Identifying Sequential Residue Patterns in Bitter and Umami Peptides

Dutta, A.; Bereau, T.; Vilgis, T.A.

DOI
10.1021/acsfoodscitech.2c00251

Publication date
2022

Document Version
Final published version

Published in
ACS Food Science and Technology

ACS FOOD SCIENCE & TECHNOLOGY

pubs.acs.org/acsfoodscitech

Article

Identifying Sequential Residue Patterns in Bitter and Umami Peptides

Arghya Dutta,* Tristan Bereau, and Thomas A. Vilgis*

Cite This: ACS Food Sci. Technol. 2022, 2, 1773−1780

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: A peptide's amino acid sequence affects its taste, but how? A rigorous structure−property connection is challenging to determine because of both the exponentially growing peptide sequence space and the scarcity of experimental measurements compared to the size of that space. By sensory methods, many peptides have been identified as tasting bitter or umami. Baselines have been determined but relate only single amino acid characteristics, in particular hydrophobicity in bitter peptides and negative charges for umami. In this work, we refine this picture by extracting sequential amino acid patterns. Our method coarse-grains the peptide sequence space to facilitate the systematic identification of common residue patterns. We identify optimal patterns for both bitter and umami peptides: one hydrophobic followed by four polar residues and two negative followed by three polar residues, respectively. We find systematic improvements compared to both random and the baselines mentioned above. Our method complements quantitative structure−activity relationship methods by leveraging sequential information to help locate taste-specific characteristics in peptides and proteins.

KEYWORDS: *bitterness, umami, peptide taste, pattern finding, plant-based proteins, taste generation*

INTRODUCTION

Special compounds evoke specific tastes: sodium chloride (salty), sugars (sweet), acids (sour), phenols (bitter, astringent, and sour), alkaloids (bitter), glutamic acids (umami), and nucleotides (umami enhancers). Tastes are crucial because the gustatory system, the sensory system that helps in perceiving taste, often informs us about safe and harmful foods through their tastes.[1] In addition, taste determines most of our food preferences.[2] For example, vegetables such as cabbage, cucumber, and spinach often taste bitter because they contain plant alkaloids, which can be toxic if consumed in large amounts and are known to have excessive bitter taste;[2,3] consequently, we avoid them. Because we tend to avoid bitter foods and seek savory ones, classifying foods on the basis of the taste responses they evoke and modulate and finding the physicochemical reasons behind those responses are indispensable steps in designing new nutritional and palatable foods. The growing number of curated databases of bitter- and umami-tasting foods[4,5] indicates recent progress in this direction.

Bitter and umami represent two major taste modalities of peptides. Bitter peptides are often found in fermented foods,[6,7] protein hydrolysates,[3] and matured cheese.[8] In matured cheese, for example, they are produced during ripening because most of the bitter-tasting amino acids are hidden in the caseins. Because bitter-tasting amino acids are generally hydrophobic,[9] it would not be surprising if an abundance of hydrophobic peptides strongly affects the flavors of foods such as still-ripening and well-matured cheeses. As for savory foods, most of them result from protein hydrolysis that occurs in long-cooked foods, fermented foods such as soy, fish, oyster sauces, and miso pastes, and long-matured foods such as cheese and cured meat.[10−15] It is now well-accepted that amino acids and peptides contribute significantly to the overall taste of savory foods.[16] While single amino acids are likely to form aroma compounds during thermal and microbiological processing,[11,17] peptides may remain more stable and can contribute to taste depending on process parameters. Taken together, these observations suggest a connection between the physicochemical properties of the amino acids and their tastes.

Given the impact of individual amino acids, the challenging question, then, is to understand the role of specific residue patterns in determining a peptide's taste via physicochemical properties such as hydrophobicity, polarity, and charge. With the discovery of new taste-relevant peptides in foods from various preparations[18] and progress in plant-based foods, this question is becoming more relevant. For example, if we design surrogate products from plant proteins, it would be useful to identify which short sequences of these proteins exhibit particular flavors. These proteins can then be thermally and enzymatically treated to extract flavor peptides for use as flavor enhancers.

Traditionally, the study of the tastes of peptides relied on the quantitative structure−activity relationship (QSAR) framework that relates peptide descriptors to some desired target property using statistical and machine learning (ML) methods.[19−21] For bitter peptides, a QSAR has been used

Received: August 18, 2022
Revised: October 20, 2022
Accepted: October 25, 2022
Published: November 9, 2022

along with physicochemical descriptors, for example, to predict the threshold concentration for bitterness,[22] to predict bitter and nonbitter peptides,[23] to find residue types of bitter di- and tripeptides,[24] and to find bioactivity of bitter peptides.[25] Predicting bitter taste on the basis of only sequence information has been attempted recently by Charoenkwan et al.[26,27] Though early on it was suggested that the positions of residues of a peptide do not affect its taste,[6,9] multiple studies since then have found that the residue positions do affect the taste.[24,28]

For umami peptides, there are fewer QSAR studies compared to the number for bitter peptides. One reason for this may be that sensory evaluation, one of the primary methods used to determine the umami-ness of foods, can be subjective[29] and expensive (see ref 30 for a summary of currently used methods). Accordingly, defining the target variable for umami intensity proved to be difficult. To overcome this difficulty, the computational methods that have been generally used to predict and analyze umami peptides relied on structural analysis such as homology modeling and molecular docking[31−33] of possible umami peptides to umami taste receptors such as T1R1/T1R3.[34] Quite recently, physicochemical descriptors[35] and only sequence information[36] were used along with ML-based methods to classify umami and non-umami peptides.

As we have seen so far, QSAR and ML methods focus either on classifying peptides or on predicting values of some target variables using physicochemical descriptors or sequential information. While QSAR methods are generally easy to interpret, the physicochemical descriptors they use come from linear and nonlinear dimensionality reduction techniques;[37] this can make the final models less interpretable. In addition, often multiple descriptors are needed to achieve higher prediction accuracy;[19,25] this makes the models multidimensional and even more difficult to interpret. While ML is shown to make accurate predictions using only sequence information,[27] they inherit the low interpretability issue often found in ML methods such as deep neural networks. The scarcity of experimentally verified data on the tastes of peptides creates an additional, often rate-limiting, step for black-box ML models that generally work well only when they are trained with a large, labeled data set. Thus, while QSAR and ML methods are essential for making accurate predictions given a peptide sequence, their low interpretability becomes an issue if, for example, the aim is to design *new* (i.e., out-of-sample) umami peptides or to locate bitter-causing segments in a long protein. Instead, a systematically derived generic residue pattern, which is possibly connected to taste, can provide a better starting point and thus can substantially accelerate the realization of these aims. In this paper, we propose a method that identifies such generic coarse-grained residue patterns that are often found in bitter and umami peptides. The lower granularity of a coarse-grained model is necessary as it helps in identifying generic residue patterns by reducing the size of the peptide sequence space.

To this end, we first reduced the size of the peptide sequence space by classifying the amino acids into four coarse-grained residue types: hydrophobic (H), polar and hydrophilic (P), positively charged (+), and negatively charged (−) (Figure 1). We combinatorially constructed seven comprehensive, increasingly large libraries of peptides with coarse-grained residue patterns. We compiled a database of bitter and umami peptides from the literature. After dividing the database

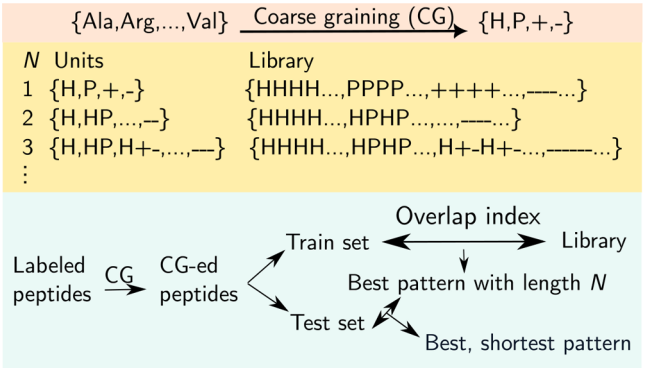

Figure 1. Schematic diagram illustrating the method we used to identify the most common coarse-grained residue patterns present in known bitter and umami peptides.

peptides into train and test sets, we compared the library peptides to the coarse-grained bitter and umami peptides from the training sets using a sequence comparison index[38] and two surrogate measures, defined using the comparison index, for bitterness and umami-ness. This comparison identified the best residue patterns that have the highest average overlaps with the bitter and umami peptides, for each library. Finally, we compared the average overlaps of the peptides, constructed from the predicted patterns from different libraries, with peptides from the test sets to find the shortest pattern that has the greatest (or close to the highest) overlap. To assess the accuracy of the predicted bitter and umami patterns, we checked if they have greater overlaps with bitter and umami peptides from test sets compared to overlaps with an all-hydrophobic bitter baseline peptide and an all-acidic umami baseline peptide, respectively, and, also, to a peptide with randomly chosen residues. We used this method to assess the accuracy of our predicted patterns because our goal is to reveal generic residue patterns rather than predicting whether an individual peptide has bitter or umami taste. Compared to QSAR and ML methods, which typically correlate aggregate physicochemical properties, our method predicts statistically robust and easily interpretable residue patterns that are connected to the tastes of peptides. In addition, our method prioritizes the role of relative sequential positions of the residues on a peptide's taste; this is difficult to capture in a QSAR framework. Taken together, our method systematically expands the currently known set of bitter and umami residue patterns and suggests a way to identify residue patterns that can be responsible for bitter or umami taste in a peptide or a protein.

METHODS

Database of Labeled Peptides. In this work, we relied on bitter and umami peptides collected from the existing literature. Our principal source is the database of 299 bitter and 140 umami peptides provided by Charoenkwan et al., who compiled the list from experimentally validated data sets and the literature.[36] To the list of bitter peptides of Charoenkwan et al. we added 24 new bitter peptides that we found in Ney's paper.[9] As for umami peptides, we added 12 more peptides from the literature[39,40] to the list of Charoenkwan et al. As more than 90% of the collected peptides are composed of 2−10 amino acid residues, we discarded single-residue peptides and peptides with more than 10 residues. This resulted in 292 bitter and 146 umami peptides that were used in this work.

Coarse-Grained Representation. The size of the sequence spaces of peptides, which are made from a combination of the 20

https://doi.org/10.1021/acsfoodscitech.2c00251
ACS Food Sci. Technol. 2022, 2, 1773−1780

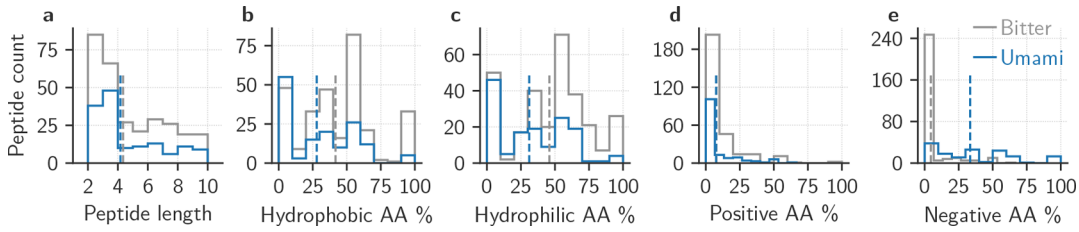

Figure 2. Physicochemical properties of bitter (gray) and umami (blue) peptides from the compiled database. Histograms of (a) peptide lengths and of percentage compositions of amino acid (AA) residue types, (b) hydrophobic, (c) hydrophilic, (d) positively charged, and (e) negatively charged, that comprise the database peptides. While bitter peptides are rich in hydrophobic residues (panel b), umami peptides mostly contain negative and polar residues (panels e and c, respectively). The vertical dashed lines indicate mean values of the distributions.

canonical amino acids, grows as 20^n, where n is the number of residues in the longest allowed peptide. As we have only 292 bitter and 146 umami peptides, we need to reduce the size of the peptide sequence space to make reliable predictions. One way of reducing the size is by coarse-graining the amino acids. We did this by classifying each amino acid into one of the four classes: hydrophobic (H), polar and hydrophilic (P), positively charged (+), or negatively charged (−). We have used the Kyte−Doolittle (KD) hydrophobicity scale to find hydrophobic (KD hydrophobicity > 0) and hydrophilic (KD hydrophobicity < 0) residues. In this scheme, at a physiological pH of 7.4, the 20 canonical amino acids are classified as follows: hydrophobic (H), {Ala, Cys, Ile, Leu, Met, Phe, Val}; hydrophilic (P), {Asn, Gln, Gly, Pro, Ser, Thr, Trp, Tyr}; positively charged (+), {Arg, His, Lys}; and negatively charged (−), {Asp, Glu}. For example, in our representation scheme, the umami peptide LLLPGELAK is represented as "HHHPP−HH+". With this representation, the size of the possible peptide sequence space decreases drastically. For example, the number of possible dipeptides decreases from 20^2 = 400 to 4^2 = 16. We converted all of the collected peptides from the literature to coarse-grained sequences and then proceeded to construct a library of coarse-grained peptides.

Library of Coarse-Grained Peptides. To extend the prediction beyond the peptide data set with which we started, we need new peptide sequences. To systematically generate new peptide sequences, we constructed seven increasingly larger peptide libraries formed by repeating a fixed set of coarse-grained patterns. While each of these libraries produced two best, i.e., most overlapped, matching patterns for bitter and umami peptides, we also compared patterns from these seven libraries. In this way, we can avoid choosing an unnecessarily large library when a smaller one performs comparably; i.e., we will not overfit. We do not seek a very small library of peptides either, as that will lead to underfitting. In the Results and Discussion, we will see how until N = 3 the libraries underfit the data while beyond N = 5 the predictive power of the libraries saturates.

We constructed each library in four steps. First, we fixed the maximum length (N) of the repeating patterns. Second, we generated a list of all $\sum_{i=1}^{N} 4^i$ possible combinatorial patterns containing N or fewer coarse-grained residues. Third, we repeated each pattern with itself to form an arbitrarily long (set to 420 residues in this work) peptide. Finally, we kept unique full peptides in the final library. Generating the library peptides in this way kept the number of peptides in each library low, which aided in deriving statistically robust patterns. Allowing for randomly selected neighbors of the repeating unit in the library would rapidly increase its size; this will lead to increased uncertainty of the predicted patterns.

For example, in the N = 1 library, there are only four repeating patterns, {H, P, + , −}, and only four unique peptides, {HH···, PP···, + +···, and −−···}. For N = 3, the library has $\sum_{i=1}^{3} 4^i$ = 84 repeating patterns such as {H, −P, +−H,···}, and they can combine to generate 76 unique peptides, {HH···, −P−P···, +−H+−H···, ···}. In this way, we went up to N = 7 and constructed seven libraries. We stopped at N = 7 because as we increase N, the number of peptides in the library increases sharply, and we will risk overfitting the data. For example, the N = 7 library has 21 844 repeating patterns and 21 736 unique peptides. Now to compare the library peptides and labeled peptides, we need an index that can measure the similarity between any two coarse-grained sequences.

Sequence Overlap Index. Following Schilling et al.,[38] we defined an overlap index between two coarse-grained peptides, X and Y, as the ratio of the number of position-dependent residue matches ($|X \cap Y|_{seq}$) and the length of the smaller peptide:

$$I(\mathrm{X}, \mathrm{Y}) = \frac{|X \cap Y|_{seq}}{\min(|X|, |Y|)} \quad (1)$$

where $|X|$ denotes the number of residues in peptide X. For example, two peptides, −H++ and +H+−−, have two position-dependent residue matches (H and + at positions 2 and 3, respectively), so I(−H ++, +H+−−) = 2/4 = 0.5.

The overlap index, I, allows us to define a surrogate measure for bitter and umami tastes of a library peptide. For a coarse-grained library peptide, we defined bitterness (umami-ness) as the average overlap between the library peptide and all coarse-grained bitter (umami) peptides from the compiled data set.

Best Patterns and Their Validation. With this surrogate measure for the bitterness (umami-ness) in our formalism, we considered each of the seven libraries in turn, computed the bitterness (umami-ness) of its constituent peptides, and then sorted them to find five peptides with the largest bitterness (umami-ness) values. (For the N = 1 library, there are only four possible peptides; we chose the best one.) We found that the bitterness (umami-ness) values of all five chosen bitter (umami) peptides were similar, so we did not assign a higher weight to the most bitter (umami) peptide when determining the best bitter (umami) pattern. At each sequence position of the repeating patterns of these five peptides, we found the most frequent residue type (from H, P, + , and −). By sequentially merging these most frequent residue types, we finally obtained the best pattern. For example, the most frequent residue types in the five most bitter peptides from the N = 3 library are H, P, and P (see Figure 3a3). As a result, our predicted N = 3 bitter pattern is "HPP". Note that this composed best pattern has the same length as the maximum length of the repeating patterns (N) of a library. Thus, the best pattern depends on the library, the taste type (bitter or umami), and the external database of peptides that we use to measure the taste (bitter- or umami-ness). Therefore, choosing a well-curated and sufficiently large (so that the prediction errors are small) set of taste-labeled peptides is crucial in our data-driven approach.

To ensure reproducibility of the predicted pattern, we performed extensive out-of-sample testing. To this end, we split the external taste-labeled peptide database into an 80% training set and a 20% test set using stratified random sampling. We used stratified sampling to keep the ratio of bitter and umami peptides roughly similar in the training and test sets. Otherwise, a random sampling will pick more bitter peptides than umami ones because we have 292 bitter and 146 umami peptides in the peptide database. This, in turn, will result in imbalanced training data for identifying the bitter and umami patterns. Finally, we obtained the best bitter (umami) pattern for each of the seven libraries by using the training set peptides. To gather enough statistics, we repeated this procedure 500 times.

Baseline Patterns. Following the literature, we set a peptide with all hydrophobic residues as the baseline bitter peptide[6,9,41] and a

https://doi.org/10.1021/acsfoodscitech.2c00251
ACS Food Sci. Technol. 2022, 2, 1773−1780

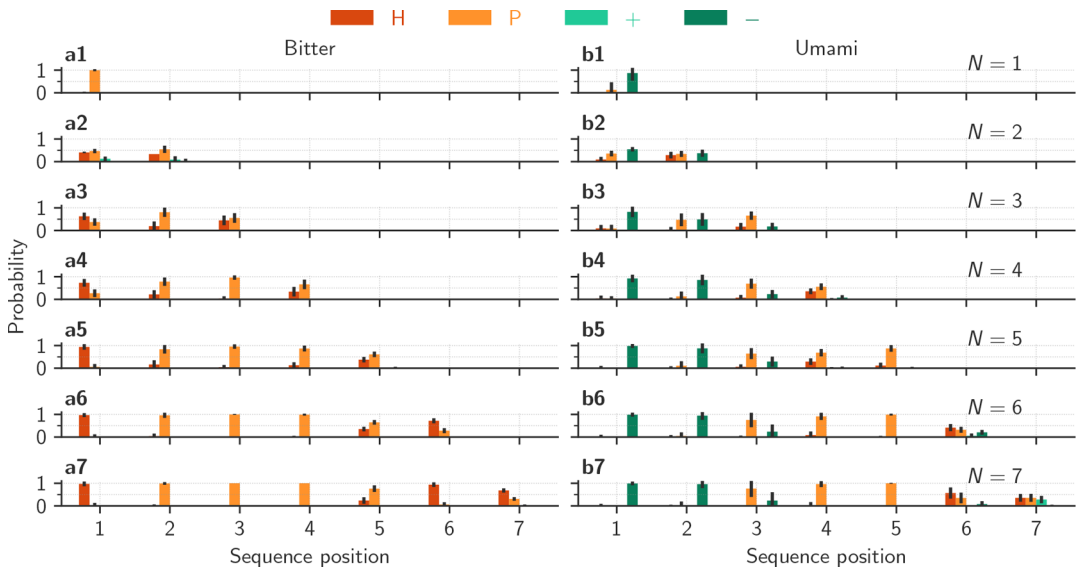

Figure 3. Predicted coarse-grained sequence patterns that have the greatest overlap with bitter (left panel, a1−a7) and umami (right panel, b1−b7) peptides. The maximum lengths of the peptide library patterns (N) increase from top to bottom in each panel. The colored bars show the probabilities of finding a specific residue type at a given sequence position of the predicted pattern. The color codes are displayed at the top. Black error bars indicate standard deviations computed over 500 training sets.

peptide with all negatively charged residues as the baseline umami peptide.[30,32,42,43] While the compositions of longer umami peptides are known to be varied,[30] the importance of the presence of negatively charged acidic residues is generally well accepted in the community. In addition, setting a baseline will allow us to quantitatively assess the conjecture that the relative locations of the residues do not affect a peptide's taste.[6,9] In Figure 1, we present the main steps of the complete pipeline that we used in this paper.

RESULTS AND DISCUSSION

Physicochemical Properties of the Database Peptides. We first analyzed the physicochemical properties of the labeled coarse-grained peptides from the compiled database. Both bitter and umami peptides from the assembled data set have approximately four residues, on average (Figure 2a). Interestingly, ∼54% of the database peptides have only two or three residues. This reflects the fact that while consensus often exists regarding the tastes of shorter peptides, there are some disagreements regarding the taste of longer peptides, especially for longer umami peptides.[30] This makes the exploration of longer peptide patterns even more relevant for the food industry because it can lead to the discovery of new bitter and umami peptides.[43]

To determine the relative abundance of the four coarse-grained residue types (i.e., H, P, +, and −), we computed the corresponding histograms (Figure 2b−e) of their presence (in percent) in the bitter and umami peptides from the compiled database of peptides. We found that in both bitter and umami peptides, the hydrophilic residues are abundantly present (Figure 2c) while positively charged residues are mostly absent (Figure 2d). On average, bitter peptides contain more hydrophobic residues than the umami peptides: ∼42% compared to ∼28% (Figure 2b). The umami peptides are richer in negatively charged amino acids (∼33%) compared to the bitter peptides (∼5%) (Figure 2e). Both of these observations are in accord with the current consensus that the hydrophobic residues dominate bitter peptides,[9,41] while negative residues dominate umami peptides.[30,32,42,43] We, however, also note the significant presence of hydrophilic residues in bitter peptides. This observation asks for a systematic analysis of residue patterns in the primary sequence of coarse-grained bitter and umami peptides. In the next section, we present our findings from such an analysis.

Predicted Bitter and Umami Patterns. Figure 3 shows the patterns that best predict bitterness (left panel, Figure 3a1−a7) and umami-ness (right panel, Figure 3b1−b7) for all seven libraries, from N = 1 to N = 7. For N = 1, we have the smallest library, and the predicted pattern simply picks the most common residue type in bitter and umami peptides from the training sets. For bitter peptides, hydrophilic residues are most common (∼46%), followed by hydrophobic residues (∼42%) (Figure 2c,b), while for umami peptides, negative residues are most common (∼33%), followed by hydrophilic residues (∼31%) (Figure 2e,c). Therefore, it is not surprising to find that for the N = 1 library our algorithm predicts hydrophilic (P) and negative residues (−) as the best patterns for bitter and umami peptides, respectively (Figure 3a1,b1). This prediction, however, is an example of underfitting the data as we did not allow for enough complexity (i.e., enough peptides) in our library.

The lack of complexity affects the results for the library with N = 2, too. We find the residue types compete closely at both sequence positions of the predicted patterns (Figure 3a2,b2). The algorithm predicts an all-hydrophilic residue pattern, "PP", as the best bitter pattern and an all-negative residue pattern, "−−", as the best umami pattern. Though for umami peptides the predicted pattern matches with the literature consensus,[30] for bitter peptides it does not, which is similar to the result from the N = 1 library. The presence of longer peptides with many hydrophilic residues in our bitter peptide data set subdues the expected "HH" pattern. Interestingly, however, the "HH" pattern is predicted to be the second-best bitter pattern (Figure 3a2). This compares well with the findings of Xu et al.,[21] who found the dominant presence of hydrophobic residues in both positions of bitter dipeptides. Observe that we derived the patterns in Figure 3 using the *surrogate measures* for tastes that we defined; we did not have labels or bitterness

https://doi.org/10.1021/acsfoodscitech.2c00251
ACS Food Sci. Technol. 2022, 2, 1773−1780

values for the combinatorially constructed library peptides. The similarity between the second-best bitter peptide pattern from our analysis and the findings of Xu et al.[21] computed with 12 amino acid descriptors on labeled bitter dipeptides data set demonstrates the *strength of the surrogate measure* we defined.

From the $N = 3$ library onward, we started to obtain more robust predictions across the training sets. We found that the best predicted bitter and umami patterns for the $N = 3$ library are "HPP" and "−−P" (with "HPH" and "−PP" as close second-best patterns), respectively (Figure 3a3,b3). Interestingly, the close second-best bitter pattern, "HPH", matches well with the result of Xu et al.,[21] who found hydrophobicity of the C-terminal residue and electronic properties of the second residue are important for bitterness in tripeptides. The bitter pattern we obtained from the $N = 4$ library, "HPPP" (Figure 3a4), also compares well at residue positions 1, 3, and 4 with the findings of Xu et al. for tetrapeptides. However, for the second position, they found that hydrophobicity plays a role; we obtained a polar residue in our predicted pattern.

We could not find systematic sequential residue type analysis for bitter peptides with more than four residues and umami peptides with more than three residues. As a consequence, the longer patterns we found, {HPPPP, HPPPPH, HPPPPHH} for bitter (Figure 3a5−a7) and {−−PP, −−PPP, −−PPPH, and −−PPPHH} for umami peptides (Figure 3b4−b7), can provide useful templates for exploring new bitter and umami peptides. Note that different residue types closely compete for the sixth and seventh residue positions of the $N = 6$ and 7 umami patterns. Also, the dominant residue type at each sequence position mostly remains conserved across the $N \geq 2$ libraries. The presence of several predictions from these libraries naturally leads to the question of which library and its predicted bitter and umami patterns should be selected if we need to pick one "best" pattern for each taste.

Selecting the Minimal Peptide Pattern. To answer this question, we computed the average overlaps of the predicted bitter and umami patterns, which we found using the training sets, with the corresponding test sets' bitter and umami peptides, for each of the seven libraries (Figure 4a,b, shown as red dots). We also computed the average overlaps of the baseline bitter ("HH···") and the baseline umami ("−−···") patterns (black dots, Figure 4a,b) and the average overlaps of a peptide with randomly chosen residues (gray dots, Figure 4a,b). All averages were computed with the test sets' bitter and umami peptides. The standard errors of the mean of the averages are smaller than the sizes of the dots.

For bitter peptides, the predicted patterns for the smallest two libraries ($N = 1$ and 2) are entirely made of hydrophilic residues (Figure 3a1,a2). The average overlaps for those predicted patterns (red dots, Figure 4a) are larger than the all-hydrophobic baseline pattern (black dots, Figure 4a). This counterintuitive result, however, is an artifact of having a large number of hydrophilic residues in the bitter peptide data set, as discussed above. For umami peptides, the smallest predicted patterns ($N = 1$ and 2) consist entirely of negative residues (Figure 3b1,b2). Because we have considered an all-negative residue as our umami baseline, the overlaps of the predicted patterns and baseline patterns with the test sets' umami peptides match (overlapped red and black dots, Figure 4b). For both bitter and umami peptides, with an increase in N, the overlaps increase until $N = 5$; then they mostly plateau. From this observation, we chose the $N = 5$ library as the minimal library that is large enough to have enough coarse-grained

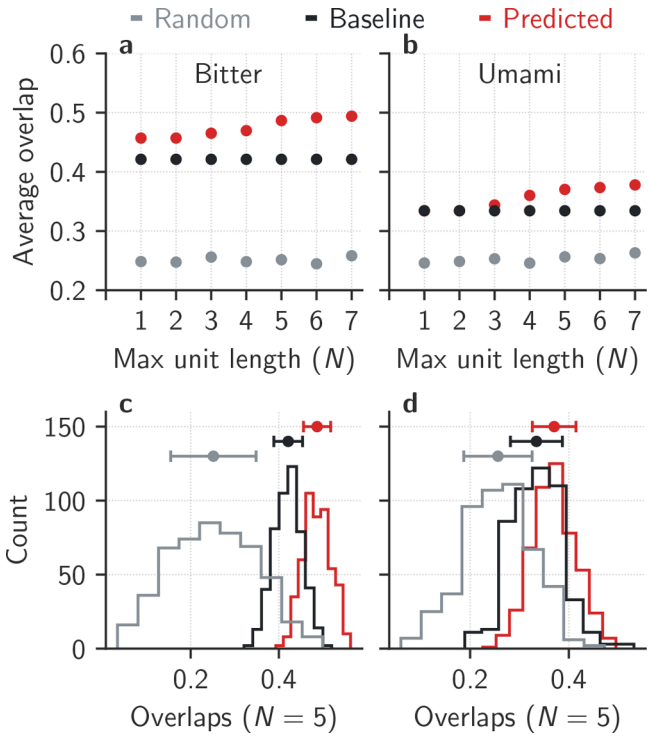

Figure 4. Increasing the maximum length, N, of the residue patterns, i.e., allowing for more complexity in pattern libraries, does not lead to a better pattern above $N = 5$. Panels a and b show average overlaps of the predicted bitter and umami patterns from each library (red dots), the baseline bitter ("HH···") and umami ("−−···") patterns (black dots), and a randomly generated residue pattern (gray dots) with bitter and umami peptides from 500 test sets, respectively. The standard errors of the mean are smaller than the sizes of the dots. Panels c and d show histograms of overlaps of the $N = 5$ library's predicted bitter ("HPPPP") and umami ("−−PPP") patterns with bitter and umami peptides from the test sets, respectively. The ordinates denote the number of test sets with overlaps in a certain range. Horizontal lines above the histograms indicate the means and standard deviations of the distributions.

peptide patterns that it neither underfits the data nor has more peptide patterns than necessary, given the model complexity. In addition, there are disagreements in the community regarding the umami tastes of some small peptides that are present in the published resources that we used[36] (see ref 44 for a review); even if those peptides are removed from the database, the predicted $N = 5$ umami pattern remains the same. See the Supporting Information for a more in-depth discussion about the disputed umami peptides.

To demonstrate how the $N = 5$ peptide library overlaps with peptides from the test sets, we computed the histograms of the average overlaps of the $N = 5$ library's predicted patterns, "HPPPP" for bitter and "−−PPP" for umami peptides, with bitter and umami peptides from the test sets (Figure 4c,d). The predicted patterns (in red) clearly are improvements over the baseline patterns (in black) and randomly chosen patterns (in gray). The improvements are more pronounced for the bitter pattern than for the umami pattern. This analysis demonstrates the accuracy of our method, which is designed to identify generic sequence patterns rather than predicting a property or classifying a new peptide. Taken together, these observations imply that the predicted bitter and umami

https://doi.org/10.1021/acsfoodscitech.2c00251
ACS Food Sci. Technol. 2022, 2, 1773−1780

patterns can act as promising design templates for bitter and umami peptides.

For bitter and umami peptides, our analysis offers a set of coarse-grained residue patterns that are possibly linked to the peptides' bitter and umami tastes. The predicted bitter patterns can be useful, for example, in finding short, possibly bitter-causing, patterns in longer proteins. As test cases, we considered two proteins that are associated with bitter taste: Patatin-T5 (UniProt ID P15478)[45] and Legumin-A (UniProt ID P02857).[46] We first converted the primary sequence of these proteins to a sequence of coarse-grained residues (see the Supporting Information for the full primary and coarse-grained sequences) and then searched for the predicted pattern from N = 5 library ("HPPPP"). The search resulted in eight five-residue-long sequence segments in Patatin-T5: {ATTNS (2), ATTSS (16), IGGTS (73), ITTPN (86), FQSSG (114), ATNTS (200), LGTGT (258), LTGTT (324)}. The numbers in parentheses denote the positions of the first residues of the segments in the protein's primary structure. For Legumin-A, we found five such short sequences: {IQQGN (93), IGPSS (347), CNGNT (418), VPQNY (436), AGTSS (468)}. The analysis provides a possible experimental way to determine where to cleave the protein to decrease its bitterness. Accordingly, it may be useful to study these short sequences further in experiments.

In summary, in this study we aim to build a simple and interpretable model that identifies generic residue patterns that are prevalent in bitter and umami peptides and, possibly, evoke those tastes. By coarse-graining the 20 canonical amino acid residues into four physicochemically relevant classes [hydrophobic (H), hydrophilic (P), positively charged (+), and negatively charged (−)], we drastically reduced the dimensionality of the peptide sequence space. With these coarse-grained residues, we systematically built seven increasingly larger, more complex, combinatorial libraries of peptides. We compiled and coarse-grained a library of bitter and umami peptides from the literature. We used a sequence overlap index[38] I (eq 1) to compare library peptides with the bitter and umami peptides, at a coarse-grained level. Importantly, the overlap index allowed us to compute the average overlaps of a library peptide with bitter and umami peptides from the compiled data set and define those average overlaps as surrogate measures for bitterness and umami-ness, respectively. For each of the seven peptide libraries, the best (i.e., most overlapping) bitter and umami patterns provided us with the predicted patterns. We checked the robustness of the predicted patterns through 80%−20% train−test splitting, and we reported the patterns that we obtained after averaging over 500 such splits. By comparing the average overlap of the predicted patterns with test set peptides, we found the minimal N = 5 library whose patterns, "HPPPP" for bitter and "−−PPP" for umami peptides, have almost equal overlap compared to larger libraries. In addition, we found that these predicted patterns represent the known bitter and umami peptides more accurately than baseline patterns—all hydrophobic residues for bitter peptides and all negatively charged residues for umami peptides—and peptides with randomly chosen residues.

A QSAR or ML model typically connects aggregate properties; they are commonly not used as a tool for identifying residue patterns in peptides. They are useful, for example, when the goal is to determine how taste is affected by specific physicochemical properties such as peptide conformations or residue charge distributions. Our method is not designed to answer those questions. Thus, instead of competing with QSAR and ML, our method complements them by providing a way to rapidly identify taste-inducing residue patterns hidden in peptides and proteins. While we could include more physicochemical descriptors, doing so increases the number of coarse-grained units, which increases the uncertainty in the predicted patterns. By coarse-graining the peptide space while retaining key physicochemical properties such as hydrophobicity and charge, our method provides statistically robust predictions with limited experimental data.

The potential of the proposed coarse-grained peptide pattern search is demonstrated in Figure 5. The amino acids

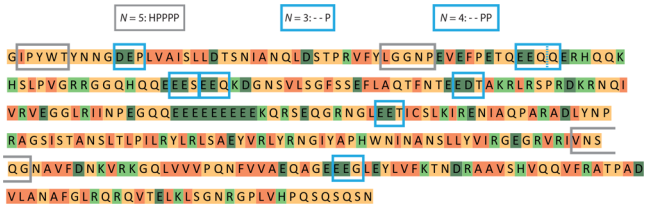

Figure 5. Illustration of the method along the protein Legumin-B (Uniprot ID P16078), which is present in broad beans (*Vicia faba*). The primary structure is colored according to our scheme. Possible bitter and umami peptides are framed in gray and blue boxes, respectively.

of the primary structure of Legumin-B (Uniprot ID P16078), a storage protein from broad beans (*Vicia faba*), are colored according to the scheme from Figure 3. Employing the results from Figure 3, a search for the pattern HPPPP, the N = 5 pattern with bitter taste, provides three peptides: IPYWT (2), LGGNP (38), and VNSQG (238) (the numbers denote the positions of the peptides in the sequence). They are marked with gray boxes in Figure 5. The umami potential of this protein is very high: a complete hydrolysis provides 38 glutamic and 10 aspartic acids; they constitute >14% of the protein's total amino acid content. While there are no matches for the N = 5 umami pattern "−−PPP" in this particular protein, a search for the N = 4 pattern "−−PP" returns one peptide, EEQQ (51), and the N = 3 umami pattern "−−P" returns seven peptides: DEP (11), EEQ (51), EES (76), EEQ (79), EDT (104), EET (157), and EEG (271). All of them are marked with blue boxes in Figure 5. Note that short bitter and umami pattern matches from the reversed and cleaved primary structure of the protein may also have bitter and umami tastes. These predicted patterns can be used, for example, in experimentally designing new umami peptides or in finding short bitter or umami segments in a long protein. Also, specially chosen or designed enzymes that can cleave proteins at defined peptide bonds may be appropriate for guiding taste profiles of hydrolysates; they may then be used as new flavors of plant origin.

In conclusion, our coarse-grained peptide search provides a simple and quick screening for the taste potential of proteins. In addition, by predicting coarse-grained residue patterns instead of specific peptides, it offers considerable latitude when designing new bitter and umami peptides. The new peptides can be validated via *in silico* methods such as molecular docking and via sensory evaluations following *in vitro* synthesis. Taken together, our method can help both in locating possible bitter and umami patterns in a long protein and in designing

https://doi.org/10.1021/acsfoodscitech.2c00251
ACS Food Sci. Technol. 2022, 2, 1773−1780

new bitter and peptides. In addition, it is not tied to only bitter or umami tastes; it can be readily applied to identify residue patterns for other taste modalities of peptides. The need to identify peptide taste profiles is becoming increasingly important for plant-based meat surrogates. However, one point remains clear: many umami peptides, particularly the longer ones, of animal origin are signatures of specific proteins, and it remains difficult and seemingly impossible to find those in plant proteins.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsfoodscitech.2c00251.

Original and coarse-grained primary structures of two bitter proteins, Patatin-T5 and Legumin-A, and discussion of a dispute regarding the umami taste of some peptides (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Thomas A. Vilgis − *Max Planck Institute for Polymer Research, 55128 Mainz, Germany;* orcid.org/0000-0003-2101-7410; Email: vilgis@mpip-mainz.mpg.de

Arghya Dutta − *Max Planck Institute for Polymer Research, 55128 Mainz, Germany;* Present Address: A.D.: Institute of Biochemistry II, Faculty of Medicine, Goethe University, 60590 Frankfurt, Germany; orcid.org/0000-0003-2116-6475; Email: argphy@gmail.com

Author

Tristan Bereau − *Van 't Hoff Institute for Molecular Sciences and Informatics Institute, University of Amsterdam, 1090 GD Amsterdam, The Netherlands;* orcid.org/0000-0001-9945-1271

Complete contact information is available at: https://pubs.acs.org/10.1021/acsfoodscitech.2c00251

Funding

Open access funded by Max Planck Society. A.D. acknowledges support by BiGmax, the Max Planck Society's Research Network on Big-Data-Driven Materials-Science. T.B. was partially supported by the Emmy Noether program of the Deutsche Forschungsgemeinschaft (DFG).

Notes

All codes and data are openly available on GitHub at https://github.com/arghyadutta/patterns-to-taste.

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors acknowledge localCIDER,[47] Matplotlib,[48] Scikit-learn,[49] and Pandas,[50] four open-source packages that were used in this work.

■ REFERENCES

(1) Chandrashekar, J.; Hoon, M. A.; Ryba, N. J.; Zuker, C. S. The receptors and cells for mammalian taste. *Nature* **2006**, *444*, 288−294.

(2) Drewnowski, A. Taste preferences and food intake. *Annual Review of Nutrition* **1997**, *17*, 237−253.

(3) Maehashi, K.; Huang, L. Bitter peptides and bitter taste receptors. *Cell. Mol. Life Sci.* **2009**, *66*, 1661−1671.

(4) Dagan-Wiener, A.; Di Pizio, A.; Nissim, I.; Bahia, M. S.; Dubovski, N.; Margulis, E.; Niv, M. Y. BitterDB: taste ligands and receptors database in 2019. *Nucleic Acids Res.* **2019**, *47*, D1179−D1185.

(5) Rojas, C.; Ballabio, D.; Pacheco Sarmiento, K.; Pacheco Jaramillo, E.; Mendoza, M.; García, F. ChemTastesDB: A curated database of molecular tastants. *Food Chem.: Mol. Sci.* **2022**, *4*, 100090.

(6) Matoba, T.; Hata, T. Relationship between bitterness of peptides and their chemical structures. *Agric. Biol. Chem.* **1972**, *36*, 1423−1431.

(7) Lemieux, L.; Simard, R. E. Bitter flavour in dairy products. I. A review of the factors likely to influence its development, mainly in cheese manufacture. *Le Lait* **1991**, *71*, 599−636.

(8) Toelstede, S.; Hofmann, T. Sensomics mapping and identification of the key bitter metabolites in Gouda cheese. *J. Agric. Food Chem.* **2008**, *56*, 2795−2804.

(9) Ney, K. H.*Bitterness of Peptides: Amino Acid Composition and Chain Length*; ACS Symposium Series; Food Taste Chemistry; American Chemical Society: Washington, DC, 1979; Vol. *115*; pp 149−173 DOI: 10.1021/bk-1979-0115.ch006.

(10) Yamaguchi, S.; Ninomiya, K. Umami and food palatability. *Journal of nutrition* **2000**, *130*, 921S−926S.

(11) Zhao, C. J.; Schieber, A.; Gänzle, M. G. Formation of taste-active amino acids, amino acid derivatives and peptides in food fermentations-A review. *Food Research International* **2016**, *89*, 39−47.

(12) Lioe, H. N.; Selamat, J.; Yasuda, M. Soy sauce and its umami taste: a link from the past to current situation. *J. Food Sci.* **2010**, *75*, R71−R76.

(13) Kusumoto, K.-I.; Yamagata, Y.; Tazawa, R.; Kitagawa, M.; Kato, T.; Isobe, K.; Kashiwagi, Y. Japanese traditional Miso and Koji making. *Journal of Fungi* **2021**, *7*, 579.

(14) Heres, A.; Toldrá, F.; Mora, L. Characterization of Umami Dry-Cured Ham-Derived Dipeptide Interaction with Metabotropic Glutamate Receptor (mGluR) by Molecular Docking Simulation. *Applied Sciences* **2021**, *11*, 8268.

(15) Wang, H.; Xu, J.; Liu, Q.; Xia, X.; Sun, F.; Kong, B. Effect of the protease from *Staphylococcus carnosus* on the proteolysis, quality characteristics, and flavor development of Harbin dry sausage. *Meat Science* **2022**, *189*, 108827.

(16) Temussi, P. A. The good taste of peptides. *Journal of Peptide Science* **2012**, *18*, 73−82.

(17) Van Boekel, M. Formation of flavour compounds in the Maillard reaction. *Biotechnology advances* **2006**, *24*, 230−233.

(18) Jünger, M.; Mittermeier-Kleßinger, V. K.; Farrenkopf, A.; Dunkel, A.; Stark, T.; Fröhlich, S.; Somoza, V.; Dawid, C.; Hofmann, T. Sensoproteomic Discovery of Taste-Modulating Peptides and Taste Re-engineering of Soy Sauce. *J. Agric. Food Chem.* **2022**, *70*, 6503−6518.

(19) Hellberg, S.; Sjoestroem, M.; Skagerberg, B.; Wold, S. Peptide quantitative structure-activity relationships, a multivariate approach. *J. Med. Chem.* **1987**, *30*, 1126−1135.

(20) Lee, E. Y.; Wong, G. C.; Ferguson, A. L. Machine learning-enabled discovery and design of membrane-active peptides. *Bioorg. Med. Chem.* **2018**, *26*, 2708−2718.

(21) Xu, B.; Chung, H. Y. Quantitative structure-activity relationship study of bitter di-, tri- and tetrapeptides using integrated descriptors. *Molecules* **2019**, *24*, 2846.

(22) Kim, H.-O.; Li-Chan, E. C. Y. Quantitative structure−activity relationship study of bitter peptides. *J. Agric. Food Chem.* **2006**, *54*, 10102−10111.

(23) Charoenkwan, P.; Nantasenamat, C.; Hasan, M. M.; Moni, M. A.; Lio, P.; Shoombuatong, W. iBitter-Fuse: A Novel Sequence-Based Bitter Peptide Predictor by Fusing Multi-View Features. *Int. J. Mol. Sci.* **2021**, *22*, 8958.

(24) Wu, J.; Aluko, R. E. Quantitative structure-activity relationship study of bitter di-and tri-peptides including relationship with angiotensin I-converting enzyme inhibitory activity. *Journal of peptide science: an official publication of the European Peptide Society* **2007**, *13*, 63−69.

(25) Yin, J.; Diao, Y.; Wen, Z.; Wang, Z.; Li, M. Studying peptides biological activities based on multidimensional descriptors (E) using support vector regression. *International Journal of Peptide Research and Therapeutics* **2010**, *16*, 111−121.
(26) Charoenkwan, P.; Yana, J.; Schaduangrat, N.; Nantasenamat, C.; Hasan, M. M.; Shoombuatong, W. iBitter-SCM: Identification and characterization of bitter peptides using a scoring card method with propensity scores of dipeptides. *Genomics* **2020**, *112*, 2813−2822.
(27) Charoenkwan, P.; Nantasenamat, C.; Hasan, M. M.; Manavalan, B.; Shoombuatong, W. BERT4Bitter: a bidirectional encoder representations from transformers (BERT)-based model for improving the prediction of bitter peptides. *Bioinformatics* **2021**, *37*, 2556−2562.
(28) Ishibashi, N.; Sadamori, K.; Yamamoto, O.; Kanehisa, H.; Kouge, K.; Kikuchi, E.; Okai, H.; Fukui, S. Bitterness of phenylalanine- and tyrosine-containing peptides. *Agric. Biol. Chem.* **1987**, *51*, 3309−3313.
(29) Smyth, H.; Cozzolino, D. Instrumental methods (spectroscopy, electronic nose, and tongue) as tools to predict taste and aroma in beverages: advantages and limitations. *Chem. Rev.* **2013**, *113*, 1429−1440.
(30) Qi, L.; Gao, X.; Pan, D.; Sun, Y.; Cai, Z.; Xiong, Y.; Dang, Y. Research progress in the screening and evaluation of umami peptides. *Comprehensive Reviews in Food Science and Food Safety* **2022**, *21*, 1462−1490.
(31) Yu, Z.; Kang, L.; Zhao, W.; Wu, S.; Ding, L.; Zheng, F.; Liu, J.; Li, J. Identification of novel umami peptides from myosin via homology modeling and molecular docking. *Food chemistry* **2021**, *344*, 128728.
(32) Wang, W.; Cui, Z.; Ning, M.; Zhou, T.; Liu, Y. *In-silico* investigation of umami peptides with receptor T1R1/T1R3 for the discovering potential targets: A combined modeling approach. *Biomaterials* **2022**, *281*, 121338.
(33) Liang, L.; Duan, W.; Zhang, J.; Huang, Y.; Zhang, Y.; Sun, B. Characterization and molecular docking study of taste peptides from chicken soup by sensory analysis combined with nano-LC-Q-TOF-MS/MS. *Food Chem.* **2022**, *383*, 132455.
(34) Zhang, F.; Klebansky, B.; Fine, R. M.; Xu, H.; Pronin, A.; Liu, H.; Tachdjian, C.; Li, X. Molecular mechanism for the umami taste synergism. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 20930−20934.
(35) Charoenkwan, P.; Nantasenamat, C.; Hasan, M. M.; Moni, M. A.; Manavalan, B.; Shoombuatong, W. UMPred-FRL: A New Approach for Accurate Prediction of Umami Peptides Using Feature Representation Learning. *International Journal of Molecular Sciences* **2021**, *22*, 13124.
(36) Charoenkwan, P.; Yana, J.; Nantasenamat, C.; Hasan, M. M.; Shoombuatong, W. iUmami-SCM: A Novel Sequence-Based Predictor for Prediction and Analysis of Umami Peptides Using a Scoring Card Method with Propensity Scores of Dipeptides. *J. Chem. Inf. Model.* **2020**, *60*, 6666−6678.
(37) Bo, W.; Chen, L.; Qin, D.; Geng, S.; Li, J.; Mei, H.; Li, B.; Liang, G. Application of quantitative structure-activity relationship to food-derived peptides: Methods, situations, challenges and prospects. *Trends in Food Science & Technology* **2021**, *114*, 176−188.
(38) Schilling, C.; Mack, T.; Lickfett, S.; Sieste, S.; Ruggeri, F. S.; Sneideris, T.; Dutta, A.; Bereau, T.; Naraghi, R.; Sinske, D.; Knowles, T. P. J.; Synatschke, C. V.; Weil, T.; Knöll, B. Sequence-Optimized Peptide Nanofibers as Growth Stimulators for Regeneration of Peripheral Neurons. *Adv. Funct. Mater.* **2019**, *29*, 1809112.
(39) Shiyan, R.; Liping, S.; Xiaodong, S.; Jinlun, H.; Yongliang, Z. Novel umami peptides from tilapia lower jaw and molecular docking to the taste receptor T1R1/T1R3. *Food Chem.* **2021**, *362*, 130249.
(40) Liu, Z.; Zhu, Y.; Wang, W.; Zhou, X.; Chen, G.; Liu, Y. Seven novel umami peptides from *Takifugu rubripes* and their taste characteristics. *Food Chem.* **2020**, *330*, 127204.
(41) Iwaniak, A.; Minkiewicz, P.; Darewicz, M.; Hrynkiewicz, M. Food protein-originating peptides as tastants - Physiological, technological, sensory, and bioinformatic approaches. *Food Research International* **2016**, *89*, 27−38.
(42) Rhyu, M.-R.; Kim, E.-Y. Umami taste characteristics of water extract of *Doenjang*, a Korean soybean paste: Low-molecular acidic peptides may be a possible clue to the taste. *Food Chem.* **2011**, *127*, 1210−1215.
(43) Yu, X.; Zhang, L.; Miao, X.; Li, Y.; Liu, Y. The structure features of umami hexapeptides for the T1R1/T1R3 receptor. *Food chemistry* **2017**, *221*, 599−605.
(44) Zhang, Y.; Venkitasamy, C.; Pan, Z.; Liu, W.; Zhao, L. Novel umami ingredients: Umami peptides and their taste. *J. Food Sci.* **2017**, *82*, 16−23.
(45) Spelbrink, R. E.; Lensing, H.; Egmond, M. R.; Giuseppin, M. L. Potato patatin generates short-chain fatty acids from milk fat that contribute to flavour development in cheese ripening. *Appl. Biochem. Biotechnol.* **2015**, *176*, 231−243.
(46) Real Hernandez, L. M.; Gonzalez de Mejia, E. Enzymatic production, bioactivity, and bitterness of chickpea (*Cicer arietinum*) peptides. *Comprehensive reviews in food science and food safety* **2019**, *18*, 1913−1946.
(47) Holehouse, A. S.; Das, R. K.; Ahad, J. N.; Richardson, M. O.; Pappu, R. V. CIDER: resources to analyze sequence-ensemble relationships of intrinsically disordered proteins. *Biophysical journal* **2017**, *112*, 16−21.
(48) Hunter, J. D. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering* **2007**, *9*, 90−95.
(49) Pedregosa, F.Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 2011, *12*, 2825−2830https://jmlr.org/papers/v12/pedregosa11a.html.
(50) McKinney, W.Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference*, 2010; pp 56−61 DOI: 10.25080/Majora-92bf1922-00a.