SUPPORTING INFORMATION
General Graph Neural Network Based Implicit Solvation Model for Organic Molecules in Water

Paul Katzberger, Sereina Riniker*

[*] *Department of Chemistry and Applied Biosciences, ETH Zurich, Vladimir-Prelog-Weg 2, 8093 Zurich, Switzerland. Email: sriniker@ethz.ch*

Additional Tables and Figures

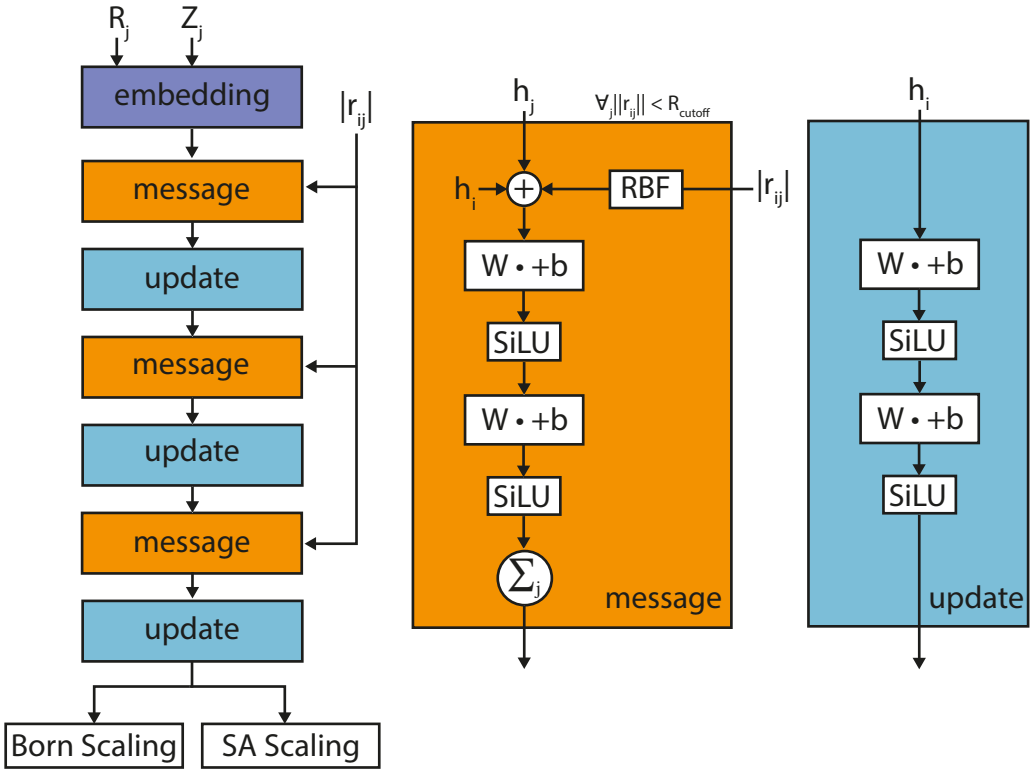

Figure S1: Schematic representation of the GNN architectures based on Ref. 1. Node-wise computations are shown in blue, while message operations are shown in orange. The message send from each node to all other nodes within a cutoff $\mathrm{R_{cutoff}}$ first concatenates the atom features of the sending and receiving node with the distance encoded by a Bessel function (RBF), and passes the features to a two-layer MLP with SiLU activation functions. Next, the messages are aggregated by summation via node-wise computations and passed to a two-layer MLP with SiLu activation functions. This process is repeated three times and the final output of shape 2xN is passed to the Born and SA scaling functions (i.e., a separate scaling value is predicted for the Born and the SA term per atom in a multitask fashion). The calculation of the final energy is performed in analogy to the classical SAGB model (not shown) and the corresponding forces are obtained by taking the derivative with respect to the positions.

Seed	Hidden layer size	MAE $[\mathrm{kJ\ mol^{-1}\ nm^{-1}}]$	RMSE $[\mathrm{kJ\ mol^{-1}\ nm^{-1}}]$	R^2	Pearson
1	48	5.311905	9.080699	0.993304	0.996650
1	64	5.065377	8.609815	0.993980	0.996995
1	96	4.784358	8.055551	0.994730	0.997389
1	128	4.585881	7.685227	0.995204	0.997616
2	48	5.338575	9.122822	0.993242	0.996628
2	64	5.091256	8.650021	0.993924	0.996973
2	96	4.763875	7.993013	0.994812	0.997413
2	128	4.579104	7.651880	0.995245	0.997630
3	48	5.350449	9.169399	0.993172	0.996602
3	64	5.080727	8.637758	0.993941	0.996977
3	96	4.775831	8.005622	0.994795	0.997405
3	128	4.593868	7.637986	0.995263	0.997633

Table S1: Performance of four GNN architectures with different hidden layer size on the external test set: mean absolute error (MAE), root-mean-square error (RMSE), R^2, and Pearson correlation coefficient.

Solvent model	Parallel simulations	Cumulative time [ns/day]
TIP3P	1	937
GNN 48	256	5'017
GNN 64	256	4'122
GNN 96	256	3'021
GNN 128	256	2'348
Vaccuum	256	304'640

Table S2: Simulation times of compound C3 simulated using a NVIDIA RTX 3090 GPU.

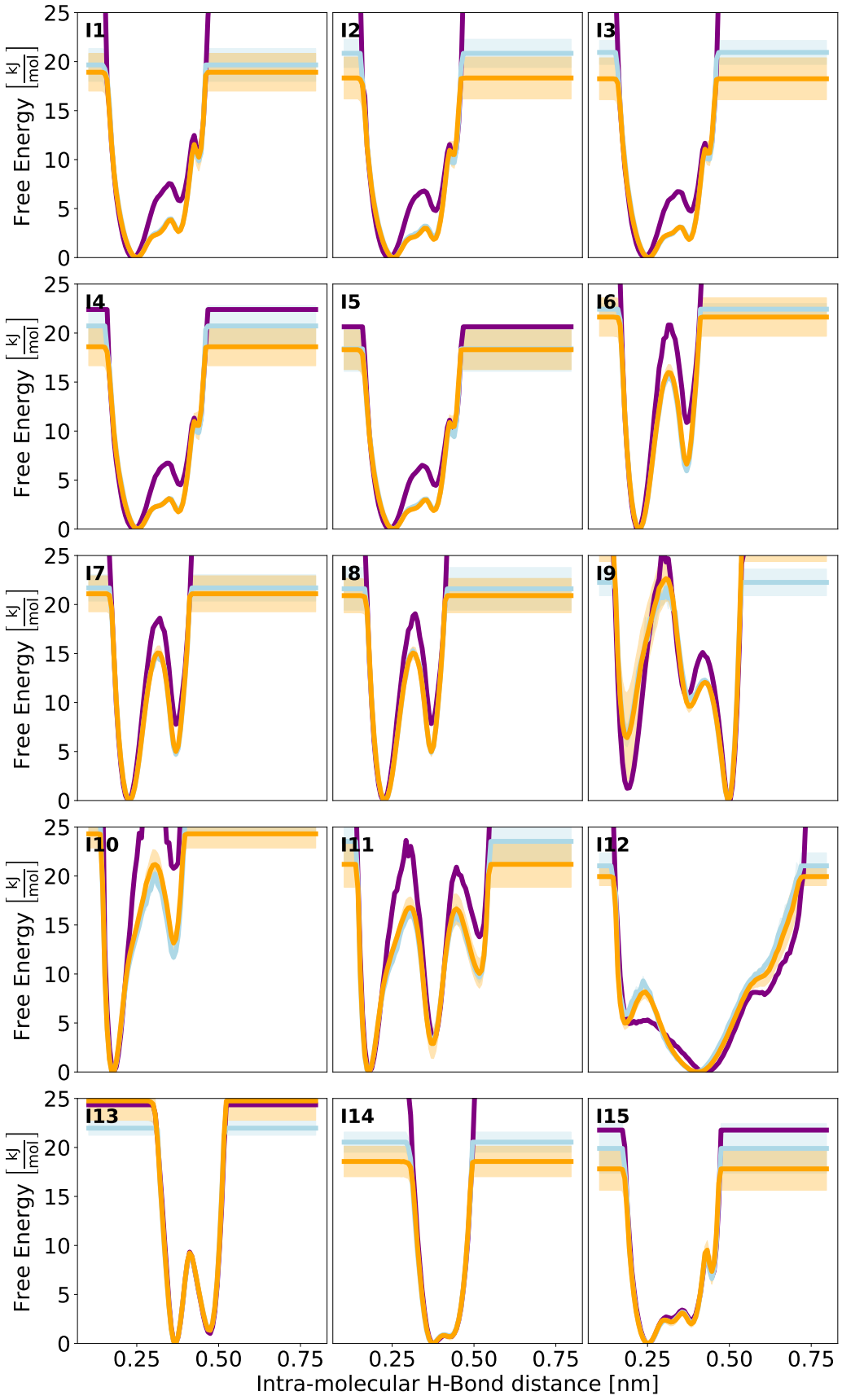

Figure S2: Free-energy profiles of the opening of the intramolecular hydrogen bonds in the compounds of set **I**: Comparison of the GNN implicit solvent model (orange, 128x5 ns) with the explicit solvent TIP3P model (blue, 1x500 ns) and the baseline implicit solvent GB-Neck2 model (purple, 3x500 ns). The studied compound is indicated by its identifier in the upper left corners of the individual plots.

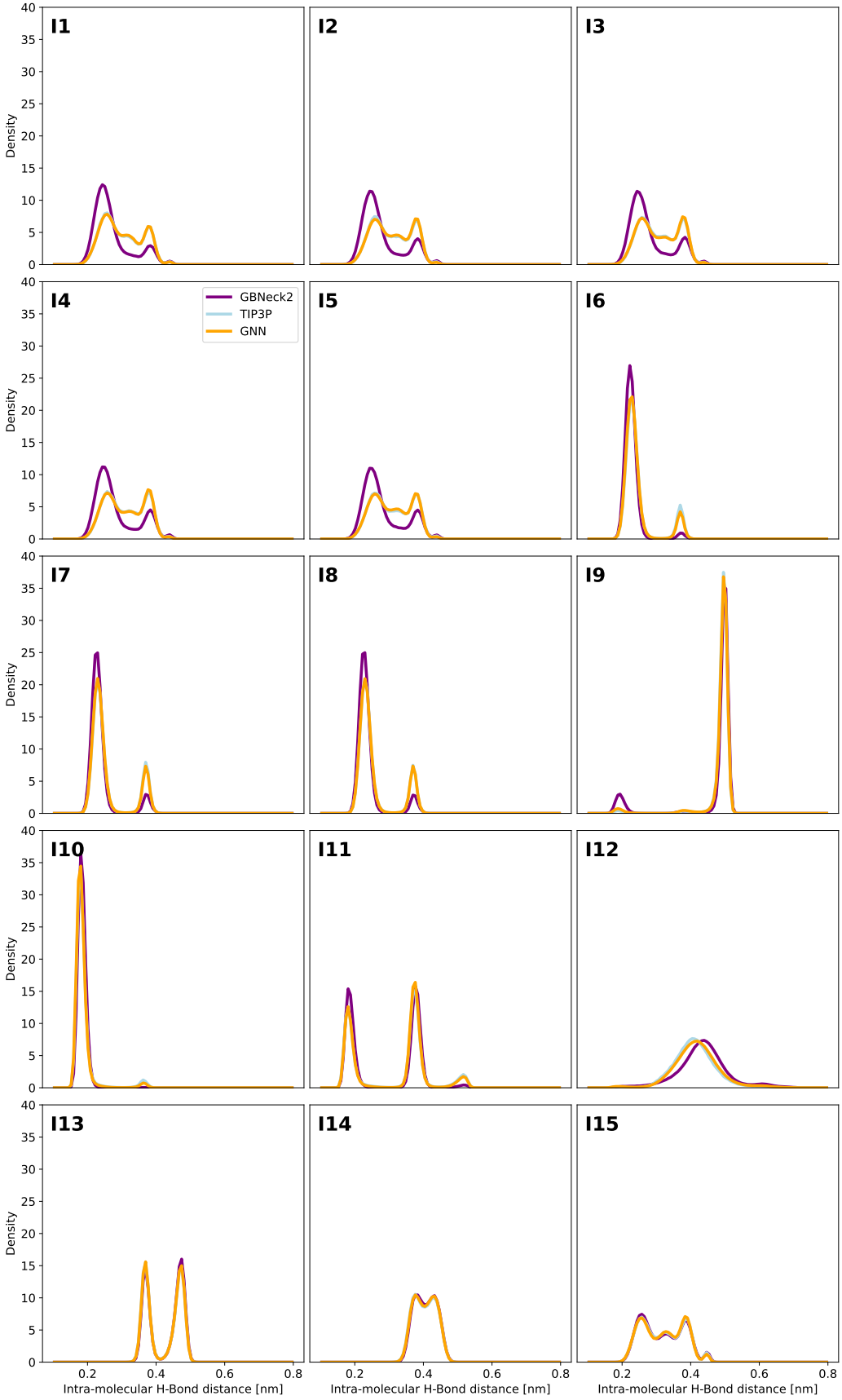

Figure S3: Probability distributions of the intramolecular hydrogen-bond distance for all compounds in set **I** using the GNN implicit solvent model (orange, 128x5 ns), the explicit solvent TIP3P reference (blue, 1x500 ns), and the baseline GB-Neck2 implicit solvent model (purple, 3x500 ns). The studied compound is indicated by its identifier in the upper left corners of the individual plots.

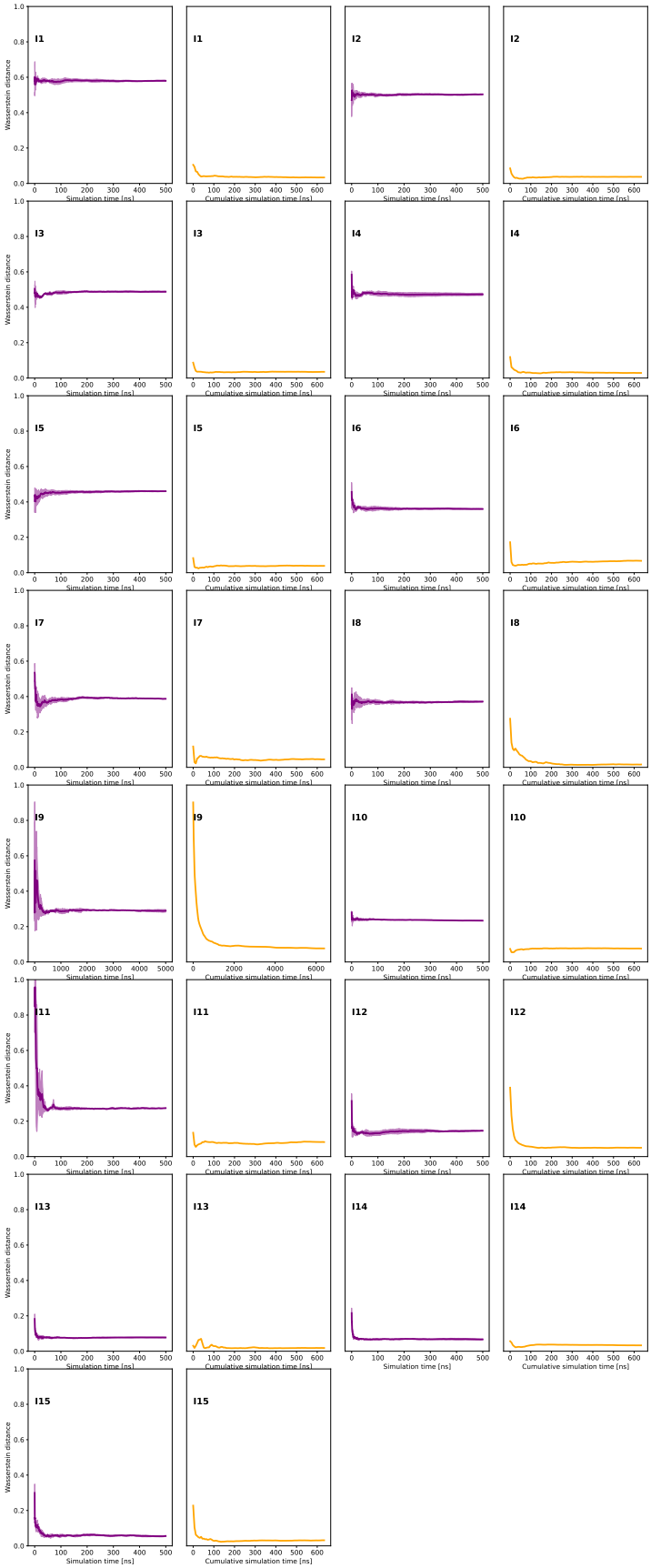

Figure S4: Convergence analysis of the compounds in set **I**: Wasserstein distance of the GB-Neck2 (purple) and GNN (orange) solvation models with respect to the explicit solvent TIP3P model as a function of simulation time. The studied compound is indicated by its identifier in the upper left corners of the individual plots.

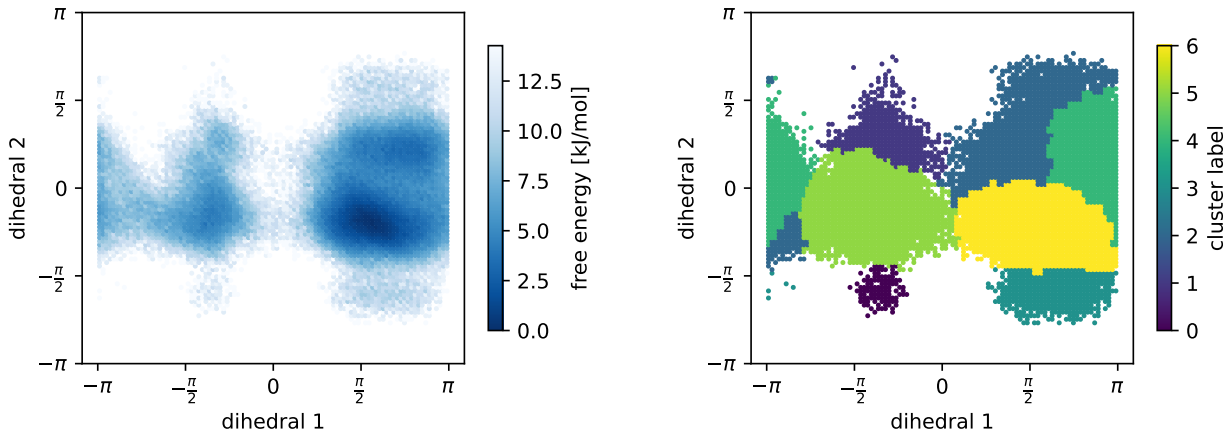

Figure S5: (**A**): 2D probability distribution along the two central torsional angles of compound C1 produced using the TIP3P solvation model. (**B**): Assigned clusters based on the EBC clustering algorithm.

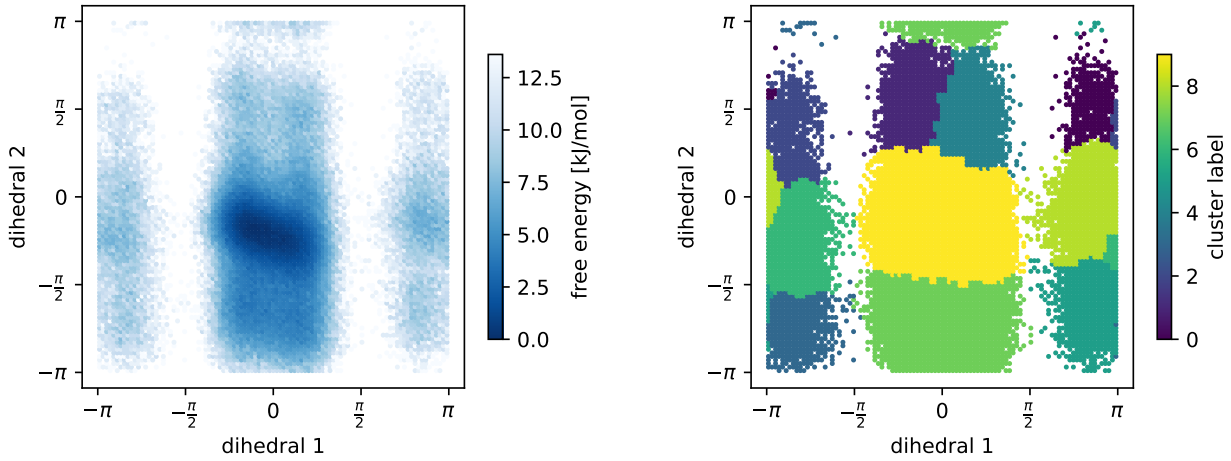

Figure S6: Compound C2. (**A**): 2D probability distribution along the two central torsional angles produced using the TIP3P solvation model. (**B**): Assigned clusters based on the EBC clustering algorithm.

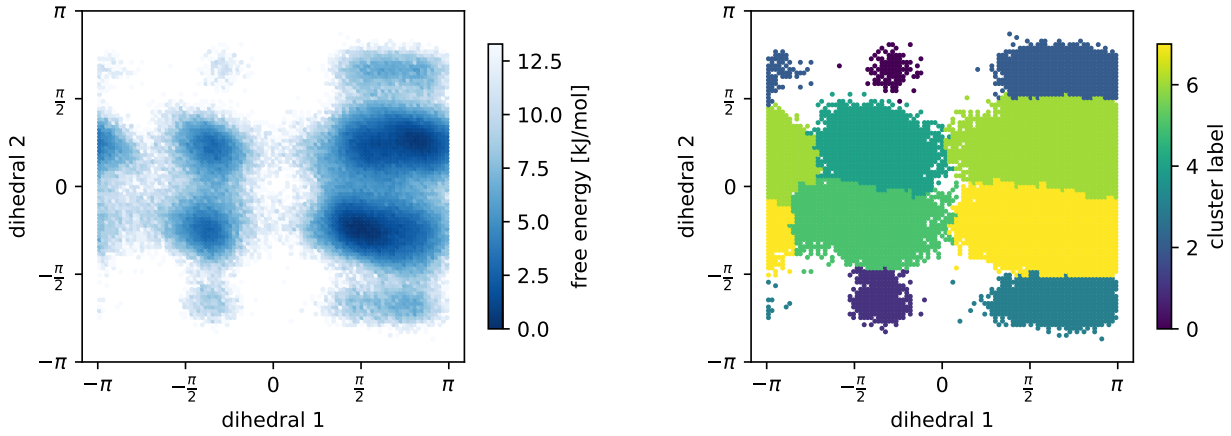

Figure S7: Compound C3. (**A**): 2D probability distribution along the two central torsional angles produced using the TIP3P solvation model. (**B**): Assigned clusters based on the EBC clustering algorithm.

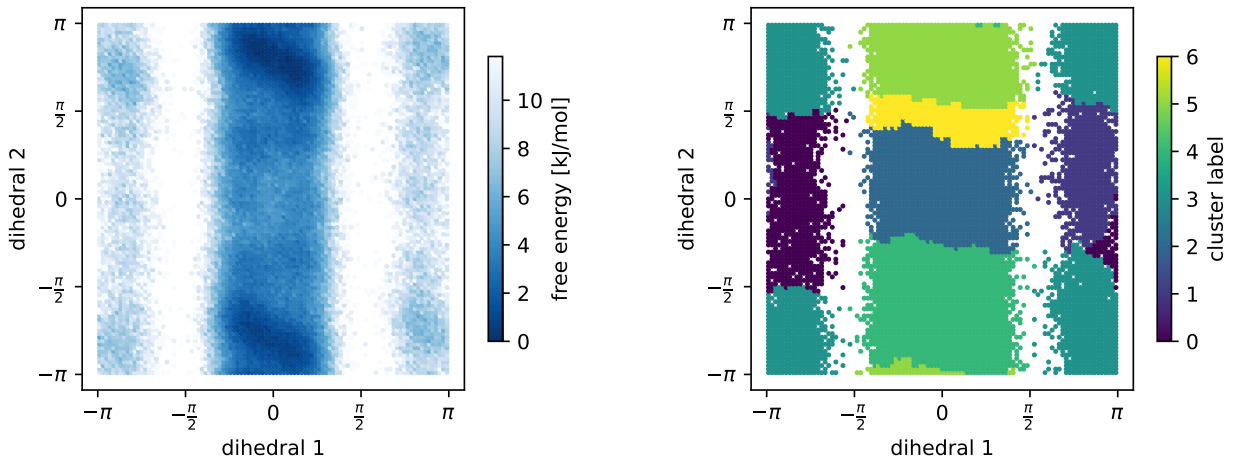

Figure S8: Compound C4. (**A**): 2D probability distribution along the two central torsional angles produced using the TIP3P solvation model. (**B**): Assigned clusters based on the EBC clustering algorithm.

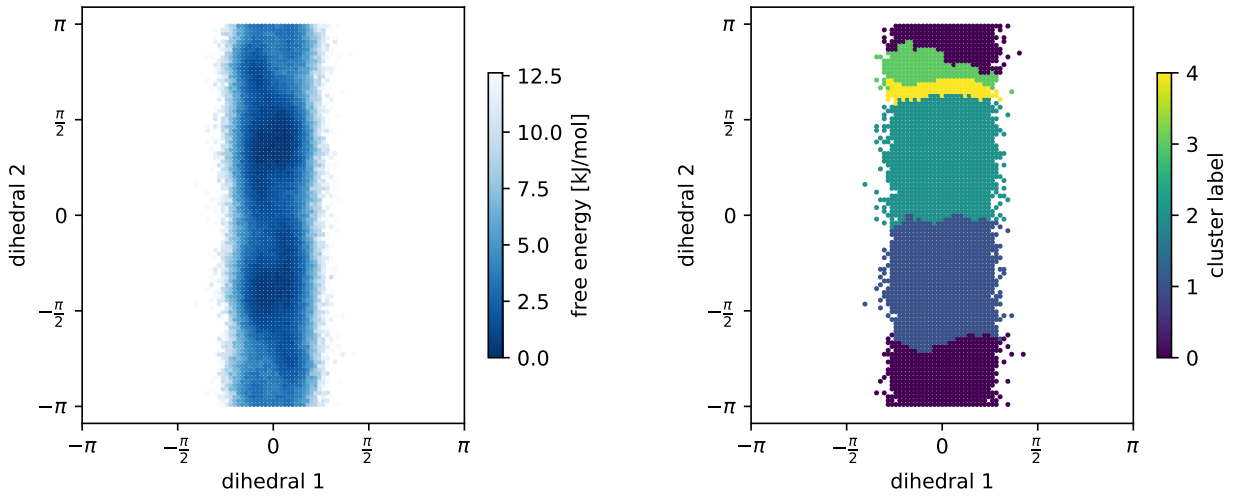

Figure S9: Compound C5. (**A**): 2D probability distribution along the two central torsional angles produced using the TIP3P solvation model. (**B**): Assigned clusters based on the EBC clustering algorithm.

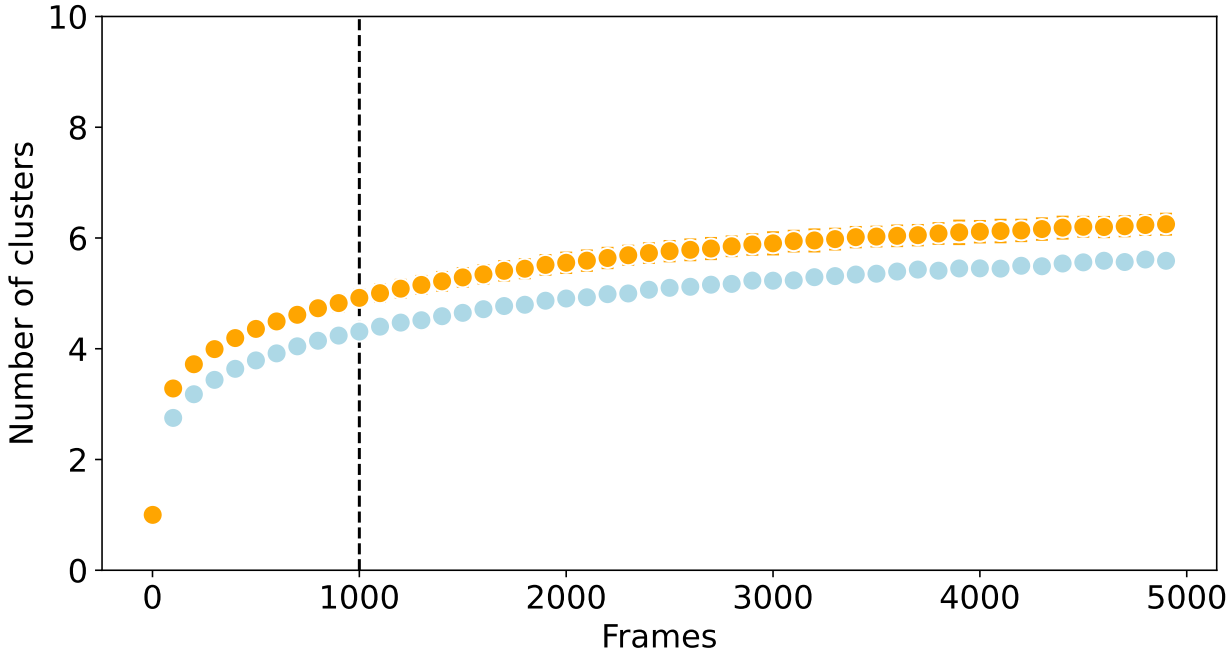

Figure S10: Comparison of effective sampling rates between the implicit GNN solvation model (orange) and the explicit TIP3P solvation model (blue). The average number of visited clusters after a given number of MD steps for compound C1 is shown.

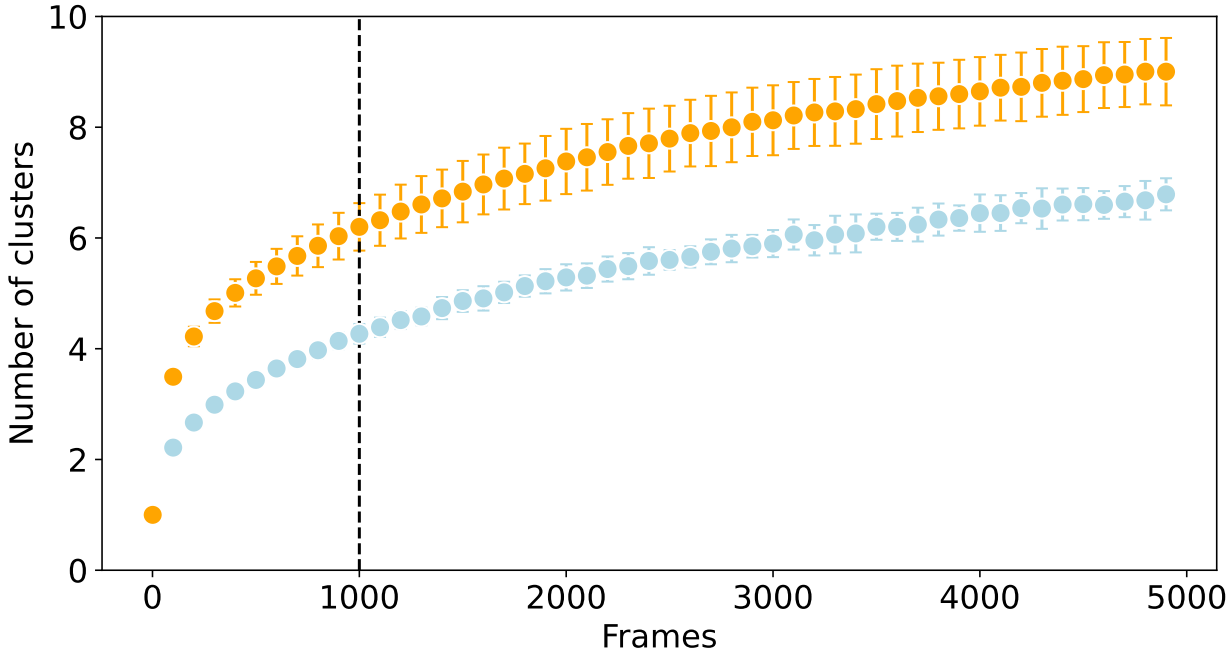

Figure S11: Comparison of effective sampling rates between the implicit GNN solvation model (orange) and the explicit TIP3P solvation model (blue). The average number of visited clusters after a given number of MD steps for compound C2 is shown.

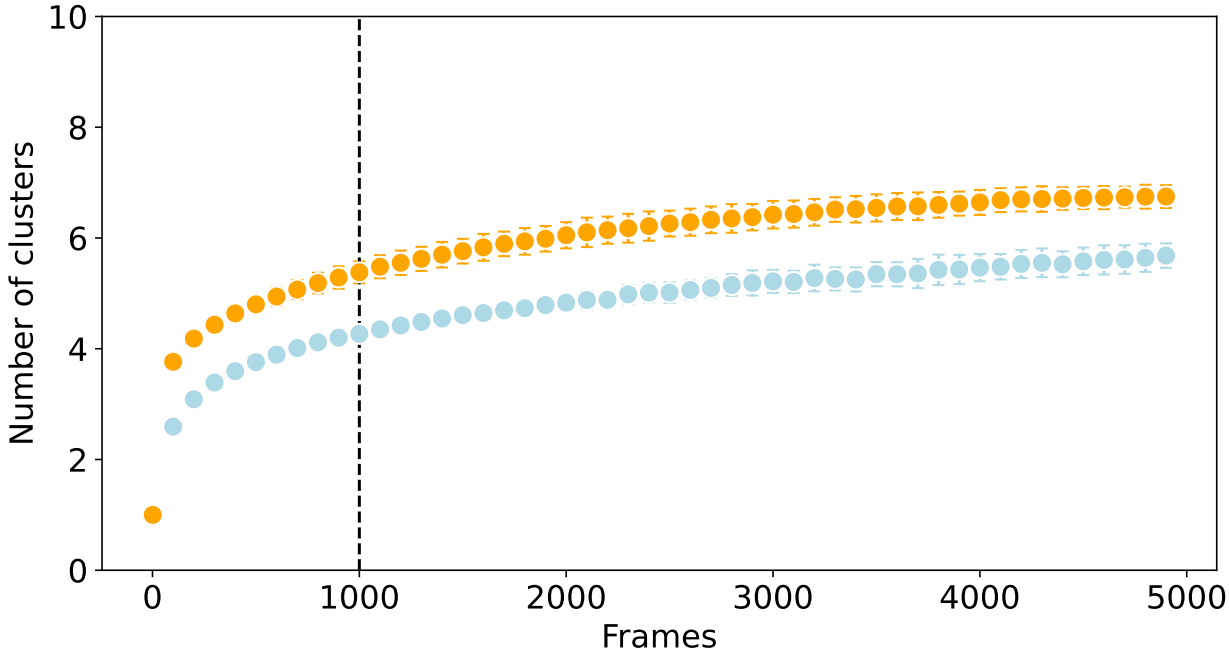

Figure S12: Comparison of effective sampling rates between the implicit GNN solvation model (orange) and the explicit TIP3P solvation model (blue). The average number of visited clusters after a given number of MD steps for compound C4 is shown.

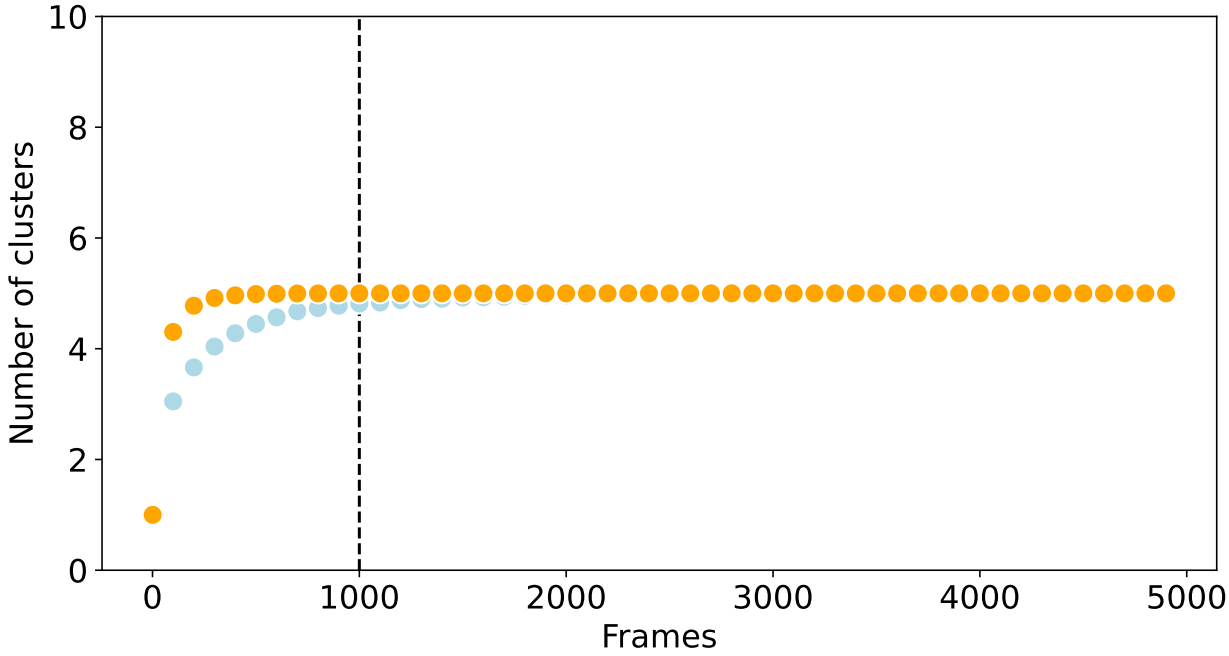

Figure S13: Comparison of effective sampling rates between the implicit GNN solvation model (orange) and the explicit TIP3P solvation model (blue). The average number of visited clusters after a given number of MD steps for compound C5 is shown.

GNN Model 2

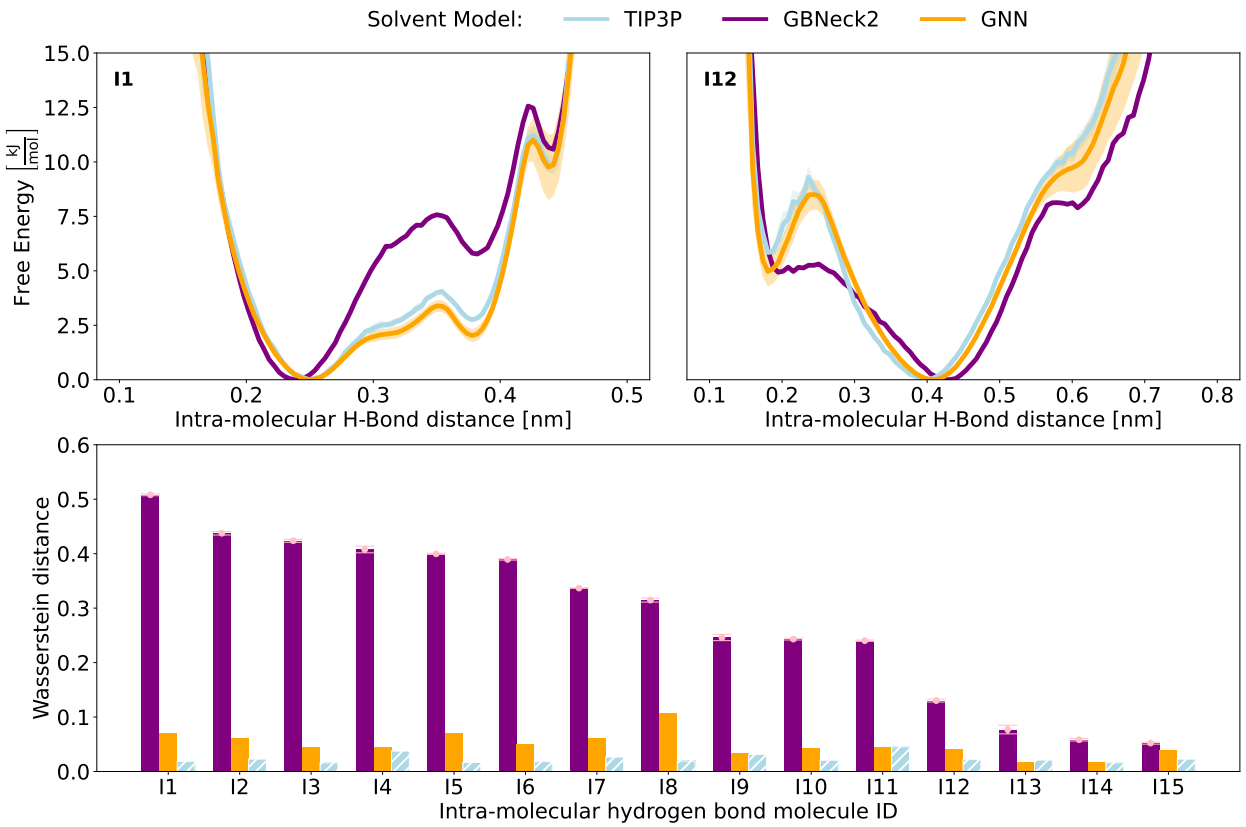

Figure S14: Comparison of the GNN (model 2) implicit solvent model (orange, 128x5 ns) with the explicit solvent TIP3P model (blue, 1x500 ns) and the baseline implicit solvent GB-Neck2 model (purple, 3x500 ns). (Upper left): Free-energy profile of the opening of the intramolecular hydrogen bond of compound I1. (Upper right): Free-energy profile of the opening of the intramolecular hydrogen bond of compound I12. (Bottom): Wasserstein distances of implicit solvent models compared to the explicit solvent TIP3P reference. For comparison, the hashed blue bar indicates the Wasserstein distance of the first half of the TIP3P simulation versus the second half. Pink error bars represent the standard deviation over multiple replicates of the GB-Neck2 model.

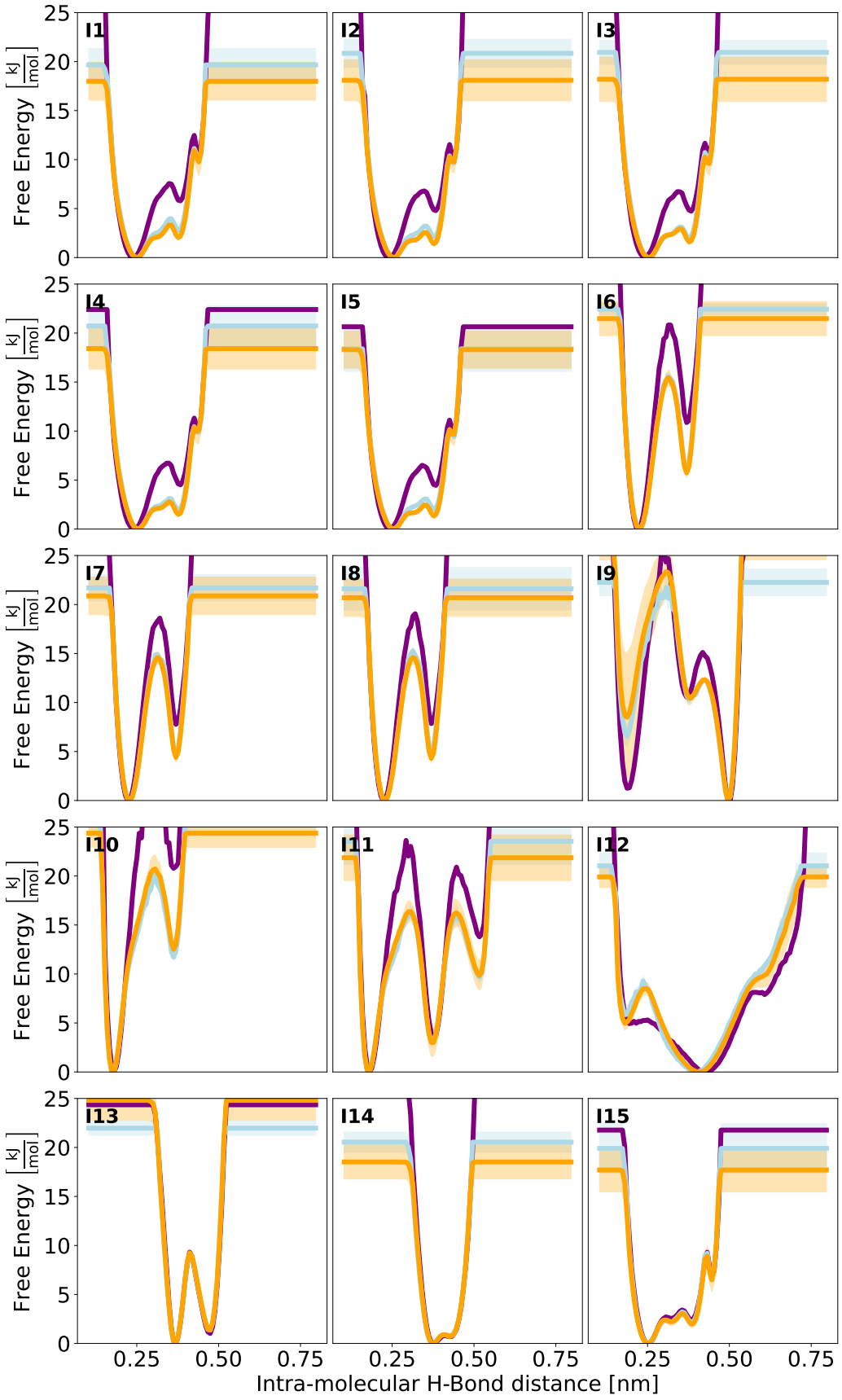

Figure S15: Free-energy profiles of the opening of the intramolecular hydrogen bonds of set **I**: Comparison of the GNN (model 2) implicit solvent model (orange, 128x5 ns) with the explicit solvent TIP3P model (blue, 1x500 ns) and the baseline implicit solvent GB-Neck2 model (purple, 3x500 ns). The studied compound is indicated by its identifier in the upper left corners of the individual plots.

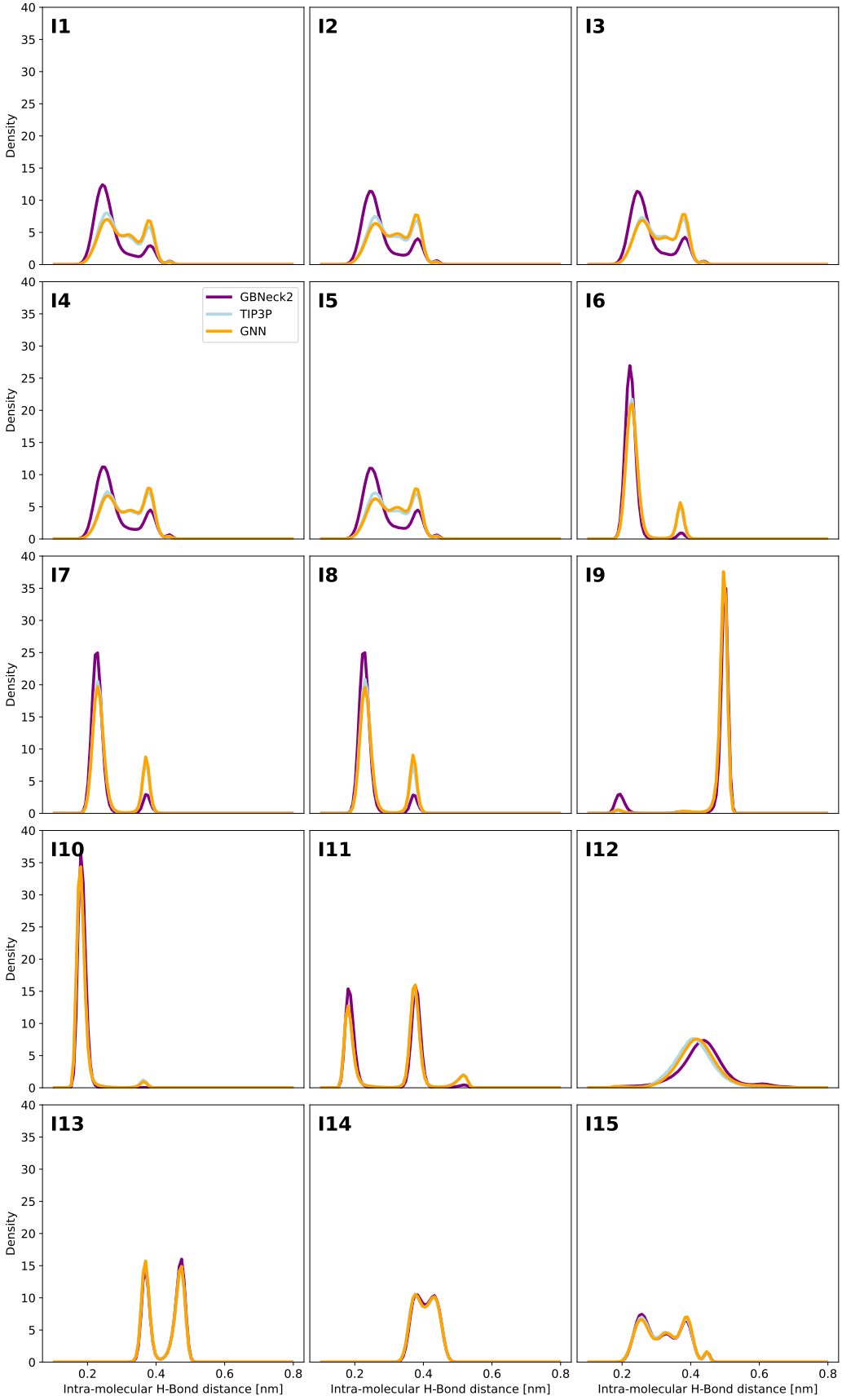

Figure S16: Probability distributions of the intramolecular hydrogen-bond distance for all compounds in set **I** using the GNN implicit solvent model (orange, 128x5 ns), the explicit solvent TIP3P reference (blue, 1x500 ns), and the baseline GB-Neck2 implicit solvent model (purple, 3x500 ns). The studied compound is indicated by its identifier in the upper left corners of the individual plots.

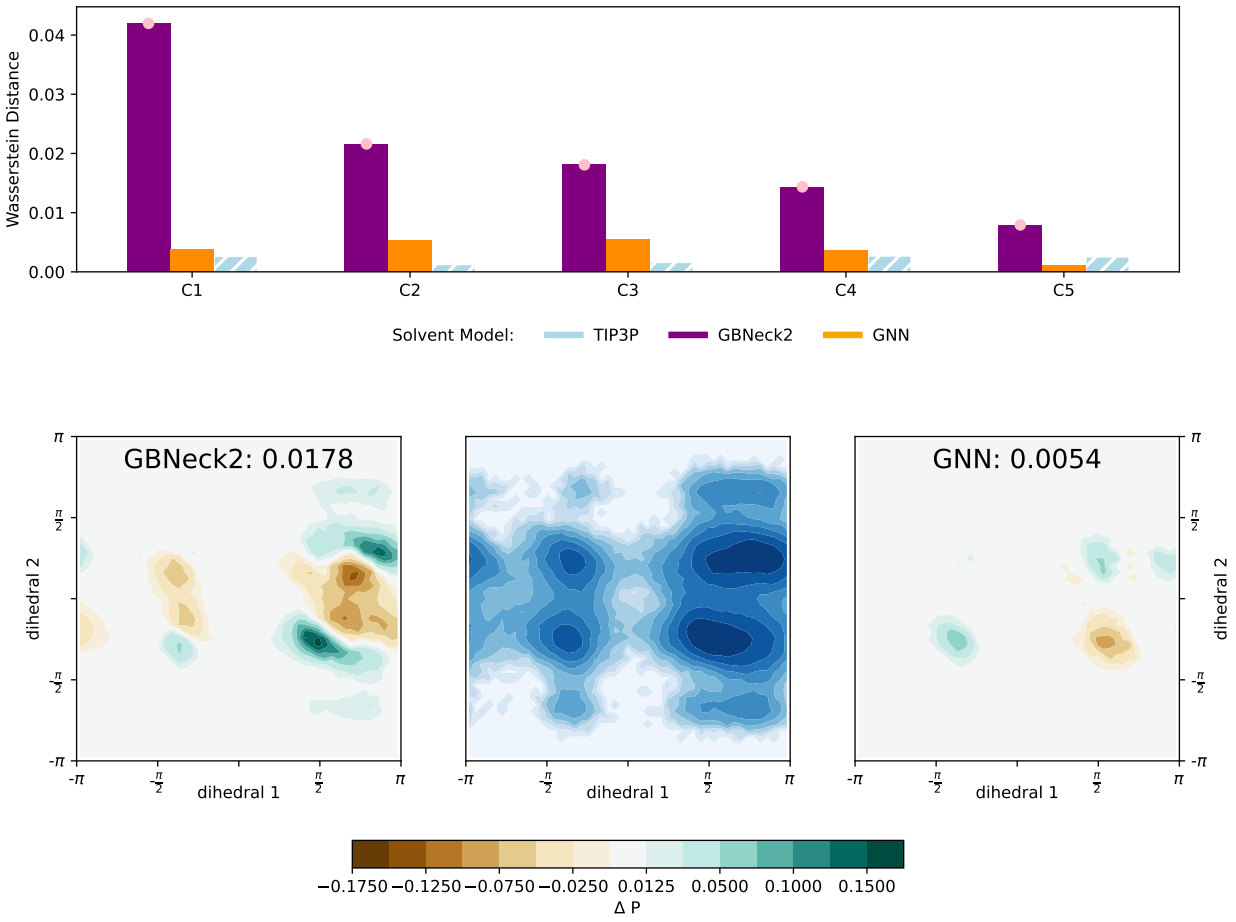

Figure S17: Comparison of the GNN (model 2) implicit solvent model (orange, 128x10 ns) with the explicit solvent TIP3P reference (blue, 1x500 ns) and the baseline implicit solvent GB-Neck2 model (purple, 3x500 ns). (Top): Wasserstein distances of implicit solvent models compared to the TIP3P reference. For comparison, the hashed blue bar indicates the Wasserstein distance of the first half of the TIP3P simulation versus the second half. (Bottom left): Difference in probability distribution for the GB-Neck2 and TIP3P simulations for compound C3. (Bottom middle): Probability distribution along the two central torsion angles (marked blue in panel A) of compound C3 produced using the TIP3P solvation model. (Bottom right): Difference in probability distribution for the GNN and TIP3P simulations for compound C3.

GNN Model 3

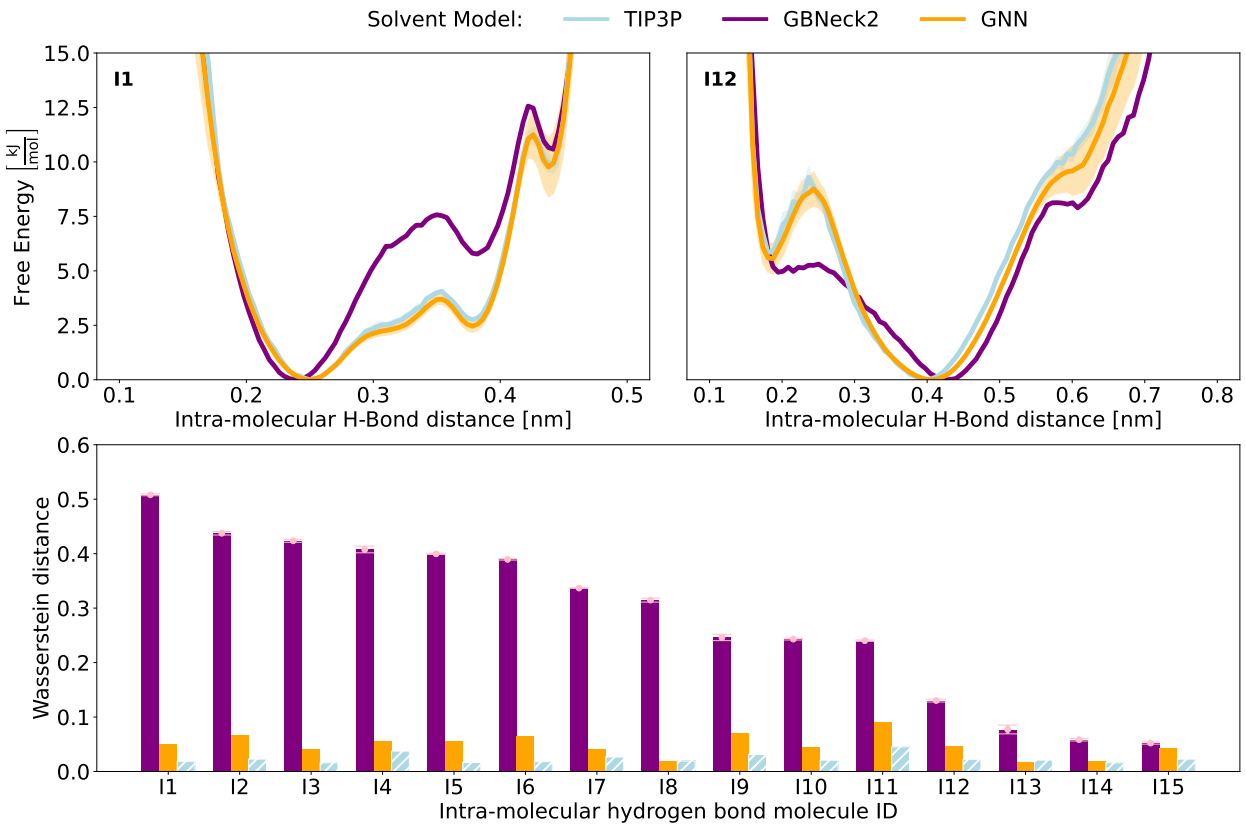

Figure S18: Comparison of the GNN (model 3) implicit solvent model (orange, 128x5 ns) with the explicit solvent TIP3P model (blue, 1x500 ns) and the baseline implicit solvent GB-Neck2 model (purple, 3x500 ns). (Upper left): Free-energy profile of the opening of the intramolecular hydrogen bond of compound I1. (Upper right): Free-energy profile of the opening of the intramolecular hydrogen bond of compound I12. (Bottom): Wasserstein distances of implicit solvent models compared to the explicit solvent TIP3P reference. For comparison, the hashed blue bar indicates the Wasserstein distance of the first half of the TIP3P simulation versus the second half. Pink error bars represent the standard deviation over multiple replicates of the GB-Neck2 model.

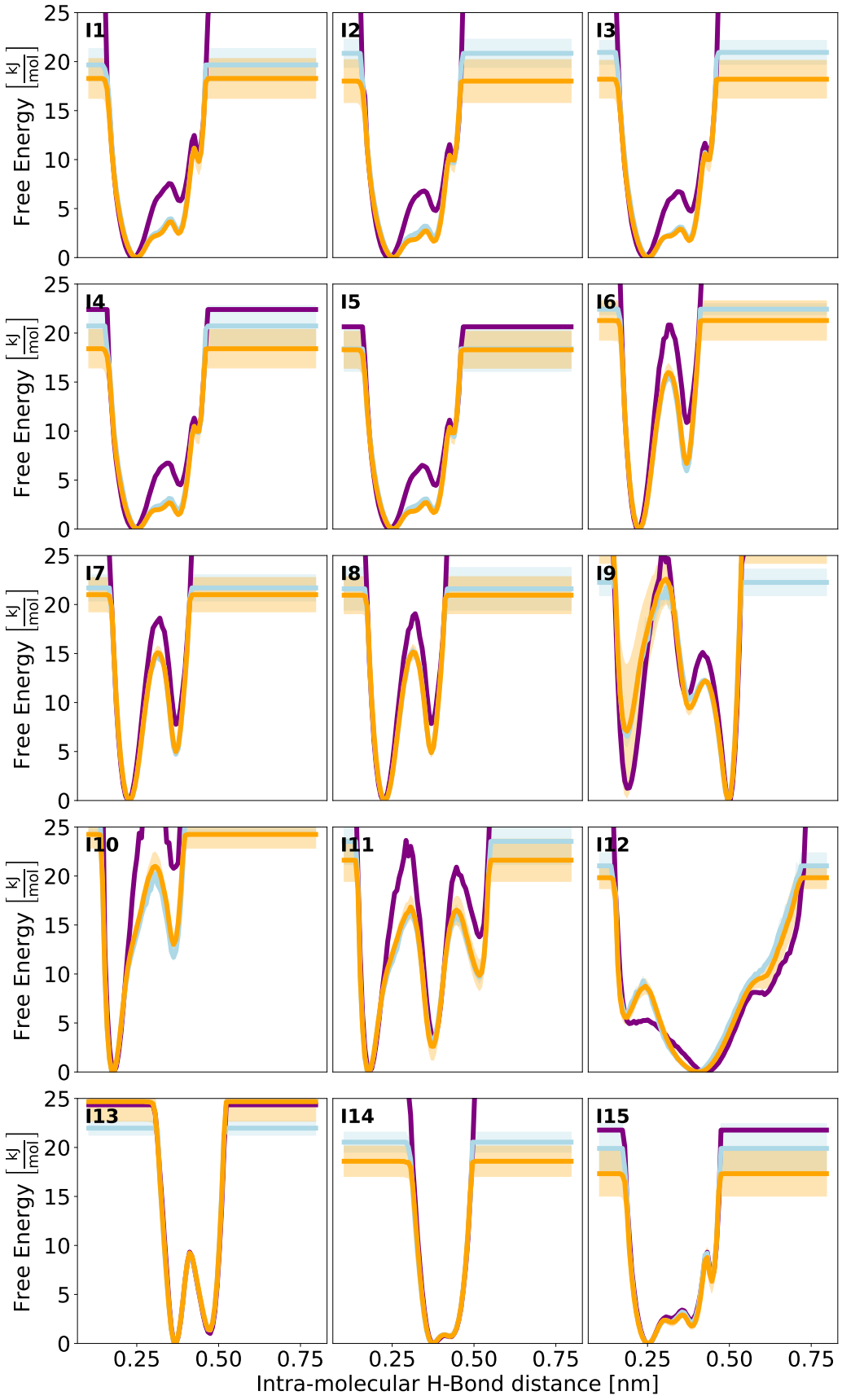

Figure S19: Free-energy profiles of the opening of the intramolecular hydrogen bonds of set **I**: Comparison of the GNN (model 3) implicit solvent model (orange, 128x5 ns) with the explicit solvent TIP3P model (blue, 1x500 ns) and the baseline implicit solvent GB-Neck2 model (purple, 3x500 ns). The studied compound is indicated by its identifier in the upper left corners of the individual plots.

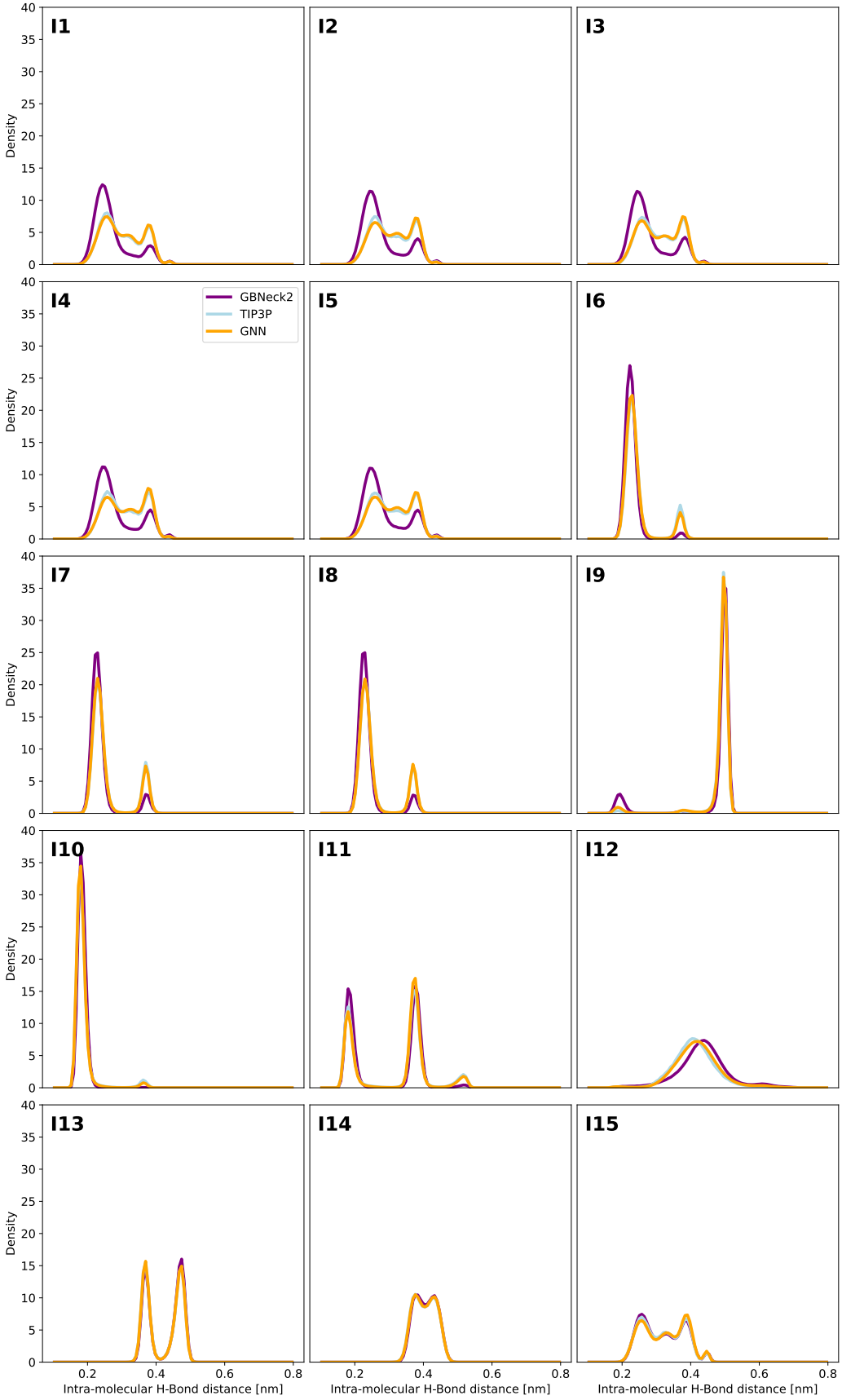

Figure S20: Probability distributions of the intramolecular hydrogen-bond distance for all compounds in set **I** using the GNN (model 3) implicit solvent model (orange, 128x5 ns), the explicit solvent TIP3P reference (blue, 1x500 ns), and the baseline GB-Neck2 implicit solvent model (purple, 3x500 ns). The studied compound is indicated by its identifier in the upper left corners of the individual plots.

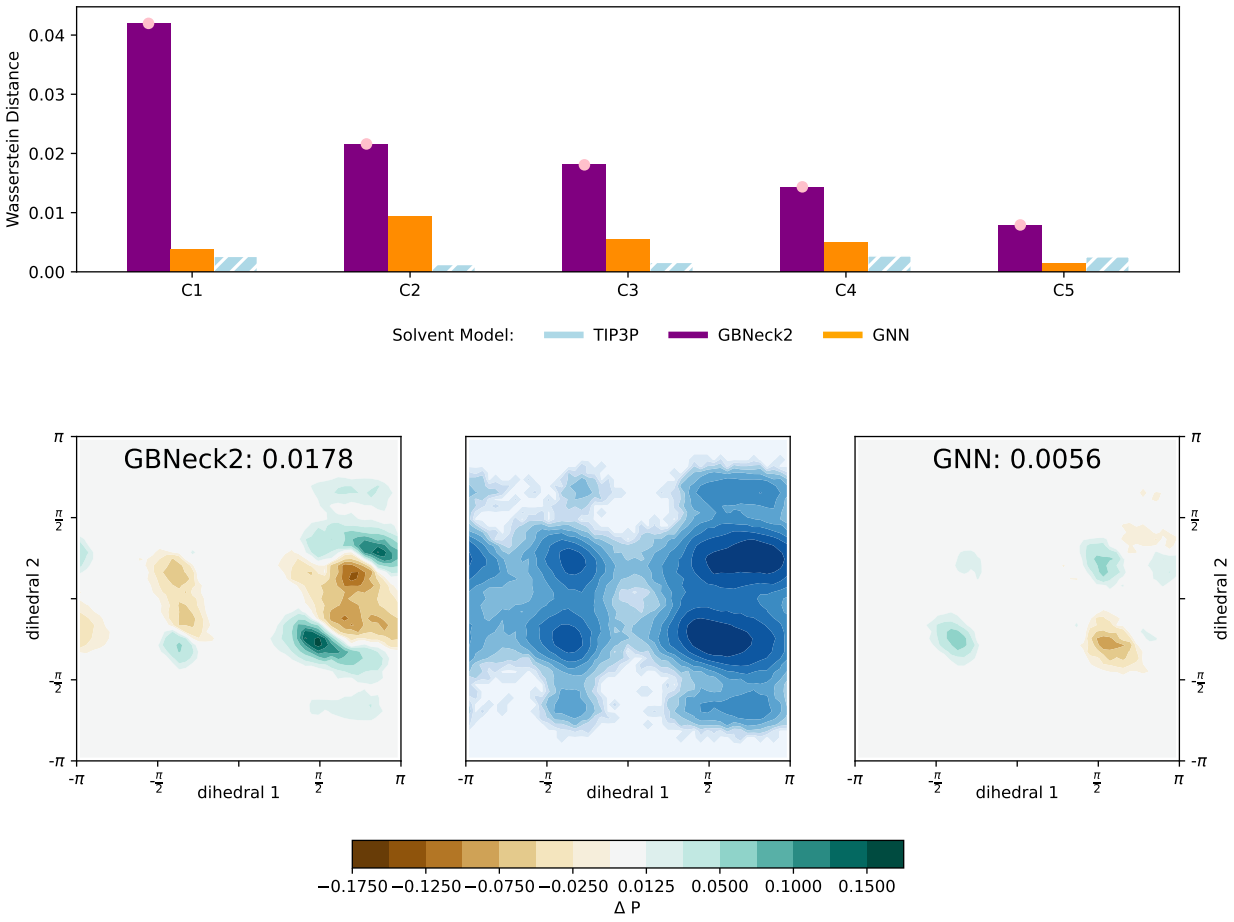

Figure S21: Comparison of the GNN (model 3) implicit solvent model (orange, 128x10 ns) with the explicit solvent TIP3P reference (blue, 1x500 ns) and the baseline implicit solvent GB-Neck2 model (purple, 3x500 ns). (Top): Wasserstein distances of implicit solvent models compared to the TIP3P reference. For comparison, the hashed blue bar indicates the Wasserstein distance of the first half of the TIP3P simulation versus the second half. (Bottom left): Difference in probability distribution for the GB-Neck2 and TIP3P simulations for compound C3. (Bottom middle): Probability distribution along the two central torsion angles (marked blue in panel A) of compound C3 produced using the TIP3P solvation model. (Bottom right): Difference in probability distribution for the GNN and TIP3P simulations for compound C3.

References

[1] Katzberger, P.; Riniker, S. Implicit solvent approach based on generalized Born and transferable graph neural networks for molecular dynamics simulations. *J. Chem. Phys.* **2023**, *158*, 204101.