From Canonical to Unique: Extension of A Lipophilicity Scale of Amino Acids to Non-Standard Residues

Antonio Viayna[1,2*], Paulina Matamoros[3,4], David Blázquez-Ruano[5], William J. Zamora[3,4,6*]

[1] Departament de Nutrició, Ciències de l'Alimentació i Gastronomia, Facultat de Farmàcia i Ciències de l'Alimentació, Universitat de Barcelona (UB), Av. Prat de la Riba 171, 08921 Santa Coloma de Gramenet, Spain.

[2] Institut de Química Teòrica i Computacional (IQTC-UB), Universitat de Barcelona (UB), Barcelona, Spain.

[3] CBio[3] Laboratory, School of Chemistry, University of Costa Rica, San Pedro, San José, Costa Rica.

[4] Laboratory of Computational Toxicology and Biological Testing Laboratory (LEBi), University of Costa Rica, San Pedro, San José, Costa Rica.

[5] CICbioGUNE, Basque Research and Technology Alliance, Derio, Spain.

[6] Advanced Computing Lab (CNCA), National High Technology Center (CeNAT), Pavas, San José, Costa Rica.

* Correspondence: toniviayna@ub.edu (AV), william.zamoraramirez@ucr.ac.cr (WJZ)

ORCID

AV: 0000-0002-2112-5828

PM: 0009-0009-6719-0746

DBR: 0009-0000-8370-2515

WJZ: 0000-0003-4029-4528

Keywords

Partition coefficient, Lipophilicity, Non-natural amino acids, Solvation, Post-translational modifications, Acetyl lysine, Proteomics

Abstract

The lipophilicity of amino acids plays a pivotal role in determining their physicochemical properties as it gives an estimate of solubility, binding propensity, and bioavailability. Herein, we applied the IEFPCM/MST implicit solvation model to compute the n-octanol/water partition coefficient as lipophilic descriptor for non-standard amino acids. Thus, extending our previous work on the hydrophobicity scale of amino acids. To this end, we employed two structural models, named Model 1 and 2, differentiated solely by their C-terminal capping groups using an N- or O- methyl substituent, respectively. Our findings revealed substantial similarities between the models, validating the lipophilicity values for the non-standard side chains. Differences were observed in fewer cases, indicating an effect of the capping group on the side chain hydrophobicity. This effect is expected as one model contains an hydrogen bond donor (Model 1) while the other one uses an hydrogen bond acceptor (Model 2). Overall, both models exhibit strong correlations with the experimental values, with Model 1 showing lower statistical errors. In addition, our predictions were able to correctly predict the experimental

hydrophobicity change due to the number of acetylated lysines in peptide pairs determined by HPLC, suggesting that our scale can be employed for proteomics studies that include post-translational modifications beyond acetylation.

Introduction

Amino acids (AAs) are organic molecules that constitute the basic building blocks of proteins. From a functional point of view, mainly directed by their sequence and 3D arrangement, they play a fundamental role in a multitude of biological processes and functions in living organisms, such as enzymatic catalysis, cell signaling, structural support, or immune response, among others.[1] The 20 canonical AAs are encoded by the genetic code and, structurally, they share some patterns: an amino group, a carboxyl group, and a side chain all attached to the α-carbon. The latter feature is the one that varies according to the amino acid, giving to each one their unique properties.[2]

One of the main properties of proteins is their lipophilicity, which is a fundamental property with a clear impact on biology, pharmacology, and medicinal chemistry and drug discovery.[3,4] In the context of proteins is important for understanding processes such as protein folding, where hydrophobic amino acids tend to cluster in the protein interior away from the aqueous environment. It also influences ligand binding, affecting the binding affinity and specificity of proteins and contributing to the formation of receptor-ligand interactions and also in protein-protein interactions, promoting the formation of protein complexes, among others. In addition, recent studies have created energy functions based on lipophilicity for membrane-protein studies of receptors, channels, and transporters.[5] Given these reasons, it is crucial to have tools that permit the quantification of the degree of hydrophobicity of proteins.

For proteins, lipophilicity is primarily influenced by the specific features of the amino acid side chains. Consequently, one of the main strategies, involves quantifying the individual hydrophobicity of each amino acid, leading to the development of lipophilicity scales. These scales consider various properties such as partitioning of small molecules in a bulk solvent, employing knowledge-based techniques based on structural data and/or using experimental information coming from biological assays.[6-8]

By employing these scales, it is possible to generate lipophilicity profiles of peptides and/or proteins based on the individual hydrophobicity values of residues, assuming an additivity principle. However, depending on the employed scale, variations can occur not only in the absolute magnitude of residues but also in their relative values. These variations pose difficulties in correlating different scales, as well as reflect discrepancies between materials, methods, and experimental conditions that permit the definition of each scale.

In this line, in our previous study,[9] we developed an extensive lipophilicity scale of the 20 standard amino acids based on theoretical computations that took into account the local context of each amino acid in the proteins deposited in the Dunbrak's rotamer library.[10] Thus, this scale incorporated the structural features of the conformational landscape of amino acids, as well as the impact of pH, providing a reliable depiction of pH-adapted lipophilicity profile in peptides and proteins.

However, when we move to non-standard amino acids, derivatives that differ in structure or composition from the 20 standard ones usually found in proteins, set a

challenge to have new adaptations of the classical lipophilicity scales to be reliably standardized to be applied to those biomolecules with non-canonical modifications.

Recent efforts have focused on the impact of the presence of non-canonical amino acids on peptide and protein structure and function. In fact, this new class of amino acids has found an excellent opportunity for use in the design of peptidomimetics. This is mainly because they have been identified naturally and have been found to improve both the stability of the structures and their bioactivity.[11] Concerning structure, it has been shown that the presence of such residues decreases the accuracy of structure prediction tools, so it has been recommended to simulate first using the proteinogenic amino acids and then perform the modification to carry out molecular dynamic's studies.[12] Regarding function, non-canonical amino acids have emerged in the field of synthetic biology, focusing mainly on the research of biomaterials looking for adhesion capabilities, also in the design of antimicrobial peptides improving their protease resistance, solubility, and half-life.[13] Such efforts have led to novel structure/activity studies on modified peptides that present chemoinformatics tools to efficiently characterize the chemical space of these new peptides and thereby better understand their activities, e.g., their antimicrobial activity against multidrug-resistant bacteria.[14]

In the context of the lipophilicity for non-canonical amino acids, prominent examples highlighting the significance and relevance of this topic include recent studies by Kubyshkin (2021) [15] and Oeller *et al*. (2023). [16] The computational work of Oeller *et al*. introduced the CamSol-PTM tool, which offers a rapid and accurate methodology for predicting the solubility of peptides containing non-standard amino acids. Regarding to the experimental work of Kubyshkin, it aimed to develop an experimental lipophilicity

scale incorporating both coded and non-coded amino acids, using the *n*-octanol/water partition coefficient. This work, based on *N*-acetyl and *O*-methyl amino acid analogs, determined the log*P* for this synthetic compounds using the NMR technique, which provides a valuable opportunity to validate computational tools for lipophilicity determination. However, it has time constraints in case of generating new chemical modifications due to the experimental protocol to be implemented. Thus, a computational strategy with adequate accuracy to reproduce these experimental values can alleviate the laborious and time-consuming process of the experimental techniques and can offer the advantage of being able to apply a rapid and straightforward strategy to calculate the lipophilicity upon any modification to create a new non-standard amino acid.

Therefore, the present work aims to expand our previous work on pH-dependent lipophilicity scale of amino acids,[9] specifically the scale that reproduces the behavior of residues in solvent-like environments (SolvL scale), by extending it to a set of non-standard amino acids presented and experimentally measured by Kubyshkin in 2021. The objective is to test, validate, and update our lipophilicity scale to properly account for this descriptor on non-coded amino acids.

Materials and methods

Dataset

In the present article, we selected different non-canonical amino acids (see **Tables S1-S6**) which had been previously investigated and published in an experimental study.[15] The work presented by Kubyshkin focused on examining the experimental lipophilicity of non-standard amino acid derivatives originating from methionine, phenylalanine, tyrosine, tryptophan, lysine, and proline using the n-octanol/water system. In our study,

non-taking into account the standard versions of amino acids, a total of 57 non-canonical amino acids have been investigated. This includes 7 modifications of methionine, 4 of lysine, 9 of phenylalanine, 4 of tyrosine, 25 of proline, and 8 of tryptophan.

Regarding the descriptors measured using the DataWarrior software, [17] within the framework of Model 1, the molecular weights of the studied molecules lie in the range of 168.20 to 338.20 g/mol (see **Figure 1**). The total number of rotatable bonds varies from one to nine (see **Figure 2**). Additionally, the count of hydrogen bond acceptors (HBA) spans from four to seven (see **Figure 3**), while hydrogen bond donors (HBD) range from one to four (see **Figure 4**). In the context of Model 2, which involves substituting the NH group with an oxygen atom in one of the capping groups, there is an approximate one-unit increase in molecular weight. Simultaneously, there is a decrease of one unit in count of HBD, while the number of acceptors remains unchanged. For more detailed information check **Tables S7-S12** at the Supporting Information.

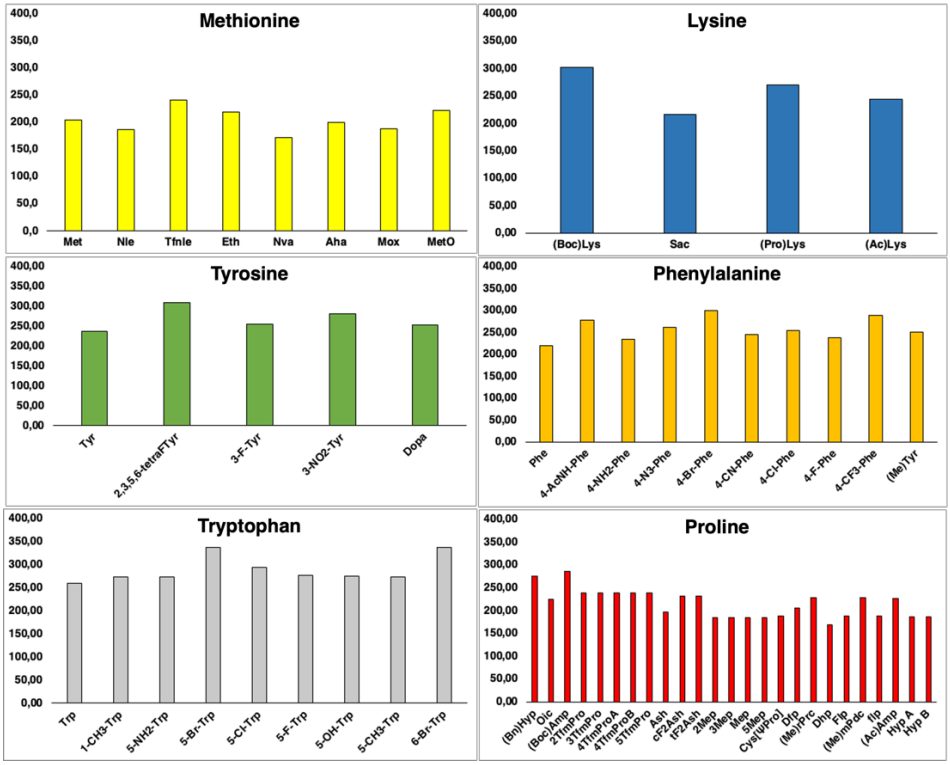

Figure 1. Distribution of values of molecular weight (MW) from the studied compounds regarding Model 1, estimated with DataWarrior.

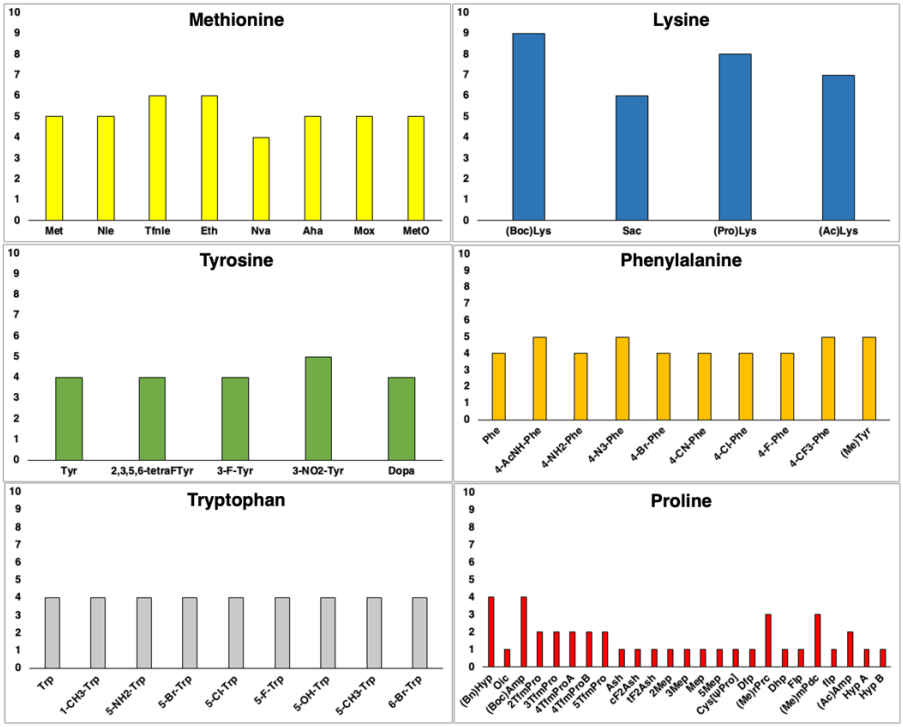

Figure 2. Distribution of values of rotatable bonds (RB) from the studied compounds, regarding Model 1, estimated with DataWarrior.

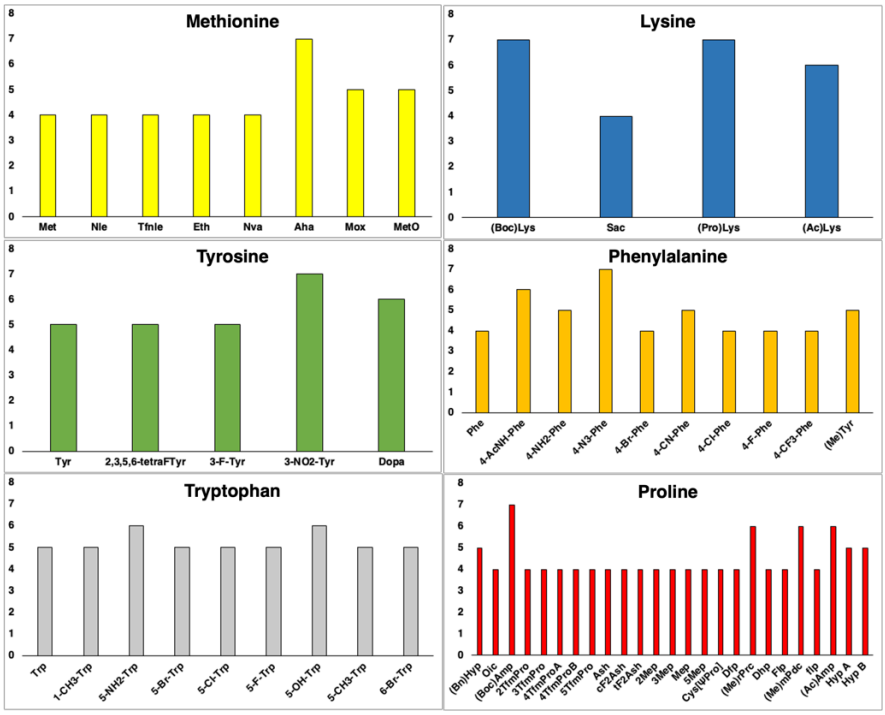

Figure 3. Distribution of values of hydrogen bond acceptors (HBA) from the studied compounds regarding Model 1, estimated with DataWarrior.

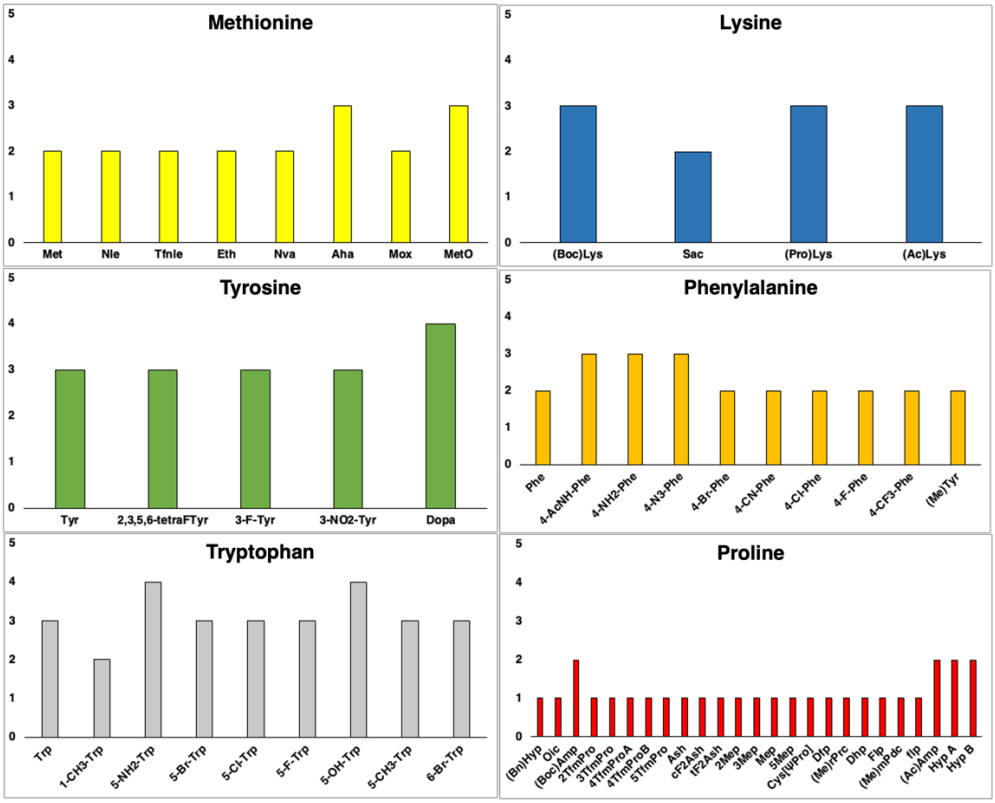

Figure 4. Distribution of values of hydrogen bond donors (HBD) from the studied compounds regarding Model 1, estimated with DataWarrior.

Concerning their lipophilicity, if we consider the difference of each non-standard amino acid compared to the original, based on experimental values reported by Kubyshkin for methionine derivatives, four of them are slightly more lipophilic, and three are more hydrophilic, with variations ranging from plus 0.80 units (most lipophilic) to minus 1.92 (most hydrophilic). In the case of lysine, all derivatives are more hydrophilic than the canonical, with the most marked difference being 2.43 units. Moving to tyrosine, among the four cases, except for a single case (Dopa) that is slightly more hydrophilic, all others

are more lipophilic, despite moving in a narrow range from 0.50 (most hydrophilic) to 0.69 (most lipophilic).

For phenylalanine, five derivatives are more lipophilic, three are more hydrophilic, and one has the same experimental log*P* value. The variation range spans 2.29 units, from the most hydrophilic to the most lipophilic. A similar situation is observed for tryptophan derivatives, where out of the eight cases, only two derivatives with polar groups (5-amino and 5-hydroxy) are more hydrophilic than the standard residue (1.46 units less the most hydrophilic and 1.17 units more the most lipophilic, resulting in a range of 2.63 units). Among the 25 proline cases, except for six instances, all are more lipophilic, with the most lipophilic being 1.59 units greater than standard proline and the most hydrophilic being 0.93 units less lipophilic.

We decided not to include some tyrosine derivatives containing iodine atoms in our study. This decision was based on the limitations of the DFT-based IEFPCM/MST continuum solvation method used for estimating solvation energies, as it lacks parameterization for iodine atoms. However, this method does include parameterization for other halogen atoms like fluorine, chlorine, and bromine, which present minimal differences experimentally when compared to iodine derivatives. Hence, we included molecules containing these three halogen atoms in our study. A similar criterion was taken in the exclusion of selenomethionine from the analysis since the selenium atom is also not included in the IEFPCM/MST current parametrization.

For each molecule, we considered two variants regarding the *N*- and *C*- terminal capping groups. These end fragments are responsible for mimicking the peptide bond which confers rigidity to these regions, as well as, aiming to mimic the physicochemical behavior of the amino acid when present inside a protein, rather than in an individual state. Our study included in parallel both variants for all amino acids, in order to preserve the original capping groups from our previous study, [9] but also to compare with those used by Kubyshkin [15] in his experimental study (see **Figure 5**).

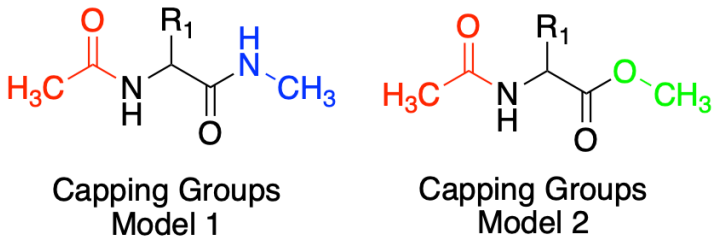

Figure 5. Chemical scaffolds of the capping group models used for the amino acids studied in this article: *N*-methyl (in blue), *O*-methyl (in green), and Acetyl group (in red). Model 1 represents the original capping groups from our previous study [9] while Model 2 are those used by Kubyshkin [15] in his experimental study. R_1 label represents any non-canonical side chain.

Figure 5 shows the first variant, known as "Model 1", which involves the introduction of an *N*-methyl (NME) group at the *N*-terminal end and an acetyl (ACE) group at the *C*-terminal end of the derivatives. "Model 2" uses the capping groups of the experimental data published in Kubyshkin's article, which presents slight modifications. While the C-terminal group remains the same, the *N*-terminal end features an oxy-methyl (OME) group instead of the NME group.

This a priori small change, presumes to have a reasonable impact in the hydrophobicity of the studied compounds. Since the NH to O modification supposes the loss of an hydrogen bond acceptor interaction and translates in an increase of lipophilicity, like the experimental values of N-methylacetamide (logP = -1.05) and Methyl acetate (logP = 0.18) reported by C. Hansch and collaborators in 1995 support. [18]

Conformational studies and log*P* estimations

All molecules were designed using Avogadro software (version 1.1.1). [19] Then, we employed OpenBabel 2.4.0 genetic algorithm to stochastically conduct a preliminary generation of the preferred conformations of the amino acids based on energy score. [20] Due to the structural complexity of some molecules (with a number of rotatable bonds ranging from 1 to 9), we limited the generation of conformers to a maximum of 100 structures, to make a balance between a complete conformational landscape of them, but at the same time deal with an acceptable number of conformers.

Then, generated geometries of the conformers in both water and n-octanol were optimized using the B3LYP/6-31G(d) level of theory. [21-23] The influence of solvent on the geometric parameters was considered by employing the IEFPCM/MST model, [24-26] integrated into a local version of Gaussian16. [27] Minimum energy state of optimized geometries in each solvent was confirmed by inspecting the vibrational frequencies, excluding those conformations presenting negative ones. Afterward, thermal corrections were introduced to estimate the relative free energy of the conformers in water and *n*-octanol. Also, single-point energy calculations were carried out in the gas phase to evaluate the solvation free energy of each conformation. Those redundant conformers that after visual inspection converged in the same geometry were eliminated to avoid

weight imbalance between both solvents. Finally, the log*P* value was estimated by considering the Boltzmann-weighted distribution of the conformational families obtained in water and n-octanol.

Capping group reference value based on glycine residue

To ensure that the log*P* values obtained from our computations were exclusively influenced by the inherent characteristics of their side chains rather than the capping groups, a reference framework was implemented. This involved considering the computationally predicted log*P* value for glycine, a molecule lacking of a heavy atom side chain, but still marked by the influence of capping groups. In the context of the derivatives measured in Model 1 (incorporating the ACE and NME capping groups), an adjustment was introduced by adding +0.17 log*P* units to the calculated value. This value originated from the disparity observed between the glycine amino acid value as reported by Zamora et al. in their 2019 publication and the experimental value documented by Fauchere and Pliska on their published scale. [28] Conversely, within the framework of Model 2 (incorporating the ACE and OME capping groups), a correction was made by subtracting -0.78 logP units. This value reflected the difference existing between the log*P* value computed using the IEFPCM/MST approach and the experimental value detailed in Kubyshkin's article.

Results and Discussion

This work focuses on the reproduction of the experimental values obtained for Kubyshkin [15] using our continuous solvation model. The discussion will be done by amino acid type as follows:

Methionine derivates

The canonical residue methionine is an essential amino acid for its antioxidant effect by reacting with oxidizing species, [29] therefore, the tuning of its properties, e.g., lipophilicity, may be relevant to enhance its bioactivity. To this end, **Figure 6** shows a consistent behavior between Models 1 and 2. Notably, most lipophilic moieties exhibit a congruent (response about the standard methionine residue values. In Model 1, the log*P* value is near to zero, while in Model 2 there is a slightly augmented lipophilicity (0.27). More detailed values can be observed in **Table S13**.

Nonstandard residues ethionine (Eth), norvaline (Nva), norleucine (Nle) and trifluoronorleucine (Tfnle) exhibited a more lipophilic profile than Methionine (Met), with log*P* values moving between 0.20 and 1.82, considering both models. This behavior is logical, attributable to the aliphatic nature of these derivatives (Nle, Nva and Eth), or the addition of halogen moieties, exemplified by Tfnle.

In the case of hydrophilic derivatives, methionine sulfoxide (MetO) and azidohomoalanine (Aha) a consistent pattern is observed. Introduction of functional groups such as azide or sulfoxide provoked a discernible alteration in the lipophilic profile of methionine, culminating in marked negative values, moving between -0.90 and -1.20, considering both models. This is to be expected due to the high polarity of the oxygen and nitrogen atoms that confer hydrogen bond acceptor properties.

One of the most evident divergences between both approximations arises in the case of methoxinine (Mox), characterized by the replacement of the sulfur atom with an oxygen moiety relative to the standard methionine structure. Model 1 gives a sub-zero

value of -0.65, whereas Model 2 manifests a migration towards an apolar value of 0.46. This small incongruity may be ascribed, at least in part, to the presence of an *O*-methyl capping group in Model 2, different from the *N*-methyl capping group featured in Model 1. The computational method IEFPCM/MST elucidates a propensity to increase lipophilicity concerning Model 1, accentuated by the alteration of a nitrogen-hydrogen moiety to oxygen, resulting in the loss of a donor hydrogen bond interaction, a structural modification that IEFPCM/MST tends to penalize towards a more lipophilic value.

Although the influence of capping groups is notably adjusted by the previously commented corrections in the methodological section, certain Model 2 values are corrected starting from an overestimated lipophilic value, and therefore exhibiting an inclination towards greater lipophilicity. This tendency is also observed in ethionine and trifluoronorleucine cases, that present a difference of 1-1.2 log*P* units between models 1 and 2.

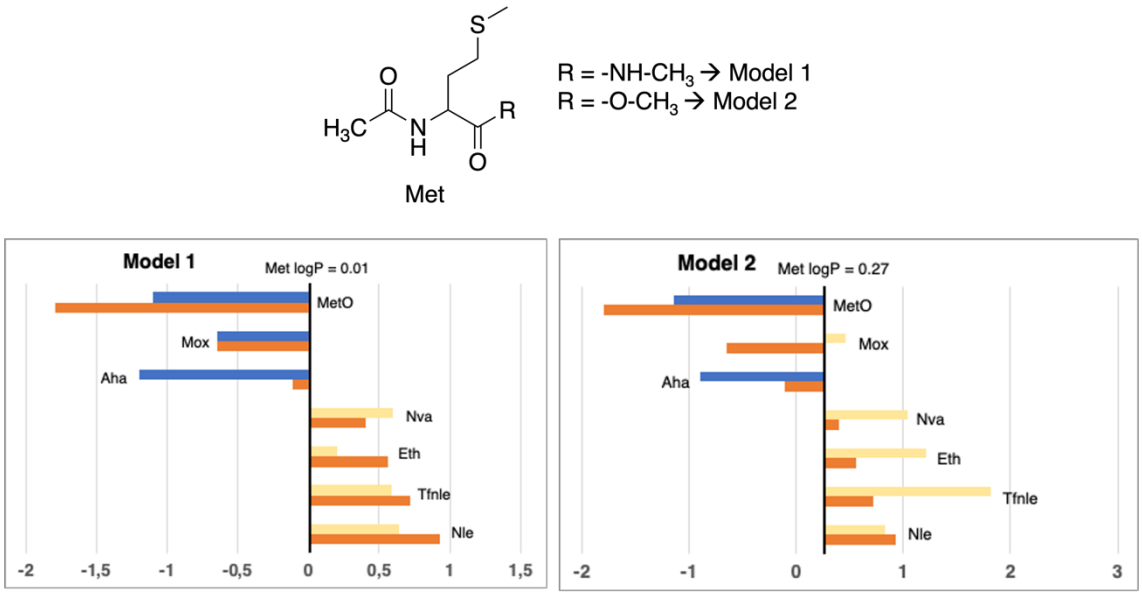

Figure 6. Partition coefficients for methionine (Met) derivatives using Models 1 (left) and Models 2 (right). Standard methionine (Met) residue value is present in the central line of each representation. Nonstandard residue values more lipophilic and more

hydrophilic than the original one, are represented in yellow and blue bars, respectively. Experimental values are represented in orange bars. Detailed experimental data can be found in **Table S13**.

According to correspondence with experimental data (**Table S13**), in Model 1, all methionine cases maintain in differences lower than 1 logP unit. Most slightly deviated case is azidohomoalanine (+1.09 logP units), which contain a chemical group that tend to present difficulties in their estimation. Instead, in Model 2, two cases are above 1 unit of difference, more specifically trifluoronorleucine (1.10 units) and methoxinine (1.11 units), probably ascribed to the presence of groups that exaggerate their lipophilic profile.

Aromatic derivates

In our study we analyzed modifications of the main aromatic residues, tyrosine, phenylalanine, and tryptophan. These residues have several functions that maintain the structure and function of proteins. Interactions with cations are essential for maintaining bioactive protein conformations. [30] Thus, any structural modification of these residues will have an impact on both their lipophilicity and aromaticity and thus on the structure/activity relationships

Tyrosine derivates

Turning our attention to tyrosine derivatives (depicted in **Figure 7**), a similar pattern to methionine is observed, the log*P* value tends to augment lipophilicity in Model 2 (0.52), while registering a slightly hydrophilic quotient of -0.02 in Model 1. More detailed values can be consulted in **Table S14**.

Delving into the assessment of the most lipophilic derivatives, a discernible hierarchy emerges, with 3-fluorotyrosine (3-F-Tyr), 3-nitrotyrosine (3-NO2-Tyr), and

2,3,5,6-tetrafluorotyrosine (2,3,5,6-tetraFTyr) exhibiting proportional lipophilic tendency. Within Model 1, their log*P* values span from 0.36 to 1.66, while Model 2 assigns values between 0.59 and 2.10 – a correlation that is consistently maintained across both models, with the latter consistently indicating a slightly higher lipophilicity.

Conversely, a notable disagreement emerges in the characterization of the Dopa derivative. While Model 1 classifies it as a hydrophilic residue compared to the standard tyrosine (-0.78), Model 2 designates an equivalent lipophilicity to the reference amino acid (0.52). Once again, Model 2 tends to give more lipophilic values in certain cases with respect to Model 1. A trend was observed to a greater or lesser extent in the other tyrosine derivatives. From a chemical point of view, the main difference between tyrosine and Dopa is that the latter has an additional hydroxyl group, a hydrophilic group. Therefore, it is expected that Dopa should have a log*P* value more similar to that of Model 1 (-0.78), which is more hydrophilic and closer to the experimental value reported by Kubyshkin (-0.21).

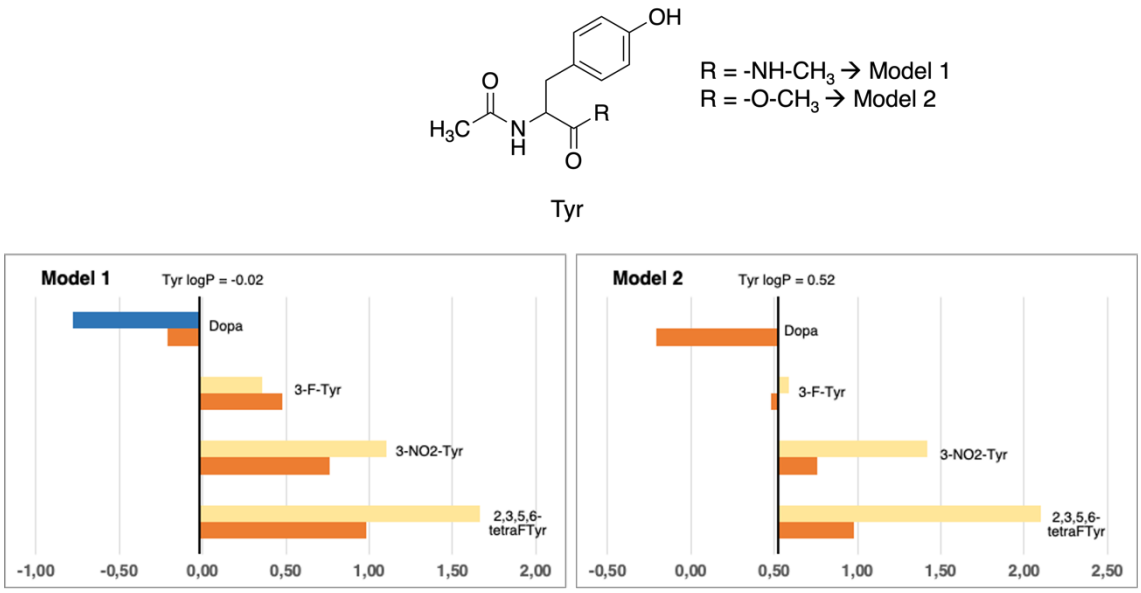

Figure 7. Partition coefficients for tyrosine (Tyr) derivatives using Models 1 (left) and Models 2 (right). Standard tyrosine (Tyr) residue value is present in the central line of

each representation. Nonstandard residues values more lipophilic and more hydrophilic than the original one, are represented in yellow and blue bars, respectively. Experimental values are represented in orange bars. Detailed experimental data can be found in **Table S14**.

In this case, all Model 1 estimations clearly maintain under a 1 logP unit difference with respect to the experimental value. The only case with greater deviation is 2,3,5,6-tetrafluorotyrosine, present in Model 2, that shows an overestimation of 1.12 units in logP, probably due to the simultaneously presence of 4 fluorine atoms in their structure.

Phenylalanine derivates

In a similar line with observations in other derivatives, the logP value associated with standard phenylalanine (see **Figure 8**) reveals a discernible contrast between Model 2 (1.60) and Model 1 (0.61), reflecting a notable increment of one unit in lipophilic propensity within the former. Detailed values can be checked in **Table S15**.

The most lipophilic non-standard residues coincide between both models, encompassing 4-fluorophenylalanine, 4-chlorophenylalanine, 4-trifluoromethyl-phenylalanine, and 4-bromophenylalanine. In Model 1, this subset gives log*P* values ranging from 1.35 to 2.75, while Model 2 attributes values spanning 1.79 to 3.92. Despite not being an identical range, a certain proportionality is maintained between the 4 residues (4-F-Phe < 4-Cl-Phe < 4-CF3-Phe < 4-Br-Phe). All of them have in common that they are residues with halogen groups where those more lipophilic halogen residues (Br) correspond to those that are less lipophilic (F).

Clear disparities appear in two cases. The characterization of methyltyrosine within Model 1 denotes an apolar amino acid, diverging from Model 2's classification as

slightly polar relative to the reference phenylalanine value. Although the absolute values in both models (1.42 vs 1.48) closely approximate the experimental value of 0.92 – which unmistakably denotes an apolar nature – the discrepant classifications stem from Model 2's overestimation of the lipophilic propensity of standard phenylalanine (valued at 1.60), influencing the classification. Also, a slight discrepancy between models of the amino acid 4-acetylamido-phenylalanine is observed, being more hydrophilic in model 1 than in model 2 with a difference of 1.53 units between them.

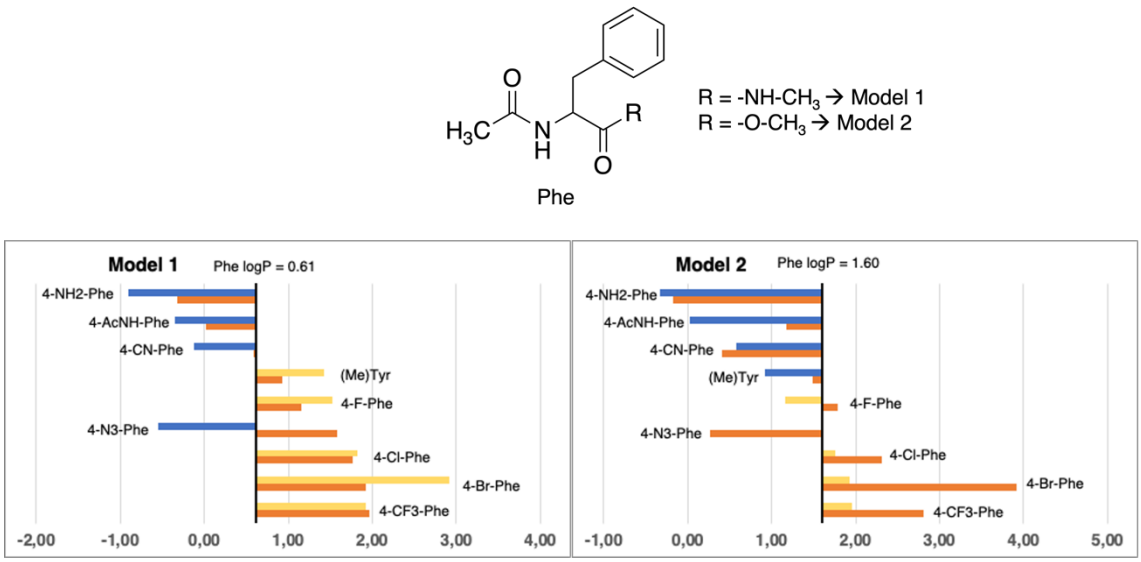

Figure 8. Partition coefficient for phenylalanine (Phe) derivatives using Models 1 (left) and Models 2 (right). Standard phenylalanine (Phe) residue value is present in the central line of each representation. Nonstandard residues values more lipophilic and more hydrophilic than the original one, are represented in yellow and blue bars, respectively. Experimental values are represented in orange bars. Detailed experimental data can be found in **Table S15**.

Delving into experimental values consonance, just some specific cases present deviations greater than 1 unit of logP: 4-azidophenylalanine (2.13 units in Model 1 and 1.32 in Model 2), 4-bromophenylalanine (2.00 units Model 2) and 4-acetamidophenylalanine (1.16 units in Model 2). All cases are in line with other situations observed.

Tryptophan derivates

In this specific scenario, differences between Model 1 and Model 2 about the log*P* value assessment for standard tryptophan exhibit a nearly negligible difference. As evidenced in **Figure 9**, Model 1 assigns a value of 1.37, while its Model 2 counterpart presents a value of 1.71, illustrating a marginal deviation of merely 0.34 units. It is worth noting that, both of these values closely approximate the experimental measurement (1.20), differing between 0.17 to 0.51 logP units, denoting a modest shift towards heightened lipophilicity. More exact values are available in **Table S16**.

A clear trend emerges, demonstrating the congruence in residue classification across both models. On one hand, residues exhibiting enhanced polarity relative to standard phenylalanine are characterized by the introduction of aliphatic or halogen substituents. Examples encompass 5-methyltryptophan (5-CH3-Trp), 5-fluorotryptophan (5-F-Trp), 5-chlorotryptophan (5-Cl-Trp), 1-methyltryptophan (1-CH3-Trp), 6-bromotryptophan (6-Br-Trp), and 5-bromotryptophan (5-Br-Trp), characterized by absolute values ranging from 1.37 to 4.07 in Model 1, and 1.79 to 4.66 in Model 2. On the other hand, polar residues are exemplified by 5-aminotryptophan (5-NH2-Trp) and 5-hydroxytryptophan (5-OH-Trp), exhibiting values of -0.24 and 0.42 in Model 1, and 0.01 and 0.28 in Model 2, respectively. While the trend holds true for hydrophilic residues, certain alterations in the order become evident in the case of hydrophobic derivatives, particularly noticeable in the order of bromine and methyl derivatives.

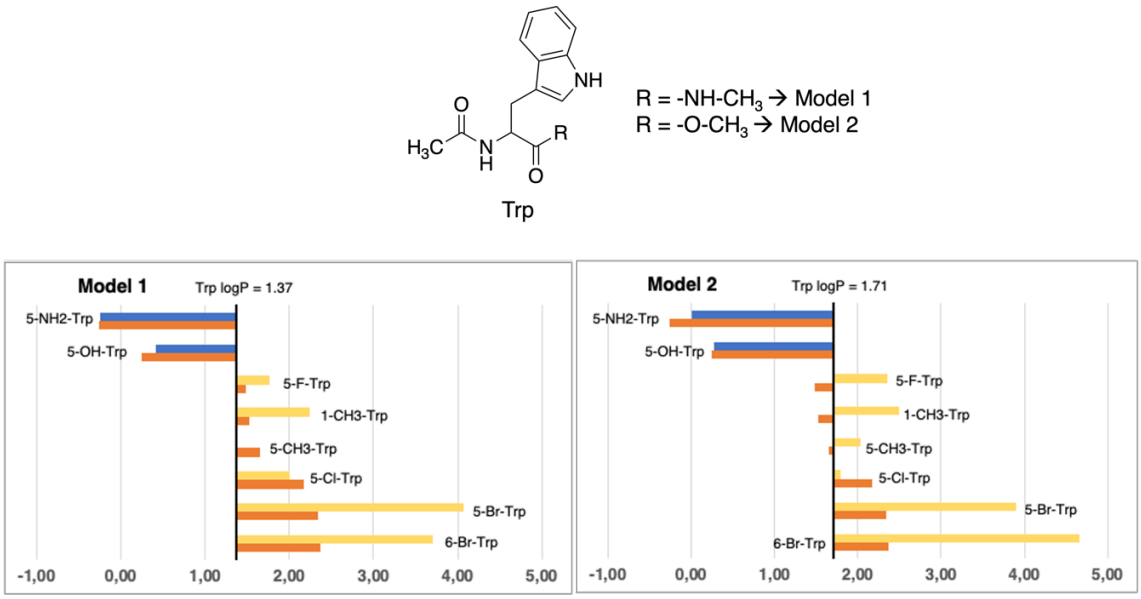

Figure 9. Partition coefficient for tryptophan (Trp) derivatives using Models 1 (left) and Models 2 (right). Standard tryptophan (Trp) residue value is present in the central line of each representation. Nonstandard residues values more lipophilic and more hydrophilic than the original one, are represented in yellow and blue, respectively. Experimental values are represented in orange bars. Detailed experimental data can be found in **Table S16**.

In this case, divergent values compared to experimental ones are, in both models, the two compounds with bromine atoms: 5-bromo and 6-bromotryptophan, being always too lipophilic. For 5-bromotryptophan, the difference is between 1.56 and 1.73 logP units and for 6-bromotryptohan is between 1.34 and 2.29. The rest of cases maintain in values lower than 1 logP unit.

Lysine derivates

This amino acid represents the most frequently post-translationally modified so it is no coincidence its impact on protein regulation and function. [31] The reference log*P* value for standard lysine is taken from Zamora et al.,[9] which gives an approximate logP of -

0.40. (see **Figure 10**). Based on that, most lipophilic nonstandard residues encompass tert-butoxycarbonyl-lysine ((Boc)Lys) and S-alylcysteine (Sac), which have almost identical values in one each model, around ~0.90 and ~1.30, respectively. In both cases, we are introducing aliphatic carbons into the chain and/or reducing the polarity of the distal amino group of standard lysine, which could explain the increase in 1.3-1.7 log*P* units. Absolute values can be checked in **Table S17**.

However, discrepancies manifest in the two other cases. Concerning *N*-propargyloxycarbonyl-lysine ((Pro)Lys), Model 1 considers it slightly polar (-0.45), whereas Model 2 positions it distinctly as more apolar (0.57) compared to coded lysine (-0.40). Although the side chain retains a bit of polarity, due to its amide group, the introduction of four aliphatic carbons, similar to the previously noted (Boc)Lys, logically places the most reasonable value within Model 2 (predicted value of ~0.90).

According to N-acetyl-lysine ((Ac)Lys), Model 1 denotes an extremely hydrophilic characterization of -1.64, while in Model 2 persist in -0.39, almost identical than the standard residue. Chemically insight underscores the incorporation of an acetyl group, analogous to (Boc)Lys and ((Pro)Lys), that should provoke a slight increase in the lipophilicity. Consequently, Model 1 tends to excessively accentuate the hydrophilic trait of this residue, while Model 2's depiction is more aligned with an hydrophobic profile.

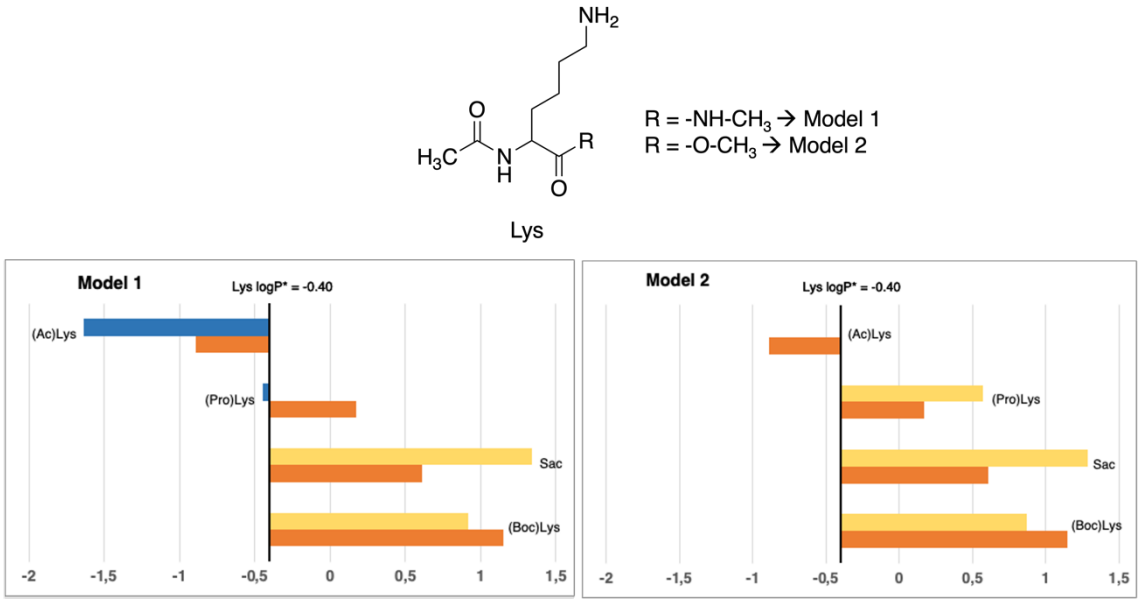

Figure 10. Partition coefficients for lysine (Lys) derivatives using Models 1 (left) and Models 2 (right). Standard lysine (Lys) residue value is present in the central line of each representation. Nonstandard residues values more lipophilic and more hydrophilic than the original one, are represented in yellow and blue, respectively. Experimental values are represented in orange bars. Detailed experimental data can be found in **Table S17**. .*Value measured in the article of Zamora et al. in 2019.

In this case, the comparison with experimental data, show that all cases in both Models almost have a perfect fitting, presenting a difference of lower that 1 unit. In Model 1, the most deviated case has a 0.75 units divergence while in Model 2 is 0.68 units.

Proline derivates

In the context of proline residues, their examination was previously undertaken by Matamoros et al. in 2022, [32] employing the SMD solvation approach. [33] In the current study, we subject these residues to analysis via our IEFPCM/MST methodology. As illustrated in **Figure 11** and detailed in **Table S18**, the outcomes affirm the comparable efficacy of our approach, thus rendering it well-suited and eminently applicable for extending our scale of lipophilicity.

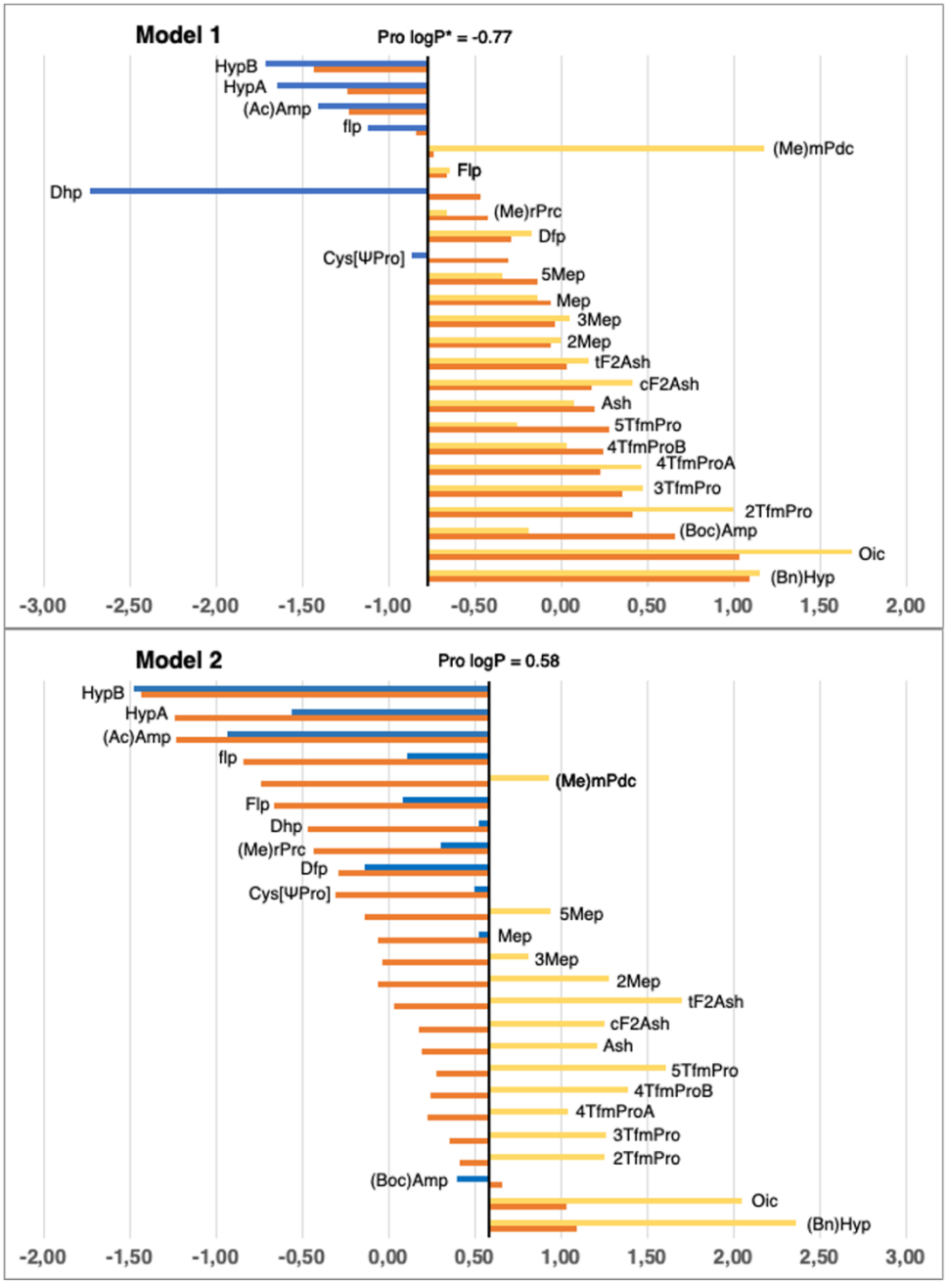

Figure 11. Partition coefficient for proline (Pro) derivatives using Models 1 (up) and 2 (down). Standard proline (Pro) residue value is present in the central line of each representation. Nonstandard residues values more lipophilic and more hydrophilic than the original one, are represented in yellow and blue, respectively. Detailed experimental values can be found in **Table S18**.*Value measured in the article of Zamora et al. in 2019.

Regarding proline, in general all cases maintain in a difference with respect the experimental value, lower than 2 units of logP in both models. Only dehydroproline (Dhp) in Model 1, presents a deviation of 2.26 logP units, presenting an excessive hydrophilic profile due to the capacity of performing hydrogen bond interaction of the NH group.

General correlation between experimental and computational data

Regarding the correlation between experimental and computational estimations, **Figures 12 and 13** reveal a noteworthy similarity between the two models. Both exhibit a correlation coefficient (R^2) that is highly acceptable - 0.73 for Model 1 and 0.71 for Model 2. Also, the root mean square error (RMSE) for Model 1 is 0.7, while Model 2 stands on 0.9. Furthermore, additional statistical parameters such as mean square error (MSE) and mean unsigned error (MUE) also indicate a consistent pattern across both models. Despite the subtle differences, Model 1 demonstrates a stronger correlation and exhibits lower error when compared to the experimental values. This observation aligns coherently with our previously published findings that utilized the same capping groups as those in Model 1.

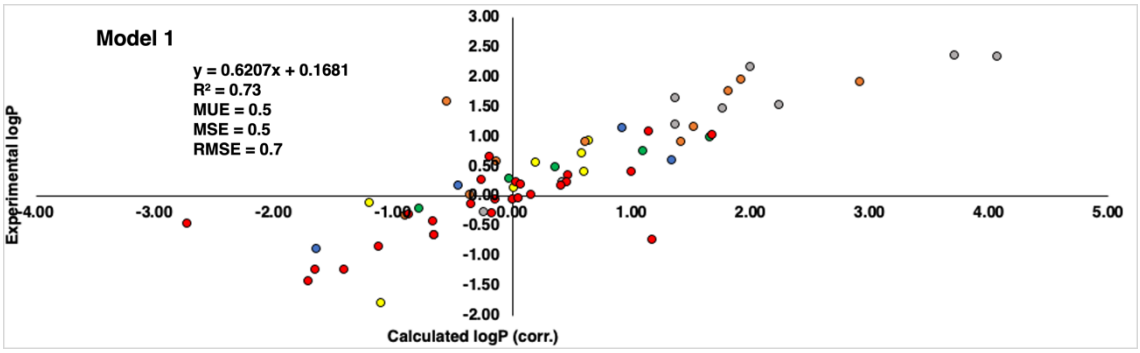

Figure 12. Correlation between calculated logP (corrected by eliminating the influence of capping groups of **Model 1**) (axis X) and experimental log*P* values reported by Kubyshkin (axis Y). Groups of residues are represented in different pattern of colors: methionine (yellow), tyrosine (green), phenylalanine (orange), tryptophan (grey), lysine (blue) and proline (red).

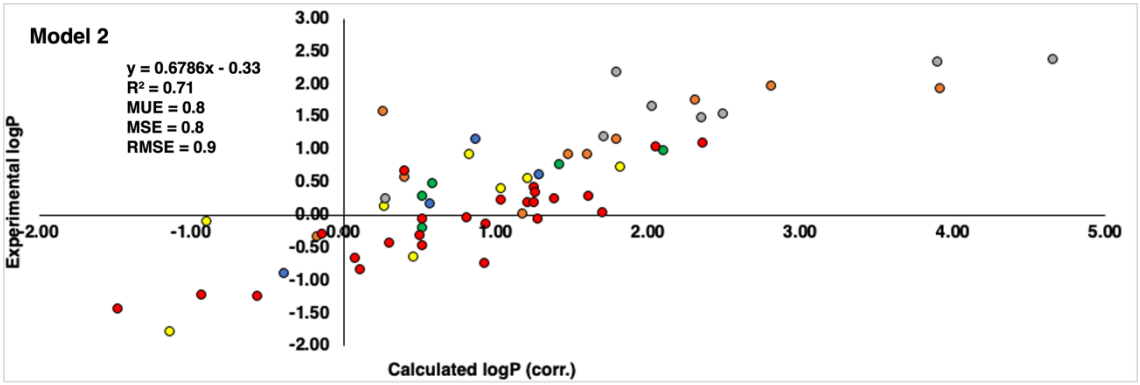

Figure 13. Correlation between calculated logP (corrected by eliminating the influence of capping groups of **Model 2**) (axis X) and experimental logP values reported by Kubyshkin (axis Y). Groups of residues are represented in different pattern of colors: methionine (yellow), tyrosine (green), phenylalanine (orange), tryptophan (grey), lysine (blue) and proline (red).

Delving into specifics, it becomes evident that indistinctly of the model considered, all methionine and tyrosine residues (yellow and green colored dots in **Figures 12 and 13**, respectively) consistently exhibit present values with a discrepancy

not higher than 1.2 logP units. Similarly, in the case of lysine is even better, remaining below than 1 logP unit in all cases (blue colored dots in **Figures 12 and 13**). This can be translated into the fact that there is hardly any difference between the experimental values and those reproduced computationally. It's worth noting that a deviation near 1 log*P* unit can often be attributed to inherent errors within the method itself.

As regards the other derivatives, in general trend is maintained, however, certain cases with notably substantial discrepancies have come to light. In the case of the Phenylalanine derivatives (orange-colored dots in **Figures 12 and 13**), the azidohomoalanine derivative (4-N3-Phe) in Model 1, presents a deviation of 2.13 units from the experimental value, though this deviation is somewhat mitigated in Model 2 (1.32 units). An underlying explanation could be associated with the azide group present within the side chain. Despite the group's net charge being neutral, there exists a subtle polarization distributed across its nitrogen atoms. This charge distribution potentially contributes to deviations and fluctuations in accurately estimating the compound's lipophilicity. The difference of 0.81 units in the assessment of the identical residue across Model 1 and Model 2 might arise from the distinct origins of their unadjusted initial values. In Model 1, the original value comes from a measurement involving the NME capping group, introducing a hydrogen bond interaction that doesn't occur with the OME capping in Model 2. Consequently, this residue could potentially be overestimated as hydrophilic, driven by the presence of the NME capping group. This last aspect is also observed in the most deviated case from proline dataset (red dots in **Figures 12 and 13**), Dehydroproline (Dhp) that in the case of Model 1 deviates 2.26 units more hydrophilic than the experimental one.

Another pattern observed in deviated cases is the presence of bromine atoms, more specifically, in the case of Phenylalanine and Tryptophan residues (orange and grey colored dots in **Figures 12 and 13**, respectively), 4-bromo-phenylalanine (4-Br-Phe), 5-Br-Trp (5-bromotryptophan) and 6-Br-Trp (6-bromotryptophan) present a more marked difference. While the rest of residues present a deviation around 1 logP unit, these cases move around 1.34 and 2.29, between both models. Bromine atom is a relatively heavy atom compared to other lighter halogens like fluor and chlorine (deviations around 0-0.8 logP units) and it has unique electronic properties due to its larger size and higher atomic number. These factors can influence the interactions of bromine with its surrounding environment, including solvent molecules and other atoms in a molecule.

Application of non-standard amino acids in proteomics

Acetylation is a relevant post-translational modification (PTM) that is mainly carried out in both the ε-amine on the side chain of lysine and in the α-amine of the N-terminus in peptides and proteins. [34,35] In health and pathological states, the reversible acetylation of lysine residues plays a crucial role in regulating cellular and developmental processes. These facts make the correct identification of acetylated peptides and proteins a constantly developing field in modern proteomics. [36] In this context, experimental studies have analyzed the impact of peptide acetylation on chromatographic retention time in RP HPLC in order to create predictive models that suggest an efficient way for the separation of these peptides that will allow their identification. [36]

Table 1 reports the experimental hydrophobicity index (ΔHI) obtained by Mizero and collaborators, [36] calculated as the difference between the hydrophobicity index for modified (acetylated) and non-modified peptide pairs expressed in % acetonitrile (% ACN). As can be seen, it is evident that the greater the change in the degree of acetylation

in the lysines of the peptide pairs, the greater the amount of organic solvent (% ACN) will be necessary.

Table 1. Variation of the experimental hydrophobicity index (*HI*) using as separation mode acetonitrile (% ACN) in RP HPLC 0.1 % of formic acid for modified (acetylated) and non-modified peptide pairs.

Acetylated lysine residues in peptides (number of peptide pairs)	**Experimental hydrophobicity index (Δ*HI*) in RP HPLC 0.1% formic acid (% ACN)**	**Δ(Ac)Lys**	**ΔΔ*HI*/Δ(Ac)Lys**
0 (10632)	5.02	-	-
1 (13791)	9.14	1	4.12
2 (2390)	11.20	2	6.18
3 (316)	12.54	3	7.52

Figure 14 depicts the derivative of the change in the hydrophobic index with respect to the change in (Ac)Lys residues in the modified/non-modified pairs (ΔΔ*HI*/Δ(Ac)Lys) from which the slope permit to obtain the increase in hydrophobicity due to the acetylation of a lysine residue (ca. 1.70 units)

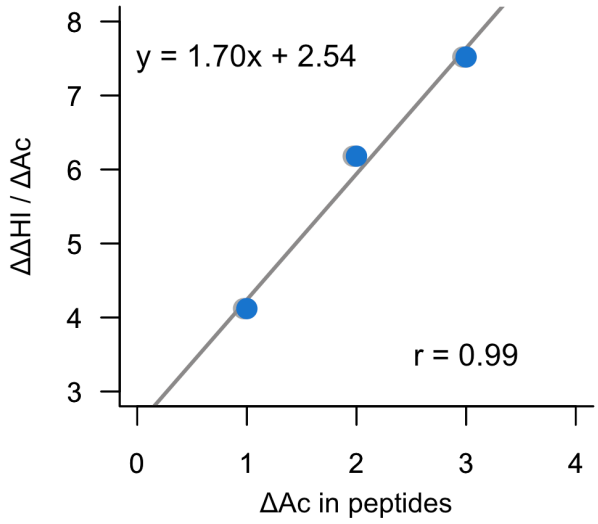

Figure 14. Representation of the variation of the experimental hydrophobic index (% ACN) by the change in the number of acetylated lysines in peptide pairs as a function of the change of acetylated lysines in peptide pairs.

In order to further evaluate the reliability of the predictions of the calculations performed in this work, **Table 2** reports the increase of hydrophobicity under the acetylation of lysine residues ($\Delta\log P_{Ac}$) using our computations but also those obtained of ChemAxon [37] and milogP. [38] Since our prediction (ca. 1.43 logP units) is close to that obtained experimentally (see **Figure 14**), it can be used to efficiently predict experimental HPLC conditions to separate acetylated peptides and thus, be used for proteomic studies.

Table 2. Partition coefficients (log*P*) values for lysine (Lys) and acetyllysine ((Ac)Lys) using the capping groups of the Model 1 and change in lipophilicity due to acetylation ($\Delta\log P_{Ac}$) of lysine residues.

Amino acid code	**log*P***		**$\Delta\log P_{Ac}$**
	Lys	**(Ac)Lys**	
This work	-3.07	-1.64	1.43

ChemAxon	-4.51	-1.66	2.85
milogP	-3.73	-1.25	2.48

Conclusions

The lipophilicity of amino acids is one of the main physicochemical properties of these biomolecules as it gives an estimate of solubility, binding propensity, and bioavailability. In this work we show how several structural modifications in residues such as methionine, aromatic, lysine, and proline can tune the hydrophobic properties of these residues, opening a window of possibilities to be used as a guide for the design of peptides and proteins with tailor-made characteristics. The structural models used, based on differences in capping groups showed mostly important similarities, validating the lipophilicity values obtained for the non-standard side chains. Differences were found in fewer cases, indicating an effect of the capping group on the side chain hydrophobicity which can be expected as one model uses a hydrogen bond donor (Model 1) while the other uses a hydrogen bond acceptor (Model 2). In general, both models correlate well with the experimental values, obtaining statistical errors less in the case of Model 1. Finally, our predictions were able to efficiently predict the experimental hydrophobicity change due to the number of acetylated lysines in peptide pairs determined by HPLC, opening up the possibility that our scale can be employed for proteomics studies that include post-translational modifications beyond acetylation.

Abbreviations

(Ac)Lys: N-acetyllysine
(Boc)Lys: tert-butoxycarbonyllysine
(Pro)Lys: N-propargyloxycarbonyllysine

1-CH3-Trp: 1-methyltryptophan

2,3,5,6-tetraFTyr: 2,3,5,6-tetrafluorotyrosine

3-F-Tyr: 3-fluorotyrosine

3-NO2-Tyr: 3-nitrotyrosine

4-Br-Phe: 4-bromophenylalanine

4-CF3-Phe: 4-trifluoromethylphenylalanine

4-Cl-Phe: 4-chlorophenylalanine

4-F-Phe: 4-fluorophenylalanine

4-N3-Phe: 4-azidophenylalanine

5-Br-Trp: 5-bromotryptophan

5-CH3-Trp: 5-methyltryptophan

5-NH2-Trp: 5-aminotryptophan

5-OH-Trp: 5-hydroxytryptophan

6-Br-Trp: 6-bromotryptophan

AAs: amino acids

ACE: Acetyl

ACN: Acetonitrile

Aha: Azidohomoalanine

CamSol-PTM: Cambridge Solvation Post Traductional Modifications

DFT: Density Functional Theory

Dhp: Dehydroproline

Eth: Ethionine

HBA: Hydrogen Bond Acceptor

HBD: Hydrogen Bond Donor

HI: Hydrophobicity index

IEFPCM/MST: Integral Equation Formalism Polarizable Continuum Model / Miertus–Scrocco–Tomasi

Lys: Lysine

Met: Methionine

MetO: Methionine Sulfoxide

Mox: Metoxinine

MSE: Mean square error

MUE: Mean unsigned error

MW: Molecular Weight

Nle: Norleucine

NME: N-Methyl

NMR: Nuclear Mangetic Resonance

Nva: Norvaline

OME: O-Methyl

Phe: Phenylalanine

Pro: Proline

PTM: Post-translational modification

RB: Rotatable Bond

RMSE: Root Mean Square Error

RP HPLC: Reversed Phase-High Performance Liquid Chromatography

Sac: S-alylcysteine

SMD: Solvation Model Density

Tfnle: Trifluoronorleucine

Trp: Tryptophan

Tyr: Tyrosine

Declarations

Author contributions

AV: Conceptualization, Data curation, Formal analysis, Funding-acquistion, Investigation, Methodology, Project administration, Resources, Supervision, Writing-original draft, Writing-review & editing

PM: Investigation, Data curation, Visualization, Writing-review & editing

DBR: Investigation, Data curation, Visualization, Writing-review & editing

WJZ: Conceptualization, Data curation, Formal analysis, Funding-acquistion, Investigation, Methodology, Project administration, Resources, Supervision, Writing-original draft, Writing-review & editing

Conflicts of interest

The authors declare that they have no conflicts of interest

Ethical approval

Not applicable

Consent to participate

Not applicable

Consent to publication

Not applicable

Funding

The authors thank the Vice Chancellor for Research of the University of Costa Rica for its support work via the research projects 115-C2-126 and 115-C1-450**.** The Consorci de Serveis Universitaris de Catalunya (CSUC) is acknowledged for providing computational resources.

References

1. von Heijne G. Protein Evolution and Design. Annu Rev Biochem. 2018;87:101-3.
2. Doig AJ. Frozen, but no accident – why the 20 standard amino acids were selected. FEBS J. 2017;284:1296-1305.
3. Efremov RG, Chugunov AO, Pyrkov TV, Priestle JP, Arseniev AS, Jacoby E. Molecular lipophilicity in protein modeling and drug design. Curr Med Chem. 2007;14:393-415.
4. Tang S, Li J, Huang G, Yan L. Predicting Protein Surface Property with its Surface Hydrophobicity. Protein Pept Lett. 2021;28:938-44.
5. Weinstein JY, Elazar A, Fleishman SJ. A lipophilicity-based energy function for membrane-protein modelling and design. PLoS Comput Biol. 2019;15:e1007318.

6. Simm S, Einloft J, Mirus O, Schleiff E. 50 Years of Amino Acid Hydrophobicity Scales: Revisiting the Capacity for Peptide Classification. Biol Res. 2016;49:31.
7. Peters C, Elofsson A. Why is the Biological Hydrophobicity Scale More Accurate than Earlier Experimental Hydrophobicity Scales?. Proteins. 2014; 82:2190-8.
8. MacCallum JL, Tieleman DP. Hydrophobicity Scales: A Thermodynamic Looking Glass into Lipid-Protein Interactions. Trends Biochem Sci. 2011;36:653-62.
9. Zamora WJ, Campanera JM, Luque FJ. Development of a Structure-Based, pH-Dependent Lipophilicity Scale of Amino Acids from Continuum Solvation Calculations. J Phys Chem Lett. 2019;10:883-9.
10. Dunbrack Jr RL, Karplus, M. Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains. Nat Struct Mol Biol. 1994;1:334-40.
11. Castro TG, Melle-Franco M, Sousa CEA, Cavaco-Paulo A, Marcos JC. Non-Canonical Amino Acids as Building Blocks for Peptidomimetics: Structure, Function, and Applications. Biomolecules. 2023;13:981.
12. Ochoa R, Fox T. Assessing the fast prediction of peptide conformers and the impact of non-natural modifications. J Mol Graph Model. 2023;125:108608.
13. Jin X, Park OJ, Hong SH. Incorporation of non-standard amino acids into proteins: challenges, recent achievements, and emerging applications. Appl Microbiol Biotechnol. 2019;103:2947-58.

14. López-López E, Robles O, Plisson F, Medina-Franco JL. Mapping the structure–activity landscape of non-canonical peptides with MAP4 fingerprinting. Digit Discov. 2023;2:1494-1505.
15. Kubyshkin V. Experimental lipophilicity scale for coded and non coded amino acid residues. Org Biomol Chem. 2021;19:7031-40.
16. Oeller M, Kang RJD, Bolt HL, Gomes Dos Santos AL, Weinmann AL, Nikitidis A**, et al.** Sequence-based prediction of the intrinsic solubility of peptides containing non-natural amino acids. Nat Commun. 2023;14:7475.
17. Sander C, Freyss J, van Korff M, Rufener C. DataWarrior: An Open-Source Program For Chemistry Aware Data Visualization And Analysis. J Chem Inf Model. 2015;55:406-73.
18. Hansch C, Leo A, Hoekman D. Exploring QSAR: Hydrophobic, electronic, and steric constants. In: Hansch C, Leo A, Hoekman D, editors. Exploring QSAR: Hydrophobic, electronic, and steric constants. Washington, DC: American Chemcial Society; 1995. p. 6.
19. Hanwell MD, Curtis DE, Lonie DC, Vandermeersch T, Zurek E, Hutchison GR. vogadro: An advanced semantic chemical editor, visualization, and analysis platform. J Cheminformatics. 2012;3:17.
20. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: An open chemical toolbox. J Cheminformatics. 2011;3:33.
21. Lee CT, Yang WT, Parr RG. Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron-Density. Phys Rev B. 1988;37:785-9.
22. Becke AD. Density functional thermochemistry. III. The role of exact exchange. J Chem Phys. 1993; 98:5648-52.

23. Stephens PJ, Devlin FJ, Chabalowski CF, Frisch MJ. Ab-Initio Calculation of Vibrational Absorption and Circular-Dichroism Spectra Using Density-Functional Force-Fields. J Phys Chem-Us. 1994;98:11623-27.
24. Curutchet C, Orozco M, Luque FJ. Solvation in octanol: Parametrization of the continuum MST model. J Comput Chem. 2001;22:1180-93.
25. Curutchet C, Bidon-Chanal A, Soteras I, Orozco M, Luque FJ. MST continuum study of the hydration free energies of monovalent ionic species. J Phys Chem B. 2005;109:3565-74.
26. Soteras I, Curutchet C, Bidon-Chanal A, Orozco M, Luque FJ. Extension of the MST model to the IEF formalism: HF and B3LYP parametrizations. J Mol Struc-Theochem. 2005;727:29-40.
27. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR**, et al.** Gaussian 16, Revision C.01, Gaussian, Inc., Wallingford CT, 2016.
28. Fauchère JL, Pliska V. Hydrophobic Parameters Pi of Amino Acid Side Chains from the Partitioning of N-Acetyl-Amino Acid Amides. Eur J Med Chem. 1983;18:369-75.
29. Lim JM, Kim G, Levine RL. Methionine in Proteins: It's Not Just for Protein Initiation Anymore. Neurochem Res. 2019;44:247-57.
30. Pinheiro S, Soteras I, Gelpí JL, Dehez F, Chipot C, Luque FJ, et al. Structural and energetic study of cation-π-cation interactions in proteins. Phys Chem Chem Phys. 2017;19:9849-61.
31. Zhang Z, Tan M, Xie Z, Dai L, Chen Y, Zhao Y. Identification of lysine succinylation as a new post-translational modification. Nat Chem Biol. 2011;7:58-63.

32. Matamoros P, Pinheiro S, Viayna A, Zamora WJ. Towards an understanding of the lipophilicity of non-coded amino acids: computational simulations of proline analogs. IEEE 4th International Conference on BioInspired Processing (BIP); 2022 Nov 15-17; Cartago, Costa Rica. p. 1-5.
33. Marenich AV, Cramer CJ, Truhlar DG. Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. J Phys Chem B. 2009;113:6378-96.
34. Ree R, Varland S, Arnesen T. Spotlight on protein N-terminal acetylation. Exp Mol Med. 2018;50:1-13.
35. Yan K, Mousavi N, Yang XJ. Analysis of Lysine Acetylation and Acetylation-like Acylation In Vitro and In Vivo. Curr Protoc. 2023;3:e738.
36. Mizero B, Yeung D, Spicer V, Krokhin OV. Peptide retention time prediction for peptides with post-translational modifications: N-terminal (α-amine) and lysine (ε-amine) acetylation. J Chromatogr A. 2021; 1657:462584.
37. Marvin version 23.16, ChemAxon (https://chemaxon.com)
38. molinspiration.com/cgi/properties [Internet]. Slovakia: Slovensky Grob; c2024 [cited 2024 Jan 26]. Available from: https://www.molinspiration.com/cgi/properties