Electronic Supplementary Information

For

A first-principle exploration of the conformational space of sodiated pyranose assisted by neural network potentials

Huu Trong Phan,[a,b,c] Pei-Kang Tsou,[a] Po-Jen Hsu,[a] and Jer-Lai Kuo,[a,b,c,d]

a. Institute of Atomic and Molecular Sciences, Academia Sinica, Taipei, 10617, Taiwan

b. Molecular Science and Technology Program, Taiwan International Graduate Program, Academia Sinica, Taipei, 11529, Taiwan

c. Department of Chemistry, National Tsing Hua University, Hsinchu 30013, Taiwan.

d. International Graduate Program of Molecular Science and Technology (NTU-MST), National Taiwan University, Taipei 10617, Taiwan

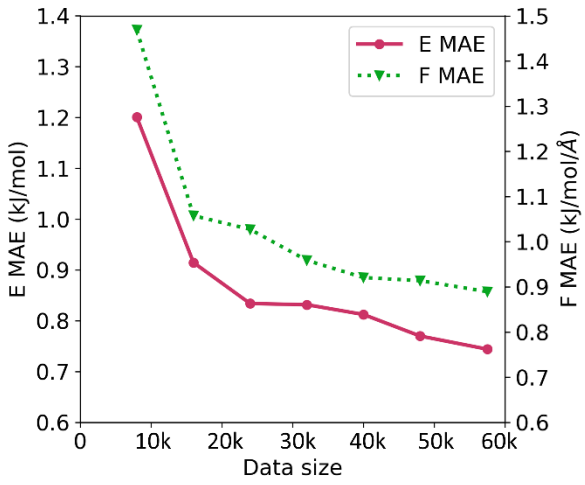

Figure S1. The benchmark of the correlation between the number of data points with the predictive performance of the obtained NNP models. Different amount of data of hexoses in the AH-0 group is used as the training sets for this benchmark.

Table S1. The details of the predictive performance of different NNP generations on the individual test set of each type of hexose (including both anomeric forms). The unit for E-MAE and F-MAE is kJ/mol and kJ/mol/Å, respectively.

		Test set size	NNP-0		NNP-1		NNP-2	
			E MAE	F MAE	E MAE	F MAE	E MAE	F MAE
AH-0	α-Glc	2138	0.87	1.11	0.81	0.98	0.87	0.96
	β-Glc	2556	0.92	1.09	0.92	0.96	0.89	0.95
	α-Gal	3015	0.97	1.20	0.98	0.97	0.98	0.95
	β-Gal	3303	0.99	1.25	0.85	1.03	0.77	1.01
	α-Man	2265	0.78	1.06	0.89	0.97	0.87	0.97
	β-Man	2439	0.67	1.06	0.71	0.99	0.72	0.99
AH-1	α-All	1594	4.37	2.74	0.90	0.99	0.88	0.97
	β-All	1463	4.41	2.51	0.82	0.95	0.80	0.92
	α-Alt	1664	2.88	2.57	0.91	1.03	0.88	1.00
	β-Alt	1565	3.06	2.50	0.91	1.01	1.00	0.98
	α-Gul	1620	2.36	2.55	0.96	1.00	0.99	0.96
	β-Gul	1615	2.11	2.55	0.86	1.03	0.85	1.00
	α-Ido	1667	2.26	2.52	0.89	1.04	0.83	1.01
	β-Ido	1713	2.04	2.55	0.86	1.02	0.85	1.00
	α-Tal	1546	2.55	2.34	0.90	1.02	0.83	0.99
	β-Tal	1754	2.72	2.50	0.85	1.10	0.86	1.06
KH	α-Fru	1998	53.49	15.83	0.99	1.40	0.82	1.08
	β-Fru	1793	53.36	15.72	1.18	1.39	1.02	1.07
	α-Psi	1576	57.56	16.41	1.01	1.43	0.88	1.10
	β-Psi	1646	56.92	16.33	1.21	1.44	1.14	1.13
	α-Tag	1758	51.13	15.84	0.95	1.41	0.83	1.10
	β-Tag	1710	55.44	15.97	1.07	1.37	0.97	1.06
	α-L-sor	1722	52.88	15.45	1.01	1.40	0.89	1.10
	β-L-sor	1714	52.19	15.73	0.90	1.41	0.83	1.17

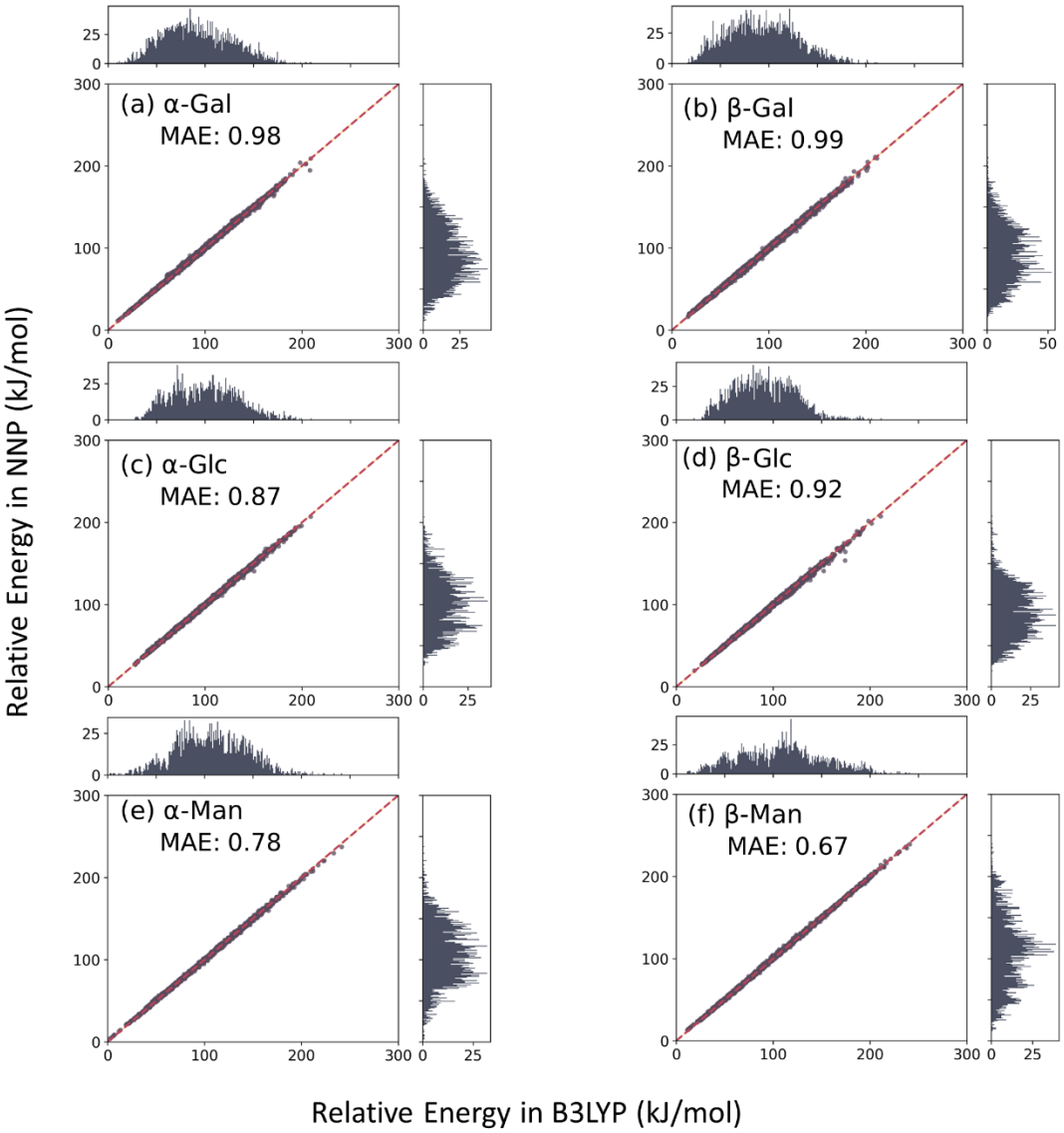

Figure S2. The correlation in relative energy between B3LYP and NNP-0 model on the AH-0 group. The zero of energy is set as the energy of the global minimum of sodiated β-Mannose (-849.574341765 Ha).

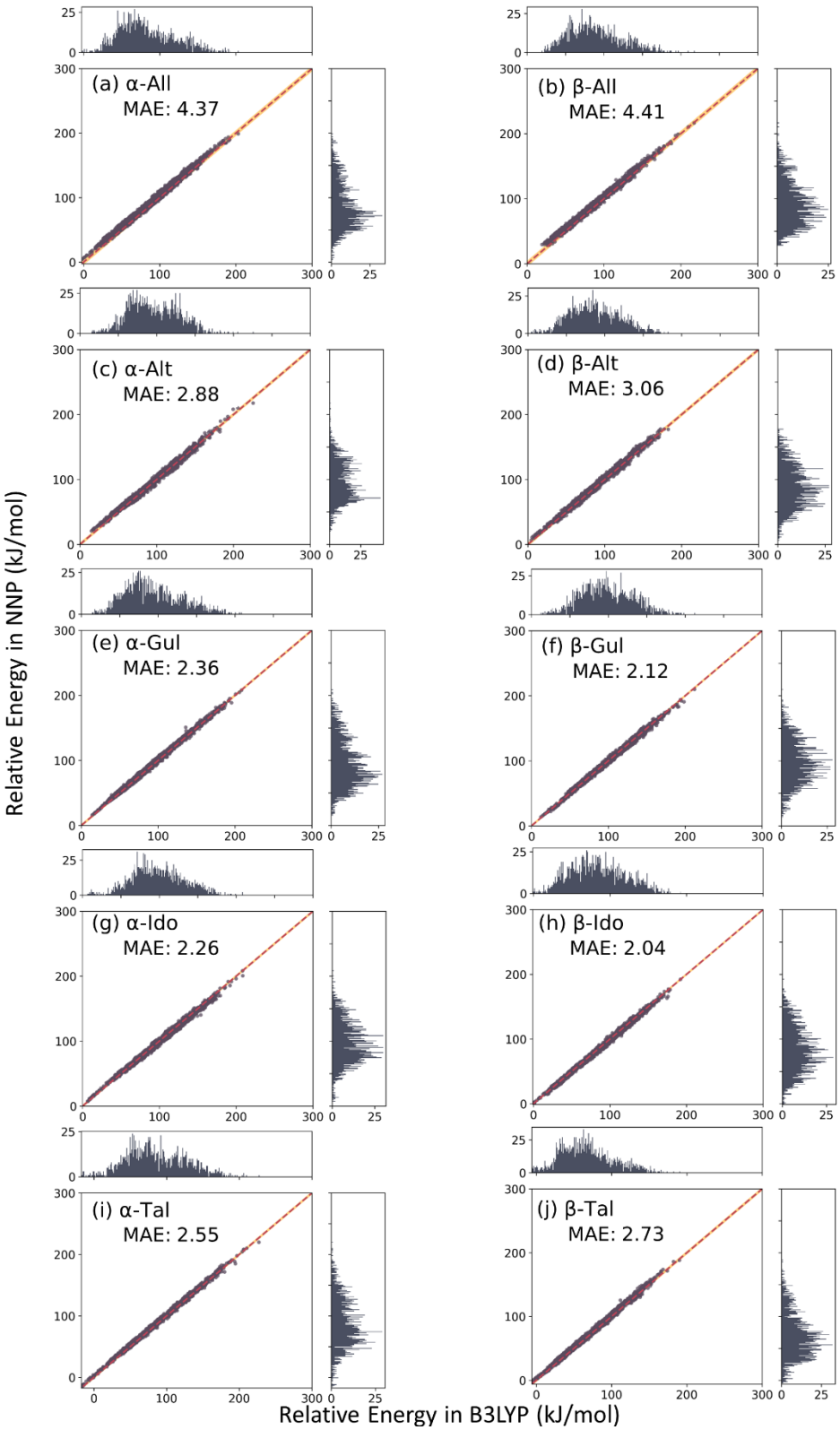

Figure S3. The correlation in relative energy between B3LYP and NNP-0 model on the AH-1 group. The zero of energy is set as the energy of the global minimum of sodiated β-Mannose (-849.574341765 Ha).

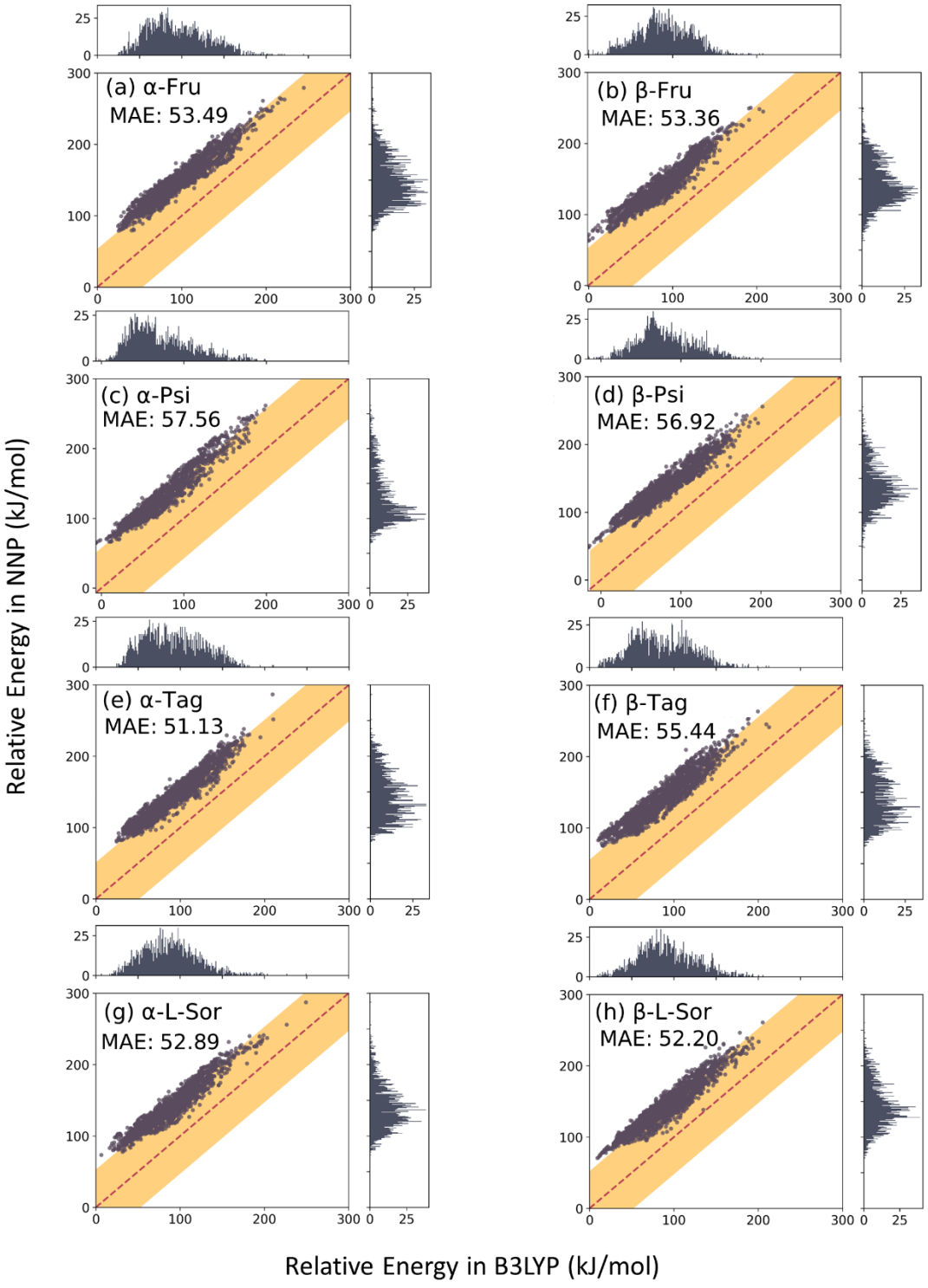

Figure S4. The correlation in relative energy between B3LYP and NNP-0 model on the KH group. The zero of energy is set as the energy of the global minimum of sodiated β-Mannose (-849.574341765 Ha).

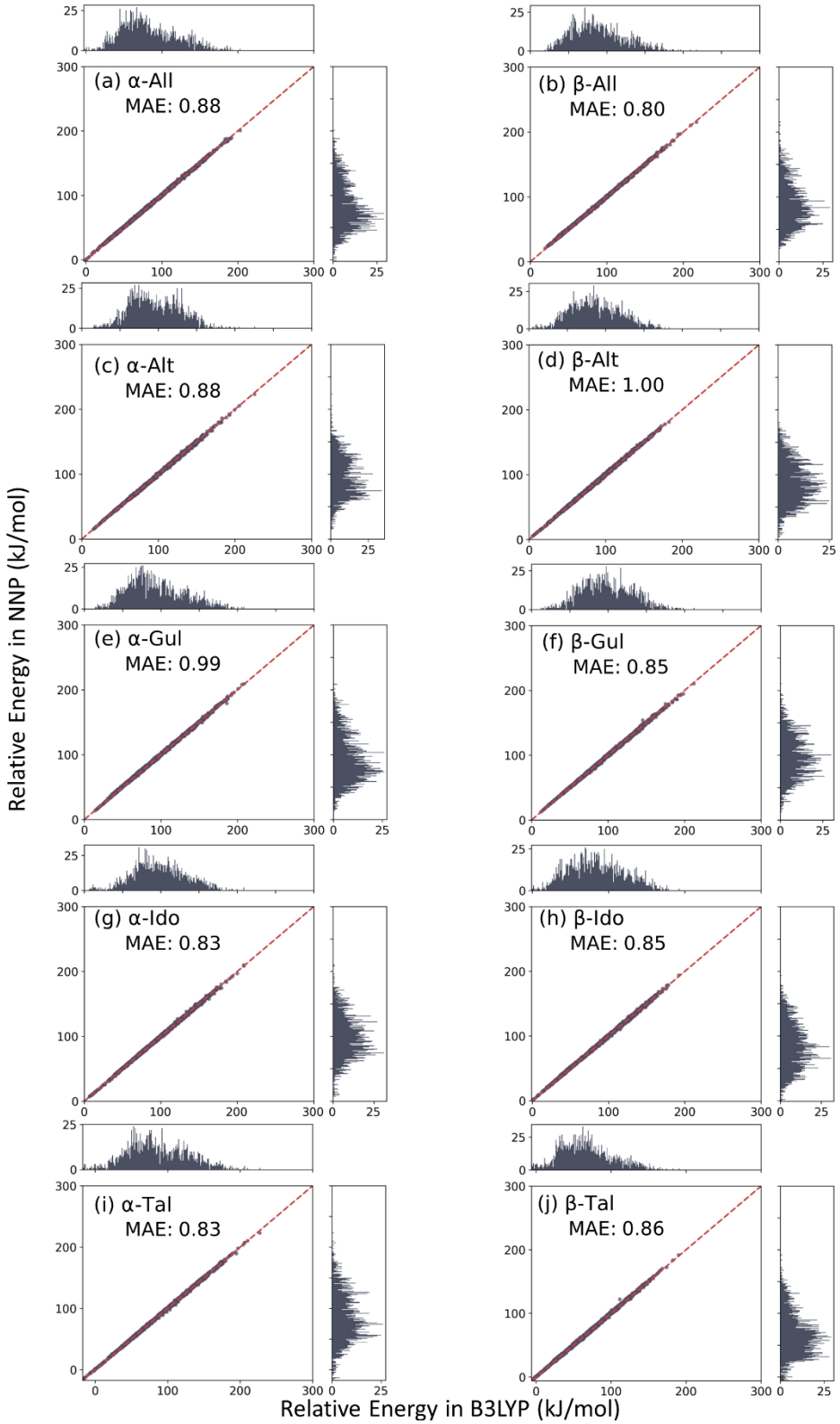

Figure S5. The correlation in relative energy between B3LYP and NNP-2 model on the AH-1 group. The zero of energy is set as the energy of the global minimum of sodiated β-Mannose (-849.574341765 Ha).

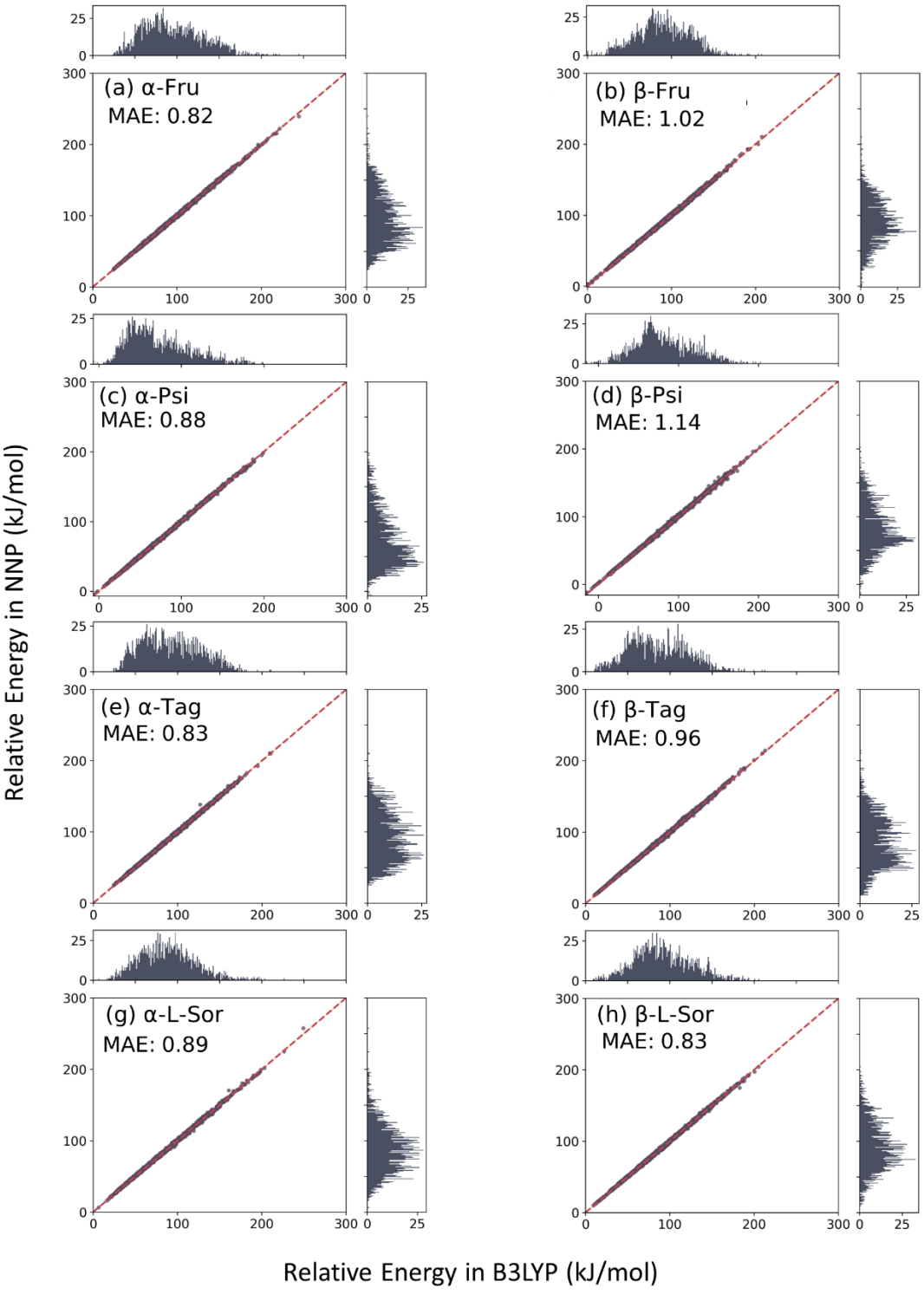

Figure S6. The correlation in relative energy between B3LYP and NNP-2 model on the KH group. The zero of energy is set as the energy of the global minimum of sodiated β-Mannose (-849.574341765 Ha).

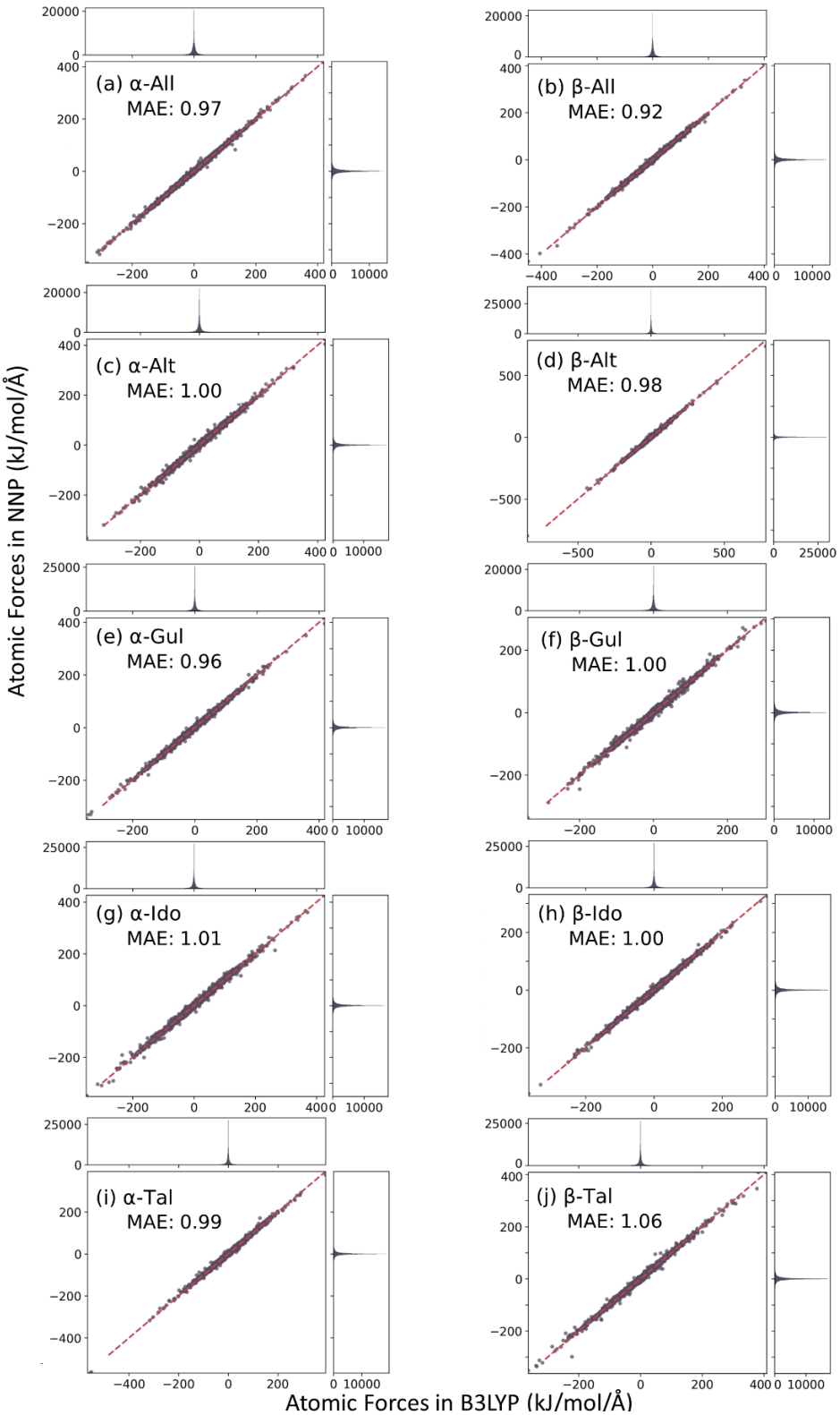

Figure S7. The correlation of the atomic force between B3LYP and NNP-2 model on the AH-1 group. The Cartesian force vector components $\boldsymbol{F_x}$, $\boldsymbol{F_y}$, $\boldsymbol{F_z}$ are included in a single plot. Each point represents a pair of force vector components $F_i^{(B3LYP)}$ and $F_i^{(NNP)}$ with $i \in (x, y, z)$.

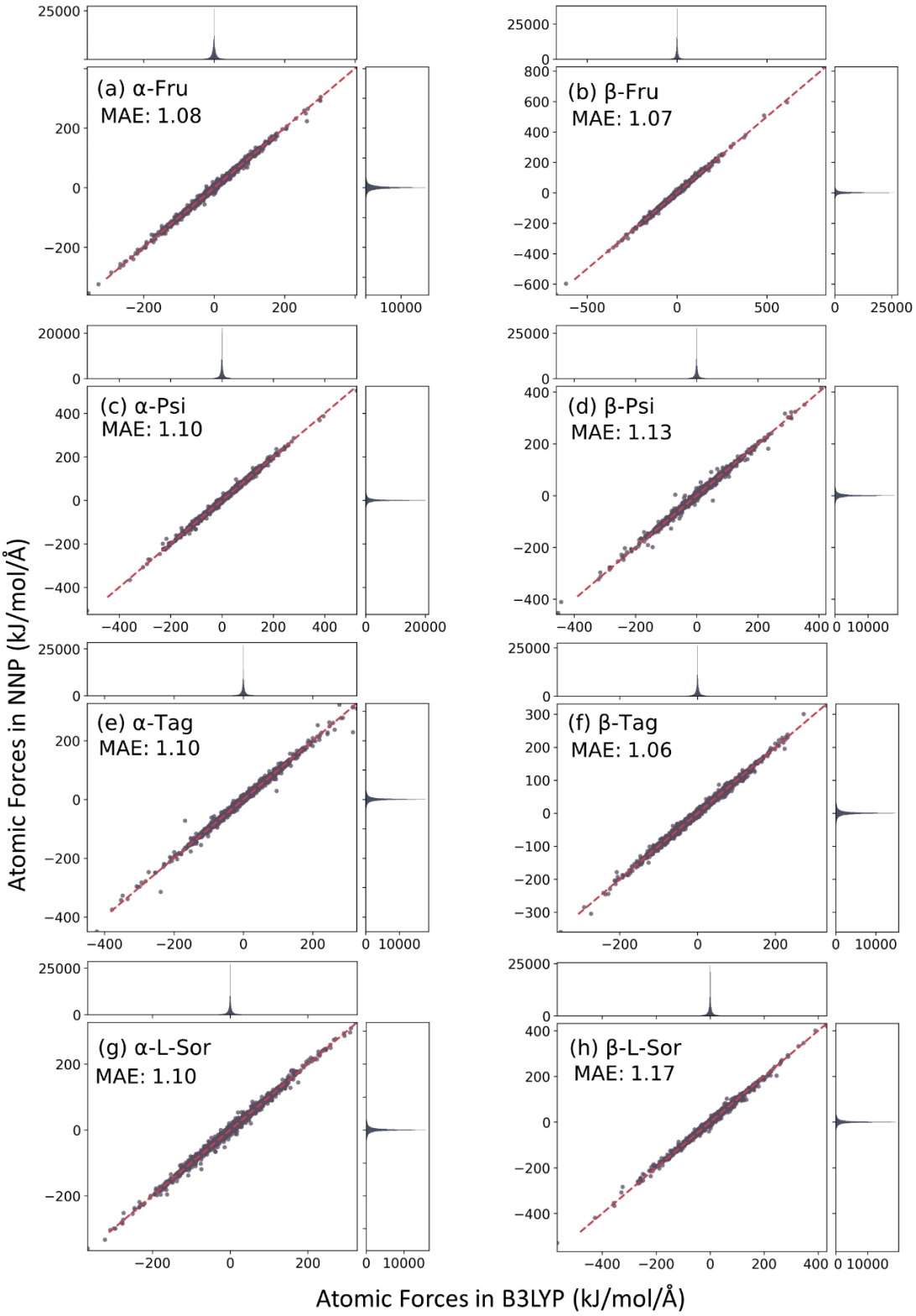

Figure S8. The correlation of the atomic forces between B3LYP and NNP-2 model on the KH group. The Cartesian force vector components $\boldsymbol{F_x}$, $\boldsymbol{F_y}$, $\boldsymbol{F_z}$ are included in a single plot. Each point represents a pair of force vector components $F_i^{(\mathrm{B3LYP})}$ and $F_i^{(\mathrm{NNP})}$ with $i \in (x, y, z)$.

Table S2 The summary of the data points created "on-the-fly" of pyranoses in AH-1 and KH when applying the Active Learning scheme. Since the NNP-1 was used to perform the geometry optimization calculations on the KH group only, there is no data created on the AH-1 so it is denoted as N/A.

Group	Types of monosaccharides	Data created by NNP-0		Data created by NNP-1	
		Initial geometries	Number of data points generated	Initial geometries	Number of data points generated
AH-1	α-All	200	1804	N/A	N/A
	β-All	200	1764	N/A	N/A
	α-Alt	200	2033	N/A	N/A
	β-Alt	200	2145	N/A	N/A
	α-Gul	200	2207	N/A	N/A
	β-Gul	200	2154	N/A	N/A
	α-Ido	200	2125	N/A	N/A
	β-Ido	200	2326	N/A	N/A
	α-Tal	200	1993	N/A	N/A
	β-Tal	200	2208	N/A	N/A
KH	α-Fru	200	2197	200	1040
	β-Fru	200	2285	200	950
	α-Psi	200	2107	200	1013
	β-Psi	200	2274	200	961
	α-Tag	200	2663	200	1038
	β-Tag	200	2327	200	867
	α-L-sor	200	2240	200	895
	β-L-sor	200	2396	200	970

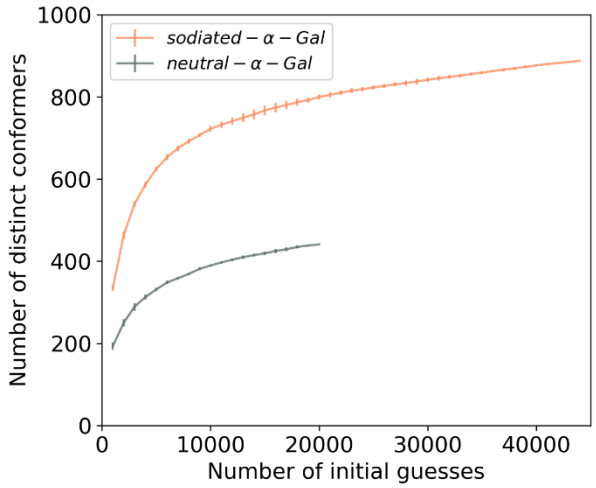

Figure S9. The correlation between the number of initial guesses and the number of distinct conformers in the structure sampling of neutral and sodiated conformers of α-Gal using DFTB3 method.[1-3]

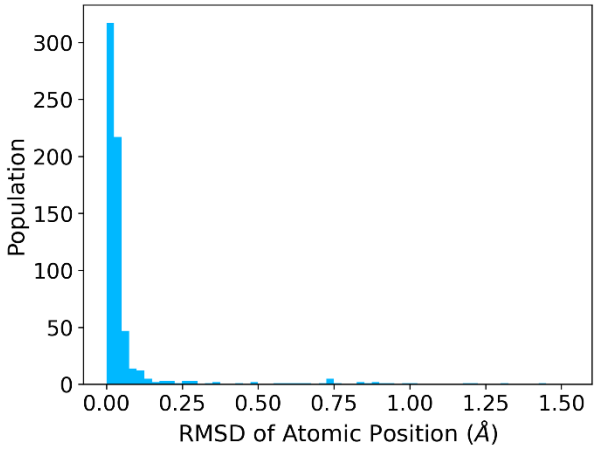

Figure S10. The histogram of the Root Mean Square Deviation of all the pairs of local minima of sodiated α-Gal, one is the NNP local minima and another is the B3LYP local minima derived from the NNP local minima by performing the re-optimization.

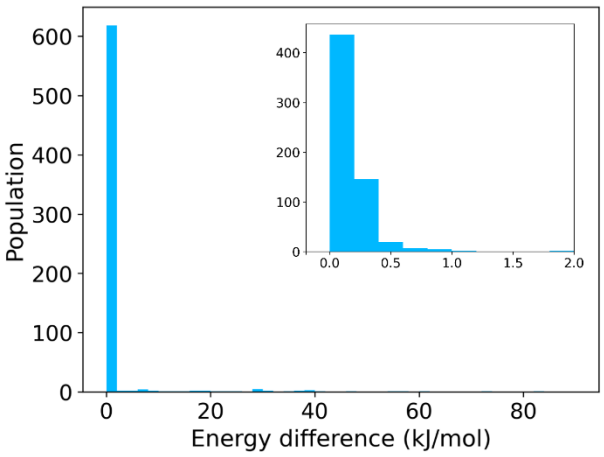

Figure S11. The histogram of the energy difference (kJ/mol) evaluated in B3LYP level between the NNP-2 and B3LYP local minima dataset of sodiated α-Gal, the energy bin width is set as 2 kJ/mol for the outer figure, and 0.2 kJ/mol for the inner figure.

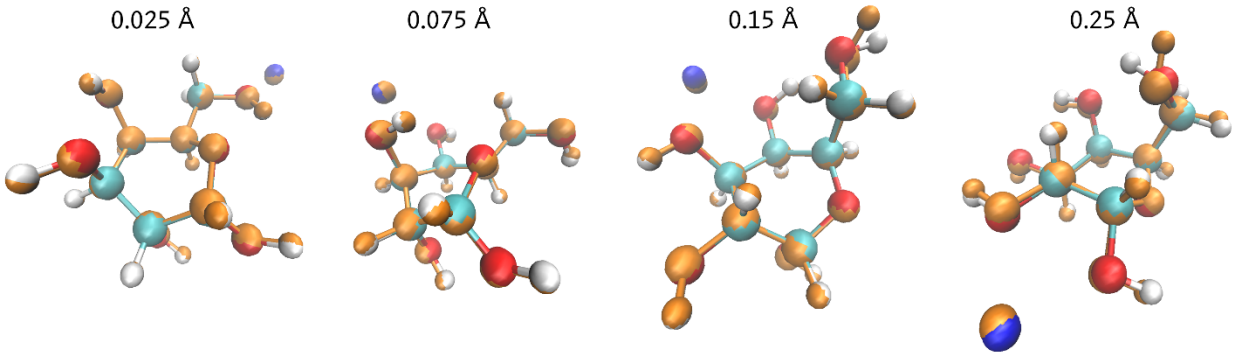

Figure S12. Visualizations of different pairs in which an NNP local minima (in orange color) superposing on the corresponding B3LYP local minima (in cyan, red and white colors). The RMSD values are shown on the top. All the geometries are the sodiated α-Gal conformers.

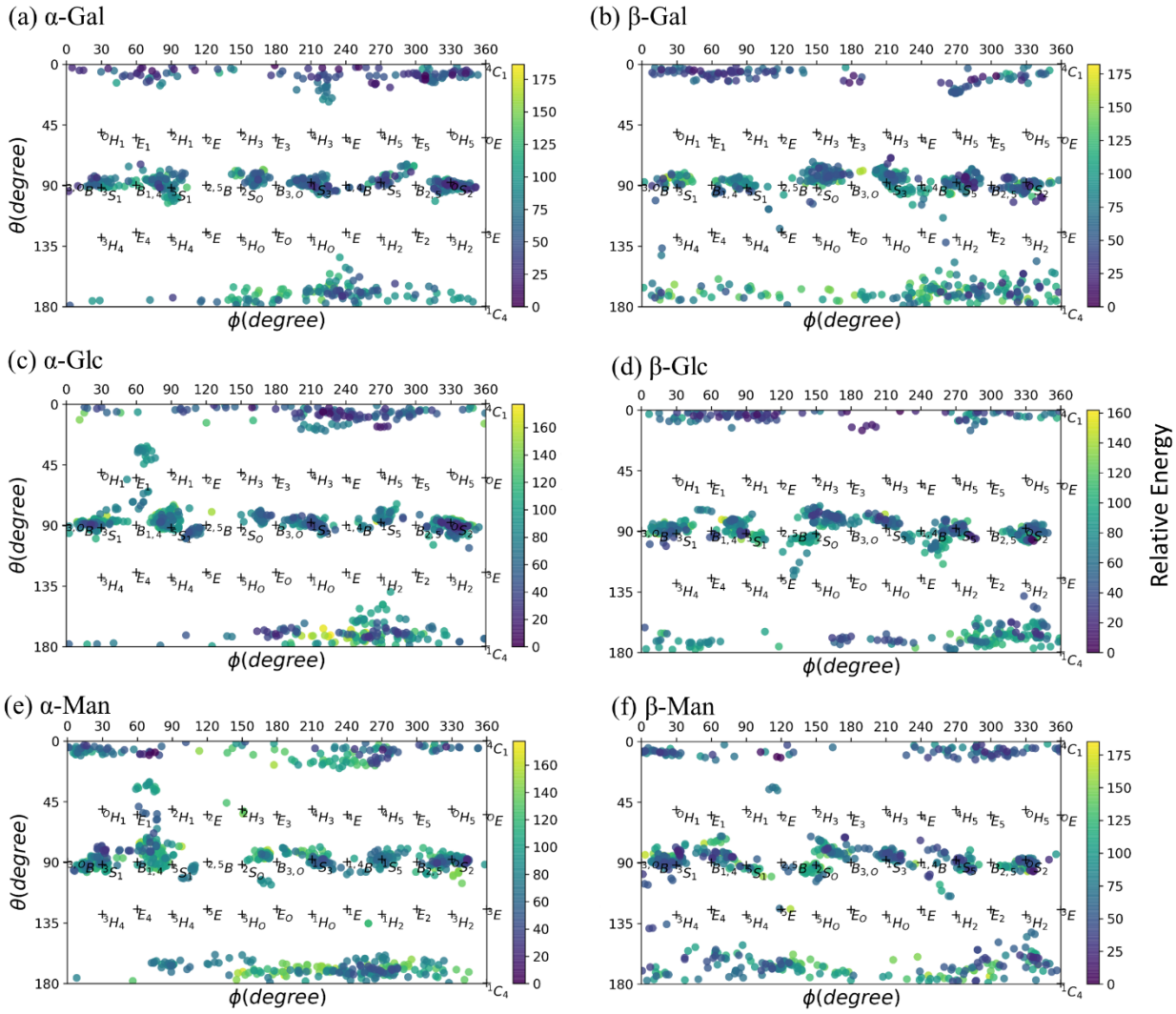

Figure S13. The Mercator projection of the NNP local minima in the AH-0 group. The colors indicate the relative energy shown in the sidebar with the zero of energy set as the global minimum energy of each type.

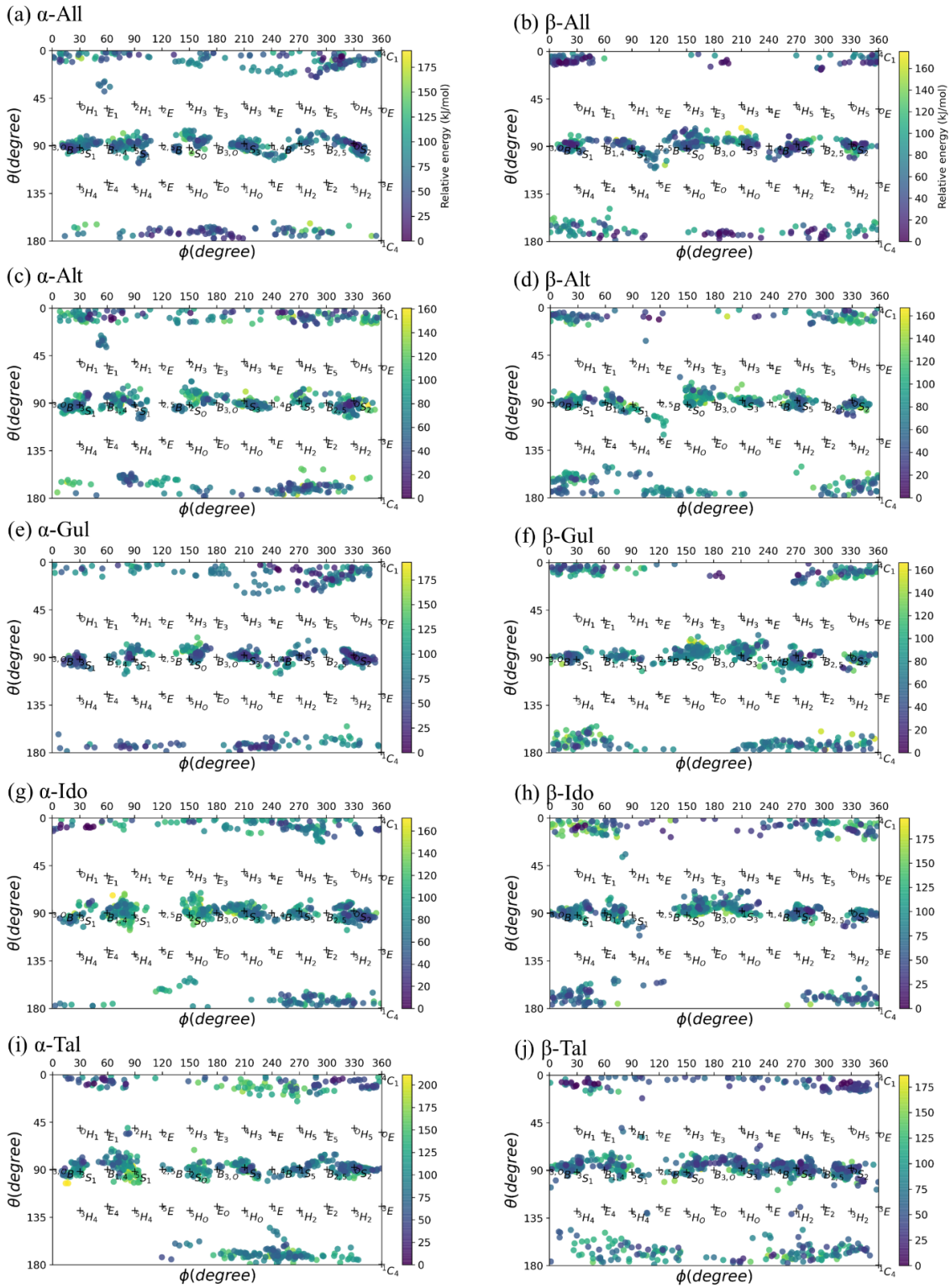

Figure S14. The Mercator projection of the NNP local minima in the AH-1 group. The colors indicate the relative energy shown in the sidebar with the zero of energy set as the global minimum energy of each type.

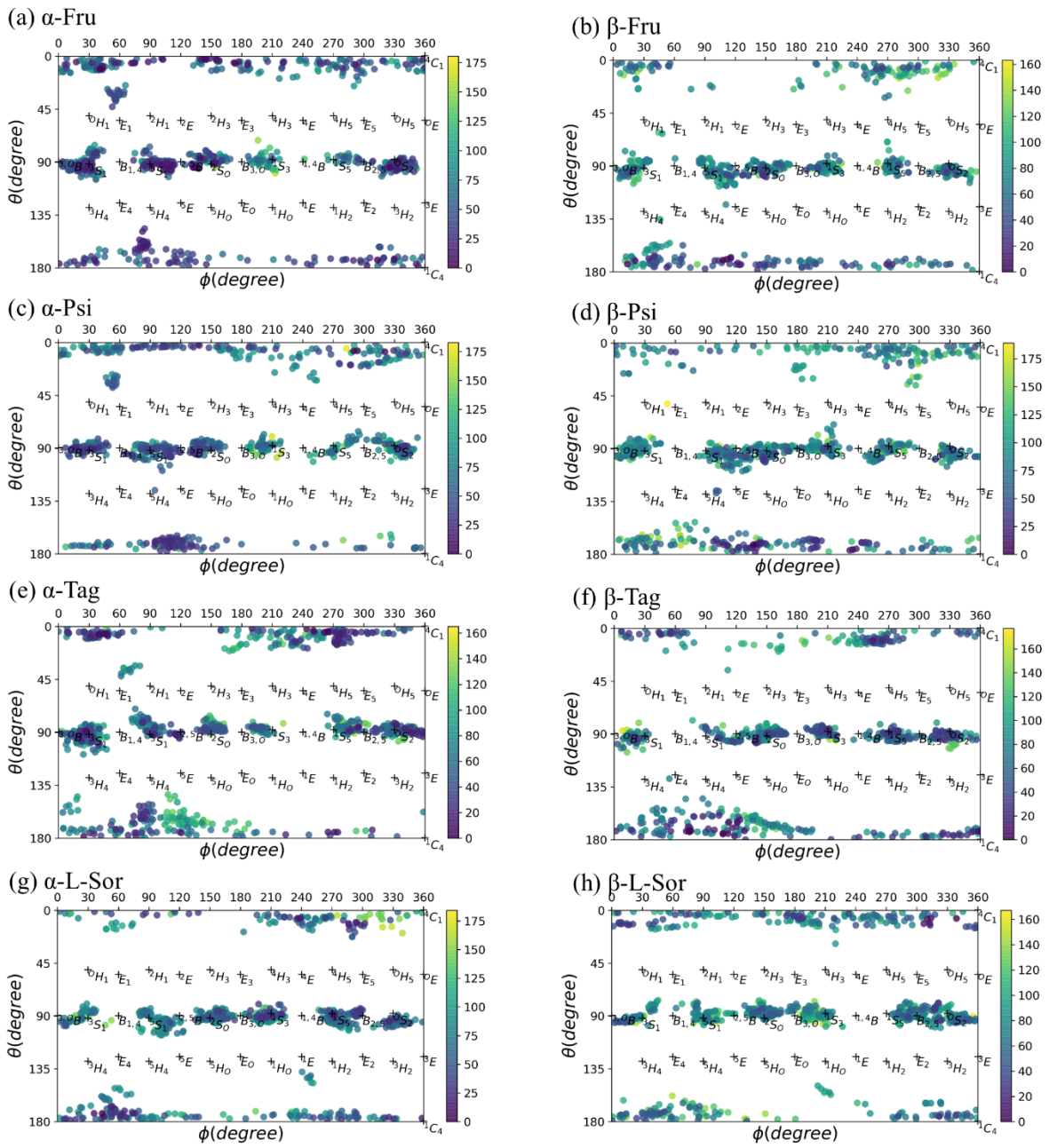

Figure S15. The Mercator projection of the NNP local minima in the KH group. The colors indicate the relative energy shown in the sidebar with the zero of energy set as the global minimum energy of each type.

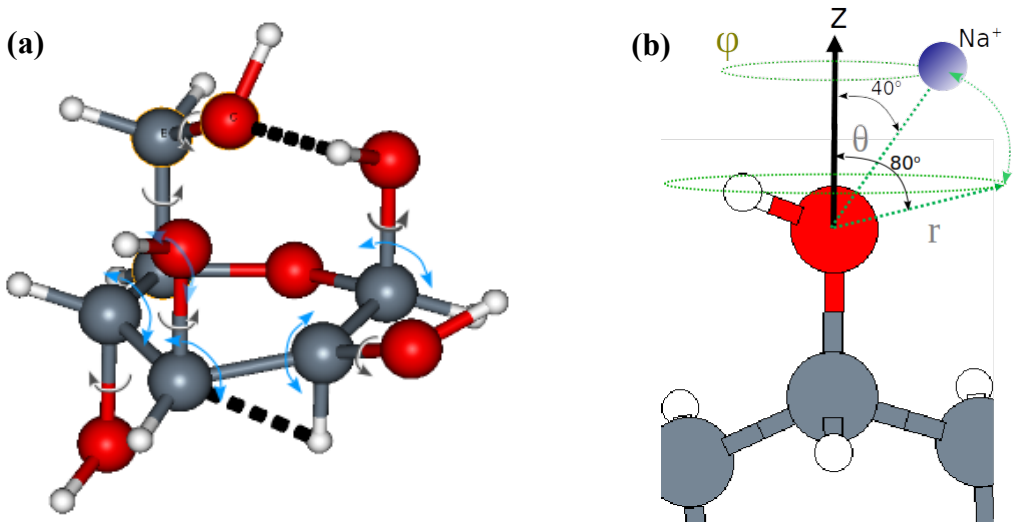

Figure S16. The visualization of the structural sampling of monosaccharides. (a) the sampling of neutral conformers with two kinds of operations, ring mutation with the atoms rotating following the blue arrows, the rotations of exo-cyclic groups following the brow arrows (b) the sampling of sodiated conformers. A set of 20 positions of placing Na^+ in the spherical coordinate with O atoms is the center, two θ angles (40°, 80°), and 10 φ angles (evenly spaced in the range of 0-360°)

Structure sampling of the neutral conformers

The ring form of monosaccharide includes the ring and the attached exocyclic groups. There are two types of operations performed to generate the initial guess of the neutral form: the ring mutation and the rotation of exo-cyclic groups.

The visualization of the sampling scheme of neutral conformers is shown in Figure S16a. In the ring mutation operation, the ring is twisted so that each of the ring atoms is rotated around the axis defined by its two neighbor atoms in the ring while the bond length is kept fixed. The rotational degree is randomly chosen with the range from -72° to 72° (normal sampling mode) or -18° to 18° (gentle sampling mode). For the exo-cyclic group, the rotation axis is defined as the C-O (-OH) or C-C bond (for $-CH_2OH$) with the rotational range is 0-360°.

We generated a set of 20000 initial guesses of neutral conformers and perform the optimization using DFTB3[1–3] method. Afterward, the screening step is performed on the obtained geometries to remove the local minima with unphysical geometries, having imaginary frequencies and duplicates by the TSCA algorithm[4] with the similarity threshold set as 0.99. The screened set of local minima is used to sample the sodiated structures.

Structure sampling of the sodiated conformers

As shown in Figure S16b, for each neutral conformer, around each O atom in the exocyclic group, 20 grid points are determined by the combination of 2 polar angles (θ) (40°, 80°) with 10 azimuthal angles (φ) (0°, 36°, 72°, 108°, 144°, 180°, 216°, 252°, 288°, 324°), and the fixed radial distance (r) is 2.4 Å. The C-O vector is set as the referenced z-axis. For the ring O, we set the C1-C6 vector as the z-axis, there are 5 azimuthal angles (0, 72°, 144°, 216°, 288°) and 1 polar angle the polar angle is set as 90°. After the initial guesses of sodiated candidates are generated, the optimization is

processed. The obtained set of conformers will then be refined in the same procedure as refining neutral conformers.

References

1 M. Gaus, Q. Cui and M. Elstner, DFTB3: Extension of the Self-Consistent-Charge Density-Functional Tight-Binding Method (SCC-DFTB), *J. Chem. Theory Comput.*, 2011, **7**, 931–948.

2 M. Gaus, A. Goez and M. Elstner, Parametrization and Benchmark of DFTB3 for Organic Molecules, *J. Chem. Theory Comput.*, 2013, **9**, 338–354.

3 M. Kubillus, T. Kubař, M. Gaus, J. Řezáč and M. Elstner, Parameterization of the DFTB3 Method for Br, Ca, Cl, F, I, K, and Na in Organic and Biological Systems, *J. Chem. Theory Comput.*, 2015, **11**, 332–342.

4 P.-J. Hsu, K.-L. Ho, S.-H. Lin and J.-L. Kuo, Exploration of hydrogen bond networks and potential energy surfaces of methanol clusters using a two-stage clustering algorithm, *Phys. Chem. Chem. Phys.*, 2016, **19**, 544–556.