Supplementary Information

Discovery of β-Sheet Peptide Assembly Codes via an Experimentally Validated Predictive Computational Platform

Hang Zheng,[a,b,1] Ke Huang,[c,1] Chi Sing Lee,[c,*] Wei Han[c,d,*]

[1] These authors contribute equally to this study.

[a] School of Chemical Biology and Biotechnology, Peking University Shenzhen Graduate School, Shenzhen 518055, China

[b] DP Technology, Beijing 100089, China

[c] Department of Chemistry, Faculty of Science, Hong Kong Baptist University, Hong Kong SAR, China

[d] Institute of Chemical Biology, Shenzhen Bay Laboratory, Shenzhen 518132, China

[*]Corresponding author, E-mail: hanw_chem@hkbu.edu.hk (W. H.); cslee-chem@hkbu.edu.hk (C.S. L.)

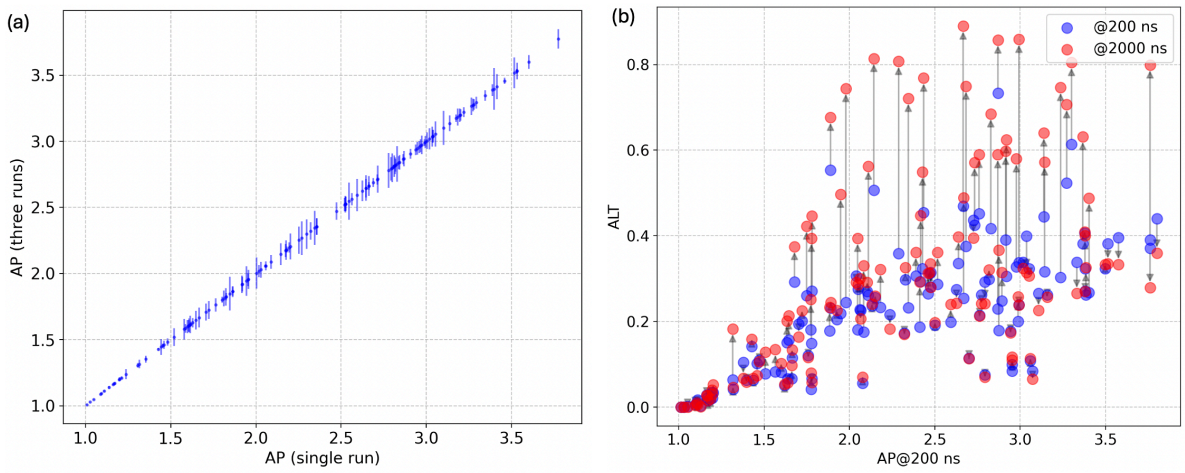


Figure S1. Validation of the two-step simulation protocol. (a) Comparison of Aggregation Propensity (AP) values derived from a single run versus the average of three independent runs. Each data point represents one of the 130 selected tripeptides. Error bars indicate the standard error of the mean (SEM) across the three runs. (b) Assessment of Amyloid-like Tendency (ALT) stability over extended simulations. Blue dots represent ALT values at the initial simulation time (t=200 ns), while red dots represent values after extended simulation (t=2000 ns). Both are plotted against the corresponding AP at t = 200 ns. Gray arrows illustrate the shift in ALT for each tripeptide over the extended timeframe.

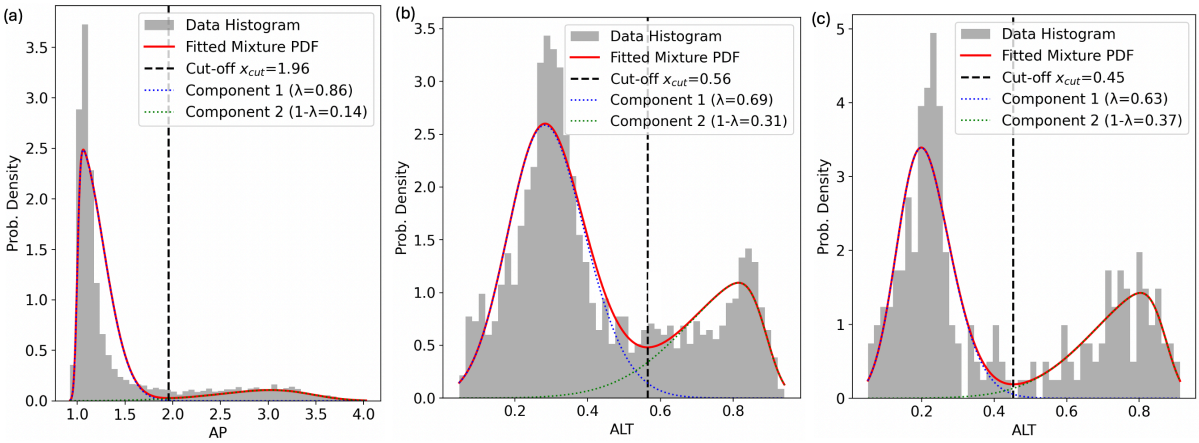


Figure S2. Determination of cutoff thresholds for peptide scalar properties. Histograms (gray bars) display the distributions of (a) tripeptide AP values, (b) tripeptide ALT values, and (c) tetrapeptide ALT values. Red curves represent the bimodal distributions fitted to the distributions according to $P(x) = \lambda f_1(x) + (1-\lambda) f_2(x)$, f_1 and f_2 are two independent skew-normal distributions and λ is the mixing coefficient. Vertical dashed lines indicate the cutoff thresholds (x_{cut}), defined as the point where $dP/dx = 0$ and $d^2P/dx^2 > 0$.

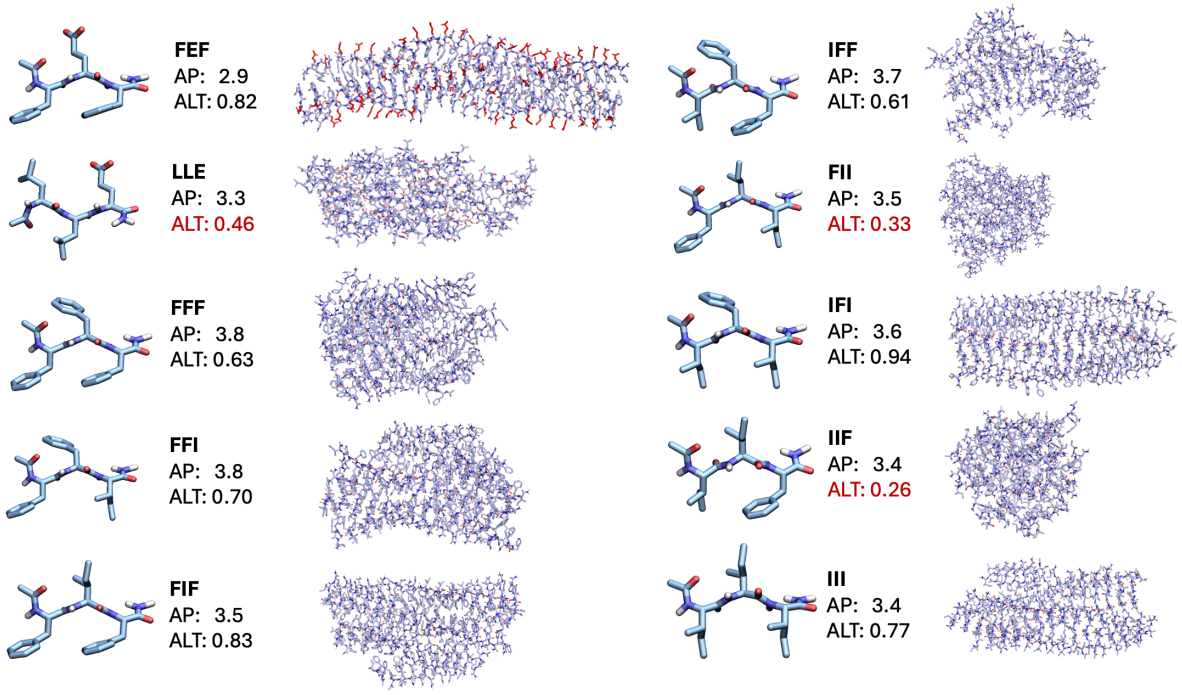


Figure S3. Simulation snapshots of amyloid-forming tripeptides. Representative images of ten end-capped tripeptides known for amyloid-like assembly. Left: Stick representation of the monomeric peptide. Right: Final snapshot of the simulation trajectory showing the terminal configuration.

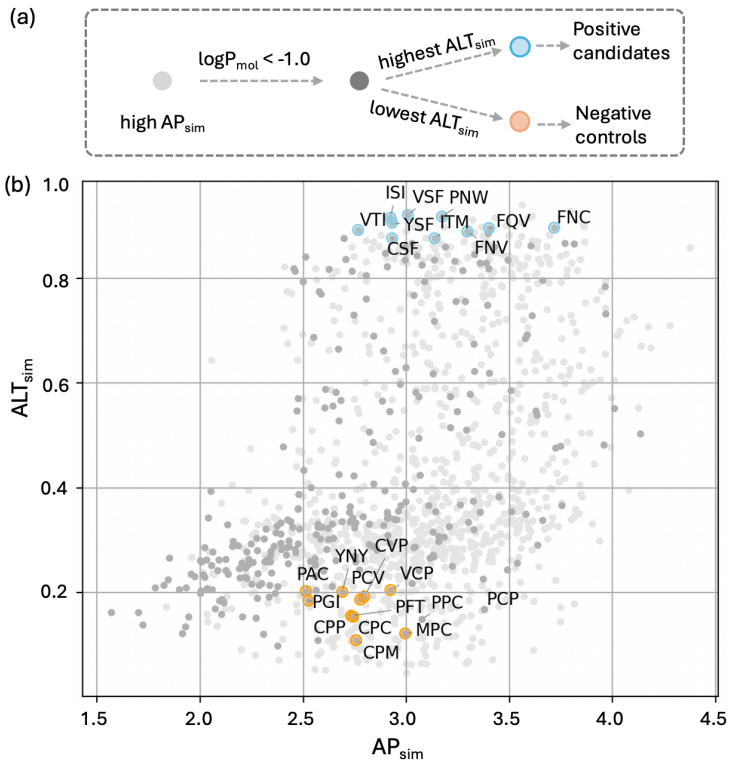


Figure S4. Tripeptide selection for experimental validation. (a) Schematic workflow of the tripeptide selection process based on simulated AP and ALT descriptors. (b) Distribution of selected tripeptides. Each data point represents a peptide system, with the simulated ALT value plotted against the simulated AP value. Points are colored light gray for peptides with $\log P_{mol} > -1.0$ and dark gray otherwise. Peptides selected as positive cases and negative controls for experimental validation are highlighted with light blue and sandy brown circles, respectively.

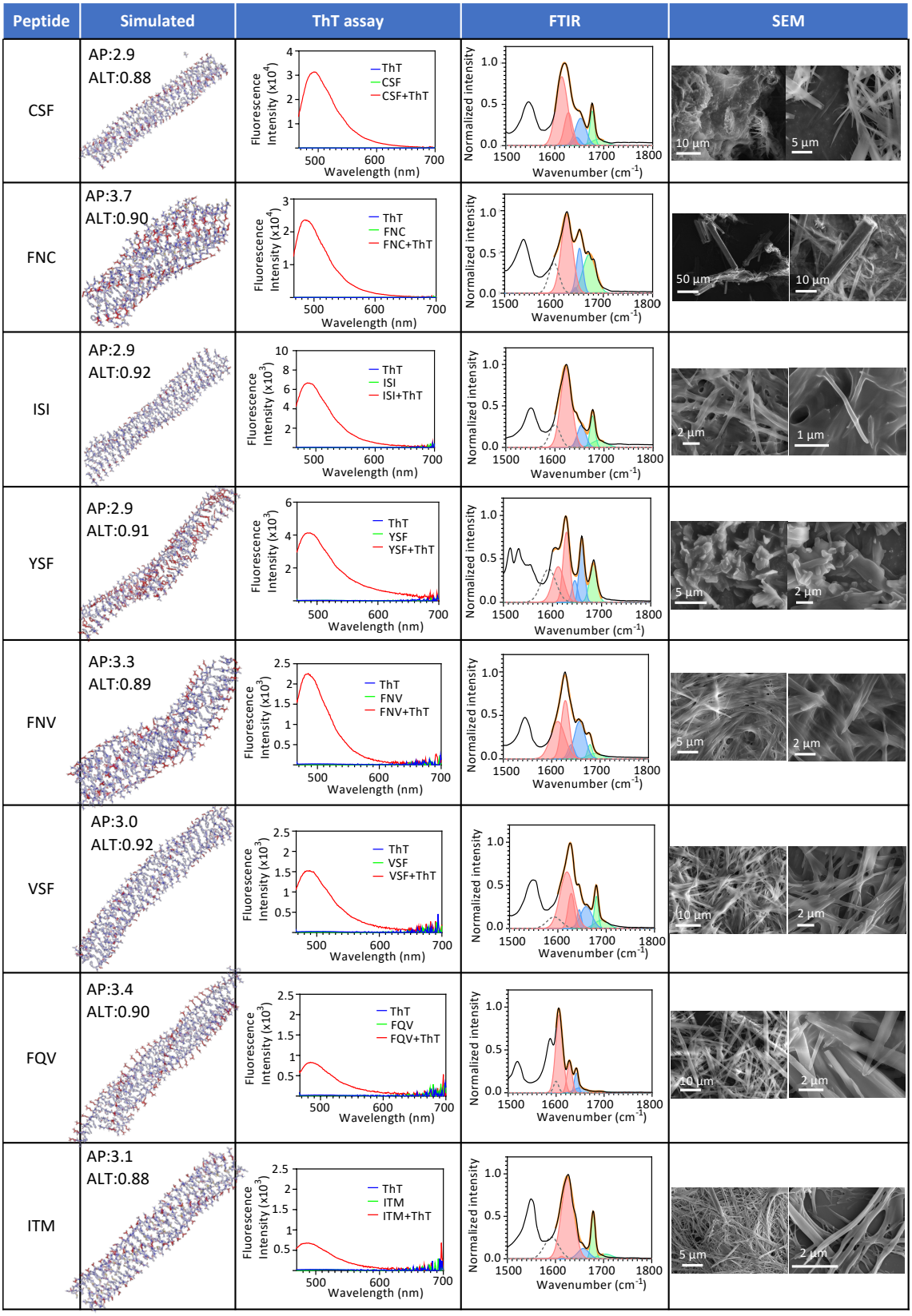
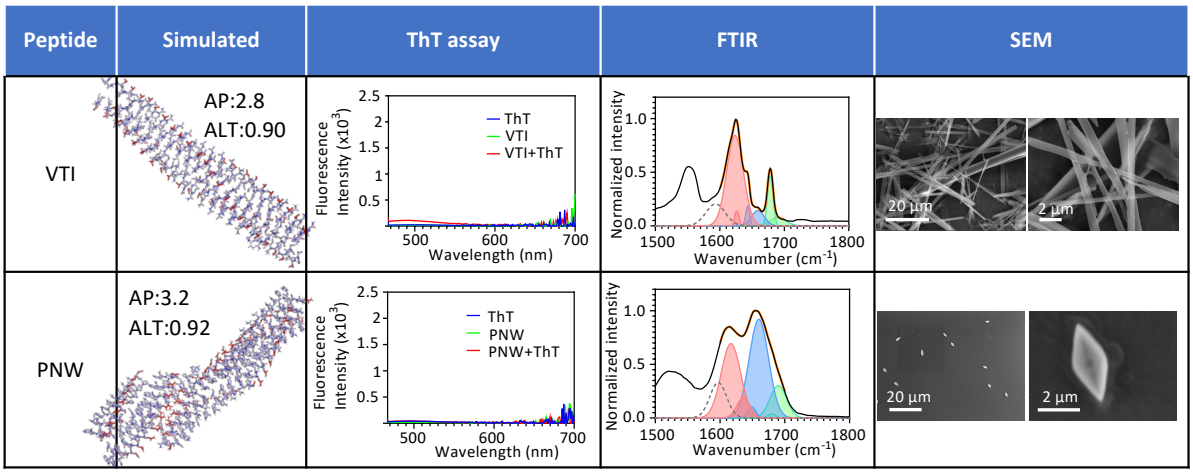


Figure S5. Grid display of simulation and experimental validation results for predicted amyloid-forming tripeptides. Column 1: Peptide sequence. Column 2: Final snapshot from simulations. Column 3: ThT florescence spectra. Column 4: FTIR spectra characterizing β-sheet contents. Column 5: SEM micrographs showing morphologies of assemblies.

Figure S5 (continued)



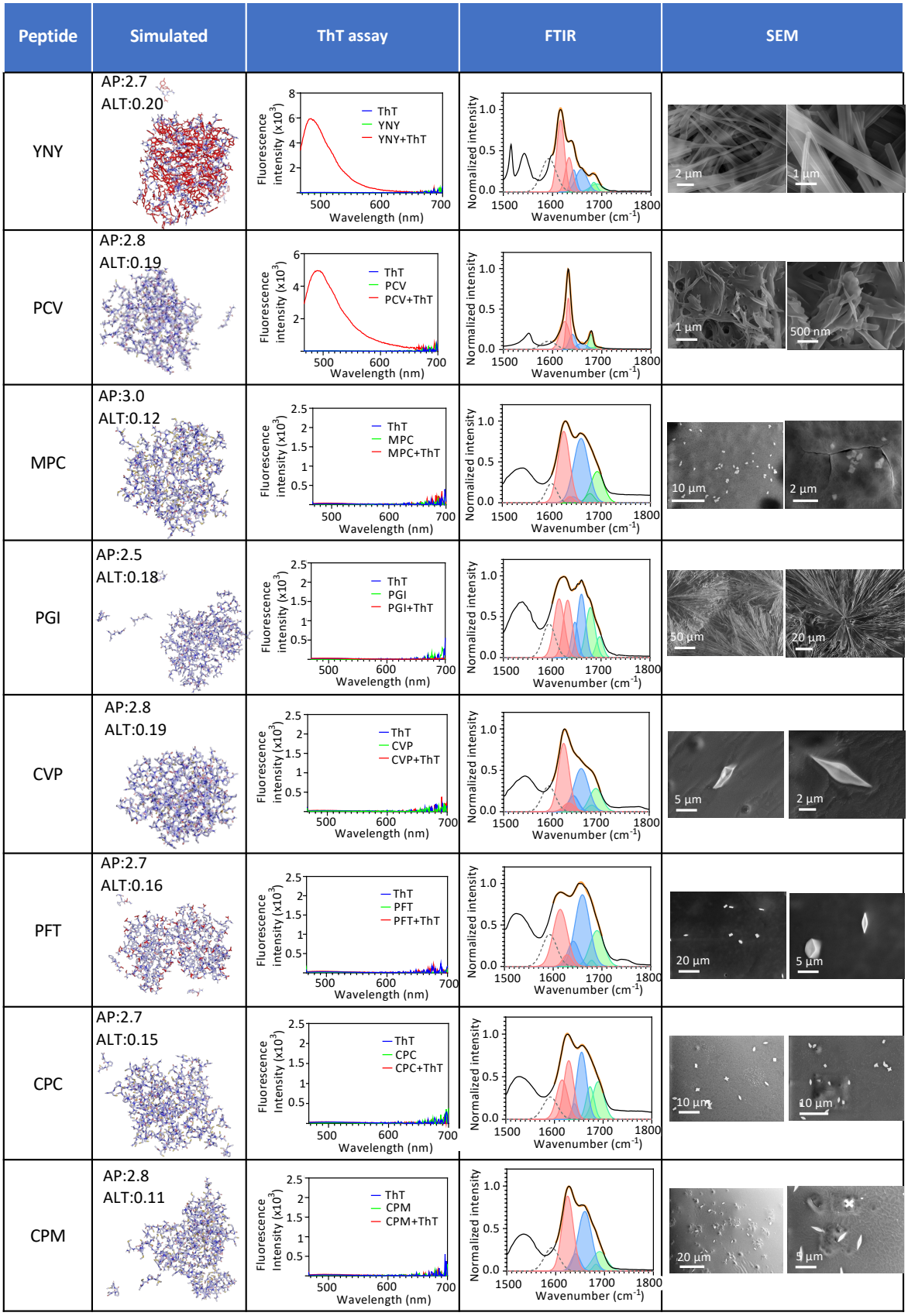
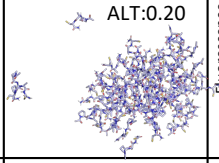
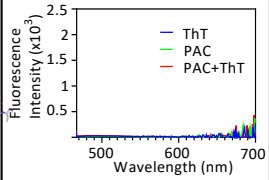
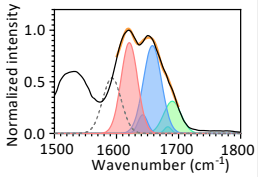
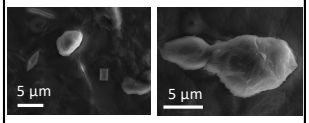
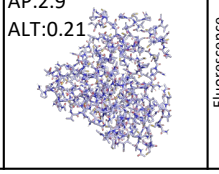
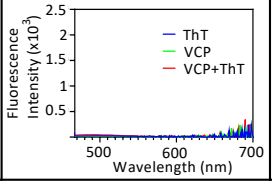
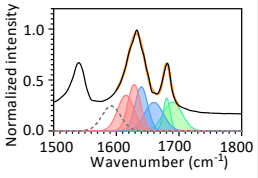
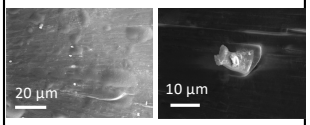


Figure S6. Grid display of simulation and experimental validation results for predicted tripeptides forming non-amyloid structures.

Figure S6 (continued)

Peptide	Simulated	ThT assay	FTIR	SEM
PAC				
VCP				

(a)

	α=0.5	**α=0.7**	α=0.9
MSE	0.0049	**0.0048**	0.0056
R^2	0.8623	**0.8636**	0.8424
AUPRC	0.74	**0.77**	0.71

(b)

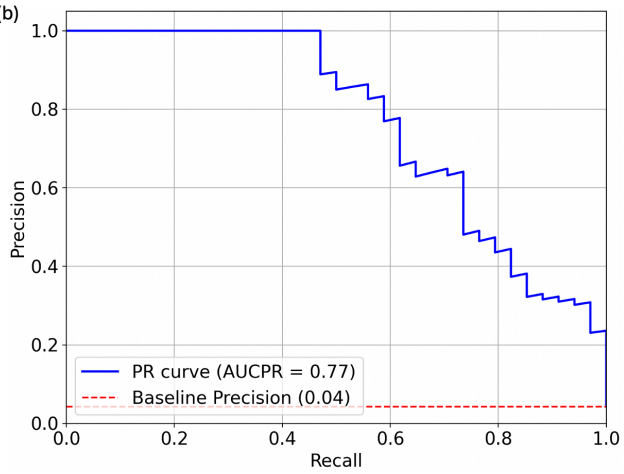


Figure S7. Performance assessment of the fine-tuned Uni-Mol model for ALT prediction. Evaluation was performed on a test set of 800 tripeptides. (a) Key performance metrics across different values of the hyperparameter α. The optimal model (α=0.7) is highlighted in bold. (b) Precision-recall curve for the optimized model (α=0.7), demonstrating its predictive robustness.

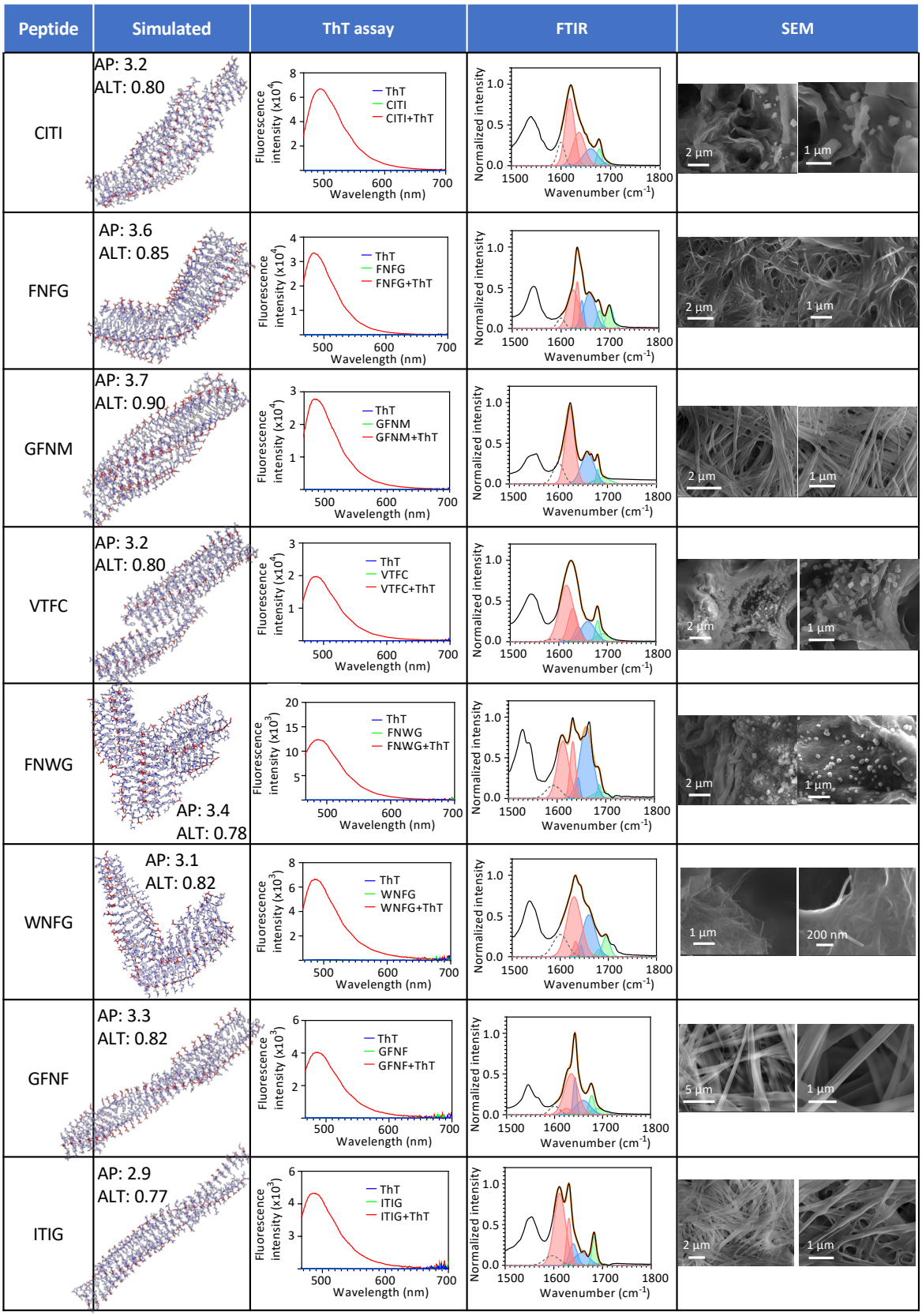
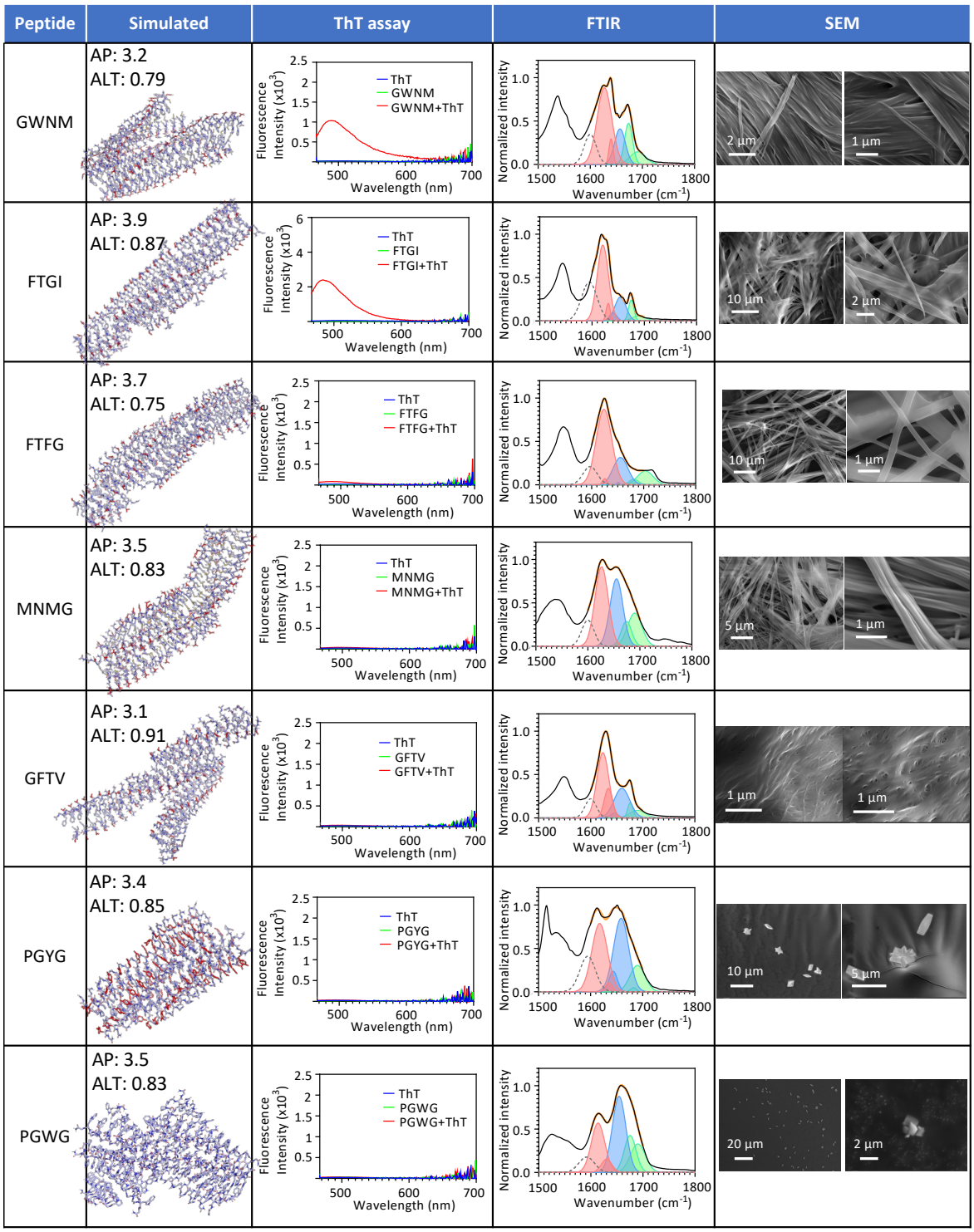


Figure S8. Grid display of simulation and experimental validation results for predicted amyloid-forming tetrapeptides.

Figure S8 (continued)



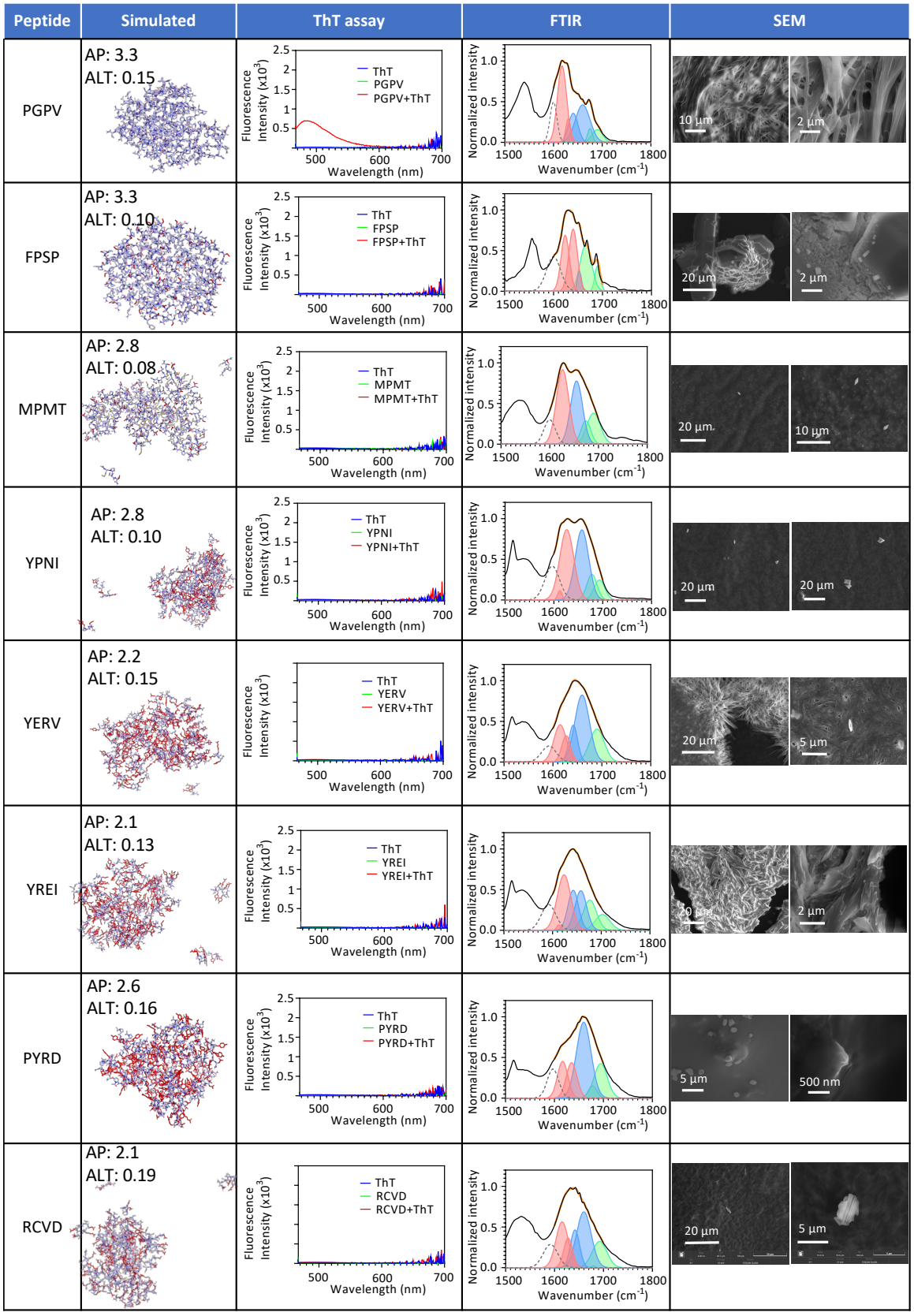
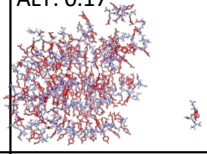
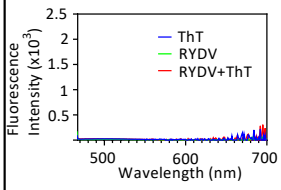
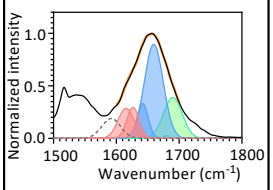
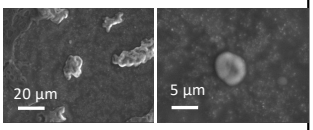
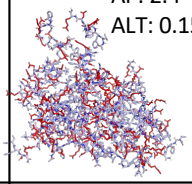
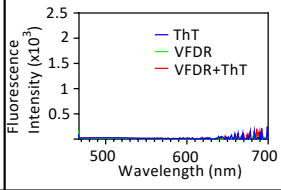
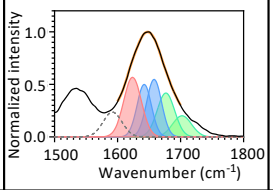
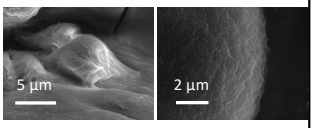
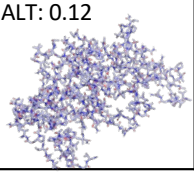
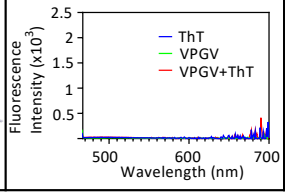
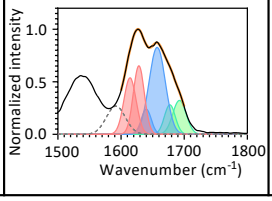
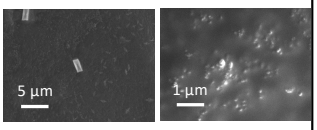


Figure S9. Grid display of simulation and experimental validation results for predicted tetrapeptides forming non-amyloid structures.

Figure S9 (continued)

Peptide	Simulated	ThT assay	FTIR	SEM
RYDV	AP: 2.4 ALT: 0.17 			
VFDR	AP: 2.4 ALT: 0.15 			
VPGV	AP: 3.3 ALT: 0.12 			

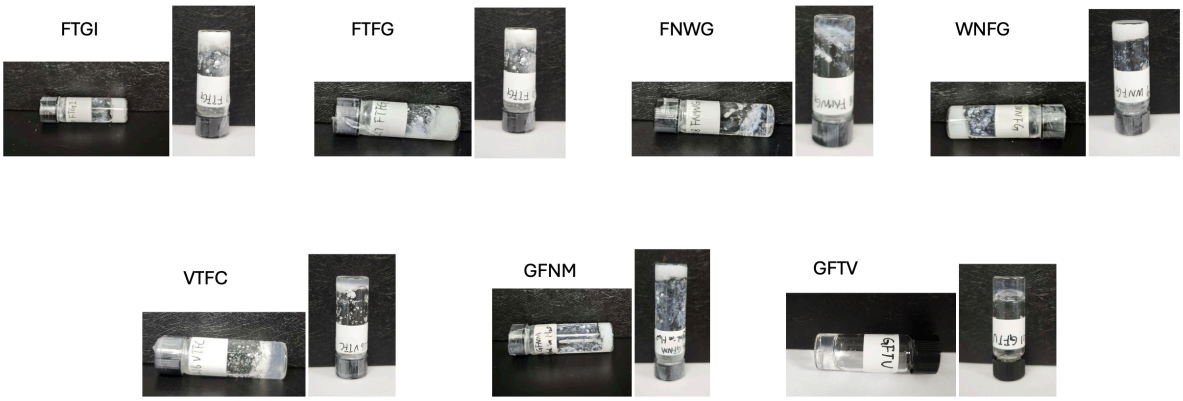


Figure S10. Photographs of tetrapeptide hydrogels discovered in this study.

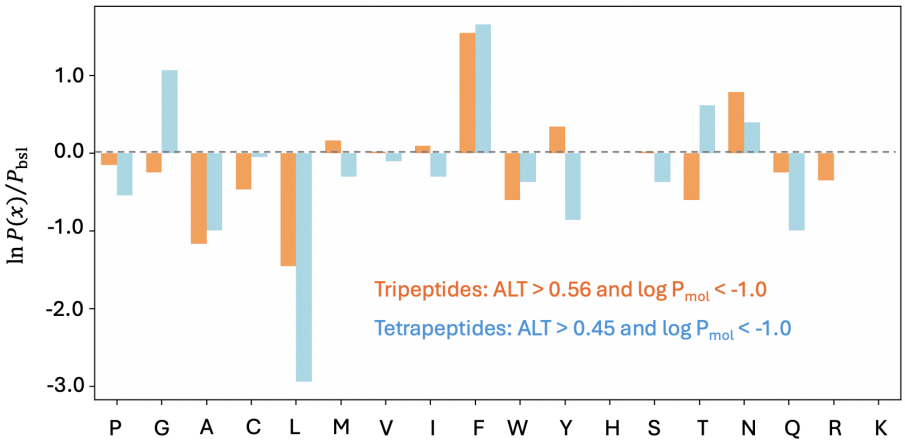


Figure S11. Amino acid preference in amyloid-forming peptides. Frequency of individual amino acids in tripeptide and tetrapeptide sequences are shown as sandy brown and light blue bars, respectively. Analysis was restricted to peptides with $\log P_{mol} < -1.0$. Peptides were classified as amyloid-forming based on simulated ALT threshold of > 0.56 for tripeptides and > 0.45 for tetrapeptides.

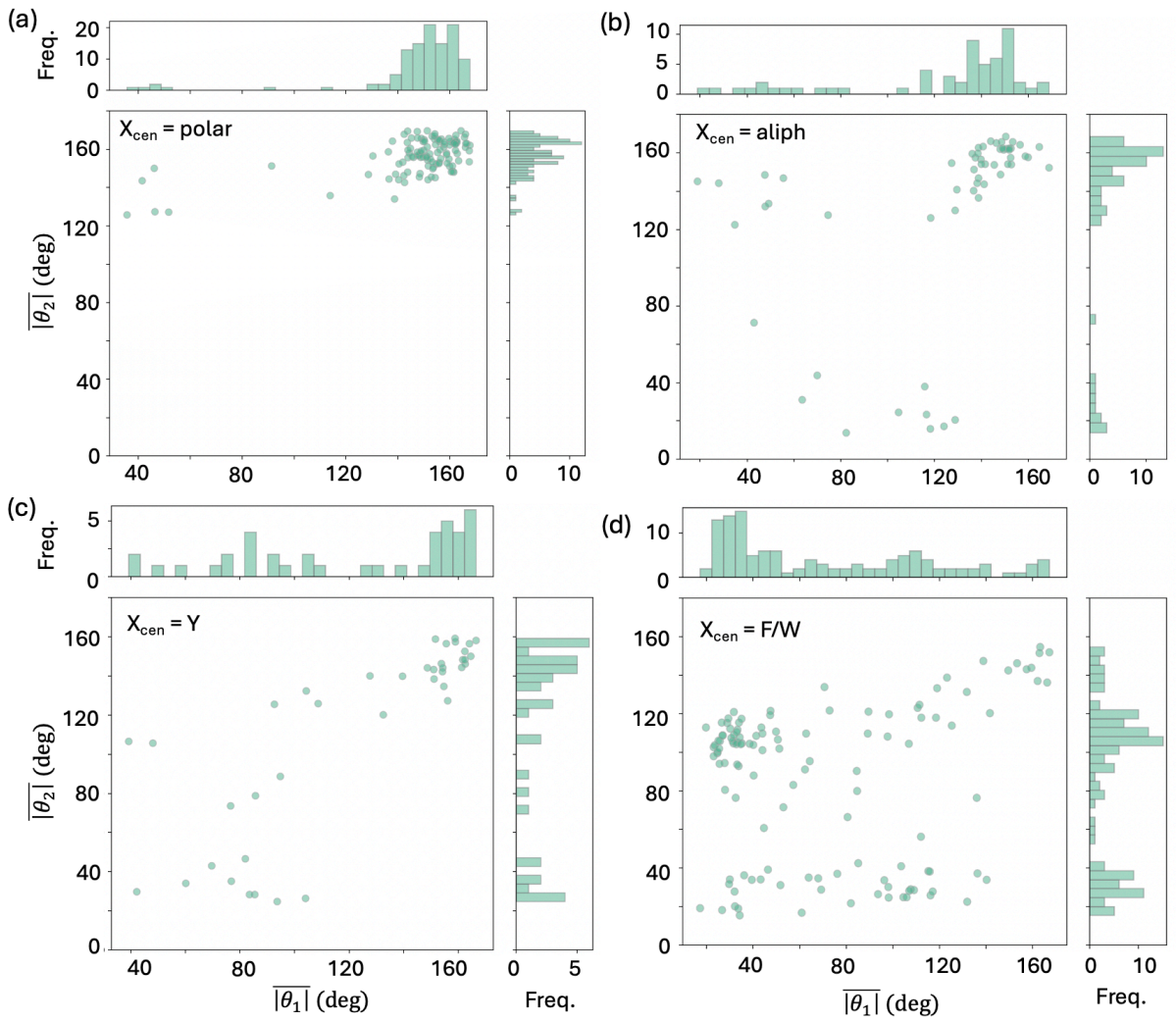


Figure S12. The (θ_1, θ_2) Plots for amyloid-forming tripeptide sequences with their central positions occupied by polar (a), aliphatic (b), tyrosine (c), or phenylalanine/tryptophan (d) residues. Histograms of marginal distributions of two θ angles are also provided.

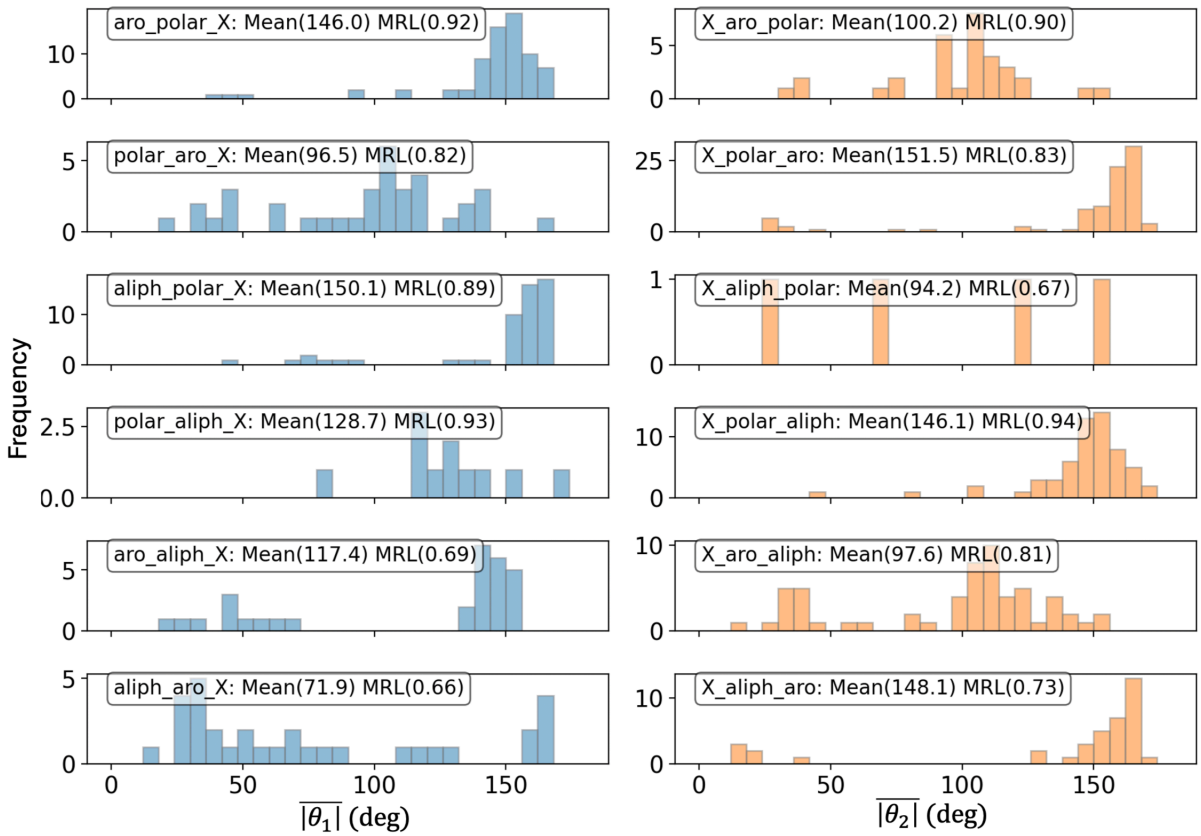


Figure S13. Supplemental distributions of θ angle between neighboring residues for various amino acid combinations identified in amyloid-forming sequences.