Intramolecular and water mediated tautomerism of solvated glycine

Pengchao Zhang[1,2], **Axel Tosello Gardini**[2,3], **Xuefei Xu**[1,*], and **Michele Parrinello**[2,*].

[1]Center for Combustion Energy, Department of Energy and Power Engineering, and Key Laboratory for Thermal Science and Power Engineering of Ministry of Education, Tsinghua University, Beijing 100084, China

[2]Atomistic Simulations, Italian Institute of Technology, Genova 16152, Italy

[3]Department of Materials Science, Università di Milano-Bicocca, 20126 Milano, Italy

[*]Co-corresponding author e-mail: xuxuefei@tsinghua.edu.cn, michele.parrinello@iit.it

November 13, 2023

Abstract

The understanding of prototropic tautomerism in water and the characterization of solvent effects on protomeric equilibrium pose significant challenges. Using molecular dynamics simulations based on state-of-the-art deep learning potential and enhanced sampling methods, we provide a comprehensive description of all configurational transformations in glycine solvated in water and determine accurate free energy profiles of these processes. We observe that the tautomerism between the neutral and zwitterionic forms of solvated glycine can occur by both intramolecular proton transfer in glycine and intermolecular proton transfer in the contact ion pair (anionic glycine and hydronium ion) or the separated ion pair (cationic glycine and hydroxide ion).

1 Introduction

Prototropic tautomerism plays a crucial role in determining the thermodynamic and kinetic properties of amino acids and their derivatives, and consequently their reactivity and interaction with other biomolecules.[1–7] The simplest amino acid, glycine, can exist in the zwitterionic [Z], neutral [N], anionic [A] or cationic [C] form (Fig. 1) depending on the pH of the solution,[8–10] providing a simple yet relevant model for the study of tautomeric behavior. Here we will focus on the transformation of glycine in water between the [Z] and [N] forms.

Experimental studies have been carried out to determine the protonation states of glycine under various conditions.[9,11,12] It has been found that in the gas phase the [N] tautomer is the more stable. Under microsolvation conditions,[13,14] it has been demonstrated that proton transfer can occur and the [Z] form is favoured. For fully solvated glycine,[8–11,15] the [Z] tautomer has also been shown to be more stable than the [N] tautomer. However, little is experimentally known on the dynamic processes of transformation from [N] to [Z], and these microscopic processes can be investigated by theoretical simulations at the atomic level.

The gas phase experiments have been accompanied by a number of static calculations.[16–18] However, the glycine tautomerism is difficult to achieve in the gas phase, but it is facilitated by the dynamics of the water environment and the fluctuation of the hydrogen bond (HB) network. In an effort to include water effects, various approaches have been taken, from describing water as polarizable continuum models (PCMs),[9,10,19–23] to quantum mechanical/molecular mechanical (QM/MM) calculations,[23,24] and to the use of reactive force fields.[25] Most studies have come to the same conclusion as experiments, i.e., the [Z] tautomer is stable in water. However, only a few simulations have explored the tautomeric dynamics of glycine, and revealed the mechanism of short-range proton transfer, i.e., the transition from [N] to [Z] occurs intramolecularly or sometimes via a nearby solvent water molecule.[18,24,25] Here we aim at describing the static and

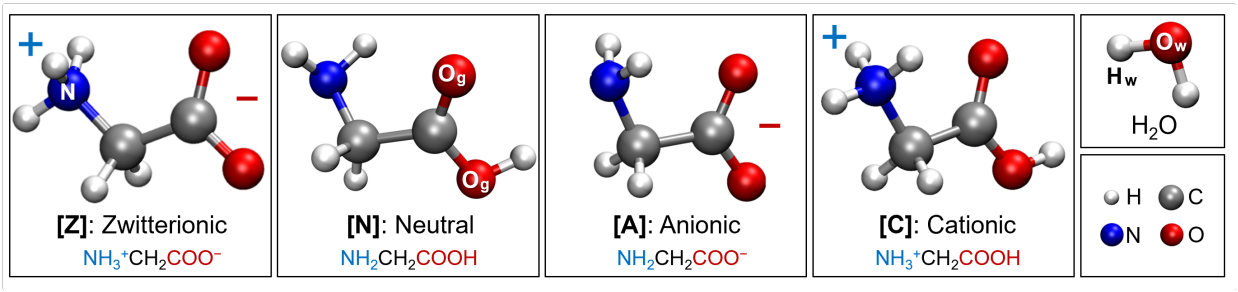

Figure 1: **The schematic diagrams of four glycine states and a water molecule.** [Z], [N], [A], and [C] denote zwitterionic, neutral, anionic, and cationic forms, respectively. Atomic notation is used to distinguish the atoms present in the molecules of glycine and water. X_y represents the atom X (H or O) that originates from molecule y (glycine or water).

dynamic behavior of glycine in water taking into account the full complexity of water dynamics and chemistry. The most appropriate theoretical framework for such a study is that of *ab initio* molecular dynamics (AIMD).[18,26–29]

Unfortunately, AIMD simulations are computationally expensive, and this has limited the size of the system studied and the time scale of the simulations. In pioneering work, Behler and Parrinello have shown that the cost of *ab initio*-quality simulations can be greatly reduced without compromising accuracy.[30] Their solution was to represent the potential energy surface as a suitably designed neural network whose inputs are a set of descriptors chosen so as to best describe the local atomic environment, while enforcing the symmetry of the problem and allowing scaling up to large systems. The neural network is then trained on a set of DFT (density functional theory) energies and forces performed on a carefully selected set of configurations. More recently, many variants of this approach have been proposed, such as Gaussian approximation potential,[31] deep potential (DP),[32–34] SchNet,[35] and other equivariant approaches.[36–38] These developments benefit greatly from easy access to efficient machine learning libraries and from efficient neural network training strategies. Here, we use the DP model to perform molecular dynamics (MD) simulations, which has proven its usefulness in many applications, including the studies of water and reactions in water.[39–44]

During the DP model training, it is important to present the network with an appropriate set of configurations, especially when we want to study reactive processes such as the protonation and deprotonation of glycine. Since these processes take place on a time scale that is much larger than the one accessible to standard MD simulations, the use of enhanced sampling methods is necessary to sample such rare events. Here we employ the recently developed on-the-fly probability enhanced sampling (OPES) method,[45,46] which is an evolution of metadynamics.[47,48] Like many other enhanced sampling methods, OPES uses collective variables (CVs) to accelerate the fluctuations that eventually lead to the reactive process. In this work we adapt the Voronoi CVs introduced by Grifoni et al.[49,50] so as to be able to simultaneously describe the [Z], [N], [A], [C] forms of glycine and the intermediate structures connecting these forms.

Then we follow the deep Kohn-Sham (DeePKS)[51–54] strategy to efficiently build a very accurate training dataset. In particular, we want to perform simulations in which the energies and forces are as accurate as those predicted by the hybrid meta-GGA functional M06-2X, which has been shown to give an accurate representation of the thermochemistry and kinetics of main group elements.[55] However, the computational cost of performing single point calculations at the M06-2X level for the current system is too high. In simple terms, the way DeePKS gets around this difficulty is to train a neural network that expresses the difference E_δ between the target energy and the baseline energy. In our case the target energy is $E_{\mathrm{M06-2X}}$, and the baseline energy E_{PBE} is from the cheaper GGA functional PBE.[56] The E_δ is calculated as the sum of atomic energy differences, determined using a neural network that takes atomic coordinates, density matrices, and orbitals as inputs. The remarkable finding is that to train E_δ only a very small number of training data are needed. Once the PBE-based DeePKS model is trained, energy is calculated from $E_{\mathrm{PBE}} + E_\delta$, and forces can be driven from the energy. The energies and forces generated using such a DeePKS model will have an accuracy close to that of the full M06-2X calculation at a much lower cost, and will be used to train the DP model.

We use the generated DP model to explore the tautomeric free energy landscape of glycine in water, and find that

transitions between the [N] and [Z] forms of glycine can occur via multiple proton transfer (PT) processes.

2 Computational methods

2.1 Potential model generation

2.1.1 Building training datasets

Key to the potential model training is the construction of appropriate training datasets. As discussed in the Introduction section, we need to train two potential models: One is the DP model to be used in molecular dynamics simulations for investigating the glycine tautomerism in water; the other is the PBE-based DeePKS model with the accuracy approaching the M06-2X functional to be used in the labeling of the DP training datasets. In either case we use pure water and solvated glycine configurations (see Table S1 in the SI). These configurations are generated in multiple independent MD simulations with enhanced sampling to generate as much uncorrelated data as possible and to include reactive processes. The details of the MD simulations are given below.

The numbers of configurations in the training sets of the two models are much different due to the difference between the descriptors. The similarities are that the two model descriptors all satisfy the physical symmetry and the locality. The differences are that the neural network descriptors in the DP model contain only the angular and radial atomic environment, while the descriptors of the DeePKS model additionally include density matrices projected on the atomic orbitals and satisfy the gauge invariance symmetry.[33,52,54] This means that only a few hundred configurations are needed to train the DeePKS model. In contrast, more than tens of thousands of configurations are needed for training the DP model to have an accurate modeling of the reactive process. Finally, a total of 300 and 55,498 configurations are collected in the DeePKS and DP model training datasets, respectively.

2.1.2 DFT calculations

To prepare DeePKS training datasets and all test datasets, the M06-2X[55] energies and forces are calculated using the CP2K[57] package. In the calculations, the Goedecker-Teter-Hutter pseudopotentials[58,59] are used together with a quadruple-zeta valence basis set with polarization[60] (QZV3P). The multi-grid level utilizes a plane-wave cutoff of 1000 Ry for the total density and 70 Ry for the Kohn-Sham orbitals. To accelerate the convergence of self-consistent field (SCF) iterations, the auxiliary density matrix method[61] is utilized.

2.1.3 DeePKS model training

The DeePKS model is then generated using an iterative approach, in which we alternately train the network of the correction term from the PBE baseline to the M06-2X target using the DeePKS-kit[51–53] package, and solve the resulting DeePKS model in ABACUS.[54,62,63] In the ABACUS calculations, the optimized norm-conserving Vanderbilt[64] pseudopotentials are used together with numerical atomic orbital[65] basis. The kinetic energy cutoff is set at 100 Ry, and the SCF convergence threshold for the density error is 1×10^{-7} Ry.

The above iterations are no longer needed once the DeePKS model has been trained. One performs single point calculations based on the PBE functional corrected with the E_δ term to obtain total energies and forces. Relative to an independent set of testing data obtained by performing ordinary M06-2X/QZV3P calculations, the PBE-based DeePKS model shows root mean square errors (RMSEs) of 0.53 and 0.61 meV/atom for the energies and of 43 and 52 meV/Å for the forces, where we distinguish between errors relative to pure water and solvated glycine systems. In terms of efficiency, the PBE-based DeePKS model saves about an order of magnitude in time compared to using the standard M06-2X functional (Table S3 in the SI).

2.1.4 DP model training

The DP model is then trained in the DeePMD-kit[34,66] package using the training dataset of M06-2X quality generated as described above. As discussed earlier, at this stage we needed 55,498 configurations for an accurate result. Then, the trained DP model is examined by testing the prediction on an independently generated dataset containing 6,020 configurations for which standard M06-2X energy and force calculations have been performed. The DP model exhibits energy RMSEs of 0.79 and 1.72 meV/atom and atomic force RMSEs of 58 and 63 meV/Å for water and solvated glycine systems, respectively, suggesting that the DP model has achieved a level of precision comparable to M06-2X. Compared to an M06-2X functional based simulation, the DP model is five orders of magnitude faster (Table S3).

2.2 Molecular dynamics simulations

2.2.1 MD simulation for the training dataset generation

The configurations required to train the DeePKS and DP models are sampled from MD simulations using CP2K[57] as the driver. To speed up the sampling, we use the semi-empirical GFN1-xTB[67] to derive potential energy

surface. This GFN1-xTB method has a lower computational cost, but is still accurate enough to describe the structures of different glycine forms in water. The simulations are performed in constant volume constant temperature (NVT) ensemble with a time step of 1.0 fs. The temperature of 300 K is enforced using the velocity rescaling thermostat[68] with a damping time of 0.2 ps.

2.2.2 DPMD simulation for studying glycine tautomerism

To study the glycine tautomerism in water, deep potential molecular dynamics (DPMD) simulations are performed using the LAMMPS[69] software. Before the formal simulation, we have investigated the size effect by placing a glycine molecule in varying amounts of water. Based on the convergence test result (Fig. S9), the system C (Table S1) with one glycine molecule in 128 H_2O molecules is used in the final DPMD simulation to strike a balance between efficiency and precision. In order to sufficiently sample all possible tautomeric processes, simulations are run for 30 ns in the NVT ensemble with an integration time step of 1.0 fs. The temperature is controlled at 300 K using again the velocity rescaling thermostat with a damping time of 0.04 ps.

2.3 Enhanced sampling settings

2.3.1 Bias potential

The OPES method using collective variables (CVs) is employed in both MD simulations for constructing the datasets and the DPMD simulation of glycine tautomerism. The CVs $\mathbf{s}(\mathbf{R})$, which are functions of atomic coordinates ($\mathbf{R}$), skilfully capture the slow modes associated with rare events. During MD simulations, the potential energy ($U(\mathbf{R})$) of the system is modified by adding the external bias potential ($V(\mathbf{s})$) through the utilization of the PLUMED[70] plugin. Specifically, $V(\mathbf{s})$ at the nth step in OPES[45,46] is defined by:

$$V_n(\mathbf{s}) = (1 - \frac{1}{\gamma})\frac{1}{\beta}\log(\frac{P_n(\mathbf{s})}{Z_n} + \epsilon) \tag{1}$$

where $\beta = 1/k_BT$ is the inverse Boltzmann factor. Probability $P(\mathbf{s})$ is the unbiased marginal distribution, and parameter Z is the normalization factor. Bias factor $\gamma = \beta\Delta E_{\text{bias}}$ ensures the reshaping of the original probability. The regularization term $\epsilon = e^{-\gamma/(1-1/\gamma)}$ not only guarantees a positive argument for the logarithm, but also imposes a bias constraint, thereby limiting sampling within the specified region of interest. In particular, the bias update is carried out every 500 steps with the adaptive kernel width, and the value of ΔE_{bias} is set at 35 kJ/mol.

Notably, the "explore" variant of OPES (OPES-explore) is used in the configuration collection of datasets due to its ability to accelerate the exploration of phase space, while the OPES is performed to generate well-converged free energies during the final DPMD simulation. The reason is that the ways of estimating the probability distribution are different in the OPES and OPES-explore methods, although the idea of defining the bias potential is similar. Specifically, unbiased probability $P(\mathbf{s})$ is estimated on-the-fly using weighted kernel density estimation (KDE) in OPES, while the well-tempered probability $p^{\text{WT}}(\mathbf{s})(\propto [P(\mathbf{s})]^{\frac{1}{\gamma}})$ is estimated based on averaged KDE in OPES-explore.[45,46]

2.3.2 Collective variables

Proton can diffuse through water via the Grotthuss mechanism,[44,71,72] in which it is not a well-specified proton that moves, but rather a charge defect that migrates through the water hydrogen network. Thus a CV that is used to describe proton diffusion has to be able to identify charge defects without making reference to a specific set of atomic coordinates. In the present case, several charge defects are possible: the $–NH_3^+$ and $–COO^-$ groups in glycine, and the two water self-ions (hydronium H_3O^+ and hydroxide OH^-). To identify automatically these charge defects we follow the strategy of References[49,50] and tessellate the space with Voronoi polyhedra centered on the O and N atoms. We then sum the charge contained in each polyhedron. If the charge in a polyhedron is different from zero, we attribute the charge defect to the atom N or O that is at the polyhedron center, and we distinguish the O defects according to whether they are centered on water or glycine oxygen (O_w and O_g in Fig. 1).

To count the number of H atoms n_i centered on O or N atom i, we use the formula

$$n_i = \sum_{j=1}^{\text{Num}_\text{H}} \frac{e^{-\lambda|\mathbf{R}_i - \mathbf{R}_j|}}{\sum_{m=1}^{\text{Num}_\text{O\&N}} e^{-\lambda|\mathbf{R}_m - \mathbf{R}_j|}} \tag{2}$$

where the first sum is over all H atoms with index j, and m is the index of Voronoi centers. The parameter λ regulates the smoothness of the function.

The charge defect number δ_i of the Voronoi center i is calculated by subtracting the reference number n_i^0 (the original proton number connecting to the atom at the polyhedron center) from H number n_i

$$\delta_i = n_i - n_i^0 \tag{3}$$

where we take $n_i^0 = 2$ for the water oxygen (O_w) and the glycine nitrogen (N), while for the carboxylic oxygens (O_g) we set $n_i^0 = \frac{1}{2}$ on account of the symmetry between two oxygen atoms.

As the system is neutral overall, the [C] and [A] forms of glycine are compensated by OH^- and H_3O^+, respectively. Then, to distinguish between the [C]−OH^- pair, [Z]&[N] states and [A]−H_3O^+ pair, we define a CV $\mathbf{s}_p$ that can identify these protonation states, by combining the charge defects of both glycine and water.

$$\mathbf{s}_p = \sum_{i=1}^{\mathrm{Num_{O_w}}} \delta_i + 2\Big(\sum_{j=1}^{\mathrm{Num_N}} \delta_j + \sum_{k=1}^{\mathrm{Num_{O_g}}} \delta_k\Big) \tag{4}$$

where $\mathbf{s}_p \approx 1$ for [C]−OH^-, $\mathbf{s}_p \approx -1$ for [A]−H_3O^+, and $\mathbf{s}_p \approx 0$ in the other two cases. Since $\mathbf{s}_p$ cannot distinguish the [Z] and [N] forms, we introduce another CV $\mathbf{s}_d$ to estimate the charge-charge distance that is capable of separating the two cases. The $\mathbf{s}_d$ is measured in terms of the distances between O_w and N and between O_w and O_g, as well as the average distance between N and O_g, as follows:

$$\begin{aligned}\mathbf{s}_d = &- \sum_{i=1}^{\mathrm{Num_{O_w}}} \sum_{j=1}^{\mathrm{Num_N}} r_{i,j}\delta_i\delta_j - \sum_{i=1}^{\mathrm{Num_{O_w}}} \sum_{k=1}^{\mathrm{Num_{O_g}}} r_{i,k}\delta_i\delta_k \\ &- \sum_{j=1}^{\mathrm{Num_N}} \sum_{k=1}^{\mathrm{Num_{O_g}}} r_{j,k}\delta_j\delta_k\end{aligned} \tag{5}$$

where r is the modulus distance between the two Voronoi centers. The CV $\mathbf{s}_d$ will approximately be 0 and 3 Å in the [N] and [Z] forms, respectively, and will become larger than ~3 Å in the other two cases.

During the OPES simulations, both CVs $\mathbf{s}_p$ and $\mathbf{s}_d$ use the value of $\lambda = 5$ so as to have a smoother definition of the Voronoi polyhedra.

Since H_3O^+ and OH^- can be present in water, we have added configurations related to the autoionization process of water into our training datasets. In the potential training process we use the CV

$$\mathbf{s}_a = \sum_{i=1}^{\mathrm{Num_{O_w}}} \delta_i^2 \tag{6}$$

to promote the self ionization processes, and $\mathbf{s}_a$ varies from 0 to 2.

The CV $\mathbf{s}_a$ is supplemented by a CV $\mathbf{s}_t$ that represents the distance between H_3O^+ and OH^-:

$$\mathbf{s}_t = - \sum_{i=1}^{\mathrm{Num_{O_w}}} \sum_{j>i}^{\mathrm{Num_{O_w}}} r_{i,j}\delta_i\delta_j \tag{7}$$

While $\mathbf{s}_t$ can distinguish between the pure water state ($\mathbf{s}_t \approx 0$) and autoionization state ($\mathbf{s}_t > 0$), it is hardly to identify the initial proton transfer which corresponds to a $\mathbf{s}_t$ value nearly concentrated around zero. Therefore, we use a piecewise logarithmic function:

$$\mathbf{s}_t' = \begin{cases} \log(\mathbf{s}_t + \epsilon), & 0 \le \mathbf{s}_t < 1 \\ \mathbf{s}_t - 1 + \log(1+\epsilon), & \mathbf{s}_t \ge 1 \end{cases} \tag{8}$$

where $\epsilon = 0.03$ is a regularization parameter.

2.3.3 Free energy calculation

After the DPMD simulation, the free energy surface (FES) along a given CV can be calculated as follows:

$$F(\mathbf{s}) = -\frac{1}{\beta}\log P(\mathbf{s}) \tag{9}$$

In the regime where the bias is quasi-static, $P(\mathbf{s})$ can be reweighted[45] as an average over the biased ensemble.

$$P(\mathbf{s}) = \frac{\langle \delta[\mathbf{s} - \mathbf{s}(\mathbf{R})] e^{\beta V(\mathbf{s})} \rangle_V}{\langle e^{\beta V(\mathbf{s})} \rangle_V} \tag{10}$$

See the supporting information for the full simulation workflow and more computational details.

3 Results and discussion

3.1 Free energy surfaces

The converged FES is plotted in Fig. 2a as a function of the two CVs $\mathbf{s}_p$ and $\mathbf{s}_d$, which reflect the glycine protonation state and the charge-charge distance, respectively. The minima corresponding to state [N] and state [Z] are easily identified. Less evident is the presence of two other metastable states, [A] and [C]. Their existence can be made more clear if we project the free energy along the CV $\mathbf{s}_p$ as shown in Fig. 2b. In this one-dimensional representation, the [N] and [C] forms cannot be resolved and are part of one single central minimum, but clearly two local minima that correspond to [A] and [C] can be detected, and these states lie higher in energy relative to the minimum by 44.1 kJ/mol ([A]) and 49.2 kJ/mol ([C]), respectively. To facilitate reading the result in a one dimensional projection, we also plot the FES along the CV $\mathbf{s}_d$ (Fig. 2c). In this projection, [A] and [C] cannot be identified, while the [Z] and [N] states are now clearly visible. The energy difference between the lowest free energy [Z] state and the [N] state is 32.6 kJ/mol, in good agreement with experiments, which have reported values ranging from 30.4 kJ/mol to 32.1 kJ/mol.[73–75]

It is interesting to analyze in some detail the nature of the [A] and [C] states. Because of the requirement that the system must be neutral, the charged glycine protomers in these two states are accompanied by a counterion, which is H_3O^+ in [A] and OH^- in [C]. The behavior of these two ion pairs is different; in [A] the hydronium remains close to the glycine, while in [C] the ion pair can separate more easily. This different behavior is reflected in the different

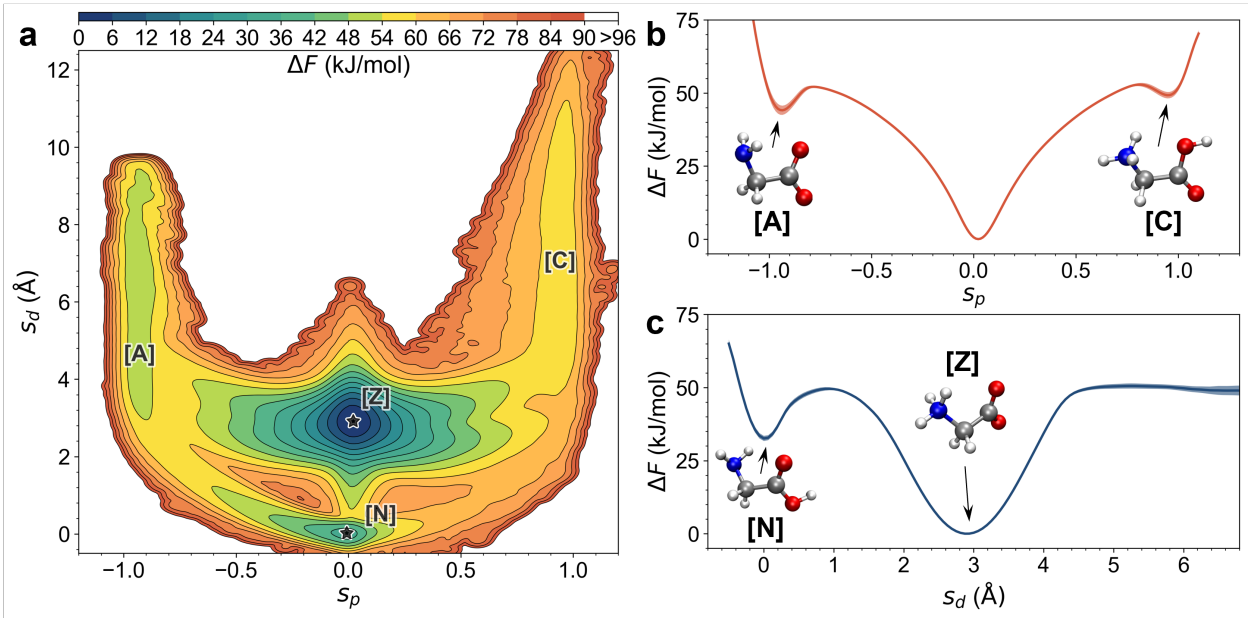

Figure 2: **Free energy profiles. a**. Two-dimensional FES as a function of the glycine protonation coordinate $\mathbf{s}_p$ and the charge-charge distance $\mathbf{s}_d$. **b** and **c**. One-dimentional projections of the FES along the CVs $\mathbf{s}_p$ and $\mathbf{s}_d$. The standard deviations of one-dimentional FESs are represented by transparent colors.

$\mathbf{s}_d$ distributions. The $\mathbf{s}_d$ distribution in [C]–OH^- pair is much broader than that of [A]–H_3O^+ pair, and as a consequence the FES exhibits a slight left-right asymmetry of Fig. 2a.

3.2 Prototropic tautomerism pathways

We now analyze the pathways leading from [N] to [Z]. We have identified three possible proton transfer pathways as shown in Fig. 3a,b. The most obvious is a direct intramolecular proton transfer (Intra-PT), which is consistent with the gas phase studies.[76–78] The $–NH_2$ group first rotates so that its lone pair helps to accept the –COOH proton to form the ammonium group $–NH_3^+$, which later rotates to assume a more stable [Z] conformation.

In addition to the direct route, there are two other pathways in which the proton is transferred via a Grotthuss-like mechanism[44,71,72] with the aid of the water molecules in the solvent. These two mechanisms have either [A] or [C] as an intermediate; in the first case, a contact ion pair [A] – H_3O^+ is formed, whereas in the second one passes through the separated ion pair [C] – OH^-.

In the anionic pathway, the carboxylic proton is transferred to a nearby solvent water, which transfers a positive charge to another water molecule via a Zundel intermediate to end up on the glycine, forming the ammonium group of the [Z] state. This H_3O^+ mediated proton transfer (H_3O^+-PT) process is also found at the microhydration limit of glycine-water clusters.[18,76,77]

Similar but different is the new pathway that leads from [N] to [Z] via an intermediate [C]–OH^- pair mediated proton transfer (OH^--PT). The first step is the transfer of a proton from a water molecule to the amine group of glycine to form an ammonium cation and an OH^- ion. The OH^- ion is then solvated in water and diffuses a relatively long distance via a Grotthuss mechanism, eventually ending up back at the glycine to abstract the carboxylic proton to form the [Z] protomer and a water molecule.

The reason for the distinct charge-charge distance behavior comes from the different amphipathy of H_3O^+ and OH^- ions as depicted in Fig. 3c,d, where the distance between the contact ion pair [A] – H_3O^+ is in the short range (< ~4.5 Å), while the distance between the separated ion pair [C] – OH^- is in the relatively long range (> ~4.5 Å). Since glycine disrupts the HB network of water and provides a hydrophobic environment, H_3O^+ tends to stay in proximity to glycine due to its hydrophobic O atom. Compared to H_3O^+, the O atom in OH^- is hydrophilic and can easily form HB with surrounding water molecules as HB acceptors,[44,79] resulting in a more extensive HB network around OH^-. This indicates that OH^- can diffuse further away from glycine towards the outer water solvation shell compared to H_3O^+, and thus the outer solvation shell of glycine is involved in the prototropic tautomerism and cannot be ignored.

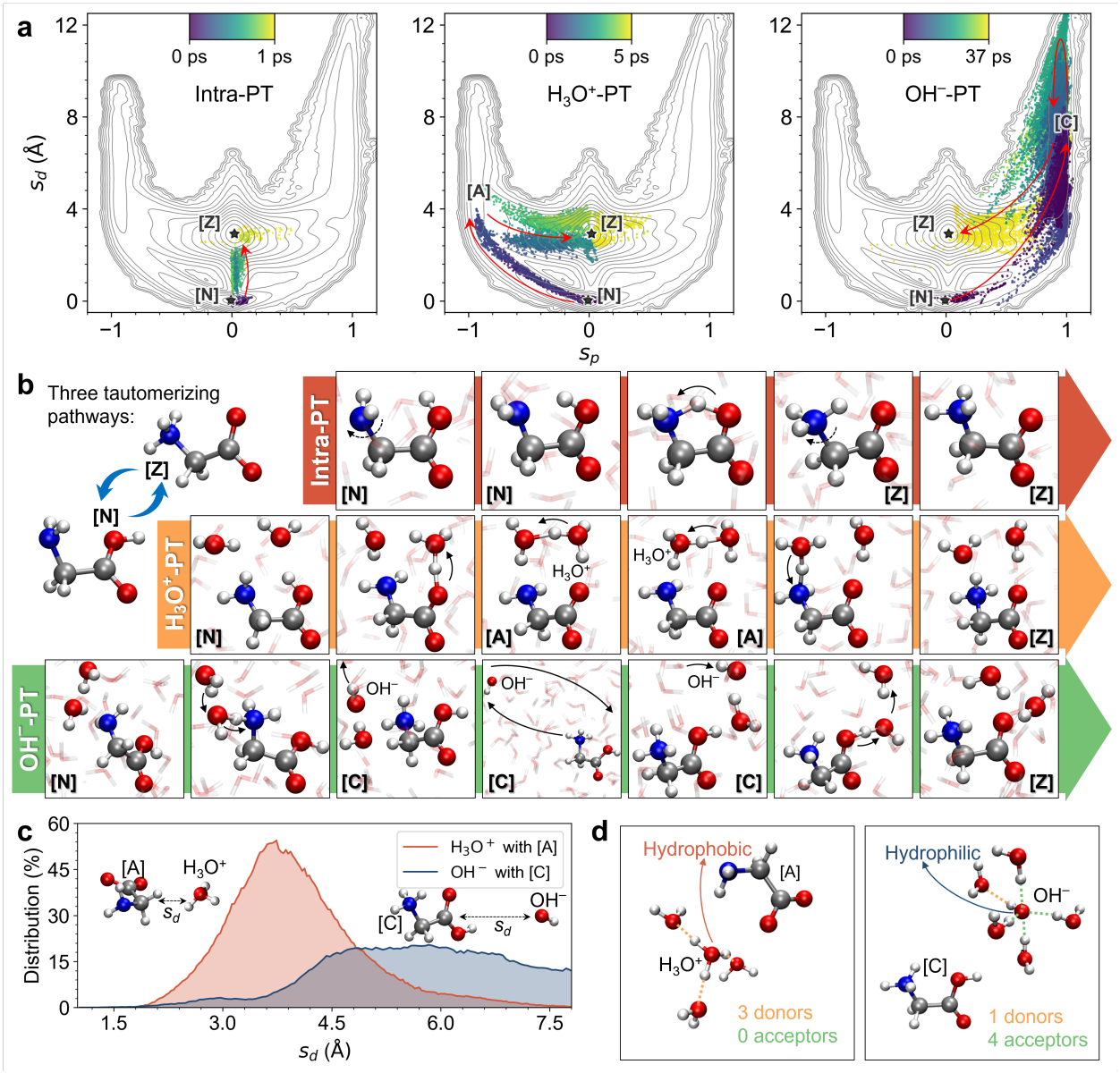

Figure 3: **Three tautomeric pathways. a**. Continuous sampling processes of tautomeric pathways with the CVs $\mathbf{s}_p$ and $\mathbf{s}_d$, where the color bar represents the relative simulation time. **b**. Trajectories of [N]−[Z] tautomerism along the Intra-PT, H_3O^+-PT, and OH^--PT pathways. The molecular configurations involved in the reaction pathways are highlighted in the sphere models, and other surrounding water molecules are shown in the transparent stick models. **c**. The distribution of the CV $\mathbf{s}_d$, i.e., intermolecular charge-charge distance (no intramolecular distance in the [A] and [C] cases). **d**. The hydrogen-bond schematic diagrams of H_3O^+ and OH^- ions near [A] and [C].

Note that experimental investigations[75] on the glycine protonation reactions were mostly conducted in the water with glycine concentrations falling within the interval of 0.03-0.25 M at room temperature. The present simulation results are obtained based on a neutral aqueous simulation system with a little bit higher glycine concentration of 0.43 M. Nevertheless, as documented in the literature, the solubility of glycine in water spans a range of 0.36-0.56 M,[80] and therefore the present simulation ensures a solvated glycine molecule. In addition, it is noteworthy that the discovery of reaction pathways involving ion pairs in the simulation, accompanied by the presence of H_3O^+ or OH^- in the aqueous environment, indicates an instantaneous change of the pH, leading to values of 0.4 or 13.6, respectively. Under the present simulated condition, we have clearly delineated three distinct [N]−[Z] tautomerization pathways for the first time, elucidated the roles and characters of [C]−OH^- and [A]−H_3O^+ ion pairs in these processes, and indicated a way by which the pH of the solution can influence the protomeric equilibrium. This potentially have broader applicability in chemistry, biology, engineering and other proton transfer processes. We look forward to the upcoming advanced experiments that will serve to validate our findings. It is also important to emphasize that the quantitative results of these experiments will depend on the glycine concentration and, more interestingly, the pH of the system.

4 Conclusion

In summary, the present study provides novel insights into the prototropic tautomerism of glycine in water and covers all possible configurational transformations using an accurate description of the interaction potential and a thorough sampling of the potential energy surface. We discover three pathways for tautomerization between the neutral and zwitterionic forms of solvated glycine; one is via intramolecular proton transfer in glycine, the second one involves short-range intermolecular proton transfer in the contact ion pair between anionic glycine and hydronium ion, and the third one has the aid of long-range intermolecular proton transfer in the separated ion pair between cationic glycine and hydroxide ion. In the two intermolecular proton transfer pathways, the observed remarkably distinct charge-charge distance of the intermediate ion pairs is attributed to the different amphipathy of water self-ions.

The combination of our computational technologies, including DeePKS, DeePMD, OPES, and Voronoi CVs, not only deepens our understanding of glycine tautomerism in water, but also provides a comprehensive framework and methodology for facilitating further research into the intricate dynamics of proton transfer.

Acknowledgments

This work received partial support from the National Natural Science Foundation of China under Grant No. [21973053]. The authors would like to express their gratitude to Umberto Raucci, Enrico Trizio, Sudip Das, Linfeng Zhang, Qi Ou, Andrea Rizzi and Francesco Mambretti for their valuable discussions. Computational resources were provided by the High Performance Computing (HPC) platform at Tsinghua University and the HPC Franklin at Fondazione Istituto Italiano di Tecnologia.

Author contributions

All authors contributed to the conception and design of the study. P.Z. and A.T.G. conducted the testing of the enhanced sampling method. P.Z. performed the simulation workflow and data analysis, with the assistance of all authors. X.X. and M.P. provided supervision throughout stages of the project. The initial draft of the manuscript was prepared by P.Z., and all authors participated in editing and reviewing the final version.

References

[1] Elguero, J.; Katritzky, A. R.; Denisko, O. V. Prototropic tautomerism of heterocycles: Heteroaromatic tautomerism-General overview and methodology. *Advances in Heterocyclic Chemistry* **2000**, *76*, P1–P84.

[2] Singh, V.; Fedeles, B. I.; Essigmann, J. M. Role of tautomerism in RNA biochemistry. *Rna* **2015**, *21*, 1–13.

[3] Pospisil, P.; Ballmer, P.; Scapozza, L.; Folkers, G. Tautomerism in computer-aided drug design. *Journal of Receptors and Signal Transduction* **2003**, *23*, 361–371.

[4] Raczyńska, E. D.; Kosińska, W.; Ośmiałowski, B.; Gawinecki, R. Tautomeric equilibria in relation to pi-electron delocalization. *Chemical reviews* **2005**, *105*, 3561–3612.

[5] Douhal, A.; Kim, S.; Zewail, A. Femtosecond molecular dynamics of tautomerization in model base pairs. *Nature* **1995**, *378*, 260–263.

[6] Rodriquez, C. F.; Cunje, A.; Shoeib, T.; Chu, I. K.; Hopkinson, A. C.; Siu, K. M. Proton migration and tautomerism in protonated triglycine. *Journal of the American Chemical Society* **2001**, *123*, 3006–3012.

[7] Nemeria, N. S.; Chakraborty, S.; Balakrishnan, A.; Jordan, F. Reaction mechanisms of thiamin diphosphate enzymes: defining states of ionization and tautomerization of the cofactor at individual steps. *The FEBS journal* **2009**, *276*, 2432–3446.

[8] Sheinblatt, M.; Gutowsky, H. A nuclear magnetic resonance study of the protolysis kinetics of glycine. *Journal of the American Chemical Society* **1964**, *86*, 4814–4820.

[9] Ottosson, N.; Børve, K. J.; Spangberg, D.; Bergersen, H.; Sæthre, L. J.; Faubel, M.; Pokapanich, W.; Oöhrwall, G.; Bjoörneholm, O.; Winter, B. On the Origins of Core- Electron Chemical Shifts of Small Biomolecules in Aqueous Solution: Insights from Photoemission and *ab initio* Calculations of Glycineaq. *Journal of the American Chemical Society* **2011**, *133*, 3120–3130.

[10] Hernández, B.; Pflüger, F.; Kruglik, S. G.; Ghomi, M. Protonation–deprotonation of the glycine backbone as followed by Raman scattering and multiconformational analysis. *Chemical Physics* **2013**, *425*, 104–113.

[11] Locke, M. J.; McIver Jr, R. T. Effect of solvation on the acid/base properties of glycine. *Journal of the American Chemical Society* **1983**, *105*, 4226–4232.

[12] Kumar, S.; Rai, A. K.; Singh, V.; Rai, S. Vibrational spectrum of glycine molecule. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **2005**, *61*, 2741–2746.

[13] Alonso, J. L.; Cocinero, E. J.; Lesarri, A.; Sanz, M. E.; López, J. C. The glycine–water complex. *Angewandte Chemie* **2006**, *118*, 3551–3554.

[14] Schwaab, G.; de Tudela, R. P.; Mani, D.; Pal, N.; Roy, T. K.; Gabas, F.; Conte, R.; Caballero, L. D.; Ceotto, M.; Marx, D.; others Zwitter

ionization of glycine at outer space conditions due to microhydration by six water molecules. *Physical Review Letters* **2022**, *128*, 033001.
[15] Sun, J.; Niehues, G.; Forbert, H.; Decka, D.; Schwaab, G.; Marx, D.; Havenith, M. Understanding THz spectra of aqueous solutions: Glycine in light and heavy water. *Journal of the American Chemical Society* **2014**, *136*, 5031–5038.
[16] Peŕez de Tudela, R.; Marx, D. Water-induced zwitterionization of Glycine: stabilization mechanism and spectral signatures. *The journal of physical chemistry letters* **2016**, *7*, 5137–5142.
[17] Bachrach, S. M. Microsolvation of glycine: a DFT study. *The Journal of Physical Chemistry A* **2008**, *112*, 3722–3730.
[18] Tripathi, R.; Durań Caballero, L.; Peŕez de Tudela, R.; Hoölzl, C.; Marx, D. Unveiling Zwitterionization of Glycine in the Microhydration Limit. *ACS omega* **2021**, *6*, 12676–12683.
[19] Wood, G. P.; Gordon, M. S.; Radom, L.; Smith, D. M. Nature of Glycine and Its α-Carbon Radical in Aqueous Solution: A Theoretical Investigation. *Journal of Chemical Theory and Computation* **2008**, *4*, 1788–1794.
[20] Senn, H. M.; Margl, P. M.; Schmid, R.; Ziegler, T.; Blöchl, P. E. *Ab initio* molecular dynamics with a continuum solvation model. *The Journal of chemical physics* **2003**, *118*, 1089–1100.
[21] Aikens, C. M.; Gordon, M. S. Incremental solvation of nonionized and zwitterionic glycine. *Journal of the American Chemical Society* **2006**, *128*, 12835–12850.
[22] Kayi, H.; Kaiser, R. I.; Head, J. D. A theoretical investigation of the relative stability of hydrated glycine and methylcarbamic acid—from water clusters to interstellar ices. *Physical Chemistry Chemical Physics* **2012**, *14*, 4942–4958.
[23] Valverde, D.; da Costa Ludwig, Z. M.; Da Costa, C. R.; Ludwig, V.; Georg, H. C. Zwitterionization of glycine in water environment: Stabilization mechanism and NMR spectral signatures. *The Journal of Chemical Physics* **2018**, *148*, 024305.
[24] Choi, C. H.; Re, S.; Feig, M.; Sugita, Y. Quantum mechanical/effective fragment potential molecular dynamics (QM/EFP-MD) study on intramolecular proton transfer of glycine in water. *Chemical Physics Letters* **2012**, *539*, 218–221.
[25] Rahaman, O.; Van Duin, A. C.; Goddard III, W. A.; Doren, D. J. Development of a ReaxFF reactive force field for glycine and application to solvent effect and tautomerization. *The Journal of Physical Chemistry B* **2011**, *115*, 249–261.
[26] Car, R.; Parrinello, M. Unified approach for molecular dynamics and density-functional theory. *Physical review letters* **1985**, *55*, 2471.
[27] Born, M.; Heisenberg, W. Zur quantentheorie der molekeln. *Original Scientific Papers Wissenschaftliche Originalarbeiten* **1985**, 216–246.
[28] Sun, J.; Bousquet, D.; Forbert, H.; Marx, D. Glycine in aqueous solution: solvation shells, interfacial water, and vibrational spectroscopy from *ab initio* molecular dynamics. *The Journal of chemical physics* **2010**, *133*, 09B609.
[29] Leung, K.; Rempe, S. B. *Ab initio* molecular dynamics study of glycine intramolecular proton transfer in water. *The Journal of chemical physics* **2005**, *122*, 184506.
[30] Behler, J.; Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical review letters* **2007**, *98*, 146401.
[31] Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Physical review letters* **2010**, *104*, 136403.
[32] Zhang, L.; Han, J.; Wang, H.; Saidi, W.; Car, R.; others End-to-end symmetry preserving inter-atomic potential energy model for finite and extended systems. *Advances in neural information processing systems* **2018**, *31*.
[33] Zhang, L.; Han, J.; Wang, H.; Car, R.; Weinan, E. Deep potential molecular dynamics: a scalable model with the accuracy of quantum mechanics. *Physical review letters* **2018**, *120*, 143001.
[34] Wang, H.; Zhang, L.; Han, J.; E, W. DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics. *Computer Physics Communications* **2018**, *228*, 178–184.
[35] Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. Schnet–a deep learning architecture for molecules and materials. *The Journal of Chemical Physics* **2018**, *148*, 241722.
[36] Schütt, K.; Unke, O.; Gastegger, M. Equivariant message passing for the prediction of tensorial properties and molecular spectra. International Conference on Machine Learning. 2021; pp 9377–9388.
[37] Batzner, S.; Musaelian, A.; Sun, L.; Geiger, M.; Mailoa, J. P.; Kornbluth, M.; Molinari, N.; Smidt, T. E.; Kozinsky, B. E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature communications* **2022**, *13*, 2453.
[38] Hoogeboom, E.; Satorras, V. G.; Vignac, C.; Welling, M. Equivariant diffusion for molecule generation in 3d. International conference on machine learning. 2022; pp 8867–8887.
[39] Zhang, L.; Wang, H.; Car, R.; Weinan, E. Phase diagram of a deep potential water model. *Physical review letters* **2021**, *126*, 236001.
[40] Piaggi, P. M.; Weis, J.; Panagiotopoulos, A. Z.; Debenedetti, P. G.; Car, R. Homogeneous ice nucleation in an *ab initio* machine-learning model of water. *Proceedings of the National Academy of Sciences* **2022**, *119*, e2207294119.
[41] de la Puente, M.; David, R.; Gomez, A.; Laage, D. Acids at the Edge: Why Nitric and Formic Acid Dissociations at Air–Water Interfaces Depend on Depth and on Interface Specific Area. *Journal of the American Chemical Society* **2022**, *144*, 10524–10529.
[42] Galib, M.; Limmer, D. T. Reactive uptake of N_2O_5 by atmospheric aerosol is dominated by interfacial processes. *Science* **2021**, *371*, 921–925.
[43] Yang, M.; Bonati, L.; Polino, D.; Parrinello, M. Using metadynamics to build neural network potentials for reactive events: the case of urea decomposition in water. *Catalysis Today* **2022**, *387*, 143–149.
[44] Zhang, P.; Feng, M.; Xu, X. Double-layer distribution of hydronium and hydroxide ions in the air-water interface. *ChemRxiv* **2023**,
[45] Invernizzi, M.; Parrinello, M. Rethinking metadynamics: From bias potentials to probability distributions. *The journal of physical chemistry letters* **2020**, *11*, 2731–2736.
[46] Invernizzi, M.; Parrinello, M. Exploration vs convergence speed in adaptive-bias enhanced sampling. *Journal of Chemical Theory and Computation* **2022**, *18*, 3988–3996.
[47] Laio, A.; Parrinello, M. Escaping free-energy minima. *Proceedings of the national academy of sciences* **2002**, *99*, 12562–12566.
[48] Barducci, A.; Bussi, G.; Parrinello, M. Well-tempered metadynamics: a smoothly converging and tunable free-energy method. *Physical review letters* **2008**, *100*, 020603.
[49] Grifoni, E.; Piccini, G.; Parrinello, M. Microscopic description of acid–base equilibrium. *Proceedings of the National Academy of Sciences* **2019**, *116*, 4054–4057.
[50] Grifoni, E.; Piccini, G.; Parrinello, M. Tautomeric equilibrium in condensed phases. *Journal of Chemical Theory and Computation* **2020**, *16*, 6027–6031.
[51] Chen, Y.; Zhang, L.; Wang, H.; E, W. Ground state energy functional with Hartree–Fock efficiency and chemical accuracy. *The Journal of Physical Chemistry A* **2020**, *124*, 7155–7165.
[52] Chen, Y.; Zhang, L.; Wang, H.; E, W. DeePKS: A comprehensive data-driven approach toward chemically accurate density functional theory. *Journal of Chemical Theory and Computation* **2020**, *17*, 170–181.
[53] Chen, Y.; Zhang, L.; Wang, H.; Weinan, E. DeePKS-kit: A package for developing machine learning-based chemically accurate energy and density functional models. *Computer Physics Communications* **2023**, *282*, 108520.
[54] Li, W.; Ou, Q.; Chen, Y.; Cao, Y.; Liu, R.; Zhang, C.; Zheng, D.; Cai, C.; Wu, X.; Wang, H.; others DeePKS+ABACUS as a Bridge between Expensive Quantum Mechanical Models and Machine Learning Potentials. *The Journal of Physical Chemistry A* **2022**, *126*, 9154–9164.
[55] Zhao, Y.; Truhlar, D. G. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theoretical chemistry accounts* **2008**, *120*, 215–241.
[56] Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized gradient approximation made simple. *Physical review letters* **1996**, *77*, 3865.
[57] Kühne, T. D.; Iannuzzi, M.; Del Ben, M.; Rybkin, V. V.; Seewald, P.; Stein, F.; Laino, T.; Khaliullin, R. Z.; Schütt, O.; Schiffmann, F.; others CP2K: An electronic structure and molecular dynamics software package-Quickstep: Efficient and accurate electronic structure calculations. *The Journal of Chemical Physics* **2020**, *152*, 194103.
[58] Goedecker, S.; Teter, M.; Hutter, J. Separable dual-space Gaussian pseudopotentials. *Physical Review B* **1996**, *54*, 1703.
[59] Hartwigsen, C.; Gœdecker, S.; Hutter, J. Relativistic separable dual-space Gaussian pseudopotentials from H to Rn. *Physical Review B* **1998**, *58*, 3641.
[60] VandeVondele, J.; Krack, M.; Mohamed, F.; Parrinello, M.; Chassaing, T.; Hutter, J. Quickstep: Fast and accurate density functional calculations using a mixed Gaussian and plane waves approach. *Computer Physics Communications* **2005**, *167*, 103–128.
[61] Guidon, M.; Hutter, J.; VandeVondele, J. Auxiliary density matrix methods for Hartree- Fock exchange calculations. *Journal of chemical theory and computation* **2010**, *6*, 2348–2364.
[62] Chen, M.; Guo, G.; He, L. Systematically improvable optimized atomic basis sets for *ab initio* calculations. *Journal of Physics: Condensed*

Matter **2010**, *22*, 445501.
[63] Li, P.; Liu, X.; Chen, M.; Lin, P.; Ren, X.; Lin, L.; Yang, C.; He, L. Large-scale *ab initio* simulations based on systematically improvable atomic basis. *Computational Materials Science* **2016**, *112*, 503–517.
[64] Schlipf, M.; Gygi, F. Optimization algorithm for the generation of ONCV pseudopotentials. *Computer Physics Communications* **2015**, *196*, 36–44.
[65] Lin, P.; Ren, X.; He, L. Strategy for constructing compact numerical atomic orbital basis sets by incorporating the gradients of reference wavefunctions. *Physical Review B* **2021**, *103*, 235131.
[66] Zeng, J.; Zhang, D.; Lu, D.; Mo, P.; Li, Z.; Chen, Y.; Rynik, M.; Huang, L.; Li, Z.; Shi, S.; others DeePMD-kit v2: A software package for deep potential models. *The Journal of Chemical Physics* **2023**, *159*.
[67] Grimme, S.; Bannwarth, C.; Shushkov, P. A robust and accurate tight-binding quantum chemical method for structures, vibrational frequencies, and noncovalent interactions of large molecular systems parametrized for all spd-block elements (Z= 1–86). *Journal of chemical theory and computation* **2017**, *13*, 1989–2009.
[68] Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *The Journal of chemical physics* **2007**, *126*, 014101.
[69] Plimpton, S. Fast parallel algorithms for short-range molecular dynamics. *Journal of computational physics* **1995**, *117*, 1–19.
[70] Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. PLUMED 2: New feathers for an old bird. *Computer physics communications* **2014**, *185*, 604–613.
[71] von Grotthuß, T. *Mémoire sur la décomposition de l'eau et des corps qu'elle tient en dissolution à l'aide de l'électricité galvanique*; 1805.
[72] Marx, D. Proton transfer 200 years after von Grotthuss: Insights from *ab initio* simulations. *ChemPhysChem* **2006**, *7*, 1848–1870.
[73] Wada, G.; Tamura, E.; Okina, M.; Nakamura, M. On the ratio of zwitterion form to uncharged form of glycine at equilibrium in various aqueous media. *Bulletin of the Chemical Society of Japan* **1982**, *55*, 3064–3067.
[74] Haberfield, P. What is the energy difference between $H_2NCH_2CO_2H$ and $^+H_3NCH_2CO_2^-$? *Journal of Chemical Education* **1980**, *57*, 346.
[75] Slifkin, M.; Ali, S. Thermodynamic parameters of the activation of glycine zwitterion protonation reactions. *Journal of molecular liquids* **1984**, *28*, 215–221.
[76] Jensen, J. H.; Gordon, M. S. On the number of water molecules necessary to stabilize the glycine zwitterion. *Journal of the American Chemical Society* **1995**, *117*, 8159–8170.
[77] Tang, W.; Cai, C.; Zhao, S.; Liu, H. Development of reaction density functional theory and its application to glycine tautomerization reaction in aqueous solution. *The Journal of Physical Chemistry C* **2018**, *122*, 20745–20754.
[78] Tuñón, I.; Silla, E.; Ruiz-López, M. F. On the tautomerization process of glycine in aqueous solution. *Chemical Physics Letters* **2000**, *321*, 433–437.
[79] Chen, M.; Zheng, L.; Santra, B.; Ko, H.-Y.; DiStasio Jr, R. A.; Klein, M. L.; Car, R.; Wu, X. Hydroxide diffuses slower than hydronium in water because its solvated structure inhibits correlated proton transfer. *Nature chemistry* **2018**, *10*, 413–419.
[80] Bowden, N. A.; Sanders, J. P.; Bruins, M. E. Solubility of the proteinogenic α-amino acids in water, ethanol, and ethanol–water mixtures. *Journal of Chemical & Engineering Data* **2018**, *63*, 488–497.

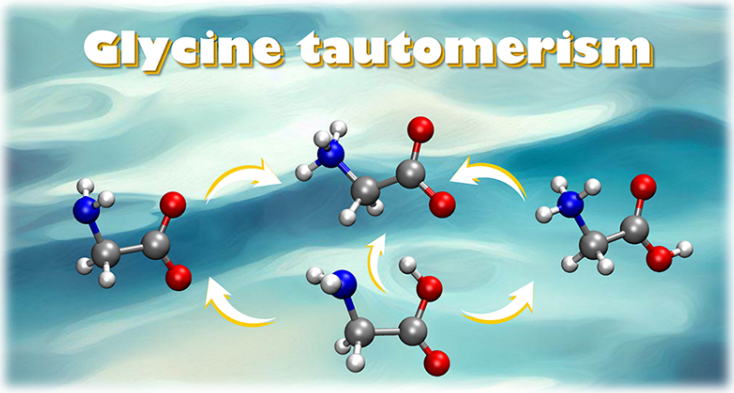

Table of contents (TOC)

–Supporting infomation–

Intramolecular and water mediated tautomerism of solvated glycine

Pengchao Zhang[1,2], **Axel Tosello Gardini**[2,3], **Xuefei Xu**[1,*], and **Michele Parrinello**[2,*].

[1]Center for Combustion Energy, Department of Energy and Power Engineering, and Key Laboratory for Thermal Science and Power Engineering of Ministry of Education, Tsinghua University, Beijing 100084, China

[2]Atomistic Simulations, Italian Institute of Technology, Genova 16152, Italy

[3]Department of Materials Science, Università di Milano-Bicocca, 20126 Milano, Italy

[*]Co-corresponding author e-mail: xuxuefei@tsinghua.edu.cn, michele.parrinello@iit.it

November 13, 2023

DeePKS model training

The DeePKS model employs descriptors derived from the projected density matrices, necessitating a predefined set of projectors with a maximum angular momentum of 2. The number of Bessel functions is determined by the radial and wavefunction cutoffs, specifically 5 Bohr and 100 Ry, respectively. The number of neurons is set to $[32, 32, 32]$ in hidden layers and the non-linear activation function between hidden layers is "gelu". The train step is set to 10 thousand with a learning rate of 10^{-4} to 10^{-7}. A weighted mean square loss function $\mathcal{L}^{\mathrm{DeePKS}}$ is applied to energy and atomic forces.

$$\mathcal{L}^{\mathrm{DeePKS}} = \left|E^{\mathrm{M06-2X}} - E^{\mathrm{DeePKS}}\left(\varphi_i|w\right)\right|^2 + p\left|\mathbf{F}^{\mathrm{M06-2X}} - \mathbf{F}^{\mathrm{DeePKS}}\left(\varphi_i|w\right)\right|^2 \tag{S1}$$

Energy and force are determined by the Hamiltonian. The total Hamiltonian of the DeePKS model ($\hat{H}^{\mathrm{DeePKS}}$) is expressed as follows:

$$\hat{H}^{\mathrm{DeePKS}} = \hat{H}^{\mathrm{PBE}} + \hat{H}^{\delta} \tag{S2}$$

where $\hat{H}^{\mathrm{DeePKS}}$ incorporates a correction term ($\hat{H}^{\delta}$) that is added onto the baseline functional ($\hat{H}^{\mathrm{PBE}}$). φ_i is the eigenstate of Hamiltonian $\hat{H}^{\mathrm{DeePKS}}$ depending on the correction term $\hat{H}^{\delta}$. $\hat{H}^{\delta}$ is determined by projected density matrices, localized orbitals, and the DeePKS descriptor. $\hat{H}^{\mathrm{DeePKS}}$ after adding $\hat{H}^{\delta}$ corresponds to different ground states, so the model training and SCF solution are performed in turn, iterating until convergence.[1,2] $\hat{H}^{\mathrm{DeePKS}}$ is approximately equal to the M06-2X Hamiltonian ($\hat{H}^{\mathrm{M06-2X}}$) when the wave function is converged, then the corresponding energy and force can be solved. In addition, w is the neural network parameter. p is the pre-factor to balance the errors.

DP model training

The descriptor of Deep Potential Smooth Edition (DeepPot-SE) is used,[3,4] where the embedding net size is set to $[25, 50, 100]$ per layer and the non-linear activation function between hidden layers is "tanh". The submatrix size in the embedding net is set to 16. A neighbor atom cutoff radius of 6.0 Å with smoothing beginning at 0.5 Å, and the maximum neighbor number of $[\mathrm{H}, \mathrm{O}, \mathrm{N}, \mathrm{C}]$ is $[85, 45, 1, 2]$. The fitting net is connected after the embedding net, with a size of $[240, 240, 240]$ per layer and the non-linear activation function between hidden layers is "tanh". The ResNet[5] architecture is built between the fitting net. The train step is set to 10 million with a learning rate of 10^{-3} to

10^{-8}. A weighted mean square loss function $\mathcal{L}^{\mathrm{DP}}$ is applied to energy and atomic forces.

$$\mathcal{L}^{\mathrm{DP}} = \frac{p_e}{N}\left|E^{\mathrm{DeePKS}} - E^{\mathrm{DP}}(w)\right|^2 + \frac{p_f}{3N}\sum_{i\alpha}\left|F_{i\alpha}^{\mathrm{DeePKS}} - F_{i\alpha}^{\mathrm{DP}}(w)\right|^2 \tag{S3}$$

where N is the number of atoms. p_e and p_f are the pre-factor of energy E and force F, respectively. w is the neural network parameter. $F_{i\alpha}$ means the force of the atom i along the αth direction.

Derivatives of collective variables

When adapting Voronoi CVs ($\mathbf{s}(\mathbf{R})$) to distinguish glycine states, it is important to note that these CVs can be derived with respect to the atomic coordinates ($\mathbf{R}$), as indicated in the following equation S4:

$$\frac{\partial V(\mathbf{s})}{\partial \mathbf{R}} = \frac{\partial V(\mathbf{s})}{\partial \mathbf{s}(\mathbf{R})}\frac{\partial \mathbf{s}(\mathbf{R})}{\partial \mathbf{R}} \tag{S4}$$

where the implementation of the first term $\frac{\partial V(\mathbf{s})}{\partial \mathbf{s}(\mathbf{R})}$ has been completed within the PLUMED[6] plugin, we will proceed to describe the second term $\frac{\partial \mathbf{s}(\mathbf{R})}{\partial \mathbf{R}}$ in detail and make corresponding changes to the code. For clarity and ease of understanding, we will focus on the relatively complicated part of the Voronoi CVs, as represented by the following equation S5. If there is a need to obtain derivatives for other CVs ($\mathbf{s}_p$, $\mathbf{s}_d$, $\mathbf{s}_a$, and $\mathbf{s}_t$), such calculations can be easily achieved by applying the derivative chain rule.

$$\mathbf{s}(\mathbf{R}) = \sum_{i\in\mathrm{group}} \delta_i \tag{S5}$$

Next, we will illustrate the process using the example of distinguishing autoionization in pure water without any prior information. In this scenario, all O atoms act as Voronoi centers and form a defined group. To identify the protonation states of all O atoms, all nearby H atoms around the target O_i atom (labeled as $\mathrm{O}_i \rightarrow \mathrm{H}_{\mathrm{all}}$) and all nearby O atoms around the target H_j atom (labeled as $\mathrm{O}_i \rightarrow \mathrm{H}_j \rightarrow \mathrm{O}_{\mathrm{all}}$) are searched.

In $\mathrm{O}_i \rightarrow \mathrm{H}_{\mathrm{all}}$ stage, if H_j (selected in $\mathrm{H}_{\mathrm{all}}$) is a neighbor of a given target O_i, the derivative for H_j coordinate is governed by the following equation S6. For simplicity, some small terms are ignored.

$$\frac{\partial \mathbf{s}(\mathbf{R})}{\partial \mathbf{R}_{\mathrm{H}_j}} \approx \sum_{i\in\mathrm{O}}\sum_{j\in\mathrm{H}} -\lambda \frac{e^{-\lambda\left|\mathbf{R}_{\mathrm{O}_i}-\mathbf{R}_{\mathrm{H}_j}\right|}}{\sum_{m\in\mathrm{O}} e^{-\lambda\left|\mathbf{R}_{\mathrm{O}_m}-\mathbf{R}_{\mathrm{H}_j}\right|}}\left[1 - \frac{e^{-\lambda\left|\mathbf{R}_{\mathrm{O}_i}-\mathbf{R}_{\mathrm{H}_j}\right|}}{\sum_{m\in\mathrm{O}} e^{-\lambda\left|\mathbf{R}_{\mathrm{O}_m}-\mathbf{R}_{\mathrm{H}_j}\right|}}\right]\frac{\mathbf{R}_{\mathrm{H}_j}-\mathbf{R}_{\mathrm{O}_i}}{\left|\mathbf{R}_{\mathrm{O}_i}-\mathbf{R}_{\mathrm{H}_j}\right|} \tag{S6}$$

if H_j (selected in $\mathrm{H}_{\mathrm{all}}$) is a neighbor of a non-target Voronoi center O_n ($n \neq i$), the derivative for H_j coordinate is given by the following equation S7. For simplicity, some small terms are ignored.

$$\frac{\partial \mathbf{s}(\mathbf{R})}{\partial \mathbf{R}_{\mathrm{H}_j}} \approx \sum_{i\in\mathrm{O}}\sum_{j\in\mathrm{H}} -\lambda \frac{e^{-\lambda\left|\mathbf{R}_{\mathrm{O}_n}-\mathbf{R}_{\mathrm{H}_j}\right|}}{\sum_{m\in\mathrm{O}} e^{-\lambda\left|\mathbf{R}_{\mathrm{O}_m}-\mathbf{R}_{\mathrm{H}_j}\right|}}\left[0 - \frac{e^{-\lambda\left|\mathbf{R}_{\mathrm{O}_i}-\mathbf{R}_{\mathrm{H}_j}\right|}}{\sum_{m\in\mathrm{O}} e^{-\lambda\left|\mathbf{R}_{\mathrm{O}_m}-\mathbf{R}_{\mathrm{H}_j}\right|}}\right]\frac{\mathbf{R}_{\mathrm{H}_j}-\mathbf{R}_{\mathrm{O}_n}}{\left|\mathbf{R}_{\mathrm{O}_n}-\mathbf{R}_{\mathrm{H}_j}\right|} \tag{S7}$$

In $\mathrm{O}_i \rightarrow \mathrm{H}_j \rightarrow \mathrm{O}_{\mathrm{all}}$ stage, if O_i (selected in $\mathrm{O}_{\mathrm{all}}$) is a neighbor of H_j, the derivative for O_i coordinate is governed by the following equation S8.

$$\frac{\partial \mathbf{s}(\mathbf{R})}{\partial \mathbf{R}_{\mathrm{O}_i}} = \sum_{i\in\mathrm{O}}\sum_{j\in\mathrm{H}} -\lambda \frac{e^{-\lambda\left|\mathbf{R}_{\mathrm{O}_i}-\mathbf{R}_{\mathrm{H}_j}\right|}}{\sum_{m\in\mathrm{O}} e^{-\lambda\left|\mathbf{R}_{\mathrm{O}_m}-\mathbf{R}_{\mathrm{H}_j}\right|}}\left[1 - \frac{e^{-\lambda\left|\mathbf{R}_{\mathrm{O}_i}-\mathbf{R}_{\mathrm{H}_j}\right|}}{\sum_{m\in\mathrm{O}} e^{-\lambda\left|\mathbf{R}_{\mathrm{O}_m}-\mathbf{R}_{\mathrm{H}_j}\right|}}\right]\frac{\mathbf{R}_{\mathrm{O}_i}-\mathbf{R}_{\mathrm{H}_j}}{\left|\mathbf{R}_{\mathrm{O}_i}-\mathbf{R}_{\mathrm{H}_j}\right|} \tag{S8}$$

if O_k (selected in $\mathrm{O}_{\mathrm{all}}$, $k \neq i$) is a neighbor of H_j, the derivative for O_k coordinate is given by the following equation S9.

$$\frac{\partial \mathbf{s}(\mathbf{R})}{\partial \mathbf{R}_{\mathrm{O}_k}} = \sum_{i\in\mathrm{O}}\sum_{j\in\mathrm{H}} -\lambda \frac{e^{-\lambda\left|\mathbf{R}_{\mathrm{O}_k}-\mathbf{R}_{\mathrm{H}_j}\right|}}{\sum_{m\in\mathrm{O}} e^{-\lambda\left|\mathbf{R}_{\mathrm{O}_m}-\mathbf{R}_{\mathrm{H}_j}\right|}}\left[0 - \frac{e^{-\lambda\left|\mathbf{R}_{\mathrm{O}_i}-\mathbf{R}_{\mathrm{H}_j}\right|}}{\sum_{m\in\mathrm{O}} e^{-\lambda\left|\mathbf{R}_{\mathrm{O}_m}-\mathbf{R}_{\mathrm{H}_j}\right|}}\right]\frac{\mathbf{R}_{\mathrm{O}_k}-\mathbf{R}_{\mathrm{H}_j}}{\left|\mathbf{R}_{\mathrm{O}_k}-\mathbf{R}_{\mathrm{H}_j}\right|} \approx 0 \tag{S9}$$

At this point, all derivatives with respect to the coordinates have been determined. In addition, we incorporate the neighbor list function, which efficiently searches for nearby atoms, reducing the computational complexity from $O(N[\mathrm{O}_{\mathrm{all}} \times \mathrm{H}_{\mathrm{all}} \times \mathrm{O}_{\mathrm{all}}])$ to $O(N[\mathrm{O}_{\mathrm{all}} \times \mathrm{H}_{\mathrm{neighbor}} \times \mathrm{O}_{\mathrm{neighbor}}])$. In particular, a neighbor list with a cutoff 2.4 Å is set, updated at each step, to search for atoms in groups.

Polarization calculation

In the realm of polarization analysis, a well-established formalism extensively relies on the utilization of maximally localized Wannier functions (MLWFs).[7,8] In a molecular or condensed matter system, polarization arises due to the separation of positive and negative charges. It can be characterized by the dipole moment, which is a measure of the overall charge distribution within the system. When dealing with finite systems, the dipole $\boldsymbol{\mu}$ of the glycine molecule can be calculated as follows:[9,10]

$$\boldsymbol{\mu} = e\left(\sum_{i\in\mathrm{H}}\mathbf{R}_i + 4\sum_{j\in\mathrm{C}}\mathbf{R}_j + 5\sum_{k\in\mathrm{N}}\mathbf{R}_k + 6\sum_{l\in\mathrm{O}}\mathbf{R}_l - 2\sum_{m\in\mathrm{W}}\mathbf{R}_m\right) \tag{S10}$$

where e represents the electronic charge. The positions of the nuclei (H, C, N, and O) and the MLWF centers (W) of glycine are considered in the calculation. Each nucleus without outer valence electrons (H, C, N, and O) carries a positive charge of 1, 4, 5, and 6, respectively. Additionally, each MLWF center is associated with 2 valence electrons. The MLWF centers are determined through the calculation in CP2K.[11] The structures of solvated glycine are obtained from the enhanced sampling trajectory of system A (Table S1). The remaining settings are kept consistent with the DFT labeling for the DeePKS dataset.

Other computational details

Hydrogen bond (HB) analysis is performed using the MDAnalysis[12,13] library, where the distance cutoff between the donor and the acceptor is 3.5 Å, and the donor-hydrogen-acceptor angle cutoff is set to 140°. To compare the explicit solvation effect, the glycine clusters are also modeled using the implicit solvation model (SMD)[14] at the M06-2X/def2-TZVP level[15–17] in the Gaussian.[18] The simulation results are visualized using the VMD[19] and Matplotlib[20] packages. The diffusion coefficient is calculated from the mean square displacement (MSD) using Einstein's relation. The diffusion coefficient of water is $2.36 \pm 0.09 \times 10^{-9}$ m^2/s at 300 K, which is consistent with experiments (~2.3×10^{-9} m^2/s). The general simulation workflow (Fig. S1) comprises six steps mentioned in manuscript.

Solvation and polarization

Notably, the ongoing research into the minimum number of water molecules required to stabilize the [Z] form has remained controversial, with debates ranging from two to ten water molecules, and means that effective treatment and accurate description of the solvent effect remains a crucial and challenging task.[21–28] We discuss the solvation and polarization of glycine in the SI, which in turn fine-tune the thermodynamics and dynamic behavior.[29]

The gradual expansion of the solvent shell is studied by the radial distribution function (RDF)[30,31] to understand the structural distribution between glycine and water molecules, as depicted in the Fig. S17a. In the region of the first solvent shell (FSS), the RDFs of $O_g - O_w$ display a peak shift between 2.4 and 2.8 Å in the [A]↔[N] or [Z]↔[C] path, which is relative to the protonation of the $–COO^-$ group. The situation is similar to the protonation of the $–NH_2$ group in the [A]↔[Z] or [N]↔[C] path, where the RDFs of $N - O_w$ have a peak shift between 2.5 and 2.9 Å. In the outer solvent shell (OSS), RDFs show slight changes between 3.6 Å and 6 Å depending on changes in FSS. Minimal variations are observed beyond the homogeneous region that extends beyond 6 Å, suggesting that FSS predominantly influences the solvation of glycine, and this is also supported by the THz spectra.[32]

In the FSS, the [Z], [N], [A], and [C] forms are associated with approximately 12, 11, 10, and 12 water molecules, and exhibit 5.4, 4.1, 6.2, and 3.4 HBs, respectively (Fig. S17b). Compared to other forms, the stability of the [Z] form can be attributed to the greater abundance of water molecules within the FSS and the formation of a more extensive HB network, which also contribute to the polarization of [Z].

In full solution, the [Z] form exhibits an average dipole moment of 16.3 ± 0.8 D, whereas in its isolated form, the dipole moment amounts to only 12.6 ± 0.5 D (Fig. S17c). These results align with previous calculation.[31] Remarkably, when exclusively considering the FSS, a dipole moment of 15.4 ± 0.7 D is emulated, effectively mirroring the characteristic polarization of the [Z] form. The [N] form, whether in a gaseous phase or within an aqueous environment, exhibits dipole moments typically ranging from 1.6-2.3 D. This is attributed to the non-ionized state of both the amino and carboxyl groups in the [N] form. In contrast, the presence of water molecules in the FSS mainly stabilizes the charged $–COO^-$ and $–NH_3^+$ groups in the [Z] form, leading to an augmentation of the overall dipole moment.

Specifically, the average distribution value of the electron pair (represented by maximally localized Wannier centers (MLWCs)[7,8]) surrounding O_g^H is approximately 0.15 Å smaller than that of N, and the electron pairs near O_g^H exhibit greater localization compared to that near N (Fig. S17d). This indicates that $–COO^-$ primarily influences the larger dipole moment as opposed to $–NH_3^+$, due to the higher electronegativity of O_g^H.

Table S1: **Details of the simulation systems.** systems A and B are used for building dataset, as well as system C are used for conducting molecular dynamics simulations and subsequent data analysis.

System	Usage	Box size (Å^3)	No. of H_2O	No. (Molar) of glycine
A	Training, testing dataset	$12 \times 12 \times 12$	54	1 (1.03 M)
B	Training, testing dataset	$12 \times 12 \times 12$	58	0
C	Final MD and analysis	$15.8 \times 15.8 \times 15.8$	128	1 (0.43 M)

Table S2: **A convergence test of the single point calculation.** The plane-wave cutoff test of DFT labeling for the DeePKS dataset. The calculation of ΔE involves subtracting previous energy in order to determine the difference, for example, $\Delta E(700\text{Ry}, 60\text{Ry}) = E(700\text{Ry}, 60\text{Ry}) - E(600\text{Ry}, 60\text{Ry})$, $\Delta E(800\text{Ry}, 60\text{Ry}) = E(800\text{Ry}, 60\text{Ry}) - E(700\text{Ry}, 60\text{Ry})$ and so on.

Cutoff (Ry)	Rel_Cutoff (Ry)	ΔE (a.u.)	ΔE (a.u./atom)
600	60	Basis_1	Basis_1/atom
700	60	-3.23×10^{-3}	-1.88×10^{-5}
800	60	5.44×10^{-4}	3.16×10^{-6}
900	60	-5.89×10^{-5}	-3.43×10^{-7}
1000	60	-9.65×10^{-5}	-5.61×10^{-7}
1100	60	8.75×10^{-8}	5.09×10^{-10}
1000	60	Basis_2	Basis_2/atom
1000	70	1.29×10^{-7}	7.51×10^{-10}
1000	80	-2.20×10^{-9}	-1.28×10^{-11}
1000	90	4.00×10^{-10}	2.33×10^{-12}
1000	100	1.00×10^{-10}	5.82×10^{-13}
1000	110	3.00×10^{-10}	1.74×10^{-12}

Table S3: **The runtime of calculations.** DFT labelling and DeePKS labelling are performed on CPU processor with parallelization, where the runtime is for a configuration to complete a self-consistent field iteration. The runtime of DPMD simulation with OPES on GPU is also given.

Step	System	Machine	Time
DFT labeling for the DeePKS dataset	A	24-core CPU	40 min/frame
DFT labeling for the DeePKS dataset	B	24-core CPU	35 min/frame
DeePKS labeling for the DP dataset	A	24-core CPU	5 min/frame
DeePKS labeling for the DP dataset	B	24-core CPU	4 min/frame
DPMD simulation with OPES	A	1 Tesla V100	70 timesteps/s
DPMD simulation with OPES	C	1 Tesla V100	12 timesteps/s

Table S4: **The component of the datasets and the accuracy of the models.** The number of configurations in the train and testing datasets for systems A and B (the DeePKS training datasets and all test datesets are performed by M06-2X/QZV3P, and the DP training datasets are performed by the DeePKS models). The root mean square errors (RMSEs) for the DeePKS and DP models on the test data sets are also given.

Infomation	DeePKS (syst. A)	DeePKS (syst. B)	DP (syst. A)	DP (syst. B)
training dataset	150	150	42,248	13,250
testing dataset	205	165	5,560	460
E_{RMSE} (meV/atom)	0.53	0.61	0.79	1.72
F_{RMSE} (meV/Å)	43	52	58	63

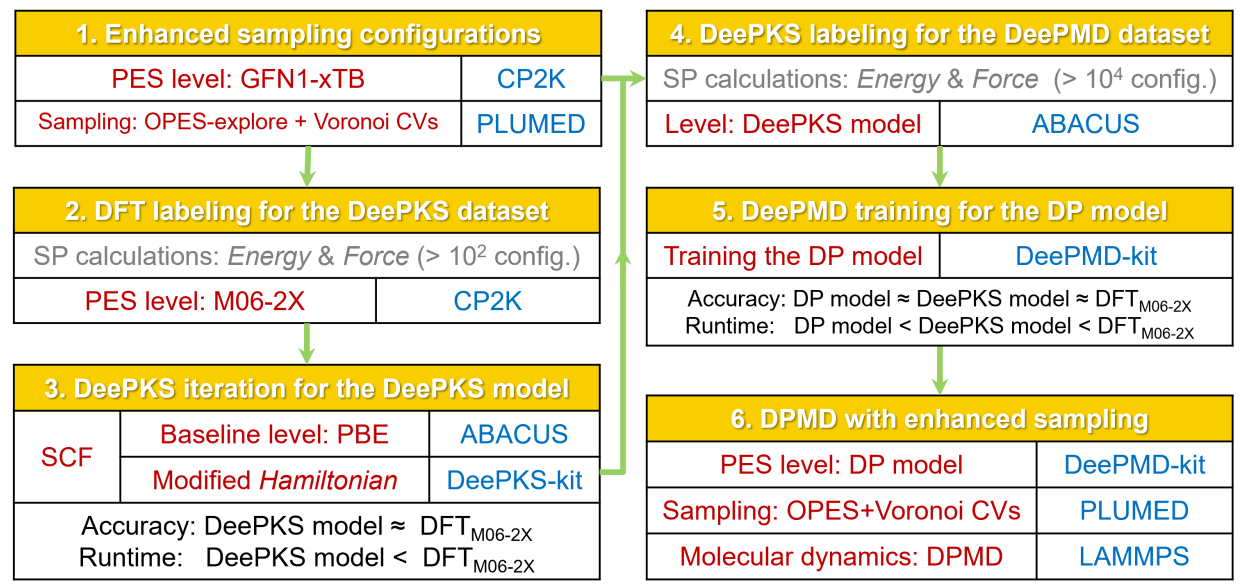

Figure S1: **Simulation workflow in detail.** The comprehensive workflow includes enhanced sampling, DFT labeling, DeePKS iteration, DeePKS labeling, DP training, and DPMD with OPES for efficient and accurate exploration of glycine tautomerism in water.

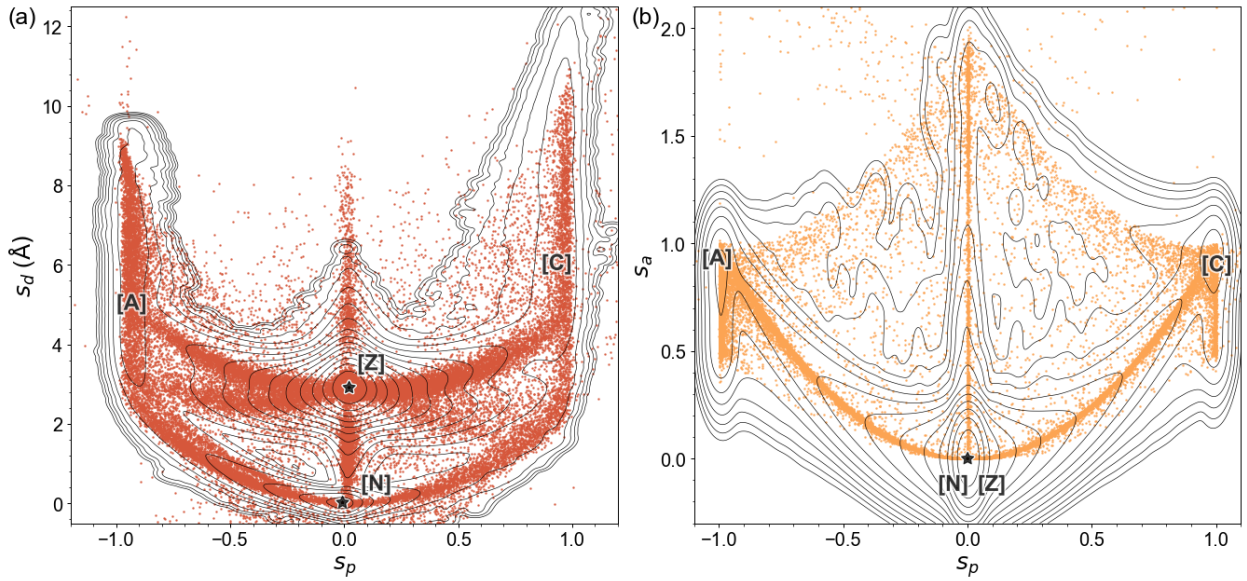

Figure S2: **The distribution of the configuration in the training dataset.** The configurational distribution with respect to (a) glycine protonation ($\mathbf{s}_p^{\lambda=5}$) and charge-charge distance ($\mathbf{s}_d^{\lambda=5}$), (b) glycine protonation ($\mathbf{s}_p^{\lambda=8}$) and number of self-ions ($\mathbf{s}_a^{\lambda=8}$). The origin of Contours is attributed to the utilization of free energy reweighting. The training dataset contains sufficient coverage of a satisfactory phase space using the Voronoi CVs.

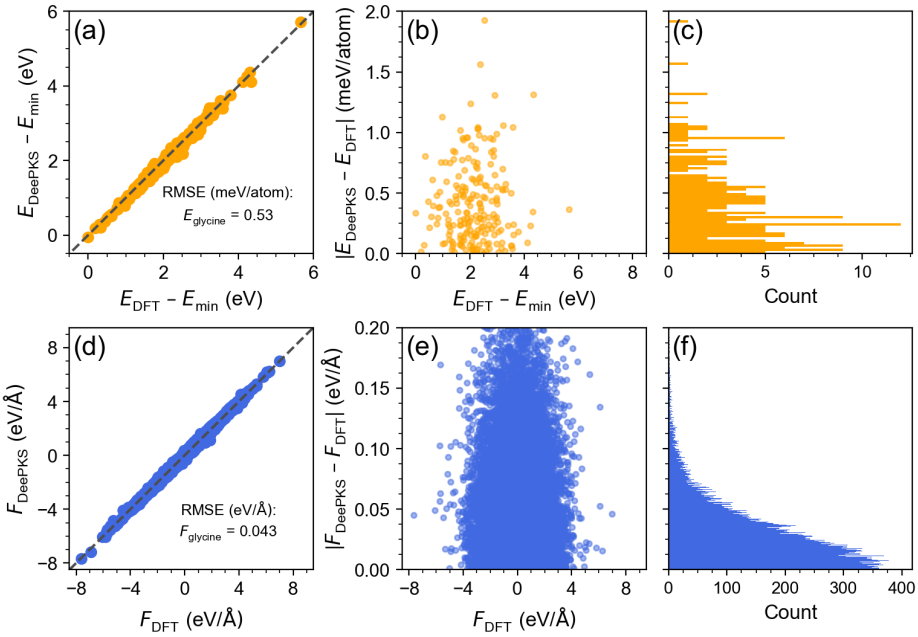

Figure S3: **Error distribution of the DeePKS model on the testing dataset of glycine in water (system A).** (a) The distribution of $E_{\mathrm{DeePKS}} - E_{\mathrm{min}}$ with respect to $E_{\mathrm{DFT}} - E_{\mathrm{min}}$. (b) Left: the distribution of $|E_{\mathrm{DeePKS}} - E_{\mathrm{DFT}}|$ with respect to $E_{\mathrm{DFT}} - E_{\mathrm{min}}$. (c) Right: the histogram of $|E_{\mathrm{DeePKS}} - E_{\mathrm{DFT}}|$. (d) The distribution of F_{DeePKS} with respect to F_{DFT}. (e) Left: the distribution of $|F_{\mathrm{DeePKS}} - F_{\mathrm{DFT}}|$ with respect to F_{DFT}. (f) Right: the histogram of $|F_{\mathrm{DeePKS}} - F_{\mathrm{DFT}}|$. Here, E_{DFT} and F_{DFT} are the energy and force calculated by the M06-2X functional; E_{DeePKS} and F_{DeePKS} are the energy and force performed by the present DeePKS model; and E_{min} is the minimum absolute energy in the testing dataset.

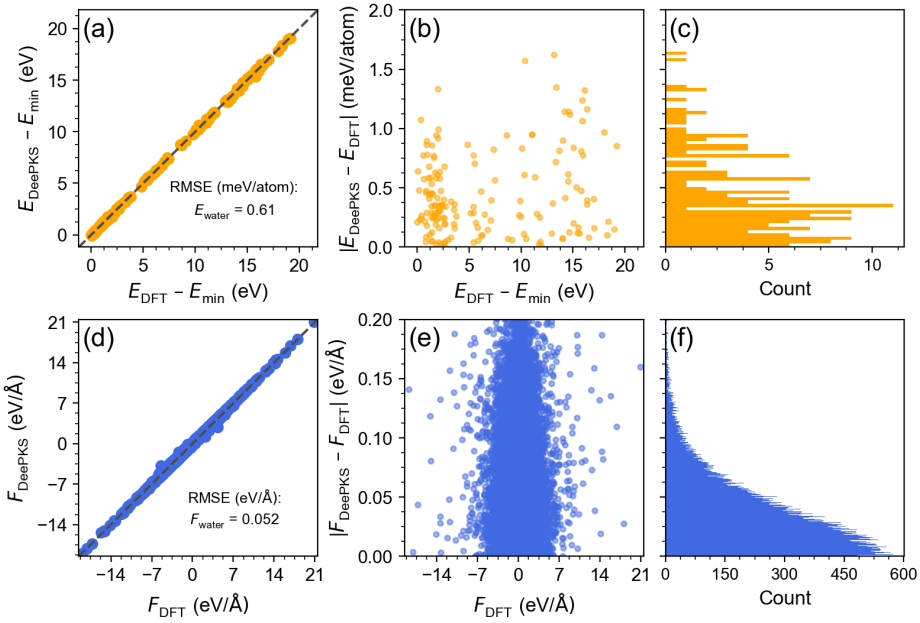

Figure S4: **Error distribution of the DeePKS model on the testing dataset of water (system B).** (a) The distribution of $E_{\mathrm{DeePKS}} - E_{\mathrm{min}}$ with respect to $E_{\mathrm{DFT}} - E_{\mathrm{min}}$. (b) Left: the distribution of $|E_{\mathrm{DeePKS}} - E_{\mathrm{DFT}}|$ with respect to $E_{\mathrm{DFT}} - E_{\mathrm{min}}$. (c) Right: the histogram of $|E_{\mathrm{DeePKS}} - E_{\mathrm{DFT}}|$. (d) The distribution of F_{DeePKS} with respect to F_{DFT}. (e) Left: the distribution of $|F_{\mathrm{DeePKS}} - F_{\mathrm{DFT}}|$ with respect to F_{DFT}. (f) Right: the histogram of $|F_{\mathrm{DeePKS}} - F_{\mathrm{DFT}}|$.

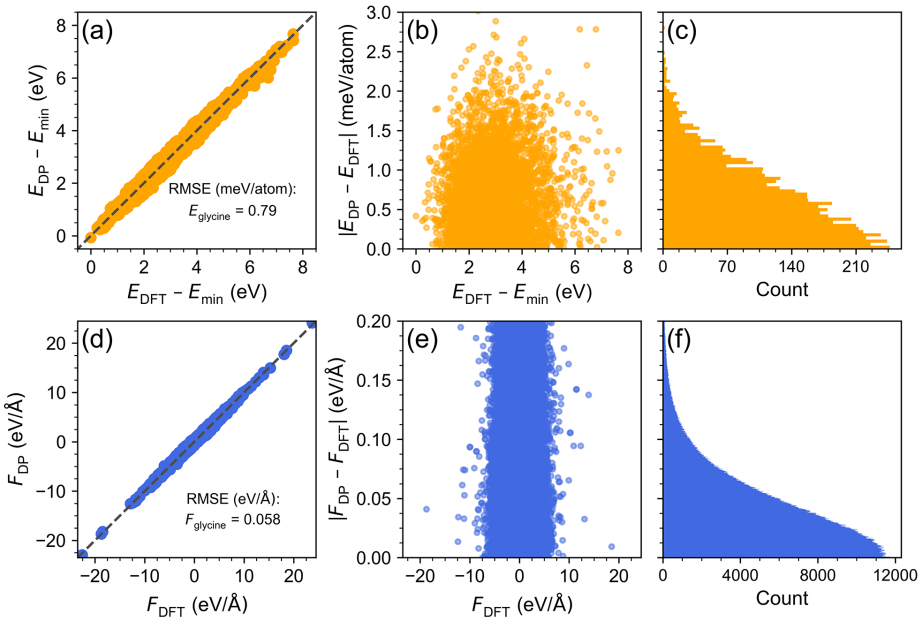

Figure S5: **Error distribution of the DP model on the testing dataset of glycine in water (system A).** (a) The distribution of $E_{DP} - E_{min}$ with respect to $E_{DFT} - E_{min}$. (b) Left: the distribution of $|E_{DP} - E_{DFT}|$ with respect to $E_{DFT} - E_{min}$. (c) Right: the histogram of $|E_{DP} - E_{DFT}|$. (d) The distribution of F_{DP} with respect to F_{DFT}. (e) Left: the distribution of $|F_{DP} - F_{DFT}|$ with respect to F_{DFT}. (f) Right: the histogram of $|F_{DP} - F_{DFT}|$. Here, E_{DFT} and F_{DFT} are the energy and force calculated by the M06-2X functional; E_{DP} and F_{DP} are the energy and force performed by the present DP model; and E_{min} is the minimum absolute energy in the testing dataset.

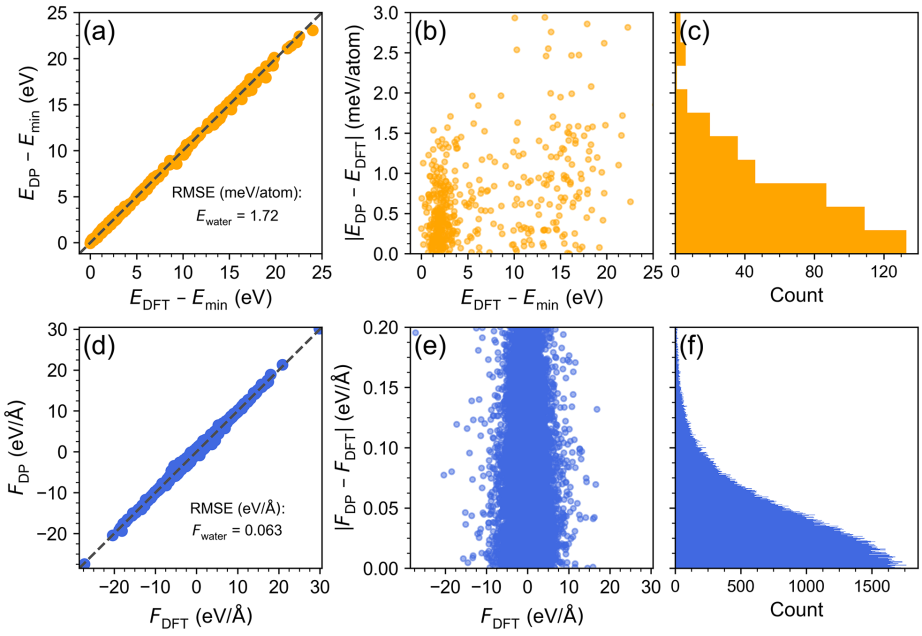

Figure S6: **Error distribution of the DP model on the testing dataset of water (system B).** (a) The distribution of $E_{DP} - E_{min}$ with respect to $E_{DFT} - E_{min}$. (b) Left: the distribution of $|E_{DP} - E_{DFT}|$ with respect to $E_{DFT} - E_{min}$. (c) Right: the histogram of $|E_{DP} - E_{DFT}|$. (d) The distribution of F_{DP} with respect to F_{DFT}. (e) Left: the distribution of $|F_{DP} - F_{DFT}|$ with respect to F_{DFT}. (f) Right: the histogram of $|F_{DP} - F_{DFT}|$.

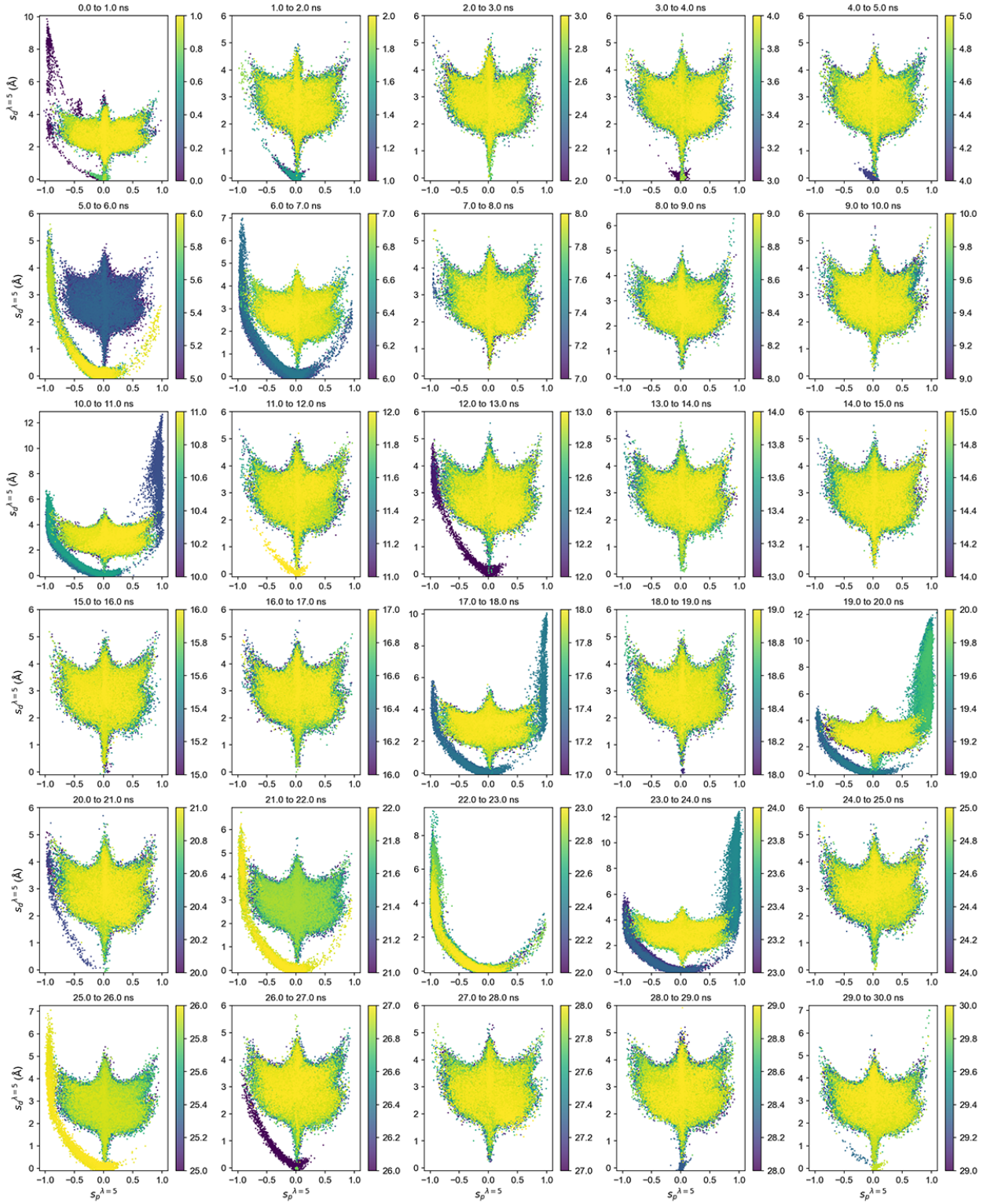

Figure S7: **The CV profiles with time.** The evolution of glycine protonation ($\mathbf{s}_p^{\lambda=5}$) and charge-charge distance ($\mathbf{s}_d^{\lambda=5}$), with each configuration assigned a color code according to the corresponding simulation time represented in the color bar. The sampling shows a discernible tendency to resemble the 'Maple Leaf' and 'U' shapes, and is multiple traversals of all possible configurations.

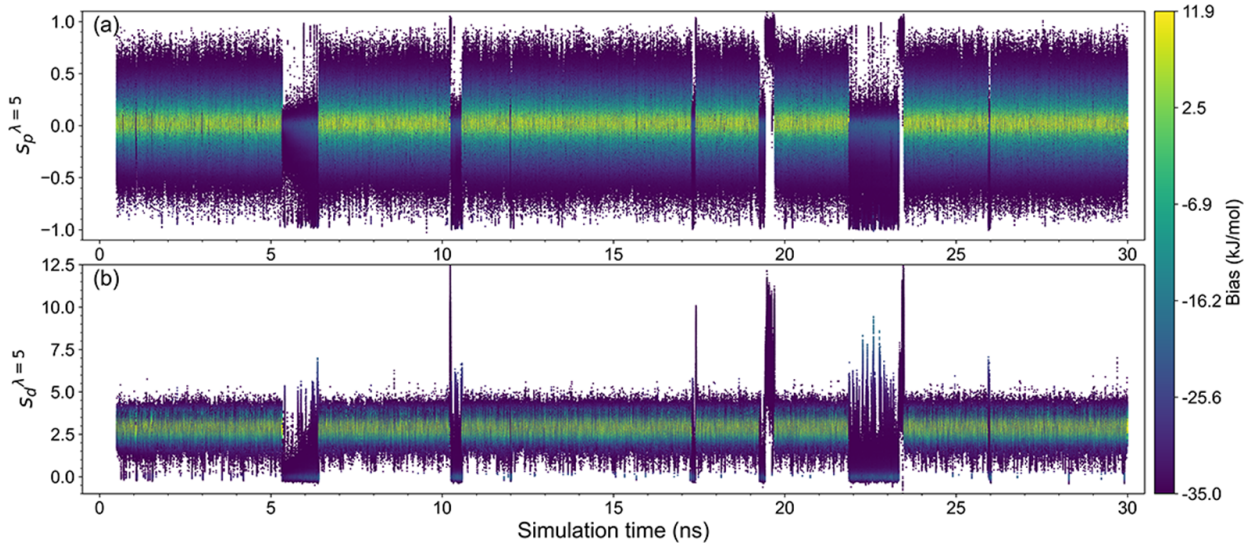

Figure S8: **The CV profiles with bias.** The evolution of (a) glycine protonation ($\mathbf{s}_p^{\lambda=5}$) and (b) charge-charge distance ($\mathbf{s}_d^{\lambda=5}$) over simulation time, with each configuration assigned a color code according to the corresponding bias value represented in the color bar. Based on extensive enhanced sampling, the bias approximately reaches a quasi-static state.

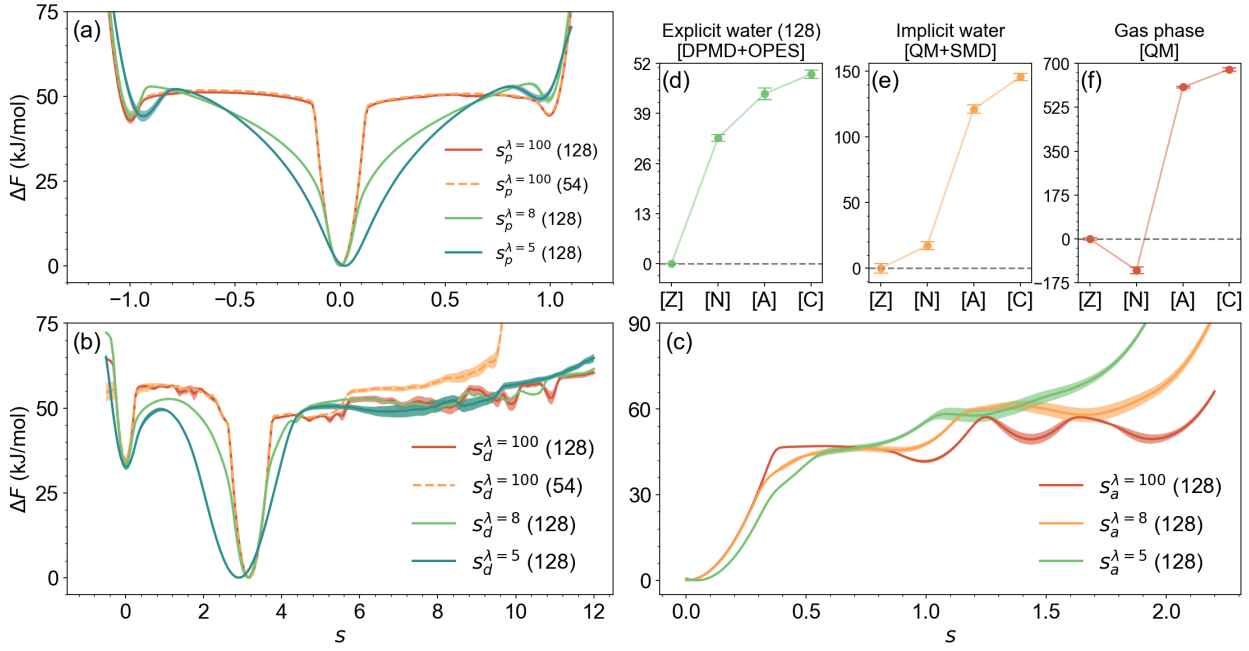

Figure S9: **The one-dimensional free energy and free energy differences of different glycine forms.** The free energy profiles with respect to CVs (a) $\mathbf{s}_p$, (b) $\mathbf{s}_d$, and (c) $\mathbf{s}_a$, where 128 and 54 represent the number of H_2O with a glycine molecule. The inclusion of 128 H_2O takes into account sufficient size effects and enables converged free energy compared to 54 H_2O. The free energy differences (kJ/mol) among glycine in [Z], [N], [A], and [C] forms under (d) explicit water solvation as simulated in this work [DPMD+OPES], (e) implicit water model [QM+SMD], and (f) gas phase [QM]. If we consider the implicit solvation of glycine, only quantitative results can be obtained, suggesting that the interactions ignored by implicit solvation models influence the stabilization of glycine. The QM calculation employed the chemical model method of M06-2X/def2-TZVP.

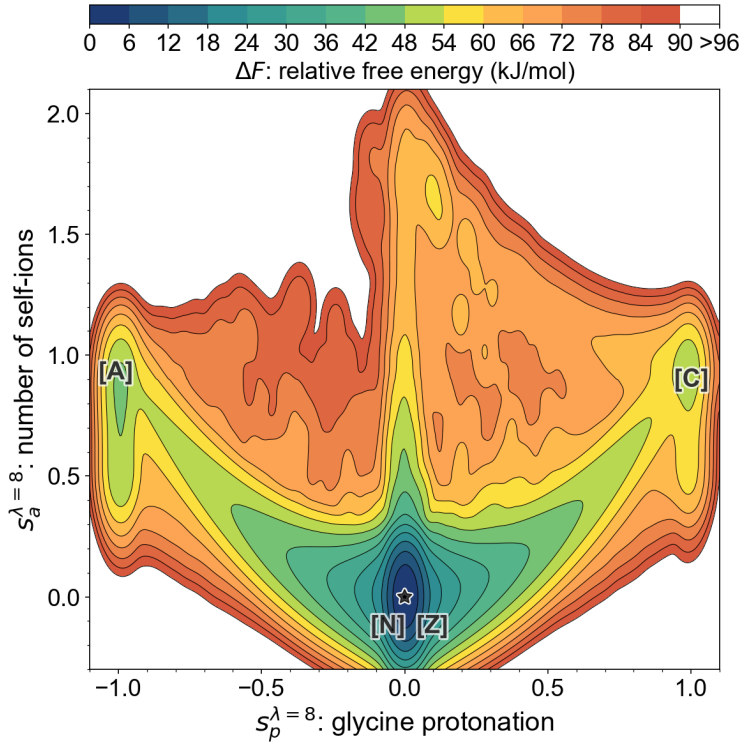

Figure S10: **Two-dimensional free energy surface (FES).** FES with respect to glycine protonation ($\mathbf{s}_p^{\lambda=8}$) and number of self-ions ($\mathbf{s}_a^{\lambda=8}$). There is a H_3O^+ with [A] states ($\mathbf{s}_a^{\lambda=8} \approx 1$), an OH^- with [C] state ($\mathbf{s}_a^{\lambda=8} \approx 1$), and no ion in [N] or [Z] state ($\mathbf{s}_a^{\lambda=8} \approx 0$).

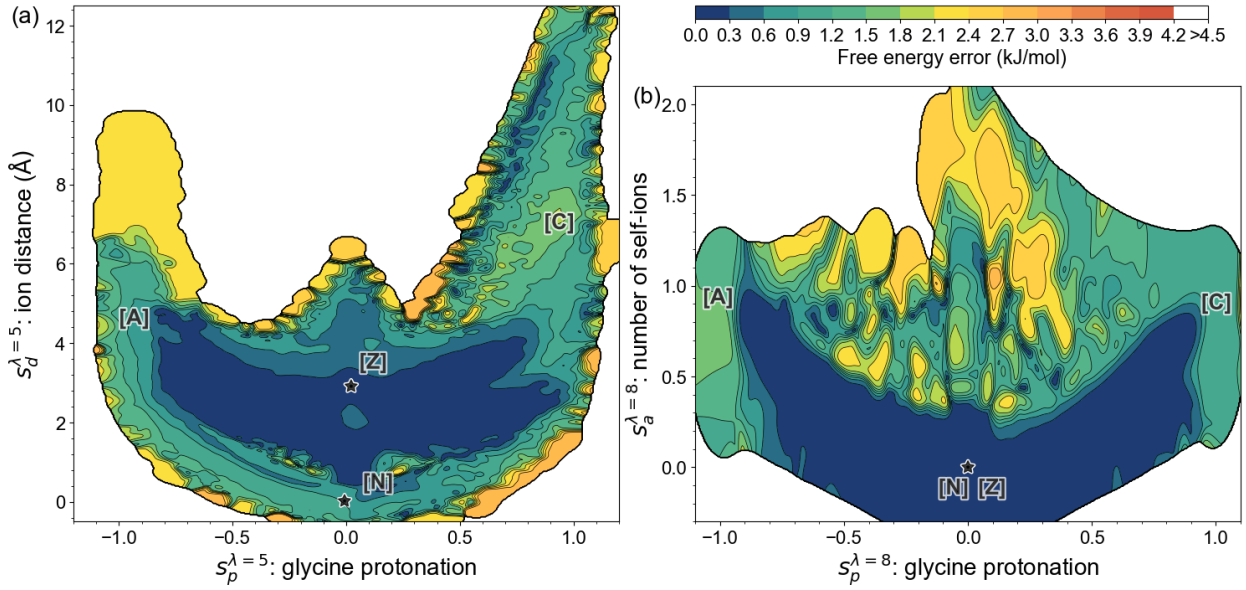

Figure S11: **The standard deviation of free energy surfaces**. Two-dimensional (b) free energy error with respect to glycine protonation ($\mathbf{s}_p^{\lambda=5}$) and charge-charge distance ($\mathbf{s}_d^{\lambda=5}$), and (c) free energy error with respect to glycine protonation ($\mathbf{s}_p^{\lambda=8}$) and number of self-ions ($\mathbf{s}_a^{\lambda=8}$).

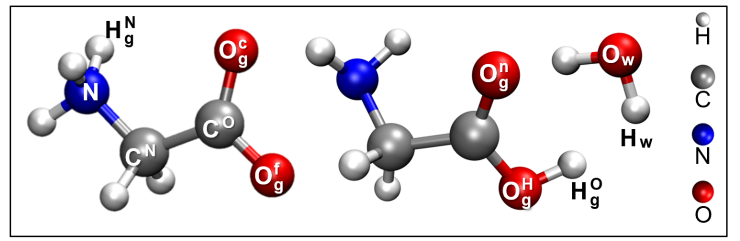

Figure S12: **Atom notations** are used to denote the atoms present in the molecules of glycine and water. X_y represents the atom X (H or O) that originates from molecule y (glycine or water). Specifically, for the $–COO^-$ group, O^c (O^f) designates O that is close to (far from) N. Similarly, for the –COOH group, O^H (O^n) denotes O that does (does not) form a covalent bond with H. And H^N (H^O) represents the H bonded to N (O).

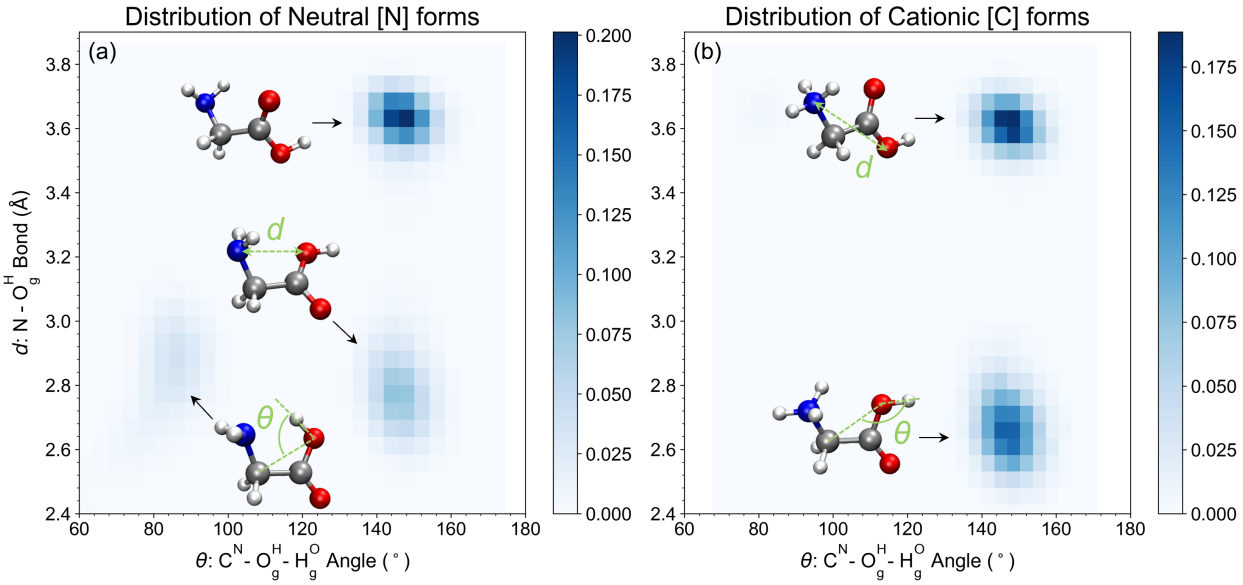

Figure S13: **Configurational distribution.** The distribution of glycine in (a) neutral [N] and (b) cationic [C] forms as a function of the bond (N-O_g^H) and the angle (C^N-O_g^H-H_g^O) as shown in the schematic diagram of molecular structures. The color bars depict the density of configuration numbers.

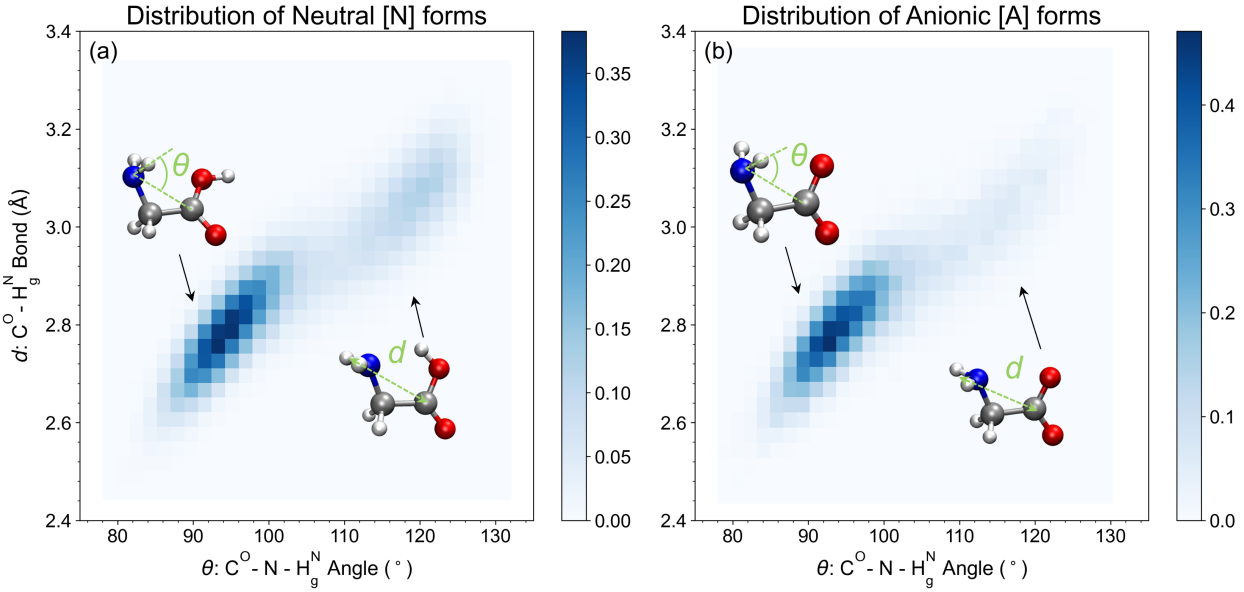

Figure S14: **Configurational distribution.** The distribution of glycine in (a) neutral [N] and (b) anionic [A] forms as a function of the bond (the average of two C^{O}-H_g^{N} bond lengths) and the angle (the average of two C^{O}-N-H_g^{N} angles) as shown in the schematic diagram of molecular structures. The color bars depict the density of configuration numbers.

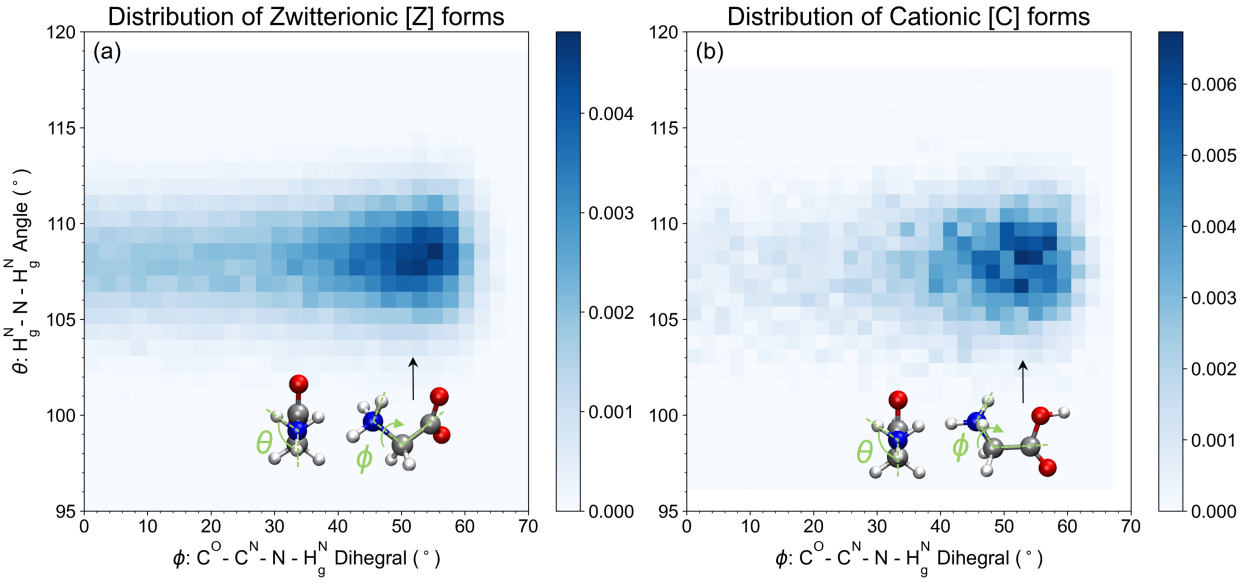

Figure S15: **Configurational distribution.** The distribution of glycine in (a) zwitterionic [Z] and (b) cationic [C] forms as a function of the angle (the average of three H_g^{N}-N-H_g^{N} angles) and the dihedral (the minimum of three C^{O}-C^{N}-N-H_g^{N} dihedrals) as shown in the schematic diagram of molecular structures. The color bars depict the density of configuration numbers.

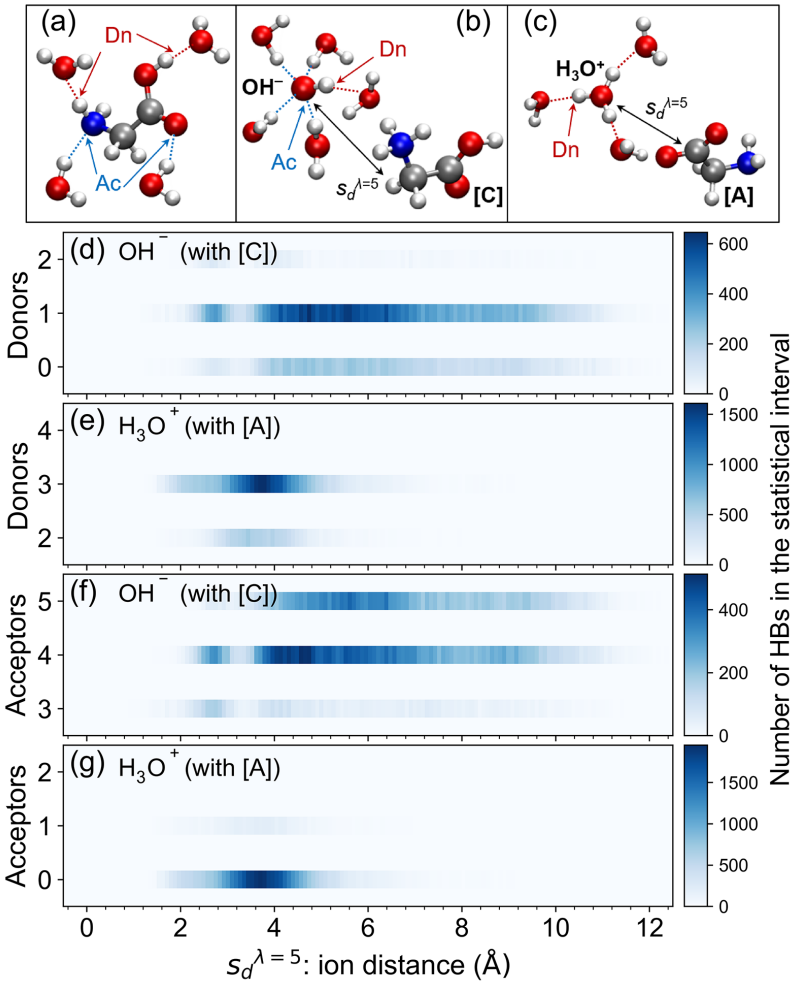

Figure S16: **Hydrogen bond of self-ions**. The schematic representation of (a) donor (Dn) and acceptor (Ac) sites in the neutral [N] form. The presence of (b) hydroxide (OH^-) with [C] form, and (c) hydronium (H_3O^+) with anionic [A] form in water. The number of donor for (d) OH^- and (e) H_3O^+, and the number of acceptor for (f) OH^- and (g) H_3O^+ are displayed as a function of charge-charge distance ($\mathbf{s}_d^{\lambda=5}$).

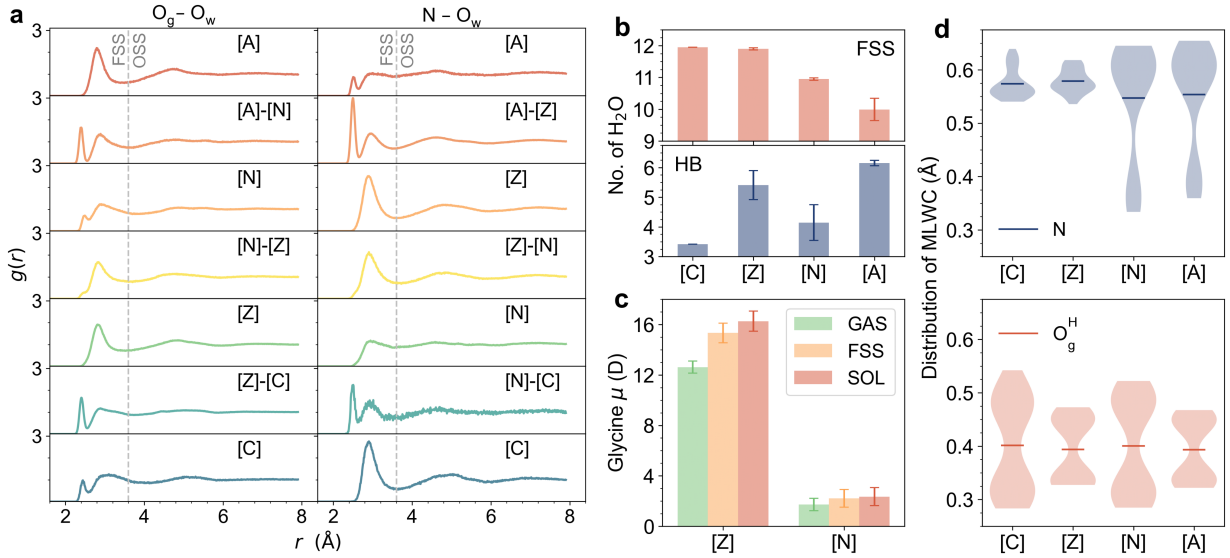

Figure S17: **The solvation and polarization of glycine. a**, RDF profiles with respect to glycine transformations for the pairs of particles comprising O_g-O_w and N-O_w targeted atoms. The region within a specified cutoff distance of 3.6 Å is defined as the FSS of glycine and the external region is described as the OSS. **b**, The number of water molecules and total HB number of glycine in the FSS. **c**, The change of the glycine dipole moment in gas phase (GAS), FSS, and full solvent (SOL), where the molecular structures of glycine in gas phase and FSS are those obtained from the fully solvated glycine by removing corresponding water molecules. **d**, The violin plots represent the distributions of the distances between the nuclei (N and O_g^H) of fully solvated glycine and the MLWCs, where the lines represent the mean value.

References

[1] Chen, Y.; Zhang, L.; Wang, H.; E, W. DeePKS: A comprehensive data-driven approach toward chemically accurate density functional theory. *Journal of Chemical Theory and Computation* **2020**, *17*, 170–181.

[2] Li, W.; Ou, Q.; Chen, Y.; Cao, Y.; Liu, R.; Zhang, C.; Zheng, D.; Cai, C.; Wu, X.; Wang, H.; others DeePKS+ABACUS as a Bridge between Expensive Quantum Mechanical Models and Machine Learning Potentials. *The Journal of Physical Chemistry A* **2022**, *126*, 9154–9164.

[3] Zhang, L.; Han, J.; Wang, H.; Saidi, W.; Car, R.; others End-to-end symmetry preserving inter-atomic potential energy model for finite and extended systems. *Advances in neural information processing systems* **2018**, *31*.

[4] Zhang, L.; Han, J.; Wang, H.; Car, R.; Weinan, E. Deep potential molecular dynamics: a scalable model with the accuracy of quantum mechanics. *Physical review letters* **2018**, *120*, 143001.

[5] He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016; pp 770–778.

[6] Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. PLUMED 2: New feathers for an old bird. *Computer physics communications* **2014**, *185*, 604–613.

[7] Marzari, N.; Vanderbilt, D. Maximally localized generalized Wannier functions for composite energy bands. *Physical review B* **1997**, *56*, 12847.

[8] Marzari, N.; Mostofi, A. A.; Yates, J. R.; Souza, I.; Vanderbilt, D. Maximally localized Wannier functions: Theory and applications. *Reviews of Modern Physics* **2012**, *84*, 1419.

[9] Silvestrelli, P. L.; Parrinello, M. Water molecule dipole in the gas and in the liquid phase. *Physical Review Letters* **1999**, *82*, 3308.

[10] Sharma, M.; Resta, R.; Car, R. Intermolecular dynamical charge fluctuations in water: A signature of the H-bond network. *Physical review letters* **2005**, *95*, 187401.

[11] Kühne, T. D.; Iannuzzi, M.; Del Ben, M.; Rybkin, V. V.; Seewald, P.; Stein, F.; Laino, T.; Khaliullin, R. Z.; Schütt, O.; Schiffmann, F.; others CP2K: An electronic structure and molecular dynamics software package-Quickstep: Efficient and accurate electronic structure calculations. *The Journal of Chemical Physics* **2020**, *152*, 194103.

[12] Michaud-Agrawal, N.; Denning, E. J.; Woolf, T. B.; Beckstein, O. MDAnalysis: a toolkit for the analysis of molecular dynamics simulations. *Journal of computational chemistry* **2011**, *32*, 2319–2327.

[13] Gowers, R. J.; Linke, M.; Barnoud, J.; Reddy, T. J.; Melo, M. N.; Seyler, S. L.; Domanski, J.; Dotson, D. L.; Buchoux, S.; Kenney, I. M.; others MDAnalysis: a Python package for the rapid analysis of molecular dynamics simulations. Proceedings of the 15th python in science conference. 2016; p 105.

[14] Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. *The Journal of Physical Chemistry B* **2009**, *113*, 6378–6396.

[15] Zhao, Y.; Truhlar, D. G. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theoretical chemistry accounts* **2008**, *120*, 215–241.

[16] Weigend, F.; Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Physical Chemistry Chemical Physics* **2005**, *7*, 3297–3305.

[17] Weigend, F. Accurate Coulomb-fitting basis sets for H to Rn. *Physical chemistry chemical physics* **2006**, *8*, 1057–1065.

[18] Frisch, M. J. et al. Gaussian 16 Revision B.01. 2016; Gaussian Inc. Wallingford CT.

[19] Humphrey, W.; Dalke, A.; Schulten, K. VMD: visual molecular dynamics. *Journal of molecular graphics* **1996**, *14*, 33–38.

[20] Hunter, J. D. Matplotlib: A 2D graphics environment. *Computing in science & engineering* **2007**, *9*, 90–95.

[21] Peŕez de Tudela, R.; Marx, D. Water-induced zwitterionization of Glycine: stabilization mechanism and spectral signatures. *The journal of physical chemistry letters* **2016**, *7*, 5137–5142.

[22] Ramaekers, R.; Pajak, J.; Lambie, B.; Maes, G. Neutral and zwitterionic glycine. H_2O complexes: A theoretical and matrix-isolation Fourier transform infrared study. *The Journal of chemical physics* **2004**, *120*, 4182–4193.

[23] Aikens, C. M.; Gordon, M. S. Incremental solvation of nonionized and zwitterionic glycine. *Journal of the American Chemical Society* **2006**, *128*, 12835–12850.

[24] Bachrach, S. M. Microsolvation of glycine: a DFT study. *The Journal of Physical Chemistry A* **2008**, *112*, 3722–3730.

[25] Jensen, J. H.; Gordon, M. S. On the number of water molecules necessary to stabilize the glycine zwitterion. *Journal of the American Chemical Society* **1995**, *117*, 8159–8170.

[26] Tripathi, R.; Durań Caballero, L.; Peŕez de Tudela, R.; Hoölzl, C.; Marx, D. Unveiling Zwitterionization of Glycine in the Microhydration Limit. *ACS omega* **2021**, *6*, 12676–12683.

[27] Kaufmann, M.; Leicht, D.; Schwan, R.; Mani, D.; Schwaab, G.; Havenith, M. Helium droplet infrared spectroscopy of glycine and glycine–water aggregates. *Physical Chemistry Chemical Physics* **2016**, *18*, 28082–28090.

[28] Kim, J.-Y.; Ahn, D.-S.; Park, S.-W.; Lee, S. Gas phase hydration of amino acids and dipeptides: effects on the relative stability of zwitterion vs. canonical conformers. *RSC Advances* **2014**, *4*, 16352–16361.

[29] Azizi, K.; Laio, A.; Hassanali, A. Solvation thermodynamics from cavity shapes of amino acids. *PNAS Nexus* **2023**, pgad239.

[30] Wood, G. P.; Gordon, M. S.; Radom, L.; Smith, D. M. Nature of Glycine and Its α-Carbon Radical in Aqueous Solution: A Theoretical Investigation. *Journal of Chemical Theory and Computation* **2008**, *4*, 1788–1794.

[31] Sun, J.; Bousquet, D.; Forbert, H.; Marx, D. Glycine in aqueous solution: solvation shells, interfacial water, and vibrational spectroscopy from *ab initio* molecular dynamics. *The Journal of chemical physics* **2010**, *133*, 09B609.

[32] Sun, J.; Niehues, G.; Forbert, H.; Decka, D.; Schwaab, G.; Marx, D.; Havenith, M. Understanding THz spectra of aqueous solutions: Glycine in light and heavy water. *Journal of the American Chemical Society* **2014**, *136*, 5031–5038.