Sampling Conformational Ensembles of Highly Dynamic Proteins via Generative Deep Learning

Talant Ruzmetov[#a], Ta I Hung[#a,b], Saisri Padmaja Jonnalagedda[#c], Si-han Chen[a], Parisa Fasihianifard[a], Zhefeng Guo*[d], Bir Bhanu*[b,c] and Chia-en A. Chang*[a,b]

[a] Department of Chemistry, [b] Department of Bioengineering, [c] Department of Electrical and Computer Engineering, University of California, Riverside, CA92521
[d] Department of Neurology, Brain Research Institute, University of California, Los Angeles, CA 90095

Email: zhefeng@ucla.edu, bhanu@ee.ucr.edu and chiaenc@ucr.edu

Abstract

Proteins are inherently dynamic, and their conformational ensembles are functionally important in biology. Large-scale motions may govern protein structure–function relationship, and numerous transient but stable conformations of intrinsically disordered proteins (IDPs) can play a crucial role in biological function. Investigating conformational ensembles to understand regulations and disease-related aggregations of IDPs is challenging both experimentally and computationally. In this paper first an unsupervised deep learning-based model, termed Internal Coordinate Net (ICoN), is developed that learns the physical principles of conformational changes from molecular dynamics (MD) simulation data. Second, interpolating data points in the learned latent space are selected that rapidly identify novel synthetic conformations with sophisticated and large-scale sidechains and backbone arrangements. Third, with the highly dynamic amyloid-$\beta_{1\text{-}42}$ (Aβ42) monomer, our deep learning model provided a comprehensive sampling of Aβ42's conformational landscape. Analysis of these synthetic conformations revealed conformational clusters that can be used to rationalize experimental findings. Additionally, the method can identify novel conformations with important interactions in atomistic details that are not included in the training data. New synthetic conformations showed distinct sidechain rearrangements that are probed by our EPR and amino acid substitution studies. The proposed approach is highly transferable and can be used for any available data for training. The work also demonstrated the ability for deep learning to utilize learned natural atomistic motions in protein conformation sampling.

Introduction

Proteins are complex and have dynamic properties. Conformational ensembles of proteins are essential for performing biological processes, including enzyme activities, protein folding, protein–ligand binding, and protein aggregation [1,2]. Characterization of protein conformations allows researchers to understand protein function, activity, and mechanisms. However, the tasks may be daunting and are even more challenging for investigating intrinsically disordered proteins (IDPs), for which numerous conformations can be functionally important. Some IDPs, such as amyloid-$\beta_{1\text{-}42}$ (Aβ42), are related to a number of protein aggregation-related diseases, including Alzheimer's disease [3] . Revealing important conformations with atomistic details is crucial to

fully understand the function of a protein and its regulation involved in basic science, therapeutics design and biotechnology.

Experimental techniques determine protein structures and probe their dynamics, and bioinformatics tools such as AlphaFold2 and RoseTTAFold can accurately predict structures from given protein sequences [4–6]. However, these techniques typically provide a handful of structures with limited information on dynamics. Advanced computational methods are widely used to investigate protein dynamics and sample protein conformational ensembles. A powerful method is all-atom molecular dynamic (MD) simulations, which is based on physical principals to sample protein conformations and model their dynamics [7,8]. However, many stable protein conformations are separated by free energy barriers: an MD run can take excessively long computational time (i.e., beyond microsecond timescale) to cross energy barriers for sampling various conformations. Other methods developed to tackle sampling limitations include conformational search methods, MD-based enhanced sampling techniques, and Monte Carlo simulations [6,9–12]. A popular analysis tool, principal component analysis (PCA), has been used to extract information from sampled protein conformations to guide conformational search [13,14]. Low-mode based conformational search methods use normal mode analysis (NMA) to perturb given structures along their low-mode eigenvectors to generate new conformations [15–18]. Although these sampling approaches effectively extend the time scale of MD simulations, efficiently obtaining adequate conformations for highly flexible protein systems and/or overcoming significant energy barriers is still challenging.

Artificial intelligence (AI) provides an alternative approach for accelerating the generation of protein conformational ensembles. Novel machine learning/deep learning (ML/DL) approaches analyze simulation results to further guide conformation search with enhanced sampling techniques [19–27]. ML/DL optimizes coarse-grain models to speed up conformation transitions with preserved atomistic interactions [28,29]. Several DL models accelerate protein conformational searches, such as training neural networks to model the distribution of conformational ensembles with generative modeling [30–39]. By training with conformations obtained from MD simulations, these approaches can quickly generate conformational ensembles and bypass the kinetic barriers, the bottleneck of MD simulation.

Generative models study conformations of chemical compounds and proteins in various settings; examples include coarse-grained models and protein conformations presented by a backbone [28,40]. Ideally, the generative models can find new and thermodynamically stable protein conformational ensembles not seen in the training dataset obtained from MD simulations, as demonstrated in recent publications focusing on sampling conformations for IDPs [41–45]. Nevertheless, proteins have complex conformational ensembles, and many conformational transitions are driven by backbone motions and, more importantly, sidechain interactions such as salt bridges, hydrogen bonds, and hydrophobicity. As a result, to find new protein conformations, features that can accurately describe all possible motions in atomistic details are critical in generative AI models.

In this study, we present a generative DL model, termed Internal Coordinate Net (ICoN), which is trained on datasets from MD simulations to identify new conformations efficiently and accurately for highly flexible protein systems. ICoN uses novel features and an atomistic bond-angle-torsion (BAT)-based vector representation, vBAT, to smoothly rotate various dihedral angles that lead to new conformations. Our model learns the physical principles that govern molecular motions, and the 3D latent space contains the transformations of various degrees of freedom (DOF), mainly dihedral rotations, to allow for efficient conformational searches. Essential protein motions analyzed by dihedral Principal Component Analysis (dPCA) are clearly revealed in the latent space, and generating new conformations is achieved by data-point interpolation within the latent space. Using a fraction (1%) of MD data as a training set, we demonstrated that ICoN rapidly found thousands of new, conformationally distinct and thermodynamically stable conformations for αB-crystallin57-69 and the Aβ42 monomer (**Figure 1**) in a few minutes using a single gaming GPU card. Post-analysis of our novel synthetic Aβ42 monomer conformations revealed various new conformations, including new Arg5-Ala42 contacts and a new Asp23-Lys28 salt bridge not seen in the MD training set but reported in existing publications. Our synthetic conformations covered reported structures that were suggested to be less toxic or more prone to oligomerization [46–52]. Our model provides a computationally efficient approach to identify new conformations for highly flexible protein systems. The novel synthetic Aβ42 monomer conformations also provide insights into Aβ42 oligomerization.

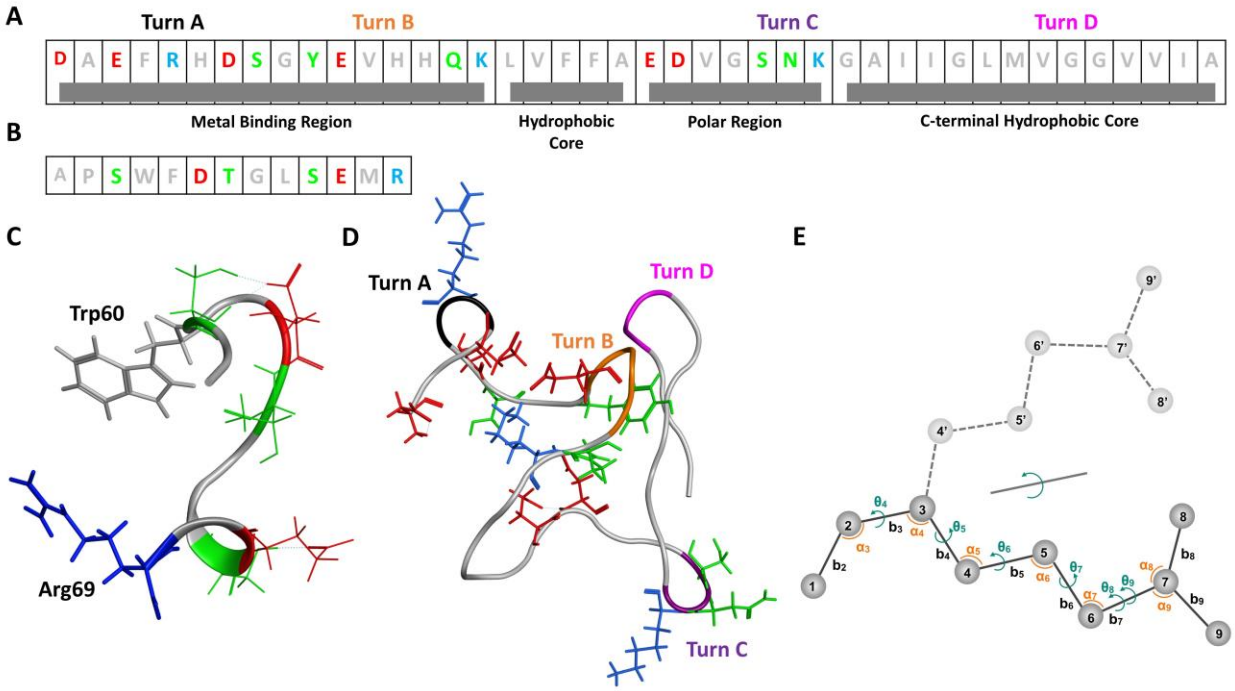

Figure 1. (A) Sequence of 42-residue amyloid beta peptide (Aβ42). (B) Sequence of 13-residue αB-crystallin57-69. (C) A synthetic conformation of αB-crystallin57-69 with intramolecular charge and aromatic ring interactions between Trp60 and Arg69. Red: negative charge residues; Blue: positive charge residues; Green: polar; Grey: neutral and non-polar residues. (D) A synthetic conformation of Aβ42 with all 4 turns A-D. (E) Classical internal bond-angle-torsion (BAT) coordinate representation (Z-matrix) for a small molecule. Atoms 1−3 are termed root atoms for this molecule. Atoms i > 3 are defined by (bi, ai, θi), where b, a and t are bond, angle and torsions, respectively. For example, atom 4 is presented by (b4, a4, θ4). The dashed light gray line shows the smooth rotation of the bond between atoms 2 and 3 (θ4).

Results

ICoN model architecture, training, and validation.

Molecular representation using the vBAT coordinate. The ICoN model is a latent space-guided generative AI model trained on protein conformational ensembles to learn physical properties that govern conformational transitions in MD simulations (**Figure 2**). Because the dihedral rotations are the major motions to determine conformations, using BAT to accurately describe the concerted motions of multiple dihedral rotations is critical. Dihedral rotations have a periodicity issue, so we used a vector called vBAT to avoid periodicity (**Figure 2B** and details in

Supplementary Material). All-atom vBAT internal coordinate representation is inherently equivariant to rotations and translations and can exactly convert vBAT vectors to Cartesian coordinates without any additional approximation. However, coordinate conversions can be time-consuming. Thus, our model implemented GPU acceleration to perform coordinate transformations; converting 5,000 frames from vBAT to Cartesian coordinates took < 0.2 s for Aβ42. **Figure S2** demonstrates the efficiency in coordinate transformations that successfully avoided a computation bottleneck with the use of internal coordinates.

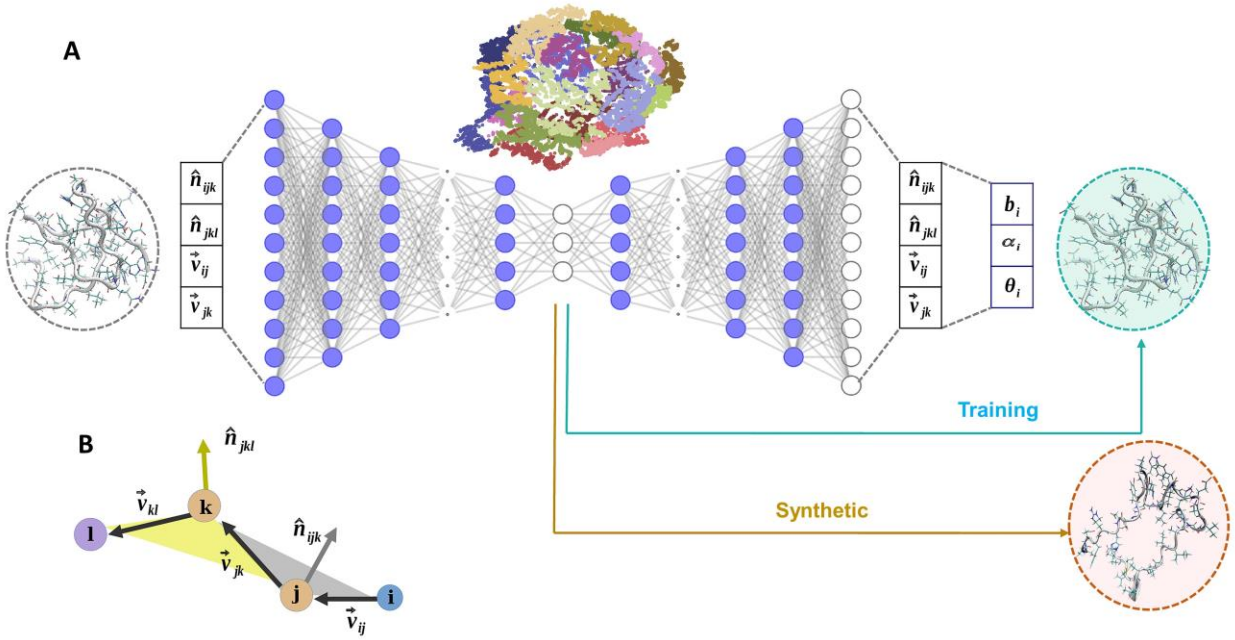

Figure 2. Overview of the Internal Coordinate Net (ICoN) network architecture. (A) The input data are thousands of protein conformations represented by BAT-based vBAT features. The molecular conformations have 3N-6 degrees of freedom and 4x3(N-3) dimension features, where N is the number of atoms. The high dimensions are compressed by a series of fully connected layers into a 3D latent space using an encoder. The entire process is then inverted back to the original dimensions using a decoder and then converted back to Cartesian coordinates. (B) An illustration of vBAT features using 4 atoms with indexes, i, j, k, and l. The vectors connecting the atoms $\vec{v}_{ij}$, $\vec{v}_{jk}$, $\vec{v}_{kl}$. . $\hat{n}_{ijk}$ and $\hat{n}_{jkl}$ are unit vectors normalized to the planes constructed by atoms i-j-k and j-k-l, respectively.

<u>ICoN network architecture and training.</u> The neural network is based on autoencoder architecture (**Figure 2A**). Classical BAT coordinates (the Z matrix) have 3N-6 DOF for presenting the internal motions of a molecule, where N is the number of atoms, and the 6

external translation and rotation DOF are eliminated. Because vBAT uses the vectors instead of one value for each bond, angle and dihedral DOF, we have 4x3x(N-3) features for each molecule. As a result, αB-crystallin57-69 and Aβ42 have 2376 and 7488 features, respectively. The hyperparameter training in the encoder resulted in 7 layers to reduce a large dimension representation into 3D, so each molecular conformation can be presented in the 3D latent space (**Figures 2A** and **S3**). The decoding process brought the representation from 3D back to the original dimensions with the same number of layers but in reverse order. Each conformation is converted from vBAT to Cartesian coordinates for further analysis. Throughout the iterative training process, in addition to learning a relationship between a vast number of features, the network also learns how to meaningfully compress the representation into the lower-dimensional 3D latent space by performing nonlinear dimensionality reduction.

Notably, although this study used 3D latent space for easy observation of conformation distribution and visualization of interpolation, the network can reduce 4x3x(N-3) features to any dimension desired by users. We used 3D to achieve efficient training and reduced memory consumption as well. It took 4 min with one NVIDIA 1080Ti GPU card to perform 15000 epochs for successful training with 10,000 frames of Aβ42.

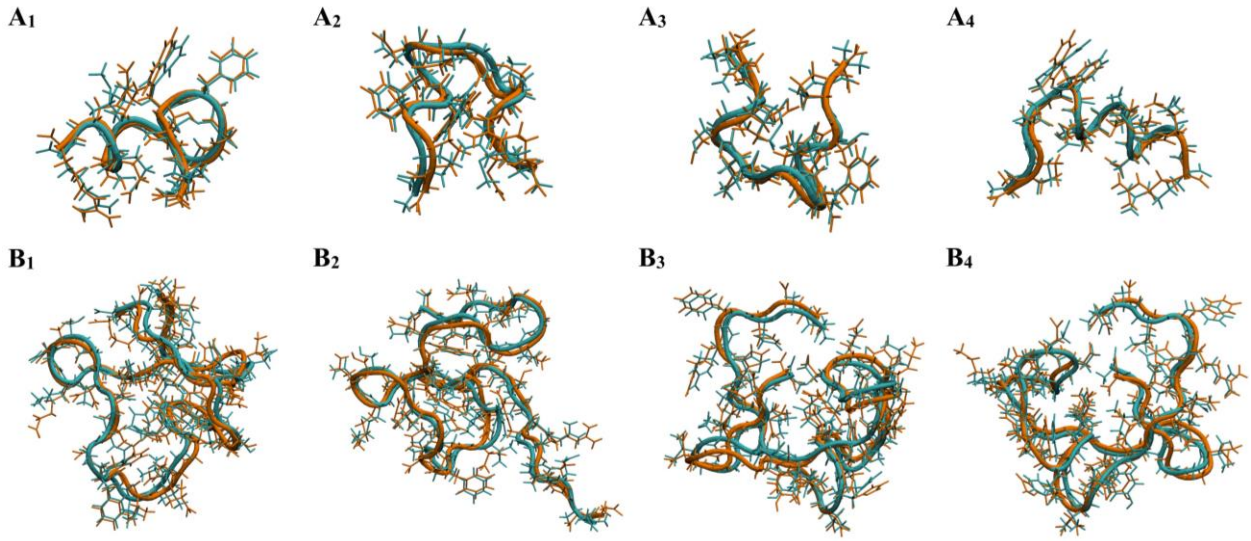

Figure 3. Comparisons of pairs of original and reconstructed conformations: four representative pairs of each system. The reconstructed (orange) and original (cyan) conformations all have heavy atom root mean square deviation (RMSD) within 0.9 Å for αB-cristallin57-69 (A1 to A4) and <1.3 Å for Aβ42 (B1 to B4).

Validation of the models by conformation reconstruction. A successful trained model should accurately reconstruct a protein conformation from the reduced 3D representation to the original conformation. To evaluate the accuracy, we presented conformations from our validation set using the 3D representation and then reconstructed them back to atomistic Cartesian representation and computed the root mean square deviation (RMSD) of all heavy atoms between the 2 conformations, the original and reconstructed ones. A few representative conformations are illustrated in **Figure 3**: αB-crystallin57-69 has RMSD < 0.9Å and Aβ42 has RMSD < 1.3Å. As shown in **Table 1**, with our model, both backbone and sidechains conformations can be accurately reconstructed back to the original input structure, ambient after large dimensionality reduction in the latent space. We also performed a more derailed comparisons between the original and reconstructed conformations, such as the distribution of each dihedral rotation and their correlations, to ensure that all the properties were accurately reproduced (**Figures S4-S7**). Notably, both αB-crystallin57-69 and Aβ42 exhibit large-scale protein motions included in our training set. The validation demonstrates the robustness of our model to learn a diverse range of conformational states with high precision, and we can describe small sidechain rotations accurately.

MD simulations		ICoN model construction/validation		Synthetic Conformations				
					Step 1	Step 2	Step 3	Step 4
MD index	force field & initial structure (PDB)	# of frames for training/validation	Validation of reconstructed conformations, average heavy atom RMSD (backbone RMSD)	Model index	# of confs from initial interpolation in the latent space	# of confs within energy cutoff	# of distinct confs (repeat eliminated)	# of novel synthetic confs (repeat with MD confs eliminated)
αB-crystalline (0.5 μs)								
Seed 1	14sb (Alpha Fold2 predicted structure)	5,000/5,000	2.58 Å (1.84 Å)	Model I	99990 (100%)	97594 (97.6%)	43484 (43.5%)	38425 (38.4%)
Seed 2		10,000 for comparison with synthetic conformations obtained by Model I						
Aβ42 (1 μs for all runs)								
MD Run1	ff14IDPSFF (2NAO)	10,000/10,000	3.9 Å (3.6 Å)	Model 1	199990 (100%)	187487 (93.7%)	150713 (75.3%)	127673 (63.8%)
MD Run2	ff14IDPSFF (1Z0Q)	10,000/10,000	6.18 Å (5.96 Å)	Model 2	199990 (100%)	174012 (87%)	148188 (74%)	145925 (72.9%)
MD Run3	ff14SB (2NAO)	5,000/5,000	3.56 Å (3.26 Å)	Model 3	99990 (100%)	91279 (91.3%)	80001 (80%)	N/A
MD Run4	ff14SB (1Z0Q)	5,000/5,000	4.7 Å (4.4 Å)	Model 4	99990 (100%)	94449 (94.4%)	66205 (66.2%)	48320 (48.3%)
MD Run1.1	ff14IDPSFF (2NAO)	10,000 for comparison with synthetic conformations obtained by Model 1						
MD Run1.2	ff14IDPSFF (2NAO)	10,000 for comparison with synthetic conformations obtained by Model 1						
MD Run1.3	ff14IDPSFF (2NAO)	10,000 for comparison with synthetic conformations obtained by Model 1						

Table 1. Summary of molecular dynamics (MD) simulations, numbers of MD frames used for ICoN model training and validation, and numbers of conformations generated from ICoN models in each step. Energy cutoff was -200 and -400 kcal/mol for αB-crystalline and Aβ42. Two conformations are treated as repeats when the computed heavy atom RMSD is smaller than 1 Å in Step 3. In Step 4, the RMSD cutoff is 1 Å and 1.2 Å for αB-crystalline and Aβ42, respectively.

Learned physics in the latent space of αB-cristallin57-69

Because our ICoN model learns important dihedral rotations that determine molecular conformations, the latent space stores information that leads to smooth conformational transitions between 2 datapoints, such as various sets of concerted dihedral rotations. Moreover, we can visualize conformation transitions and clusters and can perform data analysis in the 3D latent space (**Figures 4 and S9**). We demonstrated that the DL model learned the physics governing molecular motions, and the information captured in the latent space can be effectively utilized.

Essential motions of αB-cristallin57-69 revealed in the latent space. We selected 2 conformations in the latent space, Conf indexes #50 and #53, from **Figures 4** to examine the conformational transitions in the latent space. Note that the MD saved one frame every 1 ps, and we re-saved a frame every 100 ps when preparing our training and validation sets. Therefore, Conf indexes #50 and #53 (2 red dots in **Figure 4D**) are significantly different; they present 2 end points from 300-ps time-period MD. To illustrate the conformation transition, we superimposed 4 conformations (indexes #50 to #53 **Figure 4A**) and found that the first 6 residues were maintained in a helix secondary structure, and the middle region around Leu9 showed minor but important fluctuations to twist the center of the peptide, whereas the flexible tail (residues 7 to 13) exhibited greater dynamics. A correlated sidechain movement (i.e., sidechains of Leu9 and Met12) was observed during the transitions as well.

In the latent space, we found that the 300-ps interval MD trajectory sampled conformational jiggling during the transitions (light blue dots in **Figure 4D**). To extract the essential motions from the conformation fluctuation, we applied PCA with BAT coordinates to analyze the major motions from the 300-ps time-period MD to quantify the motions. The first Principal Component (PC) revealed the major motions, which removed unessential fluctuations (**Figure 4B**). Presenting these essential motions in the latent space resulted in a very smooth curve (green dots in **Figure 4D**). Note that in the latent space, the PC motions appeared as a non-linear curve, not a straight line. **Figure S9C** exhibits more smooth curves for motions following the second PC and the first + second PC modes, so the latent space recorded similar information obtained from PCA. As for PCA, our ICoN learned the important DOF that led to conformational transition (i.e., various sets of concerted dihedral rotations), and the latent space retained the information from DL. Therefore, interpolation between 2 dots in the latent space can efficiently sample protein motions that are led by movements from critical DOFs. As illustrated in **Figure 4**, the interpolation (black dots in **Figure 4D**) revealed the learned motions, which also showed a more rigid helix secondary structure in the first 6 residues and the fluctuations in the middle region around Leu9 to result in more distinct conformations with smooth transitions (**Figure 4C**). Importantly, our model accurately identified the concerted motions shown in both MD and PCA. These essential motions were in good agreement with those detected by dihedral PCA and our observation of the MD trajectory. Of note, our vBAT and model can precisely learn and present very small conformational changes, such as the methyl rotation of Met12.

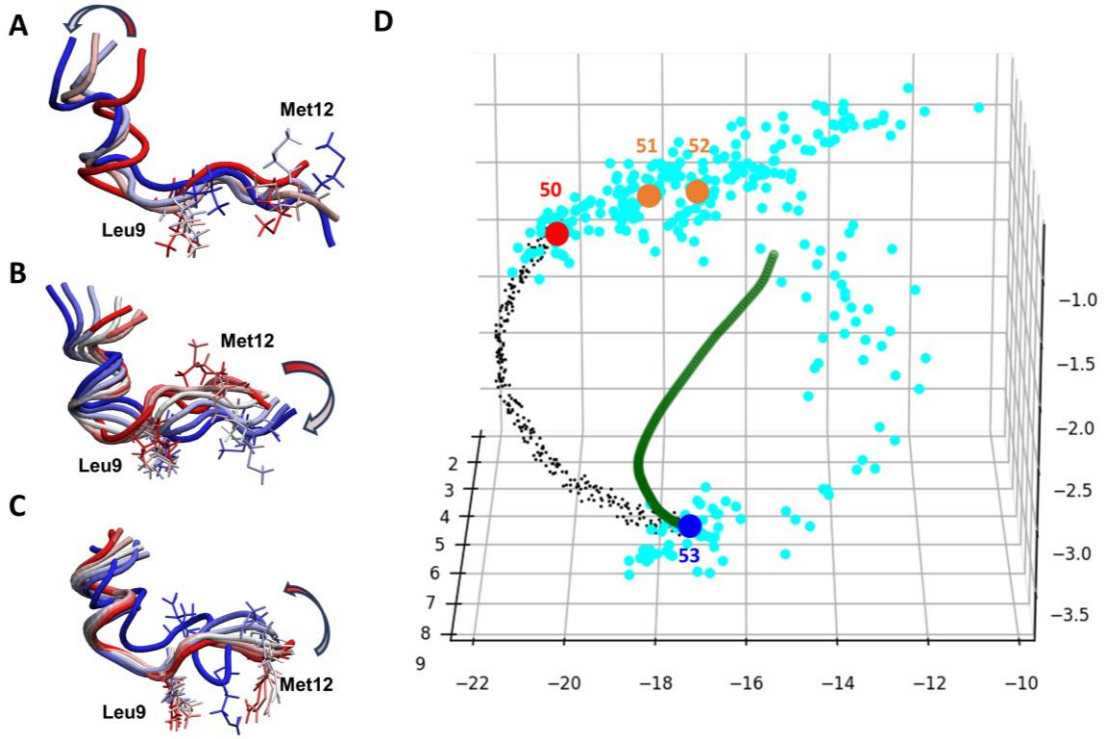

Figure 4. Conformational transitions of αB-crystallin57-69 peptide. (A) Super-imposed conformations used for DL training (Conf indexes #50 to #53). (B) Conformational transitions led by distortions using the first dihedral principal component (PC) mode. (C) Conformational transitions led by interpolation between training Conf indexes #50 (red dot) and #53 (dark blue dot). Leu9 and Met12, which have the most fluctuations, are presented in licorice. (D) 3D latent space from the ICoN model. Light blue dots present conformations saved every 1 ps in MD, with a total of 300 dots (300-ps simulation length). The MD frames were re-saved every 100 ps for training, and 4 frames are presented in the plot (red, blue and orange dots). The green dot line illustrates conformation distortions using the first PC mode, with the conformations shown in (B). Note that the PC analysis involved using the 300-ps time-period not the whole 500-ns MD run. Inspired by the green curve, non-linear interpolation between Conf indexes #50 and #53 was used for conformational search and shown as black dots.

Interpolation-generated synthetic conformations of αB-cristallin57-69. Because ICoN identified DOF that determine conformational changes, we used the latent space interpolation as a conformational search engine to find more synthetic conformations. The search was achieved by interpolation between 2 dots with consecutive indexes (i.e., Conf indexes #10 and #11). Following the essential motions, we anticipated finding both new conformations and existing ones sampled by MD. Using the same non-linear interpolation function illustrated in **Figure 4D**, 10 points were generated for each pair, yielding 99,990 synthetic conformations from

interpolating a total of 9,999 pairs of dots. We first eliminated high-energy conformations, then each new synthetic conformation was compared with its predecessors and eliminated if it was a repeat, yielding 43,484 distinct conformations (**Table 1**). We further compared these synthetic conformations with the raw MD data with 500,000 frames to identify 38,425 novel synthetic conformations.

We also examined the stability of the synthetic conformations quantified by their conformational energies using the molecular mechanics/generalized Born and surface areas (MM/GBSA) calculations. Conformations sampled from MD and our deep learning model have similar energy distribution (**Figure 5A**), which validates our synthetic conformations as being thermodynamically stable. In addition, we also compared our synthetic conformations with conformations obtained by another MD run (**Table 1**, Seed 2); none of the conformations from this MD was used in our training set. Our interpolation sampling found similar conformations found by another MD run **Figure 5B**, which demonstrates the efficiency and accuracy of the use of the latent space for conformational sampling.

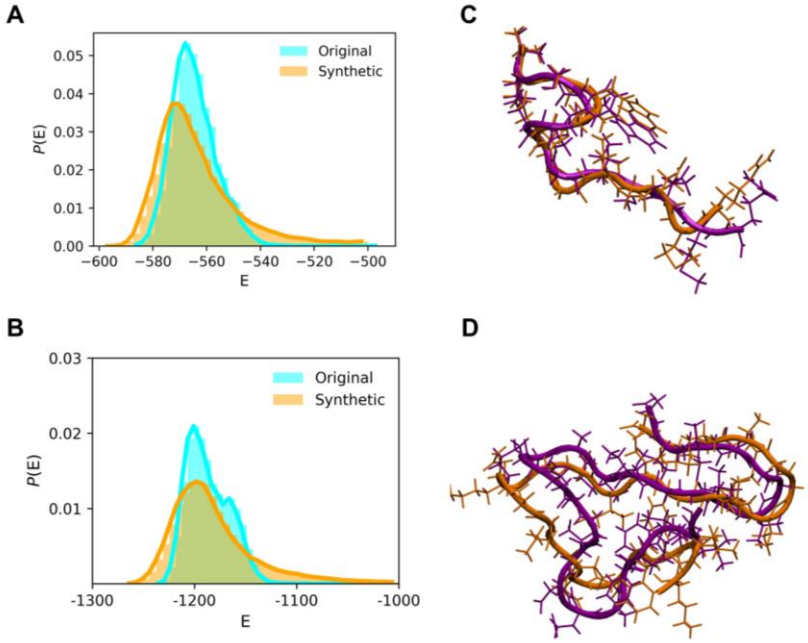

Figure 5. Examination of the stability of synthetic conformations using energy distribution and known conformations. Comparison of energy distribution from conformations of the original MD run (cyan) and synthetic (orange) conformations for (A) αB-crystallin57-69 and (B) Aβ42. Synthetic conformation (orange) similar to a known conformation sampled by other MD runs (purple) for (C) αB-crystallin57-69 and (D) Aβ42. The MD runs used for comparison but not used for training are Seed 2 and MD Run1.1-3 in Table 1.

Latent space interpolation-generated synthetic conformations of Aβ42

The Aβ42 monomer exhibits a broad spectrum of conformations, from random coil to more structured α-helical and β-sheet conformations. Sampling strategies that can accurately model sidechain motions are crucial because the sidechain arrangements govern the intramolecular attractions that lead to various local turns and pre-organized shapes for subsequent oligomerization. Using MD Runs 1 to 4 (**Table 1**), we used the same strategy used for αB-cristallin57-69 to interpolate 2 consecutive dots in the latent space for obtaining synthetic conformations. The 4 MD runs were initiated by a different Aβ42 monomer structure and/or with a different force field. To elucidate the search efficacy, we plotted the distribution with coordinates of Rg and RMSD (**Figure 6**) for MD Run1-4 and our synthetic conformations from Model 1-4 using these MD runs for training. The search using generative AI efficiently found many new conformations without the need for lengthy MD simulations (**Table 1)**.

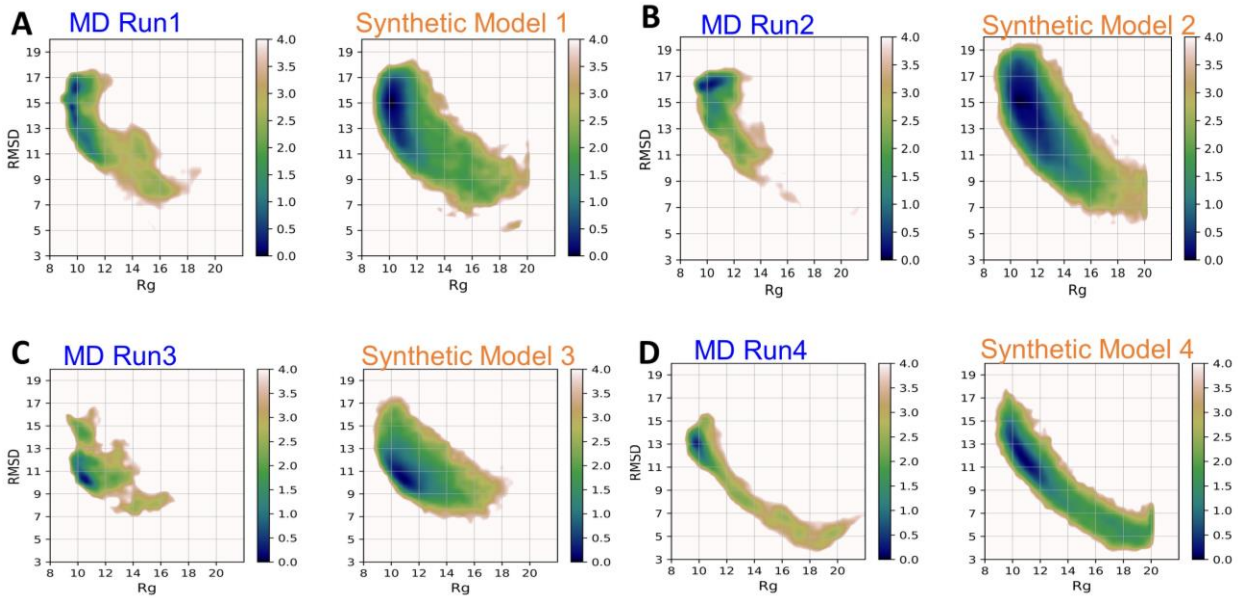

Figure 6. Free energy profile on the 2D space of RMSD and radius of gyration (Rg) for Aβ42. For each pair of comparisons, the left depicts the conformation distribution from the MD run with frames resaved every 10 ps and the right shows the distribution from distinct synthetic conformations. Models 1-4 were trained using dataset obtained from (A) MD Run1, (B) MD Run2, (C) MD Run3; and (D) MD Run4, respectively. MD Runs 1 and 2 used the ff14IDP force field, and MD Runs 3 and 4 used the classical Amber ff14SB force field. The units of the free-energy landscapes are kcal/mol.

To validate that our synthetic conformations are thermodynamically stable, we again plotted energy distribution of synthetic conformations found in Model 1 (**Figure 5C**). We also compared the synthetic conformations to another 3 MD runs using the same initial structure and IDP force field (MD Run1.1-1.3) to examine whether our search could find those sampled by other MD simulations. Although the Aβ42 monomer has numerous conformations, we could identify similar conformations (**Figure 5C**). In addition to checking the total conformational energy, we compared each energy component to verify that both bonded and non-bonded terms have energy distribution similar to those modeled by MD runs (**Figure S10 and S11**). Notably, the energy distribution from the synthetic conformations is relatively broader, thereby suggesting a greater diversity in the sampled conformations, which cover high and low energy regions. The energy comparison underscores the high quality of structural integrity of the synthetic conformations.

Analysis and biological implications of novel synthetic Aβ42 conformation ensembles

Obtaining Aβ42 monomer conformations is a critical step for understanding the mechanisms of initial encounters and interactions between monomers, which lead to subsequent oligomerization and fibrillization. Because the oligomerization steps can be highly sensitive to different environments (e.g., different membrane, existing fibrils, or ion concentrations), substantially different monomer conformations may initiate aggregation using different mechanisms in various environments. Several experiments and modeling work also suggest that some conformations are prone to be aggregated or non-toxic [46–50]. Because of various experimental results regarding the monomer structure–function relationship in Aβ42 aggregation, we focus on structures with salt bridges and R5-A42 contacts to demonstrate the utility of the conformational search results from our generative AI model.

Of note, because our method efficiently sampled numerous low-energy monomer conformations, we used synthetic conformations found using Model 1 (**Table 1**) to demonstrate the biological importance of these conformations revealed by our ICoN model.

Major conformation turns in a global tertiary structural ensemble. The 127,673 distinct conformations (**Table 1**) are presented in the pairwise residue contact map (**Figure 7A**). Four

local bends are marked as turn A (F4-H6), turn B (E11-H14), turn C (S26-K28), and turn D (V36-G38). We found novel synthetic conformations with all 4 turns (**Figure 1D**). Turns A and B are located near the N-terminal, which is highly flexible and believed to be the metal binding region under abnormal physiological conditions (residues D1 to K16) [53]. Turn C situates close to a central polar region that contains many hydrophilic residues (E22 to G29). Turn D resides in the C-terminal hydrophobic region (A30 to A42). These turns directly correlate with the local and/or overall Aβ42 conformations, where some turns are commonly reported. For example, turn regions spanning the center residues (K16-G25), in particular turn C referenced as "β-turn-β" motif, have been reported in Aβ42 aggregates [54,55]. Although the conformations/turns and their population are obtained using only Model 1, they provide an overview of spatial arrangements of the Aβ42 monomer. Notably, although Aβ42 is intrinsically disordered, the protein sequence still can lead to many low-energy and highly popular conformations.

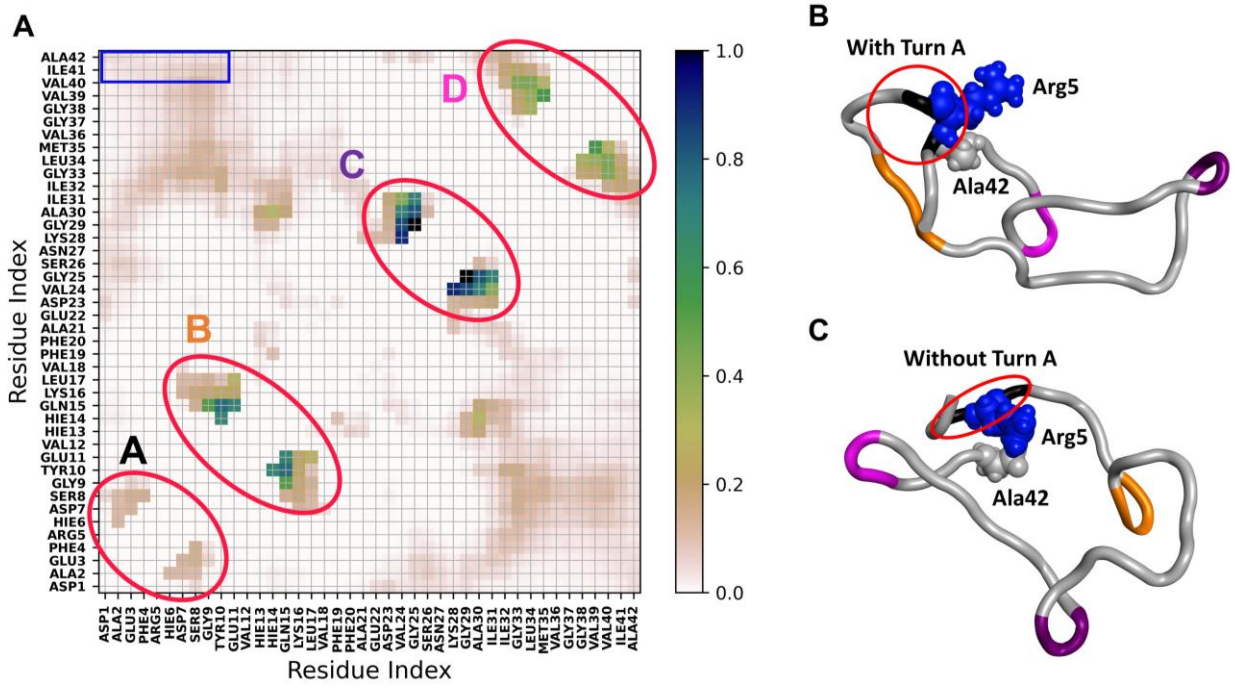

Figure 7. Probability of intramolecular contacts of synthetic conformations between Cα atoms of Aβ42 and representative conformations with Arg5-Ala42 contacts. (A) Four local bends are circled: turn A (F4-H6), turn B (E11-H14), turn C (S26-K28), and turn D (V36-G38). Residues contacting the C-terminal residues are marked in blue squares. Novel synthetic conformations depicting non-polar contacts between Arg5-Ala42 with turn A (B) and without turn A (C). Both conformations utilize turn D to stabilize the C-terminus. Residues 5 and 42 are shown in the van der Waal representation.

Turn A locates in the highly flexible N-terminus, where mutation substitutions of Arg5 suppressed the aggregation of Aβ42 [47]. Our EPR measurements also showed distinct spectral features at residue 5, indicating that Arg5 may feature a distinct structure or interactions as compared with the nearby residues (**Figure 8**). The novel synthetic conformations revealed that Arg5 can form interactions with Ala42 with or without the presence of Turn A, as illustrated in **Figures 7B** and **7C**, respectively. Of note, turn D (V36-G38) is present alongside, which may help to orient and stabilize the C-terminus to form an interaction with Arg5 (red circle in **Figures 7B-C**). Our previous work showed that substituting a nitroxide spin label compound R1 at positions G37 and G38 altered the kinetics of Aβ42 fibril formation [56], and these synthetic conformations provide a possible monomer structure to guide future experimental design.

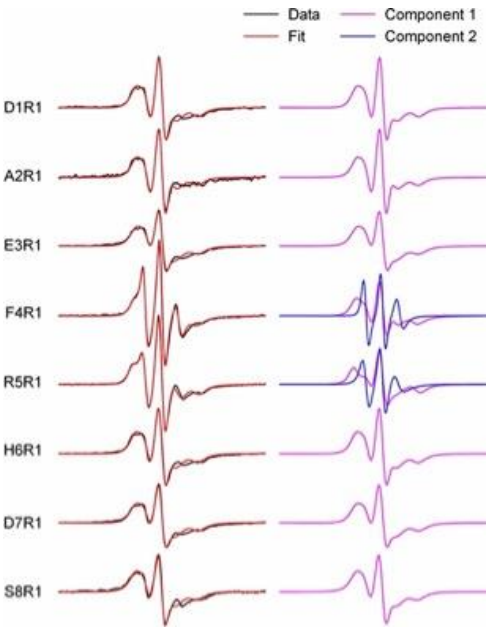

Figure 8. EPR spectra of spin-labeled Aβ42 fibrils show unique spectral features at residues 4 and 5. R1 represents the spin label. Left: EPR data are superimposed on the best fit of simulated EPR spectra. Right: Individual EPR spectral components for obtaining the best fit. Note that only the EPR spectra at residues 4 and 5 require a second spectral component. These data are consistent with a turn structure at residues 4-5.

Presence of salt bridges in toxic and less toxic monomer conformations. Salt bridges play an important role in providing intra-molecular attractions and conformational specificity that directly relate to Aβ42 aggregation. This non-covalent interaction forms when sidechains of oppositely charged residues are close enough to each other to experience electrostatic attractions. Therefore, protein structures must be described with precise atomistic details and sidechain motions sampled accurately. Although the D23-K28 salt bridge did not exist in the 1-μs MD Run1 used in our training set, by using interpolation in the latent space, the model sampled novel

synthetic conformations with a new salt bridge, D23-K28 (**Figure 9**), which is reported to promote aggregation [46]. Experiments suggested the importance of this salt bridge, but the conformation ensemble was never determined experimentally. The novel synthetic conformations revealed distinctly different rearrangements with the D23-K28 salt bridge, ranging from more extended to highly packed conformations (**Figure 9**). The D23-K28 constraint results in Turn C, which is important to oligomerization.

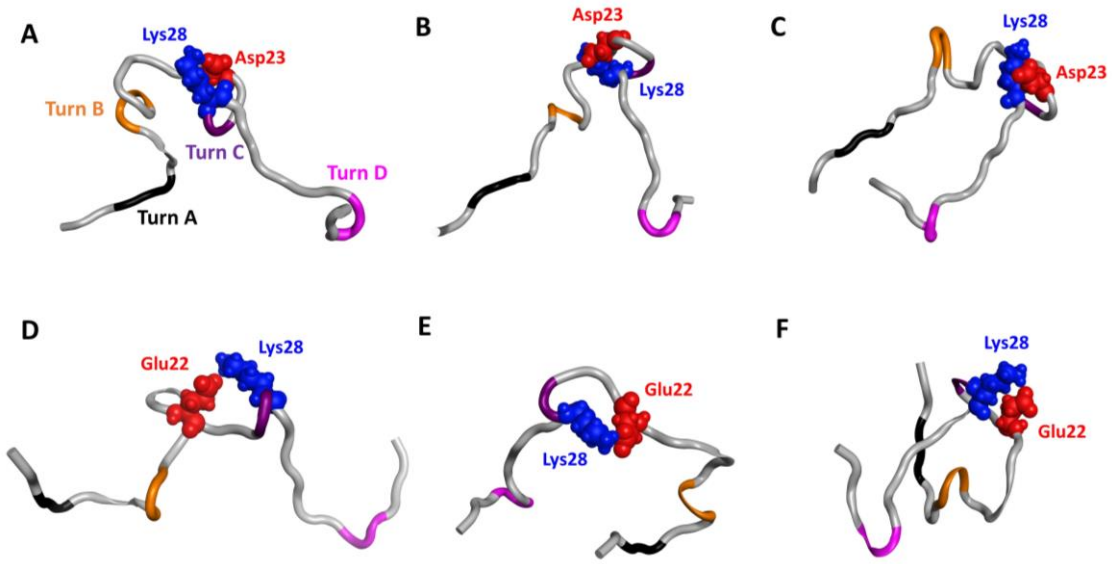

Figure 9. Novel synthetic conformations with intramolecular salt bridges. Top: structures with D23-K28 salt bridge. (A) and (B) show an open form with different local arrangements near turns B and C, and (C) reveals a U-shape, partial hairpin conformation. Bottom: structures with E22-K28 salt bridge with open (D), partial open (B) and compact (C) conformations.

The synthetic conformations revealed novel conformations with the salt bridge E22-K28 with significantly different conformations (**Figures 9D-F**). This salt bridge was reported to prevent a toxic turn at positions E22 and D23, for potentially a less-toxic monomer [57]. The synthetic conformation presents key features suggested by experiments, with no turn at positions 22 and 23 when the salt bridge E22-K28 exists (**Figure 9**). Nevertheless, other publications suggested that the salt bridge E22-K28 may help form other intramolecular contacts (e.g., contacts between the C-terminus and Arg5) to stabilize a hairpin structure and promote dimer formation [47]. Although which salt bridges promote or inhibit monomer aggregation is debatable, our novel synthetic conformations provide various possible sidechain rearrangements and salt-

bridge conformations for further interpreting experiments in understanding aggregation mechanisms.

Discussion

In this work, we demonstrated the ability of the DL model ICoN, which applies learned physics of protein motions, to directly generate new protein conformational ensembles. The transferable conformational search approach applied a neural network model using protein conformations generated by classical MD simulations with all-atom representation in explicit solvent model to train an autoencoder. Here we re-saved frames from MD trajectories without further processing the data. We reduced protein atomic coordinates to a 3D latent space, with each conformation represented by one data point in the 3D plot. Each data point can be accurately reconstructed back to a protein structure as well. ICoN uses interpolation between 2 data points in the 3D latent space to guide the search for conformational changes, with natural molecular motions identified by the ICoN model. The BAT-based (vBAT) feature overcomes problems with Cartesian coordinates that fail to smoothly present dihedral rotation during conformation sampling, which is particularly important in sampling large-scale motions such as those shown in IDPs. vBAT also effectively handles dihedral periodicity. The model is highly efficient for identifying novel synthetic Aβ42 conformation ensembles not seen in the MD trajectory used for training. These ensembles have different large-scale motions (i.e., new compact conformations) and detailed sidechain rearrangements such a forming a new salt bridge, D23-K28.

The synthetic Aβ42 conformations sampled by the ICoN model provide a comprehensive snapshot of the conformational landscape for Aβ42 monomers. For example, the ICoN model finds the formation of salt bridges involving residues E22 or D23 **(Figure 8)**, both of which are hotspots for familial mutations in Alzheimer's disease. The ICoN model also identifies four local bends at residue positions 4-6, 11-14, 26-28, and 36-38 (**Figure 1D**). Although each of these turns have been previously observed in isolation, the observation of all four turns in Aβ42 monomers suggests that the formation of these turns may precede the assembly of specific oligomeric or fibrillar structures. The synthetic conformations containing these turns provide a foundation for investigating how these turns affect the aggregation pathway or formation of specific aggregate structures.

Training of the ICoN model requires sufficient protein conformations, so the DL can reveal natural motions of the protein to guide the search toward new low-energy structures. Because experimentally determined structures typically cover only a small proportion of the overall conformation ensembles, we need to rely on MD simulations. However, force-field parameters used in MD may affect the performance in obtaining protein conformations and dynamics. Because conformation ensembles with artificial bias in the training set will affect the performance of ICoN, using a suitable force field for MD simulations is crucial. For example, conventional force fields parameterized using structured proteins can over-stabilize disordered proteins/regions, such as generating more helical structures in Aβ42 when simulated by Amber ff14SB, as compared with diverse disordered Aβ42 structures simulated by an IDP-specific force field, ff14IDPSFF [58]. The model trained by MD runs using a less flexible ff14SB affected the search results and yielded fewer novel synthetic conformations (**Table 1**). Note that our training set did not need to cover the complete conformational distribution. However, conformation fluctuation governed by key dihedral rotations should be included in the training set so that the model can identify these key elements for guiding the conformational search to generate novel synthetic conformations.

Conceptually, ICoN performs as a conformational search method rather than a deterministic simulation method. Therefore, the search results do not directly reflect the equilibrium distributions of protein conformations. However, protein conformations are populated according to a Boltzmann distribution, which can be approximated by their computed energy. Unlike structured proteins, which have highly stable global energy minima, IDPs have numerous fluctuating heterogeneous conformations with similar energy. The IDPs can be highly sensitive to their environment, and precisely reproducing physiological conditions when performing experiments or simulation is impractical. Therefore, thoroughly finding IDP conformational ensembles is more practical and useful instead of aiming to obtain the population distribution in equilibrium under a particular physiological environment. Importantly, unlike structured proteins whose native structures are functionally important, meta-stable conformations of IDPs can be crucial in performing the biological function. For example, experiments showed that Aβ42 monomer aggregation has a rate-limiting nucleation-dependent polymerization process [50], which suggests that the highly populated monomer conformations might not be the ones that drive nucleation, and the monomers are not pre-organized as a ready-to-aggregate conformation.

Therefore, as compared with the conformational search for structure proteins, thoroughly sampling meta-stable conformations for IDPs is critical. As demonstrated in this study, the novel synthetic Aβ42 monomer conformations provide atomistic details to support several experimental observations, which also bring insights into mechanisms of oligomerization.

Conformational changes of biomolecules follow principles of physics, which are consequences of different arrangements of atoms rotating around a set of single bonds, the internal degrees of freedom having the highest flexibility. Both PCA and NMA describe the natural motions of a molecule, which are commonly used to determine essential protein dynamics and guide conformational sampling. The two techniques are linear transforms that extract the most important elements in a data matrix, a covariance or Hessian matrix. If the conformations are described by BAT coordinates, concerted dihedral rotations are directly identified by the eigenvectors: new conformations can be generated via distortion using each eigenvector and their combinations. However, real protein systems can have very complicated and higher-order correlations that have intrinsic nonlinear effects and may not be well described using standard PCA or NMA. Moreover, although PC space is commonly used for data analysis, there is no function to convert manipulated data points in PC space back to its atomistic protein structure. The activation function used in the neural network adds nonlinear effects to process nonlinear features, which allows for the autoencoder to use a reduced dimension to describe protein conformations and accurately convert the perturbed reduced-dimension coordinates back to their atomistic structures. A transition between two protein conformations is governed by certain concerted rotating dihedral angles, which can be achieved by interpolating two data points (conformations) in the latent space. As illustrated in **Figure 4**, if one obtains conformations along the smooth green curve in the latent space, the conformational changes reproduce the motions following the first PC mode. Notably, various arrangements can lead to the same conformational transitions, which can be revealed by interpolating between data points in the latent space. As for the natural motions identified by PCA or NMA, these conformations follow natural motions of the protein system (black dot line in **Figure 4**), and new conformations can be found by processing the interpolations in the latent space.

Although the ICoN model is fast, taking a few minutes to train 10,000 frames of Aβ42 and hours to perform quick energy minimization for the synthetic conformations generated from

latent space interpolation using a GPU card, the training set is from a physics-based MD sampling. It typically takes days, if not weeks, to perform sufficiently long MD simulations to obtain the training set. To speed up the sampling, one can apply the ICoN model for preliminary results to select dissimilar protein conformations to seed more MD runs [19]. An advantage in using BAT-based coordinates is that one can easily select a set of dihedral angles that determine the motions of interest for training, instead of using all DOF. For small proteins or peptides such as Aβ42 and αB-crystallin57-69, our classical internal BAT (Z-matrix)-based vBAT nicely presents local sidechain rearrangements and large-scale backbone motions. However, for larger (>200 residues) proteins, rotations of some dihedral angles near the root atoms may result in unrealistic motions on the far end of the protein. The issue can be addressed easily by using multilayer BAT coordinates, assigning multiple fragments/chains for a protein or multi-protein complex. Also, reconnecting fragments with pseudo-DOF can avoid the accumulated dependence problem [59]. Future work will implement the multilayer BAT coordinates to eliminate the protein size limitation when building the ICoN model. Various interpolation and extrapolation strategies in the latent space will be examined for generating novel synthetic conformations as well.

Methods

All-atom Molecular Dynamics Simulations and Preparation of Training and Validation Data Sets

All MD simulations were performed using the AMBER20 package using either ff14SB or ff14IDPSFF force fields (Table 1) with GPU acceleration [60]. The systems were simulated using TIP3P explicit solvent model [61] at temperature of 298K with NPT ensemble. 12 Å cutoff was used for short range non-bonded interactions and the long-range electrostatic interactions were computed by the particle mesh Ewald method (PME) [62].

Aβ42

MD-Run1 was initiated using PDB: 2NAO with ff14IDPSFF force field (Table 1). Each frame was saved at 1-ps time integral which made up a total of 1,000,000 frames. Other MD

trajectories using either PDB 2NAO or 1Z0Q as the initial structure were obtained from our previous work and detailed in Table 1 [51,52,63]. Conformations of MD-Run1 to 4 were used for training, validation, and sampling conformations from latent space, while MD-Run1.1 to 3 were utilized for pairwise RMSD search for synthetic conformations obtained by Model 1 (**Table 1**).

aβ-crystallin57-69

AlphaFold2 computed structure of aβ-crystallin57-69 (PDB: AF_AFP02511F1) was used as initial structure for MD simulations. Two randomly seeded 500-ns MD simulations were obtained from our previous work, where each frame was collected at 1-ps interval which made up 500,000 frames [64]. Conformations of Seed1 (**Table 1**) were used for training, validation, and sampling conformations from latent space, while Seed2 was utilized for pairwise RMSD search for synthetic conformations.

Preparation of training and validation data sets

To construct our dataset for ICoN model training, we resaved 1 frame every 100-ps starting from 0-ps of Aβ42 MD-Run1 and 2 (10,000 conformations; 1% of raw MD) to prepare 2 training sets. Similarly, we resaved 1 frame every 100-ps starting from frame 50-ps of Aβ42 MD-Run1 and 2 for 2 validation sets. For MD-Run 3 and 4, we resaved a frame every 200 ps (0.5% of raw MD) because the classical ff14SB force field yielded less diverse conformations with favored helicity during simulations [58] (see Ref #58 for details). As a result, we resaved a frame starting from frame 0 and frame 100 of MD-Run 3 and 4 to prepare 5,000 frames for training and validation, respectively. Similar approach is used for aβ-crystallin57-69 Seed1, where 1 frame every 50-ps starting from 0-ps (5,000 conformations; 1% of raw MD) used for training, and 1 frame every 50-ps starting from 25-ps used for validation. This ensures training and validation datasets are separated by 50-ps and 25-ps time interval for Aβ42 and aβ-crystallin57-69 respectively. Details of other training and validation datasets are shown in Table 1. This approach allows training a model from minimal amount of input data, yet generate numerous conformations with extended conformational space.

Protein Structure Representation

Classical Bond-Angle-Torsion coordinates

Bond Angle Torsion (BAT) is the internal coordinate representation [65–67], where "bond" denotes the distance between a pair of bonded atoms, "angle" refers to the bond angle between pairs of bonds connected to a central atom, and "torsion" indicates the dihedral angle formed by four bonded atoms via three bonds. Specifically, the dihedral angle is defined as the angle between the plane containing atoms (i, j, k) and another plane containing atoms (j, k, l) (**Figure 1E**). BAT coordinates can be accurately transformed back to all atom Cartesian coordinates. In this, the placement of each atom i > 3 that is not bounded to atom 2 is specified by its bond length (b_i), bond angle (α_i), and dihedral angle (θ_i) with respect to three other atoms that are bonded in sequence and whose positions are already defined (**details in SI**). Three terminal atoms, designated as root-based atoms, are used to initiate conversion from BAT to Cartesian. Root based atoms carry six external degrees of freedom such as global translations and rotations.

Vector Representation of BAT (vBAT)

Although BAT coordinates effectively capture concerted motions, they suffer from a periodicity problem. Dihedral rotation from 179° to -179° is actually a 2° shift in angular space. However, the basic arithmetic operation used in computing a covariance matrix results in a large, 179°-(-179°) = 358°, rotation [68]. Hence, we used vector representation of BAT coordinates termed vBAT as input features to ICoN model. The vBAT internal coordinate representation is equivariant under global translations and rotations, and thus provided strong inductive bias for our ICoN model. Furthermore, compared to other structural representations such as Anchored Cartesian [41] or PCA based representations [32], internal coordinate representation is independent of reference structure. A detailed description of vBAT computation is provided in this section and in Supplementary Information.

For a set of four bonded atoms with atom indexes i, j, k, and l (**Figure 2B**), we defined three bond vectors v_{ij}, v_{jk}, and v_{kl}, that are the relative positions between the atoms. The unit vectors that are normal to the plane made by a pair of neighboring bond vectors defined vector features, denoted as $\hat{n}_{ijk}$ and $\hat{n}_{jkl}$.

$$\hat{n}_{ijk} = \frac{-\vec{v}_{ij} \times \vec{v}_{jk}}{|\vec{v}_{ij}||\vec{v}_{jk}|} \qquad \hat{n}_{jkl} = \frac{-\vec{v}_{jk} \times \vec{v}_{kl}}{|\vec{v}_{jk}||\vec{v}_{kl}|} \qquad (1)$$

vBAT vector features comprising $\hat{n}_{ijk}$, $\hat{n}_{jkl}$, $\vec{v}_{ij}$, and $\vec{v}_{jk}$ were used as an input to ICoN model. When sampling conformations in latent space, all external degrees of freedom (root-based coordinates) were inferred from the reference conformation. Sampled features $\hat{n}_{ijk}$, $\hat{n}_{jkl}$, $\vec{v}_{ij}$, and $\vec{v}_{jk}$ were first converted to bond angles, and torsions, and then transformed back to Cartesian coordinates in large batches using a highly optimized GPU implementation (**Figure S2**). Since changes in bond length are very minimal throughout MD simulations, we used fixed bond length inferring from the reference structure for all generated conformations.

Model Architecture and Training

Model Architecture

The ICoN model is an autoencoder type architecture based on fully connected neural network (**Figure 2A**) trained on MD data to directly output all atom protein conformations. The model is designed to compress an input representation into lower dimensional latent space (3D) with an encoder that can be directly visualized. These points representing conformations in 3D latent space are decompressed back to the original input dimension with decoder.

Initially, vBAT vectors were flattened followed by concatenation into a high dimensional array to use as input features to the ICoN model. As a result, we have 4x3x(N-3) features for each molecule 2376, and 7488 features for αB-crystallin57-69 and aβ42 respectively. The activation units were gradually reduced to 3-dimensional latent space in seven layers, where each molecular conformation can be presented in the 3D latent space, as illustrated in **Figure 2A.** For both systems, the encoding process reduced the dimension from n_f , to $n_f/4$, $n_f/8$, $n_f/16$, $n_f/32$, $n_f/64$, $n_f/64$, and 3, where n_f is the total number of input features. Then, the decoding process brought the representation from 3D to 2376 dimension with the same number of layers but reverse the order.

The LeakyReLU activation function was applied to all layers except the last layer. Layer normalization is applied in the second layer. To prevent overfitting, 10% of the weights were randomly set to zero in the third layer, which led to better convergence of validation loss. The

decoder takes the latent vector as input and decompresses it back to its original feature dimension. The decoder also featured LeakyReLU nonlinearity but did not include any layer normalization or dropout layers. The bias units are omitted aiming to minimize the model's parameter count. Meanwhile, all model parameters were initialized through the Xavier Uniform weight initialization method, which is known to enhance training stability and efficiency [69]. The total count of model parameters utilized is 3,294,964 for αB-crystallin57-69 and 32,721,156 for aβ42. Throughout the iterative training process, in addition to learning a relationship between vast number of features, the network also learns how to meaningfully compress the representation into the lower dimensional 3D latent space by performing nonlinear dimensionality reduction. Each conformation is converted from vBAT to Cartesian for further analysis.

Loss Function

The combination of L1 and L2 loss functions, the Smooth L1 loss, was used to minimize the difference between the model output (y_i) and the original input (x_i). The equation for the smooth L1 loss function, L, is given below (Eq. 2):

$$L = \begin{cases} \frac{(x_i - y_i)^2}{2\beta}, if |x_i - y_i| < \beta \\ |x_i - y_i| - 0.5\beta, otherwise \end{cases} \quad (2)$$

where, $(x_i\text{-}y_i)^2$ is L2 loss and $|x_i\text{-}y_i|$ is L1 loss.

The L1 and L2 functions intersect at β=1.0 resulting in a smooth, continuous loss function that is less sensitive to outliers compared to the L1 loss. However, for larger errors, the function transitions to an L1 loss, ensuring robustness to extreme values. As a result, using Smooth L1 loss can lead to more stable training and better convergence.

Optimization was performed using the Adam optimizer [70] for both protein systems. For Aβ42 training, the learning rate was manually reduced from 0.002 to 0.001 after 5000 iterations, resulting in a total of 10,000 training iterations. The learning rate was kept constant at 0.001 for aβ-crystallin57-69, comprising a total of 5000 training iterations. We fed 200 conformations in mini-batches per iteration to the model, randomly selecting frames from the data pool during each training iteration. Training was terminated upon convergence of both the training and

validation loss, a process that took a few minutes for both systems. The MDAnalysis Python library [71] used as the tool of choice for reading and writing trajectories.

Implementation

In this study, we employed the PyTorch library to implement all deep learning algorithms [72]. MDAnalysis, a Python library, facilitated MD trajectory input/output operations and torsion list construction[71] . Leveraging the torch tensor data structure in PyTorch, we implemented GPU-optimized BAT and vBAT calculations. The code and pre-trained ICoN model parameters can be accessed at the following link: https://github.com/chang-group/ICoN

Evaluation of reconstructed conformations

The accuracy of the ICoN model was assessed using the validation dataset. Initially, vBAT features extracted from the validation set were fed into the pretrained IcoN model, followed by a forward pass. Subsequently, the model's output was converted into fully atomistic Cartesian coordinates. Reconstruction accuracy was measured by computing the RMSD of all heavy atoms (including backbone atoms) between the original and reconstructed conformations. Heavy atom RMSD calculation involved all non-hydrogen atoms, while backbone RMSD estimation focused on N, Cα, C, and O atoms (refer to Table 1).

Furthermore, we conducted more comprehensive comparisons between the original and reconstructed conformations, examining factors such as the distribution of all dihedral angles (including both backbone and sidechain for all amino acids) and their correlations. To address periodicity issues in angular space, dihedral angle correlations were computed using a trigonometric function (see details in the Supplementary Information) [59,68,73]. These analyses were performed to verify that all properties of the original conformations are faithfully reproduced (see **Figure S3-S6**).

Generation of Novel Conformations

Step1: Non-linear Interpolation

In order to generate synthetic conformations, we first encode training and validation data set into the latent space. Then, non-linear interpolation is performed between each pair of consecutive

points in 3D latent space. Interpolation involves generation of path that moves between selected pair of points in an arc gradually changing polar angle Θ, azimuthal angle Φ, and radius R in a spherical coordinate system. The center of the sphere is located at the midpoint between two points, and random adjustments to the center position were made to generate a distinct path (see details in the Supplementary Information). 10 points were interpolated from each consecutive pair of points yielding a total of 19,9990 and 99,990 interpolated conformations for Aβ42 MD-Run1 and aβ-crystallin57-69-Seed1, respectively (**Table 1**).

Step 2: Eliminating non-physical conformations.

To avoid steric clashes or non-physical conformations, all generated structures were subjected to a short minimization using a modified Generalized Born with implicit solvent model (mbondi3, igb=8) for 1000 steps. We applied energy cutoff of -400 (kcal/mol) and -200 (kcal/mol) for Aβ42 and aβ-crystallin57-69 respectively. Energy cutoff values are selected from more positive low population regions of total energy distribution. All conformations with energy higher than the cutoff were excluded (Table 1). Energy minimization did not affect the accuracy of prediction but is essential to correct for the physically unrealistic conformations. To have a fair comparison between various energy terms, the original MD trajectories were also minimized using a similar strategy involving 50 steps.

Step3: *Elimination of repeated conformations with pairwise RMSD search.*

After energy-based elimination, we further perform pairwise heavy atom RMSD search with 1Å cutoff across all remaining conformations in order to eliminate repeated conformations. 1 Å RMSD difference can capture small variations in sidechain as well as salt bridge formation or breaking. This step is essential to ensure all conformations are unique and diverse. **Table 1** shows the number of remaining conformations for all MD runs.

Step4: *Elimination of existing conformations in raw MD.*

We further compare remaining synthetic conformations from previous step with all the conformations in raw MD using pairwise heavy atom RMSD search. Conformations with the pairwise RMSD lower than 2 Å cutoff were deemed similar and excluded from synthetic pool for both Aβ42 and aβ-crystallin57-69 (Table 1). All pairwise RMSD searches were performed using PTRAJ [74].

Dihedral PCA (dPCA)

To extract conformations with major protein motions, we utilized dihedral Principal Component Analysis (dPCA) employing torsion angles with in house code. The resulting conformations, derived from the first three principal components, were then projected onto a latent space using the IcoN model in order to get insight on conformational transition of the aβ-crystallin57-69 300-ps trajectory, composed of 300 frames. Notably, the transitions observed in the latent space, governed by principal component modes, manifested as smooth nonlinear trajectories. Consequently, we opted for nonlinear interpolation to produce synthetic conformations.

The dPCA calculation involves mapping torsion angles to a unit circle using trigonometric functions. This enables accurate estimation of differences (involving subtraction) and averages (involving summation) during the construction of the covariance matrix, thereby preventing erroneous computation of their correlation at the discontinuity margin (±180° or 360°/0°) [59,68,73].

Analysis of synthetic conformations

Free energy profiles

To derive free energy profiles, we initially constructed a two-dimensional histogram by binning the radius of gyration (Rg) and backbone root mean squared deviation (RMSD) of protein conformations with respect to a reference structure. Subsequently, we determined the normalized population (P) for each bin throughout the entire trajectory. The free energy values were then computed using the equation $F = -k_B T \log(P)$, where k_B represents the Boltzmann constant and T is a room temperature.

Contact maps

The contact maps were generated by applying a cutoff of 6.5 Å to all intramolecular distances between C-alpha positions. A contact density of one is assigned when a contact is formed, and zero otherwise, for each frame. Local contact densities of residue pairs with indexes i and j, where $|i-j| < 4$ were set to zero. The average across all frames (conformations) yielded the contact probability between the pairs of residues.

Aβ42 turn characterization

Four local turns have been identified as turn A (F4-H6), turn B (E11-H14), turn C (S26-K28), and turn D (V36-G38). To selectively isolate conformations characterized by turns occurring in specific regions, we initially identified residue pairs exhibiting a high pair contact density in proximity to the corresponding turn region, as depicted in contact map of **Figure 7**. Subsequently, a turn was deemed to be formed if the C-alpha distance between the selected pair of residues was less than 6.5 Å. Conversely, a turn was classified as not formed when the contact distance exceeded 9.0 Å. For turns A, B, C, and D, the residue pairs utilized for contact identification are Ala2:Asp7, Tyr10:Gly15, Val24:Gln29, and Gly33:Val40, respectively are selected.

Data Availability

The code, pre-trained ICoN model parameters, and data sets are available at: https://github.com/chang-group/ICoN

Acknowledgements.

We thank Dr. Hisashi Okumura and Dr. Mai Suan Li for providing their modeled Aβ42 conformations, and Dr. David Minh and Ruben Montes for helpful discussion. This study was supported by the US National Institutes of Health (R01GM-109045 to C. C. and R01AG-050687 and R21AG-082041 to Z.G.), the US National Science Foundation (MCB-1350401 to C.C.) and the Bourns Endowment funds to B.B.)

References

(1) Miller, M. D.; Phillips, G. N. Moving beyond Static Snapshots: Protein Dynamics and the Protein Data Bank. *Journal of Biological Chemistry* **2021**, *296*, 100749. https://doi.org/10.1016/j.jbc.2021.100749.
(2) Dill, K. A.; Bromberg, S. *Molecular Driving Forces: Statistical Thermodynamics in Biology, Chemistry, Physics, and Nanoscience*; Garland Science, 2011.
(3) Zhang, Y.; Chen, H.; Li, R.; Sterling, K.; Song, W. Amyloid β-Based Therapy for Alzheimer's Disease: Challenges, Successes and Future. *Signal Transduct Target Ther* **2023**, *8* (1), 248. https://doi.org/10.1038/s41392-023-01484-7.
(4) Huang, B.; Kong, L.; Wang, C.; Ju, F.; Zhang, Q.; Zhu, J.; Gong, T.; Zhang, H.; Yu, C.; Zheng, W.-M.; Bu, D. Protein Structure Prediction: Challenges, Advances, and the Shift of Research Paradigms. *Genomics Proteomics Bioinformatics* **2023**, *21* (5), 913–925. https://doi.org/10.1016/j.gpb.2022.11.014.
(5) Whitford, D. *Proteins: Structure and Function*; Wiley, 2013.
(6) Tripathi, T.; Dubey, V. K. *Advances in Protein Molecular and Structural Biology Methods*; Elsevier Science, 2022.
(7) Macuglia, D.; Roux, B.; Ciccotti, G. The Emergence of Protein Dynamics Simulations: How Computational Statistical Mechanics Met Biochemistry. *The European Physical Journal H* **2022**, *47* (1), 13. https://doi.org/10.1140/epjh/s13129-022-00043-y.
(8) Hollingsworth, S. A.; Dror, R. O. Molecular Dynamics Simulation for All. *Neuron* **2018**, *99* (6), 1129–1143. https://doi.org/10.1016/j.neuron.2018.08.011.
(9) Allison, J. R. Computational Methods for Exploring Protein Conformations. *Biochem Soc Trans* **2020**, *48* (4), 1707–1724. https://doi.org/10.1042/BST20200193.
(10) Paquet, E.; Viktor, H. L. Molecular Dynamics, Monte Carlo Simulations, and Langevin Dynamics: A Computational Review. *Biomed Res Int* **2015**, *2015*, 1–18. https://doi.org/10.1155/2015/183918.
(11) T. Daniel Crawford; Cecilia Clementi; Robert Harrison; Teresa Head-Gordon; Shantenu Jha; Anna Krylov; Ashley Ringer-McDonald; Theresa Windus; Dominika Zgid. *Molssi*. https://molssi.org/molssi-community-code-partners/ (accessed 2024-04-18).
(12) Bruce, N. J.; Ganotra, G. K.; Richter, S.; Wade, R. C. KBbox: A Toolbox of Computational Methods for Studying the Kinetics of Molecular Binding. *J Chem Inf Model* **2019**, *59* (9), 3630–3634. https://doi.org/10.1021/acs.jcim.9b00485.
(13) David, C. C.; Jacobs, D. J. Principal Component Analysis: A Method for Determining the Essential Dynamics of Proteins; 2014; pp 193–226. https://doi.org/10.1007/978-1-62703-658-0_11.
(14) Yasuda, T.; Shigeta, Y.; Harada, R. Efficient Conformational Sampling of Collective Motions of Proteins with Principal Component Analysis-Based Parallel Cascade Selection Molecular Dynamics. *J Chem Inf Model* **2020**, *60* (8), 4021–4029. https://doi.org/10.1021/acs.jcim.0c00580.
(15) Peng, C.; Wang, J.; Shi, Y.; Xu, Z.; Zhu, W. Increasing the Sampling Efficiency of Protein Conformational Change by Combining a Modified Replica Exchange Molecular Dynamics and Normal Mode Analysis. *J Chem Theory Comput* **2021**, *17* (1), 13–28. https://doi.org/10.1021/acs.jctc.0c00592.

(16) Chang, C.; Gilson, M. K. Tork: Conformational Analysis Method for Molecules and Complexes. *J Comput Chem* **2003**, *24* (16), 1987–1998. https://doi.org/10.1002/jcc.10325.

(17) Kolossváry, I.; Guida, W. C. Low Mode Search. An Efficient, Automated Computational Method for Conformational Analysis: Application to Cyclic and Acyclic Alkanes and Cyclic Peptides. *J Am Chem Soc* **1996**, *118* (21), 5011–5019. https://doi.org/10.1021/ja952478m.

(18) Kanada, R.; Terayama, K.; Tokuhisa, A.; Matsumoto, S.; Okuno, Y. Enhanced Conformational Sampling with an Adaptive Coarse-Grained Elastic Network Model Using Short-Time All-Atom Molecular Dynamics. *J Chem Theory Comput* **2022**, *18* (4), 2062–2074. https://doi.org/10.1021/acs.jctc.1c01074.

(19) Tian, H.; Jiang, X.; Xiao, S.; La Force, H.; Larson, E. C.; Tao, P. LAST: Latent Space-Assisted Adaptive Sampling for Protein Trajectories. *J Chem Inf Model* **2023**, *63* (1), 67–75. https://doi.org/10.1021/acs.jcim.2c01213.

(20) Ghorbani, M.; Prasad, S.; Klauda, J. B.; Brooks, B. R. GraphVAMPNet, Using Graph Neural Networks and Variational Approach to Markov Processes for Dynamical Modeling of Biomolecules. *J Chem Phys* **2022**, *156* (18). https://doi.org/10.1063/5.0085607.

(21) Lemke, T.; Peter, C. EncoderMap: Dimensionality Reduction and Generation of Molecule Conformations. *J Chem Theory Comput* **2019**, *15* (2), 1209–1215. https://doi.org/10.1021/acs.jctc.8b00975.

(22) Merz, K. M.; Wei, G.-W.; Zhu, F. Editorial: Machine Learning in Bio-Cheminformatics. *J Chem Inf Model* **2024**, *64* (7), 2125–2128. https://doi.org/10.1021/acs.jcim.4c00444.

(23) Bonati, L.; Zhang, Y.-Y.; Parrinello, M. Neural Networks-Based Variationally Enhanced Sampling. *Proceedings of the National Academy of Sciences* **2019**, *116* (36), 17641–17647. https://doi.org/10.1073/pnas.1907975116.

(24) Mulnaes, D.; Gohlke, H. TopScore: Using Deep Neural Networks and Large Diverse Data Sets for Accurate Protein Model Quality Assessment. *J Chem Theory Comput* **2018**, *14* (11), 6117–6126. https://doi.org/10.1021/acs.jctc.8b00690.

(25) Prašnikar, E.; Ljubič, M.; Perdih, A.; Borišek, J. Machine Learning Heralding a New Development Phase in Molecular Dynamics Simulations. *Artif Intell Rev* **2024**, *57* (4), 102. https://doi.org/10.1007/s10462-024-10731-4.

(26) Kikutsuji, T.; Mori, Y.; Okazaki, K.; Mori, T.; Kim, K.; Matubayasi, N. Explaining Reaction Coordinates of Alanine Dipeptide Isomerization Obtained from Deep Neural Networks Using Explainable Artificial Intelligence (XAI). *J Chem Phys* **2022**, *156* (15), 154108. https://doi.org/10.1063/5.0087310.

(27) Do, H. N.; Wang, J.; Bhattarai, A.; Miao, Y. GLOW: A Workflow Integrating Gaussian-Accelerated Molecular Dynamics and Deep Learning for Free Energy Profiling. *J Chem Theory Comput* **2022**, *18* (3), 1423–1436. https://doi.org/10.1021/acs.jctc.1c01055.

(28) Majewski, M.; Pérez, A.; Thölke, P.; Doerr, S.; Charron, N. E.; Giorgino, T.; Husic, B. E.; Clementi, C.; Noé, F.; De Fabritiis, G. Machine Learning Coarse-Grained Potentials of Protein Thermodynamics. *Nat Commun* **2023**, *14* (1), 5739. https://doi.org/10.1038/s41467-023-41343-1.

(29) Taneja, I.; Lasker, K. Machine-Learning-Based Methods to Generate Conformational Ensembles of Disordered Proteins. *Biophys J* **2024**, *123* (1), 101–113. https://doi.org/10.1016/j.bpj.2023.12.001.
(30) Glielmo, A.; Husic, B. E.; Rodriguez, A.; Clementi, C.; Noé, F.; Laio, A. Unsupervised Learning Methods for Molecular Simulation Data. *Chem Rev* **2021**, *121* (16), 9722–9758. https://doi.org/10.1021/acs.chemrev.0c01195.
(31) Zheng, L.-E.; Barethiya, S.; Nordquist, E.; Chen, J. Machine Learning Generation of Dynamic Protein Conformational Ensembles. *Molecules* **2023**, *28* (10), 4047. https://doi.org/10.3390/molecules28104047.
(32) Noé, F.; Olsson, S.; Köhler, J.; Wu, H. Boltzmann Generators: Sampling Equilibrium States of Many-Body Systems with Deep Learning. *Science (1979)* **2019**, *365* (6457). https://doi.org/10.1126/science.aaw1147.
(33) Ramaswamy, V. K.; Musson, S. C.; Willcocks, C. G.; Degiacomi, M. T. Deep Learning Protein Conformational Space with Convolutions and Latent Interpolations. *Phys Rev X* **2021**, *11* (1), 011052. https://doi.org/10.1103/PhysRevX.11.011052.
(34) Atz, K.; Grisoni, F.; Schneider, G. Geometric Deep Learning on Molecular Representations. *Nat Mach Intell* **2021**, *3* (12), 1023–1032. https://doi.org/10.1038/s42256-021-00418-8.
(35) Jin, Y.; Johannissen, L. O.; Hay, S. Predicting New Protein Conformations from Molecular Dynamics Simulation Conformational Landscapes and Machine Learning. *Proteins: Structure, Function, and Bioinformatics* **2021**, *89* (8), 915–921. https://doi.org/10.1002/prot.26068.
(36) Audagnotto, M.; Czechtizky, W.; De Maria, L.; Käck, H.; Papoian, G.; Tornberg, L.; Tyrchan, C.; Ulander, J. Machine Learning/Molecular Dynamic Protein Structure Prediction Approach to Investigate the Protein Conformational Ensemble. *Sci Rep* **2022**, *12* (1), 10018. https://doi.org/10.1038/s41598-022-13714-z.
(37) Sultan, M. M.; Wayment-Steele, H. K.; Pande, V. S. Transferable Neural Networks for Enhanced Sampling of Protein Dynamics. *J Chem Theory Comput* **2018**, *14* (4), 1887–1894. https://doi.org/10.1021/acs.jctc.8b00025.
(38) Mehdi, S.; Smith, Z.; Herron, L.; Zou, Z.; Tiwary, P. Enhanced Sampling with Machine Learning. *Annu Rev Phys Chem* **2024**, *75* (1). https://doi.org/10.1146/annurev-physchem-083122-125941.
(39) Hoseini, P.; Zhao, L.; Shehu, A. Generative Deep Learning for Macromolecular Structure and Dynamics. *Curr Opin Struct Biol* **2021**, *67*, 170–177. https://doi.org/10.1016/j.sbi.2020.11.012.
(40) Ward, M. D.; Zimmerman, M. I.; Meller, A.; Chung, M.; Swamidass, S. J.; Bowman, G. R. Deep Learning the Structural Determinants of Protein Biochemical Properties by Comparing Structural Ensembles with DiffNets. *Nat Commun* **2021**, *12* (1), 3023. https://doi.org/10.1038/s41467-021-23246-1.
(41) Gupta, A.; Dey, S.; Hicks, A.; Zhou, H.-X. Artificial Intelligence Guided Conformational Mining of Intrinsically Disordered Proteins. *Commun Biol* **2022**, *5* (1), 610. https://doi.org/10.1038/s42003-022-03562-y.

(42) Janson, G.; Valdes-Garcia, G.; Heo, L.; Feig, M. Direct Generation of Protein Conformational Ensembles via Machine Learning. *Nat Commun* **2023**, *14* (1), 774. https://doi.org/10.1038/s41467-023-36443-x.
(43) Teixeira, J. M. C.; Liu, Z. H.; Namini, A.; Li, J.; Vernon, R. M.; Krzeminski, M.; Shamandy, A. A.; Zhang, O.; Haghighatlari, M.; Yu, L.; Head-Gordon, T.; Forman-Kay, J. D. IDPConformerGenerator: A Flexible Software Suite for Sampling the Conformational Space of Disordered Protein States. *J Phys Chem A* **2022**, *126* (35), 5985–6003. https://doi.org/10.1021/acs.jpca.2c03726.
(44) Zhu, J.-J.; Zhang, N.-J.; Wei, T.; Chen, H.-F. Enhancing Conformational Sampling for Intrinsically Disordered and Ordered Proteins by Variational Autoencoder. *Int J Mol Sci* **2023**, *24* (8), 6896. https://doi.org/10.3390/ijms24086896.
(45) Ramanathan, A.; Ma, H.; Parvatikar, A.; Chennubhotla, S. C. Artificial Intelligence Techniques for Integrative Structural Biology of Intrinsically Disordered Proteins. *Curr Opin Struct Biol* **2021**, *66*, 216–224. https://doi.org/10.1016/j.sbi.2020.12.001.
(46) Reddy, G.; Straub, J. E.; Thirumalai, D. Influence of Preformed Asp23−Lys28 Salt Bridge on the Conformational Fluctuations of Monomers and Dimers of Aβ Peptides with Implications for Rates of Fibril Formation. *J Phys Chem B* **2009**, *113* (4), 1162–1172. https://doi.org/10.1021/jp808914c.
(47) Itoh, S. G.; Yagi-Utsumi, M.; Kato, K.; Okumura, H. Key Residue for Aggregation of Amyloid-β Peptides. *ACS Chem Neurosci* **2022**, *13* (22), 3139–3151. https://doi.org/10.1021/acschemneuro.2c00358.
(48) Khaled, M.; Rönnbäck, I.; Ilag, L. L.; Gräslund, A.; Strodel, B.; Österlund, N. A Hairpin Motif in the Amyloid-β Peptide Is Important for Formation of Disease-Related Oligomers. *J Am Chem Soc* **2023**, *145* (33), 18340–18354. https://doi.org/10.1021/jacs.3c03980.
(49) Lopez del Amo, J. M.; Fink, U.; Dasari, M.; Grelle, G.; Wanker, E. E.; Bieschke, J.; Reif, B. Structural Properties of EGCG-Induced, Nontoxic Alzheimer's Disease Aβ Oligomers. *J Mol Biol* **2012**, *421* (4–5), 517–524. https://doi.org/10.1016/j.jmb.2012.01.013.
(50) Hsu, F.; Park, G.; Guo, Z. Key Residues for the Formation of Aβ42 Amyloid Fibrils. *ACS Omega* **2018**, *3* (7), 8401–8407. https://doi.org/10.1021/acsomega.8b00887.
(51) Tomaselli, S.; Esposito, V.; Vangone, P.; van Nuland, N. A. J.; Bonvin, A. M. J. J.; Guerrini, R.; Tancredi, T.; Temussi, P. A.; Picone, D. The A-to-β Conformational Transition of Alzheimer's Aβ-(1–42) Peptide in Aqueous Media Is Reversible: A Step by Step Conformational Analysis Suggests the Location of β Conformation Seeding. *ChemBioChem* **2006**, *7* (2), 257–267. https://doi.org/10.1002/cbic.200500223.
(52) Wälti, M. A.; Ravotti, F.; Arai, H.; Glabe, C. G.; Wall, J. S.; Böckmann, A.; Güntert, P.; Meier, B. H.; Riek, R. Atomic-Resolution Structure of a Disease-Relevant Aβ(1–42) Amyloid Fibril. *Proceedings of the National Academy of Sciences* **2016**, *113* (34). https://doi.org/10.1073/pnas.1600749113.
(53) Liao, Q.; Owen, M. C.; Olubiyi, O. O.; Barz, B.; Strodel, B. Conformational Transitions of the Amyloid-β Peptide Upon Copper(II) Binding and PH Changes. *Isr J Chem* **2017**, *57* (7–8), 771–784. https://doi.org/10.1002/ijch.201600108.
(54) Ahmed, M.; Davis, J.; Aucoin, D.; Sato, T.; Ahuja, S.; Aimoto, S.; Elliott, J. I.; Van Nostrand, W. E.; Smith, S. O. Structural Conversion of Neurotoxic Amyloid-B1–42

Oligomers to Fibrils. *Nat Struct Mol Biol* **2010**, *17* (5), 561–567. https://doi.org/10.1038/nsmb.1799.

(55) Lührs, T.; Ritter, C.; Adrian, M.; Riek-Loher, D.; Bohrmann, B.; Döbeli, H.; Schubert, D.; Riek, R. 3D Structure of Alzheimer's Amyloid-β(1–42) Fibrils. *Proceedings of the National Academy of Sciences* **2005**, *102* (48), 17342–17347. https://doi.org/10.1073/pnas.0506723102.

(56) Ngo, S.; Guo, Z. Key Residues for the Oligomerization of Aβ42 Protein in Alzheimer's Disease. *Biochem Biophys Res Commun* **2011**, *414* (3), 512–516. https://doi.org/10.1016/j.bbrc.2011.09.097.

(57) Morimoto, A.; Irie, K.; Murakami, K.; Ohigashi, H.; Shindo, M.; Nagao, M.; Shimizu, T.; Shirasawa, T. Aggregation and Neurotoxicity of Mutant Amyloid β (Aβ) Peptides with Proline Replacement: Importance of Turn Formation at Positions 22 and 23. *Biochem Biophys Res Commun* **2002**, *295* (2), 306–311. https://doi.org/10.1016/S0006-291X(02)00670-8.

(58) Wu, K. Y.; Doan, D.; Medrano, M.; Chang, C. A. Modeling Structural Interconversion in Alzheimers' Amyloid Beta Peptide with Classical and Intrinsically Disordered Protein Force Fields. *J Biomol Struct Dyn* **2022**, *40* (20), 10005–10022. https://doi.org/10.1080/07391102.2021.1939163.

(59) Tang, Z.; Chang, C. A. Systematic Dissociation Pathway Searches Guided by Principal Component Modes. *J Chem Theory Comput* **2017**, *13* (5), 2230–2244. https://doi.org/10.1021/acs.jctc.6b01204.

(60) Case Ross C Walker, D. A.; Roitberg Kenneth Merz Pengfei Li, A. M. *Amber 2020 Reference Manual Principal Contributors to the Current Codes*. http://ambermd.org/contributors.html.

(61) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J Chem Phys* **1983**, *79* (2), 926–935. https://doi.org/10.1063/1.445869.

(62) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A Smooth Particle Mesh Ewald Method. *J Chem Phys* **1995**, *103* (19), 8577–8593. https://doi.org/10.1063/1.470117.

(63) Mu, J.; Liu, H.; Zhang, J.; Luo, R.; Chen, H.-F. Recent Force Field Strategies for Intrinsically Disordered Proteins. *J Chem Inf Model* **2021**, *61* (3), 1037–1047. https://doi.org/10.1021/acs.jcim.0c01175.

(64) Chen, J.; Fasihianifard, P.; Raz, A. A. P.; Hickey, B. L.; Moreno, J. L.; Chang, C.-E. A.; Hooley, R. J.; Zhong, W. Selective Recognition and Discrimination of Single Isomeric Changes in Peptide Strands with a Host : Guest Sensing Array. *Chem Sci* **2024**, *15* (5), 1885–1893. https://doi.org/10.1039/D3SC06087J.

(65) Chang, C.-E.; Potter, M. J.; Gilson, M. K. Calculation of Molecular Configuration Integrals. *J Phys Chem B* **2003**, *107* (4), 1048–1055. https://doi.org/10.1021/jp027149c.

(66) Minh, D. D. L. Alchemical Grid Dock (AlGDock): Binding Free Energy Calculations between Flexible Ligands and Rigid Receptors. *J Comput Chem* **2020**, *41* (7), 715–730. https://doi.org/10.1002/jcc.26036.

(67) Hikiri, S.; Yoshidome, T.; Ikeguchi, M. Computational Methods for Configurational Entropy Using Internal and Cartesian Coordinates. *J Chem Theory Comput* **2016**, *12* (12), 5990–6000. https://doi.org/10.1021/acs.jctc.6b00563.

(68) Altis, A.; Nguyen, P. H.; Hegger, R.; Stock, G. Dihedral Angle Principal Component Analysis of Molecular Dynamics Simulations. *J Chem Phys* **2007**, *126* (24). https://doi.org/10.1063/1.2746330.

(69) Glorot, X.; Bengio, Y. Understanding the Difficulty of Training Deep Feedforward Neural Networks. In *International Conference on Artificial Intelligence and Statistics*; 2010.

(70) Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. **2014**.

(71) Michaud-Agrawal, N.; Denning, E. J.; Woolf, T. B.; Beckstein, O. MDAnalysis: A Toolkit for the Analysis of Molecular Dynamics Simulations. *J Comput Chem* **2011**, *32* (10), 2319–2327. https://doi.org/10.1002/jcc.21787.

(72) Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Köpf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; Chintala, S. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2019.

(73) Mu, Y.; Nguyen, P. H.; Stock, G. Energy Landscape of a Small Peptide Revealed by Dihedral Angle Principal Component Analysis. *Proteins: Structure, Function, and Bioinformatics* **2005**, *58* (1), 45–52. https://doi.org/10.1002/prot.20310.

(74) Roe, D. R.; Cheatham, T. E. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J Chem Theory Comput* **2013**, *9* (7), 3084–3095. https://doi.org/10.1021/ct400341p.