

## VIP Very Important Paper

# Chemical Evolution of Early Macromolecules: From Prebiotic Oligopeptides to Self-Organizing Biosystems via Amyloid Formation

Fruzsina Bencs,<sup>[a, b]</sup> Nóra Taricska,<sup>[b, c]</sup> Zsolt Dürvanger,<sup>[b, c]</sup> Dániel Horváth,<sup>[b, c]</sup>  
Zsolt Fazekas,<sup>[a, b]</sup> Vince Grolmusz,<sup>[d]</sup> Viktor Farkas,<sup>[b, c]</sup> and András Perczel\*<sup>[b, c]</sup>

Short amyloidogenic oligopeptides (APRs) are proposed as early macromolecules capable of forming solvent-separated nanosystems under prebiotic conditions. This study provides experimental evidence that APRs, such as the aggregation-prone oligopeptide A (APR-A), can undergo mutational transitions to form distinct variants and convert to APR-B, either amyloid-like or water-soluble and non-aggregating. These transitions occur along a spectrum from strongly amyloidogenic (pro-amyloid) to weakly amyloidogenic (anti-amyloid), with the mutation sequence order

playing a key role in determining their physicochemical properties. The pro-amyloid pathway facilitates heterogeneous phase separation, leading to amyloid-crystal formation with multiple polymorphs, including the first class 3 amyloid topology. By mapping these transitions, we demonstrate the potential co-evolution of water-soluble miniproteins and insoluble amyloids, both of which could have been pivotal in early bio-nanosystem formation. These insights into amyloid modulation provide a crucial step toward understanding amyloid control mechanisms.

## 1. Introduction

Limited knowledge exists about the prebiotic chemical evolution that occurred around 4 billion years ago, before the emergence

of DNA, RNA, proteins, and ribosomes. Miller and Urey's pioneering experiments demonstrated that complex organic molecules, including amino acids, could form under early Earth conditions.<sup>[1]</sup> Further studies suggest that amino acids were present in the early universe nearly 9 billion years before life originated on Earth.<sup>[2]</sup> Leman et al. showed that carbonyl sulfide, a simple volcanic gas, can catalyze oligopeptide formation from amino acids via Leuchs' anhydride under mild aqueous conditions.<sup>[3]</sup> Exposure of  $\alpha$ -amino acids to carbonyl sulfide (COS) yields oligopeptides in minutes to hours at room temperature, with efficiencies up to 80%, depending on reaction conditions and catalysts (e.g.,  $\text{PdCl}_2$ ). These peptides could have formed at various stages of Earth's chemical evolution, as evidenced by the discovery of dipeptides in meteorites and asteroids.<sup>[4,5]</sup>

These shorter oligopeptides in the aqueous medium were unlikely to fold, but above a critical concentration, they would have interacted with each other. They could have aggregated via H-bonding through their free  $\beta$ -edges, concealing hydrophobic residues from water, thus forming more regular amyloids. Hexapeptides were chosen to model amyloid formation because they are short enough to evolve a variety of primary sequences, yet long enough to self-assemble into structured amyloids. Although not all of today's 20 proteinogenic amino acids were present 4 billion years ago, we focus on the 20 known today for clarity. From these 20 residues, 64 million ( $20^6$ ) hexapeptides were generated, with about 1500 having amyloid-forming data in the Waltz database.<sup>[6]</sup> Approximately 130 3D structures of hexapeptide amyloids have been determined by crystallography.<sup>[7]</sup> For simplicity, we selected hexapeptides from  $\alpha$ -(L)-amino acids with neutral side chains, minimizing pH sensitivity. Our aim was to show how amyloidogenic sequences could have cross-evolved by point mutations, supporting the hypothesis that amyloid nanostructures may have acted as

[a] F. Bencs, Z. Fazekas  
ELTE Hevesy György PhD School of Chemistry, ELTE Eötvös Loránd University,  
Pázmány Péter sétány 1/A, Budapest H-1117, Hungary


[b] F. Bencs, N. Taricska, Z. Dürvanger, D. Horváth, Z. Fazekas, V. Farkas,  
A. Perczel  
Laboratory of Structural Chemistry and Biology, Institute of Chemistry, ELTE  
Eötvös Loránd University, Pázmány Péter sétány 1/A, Budapest H-1117,  
Hungary  
E-mail: [perczel.andras@ttk.elte.hu](mailto:perczel.andras@ttk.elte.hu)

[c] N. Taricska, Z. Dürvanger, D. Horváth, V. Farkas, A. Perczel  
HUN-REN-ELTE Protein Modeling Research Group, ELTE Eötvös Loránd  
University, Pázmány Péter sétány 1/A, Budapest H-1117, Hungary

[d] V. Grolmusz  
Group of Protein Information Technology, Department of Computer Science,  
ELTE Eötvös Loránd University, Pázmány Péter sétány 1/C, Budapest H-1117,  
Hungary

Abbreviations: APR, aggregation-prone region; AP, antiparallel  $\beta$ -sheet; ANuPP, Aggregation Nucleation Prediction in Peptides and Proteins; B-type CD, typically associated with  $\beta$ -strands,  $\beta$ -sheets; CD spectroscopy, far-ultraviolet electronic circular dichroism (185–260 nm) spectroscopy;  $\epsilon$ , dielectric constant; EtOH, ethanol; GRAVY, grand average of hydropathy; MeCN, acetonitrile; MeOH, methanol; PA, parallel  $\beta$ -sheet; TFA, Trifluoroacetic acid; U-state, unfolded state; U-type CD, CD pattern assigned to the conformational ensemble of unstructured polypeptide backbone.

 Supporting information for this article is available on the WWW under  
<https://doi.org/10.1002/chem.202404669>

 © 2025 The Author(s). Chemistry – A European Journal published by  
Wiley-VCH GmbH. This is an open access article under the terms of the  
[Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use,  
distribution and reproduction in any medium, provided the original work is  
properly cited and is not used for commercial purposes.

structural templates during the chemical evolution of functional polypeptides 4 billion years ago.<sup>[8–10]</sup>

Under near-physiological conditions, globular proteins can adopt an amyloid-like nanostructure, a process in which amyloid-prone regions (APRs) likely undergo spontaneous restructuring.<sup>[11]</sup> APR amyloid architectures generally fall into one of Eisenberg's ten homosteric zipper classes, all of which share regular and repetitive H-bonds connecting adjacent  $\beta$ -strands.<sup>[12,13]</sup> The inherently high stability of  $\beta$ -strands makes amyloids thermodynamically superior to any other secondary structure.<sup>[14–16]</sup> The barrier height separating the folded state(s) from the amyloid state(s) provides the kinetic control<sup>[17]</sup> of the transition between them, temporarily protecting the folded structure from becoming an amyloid. Recall C. Dobson's famous quote: "globular proteins are misfolded amyloids." The water-hidden sequence bits that form the core of a globular protein folded in water may also be those that drive misfolding and amyloid formation. Additionally, APRs are short sequential elements responsible for forming protein homooligomers, binding to receptors, and interacting with other proteins. Thus, the intrinsic entanglement of protein folding, function, and amyloid formation—perhaps strongly controlled by APRs—is remarkable.<sup>[18,19]</sup> Gatekeeper residues, such as Pro, prevent amyloid formation by disrupting  $\beta$ -sheets,<sup>[20]</sup> while Lys, Arg, Asp, and Glu can either disrupt (charged) or promote (uncharged) amyloids depending on the charge state controlled by pH.<sup>[21–23]</sup> The free N- and C-terminals of APRs also act as gatekeepers, mimicking charged residues. Amyloid formation is influenced by the presence of specific residues (e.g., P, Q, N, E, D, Y, W, F), pH, temperature, salt concentration, and solvent polarity.<sup>[24–27]</sup>

Since the molecular weight of amyloids is too large to be studied by solution-state NMR spectroscopy, often too dynamic for crystallographic measurements, and too heterogeneous for routine cryoEM studies, low-resolution methods such as ECD, VCD, FTIR spectroscopies, and AFM imaging have been employed to systematically monitor real-time self-association and amyloid formation. Comprehensive chiroptical analysis of aggregation using electronic circular dichroism spectroscopy (ECD) can shed light down to a particle size of  $\sim 90 < d < 180$  nm ( $d \sim \lambda/2$ ; diffraction limit).<sup>[17,28]</sup> ECD can reveal differences between secondary structures and thus reflect to amyloid formation, when an unstructured polypeptide giving a U-type ECD curve changes to a B-type spectrum, characteristic of amyloids composed of  $\beta$ -sheets.<sup>[29–31]</sup> By measuring the build-up and interconversion of the different types of ECD spectra as a function of time (t), pH, concentration (c) and temperature (T), etc., and recording series of  $\Theta^{CD} \sim f(t, pH, c, T)$  functions, a semi-quantitative amyloid analysis becomes possible.<sup>[12,17,31,32]</sup>

The phase separation of oligopeptides in aqueous media is a prerequisite for gaining functional properties (e.g., catalysis, enrichment, size separation) during abiogenesis. Therefore, aggregation and amyloid formation of sequences could have served as evolutionary selection criteria. If a self-assembled sequence mutates, and the resulting new sequence variants are also amyloidogenic, this could be advantageous for increasing sequence diversity. Such diversification would expand the pool of explored sequences, thereby increasing the likelihood

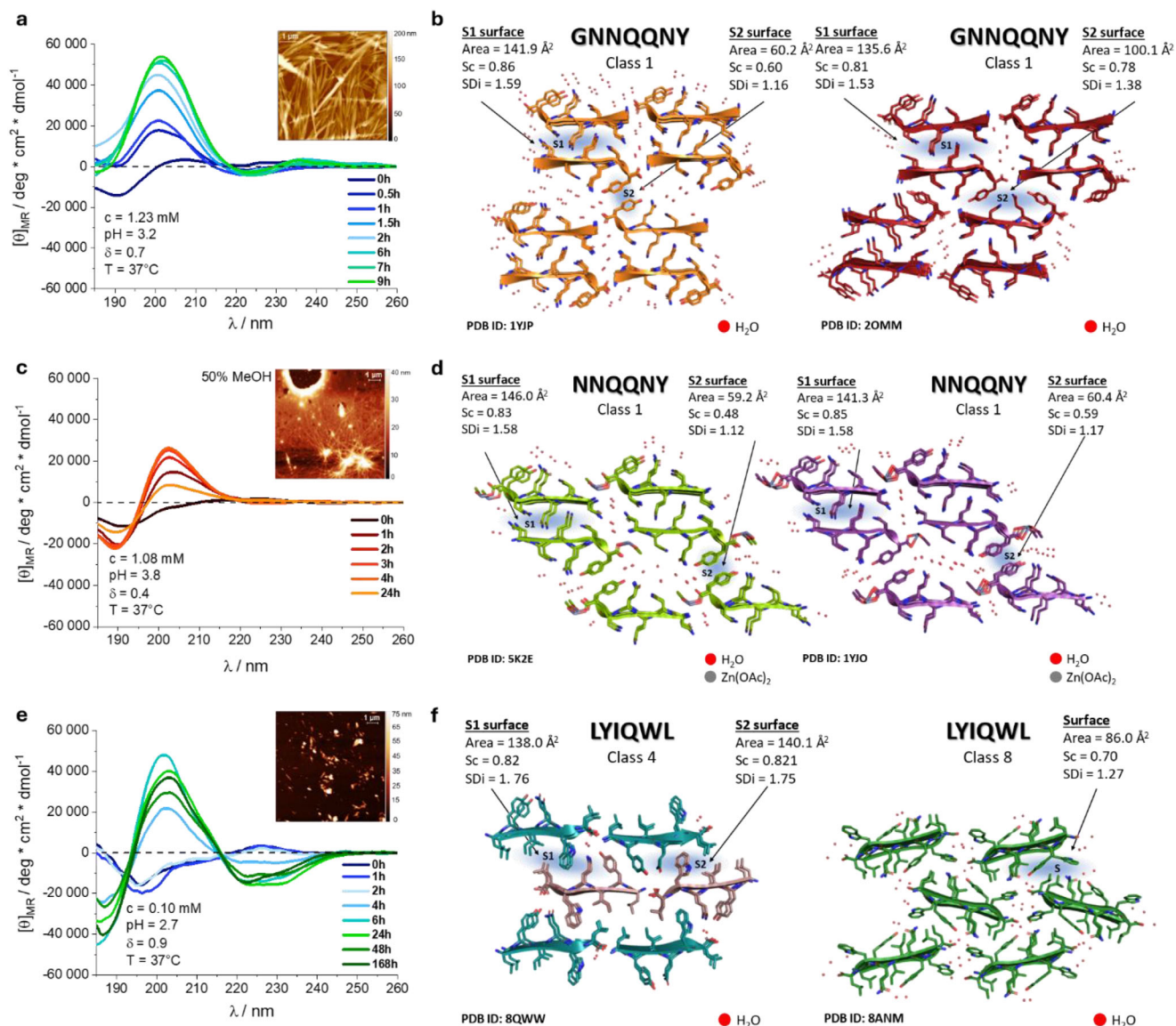
of discovering new chemical functions. On the other hand, if mutants near amyloidogenic sequences are non-amyloidogenic and remain monomeric in water, these sequences would rarely gain self-association-catalyzed functions. In this evolutionary model, there are direct pathways by which evolution could test a sequence's usefulness, and hidden pathways associated with monomeric sequences that could not be tested due to their lack of self-association. We will refer to the first pathway as the pro-amyloid pathway and the second as the anti-amyloid pathway.

Our goals were to demonstrate for two arbitrarily selected APR hexapeptides that they can interconvert via: 1) a mutational pathway in which all intermediate linker mutants are highly amyloidogenic (pro-amyloid pathway), or 2) a pathway in which all mutants are highly dynamic with little or no ability to form amyloid (anti-amyloid pathway). 3) We aim to show that both pro-amyloid and anti-amyloid pathways can coexist for pre-selected endpoints. 4) We wanted to investigate how successive point mutations influence the molecular structures of amyloids. 5) We also aimed to elucidate the contributions of specific side chains (N and Q) and aromatic residues (Y, W, F) to amyloid fibril formation. 6) Additionally, we sought to examine the influence of factors such as pH, solubility, and co-solvents on the generation of supersaturated solutions that initiate amyloid formation.

## 2. Results

### 2.1. Pinpointing the Possible Mutation Pathways

To support the hypothesis that versatile amyloid structures may have evolved to play a role in chemical evolution, we tested the interconversion possibilities between two arbitrarily chosen "endpoint" peptides of known amyloidogenicity. Although any amyloidogenic hexapeptides could have been selected to address our fundamental question, we chose two that, for the sake of interest, have biological relevance. The first APR, LYIQWL, is derived from a variant of Exendin and could form a functional amyloid like glucagon.<sup>[27]</sup> The second endpoint oligopeptide, NNQQNY,<sup>[32]</sup> is derived from Sup35, an essential protein in yeast for efficient translation termination, and also serves as a stress-sensing sequence with phase separation potential.<sup>[33–36]</sup> (PDB ID: 5K2E, 1YJO) The hydrophobicity index of LYIQWL is positive (+1.067), indicating that it is more hydrophobic, while GNNQQNY has a negative GRAVY index (-2.743), showing that it is more hydrophilic.<sup>[37,38]</sup> (Table S4). The more hydrophobic LYIQWL can reach supersaturation in water and form  $\beta$ -sheets at a lower concentration and faster (1 hour), whereas the more hydrophilic GNNQQNY requires about 10 times the concentration and more time (24 h) to achieve the same. However, the endpoint sequences share several similar molecular features: both contain at least one aromatic residue, an amide group on one of their side chains and can carry an explicit charge at their free N- and/or C-termini depending on pH, though neither has a chargeable side chain under physiological conditions. These similarities suggest that the conditions under which they form



**Figure 1.** a, c, e) Amyloid formation of GNNQQNY, NNQNNY, and LYIQWL were monitored by ECD spectroscopy as a function of the time ( $0 \leq t(h) \leq 168$ ) shows a high similarity, although their water solubility is different:  $c_{\text{GNNQQNY}} = 1.25 \text{ mg mL}^{-1}$ ,  $c_{\text{NNQNNY}} = 3\text{--}5 \text{ mg mL}^{-1}$ ,  $c_{\text{LYIQWL}} = 0.167 \text{ mg mL}^{-1}$ . NNQNNY is too hydrophilic to form  $\beta$ -sheets in pure water, so we decreased the dielectric constant by adding methanol to the solution, and  $\beta$ -sheets formed. Insets of a, c, e panels are amyloid fibrils of these endpoint oligopeptides measured by AFM. b), d), f) Amyloid-like X-ray structures of the endpoint oligopeptides with hydrophobic cores highlighted by blue shadows: b) GNNQQNY has a loosely packed structure (PDB ID: 1YJP) forming a single type of core structure and showing a more extensive network of aromatic interactions (S2 surface). The other polymorphic amyloid structure of GNNQQNY, on the other hand, has a more tightly packed double core (PDB ID: 20MM). d) Two very similar crystal structures of NNQNNY (PDB ID: 5K2E and 1YJO), each forming a very similar hydrophobic zipper core. f) The antiparallel  $\beta$ -stranded class 8 amyloid crystal (PDB ID: 8ANM) of LYIQWL formed from pure H<sub>2</sub>O with a single hydrophobic core. The parallel stranded class 4 amyloid crystal (PDB ID: 8QWW) of LYIQWL has two very similar hydrophobic cores: surface S1 and surface S2. SDi is the surface detail index which provides information about the interface detail level.<sup>[7]</sup>

amyloids may be similar. According to literature data, supported by our recent biophysical characterization (ECD, FTIR, AFM), both peptides are unstructured in water but form amyloids within a reasonable timescale (Figure 1).<sup>[28]</sup>

Amyloid formation is successfully monitored in detail by ECD when a variety of spectra are recorded as a function of time. While the pH and incubation temperature at which these endpoint oligopeptides form amyloids are almost the same (pH  $\sim$  3 and  $T = 37^\circ\text{C}$ ), the required concentration for GNNQQNY is higher than that for LYIQWL. Moreover, LYIQWL requires some

time before the characteristic B-type spectra appear, whereas for GNNQQNY, the same type of ECD spectral shift occurs within the first hour (Figure 1). The differing concentrations and incubation times required for  $\beta$ -sheet and amyloid formation are attributed to their different water solubilities. The more hydrophobic LYIQWL can reach supersaturation in water and form  $\beta$ -sheets at a lower concentration, while the more hydrophilic GNNQQNY requires about 10 times the concentration to achieve the same result. Fibrillar amyloid formation was captured by AFM images for both LYIQWL and NNQNNY, as



well as for its parent GNNQQNY (Figure 1 inlets). A comprehensive analysis of the resulting amyloid information was enabled by simultaneously recording FTIR and ECD data and measuring AFM images (SFigure 7). Each measurement was taken from a 1.5 mg/mL stock solution ( $1.67 \pm 0.05$  mM) at pH 3.8, after 24 hours of mixing at 500–600 rpm at 37 °C. While the AFM images reveal amyloid fibril morphologies, the B-type ECD spectra align with the  $\nu(\text{C}=\text{O})$  stretching vibrational modes detected by FTIR at 1629  $\text{cm}^{-1}$ . This analysis concludes that both hexapeptides form amyloids (SFigure 7e, f). Furthermore, X-ray data for LYIQWL show that it is a true APR, providing possibly the largest pool of polymorphic amyloid crystal structures to date. The different amyloid topological classes identified for it (classes 1, 4, and 8) demonstrate that both antiparallel and parallel  $\beta$ -strands can be present, as also presumed for its solution-state amyloid assemblies (Figure 1f).<sup>[28]</sup> Side chains, including Tyr, Gln, and Trp residues, are involved in the H-bond network at the zipper interface formation. The aromatic side chains of Tyr and Trp form a cluster that stabilizes the amyloid, with the side chains of Trp oriented in several directions (Figure 1f-right). The amyloid-like GNNQQNY,<sup>[33,35]</sup> along with its G-truncated variant used as an endpoint (NNQQNY), also forms polymorphs (Figure 1b,d).<sup>[38–41]</sup> To conclude, NNQQNY and LYIQWL as endpoints share several similar molecular features: both contain at least one aromatic residue, an amide group on one of their side chains, and can carry an explicit charge at their free N- and/or C-termini depending on pH, but neither has a chargeable side chain under physiological conditions. These similarities suggest that the conditions under which they form amyloid could be set similarly, and thus parameters such as temperature and pH could be adjusted accordingly.

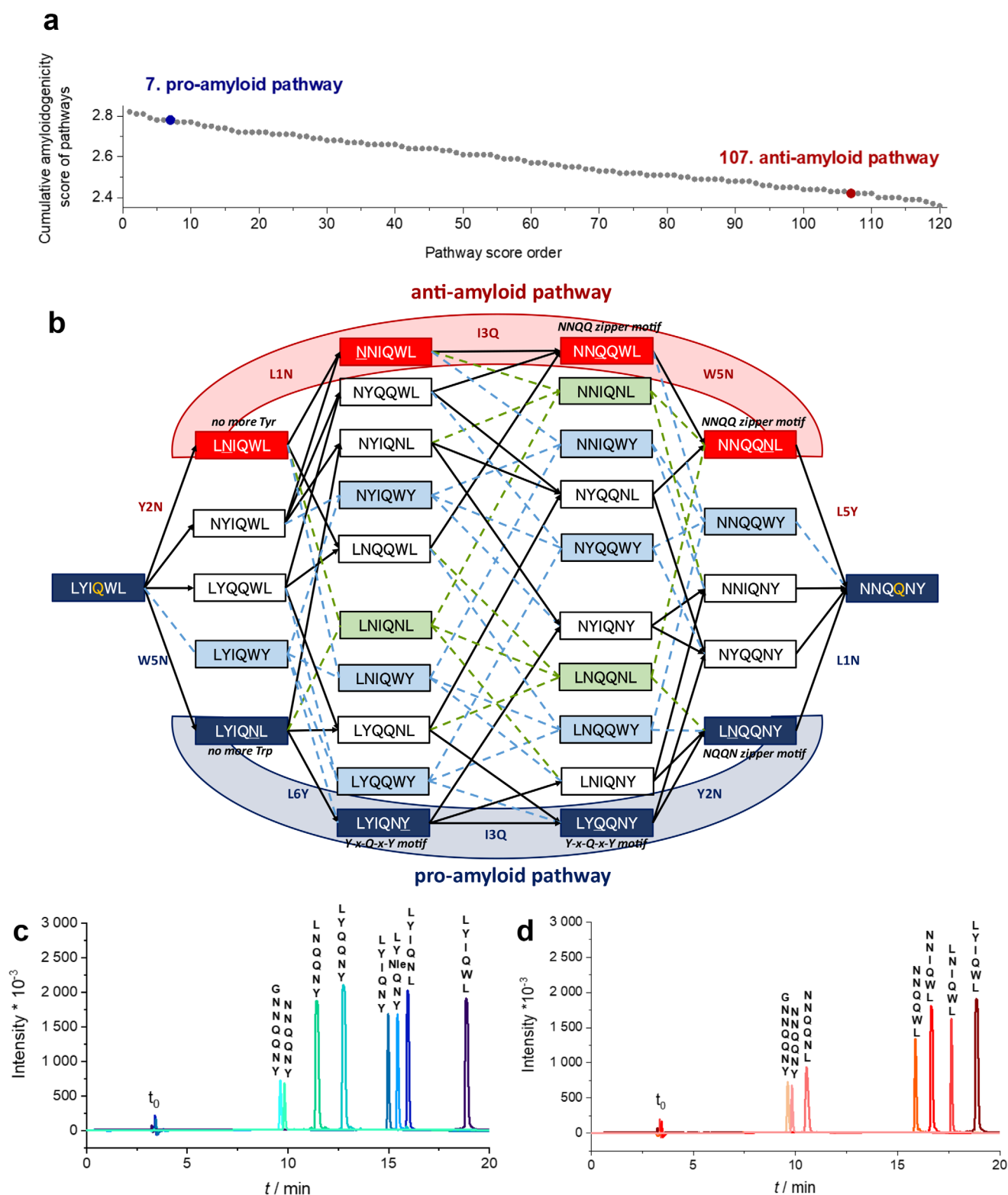
The shortest possible mutation pathways must be selected, avoiding unnecessary intermediate steps. Ideally, each mutant in the pathway should contain one, but no more than two, non-adjacent aromatic residues. One aromatic residue is sufficient for accurate concentration determination, while two aromatic side chains could complicate the ECD spectrum due to aromatic–aromatic interactions, potentially suppressing the U to B-type spectral shifts associated with amyloid formation. Among the 120 possible pathways, we reduced the possibilities to 32 by applying these criteria (Figure 2b – solid line).

We will demonstrate the paramount importance of the exact order of mutations for solubility and amyloidogenicity, as the same 5 mutations are always performed, but in a different order, to create each new pathway. The modeling of chemical evolution diversity is achieved by altering the chemical environment around each residue along each distinct pathway, which then determines their differing amyloidogenicity. To maximize the difference between pathways and identify the most pro- and anti-amyloid ones, three additional considerations were introduced. First, Trp should be eliminated in the pro-amyloid pathway, as it can cause high diversity in orientation,<sup>[44]</sup> but preserved in the anti-amyloid pathway. Since both Tyr-Xxx-Gln and Gln-Xxx-Tyr motifs stabilize  $\beta$ -strand backbone structures, the Y-x-Q / Q-x-Y triplets should preferably be introduced only for the pro-amyloid pathways. Based on these considerations, we chose to synthesize the following pro-amyloid pathway, which includes the

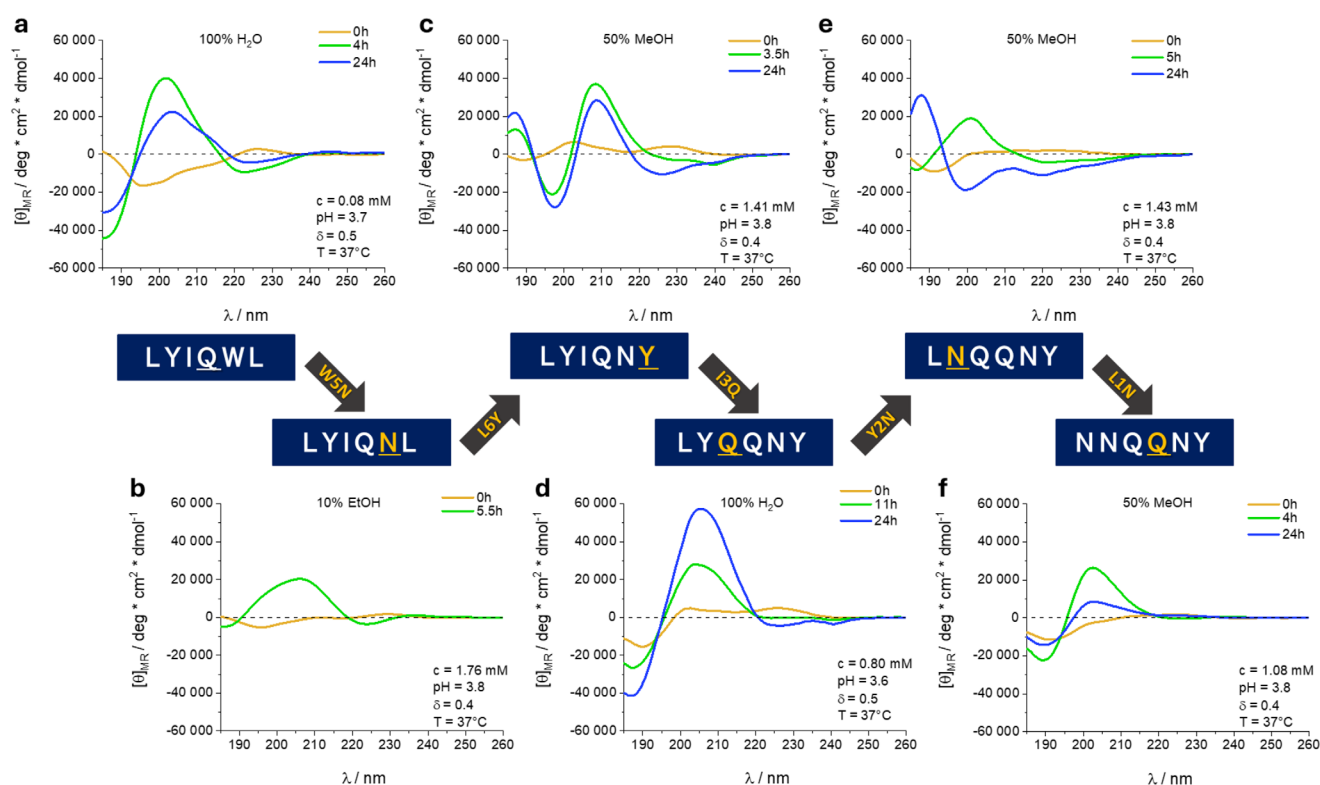
mutation order: LYIQWL→W5N, L6Y, I3Q, Y2N, L1N→NNQQNY (blue pathway in Figure 2b and #7 from STable 1), with a cumulative ANuPP predicted amyloidogenicity score of 2.78. The selected anti-amyloid pathway is: LYIQWL→Y2N, L1N, I3Q, W5N, L6Y→NNQQNY (red pathway in Figure 2b and #107 from STable 1), with a cumulative ANuPP37 score of 2.42. The 12 pro- and anti-amyloid pathway mutants were synthesized and characterized (STable 2). Retention time shifts measured by RP-HPLC show their intrinsically different hydrophobicity, which changes according to their ANuPP indices,<sup>[45,46]</sup> signaling that they will have different water solubility (Figure 2c, STable 1, 4). While the NNQQNY corner hexapeptide has the shortest retention times ( $\sim 10$  min) and is the most hydrophilic oligopeptide, LYIQWL has the longest ( $\sim 18$ –19 min). Considering hydrophobicity and solubility, oligopeptide concentration is also a determinant factor in amyloid formation, which is influenced by temperature and solvent composition. Therefore, adding co-solvents to reach supersaturation and trigger amyloid formation is a known practice to initiate self-association.<sup>[47,48]</sup> Ethanol in H<sub>2</sub>O ( $\sim 10\%$ ) shifts the dielectric constant of the solvent with respect to pure water (STable 3) and helps to solubilize hydrophobic oligopeptides (e.g., LYIQWL) as well as nascent amyloid particles. Moreover, changing the solvent composition can alter the amyloid topology, as seen for LYIQWL, where the amyloid class changes with the co-solvent used. The choice of using pure and solvent mixtures is based on literature and our experience,<sup>[43,47,48]</sup> resulting in the following systems: 100% H<sub>2</sub>O:  $\epsilon = 74.188$ , 10% EtOH:90% H<sub>2</sub>O:  $\epsilon = 68.742$ , and 50% MeOH:50% H<sub>2</sub>O:  $\epsilon = 51.347$ , respectively (see STable 3)

## 2.2. Characterization of the Pro-Monomeric Pathway

Along the pro-amyloid pathway, the first mutation, W5N (LYIQWL→LYIQNL), results in a hexapeptide that forms amyloid in pure water, similar to its parent sequence (Figure 4a). Due to its free N- and C-termini, it is positively charged ( $\delta = +1$ ) in strongly acidic conditions and negatively charged ( $\delta = -1$ ) in basic medium. These unipolar states (SFigure 2) are unsuitable for amyloid formation, but adjusting the pH to a near-neutral charge (e.g.,  $\delta \sim +0.5$  at pH  $\sim 3.8$ ) enables fibril formation (SFigure 3a).<sup>[49,50]</sup> LYIQNL forms amyloid over a broad pH range, as long as unipolarity is avoided. In addition to pH, oligopeptide concentration, temperature, and solvent composition influence amyloid formation. Adding co-solvents to achieve supersaturation is a common strategy to initiate self-association.<sup>[42,43,49]</sup> Ethanol in H<sub>2</sub>O ( $\sim 10\%$ ) shifts the dielectric constant relative to pure water (STable 3), aiding the solubilization of hydrophobic oligopeptides (e.g., LYIQWL) and nascent amyloid particles. Moreover, solvent composition impacts amyloid topology, as seen with LYIQWL, where different co-solvents yield distinct amyloid classes. Based on literature and experimental data,<sup>[42–44]</sup> we used three solvent systems: 100% H<sub>2</sub>O ( $\epsilon = 74.188$ ), 10% EtOH:90% H<sub>2</sub>O ( $\epsilon = 68.742$ ), and 50% MeOH:50% H<sub>2</sub>O ( $\epsilon = 51.347$ ) (STable 3). For the more hydrophilic NNQQNY hexapeptide, the 50% MeOH/H<sub>2</sub>O mixture was necessary for amyloid formation (see below).



**Figure 2.** a) The different mutation orders, 120 in total, give rise to different mutation pathways connecting LYIQWL and NNQQNY endpoint and *vice versa*. The cumulative ANuPP amyloidogenicity scores<sup>[42,43]</sup> characterizing the 120 different mutation pathways, ranked in decreasing order of amyloidogenicity. The blue dot (7th) represents the pro-, while the red dot (107th) represents the anti-amyloid pathway selected for analysis. b) If we consider only those pathways where mutants have at least one but at most two aromatic residues, the number of viable pathways is reduced from 120 to 32 (shown as black solid arrows), while rejected dead-end pathways are shown as dashed arrows. (Sequences with zero (light green) and with more than two or neighboring Trp (W) and Tyr (Y) (light blue) aromatic residues were marked but avoided). The pro-amyloid pathway is highlighted in dark blue, while the anti-amyloid pathway is shown in red. The reversed-phase HPLC chromatograms of the selected oligopeptides forming the c) pro- and d) anti-amyloid pathways follow their order like the GRAVY score. (Table 4).



**Figure 3.** ECD spectra of the pro-amyloid pathway elements connecting the two endpoints: **a)** LYIQWL and **f)** NNQQNY, starting with the mutations W5N, L6Y, I3Q, Y2N and L1N. All the mutants were designed as highly amyloidogenic sequences: **b)** LYIQNL, **c)** LYIQNY, **d)** LYQQNY, **e)** LNQQNY and **f)** NNQQNY and indeed all gave a B-type ECD spectrum in the corresponding solvent(s), indicating amyloid formation. The yellow, green, and blue curves correspond to the data points recorded at  $t = 0$  hours, 3–11 hours, and 24 hours, respectively.

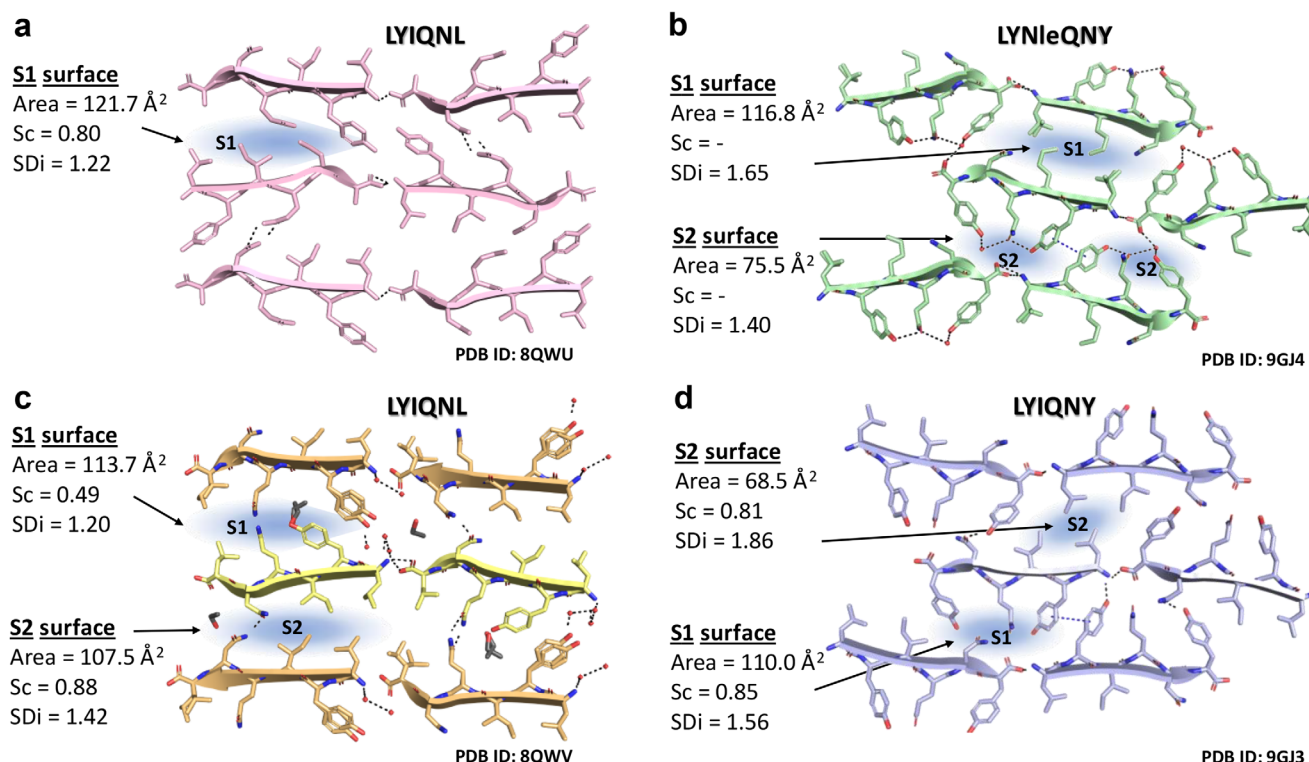
Both LYIQWL and LYIQNL form amyloid in MeOH/H<sub>2</sub>O (1:1) and EtOH/H<sub>2</sub>O (1:9) mixtures, as indicated by their similar B-type ECD spectra (Figure 3a and b; SFigure 4g, h). The initially unstructured, LYIQNL gradually adopts amyloid conformation, as evidenced by the U→B-type ECD spectral shift. In pure water, simpler  $\beta$ -layers form, leaving Tyr side chains mobile and unpacked, preventing measurable Tyr-specific ECD signals. However, adding EtOH or MeOH creates apolar interfaces, stabilizing Tyr side chains and generating strong  $\pi$ - $\pi$  interaction signals ( $\sim 235$  nm, SFigure 4g). This structural ordering increases hydrophobicity, accelerating amyloid aggregation.

FTIR data (SFigure 7a, f) reveal characteristic  $\beta$ -sheet C = O vibrational modes at  $1626\text{ cm}^{-1}$  and  $1628\text{ cm}^{-1}$ , while AFM images confirm filamentous amyloid structures. LYIQNL forms parallel  $\beta$ -stranded amyloids in these solvent mixtures, producing amyloid crystals in both ethanol-containing and ethanol-free conditions. The dielectric constant determines amyloid topology: in water, a class 4 amyloid is obtained, whereas ethanol-containing media yield a class 3 amyloid.<sup>[44]</sup> Though the  $\beta$ -strand packing differs between class 3 and class 4 amyloids, their  $\beta$ -layer interfaces are similar. As expected, intra-sheet side-chain H-bonds stabilize  $\beta$ -strands via Asn- or Gln-ladders, while inter-sheet H-bonds occur between Asn-Asn and Gln-Gln (class 3) or Asn-Gln (class 4) amyloids (Figure 4a and c).

The next mutation, L6Y (LYIQNL→LYIQNY), results in rapid amyloid formation in both water and alcohol-water mixtures. The process is so fast that the U-type ECD spectrum of the unstruc-

tured monomeric form is barely detectable at  $t \sim 0$  min (Figure 4c, SFigure 3e). The characteristic B-type ECD spectra, present from the start of incubation, confirm amyloid formation, further supported by FTIR and AFM data (SFigure 7). MeOH's significantly lower dielectric constant<sup>[51]</sup> ( $\epsilon = 29.5$  at  $37^\circ\text{C}$ ) compared to water ( $\epsilon = 74.2$  at  $37^\circ\text{C}$ ) creates a 50% MeOH-H<sub>2</sub>O mixture with  $\epsilon = 51.3$  ( $37^\circ\text{C}$ ), reducing LYIQNY solubility and promoting self-association.<sup>[48]</sup> The solvent shift induced by MeOH on the B-type ECD spectra is minimal, reinforcing that amyloid formation also occurs in 50% MeOH. Needle-like LYIQNY single crystals grown from aqueous media exhibit parallel  $\beta$ -sheets with both class 1 and class 4 topologies. These mutations preserve the sequences' intrinsic amyloidogenicity and crystallization potential. The primary change among polymorphs is the transition from class 1 to class 3 and 4 topologies, while all amyloid structures remain stabilized by Gln- and/or Asn-ladders.

LYIQWL, LYIQNL, and LYIQNY in the pro-amyloid pathway all contain the Y-x-Q motif, which is proposed to form intra-strand H-bonds between adjacent Tyr and Gln residues. These bonds stabilize an extended backbone structure, facilitating amyloid formation. Structural data from LYIQWL's seven amyloid polymorphs support this  $\beta$ -strand stabilization. Among LYIQNL's two polymorphs, only one (class 3, Figure 4c) contains this H-bond, while the other (class 4, Figure 4a) does not. In LYIQNY, the "tandem" Y-x-Q-x-Y motif creates steric interference, preventing simultaneous formation of both H-bond networks (Figure 4d). This motif generates a polar surface that is easily hydrated,



**Figure 4.** Amyloid-like crystal structures grown from peptides along the pro-amyloid pathway. **a, c)** Crystal structures of LYIQNL grown from aqueous and ethanol-containing media, respectively. **b)** Structure of crystals of LYNleQNY grown from ethanol containing media. **d)** Structure of crystals of LYIQNY grown from aqueous media. All sequences form parallel  $\beta$ -sheets in their crystal structures. Intra-strand Y-x-Q hydrogen bonds are shown as black dashed lines present in **b)** and intermolecular aromatic interaction with blue dashed lines **b, d)** and in the case of **c)** perhaps it could also form. Hydrophobic interfaces at the steric zipper interfaces are highlighted with blue shadows. Note, that the hydrophobic interfaces in the structure of LYNleQNY (**c)** and LYIQNY (**d)** belong to the class 1 topology. Despite their highly similar unit cell parameters, this class 1 interface is significantly more pronounced in the structure of LYNleQNY than in that of the LYIQNY.

reducing amyloid propensity. However, as seen in protonated Asp and Glu amyloids,<sup>[28]</sup> intermolecular H-bonding enhances  $\beta$ -sheet adhesion, stabilizing the amyloid. In LYIQNY, the Tyr, Gln, and Asn side chains cluster with the oppositely charged C- and N-termini, forming a large hydrophilic interface (Figure 4d). Meanwhile, Leu and Ile create a properly aligned hydrophobic interface on the opposite side, further stabilizing the amyloid. The 2:4 ratio of apolar to polar side chains keeps LYIQNY water-soluble, explaining its slower amyloid formation in water and faster assembly in the less polar 50% MeOH-water mixture.

The third mutation, I3Q (LYIQNY  $\rightarrow$  LYQQNY), induces a significant ECD spectral shift compared to its parent hexapeptide. LYQQNY exhibits a B-type spectrum after 24 hours (Figure 3a,d), closely resembling LYIQWL. The intense negative ( $\ominus \sim -42000$  at 188 nm) and positive ( $\oplus \sim +57000$  at 208 nm) maxima, along with a negative maximum pair at 225 and 237 nm, indicate amyloid formation with high  $\beta$ -strand content. Amyloid formation is confirmed in pure H<sub>2</sub>O, MeOH/H<sub>2</sub>O, and EtOH/H<sub>2</sub>O mixtures (Figure 3d, SFigure 3d). LYQQNY retains the tandem Y-x-Q-x-Y motif, which likely catalyzes amyloid formation via intra-strand H-bonds, though this remains unconfirmed in the solid state. Needle-like microcrystals of LYQQNY have been obtained from various solvent mixtures, but diffraction-quality crystals have not yet been achieved. Introducing norleucine, the isomer of leucine at position 3 (Ile3Nle) in LYNleQNY enabled successful

crystallization and X-ray analysis. Like LYIQNY, which belongs to C1-C4 amyloid classes (Figure 4d), LYNleQNY forms amyloid crystals in 10% EtOH with identical unit cell parameters. This is expected, as the isobutyl side chain of Ile closely resembles the normal butyl side chain of Nle. Both sequences exhibit similar 3D amyloid core structures, with a Leu $\leftrightarrow$ Ile $\leftrightarrow$ Ile $\leftrightarrow$ Leu zipper mirroring the Leu $\leftrightarrow$ Nle $\leftrightarrow$ Nle $\leftrightarrow$ Leu arrangement (Figure 4b,d). Additionally, aromatic-aromatic interactions stabilizing quaternary amyloid structures display comparable geometric features (SFigure 6). Despite these similarities, subtle differences emerge. Replacing Ile3 with Nle3 alters the sterically adjacent Asn5 side-chain conformation, modifying its H-bond network, which is complemented by a water molecule in the LYNleQNY amyloid crystal (Figure 4b). Furthermore, the displacement between adjacent  $\beta$ -sheets shifts, tightening the hydrophobic class 1 interface (Figure 4b,d).

The fourth mutation, Y2N, completes the NQQN unit within hexapeptide LNQQNY, known as the shortest crystallized amyloid (PDB: 2OLX, 2ONX).<sup>[35]</sup> However, LNQQNY forms amyloid very slowly in pure water and 10% EtOH. In contrast, in 50% MeOH, the initially unstructured hexapeptide transitions into amyloid within 24 h, as confirmed by ECD (Figure 3e, SFigure 3c). The characteristic B-type ECD spectrum and the emergence of a negative band at 235 nm indicate  $\pi$ - $\pi$  interactions among Tyr residues, potentially catalyzing amyloid structuring. Similar to



the parent LYQQNY, LNQQNY produces microcrystals in various solvent mixtures, but they remain too small for X-ray crystallography. Nevertheless, time-dependent ECD spectra and FTIR data (SFigure 7d) confirm its amyloidogenic properties.

The fifth mutation, L1N, yields NNQQNY, the most polar peptide of the pathway. With five amide side chains, it is highly water-soluble, preventing amyloid formation in pure water, or 10% EtOH, even at elevated concentrations, such as 2–5 mg/mL. However, in 50% MeOH, NNQQNY forms amyloid within hours, as evidenced by an intense B-type ECD signal ( $\Theta_{222}$  nm $\sim$ 53000) (Figure 3f). Interestingly, NNQQNY can crystallize into a class 1 topology amyloid in water, but only under specific conditions: high peptide concentration (30 mg mL $^{-1}$ ), 20 °C, with zinc sulfate and sodium acetate in HEPES buffer at pH 7.0. Structural analysis of this crystal (PDB: 5K2E, 1YJO) reveals stabilization via Tyr $\leftrightarrow$ Tyr contacts ( $d \sim 4.9$  Å) (Figure 1d), though this  $\pi$ - $\pi$  interaction is not detectable in solution by ECD spectroscopy.

In summary, we successfully established a pro-amyloid pathway where sequential point mutations consistently yield amyloidogenic APRs. Each of the five designed mutants exhibited characteristic B-type ECD spectra, either in water or MeOH/H $_2$ O, at low peptide concentrations ( $T = 37$  °C) within 24 h, confirming their intrinsic amyloidogenicity. The formation of  $\beta$ -fibrils is validated by FTIR C=O stretching vibrations (1626–1629 cm $^{-1}$ ) and AFM imaging (SFigure 7). All mutants contain the Y-x-Q and/or Q-x-Y motif, supporting an extended  $\beta$ -strand backbone. In LYQQNY and LYIQNY, where the tandem Y-x-Q-x-Y motif appears, aromatic-aromatic interactions are also detected by ECD. As hydrophobic residues (L, I) are replaced with polar ones (N, Q), APR solubility increases, requiring higher concentrations and cosolvent (e.g., MeOH) to lower the dielectric constant and induce amyloid formation. Trp removal reduces polymorphic amyloid diversity. LYIQWL exhibits the broadest range of  $\beta$ -strand topologies (class 1 to class 8- see SFigure 8). Mutation to LYIQNL yields class 3 and 4 amyloid crystals, while LYIQNY represents class 1 and 4. The pathway culminates in NNQQNY, which exclusively forms class 1 amyloid crystals. Aromatic side chains play a crucial role in amyloid higher-order structuring. While both pathway endpoints (LYIQWL and NNQQNY) are amyloidogenic, they differ in their  $\beta$ -sheet orientation (parallel or antiparallel), water solubility and GRAVY index<sup>37</sup>. Hydrophilic amyloids (negative GRAVY index) predominantly adopt class 1 topology with parallel  $\beta$ -strands ( $SD_i \geq 1.5$ ), accumulating Q/N residues. Conversely, hydrophobic sequences (positive GRAVY index) exhibit greater structural diversity, supporting both parallel and antiparallel  $\beta$ -strands, spanning multiple topologies (class 1, 3, 4, 8), with  $SD_i \leq 1.5$ . A detailed analysis was provided.<sup>7</sup>

### 2.3. Characterization of the Anti-Monomeric Pathway

In the anti-amyloid pathway analysis, we aimed to determine whether a mutation sequence exists between LYIQWL and NNQQNY where no intermediates exhibit amyloid formation under conditions probed above. To achieve this, we minimized or eliminated the Y-x-Q and Q-x-Y motifs that stabilize  $\beta$ -strands and have disrupted N and Q residue interactions to prevent H-

bond ladders. We maintained pH within 2.7–3.8, consistent with the pro-amyloid pathway (SFigure 3b). The following mutation sequence was designed: Y2N, L1N, I3Q, W5N, and L6Y, resulting in the expected anti-amyloid pathway: LYIQWL  $\rightarrow$  LNIQWL  $\rightarrow$  NNIQWL  $\rightarrow$  NNQQWL  $\rightarrow$  NNQQNL  $\rightarrow$  NNQQNY described below (Figure 2b, Table 1).

The first mutation (Y2N) converts highly amyloidogenic LYIQWL to LNIQWL, removing the aromatic residue from the Y-x-Q motif, which stabilizes amyloid structures via  $\pi$ - $\pi$  interactions and/or H-bonds. Introducing Asn enhances water solubility. While this mutation reduces amyloidogenicity, LNIQWL remains amyloidogenic, transitioning from a U-type to a B-type ECD spectrum within 1–2 h (Figure 5b). AFM imaging confirms “granular” amyloid formation (SFigure 7g), and FTIR reveals multiple amide I bands, with the strongest C = O vibration at 1629 cm $^{-1}$ , characteristic of  $\beta$ -strands. This indicates that a single mutation is yet insufficient to fully suppress the very strong APR amyloidogenicity of LYIQWL (SFigure 3i).

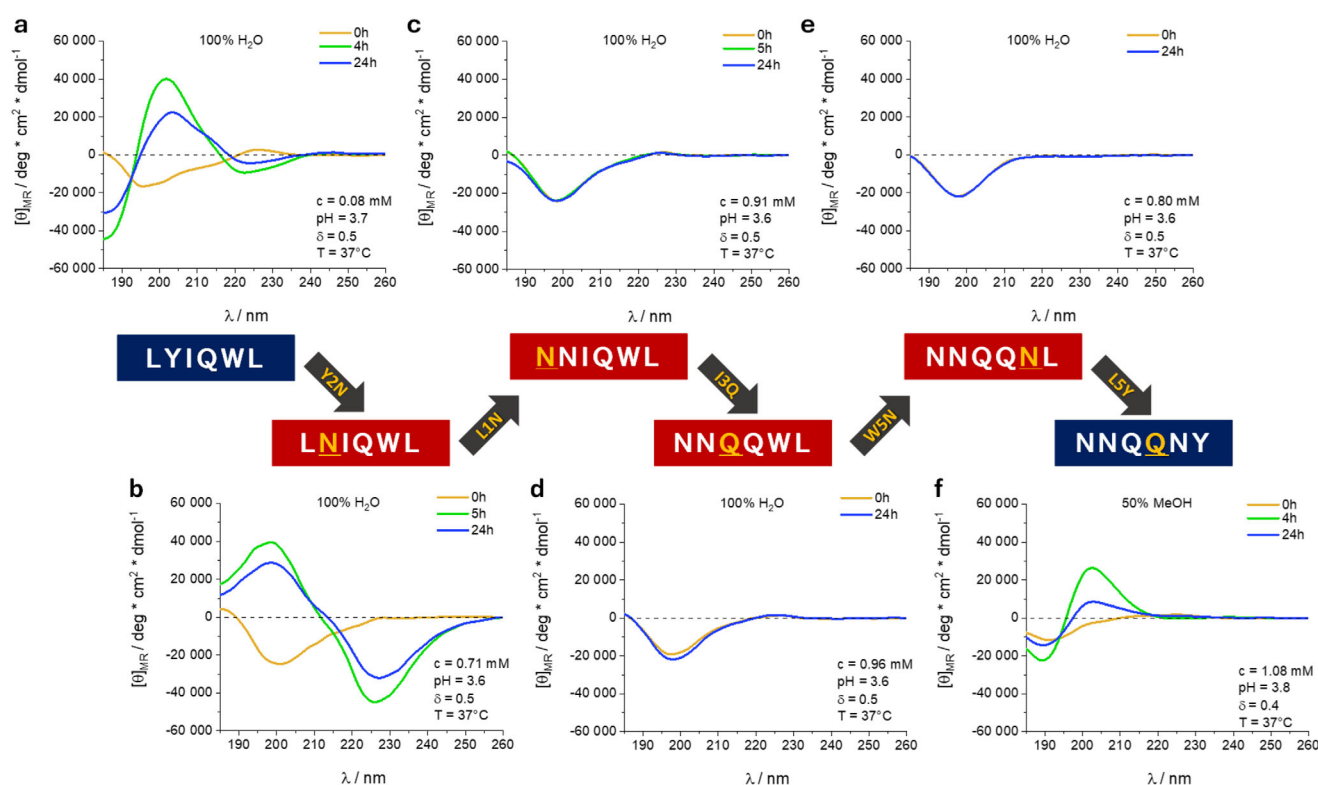
The second mutation, L1N, produces NNIQWL, a hexapeptide that did not form amyloid under any of the tested conditions (Figure 5c, SFigure 3j), as the U-type ECD spectrum remained unchanged over time, and AFM imaging showed no evidence of filament formation (SFigure 7h), confirming that NNIQWL remains monomeric and intrinsically unstructured. The parent peptide LNIQWL contains an L-x-l-x-W hydrophobic interface, while NNIQWL has a more hydrophilic N-x-l-x-W interface, which is alternately hydrated, preventing self-association and keeping the peptide monomeric in water.

The third mutation, I3Q, results in NNQQWL, a mutant that also resists amyloid formation. The U-type ECD spectrum remained unchanged, even after 144 h of incubation at 37 °C ( $c_{\text{pep}} \sim 1$  mM) (Figure 5d). No amyloid formation was observed, even in the more amyloidogenic 50% MeOH solvent mixture. Despite the potential of the Trp5 indole ring to stabilize amyloid through  $\pi$ - $\pi$  interactions, NNQQWL remained monomeric and unstructured over time (SFigure 3k). This is further supported by AFM imaging, which showed no detectable structures (SFigure 7i).

The fifth W5N mutation results in NNQQNL, where 5 consecutive amide side chains could improve water solubility but stabilize an extended amyloid backbone structure via side chain H-bond ladders. The U-type ECD spectra (Figure 5e), the C = O vibrational mode above 1650 cm $^{-1}$  in the FTIR spectrum, and the empty AFM image (SFigure 7j) indicate that NNQQNL does not form amyloid. These data suggest that amino acids with amide side chains (N and/or Q) are useful, perhaps necessary, but not sufficient on their own to form dry zipper interfaces for stabilizing amyloid nanostructures. This solubility concept is supported by our observation that NNQQNL can form amyloid only when its solubility is reduced by the use of 50% MeOH, since this solvent mixture has a significantly lower dielectric constant. (Incubation at 37 °C allows for very slow amyloid formation, which takes more than 144 h to complete (SFigure 3l).)

The final L6Y mutation results in the hexapeptide NNQQNY, where the Q-x-Y amyloid-enhancing motif is restored, allowing a Tyr-Tyr interstrand  $\pi \leftrightarrow \pi$  interaction to form. Although NNQQNY is too water-soluble to aggregate in H $_2$ O, it forms amyloid in 50% MeOH. (Figure 5f and SFigure 7e) While it took a week for the





**Figure 5.** ECD spectra of consecutive point mutations along the anti-amyloid mutation pathway, connecting endpoints a) LYIQWL and f) NNQQNY. The mutation starts with Y2N and is followed by LIN, IQ, W5N and L6Y, resulting in a series of mutations composed of low amyloidogenic sequences: b) LNIQWL, c) NNIQWL, d) NNQQL, e) NNQQL and f) NNQQNY APR hexapeptides. Yellow, green, and blue curves correspond to the data points at 0 h, 4–5 h, and 24 h, respectively.

parent NNQQL to form amyloid, just a few hours of incubation were sufficient for NNQQNY to achieve the same result. In other words, the presence of an aromatic residue (e.g., Tyr) in a hydrophilic amyloidogenic peptide can strongly enhance its amyloidogenicity.

### 3. Discussion

APRs are short linear motifs often moderated by anti-amyloidogenic residues, known as gatekeepers, which can spatially confine, separate, or silence amyloids. An APR can also be silenced by being embedded in an  $\alpha$ -helix.<sup>[47]</sup> We demonstrate that “naked” APRs, like LYIQWL, LYIQNL, LYIQNY, LYQQNY, LNQQNY, and NNQQNY, are highly amyloidogenic. However, when integrated into globular proteins (SFigure 5a), these APRs must be silenced to prevent unwanted amyloid formation. We identified all these APRs in proteins and concluded they are silenced by surrounding gatekeepers, or embedded in  $\alpha$ -helices. For example, in the Exendin-4 variant Tc5b miniprotein (PDB ID: 1L2Y), the LYIQWL APR is framed by the positively charged R(+) and K(+) gatekeeper residues and also embedded in an  $\alpha$ -helix: NLYIQWLK(+)D(-)... In the E5 variant, gatekeepers directly frame the APR: ...-R(+)LYIQWLK(+)-...<sup>[17]</sup> The LYIQNL APR in Human Palmitoleoyl-protein carboxylesterase NOTUM (PDB ID: 6ZYF) is also silenced by gatekeepers: -R(+)LYIQNLGR(+)-... Similarly, the APR in glutamyl-tRNA reductase<sup>[52]</sup> is silenced by

negatively charged residues: ...-D(-)JLYIQNLAE(-)-. Moreover, both negative and positive charges can simultaneously mask an APR, as shown in the genetic interactor of prohibitin7<sup>[41]</sup>: ...-E(-)R(+)E(-)ILNQQNYLR(+)-.

To prevent unwanted aggregation, gatekeepers must remain in their “on-state”, carrying explicit charges. If switched to the “off-state” and become uncharged, APRs could lead to amyloid formation. The amyloidogenic nature of an APR causes it to be naturally hidden as a helix in a folded protein, providing a secondary risk-reduction mechanism. However, less amyloidogenic APRs can be found in various secondary structural elements (SFigure 5). Evolutionarily, preserving APRs might be beneficial due to their enhanced ability to form hydrophobic interfaces in water, but only in their silenced state. This allows them to bind selectively and with high affinity to specific proteins, such as receptors, without self-interaction. Additionally, heterodimer formation is promoted by suppressing homo-oligomerization, as shown in the GPCR recognition of GLP1 hormone polypeptides.<sup>[28]</sup> We demonstrated that these APR endpoints can be cross-mutated via both pro-amyloid and anti-amyloid pathways. From an evolutionary standpoint, a pro-amyloid pathway could offer a selection advantage, allowing amyloidogenic intermediates to be “probed” and potentially gain function, as amyloids are phase-separated nanoparticles. Elements of the anti-amyloid pathway could evolve in the background, mutating from one amyloid APR to another through water-soluble monomers. In both pathways, identical mutations

(L1N, Y2N, I3Q, W5N, L6Y) occurred at residues 1, 2, 3, 5, and 6, but in a different order.

It is important to consider the starting points that define the scope and results of our analysis. Why did we narrow down the sequence space of millions or billions of elements to find the right mutation pathways? Here are short answers to a dozen of potential questions: 1) Why were the two endpoints chosen arbitrarily? To make generalization more plausible with two quasi-independent sequences. 2) Why is it good to have exactly 6 residues in both endpoint sequences? Because the 64 million hexapeptide sequence space is large enough to adequately represent sequences with uncharged, proteinogenic residues. 3) Why is 15–20% homology considered an indicator of sequence independence? This range is widely accepted in protein science. 4) Why was it good to have mutations at 5 positions instead of 6? Mutating all 6 residues would have resulted in 720 pathways instead of 120, leading to more, potentially unnecessary work. 5) Why is it useful that both endpoints are highly amyloidogenic? We wanted to study amyloid formation, interconversion, and phase separation in water. 6) Why choose endpoints with different levels of amyloid polymorphism? Polymorphism was observed for both LYIQWL (7 polymorphic crystals) and NNQQNY (3 polymorphic crystals), providing a better understanding of amyloid architecture diversity. 7) Why choose endpoints with different water solubilities (GRAVY index)? A wider polarity profile better represents amyloidogenic sequences while ensuring the sequences are water-soluble in the mg/ml range. 8) Why retain an aromatic residue? It may be necessary for amyloid formation, although as shown not sufficient on its own. 9) Why choose one endpoint sequence rich in Asn and Gln? These residues form H-bond ladders, an important feature in amyloids, making differential analysis promising. 10) Why include Trp in one endpoint sequence? While rare in APRs, investigating the role of the indole ring in amyloid formation was considered worthwhile. 11) Why incorporate a non-protein residue such as Nle? Chemical evolution likely involved residues beyond the 20 standard amino acids, and Nle is a small step toward broadening residue diversity. 12) Why choose the shortest path between the endpoints? Although alternatives could have been explored, the complexity of the task led us to focus on the shortest paths for clarity and efficiency.

Along the anti-amyloid pathway, where mutants remain water-soluble, unstructured, and form temporary or no intermolecular complexes, the entropy term remains virtually unchanged. However, for the nanostructures forming along the pro-amyloid pathway, as their complexity increases, entropy decreases. This is compensated by favorable enthalpic terms, lowering Gibbs-free energy ( $\Delta G$ ). Constructive interactions, such as attraction of opposite charges, formation of H-bonds, and  $\pi$ - $\pi$ , cation- $\pi$ , anion- $\pi$ , and dispersive interactions, stabilize amyloids thermodynamically. Furthermore, the complexity of amyloids allows selected residues and side chains, now separated from the aqueous phase, to acquire an evolutionarily meaningful function, such as catalysis, absorption, or phase separation. This concept of increasing complexity via amyloid formation in chemical evolution mirrors what we observe in biological evolution.

During chemical evolution, carbonyl sulfide (COS) activated amino acids formed polyamide systems, producing a variety of oligopeptides and polypeptides.<sup>[3]</sup> While the selective advantage of some oligopeptides is beyond the scope of this paper, we aim to explain which 3D polyamide structures might have favorable properties for enriching and conserving certain primary sequences. Our findings show that in aqueous media, nearby APR mutants can form amyloids with different topological classes. Thus, amyloid formation not only promotes phase separation but also enriches APR sequences selected for potential functions. The rapid mechanism of  $\beta$ -edge self-recognition and self-alignment makes enrichment of class 1–4 amyloids more likely, as parallel  $\beta$ -strands align more easily. This may explain why parallel amyloids were predominantly found in this study and why they are more common in shorter amyloid-forming polypeptides.<sup>[7]</sup>

At the beginning of chemical evolution, both the formation and disintegration, decay, and recycling of amyloid structures could have been important selection criteria. Amyloids appear to be promising nanosystems, forming stable yet not over-stabilized macromolecular systems. When comparing the number of covalent bonds between monomers to the number of weak interactions, such as backbone H-bonds, amyloids exhibit a favorable characteristic. The energy required to break an amide bond is about 100 kcal/mol or more, while breaking an H-bond requires only about 1–2 kcal mol<sup>-1</sup>. Compared to all known polyamide systems, amyloids have the largest number of H-bonds relative to covalent bonds between units. Considering only the number of H-bonds (H) between backbone atoms, crucial for stabilizing 3D structures, and the number of amide covalent bonds (k),  $\beta$ -sheets have the most favorable ratio: H/k  $\sim$ 1.00. For instance, a polypeptide with n residues and  $\sim(n-1)$  amide bonds arranged as parallel  $\beta$ -sheets can form up to n backbone-to-backbone H-bonds, yielding H/k = n/(n-1), which approximates  $\sim$ 1.00. In contrast, the basic domain types in today's globular proteins show lower H/k ratios:  $\alpha$ -fold  $\sim$ 0.60,  $\beta$ -fold  $\sim$ 0.48,  $\alpha/\beta$ -fold  $\sim$ 0.55, and  $\alpha+\beta$ -fold  $\sim$ 0.52.47 Thus, the H/k ratios in amyloids are 67–108% higher than in typical protein domains, suggesting that amyloids achieve stable 3D structures that can however be degraded and recycled with less energy. An additional advantage of amyloids is their polymorphic structures, where a single amino acid sequence can generate several amyloidogenic conformers (e.g., LYIQWL yielding 7 polymorphs), each with different sidechain-sidechain arrangements. This provides alternative local sidechain subspaces with plasticity, making amyloids adaptable. Furthermore, amyloids can serve as catalytic surfaces by forming  $\beta$ -sheets with the appropriate catalytic triad expressed on the steric zipper surface. This properties suggest that amyloids could have been the first complex systems in which peptide degradation occurred.

## 4. Conclusion

In conclusion, the structural polymorphism and latent capacity for  $\beta$ -sheet formation in amyloids have proven essential for

their adaptability and functional versatility. Their ability to adopt different conformations based on environmental variations, combined with their unique H-bond variability and flexibility, likely positioned them as one of the earliest molecular frameworks supporting functional evolution. These properties not only facilitate amyloid formation but also play a significant role in protein aggregation, phase separation, and the development of functional peptides. The insights gained from understanding these mechanisms shed light on the evolutionary significance of amyloids in the emergence of complex biological systems.

## 5. Materials and Methods

### 5.1. Amyloidogenicity Prediction of APRs, Transitional Pathway Design and Ordering

All possible mutation pathways between LYIQWL and NNQQNY were obtained after completing all individual mutations one by one. There are 32 ( $= 2^5$ ) possible peptide sequences between the two endpoints of the pathway (NNQQNY $\rightarrow$ LYIQWL), since Gln4 is conserved, and the total number of mutation pathways is 5! ( $5 \times 4 \times 3 \times 2 \times 1 = 120$ ). To characterize the amyloid propensity of all 32 hexapeptides, we utilized the ANuPP machine learning algorithm (web-based meta-classifier at: <https://web.iitm.ac.in/bioinfo2/anuppp/homeseq1/> (accession time: 2023.08.15)), which focuses on the atomic-level properties of component residues and uses structural information from various databases (CPAD 2.0<sup>[53]</sup>; Waltz-DB<sup>[6]</sup>; AmyLoad<sup>[54]</sup>).<sup>[42]</sup> This method identifies APR sequences of varying lengths in proteins, but at least six amino acids are required to test whether a given sequence contains an APR segment and to predict a score. As we have only tested hexapeptides, the score provided by the method is independent of the length of the sequence, allowing the scores obtained to be compared. Based on these results, a relative amyloid propensity can be established. We propose that the higher the predicted score, the greater the propensity for amyloid formation. In the manuscript, we refer to this score as the “amyloidogenicity score”. All possible direct paths between the starting and target hexapeptides have been constructed using a basic Python script. A path is considered direct if every residue is mutated at most one time, i.e., the peptide sequences are directly interpolated. Then, the amyloidogenicity of the path is defined as the amyloidogenicity sum of its constituting peptides. Finally, paths were ordered increasingly according to their amyloidogenicity score.

### 5.2. APR Model Peptide Synthesis and Purification

The selected model sequences were synthesized using an in-house developed flow chemistry-based solid phase peptide synthesizer<sup>[55,56]</sup> using the Fmoc/<sup>t</sup>Bu strategy. Preloaded Fmoc-AA-Wang Tentagel resins – containing the first C-terminal residue – were used, while coupling was performed with OxymaPure/DIC reagents in DMF at 80 °C (70–90 bar pressure). Oligopeptides were cleaved from the resin by using the mix-

ture of 2.5 v/v% triisopropylsilane, 2.5 v/v% water, and 95 v/v% TFA, at room temperature, with continuous stirring for 3 h. TFA was removed by rotary vacuum evaporator, and the oligopeptides were precipitated in cold diethyl ether. After sedimentation, the diethyl ether was decanted, and the sediment was washed again with fresh ether. This cycle was repeated three times, followed by vacuum drying. The raw oligopeptides were dissolved in 5:95 v/v% of MeCN:H<sub>2</sub>O and filtrated (PTFE membrane:45  $\mu$ m). Peptides were purified by reverse phase HPLC (Jasco LC-2000Plus HPLC system) using C18 column (Phenomenex Kinetex 5  $\mu$ m 150 $\times$ 21.2 mm,) with gradient elution under the following parameters: initial 10-min isocratic elution with 100% A eluent (water containing 0.1% TFA), followed by a gradient slope of 0.5 v/v % B (MeCN containing 0.08% TFA) per minute and then collected fractions were lyophilized. Analytical purity was controlled both by MS (HR MS- Orbitrap) and analytical HPLC ((Jasco LC-2000Plus HPLC system) equipped with Aeris 3.6  $\mu$ m PEPTIDE XB-C18 100 $\text{\AA}$ , 4.6 $\times$ 250 mm column. (See Supplementary Information STable 2 and SFigure 1a–l)

### 5.3. Sample Preparation for ECD Measurement

The purified and lyophilized APR model hexapeptide samples were dissolved in 100 v/v% distilled water, 10 v/v% ethanol – 90 v/v% water mixture or 50 v/v% methanol – 50 v/v% water mixture at the concentration range of 0.15–1.50 mg mL<sup>−1</sup>. The concentration of 5-fold diluted samples was determined using a NanoDrop Lite spectrophotometer (Thermo Scientific), using the molar extinction coefficients ( $\epsilon_{\text{Tyr}} = 1280 \text{ M}^{-1} \text{ cm}^{-1}$  and  $\epsilon_{\text{Trp}} = 5690 \text{ M}^{-1} \text{ cm}^{-1}$ ) of a model compound Glycyl-L-tyrosyl glycine and *N*-Acetyl-L-tryptophanamide, respectively, at 280 nm.<sup>[57]</sup> The pH of each sample was adjusted using 0.01/0.1 M NaOH and HCl solutions (Orion Star A211 pH meter (Thermo Scientific)). (SFigure 3) The samples were continuously stirred by a magnetic stirrer (500–600 rpm) and incubated (37 °C). The stock solutions were directly measured at given times (0–168 h).

### 5.4. Secondary Structure Identification with Electronic Circular Dichroism Spectroscopy (ECD) and Data Processing

The Far-UV CD measurements were performed on JASCO (Tokyo, Japan) J-810 and J-1500 spectropolarimeters equipped with a Peltier temperature controller. Each spectrum was the average of three scans taken in the far UV (185–260 nm) region with a 0.1 mm path length quartz cell. The following settings were used throughout the measurements: a temperature control system at room temperature (25 °C), a bandwidth of 1 nm, a step size of 0.2 nm, a response time of 4 s, and a scan rate of 50 nm min<sup>−1</sup>. All spectra were corrected using the Spectra Analysis function of the JASCO Spectra Manager v2.0 by subtracting the solvent spectrum obtained under identical conditions and by smoothing with a convolution value width of nine, using the Savitzky–Golay method. All corrected spectra were converted from mDeg to mean residue ellipticity ( $[\Theta]_{\text{MR}} / \text{deg cm}^2 \text{ dmol}^{-1}$ ) to account for concentration differences. (SFigure 3 and 4)

### 5.5. Sample Preparation for FTIR and AFM-Based Analysis of Amyloid Formation of APR Peptides

The two selected weak and strong amyloidogenic pathway APR peptides were dissolved in the solvent. ECD measurements were used to select the appropriate solvent composition. Peptides were dissolved with a final concentration of 1.5 mg mL<sup>-1</sup>, pH was adjusted to pH 3.8. All samples were mixed with a magnetic stirrer at 500–600 rpm, at 37 °C. Far-ECD spectra were collected every 0 h, ~4–9 h and 24h. The final 24-hour samples were measured with an FTIR spectrometer equipped with a BioATR reflection element. AFM images were taken from the 24-h solution, 5 µL of the sample was dropped on the freshly cleaved mica plate and dried in a vacuum for 24 h before measurement.

### 5.6. Fourier-Transform Infrared Spectroscopy (FTIR) Measurements and Data Processing

FTIR measurements were performed using a Bruker (Billerica, MA, USA) Equinox 55 FTIR spectrometer equipped with a Bio-ATR (attenuated total reflectance) sample cell, where the internal reflection element is made of a ZnSe crystal. The ZnSe photoelastic modulator of the instrument was set to 1600 cm<sup>-1</sup>, and an optical filter with a transmission range of 1900–1200 cm<sup>-1</sup> was used to enhance the sensitivity in the amide I–II spectral region. The MCT (mercury-cadmium-telluride) detector was cooled with liquid nitrogen. At each sampling point, 30 µL peptide solution was transferred to the ATR cavity of the IR equipment. Each FTIR spectrum was recorded by averaging 128 scans in the range of 4000 to 850 cm<sup>-1</sup> with a resolution of 4 cm<sup>-1</sup> using an aperture of 3000 microns. The IR spectra were calculated from the single-channel DC spectra, and baseline correction was applied to each measurement by subtracting the blank solvent spectrum. Software OPUS 6.5 was applied for data processing and exporting. FTIR, AFM, and ECD comparative measurements were performed simultaneously on the same stock solution. (SFigure 7.)

### 5.7. Analysis of Surface Morphology with Atomic Force Microscopy (AFM) and Data Processing

After 24 h of drying under a vacuum, the prepared samples were measured. Surface morphology analysis was performed using a FlexAFM microscope system (Nanosurf AG, Liestal, Switzerland). This system was operated in dynamic mode and was controlled by Nanosurf control software C3000 version 3.10.4. Tap150GD-G cantilevers (BudgetSensors Ltd., Sofia, Bulgaria) with a tip radius of less than 10 nm were used for the measurements. Prior to data collection, initial scans were performed at lower resolutions to ensure consistent surface topology and to identify any fibril-like structures. Data was collected once and measurements were taken from different locations on a single sample. Images were taken near densely packed areas of the surface using a window size of 10 × 10 µm and a resolution of 512 pixels per line.

Gwyddion 2.62 software<sup>[58]</sup> was used to process the AFM data and generate the images. FTIR, AFM, and ECD comparative measurements were performed simultaneously on the same stock solution.

### 5.8. Single Crystal Formation: LYIQNY

The lyophilized peptide was dissolved in water containing 0.1% TFA at a concentration of appr. 0.05 – 0.2 mg mL<sup>-1</sup> and incubated at 37 °C for several days. LYNleQNY: Lyophilized LYNleQNY was dissolved in 10 v/v% EtOH at 0.3 mg mL<sup>-1</sup> concentration and incubated at 4 °C. Thin needle-like crystals appeared after about 7 months.

### 5.9. X-Ray Diffraction Measurement

Single crystal X-ray diffraction data were collected at 100 K on a Rigaku XtaLab Synergy-R diffractometer using Cu Kα radiation. CrysAlisPro v171.42.58a (Oxford Diffraction / Agilent Technologies UK Ltd, Yarnton, England) was used for data acquisition and data reduction. The phase problem was solved by molecular replacement using a polyalanine β-strand model generated from the previously solved structure of LYIQWL (PDB ID: 8ANM). A phaser from the Phenix package<sup>[59]</sup> was used to solve the structure (Table 1). Manual model building was performed in COOT<sup>[60]</sup> and the resulting models were refined using Phenix.refine<sup>[61]</sup> and Buster.<sup>[62]</sup>

### 5.10. Data Availability

Novel crystal structures have been deposited in the PDB database under accession codes LYIQNY – PDB ID:9GJ3; LYNleQNY – PDB ID:9GJ4

### 5.11. Bioinformatic – Search ARP Sequences in the Protein Database from Genome Sequencing Project and Secondary Structure Distribution

For searching a sequential motif in the UniProt database and querying the protein sequence IDs we used the UniProt RESTful API. A custom script has been written to send a POST request with the peptide sequence payload to <https://peptidesearch.uniprot.org/asyncrest>, wait for the job to complete, and download the corresponding AlphaFold2 predicted structures from <https://alphafold.ebi.ac.uk/files>. In all cases, the file named "AF-[UniProt ID]-F1-model\_v4.pdb" has been downloaded (accession time: 2023.12.20). Then, protein sequences have been filtered by tripeptide content similarity; for every sequence a tripeptide count dictionary has been calculated (with overlapping tripeptide windows), which have been normalized to get tripeptide content distributions. These distributions have been compared through the Hellinger statistical distance to get a dissimilarity measure between any two



**Table 1.** Data collection and refinement statistics. Data for the highest resolution shell are given in parentheses.

	LYIQNY	LYNleQNY
Unit cell parameters <i>a</i> , <i>b</i> , <i>c</i> , (Å) $\alpha$ , $\beta$ , $\gamma$ (°)	4.850, 20.538, 42.326 90.0, 90.0, 90.0	4.826, 20.872, 42.707 90.0, 90.0, 90.0
Space group	P2 <sub>1</sub> 2 <sub>1</sub> 2 <sub>1</sub>	P2 <sub>1</sub> 2 <sub>1</sub> 2 <sub>1</sub>
Resolution (Å)	21.17 – 1.50 (1.55 – 1.50)	18.75 – 1.55 (1.61 – 1.55)
No. of uniq. refl./observed refl.	849 / 3811	786 / 4422
$\langle I / \sigma(I) \rangle$	8.4 (2.4)	5.4 (1.5)
$R_{\text{meas}}$	0.120 (0.432)	0.185 (0.635)
Completeness (%)	100.0 (100.0)	100.0 (100.0)
CC( $\frac{1}{2}$ )	0.995 (0.864)	0.992 (0.880)
Refinement		
Resolution range (Å)	21.16– 1.50	18.75 – 1.55
$R / R_{\text{free}}$ (No. of obs.)	0.1004 (762) / 0.1097 (87)	0.1358 (704) / 0.1543 (78)
No. of non-hydrogen atoms: peptide / solvent	58 / -	63 / 1
B-factor of peptide / solvent (Å <sup>2</sup> )	6.53 / -	9.68 / 9.72
RMS dev. bond length (Å)	0.012	0.018
RMS dev. bond angles (°)	1.387	1.915
Ramachandran fav. / all. / disall.	4 / 0 / 0	4 / 0 / 0

sequences. For a single peptide query, this sequence-sequence distance matrix has been constructed for all downloaded pdb files, which were then subsequently clustered using agglomerative clustering with average linkage and a threshold value of 0.9. In the follow-up investigations the cluster medoids have been used as cluster representatives. The pdb files filtered this way underwent dihedral angle and solvent accessible surface (SASA) analysis using the BioPython package (version 1.84.dev0). Visualizations have been done using the Matplotlib package (version 3.8.2).

## Acknowledgements

The authors thank Professors Tibor Vellai and Eörs Szathmáry for their valuable advice. This work was funded by the ELTE Thematic Excellence Program (SzintPlus), the Hungarian Ministry for Innovation and Technology, and Project no. 2018-1.2.1-NKP-2018-00005 (HunProtExc) from the National Research, Development, and Innovation Fund of Hungary. Project RRF-2.3.1-21-2022-00015 was supported by the EU's Recovery and Resilience Faci (PharmaLab). The crystallographic study received funding from project No. VEKOP-2.3.3-15-2017-00018, co-financed by the EU and the State of Hungary. The work was also sponsored by the Gedeon Richter Talentum Foundation, within the framework of the Gedeon Richter Excellence PhD Scholarship. F. B. was supported by Gedeon Richter.

## Conflict of Interests

The authors declare no conflict of interest.

## Data Availability Statement

The data that support the findings of this study are available in the supporting information of this article from the corresponding author upon reasonable request.

**Keywords:** AFM of amyloidogenic oligopeptides · aggregation-prone regions · amyloid X-ray structure · ECD- and FTIR-spectroscopy · molecular evolution

- [1] S. L. Miller, *Science* **1953**, 117, 528.
- [2] S. A. Kauffman, D. P. Jelenfi, G. Vattay, *J. Theor. Biol.* **2020**, 486, 110097.
- [3] L. Leman, L. Orgel, M. R. Ghadiri, *Science* **2004**, 306, 283.
- [4] A. Shimoyama, R. Ogasawara, *Origins of Life and Evolution of the Biosphere*, Kluwer Academic Publishers, The Netherlands, **2002**.
- [5] D. P. Glavin, J. P. Dworkin, C. M. O. Alexander, J. C. Aponte, A. A. Baczynski, J. J. Barnes, H. A. Bechtel, E. L. Berger, A. S. Burton, P. Caselli, A. H. Chung, S. J. Clemett, G. D. Cody, G. Dominguez, J. E. Elsila, K. K. Farnsworth, D. I. Foustoukos, K. H. Freeman, Y. Furukawa, Z. Gainsforth, H. V. Graham, T. Grassi, B. M. Giuliano, V. E. Hamilton, P. Haenecour, P. R. Heck, A. E. Hofmann, C. H. House, Y. Huang, H. H. Kaplan, et al., *Nat Astron* **2025**, 9, 199.
- [6] J. Beerten, J. Van Durme, R. Gallardo, E. Capriotti, L. Serpell, F. Rousseau, J. Schymkowitz, *Bioinformatics* **2015**, 31, 1698.
- [7] M. Sulyok-Eiler, V. Harmat, A. Perczel, *J. Chem. Inf. Model.* **2024**, 64, 8628.
- [8] T. Langenberg, R. Gallardo, R. van der Kant, N. Louros, E. Michiels, R. Duran-Romaña, B. Houben, R. Cassio, H. Wilkinson, T. Garcia, C. Ulens, J. Van Durme, F. Rousseau, J. Schymkowitz, *Cell Rep.* **2020**, 31, 107512.
- [9] J. Greenwald, M. P. Friedmann, R. Riek, *Angew. Chem.* **2016**, 128, 11781.
- [10] T. Dale, *J. Theor. Biol.* **2006**, 240, 337.
- [11] A.-M. Fernandez-Escamilla, F. Rousseau, J. Schymkowitz, L. Serrano, *Nat. Biotechnol.* **2004**, 22, 1302.
- [12] D. S. Eisenberg, M. R. Sawaya, *Annu. Rev. Biochem.* **2017**, 86, 69.
- [13] E. Monsellier, F. Chiti, *EMBO Rep.* **2007**, 8, 737.
- [14] A. Perczel, P. Hudáky, V. K. Pálfi, *J. Am. Chem. Soc.* **2007**, 129, 14959.
- [15] A. Perczel, Z. Gáspári, I. G. Csizmadia, *J. Comput. Chem.* **2005**, 26, 1155.
- [16] D. Sharma, G. Feng, D. Khor, G. Z. Genchev, H. Lu, H. Li, *Biophys. J.* **2008**, 95, 3935.
- [17] N. Taricska, D. Horváth, D. K. Menyhárd, H. Ákontz-Kiss, M. Noji, M. So, Y. Goto, T. Fujiwara, A. Perczel, *Chem. Eur. J.* **2020**, 26, 1968.

- [18] B. Houben, F. Rousseau, J. Schymkowitz, *Trends Biochem. Sci.* **2022**, 47, 194.
- [19] M. Landreh, M. R. Sawaya, M. S. Hipp, D. S. Eisenberg, K. Wüthrich, F. U. Hartl, *J. Intern. Med.* **2016**, 280, 164.
- [20] C. Parrini, N. Taddei, M. Ramazzotti, D. Degl'Innocenti, G. Ramponi, C. M. Dobson, F. Chiti, *Structure* **2005**, 13, 1143.
- [21] J. S. Richardson, D. C. Richardson, in *Prediction of Protein Structure and the Principles of Protein Conformation* (Ed.: G. D. Fasman), Springer US, Boston, MA, **1989**, pp. 1–98.
- [22] A. Fulara, A. Lakhani, S. Wójcik, H. Nieznańska, T. A. Keiderling, W. Dzwolak, *J. Phys. Chem. B* **2011**, 115, 11010.
- [23] F. Chiti, C. M. Dobson, *Annu. Rev. Biochem.* **2006**, 75, 333.
- [24] A. A. H. Zanjani, N. P. Reynolds, A. Zhang, T. Schilling, R. Mezzenga, J. T. Berryman, *Sci. Rep.* **2019**, 9, 15987.
- [25] I. M. Stanković, S. Niu, M. B. Hall, S. D. Zarić, *Int. J. Biol. Macromol.* **2020**, 156, 949.
- [26] M. Biancalana, K. Makabe, A. Koide, S. Koide, *J. Mol. Biol.* **2008**, 383, 205.
- [27] E. Gazit, *FASEB J.* **2002**, 16, 77.
- [28] D. Horváth, Z. Dürvanger, D. K. Menyhárd, M. Sulyok-Eiler, F. Bencs, G. Gyulai, P. Horváth, N. Taricska, A. Perczel, *Nat. Commun.* **2023**, 14, 4621.
- [29] A. Perczel, M. Hollósi, G. Tusnády, G. D. Fasman, *Protein Eng. Des. Sel.* **1991**, 4, 669.
- [30] G. D. Fasman, Ed., *Circular Dichroism and the Conformational Analysis of Biomolecules*, Springer US, Boston, MA, **1996**.
- [31] R. W. Woody, *The Peptides*, Academic Press, **1985**.
- [32] N. Sreerama, *Protein Sci.* **2003**, 12, 384.
- [33] R. Nelson, M. R. Sawaya, M. Balbirnie, A. Ø. Madsen, C. Riek, R. Grothe, D. Eisenberg, *Nature* **2005**, 435, 773.
- [34] D. R. Lyke, J. E. Dorweiler, A. L. Manogaran, *Yeast* **2019**, 36, 465.
- [35] M. R. Sawaya, S. Sambashivan, R. Nelson, M. I. Ivanova, S. A. Sievers, M. I. Apostol, M. J. Thompson, M. Balbirnie, J. J. W. Wiltzius, H. T. McFarlane, A. Ø. Madsen, C. Riek, D. Eisenberg, *Nature* **2007**, 447, 453.
- [36] M. Balbirnie, R. Grothe, D. S. Eisenberg, *Proc. Natl. Acad. Sci. USA* **2001**, 98, 2375.
- [37] GRAVY CALCULATOR. (accessed: March 2025). <https://www.gravy-calculator.de/>
- [38] A. V. Glyukina, N. S. Bogatyreva, O. V. Galzitskaya, *PLoS One* **2011**, 6, e28464.
- [39] S. Tzotzos, A. J. Doig, *Protein Sci.* **2010**, 19, 327.
- [40] Complete Sequence of Thermosipho Melanesiensis BI429. (accessed: March 2025). <https://www.uniprot.org/uniprotkb/A6LKW9/entry>
- [41] Genetic Interactor of Prohibitin 7, Mitochondrial. (accessed: March 2025). <https://www.uniprot.org/uniprotkb/C5DT30/entry>
- [42] R. Prabakaran, P. Rawat, S. Kumar, M. Michael Gromiha, *J. Mol. Biol.* **2021**, 433, 166707.
- [43] Aggregation Nucleation Prediction in Peptides and Proteins. (accessed: March 2025). <https://web.iitm.ac.in/bioinfo2/ANuPP/homeseq1/>
- [44] A. S. Reddy, M. Chopra, J. J. De Pablo, *Biophys. J.* **2010**, 98, 1038.
- [45] H. Konno, T. Watanabe-Nakayama, T. Uchihashi, M. Okuda, L. Zhu, N. Kodera, Y. Kikuchi, T. Ando, H. Taguchi, *Proc. Natl. Acad. Sci. U.S.A.* **2020**, 117, 7831.
- [46] N. Lester-Zer, M. Ghayeb, L. Chai, *Proc. Natl. Acad. Sci. U.S.A.* **2019**, 116, 22478.
- [47] J. Hu, H. Sun, H. Hao, Q. Zheng, *J. Pharmaceut. Anal.* **2020**, 10, 194.
- [48] Z. Dürvanger, F. Bencs, D. K. Menyhárd, D. Horváth, A. Perczel, *Commun. Biol.* **2024**, 7, 968.
- [49] J. Nochebuena, J. Ireta, *J. Chem. Phys.* **2015**, 143, 135103.
- [50] F. Bencs, L. Románszki, V. Farkas, A. Perczel, *Chem. Eur. J.* **2025**, e202404255.
- [51] G. Akerlof, *J. Am. Chem. Soc.* **1932**, 54, 4125.
- [52] C. Bleiholder, N. F. Dupuis, T. Wyttenbach, M. T. Bowers, *Nature Chem* **2011**, 3, 172.
- [53] P. Rawat, R. Prabakaran, R. Sakthivel, A. Mary Thangakani, S. Kumar, M. M. Gromiha, *Amyloid* **2020**, 27, 128.
- [54] P. P. Wozniak, M. Kotulska, *Bioinformatics* **2015**, 31, 3395.
- [55] V. Farkas, K. Ferentzi, K. Horváti, A. Perczel, *Org. Process Res. Dev.* **2021**, 25, 182.
- [56] K. Ferentzi, D. Nagy-Fazekas, V. Farkas, A. Perczel, *React. Chem. Eng.* **2024**, 9, 58.
- [57] S. C. Gill, P. H. Von Hippel, *Anal. Biochem.* **1989**, 182, 319.
- [58] D. Nečas, P. Klapetek, *Open Physics* **2012**, 10, 181.
- [59] D. Liebschner, P. V. Afonine, M. L. Baker, G. Bunkóczi, V. B. Chen, T. I. Croll, B. Hintze, L.-W. Hung, S. Jain, A. J. McCoy, N. W. Moriarty, R. D. Oeffner, B. K. Poon, M. G. Prisant, R. J. Read, J. S. Richardson, D. C. Richardson, M. D. Sammito, O. V. Sobolev, D. H. Stockwell, T. C. Terwilliger, A. G. Urzhumtsev, L. L. Videau, C. J. Williams, P. D. Adams, *Acta Crystallogr D Struct Biol* **2019**, 75, 861.
- [60] P. Emsley, B. Lohkamp, W. G. Scott, K. Cowtan, *Acta Crystallogr D Biol Crystallogr* **2010**, 66, 486.
- [61] P. V. Afonine, R. W. Grosse-Kunstleve, N. Echols, J. J. Headd, N. W. Moriarty, M. Mustyakimov, T. C. Terwilliger, A. Urzhumtsev, P. H. Zwart, P. D. Adams, *Acta Crystallogr D Biol Crystallogr* **2012**, 68, 352.
- [62] O. S. Smart, T. O. Womack, C. Flensburg, P. Keller, W. Paciorek, A. Sharff, C. Vonnrhein, G. Bricogne, *Acta Crystallogr D Biol Crystallogr* **2012**, 68, 368.

Manuscript received: December 19, 2024  
Revised manuscript received: March 3, 2025  
Version of record online: May 2, 2025