PCCP

PAPER

View Article Online
View Journal | View Issue

Cite this: *Phys. Chem. Chem. Phys.*, 2025, **27**, 11780

Unraveling the low-energy conformers of neutral and charged mono- and di-saccharides with first-principles accuracy assisted by neural network potentials†

Huu Trong Phan,[abc] Pei-Kang Tsou,[a] Hieu Cao Dong,[abd] Po-Jen Hsu[a] and Jer-Lai Kuo *[abcd]

Received 1st January 2025,
Accepted 4th April 2025

DOI: 10.1039/d5cp00005j

rsc.li/pccp

We present an efficient structure sampling algorithm that combines neural network potentials (NNPs) and a well-sampled local minima dataset to explore the conformational space of mono-saccharides and di-saccharides. The structure sampling methodology leverages a "pattern transfer" approach, in which molecular initial guesses are created by utilizing existing local minima from a structurally relevant molecular system as a template. The NNP models are integrated into the structure sampling scheme to efficiently identify low-energy conformer candidates. Vibrational spectra simulated from these identified structures show qualitative alignment with experimental infrared spectra. This study demonstrates the potential of combining NNP models in an efficient structure sampling scheme to investigate flexible molecular systems, advancing understanding of carbohydrate structure–property relationships.

1. Introduction

Carbohydrates, also known as sugars, play a crucial role in a wide range of biological functions in living organisms.[1] Mono-saccharides, the building blocks of carbohydrates, and di-saccharides, which consist of two linked monomer units, have been extensively studied using various analytical techniques, such as mass spectroscopy,[2–5] rotational spectroscopy[6–8] or vibrational spectroscopy.[9–21] These techniques have proven valuable in elucidating the structural and conformational properties of mono-saccharides, which are essential for understanding their biological roles. For example, Ni *et al.* conducted collision-induced dissociation (CID) experiments to differentiate different types of carbohydrates.[3–5,22,23] Double resonance IR spectroscopy was applied to resolve the glucose and glucosamine conformations in the studies of Voss *et al.*[10] and Scutelnic *et al.*[12] Infrared multiple photon dissociation (IRMPD) was employed to study *N*-acetyl hexosamine by Tan *et al.*[13] Experimental studies on carbohydrates have made significant progress, advancing to a level where they can now handle oligosaccharides.[14] The advancements in experimental techniques have created a high demand for theoretical calculations, as such a database is essential for providing structural insights and interpreting experimental observables. However, establishing a comprehensive and accurate structure database for an extended set of mono-saccharides or di-saccharides remains a challenging task. The traditional cascade sampling scheme usually involves a multi-level approach, where increasingly accurate computational methods are applied in successive steps. Initially, empirical or semi-empirical methods, such as the third-order density functional tight-binding method (DFTB3),[24–26] are used for large-scale conformational sampling. Then, more accurate density functional theory (DFT) methods refine these preliminary structures to locate local minima. However, this scheme has limitations, including the risk of possibly losing important conformers due to inherent errors in the empirical/semi-empirical methods and the re-optimization process. To address these limitations, our approach implements a comprehensive two-stage sampling protocol, as detailed in later sections.

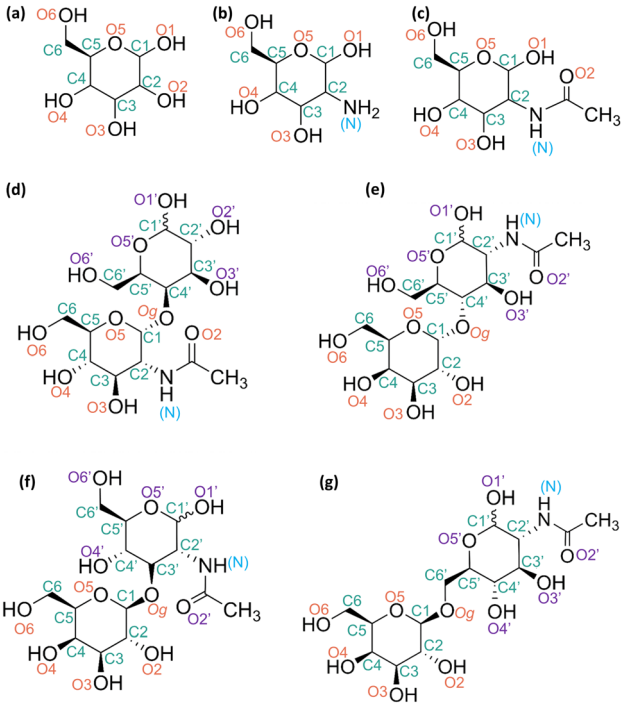
The advent of neural network potentials (NNP) has brought a significant change in the field of computational chemistry, providing an efficient and accurate alternative to quantum chemistry calculations. NNPs have been successfully employed to explore the potential energy surface (PES) in the vicinity of local minima and transition states.[27–29] In the present study, we train the NNP to describe the PES surrounding the local minima of mono-saccharides in various forms, including

[a] *Institute of Atomic and Molecular Sciences, Academia Sinica, Taipei, 10617, Taiwan. E-mail: jlkuo@pub.iams.sinica.edu.tw*
[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*

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d5cp00005j

neutral, protonated, lithiated, and sodiated species. Furthermore, we extended the description of NNPs to encompass protonated disaccharides, which are known to be more complex, highly flexible molecular systems. This extension demonstrates the versatility and robustness of NNPs in handling increasingly intricate carbohydrate structures, paving the way for more accurate and efficient conformational sampling and analysis.

Low-energy conformers are critical for the interpretation of experimental infrared (IR) spectra, as they predominantly populate thermal equilibrium states. In this work, we propose an approach for efficiently sampling the low-lying energy conformers of an extended set of mono-saccharides and di-saccharides through the novel concept of "pattern transfer". This structure sampling scheme capitalizes on the structural similarities observed across related molecular systems, particularly targeting conserved structural motifs such as hydrogen-bonding networks, ring puckering conformations, and favored dihedral angles along glycosidic bonds. Our sampling approach consists of two stages: first, initial structures are generated through "pattern transfer" and optimized using NNPs; second, a supplementary random sampling is performed on the identified local minima to explore additional regions of the conformational space and mitigate the potential limitation of overlooking low-energy conformers during pattern transfer. This two-stage structure sampling approach, facilitated by the efficiency of NNP models, ensures a comprehensive exploration of the molecular conformational space while reducing the computational cost required for traditional structure sampling schemes.



Scheme 1 (a) Schematic illustration of the numbering convention on the O (in orange and violet), C (in dark cyan), and N atoms (blue) on (a) hex, (b) hexN, (c) hexNAc, (d) α-GlcNAc-(1 → 4)-Gal, (e) α-Gal-(1 → 4)-GlcNAc, (f) β-Gal-(1 → 3)-GlcNAc, and (g) β-Gal-(1 → 6)-GlcNAc.

2. Methodology and computational details

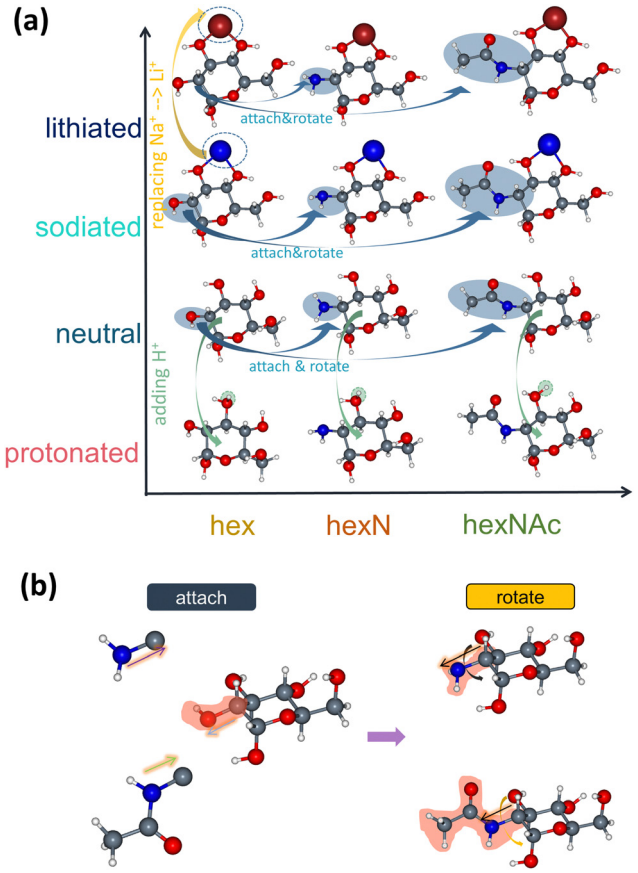
2.1. Structure sampling scheme

The numbering scheme for C, O, and N atoms on the molecular systems of hex, hexN and hexNAc, and di-saccharides follows a similar convention to that used in our previous studies,[27,29] and is depicted in Scheme 1. The N atom in the functional groups (–NH_2 or –NAc) is labelled as N, while the carbonyl oxygen atom is labelled as O2 as the –NAc moiety is attaching with the C2 position.

Similar to the sampling procedure applied in previous studies,[27,29] the structure sampling scheme to explore the conformational space consists of two stages. In the first stage, initial structures are generated using various methods based on the "pattern transfer" concept and optimized using the NNP model. These optimized structures then undergo a subsequent structure screening step to remove duplicates and unphysical local minima. In the second stage, a supplementary random sampling scheme is applied to the local minima newly obtained from the first stage to minimize the possible loss of low-energy conformers in the first stage. Specifically, all functional groups in a structure are simultaneously rotated by random angles between -90° and 90°. The generated geometries are also optimized using an NNP model. The obtained local minima from this stage are screened together with the local minima obtained in the previous stage to generate a combined set of NNP local minima. This two-stage approach ensures a thorough exploration of the conformational space, specifically targeting the search for low-energy conformers.

2.1.1. Structure sampling scheme on mono-saccharides. The structure database of neutral and sodiated forms of hexoses was extensively sampled and described in our previous work.[4,23,27] Briefly, neutral conformers were initialized *via* ring mutations and functional group rotations, resulting in 20k initial structures. These structures were optimized using the DFTB3 method, yielding approximately 500 distinct local minima. For sampling sodiated hexoses, the Na^+ ion was positioned at multiple sites: 20 grid points around the O atoms in the exocyclic groups (OH and CH_2OH) and 10 grid points around O in the ring, generating around 50–60k initial structures. This sampling scheme resulted in ~1000 distinct DFTB3 local minima of sodiated hexose. In this work, these DFTB3 local minima were further optimized using the NNP model and screened using a two-stage clustering algorithm (TSCA)[30] with a similarity threshold of 0.99 to obtain sets of NNP local minima in neutral and sodiated forms.

Scheme 2(a) illustrates various conformational sampling pathways to sample different types of mono-saccharide. The vertical direction axis shows the neutral form and different forms of cationization, where the lithiated forms are obtained by replacing Na^+ with Li^+ in the sodiated conformers, while the protonated forms are initialized from the neutral conformers. Specifically, the protonated form of hexose is generated by adding protons at specific positions around the O atoms.



Scheme 2 (a) Schematic illustration of the structure sampling scheme. The scheme follows two main directions: (1) vertically, neutral conformers are used to initialize protonated conformers, while sodiated conformers initialize lithiated conformers; (2) horizontally, hex conformers are used to sample hexN and hexNAc conformers *via* an "attach-and-rotate" algorithm. (b) Schematic illustration of the "attach-and-rotate" algorithm used to generate hexN and hexNAc conformers from hexose conformers. The algorithm involves attaching a functional group to the hexose molecule and systematically rotating the newly formed bond to sample different orientations.

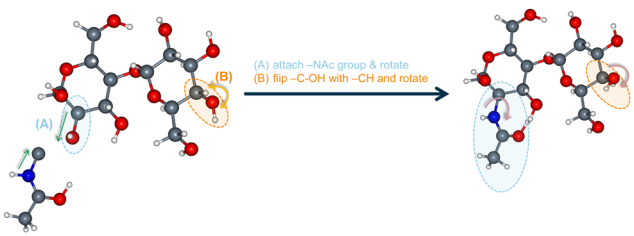
For the hydroxyl group, there are two grids for placing H^+ around hydroxyl groups which is defined by rotating O–H along the C–O bond vector by two angles (±120°). For ring oxygen atoms, the proton was placed so as to mimic the tetrahedral molecular geometry of methane, with the ring oxygen as the central atom, while C1 and C5 define two vertices. The remaining vertices are grids for place proton (with the O–H^+ length fixed as 0.98 Å). The edges of the remaining two vertices are grids for protons. For the –NH_2 group, a single proton position is defined in the plane that bisects the H–N–H angle, reflecting the pyramidal geometry of the protonated amine. For the –NAc group, both the nitrogen and carbonyl oxygen atoms are considered to place a proton. Around the N atom, two proton positions are sampled: above and below the plane formed by the N–C(O)–C atoms. For the carbonyl oxygen, five positions are evenly distributed around the O atom, with the C–O–H^+ angle fixed as 113.4°.

The horizontal axis of Scheme 2(a) shows different monosaccharide classes including hex, hexN, and hexNAc. Structures of hexN and hexNAc are created from hex local minima by the "attach-and-rotate" algorithm. This process involves replacing the OH group at the C2 position with either an –NH_2 (for hexN) or –NAc (for hexNAc) group, followed by systematic rotation of these substituents around the C–N bond at 120° (for –NH_2 group) or 60° intervals, yielding 3 or 6 new structures, respectively. Afterward, a screening step was performed to remove unphysical geometries.

2.1.2. Structure sampling on protonated di-saccharides. The conformational pattern of neutral di-saccharide (Glc-(1 → *n*)-Glc) local minima was assumed to be transferable to relevant di-saccharide types. The geometry initialization following the concept of "pattern transfer" is illustrated in Scheme 3, which later is confirmed by the similar structural pattern between the source (Glc-(1 → *n*)-Glc) molecule type to the target one (Gal-(1 → *n*)-GlcNAc). For the sequence of GlcNAc-(1 → *n*)-Gal, the source molecule type is Glc-(1 → *n*)-Gal; meanwhile, the Glc-(1 → *n*)-Glc is the source molecule for Gal-(1 → *n*)-GlcNAc. The details of the structural analysis are discussed in the "Results & discussion section". For the initialization of protonated di-saccharides, the protonation site is positioned at the carbonyl O atom in the –NAc group. This configuration is selected based on the analysis of the stability of this configuration in the protonated GlcNAc molecular system. Specifically, the analysis on the NNP local minima database reveals that this protonation site is more stable by 31.2 kJ mol^{-1} compared to other sites. The Glc to Gal transformation was performed using a "flip-and-rotate" algorithm, where the OH group at the C4 position was flipped and rotated with 3 grids (0°, 120°, 240°). The Glc to GlcNAc transformation followed the "attach-and-rotate" algorithm described earlier.

2.2. Preparing data for NNP model training

The initial dataset for training was created by optimizing the initial geometries generated from the first stage of the sampling scheme, as described earlier, at the M06-2X functional[31] and the 6-311+G(d,p) basis set.[32–35] To ensure structural diversity, the initial geometries were clustered based on their ring puckering conformation and the relative position of the cation. For the neutral form, only the puckering state was considered. From the collected set of geometry clusters, representative structures were selected through an iterative sampling scheme. In each iteration on the set of clusters, a geometry was randomly selected in each cluster, with the process repeated until accumulating an initial set of 500 geometries. This initial



Scheme 3 Transformation of a neutral disaccharide into a protonated disaccharide, with the proton assumed to bind to the O atom of the NAc group.

set of geometries was partitioned as follows: 200 structures were designated for the creation of a training set, serving as initial geometries for data generation *via* geometry optimization using either the M06-2X or NNP model. The remaining were allocated for the generation of the validation (100 structures) and test (200 structures) sets. For these two sets, data points were generated through geometry optimization using M06-2X. Subsequently, data points were selectively extracted from the optimization trajectories to minimize the redundancy of highly structurally similar geometries by collecting snapshots with energy within 100 kJ mol^{-1} relative to the corresponding local minimum, discretizing them into energy bins with a fixed bin width of 1 kJ mol^{-1}, preserving only one snapshot with the lowest energy in each bin, and further screening the extracted snapshots using the TSCA algorithm to derive a set of structurally distinct structures. When the NNP model was used for geometry optimization, the extracted data points underwent subsequent single-point calculations at the M06-2X level to obtain reference energies and forces. The creation of the validation and test sets followed a similar procedure, and the initial version of the NNP was trained on this initial training set. During the neural network training process, the trainable parameters were optimized using the data points in the training set, while the validation set was used to monitor for overfitting. The NNP model checkpoint with the lowest loss error on the validation set was selected as the final model, and its predictive performance was evaluated on the independent test set, with these results reported in our work.

2.3. Construction of vibrational spectra from quantum harmonic superposition approximation (Q-HSA) analysis

Low-energy local minima with relative energies below 25 kJ mol^{-1} are subjected to re-optimization and frequency calculations at M06-2X/6-311+G(d,p). The quantum harmonic superposition approximation (Q-HSA) method[36] was employed to evaluate the relative populations of different local minima conformers across a temperature range of 20–400 K. In this method, the set of distinct DFT conformers is treated as an ensemble of harmonic oscillators. The relative population of a conformer a is estimated as the ratio of its partition function (P_a) to the total partition function of the system. Consequently, the total intensity of the vibrational spectrum is computed as the weighted sum of all considered local minima, defined as:

$$I_{\text{total}}(\omega, T) = \sum_a I_a(\omega) P_a(T) \quad (1)$$

All the peaks are convoluted using Lorentzian functions with full width at half maximum (γ) as 5 cm^{-1}. For comparison with experimental data, temperatures of 50 K and 300 K were assumed for the IRPD and IRMPD spectra, respectively. This temperature distinction is significant as cryogenic IRPD spectra predominantly reflect the vibrational features of the global minimum structure, while room-temperature IRMPD spectra exhibit more congested features due to contributions from multiple thermally-accessible conformers. In this study, conformers with relative population exceeding 1% at the relevant temperature (50 K for IRPD or 300 K for IRMPD) are considered significantly populated. The individual spectra of these conformers were analyzed to determine their contributions to the total spectra. The simulated vibrational spectra derived from our local minima database were compared with other computational studies and experimental IR spectra to validate our database and computational approach.

2.4. Computational details

The DFT calculations were carried out at the M06-2X/6-311+G(d,p) level of theory. The M06-2X level of theory has been shown to provide accurate geometries and energies for carbohydrate systems.[37] These calculations were performed using the Gaussian 16 package.[38]

The SchNet neural network architecture[39,40] provided by the schnetpack library[41] was employed to construct the NNP model. The architecture initializes atomic features based on atom types and encodes chemical environment information through the Gaussian basis expansion of interatomic distances. The atomic feature dimension was set to 128, with 4 interaction layers, and the cutoff distance of 15 Å. The basis set comprised 75 Gaussian functions. In each interaction layer, the atomic features are refined through continuous-filter convolutions that incorporate influences of the chemical environment, followed by processing through feed-forward neural networks to generate atom-wise energy contributions. These contributions are summed to estimate the total molecular energy. The training process starts at an initial learning rate of 5×10^{-4}, regulated by a cosine annealing warm restart scheduler (T_0 = 75, T_{mult} = 1). Training was conducted over 1000 epochs with a batch size of 32, and the process was repeated several times to ensure convergence. The tradeoff parameter ρ, which controls the ratio of mean squared errors of energy and gradient in the loss function, was set to 0.01.

3. Results & discussion

3.1. The predictive performance of the NNP model

3.1.1. The NNP model for mono-saccharides. The initial NNP model was trained on a dataset derived from geometry optimizations conducted at the M06-2X level of theory. The training set comprised 2k–3k data points for each mono-saccharide type. The specific distributions of the data points in the training set, test set, and validation set are detailed in Tables S1–S4 in the ESI.† Each data point consists of atomic coordinates, total energy, and atomic forces extracted from optimization trajectories following the procedure described in the Methodology section. It is important to note that a single NNP model was trained to describe all considered molecular systems simultaneously rather than training separate models for each system. To achieve this, we combined the training sets from all individual mono-saccharide systems into one unified training set, and similarly combined all respective validation sets into a single validation set.

The predictive performance on the test sets of the obtained model, named NNP-mono, trained on the initial training set, is

Table 1 Performance evaluation of the NNP-mono model for mono-saccharides. The paired values represent E-MAE (kJ mol^{-1}) and F-MAE (kJ mol^{-1} Å^{-1}), respectively

Type	Neutral	Protonated	Lithiated	Sodiated
hex	0.85/0.97	1.22/1.52	1.09/1.18	0.92/0.99
hexN	1.12/1.11	1.19/1.47	1.17/1.22	1.09/1.13
hexNAc	1.13/1.12	1.58/1.70	1.27/1.27	1.20/1.19

summarized in Table 1 for the combined test set representing each class mono-saccharide class and in Table S5 (ESI†) for individual test sets. The model demonstrated consistent accuracy across neutral, protonated, lithiated, and sodiated conformers, with energy mean absolute errors (E-MAE) ranging from 0.8 to 1.8 kJ mol^{-1} and force mean absolute errors (F-MAE) ranging from 1.0 to 1.8 kJ mol^{-1} Å^{-1}. These results indicate that NNP-mono effectively captures both the conformational landscape of monosaccharides as well as ion–molecule interactions.

3.1.2. The NNP model for protonated di-saccharides. The data generation process started with α-Gal-(1 → 4)-α/β-GlcNAc and its reverse sequence, α-GlcNAc-(1 → 4)-α/β-Gal to generate the initial version of NNP. For each anomeric form of each sequence, 200 initial structures underwent M06-2X geometry optimizations. This procedure yielded approximately 3k data points per anomeric form, resulting in a total of ∼12k data points. To enhance the predictive performance of the NNP model, data points from relevant molecular systems collected from the NNP training data (NNP-mono in this work and other NNPs in previous studies),[27,29] including neutral α/β-Gal (∼2k points), neutral α-Glc-(1 → 4)-α/β-Glc (∼4k points), and protonated α/β-GlcNAc (∼5.5k points), were also included, resulting in an initial training set of ∼22.8k data points. The trainable weights and biases of the initial NNP model were initialized using those from the pre-trained mono-saccharide model (NNP-mono) discussed earlier, leveraging the chemical knowledge already captured in the mono-saccharide systems. The obtained model, named NNP-di-0, exhibited predictive performance on the test set of α-Gal-(1 → 4)-α/β-GlcNAc and α-GlcNAc-(1 → 4)-α/β-Gal with mean absolute errors of potential energy (E-MAE) ranging from 3.1–4.0 and 4.7–5.0 kJ mol^{-1}, and the mean absolute errors of atomic forces (F-MAE) are quite consistent ranging from 2.6–2.9 kJ mol^{-1} Å^{-1}, respectively. Subsequently, as part of generating data points for the remaining types β-Gal-(1 → *n*)-α/β-GlcNAc (*n* = 3, 6), the NNP-di-0 model was employed to perform geometry optimizations on a batch of 200 initial geometries for each type. An extracted set of data points, obtained following a procedure similar to that described above, was subjected to single-point calculations at the M06-2X level of theory. Selected data points from the NNP-di-0 optimizations that exhibited large deviations (absolute errors of energy greater than 1.5 kJ mol^{-1} and mean absolute errors for atomic forces greater than 1.5 kJ mol^{-1} Å^{-1}) from the DFT data were added to the training set. For each type, an average of 3k data points were appended to the training set. Additionally, ∼3k neutral di-saccharide data points for each of β-Glc-(1 → *n*)-α/β-Glc (*n* = 3, 6), representing (1 → 3), (1 → 6) linkages, were included. As a result, an additional ∼18.7k data points were added to the training set, making up a total of ∼42k data points. The details of the number of data points in the training set, test set and validation set for each type of protonated di-saccharide are enlisted in Table S15 (ESI†).

Table 2 Predictive performance of NNP-di-1 on individual test sets of each type of protonated di-saccharide. The units of E-MAE and F-MAE are kJ mol^{-1} and kJ mol^{-1} Å^{-1}, respectively

Protonated di-saccharides	E-MAE	F-MAE
α-GlcNAc-(1 → 4)-α-Gal	2.87	2.04
α-GlcNAc-(1 → 4)-β-Gal	2.75	2.08
α-Gal-(1 → 4)-α-GlcNAc	2.56	1.90
α-Gal-(1 → 4)-β-GlcNAc	2.35	1.94
β-Gal-(1 → 3)-α-GlcNAc	3.30	2.16
β-Gal-(1 → 3)-β-GlcNAc	3.12	2.11
β-Gal-(1 → 6)-α-GlcNAc	3.74	2.16
β-Gal-(1 → 6)-β-GlcNAc	3.41	2.26

The newly generated NNP model, labeled NNP-di-1, generally exhibited a significant improvement in predictive performance on α-Gal-(1 → 4)-α/β-GlcNAc and α-GlcNAc-(1 → 4)-α/β-Gal compared to NNP-di-0. Specifically, the errors among 4 types are consistent with E-MAEs ranging from 2.6–2.9 (kJ mol^{-1}) and F-MAEs ranging from 1.9–2.1 kJ mol^{-1} Å^{-1} (Table 2). Moreover, the prediction capability of NNP-di-1 on the newly added types, β-Gal-(1 → *n*)-α/β-GlcNAc (*n* = 3, 6), was also comparable to that of the aforementioned types. Specifically, the E-MAE ranged from 2.8–3.7 kJ mol^{-1}, while the F-MAE was 2.1–2.3 kJ mol^{-1} Å^{-1} for these types. Given this level of accuracy, NNP-di-1 was employed in the structure sampling process to identify local minimum structures. The transferability of the NNP model was validated through its systematic extension to di-saccharide systems. The trainable parameters of the model were initialized from a pre-trained mono-saccharide NNP-mono and subsequently refined using a training set comprising both mono- and di-saccharides. The successful transfer of learned features from mono- to di-saccharides suggests the potential scalability of this approach to more complex carbohydrate systems while maintaining first-principles accuracy.

3.2. Employment of the NNP models to identify low-energy conformers

3.2.1. The structure search of mono-saccharides. The NNP-mono model was deployed to perform geometry optimization. The obtained set local minima derived by this NNP model were concisely called NNP local minima in this study. The structure sampling results for identifying low-energy local minima of neutral, protonated, lithiated, and sodiated conformers are detailed in Tables S6–S9 (ESI†). A two-stage sampling approach was employed, with the initial structures in the first stage being either DFTB3 local minima (for neutral and sodiated hexoses) or geometries created from the attach-and-rotate, cation replacement scheme or grid sampling of a proton (Scheme 2). The second stage involved a random sampling as a supplementary step. In the first stage, re-optimization from DFTB3 minima for neutral hexoses yielded a success rate (ratio of distinct local minima to initial guesses) of 78%, while it was significantly

lower (54%) for sodiated hexoses, on average. The second stage sampling added approximately 200% more local minima in both neutral and sodiated forms. For other forms, the second stage random sampling only contributed an additional 30–50% of new NNP local minima. The numbers of structures and local minima generated at each step are compiled in Tables S6–S9 (ESI†).

A NNP model (NNP-mono) was applied to sample the local minima of mono-saccharides, with NNP integrated in the geometry optimization calculations. From the dataset of local minima at the NNP level, candidates with relative energies less than 25 kJ mol^{-1} were collected and re-optimized using M06-2X to localize the actual minima. The obtained set of DFT minima underwent a TSCA screening step to remove any possible duplicates. On average, the success rate of identifying distinct M06-2X minima through this re-optimization process was more than 80%. The high success rate in reproducing DFT-level minima, coupled with computational efficiency increase of at least 5000 times faster, demonstrates the potential of this approach for expediting the discovery of low-energy conformers in carbohydrate molecular systems.

3.2.2. The structure search of protonated di-saccharides. The generation of initial guesses began with the NNP local minima database of neutral di-saccharides of the form Glc-(1 → *n*)-Glc and Glc-(1 → *n*)-Gal. The carbon chain backbone of the (1 → *n*) (*n* = 3, 4, 6) glycosidic linkages served as a template to generate the corresponding Gal-(1 → *n*)-GlcNAc and GlcNAc-(1 → *n*)-Gal structures. The concept of "pattern transfer" was applied based on the structural similarity between the source and target molecules. As shown in Fig. 1, the dihedral angle distributions of NNP local minima exhibit high similarity between the source (left panel, α-Glc-(1 → 4)-α-Glc) and target molecular systems (right panel, α-Gal-(1 → 4)-α-GlcNAc), even when considering the distribution of low-energy local minima. This remarkable conformational correspondence suggests that the energetic landscapes of di-saccharides with similar backbone share common features, despite differences in their substituent groups. The structure database of neutral di-saccharides contains around 50k local minima with relative energies up to approximately 150 kJ mol^{-1}. Local minima with relative energies within 60 kJ mol^{-1} were extracted, yielding around 6k–20k structures, which were subsequently transformed into GlcNAc-(1 → *n*)-Gal and Gal-(1 → *n*)-GlcNAc configurations. These geometries underwent geometry optimization by the most updated NNP model on di-saccharides (NNP-di-1) and screening to obtain an initial set of NNP local minima. Following a scheme similar to that applied in sampling structures of mono-saccharides, a supplementary random sampling step was performed on the obtained set of local minima, followed by a TSCA screening step on a combined set of the local minima from the previous stage. Statistical details of geometries and local minima generated at each step are presented in Table 3. The number of distinct NNP local minima ranged from around 10k to 60k. Re-optimization of the selected set of low-energy local minima observed an average success rate of ~86%, with the number of distinct M06-2X local minima ranging from 26 to 157. This outcome is deemed sufficiently reasonable for the application of an NNP model in a production task.

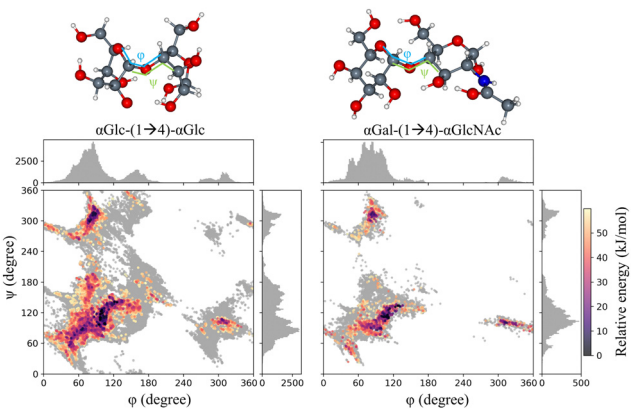


Fig. 1 The pattern of dihedral angles (ϕ,ψ) along the glycosidic bonds of the source local minima (left panel, α-Glc-(1→ 4)-α-Glc) and the target local minima (right panel, α-Gal-(1→ 4)-α-Glc). The ϕ and ψ angle is defined as the dihedral angle along O5-C1-Og-C4′ and C1-Og-C4′-C3′, respectively. The local minima with relative energy larger than 60 kJ mol^{-1} are shown in gray color, while local minima below 60 kJ mol^{-1} are shown in other colors.

3.3. Rotational constant analysis of neutral structures

The geometries of neutral conformers have been thoroughly determined through the analysis of rotational spectra. Alonso *et al.* analyzed the geometries of low-energy conformers of glucose (Glc),[42] galactose (Gal),[43] glucosamine (GlcN),[44] and *N*-acetyl galactosamine (GalNAc),[45] reporting a high level of agreement between *ab initio* calculations and experimental results. Our structure sampling scheme successfully reproduced all the low-energy conformers reported by Alonso *et al.*, as referenced in Table S10 (ESI†), and also identified one additional low-energy conformer of α-GlcN with a relative zero-point corrected energy (E_{zpe}) of +4.77 kJ mol^{-1}. The low-energy neutral local minima identified at the M06-2X/6-311+G(d,p) level in our study were further re-optimized at the MP2/6-311++G(d,p) level[46] to facilitate a fair comparison. The successful reproduction of the low-energy conformers previously reported in the literature, along with the discovery of a new low-energy conformer, validates the reliability of our structure sampling approach and provides increased confidence to explore a wider conformational space.

3.4. Results of the simulation of IR spectra

3.4.1. Molecular systems of protonated hexN and hexNAc. Our structure search on protonated hexN identified from 1.2–1.6k distinct NNP local minima ("NNP minima (2)" in Table S7, ESI†). The protonated hexN system has been thoroughly investigated both theoretically and experimentally in several studies.[12,15,16] Scutenic *et al.* employed IR–IR double resonance to elucidate the conformational spectra,[12] while Compagnon *et al.* analyzed IRMPD spectra.[15,17,18] These studies reached a consensus on the identified low-energy conformers and established a reasonable link between theoretical calculations and

Table 3 Summary of the number of structures obtained at each step in the structure sampling scheme of protonated di-saccharides. The "NNP minima (1)" were obtained from optimization using the NNP-di-1 model. The "NNP minima (2)" encompass both "NNP minima (1)" and the additional NNP local minima obtained from the supplementary random sampling step

Type (protonated)	Physical initial guesses	NNP minima (1)	Random sampled structures	NNP minima (2)	NNP minima with relative energy below 25 kJ mol^{-1}	M06-2X local minima
α-GlcNAc-(1 → 4)-α-Gal	28 604	4611	24 487	12 421	37	33
α-GlcNAc-(1 → 4)-β-Gal	108 941	19 519	29 509	27 984	50	35
α-Gal-(1 → 4)-α-GlcNAc	314 289	23 783	69 498	35 391	155	131
α-Gal-(1 → 4)-β-GlcNAc	234 909	18 594	106 615	39 737	177	157
β-Gal-(1 → 3)-α-GlcNAc	107 768	18 371	123 845	55 929	64	53
β-Gal-(1 → 3)-β-GlcNAc	156 810	22 387	73 139	50 817	29	26
β-Gal-(1 → 6)-α-GlcNAc	162 731	24 092	146 281	61 453	77	67
β-Gal-(1 → 6)-β-GlcNAc	129 083	20 118	99 062	51 413	72	66

experimental bands. In the present study, the NNP local minima established from our structure sampling scheme were compared with those found in Compagnon *et al.*'s work.[15] To facilitate a direct comparison with the work of Compagnon *et al.*, M06-2X/6-311+G(d,p) local minima were reoptimized at the CAM-B3LYP/6-311++G(d,p) level of theory[47] and evaluated the energy of the obtained local minima using MP2/6-311++G(3df,3pd) single-point calculation. The re-optimization of structures from our database successfully reproduced the majority of low-energy conformers found by Compagnon *et al.*, while also revealing several new conformers. For example, a new second-lowest energy conformer was identified for protonated β-GlcN, with a relative energy of +4.94 kJ mol^{-1} at the MP2 level. The details of conformers found in this re-optimization process are shown in Table S11 (ESI†). The inclusion of these newly found conformers potentially enhances the assignment of experimental bands.

Similarly, we conducted a thorough structural sampling of *N*-acetyl hexosamine (hexNAc) considering different protonation sites. Our results indicate that protonation at the carbonyl oxygen is the most energetically favorable. For instance, in α-GlcNAc, the 36 most stable conformers feature a proton attached to the carbonyl oxygen. According to the relative energies predicted by our NNP approach, the lowest energy conformer which features the protonation site at the amino N atom is +31.20 kJ mol^{-1} higher in energy (^{3,O}B conformation). The NNP local minima were transformed into a methylated form at the anomeric hydroxyl group and underwent optimization at CAM-B3LYP for a direct comparison with Compagnon *et al.*'s study.[19] The results show that the reoptimization at the CAM-B3LYP level added more low-energy conformers compared to the previous study (Table S12 (ESI†)).

3.4.2. Molecular systems of lithiated and sodiated hexNAc. Lithiated and sodiated mono-saccharide molecular systems exhibit more structural complexity compared to their protonated counterparts, as the cations can coordinate at various sites and multiple hydroxyl groups simultaneously. Previous studies have investigated the IRMPD spectra of these systems in both low and high-frequency regions. Contreras *et al.* explored the conformational preferences of methylated hexNAc using IRMPD in the low-frequency range (900–1800 cm^{-1}).[20] In a separate study, Tan *et al*. examined the IRMPD spectra in the frequency range from 3400–3750 cm^{-1}.[13] To compare our results with Tan *et al.*'s finding, reoptimization calculations were conducted at the B3LYP[48–50]/6-31+G(d,p) level of theory. Notably, our sampling scheme discovered new global minimum for lithiated α/β-GlcNAc and β-GalNAc, with energies 7.8, 8.8, and 8.3 kJ mol^{-1} lower, respectively, than the previously reported values (Table S13, ESI†). Moreover, many other low-energy conformers were discovered, including those exhibiting boat (B) and skew (S) conformations. These findings challenge the assumptions made in Tan *et al.*'s study, which were limited with ^{4}C$_1$ conformation. To compare our findings with the work of Contreras *et al.*, the anomeric hydroxyl group was replaced with a methyl group and optimizations were performed at the B3LYP/6-311+G(d,p) level, consistent with their study. Once again, we discovered new global minima for lithiated β-GlcNAc-OMe, α-GalNAc-OMe, and β-GalNAc-OMe, with energies lower by 11.65, 25.35, and 22.90 kJ mol^{-1}, respectively (Table S14 (ESI†)). These results highlight the significance of comprehensive conformational sampling when studying the structures and energetics of lithiated mono-saccharide systems, as previous studies may have overlooked crucial low-energy conformers.

To date, there have been no theoretical calculations on the structure sampling of sodiated hexNAc, despite the availability of experimental IR spectra reported by Martens *et al.*[21] and Pellegrinelli *et al*.[14] In this section, we provide structural insights that can contribute to the spectral features discussed in these literature reports.

The Q-HSA analysis, based on the low-energy conformers identified by the NNP model and reoptimized at the M06-2X level of theory, reveals that α-anomer conformers dominate throughout the investigated temperature range in both GlcNAc and GalNAc (Fig. S1, ESI†). The most stable β-anomer is higher in energy than the corresponding α-anomer by +12.32 and +15.08 kJ mol^{-1} (E_{zpe} at M06-2X) for GlcNAc and GalNAc, respectively. In α-GlcNAc, the two most stable conformers, ^{4}C$_1$_2-3 and ^{4}C$_1$_3-4, have a negligible zero-point corrected energy gap (ΔE_{zpe}) of 0.25 kJ mol^{-1} (Fig. S2, ESI†). In GalNAc, the four lowest energy conformers in GalNAc exhibit similar geometries, with minor differences in hydroxyl group orientation, resulting in E_{zpe} of +0.0, +0.21, +0.41, and +1.85 kJ mol^{-1} (Fig. S3, ESI†). These narrow energy gaps lead to competition in the relative population derived from the HSA analysis (Fig. S1, ESI†). In GlcNAc, the global minimum population dominates at approximately 60%, while in GalNAc, these gaps in relative population are not significant at 300 K.

The comparison of the IR spectra from the study by Martens *et al.* and our simulated spectra (at 300 K, scaled by a factor of 0.958) shows decent agreement for both sodiated sugar types (sodiated GlcNAc and GalNAc), as exhibited in Fig. 2. As observed in the experimental spectra, both types share a similar carbonyl (C═O) stretching band around 1675 cm^{-1}, while the CN stretch exhibits a red-shift of approximately 37 cm^{-1} in GalNAc compared to GlcNAc. The computed spectra agree well with the trend observed in the experimental spectra. The decomposition of the total spectra for the two types is presented in Fig. S2 and S3 (ESI†). In sodiated GlcNAc, the two most stable conformers, $^4C_1$_2-3 and $^4C_1$_3-4 (all α-anomers), compete with each other in population (Fig. S2, ESI†), resulting in two associated peaks at 1520 cm^{-1} and 1495 cm^{-1} (CN stretch) (Fig. S3, ESI†). In GalNAc, the four lowest-energy conformations, though competing in population, all share a similar conformation ($^4C_1$_2-3-4, α form), leading to a similar peak position around 1495 cm^{-1} (Fig. S3, ESI†). The shift estimated by the calculations is 25 cm^{-1}, demonstrating a decent level of agreement between the experimental and simulated spectra.

Pellegrinelli *et al.* reported the cryogenic IR spectra of sodiated GalNAc.[14] As the experiment was conducted under cryogenic conditions, the simulation of accumulated spectra considered the relative population of individual conformers at 50 K, with the peak positions scaled by a factor of 0.938. The comparison with experimental spectra shows a reasonable level of agreement (Fig. S4 and S5, ESI†). In our analysis, the M06-2X rescaled vibrational modes tend to overestimate the OH stretching frequencies below 3600 cm^{-1}, while underestimating NH stretching frequencies. Therefore, as from the analysis of the vibrational modes from the computed spectra, the experimental IR band around 3350 cm^{-1} for the α-anomer is assigned as the O4H stretch from the global minimum conformer (Fig. S4, ESI†). The strong H-bond of O4H···O6 (~1.833 Å) further red-shifts the O4H stretch to below the NH stretch region. Another band around 3465 cm^{-1} in the experimental spectra is attributed to the free NH stretch, while features beyond 3600 cm^{-1} correspond to mildly H-bonded or free OH stretching modes. The β-anomer exhibits generally red-shifted bands compared to the spectra of the α-anomer in both experimental and simulated spectra, with prominent features at around 3335 cm^{-1} and another around 3455 cm^{-1} (Fig. S5, ESI†) in the experimental spectrum. Analysis of the individual vibrational spectra reveals similar properties for these features in both α- and β-anomers: the lowest-frequency band corresponds to O4H stretches from strong hydrogen bonding interactions in the global minima structures, while the free NH stretch appears at 3455 (β form) and 3465 (α form) cm^{-1}. The decent agreements in computed and experimental spectra demonstrate that our discovered local minima provide valuable molecular insights into the experimental spectra.

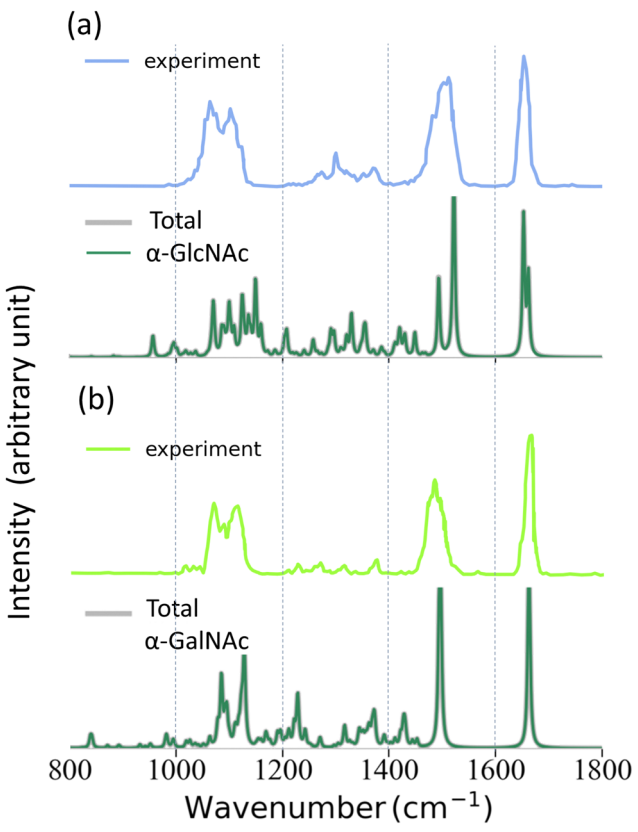


Fig. 2 Comparison of experimental spectra (top) and the simulated vibrational spectra (bottom) for (a) sodiated GlcNAc and (b) sodiated GalNAc in the frequency range 800–1800 cm^{-1}. The harmonic frequencies are scaled by a factor of 0.958.

3.5. The HSA simulation and vibrational spectra analysis of protonated di-saccharides

Schindler *et al.* have promoted a systematic scheme to unambiguously identify saccharide isomerism after sequencing *via* their IR fingerprint.[17] However, due to the unique complexity of di-saccharides, there have been no attempts to resolve the three-dimensional molecular structures that give rise to the experimental spectra. The local minima structures of neutral di-saccharides identified using NNP were used to efficiently sample the structures of protonated di-saccharides. The examples investigated in this study include the di-saccharides Gal-(1 → *n*)-GlcNAc (*n* = 3, 4, 6) and an inverse sequence α-GlcNAc-(1 → 4)-Gal. By investigating these specific examples, we aim to demonstrate the applicability and effectiveness of our proposed method in elucidating the three-dimensional molecular structures that contribute to the experimental IR spectra of di-saccharides.

The relative population evolution of four types of protonated di-saccharides, regardless of anomeric form, along the temperature range of 20–400 K is shown in Fig. S6 (ESI†). It is observed that the $^4C_1$–$^4C_1$ conformation is the predominantly stable conformer. In protonated α-Gal-(1 → 4)-GlcNAc, the energy gap between low-energy conformers is very narrow, with 8 conformers found within a range of 5 kJ mol^{-1} (Fig. S10–S12, ESI†). Due to this narrow energy gap, the relative population between conformers at 300 K is relatively closer compared to other types (Fig. S6, ESI†). In the remaining types, the energy difference between the first and second most stable conformers normally ranges from 4–6 kJ mol^{-1}.

The simulated spectra constructed from the collected set of low-energy local minima and the corresponding experimental IRMPD spectra are provided in Fig. S7–S14 (ESI†). Fig. 3 compares the experimental and simulated spectra (at 300 K, scaling factor 0.938) of four investigated types of protonated di-saccharides (regardless of anomeric forms) and displays the representative global minimum structure for each type. As observed in Fig. 3, the experimental spectra reflect an apparent contrast between protonated α-GlcNAc-(1 → 4)-Gal (Fig. 3(a)) and α-Gal-(1 → 4)-GlcNAc (Fig. 3(c)). The former exhibits a broad, relatively featureless spectral pattern in the range between 3000–3580 cm^{-1}, while the spectral pattern in the latter appears busy. The features from the simulated spectra generally agree with this observation. The analysis of vibrational modes from the individual spectra of protonated α-Gal-(1 → 4)-GlcNAc (Fig. 3c and Fig. S10–S12, ESI†) reveals that the spectral features centering around 3400 cm^{-1} (experimental band) correspond to the free NH stretch. The features around 2900–3300 cm^{-1} are congested due to the O3′H stretch contributed by the strong H-bonding interaction O3′H···O2 (red arrow in S10–S12, ESI†) from multiple low-energy conformers. The distance of this H-bond is relatively short, which ranges from 1.64 to 1.78 Å. Another type of strong H-bond interaction, which is either inter-residue (O6′H···O6H-bond interaction) or intra-residue (O4H···O6 H-bond interaction) also featuring relatively strong OH interactions, has a distance generally ranging from 1.89–2.05 Å. These vibrational modes are responsible for the peak centered around 3520 cm^{-1} (experimental band). The spectral features beyond 3560 cm^{-1} are attributed to OH stretches from less strong H-bonds and free OH groups. In the case of α-GlcNAc-(1 → 4)-Gal (Fig. 3a and Fig. S7, S8, ESI†), since the NAc group is at the non-reducing end, the NH group can participate in strong hydrogen bonding interaction with OH groups on the reducing-end (NH···O6′) with lengths ranging from 1.72 to 1.75 Å. The features around 3400 cm^{-1} in the experimental spectra (which is assigned as NH stretch region in α-Gal-(1 → 4)-GlcNAc) are absent as it is significantly red-shifted to the 2900–3000 cm^{-1} region (Fig. 3). According to Fig. S7 and S8 (ESI†), the individual vibrational analysis of the three most stable conformers indicates that the NH stretch and OH^+ stretch contribute to the two peaks ~2900 and 2950 cm^{-1} in the computed spectra. Other conformers with higher energy also have two peaks concentrating around 2800–3000 cm^{-1}. Hence, the congested spectral region around 2800–3000 cm^{-1} is largely contributed by NH and OH^+ stretching modes. The only exception is the fifth lowest-energy conformer ($^4C_1$–$^4C_1$ in the α form with energy E_{zpe} = +10.07 kJ mol^{-1}), as its geometry has a slightly larger H-bond NH···O6′ around 1.89 Å, causing the NH stretch to blue-shift to around 3165 cm^{-1} in the theoretical band, which explains the appearance of a flat and broadened hump around 3165 cm^{-1} in the experimental spectrum. The spectral features beyond 3545 cm^{-1} correspond to the OH stretch, similar to the abovementioned α-Gal-(1 → 4)-GlcNAc. The strongest intramolecular H-bond for this type is O4H···O6 or O6′H···O5′, with lengths ranging from 1.94 to 2.06 Å.

In β-Gal-(1 → 3)-GlcNAc (Fig. 3b and Fig. S9, ESI†), the global minimum conformer is the β anomer (β form at the reducing-end), which differs from the remaining three types. In this structure, the NH group forms H-bonds with the ether oxygen (O5) and O6H simultaneously, with lengths of 1.94 and 2.23 Å, respectively. These distances are longer compared to α-GlcNAc-(1 → 4)-Gal; therefore, the peak associated with this stretching mode is blue-shifted to 3110 cm^{-1} in the theoretical band. The OH^+···O1′ stretch also has a peak at 3230 cm^{-1} (theoretical band). The strong OH stretch originates from the O4′H···O6′H-bond, which has a relatively short length (1.89 Å) and an associated peak appearing around 3520 cm^{-1} (theoretical band). The second most stable conformer (α anomer with energy E_{zpe} = +5.66 kJ mol^{-1}) has free NH stretch peaks (3397 cm^{-1} on theoretical band) and very strong OH stretching peaks: O6H···O4 (3411 cm^{-1}), O4′H···O6′ (3473 cm^{-1}), and O2H···O4′ (3512 cm^{-1}), with lengths ranging from 1.85–1.86 Å. The detailed spectra of the remaining conformers are shown in Fig. S9 (ESI†). Therefore, the congested experimental spectral features around 2900–3200 cm^{-1} are largely contributed by the NH and OH^+ stretching. A minor and less intense peak centered

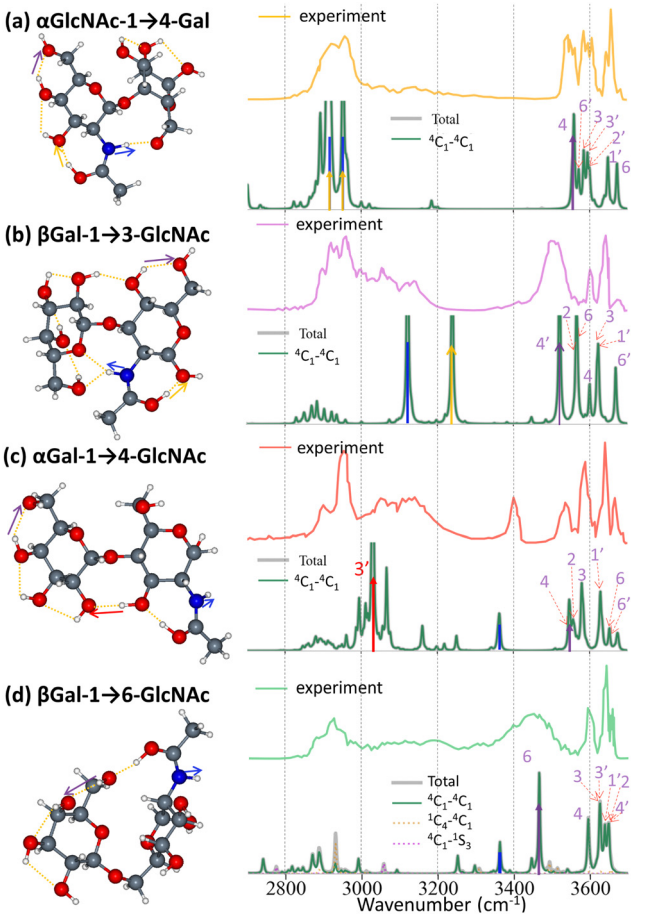


Fig. 3 Global minimum structures and comparison of experimental spectra with simulated spectra (at 50 K, scaling factor 0.938) for four disaccharides: (a) α-GlcNAc-(1 → 4)-Gal, (b) β-Gal-(1 → 3)-GlcNAc, (c) α-Gal-(1 → 4)-GlcNAc, and (d) β-Gal-(1 → 6)-GlcNAc. In the simulated spectra, the NH stretch is represented by blue columns, while OH stretch from strong hydrogen bonds is indicated by red, yellow, or purple arrows. The vertical dashed line indicates our definition of the region beyond which are less strong H-bond or free OH stretching bands.

at 3400 cm^{-1} is from the mixture of free NH stretch and strong H-bonding O6H···O4 interaction of the second lowest energy conformer (Fig. S9 (ESI†)). The intense and broadened peaks centered at 3500 cm^{-1} are from relatively strong OH stretches from multiple conformers. The features beyond 3550 cm^{-1} are OH stretches from less strong H-bonds or free OH groups.

In the β-Gal-(1 → 6)-GlcNAc molecular system (Fig. 3d and Fig. S13, S14, ESI†), the NH group is not involved in H-bonding interaction, which explains the appearance of free NH stretch ranging from 3346–3392 cm^{-1} (theoretical bands). The only exception is the second lowest energy conformer ($^4C_1$–$^1C_4$ with energy +4.07 kJ mol^{-1}), where the NH group is involved in a hydrogen bond (NH···O1′ (1.974 Å)), and the NH stretch is red-shifted to 3292 cm^{-1}. The OH stretches from strong H-bonds exhibit several peaks in the 3200–3500 cm^{-1} range (Fig. S13 and S14, ESI†). Exceptionally, there exists a peak from a strong OH stretching band that gives rise to an intense peak at 2922 cm^{-1} (theoretical band) in the second most stable conformer ($^4C_1$–$^1C_4$ in α form with E_{zpe} = + 4.07 kJ mol^{-1}), which associates with O1′H···O6. Another peak, though minor in intensity, at 3073 cm^{-1}, also originates from strong OH stretching from another conformer ($^4C_1$–$^4C_1$) with a relative energy of +10.58 kJ mol^{-1}. The CH stretching band, though normally less intense in other conformers, gives rise to an intense peak at 2881 cm^{-1} in the second most stable conformer. From the analysis of the vibrational modes of the low-energy conformers, the congested features around 2850–3000 cm^{-1} are largely contributed by a mixture of strong CH and OH stretches. The flat and broadened features around 3000–3300 cm^{-1} in the IRMPD spectrum are contributed by strong OH stretching bands from multiple low-energy conformers (Fig. S13 and S14, ESI†) and mix with NH stretches if the NH group is involved in the H-bonding network, such as the second lowest energy conformer. The stretching band from free NH groups gives rise to a shoulder around 3400 cm^{-1} (experimental band). The intense and broadened band ranging between 3350–3550 cm^{-1} comes from strong OH stretches from multiple conformers. The spectral features beyond 3550 cm^{-1} are from less strong OH and free OH groups, also from multiple conformers.

In summary, this section provides an interpretation of the experimental spectra relying on the molecular resolution from our structure sampling scheme assisted by NNP and the Q-HSA analysis. The low-energy conformers are informative in providing insights into the vibrational modes contributing to the experimental spectra and help explain why each molecular system has unique spectral features.

4. Conclusions

In this study, a structure sampling scheme was developed that utilizes established local minima datasets to efficiently sample relevant molecular systems. The approach implements a two-stage approach, with pattern transfer serving as the primary sampling scheme. The first stage systematically propagates conformational patterns between molecular systems through targeted structural modifications. The second stage employs supplementary random sampling to local minima identified in the first stage, ensuring a comprehensive exploration of the conformational landscape. Integration of the NNP model facilitates the efficient identification of energetically favorable conformers in our structure sampling approach. The pattern transfer methodology demonstrates systematic scalability across molecular complexity levels. For mono-saccharides, the scheme utilizes existing relevant local minimum structures to transform to target molecular systems through operations such as functional group substitutions or cation replacements. This approach extends to di-saccharide systems, in which established neutral conformer databases function as structural templates for sampling protonated species. The simulated vibrational spectra derived from the Q-HSA analysis show qualitative agreement with the experimental IR spectra. This agreement can be attributed to the effectiveness of the NNP-assisted sampling protocol, as evidenced by the high success rate (>80% on average) in identifying distinct DFT local minima when re-optimizing from NNP local minima. For mono-saccharides, the sampling scheme not only reproduces previously characterized stable conformers but also reveals additional low-energy local minima. Extension to di-saccharides enabled efficient conformational sampling of protonated species, with simulated vibrational spectra exhibiting a decent match with experimental observations. This approach establishes a robust framework for investigating increasingly complex molecular architectures, providing molecular-level structural insights that complement spectroscopic characterization.

Data availability

The data supporting this article have been included as part of the ESI.† The M06-2X/6-311+G(d,p) local minima structures of molecular systems described in this work are available at **https://drive.google.com/file/d/1poMYeBCTB_vZQHl3qKdFgaqeQAstEwkx/view?usp=sharing**.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This work is supported by the National Science and Technological Council of Taiwan through various grants (NSTC112-2113-M-001-034, NSTC113-2639-M-A49-002-ASP, NSTC113-2113-M-001-032-MY3) and Academia Sinica. Computational resources were supported in part by the National Center for High-Performance Computing (NCHC), Taiwan. We thank Bui Vuong Chan Dong for his valuable contribution to the results presented in this work. HTP and DCH have been supported by PhD scholarships *via* the Taiwan International Graduate Program (TIGP) in Academia Sinica. P. K. T. has been supported by an Academia Sinica post-doctoral fellowship and IAMS junior fellow. J. L. K. acknowledges the Academia Sinica Presidential Scholars Program.

References

1 A. Varki, *Glycobiology*, 2017, **27**, 3–49.
2 J. Zaia, *Mass Spectrom. Rev.*, 2004, **23**, 161–227.
3 H. T. Huynh, H. T. Phan, P.-J. Hsu, J.-L. Chen, H. S. Nguan, S.-T. Tsai, T. Roongcharoen, C. Y. Liew, C.-K. Ni and J.-L. Kuo, *Phys. Chem. Chem. Phys.*, 2018, **20**, 19614–19624.
4 C.-c Chiu, H. T. Huynh, S.-T. Tsai, H.-Y. Lin, P.-J. Hsu, H. T. Phan, A. Karumanthra, H. Thompson, Y.-C. Lee, J.-L. Kuo and C.-K. Ni, *J. Phys. Chem. A*, 2019, **123**, 6683–6700.
5 H.-S. Nguan and C.-K. Ni, *J. Phys. Chem. A*, 2022, **126**, 8799–8808.
6 E. J. Cocinero, A. Lesarri, P. Écija, Á. Cimas, B. G. Davis, F. J. Basterretxea, J. A. Fernández and F. Castaño, *J. Am. Chem. Soc.*, 2013, **135**, 2845–2852.
7 E. J. Cocinero, A. Lesarri, P. Écija, F. J. Basterretxea, J.-U. Grabow, J. A. Fernández and F. Castaño, *Angew. Chem., Int. Ed.*, 2012, **51**, 3119–3124.
8 I. Peña, E. J. Cocinero, C. Cabezas, A. Lesarri, S. Mata, P. Écija, A. M. Daly, Á. Cimas, C. Bermúdez, F. J. Basterretxea, S. Blanco, J. A. Fernández, J. C. López, F. Castaño and J. L. Alonso, *Angew. Chem., Int. Ed.*, 2013, **52**, 11840–11845.
9 C. Masellis, N. Khanal, M. Z. Kamrath, D. E. Clemmer and T. R. Rizzo, *J. Am. Soc. Mass Spectrom.*, 2017, **28**, 2217–2222.
10 J. M. Voss, S. J. Kregel, K. C. Fischer and E. Garand, *J. Am. Soc. Mass Spectrom.*, 2018, **29**, 42–50.
11 S. Warnke, A. Ben Faleh, V. Scutelnic and T. R. Rizzo, *J. Am. Soc. Mass Spectrom.*, 2019, **30**, 2204–2211.
12 V. Scutelnic and T. R. Rizzo, *J. Phys. Chem. A*, 2019, **123**, 2815–2819.
13 Y. Tan, N. Zhao, J. Liu, P. Li, C. N. Stedwell, L. Yu and N. C. Polfer, *J. Am. Soc. Mass Spectrom.*, 2017, **28**, 539–550.
14 R. P. Pellegrinelli, L. Yue, E. Carrascosa, S. Warnke, A. Ben Faleh and T. R. Rizzo, *J. Am. Chem. Soc.*, 2020, **142**, 5948–5951.
15 L. Barnes, A.-R. Allouche, S. Chambert, B. Schindler and I. Compagnon, *Int. J. Mass Spectrom.*, 2020, **447**, 116235.
16 C. Fraschetti, L. Guarcini, C. Zazza, L. Mannina, S. Circi, S. Piccirillo, B. Chiavarino and A. Filippi, *Phys. Chem. Chem. Phys.*, 2018, **20**, 8737–8743.
17 B. Schindler, L. Barnes, G. Renois, C. Gray, S. Chambert, S. Fort, S. Flitsch, C. Loison, A.-R. Allouche and I. Compagnon, *Nat. Commun.*, 2017, **8**, 973.
18 B. Schindler, G. Laloy-Borgna, L. Barnes, A.-R. Allouche, E. Bouju, V. Dugas, C. Demesmay and I. Compagnon, *Anal. Chem.*, 2018, **90**, 11741–11745.
19 L. Barnes, B. Schindler, S. Chambert, A.-R. Allouche and I. Compagnon, *Int. J. Mass Spectrom.*, 2017, **421**, 116–123.
20 C. S. Contreras, N. C. Polfer, J. Oomens, J. D. Steill, B. Bendiak and J. R. Eyler, *Int. J. Mass Spectrom.*, 2012, **330–332**, 285–294.
21 J. Martens, G. Berden, R. E. van Outersterp, L. A. J. Kluijtmans, U. F. Engelke, C. D. M. van Karnebeek, R. A. Wevers and J. Oomens, *Sci. Rep.*, 2017, **7**, 3363.
22 J.-L. Chen, H. S. Nguan, P.-J. Hsu, S.-T. Tsai, C. Y. Liew, J.-L. Kuo, W.-P. Hu and C.-K. Ni, *Phys. Chem. Chem. Phys.*, 2017, **19**, 15454–15462.
23 C.-c Chiu, S.-T. Tsai, P.-J. Hsu, H. T. Huynh, J.-L. Chen, H. T. Phan, S.-P. Huang, H.-Y. Lin, J.-L. Kuo and C.-K. Ni, *J. Phys. Chem. A*, 2019, **123**, 3441–3453.
24 M. Gaus, Q. Cui and M. Elstner, *J. Chem. Theory Comput.*, 2011, **7**, 931–948.
25 M. Gaus, A. Goez and M. Elstner, *J. Chem. Theory Comput.*, 2013, **9**, 338–354.
26 M. Kubillus, T. Kubař, M. Gaus, J. Řezáč and M. Elstner, *J. Chem. Theory Comput.*, 2015, **11**, 332–342.
27 H. T. Phan, P.-K. Tsou, P.-J. Hsu and J.-L. Kuo, *Phys. Chem. Chem. Phys.*, 2023, **25**, 5817–5826.
28 P.-K. Tsou, H. T. Huynh, H. T. Phan and J.-L. Kuo, *Phys. Chem. Chem. Phys.*, 2023, **25**, 3332–3342.
29 H. T. Phan, P.-K. Tsou, P.-J. Hsu and J.-L. Kuo, *Phys. Chem. Chem. Phys.*, 2024, **26**, 9556–9567.
30 P.-J. Hsu, K.-L. Ho, S.-H. Lin and J.-L. Kuo, *Phys. Chem. Chem. Phys.*, 2016, **19**, 544–556.
31 Y. Zhao and D. G. Truhlar, *Theor. Chem. Acc.*, 2008, **120**, 215–241.
32 A. D. McLean and G. S. Chandler, *J. Chem. Phys.*, 1980, **72**, 5639–5648.
33 T. Clark, J. Chandrasekhar, G. W. Spitznagel and P. V. R. Schleyer, *J. Comput. Chem.*, 1983, **4**, 294–301.
34 M. J. Frisch, J. A. Pople and J. S. Binkley, *J. Chem. Phys.*, 1984, **80**, 3265–3269.
35 R. Krishnan, J. S. Binkley, R. Seeger and J. A. Pople, *J. Chem. Phys.*, 1980, **72**, 650–654.
36 D. J. Wales and I. Ohmine, *J. Chem. Phys.*, 1993, **98**, 7245–7256.
37 M. Marianski, A. Supady, T. Ingram, M. Schneider and C. Baldauf, *J. Chem. Theory Comput.*, 2016, **12**, 6157–6168.
38 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery Jr, J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, T. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman and D. J. Fox, *Gaussian 16*, Gaussian Inc., Wallingford CT, 2009.
39 K. Schütt, P.-J. Kindermans, H. E. Sauceda Felix, S. Chmiela, A. Tkatchenko and K.-R. Müller, *Adv. Neural Inf. Process. Syst.*, 2017, **30**, 991–1001.
40 K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko and K.-R. Müller, *J. Chem. Phys.*, 2018, **148**, 241722.
41 K. T. Schütt, P. Kessel, M. Gastegger, K. A. Nicoli, A. Tkatchenko and K.-R. Müller, *J. Chem. Theory Comput.*, 2019, **15**, 448–455.

42 J. L. Alonso, M. A. Lozoya, I. Peña, J. C. López, C. Cabezas, S. Mata and S. Blanco, *Chem. Sci.*, 2013, **5**, 515–522.
43 I. Peña, C. Cabezas and J. L. Alonso, *Chem. Commun.*, 2015, **51**, 10115–10118.
44 I. Peña, L. Kolesniková, C. Cabezas, C. Bermúdez, M. Berdakin, A. Simão and J. L. Alonso, *Phys. Chem. Chem. Phys.*, 2014, **16**, 23244–23250.
45 R. Aguado, M. Sanz-Novo, S. Mata, I. León and J. L. Alonso, *J. Phys. Chem. A*, 2022, **126**, 7621–7626.
46 C. Møller and M. S. Plesset, *Phys. Rev.*, 1934, **46**, 618–622.
47 T. Yanai, D. P. Tew and N. C. Handy, *Chem. Phys. Lett.*, 2004, **393**, 51–57.
48 S. H. Vosko, L. Wilk and M. Nusair, *Can. J. Phys.*, 1980, **58**, 1200–1211.
49 C. Lee, W. Yang and R. G. Parr, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1988, **37**, 785–789.
50 A. D. Becke, *J. Chem. Phys.*, 1993, **98**, 5648–5652.