

# Aggregation Rules of Short Peptides

Jiaqi Wang,<sup>▽</sup> Zihan Liu,<sup>▽</sup> Shuang Zhao,<sup>▽</sup> Yu Zhang, Tengyan Xu,\* Stan Z. Li,\* and Wenbin Li\*

Cite This: *JACS Au* 2024, 4, 3567–3580

Read Online

ACCESS |

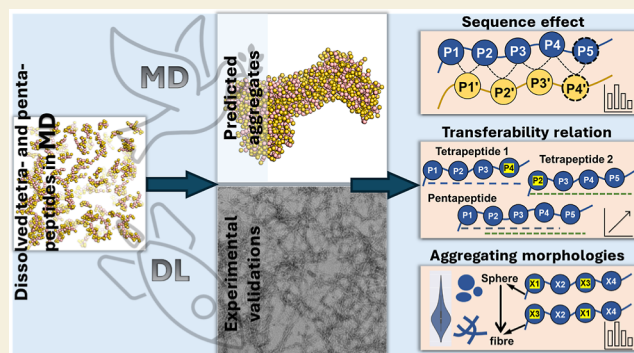
Metrics & More

Article Recommendations

Supporting Information

**ABSTRACT:** The elucidation of aggregation rules for short peptides (e.g., tetrapeptides and pentapeptides) is crucial for the precise manipulation of aggregation. In this study, we derive comprehensive aggregation rules for tetrapeptides and pentapeptides across the entire sequence space based on the aggregation propensity values predicted by a transformer-based deep learning model. Our analysis focuses on three quantitative aspects. First, we investigate the type and positional effects of amino acids on aggregation, considering both the first- and second-order contributions. By identifying specific amino acids and amino acid pairs that promote or attenuate aggregation, we gain insights into the underlying aggregation mechanisms. Second, we explore the transferability of aggregation propensities between tetrapeptides and pentapeptides, aiming to explore the possibility of enhancing or mitigating aggregation by concatenating or removing specific amino acids at the termini. Finally, we evaluate the aggregation morphologies of over 20,000 tetrapeptides, regarding the morphology distribution and type and positional contributions of each amino acid. This work extends the existing aggregation rules from tripeptide sequences to millions of tetrapeptide and pentapeptide sequences, offering experimentalists an explicit roadmap for fine-tuning the aggregation behavior of short peptides for diverse applications, including hydrogels, emulsions, or pharmaceuticals.

**KEYWORDS:** short peptides, complete sequence space, molecular dynamics, deep learning, aggregation rules, transferability relation, aggregating morphologies



## 1. INTRODUCTION

Peptide aggregation is a widely observed phenomenon in both natural contexts and the human environment. It refers to the association of peptide monomers into oligomers and the subsequent combination into exquisite or amorphous supra-molecular assemblies, a process primarily driven by non-covalent interactions such as hydrogen bonding, electrostatic interactions, and van der Waals forces.<sup>1,2</sup> The importance of peptide aggregation has been recognized for a long time, and it has a 2-fold nature. On the one hand, peptide aggregates can be developed into various applications such as semiconductors and batteries,<sup>3,4</sup> fluorescent probes and ligands,<sup>5,6</sup> and drugs and nutraceuticals.<sup>7–9</sup> On the other hand, the aggregation of peptides is involved in more than 20 diseases, such as Alzheimer's disease,<sup>10</sup> Parkinson's disease,<sup>11</sup> and type II diabetes.<sup>11</sup> Elucidating the aggregation mechanisms of peptides under aqueous or cellular conditions is of great significance for controlling the degree of aggregation (promoting or mitigating), which in turn has tremendous biomedical and practical relevance.

The primary sequences driving aggregation in polypeptides and proteins are short peptides of less than 10 amino acids. For example, the A $\beta$ <sub>16–22</sub> peptides are the main driving force for the aggregation of full-length amyloids.<sup>12</sup> Therefore, in this research, we focus on deducing the aggregation rules of short

peptides, i.e., impact of the type and position of individual amino acids on aggregations within the entire sequence space of tetrapeptides and pentapeptides comprising 160,000 and 3,200,000 sequences, respectively. We constrain our investigation conditions to aqueous solutions (omitting the complexity of cellular environments with varying pH levels, solvents, or the presence of cellular substances such as proteins) in order to provide peptide aggregation rules under a set of experimental parameters that are readily reproducible and easily accessible.

The peptide sequence, which contains information on the type and position of the amino acids in a peptide, not only determines whether the peptide can aggregate or not but also governs the morphologies of the resulting aggregates.<sup>13</sup> Research into the formation mechanism and exquisite control of morphologies at the molecular level is crucial for modulating various properties of aggregates (such as mechanical, optical, and electronic properties) for the development of versatile

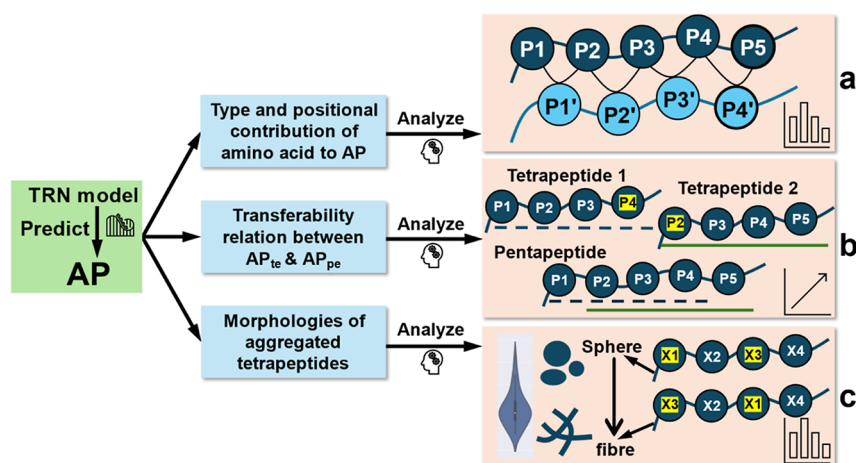
Received: June 12, 2024

Revised: August 1, 2024

Accepted: August 1, 2024

Published: September 3, 2024





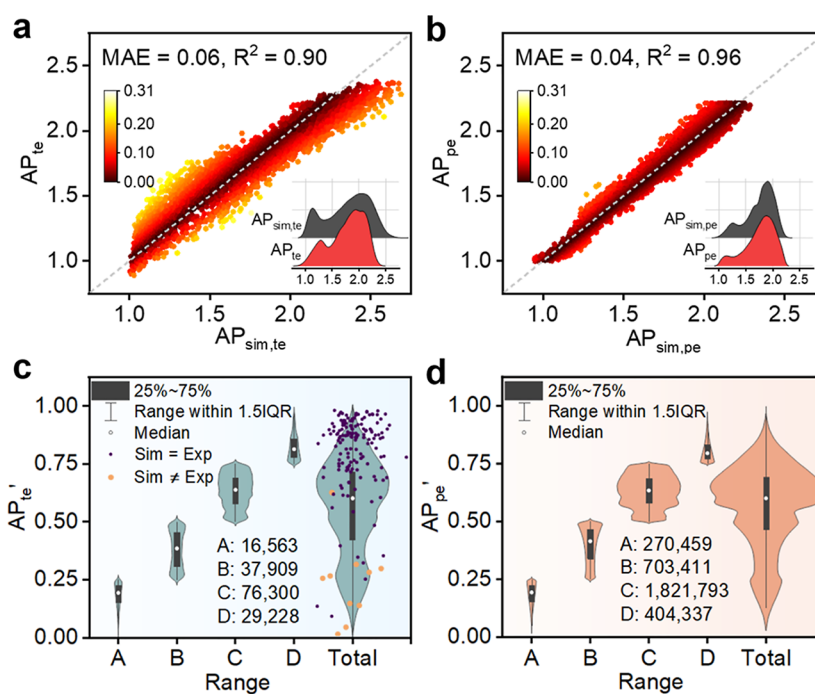
**Figure 1.** Analyses of the aggregation rules of short peptides from three perspectives. A deep learning-based TRN model predicts the AP values of tetrapeptides ( $AP_{te}$ ) and pentapeptides ( $AP_{pe}$ ) across the entire sequence space (i.e., 160,000 tetrapeptides and 3,200,000 pentapeptides). Based on the predicted AP values, we analyze the (a) type and positional contribution of 20 amino acids to the AP, (b) transferability relationship between the AP of a pentapeptide and the averaged AP of two corresponding tetrapeptides ( $AP_{pe}$  versus  $AP_{ave}$ ), and (c) distribution of the morphologies of aggregated tetrapeptides as well as the type and positional contributions of 20 amino acids to different morphologies.

applications.<sup>14–16</sup> Therefore, in this research, we also perform over 20,000 coarse-grained molecular dynamics (CGMD) simulations of tetrapeptide aggregation for 1.25  $\mu$ s. By comparing computational and experimental data, we provide a comprehensive analysis of the type and position effects of amino acids on the morphologies of the resulting aggregates.

With the development of simulation and machine learning approaches, pioneering work on sequence-based peptide aggregation is on the rise. In 2014, Frederix et al.<sup>17</sup> used CGMD to study the aqueous aggregation propensities and morphologies of 8000 tripeptides. Since the design rules generated by Frederix et al. are based on tripeptides containing only three amino acids, it would be challenging to predict which termini are the preferred positions of the 20 amino acids for aggregation in longer peptides. Therefore, it is necessary to extend the investigation of the design rules to longer peptides, such as tetrapeptides and pentapeptides. Beyond tripeptides, van Teijlingen and Tuttle<sup>18</sup> developed an active machine learning method for the search of aggregating peptides in 2021. The method they developed pushed the limit of aggregation prediction to hexapeptides, but no aggregation design rule was presented. More recently, Batra et al.<sup>19</sup> developed a machine learning workflow that integrates Monte Carlo tree search and random forest to autonomously search for aggregating peptides and demonstrated that the predictive power of machine learning overcomes human biases in the discovery of aggregating peptides, but design rules were still insufficiently investigated. Inspired by the aforementioned works, we developed a deep learning model based on a transformer regression network (TRN) that is capable of predicting the aggregation propensity (AP) of any oligopeptide (peptides with less than or equal to 10 amino acids).<sup>20,21</sup> The AP values reported in this paper are all predicted by our TRN model trained on the CGMD-generated AP data. Due to the limited spatial and temporal scale of CGMD and the coarse-grained force field model, we focus on aggregation without distinguishing between aggregation, self-assembly, crystallization, or precipitation in this work.

Starting from the predicted and validated AP values against experimental results (the validation details can be found in Section 2.1), we derive aggregation rules based on the AP of

tetrapeptides and pentapeptides from three perspectives (Figure 1), aiming to generalize the aggregation rules to longer oligopeptides (beyond tripeptides): **first**, we investigate the type and positional contribution of each amino acid to the AP within the complete sequence space of tetrapeptides and pentapeptides. In addition to the first-order aggregation rules considering a single amino acid (P1, P2, P3, P4, and P5 refer to one of the 20 amino acids at positions 1 to 5 of a pentapeptide, see Figure 1a), we also analyze the type and positional effect of 400 amino acid pairs (P1', P2', P3', and P4' refer to one of the 400 amino acid pairs, i.e., pairs of adjacent amino acids) on aggregation, referred to as the second-order effect (Figure 1a); **second**, we investigate the transferability relationship between  $AP_{te}$  and  $AP_{pe}$ , where  $AP_{te}$  and  $AP_{pe}$  represent the AP values of tetrapeptides and pentapeptides, respectively. A pentapeptide P1–P2–P3–P4–P5 can be considered as a “concatenation” of two tetrapeptides, i.e., P1–P2–P3–P4 and P2–P3–P4–P5, with the acylation of P4 of the first tetrapeptide or P2 of the second tetrapeptide (Figure 1b). By analyzing the relationship between the averaged  $AP_{te}$  of two tetrapeptides and the  $AP_{pe}$  of the corresponding pentapeptide, the AP of longer oligopeptides can potentially be obtained without performing molecular dynamics (MD) simulations or even machine learning prediction. We analyze the peculiar cases where  $AP_{te}$  and  $AP_{pe}$  show poor transferability by examining the contribution of each amino acid in the AP determination, especially the amino acids located at P4 of the first tetrapeptide or P2 of the second tetrapeptide; **finally**, we examine the morphologies of over 20,000 aggregating tetrapeptides, in terms of morphology distribution and effect of type and position of amino acids (Figure 1c). The standard deviation of parallel CGMD run for morphology and the effect of initial secondary structure in Martini force field version 2.2 are also investigated. In addition, we compare the morphologies of 66 tetrapeptide aggregates between CGMD simulations and experimental results, aiming to assess their discrepancies and suggest possible solutions to achieve more accurate computational morphologies.



**Figure 2.** Prediction and distribution of  $AP_{te}$  (and  $AP_{te}'$ ) and  $AP_{pe}$  (and  $AP_{pe}'$ ). (a,b) Predicted AP values of tetrapeptides ( $AP_{te}$ ) and pentapeptides ( $AP_{pe}$ ) compared to simulation-generated AP of tetrapeptides ( $AP_{sim,te}$ ) and pentapeptides ( $AP_{sim,pe}$ ). (c,d) Violin distribution of  $AP_{te}'$  (normalized  $AP_{te}$ ) and  $AP_{pe}'$  (normalized  $AP_{pe}$ ) within four ranges of A  $\in [0.00, 0.25)$ , B  $\in [0.25, 0.50)$ , C  $\in [0.50, 0.75)$ , and D  $\in [0.75, 1.00)$ , with the number of peptides counted in each range. The purple and yellow dots overlapping the “Total” distribution in (c) indicate the comparison results with experimental TEM images,<sup>23</sup> i.e., consistent and inconsistent on a qualitative level, respectively.

## 2. RESULTS AND DISCUSSION

### 2.1. Prediction of $AP_{te}$ and $AP_{pe}$

AP is utilized as a target in the training of the machine learning model, calculated as the ratio of accessible surface area at the beginning and at the end of a CGMD simulation (calculation details can be found in the Experimental Section 4.1 and Figure S1 of Supporting Information-1).<sup>20</sup> The  $AP_{te}$  and  $AP_{pe}$  predicted by the TRN Combo model<sup>20</sup> (see Section 4.2 for access of the TRN model) are compared to the ground truth data generated by the CGMD simulations, i.e., 5000  $AP_{sim,te}$  data and 10,000  $AP_{sim,pe}$  data (Figure 2a,b). In terms of prediction performance regarding pentapeptides, the model has a mean absolute error (MAE) of 0.04 and a coefficient of determination ( $R^2$ )<sup>22</sup> of 0.96, while for tetrapeptides, the MAE is 0.06 and  $R^2 = 0.90$ . Although the training data does not include  $AP_{te}$ , the TRN model exhibits reasonable extrapolation capability, corroborating the work in the field of computational chemistry regarding “out-of-distribution” prediction.<sup>18</sup>

Furthermore,  $AP_{te}$  and  $AP_{sim,te}$  (also  $AP_{pe}$  and  $AP_{sim,pe}$ ) exhibit similar distributions, as shown in the insets of Figure 2a (also 2b). The maximum error between  $AP_{te}$  and  $AP_{sim,te}$  is 0.32 for tetrapeptides (0.19 for pentapeptides), and the relatively large errors are mainly distributed in the small  $AP_{sim,te}$  and  $AP_{sim,pe}$  regimes (i.e.,  $[1.0, 1.5]$ ), which has an infinitesimal influence on the selection of aggregating peptides. Therefore, we conclude that the TRN model predicts  $AP_{te}$  and  $AP_{pe}$  with satisfactory accuracy.

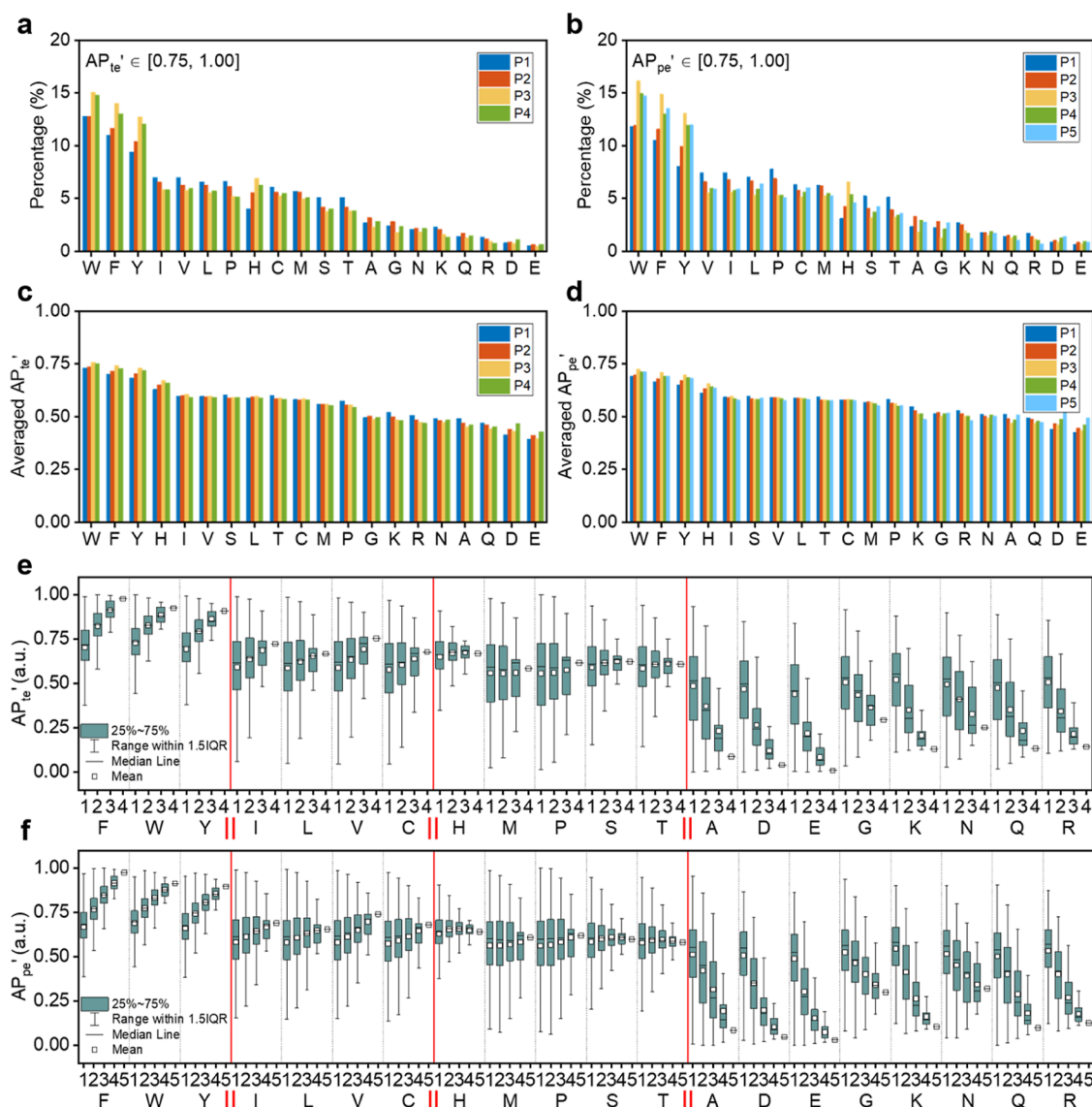
To gain an overview of  $AP_{te}$  and  $AP_{pe}$ , we plot the violin distribution of normalized  $AP_{te}$  and  $AP_{pe}$  (i.e.,  $AP_{te}'$  and  $AP_{pe}'$ ) of a total of 160,000 tetrapeptides and 3,200,000 pentapeptides (Figure 2c,d). The  $AP_{te}'$  and  $AP_{pe}'$  is calculated by  $AP' = (AP - AP_{min}) / (AP_{max} - AP_{min})$ , and they are categorized into four

ranges of A  $\in [0.00, 0.25)$ , B  $\in [0.25, 0.50)$ , C  $\in [0.50, 0.75)$ , and D  $\in [0.75, 1.00)$ .  $AP_{te}'$  and  $AP_{pe}'$  exhibit considerable similarity in distribution in each range, with an approximate peptide number ratio of 9:23:53:15. It can be inferred that it is an intrinsic nature that most of the peptides (>90%, summed over range B to D) tend to form aggregates (including precipitation) due to various interactions induced by the side chains, offering great potential for the development of versatile applications.

To further verify the accuracy of the predicted AP values, we qualitatively compare the 165 predicted AP values of tetrapeptides (whose distribution is shown as purple and yellow dots in Figure 2c) with experimental TEM results.<sup>23</sup> We categorize the peptides with a normalized AP of less than 0.34 (corresponding to 1.4 before normalization) in the simulations as non-aggregating peptides. Among the 165 AP values, 155 of them (e.g., purple dots) agree with the experiments, yielding an accuracy percentage of  $\sim 94\%$ . Therefore, we conclude that the prediction of aggregation is in reasonable agreement with experiments. It should be noted that the experimental data are not involved in the machine learning training process. Details of the comparison are shown in Table S1 in Supporting Information-1.

### 2.2. First-Order Aggregation Rules

To obtain a comprehensive and self-consistent view of the effect of amino acid type and position on aggregation, we perform statistical analyses on the distribution of individual amino acids from three aspects, referred to as first-order aggregation rules: **first**, we divide the AP values into four ranges (i.e., A, B, C, and D, Figure 2c,d) and calculate the percentage of each amino acid at each position within the four ranges (Figure 3a,b for range D of tetrapeptides and pentapeptides, respectively; Figure S2a–c of Supporting

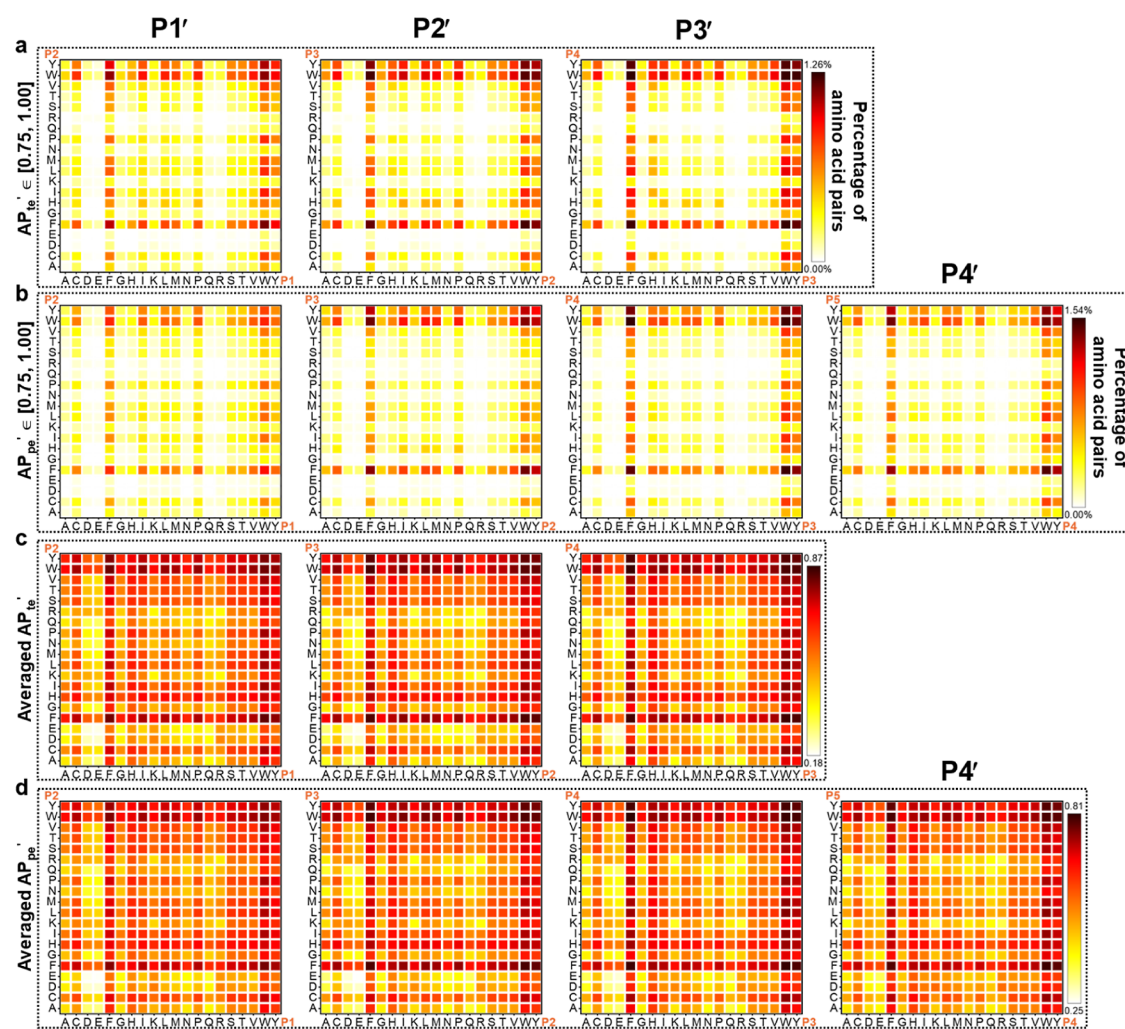


**Figure 3.** Aggregation rules in terms of individual amino acids (i.e., first-order aggregation rules). (a,b) Percentage of each amino acid at each position within the pool of aggregating peptides (i.e., the normalized AP values within the range  $D \in [0.75, 1.00]$ ) for (a) tetrapeptides and (b) pentapeptides. (c,d) Averaged AP values of the peptides with a fixed amino acid at a specific position for (c) tetrapeptides and (d) pentapeptides. (e,f) Averaged AP values of peptides containing (e) 1–4 and (f) 1–5 specific amino acids.

**Information-1** for ranges A, B, and C of tetrapeptides; and Figure S2d–f of **Supporting Information-1** for ranges A, B, and C of pentapeptides). Next, we compute the averaged AP values of the peptides in which a given amino acid is fixed at the one of the four (or five) positions (Figure 3c for the tetrapeptides and Figure 3d for the pentapeptides). For each amino acid at each position, a total of 8000 and 16,000 tetrapeptide and pentapeptide sequences, respectively, are enumerated for calculating the averaged AP values. For example, for tetrapeptides with F fixed at position 1 (or positions 2, 3, and 4), we average the AP values of the 8000 peptides in which the remaining three positions are occupied by any of the 20 amino acids including F. Finally, we examine the distribution of AP values with respect to different contents of a particular amino acid in the peptides, regardless of its position (Figure 3e for tetrapeptide and 3f for pentapeptide). For tetrapeptides, the numbers of peptides containing one, two, three, and four specific amino acid(s) are 27,436 ( $=19 \times 19 \times 19 \times 4$ ), 2166 ( $=19 \times 19 \times 6$ ), 76 ( $=19 \times 4$ ), and 1, respectively. For

pentapeptides, the corresponding numbers of peptides containing one, two, three, four, and five specific amino acid(s) are 651,605 ( $=19 \times 19 \times 19 \times 19 \times 5$ ), 68,590 ( $=19 \times 19 \times 19 \times 10$ ), 3610 ( $=19 \times 19 \times 10$ ), 95 ( $=19 \times 5$ ), and 1, respectively. The bold numbers in parentheses indicate the number of possible positions of a single amino acid and its doublet, triplet, and quadruplet (quadruplet for pentapeptides only). The distributions of the AP values of peptides with a single amino acid and its doublet, triplet, and quadruplet at different positions are shown in **Supporting Information-2**: Figures S1–20 for tetrapeptides and Figures S21–40 for pentapeptides.

As observed from the aggregation range (i.e., range D,  $AP'_{te}$  or  $AP'_{pe} \in [0.75, 1.00]$ ), the aromatic amino acids W, F, and Y contribute most when they are located at positions from the middle to the C-terminus (Figure 3a,b), especially at the middle position in pentapeptides (Figure 3b). To provide a physical rationale of this observation, we analyze the molecular level interactions in three different scenarios as the aromatic



**Figure 4.** Aggregation rules in terms of 400 amino acid pairs (i.e., second-order aggregation rules). (a,b) Percentage of amino acid pairs in the high AP range of D ( $AP' \in [0.75, 1.00]$ ) for (a) tetrapeptides and (b) pentapeptides. (c,d) Averaged AP values of (c) tetrapeptides and (d) pentapeptides with amino acid pairs fixed at specific “positions”.

amino acids are located at the N-terminus, C-terminus, and the middle position, respectively: (1) when the aromatic amino acids are located at the N-terminus, the  $\text{NH}_3^+$  groups of the amino acids act as hydrogen donors, which is prescribed in the Martini force field.<sup>24,25</sup> This contributes to the formation of hydrogen bonds and specific angles between the nitrogen atom at the terminus and the backbone carbon atoms of the aromatic amino acids, which are not conducive to strong  $\pi$ - $\pi$  interactions that require benzene rings in a parallel direction; (2) as the aromatic amino acids are located at the C-terminus, the zwitterionic-state carboxyl groups of the amino acids act as hydrogen acceptors. Since the C-terminus is predisposed to the zwitterionic state, the terminus would interact strongly with other peptides through Coulombic forces and hydrogen bonding. However, the formation of specific structures through Coulombic interactions would also possibly incur the steric effect for effective  $\pi$ - $\pi$  interactions; and (3) as the aromatic amino acids are located in the middle of a peptide chain, they have more degrees of freedom to attract each other through  $\pi$ - $\pi$  interactions, leading to the strongest AP.

Other statistical results also confirm the contributions of W, F, and Y to aggregation. The averaged  $AP'$  is the highest when W, F, and Y are located at the middle to C-termini positions (Figure 3c,d), and increasing the number of W, F, and Y in

peptide chains can significantly increase the  $AP'$  at the statistical level (Figure 3e,f), especially when their doublets, triplets, or quadruplets are located at the middle to C-termini positions within the chain (Supporting Information-2: Figures S5 and S25 for F, Figures S19 and S39 for W, and Figures S20 and S40 for Y).

Second to aromatic amino acids, amino acids I, L, V, P, H, C, M, S, and T also contribute positively to aggregation (Figure 3). Except for H, all other amino acids I, L, V, P, C, M, S, and T contribute more to aggregation when they occupy positions close to the two termini, especially at the N-terminus (Figure 3a,b). As I, L, V, and M are located at the N-terminus, they act as hydrogen donors and contribute to the interactions with the C-terminus of other peptides; as they are located at the C-terminus, they act as hydrogen acceptors, which incurs a strong tendency to interact with water, potentially reducing the hydrophobic effect of the side chains and thus inducing less aggregation than at the N-terminus. It should be noted that the exposure of hydrophobic side chains at both termini is conducive to aggregation. For the amino acid P, the side chains of the five-membered rings remain intact when located at the N-terminus, while the backbone structure is altered for the formation of the amino bond when located at the C-terminus, possibly allowing special packing due to its unique “kink

structure” at the N-terminus, consistent with the reported results.<sup>17</sup> The amino acid C is a relatively simple case, as the sulfur bond is not specifically parametrized in the Martini force field, and the only difference is that the backbone is prone to interact with water due to the “hydrogen acceptor” nature at the C-terminus, which reduces the AP. S and T with polar side chains can also contribute to aggregation by forming hydrogen bonds with each other, and for the same reason as amino acid C, the “hydrogen acceptor” nature at the C-terminus induces less aggregation than that at the N-terminus. Similar to aromatic amino acids, the amino acid H prefers the middle position to the C-terminus and especially to the middle position, as peptides with H in the middle position have more degrees of freedom to attract each other through  $\pi$ - $\pi$  interactions, leading to the strongest AP.

The remaining amino acids A, D, E, G, K, N, Q and R generally have a statistically negative effect on aggregation due to their strong hydrophilicity and the repulsion induced by Coulombic interactions. This is evidenced by their scarcity in the aggregation range D (Figure 3a,b) and the decreasing AP values with increasing content of these amino acids (Figure 3e,f). It should be noted, however, that peptides containing these amino acids can still achieve a strong tendency to aggregate when they are placed at positions that are favorable for aggregation. For example, negatively charged amino acids D and E and positively charged amino acids K and R could promote aggregation when placed at the C- and N-terminus, respectively, due to intermolecular alignment by repulsion of equal charges and possible formation of salt bridges<sup>17</sup> (Figure 3a–d; Supporting Information-2 Figures S3, S4, S9, and S15 for D, E, K, and R in tetrapeptides; and Figures S23, S24, S29, and S35 for D, E, K, and R in pentapeptides, respectively).

In summary, hydrophobicity (additional discussion regarding the correlation between hydrophobicity and AP is detailed in Supporting Information-1), hydrogen bonding, unique kink structure, and polarity all have a secondary effect on aggregation compared to  $\pi$ -stacking. It is challenging to quantify the effect of each type of interaction on aggregation using a specific number; alternatively, our quantitative measurement of the percentage of 20 standard amino acids at each position within each AP range should provide a qualitative guide for experimentalists to manipulate aggregation of peptides.

### 2.3. Second-Order Aggregation Rules

Since the analysis of AP in terms of the distribution of individual amino acids does not provide sufficient information about which an amino acid should be placed next to the desired amino acid to promote aggregation, we therefore perform analyses of the type and positional distribution of a total of 400 ( $=20^2$ ) amino acid pairs, termed second-order aggregation rules (Figure 4). For each amino acid pair in a tetrapeptide, they can occupy three possible “positions”: P1' (P1–P2), P2' (P2–P3), and P3' (P3–P4). For pentapeptides, each amino acid pair can occupy an additional fourth position, P4' (P4–P5). Similar to the analysis with respect to individual amino acids, we analyze the percentage of amino acid pairs with respect to each position in each given AP range [specifically, Figure 4a,b for the AP range of D  $\in$  [0.75, 1.00) and Supporting Information-1 Figures S3 and S4 for the ranges of A  $\in$  [0.00,0.25), B  $\in$  [0.25,0.50), and C  $\in$  [0.50,0.75)], for tetrapeptides and pentapeptides, respectively]. In addition, we calculate the averaged AP values of peptides

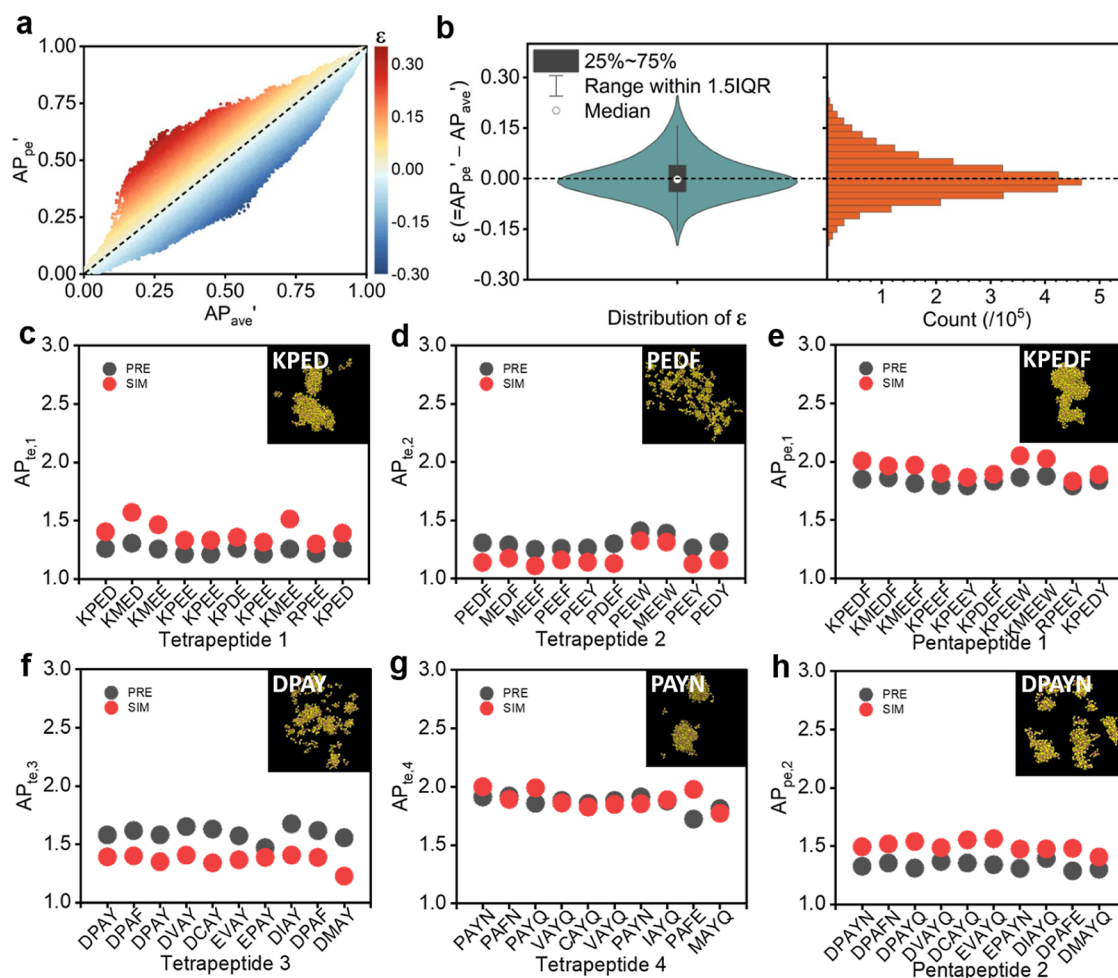
with a specific amino acid pair placed at each position (Figure 4c,d). For example, if FF is fixed at P1' of a tetrapeptide, we average the AP values of 400 tetrapeptides with P3 and P4 occupied by any of the 20 amino acids.

Within the aggregating range (i.e., range D) of tetrapeptides and pentapeptides, it can be observed that the F-, W-, and Y-containing doublets (e.g., FF, FW, and FY) are the most prominent, especially when the doublets are located at P3'/P4' near the middle to the C-terminus. Second to the aromatic doublets, the doublets of aromatic amino acids coupled with C, H, I, L, M, P, S, T, and V can also promote aggregation, which is congruent with the aggregation rules regarding single amino acids. Finally, the doublets formed by A, D, E, G, K, N, Q, and R are rarely found in the aggregating range. It should be noted that the doublets formed by the above amino acids and F, W, and Y may still occur (such as AF, AW, GF, etc.) in aggregating peptides.

It is important to notice that the second-order effect is not a mere superposition of the first-order aggregation rules. To distinguish second-order aggregation rules from the superposition of two individual first-order aggregations, we compare the AP rankings of tetrapeptides with F and H at the third and fourth positions (Table S2 in Supporting Information-1), respectively. NNFN ranks lower than NNHN (NNFN has a lower AP), and NNNF ranks lower than NNNH. However, when FF acts synergistically at the end of the tetrapeptide, it significantly increases the AP value. Consequently, NNFF ranks higher than NNHH, with a ranking difference of 34,576. This implies that when analyzing NNFF, we cannot simply consider it as the sum of the effects of F at the third position and F at the fourth position. Instead, we must take into account the synergistic effect of FF, including hydrogen bonding,  $\pi$ - $\pi$  stacking interactions, hydrophobic interactions, and possible changes in the secondary structure incurred by the roles of amino acids at different positions within the peptide chain. Similarly, when F and H are positioned in the middle, we observe the same phenomenon. NFNN ranks 10,780 positions lower than NHNN, and NNFN ranks 6745 positions lower than NNHN. However, when FF acts synergistically, NFFN ranks 32,761 positions higher than NHNN. The same pattern is observed when F and H are at the beginning of the tetrapeptide. FNHN ranks 14,454 positions lower than HNNN, and NFNN ranks 10,780 positions lower than NHNN. However, the synergistic effect of FF results in FFNN having a higher AP value and ranking 30,303 positions higher than HHNN.

### 2.4. Transferability of AP

Peptide aggregation should be promoted or prevented depending on different situations. For example, aggregation should be promoted for hydrogel formation,<sup>23</sup> while it should be mitigated for possible prevention of neurodegenerative diseases.<sup>26</sup> Thus, understanding the mechanism of aggregation and developing tuning approaches/laws are of great significance for the design of desirable peptides. To achieve the goal of manipulating aggregation, we have undertaken a detailed investigation into the transferability relationship of APs' between tetrapeptides and pentapeptides, through which we hope to tune the aggregation tendency by simply adding or removing specific amino acids within the sequences, inspired by our previous work that the concatenation of pentapeptides to decapptides can promote or prevent aggregation.<sup>20</sup>



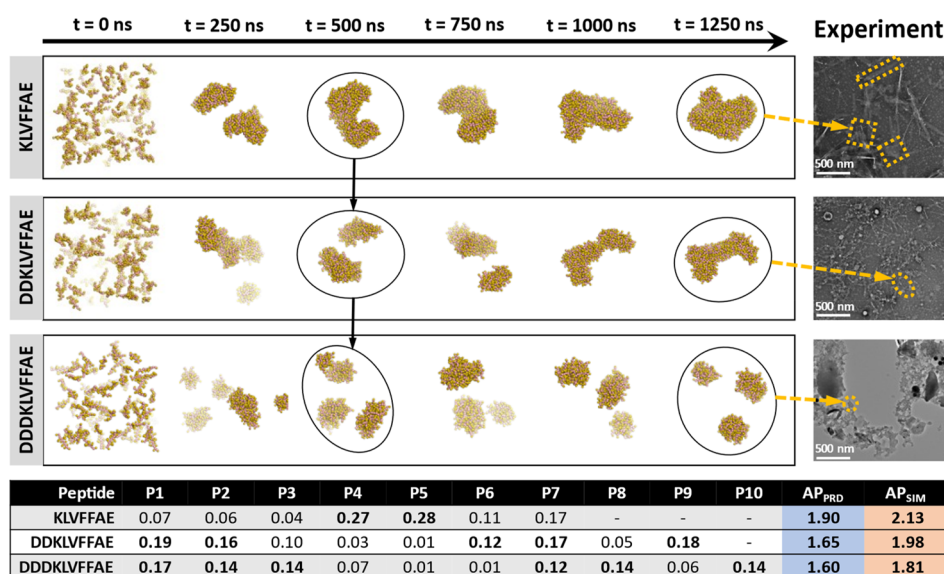
**Figure 5.** Transferability relationship between  $AP_{pe}'$  and averaged  $AP_{ave}'$  ( $AP_{ave}'$  are the average of two  $AP_{te}'$ ). (a) Relationship between  $AP_{pe}'$  and  $AP_{ave}'$ , with the color indicating the difference  $\varepsilon$  between  $AP_{pe}'$  and the averaged  $AP_{ave}'$  ( $\varepsilon = AP_{pe}' - AP_{ave}'$ ). (b) Violin distribution (left) and number distribution (right) of  $\varepsilon$ . (c–e) AP values of 10 groups of peptides with positive maximum  $\varepsilon$  between  $AP_{pe}$  and  $AP_{ave}$ . (f–h) AP values of the 10 groups of peptides with negative maximum  $\varepsilon$  between  $AP_{pe}$  and  $AP_{ave}$ .

We first calculate the averaged AP' ( $AP_{ave}'$ ) values of two tetrapeptides [ $AP_{ave}' = (AP_{te,1}' + AP_{te,2}')/2$ ] and then examine the difference between the  $AP_{ave}'$  and  $AP_{pe}'$  of pentapeptides composed of these two tetrapeptides, such as KPED, PEDF, and KPEDF. Examining the sequence relationship of the peptides KPED, PEDF, and KPEDF, it is straightforward to observe that the pentapeptide KPEDF is one amino acid addition of the tetrapeptides by the concatenation of F at the C-terminus of KPED or K at the N-terminus of PEDF, through which we hope to alter the aggregation tendency of the matrix peptides. Figure 5a illustrates the correlation between 3.2 million groups of  $AP_{pe}'$  and  $AP_{ave}'$ , and Figure 5b shows the distribution of the difference between  $AP_{pe}'$  and  $AP_{ave}'$ , as denoted by  $\varepsilon$ . Although most of the  $AP_{pe}'$  ( $\sim 80\%$ ) values are similar to the  $AP_{ave}'$  ( $|\varepsilon| \leq 0.15$ ) values, the aggregation tendency of pentapeptides can be significantly altered compared to the corresponding tetrapeptides, as indicated by the fact that the largest absolute difference  $|\varepsilon|$  can reach over 0.3.

To understand the causes of the altered aggregation tendency, we then select 10 groups of peptides with the largest  $\varepsilon$  (Figure 5c–e) and another 10 groups with the smallest  $\varepsilon$  (Figure 5f–h) to determine the amino acids that have the dominant contributions to the AP values by

performing attribution analysis of each amino acid to AP (Section 4.3 for calculation details of attribution). To validate the predicted results here, we perform additional CGMD simulations on the selected tetrapeptides and pentapeptides for 200 ns (“effective time” of 800 ns) to generate ground-truth AP values. 1250 ns CGMD simulations are also performed on tetrapeptides to confirm convergence (Figure S5 of Supporting Information-1). The calculated attribution of each amino acid and the corresponding predicted and simulated AP values (denoted by  $AP_{PRD}$  and  $AP_{SIM}$ , respectively) are shown in Table S3 in Supporting Information-1.

Examining Figure 5c,e (also rows Te1 and Pe1, i.e., tetrapeptide 1 and pentapeptide 1, of Table S3 in Supporting Information-1), it is observed that the AP of tetrapeptides can be increased by placing an aromatic amino acid at the last position of the tetrapeptide sequence, e.g., KPED  $\rightarrow$  KPEDF. This is consistent with the first-order aggregation rules that aromatic amino acids play a more dominant role in contributing to aggregation when located at the C-terminus, which is also evidenced by the fact that the attribution of the aromatic amino acids at the last position of a pentapeptide is generally the largest (e.g., 0.138, 0.012, 0.225, 0.177, and 0.229 for each amino acid in KPEDF, as shown in the row Pe1 of Table S3 in Supporting Information-1). Examining Figure 5d,e



**Figure 6.** Computational and experimental results of AP and morphologies of peptide  $A\beta_{16-22}$  with the addition of amino acids DD and DDD to the N-terminus. The table lists the attribution of each amino acid with the increasing number of D, as well as the predicted ( $AP_{PRD}$ ) and simulation AP ( $AP_{SIM}$ ).

(also rows Te2 and Pe1, i.e., tetrapeptide 2 and pentapeptide 1, of Table S3 in [Supporting Information-1](#)), we observe that the AP can also be significantly increased by adding a positively charged amino acid into the sequence of a tetrapeptide with a net negative charge to mitigate electrostatic repulsion, e.g., PEDF with  $2e^- \rightarrow$  KPEDF with  $1e^-$ . This is also consistent with the first-order aggregation rules that the positively charged amino acids yield a more positive contribution to aggregation when they are located at the N-terminus, even though they do not have the largest attribution (i.e., 0.138 in KPEDF).

In addition to promoting aggregation by adding aromatic or charged amino acids, the aggregation tendency can also be attenuated by simply adding a negatively charged amino acid (or possibly positively charged amino acids) at the N-terminus, even if aromatic amino acids are present in the sequence. The peptides in [Figure 5f,h](#) (also rows Te3 and Pe2, i.e., tetrapeptide 3 and pentapeptide 2, of Table S3 in [Supporting Information-1](#)) all have relatively low AP values due to the presence of negatively charged amino acids at the N-terminus. Also, the attributions of the negatively charged amino acids in these peptides are the largest. This is further evidenced by the fact that the AP values of the peptides in [Figure 5g](#) (also row Te4, i.e., tetrapeptide 4, of Table S3 in [Supporting Information-1](#)) are significantly reduced by the addition of negatively charged amino acids at the N-terminus ([Figure 5h](#)).

In summary, the competition between the  $\pi-\pi$  interactions and Coulombic interactions between amino acids at different positions dominates the aggregation tendency: the  $\pi-\pi$  interactions at the C-terminus induced by aromatic amino acids can compete with the electrostatic repulsion induced by only one positive charge at the N-terminus, while not a negative charge. For example, PEDF has a low AP value because this peptide has two charges ( $2e^-$ ), and DPAY has a low AP value because the N-terminus is occupied by a negative charge. The rules summarized here can be seen as complementary to the aggregation rules in [Section 2.2](#) above, emphasizing that placing aromatic and charged amino acids at

certain positions or tuning the overall charge of peptides can potentially promote or mitigate peptide aggregation.

Inspired by the above rules, we concatenate the negatively charged amino acids DD and DDD to the N-terminus of the  $A\beta_{16-22}$  peptide, i.e., KLVFFAE, trying to inhibit the aggregation of these hydrophobic cores. In addition, we calculate the attribution of each amino acid in peptides KLVFFAE, DDKLVFFAE, and DDDKLVFFAE, aiming to compare the significance of FF to the aggregation with increasing number of amino acids D. It is found that with increasing number of D at N-termini, the aggregation process is decelerated ([Figure 6](#)). For example, at the simulation time of 500 ns, KLVFFAE has aggregated into one big cluster, while peptides DDKLVFFAE and DDDKLVFFAE are still in the processing of aggregating, and at 1250 ns, the peptide DDDKLVFFAE has reached convergence in aggregating ([Figure 6](#)) and formed clusters with much smaller size compared with KLVFFAE.

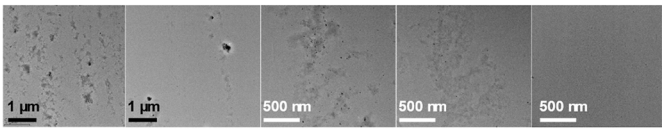
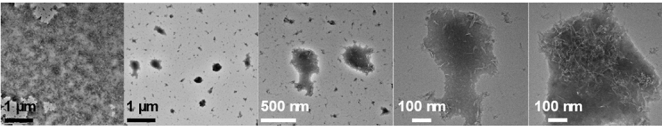
The experimental results suggest that the peptide KLVFFAE can form large nanoribbons with piece size around 200–500 nm; the peptide DDKLVFFAE can form nanofibers with much smaller size in diameter and length, while the peptide DDDKLVFFAE barely forms amorphous aggregates, all consistent with our prediction and simulation results. The attribution analysis suggests that the contribution of the FF moiety to the aggregation is significantly decreased (table in [Figure 6](#)), consolidating the aggregation rules summarized above. The computational and experimental details can be found in Experimental [Section 4.1](#).

### 2.5. Attribution Analysis of Decapeptides

In the previous section, attribution analysis is used to understand the importance of each amino acid in the transferability relationship between the aggregation tendency of tetrapeptides and pentapeptides. The attribution analysis can also be extended to the complete sequence space of oligopeptides, where the percentage analysis (as in [Figures 3 and 4](#)), requiring the prior known AP values of all peptides, cannot be performed. For example, it is extremely challenging



**Table 1. Attribution of Each Amino Acid to the Aggregation Tendency in Decapeptides and the TEM/Simulation Images of a Decapeptide ARRERVGNGKR with a Low AP<sub>PRD</sub> Value (=1.26) and a Decapeptide TFFFLYWVHFV with a High AP<sub>PRD</sub> Value (=2.18)<sup>a</sup>**

Property	Peptide	P1-N	P2	P3	P4	P5	P6	P7	P8	P9	P10-C	AP <sub>PRD</sub>
Non- to mid-tier aggregating decapeptides	VMAQNEDVLA	0.02	0.01	0.09	<b>0.11</b>	<b>0.14</b>	<b>0.26</b>	<b>0.22</b>	0.01	0.02	0.12	1.39
	VLELQLAEAP	0.04	0.04	<b>0.28</b>	0.03	<b>0.11</b>	0.03	0.09	<b>0.25</b>	<b>0.12</b>	0.01	1.50
	NNAQGNAASC	<b>0.14</b>	<b>0.14</b>	0.12	0.11	0.08	<b>0.13</b>	0.12	<b>0.14</b>	0.02	0.00	1.42
	KEMALELDQC	<b>0.14</b>	<b>0.23</b>	0.02	0.08	0.01	<b>0.18</b>	0.00	<b>0.19</b>	0.13	0.01	1.50
	CEGNGNDAP	0.01	0.00	<b>0.25</b>	0.09	<b>0.11</b>	<b>0.11</b>	0.09	<b>0.19</b>	<b>0.12</b>	0.04	1.34
	VVQQMVCENE	0.03	0.02	0.10	<b>0.11</b>	0.02	0.00	0.01	<b>0.27</b>	<b>0.17</b>	<b>0.26</b>	1.53
	TDMMPLSEEE	0.00	<b>0.26</b>	0.01	0.02	0.01	0.01	0.02	<b>0.24</b>	<b>0.22</b>	<b>0.21</b>	1.39
	DNICSEPPDM	<b>0.32</b>	<b>0.11</b>	0.03	0.01	0.00	<b>0.25</b>	0.02	0.04	<b>0.21</b>	0.00	1.38
	AEVNTCDAML	<b>0.15</b>	<b>0.34</b>	0.01	<b>0.11</b>	0.00	0.00	<b>0.25</b>	<b>0.11</b>	0.01	0.03	1.42
	ARRERVGNGKR	0.10	<b>0.12</b>	<b>0.12</b>	0.01	<b>0.12</b>	0.02	0.08	0.09	<b>0.15</b>	<b>0.18</b>	<b>1.26</b>
TEM images of ARRERVGNGKR with different length bar												
Strong-tier aggregating decapeptides	ACFQWIVYWN	0.11	0.02	<b>0.12</b>	0.05	<b>0.16</b>	0.03	0.02	<b>0.16</b>	<b>0.24</b>	0.11	1.95
	WRYLTGWGVCV	<b>0.20</b>	0.05	<b>0.13</b>	0.06	0.00	<b>0.16</b>	0.05	<b>0.23</b>	0.04	0.09	2.03
	QWSMFIVFVK	<b>0.15</b>	<b>0.25</b>	0.02	0.05	<b>0.19</b>	0.03	0.02	<b>0.22</b>	0.03	0.04	1.92
	LMFWSCHSFS	0.09	0.08	<b>0.18</b>	<b>0.21</b>	0.00	0.05	<b>0.11</b>	0.01	<b>0.26</b>	0.01	2.04
	KYGFWYFHCI	0.01	0.10	0.08	<b>0.14</b>	<b>0.16</b>	<b>0.13</b>	<b>0.17</b>	0.07	0.05	0.09	2.06
	ALYWCYNEFT	0.10	0.06	<b>0.13</b>	<b>0.14</b>	0.03	<b>0.14</b>	0.06	<b>0.13</b>	<b>0.20</b>	0.00	1.91
	CIYKCYIVWW	0.04	0.06	<b>0.13</b>	0.01	0.04	<b>0.14</b>	0.05	0.03	<b>0.22</b>	<b>0.27</b>	2.13
	CFPFWGLCFF	0.03	0.12	0.05	<b>0.13</b>	<b>0.15</b>	0.02	0.05	0.03	<b>0.20</b>	<b>0.21</b>	2.18
	VPCYYSMLWY	0.06	0.07	0.03	<b>0.14</b>	<b>0.14</b>	0.00	0.04	0.04	<b>0.25</b>	<b>0.22</b>	2.13
TFFFLYWVHFV	0.00	<b>0.12</b>	<b>0.12</b>	<b>0.12</b>	0.06	0.11	<b>0.16</b>	0.04	<b>0.18</b>	0.09	<b>2.18</b>	
TEM images of TFFFLYWVHFV with different length bar												

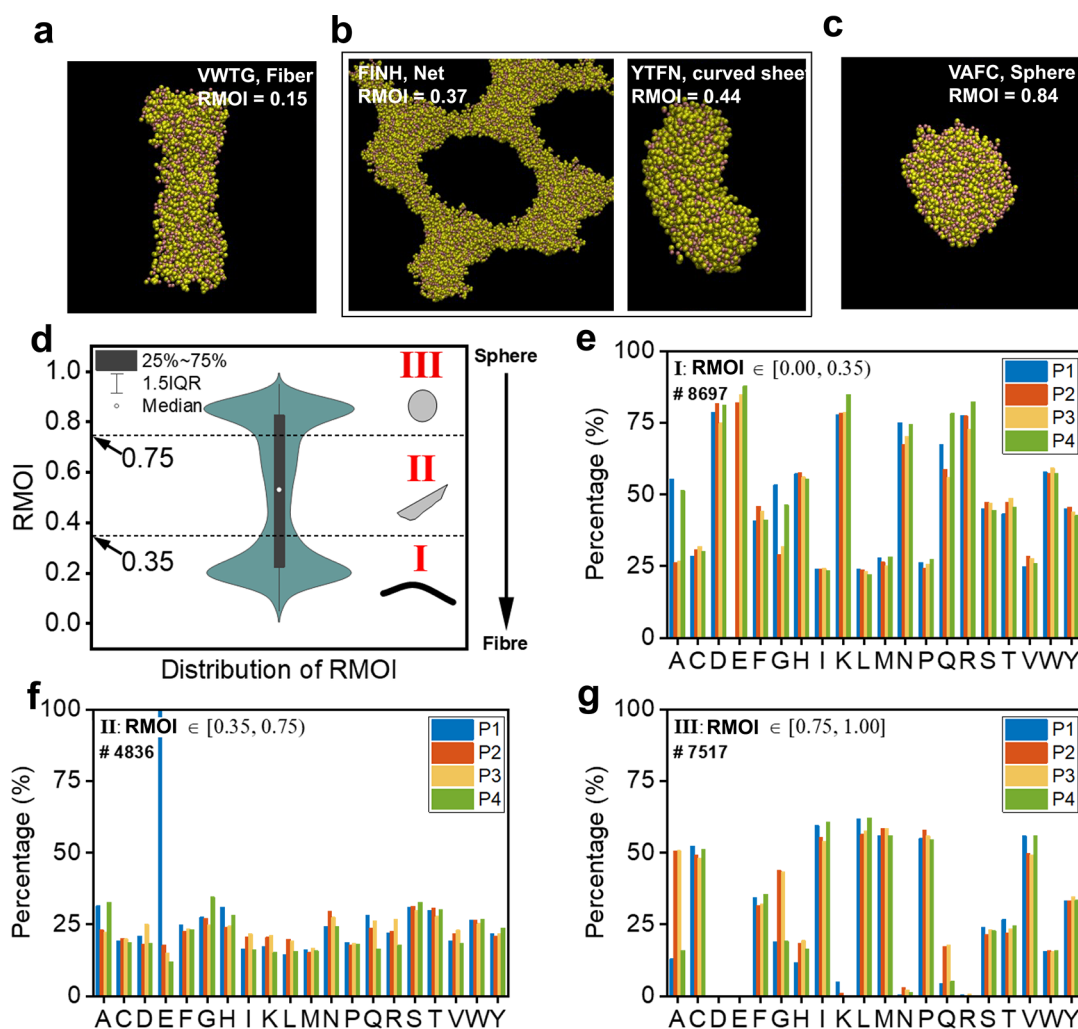
<sup>a</sup>The amino acids in bold (and the associated attributions) are the top 4 amino acids at specific positions that dominate the aggregation in selected decapeptides. The P1 (i.e., N-terminus) to P10 (i.e., C-terminus) indicate 10 positions in the decapeptide sequence.

to analyze the dominance of each amino acid in the sequence of decapeptides since it is impossible to list the more than 10 trillion APs of all decapeptides. However, the attribution predicted based on, for example, 10,000 data can provide a quantitative measure of the contribution of each amino acid to the aggregation tendency, with which we can potentially tune the type or position of the dominant amino acid(s) for controlling the aggregation tendency of peptides. We selected 20 decapeptides to verify the accuracy of the attribution of amino acids (Table 1). It is suggested that highly polar (e.g., N and Q) and charged amino acids play dominant roles in reducing the aggregation tendency, while aromatic amino acids can significantly promote aggregation, which is congruent with all of the first-order and second-order aggregation rules mentioned above, corroborating the accuracy of the attribution analysis. Specifically, the decapeptide ARRERVGNGKR is predicted to be non-aggregating, while TFFFLYWVHFV can aggregate due to the large attribution of charged and aromatic amino acids, respectively, congruent with the experimental results. The experimental transmission electron microscopy (TEM) results corroborate the observation that, for the decapeptide sequence TFFFLYWVHFV, an increased number of particles are discernible across various length scales within the different visual fields.

## 2.6. Morphologies of Aggregated Peptides

In addition to the aggregation tendency of short peptides, the morphologies after aggregation also play a critical role in determining the various properties of the aggregates and the development of subsequent applications.<sup>14,27–30</sup> For example, aligned peptide tubes could increase the flexural stiffness, while nanospheres could increase the compressive stiffness,<sup>30</sup> analogous to the role of aligned metal bars and crushed stones in macroscale-reinforced concrete. Based on the predicted AP values, we select 21,050 aggregating tetrapeptides with predicted AP values larger than 1.44, and 93% of them have an AP larger than 1.8. The distribution of predicted AP values of the selected peptides is shown in Figure S6a of Supporting Information-1. We perform CGMD simulations on each type of peptide with a duration of 1.25  $\mu$ s (an “effective” time of 5  $\mu$ s due to the coarse-graining effect), for assessing the distribution and accuracy of computational morphologies.

We characterize the morphologies using the ratio of the moments of inertia (RMOI) along the principal axes of the largest cluster of aggregates in the system. Additional calculation details of the RMOI can be found in Section 4.5, and the correlation between the RMOI and AP values of the selected tetrapeptides is shown in Figure S6b of Supporting



**Figure 7.** Distribution of morphologies after aggregation. (a–c) Possible morphologies formed in simulations, such as fibers or tubes, intermediate structures such as net and curved sheet, and spherical or vesicle structures. (d) Violin distributions of morphologies. (e–g) Positional percentage of each amino acid within each RMOI range, calculated through dividing the number of each amino acid at each position within a RMOI range, by the total number of the amino acid at the position across all ranges.

**Information-1.** We then categorize the RMOI values into three classes (i.e., I, II, and III), which are I fibers and possibly tubes with  $\text{RMOI} \in [0.00, 0.35)$  (I, Figure 7d); II intermediate shapes with  $\text{RMOI} \in [0.35, 0.75)$ , such as net, rod, and curved sheet (II, Figure 7d); and III spheres and possibly vesicles with  $\text{RMOI} \in [0.75, 1.00]$  (III, Figure 7d). The violin distribution of the RMOI values of the 21,050 aggregating peptides show a dumbbell distribution: approximately 41.6% of peptides form fibers ( $\text{RMOI} < 0.35$ ), while 35.3% form spheres and possibly vesicles ( $\text{RMOI} \geq 0.75$ ), and the remaining 23.1% exhibit a shape in between of fibers and spheres ( $0.35 \leq \text{RMOI} < 0.75$ ), such as rods, sheets, or nets, illustrating the diversity of the morphologies that tetrapeptides can form. Additional discussions regarding the scarcity of peptides with intermediate RMOI values around 0.4–0.5 are also included in Supporting Information-1.

Within each RMOI class, we quantitatively evaluate the effect of the type and position of each amino acid on the morphology. It should be noted that those 21,050 peptides are selected based on AP values; thus, there exist original amino acid percentage difference in each position. To eliminate the effect of the original percentage difference, we calculate the

ratio of the number of each amino acid at each position within each RMOI range to the total number of the amino acid at the position across all ranges (i.e., 21,050 tetrapeptides), as shown in Figure 7e–g, while the percentage of each amino acid at each position of all 21,050 tetrapeptides, as well as that within each range of RMOI, is shown in Figure S7 of Supporting Information-1.

All amino acids are found in the fiber-forming sequence, except that amino acid E is absent at the P1 position, i.e., the N-terminus (Figure 7e), indicating that peptides with amino acid E at the N-terminus generally do not aggregate into fibers (or do not aggregate at all). The charged amino acids D, E, K, and R are rarely found in sphere-forming sequences. Instead, they mainly contribute to fiber formation, especially when they are located at the C-terminus. This is likely because a fibrous shape can reduce the Coulombic energy in the aggregate, suggesting that electrostatic force may play an important role in tuning the morphology during the peptide self-assembly process. Comparing the ratios of F, W, and Y in Figure 7e,g, it can be inferred that aromatic amino acids F, W, and Y generally contribute more to fiber formation (especially amino acid W). Among the three, F and W both prefer P3 for fiber

formation, while Y prefers P2. The amino acids with polar uncharged side chains, including S, T, N, and Q, contribute more to fiber formation than to sphere formation (Figure 7e), which can be attributed to the directional nature of hydrogen bonding between peptides containing these amino acids. Among the four, S and T prefer P2 and P3 for fiber formation, while N and Q prefer the N- or C-termini. Amino acids A and G have a similar overall propensity to form fibers or spheres, but when located at the N or C terminus, fiber formation is preferred (Figure 7e). Amino acids with hydrophobic side chains, including I, L, V, M, C, and P, contribute more to sphere formation than to fiber formation (Figure 7g). This can be attributed to the tendency of hydrophobic interactions to reduce the surface area of aggregates in an aqueous solution, resulting in a tendency toward spherical morphologies. Nevertheless, it is challenging to provide a complete rationale for the effect of type and position on the morphology for all amino acids. In the future, all-atom simulations and experiments may be performed to provide further atomistic details of the aggregation process and to determine the full mechanisms of the formation of different morphologies.

Additional discussions regarding the standard deviation of computational morphology, effect of the initial setting of secondary structure on aggregation, and comparisons of the morphologies between simulations and experiments can be found in the discussions of Supporting Information-1.

### 3. CONCLUSIONS

This study addresses the need for an enhanced understanding of the aggregation rules governing short peptides and has pushed the limit of aggregation rules within 8000 tripeptides to millions of pentapeptides. The aggregation rules are derived quantitatively based on the AP values, predicted by a transformer-based regression network (TRN) model trained on the data produced by MD simulations. The achievement of approximately 94% accuracy in AP prediction compared to 165 experimental results<sup>23</sup> demonstrates the reliability of the MD data for machine learning training and the effectiveness of the proposed TRN model.

By deriving comprehensive aggregation rules, this study contributes to the precise manipulation of the aggregation of short peptides, which is relevant for both the development of peptide-based applications (such as hydrogels, emulsions, and pharmaceuticals) and the understanding of pathological conditions related to peptide aggregation (such as Alzheimer's and Parkinson's disease). The aromatic amino acids W, F, and Y are found to contribute most to aggregation when located at the middle to C-terminus due to the greater degree of freedom (less steric effect) for the  $\pi$ - $\pi$  interactions of the aromatic rings. Second to the aromatic amino acids, the amino acids I, L, V, P, H, C, M, S, and T also contribute positively to the aggregation due to various effects such as hydrophobicity and hydrogen bonding, while the amino acids A, D, E, G, K, N, Q, and R generally have a statistically negative effect due to strong hydrophilicity or Coulombic repulsion. The same rules apply to the amino acid pairs (e.g., FF > FI > FA).

In addition, this work investigates the transferability relationship between the AP values of tetrapeptides and pentapeptides, offering a possible approach for tuning the aggregation behavior by concatenating or reducing aromatic or charged amino acids at the N- and C-termini. AP can be increased by placing an aromatic amino acid at the C-termini of the tetrapeptide sequences, while it can be decreased by

simply concatenating a negatively charged amino acid (or possibly positively charged amino acids) at the N-terminus, even in the presence of aromatic amino acids in the sequence. In summary, the  $\pi$ - $\pi$  interactions at the C-terminus induced by aromatic amino acids can compete with the electrostatic repulsion induced by a single charged amino acid if the charged amino acid is not located at the N-terminus.

Furthermore, this study provides a comprehensive statistical analysis of the morphologies of the aggregates and relates them to the type and positional contributions of individual amino acids. It is found that charged amino acids D, E, K, and R tend to promote the formation of fibrous aggregates, especially when they are located at the C-terminus, as a fibrous shape could reduce the Coulombic energy in the aggregates. In contrast, amino acids with hydrophobic side chains, including I, L, V, M, C, and P, contribute more to the formation of spherical aggregates, which can be rationalized, given that hydrophobic interactions tend to reduce the surface area of aggregates in an aqueous solution.

Overall, the elucidation of aggregation rules and the ability to predict and control aggregation behavior (i.e., aggregation degree and morphologies) opens up new possibilities for the design and development of peptide-based materials and therapeutics.

## 4. METHODS

### 4.1. CGMD Simulations of Peptide Aggregation

The CGMD simulations are carried out for generating the training data of AP of pentapeptides using the open-source package GROMACS<sup>31</sup> and the version 2.2 of the Martini force field.<sup>24</sup> The AP value is defined as the ratio of solvent-accessible surface area (SASA) at the beginning ( $SASA_{\text{initial}}$ ) and end ( $SASA_{\text{final}}$ ) of a CGMD simulation<sup>23</sup>

$$AP = \frac{SASA_{\text{initial}}}{SASA_{\text{final}}}$$

For aggregating peptides, the SASA will gradually decrease as the simulation proceeds, and the AP value will increase over 1, while for non-aggregating peptides, the AP value will remain as a value close to 1.

Before performing simulations, 150 coarse-grained pentapeptides are solvated randomly in a 15 nm  $\times$  15 nm  $\times$  15 nm cubic box with 28,400 water beads (water density  $\approx$  1 g cm<sup>-3</sup>), resulting in a solute concentration of  $\sim$ 74 mmol/L. The simulation box is then energy-minimized and subsequently run with an NPT ensemble and a time step of 25 fs for  $5 \times 10^6$  steps, corresponding to a total simulation time of 125 ns ("effective time" of 600 ns due to the simulation acceleration from the coarse-graining of four atoms into one bead). The Berendsen algorithm is utilized for controlling the temperature and pressure at approximately 300 K and 1 bar, respectively.

Within the simulation time of 125 ns, the AP values are close to convergence (Figure S1 and Supporting Information-1). However, it should be noted that the simulation for AP generation within 125 ns does not strictly reach equilibrium. Consequently, should additional simulation time be granted, a potentially larger AP value might emerge. We have rigorously examined the impact of simulation duration on AP<sup>20,21</sup> and have found no instances where peptides initially identified as aggregating were later deemed non-aggregating, or the reverse. This leads us to conclude that a 125 ns simulation duration is a judicious choice, yielding reliable AP values while optimizing the computational efficiency.

For the investigation of the aggregate propensities of the longer peptides KLVFFAE, DDKLVFFAE, and DDDKLVFFAE, the number of solvated peptides in the simulation box is reduced to 100 instead of 150 in order to accelerate the simulations, corresponding to a solute concentration of  $\sim$ 49 mmol/L. After energy minimization, the

simulation box is run with an *NVT* ensemble for 125 ns and an *NPT* ensemble for 125 ns and another 1250 ns equilibration run for obtaining stable morphologies (equivalently 5000 ns due to coarse-graining smoothing).

#### 4.2. TRN Model

The transformer-based regression network (TRN) model applied in this research is developed in our previous research<sup>21</sup> and can be accessed from the github: [https://github.com/Zihan-Liu-00/DL\\_for\\_Peptide](https://github.com/Zihan-Liu-00/DL_for_Peptide). It should be noted that the model used in this research is termed as a "Combo" model, trained with AP values of penta- to decapeptides, without AP values of tetrapeptides.

#### 4.3. Attribution Analysis

Explainability in AI refers to the capacity to understand the decisions or predictions made by AI. Attribution,<sup>32–34</sup> a gradient-based approach for post hoc explanation, is employed to evaluate the importance of each input element to neural network prediction. Integrated gradient<sup>35–37</sup> has been widely recognized as an effective and reliable attribution method. Through back-propagation, the value of the gradient can reflect the activation of the input feature, hidden feature, or neurons in the model during inference.

The AP prediction model, denoted as  $f(X)$ , is trained through the model architecture of TRN, which includes a transformer encoder and a multilayer perceptron (MLP) decoder. The transformer encoder can be further decomposed into input embedding, positional encoding, and the encoder block. Input embedding is a learnable linear mapping function that maps discrete dimensions of amino acids to a high-dimensional continuous embedding space. This process is represented as

$$H^d = \text{Embedding}(X)$$

where  $X$  represents the input peptide sequence and  $d$  represents the dimension of the embedding layer. By mapping to a continuous embedding space, the gradient-based attribution can then be applicable to discrete peptide data. In deep-learning-based attribution methods, another important factor is the loss function, which is used to evaluate the gradients on the embedding layer after back-propagation. For attribution analysis of the AP regression model, we use the loss function

$$L(X) = (f(X) - \text{AP}_{\min})^2$$

$\text{AP}_{\min}$  is the minimum value of the AP values in the training data. The objective of this loss function is to bring down the AP prediction  $f(X)$ , as we intend to find the most important dimension determining the model prediction. We denote the back-propagated gradient in the hidden space  $H^d$  as

$$\text{grad}_{H^d}(X) = \frac{\partial L(X)}{\partial H^d}$$

Gradient can be interpreted as the effect on  $L(X)$  by changes in each dimension of  $H^d$ . A significant gradient value to a dimension indicates that this dimension is highly activated during the forward process, i.e., this dimension is important for the loss  $L(X)$ .

Integrated gradients (IG) gradually change a nonsemantic sample  $\hat{X}$  to a real sample  $X$ , and we calculate the gradient integral during this gradual process and the element product with the corresponding input values as the importance of input dimensions. Since a peptide input  $X$  is a discrete amino acid sequence, we use input embedding  $H^d$  to replace  $X$  during the gradual process. The mathematical expression of this process is<sup>35</sup>

$$\text{IG}_{H^d}(X) = (H^d - \hat{H}^d) \sum_{k=1}^m \frac{\partial L\left(\frac{k}{m}H^d\right)}{\partial H^d}$$

where  $L\left(\frac{k}{m}H^d\right)$  denotes the loss obtained by scaling the representation of  $L(X)$  in the embedding space to the original size

of  $\frac{k}{m}$  and  $m$  is the number of steps in the Riemann approximation of the integral.

For the  $i$ -th amino acid  $X_i$  of input peptide sequence  $X$ , gradient  $\text{IG}_{H^d}(X)$  is presented as a vector of equal dimensionality to  $H^d$ . We sum the absolute values of the elements in vector  $\text{IG}_{H^d}(X)$  and then divide it by the sum of the gradients across the entire sequence as the importance of amino acid  $X_i$  on predicting the AP value in the peptide sequence, expressed as

$$\text{Saliency}(X_i) = \frac{\text{sum}(\text{IG}_{H^d}(X_i))}{\sum_{j=0}^m \text{sum}(\text{IG}_{H^d}(X_j))}$$

#### 4.4. Peptide Synthesis and TEM Characterization

Short peptides are synthesized by solid-phase peptide synthesis (SPPS) using 2-chlorotriptyl chloride resin. The Fmoc protection groups are removed by 20% piperidine in anhydrous *N,N'*-dimethylformamide (DMF). Then, the Fmoc-protected amino acids are coupled to the free amino group by using HBTU (*O*-(benzotriazol-1-yl)-*N,N,N',N'*-tetramethyluronium hexafluorophosphate) as the coupling reagent. The peptides are cleaved from the resin using a cleavage reagent including 95% of trifluoroacetic acid (TFA), 2.5% of triisopropylsilane (TIS), and 2.5% of water for 1 h at room temperature. After TFA is removed using a rotary evaporator, the peptides are precipitated using cold ether, and the crude peptides are further purified by reversed-phase high-performance liquid chromatography.

The negative staining technique is used to observe the morphologies formed by peptides. A micropipette is used to load 10  $\mu\text{L}$  of sample solution to a carbon-coated copper grid, and filter paper is used to remove the excess solution. After rinsing the grid with deionized water, we use uranyl acetate to stain the sample for 1 min and then rinse the grid with deionized water again. The excess liquid is drained with filter paper, which is conducted on a Talos L120C system operating at 120 kV.

#### 4.5. Calculation of RMOI

For the calculation of RMOI, we solvate 300 peptides in one cubic simulation box with a length of 15 nm. For each type of peptide, a simulation is run for 1250 ns under an *NPT* ensemble ( $P = 1$  bar and  $T = 300$  K). It should be noted that 1250 ns is designed to capture the initial morphologies that are critical for subsequent assembly steps, despite being limited to microseconds. The selected time scale has also been validated by other researchers in the field,<sup>17</sup> and reasonable consistency between the experimental and computational morphologies have been achieved.<sup>17</sup> Consequently, we posit that a simulation time of 1250 ns (equivalently 5000 ns considering the coarse-graining effect) is a feasible and practical duration setting for generating reliable morphologies that serve as a foundation for further assembly studies.

For assessing the standard deviation, we run 5–10 parallel simulations on 800 peptides. We then determined the largest cluster of molecules and then aligned the cluster along its principal axes to determine its moments of inertia. The RMOI is calculated as the ratio of the moments of inertia ( $\text{RMOI} = L_x/L_z$ ) along the principal axes of the largest cluster in the system. A RMOI value close to 1 indicates a ball-like structure, while a value of 0 indicates a fibrous structure. Other than ball-like or fibrous structure, there could be other morphologies such as vesicle or net structure, and examining those structures would require visual inspection.

### ■ ASSOCIATED CONTENT

#### Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/jacsau.4c00501>.

Supporting Information-1: Supporting Information for the main text (PDF)

Supporting Information-2: AP distributions of tetra- (Figures S1-S20) and pentapeptides (Figures S21-S40) with different content of specific amino acids (PDF)

Supporting Information-3: Comparison of TEM images and computational morphologies of 165 aggregating tetrapeptides (PDF)

## AUTHOR INFORMATION

### Corresponding Authors

**Tengyan Xu** – Zhejiang Key Laboratory of Intelligent Cancer Biomarker Discovery and Translation, First Affiliated Hospital, Wenzhou Medical University, Wenzhou 325035, China; Email: [xutengyan@wmu.edu.cn](mailto:xutengyan@wmu.edu.cn)

**Stan Z. Li** – School of Engineering and AI Lab, Research Center for Industries of the Future, Westlake University, Hangzhou, Zhejiang 310030, China; Email: [Stan.ZQ.Li@westlake.edu.cn](mailto:Stan.ZQ.Li@westlake.edu.cn)

**Wenbin Li** – Research Center for Industries of the Future and School of Engineering, Westlake University, Hangzhou, Zhejiang 310030, China; [orcid.org/0000-0002-1240-2707](https://orcid.org/0000-0002-1240-2707); Email: [liwenbin@westlake.edu.cn](mailto:liwenbin@westlake.edu.cn)

### Authors

**Jiaqi Wang** – Research Center for Industries of the Future and School of Engineering, Westlake University, Hangzhou, Zhejiang 310030, China; Wisdom Lake Academy of Pharmacy and Jiangsu Province Higher Education Key Laboratory of Cell Therapy Nanoformulation, Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu 215123, China

**Zihan Liu** – School of Engineering and AI Lab, Research Center for Industries of the Future, Westlake University, Hangzhou, Zhejiang 310030, China; [orcid.org/0000-0001-6224-3823](https://orcid.org/0000-0001-6224-3823)

**Shuang Zhao** – School of Engineering, Westlake University, Hangzhou, Zhejiang 310030, China; State Key Laboratory of Precision Measurement Technology and Instruments, Department of Precision Instrument, Tsinghua University, Beijing 100084, China

**Yu Zhang** – Zhejiang Key Laboratory of Intelligent Cancer Biomarker Discovery and Translation, First Affiliated Hospital, Wenzhou Medical University, Wenzhou 325035, China

Complete contact information is available at: <https://pubs.acs.org/10.1021/jacsau.4c00501>

### Author Contributions

<sup>†</sup>J.W., Z.L., and S.Z. are co-first authors. CRediT: **Jiaqi Wang** conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, writing-original draft; **Zihan Liu** conceptualization, data curation, formal analysis, investigation, methodology, writing-original draft; **Shuang Zhao** conceptualization, data curation, formal analysis, methodology, writing-original draft; **Yu Zhang** data curation, investigation; **Tengyan Xu** conceptualization, data curation, methodology, writing-original draft; **Stan Z. Li** conceptualization, investigation, project administration, resources, supervision; **Wenbin Li** conceptualization, formal analysis, project administration, resources, software, supervision, validation, writing-original draft, writing-review & editing.

## Notes

The authors declare no competing financial interest.

## ACKNOWLEDGMENTS

The authors acknowledged the HPC Center of Xi'an Jiaotong-Liverpool University, HPC Center of Westlake University, and Beijing PARATERA Tech CO., Ltd. for computational support. J.W., S.Z., and W.L. are supported by the Research Center for Industries of the Future at Westlake University under award no. WU2022C041. J.W. also acknowledges the support of the National Natural Science Foundation of China (no. 52101023). Z.L. and S.Z.L. are supported by the Ministry of Science and Technology of the People's Republic of China (no. 2021YFA1301603) and the National Natural Science Foundation of China (no. U21A20427).

## REFERENCES

- (1) Zapadka, K. L.; Becher, F. J.; Gomes dos Santos, A.; Jackson, S. E. Factors affecting the physical stability (aggregation) of peptide therapeutics. *Interface Focus* **2017**, *7* (6), 20170030.
- (2) Nguyen, P. H.; Sterpone, F.; Derreumaux, P. Aggregation of disease-related peptides. *Prog. Mol. Biol. Transl. Sci.* **2020**, *170*, 435–460.
- (3) Tao, K.; Makam, P.; Aizen, R.; Gazit, E. Self-assembling peptide semiconductors. *Science* **2017**, *358* (6365), No. eaam9756.
- (4) Nguyen, T. P.; Easley, A. D.; Kang, N.; Khan, S.; Lim, S.-M.; Rezenom, Y. H.; Wang, S.; Tran, D. K.; Fan, J.; Letteri, R. A.; et al. Polypeptide organic radical batteries. *Nature* **2021**, *593* (7857), 61–66.
- (5) Wen, Q.; Zhang, Y.; Li, C.; Ling, S.; Yang, X.; Chen, G.; Yang, Y.; Wang, Q. NIR-II fluorescent self-assembled peptide nanochain for ultrasensitive detection of peritoneal metastasis. *Angew. Chem.* **2019**, *131* (32), 11117–11122.
- (6) Cheng, Z.; Kuru, E.; Sachdeva, A.; Vendrell, M. Fluorescent amino acids as versatile building blocks for chemical biology. *Nat. Rev. Chem.* **2020**, *4* (6), 275–290.
- (7) Muttenthaler, M.; King, G. F.; Adams, D. J.; Alewood, P. F. Trends in peptide drug discovery. *Nat. Rev. Drug Discovery* **2021**, *20* (4), 309–325.
- (8) Cooper, B. M.; Iegre, J.; O'Donovan, D. H.; Ölwegård Halvarsson, M.; Spring, D. R. Peptides as a platform for targeted therapeutics for cancer: Peptide–drug conjugates (PDCs). *Chem. Soc. Rev.* **2021**, *50* (3), 1480–1494.
- (9) Ashaolu, T. J. Antioxidative peptides derived from plants for human nutrition: their production, mechanisms and applications. *Eur. Food Res. Technol.* **2020**, *246*, 853–865.
- (10) Löhr, T.; Kohlhoff, K.; Heller, G. T.; Camilloni, C.; Vendruscolo, M. A kinetic ensemble of the Alzheimer's A $\beta$  peptide. *Nat. Comput. Sci.* **2021**, *1* (1), 71–78.
- (11) Knowles, T. P.; Vendruscolo, M.; Dobson, C. M. The amyloid state and its association with protein misfolding diseases. *Nat. Rev. Mol. Cell Biol.* **2014**, *15* (6), 384–396.
- (12) Pal, S.; Paul, S. ATP Controls the Aggregation of A $\beta$ <sub>16–22</sub> Peptides. *J. Phys. Chem. B* **2020**, *124* (1), 210–223.
- (13) Li, T.; Lu, X.-M.; Zhang, M.-R.; Hu, K.; Li, Z. Peptide-based nanomaterials: Self-assembly, properties and applications. *Bioact. Mater.* **2022**, *11*, 268–282.
- (14) 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* (45), 39174–39184.
- (15) Sun, B.; Tao, K.; Jia, Y.; Yan, X.; Zou, Q.; Gazit, E.; Li, J. Photoactive properties of supramolecular assembled short peptides. *Chem. Soc. Rev.* **2019**, *48* (16), 4387–4400.
- (16) Helen, W.; De Leonardis, P.; Ulijn, R. V.; Gough, J.; Tirelli, N. Mechanosensitive peptide gelation: mode of agitation controls

mechanical properties and nano-scale morphology. *Soft Matter* **2011**, *7* (5), 1732–1740.

(17) 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 (tri-) peptide self-assembly to design and discover new hydrogels. *Nat. Chem.* **2015**, *7* (1), 30–37.

(18) van Teijlingen, A.; Tuttle, T. Beyond tripeptides two-step active machine learning for very large data sets. *J. Chem. Theory Comput.* **2021**, *17* (5), 3221–3232.

(19) Batra, R.; Loeffler, T. D.; Chan, H.; Srinivasan, S.; Cui, H.; Korendovych, I. V.; Nanda, V.; Palmer, L. C.; Solomon, L. A.; Fry, H. C.; et al. Machine learning overcomes human bias in the discovery of self-assembling peptides. *Nat. Chem.* **2022**, *14* (12), 1427–1435.

(20) Wang, J.; Liu, Z.; Zhao, S.; Xu, T.; Wang, H.; Li, S. Z.; Li, W. Deep Learning Empowers the Discovery of Self-Assembling Peptides with Over 10 Trillion Sequences. *Adv. Sci.* **2023**, *10*, 2301544.

(21) Liu, Z.; Wang, J.; Luo, Y.; Zhao, S.; Li, W.; Li, S. Z. Efficient prediction of peptide self-assembly through sequential and graphical encoding. *Briefings Bioinf.* **2023**, *24* (6), bbad409.

(22) Chicco, D.; Warrens, M. J.; Jurman, G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Comput. Sci.* **2021**, *7*, No. e623.

(23) Xu, T.; Wang, J.; Zhao, S.; Chen, D.; Zhang, H.; Fang, Y.; Kong, N.; Zhou, Z.; Li, W.; Wang, H. Accelerating the prediction and discovery of peptide hydrogels with human-in-the-loop. *Nat. Commun.* **2023**, *14* (1), 3880.

(24) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; De Vries, A. H. The MARTINI force field: coarse grained model for biomolecular simulations. *J. Phys. Chem. B* **2007**, *111* (27), 7812–7824.

(25) Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S.-J. The MARTINI coarse-grained force field: extension to proteins. *J. Chem. Theory Comput.* **2008**, *4* (5), 819–834.

(26) Ross, C. A.; Poirier, M. A. Protein aggregation and neurodegenerative disease. *Nat. Med.* **2004**, *10* (S7), S10–S17.

(27) Yan, C.; Pochan, D. J. Rheological properties of peptide-based hydrogels for biomedical and other applications. *Chem. Soc. Rev.* **2010**, *39* (9), 3528–3540.

(28) Sivagnanam, S.; Arul, A.; Ghosh, S.; Dey, A.; Ghorai, S.; Das, P. Concentration-dependent fabrication of short-peptide-based different self-assembled nanostructures with various morphologies and intracellular delivery property. *Mater. Chem. Front.* **2019**, *3* (10), 2110–2119.

(29) Ke, D.; Zhan, C.; Li, A. D.; Yao, J. Morphological Transformation between Nanofibers and Vesicles in a Controllable Bipyridine–Tripeptide Self-Assembly. *Angew. Chem.* **2011**, *123* (16), 3799–3803.

(30) Adler-Abramovich, L.; Gazit, E. The physical properties of supramolecular peptide assemblies: from building block association to technological applications. *Chem. Soc. Rev.* **2014**, *43* (20), 6881–6893.

(31) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1–2*, 19–25.

(32) Zhang, Y.; Tino, P.; Leonardis, A.; Tang, K. A survey on neural network interpretability. *IEEE Trans. Emerg. Top. Comput. Intell.* **2021**, *5* (5), 726–742.

(33) Ancona, M.; Ceolini, E.; Öztireli, C.; Gross, M. Towards better understanding of gradient-based attribution methods for deep neural networks. arXiv preprint arXiv:1711.06104, **2017**.

(34) Linardatos, P.; Papastefanopoulos, V.; Kotsiantis, S. Explainable ai: A review of machine learning interpretability methods. *Entropy* **2020**, *23* (1), 18.

(35) Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic attribution for deep networks. *International Conference on Machine Learning*; PMLR, 2017; pp 3319–3328.

(36) Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115.

(37) Qi, Z.; Khorram, S.; Li, F. Visualizing Deep Networks by Optimizing with Integrated Gradients. *CVPR Workshops*, 2019; Vol. 2, pp 1–4.

#### NOTE ADDED AFTER ASAP PUBLICATION

This paper was published ASAP on September 3, 2024, with incomplete images in Figure 6. The corrected version was reposted on September 4, 2024.