

Decoding Solubility Signatures from Amyloid Monomer Energy Landscapes

Published as part of *Journal of Chemical Theory and Computation* special issue "Machine Learning and Statistical Mechanics: Shared Synergies for Next Generation of Chemical Theory and Computation".

Patryk Adam Wesolowski,* Bojun Yang, Anthony J. Davolio, Esmee J. Woods, Philipp Pracht, Krzysztof K. Bojarski, Krzysztof Wierbilowicz, Mike C. Payne, and David J. Wales*



Cite This: *J. Chem. Theory Comput.* 2025, 21, 2736–2756



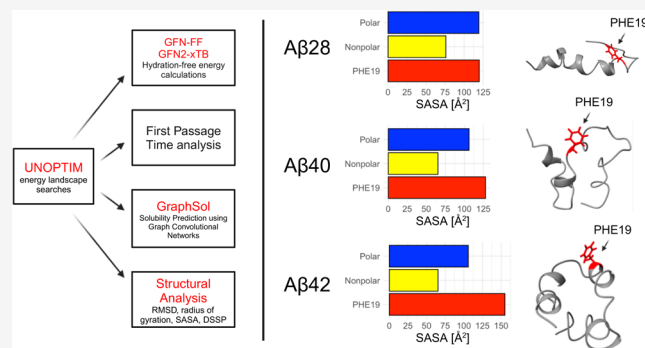
Read Online

ACCESS |

Metrics & More

Article Recommendations

ABSTRACT: This study investigates the energy landscapes of amyloid monomers, which are crucial for understanding protein misfolding mechanisms in Alzheimer's disease. While proteins possess inherent thermodynamic stability, environmental factors can induce deviations from native folding pathways, leading to misfolding and aggregation, phenomena closely linked to solubility. Using the UNOPTIM program, which integrates the UNRES potential into the Cambridge energy landscape framework, we conducted single-ended transition state searches and employed discrete path sampling to compute kinetic transition networks starting from PDB structures. These kinetic transition networks consist of local energy minima and the transition states that connect them, which quantify the energy landscapes of the amyloid monomers. We defined clusters within each landscape using energy thresholds and selected their lowest-energy structures for the structural analysis. Applying graph convolutional networks, we identified solubility trends and correlated them with structural features. Our findings identify specific minima with low solubility, characteristic of aggregation-prone states, highlighting the key residues that drive reduced solubility. Notably, the exposure of the hydrophobic residue Phe19 to the solvent triggers a structural collapse by disrupting the neighboring helix. Additionally, we investigated selected minima to determine the first passage times between states, thereby elucidating the kinetics of these energy landscapes. This comprehensive approach provides valuable insights into the thermodynamics and kinetics of A β monomers. By integration of multiple analytical techniques to explore the energy landscapes, our study investigates structural features associated with reduced solubility. These insights have the potential to inform future therapeutic strategies aimed at addressing protein misfolding and aggregation in neurodegenerative diseases.



1. INTRODUCTION

Proteins are among the most abundant biomolecules in the human body and intricately govern a wide range of physiological functions. Their precise folding into a compact conformation, termed the native state, is essential for proper functioning. This folding process, corresponding to a descent down an energy funnel, has been conceptualized by Leopold et al.¹ The structural transformations that occur during protein folding result in concurrent decreases in the enthalpy and internal entropy. At or below optimal folding temperatures, the enthalpy change (ΔH) outweighs the entropy term ($T\Delta S$) leading to a thermodynamically favorable process characterized by a reduction in Gibbs free energy, resulting in the acquisition of the native state.^{1,2} Despite the inherent thermodynamic stability of protein folding, specific environmental conditions may divert proteins from their native state, causing trapping in local energy

minima.^{3,4} These external factors include, among others, temperature, pH, and molecular crowding. Internal factors, such as mutations or chemical modifications including oxidation or glycation, can also affect folding.⁵ These diversions from native folding may involve energy barriers, which impede proteins from regaining their native conformation.² Consequently, misfolding events can lead to the formation of aggregates (Figure 1).⁶

Received: November 28, 2024

Revised: February 17, 2025

Accepted: February 18, 2025

Published: February 24, 2025



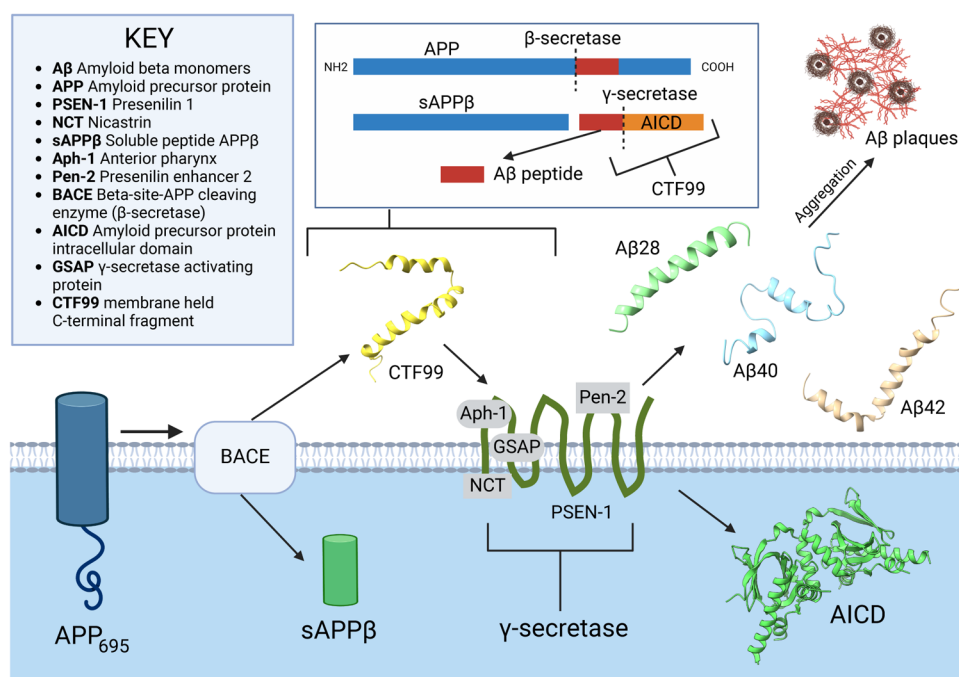


Figure 1. Cleavage of APP695 residue protein first by β -secretase to form CTF-99 (PDB ID: 2LP1), then by γ -secretase subsequently to form A β monomers of A β 40 (PDB ID: 1AML) and A β 42 (PDB ID: 1IYT) residues. A β 28 (PDB ID: 1AMC) is a monomer consisting of the same residues as the first 1–28 residues from the full A β 42 peptide, and it has been extensively explored in previous work.⁷ This sequence is followed by the AICD domain (PDB ID: 3DXC) at the C-terminal end. The figure was created with BioRender.com.

The formation of protein aggregates, commonly referred to as amyloids, is important in various scientific disciplines, including chemistry, biology, and medicine.⁸ Amyloid β (A β) monomers are not inherently problematic and may serve functional purposes,⁹ such as protecting mature neurons from excitotoxic death.¹⁰ However, amyloid formation disorders are a great concern for the aging human population, with Alzheimer's disease (AD) serving as a prominent example. AD is a progressive brain disorder that affects memory, thinking skills, and the ability to perform daily tasks. The median survival of AD is 5.8 years, with approximately 55 million people affected worldwide in 2019.¹¹ AD is characterized by the presence of plaques in the brain primarily composed of A β proteins and Tau fibrils, collectively termed "A β -plaques" (Figure 1).¹² Maintaining the delicate balance of protein damage levels is ordinarily regulated by adjusting chaperone and protease levels.⁸ However, when the generation of misfolded proteins surpasses the rate at which they are destroyed or refolded, aggregates can accumulate.^{13–15} In AD, A β peptides result from the cleaving of the amyloid precursor protein (APP), a Type I integral membrane glycoprotein composed of 695–770 residues¹⁶ with diverse metabolic functions.¹²

APP695, the predominantly expressed protein found in the human brain at significantly higher levels,^{17,18} is sequentially cleaved by β -secretase (e.g., BACE1) and γ -secretase (a multiunit integral membrane protease¹⁶) through proteolytic processing, yielding a range of A β proteins that form the key components of amyloid fibrils in AD patients.¹⁹ A β peptides are secreted outside the cell membrane, aggregating to form neurotoxic plaques (Figure 1).²⁰ In familial Alzheimer's disease (FAD), an inherent mutation in APP or presenilin (a component of the γ -secretase complex) causes increased cleavage-producing A β monomers.²⁰ In sporadic diseases, the buildup of A β monomers may be attributed to a decrease in A β

degrading enzymes.²¹ The primary amino acid sequence of A β 40 and A β 42 (Figure 1), elucidated in 1984,²² provided insights into the accumulation mechanisms. These proteins, which belong to the class of intrinsically disordered proteins,²³ undergo various structural transitions, and their aggregation into forms such as oligomers, protofibrils, and amyloid fibrils contributes to the pathology of AD.¹² A β 28 is composed of the first 28 residues of the extracellular domain²⁴ and lacks the anchoring transmembrane domain compared to the more abundant A β 40 and A β 42 monomers^{25,26} and thus exhibits significant structural differences. Its fibrillogenic properties in terms of plaque formation are also explored and compared in this paper as a benchmark system.

After cleavage from APP695, monomeric forms of A β 40 and A β 42 have been proposed to adopt unstructured conformations,²⁷ which then normally change to an α -helical structure upon binding with a negatively charged biological lipid.²⁷ Both the random coil and the α -helical structures are relatively soluble in solution. The multifunnel energy landscape^{23,28} describes multiple competing folding pathways, where A β peptides can settle into less soluble, metastable states, often forming partially folded intermediates resembling native-like structures.²⁹ Specifically, this process includes a rearrangement of a soluble random coil and α -helical structures into insoluble β -sheet aggregated structures that are fibrillar.²⁷ Proteins do not spontaneously transform from a state of lower energy to higher energy. Thus, they aggregate only when the amyloid state has lower free energy, making the process thermodynamically favorable.³⁰ For example, when A β peptides are present at high concentrations, they become highly unstable thermodynamically, which causes them to undergo aggregation. This process ultimately results in the formation of insoluble plaques.³¹

Investigation into the solubility of A β monomers is essential for the diagnosis of patients with Alzheimer's disease, since both

the soluble and insoluble forms of the A β peptide affect the brain. For instance, measuring only the concentration of insoluble A β peptides cannot distinguish between patients with high pathology control and those with Alzheimer's.³² In contrast, the concentration of soluble A β peptides, particularly A β 40, shows a clear inverse correlation with synapse loss, providing a more effective marker for differentiation.³² While soluble A β is better at predicting change in synapses of Alzheimer's patients,³² insoluble A β forms neurotoxic plaques and aggregates. Amyloid fibrils form because of a decrease in the solubility of the monomer, controlled by supersaturation. Thus, lower solubility and a lower saturation barrier for a particular conformation of these amyloids may imply higher, or earlier, contribution to plaque formation.^{33–35}

In a preceding study, we successfully integrated the UNRES (UNited RESidue) coarse-grained potential^{36,37} into the OPTIM³⁸ program, resulting in the development of the UNOPTIM program.³⁹ In this study, we employ the UNOPTIM³⁹ program alongside the discrete path sampling (DPS) approach^{40,41} to construct the kinetic transition network.^{42–44} Through this approach, we analyze the potential energy landscapes of three A β monomer proteins: A β 28 (PDB ID: 1AMC), A β 40 (PDB ID: 1AML), and A β 42 (PDB ID: 1IYT). We identified clusters of UNRES energy minima by applying specific energy thresholds, ensuring a comparable number of structures for analysis across all systems. Using specific clustering energy thresholds, the clusters are defined as the sets of local minima that can interconvert without exceeding that threshold value above their lowest member. The lowest-energy structures within these clusters were further examined and are referenced throughout the text as selected minima. Using Graph Convolutional Networks (GCN), we identified solubility trends facilitated by a detailed examination of the energy landscape of amyloid monomers. Additionally, we analyzed transitions between selected minima by employing first passage time (FPT) distributions to investigate the kinetics of these landscapes. By integration of multiple analytical techniques, this study elucidates structural features associated with reduced solubility, potentially contributing to aggregation.

II. THEORY AND TECHNICAL DETAILS

Here we describe the methodology used in this study. The flowchart in Figure 2 summarizes the flow of data and information between different levels of calculation. All components of the flowchart are further detailed within this section.

II.A. UNOPTIM Program. The interface between the UNRES^{36,37} coarse-grained potential and the OPTIM program³⁸ allows fast calculations with useful accuracy for exploring protein energy landscapes.³⁹ The UNRES model was chosen because it remains under constant development, with new extensions added frequently.^{45,46} These refinements help to improve accuracy in simulations of various protein structures with extension to large systems.^{47,48} Furthermore, the structural prediction capabilities of the UNRES model are tested every two years^{49,50} during the Critical Assessment of protein Structure Prediction (CASP) experiments. The model employs two interaction sites per amino acid, which are the united side chains (SCs) and united peptide groups (p), respectively (Figure 3).³⁶ The α -carbons (C^α) are not interaction sites but are instead used to define the geometry of the protein main chain. The united peptide groups are found halfway between two consecutive C^α atoms⁵¹ (Figure 3). Implicit solvent is used and is computed

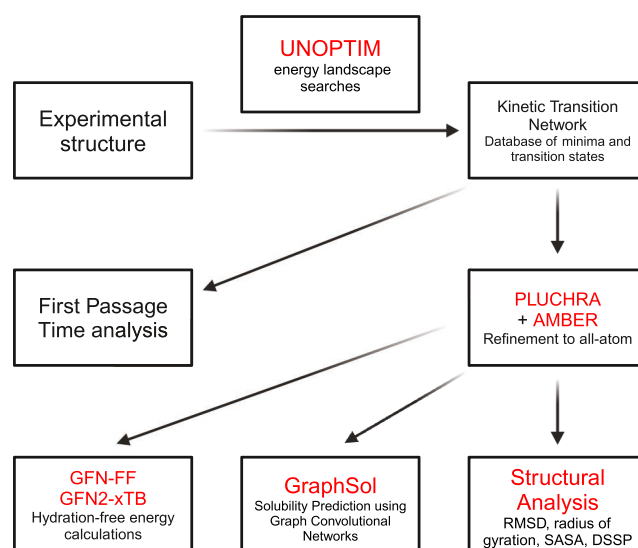


Figure 2. Flowchart illustrating the organization of the calculations used in this work, with the programs highlighted in red. All aspects of the methodology are further detailed in this section.

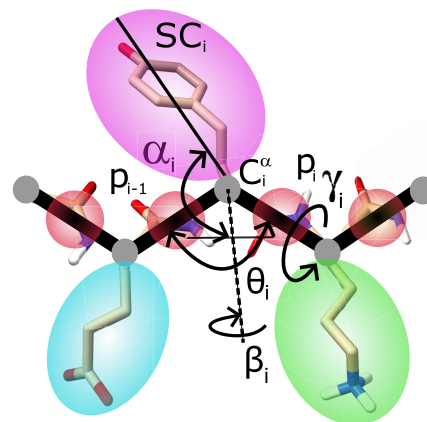


Figure 3. Representation of polypeptide chains with the UNRES model. United side chains are depicted as colored ellipsoids, while united peptide groups appear as red spheres. The specific angles referenced in the image are detailed in the text. The illustration is adapted from ref 51.

using interaction potentials involving SCs, coefficients of friction, and stochastic forces.⁵²

The UNRES energy function is

$$\begin{aligned}
 U = & w_{SC} \sum_{i < j} U_{SC_i SC_j} + w_{SCp} \sum_{i \neq j} U_{SC_i p_j} + w_{pp}^{VDW} \sum_{i < j-1} U_{p_i p_j}^{VDW} \\
 & + w_{ppf_2}^{el} \sum_{i < j-1} U_{p_i p_j}^{el} + w_{torf_2} \sum_i U_{tor}(\gamma_i, \theta_i, \theta_{i+1}) \\
 & + w_b \sum_i U_b(\theta_i) + w_{rot} \sum_i U_{rot}(\theta_i, \alpha_{SC_i}, \beta_{SC_i}) \\
 & + w_{bond} \sum_i U_{bond}(d_i) + w_{ssbond} \sum_{n_{ss}} U_{ssbond}(d_{ss}) \\
 & + w_{corr_3}^{(3)f}(T) U_{corr}^{(3)} + w_{turn_3}^{(3)f}(T) U_{turn}^{(3)} \quad (1)
 \end{aligned}$$

The formula is composed of long-range intersite terms ($U_{SC_i SC_j}$, $U_{SC_i p_j}$, $U_{p_i p_j}$), and local terms. Local terms include the virtual-bond-deformation (U_{bond}), virtual-bond-angle terms

(U_b), side-chain rotamer terms (U_{rot}), and third-order multi-body terms ($U_{corr}^{(3)}$, $U_{turn}^{(3)}$). These terms account for coupling between backbone-local and backbone-electrostatic interactions.³⁷ Virtual-bond vectors are expressed as $C^\alpha \cdots C^\alpha$ and $C^\alpha \cdots SC$. The polypeptide chain backbone geometry is defined by the virtual-bond angles $C^\alpha \cdots C^\alpha \cdots C^\alpha$ and virtual-bond-dihedral angles $C^\alpha \cdots C^\alpha \cdots C^\alpha \cdots C^\alpha$, referred to as θ and γ , respectively. The virtual-bond angle and virtual-bond-dihedral angle for the i^{th} C^α (C_i^α) are represented as θ_i and γ_i . The α_{SC} and β_{SC} angles define the orientation of an SC center with respect to the backbone. Therefore, the SC_i center local geometry can be defined by the spherical angles α_β , referred to as the angle between bisection of θ_i and the $C_i^\alpha \cdots SC_i$ vector, and the angle of rotation of the $C_i^\alpha \cdots SC_i$ vector from the $C_{i-1}^\alpha \cdots C_i^\alpha \cdots C_{i+1}^\alpha$ plane, described as β_i (Figure 3).⁵¹

Each energy term is multiplied by an appropriate weight w_x ,⁵³ and the values corresponding to factors greater than order one are additionally multiplied by temperature coefficients.⁵⁴ These coefficients reflect the influence of the first generalized-cumulant term⁵⁵ on temperature, and are defined by $f_n(T)$, where $T_0 = 300$ K:^{54,56}

$$f_n(T) = \frac{\ln[\exp(1) + \exp(-1)]}{\ln\{\exp[(T/T_0)^{n-1}] + \exp[-(T/T_0)^{n-1}]\}} \quad (2)$$

II.B. Energy Landscape Exploration and the Discrete Path Sampling Approach. The molecular energy landscape defines thermodynamic and kinetic properties. Here, we employ discrete path sampling (DPS)^{40,41} to explore the landscape using geometry optimization to locate minima of the UNRES energy function and the transition states that connect them.⁵⁷ To initiate DPS, two minima must be first linked by one discrete path, which usually encounters intervening minima.⁴¹ The set of minima and transition states constitutes a kinetic transition network,^{58–60} which allows us to estimate the thermodynamic and dynamic properties of $\alpha\beta$ monomers.

The OPTIM program uses the Limited-memory Broyden-Fletcher-Goldfarb-Shanno minimization algorithm (LBFGS algorithm)⁶¹ for minimization. Transition states are refined using hybrid eigenvector-following^{62–66} from candidates found by the doubly nudged^{67,68} elastic band method^{69–72} (DNEB). A transition state is defined as a stationary point where the Hessian matrix has exactly one negative eigenvalue.⁷³ To find candidate transition states using the DNEB method, two minima are interpolated by images connected by harmonic springs in an elastic band, and the energy of the bands is minimized by the LBFGS algorithm.^{67,68}

The energy landscapes are visualized using disconnectivity graphs,^{74,75} which represent local minima and the energy barriers between them and illustrate the organization of the landscape.^{1,76} Disconnectivity graphs simplify the multidimensional character of the energy landscape and provide insight into how molecular properties are encoded.³⁹

We identified clusters of energy minima by applying specific energy thresholds, ensuring a comparable number of structures for analysis across all systems. The lowest-energy structures within these clusters were further examined and are referenced throughout the text as selected minima. These minima were reconstructed into all-atom structures using the Protein Chain Reconstruction Algorithm (PULCHRA),⁷⁷ designed to convert coarse-grained protein models into all-atom representations. PULCHRA reconstructs the backbone atoms based on geometric principles that ensure realistic bond angles and

dihedral distributions, while side-chain atoms are added using a rotamer library optimized for accuracy and efficiency. The algorithm also applies energy-based adjustments to minimize steric clashes, ensuring that the final structures are physically plausible and suitable for downstream computational analyses.

Additionally, all reconstructed minima, represented by an AMBER potential, were solvated in a TIP3P octahedral periodic box with a layer of water molecules of 6 Å from the border of the periodic box to the solute and neutralized with counterions (3 Na⁺ ions as each system had a net charge of -3). Energy minimization was carried out in two steps: first we performed 0.5×10^3 steepest descent cycles and 10^3 conjugate gradient cycles with harmonic force restraints of 100 kcal/(mol Å²) on the solute atoms, and then we used 3×10^3 steepest-descent cycles and 3×10^3 conjugate gradient cycles without restraints. Afterward, the system was simulated at 300 K for 10 ps with harmonic force restraints of 100 kcal/(mol Å²) on solute atoms and equilibrated for 100 ps at 300 K and 10^5 Pa in the isothermal isobaric ensemble (NPT). This minimization step was performed using the AMBER ff14SB force field.⁷⁸

For the reconstructed structures, we performed conformational analysis, taking into account properties such as the Root Mean Squared Deviation (RMSD) from a reference structure, Root Mean Square Fluctuations (RMSF), the radius of gyration, solvent-accessible surface area (SASA), and secondary structure properties as well as the contact maps. RMSD calculations were performed using Biopython⁷⁹ after the alignment of all structures. RMSF values were calculated using Biopython by aligning each structure to a reference structure based on main chain atoms and carbon atoms in side chains and then computing the positional deviations across all aligned structures for each residue. Radii of gyration were calculated using the “rgyr” cpptraj module from the AMBER package. Distances for contact maps were determined with the application of the “distance” cpptraj module. For each analyzed group of structures corresponding to minima derived from the energy landscape, we compute the mean and standard deviations of these distances. SASAs were calculated using the “FreeSASA” library.⁸⁰ To define secondary structure elements for residues we employed the Define Secondary Structure of Proteins (DSSP) algorithm within the “mkdssp” software.⁸¹ The structures corresponding to selected minima were visualized using the UCSF ChimeraX software.⁸²

II.C. First Passage Time Analysis. In order to analyze the dynamics of the energy landscape, we must first compute the transfer rates between minima directly connected by a transition state. In the harmonic approximation to transition state theory the transfer rate from state j to state i is given by,^{83–87}

$$K_{ij}(T) = \frac{(\bar{\nu}_j)^\kappa}{(\bar{\nu}^\dagger)^{\kappa-1}} e^{-(V^\dagger - V_j)/k_B T} \quad (3)$$

where the subscripts correspond to minima indexes and \dagger refers to a transition state; V is the potential energy; $\bar{\nu}$ is the geometric mean of the normal-mode frequencies; and κ is the number of vibrational degrees of freedom. k_B is Boltzmann’s constant and T is temperature. We do not use the normal-mode frequencies within the UNRES potential because it is a coarse-grained formulation. Vibrational analysis is possible within this framework,^{88,89} but for the estimate in the present survey we simply use $\bar{\nu}_j = \bar{\nu}^\dagger = \nu_{av}$ for all minima and transition states, where ν_{av} is an average frequency factor. It is then convenient to consider the time in units of $1/\nu_{av}$.

The master equation describing the time evolution of the system under the assumption of Markovian dynamics between the directly connected local minima is^{90–92}

$$\frac{d\mathbf{P}(t)}{dt} = \mathbf{Q}\mathbf{P}(t) \quad (4)$$

where $\mathbf{P}(t)$ is the time-dependent vector of occupation probabilities for the local minima. The transition matrix is $\mathbf{Q} = \mathbf{K} - \mathbf{D}$, where \mathbf{K} is the rate matrix containing elements K_{ij} and \mathbf{D} is a diagonal matrix of escape rates, with elements $D_{jj} = \sum_i K_{ji}$.

We focus on the computation of first passage times (FPT), which are defined as the time taken to first reach the sink state (i.e., the product), from a given initial starting state (i.e., the reactant). As we are interested only in the first hitting time, no probability can escape the sink, and the corresponding escape rates are set to zero. We can then work within the reduced state space \mathcal{S} , where $\Omega = \mathcal{A} \cup \mathcal{S}$, Ω is the full state space, and \mathcal{A} is the state space of the sink. $\mathbf{Q}_\mathcal{S} = \mathbf{K}_\mathcal{S} - \mathbf{D}_\mathcal{S}$ is the subset of the full transition matrix \mathbf{Q} containing the interstate transition rates within \mathcal{S} . $\mathbf{D}_\mathcal{S}$ is the corresponding subset of \mathbf{D} including the escape rates to \mathcal{A} . The master equation describing the modified dynamics becomes

$$\frac{d\mathbf{P}_\mathcal{S}(t)}{dt} = \mathbf{Q}_\mathcal{S}\mathbf{P}_\mathcal{S}(t) \quad (5)$$

where $\mathbf{P}_\mathcal{S}(t)$ is the occupation probability vector for all minima within \mathcal{S} .

To compute FPT distributions, we first decompose the transition matrix into its constituent eigenmodes,

$$\mathbf{Q}_\mathcal{S} = -\sum_l^{\text{ISI}} \lambda_l \mathbf{w}_l^R \otimes \mathbf{w}_l^L \quad (6)$$

where \mathbf{w}_l^L and \mathbf{w}_l^R are the left and right eigenvectors and \otimes is the outer product. \mathbf{w}_l^L is a row vector, and \mathbf{w}_l^R is a column vector. All eigenvalues are real and negative, $-\lambda_l < 0$. Using the above decomposition and the master equation, we can also write the FPT distribution as a sum over eigenmodes,

$$p(t) = \sum_{l=1}^{\text{ISI}} \lambda_l e^{-\lambda_l t} \mathbf{1}_\mathcal{S}(\mathbf{w}_l^R \otimes \mathbf{w}_l^L) \mathbf{P}_\mathcal{S}(0) \equiv \sum_l^{\text{ISI}} \lambda_l e^{-\lambda_l t} A_l \quad (7)$$

which defines the amplitude of eigenmode l , $A_l = \mathbf{1}_\mathcal{S}(\mathbf{w}_l^R \otimes \mathbf{w}_l^L) \mathbf{P}_\mathcal{S}(0) \equiv [\mathbf{1}_\mathcal{S} \mathbf{w}_l^R][\mathbf{w}_l^L \mathbf{P}_\mathcal{S}(0)]$, a product of dot products. Here, $\mathbf{1}_\mathcal{S}$ is a row vector of ones and $\mathbf{P}_\mathcal{S}(0)$ is the initial occupation probability in \mathcal{S} , at $t = 0$. As rates are exponentially sensitive to energy barriers, we make the transformation $y = \ln t$, to produce the probability distribution $\mathcal{P}(y)$,

$$\begin{aligned} \mathcal{P}(y) &= \sum_{l=1}^{\text{ISI}} \lambda_l e^{y - \lambda_l e^y} \mathbf{1}_\mathcal{S}(\mathbf{w}_l^R \otimes \mathbf{w}_l^L) \mathbf{P}_\mathcal{S}(0) \\ &\equiv \sum_l^{\text{ISI}} \lambda_l e^{y - \lambda_l \exp(y)} A_l \end{aligned} \quad (8)$$

As $p(t)$ and $\mathcal{P}(y)$ are normalized distributions, $\sum_l A_l = 1$.

Each individual eigenmode makes a peak contribution of $\lambda_l e^{y - \lambda_l \exp(y)} A_l$ to the FPT distribution, which has a peak height of A_l/e at $y = -\ln \lambda_l$. However, as many amplitudes are small in

magnitude, there are normally only a few dominant modes that significantly contribute to the FPT distribution. Additionally, we use the right eigenvector components of each eigenmode to determine the relative contribution of each state to the time scale of each individual eigenmode peak. This approach enables us to assign peaks in FPT distributions to sets of minima in the landscape.⁹³

In practice, performing eigendecomposition of $\mathbf{Q}_\mathcal{S}$ can be challenging due to computational limits and especially numerical precision. For the energy landscapes considered in this manuscript, many FPT distributions can be computed using standard eigendecomposition procedures within LAPACK for the full transition matrix $\mathbf{Q}_\mathcal{S}$ at $T = 300$ K. However, this method breaks down for some transitions, particularly for slow transitions between minima at the bottom of different funnels that constitute kinetic traps. The eigenvalues corresponding to the long time peaks are then many orders of magnitude smaller than other eigenmodes within the system, which results in a separation of time scales and loss of precision in eigendecomposition. This problem causes long time behavior to be computed incorrectly. To overcome this limitation, we use partial Graph Transformation (pGT),^{93,94} a network reduction procedure that removes states from the system, while preserving FPT distributions. This reduction procedure can improve the condition number of the network and enable the accurate computation of eigenmodes corresponding to the slow, rate-determining transitions, which are often the most interesting events. Therefore, to compute FPTs between minima at the bottom of different funnels that are separated by large energy barriers, we use pGT to retain only the states at the bottom of competing funnels before performing eigendecomposition. The pGT procedure also facilitates the computation of FPTs for larger landscapes.

II.D. Hydration Free Energy. Hydration free energy (HFE) is an important thermodynamic property that quantifies the energy change associated with a gas phase solute being dissolved in water.^{95,96} It depends on the interactions between the solute molecules and water molecules and is calculated as the Gibbs free energy difference between the solvated state of a molecule and the gas phase.

$$\Delta G_{\text{hyd}} = G_{\text{solv}} - G_{\text{gas}} \quad (9)$$

In a broader sense, the HFE demonstrates how favorable it is for a molecule to be surrounded by water compared to being in a vacuum, which correlates directly to its solubility and stability in aqueous solutions. This property is important for predicting how molecules behave in biological systems where water is the predominant solvent.

For our A β 28, A β 40, and A β 42 databases we performed computations of ΔG_{hyd} at the classical GFN-FF level⁹⁷ and with the semiempirical quantum mechanical (SQM) method GFN2-xTB.⁹⁸ Input structures were selected from the A β 28, A β 40, and A β 42 energy landscapes. For simplicity and computational efficiency, all of the investigated structures were optimized at the faster GFN-FF level and rotational–vibrational free energy contributions were obtained thereafter in the modified rigid-rotor harmonic-oscillator (mRRHO) approximation.^{99,100} This procedure has proved to perform reasonably well for macromolecules.^{101,102} The calculations were repeated for both the gas phase and with implicit solvation, modeling solvation free energy effects with the analytical linearized Poisson–Boltzmann (ALPB) model,¹⁰³ parametrized for water. GFN2-xTB calcu-

lations with and without ALPB implicit solvation were performed as single-point evaluations of the GFN-FF optimized structures. All calculations were performed with version 3.0 of the CREST code.^{104,105} Since more than one structure for each system was investigated, the final free energies were Boltzmann-weighted to obtain \bar{G}_{solv} and \bar{G}_{gas} , which is standard practice for supramolecular calculations.¹⁰⁶ Subtracting these Gibbs free energies for the two phases finally provides an estimate of $\Delta\bar{G}_{\text{hyd}}$ according to eq 9, which corresponds directly to solubility.

II.E. Solubility Prediction Using Graph Convolutional Networks. The GCN is a type of neural network designed to process graph-structured data. With each layer of a GCN, the hidden representation of each node is updated according to the hidden representation of its neighbors. The precise propagation protocol is written as¹⁰⁷

$$\mathbf{H}^{(l+1)} = \sigma(\mathbf{D}^{-1/2} \mathbf{A} \mathbf{D}^{-1/2} \mathbf{H}^{(l)} \mathbf{W}^{(l)}) \quad (10)$$

Here, \mathbf{H}^l represents the hidden representation matrix at layer l . The hidden representation is updated with each layer of the GCN. At layer 0, the hidden representation of each node is equivalent to the input node feature vector. \mathbf{A} is an adjacency matrix, or edge matrix, of the graph with self-connections; it weights the importance of edges for propagating representations across neighbors. \mathbf{D} is a diagonal matrix used to normalize the adjacency matrix so that each row sums to 1, $D_{ii} = \sum_j A_{ij}$. $\mathbf{W}^{(l)}$ is a trainable weight matrix specific to layer l . σ represents a nonlinear activation function, in our case the rectified linear unit function, $\text{ReLU}(x) = \max(0, x)$.

To convert a protein structure to a graph representation, each amino acid is represented by one node. In the case of a binary edge matrix, an edge between node i and j is defined ($A_{ij} = 1$) if the residues they represent are determined to be in contact, with $A_{ij} = 0$ otherwise. Following the Critical Assessment of protein Structure Prediction (CASP) definition, contact means that the inter-residue distance between the C^β (or C^α for Glycine) is 8 Å or less.¹⁰⁸

The GCN used here is a previously published model with pretrained parameters called *GraphSol*.¹⁰⁹ For the solubility prediction, *GraphSol* reports an accuracy of $R^2 = 0.48$. For context, other machine learning (ML) methods achieve lower accuracy on the *eSol* data set¹¹⁰ ($R^2 = 0.44$ for *SOLart*,¹¹¹ $R^2 = 0.45$ for *ProGAN*¹¹²). *GraphSol* is a framework to predict the solubility of proteins using only the sequence. Reliance on sequence data is common for ML methods because the database of protein sequences vastly outnumbers the database of experimentally characterized protein structures. A graph representation is constructed, with the edge values being the probability of contact between residues as predicted by SPOT-contact, an ML method to predict contact maps from sequence.¹¹³ In addition to two GCN layers, *GraphSol* adopts a self-attention layer to pool the variably sized hidden state matrix into a fixed-size protein representation, followed by a sigmoidal multilayer perceptron to map this representation to an output $\in (0, 1)$.

The node feature matrix contains 94 features for each node, or $\mathbf{H}^0 \in \mathbb{R}^{N \times 94}$, with N number of nodes/amino acid residues. The first 20 dimensions of each node feature vector, H_i^0 for node i , is derived from the BLock SUBstitution Matrix BLOSUM62.¹¹⁴ This is a 20×20 matrix that gives a score for the similarity between each amino acid pair. This encoding has been shown to outperform simple one-hot encoding,¹¹⁵ which would imply that each amino acid is strictly orthogonal. The next 50 dimensions

refer to 20 features from the Position-Specific Scoring Matrix (PSSM)¹¹⁶ and 30 features from the Hidden Markov Matrix (HMM).¹¹⁷ These are methods of encoding evolutionary conservation in residues with multiple sequence alignment performed on existing databases, Universal Protein resource (UniProt) Reference clusters UniRef90¹¹⁸ and Uniclust30,¹¹⁹ respectively. The first 20 dimensions of the PSSM and HMM represent the probability of an amino acid existing at a given position in evolutionarily similar proteins, while the remaining 10 dimensions of the HMM represent aggregate probabilities of insertions or deletions. Additionally, seven physiochemical properties of each amino acid¹²⁰ and predicted structural features from Structural Property prediction with Integrated DEep neuRal network 3 (SPIDER3)¹²¹ (14 dimensions) were included. SPIDER3 is trained to predict secondary structure and solvent-accessible surface area from sequence-only information, but we can also obtain this information directly from molecular coordinates after our coarse-grained structures were mapped to all-atom representations.

For amyloid solubility prediction, the three features that encode the probability of secondary structure were replaced as one-shot encodings; the presence of a coil, sheet, or helix was calculated with the DSSP algorithm.⁸¹ The feature that represented the solvent-accessible surface area was also replaced by the calculation from DSSP. Furthermore, the edge feature matrix was replaced by a binary contact map, as there is no longer a probability associated with contact once a structure is defined. These feature replacements were made only during amyloid solubility prediction, and not during training, as the training data set only contains sequence information.

The training set for *GraphSol* was the solubility database of ensemble *Escherichia coli* (*E. coli*) proteins (*eSol*).¹¹⁰ The solubility, s , of proteins in the *eSol* database was measured as the ratio of the supernatant fraction to the total fraction after centrifugation in a cell-free translation system that only contains the essential *E. coli* factors responsible for protein synthesis,¹²² resulting in a value between 0 and 1. This solubility measurement is different from the physical solubility, defined as the concentration of a protein in a saturated solution. The physical solubility of proteins is heavily influenced by pH, salt concentration, etc., and requires higher experimental complexity to measure; thus, it is difficult to construct large data sets with physical solubility.

Because *GraphSol* utilizes a self-awareness layer, we can investigate the self-awareness weights to understand how much each node contributes to the final solubility prediction. This structure enables us to estimate the relative contribution of each amino acid to solubility. *GraphSol* employs two GCN layers; because of these layers, a node's hidden feature vector has been affected by its neighbors up to two steps away. A self-attention layer with four attention heads then pools these hidden node features into a single hidden feature vector, to which a final convolution is applied to output a single solubility score. We examined the relative contribution of each node to the end solubility prediction by applying attention layer weights to their corresponding hidden node vectors, followed by the final convolution. We did not apply the sigmoidal function during the final convolution, as it is nonadditive and monotonic. Instead, we analyzed the linear outputs before the sigmoid activation, which preserves the relative contributions of each node to the predicted solubility.

Table 1. Hydration Energies and Hydration Free Energies, Calculated for the Selected Minima from A β 28, A β 40, and A β 42 Energy Landscapes at the GFN-FF Level of Theory and Boltzmann-Weighted^a

database	method	$\Delta \bar{E}_{\text{hyd}}$		$\frac{\Delta \bar{E}_{\text{hyd}}}{N^*}$	$\Delta \bar{G}_{\text{hyd}}$		$\frac{\Delta \bar{G}_{\text{hyd}}}{N^*}$
		$[E_h]$	$[\text{kcal mol}^{-1}]$	$[\text{kcal mol}^{-1}]$	$[E_h]$	$[\text{kcal mol}^{-1}]$	$[\text{kcal mol}^{-1}]$
Aβ28	GFN-FF	-0.3851056	-241.66	-8.63	-0.3992255	-250.52	-8.95
	GFN2-xTB	-0.8219843	-515.80	-18.42	-0.8549519	-536.49	-19.16
Aβ40	GFN-FF	-0.3867420	-242.68	-6.07	-0.4044057	-253.77	-6.34
	GFN2-xTB	-0.8687382	-545.14	-13.63	-0.9117551	-572.14	-14.30
Aβ42	GFN-FF	-0.4083824	-256.26	-6.10	-0.4323021	-271.27	-6.46
	GFN2-xTB	-0.9814659	-615.88	-14.66	-0.9790356	-614.35	-14.63

^aSolvent effects were modeled by the ALPB(water) implicit solvation potential. (*) Additionally, values have been divided per number of residues (N) for each system.

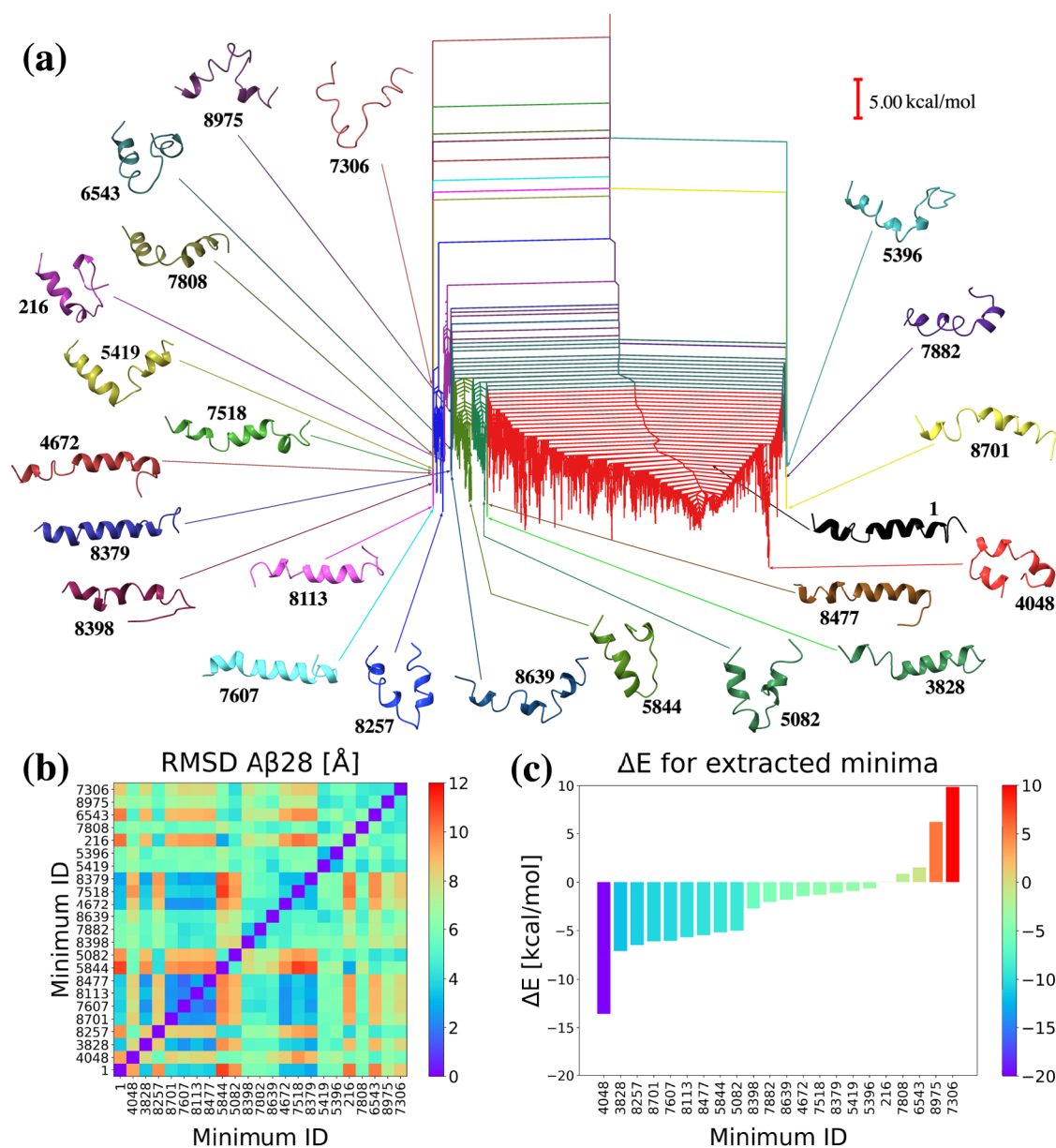


Figure 4. (a) Disconnectivity graph for the A β 28 UNRES energy landscape, with 22 distinctive groups of minima in different colors. The selected minima are shown and are colored to match their respective clusters and labeled with numbers that refer to the order from the database. Minimum of 1 is the PDB structure, and the global minimum is 4048. (b) RMSD values for the 22 minima after alignment with the PDB structure. (c) ΔE for the 22 minima relative to the PDB structure.

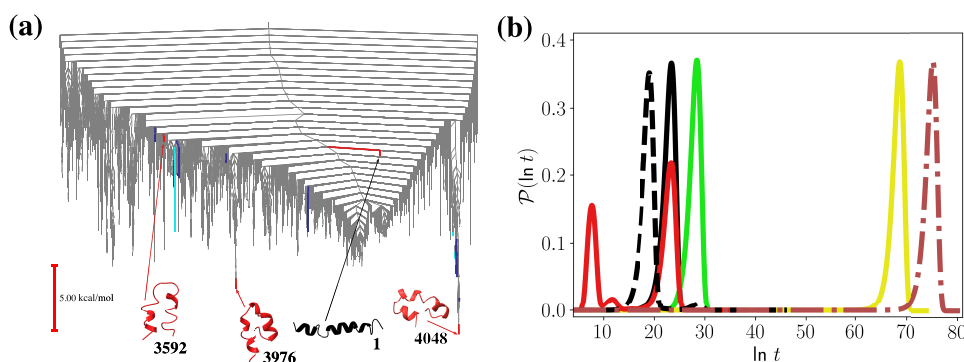


Figure 5. (a) Disconnectivity graph showing the 7525 minima in the red funnel illustrated in Figure 4, for the A β 28 energy landscape. Minima shown in dark and light blue correspond to the states that contribute most to the time scale of the fastest and the second fastest peaks, respectively, for the FPT distribution for the 4048 \leftarrow 3592 transition, which is shown in red in (b). (b) First passage time distributions for transitions within the A β 28 landscape at $T = 300$ K, 4048 \leftarrow 1 (solid black), 4048 \leftarrow 3828 (solid green), 4048 \leftarrow 8701 (solid yellow), 4048 \leftarrow 3592 (solid red), 3976 \leftarrow 1 (dashed black) and 7306 \leftarrow 1 (dot-dash reddish-brown).

III. RESULTS AND DISCUSSION

III.A. Hydration Free Energies. The hydration free energies were calculated for selected minima from the A β energy landscapes and then Boltzmann-weighted. The corresponding values are presented in Table 1. In general, more negative ΔG_{hyd} values indicate a better solubility of the investigated system. When calculating these values using implicit solvation models, it is essential to account for size effects introduced by the implicit solvation potential. Specifically, the solvent-accessible surface area is proportional to the nonpolar surface interaction energy and is directly related to the overall system size. This consideration is particularly important for A β monomers, as their size correlates with an increasing proportion of hydrophobic residues: 11 hydrophobic amino acids in A β 28 (39%), 23 hydrophobic amino acids in A β 40 (58%), and 25 hydrophobic residues in A β 42 (60%). Therefore, for comparison, we normalized the overall values by the number of residues to address the impact of increasing the system size and the influence of the hydrophobic surface energy on the analysis. In the context of A β monomers, this normalization is crucial, given that hydrophobicity increases significantly with the number of residues, particularly as residues 29–42 are known to be hydrophobic.

Analysis at the GFN-FF level reveals a clear trend in both the hydration energy (ΔE_{hyd}), excluding rovibrational free energy contributions, and the HFE. According to the force field predictions, A β 28 is expected to be the least soluble, while A β 40 shows slightly better solubility, and A β 42 demonstrates the highest solubility among the three. However, when the results are considered normalized by the number of residues, a different trend emerges. Specifically, A β 28 appears to be the most soluble, whereas A β 40 and A β 42 exhibit very similar solubilities, however significantly lower than A β 28. This observation supports the well-established concept regarding solubility for A β monomers, i.e., that the solubility decreases with the system size. This result underscores the importance of normalizing values to account for size effects. Interestingly, HFE values at the GFN2-xTB level exceed those of GFN-FF by more than a factor of 2, which hints at an underestimation of the intramolecular interactions by the force field. A probable cause for this issue is the description of electrostatics, which become important, especially in combination with the polar energy contributions of implicit solvation and which are typically much better described at the GFN2-xTB level compared to GFN-FF.⁹⁸ Nonetheless,

the SQM method exhibits the same general trend as results using HFE: post-normalization, it predicts solubilities for A β 40 and A β 42 far lower than that of A β 28.

III.B. The UNRES A β 28 Energy Landscape. First, we performed potential energy landscape explorations for A β 28 by conducting single-ended searches. This process generated a kinetic transition network with 9072 minima and 11121 transition states. We defined a threshold energy of 14 kcal/mol to cluster the minima into 22 different funnels within the energy landscape, as mentioned in the introduction. These clusters, alongside the selected minima, are shown in different colors in Figure 4a. We calculated UNRES potential energies for the lowest minimum within each cluster relative to the PDB structure, as shown in Figure 4c. Additionally, we calculated the RMSD matrices (Figure 4b) for all of the structures with the lowest energy within each cluster. Finally, Figure 4a illustrates the disconnectivity graph^{74,75} for the A β 28 UNRES potential energy landscape, with the selected minima and their positions indicated by arrows. Structures are assigned numerical identifiers corresponding to the order in which they were found in the database, with the PDB structure labeled as 1.

Our investigation identified several minima with distinct structures, frequently characterized by more collapsed conformations compared to the straight helical PDB structure. For example, the global minimum of 4048, which is 13.608 kcal/mol lower in energy than the PDB structure (Figure 4c), comprises two parallel segments. The first segment includes the Ala2-Ser8 helical fragment, followed by a turn in the Ser8-Glu11 region. The second segment contains the Glu11-Val24 helix, which is interrupted by a hydrophobic fragment from Leu17 to Phe19, with the characteristic Phe19 residue exposed to the solvent (Figure 4a).

The structure that differs the most from minimum 1 is a minimum of 5844 (Figure 4b), with an RMSD of 10.99 Å after alignment with the PDB structure. This structure begins with the Ala2-Glu15 helical fragment, followed by a sharp turn in the Gln15-Leu17 region. The remaining residues form a coil, with the characteristic Phe19 residue exposed to the solvent, and a short Asp23-Val24 turn. In Figure 4b, there is a visible cluster of four minima (8701, 7607, 8113, and 8477) that are similar in terms of RMSD, each with an RMSD lower than 3 Å after alignment with the PDB structure. The main structural difference within this set of minima is the length of the helix, which is longest in minimum 7607, extending from Ala2 to

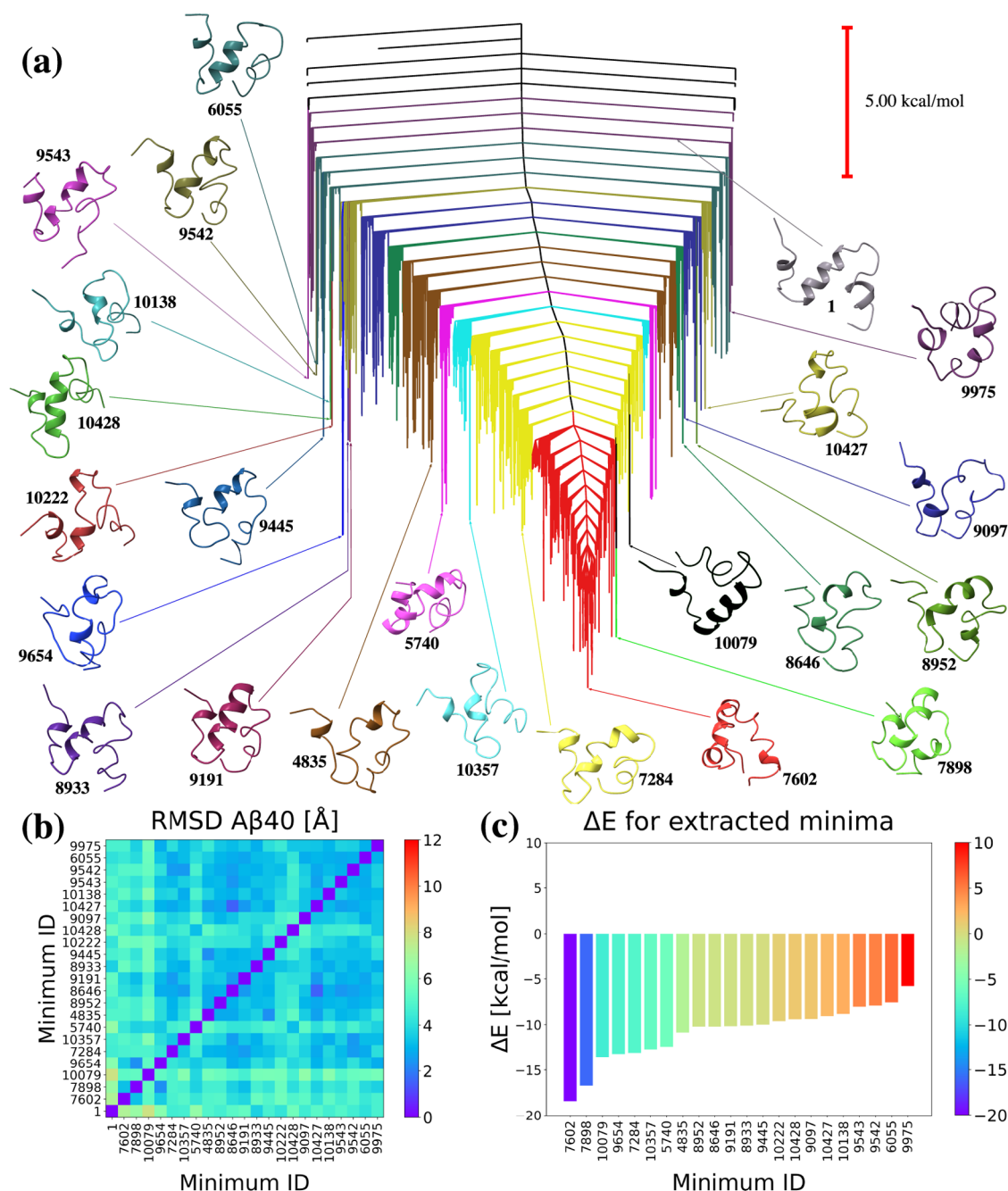


Figure 6. (a) Disconnectivity graph for the Aβ40 UNRES energy landscape, showing 21 distinctive groups of minima in different colors. The selected minima are shown and are colored to match their respective clusters and labeled with numbers that refer to the order from the database. The minimum of 10079, which does not belong to any of these groups, is also included for further analysis. Minimum 1 is the PDB structure, and the global minimum is 7062. (b) RMSD values for the 22 minima after alignment with the PDB structure. (c) ΔE for the 22 minima relative to the PDB structure.

Val24. Most minima in this set exhibit a break in the helix at Val24. The only minimum where the helix breaks earlier is 8701, which features a Val12-Glu22 helix. Out of these four minima, minimum 8701 differs the most in structure, as it also includes a random coil in the His6-Val12 segment and a short helix from Ala2 to His6.

Some of the selected minima, such as 8257, 5844, and 5082, exhibit significant structural changes and have energies lower than those of the PDB structure (Figure 4a). The length of the random coil separating helices with sharp turns varies among these minima. In minimum 8257, the Ser8-Val18 sequence of turns shifts the structure away from a straight helix, with Val24

exposure to the solvent breaking the helix. In minimum 5844, the Gln15-Lys28 fragment lacks a well-defined secondary structure, featuring turns caused by the hydrophobic Val24 collapsing into the center of the structure. This minimum retains a helical segment from Ala2 to Gln15. Minimum 5082 exhibits a central turn at His14-Gln15, shifting the structure into a V-shape, with two helical fragments, Ala2-Gly9 and Ala21-Val24. Similar to the other minima, the exposed Val24 breaks the helix at the C-terminal end. All three minima seem to exhibit turns associated with Gln15.

Interestingly, a low ΔE does not necessarily correlate with a high RMSD relative to the PDB coordinates. This behavior is

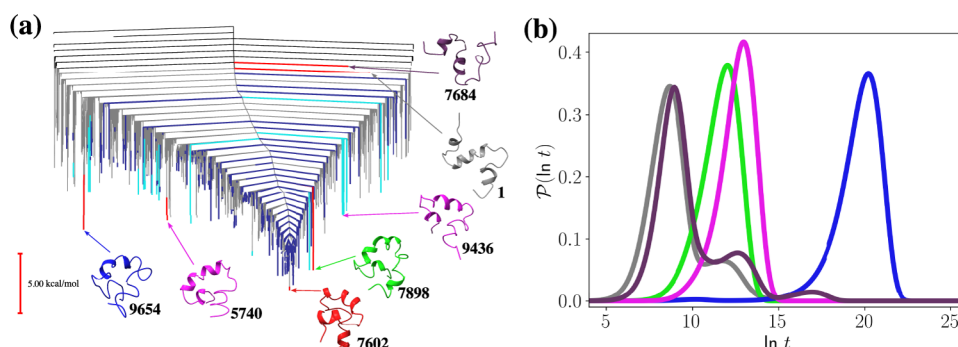


Figure 7. (a) Disconnectivity graph highlighting specific minima within the $A\beta 40$ energy landscape from Figure 6. Minima shown in dark and light blue correspond to the states that contribute the most to the time scale of the fastest and the slowest peaks, respectively, for the FPT distribution for the $7602 \leftarrow 1$ transition, which is shown in gray in (b). (b) First passage time distributions for transitions within the $A\beta 40$ landscape at $T = 300$ K, $7602 \leftarrow 1$ (gray), $7602 \leftarrow 7898$ (green), $7602 \leftarrow 9654$ (blue), $7602 \leftarrow 5740$ (pink) and $7602 \leftarrow 7684$ (purple).

consistent with the properties of amyloid peptides, which are characterized by multifunnel energy landscapes. Previous studies^{28,123} have demonstrated that these landscapes lack a well-defined global minimum. Instead, they feature many competing structures in the low-energy regions with distinct secondary structures. Some of these structures may resemble or differ significantly from the PDB structure, resulting in arrangements that appear somewhat random and do not directly correlate with a high RMSD relative to the PDB coordinates.

To highlight distinct dynamical time scales and probe competing kinetic pathways, we compute FPT distributions between selected minima (Figure 5). First passage time distributions for transitions between minima at the bottom of funnels in the energy landscape are single-peaked, as the time scales for these transitions are dominated by the time taken to leave the initial trapping basin. The larger the energy barrier between the source and sink, the slower the transition. The slowest transition from the PDB structure is to a minimum of 7306, which is the only minimum with no well-defined secondary structure. Each individual peak in an FPT distribution arises due to a set of pathways between the source and sink that occur on similar time scales. The FPT distribution for the transition from the PDB structure to the global minimum, i.e., $4048 \leftarrow 1$, is also a single peak. Pathways for this transition generally spend time in state 3976 en route to sink state 4048.

However, multiple peaks can be seen in the FPT distributions for certain starting states. For example, the FPT distribution for the $4048 \leftarrow 3592$ transition exhibits three peaks. The peak at longer times corresponds to pathways that visit the small trapping basin containing state 3976. The two faster peaks come from pathways that do not enter the 3976 kinetic trap. The states that contribute most to the time scale of the first two peaks are shown in blue in Figure 5.

III.C. The UNRES $A\beta 40$ Energy Landscape. The $A\beta 40$ monomer contains 12 additional residues compared to $A\beta 28$, all of which are hydrophobic. Using the same methodology, we first conducted single-ended searches of the potential energy landscape of $A\beta 40$, constructing a kinetic transition network with 11451 minima and 13834 transition states. We defined a threshold energy of 5 kcal/mol and identified 21 clusters of minima based on this criterion. The energy landscape of $A\beta 40$ differs significantly from that of $A\beta 28$, exhibiting a single-funnel topology that leads to a well-defined global minimum. Consequently, the identified clusters of minima represent distinct subsets within the same funnel (Figure 6a). The selected energy threshold was set to 5 kcal/mol to ensure a

comparable number of structures for structural analysis. Some minima, such as 10079, do not reside within a trapping basin and are indicated in black. We, therefore, conducted a detailed analysis of a minimum of 10079 and added it to selected minima for $A\beta 40$ (Figure 6).

Our investigation revealed that many minima have energies lower than those of the PDB structure; interestingly, the PDB structure is significantly higher on the $A\beta 40$ landscape than on the $A\beta 28$ landscape. Furthermore, the RMSD matrices (Figure 6b) indicate much smaller structural differences compared with the structures analyzed for the $A\beta 28$ landscape. Despite this trend, several minima exhibited distinctive structures, often more collapsed than the PDB structure. For example, the global minimum, 7602, with an energy 18.440 kcal/mol lower than the PDB structure (Figure 6c), is much more compact with three short helices: Phe4-Gly9, Hie14-Ala21 with Phe19 exposed to the solvent, and Ile31-Gly37. The structure contains two turns significantly influencing its shape: Glu11-Hie13 and Asn27-Lys28. The rest of the global minimum is mostly random coil.

The structure that differs the most from the PDB structure is a minimum of 10079, which is not within the trapping basin. This structure exhibits a very short helical Phe4-Asp7 fragment and a longer Val12-Asp23 helical fragment. The hydrophobic C-terminal end is mostly disordered but partially shifts into contact with the main helix; specifically, Leu17 and Ala30 are intriguingly close in this structure. Phe19 is within the helix and is exposed to the surface. Examining Figure 6b, we find that structures 9445, 8933, 9191, 8646, 8952, and 4835 form one set with smaller RMSD values, below 3 Å, while structures 9097, 10427, 10138, 9543, 9542, and 6055 form another set. Most structures within these clusters exhibit a short helical Phe4-Asp7 fragment. Phe19 is within the helix in most structures; however, the length of this helix varies among different minima, ranging from Hie14-Val24 to shorter Val18-Val24 fragments. Other regions of the structures are rather disordered and lack a well-defined secondary structure.

As for $A\beta 28$, we compute FPT distributions between selected minima (Figure 7). As the $A\beta 40$ landscape is single-funneled, there is a smaller range of transition time scales within the network, compared to $A\beta 28$. Once again, transitions between the lowest minima of trapping basins give single-peaked FPT distributions. Transitions from high energy minima to low energy minima can result in multi-peaked FPT distributions, such as those shown for the $7602 \leftarrow 1$ and $7602 \leftarrow 7684$ transitions. The states that contribute the most to the time scale of the fastest and slowest peaks for the $7602 \leftarrow 1$ transition are

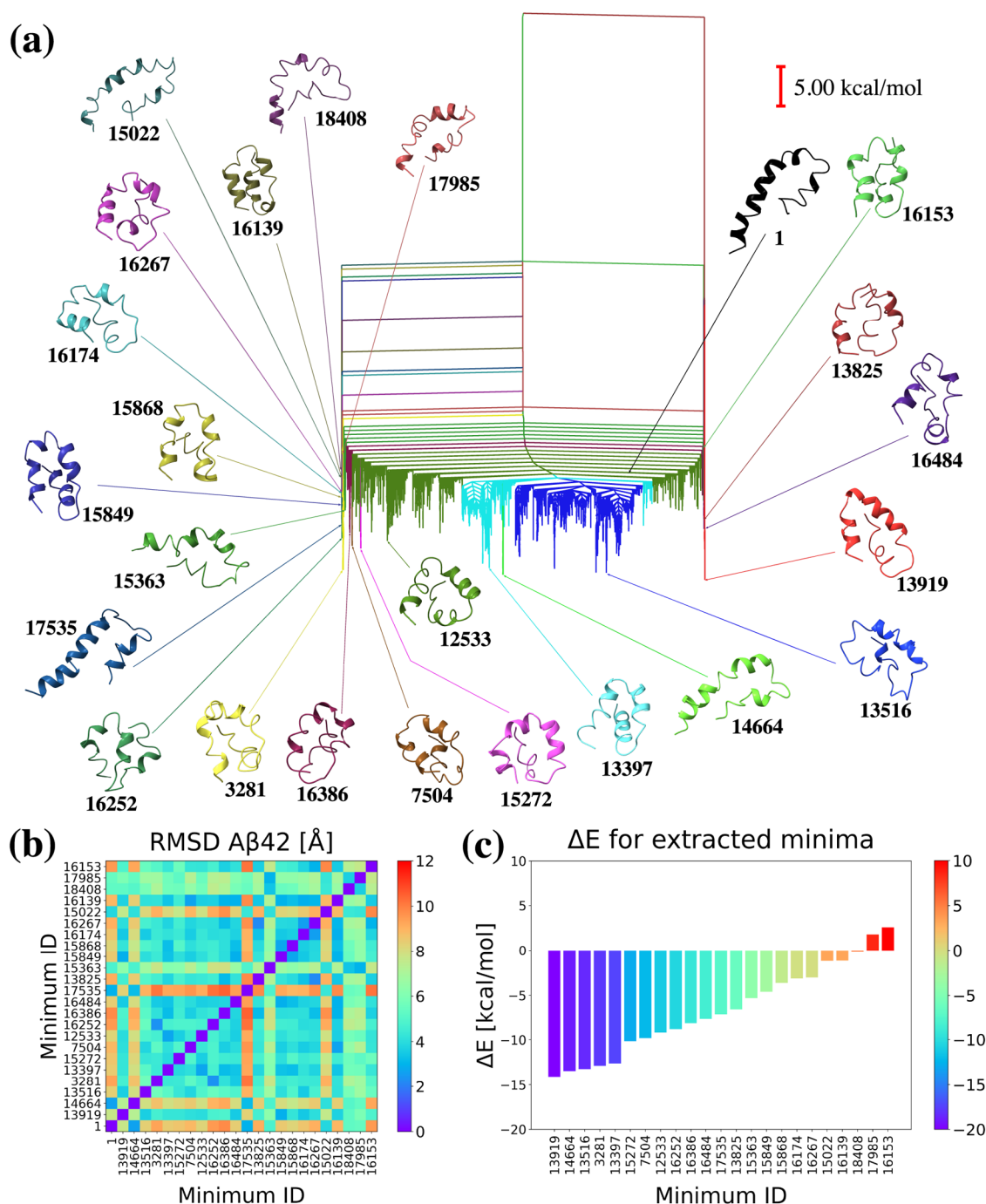


Figure 8. (a) Disconnectivity graph for the Aβ42 UNRES energy landscape, with 23 distinctive groups of minima in different colors. The selected minima are shown and are colored to match their respective groups and labeled with numbers that refer to the order in the database. Minimum 1 is the PDB structure, and the global minimum is 13919. (b) RMSD values for the 23 minima after alignment with the PDB structure. (c) ΔE for the 23 minima relative to the PDB structure.

shown in dark and light blue, respectively, in the disconnectivity graph in Figure 7. Unlike Aβ28, the time scales of these peaks are due to contributions from many different minima, rather than only a few, as the UNRES landscape has a single funnel. Despite the landscape exhibiting a single funnel, some minima have larger barriers to escape the funnel than others, e.g., the minima labeled in light blue in Figure 7a. Pathways that visit these light blue minima during the 7602 \leftarrow 1 transition contribute to the slower peak in the FPT distribution, as it takes longer to escape from these minima. Because some minima have slower escape

times, multip peaked FPT distributions appear for many different transitions between high-energy and low-energy states.

III.D. The UNRES Aβ42 Energy Landscape. Aβ42 has two additional hydrophobic residues Ile41 and Ala42 compared to Aβ40. We employed the same methodology and first conducted single-ended searches of the potential energy landscape of Aβ42, establishing a kinetic transition network comprising 18429 minima and 24919 transition states. We then defined a threshold energy of 11 kcal/mol and grouped the minima into 23 disjointed trapping sets based on this criterion. We performed the same analysis as before for structures with the smallest

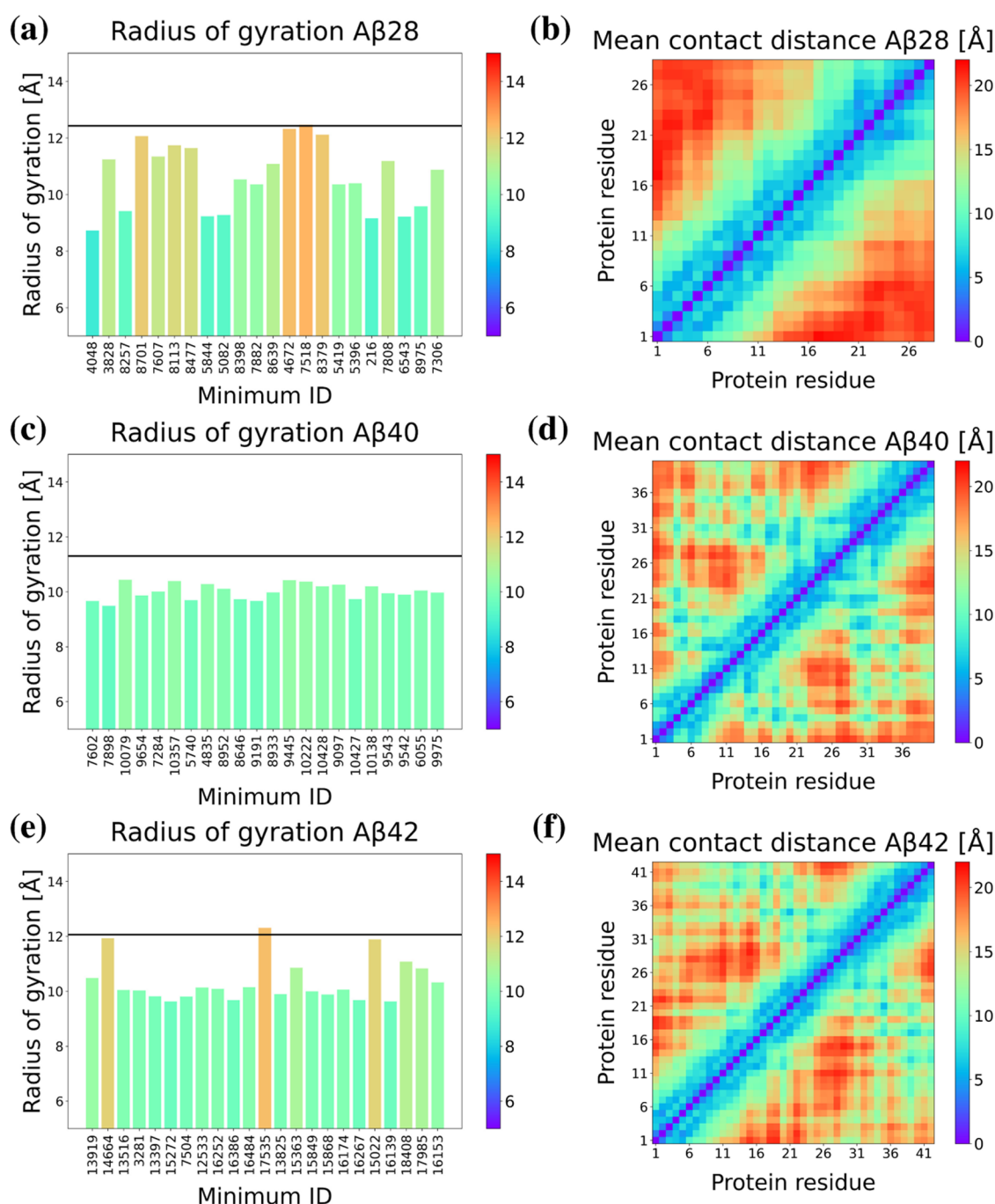


Figure 9. (a, c, e) Radius of gyration and (b, d, f) mean contact maps for selected minima from Aβ28, Aβ40, and Aβ42 databases, respectively. Mean contact maps were averaged over all selected minima from each database.

energy of each set (Figure 8). Interestingly, the multifunnel energy landscape of Aβ42 differs significantly from that of Aβ40 and the organization is like the UNRES results for Aβ28. Most of the selected minima lack secondary structure and have high potential energy barriers between them, as observed for different atomistic potentials in previous studies.^{28,124}

The global minimum, 13919, which has an energy similar to that of the next lowest energy minimum, is separated by the highest transition state (Figure 8a), highlighting the substantial energy increase required to reach other competing minima with similar energies (Figure 8c). This result demonstrates the importance of thorough exploration of the complex amyloid energy landscape (Figure 8a), as minima with similar energies

can differ significantly in structure when separated by substantial barriers.

The RMSD matrix (Figure 8b) shows that the Aβ42 structures differ much more significantly than those in the Aβ40 energy landscape. For instance, all selected minima except 17535 and 14664 differ significantly from the PDB structure, with only three minima maintaining the straight helix fragment at the N-terminal end, albeit with different fragment lengths: Ala2-Ser26 for minimum 1, Ala2-Val24 for minimum 17535, and Glu3-Val24 for minimum 14664. These minima exhibit the greatest divergence from other structures, which often break this helix in the hydrophobic C-terminal end, attempting to collapse inside the structure and leaving the N-terminal end without a

Table 2. Averaged SASA for Selected Minima from the A β Databases

database	total [\AA^2]	polar [\AA^2]	hydrophobic [\AA^2]	polar [%]	hydrophobic [%]
A β 28	2908.69 \pm 143.28	2029.01 \pm 113.63	879.68 \pm 94.63	69.78	30.22
A β 40	3370.32 \pm 124.09	1806.19 \pm 99.70	1564.13 \pm 94.93	53.59	46.41
A β 42	3483.86 \pm 130.67	1843.88 \pm 108.10	1639.98 \pm 141.98	52.96	47.04

well-defined straight helix, as observed in the mentioned structures. Additionally, other clusters with moderate RMSD values below 4 \AA are present (Figure 8b), including minima 16267, 16174, 15868, 15849, and 16139. In most of these structures, Phe19 is within the helix and consistently exposed to the solvent, which may further influence the breaking of the helix within the Phe20-Val24 region. This region, rich in hydrophobic residues, appears to cause further structural changes and shifts the rest of the hydrophobic C-terminal end into the center of the structure in most cases (Figure 8a). The A β 42 network is ill-conditioned and more complex than the other landscapes, making FPT analysis challenging. However, we plan to conduct this analysis in future work.

III.E. Further Structural Analysis. We examined selected minima in terms of the radius of gyration (Figure 9a,c,e) to identify trends and assess the compactness of the structures. The results are largely consistent with the RMSD matrices. In each case, the radius of gyration for the minima is lower than for the initial PDB structure. This observation aligns with expectations for the experimentally determined PDB structures (1AMC, 1AML), which were obtained using solid-state NMR. Solvation during simulation induces further disruption of the straight helical fragment observed in the experimental structure. For 1IYT, hexafluoroisopropanol (HFIP) was used, during solution NMR, in structure determination. HFIP is known to stabilize helices, similar to trifluoroethanol; therefore, helix disruption was anticipated when our computations were performed without HFIP.¹²⁵ For A β 28, structures with the highest radius of gyration also showed the least deviation from the initial PDB structure (Figure 9a). Notably, the radius of gyration varies significantly across the minima studied in the A β 28 energy landscape. In contrast, the results for the A β 40 minima are much more consistent (Figure 9c), displaying only minor changes in the radius of gyration. This observation suggests that most of these structures are more compact, which was also evident during the structural analysis of the minima. It is consistent with the fact that the first minimum obtained by minimizing a PDB structure lies relatively high on the energy landscape, indicating that more compact structures with less exposed hydrophobic regions are energetically favorable according to UNRES. Similarly, structures identified on the A β 42 landscape with a straight helix also have a significantly higher radius of gyration (Figure 9e).

Additionally, we constructed a map of contacts (Figure 9b,d,f) averaged over all extracted minima for each system. The contact map for A β 28 (Figure 9b) reveals a moderately dispersed pattern of contacts, indicating that the peptide adopts structures that are not uniformly compact. Several off-diagonal contacts suggest interactions between residues that are not sequentially close, pointing to the presence of secondary structural elements such as loops or turns. There are regions with denser contacts, which likely correspond to stable secondary structures, such as α helices. This dispersed contact pattern is consistent with the higher and more variable radii of gyration values observed for A β 28 (Figure 9a), suggesting less compact and more structurally diverse conformations. The contact map for A β 40 (Figure 9d)

exhibits a more uniform and denser pattern of contacts compared to that of A β 28. This result indicates that A β 40 adopts more compact and stable structures with UNRES. Several contiguous blocks of contacts are evident, indicating regions where residues are in close proximity and likely form stable secondary structures. Diagonal regions with high contact density suggest the presence of α -helices, while off-diagonal blocks indicate turns. This uniformity and density of contacts correlate with the more consistent and lower radius of gyration values (Figure 9c), indicating that the structures are more compact. The contact map for A β 42 (Figure 9f) displays a mix of dense and sparse regions, reflecting both compact and more extended structural elements. Regions with high contact density, similar to A β 40, suggest the presence of stable secondary structures. Additionally, there are more dispersed contact regions, corresponding to the straight helical structures, which contribute to higher radii of gyration values. This variability in contact density suggests a combination of compact structures and extended helical elements. The mixed contact pattern aligns with the higher and more variable radius of gyration values observed for A β 42 (Figure 9e), reflecting the structural diversity of A β 42 with both stable folded regions and extended helices.

We further investigated the significance of hydrophobic and polar regions given that these regions are known to be the primary structural differences among A β monomer structures. This analysis is particularly crucial for A β monomers, as their size correlates with an increasing proportion of hydrophobic residues: 11 hydrophobic amino acids in A β 28 (39%), 23 hydrophobic amino acids in A β 40 (58%), and 25 hydrophobic residues in A β 42 (60%). Table 2 reflects this trend, showing that with increasing monomer size, there is a rise in total SASA, and notably, a significant increase in hydrophobic contribution from 30.22% for A β 28 to 47.04% for A β 42. For A β 42, the hydrophobic and polar SASA contributions are nearly equal, which significantly affects the structure, with a significant number of hydrophobic residues exposed to the solvent.

Next, we investigated the contribution of specific residues to the SASA across different databases. Figure 10a shows the average SASA per residue for the various databases. The trends vary with the sequence length, with residues at the termini generally exhibiting the highest SASA. Hydrophobic residues adjacent to polar ones tend to be buried, resulting in a decreased SASA. This trend is particularly evident in the increasing SASA for the hydrophobic C-terminal end of A β 40 and A β 42. An interesting observation arises for hydrophobic Phe19, which consistently has the highest SASA among all hydrophobic residues, increasing significantly with the size of the A β monomer. This residue is often solvent-exposed in the selected minima. Additionally, it is apparent that with increasing system size, the contribution of polar residues to the SASA decreases. The lowest SASA values were consistently observed for Ser8, Gly9, and Ala21 across all systems.

In Figure 10b, we analyzed the RMSF of residues, calculated as the positional deviations across all aligned structures for each residue within the given system. Notably, the residues with the highest fluctuations were consistently associated with the

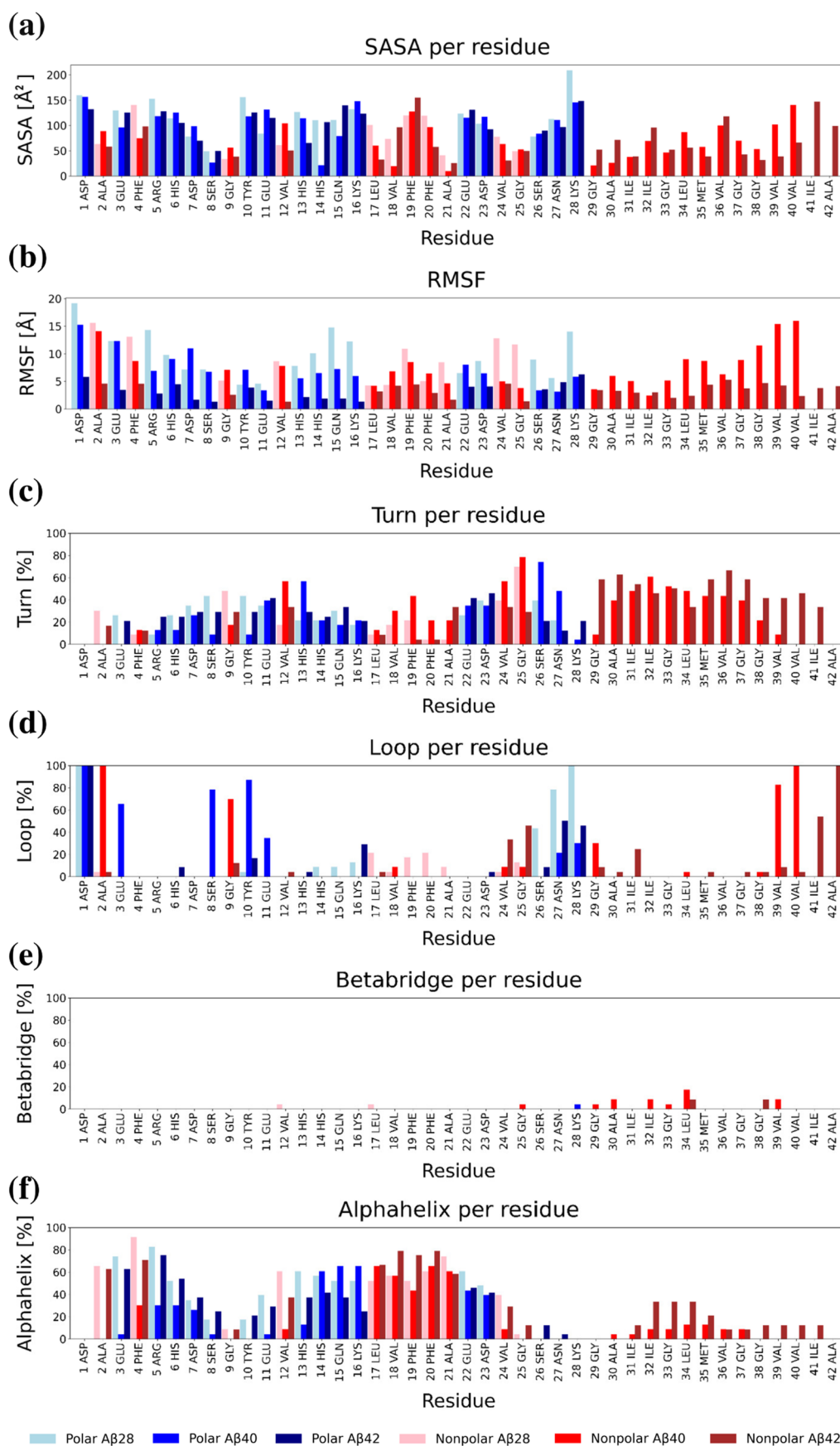


Figure 10. (a) SASA per residue. (b) RMSF per residue. (c–f) Secondary structure contribution per residue calculated with the DSSP algorithm. All data have been calculated after averaging results for all of the selected minima. The figure contains a legend with different color codes for polar and hydrophobic residues for different databases.

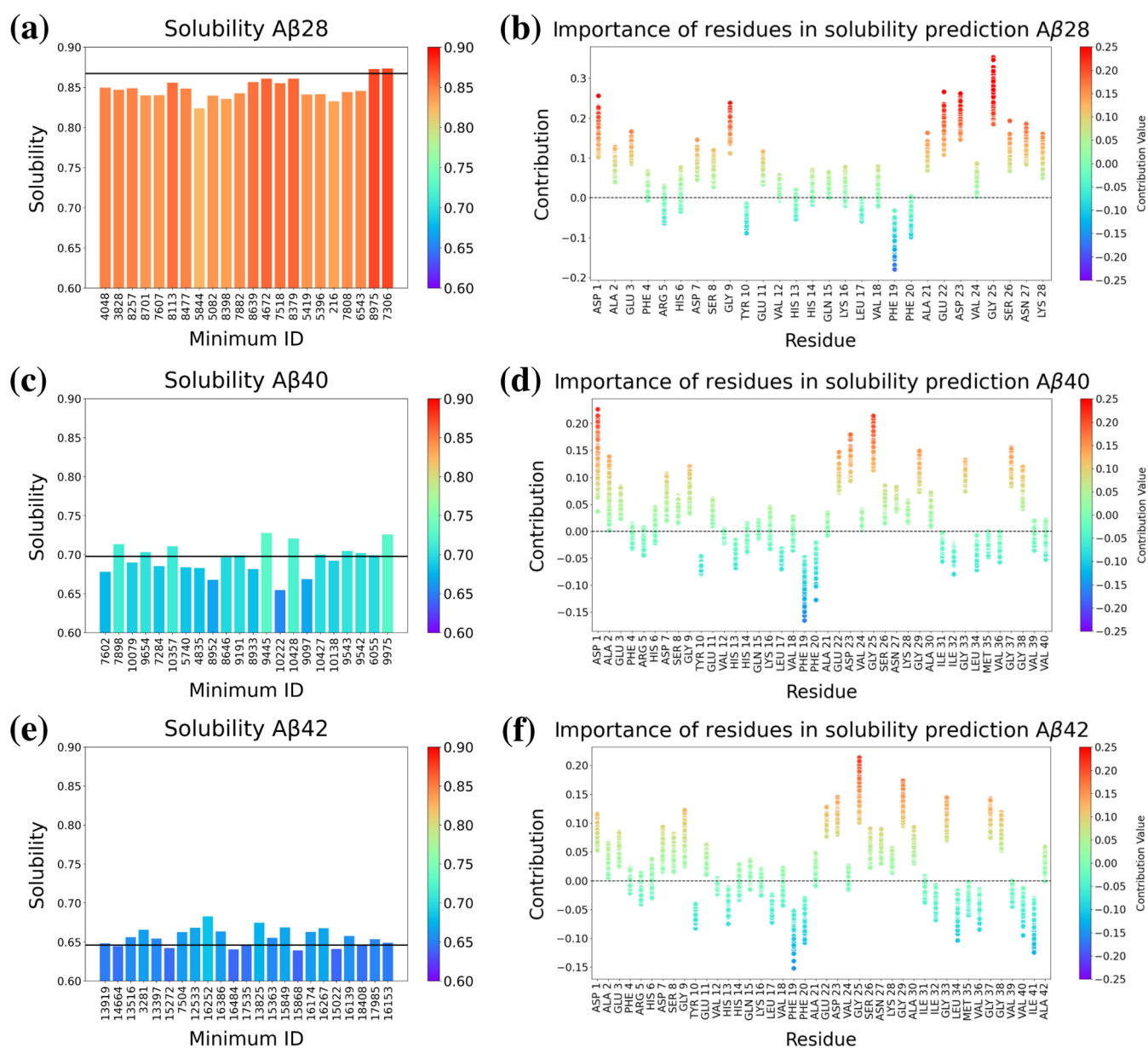


Figure 11. (a, c, e) Solubility predicted by GraphSol for selected minima of Aβ28, Aβ40, and Aβ42. In each panel, the black lines are the predicted solubilities of the PDB structures. (b, d, f) Importance of residues in solubility prediction.

termini. Interestingly, the fluctuation also appears to correlate with system size, as residues in Aβ42 exhibited the lowest RMSF compared to Aβ28 and Aβ40. Specifically, at the C-terminal end, the small structural diversity of the hydrophobic region is more pronounced in Aβ42 compared to Aβ40.

We also predicted the secondary structure distribution using the DSSP algorithm (Figure 10c–f). This analysis allowed us to identify residues in different systems, based on averaged structures, that are particularly prone to forming helices, turns, β-bridges, and random loops. Interestingly, β-bridges are not evident upon visual inspection of the structures, yet the algorithm defined some structures as particularly prone to developing these interactions. This discrepancy could be attributed to various factors, such as the influence of the UNRES potential on secondary structure development or undersampling of parts of the energy landscape where β structures would be present. It is also important to realize that visualization programs may define and display secondary

structures differently. The assignments are therefore interpreted as suggestions.

For Aβ28, Val12 and Leu17 appear prone to β-interaction at one minimum (8975). For Aβ40, several residues were identified in this category: Gly25, Lys28 for 8646; Gly29, Ile33, Gly33, Leu34, and Val39 for 8933; Ala30, Leu34 for 9543; Ala30, Leu34 for 10222; Ile32 for 5740; and Leu34, Val39 for 9975. In Aβ42, two residues, Leu34 and Gly38, were identified as being prone to form β sheets in the monomer form for minima 15272 and 18404. However, the participation of these residues in β-bridge formation is minor, contributing less than 10% of the β-bridges across all minima in the databases and only for the particular structures mentioned above. Most of these residues participate in turn in random coil formation (Figure 10c,d). The terminal residues are typically assigned as loops by DSSP (Figure 10d), and much of the C-terminal hydrophobic region is characterized by a sequence of turns (Figure 10c). The largest contribution of α-helices is visible in Figure 10f, where the

His14-Val24 fragment consistently contributes to the α -helix formation. The length of this helical fragment varies, being Val12-Val24 on average for A β 28. Generally, the helix in this region shortens with an increasing system size. Interestingly, the helix breaks at Val24 where a turn (Gly25-Ser26) occurs and the Asn27-Lys28 region primarily forms loops. The Glu3-Asp7 fragment predominantly forms α -helices, although other types of helical structures, especially 3_{10} -helix, are also detected for Phe4-His6 fragment in the A β 40 databases.

III.F. Solubility. To provide predictions of estimated solubility, we employed GCN fits to data, as discussed in Section II.E. The results are shown in Figure 11a,c,e. Our findings are consistent with the well-established experience in the literature that solubility decreases with increasing size.¹²⁶ The predicted solubility values for different minima did not differ significantly within each database. However, the A β 28 minima exhibited lower values (Figure 11a) compared to the predicted solubility of the PDB structure for this molecule, while the predicted solubility for A β 40 and A β 42 minima exhibit values both below and above the predicted solubility of the original PDB structure (Figure 11c,e).

To further investigate the contributions of specific residues to solubility predictions, we analyzed their importance within the model (Figure 11b,d,f). Note that the contributions for the hidden state of a given residue include contributions from neighbors up to two edges away, with each edge ranging up to 8 Å in distance. The weights in our analysis represent the relative contributions of each residue to the overall solubility prediction of the peptide. It is important to note that we have not applied the final sigmoidal layer to these weights, which would distort the values due to the nonlinear transformation. The importance weights should be understood as relative contributions to the solubility prediction. They can be interpreted in terms of how much each residue (and its surrounding region) contributes, positively or negatively, to the solubility prediction prior to the final sigmoidal transformation. This approach should highlight the individual impact of residues on solubility, providing insights into which of them are most influential in the peptide solubility profile.

We can immediately identify the importance of the Phe19 residue and its surrounding region with the most negative contribution to solubility while exhibiting a high variance. This result means that fluctuations in this region play a key role in GraphSol's prediction of overall protein solubility. In Figure 12,

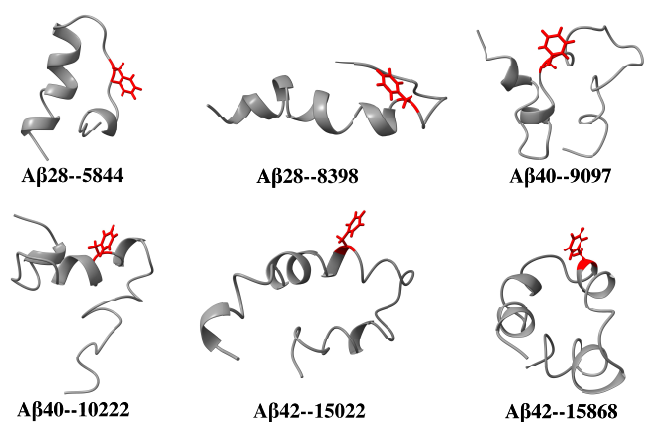


Figure 12. Illustration of various low-solubility minima, with the Phe19 residue highlighted in red.

we illustrate different low solubility minima with the hydrophobic Phe19 residue exposed to the solvent, resulting in disruption of the helical structure in the surrounding region. Bernstein et al.¹²⁷ experimentally demonstrated the critical role of Phe19 in A β aggregation by comparing wild-type A β 42 to an F19P mutant. While wild-type A β 42 rapidly formed larger oligomers and aggregates, including hexamers and pairs of hexamers thought to be early steps in protofibril formation, the Pro19 alloform only formed smaller oligomers (dimers, trimers, and tetramers) and did not progress to larger assemblies. For A β 42, Ile41 also exhibits a substantial negative effect on the solubility.

On the other hand, Gly25 contributed most positively to the predicted solubility across all data sets, with Asp1 also making a notable contribution in the case of A β 40. Remarkably, the anticorrelation from the C-terminal residues appears to be more pronounced in A β 42, suggesting a distinct solubility profile influenced by the terminal residues. Previous work indicates that the C-terminal of A β monomers is more active in aggregation, fibrillation, and β -strand formation.¹²⁸ This expectation is further corroborated by the minor but still significant β -bridge interaction observed at the C-terminal end (Figure 10e).

All correlations presented here were obtained using Pearson linearization. There is a moderate correlation ($r = 0.40$) between the radius of gyration and solubility for A β 28, no correlation for A β 40 ($r = -0.05$), and a moderate anticorrelation for A β 42 ($r = -0.49$). For A β 28, less compact structures exhibit a higher solubility, suggesting they would be less prone to aggregation. In the case of A β 40, the lack of correlation is likely due to the very similar radii of gyration across all of the analyzed structures. For A β 42, the pattern shifts, with more compact structures featuring a buried hydrophobic core demonstrating higher solubility and, potentially, a reduced tendency to aggregate. Furthermore, there is a low anticorrelation ($r = -0.29$) for A β 28 between the RMSD after alignment with the PDB structure and solubility. This result seems logical, as the extended PDB structure has a large radius of gyration; thus, structures most similar to the PDB structure will have the highest radius of gyration and the smallest RMSD. For A β 40, no significant correlation ($r = 0.16$) is observed, likely because the structures do not differ significantly in terms of both radius of gyration and RMSD. However, for A β 42, there is a moderate correlation ($r = 0.53$), indicating that an increase in RMSD aligns with an increase in solubility. As the PDB structures contain a long α -helix, their radius of gyration is significantly higher. Therefore, more compact structures, identified by a lower radius of gyration, are associated with the highest solubility.

A β 42 monomers aggregate more readily compared to A β 40.¹²⁹ The presence of two additional hydrophobic residues in A β 42 (Ile41 and Ala42) facilitates the formation of inter-residue contacts absent in A β 40, including hydrophobic clustering between residues Val39-Ile41 and increased clustering among residues Gly37-Gly38 and Val12-Lys16.¹²⁹ These interactions are evident in Figure 10b, where the fluctuations at the C-terminal end of A β 42 are significantly smaller than those for A β 40, indicating more stable interactions that predispose A β 42 to aggregation. Literature indicates that the most rigid structure within both monomers is the hydrophobic cluster around residues Leu17-Ala21.¹³⁰ Rojas et al.¹³¹ conducted coarse-grained molecular dynamics simulations of five A β 28 monomers to investigate the dynamics of aggregation. They found that when the distance between residues Leu17 and Ala21 was constrained to be inflexible, β -rich aggregates did not form.

This observation is consistent with the results of Kapurniotu et al.,¹³² who demonstrated that the A β 28 (L17K, A21D) peptide mutant, featuring a helix-stabilizing lactam bridge between Lys17 and Asp21, did not aggregate. Through our structural analysis, we found that a disrupted α -helix surrounding Phe19 produces a predicted solubility. Both contact maps (Figure 9b,d,f) and minimal fluctuations for Leu17 and Ala21, along with a small SASA for these residues, indicate the critical role of stabilization in this hydrophobic region. Disruption, by the exposure of Phe19 to the solvent, results in helix disruption, further emphasizing the importance of this region in maintaining solubility.

Previous studies have identified a key hydrophobic interaction between Phe19 and Leu34 in A β 40 to produce β -bridges, a feature also observed in our secondary structure contributions per residue (Figure 10e) for Leu34. This contact is consistently reported in structural studies of A β fibrils¹³³ and has been implicated in oligomer toxicity.^{134,135} Disrupting this interaction significantly reduces membrane affinity for A β oligomers,¹³⁴ further highlighting its role in early aggregation. Computational studies also suggest a high probability of intermolecular Phe19–Leu34 contacts in both A β 40 and A β 42,¹³⁶ supporting its relevance in β -bridge formation and potential neurotoxicity. Additionally, the observed His14–Val24 propensity for α -helical conformations is consistent with experimental solution structures of micelle-bound A β 40 and A β 42, which identify the Gln15–Val24 region as predominantly α -helical.²⁷

IV. CONCLUSIONS

Our study provides insight into the solubility, structural dynamics, and aggregation tendencies of A β 28, A β 40, and A β 42 peptides using a variety of computational approaches. Hydration free energy (HFE) calculations using the GFN-FF and GFN2-xTB methods revealed subtle solubility trends. Initially, A β 42 appeared to have the highest solubility; however, after normalizing for size, A β 28 exhibits the highest solubility, while A β 40 and A β 42 exhibit lower solubilities, consistent with the established size-dependent solubility trends in amyloid peptides. The discrepancy between HFE values from the GFN-FF and GFN2-xTB methods highlights possible underestimates of intramolecular interactions by force fields, particularly in implicit solvation models affecting electrostatics and polar energy contributions. This result highlights the importance of methodological considerations in accurately assessing peptide solubility.

Exploration of the energy landscapes of A β 28, A β 40, and A β 42 highlights distinct structural features. A β 28 and A β 42 exhibit multifunnel landscapes, associated with disorder.^{23,28} This structural complexity has been associated with the transition from α -helices to β -hairpins, a hallmark of amyloid aggregation and plaque formation.¹³⁷ In contrast, A β 40 presents a simpler, single-funnel landscape. This result could reflect undersampling of A β 40 if alternative funnels have been missed, and further investigation is warranted in future work.

Structural analysis involving the radius of gyration and contact maps further highlights the differences between the peptides. A β 28 displayed variable radius of gyration profiles, suggesting structural diversity and flexibility, whereas A β 40 appeared consistently compact, indicative of stable configurations. A β 42 exhibited a mix of compact and extended structures, reflected in its higher and more variable radii of gyration values, correlating with diverse contact patterns indicating both folded and extended helical elements. SASA analysis highlighted increasing

hydrophobic residue exposure with peptide size, influencing solubility and aggregation propensity. Hydrophobic Phe19 consistently exhibited high SASA values for all systems, linking its solvent exposure with increased aggregation tendency. Residue-specific analyses identified Phe19, Ile41, and Ala42 as key determinants impacting solubility through their roles in inter-residue interactions and structural stability. Correlations with the radius of gyration and RMSD are consistent with these structural determinants, showing a positive correlation between structural compactness and solubility in A β 42, where deviations from helical structures increased the aggregation propensity.

Our study employing the GCN method provides further insight into the solubility and structural characteristics of the A β 28, A β 40, and A β 42 peptides, highlighting the aggregation propensity. Consistent with existing literature, A β 28 is predicted to have the highest solubility, and A β 42 the lowest. Structural analyses highlight distinct features influencing solubility. A β 28, despite its smaller size, exhibits configurations with differing radius of gyration, and some structures with higher radius of gyration are predicted to have greater solubility. In contrast, A β 40 structures, characterized by consistent compactness and minimal structural variance, exhibit solubility profiles that are not strongly correlated with structural changes. For A β 42, a moderate anticorrelation between solubility and structural compactness was observed, indicating that more compact conformations with buried hydrophobic cores will have a higher solubility and be less prone to aggregate. Key structural determinants impacting solubility were identified through residue-specific analyses. The hydrophobic Phe19 emerges as a key residue, with a significant negative impact on solubility predictions for the three monomers, consistent with experimental results that link Phe19 exposure to solvent with increased aggregation propensity. Regarding dynamics, our study identified a wide range of time scales, with A β 28 exhibiting the slowest transformations, due to its diverse structural transitions. Furthermore, we observed multi-peaked first passage time distributions for relaxation to the global minimum, depending on the initial starting minimum.

Overall, our findings agree with conventional size-based solubility trends in amyloid peptides, providing detailed insights into how structural variations influence solubility and aggregation dynamics. Machine learning methods such as GCN offer predictive capabilities based on structural features, complementing experimental approaches in advancing therapeutic strategies against protein aggregation disorders. Our results highlight the central role of Phe19 in early amyloid assembly, consistent with previous studies identifying Phe19 involvement in hydrophobic interactions critical for aggregation. However, amyloid formation is a multifaceted process, and the aggregation of other intrinsically disordered proteins is driven by additional effects, including electrostatics and sequence-specific interactions. Additionally, emerging evidence suggests that amyloidogenic proteins can form liquid-like condensates, which may serve as intermediates in fibril formation and contribute to proteopathies. Future studies integrating these alternative aggregation pathways could provide a more comprehensive understanding of amyloid self-assembly. Looking ahead, future research will explore the aggregation propensity of selected low-energy structures and investigate the mutational effects on solubility and aggregation. Utilizing the UNRES potential for accelerated simulations, combined with side-chain refinement techniques such as the Pulchra program and relaxation using all-atom potentials like AMBER mini-

mization, provides a valuable balance of precision and efficiency in exploring the global structural landscapes of amyloids. These methods will be further employed in future work.

AUTHOR INFORMATION

Corresponding Authors

Patryk Adam Wesolowski – Yusuf Hamied Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, U.K.; orcid.org/0000-0002-7751-980X; Email: paw61@cam.ac.uk

David J. Wales – Yusuf Hamied Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, U.K.; orcid.org/0000-0002-3555-6645; Email: dw34@cam.ac.uk

Authors

Bojun Yang – Shenzhen College of International Education, Shenzhen 518040, China

Anthony J. Davolio – Theory of Condensed Matter Group, Cavendish Laboratory, Department of Physics, University of Cambridge, Cambridge CB3 0HE, U.K.; orcid.org/0009-0003-4766-8817

Esmee J. Woods – Department of Engineering, University of Cambridge, Cambridge CB2 1PZ, U.K.; orcid.org/0000-0001-9614-0865

Philipp Pracht – Yusuf Hamied Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, U.K.; orcid.org/0000-0002-8495-9504

Krzysztof K. Bojarski – Department of Physical Chemistry, Gdansk University of Technology, Gdansk 80-233, Poland; orcid.org/0000-0002-2066-1522

Krzysztof Wierbilowicz – Department of Biochemistry and Molecular Genetics, University of Virginia School of Medicine, Charlottesville, Virginia 22908, United States

Mike C. Payne – Theory of Condensed Matter Group, Cavendish Laboratory, Department of Physics, University of Cambridge, Cambridge CB3 0HE, U.K.

Complete contact information is available at:
<https://pubs.acs.org/10.1021/acs.jctc.4c01623>

Author Contributions

P.A.W. conducted energy landscape calculations, structural analysis, and data analysis; wrote the original draft; and edited the manuscript. **B.Y.** performed literature review, prepared figures, drafted the introduction, and edited the manuscript. **A.J.D.** conducted solubility prediction and data analysis, described the GCN methodology, and edited the manuscript. **E.J.W.** performed FPT calculations, described the FPT methodology, contributed to the Results and Discussion section, and edited the manuscript. **P.P.** conducted HFE calculations, described the HFE methodology, contributed to the Results and Discussion section, and edited the manuscript. **K.K.B.** performed structural analysis and data analysis, and described the part of methodology. **K.W.** conducted literature review and edited the paper. **M.C.P.** supervised the GCN portion of the project. **D.J.W.** supervised the project, wrote and maintained the Cambridge software for energy landscape exploration, and edited the manuscript. All authors read and accepted the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

P.A.W. acknowledges the Engineering and Physical Sciences Research Council (EPSRC) for funding his studentship through Doctoral Training Partnership EP/W524633/1. **P.P.** acknowledges the support by the Alexander von Humboldt Foundation for a Feodor Lynen Research Fellowship. **K.K.B.** acknowledges the National Science Centre for financial support of his research (Narodowe Centrum Nauki, grant number: UMO-2023/48/C/ST4/00163).

REFERENCES

- (1) Leopold, P. E.; Montal, M.; Onuchic, J. N. Protein folding funnels: a kinetic approach to the sequence-structure relationship. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 8721–8725.
- (2) Privalov, P. L. Thermodynamics of protein folding. *J. Chem. Thermodyn.* **1997**, *29*, 447–474.
- (3) Veitshans, T.; Klimov, D.; Thirumalai, D. Protein folding kinetics: timescales, pathways and energy landscapes in terms of sequence-dependent properties. *Fold. Des.* **1997**, *2*, 1–22.
- (4) Dobson, C. M. Protein misfolding, evolution and disease. *Trends. Biochem. Sci.* **1999**, *24*, 329–332.
- (5) Stefani, M. Protein folding and misfolding on surfaces. *Int. J. Mol. Sci.* **2008**, *9*, 2515–2542.
- (6) Soto, C.; Pritzkow, S. Protein misfolding, aggregation, and conformational strains in neurodegenerative diseases. *Nat. Neurosci.* **2018**, *21*, 1332–1340.
- (7) Ntarakas, N.; Ermilova, I.; Lyubartsev, A. P. Effect of lipid saturation on amyloid-beta peptide partitioning and aggregation in neuronal membranes: molecular dynamics simulations. *Eur. Biophys. J.* **2019**, *48*, 813–824.
- (8) Alam, P.; Siddiqi, K.; Chaturvedi, S. K.; Khan, R. H. Protein aggregation: from background to inhibition strategies. *Int. J. Biol. Macromol.* **2017**, *103*, 208–219.
- (9) Chiti, F.; Dobson, C. M. Protein misfolding, functional amyloid, and human disease. *Annu. Rev. Biochem.* **2006**, *75*, 333–366.
- (10) Giuffrida, M. L.; Caraci, F.; Pignataro, B.; Cataldo, S.; De Bona, P.; Bruno, V.; Molinaro, G.; Pappalardo, G.; Messina, A.; Palmigiano, A.; et al. β -amyloid monomers are neuroprotective. *J. Neurosci.* **2009**, *29*, 10582–10587.
- (11) Long, S.; Benoist, C.; Weidner, W. World Alzheimer Report 2023: Reducing dementia risk: never too early, never too late, *London Engl. Alzheimer. Dis. Int.* (2023).
- (12) Chen, G.-f.; Xu, T.-h.; Yan, Y.; Zhou, Y.-r.; Jiang, Y.; Melcher, K.; Xu, H. E. Amyloid beta: structure, biology and structure-based therapeutic development. *Acta Pharmacol. Sin.* **2017**, *38*, 1205–1235.
- (13) Tyedmers, J.; Mogk, A.; Bukau, B. Cellular strategies for controlling protein aggregation. *Nat. Rev. Mol. Cell Biol.* **2010**, *11*, 777–788.
- (14) Gsponer, J.; Babu, M. M. Cellular strategies for regulating functional and nonfunctional protein aggregation. *Cell Rep.* **2012**, *2*, 1425–1437.
- (15) Siddiqi, M. K.; Alam, P.; Chaturvedi, S. K.; Khan, R. H. Anti-amyloidogenic behavior and interaction of diallylsulfide with human serum albumin. *Int. J. Biol. Macromol.* **2016**, *92*, 1220–1228.
- (16) Baranello, R. J.; Bharani, K. L.; Padmaraju, V.; Chopra, N.; Lahiri, D. K.; Greig, N. H.; Pappolla, M. A.; Sambamurti, K. Amyloid-beta protein clearance and degradation (ABCD) pathways and their role in Alzheimer's disease. *Curr. Alzheimer Res.* **2015**, *12*, 32–46.
- (17) Dwane, S.; Durack, E.; Kiely, P. A. Optimising parameters for the differentiation of SH-SY5Y cells to study cell adhesion and cell migration. *BMC Res. Notes* **2013**, *6*, No. 366.
- (18) Belyaev, N. D.; Kellett, K. A.; Beckett, C.; Makova, N. Z.; Revett, T. J.; Nalivaeva, N. N.; Hooper, N. M.; Turner, A. J. The Transcriptionally Active Amyloid Precursor Protein (APP) Intracellular Domain Is Preferentially Produced from the 695 Isoform of APP in a β -Secretase-dependent Pathway* \diamond . *J. Biol. Chem.* **2010**, *285*, 41443–41454.

- (19) Nunan, J.; Small, D. H. Regulation of APP cleavage by α -, β - and γ -secretases. *FEBS Lett.* **2000**, *483*, 6–10.
- (20) Chasseigneaux, S.; Allinquant, B. Functions of A β , sAPP α and sAPP β : similarities and differences. *J. Neurochem.* **2012**, *120*, 99–108.
- (21) Miners, J. S.; Baig, S.; Palmer, J.; Palmer, L. E.; Kehoe, P. G.; Love, S. SYMPOSIUM: Clearance of A β from the Brain in Alzheimer's Disease: A β -Degrading Enzymes in Alzheimer's Disease. *Brain Pathol.* **2008**, *18*, 240–252.
- (22) Glenner, G. G.; Wong, C. W. Alzheimer's disease: initial report of the purification and characterization of a novel cerebrovascular amyloid protein. *Biochem. Biophys. Res. Commun.* **1984**, *120*, 885–890.
- (23) Chebaro, Y.; Ballard, A. J.; Chakraborty, D.; Wales, D. J. Intrinsically disordered energy landscapes. *Sci. Rep.* **2015**, *5*, No. 10386.
- (24) Fraser, P. E.; Nguyen, J. T.; Surewicz, W. K.; Kirschner, D. A. pH-dependent structural transitions of Alzheimer amyloid peptides. *Biophys. J.* **1991**, *60*, 1190–1201.
- (25) Yip, C. M.; McLaurin, J. Amyloid- β peptide assembly: a critical step in fibrillogenesis and membrane disruption. *Biophys. J.* **2001**, *80*, 1359–1371.
- (26) Burdick, D.; Soreghan, B.; Kwon, M.; Kosmoski, J.; Knauer, M.; Henschen, A.; Yates, J.; Cotman, C.; Glabe, C. Assembly and aggregation properties of synthetic Alzheimer's A4/beta amyloid peptide analogs. *J. Biol. Chem.* **1992**, *267*, 546–554.
- (27) Shao, H.; Jao, S.-c.; Ma, K.; Zagorski, M. G. Solution structures of micelle-bound amyloid β -(1–40) and β -(1–42) peptides of Alzheimer's disease. *J. Mol. Biol.* **1999**, *285*, 755–773.
- (28) Röder, K.; Wales, D. J. Energy Landscapes for the Aggregation of A β 17–42. *J. Am. Chem. Soc.* **2018**, *140*, 4018–4027.
- (29) Fink, A. L. Protein aggregation: folding aggregates, inclusion bodies and amyloid. *Fold. Des.* **1998**, *3*, R9–R23.
- (30) Knowles, T. P. J.; Vendruscolo, M.; Dobson, C. M. The amyloid state and its association with protein misfolding diseases. *Nat. Rev. Mol. Cell. Biol.* **2014**, *15*, 384–396.
- (31) Vendruscolo, M.; Dobson, C. M. Protein dynamics: Moore's law in molecular biology. *Curr. Biol.* **2011**, *21*, R68–R70.
- (32) Lue, L.-F.; Kuo, Y.-M.; Roher, A. E.; Brachova, L.; Shen, Y.; Sue, L.; Beach, T.; Kurth, J. H.; Rydel, R. E.; Rogers, J. Soluble amyloid β peptide concentration as a predictor of synaptic change in Alzheimer's disease. *Am. J. Pathol.* **1999**, *155*, 853–862.
- (33) So, M.; Hall, D.; Goto, Y. Revisiting supersaturation as a factor determining amyloid fibrillation. *Curr. Opin. Struct. Biol.* **2016**, *36*, 32–39.
- (34) Noji, M.; Samejima, T.; Yamaguchi, K.; et al. Breakdown of supersaturation barrier links protein folding to amyloid formation. *Commun. Biol.* **2021**, *4*, No. 120, DOI: 10.1038/s42003-020-01641-6.
- (35) Guo, Z. Amyloid hypothesis through the lens of A β supersaturation. *Neural Regen. Res.* **2021**, *16*, 1562–1563.
- (36) Czaplewski, C.; Karczyńska, A.; Sieradzan, A. K.; Liwo, A. UNRES server for physics-based coarse-grained simulations and prediction of protein structure, dynamics and thermodynamics. *Nucleic Acids Res.* **2018**, *46*, W304–W309.
- (37) Liwo, A.; Sieradzan, A. K.; Lipska, A. G.; Czaplewski, C.; Joung, I.; Żmudzińska, W.; Hałabis, A.; Oldziej, S. A general method for the derivation of the functional forms of the effective energy terms in coarse-grained energy functions of polymers. III. Determination of scale-consistent backbone-local and correlation potentials in the UNRES force field and force-field calibration and validation. *J. Chem. Phys.* **2019**, *150* (2019).
- (38) OPTIM: A program for geometry optimization and pathway calculations. <http://www-wales.ch.cam.ac.uk/software.html>, accessed: (accessed: February 17, 2025).
- (39) Wesolowski, P. A.; Sieradzan, A. K.; Winnicki, M. J.; Morgan, J. W.; Wales, D. J. Energy landscapes for proteins described by the UNRES coarse-grained potential. *Biophys. Chem.* **2023**, *303*, No. 107107, DOI: 10.1016/j.bpc.2023.107107.
- (40) Wales, D. J. Discrete path sampling. *Mol. Phys.* **2002**, *100*, 3285–3306.
- (41) Wales, D. J. Some further applications of discrete path sampling to cluster isomerization. *Mol. Phys.* **2004**, *102*, 891–908.
- (42) Noé, F.; Fischer, S. Transition networks for modeling the kinetics of conformational change in macromolecules. *Curr. Opin. Struct. Biol.* **2008**, *18*, 154–162.
- (43) Prada-Gracia, D.; Gómez-Gardenes, J.; Echenique, P.; Faló, F. Exploring the Free Energy Landscape: From Dynamics to Networks and Back. *PLoS Comput. Biol.* **2009**, *5*, No. e1000415.
- (44) Wales, D. J. Energy Landscapes: Some New Horizons. *Curr. Opin. Struct. Biol.* **2010**, *20*, 3–10.
- (45) Roterman, I.; Sieradzan, A.; Stapor, K.; Fabian, P.; Wesolowski, P.; Konieczny, L. On the need to introduce environmental characteristics in ab initio protein structure prediction using a coarse-grained UNRES force field. *J. Mol. Graphics Modell.* **2022**, *114*, No. 108166.
- (46) Lipska, A. G.; Antoniuk, A. M.; Wesolowski, P.; Warszawski, A.; Samsonov, S. A.; Sieradzan, A. K. Coarse-grained modeling of the calcium, sodium, magnesium and potassium cations interacting with proteins. *J. Mol. Model.* **2022**, *28*, No. 201.
- (47) Liwo, A.; Czaplewski, C.; Sieradzan, A. K.; Lipska, A. G.; Samsonov, S. A.; Murarka, R. K. Theory and practice of coarse-grained molecular dynamics of biologically important systems. *Biomolecules* **2021**, *11*, 1347.
- (48) Sieradzan, A. K.; Czaplewski, C. R.; Lubecka, E. A.; Lipska, A. G.; Karczyńska, A. S.; Gieldon, A. P.; Slusarz, R.; Makowski, M.; Krupa, P.; Kogut, M.; et al. Extension of the Unres package for physics-based coarse-grained simulations of proteins and protein complexes to very large systems. *Biophys. J.* **2021**, *120*, 83a–84a.
- (49) Antoniuk, A.; Biskupek, I.; Bojarski, K. K.; Czaplewski, C.; Gieldon, A.; Kogut, M.; Kogut, M. M.; Krupa, P.; Lipska, A. G.; Liwo, A.; et al. Modeling protein structures with the coarse-grained UNRES force field in the CASP14 experiment. *J. Mol. Graphics Modell.* **2021**, *108*, No. 108008.
- (50) Lensink, M. F.; Brysbaert, G.; Mauri, T.; Nadzirin, N.; Velankar, S.; Chaleil, R. A.; Clarence, T.; Bates, P. A.; Kong, R.; Liu, B.; et al. Prediction of protein assemblies, the next frontier: The CASP14-CAPRI experiment. *Proteins: Struct., Funct., Bioinf.* **2021**, *89*, 1800–1823.
- (51) Zaborowski, B.; Jagiela, D.; Czaplewski, C.; Hałabis, A.; Lewandowska, A.; Żmudzińska, W.; Oldziej, S.; Karczyńska, A.; Omieczynski, C.; Wirecki, T.; Liwo, A. A maximum-likelihood approach to force-field calibration. *J. Chem. Inf. Model.* **2015**, *55*, 2050–2070.
- (52) Liwo, A.; Baranowski, M.; Czaplewski, C.; Gołaś, E.; He, Y.; Jagiela, D.; Krupa, P.; Maciejczyk, M.; Makowski, M.; Mozolewska, M. A.; et al. A unified coarse-grained model of biological macromolecules based on mean-field multipole-multipole interactions. *J. Mol. Model.* **2014**, *20*, No. 2306.
- (53) Liwo, A.; He, Y.; Scheraga, H. A. Coarse-grained force field: general folding theory. *Phys. Chem. Chem. Phys.* **2011**, *13*, 16890–16901.
- (54) Liwo, A.; Khalili, M.; Czaplewski, C.; Kalinowski, S.; Oldziej, S.; Wachucik, K.; Scheraga, H. A. Modification and optimization of the united-residue (UNRES) potential energy function for canonical simulations. I. Temperature dependence of the effective energy function and tests of the optimization method with single training proteins. *J. Phys. Chem. B* **2007**, *111*, 260–285.
- (55) Kubo, R. Generalized cumulant expansion method. *J. Phys. Soc. Jpn.* **1962**, *17*, 1100–1120.
- (56) Shen, H.; Liwo, A.; Scheraga, H. A. An improved functional form for the temperature scaling factors of the components of the mesoscopic UNRES force field for simulations of protein structure and dynamics. *J. Phys. Chem. B* **2009**, *113*, 8738–8744.
- (57) Neelamraju, S.; Wales, D. J.; Gosavi, S. Protein energy landscape exploration with structure-based models. *Curr. Opin. Struct. Biol.* **2020**, *64*, 145–151.
- (58) Wales, D. J. Energy landscapes: some new horizons. *Curr. Opin. Struct. Biol.* **2010**, *20*, 3–10.
- (59) Noé, F.; Fischer, S. Transition networks for modeling the kinetics of conformational change in macromolecules. *Curr. Opin. Struct. Biol.* **2008**, *18*, 154–162.

- (60) Prada-Gracia, D.; Gómez-Gardeñes, J.; Echenique, P.; Falo, F. Exploring the free energy landscape: from dynamics to networks and back. *PLoS Comput. Biol.* **2009**, *5*, No. e1000415.
- (61) Liu, D. C.; Nocedal, J. On the limited memory BFGS method for large scale optimization. *Math. Program.* **1989**, *45*, 503–528.
- (62) Cerjan, C. J.; Miller, W. H. On finding transition states. *J. Chem. Phys.* **1981**, *75*, 2800–2806.
- (63) Henkelman, G.; Jónsson, H. A dimer method for finding saddle points on high dimensional potential surfaces using only first derivatives. *J. Chem. Phys.* **1999**, *111*, 7010–7022.
- (64) Munro, L. J.; Wales, D. J. Defect migration in crystalline silicon. *Phys. Rev. B* **1999**, *59*, 3969.
- (65) Kumeda, Y.; Wales, D. J.; Munro, L. J. Transition states and rearrangement mechanisms from hybrid eigenvector-following and density functional theory: application to C10H10 and defect migration in crystalline silicon. *Chem. Phys. Lett.* **2001**, *341*, 185–194.
- (66) Zeng, Y.; Xiao, P.; Henkelman, G. Unification of algorithms for minimum mode optimization. *J. Chem. Phys.* **2014**, *140*, No. 044115, DOI: 10.1063/1.4862410.
- (67) Trygubenko, S. A.; Wales, D. J. A doubly nudged elastic band method for finding transition states. *J. Chem. Phys.* **2004**, *120*, 2082–2094.
- (68) Sheppard, D.; Terrell, R.; Henkelman, G. Optimization methods for finding minimum energy paths. *J. Chem. Phys.* **2008**, *128*, No. 134106, DOI: 10.1063/1.2841941.
- (69) Mills, G.; Jónsson, H.; Schenter, G. K. Reversible work transition state theory: application to dissociative adsorption of hydrogen. *Surf. Sci.* **1995**, *324*, 305–337.
- (70) Jónsson, H.; Mills, G.; Jacobsen, K. W. Nudged elastic band method for finding minimum energy paths of transitions. In *Classical and quantum dynamics in condensed phase simulations*; World Scientific, 1998; pp 385–404.
- (71) Henkelman, G.; Uberuaga, B. P.; Jónsson, H. A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *J. Chem. Phys.* **2000**, *113*, 9901–9904.
- (72) Henkelman, G.; Jónsson, H. Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points. *J. Chem. Phys.* **2000**, *113*, 9978–9985.
- (73) Murrell, J. N.; Laidler, K. J. Symmetries of activated complexes. *Trans. Faraday Soc.* **1968**, *64*, 371–377.
- (74) Becker, O. M.; Karplus, M. The topology of multidimensional potential energy surfaces: Theory and application to peptide structure and kinetics. *J. Chem. Phys.* **1997**, *106*, 1495–1517.
- (75) Wales, D. J.; Miller, M. A.; Walsh, T. R. Archetypal energy landscapes. *Nature* **1998**, *394*, 758–760.
- (76) Ballard, A. J.; Das, R.; Martiniani, S.; Mehta, D.; Sagun, L.; Stevenson, J. D.; Wales, D. J. Energy landscapes for machine learning. *Phys. Chem. Chem. Phys.* **2017**, *19*, 12585–12603.
- (77) Rotkiewicz, P.; Skolnick, J. Fast procedure for reconstruction of full-atom protein models from reduced representations. *J. Comput. Chem.* **2008**, *29*, 1460–1465.
- (78) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696–3713.
- (79) Chapman, B.; Chang, J. Biopython: Python tools for computational biology. *ACM SIGBio Newsletter* **2000**, *20*, 15–19.
- (80) Mitternacht, S. FreeSASA: An open source C library for solvent accessible surface area calculations. *F1000Research* **2016**, *5*, 189 DOI: 10.12688/f1000research.7931.1.
- (81) Touw, W. G.; Baakman, C.; Black, J.; Te Beek, T. A.; Krieger, E.; Joosten, R. P.; Vriend, G. A series of PDB-related databanks for everyday needs. *Nucleic Acids Res.* **2015**, *43*, D364–D368.
- (82) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Meng, E. C.; Couch, G. S.; Croll, T. I.; Morris, J. H.; Ferrin, T. E. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci.* **2021**, *30*, 70–82.
- (83) Pelzer, H.; Wigner, E. Über die Geschwindigkeitskonstante von Austauschreaktionen. *Z. Phys. Chem.* **1932**, *15B*, 445–463.
- (84) Eyring, H. The Activated Complex and the Absolute Rate of Chemical Reactions. *Chem. Rev.* **1935**, *17*, 65.
- (85) Evans, M. G.; Polanyi, M. Some applications of the transition state method to the calculation of reaction velocities, especially in solution. *Trans. Faraday Soc.* **1935**, *31*, 875.
- (86) Forst, W. *Theory of Unimolecular Reactions*; Academic Press: New York, 1973.
- (87) Laidler, K. J. *Chemical Kinetics*; Harper & Row: New York, 1987.
- (88) Khalili, M.; Liwo, A.; Rakowski, F.; Grochowski, P.; Scheraga, H. A. Molecular dynamics with the united-residue model of polypeptide chains. I. Lagrange equations of motion and tests of numerical stability in the microcanonical mode. *J. Phys. Chem. B* **2005**, *109*, 13785–13797.
- (89) Sieradzian, A. K.; Sans-Duñó, J.; Sans-Duñó, J.; Lubecka, E. A.; Czaplewski, C.; Lipska, A. G.; Leszczyński, H.; Ocetkiewicz, K. M.; Proficz, J.; Czarnul, P.; Krawczyk, H. Optimization of parallel implementation of UNRES package for coarse-grained simulations to treat large proteins. *J. Comput. Chem.* **2023**, *44*, 602–625.
- (90) Redner, S. *A Guide to First-Passage Processes*; Cambridge University Press, 2001.
- (91) van Kampen, N. G. *Stochastic Processes in Physics and Chemistry*; Elsevier: North-Holland, Amsterdam, 1981.
- (92) Gillespie, D. T. *Markov Processes: An Introduction for Physical Scientists*; Academic Press: New York, USA, 1992.
- (93) Woods, E. J.; Wales, D. J. Analysis and interpretation of first passage time distributions featuring rare events. *Phys. Chem. Chem. Phys.* **2024**, *26*, 1640–1657.
- (94) Woods, E. J.; Kannan, D.; Sharpe, D. J.; Swinburne, T. D.; Wales, D. J. Analysing ill-conditioned Markov chains. *Philos. Trans. R. Soc. A* **2023**, *381*, No. 20220245.
- (95) Palmer, D. S.; Llinàs, A.; Morao, I.; Day, G. M.; Goodman, J. M.; Glen, R. C.; Mitchell, J. B. O. Predicting Intrinsic Aqueous Solubility by a Thermodynamic Cycle. *Mol. Pharmaceutics* **2008**, *5*, 266–279.
- (96) Skyner, R. E.; McDonagh, J. L.; Groom, C. R.; van Mourik, T.; Mitchell, J. B. O. A review of methods for the calculation of solution free energies and the modelling of systems in solution. *Phys. Chem. Chem. Phys.* **2015**, *17*, 6174–6191.
- (97) Spicher, S.; Grimme, S. Robust Atomistic Modeling of Materials, Organometallic, and Biochemical Systems. *Angew. Chem.* **2020**, *132*, 15795–15803.
- (98) Bannwarth, C.; Caldeweyher, E.; Ehlert, S.; Hansen, A.; Pracht, P.; Seibert, J.; Spicher, S.; Grimme, S. Extended tight-binding quantum chemistry methods. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2020**, *11* (2), No. e01493.
- (99) Grimme, S. Supramolecular binding thermodynamics by dispersion corrected density functional theory. *Chem.—Eur. J.* **2012**, *18*, 9955–9964.
- (100) Pracht, P.; Grimme, S. Calculation of absolute molecular entropies and heat capacities made simple. *Chem. Sci.* **2021**, *12*, 6551–6568.
- (101) Spicher, S.; Grimme, S. Efficient Computation of Free Energy Contributions for Association Reactions of Large Molecules. *J. Phys. Chem. Lett.* **2020**, *11*, 6606–6611.
- (102) Wesołowski, P. A.; Wales, D. J.; Pracht, P. Multilevel Framework for Analysis of Protein Folding Involving Disulfide Bond Formation. *J. Phys. Chem. B* **2024**, *128*, 3145–3156.
- (103) Ehlert, S.; Stahn, M.; Spicher, S.; Grimme, S. Robust and Efficient Implicit Solvation Model for Fast Semiempirical Methods. *J. Chem. Theory Comput.* **2021**, *17*, 4250–4261.
- (104) Pracht, P.; Bohle, F.; Grimme, S. Automated exploration of the low-energy chemical space with fast quantum chemical methods. *Phys. Chem. Chem. Phys.* **2020**, *22*, 7169–7192.
- (105) Pracht, P.; Grimme, S.; Bannwarth, C.; Bohle, F.; Ehlert, S.; Feldmann, G.; Gorges, J.; Müller, M.; Neudecker, T.; Plett, C.; Spicher, S.; Steinbach, P.; Wesołowski, P. A.; Zeller, F. CREST—A program for the exploration of low-energy molecular chemical space. *J. Chem. Phys.* **2024**, *160*, No. 114110.
- (106) Grimme, S.; Bohle, F.; Hansen, A.; Pracht, P.; Spicher, S.; Stahn, M. Efficient Quantum Chemical Calculation of Structure Ensembles

and Free Energies for Nonrigid Molecules. *J. Phys. Chem. A* **2021**, *125*, 4039–4054.

(107) Kipf, T. N.; Welling, M. Semi-supervised classification with graph convolutional networks, arXiv:1609.02907 (2016).

(108) Ezkurdia, I.; Grana, O.; Izarzugaza, J. M.; Tress, M. L. Assessment of domain boundary predictions and the prediction of intramolecular contacts in CASP8. *Proteins: Struct., Funct., Bioinf.* **2009**, *77*, 196–209.

(109) Chen, J.; Zheng, S.; Zhao, H.; Yang, Y. Structure-aware protein solubility prediction from sequence through graph convolutional network and predicted contact map. *J. Cheminf.* **2021**, *13*, No. 7.

(110) Niwa, T.; Ying, B.-W.; Saito, K.; Jin, W.; Takada, S.; Ueda, T.; Taguchi, H. Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of *Escherichia coli* proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 4201–4206.

(111) Hou, Q.; Kwasigroch, J. M.; Rooman, M.; Pucci, F. SOLart: a structure-based method to predict protein solubility and aggregation. *Bioinformatics* **2020**, *36*, 1445–1452.

(112) Han, X.; Zhang, L.; Zhou, K.; Wang, X. ProGAN: Protein solubility generative adversarial nets for data augmentation in DNN framework. *Comput. Chem. Eng.* **2019**, *131*, No. 106533.

(113) Hanson, J.; Paliwal, K.; Litfin, T.; Yang, Y.; Zhou, Y. Accurate prediction of protein contact maps by coupling residual two-dimensional bidirectional long short-term memory with convolutional neural networks. *Bioinformatics* **2018**, *34*, 4039–4045.

(114) Mount, D. W. Using BLOSUM in sequence alignments. *Cold Spring Harbor Protoc.* **2008**, *2008*, No. pdb.top39, DOI: 10.1101/pdb.top39.

(115) Taherzadeh, G.; Zhou, Y.; Liew, A. W.-C.; Yang, Y. Sequence-based prediction of protein-carbohydrate binding sites using support vector machines. *J. Chem. Inf. Model.* **2016**, *56*, 2115–2122.

(116) Altschul, S. F.; Madden, T. L.; Schäffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402.

(117) Steinegger, M.; Meier, M.; Mirdita, M.; Vöhringer, H.; Haunsberger, S. J.; Söding, J. HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinf.* **2019**, *20*, No. 473.

(118) Suzek, B. E.; Huang, H.; McGarvey, P.; Mazumder, R.; Wu, C. H. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* **2007**, *23*, 1282–1288.

(119) Mirdita, M.; Von Den Driesch, L.; Galiez, C.; Martin, M. J.; Söding, J.; Steinegger, M. Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Res.* **2017**, *45*, D170–D176.

(120) Meiler, J.; Müller, M.; Zeidler, A.; Schmäsche, F. Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks. *J. Mol. Model.* **2001**, *7*, 360–369.

(121) Heffernan, R.; Yang, Y.; Paliwal, K.; Zhou, Y. Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility. *Bioinformatics* **2017**, *33*, 2842–2849.

(122) Shimizu, Y.; Inoue, A.; Tomari, Y.; Suzuki, T.; Yokogawa, T.; Nishikawa, K.; Ueda, T. Cell-free translation reconstituted with purified components. *Nat. Biotechnol.* **2001**, *19*, 751–755.

(123) Strodel, B. Energy landscapes of protein aggregation and conformation switching in intrinsically disordered proteins. *J. Mol. Biol.* **2021**, *433*, No. 167182.

(124) Wales, D. J. Exploring energy landscapes. *Annu. Rev. Phys. Chem.* **2018**, *69*, 401–425.

(125) Crescenzi, O.; Tomaselli, S.; Guerrini, R.; Salvadori, S.; D'Ursi, A. M.; Temussi, P. A.; Picone, D. Solution structure of the Alzheimer amyloid β -peptide (1–42) in an apolar microenvironment: Similarity with a virus fusion domain. *Eur. J. Biochem.* **2002**, *269*, 5642–5648.

(126) Lattanzi, V.; Bernfur, K.; Sparr, E.; Olsson, U.; Linse, S. Solubility of A β 40 peptide. *JCIS Open* **2021**, *4*, No. 100024.

(127) Bernstein, S. L.; Wyttenbach, T.; Baumketner, A.; Shea, J.-E.; Bitan, G.; Teplow, D. B.; Bowers, M. T. Amyloid β -protein: monomer structure and early aggregation states of A β 42 and its Pro19 alloform. *J. Am. Chem. Soc.* **2005**, *127*, 2075–2084.

(128) Han, M.; Hansmann, U. H. Replica exchange molecular dynamics of the thermodynamics of fibril growth of Alzheimer's A β 42 peptide. *J. Chem. Phys.* **2011**, *135*, No. 065101, DOI: 10.1063/1.3617250.

(129) Ball, K. A.; Phillips, A. H.; Wemmer, D. E.; Head-Gordon, T. Differences in β -strand populations of monomeric A β 40 and A β 42. *Biophys. J.* **2013**, *104*, 2714–2724.

(130) Thirumalai, D.; Klimov, D.; Dima, R. Emerging ideas on the molecular basis of protein and peptide aggregation. *Curr. Opin. Struct. Biol.* **2003**, *13*, 146–159.

(131) Rojas, A. V.; Liwo, A.; Scheraga, H. A. A study of the α -helical intermediate preceding the aggregation of the amino-terminal fragment of the β amyloid peptide (A β 1–28). *J. Phys. Chem. B* **2011**, *115*, 12978–12983.

(132) Kapurniotu, A.; Buck, A.; Weber, M.; Schmauder, A.; Hirsch, T.; Bernhagen, J.; Tatarek-Nossol, M. Conformational restriction via cyclization in β -amyloid peptide A β (1–28) leads to an inhibitor of A β (1–28) amyloidogenesis and cytotoxicity. *Chem. Biol.* **2003**, *10*, 149–159.

(133) Schwarze, B.; Korn, A.; Höfling, C.; Zeitschel, U.; Krueger, M.; Roßner, S.; Huster, D. Peptide backbone modifications of amyloid β (1–40) impact fibrillation behavior and neuronal toxicity. *Sci. Rep.* **2021**, *11*, No. 23767.

(134) Das, A. K.; Rawat, A.; Bhowmik, D.; Pandit, R.; Huster, D.; Maiti, S. An early folding contact between Phe19 and Leu34 is critical for amyloid- β oligomer toxicity. *ACS Chem. Neurosci.* **2015**, *6*, 1290–1295.

(135) Korn, A.; McLennan, S.; Adler, J.; Krueger, M.; Surendran, D.; Maiti, S.; Huster, D. Amyloid β (1–40) toxicity depends on the molecular contact between phenylalanine 19 and leucine 34. *ACS Chem. Neurosci.* **2018**, *9*, 790–799.

(136) Itoh, S. G.; Yagi-Utsumi, M.; Kato, K.; Okumura, H. Key residue for aggregation of amyloid- β peptides. *ACS Chem. Neurosci.* **2022**, *13*, 3139–3151.

(137) Chakraborty, D.; Chebaro, Y.; Wales, D. J. A multifunnel energy landscape encodes the competing α -helix and β -hairpin conformations for a designed peptide. *Phys. Chem. Chem. Phys.* **2020**, *22*, 1359–1370.