Journal of Solution Chemistry
https://doi.org/10.1007/s10953-026-01561-9

Meta-analysis on the Salt Effect on Glycine Solubility Applying Gaussian Processes

Christopher A. Piske[1,2] · Priscilla G. Leite[2] · Mónia A. R. Martins[1] · Olga Ferreira[1] · João A. P. Coutinho[3] · Dinis O. Abranches[3] · Simão P. Pinho[1]

Received: 4 October 2025 / Accepted: 3 January 2026
© The Author(s) 2026

Abstract
In aqueous solutions containing electrolytes, ions influence both the solubility and the stability of biomolecules. However, inconsistencies across published data highlight the need for a critical review. To address this, a database was constructed on the solubility of glycine in electrolyte solutions spanning from 1996 to 2024, and the experimental data were critically evaluated. Gaussian Process (GP) models were implemented to analyze, predict, and validate solubility behavior. The GP model successfully captures salting-in and salting-out trends, along with specific ion effects reported in the literature. It also provides predictive uncertainty estimates that help identify potentially inconsistent data points or sets. This uncertainty-based analysis enables the reconciliation of conflicting datasets and helps prioritize new experimental measurements in regions where data are sparse or less reliable. By applying a data-filtering method that removes experimental points falling outside the uncertainty range of the model, the influence of inconsistent values is reduced. This results in a more robust model fit and improved prediction accuracy. Therefore, the GP establishes a quantitative foundation for consolidating the current knowledge on the solubility of glycine in saline solutions, identifying methodological inconsistencies in the literature.

Keywords Electrolyte solutions · Solubility · Gaussian Process · Uncertainty · Data reliability

✉ Dinis O. Abranches
jdinis@ua.pt

✉ Simão P. Pinho
spinho@ipb.pt

[1] CIMO-Mountain Research Center, LA Sus TEC, Bragança Polytechnic University, Campus de Santa Apolónia Campus, Bragança, Portugal

[2] UTFPR – Federal Technological University of Paraná, Ponta Grossa 84016-210, Brazil

[3] CICECO – Aveiro Institute of Materials, University of Aveiro, 3810-193 Aveiro, Portugal

Published online: 04 February 2026

1 Introduction

Aqueous solutions containing electrolytes constitute the natural environment for many biomolecules, such as proteins, nucleic acids, and enzymes [1, 2]. These aqueous systems are fundamental for maintaining the three-dimensional structure of these molecules and their functionality [1]. Ions in solution can either stabilize or disrupt intermolecular interactions, thereby influencing molecular behavior [2–4]. As a result, electrolytes play a vital role in regulating the physicochemical properties of biomolecules and are involved in numerous biological processes [1, 2].

Current contributions to the knowledge of the solubility of amino acids in the presence of salts are extremely important in several aspects. From the biological point of view, it directly influences the stability of these molecules in aqueous medium [2, 5]. Such knowledge is also essential in many industrial applications, especially in the food and pharmaceutical sectors, where product formulation requires precise control of the interactions between biomolecules and ions [2, 6].

Even though relevant data on the solubility of amino acids in aqueous electrolyte solutions can be found, inconsistencies between the data are evident [5, 7], and there is a considerable lack of information on aromatic amino acids or those with more than one amino or carboxylic acid group [2, 3, 8]. These limitations make it necessary to take a critical look and review the experimental approaches, checking for their reliability [5, 7]. The lack of consensus among the data highlights the importance of solving these evident inconsistencies [5, 7].

In this context, machine learning models emerge as a promising alternative for addressing the evaluation of the experimental data [9]. These models are capable of identifying patterns within dispersed and incomplete data sets, enhancing the interpretation of trends and enabling the analysis of inconsistencies [9]. Among the available approaches, Gaussian processes (GPs) stand out not only due to their ability to model a large volume of observed data but also due to their stochastic nature, i.e., their ability to provide a quantification of the uncertainty associated with their predictions [9]. This feature is especially valuable for detecting inconsistencies between different experimental datasets, reinforcing the importance of critically assessing both the methodologies employed and the reliability of reported results [9]. Therefore, GP is a strategic tool for consolidating existing data and guiding future experimental validation efforts in a more rational way.

This study proposes a modeling framework based on GPs to analyze, predict, and validate the solubility of glycine in aqueous solutions containing electrolytes, using an experimental database compiled from multiple literature sources [1–8, 10–20]. The GP model enables the identification of inconsistencies among datasets, capturing salting-in/salting-out phenomena and ion-specific effects. It provides an assessment of the uncertainty, helping in the detection of data points deviating from expected trends. In this way, the GP offers a robust quantitative basis for validating existing data, resolving experimental inconsistencies, and guiding the design of future experiments more assertively.

2 Methodology

The structuring of a database related to the solubility of amino acids in aqueous systems containing salts represented the initial and fundamental stage of this research, requiring careful analysis of each author and the respective data reported, especially in the methodologies used, the results achieved, and their relative errors.

Solubility data collection was carried out for glycine in various electrolyte aqueous solutions to map and compare the different salt effects. To ensure the standardization of the experimental conditions, only studies carried out at a temperature of 298.2 K were selected. The database, organized in a spreadsheet, gathers salt molality, solubility, relative solubility, uncertainty, and temperature, in addition to the identification of the salt studied and the respective bibliographic reference. At the end of this stage, it was evident that many of the data collected presented inconsistencies in relation to the others, indicating the presence of possibly flawed or methodologically questionable experimental results, and it was necessary to perform a structural chemical analysis of each salt. The analysis is based on relative solubility, defined as the ratio of glycine solubility at a given salt concentration to its solubility in pure water reported by the same authors. This approach also minimizes the impact of variability in the solid crystal structure (α, β, or γ), which remains unidentified in most studies.

2.1 Sigma Profiles

Sigma profiles are molecular descriptors derived from quantum chemistry (DFT) calculations in which a molecule is placed in a continuum-solvent model and its screened surface charge density is computed. By partitioning the molecular surface into segments and constructing an unnormalized histogram of these local charge values, the sigma profile encodes the distribution of polarity, electron density, and solvation-relevant features in a size-independent vector. This representation captures chemically meaningful information that is often inaccessible to descriptors based solely on atom types, connectivity, or graph topology.

Abranches and co-workers [21] have shown that sigma profiles can function as universal molecular descriptors for machine learning workflows. Their studies demonstrate that models trained exclusively on sigma profiles are capable of predicting diverse physicochemical properties such as boiling point, vapor pressure, density, and solubility, with excellent accuracy. Because sigma profiles compactly summarize rich electronic information while maintaining fixed dimensionality, they reduce the need for large datasets, simplify model architectures, and avoid the scaling issues typical of size-dependent descriptors.

To obtain sigma profiles in this work, each target ion was subjected to a DFT calculation in the software package TURBOMOLE [22], using the def-TZVP basis set and the BP-86 functional, under the continuum-solvent model COSMO. This DFT calculation involves the geometry optimization of the ion in the COSMO solvation environment (with infinite permittivity), yielding the so-called sigma surface, a tessellated molecular surface with associated screened charge (surface charge density) on each surface patch. These calculations were complemented with basic structures from crystallographic data, ensuring accuracy through the comparison of experimental measurements. The basis set and functional chosen are the defaults used in COSMO-related calculations.

Sigma surfaces are processed into sigma profiles by binning the surface patches according to their local charge densities. The result is an unnormalized histogram (or distribution) of surface area vs. charge density across the molecular surface. In other words, sigma profiles indicate how much surface area of the molecule has a given range of screened charge. In this work, the conversion of sigma surfaces to sigma profiles was carried out using the software package COSMOtherm [23].

2.2 *Gaussian Process* (GP)

A GP is a set of multivariate normal distributions capable of predicting values of an unknown function, being used as a nonparametric probabilistic model. A GP assumes that the values of the function follow a joint normal distribution and that the relationship between the points is described by a covariance function $k\left(x, x'\right)$ (kernel) and a mean function *M(x)*.

$$f(x) \sim GP\left(M(x), k\left(x, x'\right)\right) \tag{1}$$

That is, the values of the $f(x)$ function are treated as random variables with joint normal distribution. For a set of N entry points, the joint distribution of the function values can be written as:

$$\begin{bmatrix} f_1 \\ \vdots \\ f_N \end{bmatrix} \sim N(\mu, \Sigma) = N\left(\begin{bmatrix} \mu_1 \\ \vdots \\ \mu_N \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \cdots & \Sigma_{1N} \\ \vdots & \ddots & \vdots \\ \Sigma_{N1} & \cdots & \Sigma_{NN} \end{bmatrix}\right) \tag{2}$$

$$\mu_i = M\left(x_i\right) \tag{3}$$

$$\Sigma_{ij} = k\left(x_i, x_j\right) \tag{4}$$

Therefore, μ represents the vector of means and Σ is the covariance matrix constructed from the kernel between all pairs of points. When the GP is conditioned to the known data of the function $y(x)$, whether these are represented by y (training set), and the unknown values to be estimated represented by f_* (test set at the points), it is possible to establish a joint normal distribution according to the following equation:

$$\begin{bmatrix} y \\ f_* \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_y \\ \mu_* \end{bmatrix}, \begin{bmatrix} \Sigma_y & \Sigma_{y*} \\ \Sigma_{y*}^T & \Sigma_* \end{bmatrix}\right) \tag{5}$$

Therefore, for a given test point, the GP allows the predictive distribution of f_*, conditioned to the observed data y. This distribution is also Gaussian, with mean and variance updated according to the data.

$$f_* | y \sim N\left(\mu', \Sigma'\right) \tag{6}$$

This equation represents the prediction made by the GP, where μ' is the predicted mean and Σ' is the associated uncertainty. This characteristic makes Gaussian Processes especially valuable in problems where the estimation of uncertainty is crucial for a detailed analysis.

The Python GPflow package (version 2.5.2) was the fundamental tool used to perform all GP-related calculations. The input (features) of the GP model trained in this work are the sigma profiles of each salt (SP_{salt}), obtained by summing the sigma profiles of each individual ion of the salt (SP_{cation} and SP_{anion}), followed by multiplying by its mole fraction in water (x_{salt}):

$$SP_{salt} = x_{salt}\left(SP_{cation} + SP_{anion}\right) \tag{7}$$

Because the sigma profiles obtained in this work range from − 0.062 e/Å^2 to 0.062 e/Å^2, in intervals of 0.001 e/Å^2, SP_{salt} represents a vector of size 125. The output (labels) of the GP model are the relative solubility of glycine in the specific water/salt solution. Here, relative solubility means the ratio between the solubility of glycine in the water/salt solution and the solubility of glycine in pure water.

The variables were normalized using methods based on logarithmic transformations followed by standardization (normalization). For the independent variables, the Log + bStand transformation was applied, while for the dependent variable, the LogStand transformation was applied. In both cases, the resulting variables were rescaled to a range close to the standardized range with zero mean and unit standard deviation. In the GP model, the zero-mean function was implemented, while the variance of the Gaussian probability function (likelihood) was initially set at a reduced value (10^{-3}). The kernel parameters were adjusted by maximizing the marginal log-likelihood using the L-BFGS-B algorithm. In addition, a White kernel was used, which adds an overall estimate of the uncertainty associated with the data.

The datasets and Python code used in this work are freely available in the following GitHub repository: https://github.com/dinisAbranches/Glycine_GPs.

3 Results and Discussion

To enable a thorough analysis of glycine solubility in the presence of different electrolytes, the database was developed to cover the widest possible variety of salts. This diversity facilitates a more comprehensive analysis about the influence of both cations and anions on solubility trends. The dataset encompasses a wide array of anions (F^-, Cl^-, Br^-, I^-, NO_3^-, SCN^-, CH_3COO^-, and SO_4^{2-}) and cations (Na^+, K^+, NH_4^+, Ba^{2+}, Ca^{2+}, and Mg^{2+}). This ensures the analysis is broad, allowing for detailed insights into the chemical interactions governing the solubility of amino acids in saline environments.

3.1 Dataset

Our dataset, available in the repository associated with this work, contains the solubility of glycine in aqueous solutions of different salts, as a function of salt concentration. The dataset contains a total of 262 entries. With the database systematically organized, the distribution of collected data for each salt investigated can be clearly visualized in Fig. 1. It presents the amount of data available for each salt and an overview of the number of different datasets within a given salt. The dataset also contains information on the reported experimental uncertainty of each data point, which is used to produce experimental error bars throughout this manuscript.

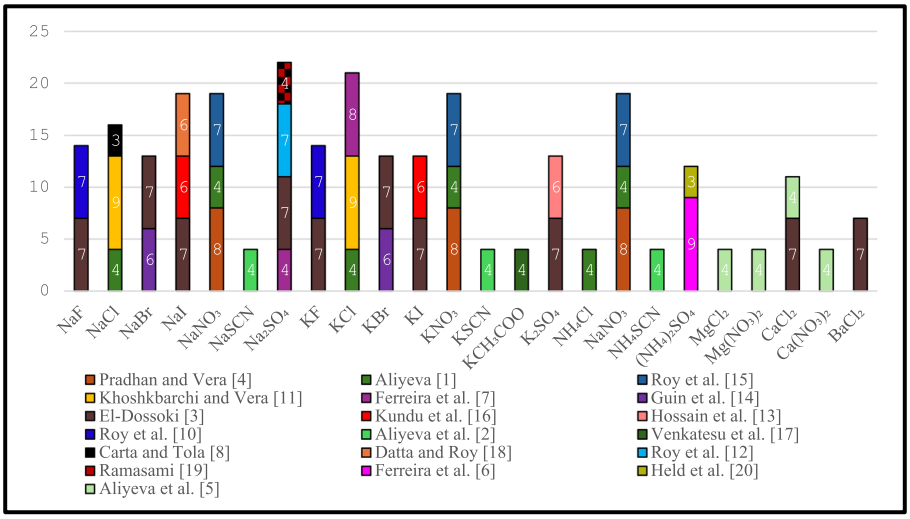

Fig. 1 Experimental data points (numbers) and datasets (different colors) collected for the salt effect on the solubility of glycine in aqueous solutions at 298.2 K

For the construction of the GP model, it was necessary to calculate the sigma profiles of the ions included in the database. These sigma profiles, obtained as explained in Sect. 2.1, allowed us to define the molecular structures of the salts and to distinguish each system within the model. By considering the sigma profiles of both cations and anions, the GP can account for ion-specific effects, giving a more faithful description of the different physicochemical behaviors associated with each salt. This approach also supports more reliable predictions of the systems under study. Figure 2 presents the calculated sigma profiles for the cations (top) and anions (bottom) that constitute the investigated salts.

To provide a comprehensive overview, a global evaluation of the model's performance was conducted. Specifically, as illustrated in Fig. 3, the experimental relative solubility values (S/S_0) are compared with the corresponding predictions by the GP model, which indicates a good predictive capacity of the model. In this case, most of the data points cluster around the diagonal line, indicating a satisfactory performance of the GP model. This statement is supported by the coefficient of determination ($R^2 = 0.750$), which confirms that the model captures the experimental data trends with reasonable accuracy, even considering the diversity of salts included in the database.

As can be seen in Fig. 4, a detailed analysis of the data reveals that, for salt systems with data reported by multiple authors, there is a lack of consensus regarding their effect on glycine solubility. This variability underscores the importance of applying the GP model to evaluate and validate the data. Such an approach enables the identification of inconsistencies and contributes to enhancing the overall reliability of the dataset.

For the aqueous solution containing KNO_3 (Fig. 4A) or $NaNO_3$ (Fig. 4C), it is observed that the results reported by Pradhan and Vera [4] present relative solubility values significantly higher than those of the other authors. This discrepancy suggests that methodological differences or specific experimental conditions may have influenced the results, causing them to deviate from the general trend observed in the other datasets.

For the KCl and NaCl systems (Figs. 4B and D), Khoshkbarchi and Vera [11] report a salting-out effect at low salt mole fractions for KCl, whereas Ferreira et al. [7] and Aliyeva

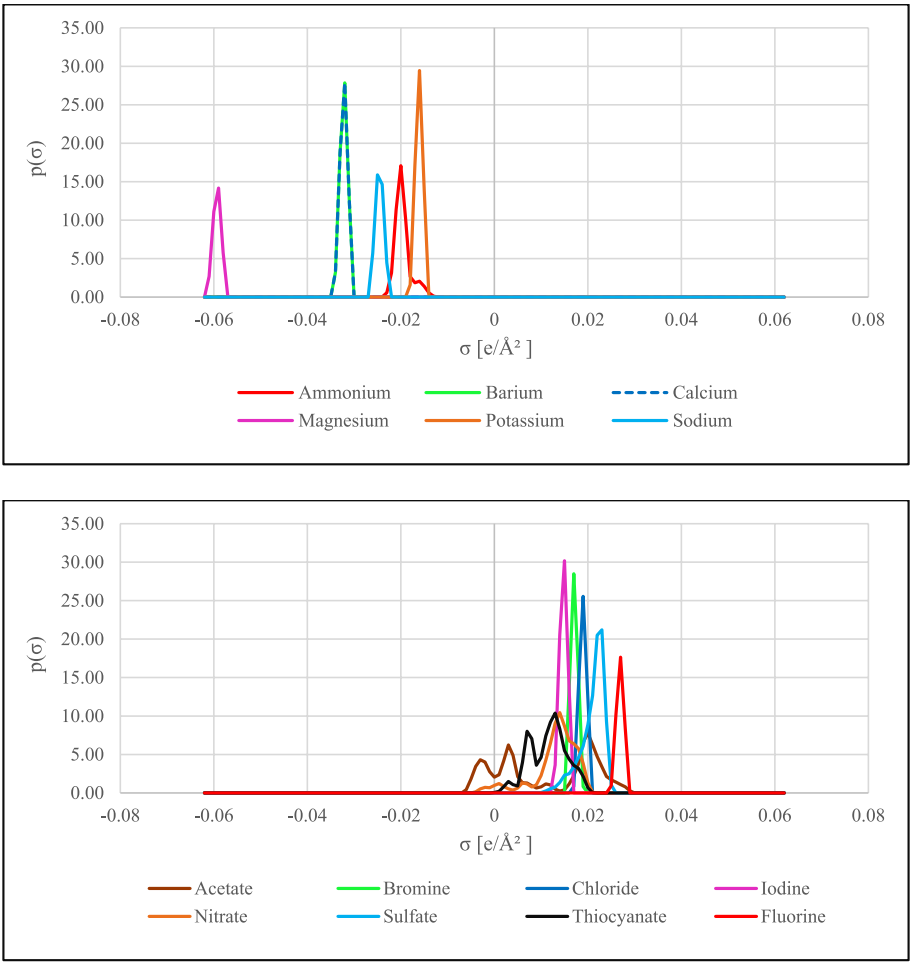

Fig. 2 Calculated sigma profiles for cations (top) and anions (bottom) forming the salts of the studied database, obtained from COSMO-RS surface charge density distributions

[1] report the opposite salting-in effect. For NaCl, Carta and Tola [8] present only two values that deviate from the trend shown by the other authors.

In the case of Na_2SO_4 (Fig. 4E), there is a good general agreement among most authors, except for El-Dossoki's data [3], which present higher values of relative solubility throughout the concentration range. As a last example, for $CaCl_2$ (Fig. 4F), Aliyeva et al. [5] report a clear, near-linear increase in relative solubility with salt mole fraction, whereas El-Dossoki [3] observes only a slight increase across the measured range.

The combined global and individual analyses of the experimental data enabled the identification of both consistent patterns and discrepancies among datasets for each salt studied. Understanding these variations is crucial for the proper curation of the database and for the development of reliable predictive models. This approach reinforces the usefulness of the GP as a tool for assessing the reliability of experimental results and revealing potential methodological biases or variations in measurement conditions. It further emphasizes the

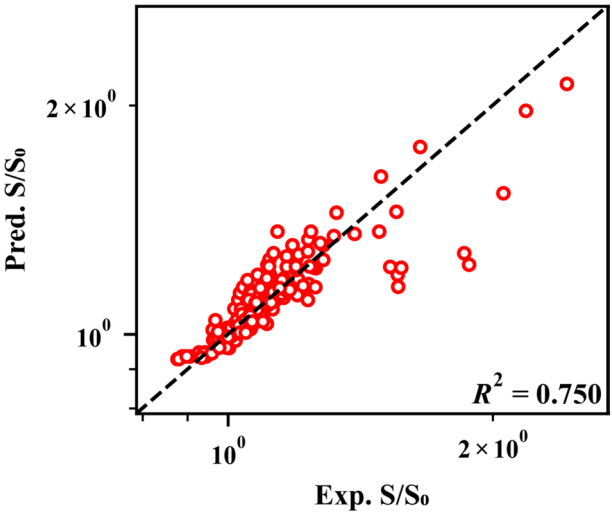

Fig. 3 Comparison between experimental (Exp. S/S_0) and predicted (Pred. S/S_0) relative solubility values for glycine in salt aqueous solutions at 298.2 K, using the GP model

importance of implementing models capable of identifying which data are closest to the expected behavior.

3.2 GP Regression

The development of the GP model implies establishing a nonlinear and probabilistic relationship between the physicochemical description of the system and the relative solubility of glycine. To this end, the input independent variables were defined as the product between the mole fraction of the salt and its sigma profile, a parameter that describes the distribution of molecular charge density. The relative solubility of glycine was used as the dependent variable.

The output variable of the model corresponds to the relative solubility predicted by the GP model. As the data are sequentially introduced from the database, the model dynamically updates its predictions, estimating the expected values for each combination of mole fraction and sigma profile. In addition to point predictions, the GP model provides a confidence interval for each estimate, allowing not only to assess global and individual trends, but also to assess the uncertainty associated with the results quantitatively. This feature is fundamental for interpreting the results, as it facilitates direct comparison between the GP predictions and the experimental values.

An analysis of the model output for the KNO_3 containing system (Fig. 5A) reveals that the prediction curve (red line) captures the overall trend of the experimental data, illustrating an increase in the relative solubility of glycine with increasing mole fraction of the salt. However, the shaded region, which represents the model confidence interval, indicates that a substantial portion of the experimental points lie outside this interval, with only a limited number falling within the predicted uncertainty bounds and discarding most of the data. The results reported by Pradhan and Vera [4] predominantly fall outside, presenting the larger distances to the uncertainty range predicted by the GP model, suggesting that these measurements are inconsistent with the expected pattern for the relative solubility of glycine in this system. A similar situation occurs in the systems containing $NaNO_3$ (Fig. 5C).

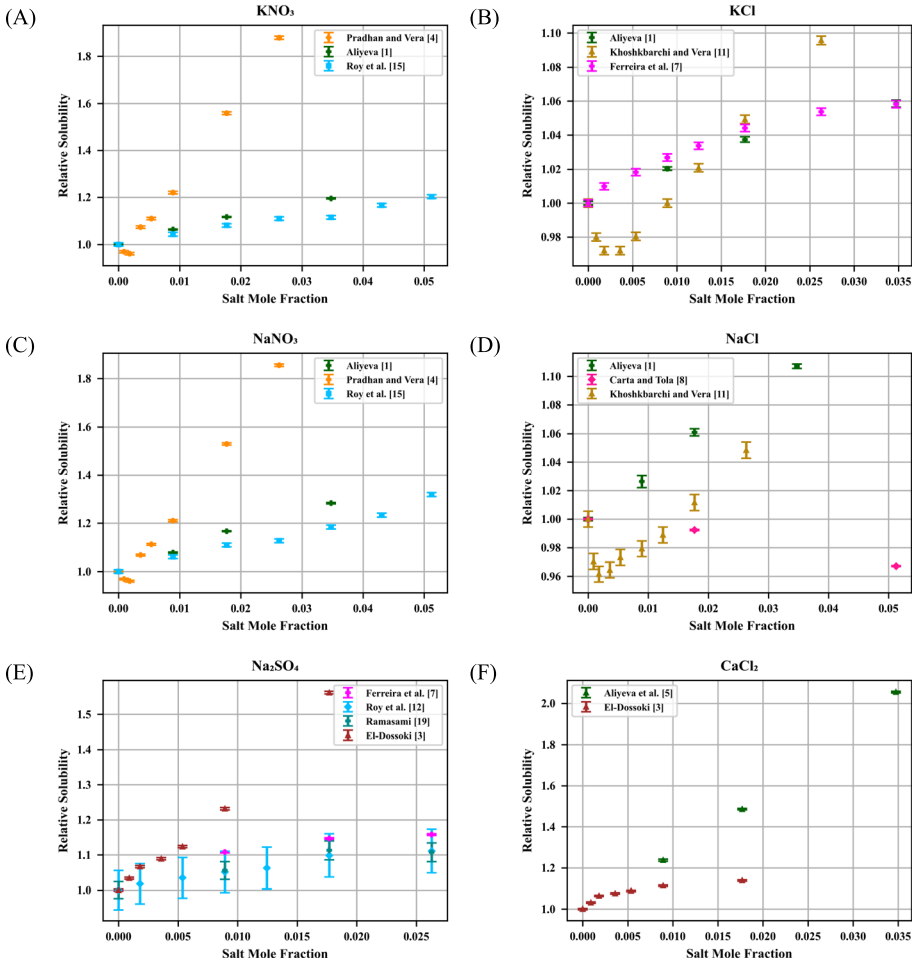

Fig. 4 Experimental relative solubility of glycine in KNO_3 (A), KCl (B), $NaNO_3$ (C), NaCl (D), Na_2SO_4 (**E**), $CaCl_2$ (**F**) aqueous solutions at 298.2 K, from different authors

In stark contrast, the systems containing KCl and NaCl (Figs. 5B and D) exhibit markedly different behavior from the nitrates. One of the key findings in these cases is the presence of distinct salting-in and salting-out effects in the dilute solutions, which introduces significant variability into the data. Despite the relative proximity of the experimental values, this variability leads to broader uncertainty intervals in the GP predictions, wide enough that most data points fall within the model's confidence limits. Specifically, for the NaCl system, the data reported by Aliyeva [1] and Carta and Tola [8] include several values, particularly at higher mole fractions, that fall outside the model's predicted uncertainty range. This suggests that portions of these datasets may be inconsistent with the expected solubility behavior. Moreover, none of the datasets examined align well with the trend predicted by the GP model, reinforcing the need for a more critical assessment of the experimental methodologies and conditions used in this system.

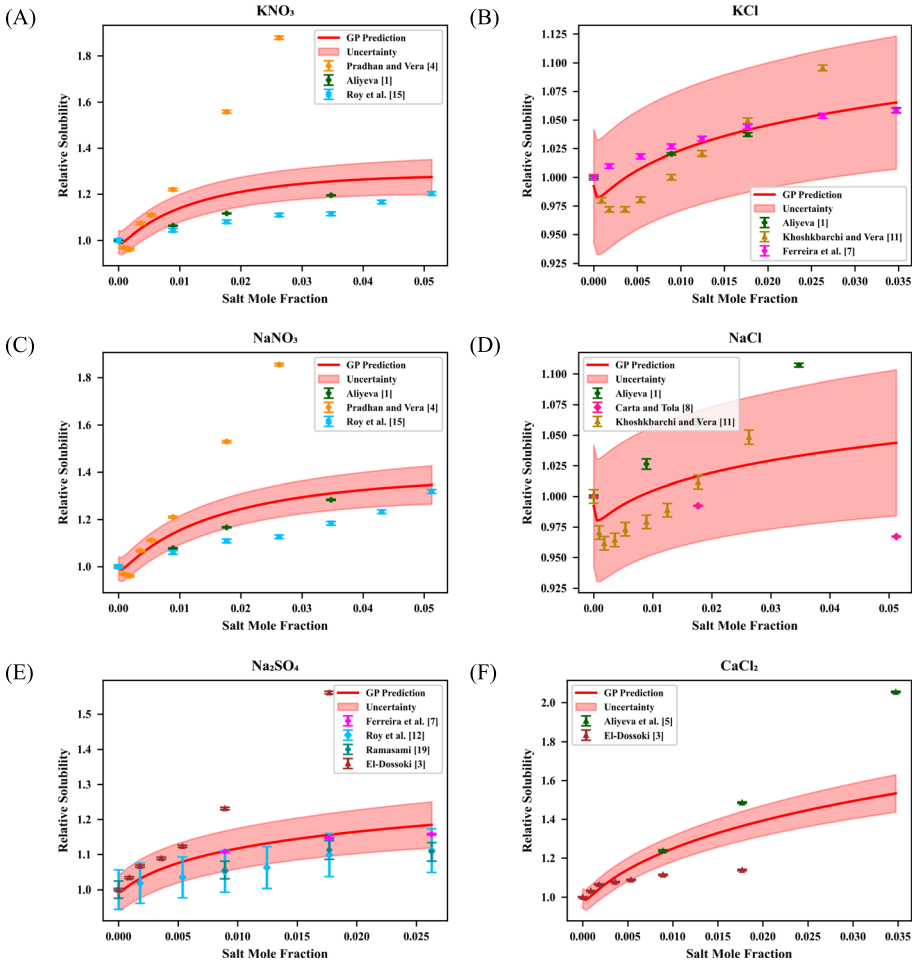

Fig. 5 GP model prediction (red line) for the relative solubility of glycine in KNO_3 (**A**), KCl (**B**), $NaNO_3$ (**C**), NaCl (**D**), Na_2SO_4 (**E**), $CaCl_2$ (**F**) aqueous solutions at 298.2 K, and the associated prediction uncertainty

For the Na_2SO_4 containing system (Fig. 5E), the model predicts a moderate and continuous increase in the relative solubility of glycine with increasing salt mole fraction. The uncertainty region encompasses many data points, as three independent studies report values within a similar range. However, the dataset from El-Dossoki [12] reports significantly higher values than the others, particularly at higher salt concentrations. This discrepancy contributes to an expanded uncertainty interval and results in a more noticeable deviation between the model prediction and the remaining experimental data at higher mole fractions. Among the sets evaluated, only the data from Ferreira et al. [7] are fully within the model's uncertainty region, being most consistent with the predicted trend.

As a final remark, for the $CaCl_2$ system (Fig. 5F), the GP model predicts a smooth increase in the relative solubility of glycine with rising salt mole fraction. The uncertainty band covers most measurements at low to intermediate mole fractions, particularly those from El-Dossoki [3], which cluster around the prediction. In contrast, the values reported

by Aliyeva et al. [5] tend to lie above the trend, and the highest concentration point clearly falls outside the uncertainty region, yielding the largest deviation. The model captures the global tendency, but the high-concentration behavior remains weakly constrained due to sparse and inconsistent data within the sampled range. For this system, El-Dossoki's measurements are the most consistent with the predicted trend.

Overall, the predictions of the GP model effectively captured the trends in the relative solubility of glycine across the various saline systems, even in cases where discrepancies existed among experimental datasets. The model showed sensitivity to the individual contributions of the constituent ions of each salt, reflecting their influence on solubility behavior. This allowed for the identification of consistent patterns among systems containing common ions, further reinforcing the model's capacity to reveal underlying physicochemical relationships. Considering the other systems in the database, a similar analysis can be carried out by observing Section I: Figures in the supporting information.

Analysis of the results indicates that the model adapts its predictive behavior to align with the observed trends. However, a further refinement is needed to improve its ability to distinguish which data points are most consistent with the predicted values. In addition, the assessment of predictive uncertainties helps identify regions of greater reliability, guiding the selection of new experimental points to determine and improve the model, thereby increasing the accuracy of the predictions.

3.3 GP Meta-Analysis

The implementation of the data-filtering strategy, removing the experimental points that are outside the uncertainty bounds of the predictions, emerges as a promising approach to improve the GP model performance. This method reduces the impact of potentially inconsistent data, enabling a more accurate and robust fit. By delimiting the data that present better agreement with the predicted behavior, the model can refine its parameters more effectively, thereby improving both its predictive accuracy and the overall reliability of the results. Figure 6 illustrates the workflow proposed for this modeling approach.

The workflow follows a simple, iterative cycle: first, the model is trained with the initial dataset; next, the predictions of relative solubility and the respective uncertainty are computed for each salt; then, these predictions are compared with the corresponding experimental measurements at each salt mole fraction; finally, any points falling clearly outside the uncertainty range of the model are excluded from subsequent training iterations. The procedure is applied to all systems in a synchronized manner: at each iteration, all inconsistent points are removed from each system, ensuring balance and preventing favoritism. In this way, outliers from one system do not affect the others.

It is important to note that the filtering procedure was applied using a 99.5% confidence interval. No experimental data points are permanently removed from the database. Instead, the exclusions are made only in a copy of the dataset used for model predictions. In the Figures, points excluded from new GP predictions and calculations are marked with a red circle, clearly indicating that they were filtered out. This iterative filtering process contributes to greater model stability, improved predictive accuracy, and increased confidence in the results, while ensuring that the data selection remains traceable and scientifically robust.

By applying the iterative filtering method, the GP model was able to identify inconsistent points within the dataset. A detailed account of the number of excluded data points for each system is provided in the Supporting Information (Table S1).

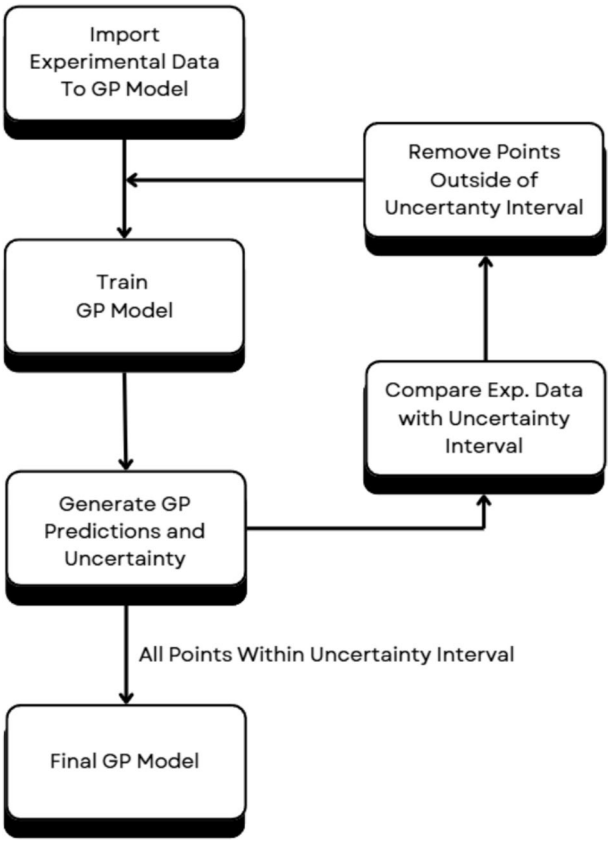

Fig. 6 A data-filtering workflow for training GPs to describe the relative solubility of glycine in different salt solutions

Furthermore, Fig. 7 illustrates the systems after the application of the filtering procedure, highlighting the improvement in model consistency once the inconsistent data points are removed from the GP predictions.

When analyzing the behavior of the model for the KNO_3 (Fig. 7A) and $NaNO_3$ (Fig. 7C) systems, it is observed that, after applying the iterative filtering procedure, the results of Roy et al. [15] consistently showed the highest agreement with the GP predictions. In both systems, the narrowing of the uncertainty interval reinforces the quality of the fit and the high predictive confidence of the model. For Pradhan and Vera [4], the results at lower mole fractions were considered consistent with the model, while the higher values were excluded during the filtering process. The data of Aliyeva [1] were also found to be consistent overall, with only the final value for the $NaNO_3$ system being disregarded after the iterative filtering.

In the evaluation of the GP model applied to the KCl (Fig. 7B) and NaCl (Fig. 7D) systems, it is observed that most of the experimental data reported by different authors are consistent with the predictions. For KCl, the model tends to follow more closely the results of Khoshkbarchi and Vera [11] at lower mole fractions, while the data of Ferreira et al. [7] become more consistent as the mole fraction increases. In contrast, for NaCl, the predictions are predominantly aligned with Khoshkbarchi and Vera [11], and only the final points

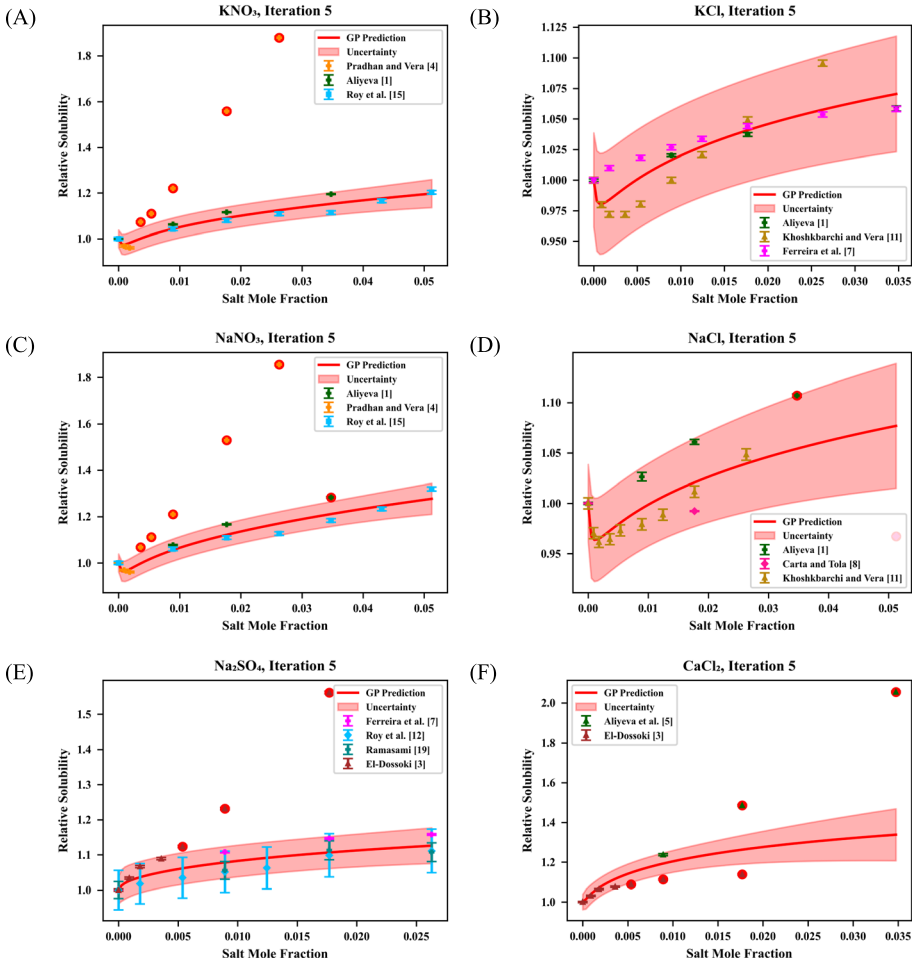

Fig. 7 GP prediction of glycine relative solubility in KNO_3 (**A**), KCl (**B**), $NaNO_3$ (**C**), NaCl (**D**), Na_2SO_4 (**E**), $CaCl_2$ (**F**) aqueous solutions at 298.2 K, with uncertainty bands after iterative removal of points outside the uncertainty interval

of Carta and Tola [8] and Aliyeva [1] were excluded by the uncertainty criterion, leading to greater dispersion of the model and broader uncertainty at higher concentrations.

For the Na_2SO_4 system (Fig. 7E), the results of El-Dossoki [3] showed some inconsistencies at higher mole fractions, whereas the data reported by Ferreira et al. [7], Ramasami [19], and Roy et al. [12] remained consistent with the predictions of the GP model across the evaluated range. However, among these authors, it is not evident which dataset aligns most closely with the model predictions.

Finally, for the $CaCl_2$ system (Fig. 7F), the GP predictions indicate that the expected effect is a slight increase in relative solubility with increasing mole fraction. This outcome contrasts with the experimental results of both El-Dossoki [3] and Aliyeva et al. [5], whose reported trends at higher concentrations are inconsistent with the model, leading to the removal of these points during filtering.

After applying the iterative filtering of inconsistent data, the uncertainty regions are significantly reduced, and the datasets reported by different authors become more coherent with each other. This process allows the model to emphasize which points are truly consistent, resulting in greater stability and improved predictive accuracy. Figure 8 illustrates this step-by-step filtering procedure for the KNO_3 system, showing how the progressive exclusion of inconsistent values strengthens the reliability of the model outcomes.

From Fig. 8A, B, corresponding to the first iteration, the GP model already shows a noticeable reduction in the uncertainty band. Two points from Pradhan and Vera [4], located outside the confidence interval, are removed, which increases the precision of the prediction. At this stage, the datasets of Aliyeva [1] and Roy et al. [15] become clearly aligned with the model trend, reinforcing their consistency. In the subsequent steps (Figs. 8C and D), corresponding to iterations 2 and 3, additional points from Pradhan and Vera [4] are progressively removed. This indicates that, for the most part, their reported

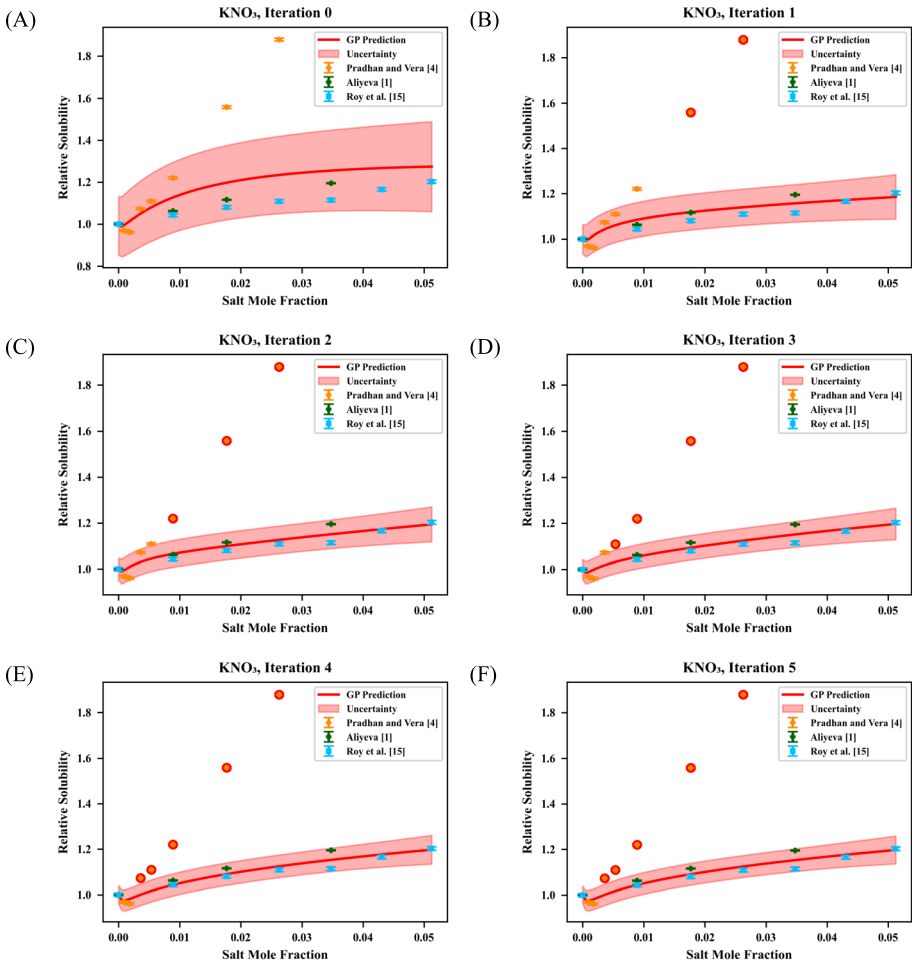

Fig. 8 Stepwise iterative filtering of inconsistent experimental data in the KNO_3 aqueous system at 298.2 K, using the GP model (confidence interval of 99.5%)

results diverge from the predicted behavior. Meanwhile, the uncertainty interval gradually adjusts around the datasets of Aliyeva [1] and Roy et al. [15], as well as the low mole fraction values of Pradhan and Vera [4], which remain consistent with the GP predictions. During the final stages (Figs. 8E and F), only minor refinements occur, with a slight narrowing of the uncertainty band. This confirms the stability of the model and the consistency of the results reported by Aliyeva [1] and Roy et al. [15]. By the fifth iteration, all inconsistent points have been filtered out based on the 99.5% confidence criterion, leaving only the data that best represent the expected solubility trend for glycine in the KNO_3 system.

Therefore, we can conclude that the application of the iterative method of filtering by the uncertainty area proved to be effective in making the GP model more reliable and accurate. By removing the datapoints lying outside the bands, it reduces the influence of inconsistent results between authors and stabilizes the model adjustment process, allowing the GP to clearly identify the data set that is most consistent with their predictions, resulting in a model with minimal uncertainty areas and improved prediction. In addition, data filtering highlights the need to verify potentially inconsistent data and guides new studies in regions of greater uncertainty.

4 Conclusions

This research demonstrated that, even when starting from a database with inconsistencies and contradictory results between different saline solutions, the GP model allowed us to quantify and interpret the specific influence of cations and anions on the relative solubility of glycine in aqueous solutions. Despite these adversities, the GP adequately reproduced the main experimental behaviors and identified which datasets are most consistent.

The proposed meta-analysis, based on iterative filtering of points outside the uncertainty area of the model itself, proved to be crucial to improve the predictive capacity and identify inconsistencies. This procedure narrowed the area of uncertainty and highlighted the data most compatible with the model, without definitively removing the results from the database. As a result, the GP has become more stable, predictive, and reliable to support conclusions and guide new studies.

It is therefore concluded that the GP, when properly modeled and continuously improved by methods sensitive to uncertainty and data quality, helps the validation of different databases, even in the presence of inconsistencies. The model quantifies the confidence in measurements, anticipates the different effects on the systems, and accurately locates inconsistent data and poorly sampled regions. Our GP-based methodology also offers practical guidance for future experimental work. By quantifying predictive uncertainty across salts and compositions, the model highlights regions where existing solubility data are sparse, inconsistent, or otherwise unreliable. These high-uncertainty areas represent the most valuable targets for new measurements, enabling experimental efforts to focus on conditions where additional data would substantially improve confidence in the solubility surface. Conversely, regions with low uncertainty correspond to well-supported measurements that are unlikely to benefit from further experimentation. Thus, the approach provides a systematic way to prioritize remeasurement and to allocate experimental resources more efficiently.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10953-026-01561-9.

Acknowledgements This work was developed within the scope of the project CIMO-Centro de Investigação de Montanha, UIDB/00690/2020 (https://doi.org/10.54499/UIDB/00690/2020), UIDP/00690/2020 (https://doi.org/10.54499/UIDP/00690/2020); and SusTEC, LA/P/0007/2020 (https://doi.org/10.54499/LA/P/0007/2020), and CICECO-Aveiro Institute of Materials, UIDB/50011/2020 (https://doi.org/10.54499/UIDB/50011/2020), UIDP/50011/2020 (https://doi.org/10.54499/UIDP/50011/2020) and LA/P/0006/2020 (https://doi.org/10.54499/LA/P/0006/2020), all financed by national funds through the FCT/MCTES (PIDDAC). The financial support from IUPAC Project No. 2022-002-2-500 is highly acknowledged.

Author Contributions Christopher A. Piske: Investigation, Methodology, Formal Analysis, Data Curation, Writing – Original Draft. Priscilla G. Leite: Writing – Review and Editing. Mónia A. R. Martins: Resources, Writing – Review and Editing. Olga Ferreira: Resources, Writing – Review and Editing. João A. P. Coutinho: Formal Analysis, Writing – Review and Editing. Dinis O. Abranches: Conceptualization, Investigation, Methodology, Formal Analysis, Writing – Review and Editing, Supervision. Simão P. Pinho: Conceptualization, Investigation, Methodology, Formal Analysis, Writing – Review and Editing, Supervision.

Funding Open access funding provided by FCT|FCCN (b-on).

Data Availability The datasets and Python code used in this work are freely available in the following GitHub repository: https://github.com/dinisAbranches/Glycine_GPs.

Declarations

Competing Interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Aliyeva, M.: Ion effects on protein model compounds in aqueous systems: experimental and computational studies. Universidade de Aveiro, Aveiro (2022)
2. Aliyeva, M., Brandão, P., Coutinho, J.A.P., Ferreira, O., Pinho, S.P.: Solubilities of amino acids in the presence of chaotropic anions. J. Sol. Chem. **53**(4), 527–537 (2024). https://doi.org/10.1007/s10953-023-01282-3
3. El-Dossoki, F.I.: Effect of the charge and the nature of both cations and anions on the solubility of zwitterionic amino acids, measurements and modeling. J. Sol. Chem. **39**(9), 1311–1326 (2010). https://doi.org/10.1007/s10953-010-9580-3
4. Pradhan, A.A., Vera, J.H.: Effect of anions on the solubility of zwitterionic amino acids. J. Chem. Eng. Data **45**(1), 140–143 (2000). https://doi.org/10.1021/je9902342
5. Aliyeva, M., Brandão, P., Gomes, J.R.B., Coutinho, J.A.P., Ferreira, O., Pinho, S.P.: Solubilities of amino acids in aqueous solutions of chloride or nitrate salts of divalent (Mg^{2+} or Ca^{2+}) cations. J. Chem. Eng. Data **67**(6), 1565–1572 (2022). https://doi.org/10.1021/acs.jced.2c00148
6. Ferreira, L.A., Macedo, E.A., Pinho, S.P.: The effect of ammonium sulfate on the solubility of amino acids in water at (298.15 and 323.15) K. J. Chem. Thermodyn. **41**(2), 193–196 (2009). https://doi.org/10.1016/j.jct.2008.09.019
7. Ferreira, L.A., Macedo, E.A., Pinho, S.P.: Effect of KCl and Na_2SO_4 on the solubility of glycine and DL-alanine in water at 298.15 K. Ind. Eng. Chem. Res. **44**(23), 8892–8898 (2005). https://doi.org/10.1021/ie050613q

8. Carta, R., Tola, G.: Solubilities of L-cystine, L-tyrosine, L-leucine, and glycine in aqueous solutions at various pHs and NaCl Concentrations. Ind. Eng. Chem. Res. **41**, 414–417 (1996). https://doi.org/10.1021/je9501853
9. Abranches, D.O., Maginn, E.J., Colón, Y.J.: Activity coefficient acquisition with thermodynamics-informed active learning for phase diagram construction. AIChE J. **69**, 8 (2023). https://doi.org/10.1002/aic.18141
10. Roy, S., Guin, P., Mahali, K., Dolui, B.: Solubility and transfer Gibbs free energetics of glycine, DL-alanine, DL-nor-valine and DL-serine in aqueous sodium fluoride and potassium fluoride solutions at 298.15 K. Ind J of Chem **56**, 399–406 (2017)
11. Khoshkbarchi, M.K., Vera, J.H.: Effect of NaCl and KCl on the Solubility of amino acids in aqueous solutions at 298.2 K: measurements and modeling. Ind. Eng. Chem. Res. **36**, 2445–2451 (1997). https://doi.org/10.1021/ie9606395
12. Roy, S., Guin, P.S., Mahali, K., Hossain, A., Dolui, B.K.: Evaluation and correlation of solubility and solvation thermodynamics of glycine, DL-alanine and DL-valine in aqueous sodium sulphate solutions at two different temperatures. J. Mol. Liq. **234**, 124–128 (2017). https://doi.org/10.1016/j.molliq.2017.03.068
13. Hossain, A., Mahali, K., Dolui, B.K., Guin, P.S., Roy, S.: Solubility analysis of homologous series of amino acids and solvation energetics in aqueous potassium sulfate solution. Heliyon **5**, 8 (2019). https://doi.org/10.1016/j.heliyon.2019.e02304
14. Guin, P.S., Mahali, K., Dolui, B.K., Roy, S.: Solubility and thermodynamics of solute-solvent interactions of some amino acids in aqueous sodium bromide and potassium bromide solutions. J. Chem. Eng. Data **63**(3), 534–541 (2018). https://doi.org/10.1021/acs.jced.7b00647
15. Roy, S., Guin, P.S., Mondal, S., Ghosh, S., Dolui, B.K.: Solubility of glycine and DL-nor-valine in aqueous solutions of $NaNO_3$ and KNO_3 and measurements of transfer thermodynamics. J. Mol. Liq. **222**, 313–319 (2016). https://doi.org/10.1016/j.molliq.2016.07.050
16. Kundu, S., Mahali, K., Roy, S.: Solvation thermodynamics of four amino acids in electrolytic solutions of sodium and potassium iodide salts at 298.15 K. Can. J. Chem. **101**(4), 224–234 (2023). https://doi.org/10.1139/cjc-2022-0251
17. Venkatesu, P., Lee, M.J., Lin, H.M.: Transfer free energies of peptide backbone unit from water to aqueous electrolyte solutions at 298.15 K. Biochem. Eng. J. **32**(3), 157–170 (2006). https://doi.org/10.1016/j.bej.2006.09.015
18. Datta, A., Roy, S.: Thermodynamics of solute-solvent interactions and solubility of some amino acids in aqueous sodium iodide solutions at T = 298.15 K. Russ. J. Phys. Chem. A **95**, S62–S70 (2021). https://doi.org/10.1134/S0036024421140041
19. Ramasami, P.: Solubilities of amino acids in water and aqueous sodium sulfate and related apparent transfer properties. J. Chem. Eng. Data **47**(5), 1164–1166 (2002). https://doi.org/10.1021/je025503u
20. Held, C., Reschke, T., Müller, R., Kunz, W., Sadowski, G.: Measuring and modeling aqueous electrolyte/amino-acid solutions with ePC-SAFT. J. Chem. Thermodyn. **68**, 1–12 (2014). https://doi.org/10.1016/j.jct.2013.08.018
21. Abranches, D.O., Maginn, E.J., Colón, Y.J.: Stochastic machine learning via sigma profiles to build a digital chemical space. Proc. Nat. Acad. Sci. **121**, 31 (2024). https://doi.org/10.1073/pnas.2404676121
22. TURBOMOLE V7.1 2016, A development of University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 1989–2007, TURBOMOLE GmbH, since 2007; available from \\ http://www.turbomole.com."
23. "BIOVIA COSMOtherm, Release 2021," *Dassault Systèmes.* http://www.3ds.com

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.