Received: 9 June 2022 | Revised: 17 August 2022 | Accepted: 22 August 2022

DOI: 10.1002/jcc.27000

Journal of COMPUTATIONAL CHEMISTRY WILEY

RESEARCH ARTICLE

# Modeling pyranose ring pucker in carbohydrates using machine learning and semi-empirical quantum chemical methods

Linghan Kong | Richard A. Bryce

Division of Pharmacy and Optometry, School of Health Sciences, Manchester Academic Health Sciences Centre, University of Manchester, Manchester, UK

**Correspondence**
Richard A. Bryce, Division of Pharmacy and Optometry, School of Health Sciences, University of Manchester, Manchester M13 9PT, UK.
Email: r.a.bryce@manchester.ac.uk

## Abstract

Pyranose ring pucker is a key coordinate governing the structure, interactions and reactivity of carbohydrates. We assess the ability of the machine learning potentials, ANI-1ccx and ANI-2x, and the GFN2-xTB semiempirical quantum chemical method, to model ring pucker conformers of five monosaccharides and oxane in the gas phase. Relative to coupled-cluster quantum mechanical calculations, we find that ANI-1ccx most accurately reproduces the ring pucker energy landscape for these molecules, with a correlation coefficient $r^2$ of 0.83. This correlation in relative energies lowers to values of 0.70 for ANI-2x and 0.60 for GFN2-xTB. The ANI-1ccx also provides the most accurate estimate of the energetics of the $^4C_1$-to-$^1C_4$ minimum energy pathway for the six molecules. All three models reproduce chair more accurately than non-chair geometries. Analysis of small model molecules suggests that the ANI-1ccx model favors puckers with equatorial hydrogen bonding substituents; that ANI-2x and GFN2-xTB models overstabilize conformers with axially oriented groups; and that the *endo*-anomeric effect is overestimated by the machine learning models and underestimated via the GFN2-xTB method. While the pucker conformers considered in this study correspond to a gas phase environment, the accuracy and computational efficiency of the ANI-1ccx approach in modeling ring pucker in vacuo provides a promising basis for future evaluation and application to condensed phase environments.

KEYWORDS
ANI, carbohydrates, GFN2-xTB, machine learning, ring pucker

## 1 | INTRODUCTION

Carbohydrates play a range of key roles in biology, including in the mediation of cell–cell and cell–pathogen interactions.[1] For example, interaction of viral surface proteins with host cell carbohydrates enables infection and disease, as in the case of Influenza, Dengue virus, and Rotavirus.[2] The glycosylation of viral proteins can also play a role in infection: a recent example is provided by the spike protein of SARS-CoV-2, which is rich in complex N-glycans at 22 amino acid sites.[3] These conjugated carbohydrates have been shown to play a dual role, in shielding the spike protein amino acids from recognition by the host immune system; and also in promoting binding of the virus to the host ACE2 protein receptor.[3]

Structural characterization of the structure and mechanism of carbohydrates is challenging however, due to their diversity in covalent connectivity and conformation.[4] In addition to the flexibility associated with the

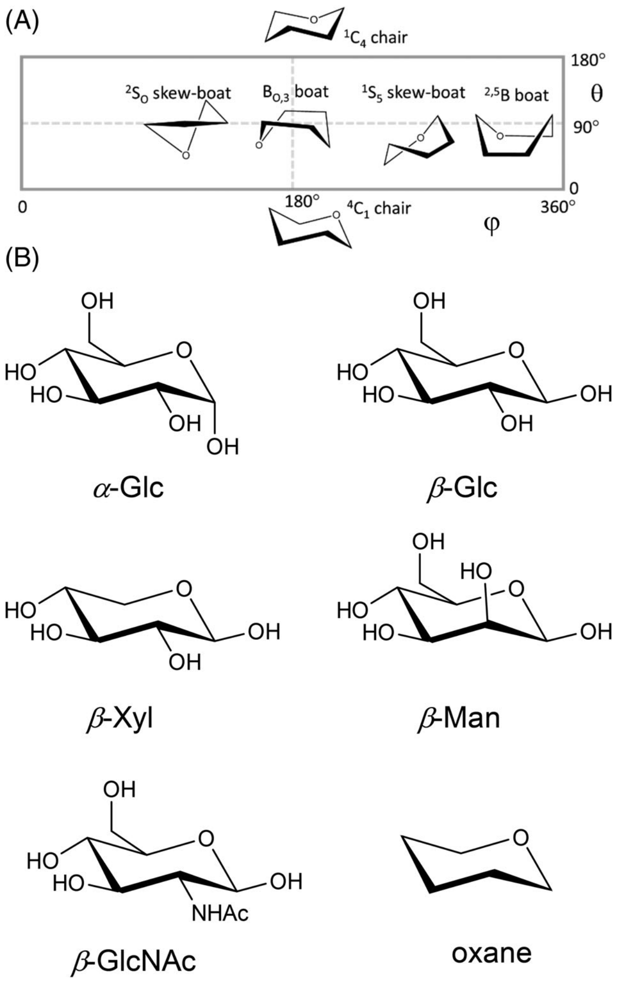

**FIGURE 1** (A) Schematic diagram of location of selected pyranose ring conformers on Cremer–Pople ($\theta$,$\varphi$) surface. (B) Monosaccharides α-D-glucose (α-Glc), β-D-glucose (β-Glc), β-D-xylose (β-Xyl), β-D-mannose (β-Man); N-acetyl β-glucosamine (β-GlcNAc), and oxane

glycosidic linkage that joins monosaccharide residues within a carbohydrate oligomer or polymer, the hexopyranose ring of each residue can adopt different shapes, called puckers. While the pyranose ring is most commonly chair (C) in pucker, other conformations can be adopted, denoted boat (B), half-chair (H), skew-boat (S) and envelope (E). The manifold of puckers can be conveniently represented on the hypersurface proposed by Cremer and Pople,[5] described by angles, $\theta$ and $\varphi$ (Figure 1A). Non-chair conformations have been found to play an important role in carbohydrate interactions and reactivity.[6] For example, a skew-boat conformation has been observed for residues of heparin substrate when non-covalently bound to fibroblast growth factor;[7] and a boat pucker is formed in the covalent intermediate of a xylanase enzyme.[8]

Given the experimental challenges in determining carbohydrate conformations in receptor bound and unbound states, computational modeling is an invaluable complementary tool. Approaches typically employ classical force fields in combination with molecular simulation techniques.[4] Recently, for example, we applied an enhanced sampling molecular dynamics scheme[9] with the GLYCAM carbohydrate force field[10] to evaluate the ring pucker free energy landscapes of a range of glycosaminoglycan monosaccharides.[11] Carbohydrate force fields have undergone numerous refinements over the years in an effort to improve the level of accuracy in modeling the subtleties of carbohydrate structure and dynamics.[12]

An alternative approach has been to simulate the conformational behavior of carbohydrates using semi-empirical quantum mechanics (SQM), linking to the condensed phase via a quantum mechanical (QM)/molecular mechanical (MM) framework.[13,14] In this regard, we note the recent development of the tight binding density function method, GFN2-xTB,[15] which has shown promise in the evaluation[15] of the relative energetics of α- and β-glucose and α-maltose conformers that feature in the data set of Marianski et al.[16] In this study, reference energies were computed using the domain-based local PNO (DLPNO) local correlation method,[17] providing DLPNO-CCSD(T) energies with extrapolation to the complete basis set (CBS) limit. The GFN2-xTB method was found to give a mean absolute deviation in conformer energy differences of 3.2 kcal/mol compared to the DLPNO-CCSD(T)/CBS level of theory. In this analysis, it was found that relative energies of conformations were underestimated on average by the GFN2-xTB method. Another smaller scale study[18] evaluated the ability of GFN2-xTB to model the conformers of the SCONF set[19,20] of β-glucose and 3,6,-anhydro-4-*O*-methyl-D-galactitol. Here, the mean absolute deviation in relative energy compared to DLPNO-CCSD(T)/CBS calculations was computed to be 1.7 kcal/mol.

An emerging route to accurate and efficient molecular potentials is via machine learning (ML), using techniques such as kernel-based methods and neural networks.[21] With a suitable training set of molecular geometries and energies, ML methods can learn to directly and rapidly predict energy as a function of molecular geometry rather than be used to fit parameters of a predetermined functional form of the potential. The approach is typified by the ANI suite of methods,[22] which are based on fitting a neural network to reproduce quantum chemical geometries and energies of diverse organic molecules. To capture the environment around each atom, modified symmetry functions are used as descriptors.[22,23]

In its first implementation, ANI-1, a training set of 22 M non-equilibrium molecular energies and geometries of 57 k molecules was used, computed at the wB97X/6-31G* level.[22] The ANI-1 method was further refined using a data set of 5.5 M molecular conformations selected by an active learning approach, to give the ANI-1x model.[24] This training set contained only C, H, N, and O atoms; subsequently the approach was extended to the elements S, F, and Cl, and increased to 8.9 M molecular conformations, again via active learning, to yield the ANI-2x model.[25] A further refinement to the model, using a smaller set of ∼500 k reference training datapoints but computed at the DLPNO-CCSD(T)/CBS level, led to the ANI-1ccx model (for C, O, N, and H elements).[26] These models have shown promise in the ranking of energetics of the conformations of small organic molecules;[25,27] for example, for a study of ∼700 drug-like molecules,[27] ANI-1ccx single point calculations at B3LYP-D3BJ/def2-SVP geometries afforded similar accuracy and calculation efficiency to the GFN2-xTB method: relative to reference DLPNO-CCSD(T)/CBS energy calculations, both ANI-1ccx and GFN2-xTB yielded correlation coefficients $r^2$ of 0.64.[27]

In this work, we evaluate the ability of the ANI-1ccx, ANI-2x and GFN2-xTB methods to characterize the in vacuo potential energy landscape of pyranose ring pucker in carbohydrates. To achieve this, we employ the benchmark dataset of Mayes et al.,[28] which characterizes, at the CCSD(T)/6-311+G(d,p) level of theory, a wide range of ring puckers for five monosaccharide molecules and the undecorated pyranose ring model, oxane (also known as tetrahydropyran). Specifically the five monosaccharides (Figure 1B) are the anomers, α-D-glucose (α-Glc) and β-D-glucose (β-Glc); the 5-deoxymethyl analog of the β-anomer, namely β-D-xylose (β-Xyl); the C2 epimer of β-D-glucose, β-D-mannose (β-Man); and the N-acetylated form of β-glucose, N-acetyl β-glucosamine (β-GlcNAc). We examine the ability of the machine learning and SQM methods to reproduce the potential energy of this set of 299 conformers, which represent a range of ring pucker and rotameric states for each molecule.

## 2 | METHODS

For this study, we employ the dataset of Mayes et al.,[28] which comprises 918 conformers of oxane, α-Glc, β-Glc, β-Xyl, β-Man and β-GlcNAc, exhibiting a broad range of hydroxyl and hydroxymethyl rotamers and ring puckers. From this set, here we use the subset of 299 local energy minima as characterized via vibrational frequency calculations, thus omitting transition states. In the dataset, Mayes et al.[28] obtained geometries at the B3LYP/6-311+G(2df,p) level of theory, with an ultrafine integration grid and tight convergence. Subsequent potential energies were computed at these geometries using the CCSD(T)/6-311+G(d,p) level of theory. These are referred to subsequently here as the reference geometries and energies, respectively. Three molecules additional to the Mayes et al. set were also considered (glycerol, 4,6-dimethyloxane and oxan-2-ol); we determined geometries at the B3LYP/6-311+G(2df,2p) level and energies via the DLPNO-CCSD(T) method,[17] using a 3-point extrapolation to obtain the complete basis set (CBS) estimate.[29] These calculations employed the ORCA quantum chemistry package.[30]

For these structures, we compute energies and geometries using ANI-1ccx and ANI-2x methods.[25,26] The calculations were carried out with ASE interface of TorchANI.[31] The geometry convergence criterion was set such that there was a force magnitude of less than 0.01 eV/Å on every atom. We also apply the tight binding semiempirical method, GFN2-xTB, using the xtb program.[32] The self-consistent charge convergence cutoff was set to $1.0 \times 10^{-6}$ $E_h$ and the geometry convergence to an iterative energy difference of less than $1.0 \times 10^{-6}$ $E_h$.

## 3 | RESULTS AND DISCUSSION

We first consider the overall ability of the ANI-1ccx, ANI-2x and GFN2-xTB methods to rank the 299 conformers of the Mayes et al.,[28] according to the potential energy relative to the lowest energy reference conformer ($\Delta E$). For the five monosaccharides (Figure 1B), the lowest energy structure corresponds to a $^4C_1$ puckered conformation. When ranking of the set is performed using the ANI-1ccx energy at the reference geometry, we observe a good correlation with reference CCSD(T)/6-311+G(d,p) relative energies across the range of ring puckers (Figure 2A). The correlation coefficient $r^2$ for the conformer set is 0.83, with a mean absolute error (MAE) in relative energy of 1.1 kcal/mol (Table 1). On geometry optimization via the ANI-1ccx model, the correlation and MAE in relative energies do not change significantly, with a small reduction in $r^2$ to 0.81 (Figure 2B) and increase in MAE of energy to 1.3 kcal/mol (Table 1). The highest MAE is found for conformers of β-GlcNAc with a value of 1.6 kcal/mol (Table 1).

A rather lower correlation is observed between ANI-2x and reference energies, with a $r^2$ of 0.70 and MAE in $\Delta E$ of 2.0 kcal/mol at the reference geometries (Figure 2C, Table 1); this agreement decreases to a $r^2$ value of 0.65 and MAE of 2.2 kcal/mol on geometry optimization via the ANI-2x model (Table 1). In particular, it appears that stability of the $^1C_4$ monosaccharide conformers are significantly overestimated relative to the $^4C_1$ conformation by the ANI-2x potential (red, Figure 2C,D). Regarding specific molecules, β-Glc and β-Man have the largest MAEs in $\Delta E$, with values of 2.1 and 3.9 kcal/mol, respectively (Table 1).

Finally, for the semiempirical GFN2-xTB Hamiltonian, the overall correlation in relative energies with reference CCSD(T)/6-311+G(d,p) values is the lowest of the three approaches, with a $r^2$ of 0.60 and MAE of 3.8 kcal/mol at reference geometries (Figure 2E, Table 1), and $r^2$ of 0.47 (Figure 2F) and MAE of 3.8 kcal/mol on geometry optimization. In this case, there appears to be a systematic disparity in the stability of $^4C_1$ versus other conformers, across molecules (Figure 2E,F; Table S1). We note that the highest MAE is found for β-Glc with a value of 5.1 kcal/mol (Table 1).

### 3.1 | Chair conformer energies

We turn now to examine in more detail the energetic landscape as a function of ring pucker for these six molecules, considering first the chair conformers, $^4C_1$ and $^1C_4$. At the reference coupled-cluster level of theory, the lowest energy conformer for each molecule is predicted to be in a $^4C_1$ pucker in the gas phase (Table 2), except for the symmetric oxane molecule, where the $^4C_1$ and $^1C_4$ conformers are degenerate. We note that the structures of this study are derived in the gas phase, where intramolecular hydrogen bonding is highly favored (e.g., Figure 3). In aqueous solution, competition for interaction with water would disrupt these internal hydrogen bonding networks. In vacuo, the monosaccharides range in relative energy of the $^1C_4$ local minimum (Table 2), from 1.0 kcal/mol for β-Xyl; to 4.7 kcal/mol for β-Glc; and to 4.9 kcal/mol for both α-Glc for β-GlcNAc. The higher $\Delta E$ of the $^1C_4$ conformation of glucose anomers compared to that of β-Xyl is in large part a reflection of the energetic cost of an equatorial-to-axial transition of the bulky $CH_2OH$ group (compare Figure 3A,B with Figure 3C,D).

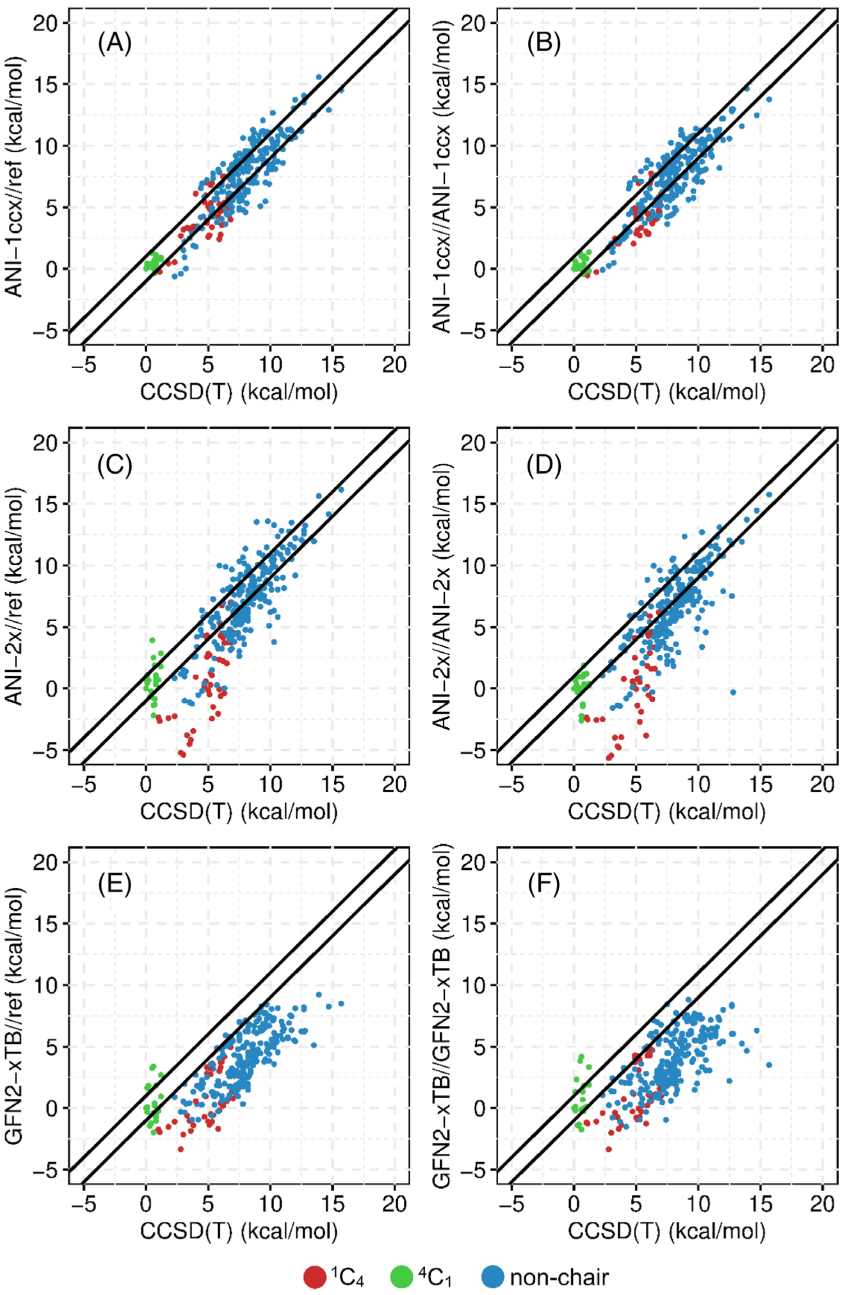

**FIGURE 2** Correlation between relative energy Δ*E* computed by (A) and (B) ANI-1ccx; (C) and (D) ANI-2x; (E) and (F) GFN2-xTB models, at reference and model optimized geometries, respectively, with relative energies computed at CCSD(T)/6-311+G(d,p) level of theory, across conformers of five monosaccharides and oxane. Energies in kcal/mol. Solid lines denote deviations of ±1 kcal/mol

**TABLE 1** Mean absolute error (MAE) in relative energy Δ*E* (in kcal/mol) and correlation coefficient $r^2$ for molecules computed via the ANI-1ccx, ANI-2x, and GFN2-xTB methods

| Method | Oxane | β-Xyl | α-Glc | β-Glc | β-Man | β-GlcNAc | tot | $r^2$ |
|---|---|---|---|---|---|---|---|---|
| *N* | 8 | 26 | 53 | 85 | 58 | 69 | 299 | |
| ANI-1ccx | 0.26 (0.39) | 1.19 (1.36) | 1.39 (1.12) | 0.98 (1.33) | 0.86 (1.09) | 1.58 (1.58) | 1.14 (1.28) | 0.83 (0.81) |
| ANI-2x | 0.23 (0.23) | 1.83 (1.97) | 1.28 (1.81) | 1.93 (2.08) | 3.27 (3.90) | 1.64 (1.47) | 1.96 (2.19) | 0.70 (0.65) |
| GFN2-xTB | 0.91 (0.92) | 4.40 (4.64) | 3.04 (3.17) | 5.06 (5.14) | 4.21 (4.48) | 2.37 (1.91) | 3.76 (3.77) | 0.60 (0.47) |

*Note*: Number of conformers for a given molecule, *N*, also shown. Values of MAE in relative energy in parentheses are computed for geometries optimized via same method as the energy calculation.

On application of the ANI-1ccx model, the overall MAE in predicted energy of chair puckers $^4C_1$ and $^1C_4$ relative to the $^4C_1$ global minimum is well described, with an overall value of 0.9 kcal/mol, increasing to 1.2 kcal/mol at the ANI-1ccx geometry (Table 3); at both reference and relaxed geometries, the MAE in Δ*E* for $^1C_4$ conformers is larger than the value found for $^4C_1$ structures (Table 3). This is expected given the latter differ from the global minimum structure only in substituent rotamers not ring pucker.

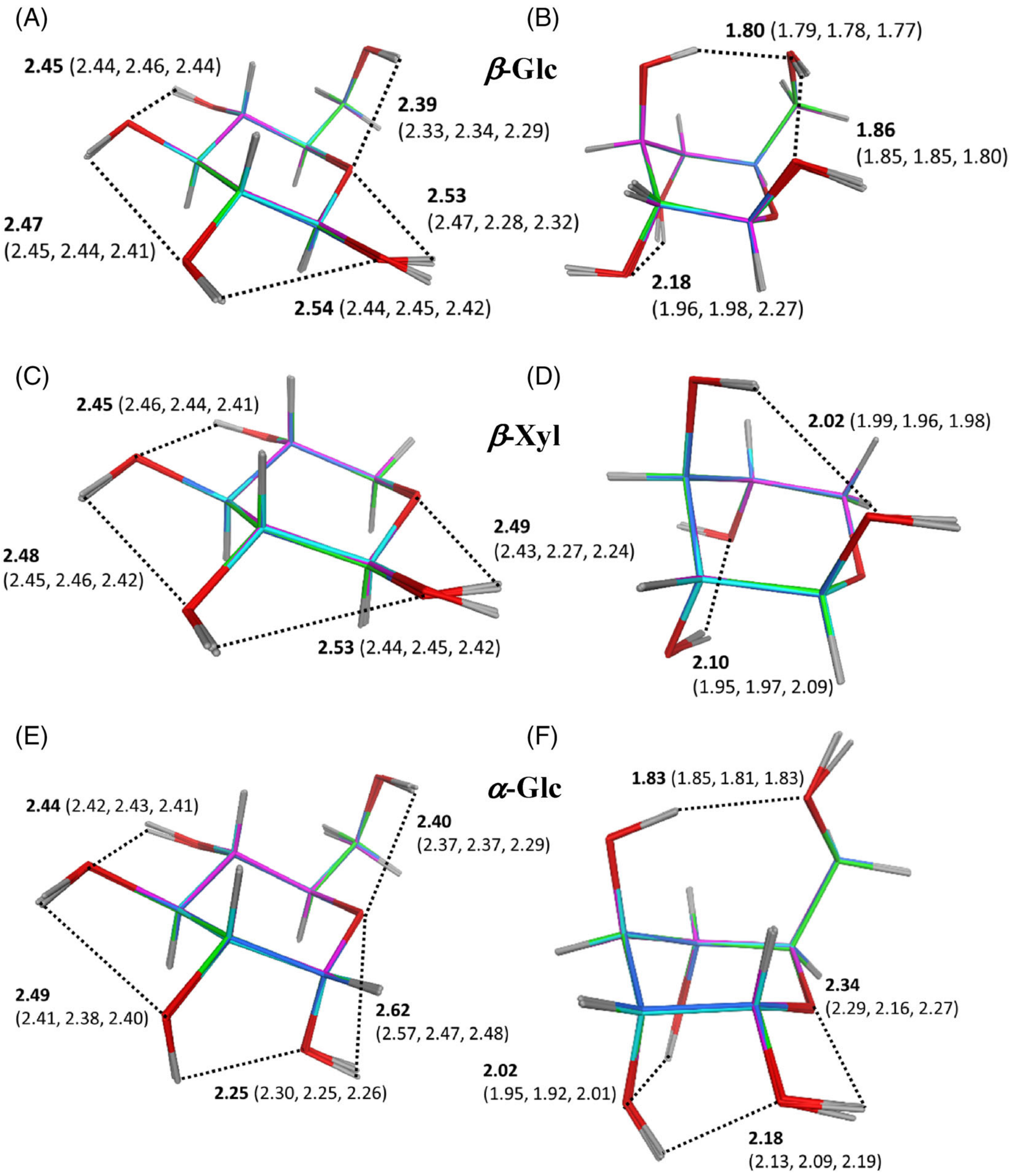

**FIGURE 3** Optimized structures of β-Glc in (A) $^4C_1$ (conformer id #BG-51 of Supporting Information) and (B) $^1C_4$ pucker (#BG-6); of β-Xyl in (C) $^4C_1$ (#BX-15) and (D) $^1C_4$ pucker (#BX-2); of α-Glc in (E) $^4C_1$ (#AG-31) and (F) $^1C_4$ pucker (#AG-3); at the reference geometry (cyan), ANI-1ccx (magenta), ANI-2x (blue) and GFN2-xTB (green). Hydrogen bond distances marked (black dotted lines) and values indicated in Å, for reference geometry (bold) and in parentheses, for ANI-1ccx, ANI-2x and GFN2-xTB methods, respectively

**TABLE 2** Relative energies Δ*E* (in kcal/mol) for selected lowest energy local puckers relative to global minimum $^4C_1$ conformer, computed at CCSD(T)/6-311+G(d,p) level of theory

| Mol | Ring pucker | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $^4C_1$ | $^1C_4$ | $^3S_1$ | $^2S_O$ | $^1S_3$ | $B_{O,3}$ | $^1S_5$ | $^{O,3}B$ | $^OS_2$ | $^5S_1$ | $B_{1,4}$ | $^{1,4}B$ | $B_{2,5}$ |
| Oxane | 0.0 | 0.0 | 5.7 | 5.7 | 5.7 | | 6.8 | | 5.7 | 6.8 | | | |
| β-Xyl | 0.0 | 1.0 | 7.4 | 4.1 | 6.7 | 6.7 | 7.5 | 7.8 | 8.2 | 8.7 | 10.6 | 12.5 | |
| α-Glc | 0.0 | 4.9 | 7.5 | 11.9 | 8.2 | 9.4 | 7.5 | 7.8 | 4.4 | 7.2 | 7.9 | 9.9 | 9.1 |
| β-Glc | 0.0 | 4.7 | 7.2 | 6.6 | 5.7 | 4.1 | 7.0 | 7.2 | 6.1 | 9.5 | 9.2 | 8.2 | 11.0 |
| β-Man | 0.0 | 2.8 | 4.7 | 8.2 | 3.1 | 5.3 | 5.5 | 6.7 | 7.0 | 10.6 | 6.3 | 10.0 | 6.7 |
| β-GlcNAc | 0.0 | 4.9 | 6.9 | 3.8 | 3.0 | 2.3 | 6.4 | 5.9 | 6.2 | 4.4 | 4.9 | 9.2 | 9.7 |

The ANI-1ccx potential predicts an increase in range of stability of $^1C_4$ conformers relative to the reference potential energy surface: for β-Xyl, the stability of the lowest energy $^1C_4$ conformer increases by 1.3 kcal/mol via ANI-1ccx (Table 4), to give a Δ*E* value of 0.3 kcal/mol (Table S2). At the other end of the range, the α-Glc lowest energy $^1C_4$ conformer reduces in stability by 1.9 kcal/mol (Table 4), to lie 6.8 kcal/mol above the $^4C_1$ minimum (Table S2 and Figure 3E,F). These relative energies are rather similar to those obtained at ANI-1ccx optimized geometries, with a Δ*E* of −0.5 and 6.9 kcal/mol for β-Xyl and α-Glc, respectively (Table S2).

For ANI-2x, the mean absolute errors in relative energy are significantly larger: we observe an overall MAE in Δ*E* for chair pucker of 3.0 kcal/mol, increasing to 3.2 kcal/mol at the ANI-2x geometry (Table 3). As for ANI-1ccx, the $^1C_4$ conformers have a higher error than $^4C_1$ conformers, although the MAE in Δ*E* of the lowest energy $^4C_1$ minimum increases from ANI-1ccx to ANI-2x from 0.5 to 1.0 kcal/mol at relaxed geometries (Table 3). Also, we find that the lowest energy $^1C_4$ conformer of β-Xyl is stabilized even further than for ANI-1ccx, by 3.6 kcal/mol (Table 4), corresponding to a Δ*E* value of −2.6 kcal/mol predicted by ANI-2x at both the reference and ANI-2x geometries (Table S2). The most pronounced overstabilization is for β-Glc and β-Man, which reduce in Δ*E* by 7.0 and 6.0 kcal/mol respectively at the reference geometry; and by 8.5 and 5.9 kcal/mol on optimization (Table 4). These are significantly larger deviations in relative energy than those obtained via ANI-1ccx.

The GFN2-xTB quantum chemical method predicts a similar mean absolute error in Δ*E* of chair conformers as for ANI-2x, with reference and relaxed values of 3.0 and 2.7 kcal/mol, respectively (Table 3). The MAE in Δ*E* for $^1C_4$ conformers is also larger than for ANI-1ccx, with a value of 4.0 kcal/mol, decreasing to 3.5 kcal/mol on geometry optimization (Table 3). At this level of theory, there is again overstabilization of $^1C_4$ conformers, with for example geometry optimized ΔΔ*E* values of −4.7 and −4.5 kcal/mol for β-Glc and β-Man, respectively (Table 4). In general, GFN2-xTB appears able to model the stability of $^1C_4$ conformers relative to the $^4C_1$ global minimum, across the monosaccharides more accurately than ANI-2x but not ANI-1ccx (Table 4).

## 3.2 | Non-chair conformer energies

For prediction of the relative energies of non-chair puckers, we observe a similar trend as for chair conformers, with lower errors in Δ*E* for ANI-1ccx relative to ANI-2x and GFN2-xTB (Table 3). For ANI-1ccx, the MAE in non-chair pucker Δ*E* is 1.2 kcal/mol at the reference geometries, and 1.3 kcal/mol on geometry optimization (Table 3). While ANI-1ccx values of MAE in Δ*E* for non-chair and chair conformers are similar, for ANI-2x, the MAE in energy estimates for non-chair conformers is significantly improved over chair structures, by 1.4 and 1.3 kcal/mol, respectively (Table 3). For the GFN2-xTB method, the MAE in non-chair Δ*E* is 3.9 and 4.0 kcal/mol, respectively (Table 3), which is 0.9 and 1.3 kcal/mol higher than for chair conformers.

**TABLE 3** Mean absolute error (MAE) in relative energy Δ*E* (in kcal/mol) of selected sets of *N* pucker conformers (see Table S1 for full range of puckers)

| Method | Ring pucker | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | $^4C_1$ | $^3S_1$ | $^{2,5}B$ | $^2S_O$ | $B_{O,3}$ | $^1S_3$ | $^1S_5$ | $^OS_2$ | $^1C_4$ | Chair | Non-chair |
| *N* | 26 | 17 | 7 | 14 | 16 | 23 | 40 | 27 | 43 | 69 | 230 |
| ANI-1ccx | 0.32 (0.48) | 1.25 (1.39) | 1.68 (1.71) | 1.00 (0.63) | 1.55 (1.46) | 1.28 (1.29) | 1.32 (1.48) | 0.87 (0.89) | 1.22 (1.64) | 0.88 (1.20) | 1.23 (1.28) |
| ANI-2x | 1.12 (0.98) | 2.46 **(3.22)** | 0.85 (0.53) | 0.45 (2.15) | 1.68 (2.00) | 1.87 (1.93) | 1.36 (1.42) | 1.38 (1.42) | **4.18 (4.48)** | **3.02 (3.16)** | 1.61 (1.87) |
| GFN2-xTB | 1.35 (1.30) | **3.57 (3.58)** | **6.03 (7.40)** | **4.27 (4.25)** | **4.10 (4.29)** | 2.82 **(3.20)** | **4.13 (4.09)** | **3.71 (3.40)** | **4.02 (3.50)** | **3.01** (2.67) | **3.92 (4.03)** |

*Note*: Values exceeding 3 kcal/mol in bold.

**TABLE 4** Deviation in relative energy, $\Delta\Delta E$ (in kcal/mol), for ANI-1ccx, ANI-2x, and GFN2-xTB methods, for lowest energy local pucker conformer relative to global minimum $^{4}C_{1}$ conformer at reference level of theory

| Mol | Ring pucker | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | ANI-1ccx | | | | | | | | | |
| | $^{4}C_{1}$ | $^{1}C_{4}$ | $^{3}S_{1}$ | $^{2}S_{O}$ | $^{1}S_{3}$ | $^{1}S_{5}$ | $^{O}S_{2}$ | $B_{1,4}$ | $^{1,4}B$ | $B_{2,5}$ |
| Oxane | 0.0 (0.0) | 0.0 (0.0) | 0.0 (−0.1) | 0.0 (−0.1) | 0.0 (−0.1) | −0.9 (−1.1) | 0.0 (−0.1) | | | |
| β-Xyl | 0.0 (0.0) | −1.3 (−1.5) | 0.8 (0.5) | −0.6 (−0.7) | −2.4 (−2.6) | −1.9 (−1.9) | −0.3 (−0.6) | −0.2 (−0.7) | −1.3 (−1.7) | |
| α-Glc | 0.0 (0.0) | 1.9 (2.1) | 1.1 (0.6) | 1.2 (0.3) | 1.0 (1.0) | 1.0 (1.3) | 2.2 (2.1) | 0.9 (0.2) | 0.6 (0.5) | 0.1 (0.0) |
| β-Glc | 0.0 (0.0) | −1.3 (−2.0) | 1.5 (0.9) | −0.5 (−1.0) | −1.9 (−2.0) | −1.5 (−2.0) | 0.7 (0.4) | −0.2 (−0.7) | −1.8 (−2.2) | −1.3 (−2.8) |
| β-Man | −0.3 (−0.4) | 0.2 (−0.6) | 0.2 (−0.1) | 0.1 (−0.3) | −0.6 (−1.0) | 0.4 (−0.1) | 1.2 (0.1) | −1.2 (−1.3) | −1.2 (−1.6) | 0.8 (−0.5) |
| β-GlcNAc | 0.0 (−0.3) | −2.5 (−1.8) | −0.7 (**−3.0**) | −2.0 (−1.1) | −2.1 (−1.3) | −2.8 (−1.7) | 0.1 (1.2) | 2.3 (2.7) | **−4.1 (−3.3)** | −0.3 (−0.1) |
| | ANI-2x | | | | | | | | | |
| | $^{4}C_{1}$ | $^{1}C_{4}$ | $^{3}S_{1}$ | $^{2}S_{O}$ | $^{1}S_{3}$ | $^{1}S_{5}$ | $^{O}S_{2}$ | $B_{1,4}$ | $^{1,4}B$ | $B_{2,5}$ |
| Oxane | 0.0 (0.0) | 0.0 (0.0) | 0.4 (0.4) | 0.4 (0.4) | 0.4 (0.4) | 0.0 (−0.1) | 0.4 (0.4) | | | |
| β-Xyl | 0.0 (0.0) | **−3.6 (−3.6)** | −1.5 (**−3.2**) | 0.1 (0.1) | −1.9 (−1.7) | −2.2 (−2.2) | −2.7 (−2.5) | 1.8 (−1.0) | −0.4 (−0.4) | |
| α-Glc | −0.7 (−0.3) | **−3.1 (−4.1)** | **−4.2 (−5.2)** | 0.3 (**−11.9**) | −0.4 (−0.7) | 1.2 (−0.1) | 0.6 (−0.1) | **−3.6 (−5.4)** | 1.2 (0.7) | 0.9 (0.1) |
| β-Glc | −0.2 (0.0) | **−7.0 (−8.5)** | −0.2 (−1.0) | 0.1 (−1.1) | −2.0 (−1.9) | −2.4 (−2.3) | −1.5 (−2.4) | −2.7 (−2.8) | −2.2 (−2.4) | −1.8 (−1.7) |
| β-Man | −2.2 (−2.6) | **−6.0 (−5.9)** | −2.2 (−2.1) | 2.0 (1.1) | −2.1 (−2.1) | −0.1 (−0.4) | 0.6 (0.7) | **−4.1 (−4.0)** | −1.6 (−1.7) | −0.5 (−0.9) |
| β-GlcNAc | 0.0 (−0.1) | −2.6 (**−4.0**) | −2.8 (**−4.1**) | −1.2 (−0.4) | −2.0 (−1.2) | 1.8 (2.0) | 0.8 (0.6) | 1.1 (1.2) | −2.5 (−2.8) | 1.8 (1.2) |
| | GFN2-xTB | | | | | | | | | |
| | $^{4}C_{1}$ | $^{1}C_{4}$ | $^{3}S_{1}$ | $^{2}S_{O}$ | $^{1}S_{3}$ | $^{1}S_{5}$ | $^{O}S_{2}$ | $B_{1,4}$ | $^{1,4}B$ | $B_{2,5}$ |
| Oxane | 0.0 (0.0) | 0.0 (0.0) | −0.8 (−0.8) | −0.8 (−0.8) | −0.8 (−0.8) | −1.6 (−1.6) | −0.8 (−0.8) | | | |
| β-Xyl | −0.6 (−0.1) | −2.3 (−2.2) | **−5.1 (−5.4)** | −1.7 (−1.9) | **−4.4 (−4.4)** | **−4.9 (−5.1)** | **−5.7 (−5.9)** | **−3.0 (−7.4)** | **−5.7 (−10.2)** | |
| α-Glc | −1.5 (−1.4) | −0.6 (−0.4) | **−4.2 (−4.3)** | **−5.0 (−5.1)** | −0.5 (−0.8) | 0.0 (0.1) | −1.3 (−1.0) | **−3.7 (−3.4)** | −0.9 (−0.9) | −0.7 (−0.4) |
| β-Glc | −1.2 (−1.3) | **−5.2 (−4.7)** | **−3.4 (−4.4)** | −2.6 (−2.8) | **−3.8 (−3.5)** | **−5.5 (−5.1)** | **−3.9 (−3.6)** | −2.7 (−2.8) | **−4.5 (−5.4)** | **−4.0 (−4.4)** |
| β-Man | −2.0 (−1.7) | **−4.1 (−4.5)** | −1.7 (−1.8) | **−4.7 (−5.3)** | −2.0 (−2.9) | **−3.8 (−4.8)** | −2.2 (−2.7) | −2.3 (−2.6) | −1.6 (**−5.4**) | −2.1 (**−3.0**) |
| β-GlcNAc | −1.1 (0.0) | −1.1 (−0.9) | −2.7 (**−3.7**) | −1.4 (−1.7) | −1.4 (−1.6) | −1.3 (−1.9) | −0.1 (−0.1) | −1.3 (−1.7) | −2.2 (−2.6) | **−3.4 (−4.2)** |

*Note*: Values for geometries optimized at level of theory in parentheses. $\Delta\Delta E$ exceeding an absolute value of 3 kcal/mol shown in bold. Corresponding $\Delta E$ for molecules in Tables S3–S8.

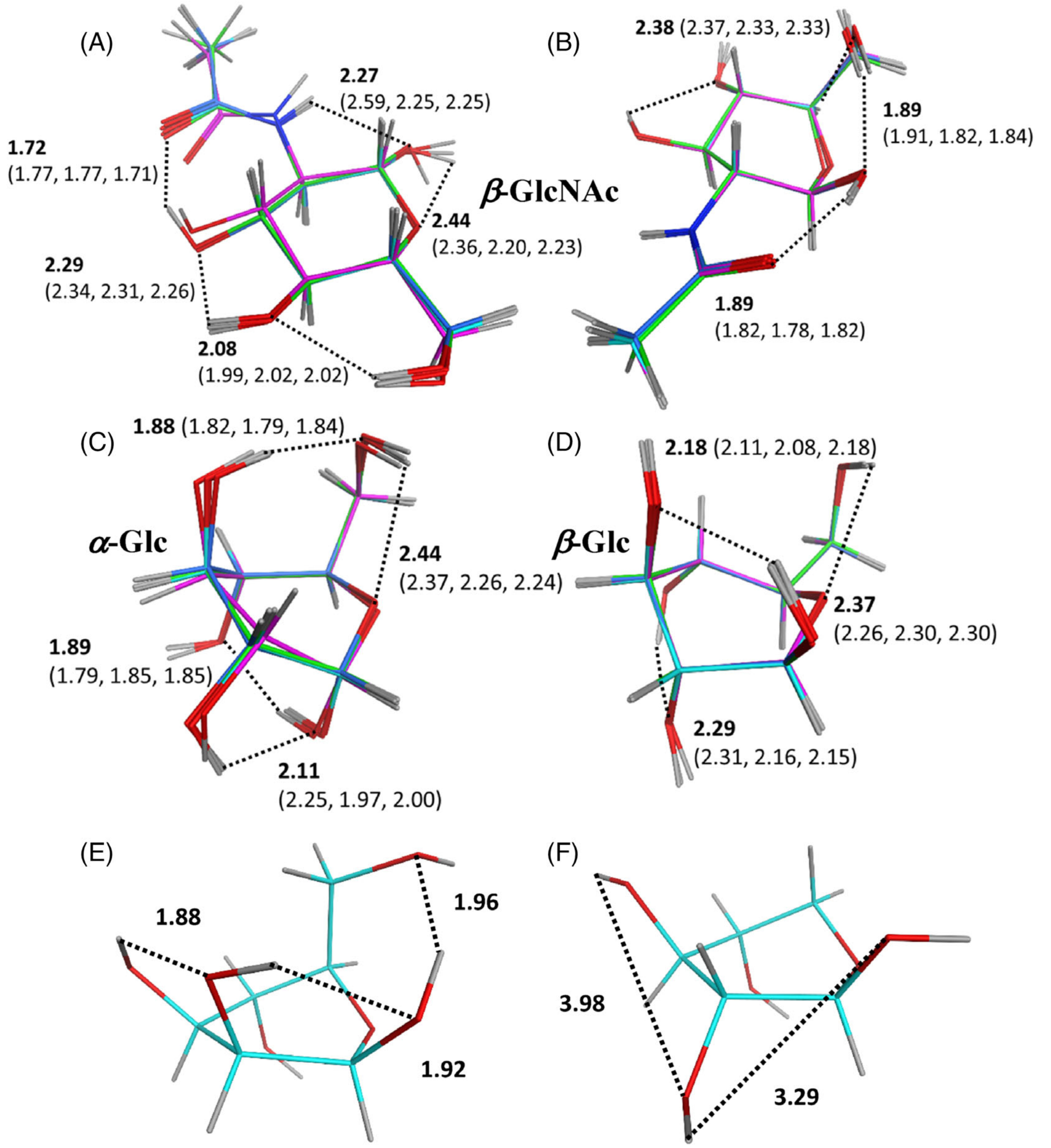

**FIGURE 4** Optimized structures of (A) β-GlcNAc in $^4C_1$ (conformer id #BN-34) pucker; (B) of β-GlcNAc in $B_{O,3}$ (#BN-55) pucker; (C) of α-Glc in $^3S_1$ (#AG-30) pucker; (D) of β-Glc in $^OS_2$ (#BG-80) pucker; at the reference level (cyan), ANI-1ccx (magenta), ANI-2x (blue) and GFN2-xTB (green). Transition state structures of (E) β-man and (F) β-Xyl in $^5E$ pucker at reference level of theory. Hydrogen bond distances marked (black dotted lines) and values indicated in Å, for reference geometry (bold) and in parentheses, for ANI-1ccx, ANI-2x and GFN2-xTB methods, respectively

For ANI-1ccx predictions at its optimized geometries, the MAE in relative energy of non-chair ring pucker ranges up to 2.4 kcal/mol, found for $^1H_2$ puckers (Table S1). For the highly populated equatorial region, it appears that boat conformers are slightly less well modeled than skewboat puckers (Table 3), with an average overall error of 1.4 kcal/mol for boat and 1.1 kcal/mol for skewboat structures. As an example, the lowest energy $B_{O,3}$ ring pucker of β-GlcNAc appears problematic, with an overstabilization via ANI-1ccx of 2.4 kcal/mol relative to the $^1C_4$ lowest energy conformer, such that the two structures are effectively degenerate (Figure 4A, B and Table S3). In general, however, ANI-1ccx performs well in modeling non-chair puckered conformer energetics, providing Δ*E* estimates typically to within 1–2 kcal/mol of reference values (Table 4).

For ANI-2x, overall the error is larger than for ANI-1ccx, with a MAE in Δ*E* for non-chair conformers of 1.6 and 1.9 kcal/mol at respective reference and optimized geometries (Table 3). The MAE in relative energies of the $^3S_1$ conformations appears particularly large, with values of 2.5 and 3.2 kcal/mol at reference and ANI-2x geometries, respectively (Table 3); as a potential factor, in a $^3S_1$ conformer of α-Glc (Figure 4C), we note the presence of three short hydrogen bonds involving axially oriented substituents, with H···O distances of 1.79, 1.85, and 1.97 Å in the ANI-2x geometry (Figure 4C). The relative energy of this conformer is reduced from 7.5 kcal/mol at the CCSD(T)/6-311+G(d,p)

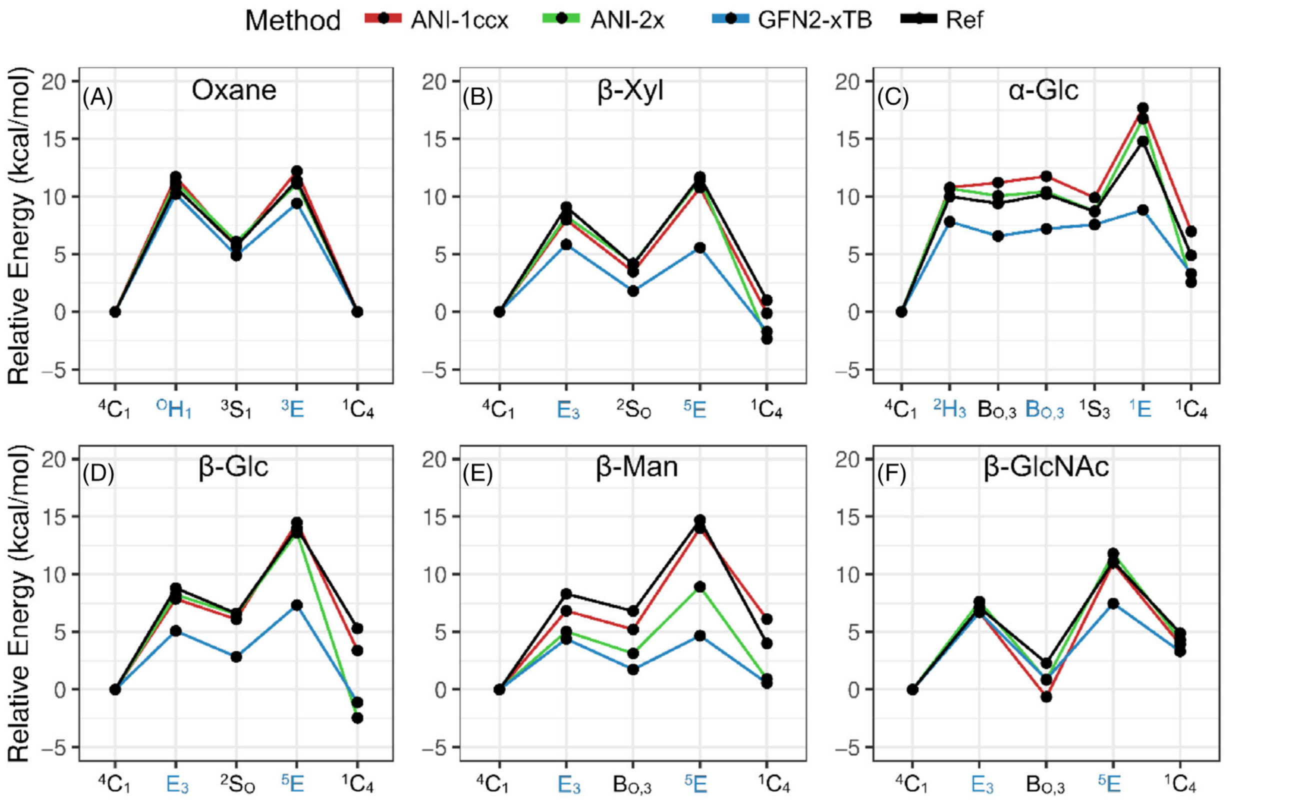

**FIGURE 5** Minimum energy conformer pathway from $^4C_1$ through transition states conformers (blue) to $^1C_4$ for oxane and five monosaccharide molecules, via CCSD(T)/6-311+G(d,p) level of theory (black), ANI-1ccx (red), ANI-2x (green) and GFN2-xTB (blue) methods

**TABLE 5** Mean RMSD in angle $\zeta$ over chair and non-chair conformers (in degrees), for molecule and across molecules (tot)

| | Chair | | | Non-chair | | |
|---|---|---|---|---|---|---|
| Mol | ANI-1ccx | ANI-2x | GFN2-xTB | ANI-1ccx | ANI-2x | GFN2-xTB |
| Oxane | 0.9 | 0.7 | 2.5 | 3.9 | 0.3 | 3.1 |
| β-Xyl | 2.6 | 1.4 | 1.7 | 4.9 | 8.2 | 15.0 |
| α-Glc | 2.2 | 2.1 | 2.9 | 7.0 | 8.9 | 6.5 |
| β-Glc | 4.4 | 3.6 | 2.8 | 4.0 | 4.3 | 7.4 |
| β-Man | 6.9 | 4.0 | 3.3 | 6.3 | 5.1 | 9.1 |
| β-GlcNAc | 5.1 | 4.9 | 2.1 | 5.7 | 6.4 | 5.1 |
| tot | 4.5 | 3.5 | 2.6 | 5.5 | 6.0 | 7.6 |

level, to 2.0 kcal/mol via ANI-2x (Table S4), thus incorrectly predicting this high energy conformation as thermally accessible at room temperature.

For GFN2-xTB, there is a larger still overall MAE in relative energy of non-chair conformers, with values of 3.9 and 4.0 kcal/mol for reference and relaxed geometries, respectively (Table 3). These errors in $\Delta E$ seem rather uniformly spread across molecules and ring puckers (Table 4), consistent with the systematic deviation indicated by the scatter plots of relative energies (Figure 2E,F). For the archetypal monosaccharide β-Glc, absolute deviations in $\Delta E$ for relaxed pucker conformers range from 2.8 kcal/mol for $^2S_O$ to 5.4 kcal/mol for $^{1,4}B$ (Table 4). An example of a non-chair structure problematic for GFN2-xTB is the $^OS_2$ conformer of β-Glc in Figure 4D, which has a $\Delta E$ that is underpredicted by 4.9 kcal/mol (Table S5). Even for the basic ring model, oxane, we note that the relative energies of skewboat conformers are underpredicted via GFN2-xTB, by 0.8 kcal/mol or more (Table 4), pointing to an underlying issue with ring strain. This and other factors are discussed in more detail in Section 3.5.

## 3.3 | Minimum energy pathways

We note that envelope conformers are only infrequently found as minima in the Mayes et al. set dataset; the five envelope conformers considered here are reproduced with reasonable accuracy by the ANI-1ccx, ANI-2x and GFN2-xTB models, with mean absolute errors in optimized $\Delta E$ of 1.5, 2.2, and 3.5 kcal/mol, respectively (Table S9).

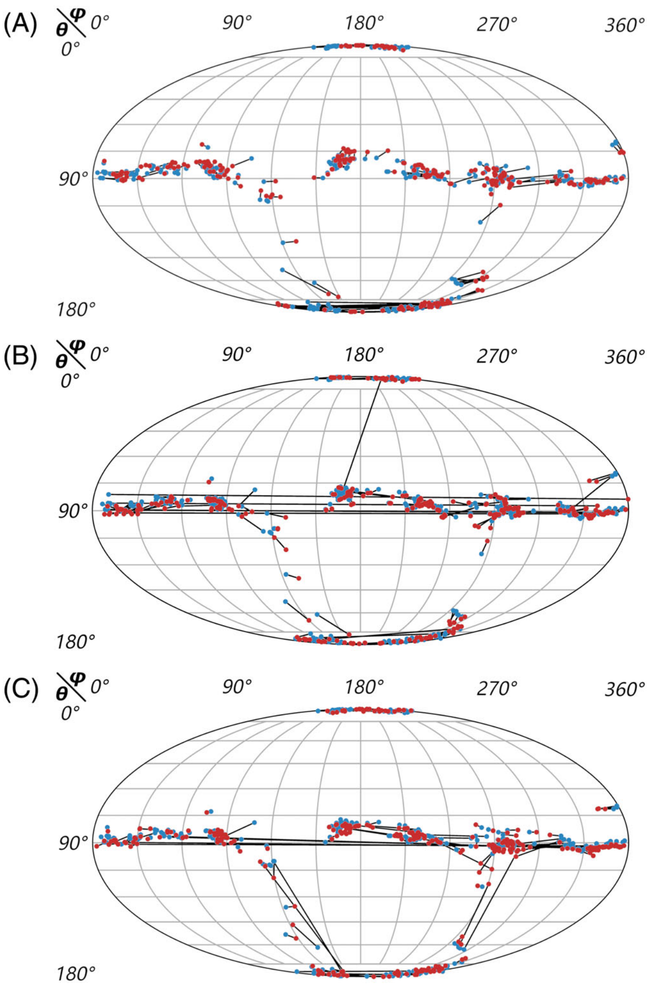

**FIGURE 6** Cremer–Pople puckering hypersurface of angles $\theta\varphi$ (in degrees) computed on reference B3LYP/6-311+G(2df,p) (blue) and (A) ANI-1ccx (red), (B) ANI-2x (red) and (C) GFN2-xTB (red) levels of theory geometries. Black lines indicate puckering change on optimization

However, the envelope conformers do appear as transition states along the minimum energy pathway from $^4C_1$ to $^1C_4$ via the boat/skewboat equator of the Cremer-Pople hypersurface. Such a pathway is of particular interest in computational glycobiology, sampled to various extents in the course of catalysis by some enzymes. Although the characterization of transition states is beyond the scope of this present study, we do assess the minimum energy itinerary from $^4C_1$ to $^1C_4$ for each of the five monosaccharides and oxane, using the reference minima and transition state structures of the Mayes et al. dataset.

For each of the six molecules, we find that the ANI-1ccx model closely follows the CCSD(T)/6-311+G(d,p) energetics of the minimum energy pathway (Figure 5), for both transition states and minima. The MAE in conformational energy across the six energy profiles is 0.9 kcal/mol; the highest error is found for the $B_{O,3}$ conformer of β-GlcNAc, with an overstabilization of 2.9 kcal/mol (Figure 5F). ANI-2x performs comparably with ANI1-1ccx in most cases, with an overall MAE in Δ*E* of 1.2 kcal/mol across profiles. The largest deviation is observed for the $^5E$ conformer of β-Man (Figure 4E), with an overestimated stability by 5.8 kcal/mol (Figure 5E), although an error of only 0.2 kcal/mol is found for the β-Xyl $^5E$ conformer (Figures 4F and 5B). Interestingly, the two $^5E$ structures differ in that the β-Man structure possesses several short hydrogen bonds (Figure 4E); however the β-Xyl conformer lacks this hydrogen bond network due to its pseudo-axial 2-OH substituent (Figure 4F).

Finally, GFN2-xTB is found to systematically overestimate the stability of conformers relative to $^4C_1$ along the minimum energy pathway with an overall MAE in Δ*E* of 2.6 kcal/mol; the lowest errors are found for the pathway of the undecorated oxane molecule (Figure 5A). As with ANI-2x, the GFN2-xTB method exhibits the largest error for the $^5E$ conformer of β-Man, with a value of 10.0 kcal/mol (Figure 5E). The potential over-preference for hydrogen bonds via ANI-2x and GFN2-xTB suggested by these results is considered in more detail in Section 3.5.

### 3.4 | Geometry of ring pucker conformers

To assess the effect of geometry optimization by the ML and SQM methods on pucker conformation, we define angle $\zeta$ as the distance in $\theta\varphi$ space between conformers before and after relaxation. We find that the chair puckered minima of the six molecules are retained on geometry optimization via the ANI-1ccx approach: the average RMSD in $\zeta$ on optimization by the ANI-1ccx model is 4.5° (Table 5). On inspection of ANI-1ccx conformers on Cremer–Pople hypersurface, we indeed observe that energy minimization at this level induces only local changes to pucker coordinates in the polar chair regions (Figure 6A). The chair conformers are also well reproduced by ANI-2x and GFN2-xTB, where the average RMSD in $\zeta$ is slightly lower than for ANI-1ccx, with values of 3.5° and 2.6° (Table 5).

For non-chair conformers, in most cases, there is generally a small change in structure on optimization; these shifts are reflected by RMSDs in angle $\zeta$ of 5.5°, 6.0° and 7.6° for ANI-1ccx, ANI-2x and GFN2-xTB, respectively (Table 5); across puckers, the corresponding RMSD in Cartesian coordinates do not exceed 0.2, 0.5, and 0.5 Å (Table S10). The most notable change in pucker with ANI-1ccx is for a $^OS_2$ conformation of β-Man, which adjusts to a $^1S_5$ conformation (Figures 7A); the relative energy changes from 8.1 to 6.7 kcal/mol on this transition (Table S6).

Inspection of the puckering hypersurface for ANI-2x and GFN2-xTB, however, indicates several major shifts in stationary point, with instances of non-chair to chair transitions (Figure 6B,C). There are also shifts around the equator of the hypersurface: for ANI-2x, a significant change in pucker is observed for the $^2S_O$ conformation of α-Glc, which shifts to a $^1S_3$ pucker (Figure 7B), with a reduction in Δ*E* of 12.6–7.5 kcal/mol (Table S4).

For the GFN2-xTB method, there are more numerous large shifts in pucker geometry (Figure 6C). For example, for β-Xyl, there is a transition from $^{1,4}B$ to $^1S_3$ pucker on geometry optimization, a change not found for either ANI model. The energy of this structure via GFN2-xTB changes from 6.2 to 2.2 kcal/mol on minimization

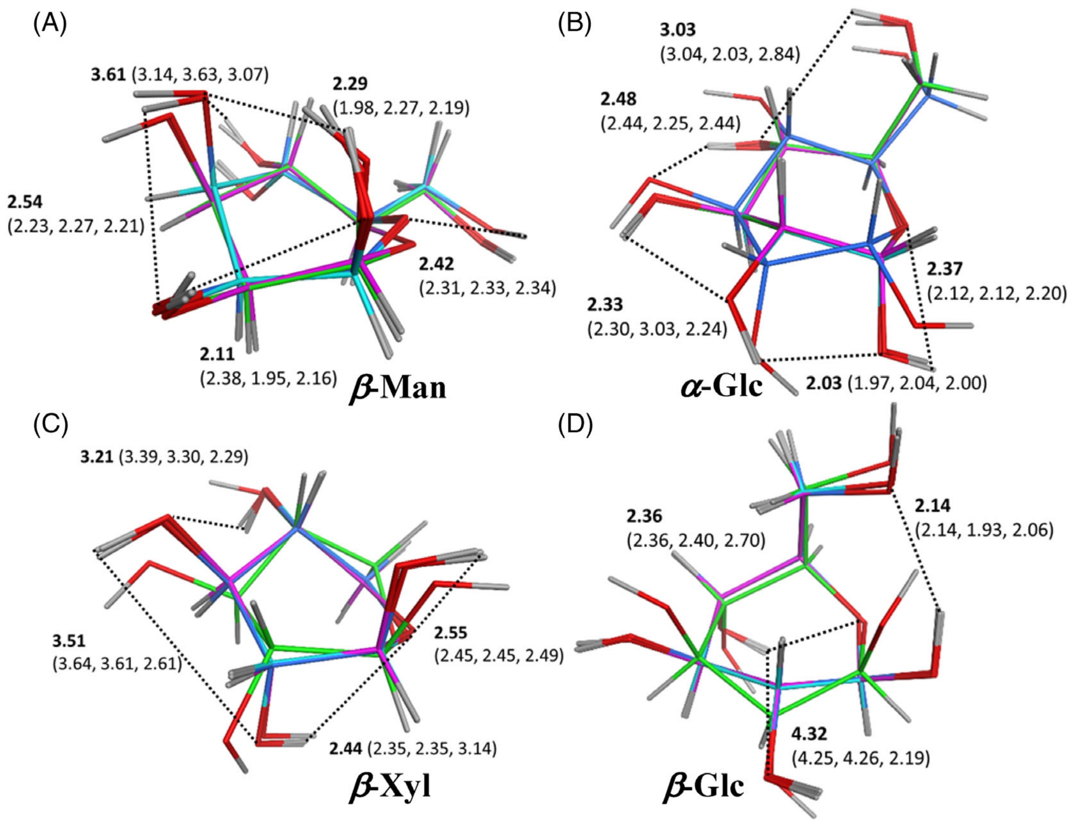

**FIGURE 7** Optimized structures of (A) of β-Man in $^{O}S_2$ (conformer id #BM-54) pucker; (B) of α-Glc in $^{2}S_O$ (#AG-26) pucker; (C) of β-Xyl in $^{1,4}B$ (#BX-1) pucker; (D) of β-Glc in $^{2,5}B$ (#BG-44) pucker; at the reference geometry (cyan), ANI-1ccx (magenta), ANI-2x (blue) and GFN2-xTB (green). Hydrogen bond distances marked (black dotted lines) and values indicated in Å, for the reference geometry (bold) and in parentheses, for ANI-1ccx, ANI-2x and GFN2-xTB methods, respectively

(Table S7), with formation of a O4H···O2 hydrogen bond of distance 2.3 Å (Figure 7C). For β-Glc, a $^{2,5}B$ conformer shifts to a $^{1}C_4$ structure on optimization at the GFN2-xTB level (Figure 7D). The relative energy of the optimized conformer is underestimated by 8.4 kcal/mol compared to the reference coupled-cluster value (Table S5). This again reflects the tendency of GFN2-xTB to overestimate non-chair pucker stability.

## 3.5 | Analysis of model compounds

In order to discern factors contributing to errors in prediction of pucker landscapes for the ML and GFN2-xTB methods, we examine the relative energetics of three model compounds: glycerol, 4,6-dimethyloxane and oxan-2-ol (Figure 8). For glycerol, we estimate the strength of an axial-axial (ax–ax) hydrogen bond, $E^{\text{HB}}_{\text{ax}}$, as the energetic difference between conformers (A) and (B) in Figure 8; and for an equatorial–equatorial (eq–eq) hydrogen bond, $E^{\text{HB}}_{\text{eq}}$, we compare the energies of conformers (C) and (D) in Figure 8. We compute the DLPNO-CCSD(T)/CBS energy for these conformers using geometries obtained at the B3LYP/6-311+G(2df,2p) level of theory.

Using this approach, we find that the ax–ax hydrogen bond is 2.0 kcal/mol more favorable than an eq–eq hydrogen bond at the reference level of theory ($\Delta E^{\text{HB}}_{\text{ax-eq}}$, Table 6). However, the value of $\Delta E^{\text{HB}}_{\text{ax-eq}}$ computed using the ANI-1ccx energy and geometry is only 0.6 kcal/mol (Table 6). For ANI-2x and GFN2-xTB, the preference for an ax–ax hydrogen bond is overestimated, with $\Delta E^{\text{HB}}_{\text{ax-eq}}$ values at relaxed geometries of 2.6 and 2.8 kcal/mol, respectively (Table 6). Indeed, the latter two methods overestimate the strength of the ax–ax hydrogen bond, $\Delta E^{\text{HB}}_{\text{ax}}$, whereas ANI-1ccx underestimates its strength by 1.2 kcal/mol (Table 6).

The second model molecule, 4,6-dimethyloxane, enables prediction of the *syn*-diaxial repulsion energy between two axially oriented methyl groups (Figure 8E,F). This represents something of an upper limit estimate for the type of *syn*-diaxial interactions featuring in the monosaccharides of this study. The DLPNO-CCSD(T)/CBS reference energy for the repulsion in 4,6-dimethyloxane is 4.3 kcal/mol ($\Delta E^{\text{steric}}_{\text{ax-eq}}$, Table 6); the ANI-1ccx and ANI-2x methods are in reasonable agreement with this value, giving $\Delta E^{\text{steric}}_{\text{ax-eq}}$ values at their relaxed geometries of 4.6 and 3.9 kcal/mol, respectively (Table 6). However, the value of $\Delta E^{\text{steric}}_{\text{ax-eq}}$ computed via GFN2-xTB is 2.9 kcal/mol, underestimating the inter-methyl repulsion by 1.4 kcal/mol (Table 6).

The third model molecule, oxan-2-ol, provides an estimate of the strength of the *endo*-anomeric effect, a stereoelectronic effect in carbohydrates that favors axial electronegative substituents at the C1 position of the pyranose ring. At the reference level, a

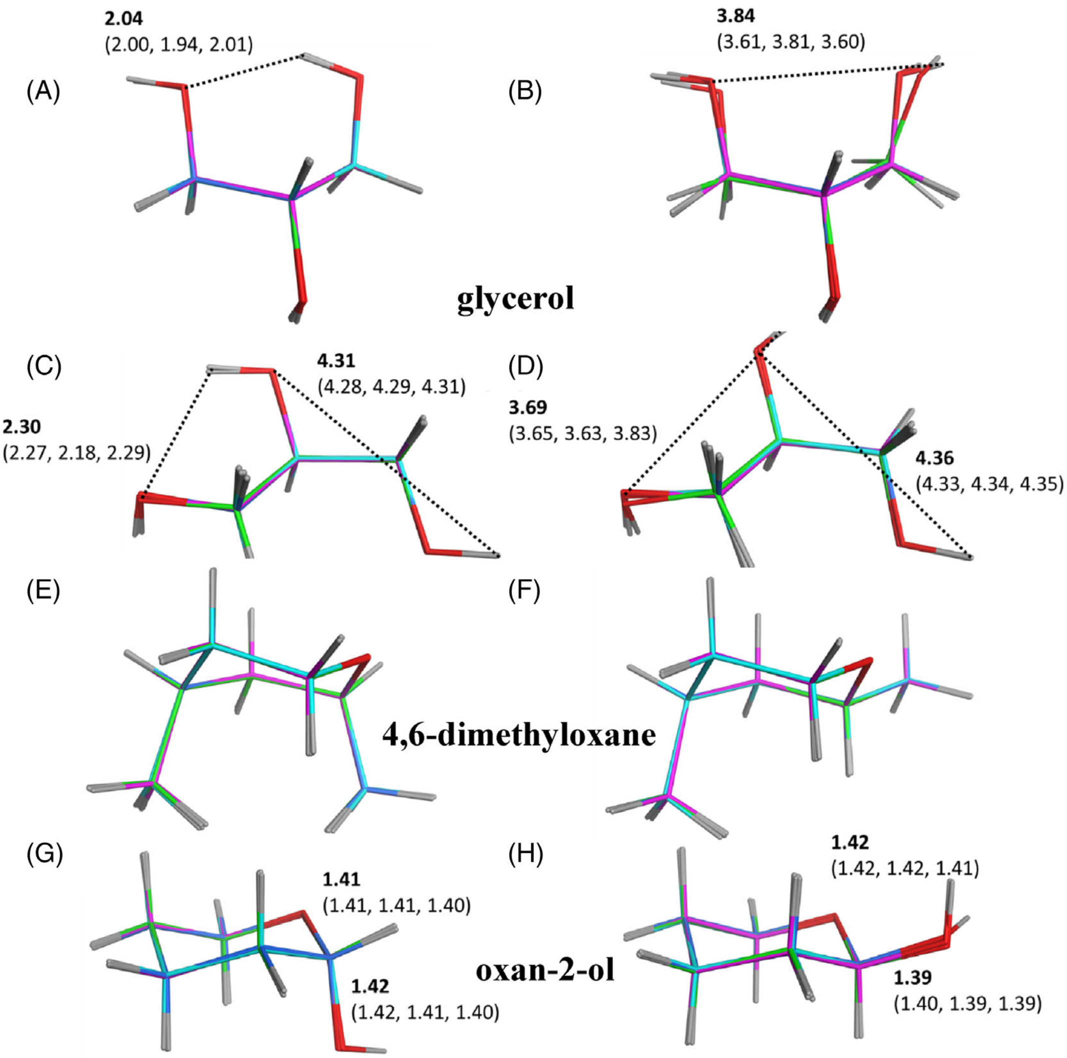

**FIGURE 8** Optimized conformers of glycerol with (A) and (B) without axial–axial hydrogen bond; or (C) with and (D) without equatorial–equatorial hydrogen bond. Optimized conformers of 4,6-dimethyloxane (E) with and (F) without *syn*-diaxial methyl repulsion present. Optimized conformers of oxan-2-ol with (C) axial and (B) equatorial 1-OH group. Geometries at B3LYP/6-311++G(2df,2p) level (cyan), ANI-1ccx (magenta), ANI-2x (blue) and GFN2-xTB (green). Distance values in Å, for reference geometry (bold) and in parentheses, for ANI-1ccx, ANI-2x and GFN2-xTB methods, respectively

**TABLE 6** Computed estimates of intramolecular hydrogen bond energy between axial ($E^{HB}_{ax}$) and equatorial groups ($E^{HB}_{eq}$), and energy difference between them, $\Delta E^{HB}_{ax-eq}$; steric energy associated with *syn*-diaxial clash of methyl groups, $\Delta E^{steric}_{ax-eq}$; and energy of *endo*-anomeric effect, $\Delta E^{AE}_{ax-eq}$

| Method | $E^{HB}_{ax}$ | $E^{HB}_{eq}$ | $\Delta E^{HB}_{ax-eq}$ | $\Delta E^{steric}_{ax-eq}$ | $\Delta E^{AE}_{ax-eq}$ |
|---|---|---|---|---|---|
| Ref. | −6.0 | −4.0 | −2.0 | 4.3 | −1.5 |
| ANI-1ccx | −5.0 (−4.8) | −4.4 (−4.2) | −0.6 (−0.6) | 4.7 (4.6) | −3.4 (−3.5) |
| ANI-2x | −7.5 (−7.6) | −5.1 (−5.0) | −2.4 (−2.6) | 3.9 (3.9) | −3.4 (−2.5) |
| GFN2-xTB | −6.6 (−6.3) | −3.6 (−3.5) | −3.0 (−2.8) | 2.9 (2.9) | −0.7 (−0.7) |

*Note*: Energy values based on calculations for model compounds glycerol, 4,6-dimethyloxane and oxan-2-ol chair conformers, via reference (DLPNO-CCSD(T)/CBS), ANI-1ccx, ANI-2x and GFN2-xTB methods.

$\Delta E^{AE}_{ax-eq}$ value of −1.5 kcal/mol is found, favoring the axial orientation of the 1-OH group as expected. For both ANI-1ccx and ANI-2x, the *endo*-anomeric effect seems overestimated at their optimized geometries, with this energy difference increased to −3.5 and −2.5 kcal/mol, respectively (Table 6); and for GFN2-xTB, the effect is underestimated and the energy difference is predicted as only −0.7 kcal/mol (Table 6).

In general terms, these discrepancies in model prediction appear to imply that the ANI-1ccx potential would favor carbohydrate conformers with equatorial substituents that form eq–eq hydrogen bonds, that do not feature *syn*-diaxial steric repulsion but that possess an axial 1-OH substituent. Correspondingly, the ANI-1ccx method overestimates the stability of the α-Glc conformer, with its axial 1-OH group, relative to the β-anomer by 2.0 kcal/mol. Also according to these model calculations, ANI-2x would also seem to favor axial 1-OH groups but prefer geometries with ax–ax hydrogen bonds. Finally, GFN2-xTB, while underestimating the *endo*-anomeric effect in oxan-2-ol, would nevertheless appear to strongly favor ax–ax hydrogen bonds and underestimate associated *syn*-diaxial repulsions. We note that the anisotropic electrostatic energy term introduced in the GFN2-xTB model allows for improved accuracy of hydrogen bonding compared to the monopolar approximation of the preceding GFN-xTB model, with hydrogen bond strengths overestimated by ~1 kcal/mol on average for the hydrogen bonded bimolecular complexes of the S66 data set.[15] For the dispersion-dominated complexes of this set, the interaction potential energy was in general slightly overestimated by an error on the order of a kcal/mol.[15]

Indeed, these features do appear to correlate with the observed preferences for chair conformers of β-Glc, β-Xyl, and α-Glc (Figure 3). For β-Glc, the $^4C_1$ conformer (Figure 3A) is preferred by 5.0 kcal/mol over the $^1C_4$ conformer (Figure 3B), using reference geometries and energies (Table S5). However, the corresponding Δ*E* values for ANI-1ccx, ANI-2x and GFN-xTB are 4.2, 0.0 and −0.8 kcal/mol respectively on optimization (Table S5). The $^1C_4$ conformer features an axial 1-OH group (Figure 3B) and thus would be expected to be overstabilized by the ANI methods. Furthermore, the presence of additional ax–ax hydrogen bonds and *syn*-diaxial repulsions in the $^1C_4$ form would lead to overestimation of the stability of this conformer by ANI-2x and GFN2-xTB methods.

For β-Xyl chair conformers (Figure 3C,D), the agreement between the reference level and other methods is somewhat improved: the reference Δ*E* of 1.0 kcal/mol compares to values of −0.5, −2.4, and −1.2 kcal/mol computed by ANI-1ccx, ANI-2x, and GFN2-xTB respectively at their relaxed geometries (Table S7). β-Xyl lacks the $CH_2OH$ group of β-Glc and this results in one less axial substituent in the $^1C_4$ conformer, with fewer hydrogen bonds and reduced *syn*-diaxial repulsion (Figure 3D). Thus, overstabilization of the $^1C_4$ conformer is less pronounced by ANI-2x and by GFN2-xTB.

Finally, for α-Glc conformers (Figure 3E,F), the reference level of theory predicts the $^1C_4$ pucker as 4.9 kcal/mol higher in energy than the $^4C_1$ conformer, whereas ANI-1ccx, ANI-2x, and GFN2-xTB predict values of 7.4, 2.5, and 3.7 kcal/mol, respectively (Table S4). Compared to β-Glc, the one less axial substituent in the $^1C_4$ structure of α-Glc leads to closer agreement of ANI-2x and GFN2-xTB with the reference estimate. However, for ANI-1ccx, the $^4C_1$ conformer of α-Glc is overstabilized by 2.5 kcal/mol relative to $^1C_4$, at least in part due to the exaggerated influence of the *endo*-anomeric effect in the $^4C_1$ structure. Finally, we note that similar arguments can assist in understanding the observed preference of non-chair conformers, for example the overstabilization of (i) the $^3S_1$ conformer of α-Glc by ANI-2x (Figure 4C); and (ii) the $^5E$ conformer of β-Man by GFN2-xTB (Figure 4E); although we note that, due to the nature of the non-chair puckers, the definition of axial and equatorial orientation is more nuanced.

## 4 | CONCLUSIONS

In this assessment of machine learning models and SQM for describing the ring pucker of monosaccharides in the gas phase, we find that the highest correlation in predicted relative energetics is obtained via the ANI-1ccx model, with a $r^2$ of 0.83 using optimized geometries at that level (Figure 2B). Chair and non-chair geometries are generally well reproduced; relatively small changes in pucker location on the Cremer–Pople equator are found on geometry optimization, the largest being for β-Man in a $^OS_2$ conformation. It is perhaps unsurprising that ANI-1ccx agrees most closely with the reference coupled-cluster level of theory, as indeed ANI-1ccx was fitted to ~500 k estimates of molecular energy at the DLPNO-CCSD(T)/CBS level of theory; these structures were selected via active learning from a larger ANI-1x data set, within which there is representation of ring systems, although carbohydrates did not appear specifically featured. ANI-1ccx is able to reproduce pyranose ring energetics generally well: for the most basic ring model, oxane, the $^2S_O$ and $^3S_1$ puckers are accurate to within 0.1 kcal/mol, although the $^1S_5$ conformer deviates by 1.1 kcal/mol (Table 4). Further work with an expanded training set to include suitable model molecules may be required to more accurately incorporate anomeric effects into the method.

For ANI-2x, a lower correlation coefficient of 0.70 is found for relative energy predictions across the 299 conformer set, with correspondingly higher errors in chair and non-chair energies. The method has a slightly improved prediction of chair geometries over ANI-1ccx but reproduces non-chair geometries less well. We note that the ANI-2x model was fitted using ωB97X/6-31G* energies and geometries. The use of diffuse functions in geometries can improve predictions of hydrogen bond strength[33] and have been shown to be important in accurate estimates within carbohydrates via density functionals with double-zeta or triple-zeta basis sets.[20] Thus, a number of high energy ring puckered geometries appear to be incorrectly predicted as thermally accessible using ANI-2x (Figure 2D). For ANI-1ccx, which is also fitted using ωB97X/6-31G* geometries, the use of a high level of theory for the energy calculation may mitigate this effect.

We also note that improving hydrogen bonding has been a subject of research for density functional-based tight binding models.[34,35] The most recent GFN2-xTB implementation appears to do well in predicting hydrogen bonding strengths in the S22 and S66 datasets, which were used in its fitting.[36] In deriving parameters for this method, extrapolated CCSD(T) energies were typically used, although structures were largely computed by the composite PBEh-3c or B97-3c functionals. In the present work, we find evidence of overweighted ax–ax hydrogen bonds and underestimated steric strain, which combine to lead to systematic problems in modeling the spectrum of ring puckers associated with monosaccharides. These low energy puckers, along with the chair-to-chair energy profiles in Figure 5, likely point to an overly accessible puckering energy landscape via a DFTB approach, an observation in accord with the study of glucose and maltose conformation by Marianski et al.[16] It is clear from the overstabilization of non-chair oxane conformers (Table 4) that there appears to be a fundamental issue with ring strain in chair versus non-chair conformers; the underestimation of *syn*-diaxial repulsion by GFN2-xTB, as commonly occurs in strained boat, skewboat, halfchair and envelope structures, is further illustrated by the 4,6-dimethyloxane conformers (Figure 8E,F).

As a machine learned potential, we observe that the ANI-1ccx model is rather robust in its ability to reproduce in vacuo pyranose ring pucker conformations and the minimum energy chair-to-chair itineraries on the puckering hypersurface. Furthermore, the computational expense of ANI-1ccx has been estimated to be many orders of magnitude faster than CCSD(T) calculations and comparable to GFN2-xTB calculations.[26,27] However, potential challenges remain in applying the approach to larger carbohydrates. ANI potentials may be less suited to simulating systems where long range electrostatic interactions play a significant role[37] as a consequence of its modified symmetry functions used to describe atomic environments; this could be an issue in computing the conformational behavior of saccharide residues across glycosidic linkages. Furthermore, in the absence of an inherent charge model, coupling of these ML methods to a condensed phase environment, as required for the pursuit of computational glycobiology, is a matter of ongoing research. The interaction with aqueous solvent will profoundly affect the pucker distributions of the monosaccharides considered in this study; obtaining an accurate balance between intramolecular and intermolecular hydrogen bonding is crucial to studies in solution, as well as the computation of protein binding affinities and enzyme reaction energetics. However, in this regard, we note recent and promising advances in combined ML/force field implementations that explore coupling of the ANI method to a condensed phase MM environment.[37–39]

## ACKNOWLEDGMENTS

The authors acknowledge the use of the Computational Shared Facility at the University of Manchester.

## DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

## ORCID

*Richard A. Bryce* https://orcid.org/0000-0002-8145-2345

## REFERENCES

[1] A. Varki, *Glycobiology* **2016**, *27*, 3.

[2] J. C. Dyason, M. von Itzstein, "Viral surface glycoproteinsin carbohydrate recognition:structure and modelling" In "*Microbial Glycobiology: Structures, Relevance and Application*", A. P. Moran, O. Holst, P. J. Brennan, M. Von Itzstein Eds., Elsevier, San Diego **2009**, p. 269

[3] L. Casalino, Z. Gaieb, J. A. Goldsmith, C. K. Hjorth, A. C. Dommer, A. M. Harbison, C. A. Fogarty, E. P. Barros, B. C. Taylor, J. S. McLellan, *ACS Centr. Sci.* **2020**, *6*, 1722.

[4] M. L. DeMarco, R. J. Woods, *Glycobiology* **2008**, *18*, 426.

[5] D. T. Cremer, J. Pople, *J. Am. Chem. Soc.* **1975**, *97*, 1354.

[6] G. Speciale, A. J. Thompson, G. J. Davies, S. J. Williams, *Curr. Opin. Struct. Biol.* **2014**, *28*, 1.

[7] S. Faham, R. Hileman, J. Fromm, R. Linhardt, D. Rees, *Science* **1996**, *271*, 1116.

[8] E. Sabini, G. Sulzenbacher, M. Dauter, Z. Dauter, P. L. Jørgensen, M. Schülein, C. Dupont, G. J. Davies, K. S. Wilson, *Chem. Biol.* **1999**, *6*, 483.

[9] I. Alibay, K. K. Burusco, N. J. Bruce, R. A. Bryce, *J. Phys. Chem. B* **2018**, *122*, 2462.

[10] K. N. Kirschner, A. B. Yongye, S. M. Tschampel, J. González-Outeiriño, C. R. Daniels, B. L. Foley, R. J. Woods, *J. Comput. Chem.* **2008**, *29*, 622.

[11] I. Alibay, R. A. Bryce, *J. Chem. Inf. Model.* **2019**, *59*, 4729.

[12] B. L. Foley, M. B. Tessier, R. J. Woods, *WIREs Comput. Mol. Sci.* **2012**, *2*, 652.

[13] C. B. Barnett, K. J. Naidoo, *J. Phys. Chem. B* **2010**, *114*, 17142.

[14] J. P. McNamara, A.-M. Muslim, H. Abdel-Aal, H. Wang, M. Mohr, I. H. Hillier, R. A. Bryce, *Chem. Phys. Lett.* **2004**, *394*, 429.

[15] C. Bannwarth, S. Ehlert, S. Grimme, *J. Chem. Theory Comput.* **2019**, *15*, 1652.

[16] M. Marianski, A. Supady, T. Ingram, M. Schneider, C. Baldauf, *J. Chem. Theory Comput.* **2016**, *12*, 6157.

[17] C. Riplinger, F. Neese, *J. Chem. Phys.* **2013**, *138*, 034106.

[18] L. Goerigk, A. Hansen, C. Bauer, S. Ehrlich, A. Najibi, S. Grimme, *Phys. Chem. Chem. Phys.* **2017**, *19*, 32184.

[19] L. Goerigk, S. Grimme, *J. Chem. Theory Comput.* **2010**, *6*, 107.

[20] G. I. Csonka, A. D. French, G. P. Johnson, C. A. Stortz, *J. Chem. Theory Comput.* **2009**, *5*, 679.

[21] J. Behler, *J. Chem. Phys.* **2016**, *145*, 170901.

[22] J. S. Smith, O. Isayev, A. E. Roitberg, *Chem. Sci.* **2017**, *8*, 3192.

[23] J. Behler, M. Parrinello, *Phys. Rev. Lett.* **2007**, *98*, 146401.

[24] J. S. Smith, B. Nebgen, N. Lubbers, O. Isayev, A. E. Roitberg, *J. Chem. Phys.* **2018**, *148*, 241733.

[25] C. Devereux, J. S. Smith, K. K. Huddleston, K. Barros, R. Zubatyuk, O. Isayev, A. E. Roitberg, *J. Chem. Theory Comput.* **2020**, *16*, 4192.

[26] J. S. Smith, B. T. Nebgen, R. Zubatyuk, N. Lubbers, C. Devereux, K. Barros, S. Tretiak, O. Isayev, A. E. Roitberg, *Nat. Commun.* **2019**, *10*, 1.

[27] D. Folmsbee, G. Hutchison, *Int. J. Quantum Chem.* **2021**, *121*, e26381.

[28] H. B. Mayes, L. J. Broadbelt, G. T. Beckham, *J. Am. Chem. Soc.* **2014**, *136*, 1008.

[29] A. Halkier, T. Helgaker, P. Jørgensen, W. Klopper, J. Olsen, *Chem. Phys. Lett.* **1999**, *302*, 437.

[30] F. Neese, *WIREs Comput. Mol. Sci.* **2012**, *2*, 73.

[31] X. Gao, F. Ramezanghorbani, O. Isayev, J. S. Smith, A. E. Roitberg, *J. Chem. Inf. Model.* **2020**, *60*, 3408.

[32] C. Bannwarth, E. Caldeweyher, S. Ehlert, A. Hansen, P. Pracht, J. Seibert, S. Spicher, S. Grimme, *WIREs Comput. Mol. Sci.* **2021**, *11*, e1493.

[33] L. A. Curtiss, P. C. Redfern, K. Raghavachari, *J. Chem. Phys.* **2005**, *123*, 124107.

[34] A. S. Christensen, M. Elstner, Q. Cui, *J. Chem. Phys.* **2015**, *143*, 084123.

[35] J. Rezac, *J. Chem. Theory Comput.* **2017**, *13*, 4804.

[36] S. Grimme, C. Bannwarth, P. Shushkov, *J. Chem. Theory Comput.* **2017**, *13*, 1989.

[37] S.-L. J. Lahey, C. N. Rowley, *Chem. Sci.* **2020**, *11*, 2362.

[38] D. A. Rufa, H. E. B. Macdonald, J. Fass, M. Wieder, P. B. Grinaway, A. E. Roitberg, O. Isayev, J. D. Chodera, *BioRxiv* **2020**. https://doi.org/10.1101/2020.07.29.227959

[39] T. J. Inizan, T. Plé, O. Adjoua, P. Ren, H. Gökcan, O. Isayev, L. Lagardère, J.-P. Piquemal, *arXiv* **2022**. https://doi.org/10.48550/arXiv.2207.14276

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** L. Kong, R. A. Bryce, *J. Comput. Chem.* **2022**, *43*(30), 2009. https://doi.org/10.1002/jcc.27000