Novel Descriptors Derived from the Aggregation Propensity of Di- and Tripeptides Can Predict the Critical Aggregation Concentration of Longer Peptides

Saeed Zanganeh, Loghman Firoozpour, Soroush Sardari, Ali Afgar, Reza Ahangari Cohan,* and Nasir Mohajel*

ABSTRACT: Self-assembling amphiphilic peptides have recently received special attention in medicine. Nonetheless, testing the myriad of combinations generated from at least 20 coded and several hundreds of noncoded amino acids to obtain candidate sequences for each application, if possible, is time-consuming and expensive. Therefore, rapid and accurate approaches are needed to select candidates from countless combinations. In the current study, we examined three conventional descriptor sets along with a novel descriptor set derived from the simulated aggregation propensity of di- and tripeptides to model the critical aggregation concentration (CAC) of amphiphilic peptides. In contrast to the conventional descriptors, the radial kernel model derived from the novel descriptor set accurately predicted the critical aggregation concentration of the test set with a residual standard error of 0.10. The importance of aromatic side chains, as well as neighboring amino acids in the self-assembly, was emphasized by analysis of the influential descriptors. The addition of very long peptides (70–100 residues) to the data set decreased the model accuracy and changed the influential descriptors. The developed model can be used to predict the CAC of self-assembling amphiphilic peptides and also to derive rules to apply in designing novel amphiphilic peptides with desired properties.

INTRODUCTION

In recent years, amphiphilic peptides have received great attention due to their ability to form a variety of nanostructures.1 These peptides are composed of distinct hydrophilic and hydrophobic parts and, depending on the sequence, they exhibit different physicochemical and biochemical properties.2,3 These peptides were first introduced by Zhang et al. at the Massachusetts Institute of Technology.4 Initial studies were focused on a better understanding of their physicochemical properties; however, in the following years, numerous potential applications have been suggested for these materials, including drug5 and gene delivery,6 vaccination,7 regenerative medicine,8 and stabilization of membrane proteins.9 Despite the proposed biomedical applications, there are still many obstacles that need to be addressed for their realization. For example, self-assembling peptides show a low efficiency for drug and gene delivery, and/or peptides with lower critical aggregation concentration (CAC) values are needed to reduce the cost for regenerative applications.10 Therefore, discovery or design strategies are still needed to obtain sequences that provide the desired physicochemical properties for each application of amphiphilic peptides.

Meeting this demand by testing all possible combinations of 20 coded and several hundreds of noncoded amino acids with different lengths is an impossible task at the laboratory level. Therefore, various design and discovery strategies have been tried by researchers. There are two approaches to discover self-assembling peptides. Either amino acids are rationally chosen based on their known properties (rational design strategy)6,11 or the desired characteristics of the final system are considered to achieve the desired attributes (directed discovery strategy).12 Directed discovery consists of three different methodologies, namely sequence editing, computation, and dynamic libraries. The computational methodology has been used by Frederix et al. They used Martini coarse-grained molecular dynamics (CG-MD) simulations to identify sequences with higher aggregation propensity (AP) values via screening of di- and tripeptides containing coded amino acids.13,14 Ultimately, they suggested several design rules by observing sequences with the highest AP values. However, extension of the computational methods to longer peptides or sequences containing noncoded amino acids is not currently achievable because of the computational cost. Quantitative structure–property relationship (QSPR) is a less computa-
tionally intensive method that uses statistical modeling to create a mathematical relationship between a measurable property and its molecular descriptors. This method has been extensively used with success for critical micelle concentration (CMC) prediction of surfactants.\(^\text{15}\) Since surfactants are structurally comparable to amphiphilic peptides and both of them have a similar property (named CAC), the same strategies were used for studying their self-assembly. Tian et al. employed QSPR for CAC modeling of 32 amphiphilic peptides. They concluded that their novel MD-based descriptor, named the molecular dynamics-based hydrophobic cross-field (MD-HCF), shows a better performance in model building than descriptors generated by the CODESSA package.\(^\text{16}\) In another study, Guo et al. developed two different models for studying the flexibility and CAC of self-assembling peptides.\(^\text{17}\) Their data showed a negative correlation between flexibility and self-assembling power. Although these initial attempts reported a high correlation coefficient between the predicted and experimental CAC values, the small size of the database prevented the authors from defining independent test sets. Moreover, testing various descriptor types and other modeling approaches like support vector machines are essential for finding the best model for accurate prediction.

Defining new descriptors having a high correlation with the desired property is a critical step to obtain an accurate predictive model. MD-driven descriptors have been shown to improve the performance of QSPR models in various areas.\(^\text{18-20}\) The computational expense of MD-driven descriptors has prevented their widespread use in QSPR models. This was the case in the Guo et al. study, where the authors omitted the use of MD-HCF descriptors mentioning their computational expense.\(^\text{17}\) This obstacle, along with a limited database of amphiphilic peptides, has restricted extensive investigations on QSPR models to predict the CAC of these peptides to just one report.\(^\text{16}\)

Therefore, in the current study, we developed a novel set of MD-driven descriptors that did not require running simulation before descriptor definition and compared their performance...
Table 1. Amphiphilic Peptides and the Measured CAC Values Collected from the Literaturea

| Peptides | CAC (−log M) | Refs | Peptides | CAC (−log M) | Refs | Peptides | CAC (−log M) | Refs |
|----------|--------------|------|----------|--------------|------|----------|--------------|------|
| V6K2GRGDS | 4.83 | 23 | A10H6 | 6.64 | 31 | Ac-GAVILEE | 3.15 | 38 |
| Ac-A6K | 4.60 | 24 | RF | 1.74 | 32 | Ac-GAVILEE-H2 | 3.10 |
| Ac-L6K2-NH2 | 4.34 | 25 | RF | 4.40 | 32 | Ac-13D | 2.96 |
| Ac-L6K3-NH2 | 3.60 | 36 | RF | 2.64 | 32 | Ac-L3D | 2.92 |
| Ac-V6K2-NH2 | 3.48 | 36 | RF | 3.52 | 32 | Ac-L3K-NH2 | 2.92 |
| Ac-V6K3-NH2 | 3.08 | 36 | RF | 4.70 | 32 | Ac-V3K-NH2 | 2.80 |
| Ac-V6K4-NH2 | 2.33 | 36 | A6YD | 2.52 | 33 | Ac-V6k-NH2 | 3.35 | 39 |
| Ac-L6K4-NH2 | 2.27 | 36 | V4WD2 | 2.12 | 34 | A3C | 3.42 | 40 |
| Ac-A6K2-NH2 | 2.10 | 36 | V4D | 2.70 | 34 | Y3C | 3.77 |
| Ac-V6D | 3.30 | 26 | V4WD | 2.32 | 34 | I3C | 4.15 |
| Ac-V6D2 | 2.96 | 36 | I4WD2 | 2.39 | 34 | I4K | 3.60 | 41 |
| A6RGD | 1.74 | 27 | L4WD2 | 2.72 | 34 | LSK | 3.89 |
| G3A3V3I3K3 | 3.23 | 28 | RFL4FR | 3.10 | 35 | L12K | 2.99 |
| K3I3V3A3G3 | 3.19 | 28 | K60L30 | 10.19 | 36 | L4K | 3.27 |
| I3Y3A3G3K3 | 3.28 | 28 | K60L20 | 6.17 | 36 | L5K | 3.85 |
| K3G3A3V3I3 | 3.55 | 28 | E60L20 | 5.46 | 36 | Ac-A6D | 3.34 | 42 |
| V3G3I3A3K3 | 3.01 | 28 | K80L20 | 4.38 | 36 | Ac-GAVILRR-NH2 | 3.09 | 43 |
| K3A3I3G3V3 | 3.15 | 28 | K60L10 | 4.01 | 36 | K4X4-gA | 3.70 | 44 |
| K-K8 | 2.10 | 29 | Ac-A9K-NH2 | 4.82 | 37 | K5X3-gA | 3.68 |
| KK8 | 2.10 | 29 | Ac-A6K-NH2 | 3.70 | 37 | K8-gA | 3.66 |
| IK-K11 | 2.58 | 30 | Ac-A3K-NH2 | 2.00 | 37 | K6X2-gA | 3.64 |
| IK-K11 | 2.66 | 30 | Ac-A6D | 3.52 | 38 | K7X1-gA | 3.62 |
| IK-K16 | 3.03 | 30 | DA6-NH2 | 3.70 | 38 | K3X5-gA | 3.77 |
| IK-K16 | 3.13 | 30 | KA6-NH2 | 3.52 | 38 | | |
| GAVILRR | 1.52 | 30 | Ac-DK-NH2 | 3.35 | | |

The CAC values were measured by fluorimetry,23−26 conductivity,27−29 or dynamic light scattering (DLS) techniques30−34 in pure water. All concentrations are represented as −log M. The defined test set is presented in bold font. Long peptides are indicated by italic format. All peptide structures and molar concentrations can be found in the supporting information Table S5.35 The value of CAC for Ac-A6k-NH2 was extracted from reference 36 as the techniques used there tended to be more accurate than those used in the other reports.

RESULTS

Here, we investigated different descriptor sets and statistical modeling approaches (Figure 1) to model the CAC of the amphiphilic peptides listed in Table 1. The first approach used descriptors derived from peptide structures (the structure-drawing commands are listed in Tables S1−S3 and the peptide structures are shown in Table S4) using the PaDEL-Descriptor (whole-peptide approach, which is depicted by the black line in Figure 1). The second approach used the principal components of PaDEL descriptors for each amino acid in the sequence (PCA approach, which is depicted by the red line in Figure 1). The third approach used z-scales for each amino acid in the sequence as defined by Jonsson et al.21 (z-scale approach, which is depicted by the blue line in Figure 1). The fourth approach used a novel set of descriptors driven from the coarse-grained simulation of di- and tripeptides (AP/APH scale approaches, which are depicted by the purple lines in Figure 1). These scales are listed in Table S5. For the last three sets, the matrix dimensions for peptides with various lengths were unified using auto cross-covariance (ACC) calculations.22 Four machine learning algorithms were applied to model the data, including the support vector machine with the radial kernel (svmRadial), support vector machine with the linear kernel (svmLinear), partial least square (PLS), and generalized boosted models (GBM). Descriptors and ACCs with the most influence in the best models were extracted and the ability of the best model to predict the CAC values of long peptides (70−100 residues) was also investigated.

Model Performance. Model performance was assessed by residual standard error (RSE) calculation for the test and train sets (Figure 2). Among all models in all approaches, the model generated by svmRadial for the AP scale 2 (containing AP scales for AA-Ala, Ala-AAa, and Ala-AAa-Ala) showed the best performance on the test set with an RSE of 0.1. The svmRadial model also showed the best performance in other approaches except for the z-scale approach, where the PLS model was the best. Generally, models fitted on APH scales, which are based on hydrophilicity-adjusted derivatives of AP scales,14 performed inferior to those fitted on the AP scale. Consequently, the svmRadial model on AP scale 2 is suitable for the CAC prediction of newly designed amphiphilic peptides.

Influential Descriptors. To identify the physicochemical properties that govern the self-assembling behavior of amphiphilic peptides, descriptors or ACCs with the most influence on the best model in each approach were extracted using the “varImp()” function of the CARET package (Table 3). Topological polar surface area (TopoPSA), centered...
Broto−Moreau autocorrelation—lag 7/weighted by van der Waals volumes (ATSC7v), and the number of hydrogen bond acceptors (nHBAcc3) were the most influential descriptors in the whole peptide approach (Table 3). The most influential ACC in the svmRadial model in the PCA approach consisted of the sum of products between PCs 1 and 3 with a lag of 1 (two adjacent amino acids in the peptide sequence), indicating the intensifying effect of neighboring amino acids on the self-assembly power of amphiphilic peptides. The other two influential ACCs contained the sum of products of PCs 4−5 and 3−8 with a lag of 1, respectively. In the PLS model in the z-scale approach, z-scale 1 was the most important z-scale as seen in ACC 1 and 3 (Table 3). This scale is related to the hydrophilicity of amino acids in the peptide sequence. Scales 3 and 2 are other influential z-scales that are related to the electronic and side-chain bulk properties of amino acids. The AP scale 3, which presented the aggregation propensity of the Ala-AA_A-Ala peptide, was the most influential ACC in the AP-scale approach. Since the performance of all models in the AP^12-scale approach is poor, none of the AP^12 scales can have physicochemical relevance. The aggregation propensity of the Ala-AA_A-Ala peptide is the most influential descriptor in the

Figure 2. Performance of svmRadial, svmLinear, PLS, and GBM models on the training and test sets: model performance on the train (green dots) and test (orange dots) sets were measured using residual standard error (RSE) calculation and visualized by plotting predicted CAC values against experimental ones. The equation for the black lines is predicted CAC = experimental CAC.
The model related descriptor was the fraction of sp3 carbons to sp2 carbon. The two descriptors related to the aggregation propensity of self-assembly of amphiphilic peptides.

Table 2. Three Descriptors or ACCs with the Most Influence on the Best Model in Each Approach

| approach                  | order 1                                      | order 2                                      | order 3                                      |
|---------------------------|----------------------------------------------|----------------------------------------------|----------------------------------------------|
| whole peptide-svmRadial   | TopoPSA                                      | ATSC7v                                       | nHBAacc3                                    |
| PCA-svmRadial             | $\sum_{i=1}^{n-1} PC_i \times PC_{i+1}$     | $\sum_{i=1}^{n-1} PC_i \times PC_{i+1}$     | $\sum_{i=1}^{n-1} PC_i \times PC_{i+1}$     |
| z-scale-PLS               | $\sum_{i=1}^{n-1} Z_i \times Z_{i+1}$       | $\sum_{i=1}^{n-1} Z_i \times Z_{i+1}$       | $\sum_{i=1}^{n-1} Z_i \times Z_{i+1}$       |
| AP-scale-svmRadial        | $\sum_{i=1}^{n-1} AP_i \times AP_{i+1}$     | $\sum_{i=1}^{n-1} AP_i \times AP_{i+1}$     | $\sum_{i=1}^{n-1} AP_i \times AP_{i+1}$     |
| AP^4-scale-svmRadial      | $\sum_{i=1}^{n-1} AP_i^4 \times AP_{i+1}^4$ | $\sum_{i=1}^{n-1} AP_i^4 \times AP_{i+1}^4$ | $\sum_{i=1}^{n-1} AP_i^4 \times AP_{i+1}^4$ |

Figure 3. Addition of long peptides to the data set reduced the model performance. Model performance was measured after (a) adding five long peptides from ref 35 to the data set and (b) after removing the shortest peptide (RF) from the data set. The model fitted on the data set containing RF and the five long peptides had an RSE value of 0.29 on the test set. However, removing RF from the data set reduced the RSE value to 0.23.

Correlation Studies. Therefore, to find the molecular descriptors related to the aggregation propensity of Ala-AAA-Ala, a Pearson correlation study was conducted between this parameter and the PaDEL descriptors defined for Ala-AAA-Ala peptides. Table 3 lists 10 molecular descriptors that had a strong positive or negative correlation with the aggregation propensity of Ala-AAA-Ala peptides. The positively related descriptors were the sum of atom-type E-State: $\equiv CH$, count of the number of occurrences of the E-state fragments, count of atom-type E-State: $\equiv CH$, count of atom-type H E-State: H on aaCH, dCH2 or dsCH, and count of atom-type H E-State: $\equiv CH$ belonging to electrotopological state indices, which measure the atom-type E-states or hydrogen atoms in $\equiv CH$ or $\equiv CH$ fragments. These states are only present in aromatic amino acids among Ala-AAA-Ala peptides. The first negatively related descriptor was the fraction of sp3 carbons to sp2 carbon. The second and third negatively correlated descriptors were Geary autocorrelations with a lag of 1 weighted by first ionization potential and polarizability, respectively. The two last negatively correlated descriptors were the largest absolute eigenvalue of Burden modified matrix $-n$ 1 and $-n$ 3 weighted by the relative intrinsic state (I-state), respectively. All negatively correlated descriptors had their lowest values in Ala-AAA-Ala peptides containing aromatic amino acids (Table 2).

Effect of Peptide Length. To understand the effect of peptide length on the best model and the most influential ACCs, the data set was expanded to include peptides between 70 and 100 amino acids (peptides from ref 36, Table 1). Then, the AP approach with svmRadial was applied to generate models for the new data set, once with and once without the shortest peptide (RF) in the data set. Generally, the addition of longer peptides reduced the accuracy of the model as the RSE value for the test set increased from 0.1 to 0.29 (Figures 2 and 3). Although removing the shortest peptide in the data set decreased the RSE value from 0.29 to 0.23 for the test set (Figure 3), it cannot be concluded that this removal improved the accuracy of the model. As seen in Figure 3b, the CAC values of the test set were located in the region that had the best fit in the train set. In other words, if the test set was selected from the regions with a CAC value of more than 5.0 or less than 2.5, the RSE would have increased drastically. The addition of the long peptides also changed the most influential ACCs (Tables 3 and S6). All influential ACCs, whether for the model on the original data set or the data set containing longer peptides with or without RF, had a lag of 1. The most influential ACC in both the original data set and the data set including long peptides and RF was the same, while the other two influential ACCs for the data set with longer peptides contained AP scale 4. Removing RF from the data set containing long peptides changed all influential ACCs to contain AP scales 4. This AP scale is different from AP scale 3 just for terminal amino acids if they are capped with acetyl or amine groups. Therefore, it can be concluded that the physicochemical properties of amphiphilic peptides differ as the peptide sizes differ.
 Nonetheless, the reported because they did not use a test set and the number of peptides of 0.832 after fourfold cross-validation. In another report, Guo et al. obtained their best-performing model by using svmRadial, was the topological polar surface area (topoPSA),46 which is calculated by the summation of tabulated surface contributions of polar fragments. Regarding this observation, it can be suggested that the molecular surface area and also polarizability of the molecule are important determinants of the self-assembling tendency. Tian et al. also observed that CODESSA descriptors, representing the static state of the molecule, performed worse than MD-HCF descriptors, representing the dynamic state of the molecule. Here, the whole-peptide approach also performed worse than other approaches, indicating that amphiphilic peptide self-assembly cannot be adequately understood by the static molecular state. Guo et al. also observed that a decrease in structural flexibility could decrease the CAC, which also emphasizes the influence of dynamic structural properties on the self-assembling tendency. None of these descriptors carry information about the importance of each amino acid type at a specific position in the self-assembly.

Interestingly, the PLS model in the z-scale approach performed better than both the whole-peptide and PCA approaches. The most important ACC in this model is the sum of products of z-scale 1 in two neighboring amino acids. z-scale 1 is related to amino-acid hydrophilicity as described by Jonsson et al. Therefore, it can be concluded that in this model, the hydrophobic interaction of two adjacent amino acids plays an important role in the CAC of the peptides. The other two z-scales are related to the side-chain bulk and electronic properties that are used to calculate the second influential ACC in the model. Therefore, one can increase the self-assembly tendency of an amphiphilic peptide by putting two highly hydrophobic amino acids in adjacent positions. The svmRadial model in the AP-scale approach outperformed the other models. The most influential ACC in this model was the sum of products of AP scales 1 and 3 with a lag of 1 (Table 3), meaning that the interaction between the aggregation propensities of Ala-AA, Ala and AAi-Ala peptides in neighboring amino acids is crucial in determining the CAC value. The AP value for AAi-Ala dipeptides (AP scale 2) is equal to 1 for all dipeptides except those containing aromatic amino acids. This finding is in accordance with the observation by Frederix et al. on the computed AP values of tripeptides that neighboring aromatic amino acids have an intensifying effect on the aggregation propensity. The existence of AP for Ala-AA, Ala peptides in two other influential ACCs makes it the

### Table 3. Correlation between AP and the PaDEL Molecular Descriptors for Ala-AA, Ala Triptides Constructed by Replacing AAi with Amino Acids in the Data Set

| number | descriptor type | definition | correlation coefficient |
|--------|-----------------|------------|------------------------|
| 1      | SdsCH           | sum of atom-type E-state: CH− | 0.96 |
| 2      | kds.aatCH       | counts the number of occurrences of the E-state fragments | 0.95 |
| 3      | ndsCH           | count of atom-type E-state: CH− | 0.95 |
| 4      | nHother         | count of atom-type H E-state: H on aaCH, dCH2 or dCH | 0.95 |
| 5      | nHdcCH          | count of atom-type H E-State: CH− | 0.95 |
| 6      | HybRatio        | hybridization ratio (fraction of sp3 carbons to sp2 carbons) | -0.87 |
| 7      | GATS1i          | Geary autocorrelation—lag 1/ weighted by first ionization potential | -0.82 |
| 8      | GATS1p          | Geary autocorrelation—lag 1/ weighted by polarizabilities | -0.81 |
| 9      | SpMax1_Bhs      | largest absolute eigenvalue of Burden modified matrix—n 1/ weighted by relative I-state | -0.80 |
| 10     | SpMax3_Bhs      | largest absolute eigenvalue of Burden modified matrix—n 3/ weighted by relative I-state | -0.80 |

“Ten descriptors with the highest positive and negative correlations are presented.”
most important descriptor in this model. However, the computed aggregation propensity of Ala-AA-Ala peptides merely ranks amino acids based on the self-assembling tendency and does not provide physicochemical information.

To understand the physicochemical meaning of the calculated aggregation propensity, we carried out a Pearson correlation between these values and the PaDEL descriptors for Ala-AA-Ala tripeptides. All electrotopological descriptors having a strong positive correlation with the aggregation propensity of Ala-AA-Ala peptides had the highest values for aromatic amino acids. Also, descriptors having a strong negative correlation with this quantity had the lowest values for aromatic amino acids. This observation indicates the importance of aromatic amino acids in the self-assembling tendency. Moreover, as ACCs with a lag of 1 had a significant influence on the best-performing model, one can conclude that two adjacent aromatic amino acids can intensely drive amphiphilic peptide self-assembly. Our results are in accordance with the study by Frederix et al. They observed that two sequential aromatic amino acids increase the self-assembly tendency. They hypothesized that this phenomenon is due to the conformation of the backbone and aromatic interactions. However, more studies are needed to get a better physicochemical interpretation on the aggregation propensity of di- and tripeptides as descriptors.

To our knowledge, this is the first report on the modeling of CAC values for amphiphilic peptides that includes long peptides (70−100 amino acids). Reduction in the accuracy on adding long peptides indicates that principles governing the self-assembly of short peptides might be less applicable for the longer ones. However, it is not clear whether inclusion or exclusion of ultra-short peptides (RF) in the data set can improve the model performance because long and also ultra-short peptides had a low frequency in our data set. Therefore, future studies on data sets with a higher number of ultra-short and long peptides are needed to determine whether self-assembling principles change by peptide length.

CONCLUSIONS

The successful prediction of CAC for amphiphilic self-assembling peptides by the novel descriptor set, reported in this study, brings forward several opportunities in the field. Besides enabling better design and screening of the amphiphilic peptides containing coded amino acids, the relationship between the AP scale of AA-Ala, Ala-AA, and Ala-AA-Ala peptides and the CAC of longer amphiphilic peptides suggests the possibility of CAC determination by CG-MD simulation of newly designed peptides that contain noncoded amino acids. Although screening of sequence space for di- and tripeptides has been reported, performing the same for longer peptides might exceed the current computational capacity. The model developed here could help us to reduce the computational burden in the screening of longer peptides for finding new self-assembling peptides.

MATERIALS AND METHODS

General Strategy. Four different approaches were used for QSPR analysis of self-assembling amphiphilic peptides (Figure 1). In the first approach (whole-peptide approach, indicated by a black line), structures for the whole-peptide molecules were generated and descriptors were defined for these structures. In the second approach (PCA approach, indicated by a red line), peptide sequences were split into Ala-AA-Ala peptides, where AAi was substituted for each amino acid in the sequence. Descriptors were defined for these tripeptides and their principal components were calculated and used for the next steps. In the other two approaches, indicated by blue (z-scale approach) and purple (AP/APH-scale approach) lines, no structure building or optimization was used. Instead, two different sets of scales were defined for each amino acid in the peptide structure.

Data Set. A database containing 74 amphiphilic peptides was constructed through a literature search. The CAC value of the included peptides was measured in pure water because the ionic strength of the solution can affect the CAC value. Moreover, when two or more reports existed for the same peptide sequence, the value measured by the more accurate method was included. Six peptides with noncoded amino acids and five long peptides (70−100 amino acids in length) were then removed to obtain a data set of 63 peptides. All CAC concentrations, which ranged from 0.01 to 2.3 × 10−7 M, were transformed into their negative logarithms before use in QSPR studies. The peptide data set is presented in Table 1 and their structures are shown in Table S4.

Structures. Peptide sequences were converted to two- and three-dimensional structures. All structures of amphiphilic peptides were built using Marvin software 19.17.0, 2019, ChemAxon (www.chemaxon.com). Briefly, the FASTA formats of peptide sequences were prepared and converted to images of one-dimensional (1D) structures by the Molconvert command. Terminal acetyl and amine groups were added manually where needed for the whole-peptide structures and also for all Ala-AA-Ala tripeptide structures. Charges were then applied to the structures at pH 7.00. The generated 1D structures were converted to two-dimensional (2D) and three-dimensional (3D) structures by Marvin software’s command-line options, where the MMFF94 force field was used to optimize the 3D structures. This step is presented as the “Structure Drawing” node in Figure 1. The commands used for the generation of peptide structures are summarized in Tables S1−S3. The generated structures are provided as a separate zip file.

Descriptors. Four sets of descriptors were defined for amphiphilic peptides in the data set. Descriptor sets for whole-peptide molecular structures and Ala-AA-Ala tripeptides (whole-peptide and PCA approaches) were generated using PaDEL descriptors software (version 2.21).47 PaDEL generated 1875 descriptors including 1444 1D and 2D descriptors, 431 3D descriptors, and 12 types of molecular fingerprints.47 In the z-scale approach, each amino acid in the peptide sequence was replaced with the three extended-scales (Figure 1−n-scales approach, each amino acid in the peptide sequence was replaced with the three extended-scales introduced by Jonsson et al.,46 generating a 3 × n matrix for each peptide, where n is the length of the peptide. In the final approach, we used a novel set of descriptors that were derived from the AP of di- and tripeptides simulated in the MARTINI force field and its derivative normalized for solubility (APPH). Similar to the z-scale approach, each amino acid in the peptide sequence was replaced with defined scales (Figure 1 and Table S5). In step 1, the scales were selected from the AP of AA-Ala and Ala-AA, dipeptides, where AAi was substituted for the amino acid in the sequence. In step 2, the scales were APs for Ala-AA-Ala peptides. In step 3, if the terminal amino acids were not capped (acylated or aminated), AA-Ala-Ala and Ala-Ala-AAi were selected for carboxyl and amine termini, respectively. The same process was performed for all AP(H).
scales. The matrices generated for all amphiphilic peptide sequences are provided in a zip file.

**Data Processing.** Different data preparation steps were taken for different QSPR approaches before fitting the same models on each data set. In the whole-peptide and PCA approaches, descriptors with near-zero variance and high correlation were removed from the data set. The resulting data set was directly used for model building in the whole-peptide approach. As for the PCA approach, principal components of the descriptor sets were calculated and 14 PCs that cover 90 percent of the variance were selected. A matrix for each peptide was constructed like matrices described for the z-scale and AP-scale approaches, by replacing 14 PCs for each amino acid in the peptide sequence. Since peptides with different sequence lengths produced matrices with different dimensions in the PCA, z-scale, and AP-scale approaches, the auto-covariance of each matrix was calculated according to eq 1 to produce a matrix for the model-building step, which was implemented in the RCPI package.

$$\text{ACC}_{j,k,i} = \sum_{l}^{n-1} z_{j,l} \times z_{k,i+l} / n - 1$$

where \(j\) and \(k\) are used for the scales (\(j = 1, 2, 3, \ldots\)), \(n\) is the number of amino acids in a sequence, and \(i\) is the amino acid position (\(i = 1, 2, \ldots, n\)). Data sets were then split into the train and test sets. Four models were fitted on the train sets in each approach: namely, support vector machine with the radial kernel (svmRadial), support vector machine with the linear kernel (svmLinear), partial least-square model (PLS), and generalized boosted models (GBM). A 10-fold cross-validation was performed to evaluate the model fitted on the train sets. Finally, the best model was used to predict the CAC values in the test set. All model-building and evaluation steps were done using the CARET package of R. Calculation of RSE was performed using eq 2 to select the model with the best performance on the test set

$$\text{residual standard error} = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{df}}$$

where \(Y_i\) is the experimental value of \(Y\), \(\hat{Y}_i\) is the predicted value by model, \(n\) is the sample size, and \(df\) is the degree of freedom.

**Influential Descriptors.** Models were further studied by investigating their important descriptors. The important descriptors in the best model for each approach were extracted using the “VarImp()” function from the CARET package of R, which sorts descriptors with the most influence on the model. The correlation between the AP for di- and tripeptides used for AP/APH-scales calculations; and the effect of long peptides on influential ACCs (PDF)

2D and 3D structures of peptide data set; 2D and 3D structures of di- and tripeptides; and AP-scale matrices (ZIP)

**AUTHOR INFORMATION**

**Corresponding Authors**
Reza Ahangari Cohan — Department of Nanobiotechnology, New Technologies Research Group, Pasteur Institute of Iran, Tehran 1316943551, Iran; Department of Hematology and Medical Laboratories Science, Faculty of Allied Medicine, Kerman University of Medical Sciences, Kerman 7616911333, Iran
Logham Firoozpour — Department of Medicinal Chemistry, Faculty of Pharmacy, Tehran University of Medical Sciences, Tehran 1416753955, Iran
Soroush Sardari — Drug Design and Bioinformatics Unit, Medical Biotechnology Department, Biotechnology Research Center, Pasteur Institute of Iran, Tehran 1316943551, Iran
Ali Afgar — Research Center for Hydatid Disease in Iran, School of Medicine, Kerman University of Medical Sciences, Kerman 7616914115, Iran

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.1c01293

**Notes**
The authors declare no competing financial interest.

**ACKNOWLEDGMENTS**
We thank Dr. Pin W. J. M. Frederix for generously providing the calculated AP and APH values for the tripeptides. We also thank Dr. Amirhossein Sakhteman for the initial assistance in model building and Dr. Saeed Yousefnejad for suggesting the use of the PCA approach. This project was financially supported by Pasteur Institute of Iran (Grant Nos. 934 and 692 and thesis number BP-9583) and Iran National Science Fund (INSF) (Grant No. 98008933).

**ABBREVIATIONS**
ACC, auto cross-covariance; AP, aggregation propensity; CAC, critical aggregation concentration; CG-MD, coarse-grained molecular dynamics; CMC, critical micelle concentration; DLS, dynamic light scattering; GA-PLS, genetic algorithm feature selection-partial least-square model; GBM, gradient boosting machine; I-state, intrinsic state; MD, molecular dynamics; MD-HCF, molecular dynamics-based hydrophobic cross-field; PCA, principal component analysis; PLS, partial least squares; QSPR, quantitative structure–property relationship; RSE, residual standard error; SVM, support vector machines; svmLinear, support vector machine with the linear kernel; svmRadial, support vector machine with radial kernel

**ASSOCIATED CONTENT**

1. Supporting Information
The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.1c01293.

Commands for generation of structures in Marvin; amphiphilic peptide data set with more detail; structures
Studies for Surfactants.

Design and Discover New Hydrogels.

Exploring the Sequence Space for (tri-) Peptide Self-Assembly to

Dynamic Hydrophobic Property.

Aggregation Concentration of Peptide Surfactants Is Predictable from

Potential Bioglasses.

J. Pept. Sci.

Screening for Dipeptide Aggregation: Toward Predictive Tools for

Structure − Supramolecular Gelation.

Assembly of Pepfactants?

against Alzheimer

Acids.

Multivariate Parametrization of 55 Coded and Non-Coded Amino

Response by Self-Aggregating Peptides.

Angew. Chem.

Cationic Dipeptide Nanotubes into Vesicles and Oligonucleotide

Delivery.

Proc. Natl. Acad. Sci. U.S.A.

Zhang, S. Designer Short Peptide Surfactants Stabilize G Protein-

Macromolecular Membrane.

Assembly of a Self-Complementary Oligopeptide to Form a Stable

Chemom. Intell. Lab. Syst.

Assembling Peptide Scaffolds in Regenerative Medicine: The Way to

Made of Natural Amino Acids

Hu, J.; Zhang, X.; Wang, Z. A Review on Progress in QSPR

(15) Hu, J.; Zhang, X.; Wang, Z. A Review on Progress in QSPR

(14) Frederix, P. W.; Scott, G. G.; Abul-Haija, Y. M.; Kalafatovic, D.;
Pappas, C. G.; Javid, N.; Hunt, N. T.; Ulijn, R. V.; Tuttle, T.
Exploring the Sequence Space for (tr) Peptide Self-Assembly to Design and Discover New Hydrosols. Nat. Chem. 2015, 7, 30.

(15) Hu, J.; Zhang, X.; Wang, Z. A Review on Progress in QSPR Studies for Surfactants. Int. J. Mol. Sci. 2010, 11, 1020−1047.

(16) Tian, F.; Wu, J.; Huang, N.; Guo, T.; Mao, C. The Critical Aggregation Concentration of Peptide Surfactants Is Predictable from Dynamic Hydrophobic Property. SAR QSAR Environ. Res. 2013, 24, 89−101.

(17) Guo, T.; Yang, J.; Zeng, L.; Wang, H.; Tong, Q.; Li, X. Does There Exist an Intrinsic Relationship between the Flexibility and Self-Assembly of Pepfactants? Mol. Simul. 2014, 40, 423−430.

(18) Linati, L.; Luwardi, G.; Malavasi, G.; Menabue, I.; Menziani, M. C.; Mustarelli, P.; Segre, U. Qualitative and Quantitative Structure−Property Relationships Analysis of Multicomponent Potential Bioglasses. J. Phys. Chem. B 2005, 109, 4989−4998.

(19) Jamal, S.; Grover, A.; Grover, S. Machine Learning from Molecular Dynamics Trajectories to Predict Caspase-8 Inhibitors against Alzheimer’s Disease. Front. Pharmacol. 2019, 10, No. 780.

(20) Van Lomme, R.; Zhao, J.; De Borggraeve, W. M.; De Proft, F.; Alonso, M. Molecular Dynamics Based Descriptors for Predicting Supramolecular Gelation. Chem. Sci. 2020, 11, 4226−4238.

(21) Jonsson, J.; Eriksson, L.; Hellberg, S.; Sjöström, M.; Wold, S. Multivariate Parameterization of S5 Coded and Non-Coded Amino Acids. Quant. Struct.-Act. Relat. 1989, 8, 204−209.

(22) Andersson, P. M.; Sjöström, M.; Lundstedt, T. Preprocessing Peptide Sequences for Multivariate Sequence-Property Analysis. Chemom. Intell. Lab. Syst. 1998, 42, 41−50.

(23) Liang, J.; Wu, W.-L.; Xu, X.-D.; Zhao, R.-X.; Zhang, X.-Z. pH Responsive Micelle Self-Assembled from a New Amphiphilic Peptide as Anti-Tumor Drug Carrier. Colloids Surf., B 2014, 114, 398−403.

(24) Qiu, F.; Chen, Y.; Zhao, X. Comparative Studies on the Self-Assembling Behaviors of Cationic and Cationic Surfactant-like Peptides. J. Colloid Interface Sci. 2009, 336, 477−484.

(25) Meng, Q.; Kou, Y.; Ma, X.; Liang, Y.; Guo, L.; Ni, C.; Liu, K. Tunable Self-Assembled Peptide Amphiphile Nanostructures. Langmuir 2012, 28, 5017−5022.

(26) Yang, S. J.; Zhang, S. Self-Assembling Behavior of Designer Lipid-like Peptides. Supramol. Chem. 2006, 18, 389−396.

(27) Castelletto, V.; Gouveia, R.; Connon, C.; Hamley, I.; Seisunen, J.; Nykänen, A.; Ruokolainen, J. Alanine-Rich Amphiphilic Peptide Containing the RGD Cell Adhesion Motif: A Coating Material for Human Fibroblast Attachment and Culture. Biomater. Sci. 2014, 2, 362−369.

(28) Cao, M.; Lu, S.; Zhao, W.; Deng, L.; Wang, M.; Wang, J.; Zhou, P.; Wang, D.; Xu, H.; Lu, J. R. Peptide Self-Assembled Nanostructures with Distinct Morphologies and Properties Fabricated by Molecular Design. ACS Appl. Mater. Interfaces 2017, 9, 39174−39184.

(29) Cao, M.; Shen, Y.; Wang, Y.; Wang, X.; Li, D. Self-Assembly of Short Elastin-like Amphiphilic Peptides: Effects of Temperature, Molecular Hydrophobicity and Charge Distribution. Molecules 2019, 24, No. 202.

(30) Kornmueller, K.; Letofsky-Papst, I.; Gradauer, K.; Mīkl, C.; Cacho-Nerin, F.; Leybold, M.; Keller, W.; Leitinger, G.;amenitsch, H.; Prassl, R. Tracking Morphologies at the Nanoscale: Self-Assembly of an Amphiphilic Designer Peptide into a Double Helix Superstructure. Nano Res. 2015, 8, 1822−1833.

(31) Hamley, I. W.; Kirkham, S.; Dehsorkhi, A.; Castelletto, V.; Adamcik, J.; Mezzenga, R.; Ruokolainen, J.; Mazzuca, C.; Gatto, E.; Venanzi, M.; et al. Self-Assembly of a Model Peptide Incorporating a Hexa-Histidine Sequence Attached to an Oligo-Alanine Sequence, and Binding to Gold NTA/nickel Nanoparticles. Biomacromolecules 2014, 15, 3412−3420.

(32) Silva, E. R.; Listik, E.; Han, S. W.; Alves, W. A.; Soares, B. M.; Reza, M.; Ruokolainen, J.; Hamley, I. W. Sequence Length Dependence in Arginine/phenylalanine Oligopeptides: Implications for Self-Assembly and Cytotoxicity. Biophys. Chem. 2018, 233, 1−12.

(33) Kornmueller, K.; Lehofer, B.; Meindi, C.; Fröhlich, E.; Leitinger, G.;amenitsch, H.; Prassl, R. Peptides at the Interface: Self-Assembly of Amphiphilic Designer Peptides and Their Membrane Interaction Propensity. Biomacromolecules 2016, 17, 3591−3601.

(34) Kornmueller, K.; Lehofer, B.; Leitinger, G.;amenitsch, H.; Prassl, R. Peptide Self-Assembly into Lamellar Phases and the Formation of Lipid-Peptide Nanostructures. Nano Res. 2018, 11, 913−928.

(35) Da Silva, E. R.; Walter, M. N. M.; Reza, M.; Castelletto, V.; Ruokolainen, J.; Connon, C. J.; Alves, W. A.; Hamley, I. W. Self-Assembled Arginine-Capped Peptide Bolaamphiphile Nanosheets for Cell Culture and Controlled Wettability Surfaces. Biomacromolecules 2015, 16, 3180−3190.

(36) Holowka, E. P.; Pochan, D. J.; Deming, T. J. Charged Polypeptide Vesicles with Controllable Diameter. J. Am. Chem. Soc. 2005, 127, 12423−12428.

(37) Xu, H.; Wang, J.; Han, S.; Wang, J.; Yu, D.; Zhang, H.; Xia, D.; Zhao, X.; Waigh, T. A.; Lu, J. R. Hydrophobic-Region-Induced Transitions in Self-Assembled Peptide Nanostructures. Langmuir 2009, 25, 4115−4123.

(38) Wang, X.; Corin, K.; Baaske, P.; Wienken, C. J.; Jerabek-Willemsen, M.; Dühr, S.; Braun, D.; Zhang, S. Peptide Surfactants for Cell-Free Production of Functional G Proteïne-Coupled Receptors. Proc. Natl. Acad. Sci. U.S.A. 2011, 108, 9049−9054.

(39) Pan, F.; Zhao, X.; Perumal, S.; Waigh, T. A.; Lu, J. R. Interfacial Dynamic Adsorption and Structure of Molecular Layers of Peptide Surfactants. Langmuir 2010, 26, 5690−5696.

(40) Cao, M.; Cao, C.; Zhou, P.; Wang, N.; Wang, J.; Xia, D.; Xu, H. Self-Assembly of Amphiphilic Peptides: Effects of the
Single-Chain-to-Gemini Structural Transition and the Side Chain Groups. Colloids Surf., A 2015, 469, 263–270.

(41) Han, S.; Cao, S.; Wang, Y.; Wang, J.; Xia, D.; Xu, H.; Zhao, X.; Lu, J. R. Self-Assembly of Short Peptide Amphiphiles: The Cooperative Effect of Hydrophobic Interaction and Hydrogen Bonding. Chem. - Eur. J. 2011, 17, 13095–13102.

(42) Nagai, A.; Nagai, Y.; Qu, H.; Zhang, S. Dynamic Behaviors of Lipid-like Self-Assembling Peptide A6D and A6K Nanotubes. J. Nanosci. Nanotechnol. 2007, 7, 2246–2252.

(43) Khoe, U.; Yang, Y.; Zhang, S. Self-Assembly of Nanodonut Structure from a Cone-Shaped Designer Lipid-like Peptide Surfactant. Langmuir 2009, 25, 4111–4114.

(44) de Bruyn Oubote, D. Rational Design of Purely Peptidic Amphiphiles for Drug Delivery Applications; University of Basel, 2011.

(45) Katritzky, A. R.; Pacureanu, L. M.; Slavov, S. H.; Dobchev, D. A.; Shah, D. O.; Karelson, M. QSAR Study of the First and Second Critical Micelle Concentrations of Cationic Surfactants. Comput. Chem. Eng. 2009, 33, 321–332.

(46) Ertl, P.; Rohde, B.; Selzer, P. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties. J. Med. Chem. 2000, 43, 3714–3717.

(47) Yap, C. W. PaDEL-Descriptor: An Open Source Software to Calculate Molecular Descriptors and Fingerprints. J. Comput. Chem. 2011, 32, 1466–1474.

(48) Cao, D.-S.; Xiao, N.; Xu, Q.-S.; Chen, A. F. Rcpi: R/Bioconductor Package to Generate Various Descriptors of Proteins, Compounds and Their Interactions. Bioinformatics 2015, 31, 279–281.