This document is confidential and is proprietary to the American Chemical Society and its authors. Do not copy or disclose without written permission. If you have received this item in error, notify the sender and delete all copies.

Site-specific charge regulation model for amino acids in aqueous solutions

Journal:	*Journal of Chemical & Engineering Data*
Manuscript ID	je-2023-00391k.R1
Manuscript Type:	Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Gallegos, Alejandro; University of California Riverside, Department of Chemical and Environmental Engineering Wu, Jianzhong; University of California Riverside, Department of Chemical and Environmental Engineering

SCHOLARONE™
Manuscripts

Site-specific charge regulation model for amino acids in aqueous solutions

Alejandro Gallegos and Jianzhong Wu*

Department of Chemical and Environmental Engineering, University of California, Riverside, CA 92521, USA

Abstract

The aim of this work is to develop a predictive model for describing the electrostatic behavior of amino acids in aqueous solutions. A three-bead coarse-grained model is proposed that distinguishes the side chain of each amino acid from its terminal amine and carboxyl groups in order to quantify the correlated ionization of multiple sites due to the intramolecular Coulomb interaction. To capture how this interaction varies with the solution condition (e.g., pH, ion types and salt concentrations), we have developed a thermodynamic model that accounts for the nonideality resulting from the interaction of amino acids with various solvated ions in an aqueous solution, such as the molecular volume exclusion effects, solvent-mediated electrostatic interactions, van der Waals attraction, and short-ranged hydrophobic and hydration forces. With a small set of parameters characterizing the intermolecular interactions and dissociation equilibrium, the thermodynamic model is able to correlate experimental data for the activity coefficients and solubility of amino acids in pure water as well as aqueous sodium chloride solutions. Moreover, it predicts apparent equilibrium constants for the charge regulation of all natural amino acids in excellent agreement with experimental data. Importantly, the thermodynamic model allows for the extrapolation of the intrinsic equilibrium constants for the ionizable functional group in *isolation* from the influence of other charged functional groups of

* To whom all correspondence should be addressed. Email:jwu@engr.ucr.edu

amino acids. As a result, the model enables to predict the charge behavior for all natural amino acids under arbitrary solution conditions.

1. Introduction

Understanding the charge behavior of amino acids is of vital importance not only because they are the building blocks of polypeptides and proteins but the charge regulation of these biomolecules plays a pivotal role in a broad range of technological applications such as bioadhesives[1, 2], smart materials[3-5], and anti-fouling coatings[6, 7]. Due to the complexity of interactions presenting within a protein as well as the multi-scale structure correlations, it can be difficult to isolate the influence of any one particular amino acid. Thus, studying the behavior of amino acids as building blocks can serve as an initial step toward comprehending the solution properties of proteins and polypeptides.

Each amino acid molecule is composed of a basic amino group ($-NH_2$), an acidic carboxyl group (-COOH), and an organic *X* group (also referred to as the side chain). Natural amino acids are distinguished by their side-chain characteristics and can be divided into three categories: those with electrically charged side chains, either positive or negative depending on the ionizable functional group, those with polar uncharged side chains, and those with hydrophobic side chains. In the latter category, glycine, cysteine and proline are special because the glycine side chain consists of a single hydrogen atom, cysteine can form disulfide bonds, and proline has a side chain that forms a ring with the nitrogen atom in the backbone. If the side chains are not ionizable, amino acids predominately exist as zwitterions because the amino and carboxyl groups are usually charged in the pH range from 3 to 9. For those with ionizable side chains, such as lysine or aspartic acid, the side chain ionization leads to a unit charge for the amino acid; however, the zwitterionic nature of the backbone persists.

In an aqueous solution, the electrostatic charge of amino acids depends on the solution pH and other thermodynamic conditions such as the salt types, ionic strength, and the solution composition. A comprehensive description of their solution behavior thus necessitates understanding the charge regulation of all ionizable groups (viz., the local environment dictates the charge status of each ionizable group). The speciation of amino acids, i.e., the fractions of functional groups according to their ionization states, is intrinsically associated with the protonation/deprotonation of the terminal amino and carboxyl groups as well as the ionizable groups in the side chains. A simple description of charge regulation is through the Henderson-Hasselbalch equation,[8] which predicts that the ionization of the functional groups depends only in the difference in the equilibrium constant and the solution pH. However, as discussed above, the ionization of each functional group also depends on its interactions with other ionizable groups in the same molecule as well as with other solvated species in the solution. The Henderson-Hasselbalch equation is problematic to describe charge regulation of amino acids as it assumes independent ionization and is applicable only to ideal solutions.

An *ad hoc* fix to the Henderson-Hasselbalch equation is to adjust the equilibrium constant according to the solution conditions[9]. This modification lumps the true equilibrium constant with the thermodynamic non-ideality resulting from the inter- and intra-molecular interactions. The so-called *apparent* equilibrium constant can be measured experimentally and, as expected, its value is found to change with the salt concentration as well as the salt type (i.e., NaCl versus KCl) [10]. Unfortunately, accurate measurements of the *apparent* equilibrium constant in arbitrary conditions can be difficult from the experimental perspective. While the apparent equilibrium constant for an isolated group is equivalent to its thermodynamic equilibrium constant in the limit of infinite dilution (viz. in an ideal solution), experimental

measurements at such conditions are not practical for molecules with multiple ionizable sites such as amino acids. Indeed, the equilibrium constants reported in the literature are usually obtained at salt concentrations at or above 10 mM[10]. In addition, the choice of experimental procedure (e.g., titration or spectroscopy) can also factor in to the difference in the equilibrium constants[11]. Theoretically, one should be able to determine the equilibrium constant based on the change in the Gibbs free energy through quantum-mechanical calculations. However, current first-principles approaches provide only estimates and are highly sensitive to the calculation setup such as system size and specific basis sets[12, 13]. Inclusion of explicit water molecules and salt ions only further complicates the theoretical calculation. Thus, the development of a robust thermodynamic model that incorporates key physics will allow for accurate description of amino acids across all solution conditions.

A conventional method to represent the activity coefficients of amino acids in their different charged states is given by the specific ion interaction theory (SIT) [14, 15]. In this approach, the activity coefficients of individual ions are given by a modified Debye-Hückel theory supplemented with an additional term to account for non-electrostatic interactions. The latter is assumed empirically to be proportional to the ionic strength, viz., $\varepsilon(I)I$ where I is the ionic strength, and $\varepsilon(I)$ is the proportionality coefficient. [16, 17] In general, $\varepsilon(I)$ is not universal, which depends on the identity of the ionic species and can be chosen to be a constant or a function of ionic strength to improve the correlation with the experimental data. Despite its simplicity, SIT is able to describe the apparent equilibrium constants because it accounts for electrostatic interactions and solvent-mediated interactions with a large number of empirical paramaters[11, 18, 19]. Pitzer's equation has also been used as an alternative approach to the Debye-Hückel theory[20]. Although the computation becomes more sophisticated, Pitzer's equation agrees

well with the experimental data for the activity coefficients and apparent equilibrium constants of several amino acid solutions. Like SIT, the practical applicability of Pitzer's equation is compromised by the large number of empirical parameters[21, 22].

In our previous work, we developed an augmented primitive model (APM) that captures the thermodynamic properties of natural amino acids in aqueous sodium chloride solutions using a 1-bead model[23]. Unlike the SIT and Pitzer's equations, the parameters employed in the APM has clear physical significance as they are related to the molecular excluded volumes and valences of the amino acids, as well as the solvent-mediated short-range interactions. Accordingly, the activity coefficients can be decomposed into a repulsive contribution due to volume exclusion, an attractive contribution due to electrostatic interactions, and an additional contribution due to water-mediated effects. However, a fundamental problem with the APM (and other one-bead models) is that it treats the charge of each amino acid molecule as a sum over the charges of its functional groups, i.e., an amino acid molecule with a positively charged amino group and negatively charged carboxyl group is considered neutral. Thus, there is no contribution to the activity coefficient from electrostatic interactions within the same molecule when the amino acid is in its zwitterionic form. Alternative models have been proposed to account for the dipole nature of the zwitterion such as the perturbation theory by Khoshkbarchi and Vera[24-26]. A major disadvantage of the one-bead model is that it neglects the intramolecular correlations that could influence the ionization behavior of different functional groups within the same amino acid molecule. For example, the presence of the positively charged amino group will favor the ionization of the carboxyl group in the vicinity. Such effect is reduced by an increase of the salt concentration due to the electrostatic screening effects.

The purpose of this work is to develop a coarse-grained model that distinguishes the side chain charge from those on the amino acid backbone, and thus allows for the predictions of the activity coefficient and apparent equilibrium constant. Importantly, this model can be utilized for the extrapolation of the intrinsic equilibrium constants for the ionizable functional group in *isolation* from the influence of other charged functional groups in the amino acid molecule. Such a thermodynamic model requires a reasonable description of both the inter- and intra-molecular interactions between the amino acid and all solvated species in the aqueous medium. While a precise description of hydration and water-mediated forces in microscopic details is beyond the scope of the present work, their contributions to thermodynamic properties can be reasonably described with semi-empirical models such that the parameters can be fixed by comparison with experimental data.

2. Thermodynamic Model and Methods

2.1 *A coarse-grained model for the aqueous solutions of amino acids*

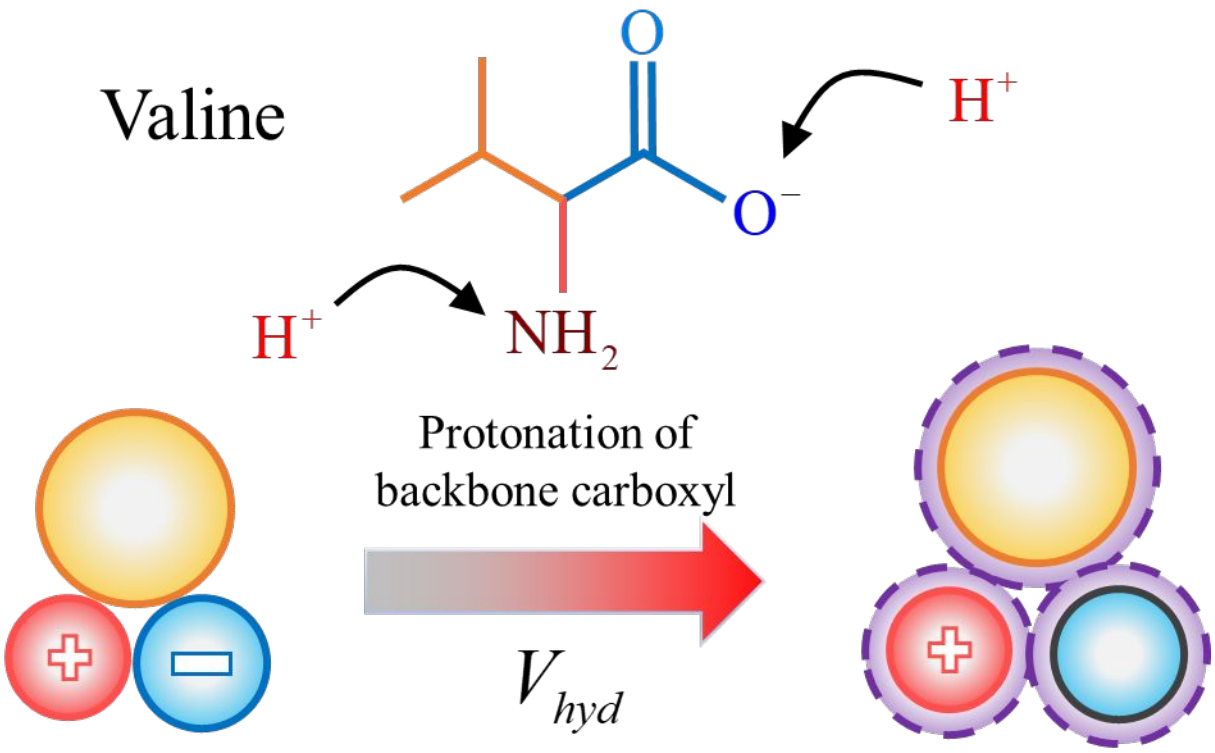

Figure 1. A coarse-grained representation of an amino acid (e.g., Valine) that accounts for the hydration change resulting from the charge regulation of the ionizable groups.

We propose a 3-bead augmented primitive model (3bAPM) to represent amino acid molecules dissolved in an aqueous solution. While the 3-bead model is not able to capture the molecular structure with atomistic details, it enables to distinguish the ionization of a side chain from that for the amino and carboxyl groups, as illustrated in Figure 1. We assume that each segment is tangentially connected to the other two segments (i.e., the side chain segment is connected to the amine and carboxyl segments while the amine and carboxyl segments are also in contact with each other). In the case of proline, we employ only a two-bead model due to the ring connecting the side chain and the amine group, respectively. Ionizable segments can exist in either uncharged or charged state, and the segment diameter depends on the ionization state of the amino acid as a whole. Other solvated ionic species (salt ions, proton, and hydroxyl ions) are considered within the framework of the augmented primitive model and are represented by charged hard spheres. Water is treated implicitly through a dielectric continuum. For simplicity, we lump all short-range interactions between amino acids and ionic species in an aqueous environment in terms of a square-well potential. While the simple model lacks atomic details and does not reflect the anisotropic nature of amino acid molecules, it provides a flexible framework to represent the thermodynamic non-ideality. Despite extensive research, a first-principles approach to predicting water-mediated interactions remains theoretically challenging. By adopting an effective potential, we are able to capture the essential physics without having to consider the microscopic details[27, 28].

Myriad of interactions can occur between amino acid and its surrounding environment. First of all, there exist non-electrostatic interactions between the charged forms of the amine and

carboxylate segments (ε_{N^+,O^-}). In addition, there are interactions between the charged amine/carboxyl segment and ionic species in the solution (e.g., ε_{N^+,Cl^-} and ε_{O^-,Na^+} in the square-well potential). Initial parameter fitting indicated that these two are close to each other; thus, we set $\varepsilon_{N^+,Cl^-} = \varepsilon_{O^-,Na^+}$, but it would likely be beneficial in future work to quantify the appropriate parameters for a wide range of salt species. Parameters ε_{N^+,O^-} and $\varepsilon_{N^+,Cl^-} / \varepsilon_{O^-,Na^+}$ are assumed the same for all amino acids (besides proline) since they all share a common backbone. Note that the backbone-backbone and backbone-salt interactions are only when the amino acid is in its zwitterion form. Besides, additional interactions must be considered between the side chains, the backbone amines, and the backbone carboxyl groups. To minimize the number of unknown parameters, we designate a general interaction parameter between the side chain and the amino acid $\varepsilon_{X,X}$. Lastly, for neutral amino acids, we consider non-electrostatic interactions between the side chain and the salt ions $\varepsilon_{X,Na^+/Cl^-}$. For amino acids with charged side chains, we treat them in a similar fashion as the neutral amino acids, except for arginine and lysine. For histidine, aspartic acid, and glutamic acid, the amino acid takes the neutral form when the backbone amino and carboxyl groups are charged. Therefore, we employ the same parameters as above. For the latter two amino acids, the neutral form is assigned when both the side chain and backbone carboxyl group are charged. Initial fittings indicated that the interaction between the charged side chain and charged carboxyl group/chloride ion was similar in value to between the charged amino group and charged carboxyl/chloride ion; therefore, we set $\varepsilon_{X^+,O^-} = \varepsilon_{N^+,O^-}$ and $\Delta\varepsilon_{X^+,Cl^-} = \varepsilon_{X^+,Cl^-} - \varepsilon_{X^0,Cl^-}$.

In an aqueous solution, amino acids can exist in multiple charge states in chemical equilibrium with one another. With the termini and side chains represented by three tangentially

connected hard spheres, we can assign spin-like state values to each segment and specify the charge state of the entire molecule by the combination of these spin values. The charge state of each amino segment is given by $s_N = [0, +1]$ where '0' denotes in its uncharged form, -NH_2, and '+1' is for the charged form, -NH_3^+. Similarly, the charge state of each carboxyl segment is given by $s_O = [0, -1]$, where '0' represents the uncharged form, -COOH, and '-1' represents the charged form, -COO^-. The non-ionizable side chains are fixed at state $s_X = 0$ while the ionizable side chains can be either positively charged, $s_X = [0, +1]$, or negatively charged, $s_X = [0, -1]$. For example, lysine contains a positively charged side chain and its charge state is specified by Lys[0,-1,+1] when its amino group in the backbone is uncharged while its carboxyl group (-1) and sidechain (+1) are oppositely charged.

2.2. Ionization reactions and equilibrium constants

An important feature of our 3-bead coarse-grained model is that we distinguish different ionizable groups in amino acids and explicitly account for the influence of both inter- and intramolecular interactions on their chemical equilibrium. Due to the short separation between ionizable groups in the amino acid, the ionization of one group is strongly influenced by the charge states of other groups in the same molecule. Thus, an amino acid must be defined by the charge states of its backbone amino (N) and carboxyl functional (O) groups as well as the charge state of its sidechain (X). For glycine, the protonation reactions are associated with that of the amino group:

$$-NH_2 + H^+ \leftrightarrow -NH_3^+ \tag{2}$$

and that of the carboxyl group:

$$-O^- + H^+ \leftrightarrow -OH\,. \tag{3}$$

In the case of Lysine, Arginine, and Histidine, their sidechain amino groups will be ionized by adding a proton, as shown in Eq. (1) while for aspartic acid and glutamic acid, their sidechain carboxyl group will be ionized by losing a proton as shown in Eq. (2).

The equilibrium constants for the protonation reaction of the ionizable groups *N* and *O* are given by

$$K^T_{X(N)} = \frac{c_X[+1, s_O, s_X]}{c_X[0, s_O, s_X]} \frac{\gamma_X[+1, s_O, s_X]}{\gamma_X[0, s_O, s_X]} \frac{1}{c_{H^+}\gamma_{H^+}}, \quad (4)$$

$$K^T_{X(O)} = \frac{c_X[s_N, 0, s_X]}{c_X[s_N, -1, s_X]} \frac{\gamma_X[s_N, 0, s_X]}{\gamma_X[s_N, -1, s_X]} \frac{1}{c_{H^+}\gamma_{H^+}}, \quad (5)$$

where $c_X[s_N, s_O, s_X]$ and $\gamma_X[s_N, s_O, s_X]$ are the molar concentration and activity coefficient of the amino acid in its charge configuration, c_{H^+} and γ_{H^+} are the concentration and activity coefficient of the proton. The activity coefficient of each species is related to its excess chemical potential, which accounts for the inter- and intramolecular interactions. The equilibrium constant for the protonation reaction in the side chain of Lysine, Arginine, or Histidine is given by

$$K^T_{X(X)} = \frac{c_X[s_N, s_O, +1]}{c_X[s_N, s_O, 0]} \frac{\gamma_X[s_N, s_O, +1]}{\gamma_X[s_N, s_O, 0]} \frac{1}{c_{H^+}\gamma_{H^+}}. \quad (6)$$

Similarly, the equilibrium constant for the sidechain in Aspartic acid or Glutamic acid is

$$K^T_{X(X)} = \frac{c_X[s_N, s_O, 0]}{c_X[s_N, s_O, -1]} \frac{\gamma_X[s_N, s_O, 0]}{\gamma_X[s_N, s_O, -1]} \frac{1}{c_{H^+}\gamma_{H^+}}. \quad (7)$$

In Eqs.(3-6), the thermodynamic equilibrium constant, $K^T_{X(i)}$, reflects the functional group *i* in the amino acid *X*; it is independent from the charge status of other ionizable groups in the amino acid. In other words, the physical contributions (i.e., the intramolecular interactions) is separated from the chemical contributions (viz., binding of proton) to the ionizable equilibria. In general, the equilibrium constant is a thermodynamic quantity defined by the change in the chemical

potentials of reactants and products at their reference states, i.e., each species in an ideal solution at unit molar concentration. Our reference states used in the 3bAPM for amino acid monomers can be easily extended to the ionization equilibria of polypeptides since we have isolated the functional groups from one another such that the formation of peptide bond has no effect on the intrinsic equilibrium constant.

Whereas the solution composition can be monitored with various analytical tools, it is not feasible to directly measure the thermodynamic equilibrium constant because the activity coefficients are often unknown. As a result, the thermodynamic equilibrium constant is typically lumped together with the activity coefficients into its apparent value as shown below for the backbone amino group:

$$K'_{X(N)} = \frac{c_X\left[+1, s_O, s_X\right]}{c_X\left[0, s_O, s_X\right]}\frac{1}{c_{H^+}} = K^T_{X(N)}\frac{\gamma_X\left[0, s_O, s_X\right]\exp\left\{\beta u^{el}_X\left[0, s_O, s_X\right]\right\}}{\gamma_X\left[+1, s_O, s_X\right]\exp\left\{\beta u^{el}_X\left[+1, s_O, s_X\right]\right\}}\gamma_{H^+} \tag{8}$$

where $\beta = 1/(k_B T)$, k_B and T are the Boltzmann constant and the absolute temperature, respectively. Note that even in the infinite dilution limit, there is still a pair interaction contribution u^{el}_X which results in a difference between the thermodynamic equilibrium constant and the apparent equilibrium constant measured in experiments.

If the ionizable groups are isolated (i.e., no intramolecular interactions) in an ideal solution, the activity coefficients are negligible and there is no difference between apparent and true equilibrium constants. In general, the *apparent* equilibrium constant is influenced by both the inter- and intramolecular correlations which vary with the solute composition. Unfortunately, it is difficult to distinguish the concentration of individual charge configurations of the amino acid such as [+1,0,0] and [+1,-1,0]. Thus, the *apparent* equilibrium constant measured in experiments is related to the average contribution of each configuration

$$K'_{X(N)} = \frac{\left\langle c_X\left[+1, s_O, s_X\right]\right\rangle}{\left\langle c_X\left[0, s_O, s_X\right]\right\rangle}\frac{1}{c_{H^+}} \tag{9}$$

where $\langle \mathsf{L} \rangle$ denotes the thermal (or Boltzmann) average (i.e., the expected value) of the quantity of interest (in this case, the concentration of the amino acid with its backbone amine group either charged or uncharged). The thermal average of the concentration is given by[29]

$$\left\langle c_X\left[s_N, s_O, s_X\right]\right\rangle = c_X^{tot}\frac{\exp\left\{-\beta\sum_i \mu_{X(i)}^H\left[s_i\right] - \beta u_X^{el}\left[s_N, s_O, s_X\right]\right\} / \gamma_X\left[s_N, s_O, s_X\right]}{\Xi} \tag{10}$$

where $\mu_{X(i)}^H(s_i) = s_i[pH - \log K_{X(i)}^T]\ln 10$ is the chemical binding energy, and $pH = -\log[c_{H^+}\gamma_{H^+}]$ is the negative logarithm of the proton activity. The electrostatic pair interaction u_X^{el} is given by the direct coulomb interaction between the segments in their respective states. The proportionality factor Ξ satisfies the condition $\sum_{s'_N}\sum_{s'_O}\sum_{s'_X}\langle c_X[s'_N, s'_O, s'_X]\rangle = c_X^{tot}$ and c_X^{tot} is the total concentration of the amino acid *X*. We can define the probability α of finding the amino acid *X* in configuration $[s_N, s_O, s_X]$ to be $\alpha[s_N, s_O, s_X] = \langle c_X[s_N, s_O, s_X]\rangle / c_X^{tot}$. Thus, by plugging Eq.(9) into Eq.(8), we arrive at the following expression for the apparent equilibrium constant:

$$K'_{X(N)} = K_{X(N)}^T\gamma_{H^+}\frac{\sum_{s'_O}\sum_{s'_X}\exp\left\{-\beta\mu_{X(O)}^H\left[s'_O\right] - \beta\mu_{X(X)}^H\left[s'_X\right] - \beta u_X^{el}\left[+1, s'_O, s'_X\right]\right\} / \gamma_X\left[+1, s'_O, s'_X\right]}{\sum_{s'_O}\sum_{s'_X}\exp\left\{-\beta\mu_{X(O)}^H\left[s'_O\right] - \beta\mu_{X(X)}^H\left[s'_X\right] - \beta u_X^{el}\left[0, s'_O, s'_X\right]\right\} / \gamma_X\left[0, s'_O, s'_X\right]} \tag{11}$$

Clearly, the experimentally measured value depends on the ionization equilibria of the amino acid by which charge configurations dominate. As a result, the proton concentration cannot be easily separated out, like that in Eq.(7), since the reported apparent equilibrium constants depend on c_{H^+} through the chemical ionization energy $\mu_{X(i)}^H$.

Due to the strong non-ideality of amino acids in aqueous solutions, the equilibrium constants for the amino group and carboxyl group may differ by almost 7 orders of magnitude[23]. When measuring the equilibrium constant for the amino group, the proton concentration will be on the order of 10^{-9} ($pH \approx 9$), and the chemical ionization energy for the carboxyl group will be approximately equal to -15 k_BT. Therefore, the ionized form of the carboxyl group will be roughly 10^7 more prevalent than the non-ionized form, meaning that we can eliminate the charge configurations that are extremely unlikely to contribute to the apparent equilibrium constant (viz., [+1,0,0] and [0,0,0]). For amino acids without an ionizable sidechain, Eq. (10) reduces to

$$K'_{X(N)} = \frac{\left\langle c_X\left[+1, s_O, s_X\right]\right\rangle}{\left\langle c_X\left[0, s_O, s_X\right]\right\rangle}\frac{1}{c_{H^+}} = K^T_{X(N)}\frac{\gamma_X\left[0,-1,0\right]\exp\left\{\beta u^{el}_X\left[0,-1,0\right]\right\}}{\gamma_X\left[+1,-1,0\right]\exp\left\{\beta u^{el}_X\left[+1,-1,0\right]\right\}}\gamma_{H^+}. \tag{12}$$

In the case of the carboxyl group, we can write the apparent equilibrium constant as

$$K'_{X(O)} = \frac{\left\langle c_X\left[s_N, 0, s_X\right]\right\rangle}{\left\langle c_X\left[s_N, -1, s_X\right]\right\rangle}\frac{1}{c_{H^+}} = K^T_{X(N)}\frac{\gamma_X\left[+1,-1,0\right]\exp\left\{\beta u^{el}_X\left[+1,-1,0\right]\right\}}{\gamma_X\left[+1,0,0\right]\exp\left\{\beta u^{el}_X\left[+1,0,0\right]\right\}}\gamma_{H^+}. \tag{13}$$

For fitting the experimental results, it is convenient to express the equilibrium constants in the logarithmic form:

$$\begin{aligned}\log K'_{X(N)} = \log K^T_{X(N)} + \log\gamma_X[0,-1,0] + \log\gamma_{H^+} \\ -\log\gamma_X[+1,-1,0] + \frac{\beta u^{el}_X\left[0,-1,0\right] - \beta u^{el}_X\left[+1,-1,0\right]}{\ln 10}\end{aligned} \tag{14}$$

and

$$\begin{aligned}\log K'_{X(O)} = \log K^T_{X(O)} + \log\gamma_X[+1,-1,0] + \log\gamma_{H^+} \\ -\log\gamma_X[+1,0,0] + \frac{\beta u^{el}_X\left[+1,-1,0\right] - \beta u^{el}_X\left[+1,0,0\right]}{\ln 10}.\end{aligned} \tag{15}$$

Amino acids with ionizable side chains like lysine or aspartic acid contain ionizable groups that have equilibrium constants close to those of the terminal N/O groups of the same charge. For those cases, the amino acids take possible configurations in which the side chain

charge is strongly correlated with that of a terminal group. As a result, Eq.(8) can no longer be simplified, and the apparent equilibrium constant remains a function of the proton concentration. One method to overcome this difficulty is to determine the apparent equilibrium constant at the pH in which the ionizable group of interest is 50% charged. Alternatively, we can assume sequential ionization, i.e., the first group ionizes before the second group. For example, the first protonation step of lysine is the ionization of the amine group in the side chain where the backbone carboxyl group is negatively charged. In the second protonation step, the backbone amine group ionizes in the presence of both the charged side chain amine group and the charged backbone carboxyl group. In the last protonation step, the carboxyl group de-ionizes while the two other amine groups are charged. We find that these two assumptions are equivalent within the precision of the correlated model parameters and we have chosen to use the second assumption due to its simplicity. Nonetheless, Eq.(10) is important from a fundamental standpoint because it sets up the appropriate framework to describe organic molecules with same charged ionizable groups that share a similar equilibrium constant.

2.3. Activity coefficients

As discussed in our recent work for the ionization of weak polyelectrolytes[29], we can specify the activity coefficient γ_i of the amino acid in different charge states as

$$k_B T \ln \gamma_X \left[s_N, s_O, s_X \right] = \mu_N^{ex}\left(s_N \right) + \mu_O^{ex}\left(s_O \right) + \mu_X^{ex}\left(s_X \right) - k_B T \ln y_X \left[s_N, s_O, s_X \right]. \qquad (16)$$

The first three terms on the right side of Eq. (15) are the excess chemical potential of each segment, μ^{ex}, i.e., the deviation from the chemical potential of species *i* in an ideal solution resulting from the intra- and inter-molecular interactions. According to 3bAPM, the excess chemical potential can be decomposed into the contributions due to the hard-sphere repulsion (hs), electrostatic correlation (el), and solvent-mediated interactions(sw), respectively,

$$\mu_i^{ex} = \mu_i^{hs} + \mu_i^{el} + \mu_i^{sw}. \tag{17}$$

The detailed expression for each contribution can be found in Supporting Information. Note that the activity coefficient of the electrolytes, protons, and hydroxyl ions are given by the APM model, $\gamma_i = \exp(\beta\mu_i^{ex})$. In Eq.(15), $-k_B T \ln y_X[s_N.s_O, s_X]$ accounts for the work required to connect the three beads of the amino acid into its given configuration $[s_N, s_O, s_X]$. It is this term that distinguishes our amino acid model from previous methods by accounting for the intrachain interactions and correlation effects.

We can decompose the cavity correlation function into two contributions to the intrachain correlation effects

$$\ln y_X\left[s_N, s_O, s_X\right] = \ln y_i^{hs}\left[s_N, s_O, s_X\right] + \ln y_i^{el}\left[s_N, s_O, s_X\right] \tag{18}$$

where the first term results from the connection of hard spheres, and the second results from the electrostatic interactions between the charges sites. The hard-sphere contribution accounts for the fact that it is favorable to have two hard spheres near one another which will reduce the total excluded volume of the two spheres. Within our three-bead model, the work is given by the contact value of the radial distribution function [30]

$$\begin{aligned}\ln y_X^{HS}\left[s_N, s_O, s_X\right] = &\ln\left[\frac{1}{1-n_3} + \frac{n_2}{4\left(1-n_3\right)^2}\tau_{N,X}\left(s_N, s_X\right) + \frac{n_2^2}{72\left(1-n_3\right)^3}\tau_{N,X}\left(s_N, s_X\right)^2\right] \\ &+ \ln\left[\frac{1}{1-n_3} + \frac{n_2}{4\left(1-n_3\right)^2}\tau_{N,O}\left(s_N, s_O\right) + \frac{n_2^2}{72\left(1-n_3\right)^3}\tau_{N,O}\left(s_N, s_O\right)^2\right] \\ &+ \ln\left[\frac{1}{1-n_3} + \frac{n_2}{4\left(1-n_3\right)^2}\tau_{O,X}\left(s_O, s_X\right) + \frac{n_2^2}{72\left(1-n_3\right)^3}\tau_{O,X}\left(s_O, s_X\right)^2\right]\end{aligned} \tag{19}$$

where the first term on the right describes the pair interaction between the amino terminus and the side chain, and the second term describes the pair interaction between the backbone amino

and carboxyl group in their respective charge states. Here, $n_2 = \sum_i \rho_i \sigma_i^2 \pi$ and $n_3 = \sum_i \rho_i \sigma_i^3 \pi / 6$ are the weighted densities related to the particle number density ρ_i and the particle diameter σ_i of the species in the system, and $\tau_{i,j}(s_i, s_j) = 2\sigma_i(s_i)\sigma_j(s_j)/[\sigma_i(s_i) + \sigma_j(s_j)]$ is the effective radius of the contact area between the hard spheres *i* and *j* in their states s_i and s_j. Because the intra-chain correlation is related to the charge configuration through the volume change of a segment upon ionization, it will influence the ionization behavior of the amino acid. The electrostatic contribution accounts for the screened interaction between charged functional groups and can be approximated by the radial distribution function of charged hard spheres from the mean spherical approximation (MSA)[31]:

$$\begin{aligned} \ln y_X^{el}\left[s_N, s_O, s_X\right] = &-\frac{l_B}{\sigma_{N,O}\left(s_N, s_O\right)}\left[Z_N^{eff}\left(s_N\right)Z_O^{eff}\left(s_O\right) - Z_N\left(s_N\right)Z_O\left(s_O\right)\right] \\ &-\frac{l_B}{\sigma_{N,X}\left(s_N, s_X\right)}\left[Z_N^{eff}\left(s_N\right)Z_X^{eff}\left(s_X\right) - Z_N\left(s_N\right)Z_X\left(s_X\right)\right] \\ &-\frac{l_B}{\sigma_{O,X}\left(s_O, s_X\right)}\left[Z_O^{eff}\left(s_O\right)Z_X^{eff}\left(s_X\right) - Z_O\left(s_O\right)Z_X\left(s_X\right)\right] \end{aligned} \tag{20}$$

where the effective valence is given by

$$Z_i^{eff}(s_i) = [s_i - \eta\sigma(s_i)^2]/[1 + \Gamma\sigma(s_i)]. \tag{21}$$

In the above equations, l_B is the Bjerrum length, η is related to the asymmetry in the size and valence of species [see Eq.(S13)] and is typically negligible for the systems of interest, κ and Γ are the Debye and MSA screening parameter, respectively, which are functions of the ionic strength[27]. In combination with the direct Coulomb energy u^{el}, Eq.(19) accounts for the screened electrostatic pair interaction between the trimer segments (N-O, N-X, and O-X) . By including Eq.(17) when determining the activity coefficients, we take into account the influence of *intramolecular* correlations on the activity coefficient, and therefore the correlated ionization

of individual functional groups in each amino acid molecule. Lastly, the activity coefficient of the amino acid is given by a weighted average of all possible configurations:

$$\gamma_X = \sum_{s'_N}\sum_{s'_O}\sum_{s'_X}\left(\gamma_X\left[s'_N, s'_O, s'_X\right]\right)^{\alpha_X\left[s'_N, s'_O, s'_X\right]} . \quad (22)$$

2.3 ***Evaluation of the model parameters***

We calibrate the activity coefficient model for amino acids in aqueous solutions of sodium chloride using experimental data for the activity coefficient, solubility, and apparent equilibrium constant. For amino acids dissolved in a sodium chloride solution, the solute species consist of the amino acid in its different charge configurations, sodium and chloride ions, as well as protons and hydroxyl ions. Based off our previous correlations with the APM[23], the hard sphere diameters for sodium ion, chloride ion, proton, and hydroxyl ion were determined to be 3.22 Å, 3.62 Å, 5.00 Å, and 3.22 Å, respectively. These parameters were able to reproduce the experimental data for the mean activity coefficients from low to moderate salt concentrations (<1.0 M). Although not considered in this work, it has been identified that there may be changes in the hydration of the solvated species as well as changes in the dielectric constant of the solution as the amino acid and salt concentration are increased [32, 33]. However, such details are not easily quantified, and they are most relevant at higher salt concentrations; thus, we have chosen to simplify our model by not considering the concentration-dependent hydration of solvated species or the variation of the dielectric constant of the solution.

The hard-sphere diameter σ_i^{hs} of the three amino-acid segments (*N*, *O*, and *X*) can be linearly correlated with their van der Waals diameters σ_i^{vdW}, $\sigma_i^{hs} = 1.71\text{Å} + 0.317\sigma_i^{vdW}$, where the coefficients were determined by fitting the experimental data for the activity coefficient of amino acids in aqueous solutions[34]. Thus, the size of the amine ($-NH_3^+$) and carboxyl group ($-COO^-$)

in their charged state is the same for *all* amino acids and that their diameters are fixed at 2.79 Å and 3.05 Å, respectively. In addition, we determine the short-range interactions $\varepsilon_{i,j}$ by comparison to experimental data for activity coefficients. Note that glycine is solely governed by the interactions mediated through its backbone (viz., $\varepsilon_{X,j}=0$); thus, the side chain interactions for the other amino acids are uniquely determined.

It can be expected that an amino acid molecule may change its hydration diameter as its functional groups protonated/deprotonated, i.e., as the amino acid changes its charge state. We determined the volume change from the neutral state to either of its charged state based off the correlation with the experimental data for the apparent equilibrium constants. The change in the segment volume and the interaction parameters ($\varepsilon_{N^+,j}$ and $\varepsilon_{O^-,j}$) serve as the fitting parameters to describe the ionization equilibria of amino acids where the interaction parameters best agree with activity coefficient data. We assume that the change in the hydration due to the ionization is uniform along the segments with a thickness of Δr. Thus, the change in the hard-sphere diameter of the segments upon the change in volume satisfies the following expression[35]:

$$\sum_{i=\{N,O,X\}} \frac{\pi}{6}\left[\left(\sigma_i+2\Delta r\right)^3-\sigma_i\right]=V_{hyd} \tag{23}$$

where σ_i refers to the hard-sphere diameter of the charged amino segment, charged carboxyl segment, and neutral side chain segment, and Δr is the change in segment radius due to the hydration change of the amino acid given by the hydration volume V_{hyd}. Thus, the change in hydration applies to the entire amino acid molecule but not to any specific functional group (i.e., segment).

The parameters for amino acids were estimated by fitting experimental data using MATLAB 2021a. The data used in this work is available in the supporting information of our

previous publication[23]. To determine the parameters, we minimize the following objective function (OF):

$$\mathrm{OF} = \sum [\xi_i^{exp} - \xi_i^{cal}]^2 \quad (24)$$

where ξ refers to the experimental data for the activity coefficients of amino acids and the apparent equilibrium constants. The uncertainty in experimental data was typically small (within symbol size) or unreported.

3. Results and Discussion

3.1 Activity coefficients for amino acids in aqueous solution

Table 1. The hard-sphere diameter and the interaction energy in units of Å and k_BT, respectively, for neutral amino acids based off correlation to experimental data for the activity coefficient.

	Segment	σ_i^{vdW} (Å)	σ_i^{hs} (Å)	ε_{N^+,O^-}	$\varepsilon_{N^+/O^-,Cl^-/N}$
	$-NH_3^+$	3.396	2.791	0.161	0.150
	$-COO^-$	4.211	3.05	0.161	0.150
				$\varepsilon_{X,X}$	$\varepsilon_{X,Na^+/Cl^-}$
Side Chain	Gly	2.628	2.548	-	-
	Ala	3.713	2.892	-0.072	0.052
	Val	4.895	3.267	-0.424	-0.179
	Leu	5.317	3.401	-0.799	-
	Cys	4.423	3.117	1.199	-
	Met	5.344	3.410	2.490	-0.148
	Trp	6.438	3.757	-0.347	-
	Phe	5.748	3.538	-0.273	-
	Ser	4.081	3.009	0.389	0.230
	Thr	4.657	3.192	0.078	0.053
	Asn	4.891	3.266	0.316	0.292
	Gln	5.314	3.400	-0.000	-

We first investigate the influence of intermolecular interactions on the activity coefficient for amino acids as a function of both the amino acid concentration as well as the salt concentration. In order to compare with experimental data, we employ fitting parameters based off the molecular size of the amino acid as well as its interactions with amino acids and salt ions through either its side chain or its backbone. The parameters for neutral amino acids are listed in Table 1. The interaction energy between amine and carboxyl segments is fixed at 0.161 k_BT while their interactions with salt ions is 0.150 k_BT. The similarity in their value reflects the short-range electrostatic nature of the interaction. By accounting for these contributions to the activity coefficient of each amino acid, we are then able to isolate the contribution due to the side chain.

Table 1 indicates that there is an increase in the repulsion between side chains from Alanine to Valine to Leucine; however, one might expect it to be attractive consider their hydrophobic nature. Because our model does not distinguish the *X*-*X* interaction from *X*-N^+ and *X*-Cl^- interactions, the repulsion must originate from the disruption of hydrogen bonding and local water structure between the amino acids. We expect that the favorable aggregation of the hydrophobic groups in the solution is outweighed by the disruption of the hydrogen bonding by the polar backbone (i.e., $\varepsilon_{X,X} > 0$ while $\varepsilon_{X,N^+/Cl^-} < 0$). On the other hand, Serine and Threonine have an attractive contribution from the side chain to other amino acids which can be attributed to the hydroxyl group in their backbone that can form hydrogen bonds. However, threonine has an additional methyl group in its side chain compared to serine which could limit the accessibility to the hydroxyl site, and therefore it explains the reduced favorable interaction. In the case of phenylalanine and tryptophan, the aromatic interactions between its side chain are attractive, but the inability to hydrogen bond would lead to some repulsion between two amino

acids. As a result, phenylamine shows almost no interaction between the side chain and another phenylalanine while the interaction between tryptophan's side chain is repulsive. For the case of cysteine and methionine, there is an attraction between the side chain and the amino acid, which is likely related to the possible sulfur interactions. Lastly, both asparagine and glutamine are attractive since their side chains can form hydrogen bonds, but the attraction between glutamine's sides chains is diminished due to the additional carbon in its side chain. Meanwhile, the side chain interactions with the salt ions are either repulsive for the hydrophobic cases (Val, and Met) or attractive for the hydrophilic side chains (Ser, Thr, Asn, and Gln). Due to the small side chain of Alanine (viz., a methyl group), it does not disrupt the backbone-salt interaction. The addition of a methyl group in the side chain of Threonine compared to Serine results in a diminished attraction to the salt ions from the side chain.

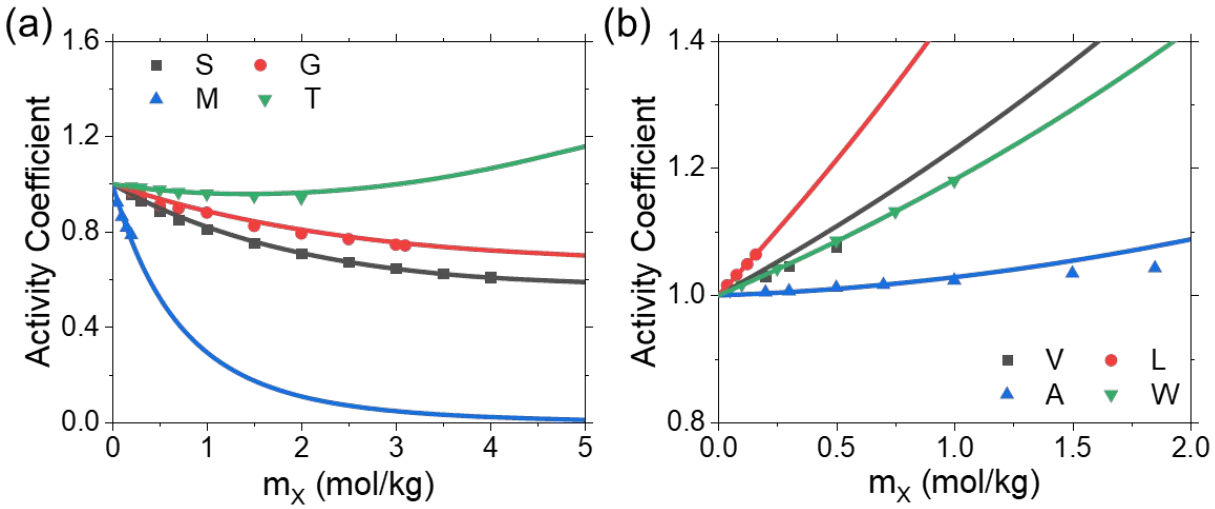

Figure 2. Activity coefficient of different amino acids in aqueous amino acid solution at 25 °C from experiments [36-38] (symbols) and theoretical correlation (lines). (a) Attractive: Serine (S), Glycine (G), Methionine (M), and Threonine (T) and (b) Repulsive: Valine (V), Leucine (L), Alanine (A), Tryptophan (W).

Neutral amino acids

We first consider neutral amino acids in aqueous sodium chloride solutions around their isoelectric points, i.e., the pH at which the molecule carries no net electrical charge. For neutral amino acids, the isoelectric point is typically around pH=6, and thus the proton and hydroxyl ions are dilute relative to the salt concentration. Figure 2 presents a comparison between the theoretical correlations and experimental data for the activity coefficients of eight neutral amino acids as a function of amino acid molality in a salt-free solution. In all cases, our model provides a quantitative description of how the activity coefficient of amino acids is influenced by the itself interactions.

Figure 2a shows four cases in which the amino acid favorably interacts with itself, leading to a decrease in the activity coefficient as the molality of amino acid is increased. On the other hand, Figure 2b shows four cases in which the activity coefficient of the amino acid increases with the rise of the molality of the amino acid. According to our 3bAPM model, the backbone-backbone interaction is the same for all amino acids, and thus the amino acids differ based on their side chain contribution. The differences in the activity coefficient can be understood as a delicate balance between the volume exclusion term (viz., hard-sphere repulsion) and the short-range interactions (viz., square-well potential) from the side chain. In general, amino acids with attractive contributions originating from the side chain favor the addition of the amino acid as indicated by the decrease in the activity coefficient. At higher amino-acid molalities, the stronger repulsion originating from the hard-sphere exclusion, thereby increasing the activity coefficient. In the case of alanine, the activity coefficient is nearly unity even up to 2.0 molality due to the weak repulsion between its side chains.

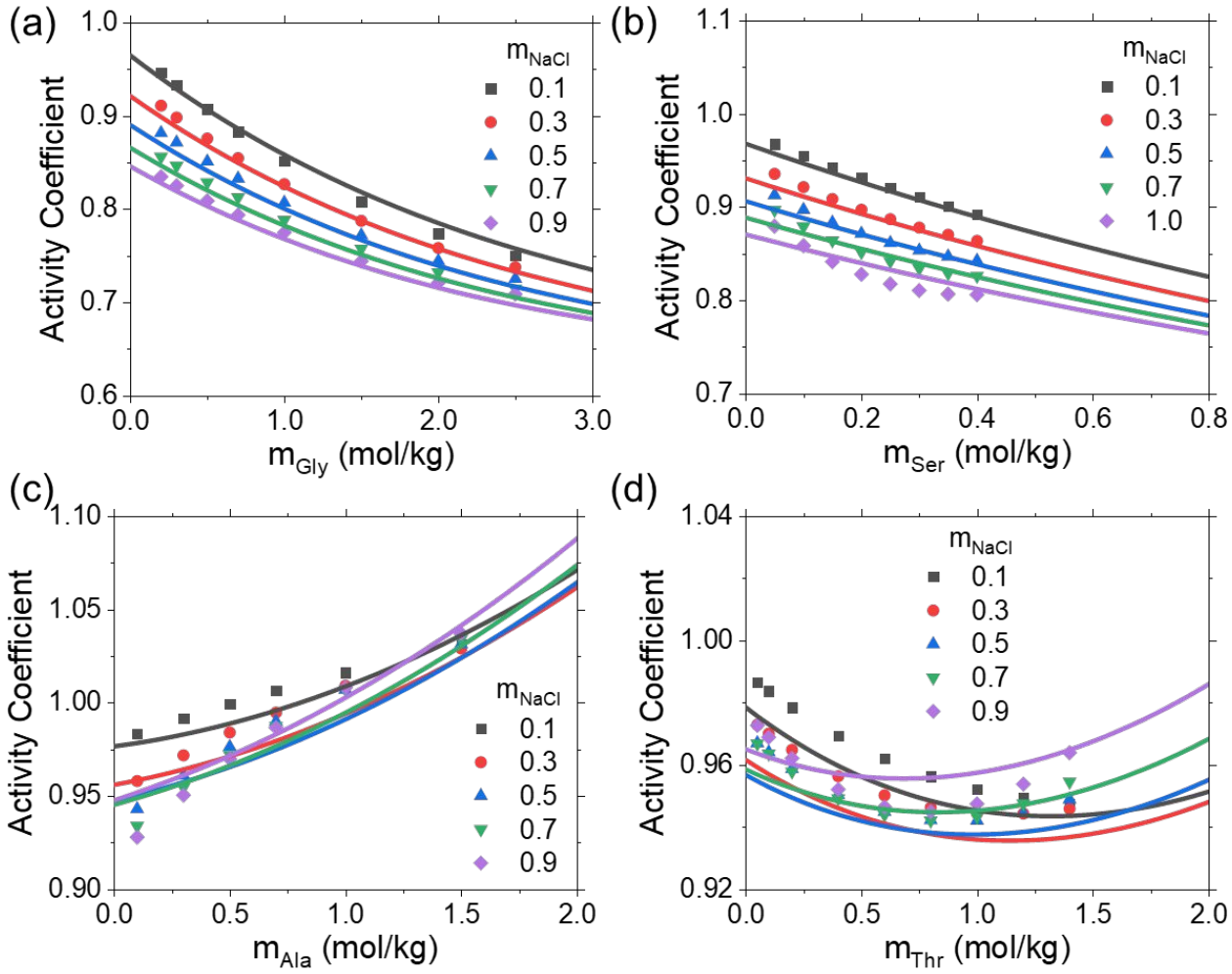

Figure 3. Activity coefficients of (a) glycine, (b) serine, (c) alanine, and (d) threonine in aqueous solutions of amino acids and sodium chloride from experiments[38-40] (symbols) and theoretical correlation (lines).

We next consider the activity coefficients of neutral amino acids in aqueous solutions at different concentrations of sodium chloride. In general, the addition of salt at a fixed amino acid molarity result in the reduction of the activity coefficient because the amino acid interacts favorably with the salt ions. As the molality of salt in the solution increases, the volume exclusion term begins to dominate the activity coefficient and leads to the large increase in activity coefficient for alanine and threonine. In Figure 3a, we see that the increase in glycine

and salt molality reduces the activity coefficient due to the favorable interactions of the backbone with other backbones and salt ions. A comparison of Figure 3a and 3b indicates that the addition of the serine side chain does not alter the trend in the reduction of the activity coefficient as the molality of serine or salt is increased, but the reduction becomes less significant because of the stronger excluded volume effect. At 0.8 molality, the higher salt curves are starting to collapse, suggesting that the volume exclusion outcompetes the attractive interactions between the amino acid. On the other hand, Figure 3c shows that the activity coefficient of alanine increases with added alanine. While the add salt reduces the activity coefficient at low alanine molality, the effect is less significant at high molality of the amino acid. Interestingly, alanine differs from glycine by only a methyl group in its side chain, yet the activity coefficients show significant differences, which indicates that even short side chains can play a major role in dictating the solvation behavior. Lastly, we consider the activity coefficient of threonine in Figure 3d, which contains a methyl and a hydroxyl group in its side chain. While the agreement between our model and the experiments is not as satisfactory as other cases, we see that our model does capture the non-monotonic effects of added threonine and salt molality on the activity coefficient. The hard-sphere repulsion plays a major role in the activity coefficient of threonine as seen by the non-monotonic effect of salt molality even at low molality of threonine.

Different from previous models for amino acids, the activity coefficients of these "neutral" amino acids contain an electrostatic component since the backbone amino and carboxyl groups are considered explicitly. The increase in salt concentration will increase the attractive electrostatic correlations resulting from the electrostatic interactions between the charged functional groups and the ions in solution. However, there is a contrasting effect as the attractive electrostatic interaction between the positively charged amino group and the negatively charged

carboxyl group will be screened by the increase in the salt concentration. As a result, the electrostatic component of the activity coefficient varies less significantly with salt concentration, but it is still overall an attractive contribution. While the comparison between our model and the experimental data is not perfect, we are able to provide at least a semi-quantitative description of the activity coefficients of neutral amino acids as a function of the concentrations of amino acid and salt ions. Importantly, our 3-bead model can predict the non-monotonic effects of added amino acid and salt due to the stronger excluded volume effects at high molarity.

Acidic and basic amino acids

The computational procedure to describe the acidic and basic amino acids is more complicated than that for the neutral amino acids. As previously discussed in Section 2.1 and 2.2, we must account for the ionization equilibria between the different configurations of the amino acid in order to determine their partial contributions to the activity coefficient. In addition, the measurements in pure water are taken at the isoelectric point of the amino acid; thus, we must self-consistently determine the isoelectric point during the ionization equilibria (i.e., adjust pH such that the charge neutrality and mass conservation is satisfied). While histidine and arginine tend to have one dominant configuration at the isoelectric point, lysine and aspartic acid both have same-charged ionizable groups with equilibrium constants that are close to one another, and therefore multiple configurations are possible. If the amino acid is added in its salted form, then typically one configuration is dominant (e.g., arginine in ArgHCl exists in its configuration [+1,-1,+1]).

Table 2. The hard-sphere diameters in units of Å for charged amino acids based off correlation to experimental data for the activity coefficient and apparent equilibrium constants.

	Hard-sphere diameter of side chain ($\sigma_{X,Z_{xNo}}$)				
Z_{nXo}	-2	-1	0	+1	+2
Arg	-	4.370	4.577	3.595	3.894
Lys	-	4.342	3.892	3.476	4.270
His	-	3.951	4.0628	3.3528	4.095
Glu	3.972	3.384	3.860	3.524	-
Asp	3.912	3.247	3.890	3.908	-

Table 3. The interaction energy in units k_BT for charged amino acids based off correlation to experimental data for the activity coefficient and apparent equilibrium constants.

	Interaction parameters	
	$\varepsilon_{X^0,X}$	$\varepsilon_{X^0,Na^+/Cl^-}$
Arg	1.408	0.410
Lys	0.499	0.447
His	0.911	0.597
Glu	-0.964	0.741
Asp	-0.964	0.741

We next investigate the influence of intermolecular interactions on the activity coefficients for acidic and basic amino acids in their neutral and charged forms. Different from that for neutral amino acids, we must calibrate our model parameters by comparison with the

experimental data for both the activity coefficient and apparent equilibrium constant. The parameters for these amino acids are listed in Table 2 and 3. In general, we find a noticeable volume change between the different charge states, although there is no clear trend between the different charge states. The amino acid is smallest when both the side chain and backbone groups are charged. For the interaction parameters, we find that arginine interacts with itself strongly, which can be attributed to the polar side chain. Lysine can form hydrogen bonds through its side chain amino group which leads to the self-attraction and attraction with other amino acids. Histidine, on the other hand, can interact with its aromatic group. We specify the aspartic acid parameters based off glutamic acid's parameters since the limited data available for aspartic acid indicates that the activity coefficient data of aspartic acid is slightly less than glutamic acid[41]. The repulsion between amino acids could be related to what is known as hydration repulsion in which the strong local hydrogen bonding network is disrupted by the nearby amino acid. Due to its high polarity, the salt effect on aspartic acid and glutamic acid is significant. On the other hand, lysine shows no favorable interaction with the salt ions, and the interactions is only weakly favorable when charged because of its long carbon chain. In contrast, arginine can interact favorably with the salt ions due to the highly polar group at the end of its side chain. Lastly, histidine can attract the salt ions via the cation-pi interaction[42].

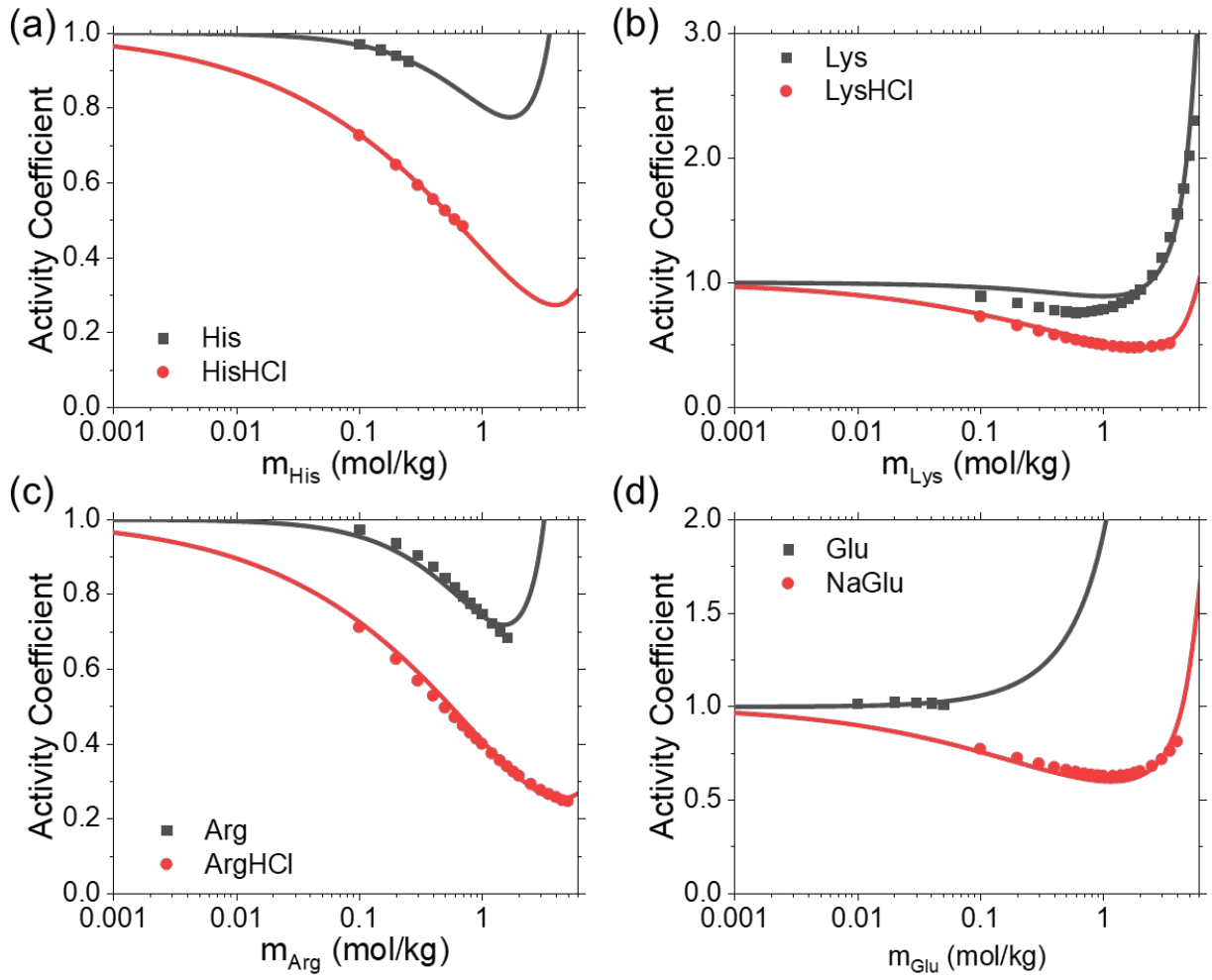

Figure 4. Activity coefficients of (a) lysine, (b) arginine, (c) histidine, and (d) glutamic acid in pure water (squares) and salted form (circles) from experiments[41, 43, 44] and theoretical correlation (solid lines).

Figure 4 compares the theoretical results with the experimental data for the activity coefficients of basic and acidic amino acids at their isoelectric point and in their salted forms. Our model shows excellent agreement for the basic and acidic amino acids in their charged and uncharged forms, indicating that the ionization equilibria are well captured by our coarse-grained model. In general, we find that the salted form has a lower activity coefficient that the neutral form. The difference can be attributed to the attractive electrostatic correlations between the

ionized functional groups and the salt in the solution. Both histidine and arginine show a significant decrease in the activity coefficient as the molality of the amino acid is increased until high molality at which the hard-sphere repulsion begins to dominate. In the case of lysine, however, the activity coefficient slightly decreases with the molality of the amino acid at low concentration. For glutamic acid, only the salted form shows a decrease in the activity coefficient whereas the neutral form of glutamic acid only increases with added glutamic acid. Our model well captures the activity coefficient data in both charged and uncharged forms. The only exception is the neutral lysine; our model predicts that the activity coefficient exhibits a minimum of 0.90 whereas the experimental minimum is 0.76. The discrepancy may arise from the negatively and positively charged forms of the lysine making a non-negligible contribution to the activity coefficient. This breaks the assumption used in the experimental measurement that lysine does not dissociate at the neutral condition[43, 44].

3.2 Apparent equilibrium constants

Unlike the thermodynamic equilibrium constants, the apparent equilibrium constants depend on the local solution conditions. As shown in Eq.(7), an apparent equilibrium constant lumps the experimentally non-measurable thermodynamic equilibrium constant with the non-ideality effects due to inter- and intramolecular interactions. The apparent value of the equilibrium constant provides insight into how the charge regulation of the amino acid is influenced (e.g., suppressed) by the changes in the solution condition. Unfortunately, the apparent equilibrium constants are only applicable within a small window of their reported salt concentrations. In addition, it is difficult to measure these constants at salt concentrations below 0.01 M. Thus, a reliable model that can incorporate the key physics governing the protonation of

amino acids will allow insight into dilute to concentrated salt conditions and the extrapolation of the thermodynamic equilibrium constant.

Table 4. The site-specific equilibrium constants for the protonation of different amino acids.

Thermodynamic Equilibrium Constants			
	N	O	X
Alanine	8.91	3.43	-
Serine	8.22	2.23	-
Aspartic acid	8.25	2.84	4.20
Glutamic acid	8.16	3.02	4.59
Lysine	9.02	3.64	10.06
Arginine	8.72	3.66	11.88

Table 4 presents the correlated equilibrium constants for the amino acids considered in Figures 5, 6, and 7. It is important to note that these constants are independent of the charge state of the amino acid since we have considered the intramolecular interactions and correlation effects explicitly. We find that the equilibrium constant for the ionization of the amino segment tends to decrease (i.e., harder to ionize) from alanine to serine. The equilibrium constant for the amino group in aspartic acid and glutamic acid are similar to that of serine, which is likely related to their polar oxygen groups in the backbone. On the other hand, lysine and arginine's equilibrium constants are comparable to that of alanine since they both have a multiple carbon chain separating their side chain polar group from the backbone. For the backbone carboxyl group, the equilibrium constant tends to increase (i.e., harder to ionize) from the acidic to neutral to basic amino acids. We can understand this trend as originating from the different electron draw of the polar groups. The equilibrium constant for the carboxyl group in arginine and lysine

are approximately the same due to the similarity in how their side chains interact with the carboxyl group.

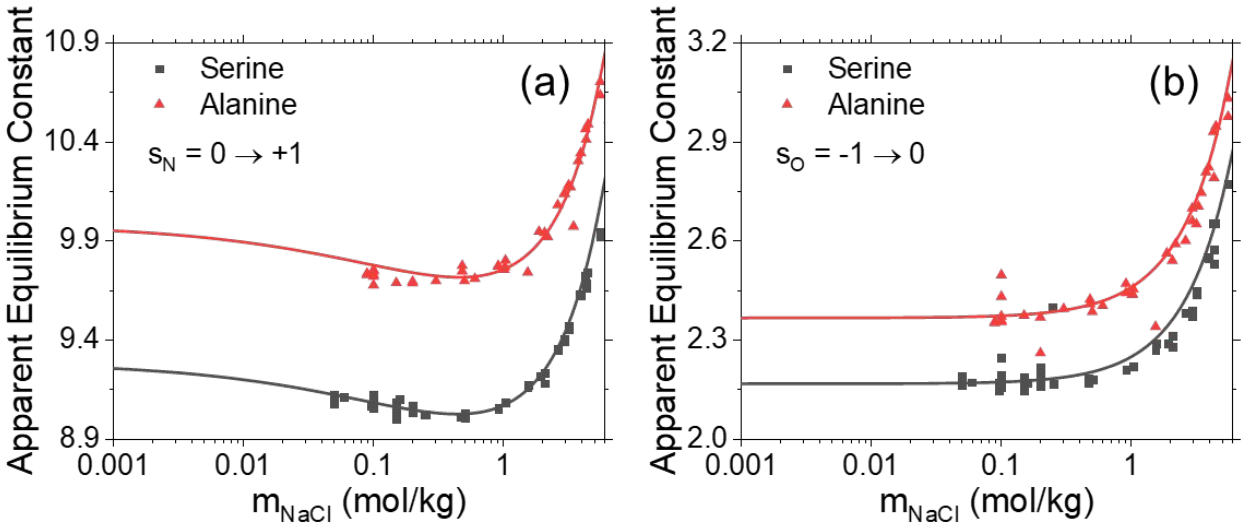

Figure 5. Apparent equilibrium constants for serine and alanine versus the sodium chloride molality from experiments[18, 22] (symbols) and theoretical prediction (lines) for (a) the protonation of the backbone amino group, and (b) the protonation of the backbone carboxyl group.

We first consider the protonation of two neutral amino acids, serine and alanine. As shown in Figure 5, the apparent equilibrium constant depends on the salt molality as well as the concentration of amino acid. While Figure 5a depicts the same functional group protonating, there is a significant difference in the apparent equilibrium constant between serine and alanine (~0.7 units at dilute salt conditions). The presence of the -OH group in serine will change the electron distribution, and thus it becomes more difficult to ionize (or protonate) the amino group. For a given curve, we fit the thermodynamic equilibrium constant and the volume change of the amino acid upon charging. In general, we find that the charged form is larger than the neutral form which may be attributed to the expansion of the water layer in the solvation shell.

From Figure 5a, we see that the addition of salt initially inhibits ionization of the amino acid. This is because the amino acid loses its total charge (i.e., it becomes a zwitterion), which is less favorable than maintaining its charge. However, as the molality is increased further, the apparent equilibrium constant will begin to increase. Similar to the activity coefficients of amino acids, the sharp rise can be attributed to the volume exclusion term. In this case, the hard-sphere repulsion favors the zwitterion form of the amino acid since it is smallest in size in this configuration.

Different from the ionization of the amino group shown in Figure 5a, Figure 5b indicates that the apparent equilibrium constant for the protonation of the backbone carboxyl group changes monotonically with salt molality. Since the protonation of the negatively carboxyl group would lead to a net positively charged amino acid and a loss in the intramolecular interaction between the carboxyl group and the amino group, the electrostatic contribution to the apparent equilibrium constant is small. Overall, the model well captures the experimental data across a large range of salt conditions for neutral amino acids.

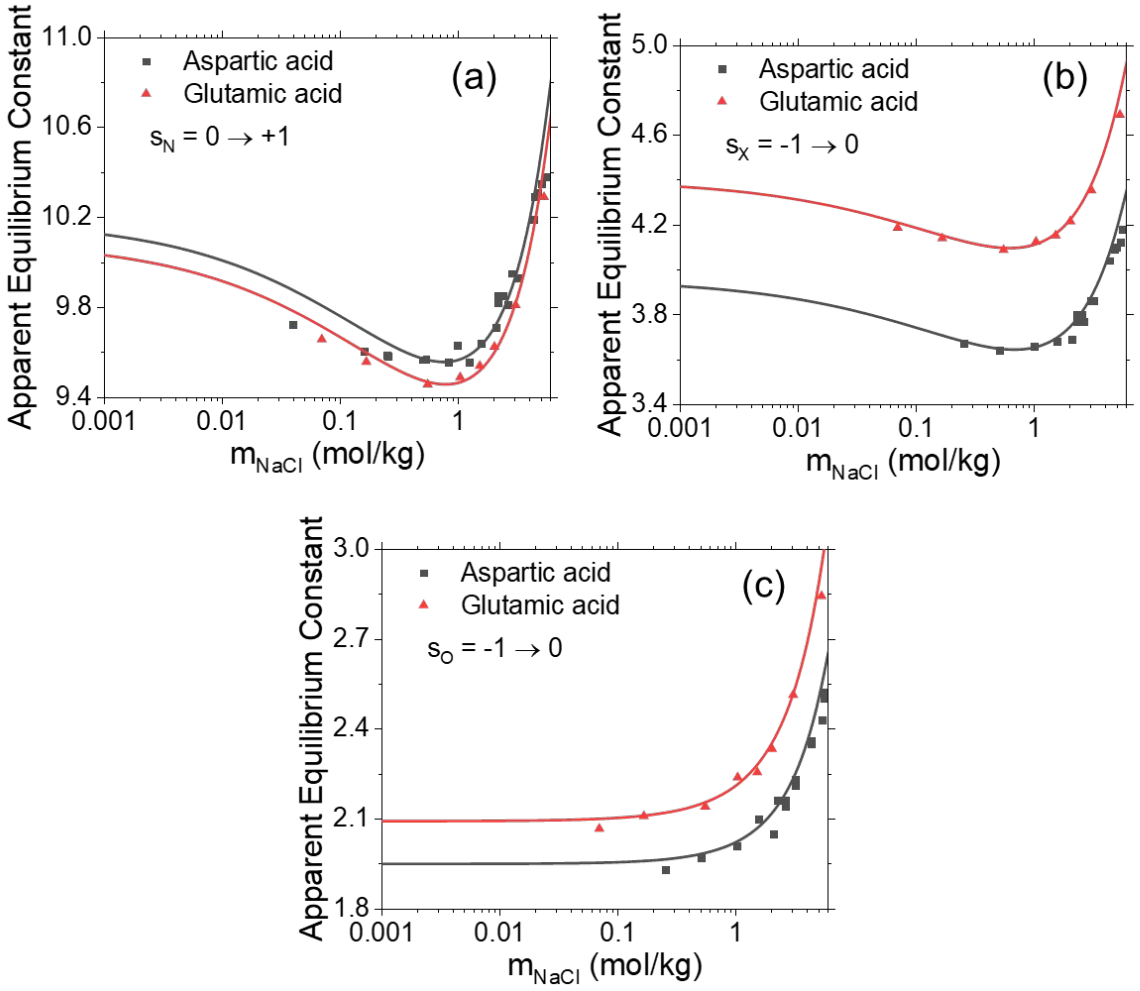

Figure 6. Apparent equilibrium constants for aspartic acid and glutamic acid versus the sodium chloride molality from experiments[18, 22, 45] (symbols) and theoretical prediction (lines) for (a) the protonation of the backbone amino group, (b) the protonation of the sidechain carboxyl group, and (c) the protonation of the backbone carboxyl group.

Figure 6 shows the apparent equilibrium constants for the protonation of the acidic amino acids, aspartic acid and glutamic acid, at different salt concentrations. These two amino acids are similar except that there is one additional carbon separating the carboxyl group in the sidechain of glutamic acid from its backbone compared to aspartic acid. Despite this small change in the structure, there is a noticeable change in the apparent equilibrium constants for all protonation

steps. We see that the shape of the curve in Figure 6a and 6b is similar to that in Figure 5a. However, Figure 6a shows that the apparent equilibrium constant changes more rapidly at low salt concentration. The reason for the stronger change with salt concentration is due to the transition from a 'divalent' amino acid (viz., the backbone carboxyl group and sidechain carboxyl group are charged) to a 'monovalent' amino acid due to the presence of the positively charged amino group. To be specific, increasing the salt concentration results in a smaller repulsion between the two charged carboxyl groups in the divalent amino acid. On the other hand, the screening by salt ions reduces the attractive interactions in the monovalent amino acid between the positively charged amino group and the negatively charged carboxyl group. Thus, we see that the as the salt molality is increased, it is at first more difficult to ionize the amino group in the backbone, but as the concentration increases, the excluded volume effect will drive the ionization of the functional group. By accounting for the individual ionizable groups within the amino acid, we are able to capture the correlated ionization between these two groups and quantitatively capture the experimental data for the apparent equilibrium constants.

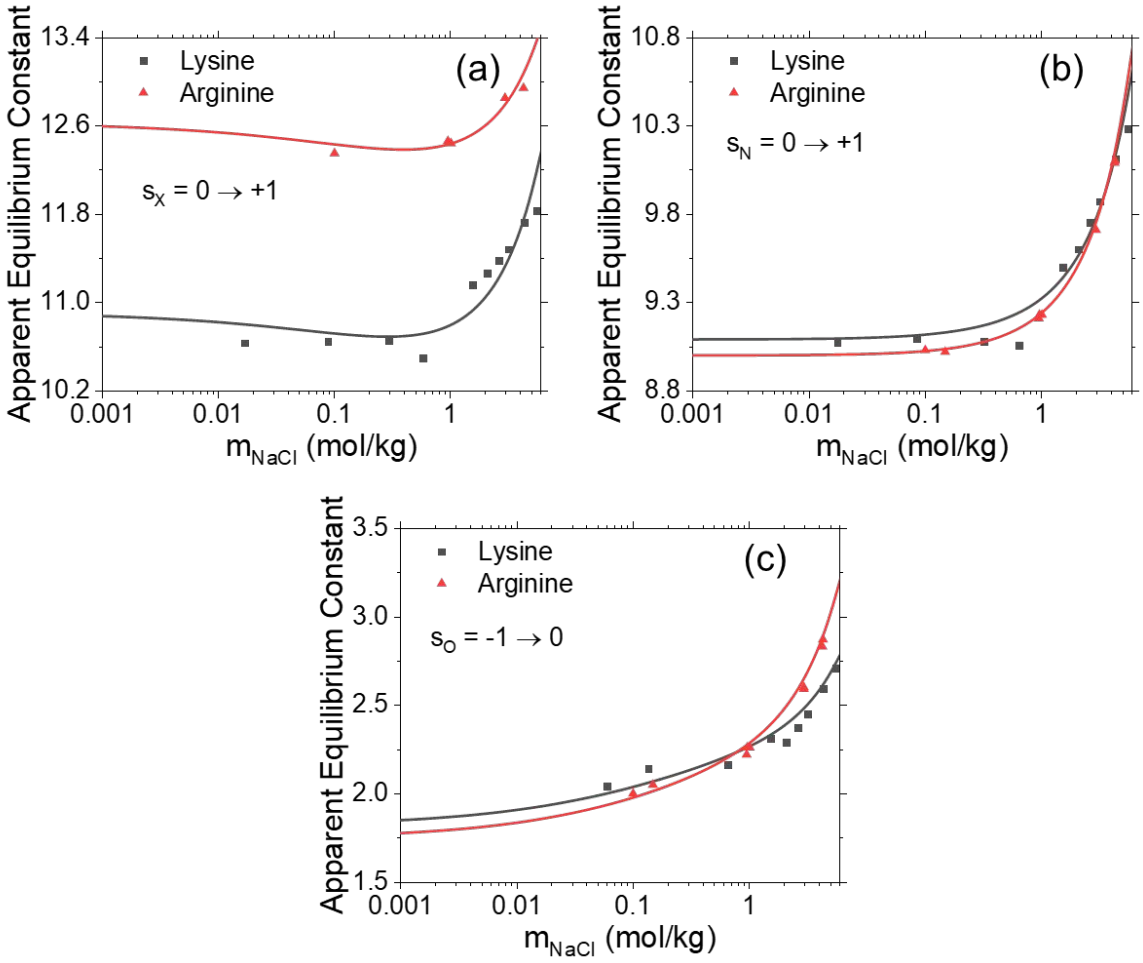

Figure 7. Apparent equilibrium constants for lysine and arginine versus the sodium chloride molality from experiments[18, 22] (symbols) and theoretical prediction (lines) for (a) the protonation of the sidechain amino group, (b) the protonation of the backbone amino group, and (c) the protonation of the backbone carboxyl group.

The last case we consider is the protonation of the basic amino acids, lysine and arginine, as a function of sodium chloride molality. As shown in Figure 7, the first protonation step of lysine or arginine is the ionization of the amino or guanidinium group, respectively, in the amino acid's side chain and the curve is similar to that of Figure 5a. The significantly higher apparent equilibrium constant for the guanidinium group (specifically its thermodynamic equilibrium constant) is because resonance between the nitrogen groups allows it to stabilize the charge. The

second protonation step involves the amino group in the backbone of the zwitterion form of the amino acid where the sidechain amino group and backbone carboxyl group are already charged. In this case, the electrostatic interactions mostly cancel with each other and the curve looks similar to that in Figure 5b where the apparent equilibrium constant is plateaued at low molality while it increases at high molality. The last protonation step is for the backbone carboxyl group, which results in the amino acid being in a positively charged divalent configuration. As shown in Figure 7c, the apparent equilibrium constant changes rapidly with salt molality because electrostatics are driving the amino acid to be in the configuration where both amino groups are charged while the carboxyl group is uncharged. In addition, we see in Figure 7b and 7c that the apparent equilibrium constant for arginine increases more with salt molality than lysine due to the differences in the excluded volume effects. The electrostatic and volume exclusion behavior in the apparent equilibrium constants is well captured by our model, demonstrating its capability in describing all steps of protonation for different amino acids.

3.2 Speciation of amino acids in aqueous solutions

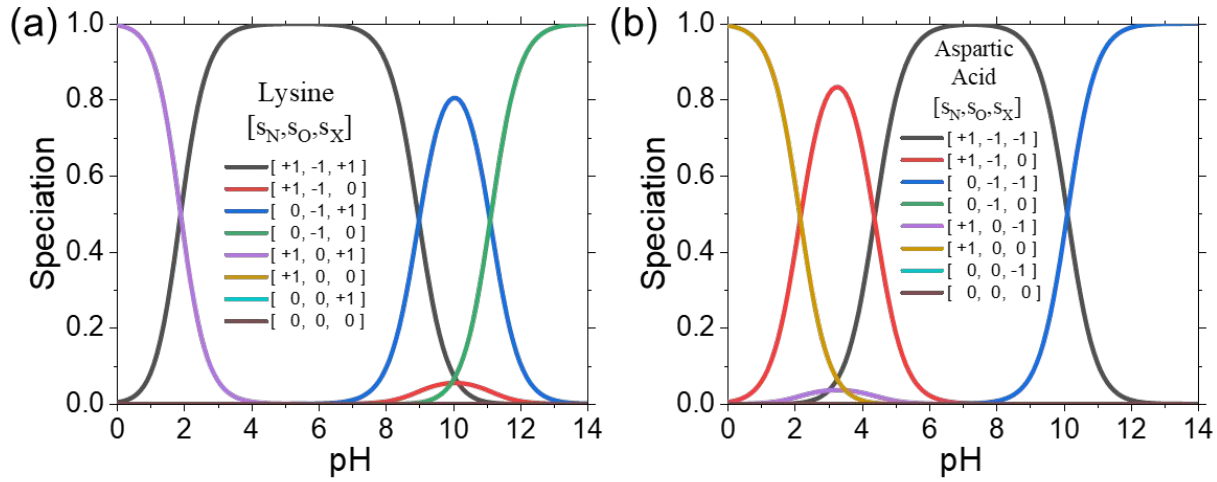

Figure 8. The speciation diagram of (a) lysine and (b) aspartic acid into their 8 possible configurations as a function of pH in a dilute amino acid and salt ($1\mu M$) aqueous solution.

As discussed previously, amino acids can exist in multiple configurations due to the protonation of their functional groups. The change between the different charge states is governed by their equilibrium constants as well as the differences in inter- and intramolecular interactions. The speciation of an amino acid can be calculated from Eq.(9) at a given set of solution conditions (viz., pH, amino acid concentration, and salt concentration). Figure 8 shows the speciation diagram of the basic amino acid lysine and the acidic amino acid aspartic acid as a function of pH in a dilute aqueous solution. We see that lysine is generally in one of four configurations: [+1,0,+1] means that the divalent cation prevalent at low pH; [+1,-1,+1] prevails when the monovalent cation present in the range of pH 2 to 9, [0,-1,+1] corresponds to the zwitterion form (sidechain amino group is charged) present in the range of pH 9 to 11, and [0,-1,0] occurs when the monovalent anion present at high pH. In the pH range from 8 to 12, we do see the presence of a fifth configuration, a zwitterion form where the backbone amino group is charged but not the sidechain group. However, this configuration has only a small presence (<5%). In addition, we see that the sidechain mostly ionizes while the amino group in the backbone is uncharged, as indicated by the relative speciation of [+1,-1,+1] and [0,-1,+1]. For example, at pH 10, the percentage of amino acids in the two configurations is 5% and 80%, respectively. Thus, the interaction between the charged amino groups in the sidechain and backbone can be neglected when calibrating the model to the experimental data for the protonation of the sidechain. A similar situation holds for aspartic acid, except in this case, the sidechain carboxyl group ionizes after the backbone carboxyl group. Thus, we validate the assumption that Eq.(8) could be reasonably approximated by considering only the transition between the two dominant states [0,-1,0] to [0,-1,+1] for lysine and [+1,-1,0] to [+1,-1,-1] for aspartic acid.

4. Conclusion

In this work, we have developed a 3-bead augmented primitive model (3bAPM) that captures the thermodynamic properties of all twenty natural amino acids in aqueous sodium chloride solutions. Our model distinguishes the amino and carboxyl group present in the backbone of each amino acid. As a result, we are able to account for the intrachain correlations on the activity coefficient and ionization equilibrium. In addition to the volume exclusion effects due to the finite size of the solute species and electrostatic correlations, 3bAPM accounts for the water-mediated attraction of amino acids with themselves and with salt ions. The activity coefficients are described in terms of the mean-spherical approximation (MSA) for the primitive model and a mean-field approximation for the short-range attraction represented by the square-well potential. Additionally, there is a contribution from the intramolecular correlations between the three beads from both electrostatics and volume exclusion.

We have demonstrated that the 3bAPM provides a semi-quantitatively description of the activity coefficients of amino acids in pure water and in aqueous sodium chloride solutions. The correlation parameters were the energy of interaction between the amino acid backbone with itself and with the salt ions while the diameter of each segment was 1.21 Å less than the van der Waals diameter based off fitting. In addition, we have obtained the interaction parameters between the side chain and itself as well as with salt ions. These parameters allow us to distinguish between the backbone mediated interactions versus the side chain interactions. When determining the equilibrium constants, the correlation parameters were the change in hydration of the amino acid which can then be related to the change in diameter for each segment due to their hydration shell. We find that, due to the hydrogen binding and polarity, hydrophilic amino acids tend to be more attractive than repulsive to itself while the opposite is generally true for

hydrophobic amino acids. In addition, amino acids are generally attracted to salt ions; however, at high concentrations of amino acid and/or salt concentration, the volume exclusion dominates thereby leading to a positive deviation in the activity coefficient.

An important aspect of our coarse-grained model is that it is able to capture the charge regulation behavior of amino acids resulting from differences in their inter- and intramolecular interactions and the intra-chain correlation effects. We find that our model well describes the activity coefficient of basic and acidic amino acids in their uncharged and charged states. Furthermore, comparison with experimental data for the apparent equilibrium constants demonstrates our 3bAPM accurately captures the physics dictating the protonation behavior of amino acids. From these correlations, we are able to determine the fundamental constants governing the protonation of the amino acid functional groups independent of the ionization state of any other functional groups. These fundamental constants will be valuable in determining the charge states of amino acids and understanding their biochemical properties under complicated solution conditions. In addition, our model sets up a clear pathway for further study of polypeptides and unstructured proteins.

Supporting Information:

The supporting information contains theoretical details including the expressions for activity coefficient and conversion from molality to molarity scale for comparison with experimental data.

Acknowledgements:

This work is financially supported by the NSF-DFG Lead Agency Activity in Chemistry and Transport in Confined Spaces under Grant No. NSF 2234013. Additional support is provided by the NSF Graduate Research Fellowship under Grant No. DGE-1326120.

References

(1) Kord Forooshani, P.; Lee, B. P. Recent approaches in designing bioadhesive materials inspired by mussel adhesive protein. *Journal of Polymer Science Part A: Polymer Chemistry* **2017**, *55* (1), 9-33.
(2) Favi, P. M.; Yi, S.; Lenaghan, S. C.; Xia, L.; Zhang, M. Inspiration from the natural world: from bio-adhesives to bio-inspired adhesives. *Journal of Adhesion Science and Technology* **2014**, *28* (3-4), 290-319.
(3) Hollingshead, S.; Lin, C. Y.; Liu, J. C. Designing smart materials with recombinant proteins. *Macromolecular bioscience* **2017**, *17* (7), 1600554.
(4) Fairman, R.; Åkerfeldt, K. S. Peptides as novel smart materials. *Current opinion in structural biology* **2005**, *15* (4), 453-463.
(5) Bauri, K.; Nandi, M.; De, P. Amino acid-derived stimuli-responsive polymers and their applications. *Polymer Chemistry* **2018**, *9* (11), 1257-1287.
(6) Saha, R.; Bhattacharya, D.; Mukhopadhyay, M. Advances in modified Antimicrobial Peptides as marine antifouling material. *Colloids and Surfaces B: Biointerfaces* **2022**, 112900.
(7) Xu, C.; Liu, X.; Xie, B.; Yao, C.; Hu, W.; Li, Y.; Li, X. Preparation of PES ultrafiltration membranes with natural amino acids based zwitterionic antifouling surfaces. *Applied Surface Science* **2016**, *385*, 130-138.
(8) Atkins, P. W.; De Paula, J.; Keeler, J. *Atkins' physical chemistry molecular thermodynamics and kinetics*; Oxford University Press, 2019.
(9) Pytkowicz, R. M. Use of apparent oceanography, equilibrium constants in chemical geochemistry, and biochemistry. *Geochemical Journal* **1969**, *3* (2-3), 181-184.
(10) Kiss, T.; Sovago, I.; Gergely, A. Critical survey of stability constants of complexes of glycine. *Pure and applied chemistry* **1991**, *63* (4), 597-638.
(11) Bretti, C.; Crea, F.; De Stefano, C.; Sammartano, S.; Vianelli, G. Some thermodynamic properties of dl-Tyrosine and dl-Tryptophan. Effect of the ionic medium, ionic strength and temperature on the solubility and acid–base properties. *Fluid Phase Equilibria* **2012**, *314*, 185-197. DOI: https://doi.org/10.1016/j.fluid.2011.10.007.
(12) Gómez-Bombarelli, R.; González-Pérez, M.; Pérez-Prior, M. T.; Calle, E.; Casado, J. Computational Calculation of Equilibrium Constants: Addition to Carbonyl Compounds. *The Journal of Physical Chemistry A* **2009**, *113* (42), 11423-11428. DOI: 10.1021/jp907209a.
(13) Kallies, B.; Mitzner, R. p K a values of amines in water from quantum mechanical calculations using a polarized dielectric continuum representation of the solvent. *The Journal of Physical Chemistry B* **1997**, *101* (15), 2959-2967.
(14) Guggenheim, E. A.; Turgeon, J. C. Specific interaction of ions. *Transactions of the Faraday Society* **1955**, *51* (0), 747-761, 10.1039/TF9555100747. DOI: 10.1039/TF9555100747.
(15) Bretti, C.; Foti, C.; Sammartano, S. A new approach in the use of SIT in determining the dependence on ionic strength of activity coefficients. Application to some chloride salts of interest in the speciation of natural fluids. *Chemical Speciation & Bioavailability* **2004**, *16* (3), 105-110. DOI: 10.3184/095422904782775036.
(16) Bickmore, B. R.; Wander, M. C. F. Activity and Activity Coefficients. In *Encyclopedia of Geochemistry: A Comprehensive Reference Source on the Chemistry of the Earth*, White, W. M. Ed.; Springer International Publishing, 2018; pp 21-23.
(17) Vilariño, T.; Fiol, S.; L. Armesto, X.; Brandariz, I.; E. Sastre de Vicente, M. Effect of ionic strength on the protonation of various amino acids analysed by the mean spherical

approximation. *Journal of the Chemical Society, Faraday Transactions* **1997**, *93* (3), 413-417, 10.1039/A605917A. DOI: 10.1039/A605917A.
(18) Bretti, C.; Cigala, R. M.; Giuffrè, O.; Lando, G.; Sammartano, S. Modeling solubility and acid-base properties of some polar side chain amino acids in NaCl and (CH3)4NCl aqueous solutions at different ionic strengths and temperatures. *Fluid Phase Equilibria* **2018**, *459*, 51-64. DOI: https://doi.org/10.1016/j.fluid.2017.11.030.
(19) Bretti, C.; Giuffrè, O.; Lando, G.; Sammartano, S. Modeling solubility and acid–base properties of some amino acids in aqueous NaCl and (CH3)4NCl aqueous solutions at different ionic strengths and temperatures. *SpringerPlus* **2016**, *5* (1), 928. DOI: 10.1186/s40064-016-2568-8.
(20) Pitzer, K. S. *Activity coefficients in electrolyte solutions*; CRC press, 2018.
(21) Kim, H. T.; Frederick, W. J. Evaluation of Pitzer ion interaction parameters of aqueous electrolytes at 25.degree.C. 1. Single salt parameters. *Journal of Chemical & Engineering Data* **1988**, *33* (2), 177-184. DOI: 10.1021/je00052a035.
(22) De Stefano, C.; Foti, C.; Gianguzza, A.; Sammartano, S. The interaction of amino acids with the major constituents of natural waters at different ionic strengths. *Marine Chemistry* **2000**, *72* (1), 61-76. DOI: https://doi.org/10.1016/S0304-4203(00)00067-0.
(23) Gallegos, A.; Wu, J. Charge Regulation of Natural Amino Acids in Aqueous Solutions. *Journal of Chemical & Engineering Data* **2020**. DOI: 10.1021/acs.jced.0c00625.
(24) Khoshkbarchi, M. K.; Vera, J. H. A simplified perturbed hard-sphere model for the activity coefficients of amino acids and peptides in aqueous solutions. *Industrial & engineering chemistry research* **1996**, *35* (11), 4319-4327.
(25) Khoshkbarchi, M. K.; Vera, J. H. A perturbed hard-sphere model with mean spherical approximation for the activity coefficients of amino acids in aqueous electrolyte solutions. *Industrial & engineering chemistry research* **1996**, *35* (12), 4755-4766.
(26) Khoshkbarchi, M. K.; Vera, J. H. A Theoretically Improved Perturbation Model for Activity Coefficients of Amino Acids and Peptides in Aqueous Solutions. *Industrial & Engineering Chemistry Research* **1998**, *37* (8), 3052-3057. DOI: 10.1021/ie970806g.
(27) Blum, L. Mean spherical model for asymmetric electrolytes. *Molecular Physics* **1975**, *30* (5), 1529-1535. DOI: 10.1080/00268977500103051.
(28) Jin, Z.; Tang, Y.; Wu, J. A perturbative density functional theory for square-well fluids. *The Journal of Chemical Physics* **2011**, *134* (17), 174702. DOI: 10.1063/1.3585677.
(29) Gallegos, A.; Wu, J. Hierarchical Model of Weak Polyelectrolytes with Ionization and Conformation Consistency. *Macromolecules* **2023**. DOI: 10.1021/acs.macromol.2c01910.
(30) Yu, Y.-X.; Wu, J. Density functional theory for inhomogeneous mixtures of polymeric fluids. *The Journal of chemical physics* **2002**, *117* (5), 2368-2376.
(31) Henderson, D.; Smith, W. R. Exact analytical formulas for the distribution functions of charged hard spheres in the mean spherical approximation. *Journal of Statistical Physics* **1978**, *19* (2), 191-200.
(32) Soares, E. d. A.; Vernin, N. S.; Santos, M. S.; Tavares, F. W. Real Electrolyte Solutions in the Functionalized Mean Spherical Approximation: A Density Functional Theory for Simple Electrolyte Solutions. *The Journal of Physical Chemistry B* **2022**, *126* (32), 6095-6101. DOI: 10.1021/acs.jpcb.2c00816.
(33) Quiñones, A. O.; Bhuiyan, L. B.; Abbas, Z.; Outhwaite, C. W. Influence of concentration and temperature dependent dielectric constants on the thermodynamics of electrolytes. *Journal of Molecular Liquids* **2023**, *371*, 121119.

(34) Zhao, Y. H.; Abraham, M. H.; Zissimos, A. M. Fast Calculation of van der Waals Volume as a Sum of Atomic and Bond Contributions and Its Application to Drug Compounds. *The Journal of Organic Chemistry* **2003**, *68* (19), 7368-7373. DOI: 10.1021/jo034808o.
(35) Marcus, Y. A simple empirical model describing the thermodynamics of hydration of ions of widely varying charges, sizes, and shapes. *Biophysical chemistry* **1994**, *51* (2-3), 111-127.
(36) Kuramochi, H.; Noritomi, H.; Hoshino, D.; Nagahama, K. Measurements of vapor pressures of aqueous amino acid solutions and determination of activity coefficients of amino acids. *Journal of Chemical & Engineering Data* **1997**, *42* (3), 470-474.
(37) Held, C.; Cameretti, L. F.; Sadowski, G. Measuring and Modeling Activity Coefficients in Aqueous Amino-Acid Solutions. *Industrial & Engineering Chemistry Research* **2011**, *50* (1), 131-141. DOI: 10.1021/ie100088c.
(38) Soto-Campos, A. M.; Khoshkbarchi, M. K.; Vera, J. H. Interaction ofDL-threonine with NaCl and NaNO3in aqueous solutions: e.m.f. measurements with ion-selective electrodes. *The Journal of Chemical Thermodynamics* **1997**, *29* (5), 609-622. DOI: https://doi.org/10.1006/jcht.1996.0182.
(39) Khoshkbarchi, M. K.; Vera, J. H. Measurement of activity coefficients of amino acids in aqueous electrolyte solutions: Experimental data for the systems H2O+ NaCl+ glycine and H2O+ NaCl+ DL-alanine at 25 C. *Industrial & engineering chemistry research* **1996**, *35* (8), 2735-2742.
(40) Khoshkbarchi, M. K.; Soto-Campos, A. M.; Vera, J. H. Interactions of DL-serine and L-serine with NaCl and KCl in aqueous solutions. *Journal of solution chemistry* **1997**, *26*, 941-955.
(41) Tsurko, E. N.; Neueder, R.; Kunz, W. Activity of water, osmotic and activity coefficients of sodium glutamate and sodium aspartate in aqueous solutions at 310.15 K. *Acta Chim. Slov* **2009**, *56*, 58-64.
(42) Liao, S.-M.; Du, Q.-S.; Meng, J.-Z.; Pang, Z.-W.; Huang, R.-B. The multiple roles of histidine in protein interactions. *Chemistry Central Journal* **2013**, *7* (1), 1-12.
(43) Bonner, O. D. Osmotic and activity coefficients of some amino acids and their hydrochloride salts at 298.15 K. *Journal of Chemical and Engineering Data* **1982**, *27* (4), 422-423.
(44) Tsurko, E. N.; Neueder, R.; Kunz, W. Water activity and osmotic coefficients in solutions of glycine, glutamic acid, histidine and their salts at 298.15 K and 310.15 K. *Journal of solution chemistry* **2007**, *36*, 651-672.
(45) Nagai, H.; Kuwabara, K.; Carta, G. Temperature dependence of the dissociation constants of several amino acids. *Journal of Chemical & Engineering Data* **2008**, *53* (3), 619-627.