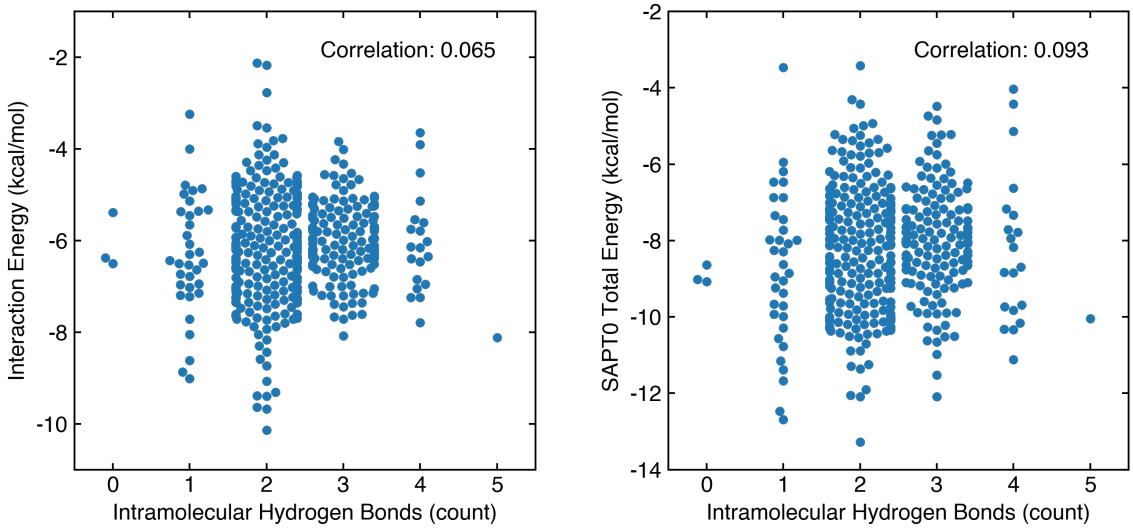

Figure S12. Comparison of intramolecular hydrogen bonds and (left) DFT interaction energy (IE) or (right) SAPT0 total energy in kcal/mol. Intramolecular hydrogen bonds have a maximum hydrogen-acceptor distance of 2.6 Å and minimum donor-hydrogen-acceptor angle of 100°. Pearson correlation coefficients for both datasets are reported.

Table S4. Population statistics of B3LYP-D3/aug-cc-pVDZ interaction energies (kcal/mol) by interaction type, multivalent stacking interactions, hydrogen bonding interactions, and other interactions that do not satisfy the criteria for the other two categories.

	Multivalent Stacking Interactions (N=410)	Hydrogen Bonding Interactions (N=33)	Other Interactions (N=304)
Average Interaction Energy (kcal/mol)	-6.1	-4.4	-3.2
Standard Deviation (kcal/mol)	1.9	1.0	1.3

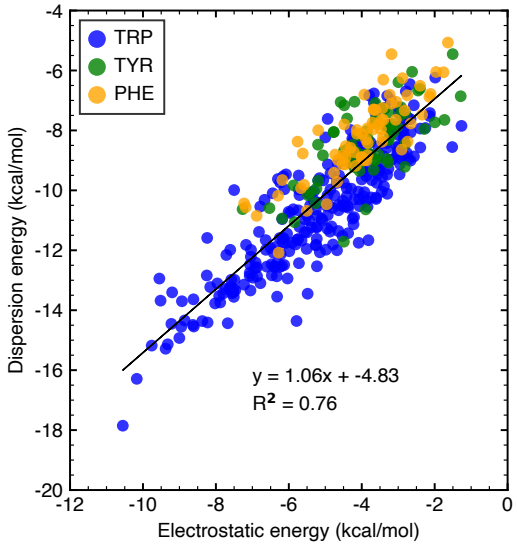

Figure S13. Comparison of SAPT0 dispersion and electrostatic energies for the CH–π interactions formed with tryptophan (blue), tyrosine (green), and phenylalanine (yellow). A best-fit line for the full dataset is shown. All energies are reported in kcal/mol. SAPT0 energies were evaluated using the aug-cc-pVDZ basis set.

Table S5. Binding interaction characterization for the 10 strongest CH-π interactions by DFT interaction energy (IE) and SAPT0 total energy. The carbohydrate chain length, number of CH-π stacking interactions and contacts, the total number of hydrogen bonds and the number of charged hydrogen bonds are all reported. Hydrogen bonds were identified by PyMol using the default criteria.

	DFT IE (kcal/mol)	SAPT0 Total Energy (kcal/mol)	Carbohydrate chain length	CH-π interactions	Hydrogen bonds	Charged Hydrogen Bonds
4a4a_0	-8.6	-13.3	2	1 stacking + 2 contacts	15	0
5gqf_1	-9	-12.5	2	1 stacking	11	8
6orh_0	-8.2	-12.1	4	2 stacking	18	1
4aw7_1	-9.1	-12.1	6	2 stacking + 1 contact	19	11
5mxh_2	-8.6	-11.4	1	1 stacking	9	6
6v1c_0	-8.4	-11.4	2	1 stacking	6	2
5e1q_0	-8.3	-11.3	3	2 stacking	10	7
3sxe_0	-8.1	-10.8	1	1 stacking	4	2
3wkh_0	-8.1	-10.5	2	2 stacking	16	4
3ah4_0	-8.1	-10	1	1 stacking	6	2

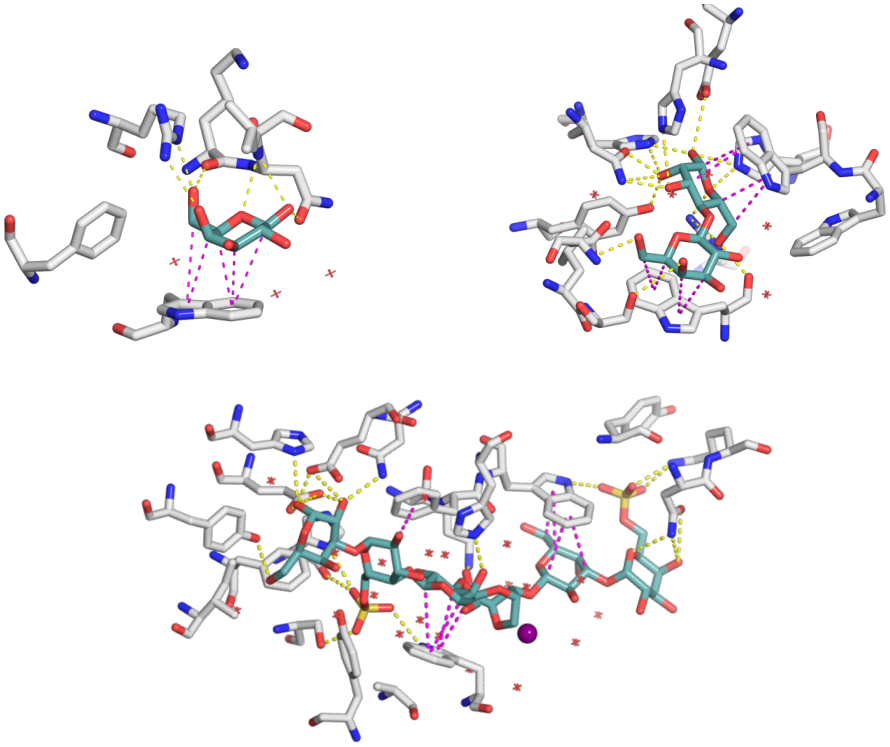

Figure S14. Visualization of the binding sites of three highly favorable CH-π stacking interactions from the following proteins, (upper left) progenitor toxin (PDB ID: 3ah4), (upper right) cellobiose epimerase (PDB ID: 3wkh), and (bottom) b-porphyranase (PDB ID: 4aw7). Atoms are colored as follows: protein carbon in light gray, carbohydrate carbon in teal, oxygen in red, nitrogen in blue, and sulphur in yellow. Hydrogen bonds are shown as yellow dashed lines and CH-π contacts are shown as magenta dashed lines.

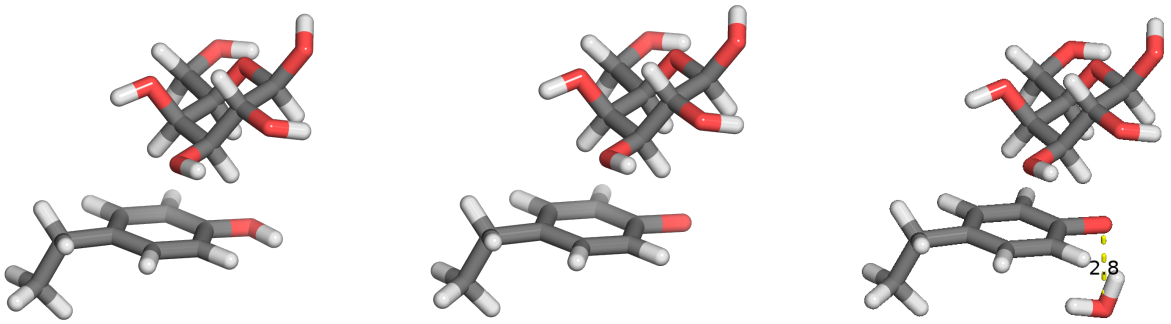

Figure S15. Example conversion of an initial (left) tyrosine close contact into a (center) deprotonated phenoxide close contact and finally, a (right) phenoxide close contact coordinated to a water molecule. The water molecule was placed beneath the phenoxide oxygen atom and optimized in Avogadro to satisfy a constraint of an O-O distance of 2.8 Å using MMFF94. All hydrogens are subsequently optimized using B3LYP-D3/aug-cc-pVDZ in TeraChem.

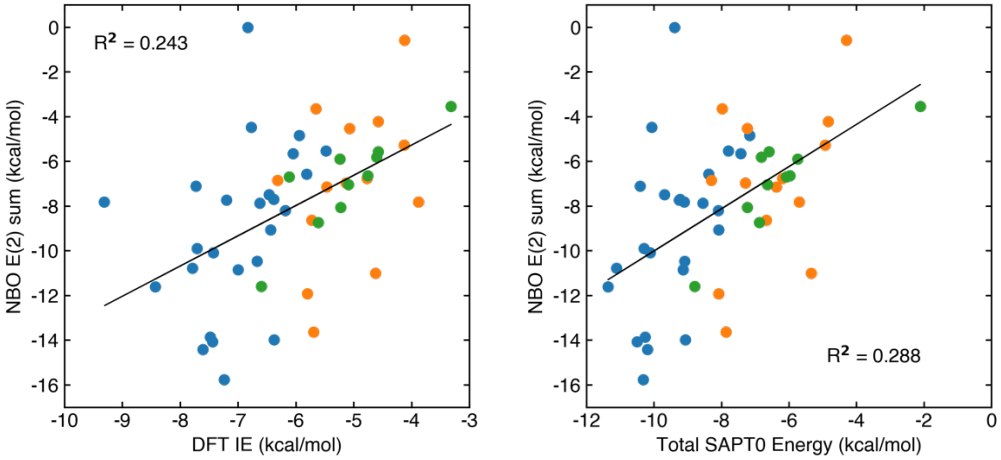

Figure S16. Comparison of the (left) solvent-corrected B3LYP-D3 DFT interaction energies (IEs) and (right) SAPT0 total energies, to the sum of all NBO perturbative E(2) energies between the carbohydrate and protein residues. All energies were evaluated for the benchmarking dataset of 50 CH–π interactions using the aug-cc-pVDZ basis set. B3LYP-D3 DFT IEs were computed using C-PCM with ε=10. Energies in kcal/mol. R^2 values are reported for both sets.

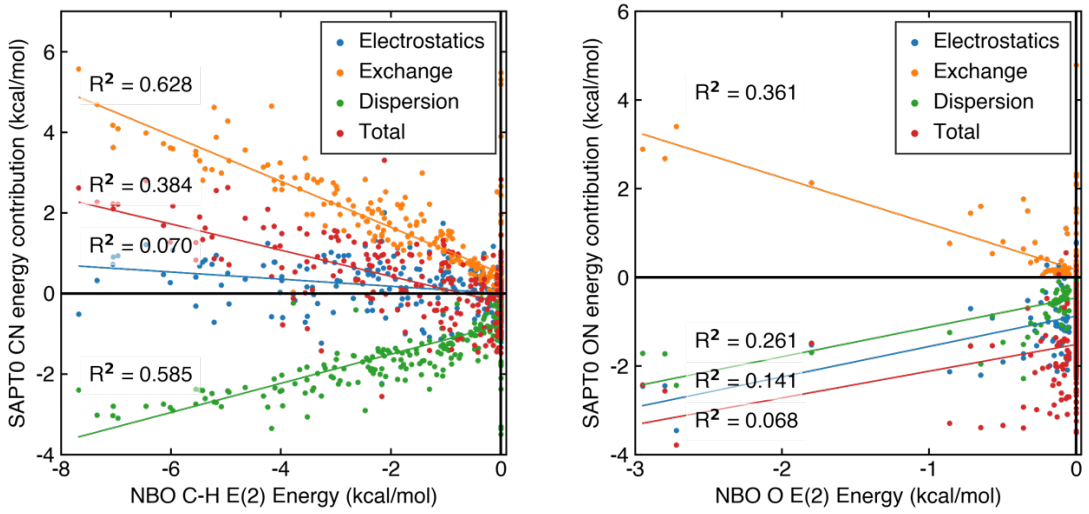

Figure S17. Comparison of the NBO perturbative E(2) energies for individual C-H groups and oxygen atoms on the carbohydrate relative to the corresponding F-SAPT energetic contributions (electrostatic, exchange, dispersion, and total energies). All energies were evaluated for the benchmarking dataset of 50 CH–π interactions using the aug-cc-pVDZ basis set. R^2 values and linear fit lines are shown for each SAPT0 energy.

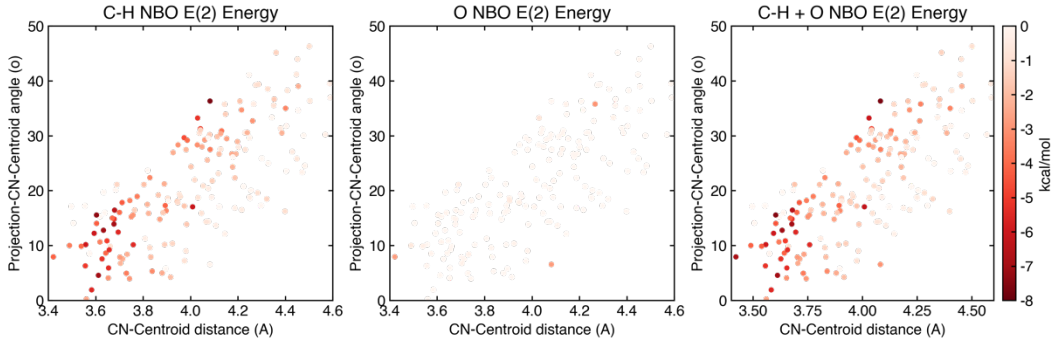

Figure S18. Scatterplot of the Cn-centroid distance ($d_{Cn\text{-}Ctr}$), the distance between carbon atom Cn on galactose and the centroid of the nearest aromatic ring (in Å), versus the projection-Cn-centroid angle ($q_{Proj\text{-}Cn\text{-}Ctr}$), the angle between the distance vector and the vector formed by the projection of Cn onto the plane of the aromatic ring system ($proj_{Cn}$) (in °). Scatter plots are colored by the perturbative NBO E(2) energy contribution of the (left) C-H group, (center) oxygen atom, and (right) both combined. This analysis was performed on the benchmarking dataset of 50 CH–π interactions and energies are reported in kcal/mol.

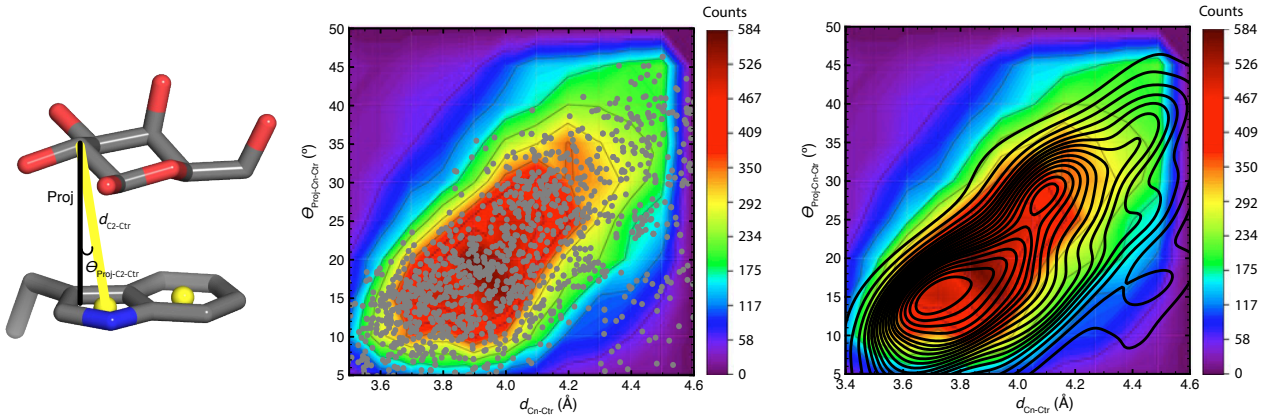

Figure S19. (left) Visualization of the distance and angle computed for galactose carbon atom 2. Atoms are colored as follows: carbon in gray, oxygen in red, and nitrogen in blue. $d_{C2\text{-}Ctr}$ is the distance between carbon atom 2 on galactose and the centroid of the nearest aromatic ring. $\theta_{Proj\text{-}C2\text{-}Ctr}$ is the angle between the distance vector and the vector formed by the projection of C2 onto the plane of the aromatic ring system ($proj_{C2}$). (center) Scatter plot and (right) KDE plot of distance and angle values computed for all galactose carbons in our full dataset overlaid on top of the heat map of distance and angle features computed by Houser, et. al. Adapted with permission from reference [1]. Copyright 2020 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim.

Table S6. Population statistics of F-SAPT energetic contributions (kcal/mol) from the three functional group sets, one containing the carbon atom (Cn) only, one containing the bound oxygen atom (On) only, and one containing the two together (Cn + On). Cn interactions are also split by whether they are endocyclic or exocyclic.

(kcal/mol)	**Maximum**	**Minimum**	**Median**	**Average**	**Standard Deviation**
Cn	4.0	-4.5	0.1	0.3	1.0
Endocyclic Cn	4.0	-4.5	0.3	0.5	1.0
Exocyclic Cn	2.5	-2.6	-0.6	-0.5	0.8
On	2.0	-4.3	-1.6	-1.6	0.8
Cn+On	1.7	-6.4	-1.4	-1.4	1.0

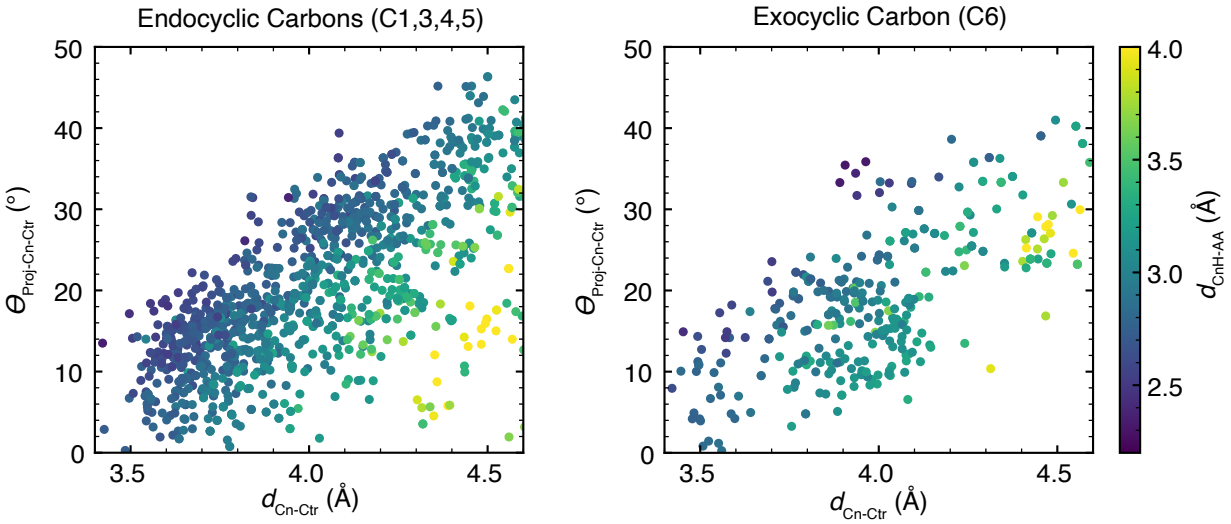

Figure S20. Scatterplot of $d_{Cn\text{-}Ctr}$, the distance between carbon atom Cn on galactose and the centroid of the nearest aromatic ring (in Å) versus $\theta_{Proj\text{-}Cn\text{-}Ctr}$, the angle between the distance vector and the vector formed by the projection of Cn onto the plane of the aromatic ring system ($proj_{Cn}$) (in °) for our dataset of CH–π stacking interactions. $d_{Cn\text{-}Ctr}$ and $\theta_{Proj\text{-}Cn\text{-}Ctr}$ computed for (left) interacting endocyclic carbon atoms and (right) exocyclic carbon atoms. Scatter plots are colored by the distance between the hydrogen atom bound to Cn (CnH) and the nearest heavy atom on the aromatic amino acid ($d_{CnH\text{-}AA}$).

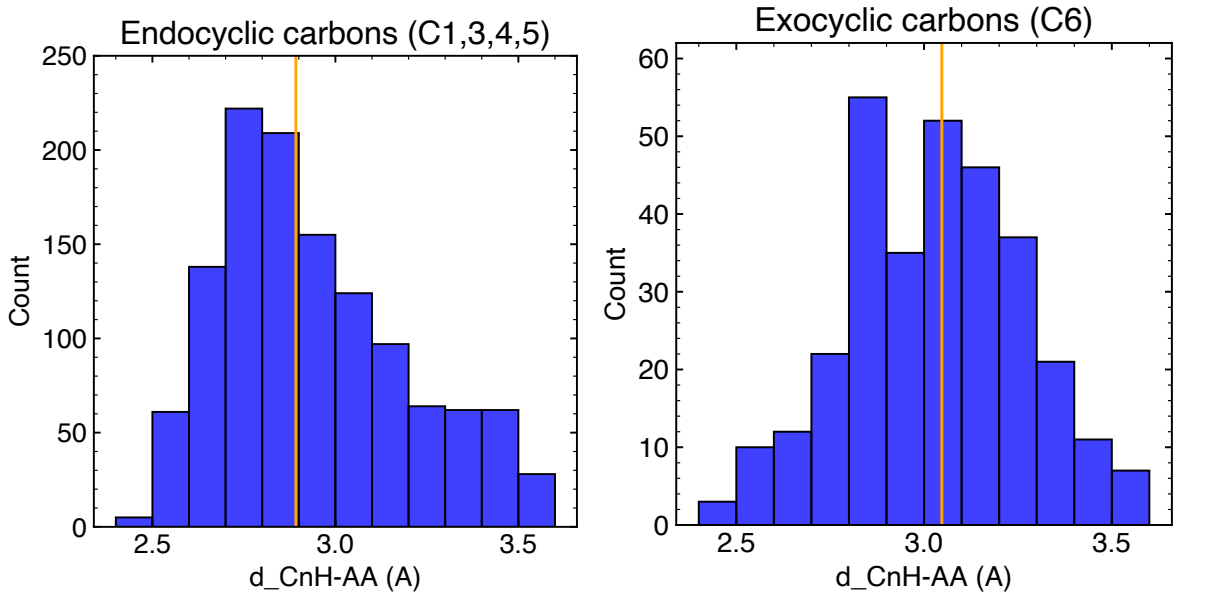

Figure S21. Histogram of the distance between the hydrogen atom bound to Cn (CnH) and the nearest heavy atom on the aromatic amino acid ($d_{CnH\text{-}AA}$) in Å. Median values of each distribution are shown as a vertical orange line. The bin width is 0.1 Å.

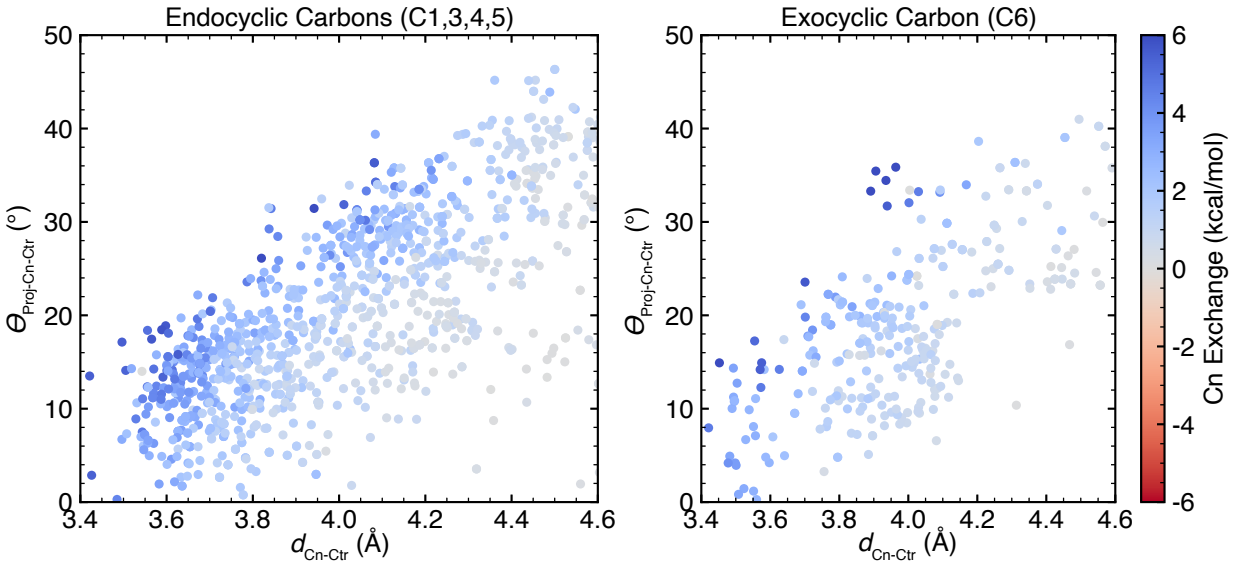

Figure S22. Scatterplot of $d_{Cn\text{-}Ctr}$, the distance between carbon atom Cn on galactose and the centroid of the nearest aromatic ring (in Å) versus $\theta_{Proj\text{-}Cn\text{-}Ctr}$, the angle between the distance vector and the vector formed by the projection of Cn onto the plane of the aromatic ring system ($proj_{Cn}$) (in °) for our full dataset of CH–π stacking interactions. Data are shown for (left) interacting endocyclic carbon atoms and (right) exocyclic carbon atoms. Scatter plots are colored by the exchange energy in kcal/mol from the interaction between Cn and the amino acid, according to the color scale shown at the far right.

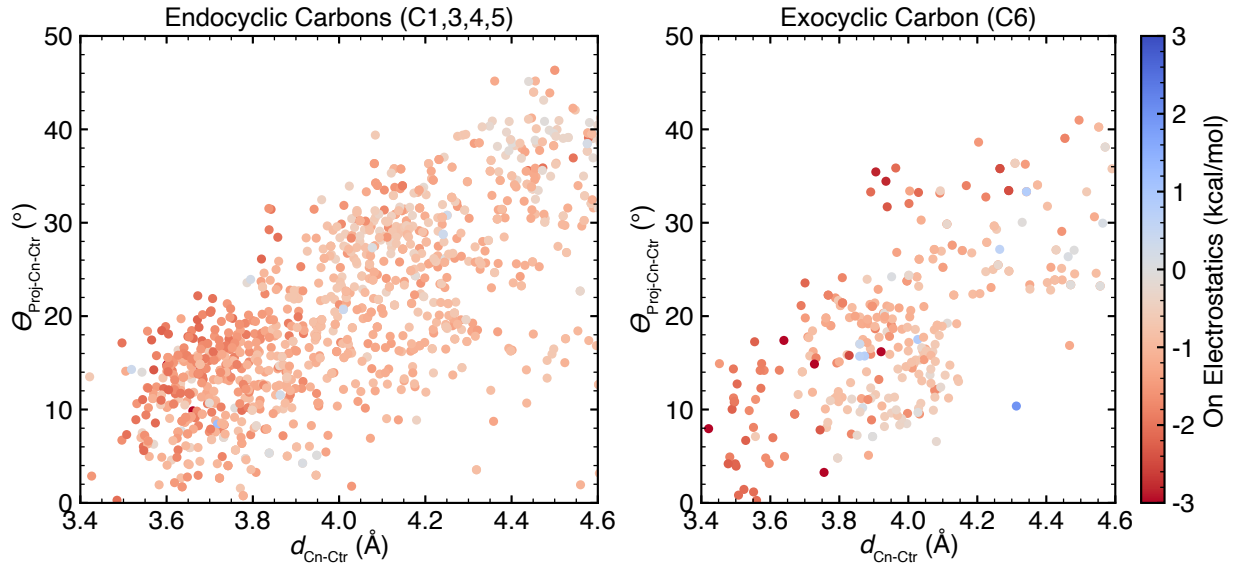

Figure S23. Scatterplot of $d_{Cn\text{-}Ctr}$, the distance between carbon atom Cn on galactose and the centroid of the nearest aromatic ring (in Å) versus $\theta_{Proj\text{-}Cn\text{-}Ctr}$, the angle between the distance vector and the vector formed by the projection of Cn onto the plane of the aromatic ring system ($proj_{Cn}$) (in °) for our full dataset of CH–π stacking interactions. Data are shown for (left) interacting endocyclic carbon atoms and (right) exocyclic carbon atoms. Scatter plots are colored by the electrostatic energy in kcal/mol from the interaction between On and the amino acid, according to the color scale at the far right.

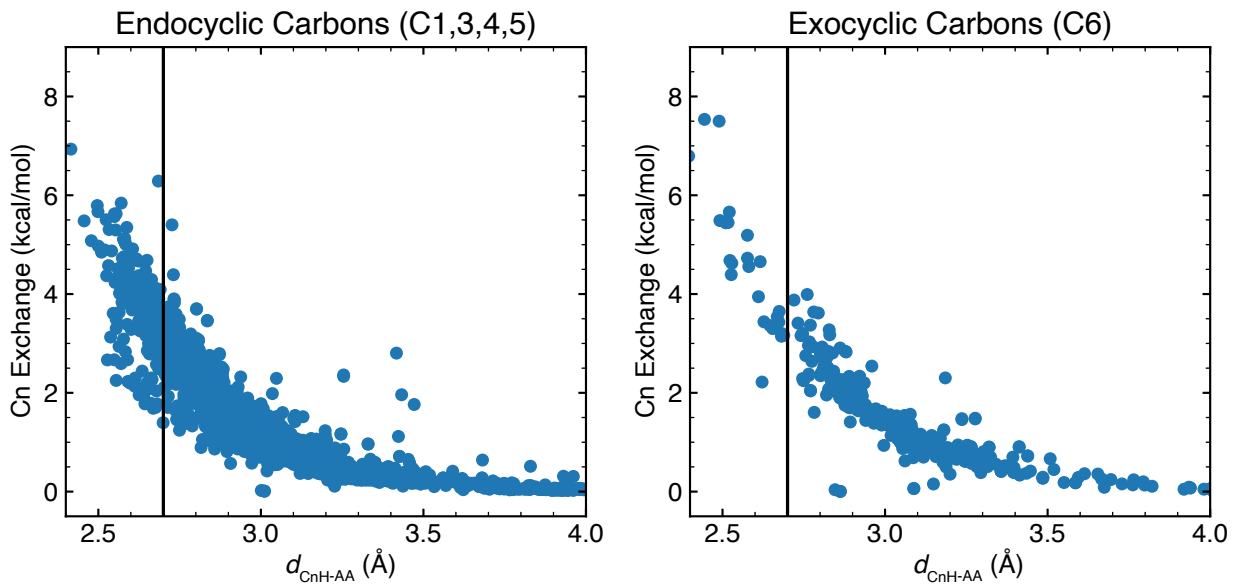

Figure S24. Scatter plots of the distance between the hydrogen atom bound to Cn (CnH) and the nearest heavy atom on the aromatic amino acid ($d_{CnH\text{-}AA}$) in Å versus the Cn exchange energy contribution in kcal/mol. Data are shown for (left) interacting endocyclic carbon atoms and (right) exocyclic carbon atoms. Black line at 2.7 Å overlaid to separate interactions with weaker overall CH–π interaction energy contributions due in part to high exchange values.

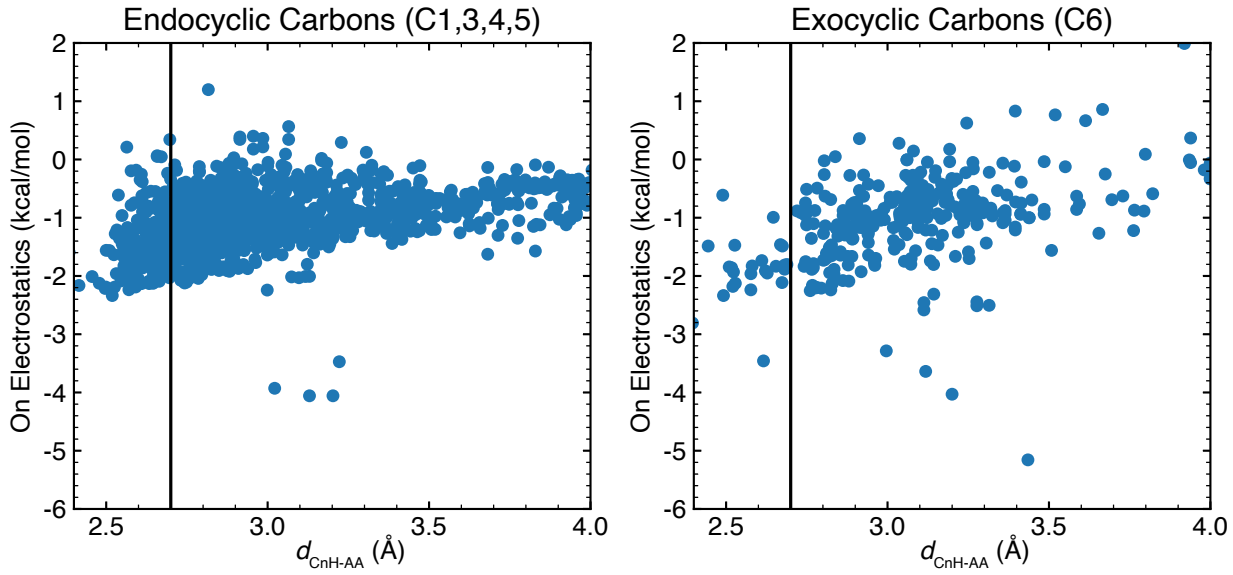

Figure S25. Scatter plots of the distance between the hydrogen atom bound to Cn (CnH) and the nearest heavy atom on the aromatic amino acid ($d_{CnH\text{-}AA}$) in Å versus the On electrostatic energy contribution in kcal/mol. Data are shown for (left) interacting endocyclic carbon atoms and (right) exocyclic carbon atoms. Black line at 2.7 Å overlaid to separate interactions with weaker overall CH–π interaction energy contributions and slightly more favorable electrostatic energies.

Table S7. Correlation coefficients for all the distance ($d_{Cn\text{-}Ctr}$) and angle ($\theta_{Proj\text{-}Cn\text{-}Ctr}$) features in feature set 1. Each cell is colored by the correlation coefficient with +1, 0, and -1, colored in red, white, and blue, respectively.

	d C1-Ctr	θ Proj-C1-Ctr	d C2-Ctr	θ Proj-C2-Ctr	d C3-Ctr	θ Proj-C3-Ctr	d C4-Ctr	θ Proj-C4-Ctr	d C5-Ctr	θ Proj-C5-Ctr	d C6-Ctr	θ Proj-C6-Ctr
d C1-Ctr	1											
θ Proj-C1-Ctr	0.47	1										
d C2-Ctr	0.73	0.59	1									
θ Proj-C2-Ctr	0.34	0.66	0.69	1								
d C3-Ctr	0.06	0.36	0.69	0.64	1							
θ Proj-C3-Ctr	0.25	0.21	0.56	0.81	0.59	1						
d C4-Ctr	-0.54	0.01	-0.02	0.16	0.64	0.14	1					
θ Proj-C4-Ctr	-0.01	-0.18	0.05	0.22	0.16	0.54	0.28	1				
d C5-Ctr	0.22	0.08	0.04	0.01	-0.02	-0.02	0.24	0.26	1			
θ Proj-C5-Ctr	0.17	0.1	-0.03	-0.04	-0.11	-0.04	0.12	0.34	0.82	1		
d C6-Ctr	-0.25	-0.19	-0.47	-0.28	-0.31	-0.29	0.31	0.2	0.78	0.63	1	
θ Proj-C6-Ctr	-0.19	-0.22	-0.4	-0.46	-0.29	-0.45	0.23	0.06	0.58	0.58	0.82	1

Table S8. Evaluation of the random forest models trained on the distance ($d_{Cn\text{-}Ctr}$) and angle ($\theta_{Proj\text{-}Cn\text{-}Ctr}$) of each carbon in galactose to the centroid of the interacting aromatic ring. Train and test R^2 and mean absolute error (MAE) as well as test mean absolute percentage error (MAPE) values are reported for each model trained.

	Train R^2	**Train MAE**	**Test R^2**	**Test MAE**	**Test MAPE**
DFT	0.71	0.38	0.47	0.51	8%
SAPT0	0.81	0.48	0.59	0.69	9%
Dispersion	0.92	0.42	0.83	0.66	7%
Electrostatics	0.87	0.48	0.73	0.69	16%
Exchange	0.88	0.80	0.75	1.18	14%
Induction	0.84	0.15	0.68	0.22	14%

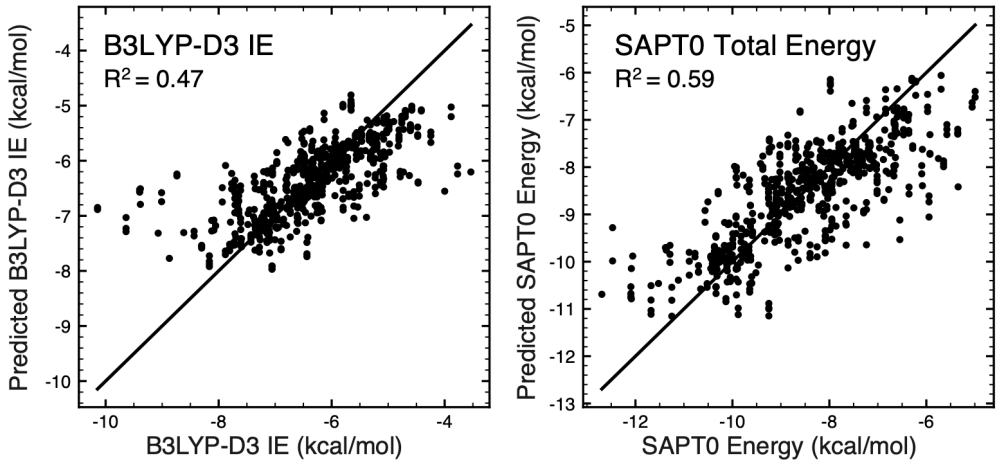

Figure S26. Parity plots of random forest model predictions on the test sets. Models were trained using the distance ($d_{Cn\text{-}Ctr}$) and angle ($\theta_{Proj\text{-}Cn\text{-}Ctr}$) features to predict the (left) B3LYP-D3 DFT interaction energy (IE) and (right) SAPT0 total energy in kcal/mol. R^2 values are listed on each plot.

Table S9. Evaluation of the random forest models trained on CH-π interactions containing tryptophan. The distance ($d_{\text{Cn-Ctr}}$) and angle ($\theta_{\text{Proj-Cn-Ctr}}$) of each carbon in galactose to the centroid of the interacting aromatic ring were used as features. Train and test R^2 and mean absolute error (MAE) as well as test mean absolute percentage error (MAPE) values are reported for each model trained.

	Train R^2	**Train MAE**	**Test R^2**	**Test MAE**	**Test MAPE**
DFT	0.64	0.36	0.36	0.51	8%
SAPT0	0.80	0.43	0.59	0.62	7%
Dispersion	0.91	0.42	0.82	0.62	6%
Electrostatics	0.87	0.48	0.71	0.70	16%
Exchange	0.87	0.85	0.73	1.23	15%
Induction	0.82	0.16	0.63	0.23	13%

Table S10. Evaluation of the random forest models trained on CH-π interactions containing tyrosine or phenylalanine. The distance ($d_{\text{Cn-Ctr}}$) and angle ($\theta_{\text{Proj-Cn-Ctr}}$) of each carbon in galactose to the centroid of the interacting aromatic ring were used as features. Train and test R^2 and mean absolute error (MAE) as well as test mean absolute percentage error (MAPE) values computed by sklearn are reported for each model trained.

	Train R^2	**Train MAE**	**Test R^2**	**Test MAE**	**Test MAPE**
DFT	0.72	0.26	0.09	0.37	7%
SAPT0	0.70	0.43	0.08	0.63	10%
Dispersion	0.77	0.39	0.15	0.66	8%
Electrostatics	0.72	0.40	0.05	0.62	17%
Exchange	0.69	0.72	-0.13	1.17	18%
Induction	0.70	0.11	-0.01	0.18	16%

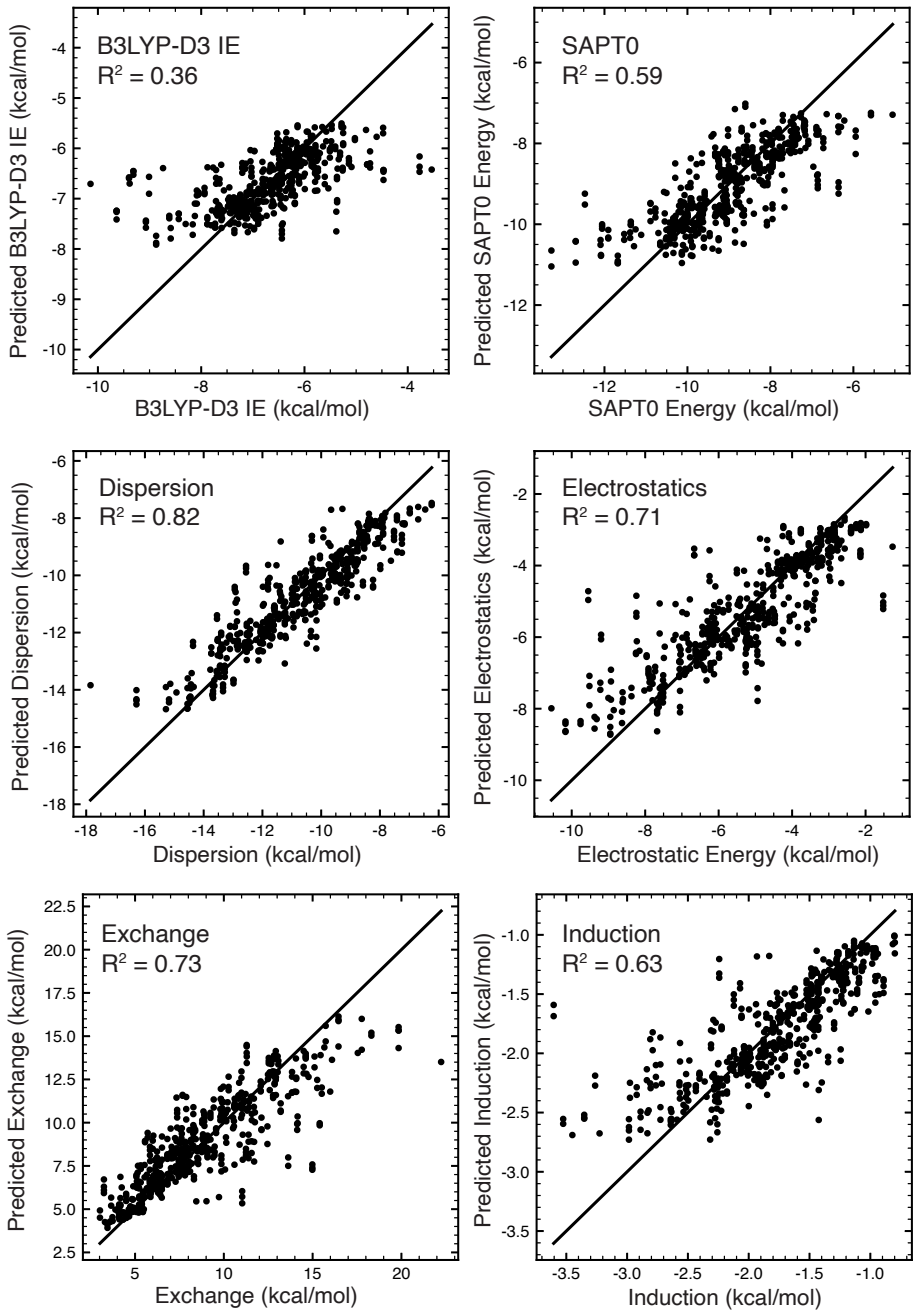

Figure S27. Parity plots of random forest model predictions on the test sets. Models were trained on CH-π interactions containing tryptophan using the distance ($d_{\text{Cn-Ctr}}$) and angle ($\theta_{\text{Proj-Cn-Ctr}}$) features to predict the (left) B3LYP-D3 DFT interaction energy (IE) and (right) SAPT0 total energy in kcal/mol. R^2 values are listed on each plot.

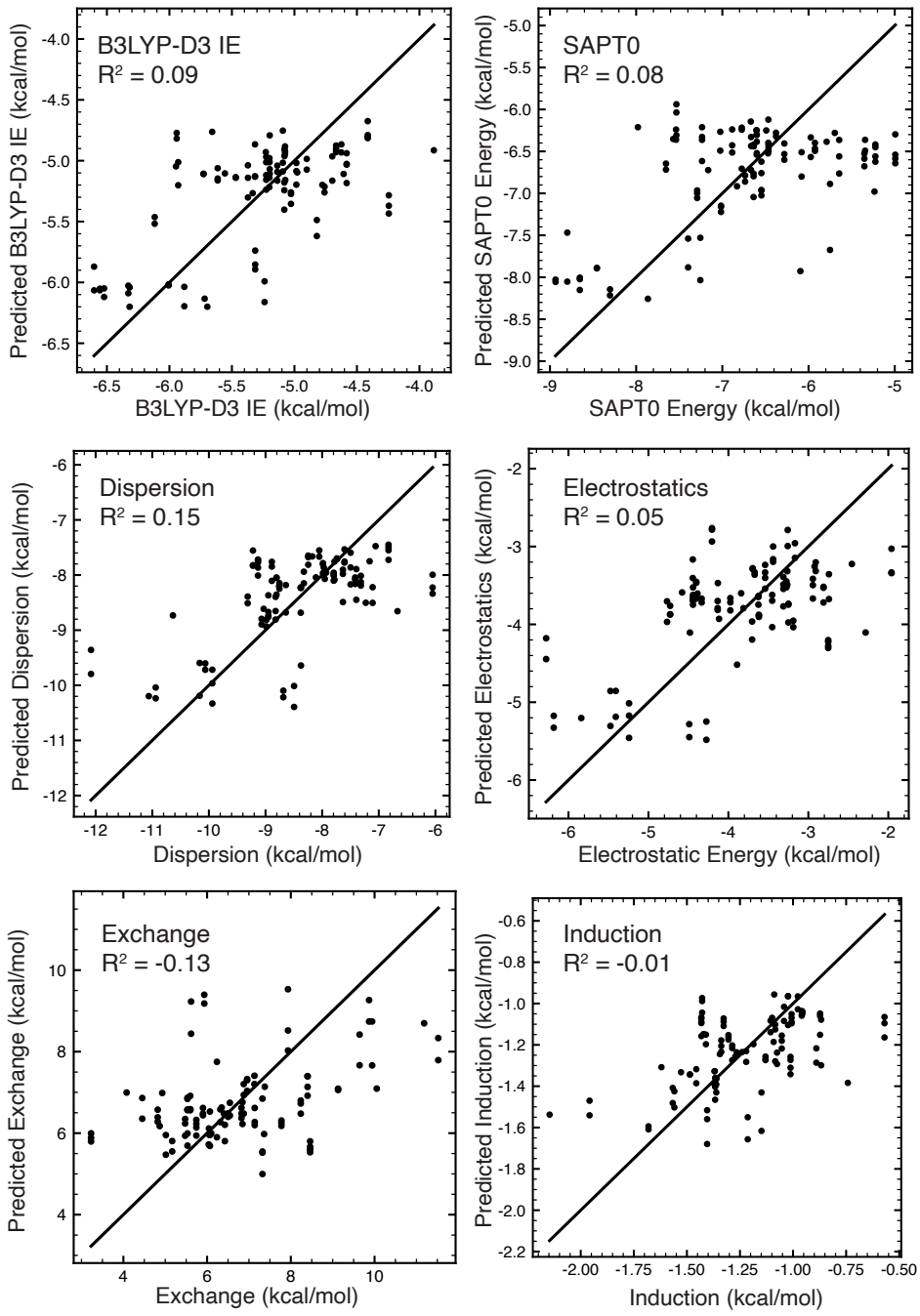

Figure S28. Parity plots of random forest model predictions on the test sets. Models were trained on CH-π interactions containing tyrosine and phenylalanine using the distance ($d_{\text{Cn-Ctr}}$) and angle ($\theta_{\text{Proj-Cn-Ctr}}$) features to predict the (left) B3LYP-D3 DFT interaction energy (IE) and (right) SAPT0 total energy in kcal/mol. R^2 values are listed on each plot.

Table S11. Feature Importance (I) values were computed from the mean decrease in impurity using the sklearn_feature_importances_ method. The top 5 features (Ft) out of the $d_{\text{Cn-Ctr}}$ (d_{Cn}) and $\theta_{\text{Proj-Cn-Ctr}}$ (θ_{Cn}) features in feature set 1 are listed with the corresponding I values. Top features are listed for the random forest models predicting DFT IE, SAPT0 total energy, and the SAPT0 energetic components, dispersion, electrostatics, exchange, and induction.

	Ft 1	I1	Ft2	I2	Ft3	I3	Ft4	I4	Ft5	I5
DFT IE	d_{C2}	0.23	d_{C5}	0.22	d_{C6}	0.09	d_{C3}	0.09	θ_{C1}	0.08
SAPT0	d_{C2}	0.30	d_{C6}	0.12	d_{C5}	0.11	d_{C3}	0.10	θ_{C3}	0.09
Dispersion	d_{C5}	0.23	d_{C3}	0.20	d_{C2}	0.20	d_{C6}	0.16	d_{C4}	0.08
Electrostatics	d_{C3}	0.19	d_{C6}	0.18	θ_{C2}	0.14	d_{C5}	0.13	d_{C2}	0.11
Exchange	d_{C3}	0.27	d_{C6}	0.20	d_{C5}	0.14	d_{C4}	0.12	d_{C2}	0.09
Induction	d_{C3}	0.24	d_{C2}	0.20	d_{C6}	0.14	d_{C5}	0.12	θ_{C2}	0.09

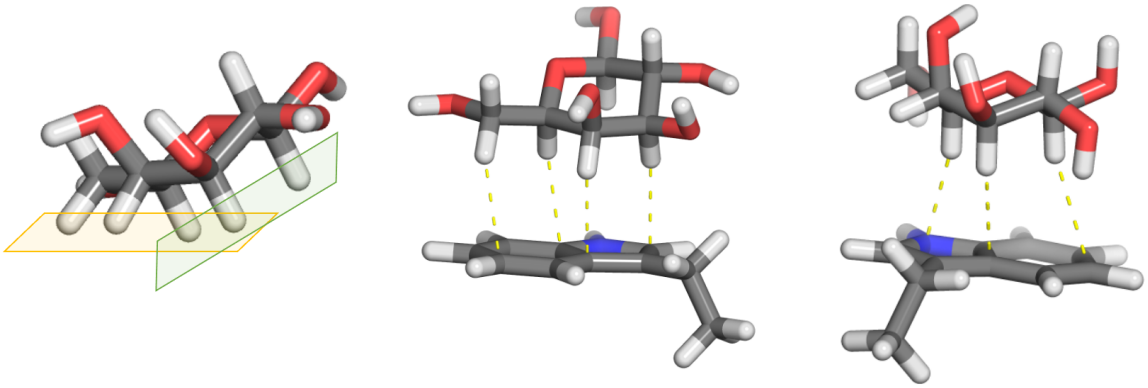

Figure S29. (left) Visualization of two sets of C-H groups capable of forming a hydrophobic face for stacking. Yellow plane: H3, H4, H5, and H6. Green plane: H1, H3, and H5. (center) Example CH–π stacking interaction formed by CH 3–6. (right) Example CH–π stacking interaction formed by CH 1,3,5. Atoms are colored as follows: carbon in gray, oxygen in red, and nitrogen in blue.

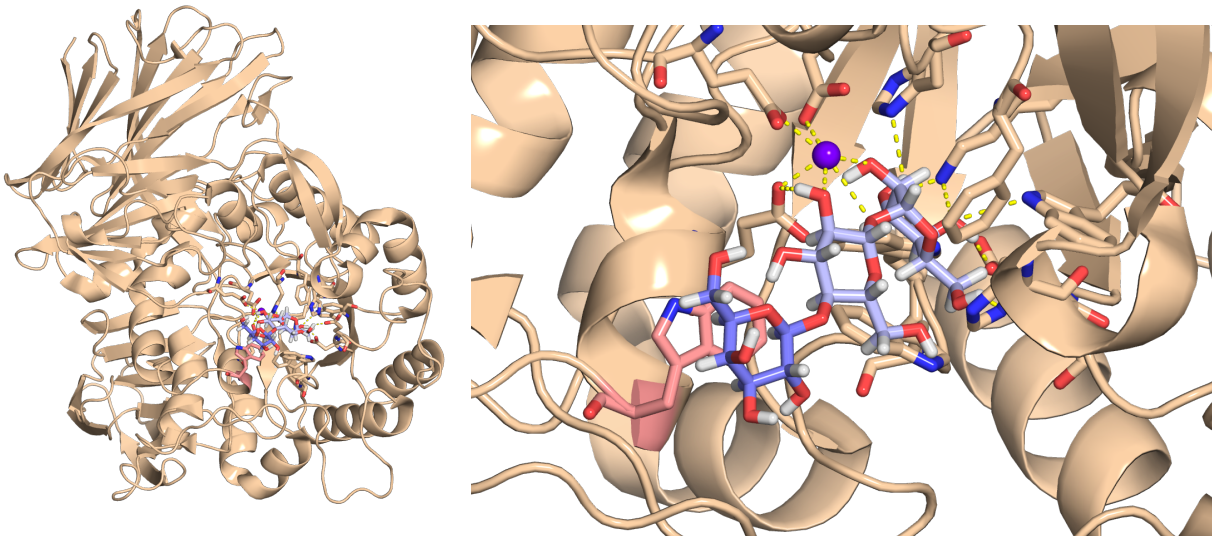

Figure S30. Visualization of the *Bacteroides thetaiotaomicrom* VPI-5482 glycoside hydrolase 97 (BtGH97 - PDB ID: 5E1Q) (left) full protein and (right) binding pocket with CH–π stacking interaction pair highlighted. Carbon atoms are colored as follows: galactose colored in purple, other carbohydrates in light purple, the interacting tryptophan colored in salmon, and all protein residues in wheat. All other atoms, regardless of molecule, are colored as follows, oxygen in red, nitrogen in blue, hydrogen in white, and calcium in dark purple. Polar contacts as classified by PyMOL that involve the carbohydrate ligand are shown with yellow dotted lines.

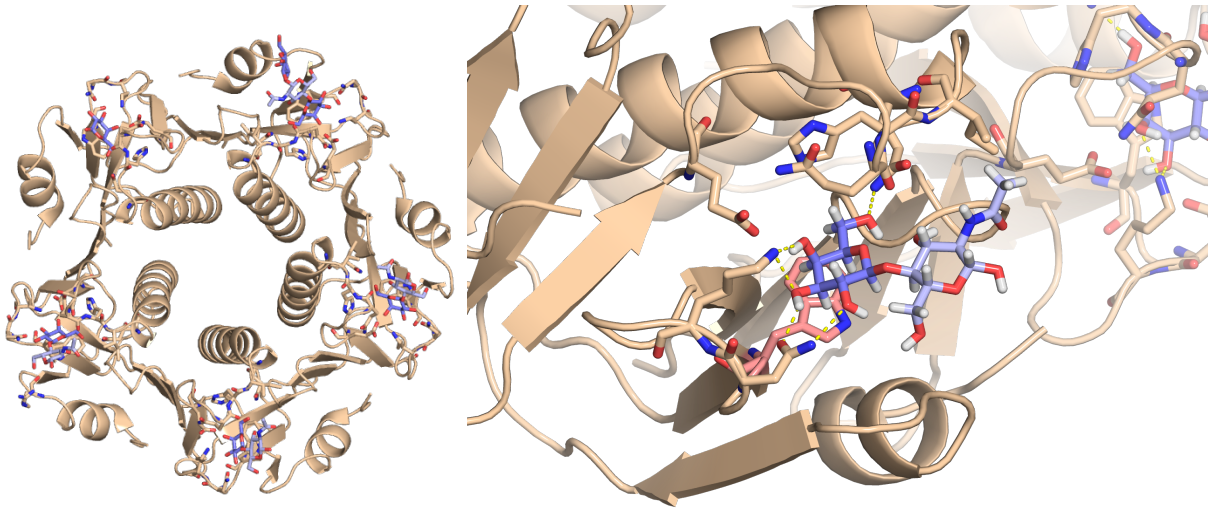

Figure S31. Visualization of the *Escherichia Coli* Enterotoxin (PDB ID: 2XRS) (left) full protein and (right) binding pocket with CH–π stacking interaction pair highlighted. Carbon atoms are colored as follows: galactose colored in purple, other carbohydrates in light purple, the interacting tryptophan colored in salmon, and all protein residues in wheat. All other atoms, regardless of the molecule, are colored as follows, oxygen in red, nitrogen in blue, and hydrogen in white. Polar contacts as classified by PyMOL that involve the carbohydrate ligand are shown with yellow dotted lines.

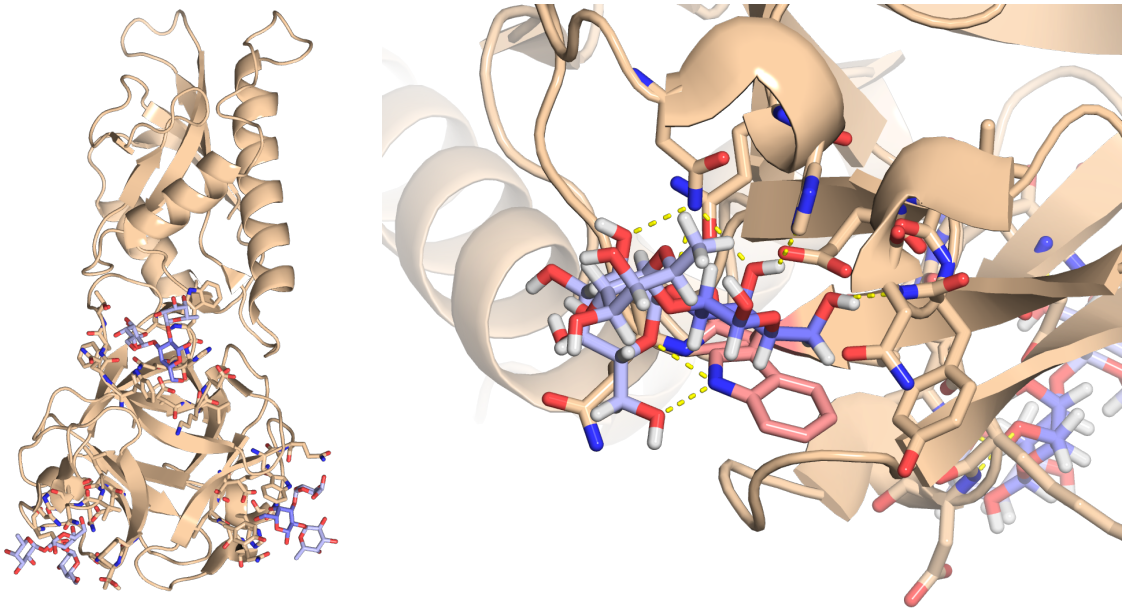

Figure S32. Visualization of the *Marasmius oreades* agglutinin lectin (PDB ID: 3EF2) (left) full protein and (right) binding pocket with CH–π stacking interaction pair highlighted. Carbon atoms are colored as follows: galactose colored in purple, other carbohydrates in light purple, the interacting tryptophan colored in salmon, and all protein residues in wheat. All other atoms, regardless of the molecule, are colored as follows, oxygen in red, nitrogen in blue, and hydrogen in white. Polar contacts as classified by PyMOL that involve the carbohydrate ligand are shown with yellow dotted lines.

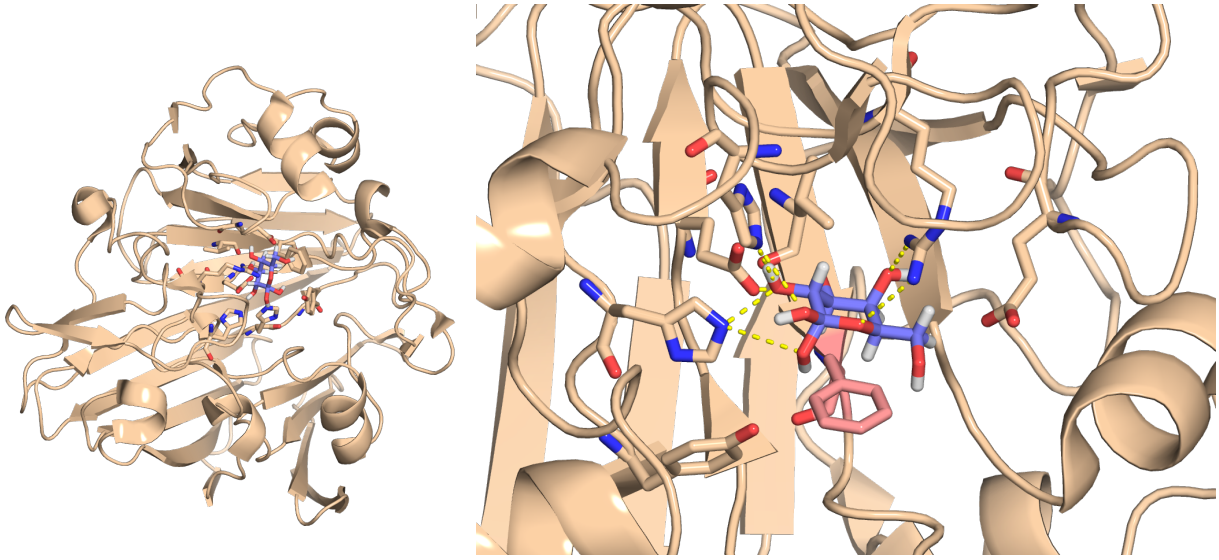

Figure S33. Visualization of the *Lactococcus lactis* galactose mutarose (PDB ID: 1NSM) (left) full protein and (right) binding pocket with CH–π stacking interaction pair highlighted. Carbon atoms are colored as follows: galactose colored in purple, the interacting phenylalanine colored in salmon, and all protein residues in wheat. All other atoms, regardless of the molecule, are colored as follows, oxygen in red, nitrogen in blue, and hydrogen in white. Polar contacts as classified by PyMOL that involve the carbohydrate ligand are shown with yellow dotted lines.

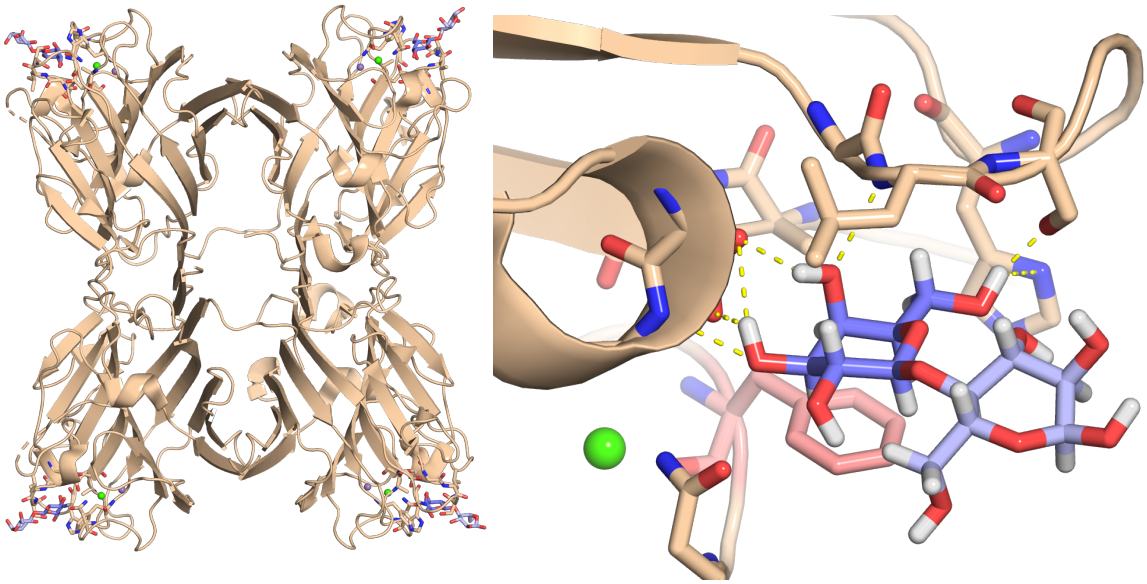

Figure S34. Visualization of the *Vatarirea macrocarpa* recombinant seed lectin (PDB ID: 4WV8) (left) full protein and (right) binding pocket with CH–π stacking interaction pair highlighted. Carbon atoms are colored as follows: galactose colored in purple, other carbohydrates in light purple, the interacting phenylalanine colored in salmon, and all protein residues in wheat. All other atoms, regardless of the molecule, are colored as follows, oxygen in red, nitrogen in blue, hydrogen in white, and chloride in green. Polar contacts as classified by PyMOL that involve the carbohydrate ligand are shown with yellow dotted lines.

References

(1) Houser, J.; Kozmon, S.; Mishra, D.; Hammerova, Z.; Wimmerova, M.; Koca, J. The Ch-Pi Interaction in Protein-Carbohydrate Binding: Bioinformatics and in Vitro Quantification. *Chem-Eur J* **2020,** *26*, 10769-10780.