An Integrated Machine Learning Approach Delineates an Entropic Expansion Mechanism for the Binding of a Small Molecule to α-Synuclein

Sneha Menon [⊥], Subinoy Adhikari[⊥], and Jagannath Mondal[*]

Tata Institute of Fundamental Research, Hyderabad 500046, India

E-mail: jmondal@tifrh.res.in,+914020203091

Abstract

The mis-folding and aggregation of intrinsically disordered proteins (IDPs) such as α-synuclein (αS) underlie the pathogenesis of various neurodegenerative disorders. However, targeting αS with small molecules faces challenges due to its lack of defined ligand-binding pockets in its disordered structure. Here, we implement a deep artificial neural network based machine learning approach, which is able to statistically distinguish fuzzy ensemble of conformational substates of αS in neat water from those in aqueous fasudil (small molecule of interest) solution. In particular, the presence of fasudil in the solvent either modulates pre-existing states of αS or gives rise to new conformational states of αS, akin to an ensemble-expansion mechanism. The ensembles display strong conformation-dependence in residue-wise interaction with the small molecule. A thermodynamic analysis indicates that small-molecule modulates the structural repertoire of αS by tuning protein backbone entropy, however entropy of the water remains unperturbed. Together, this study sheds light on the intricate interplay between small molecules and IDPs, offering insights into entropic modulation and ensemble expansion as key biophysical mechanisms driving potential therapeutics.

Introduction

Intrinsically disordered proteins (IDPs) comprise a class of proteins that lack a unique, well-defined three-dimensional structure under physiological conditions. Defying the classical structure-function paradigm,[1,2] they are involved in crucial cellular processes such as regulation of cell cycle signalling and act as hubs in interaction networks mainly due to their promiscuous binding nature.[3–5] Emergent studies present disordered proteins as an important component of biomolecular condensates that play regulatory roles in several key cellular functions.[6–9]

[⊥]These authors contributed equally to this work

Considering the broad range of functionality of IDPs, their disregulation and altered abundance can lead to a number of diseases including cancer, diabetes, cardiovascular and neurodegenerative disorders.[1,2,10] In rational structure-based drug design, exploration of the target proteins is a crucial step based on which the interactions of small molecules are optimized to have stabilising interactions with a structured ligand-binding pocket[11,12] and is typically applied to folded proteins. IDPs, however, are considered *undruggable* since they exist as an ensemble of conformations that precludes the usage of these traditional drug-design strategies.[13] Nevertheless, a few monomeric disordered proteins such as the oncogenic transcription factor c-Myc,[14] Abeta42,[15–19]Tau,[20] PTP1B[21] and osteopontin[22] have been successfully targeted by small molecules.[23]

The IDP of this article's interest, alpha-synuclein (αS) has remained an important drug-target[24–27] owing to its long-rooted association with neurodegenerative diseases. Often, the early aggregation stages of αS into small oligomers is implicated in the causation of Parkinson's disease[28–31] and a potential therapeutic strategy has been the stabilization of αS in soluble monomeric form. In particular, the small molecule fasudil (focus of the present investigation) has been shown to interact with αS and retard its aggregation.[26] The anti-aggregative effect of fasudil on αS assembly was tested both in vivo and in vitro and it was found to significantly reduce αS aggregation in both scenarios. Characterisation of the effect of small molecule binding on the ensemble properties of IDPs and the molecular details of the stabilizing interactions can provide the basis for selectivity and can be effectively calibrated to induce the desired effect to alleviate disease conditions. In particular, latest effort by Robustelli and coworkers[32] had produced an unprecedentedly long-time scale (1500 μs) all-atom molecular dynamics (MD) simulation trajectory of small molecule drug fasudil with monomeric αS, in which the predicted protein-fasudil interactions were found to be in good agreement with previously reported NMR chemical shift data. However, comparison of the global and local structural features of the αS ensemble in neat water and that in the presence of fasudil[32] (see Figure S1-S6) did not indicate a significant difference that is a customary

signature of the dynamic IDP ensemble.

Characterization of the peculiar nature of IDPs such as αS using the conventional experimental and computational approaches have proven to be challenging. IDPs can be best described by an ensemble of conformations potentially sampled by the protein along with their associated statistical/thermodynamic weights. While biophysical methods can provide an ensemble averaged conformation of an IDP, computer simulations can extensively sample the conformational ensembles of IDPs at atomic resolutions. However, biophysical methods lack the potential to provide an ensemble view and there are inherent limitations of force fields used in computer simulations to model IDPs in agreement with experiments. Integrative approaches have been exploited in studying IDPs as well as small-molecule binding to IDPs.[33–37] In recent years, Markov State Models (MSMs) have been exploited to elucidate a states-and-rates view of biomolecular dynamics i.e. finding the kinetically relevant states and the transition rates between them.[38–40] The first step in the process of building a kinetic model / MSM of any complex biomolecular ensemble is to identify a suitable metric or collective variable (CV) that efficiently captures its conformational space. CVs such as torsion angles, distances, chemical shifts and so forth, can characterize the complex molecular motion in a high number of dimensions. However, this presents a challenge to effectively analyze trajectories generated from MD simulation. In order to address this, dimension reduction is performed to obtain a set of slow variables that can be further subjected to clustering into kinetically discrete metastable states. Since the high-dimensional input features describing the heterogeneous conformational landscape of IDPs can have non-linear relationships, non-linear dimension reduction techniques are most suitable. Recent advances in the field of machine learning have shown better results in capturing such non-linear relationships.[41] In this study, we employed a deep neural network, specifically β-variational autoencoder[42,43]to perform dimension reduction and further utilized this machine-learnt latent dimension for building MSM in order to better understand the conformational landscape of IDPs in presence and in absence of a small molecule.

In the current work, to elucidate the small-molecule mediated modulatory effect on αS ensemble, we harnessed deep-learning based β-Variational AutoEncoder (β-VAE) and Markov State Models (MSM) to dissect αS ensemble in water and in aqueous fasudil solution. A projection of the atomistically simulated conformation ensemble of αS in its apo state and in presence of Fasudil, along their respective latent space describes the conformational landscape, which is further refined by developing individual MSMs. Our result reveals that, compared to the macrostates identified in water, the presence of fasudil led to an increase in the number of metastable states. Characterisation of small-molecule interaction with the protein indicates that fasudil has differential interactions with each of these macrostates. These macrostates exhibit clear distinctions as evident from their C_α contact maps which was further analysed using a convolutional VAE. An entropic perspective at the global (whole protein) and local (residue) level reveals that fasudil binding can potentially trap the protein and the resultant states may be disordered or ordered in nature. The effect is mainly manifested at the backbone level with a decrease in entropy at the fasudil binding hotspots in some states. The observations reported in this study demonstrate how the binding of a small molecule, fasudil, modulates the conformational properties of the disordered protein, αS, manifested as a change in the backbone entropy of the fasudil binding hotspots.

Results and Discussion

1. Modulation of free energy landscape of αS in presence of fasudil

Simulations of αS monomer in the presence of the small molecule fasudil and 50 mM NaCl have been reported in a previous study by Robustelli et al.[32] This study showed no large-scale differences between the bound and unbound states of αS. In our study, in order to compare the αS-fasudil ensemble with an apo ensemble at same salt concentration, we spawned multiple all-atom MD simulations of αS in neat water and 50 mM NaCl from multiple starting conformations to generate a cumulative ensemble of $\sim$62 μs.

In a clear departure from the classical view of ligand binding to a folded globular protein, the visual change in αS ensemble due to the presence of small molecule is not so strikingly apparent. In order to understand the underlying ensemble modulatory effect of the small-molecule binding events, the complex and fuzzy conformational ensemble of αS (see Figure 1a) needs to be delineated using a suitable spatial and temporal decomposition into its key sub-ensembles or metastable states. This prompted us to analyse the two ensembles using the framework of Markov State Model (MSM). As a suitable input feature to build the MSM, we estimated the set of inter-residue Cα pairwise distances, as this feature largely incorporates the conformational space of αS. In a standard MSM analysis, this high-dimensional input feature requires further processing such as dimensionality reduction, before discretisation using clustering for the construction of an MSM.

We started our investigation by searching for an optimized representation of high-dimensional and heterogeneous ensemble of monomeric αS conformation in presence of fasudil. As the datasets, particularly in the context of IDPs, are more complex and large in size, the need for effective dimensionality reduction techniques becomes more conspicuous. Dimension reduction using principal component analysis, independent component analysis,[44] singular value decomposition[45] and linear discriminant analysis[46] linearly transform the high dimensional data into a lower dimensional manifold. However, non-linear methods such as kernel PCA,[47] multidimensional scaling,[48] isomap,[49] fastmap,[50] locally linear embedding[51] etc[52] have outperformed such linear transformation methods and have been used for free energy calculations[53] and determining timescales to capture slow conformational changes.[54,55] In the present work, we opted to draw inspiration from artificial deep neural network based framework to employ model-agnostic and mostly unsupervised non-linear approach to derive optimized non-linear latent feature space that can be employed for statistical state-space decomposition of IDP such as αS.

Over the recent years, unsupervised machine learning algorithms such as autoencoders have been effective as a dimension reduction tool, owing to non-linear activations of the

neurons that allows them to capture complex relationships and non-linear patterns in the high dimensional data.[41,56] This high dimensional input is represented into a lower dimensional latent space, which is then used to reconstruct the input. However, the latent space is "*deterministic*", i.e. it provides only a fixed mapping of the input to latent space, which might not account for variations or uncertainty present in the data. Consequently, this could generate an inaccurate result for a point that does not exist in the latent space. Thus, when projecting new, similar data onto the latent space, a deterministic autoencoder will map similar inputs to nearly identical points in the latent space, potentially lacking the diversity and capturing only a single mode of the data distribution. In contrast, the latent space of a Variational Autoencoder (VAE) is "*probabilistic*" in nature, thereby avoiding this limitation of autoencoders. The technical details have been documented in the Method section. However, here we provide a brief rationale for the choice of the β-VAE model.

Variational autoencoder (VAE)[42,43] (figure 1b) is also an unsupervised machine learning algorithm, which although is closely related to autoencoder, it is inextricably linked with variational bayesian methods. A VAE consists of an encoder network and a decoder network, much like a traditional autoencoder. The encoder takes input data and maps it to a probabilistic distribution in the latent space, while the decoder reconstructs the input data from samples drawn from this distribution. The probabilistic nature of the latent space is a result of applying variational inference, which aims to estimate the actual posterior using an approximate distribution, parameterized by a mean vector and a variance vector. This estimation is achieved by minimizing the Kullback-Leibler (KL) divergence (also known as relative entropy or information gain) between the two distributions. As a result, instead of producing a single point in the latent space for each input, the encoder outputs a distribution. The latent variable is then sampled from this distribution, which is then used by the decoder to reconstruct the input. This makes the latent space continuous, thereby improving its ability to interpolate novel, unseen data. The VAE model can be further improved by enhancing the probability of generating an actual data while keeping the dis-

tance between the true and the posterior distribution under a small threshold. This results in the weighting of the KL divergence term in the loss function. The factor determining this weight is represented by β and hence the name β-VAE, which is used in this study. The implementation of the β-VAE is made publicly available in our group's Github page (`https://github.com/JMLab-tifrh/Protein_Ligand_Variational_Autoencoder`). The loss function of β-VAE is given as,

$$\text{Loss}_{VAE} = \text{Reconstruction Loss} + \beta \times \text{KL Divergence} \tag{1}$$

The β-VAE model was trained over a number of β values ranging between 5 and 1×10^{-15}. A β value of 1×10^{-12} was chosen as this value gave us the minimum RMSE between the original and the reconstructed data (see Figure 1 d). This model was further used to project the αS data in its apo state for further analysis.

We have also computed the VAMP2 score after dimension reduction of C_α pairwise distance using a linear dimension reduction method namely time-lagged independent component analysis (tICA) and compared it with the score obtained from the latent dimension of β-VAE. We found that the VAMP2 score obtained from β-VAE was consistently higher than tICA derived dimensions for the αS-Fasudil ensemble, thereby indicating that β-VAE would help identifying relatively slower CVs (see Figure S7-S8) as compared to tICA and hence better at representing slower dynamical processes in the αS and αS-Fasudil simulations.

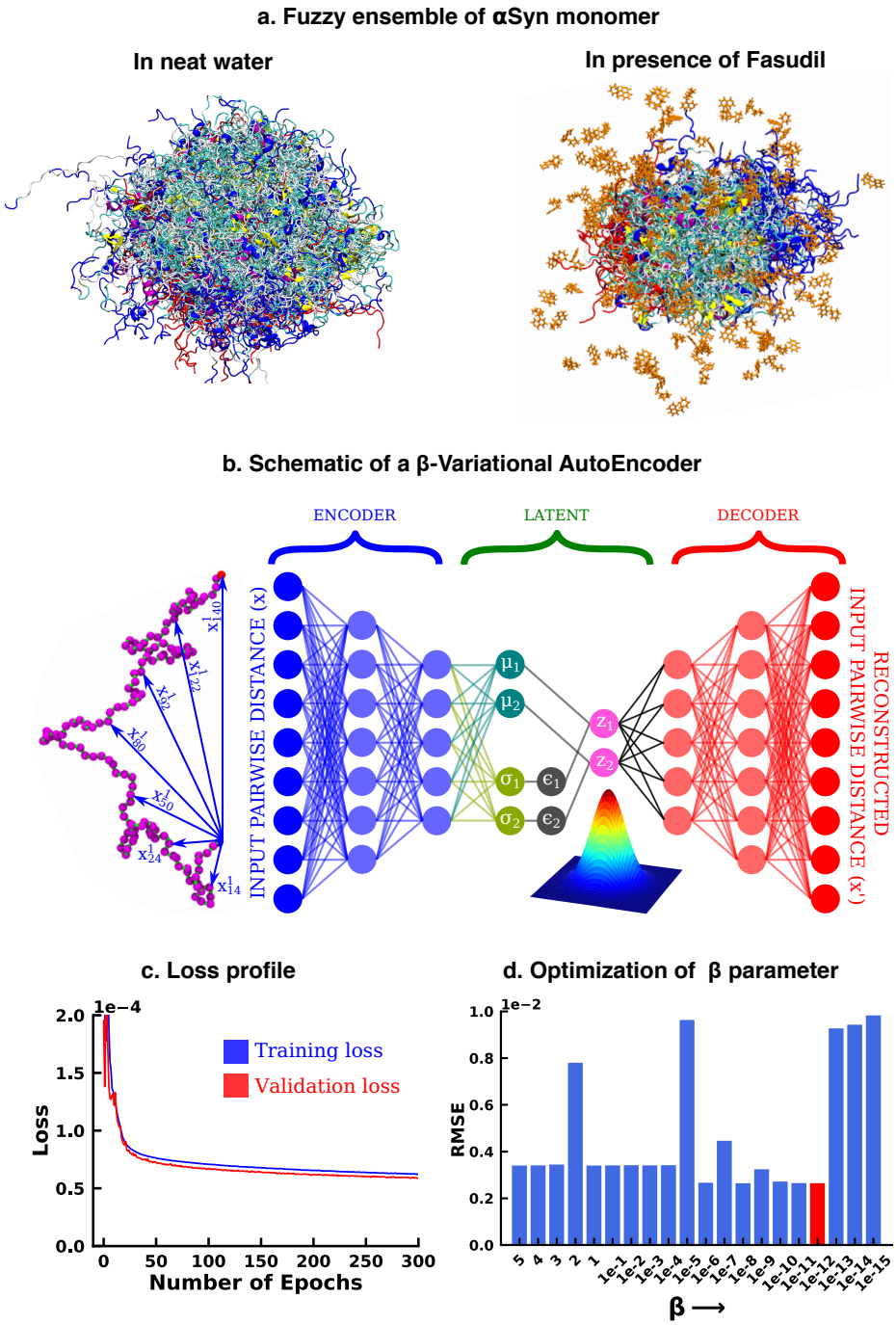


Figure 1: (a) Conformational ensemble view of αS and αS-Fasudil ensemble, (b) A schematic of β-Variational Autoencoder (β-VAE), (c) Training and validation loss for β-VAE and (d) RMSE as a function of the β parameter. The red-annotated β parameter was used for the investigation

In particular, as described in detail in the Methods section, we trained a VAE model using a large array of inter-residue pairwise distances using the β-VAE architecture. Our attempt to reconstruct the large-dimensional input feature via a β-VAE resulted in compression of the 1.5 millisecond MD simulation data into a two-dimensional latent space. The loss function steadily decreased and eventually plateaued providing a model optimized for this system (see Figure 1 c). The rugged nature of the free energy landscape of αS, is evident from its projection along the latent space dimension of simulated trajectory either in aqueous media (see Figure 2a) or in presence of fasudil (see Figure 2b). The free energy landscape represents a large number of spatially close local minima representative of energetically competitive conformations inherent in αS, which is also the source of the conformational heterogeneity and a hallmark of an IDP(see Figure 2 a-b). Next we wanted to compare the changes in the conformational landscape of the apo αS conformations. Subsequent projection of the apo state on the low-dimensional subspace also brought out a spatially heterogeneous landscape.

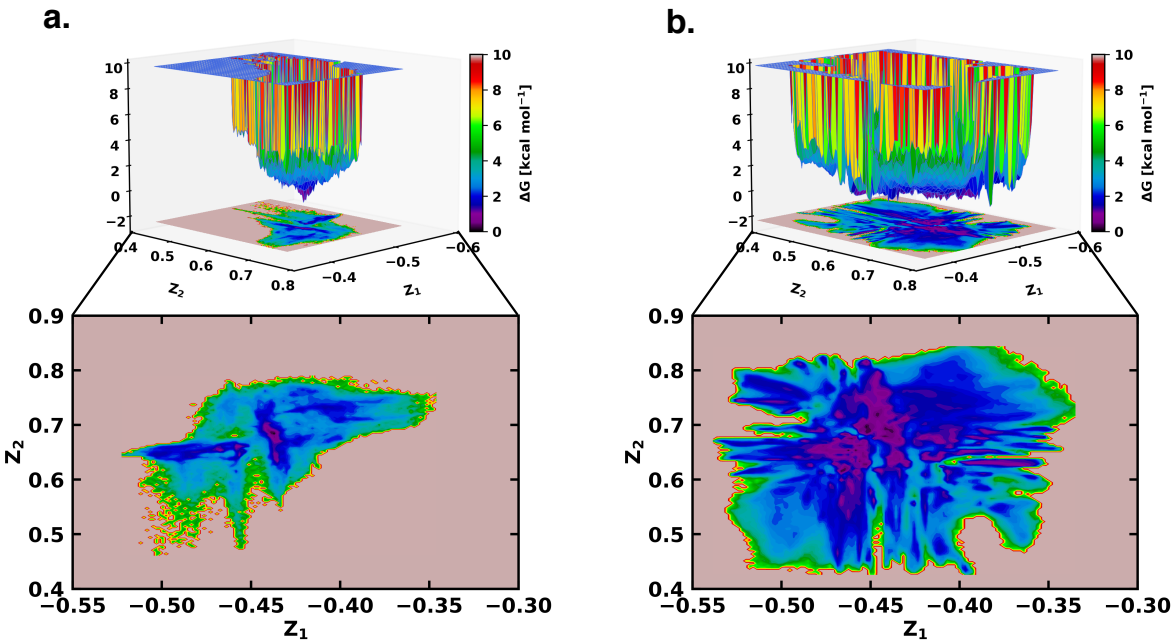


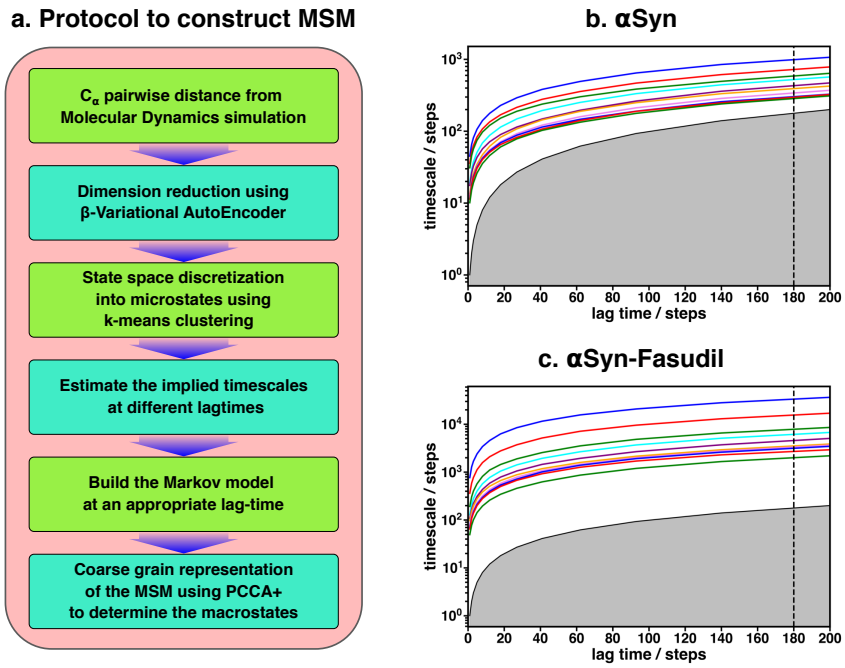
Figure 2: The free energy landscape of (a) αS and (b) αS-Fasudil system as determined from the latent dimensions of β-VAE

However, a close comparison of free energy landscapes indicates that a set of local minima representing conformations of αS are less clustered and appear in small patches in the apo state. Moreover, a relative comparison of the two free energy landscapes suggests that αS spans a significantly larger space in presence of fasudil even in this low dimensional subspace, hinting at an expansion of conformational repertoire of this IDP in presence of the small molecule (see Figure 2 a-b). Individual, one-dimensional projection of the conformational landscape along each of the two latent dimensions reveals that that the presence of the small molecule would shift part of the conformational landscape to a distinct location, suggesting that fasudil would modulate the existing conformational ensemble as well as would create new conformational space. For a more discrete and clearer state-space decomposition of the conformational ensemble of αS (in apo or in ligand-bound form), we attempt to enumerate the optimum number of distinct metastable macrostates with non-negligible equilibrium populations by constructing MSM as explained in the following section.

2. Markov State Model elucidates distinct binding competing states of αS in presence of the small-molecule drug

The two-dimensional feature obtained from β-VAE was used as the input feature for the MSM estimator provided by PyEMMA.[57] A flowchart of the implemented protocol is drawn in Figure 3a. A geometric clustering approach, k-means, was used to discretise the ensemble. Analysis of the implied time scales as a function of lag-time predicted a Markovian model of αS in water with three kinetically separated states (see Figure 3d for relative population). On the other hand, a similar analysis of αS in presence of fasudil (Figure 3c)) indicated a clear increase in the number of spatially and temporally resolved conformational metastable states. Thus, a three-state MSM was built for the αS ensemble in water whereas a six-state model was built for the αS-fasudil ensemble. Hereafter, we will refer to the macrostates created in presence of fasudil prefixed with FS and those generated in water prefixed with MS. A bootstrapping analysis was performed to estimate the mean and standard deviations

of the equilibrium populations of these metastable states (Figure 3d).



d. MSM derived macrostate populations

αSyn		αSyn-Fasudil	
Metastable State	Population (%)	Metastable state	Population (%)
MS1	10.35 ± 0.42	FS1	8.70 ± 0.37
		FS2	5.47 ± 0.12
MS2	61.21 ± 2.62	FS3	2.63 ± 0.45
		FS4	7.58 ± 0.99
MS3	28.44 ± 2.45	FS5	71.89 ± 1.33
		FS6	3.72 ± 0.53

Figure 3: (a) A flowchart of the process of building a Markov State Model (MSM). (b, c) Implied timescales plot of the αS and αS-Fasudil systems. (c) Macrostate populations of the 3-state and 6-state MSM of the αS and αS-Fasudil systems. Bootstrapping was used to estimate the mean and standard deviations of the macrostate populations.

The MSM was also further validated for Markovianity i.e. ability to predict estimates at longer time scales using the Chapman-Kolmogorov (CK) test, as depicted in Figure S9 for αS-Fasudil simulation. This test further validates the quality of the model that it estimates long-time scale behavior with reasonable accuracy. The MSM analyses essentially indicate that the small-molecule binding interactions with αS led to sampling or generation of additional conformational states than the states populated in the apo αS ensemble.

A pertinent question arises: Are the appearances of newer states in αS in presence of fasudil a result of the presence of longer trajectory in fasudil solution (1500 μs) than that of the ensemble populated in neat water (62 μs)? We verified this by building an MSM for 60 μs of data of fasudil system using 10 μs to 70 μs segment of the trajectory (see Figure S10). We find that the MSM built using lesser data (and same amount of data in the αS simulation) also indicated the presence of six states of αS in presence of fasudil, as was observed in the MSM of the full trajectory. We compared the R_g distribution of this small chunk with the αS ensemble and found to be in good agreement (see Figure S11). Addidtionally, we selected a chunk spanning from 966 μs to 1026 μs and observed a similar distribution of R_g (see Figure S12). Moreover, we built a MSM on this new chunk and identified the presence of six states (see Figure S13). Together, this exercise invalidates the sampling argument and suggests that the increase in the number of metastable macrostates of αS in fasudil solution relative to that in water is a direct outcome of the interaction of αS with the small molecule.

Another aspect that we evaluated was in the choice of the number of latent dimensions of β-VAE. Even though an increase in the number of latent dimensions may make the model more accurate, this can also result in overfitting. The model can simply memorize the pattern in the data instead of generalizing them. A higher dimensional latent space is also more difficult to interpret; therefore, we chose two dimensions. Nevertheless, we trained another β-VAE with four neurons on the latent space and built a MSM by choosing the appropriate number of microstates (see Figure S14). The implied timescales (see Figure S15) indicate the presence of six states which is consistent with the model with two latent dimensions.

3. Structural characterisation of metastable states of αS monomer in presence of fasudil

We characterised the residue-wise intramonomer contact maps of the metastable states to identify the differences in the interaction patterns in these states that can be attributed

to interactions of fasudil with the monomer. The average inter-residue contact probability maps for each of the metastable states populated in the presence of fasudil and in water are depicted in Figure 4 and Figure 5, respectively. The residue-wise percentage secondary structure in each of the macrostates are presented in Figure S16 and S17, respectively.

In state FS1, antiparallel β-sheet interactions are formed within the N-terminal region. This β-sheet network also includes long-range interactions of the H2 region of the N-terminus with the negatively charged C-terminal region. Residues 70-80 in the hydrophobic NAC region are also involved in antiparallel β-sheet interactions with residues 120-130 in the C-terminal region. Short parallel β-sheet interactions also exist between the NAC and C-terminal residues with the N-terminal residues. The residues in the C-terminus exhibit 40-90 % of β-sheet propensity (see Figure S16). Long range interactions between N-terminus and C-terminus are relatively diminished in state FS2. This state has multiple parallel β-sheet interactions between short stretches of residues in the NAC region and C-terminal region with the N-terminal region. In addition. antiparallel β-sheet interactions are also present in NAC:N-terminus, NAC:C-terminus and within the N-terminal region. Residues 8 to 13 in the N-terminus and 86 to 92 in the NAC region exhibit significant helicity. State FS3 is characterised by $\sim$20 % of β-sheet interactions across the entire protein. Hydrophobic and polar, uncharged residues in the range 61 to 80 in the NAC region are involved in β-sheet interactions with the N-terminus (residues 1 to 10) and C-terminus (100 to 120). Moreover, β-sheet interaction networks are also present within the NAC and N-terminal regions. The C-terminal region 100 to 120 forms long-range interactions with residues 1 to 10 in the N-terminal region. Extensive short and long-range β-sheet networks are prevalent in state FS4. The entire NAC region forms antiparallel β-sheet network with the N-terminus. The C-terminus forms parallel β-sheets with the NAC and N-terminus. In the most populated state FS5, the key interactions are β-sheets formed between the oppositely charged termini. The NAC region also exhibits α-helical propensity. Finally, in state FS6, the NAC region (residues 60 to 80) forms antiparallel β-sheet network with the residues 30 to 60 of the N-

terminal segment as well as the C-terminal region 100 to 120. These residues have ∼40 % β-sheet propensity.

On the other hand, the metastable states that are populated in neat water are distinct from those formed in the presence of fasudil. In state MS1, the NAC region residues 61 to 80 interacts with the N-terminal region by both parallel and antiparallel β-sheet interactions. The β-sheet network of NAC also incorporates interactions with residues 100 to 130 in the C-terminal region. These residues have ∼60 % β-sheet propensity (see Figure S17). State MS2, the most populated macrostate in neat water, is characterised by β-sheet interactions of the NAC region with H2 region (residues 30 to 60) of the N-terminus. Furthermore, the N and C-termini also form β-sheet interactions. State MS3 mainly has long-range interactions as β-sheets between the N and C-terminal regions. Moreover, a β-sheet network also exists within the NAC region and NAC:N-terminal (H2) interface. The residues in the NAC region have ∼50% β-sheet propensity in this macrostate.

We note that the differences in the structural features i.e. intra-protein contacts of the macrostates arise from the interactions of fasudil with the protein residues. For instance, in state FS2 in which fasudil interacts with multiple sites across the entire length of the protein, the secondary structure properties of this state are dictated by β-sheets formed by short stretches of residues rather than an extensive network. Furthermore, owing to the relatively stronger small-molecule binding interactions with hydrophobic residues in the N-terminus, the long-range interactions of the oppositely charged terminal regions are suppressed. Similarly, in state FS4, since fasudil primarily interacts with residues 120 to 140 of the C-terminus and other interactions are transient, the conformational properties of this metastable state is marked by a network of β-sheets across the rest of the protein. Essentially, small-molecule binding interactions via stabilisation of various intra-protein interactions modulate the conformational properties of αS.

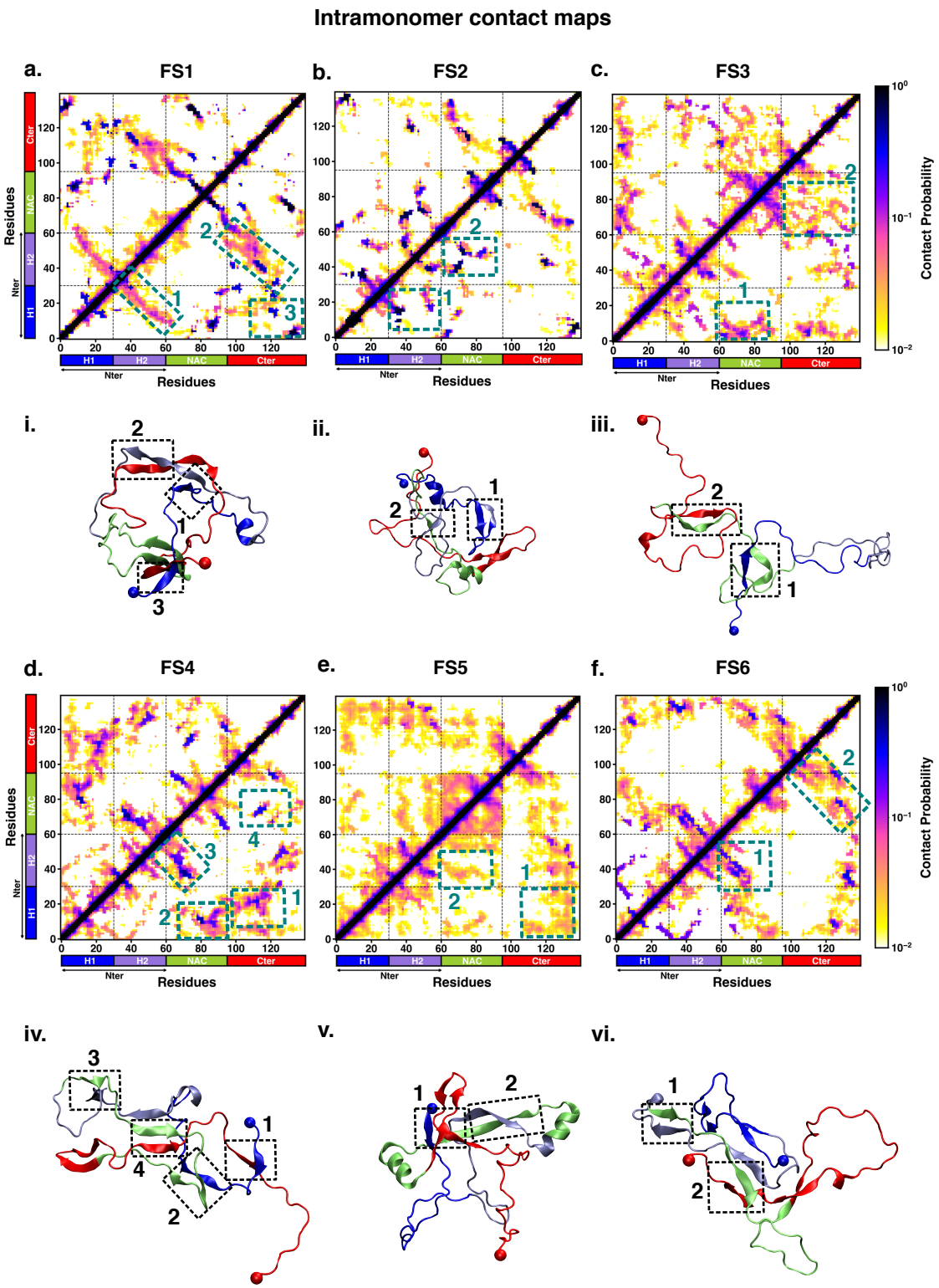


Figure 4: Intrapeptide residue-wise contact probability maps of (a) six macrostates (FS1 to FS6) in the presence of fasudil. A contact is considered if the C_α atoms of two residues are within a distance of 8 Å of eachother. Axes denote the residue numbers. The color scale for the contact probability is shown at the extreme right of each panel of maps. The color bar along the axes of the plots represents the segments in the αS monomer. Specific contact regions are marked by boxes and numbered. These contacts are illustrated by representative snapshots and the corresponding contacts are similarly marked and numbered.

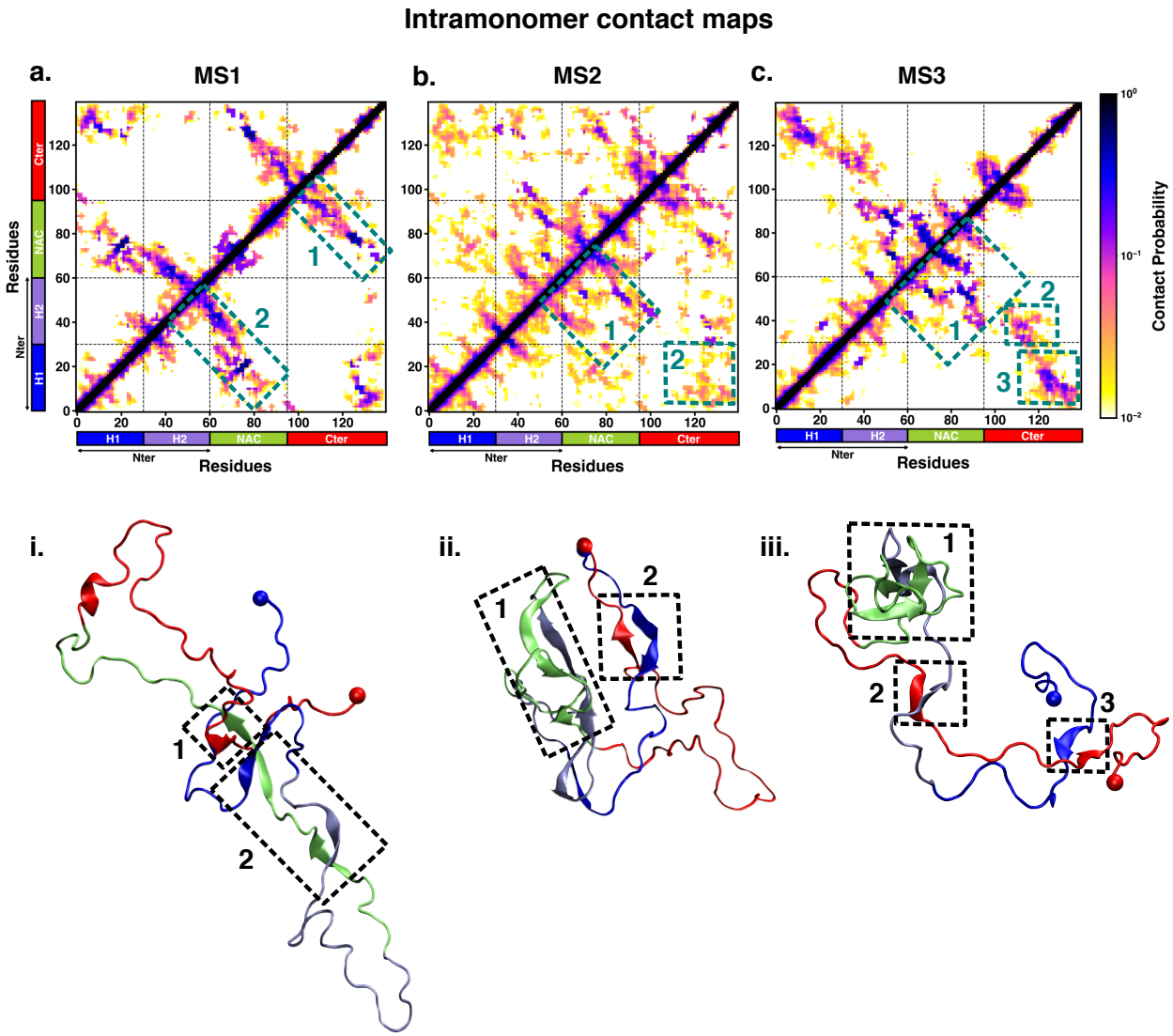
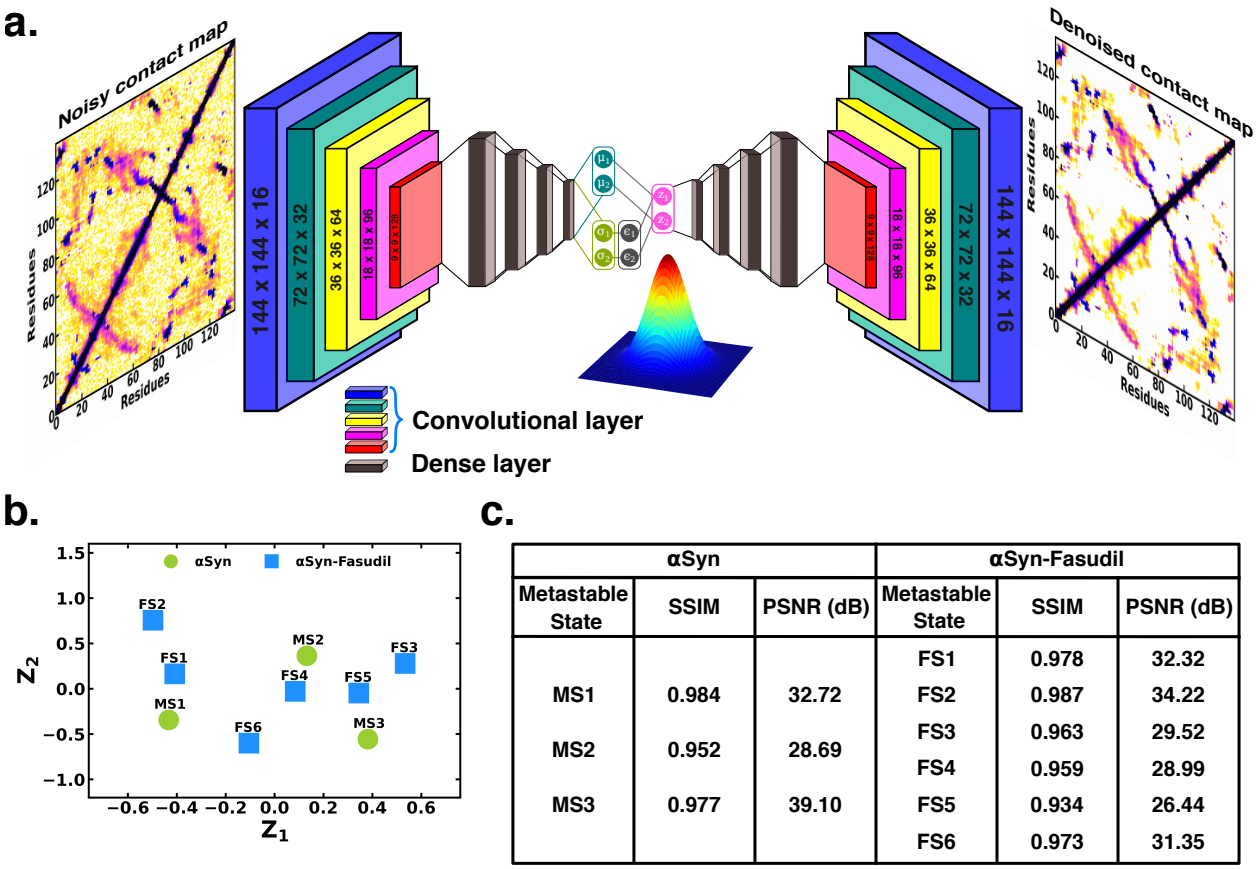


Figure 5: Intrapeptide residue-wise contact probability maps of the three macrostates (MS1 to MS3) in neat water. A contact is considered if the C_α atoms of two residues are within a distance of 8 Å of eachother. Axes denote the residue numbers. The color scale for the contact probability is shown at the extreme right. The color bar along the axes of the plots represents the segments in the αS monomer. Specific contact regions are marked by boxes and numbered. These contacts are illustrated by representative snapshots and the corresponding contacts are similarly marked and numbered.

4. Mapping the macrostates in αS and αS-Fasudil ensembles using Denoising Convolutional Variational Autoencoder

To ascertain the distinctions among these contact maps for the αS and αS-Fasudil system, we chose to filter the essential features and patterns by representing it in a lower dimensional space. Initially, this was attempted using a Convolutional Variational AutoEncoder (CVAE) with the contact maps of αS and αS-Fasudil macrostates as inputs. The model was initially trained using the six contact maps of αS-Fasudil system as a test case.

However, the model displayed a tendency to overfit by learning an identity function between the input and output, as evident from the large reconstruction error of the test dataset (see Figure S18 (a)). This overfitting issue was addressed by training a Denoising CVAE (DCVAE), where the contact map was intentionally corrupted with noise prior to training. This approach successfully mitigated the problem of overfitting as indicated by the loss profile of the test dataset (see Figure S18 (b)). An implementation is provided in GitHub (`https://github.com/JMLab-tifrh/Protein_Ligand_Variational_Autoencoder`)



αSyn			αSyn-Fasudil		
Metastable State	SSIM	PSNR (dB)	Metastable State	SSIM	PSNR (dB)
			FS1	0.978	32.32
MS1	0.984	32.72	FS2	0.987	34.22
			FS3	0.963	29.52
MS2	0.952	28.69	FS4	0.959	28.99
MS3	0.977	39.10	FS5	0.934	26.44
			FS6	0.973	31.35

Figure 6: (a) Schematic of denoising convolutional variational autoencoder (b) Latent space of the contact map of αS and αS-Fasudil metastable states. (c) SSIM and PSNR values for αS and αS-Fasudil metastable states.

The denoising performance was measured by calculating the structural similarity index measure (SSIM) between the original and denoised contact map. SSIM measures the similarity between two images as perceived by the human eye by considering the luminance, contrast and structure of the image. SSIM values can range from -1 to 1, which represents perfect anti-correlation and perfect similarity, respectively.[58] The SSIM values is tabulated in Figure 6. We find that the SSIM values for all the metastable states of αS and αS-Fasudil system are close to 1 (see Figure 6), thereby indicating that the denoised contact map as obtained from the trained DCVAE model is very close to the quality of the original contact map (see Figure S19-S21)). This is also supported by the high Peak signal-to-noise ratio (PSNR). PSNR is the ratio of the maximum possible power of a signal to the power of cor-

rupting noise that measures the fidelity of a reconstructed image compared to its original in decibels (dB).[58] A higher value of PSNR indicates smaller difference between the original and denoised contact map (see Figure 6). Upon visualizing the latent space, it became evident that the contact map of the metastable states of αS and αS-Fasudil occupy distinct locations within the latent space, which indicates that the DCVAE model effectively captured the distinct patterns among the metastable states.

5. Fasudil exhibits conformation-dependent interactions with individual metastable states

We next sought to understand the specificities of the interactions of fasudil with the different metastable states of αS. Accordingly, we calculated the contact probabilities between fasudil and each αS residue. A contact between fasudil and a protein residue is considered to be present when the minimum distance between any heavy atom of fasudil and a protein residue is less than 0.6 nm. This is calculated for the entire population of each macrostate to determine the contact probabilities. The αS-Fasudil contact maps for the six macrostates are presented in Figure 7a and snapshots of the interactions are presented in Figure 7b.

The residue-interaction profiles of each of the six individual macrostates of αS with fasudil revealed distinct motifs of interaction across the length of the protein for the different macrostates (Figure 7). Previous investigations[26,32] have reported dominant fasudil-αS interaction via the C-terminal region of the protein while it weakly interacts with the entire αS sequence. In state FS1, fasudil dominantly interacts with residues 133-139 of the C-terminal region, which consists of negatively charged residues D135 and E137 as well as the polar, neutral tyrosine residues (Y133 and Y136).

In addition, there are also non-specific interactions of fasudil with hydrophobic residues in the NAC region (residues 71 to 74). In state FS2, fasudil interactions are spread across the entirety of the protein. These comprise hydrophobic residues in the N-terminal region, neutral polar residues (62 to 66) in the NAC region and residues 130 to 136 in the C-terminal

region. Fasudil interacts strongly with the NAC region consisting of hydrophobic and polar, neutral residues in state FS3. Interactions also prevail with the negatively charged C-terminal region and residues 1 to 6 of the N-terminal region. While interacting non-specifically with residues in the amplipathic N-terminal region of state FS4, fasudil preferentially interacts with residues 121-140 of the C-terminal region via charge-charge and π-stacking interactions with the negatively charged residues (D135, E137, E139) and tyrosine side chains (Y125, Y133, Y136). In the most populated state FS5, fasudil interacts weakly with the hydrophobic residues and glutamic acid residues while interactions with the C-terminal region 107 to 140 comprising of tyrosines and acidic amino acids are dominant. Lastly, in state FS6, the strongest interactions of fasudil comprise hydrophobic and polar residues in the range 118 to 121. There are also weak interactions with hydrophobic residues in the N-terminus (11 to 16, 38, 39) and NAC regions (71 to 74, 80 to 95).

Robustelli et al.[32] have reported that in the bound state, fasudil interactions with αS are favorable charge-charge and π-stacking interactions that form and break in a mechanism, they term as *dynamic shuttling*. The stacking interactions arise from the favourable orientation of fasudil's isoquinoline ring and the aromatic ring in the side chain of tyrosine residue. Analysis of the time-series of the formation of different intermolecular interactions indicated the formation of charge-charge and π-stacking interaction with residues Y125, Y133 and Y136 and the shuttling among these interactions causes fasudil to remain localized to the C-terminal region. Notably, we observe that each of the metastable, while preferentially interacting with the C-terminus also has distinct yet significant interactions with hydrophobic residues in the N-terminus and NAC regions.

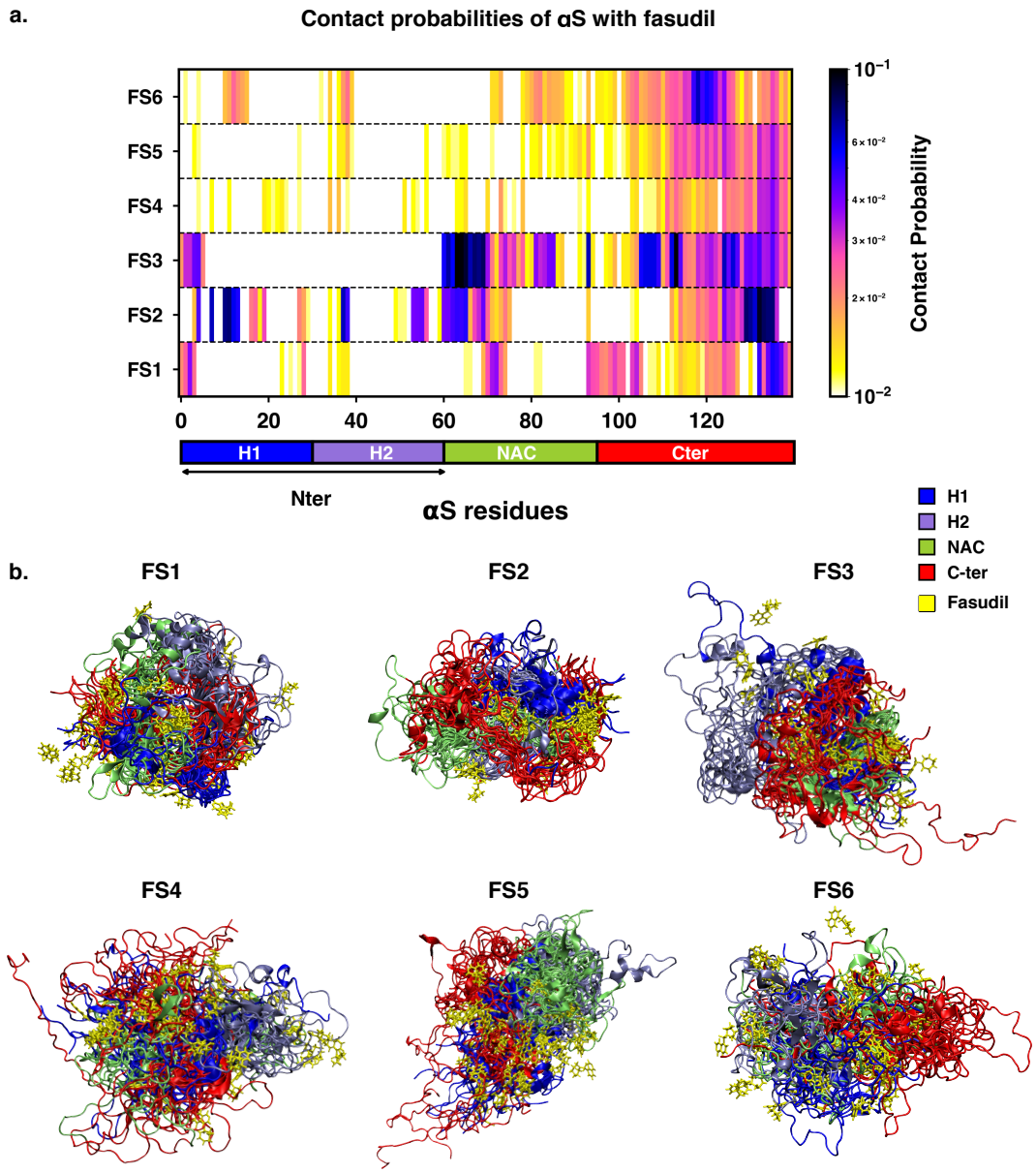


Figure 7: (a) Contact probability map of residue-wise interaction of fasudil with αS. (b) Overlay of representative conformations from the six macrostates (FS1 to FS6) along with the bound fasudil molecules. The conformations are colored segment-wise as shown in the legend. The fasudil molecules, in licorice representation, are colored yellow.

6. Entropic Signatures of Small Molecule Binding

The overall analysis of the conformational ensemble of αS in presence of fasudil demonstrates that small-molecule binding substantially modulates the ensemble of αS. Our results

indicate that the interaction of fasudil with αS residues governs the structural features of the protein. Estimation of the metastable states of αS shows that compared to aqueous conditions in which three metastable states are identified, the ensemble populated in the presence of fasudil in solution contains six metastable states. Since IDPs are described as an ensemble unlike folded proteins with a fixed structure, the effect of small molecules binding on disordered proteins or regions, is manifested as modulation of the whole ensemble by shifting the population of the states leading to a change in its conformational entropy.[59–61] The common mechanism that is generally postulated is the *entropic collapse* or *folding upon binding* model.[62] It describes the disorder-to-order transition that IDPs typically undergo upon interacting with its biological target. This transition leads to the shift in population to a more ordered, predominant state causing a loss of conformational entropy that is compensated by a corresponding enthalpic gain. Conversely, small-molecule binding may also lead to *entropic expansion* in which newer conformational states are populated that are otherwise undetectable.[60] Additionally, in cases where there is no gross change in the conformational entropy and the binding is considered *fuzzy*,[34] referred to as *isentropic shift*.

The observations reported in the present investigation for αS monomer plausibly conform with the entropy-expansion model.[59] According to this model, the interactions of small molecules with IDPs can promote the exploration of conformations with discernible probability, thereby expanding the disorderedness (conformational entropy) of the ensemble.[19,59,60,63] The small molecule interaction with αS operates transiently, and not strongly as seen in structured proteins. These transient local interactions stabilise the conformations that are inaccessible in the absence of the small molecule.

To gain deeper insights into the entropic effect of small-molecule binding to αS, we further estimated protein conformational entropy as well as water (solvent) entropy. We evaluated the entropy of water using the 2PT[64,65] method. First, we determined the molar entropy of pure water that yielded an entropy value of 55.5 J mol^{-1} K $^{-1}$, which is close to the reported value of 54.4 J mol^{-1} K $^{-1}$ from free energy perturbation calculation.[66] However, the molar

entropy is marginally reduced in all the macrostates of both αS and αS-fasudil systems, with values ranging between 55.01 J mol^{-1} K $^{-1}$ to 55.25 J mol^{-1} K $^{-1}$, with difference from that of bulk water being negligible. Upon closer investigation of the translational and rotational components of the total water entropy, we find that that the decrease of the translational component is more pronounced as compared to the rotational component. This implies that αS in its apo state or in the presence of fasudil has a comparatively greater impact on the translational component than on the rotational component of the entropy of water (see Figure S22). Nonetheless, the entropy analysis suggests that the presence of fasudil has an insignificant impact on the entropy of water.

Next, we estimated the protein conformational entropy using the PDB2ENTROPY program[67] that implements the Maximum Information Spanning Tree (MIST) approach[68,69] on sets of torsional angles of the protein. The entropy is calculated from probability distributions of torsional degrees of freedom of the protein relative to random distribution i.e. corresponding to a fully flexible protein chain. The total protein entropy values for the metastable states in water (MS1 to MS3) and fasudil-bound states (FS1 to FS6) are shown in Figure 8. While the conformational entropies of the ligand-free states, MS1 to MS3, range from -72 to -75 J mol^{-1} K^{-1}, the entropy of fasudil-bound states (FS1 to FS6) vary from -70 to -86 J mol^{-1} K^{-1}, indicating that binding of fasudil to αS leads to a broadening of the span of entropy of these states. The highest entropy among the metastable states in the two environments correspond to FS5 and MS2 with values of -70.8 J mol^{-1} K^{-1} and -72.3 J mol^{-1} K^{-1}, respectively. These states are also the most populated metastable states. MS1 and MS3 states have entropy values of -73.9 J mol^{-1} K^{-1} and -74.5 J mol^{-1} K^{-1}, respectively. Notably, the presence of fasudil and its interactions with the protein chains can elicit the exploration of states that are more ordered in nature or increasingly deviated from a random chain. The largest deviation from a fully disordered chain is attained by the state FS1 with entropy value of -85.3 J mol^{-1} K^{-1}, followed by FS2 having entropy of -83.1 J mol^{-1} K^{-1}. FS4 is also relatively more ordered than FS5 with entropy of -76.2 J mol^{-1} K^{-1}.

FS3 and FS6 are closely disordered as FS5 with entropy values of -73.2 J mol^{-1} K^{-1} and -72.5 J mol^{-1} K^{-1}, respectively. These observations of the total extent of disorderedness of the metastable states clearly indicate that the interactions of fasudil with αS residues can restrict their conformational degrees of freedom thus entrapping the protein in specific conformational states, thus manifesting as overall decrease in the protein entropy.

Considering the diffuse binding of fasudil across several residues along the entire length of the protein, we further estimated and analysed how the entropic contributions are affected at the residue level (shown in Figure 8 and Figure S23, S24). The total residue-wise entropy presented in Figure S23 shows that the residues in the N-terminal region of fasudil-bound states FS1 and FS2, exhibit greater entropic fluctuations i.e. more negative values from the apo metastable states (MS1 to MS3), suggesting more ordered or restricted fluctuations of these residues. We note that the propensity of fasudil binding to the N-terminal residues is higher in FS1 and FS2, relative to other states (see Figure 7). We further decoupled the residue entropy into backbone and sidechain components; this analysis is depicted in Figure 8 and Figure S24, respectively. Comparison of the sidechain entropies, computed from the χ torsion angles, for the unbound and bound states, does not show significant alteration upon small-molecule binding. Conversely, increased fluctuations in residue backbone entropies are observed in response to small-molecule binding (see Figure 8). The metastable states, MS1 to MS3, populated in water have minimally fluctuating residue entropies amongst the states. In contrast, significant fluctuations in residue entropies are noted predominantly in the N-terminal region (residues 1 to 60) and to some extent in the NAC region for states FS1 and FS2. These residues in the N-terminal region in FS1 and FS2 experience a drop in backbone entropy owing to the restricted degrees of freedom from

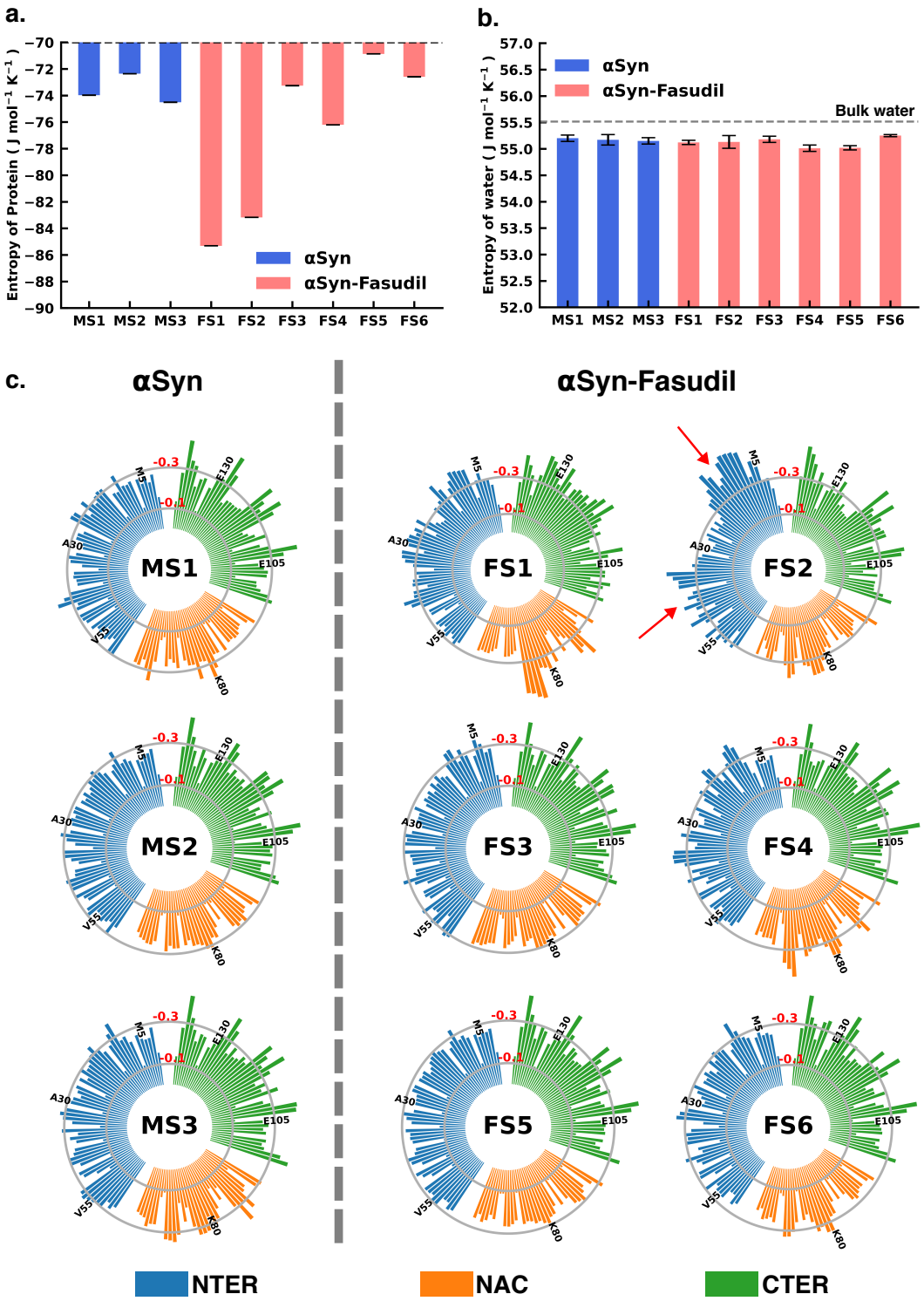


Figure 8: (a) Total Protein entropy of the metastable states of αS and αS-Fasudil calculated from the torsion angles, relative to a fully flexible chain (b) Total water entropy in αS and αS-Fasudil system and (c) Residue-wise backbone entropy within the αS and αS-Fasudil states.

fasudil binding. The decrease in residue backbone entropy correlates with fasudil binding hotspots and therefore can be ascribed to small-molecule interactions. Hydrophobic residues in the NAC region show decrease in entropy in ligand-binding locations mainly in state FS1, FS2 and FS4. However, intriguingly, we note that the correlation between fasudil interactions and residue entropy is not observed in state FS3. Finally, in the negatively charged C-terminal region, where fasudil preferentially targets the entire region in the αS protein, residue-wise entropic fluctuations are minimal compared to the apo state. Robustelli et al. refer to the interaction of fasudil with the C-terminal region as dynamic shuttling, in which breaking and formation of interactions with nearby residues occur simultaneously. This dynamic shuttling mechanism could possibly minimise the backbone entropic fluctuations compared to the relatively specific binding in the N-terminus and NAC regions.

7. Exploring interplay of conformational entropy and Mean First Passage Time in Protein Conformational Transitions

The Mean First Passage Time (MFPT) for transition among the metastable states provides valuable information for understanding the dynamics and conformational changes in the system. MFPT rate defines the average time it takes for a system to transition between one metastable state to another. A lower MFPT value suggests a favourable transition among the states. The transition timescales for αS indicate that in the presence and absence of fasudil, the transition to the most populated state is preferred (see Figure 9). For αS-Fasudil system, the transition to the most populated states takes place over timescales ranging from 74 μs and 766.5 μs, whereas for the apo state it occurs between 29.8 μs and 37.4 μs. Upon comparing these transition times with those to other macrostates, it becomes evident that these timescales represent the shortest durations. This observation suggests that the transition to the MS2 and FS5 state in αS and αS-Fasudil system is the most favoured over all other states in the system.

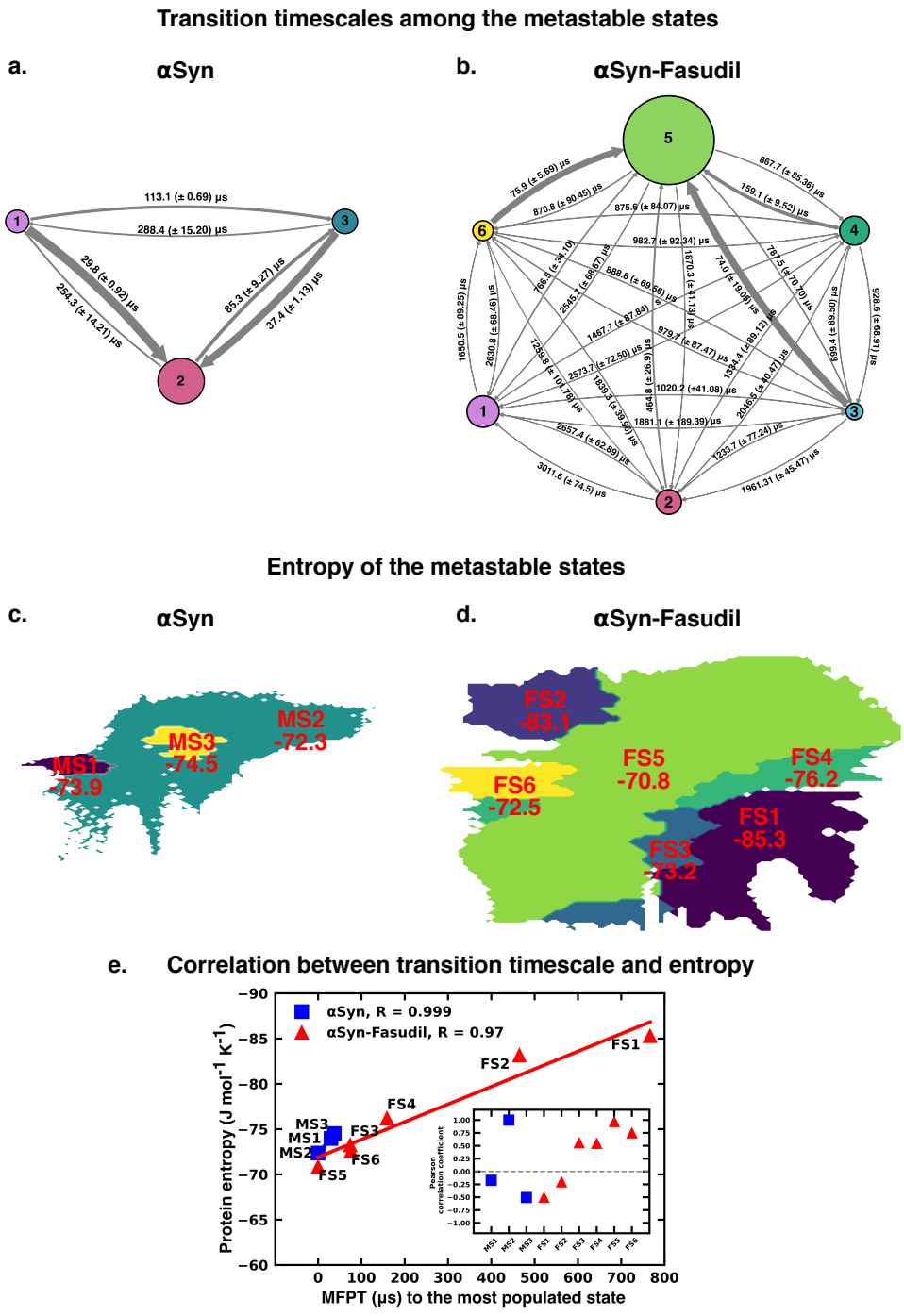


Figure 9: The rate network obtained from MFPT analysis for the (a) 3-state MSM of the αS ensemble in neat water and (b) 6-state MSM of the αS-Fasudil ensemble. The sizes of the discs is proportional to the stationary population of the respective state. The thickness of the arrows connecting the states is proportional to the transition rate (1/MFPT) between the two states and the MFPT values are shown on the arrows. (c, d) Projection of conformation sub-ensemble PCCA+ for αS and αS-Fasudil system in latent space, with respective protein entropy values (in unit of J mol^{-1} K $^{-1}$) annotated on top of it. (e) Correlation between protein entropy and transition time to the major state for αS and αS-Fasudil system. The inset plot corresponds to the correlation between entropy and transition to the macrostate for αS and αS-Fasudl system.

While MFPT offers insights into the kinetics of the metastable states, the thermodynamic property, entropy, characterizes the diversity and disorder within the system. Although a state with higher population might not always have the higher entropy, we observed that in our case, the state with the highest population also has the highest protein entropy, for both αS and αS-Fasudil system. States with higher entropy are thermodynamically favored which in turn can affect the rates of state transitions, potentially influencing MFPT values. To decipher this, we have compared the transition times of each metastable state to the most populated state with the entropy of the state, and found a strong degree of correlation (see Figure 9e). This suggests that a state with higher entropy, which also has more number of accessible microstates, potentially provide multiple pathways for other macrostates to access this state for both αS and αS-Fasudil system. However, as compared to αS in its apo state, the apparent increase ($\sim$ 2-20 times) in the time to transition to the major state in αS-Fasudil system, suggests that fasudil may potentially trap the protein conformations in the intermediate states, thereby slowing down αS in exploring the large conformational space in presence of fasudil.

Conclusions

The conformationally dynamic disordered proteins challenge the application of conventional structure-based drug-design methods that relies on the specificity of a binding pocket, based on which the drug molecule is designed and optimized. A general framework describes the effect of small-molecule binding to IDPs by modulation of the disordered ensemble by increasing or decreasing its conformational entropy. Here, we characterised the effect of the small molecule, fasudil, on the conformational characteristics of the quintessential IDP, α-synuclein. Fasudil has been experimentally reported to curb the aggregation of αS both in vivo and in vitro. The ensemble view of long time scale atomic simulations of fasudil binding to αS monomer displays a fuzzy ensemble of αS with preferential binding of fasudil

to the C-terminal region via charge-charge interactions and aromatic stacking. Using a deep neural-network based machine learning approach, variational autoencoder, we reconstructed a high dimensional feature i.e. Cα-pairwise distance into a two-dimensional latent space. We built a Markov State Model using this data to delineate the metastable conformational states of αS and identify the differences that arise from small-molecule interactions. Comparison of the metastable states of αS in the presence and absence of the small molecule fasudil revealed that the small-molecule binding navigates the structural landscape of the protein. Importantly, more number of metastable states are populated indicating that small-molecule mediated conformational modulation led to entropy expansion (see Figure 9d-e).

Atomic-level MD simulations have slowly emerged as a valuable tool for complementing experimental measurements of disordered proteins and providing detailed descriptions of their conformational ensembles.[70–73] With improvements in force fields[33,74,75] that model disordered proteins, the accuracy of MD simulations has dramatically increased, as assessed by their agreement with a wide variety of experimental measurements. Integrative approaches use a combination of experimental restraints such as chemical shifts, paramagnetic relaxation enhancement (PRE),[70,71,73,76] residual dipolar coupling constants (RDCs),[76–78] small-angle X-ray scattering (SAXS)[79,80] and computational models from MD simulations to generate a more accurate description of the ensemble of disordered proteins. More recent advances include statistical approaches such as Bayesian inference models like Metainference,[34] Experimental Inferential Structure Determination (EISD)[81,82] and Maximum Entropy formulations[83–85] that also take into account the experimental and back-calculation model errors and uncertainties. Such improved approaches to leverage MD simulations with improved force fields and experimental knowledge-based ensemble generation have shown tremendous potential for describing molecular recognition mechanisms of IDPs with other protein partners; scenarios such as folding-upon-binding[62,86] or self-dimerisation. More importantly, a set of recent initiatives have focussed on molecular characterisation of the interaction of small-molecules with IDPs via effective combination of atomistic simulation and biophysical

experiments.[19,32]

In this regard, the present investigation takes a step ahead via a quantitative and statistical dissection of the αS ensemble in presence of a small-molecule. The implementation of Markov state model in the present work clearly brings out the feasibility of metastable states that would have a finite 'stationary' or equilibrium population in presence of small-molecule or in neat water. The uniqueness of these states is supported from the projection of their inter-residue C_α contact map of the macrostates using DCVAE model. Our systematic investigation of the impact of fasudil on αS monomer ensemble reveals the altering effects of small-molecule interaction on αS ensemble by tuning the entropy of these states. Insights from studies targeting structural ensemble modulation by small-molecule can be effectively harnessed in designing drug-design strategies for aggregation diseases.

Computational Methods

Monomer Simulation Protocol

A 1500 μs long all-atom MD simulation trajectory of αS monomer in aqueous fasudil solution was simulated by D. E. Shaw Research with the Anton supercomputer that is specially purposed for running long-time-scale simulations.[32] The protein was solvated in a cubic water-box with 108 Å sides and containing 50 mM concentration of Na or Cl ions. Protein, water molecules and ions present in the system were parameterised with the a99SB-*disp* force field.[33] The small-molecule, fasudil, was parameterised using the generalised Amber force-field (GAFF). Production run of the system was performed in the NPT ensemble. Further details of the MD simulation can be obtained by referring to the original work by Robustelli et al.[32] The extensive simulation trajectories were generously made available by D. E. Shaw Research upon request.

To generate αS ensemble in water, we performed multiple simulations of αS monomer. An IDP is not represented by a fixed structure. Therefore, it would be most appropriate

to run multiple simulations starting from different initial structures and simulate the local environment around those structures; thus generating an ensemble effectively sampling the phase space. Accordingly, for initiating the apo simulations, instead of biasing the initial structure (using the starting structure used for simulations with fasudil), we randomly chose 23 different conformations from previously reported[33] multi-μs αS trajectory in water. We have also determined the R_g values corresponding to the starting structures used to start the apo simulations. It is evident that the 23 starting conformations chosen represent the entire span of the R_g space that is sampled in the fasudil ensemble (see Figure S1). In addition, we have also calculated the RMSD of the starting structures with respect to the starting structure of the fasudil simulation (see Figure S25), secondary structure content (see Figure S26-S28) and solvent accessible surface area (SASA) (see Figure S29-S30) to validate our choice of starting structures.

The simulation details are described below. Each protein conformation is solvated in a cubic box of explicit water molecules and Na or Cl ions were added to a concentration of 50 mM, to maintain the same conditions as used in the simulations with fasudil. The all-atom a99SB-*disp* force field and water model[33] were used to simulate the protein and solvent. The unbiased MD simulations were performed using the Gromacs simulation package.[87] A timestep of 2 fs and the leap-frog integrator was used. The simulation was performed in the isothermal-isobaric (NPT) ensemble at an average temperature of 300 K, maintained using the v-rescale thermostat[88] with a relaxation time of 0.1 ps and a pressure of 1 bar using the Parrinello-Rahman barostat[89] with a time constant of 2 ps for coupling. The Verlet cutoff scheme[90] with 1.0 nm cutoff was applied for Lennard-Jones interactions and short-range electrostatic interactions. Long-range electrostatic interactions were computed using the Particle-Mesh Ewald (PME) summation method[91] and covalent bonds involving hydrogen atoms were constrained using the LINCS algorithm.[92] All the systems were minimized using the steepest descent algorithm, followed by equilibration in the isothermal-isochoric (NVT) and subsequently in the NPT ensemble with position restraints on the heavy atoms

of the protein. Twenty-three simulations, each starting from different conformations, were performed. These simulation timescales are variable, ranging from 1 to 4 μs, amounting to a cumulative length of $\sim$62 μs.

We have compared the ensemble properties of the αS in the presence and absence of fasudil. As seen in figure S1 the distributions of R_g for both the simulations have a significant overlap, thereby indicating that in terms of extent of compaction, the conformational sampling of the protein is comparable in the presence and absence of fasudil. We have also compared other structural features such as end-to-end distance (R_{ee}) (see Figure S2), SASA (see Figure S3) and secondary structure content (see Figure S4-S6). We have found that these features are highly comparable for both the ensembles, which clearly indicates that the αS simulations we performed from different starting structures have effectively sampled the phase space as the single long simulation of the αS-Fasudil system.

Dimension reduction using β-Variational Autoencoder

The probability distribution $p(x)$ of a multi-dimensional variable x can be determined using various approaches,[93,94] one of them being the latent variable model. In this model, the probability distribution is modeled as a joint distribution $p(x, z)$ of the observable variable x and the hidden latent variable z which is given as $p(x, z) = p(x|z)p(z)$, where $p(z)$ is the prior distribution over the latent variables z and $p(x|z)$ is the likelihood of generating a true data, given z. The $p(x)$ is determined by marginalization over all the latent variables giving us,

$$p(x) = \int p(x|z)p(z)dz \tag{2}$$

However, this integral is often intractable thereby constraining the calculation of the true posterior distribution $p(z|x)$. This can be solved using, variational inference, which aims to estimate the true posterior using an approximate distribution $q(z|x)$, using the encoder part of VAE which parameterizes the distribution using a mean vector and a variance vector.

The goal is to determine $q(z|x)$ which is closest to the true posterior by minimizing the Kullback-Leibler (KL) divergence between them which is given as

$$D_{KL}(q(z|x)\|p(z|x)) = \int q(z|x) \log \left(\frac{q(z|x)}{p(z|x)} \right) dz \tag{3}$$

Upon expanding and rearranging, we have

$$\log p(x) - D_{KL}(q(z|x)\|p(z|x)) = \mathbb{E}_{z\sim q(z|x)} \log p(x|z) - D_{KL}(q(z|x)\|p(z)) \tag{4}$$

where $p(x)$ is the evidence. The VAE is trained so that we maximize the log-likelihood of the evidence and minimize the KL divergence between the estimated and true posterior. The loss function of VAE is given by the negation of equation 2 and is given as,

$$\text{Loss}_{VAE} = -\mathbb{E}_{z\sim q(z|x)} \log p(x|z) + D_{KL}(q(z|x)\|p(z)) \tag{5}$$

The training objective of VAE being minimizing this loss function to generate a realistic data. The VAE model can be further improved by keeping the distance between the approximated and the true posterior under a small distance, δ. This is achieved by maximizing the equation

$$\mathcal{F}(\beta; x, z) = \mathbb{E}_{z\sim q(z|x)} \log p(x|z) - \beta(D_{KL}(q(z|x)\|p(z)) - \delta) \tag{6}$$

which modifies the loss function to

$$\text{Loss}_{\beta-VAE} = -\mathbb{E}_{z\sim q(z|x)} \log p(x|z) + \beta D_{KL}(q(z|x)\|p(z)) \tag{7}$$

where β is the Lagrangian multiplier (under the Karush-Kuhn-Tucker conditions). This enhances the probability of generating an actual data and is known as β-VAE.[95] The expectation term in the loss function of β-VAE requires a point z to be sampled from the posterior

distribution. This is performed using the reparametrization trick

$$z = \mu + \sigma \odot \epsilon \tag{8}$$

where ϵ is sampled from a Standard Normal distribution. This makes the VAE model trainable using the backpropagation algorithm.

In our work we have used C_α-C_α pairwise distance (excluding i, $i+1$ and $i+2$ distances) sampled at every 9 ns of the entire 1.5 millisecond of the αS protein in presence of fasudil trajectory as our input. We have built our model using Tensorflow backend[96] with Keras library. Training was performed with 80% of the data while 20% was used for testing. We have employed a symmetric VAE where the input layer consists of 9453 nodes followed by four hidden layers having 4096, 512, 128 and 16 nodes respectively, and 2 nodes in the latent layer. Hyperbolic tangent (*tanh*) is used as our activation function with weights being initialized using the *glorot_uniform*[97] method in all the layers except for the output layer where we have used *sigmoid* activation. Adaptive Moment Estimation (Adam[98]) was used as the optimizer with a learning rate of 1×10^{-4}. The model was trained for 300 epochs with a batch size of 64. The relevant Python implementations are provided in Github (`https://github.com/JMLab-tifrh/Protein_Ligand_Variational_Autoencoder`).

Building the Denoising Convolutional Variational AutoEncoder

The Denoising Convolutional Variational Autoencoder (DCVAE) model employs five convolutional layers which incorporate filters of sizes 16, 32, 64, 96, and 128, with corresponding kernels of size 3x3. Strides are set at 1, 2, 2, 2, 2, and 1 for efficient feature extraction. This is followed by six densely connected layers with neuron counts of 4096, 2048, 1024, 256, 64 and 16 in the successive dense layers. The latent layer is comprised of 2 neurons. The decoder mirrors the encoder architecture, to reconstruct back the input contact maps. Hyperbolic tangent activation is used in all the layers, excluding the output layer where *sigmoid*

activation is used. Adam optimizer with a learning rate of 5×10^{-4} is used for training the DCVAE. The model was trained for 300 epochs with a batch size of 32 with $\beta = 1\times10^{-12}$. Training was performed with 90% of the data while 10% was used for testing.

At first, we added a noise to 200 replicas of contact map for each of the states of αS and αS-fasudil. The noise was generated stochastically from a standard normal distribution which was multiplied by a noise factor of 0.035 and added to each of the replicas, followed by scaling the dataset so that the values range between 0 and 1 using the formula

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{9}$$

where x_{min} and x_{max} are the minimum and maximum values of the input contact map. In order to assess the denoising performance, we have calculated the structural similarity index measure (SSIM) values.[58] The SSIM value can range between -1 to +1, thereby indicating the structural dissimilarity or similarity between the original (x) and denoised (y) contact maps. Mathematically, this is expressed as

$$\text{SSIM}(x, y) = [l(x, y)]^{\alpha} \cdot [c(x, y)]^{\beta} \cdot [s(x, y)]^{\gamma} \tag{10}$$

where the functions $l(x, y)$, $c(x, y)$,and $s(x, y)$ compare the luminance, contrast and structure between the contact maps and is given as

$$\begin{aligned} l(x, y) &= \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \\ c(x, y) &= \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \\ s(x, y) &= \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \end{aligned} \tag{11}$$

The variables μ_i and σ_i represent the pixel sample mean and variance of i, where $i = \{x, y\}$. σ_{xy} is the covariance of the x and y contact maps. The exponents α, β and γ controls the weighted contribution to the SSIM value and are set to 1. The constants $C_1 = (k_1L)^2$,

$C_2 = (k_2 L)^2$ are small positive values introduced to prevent instability and division by zero. They are used for numerical stability and to avoid problems when the means and variances of the contact maps are close to zero. We have set $k_1 = 0.01$, $k_2 = 0.03$ and $C_3 = C_2/2$ as per convention.[99]

In addition, we have also calculated the peak Signal-to-Noise Ratio to evaluate our denoising performance and is calculated as

$$\mathrm{PSNR} = 10 \cdot \log_{10} \left(\frac{R^2}{\mathrm{MSE}} \right) \tag{12}$$

where, R is the maximum pixel value (255 in our case) and MSE is the mean-square error between the original and denoised image and is given as

$$\mathrm{MSE} = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} [\mathcal{O}_{ij} - \mathcal{D}_{ij}] \tag{13}$$

where M is the height (number of rows) of the image, N is the width (number of columns) of the image and $\mathcal{O}_{ij}$ and $\mathcal{D}_{ij}$ is the pixel intensity at position (i, j) for the original and denoised image respectively. The relevant Python implementations are provided in Github (`https://github.com/JMLab-tifrh/Protein_Ligand_Variational_Autoencoder`).

Building a Markov State Model of monomer ensemble

The 2D latent space reconstructed using β-VAE described in the previous subsection was then used to build a Markov State Model to elicit the metastable states underlying the ensembles. We employed PyEMMA,[57] a Markov State Model (MSM) building and analysis package. The 2D data is subjected to k-means clustering. We used the VAMP-2 score to determine the number of microstates. By varying the number of microstates from 10 to 220, we found that the VAMP-2 score saturated at higher numbers for both αS and αS-Fasudil simulations. We find that the VAMP-2 score has saturated at 180 microstates (see Figure

S31-S32). A transition matrix was then built by counting the number of transitions among the microstates at a specific lag time. To choose the lag time, the implied timescale (ITS) or relaxation timescales of the systems were calculated over a range of lag times and plotted as a function of lag time. The timescale at which the ITS plot levels off is chosen as the lag time to build the final MSM. The ITS plots corresponding to the neat water and aqueous fasudil systems are presented in Figure 3. Accordingly, MSM lag times of 36 and 32 ns were selected for neat water fasudil systems, respectively. At these lag times, by identifying the gaps between the ITS, a three-state model for neat water and a six-state model for aqueous fasudil system was built. Lastly, the transition path theory[100–102] was used to ascertain the transition paths, fluxes and timescales in these models. Bootstrapping was performed to estimate the errors associated with the state populations and transition timescales. In ten iterations, the model was rebuilt after eliminating a randomly selected trajectory followed by calculation of the state populations and transition timescales. The mean value and standard deviations of the values collected by bootstrapping are reported. A Chapman-Kolmogorov test was performed for the MSMs of the neat water and small molecule ensembles, that tests the validity of the model prediction at longer time scales (see Figure S9).

Entropy of water

In order to calculate the entropy of water, simulations were performed using GROMACS (version 2022.3). A single snapshot was chosen from each macrostate of α-synuclein in its apo state and in presence of fasudil, which were then solvated in a cubic box, placed within 1 nm from the box edge. The system was neutralized using NaCl and the salt concentration was kept at 50 mM. The system was energy minimized using the steepest descent algorithm, followed by equilibration in the canonical ensemble for 2 ps. The average temperature was maintained at 300 K using velocity rescale thermostat with a time constant of 0.1 ps. Following this, NPT equilibration was performed for 2 ps to maintain the pressure at 1.0 bar using the Berendsen barostat. Finally, three independent production runs were performed

from the equilibrated structures in the NPT ensemble using the velocity-verlet integrator with a timestep of 1 fs for a timescale of 20 ps with frames saved at every 4 fs. In addition, we have performed three simulations of neat water using the parameters as described for the macrostates.

The entropy of water was estimated using the DoSPT program[103,104] that computes the thermodynamic properties from MD simulation in the framework of the 2-phase-thermodynamics (2PT) method.[64,65] This is achieved from calculating the density of state (DoS) function , $I(v)$ which is the Fourier transform of the velocity auto correlation function,

$$I(v) = \frac{1}{k_B T} \sum_{l=1}^{N} \sum_{k=1}^{3} \lim_{\tau \to \infty} \frac{m_l}{\tau} \left| \int_{-\tau}^{\tau} v_l^k(t) \exp(-2i\pi vt)\ \mathrm{d}t \right|^2 \tag{14}$$

where m_l is the mass of atom l, v_l^k is the velocity of atom l along the k direction and N is the total number of atoms in the system. DoS can be represented as a sum of translational (trn), rotational (rot) and vibrational (vib) motions.[65]

$$I(v) = I_{trn}(v) + I_{rot}(v) + I_{vib} \tag{15}$$

For water, the 2PT model overcomes the anharmonic nature of the low frequency modes (diffusive motions and libration) by decomposing the DoS into solid-like (I^s) and gas-like (I^g) contribution for the translational and rotational components. The solid-like DoS is treated using the harmonic oscillator model, whereas the gas-like DoS is described using the Enskog hard sphere (HS) theory for translation and the rigid rotor (RR) model for rotation. The entropy is calculated by integrating the DoS, weighted by the corresponding weighting functions W,

$$S_k = k_B \left[\int_0^{\infty} I_k^s(v) W_k^s(v) \mathrm{d}v + \int_0^{\infty} I_k^g(v) W_k^g \mathrm{d}v \right] \tag{16}$$

where $k =$ trn, rot, or vib and the weighting functions are given as

$$
\begin{aligned}
W^s(v) &= \frac{\beta h v}{\exp(\beta h v) - 1} - \ln[1 - \exp(-\beta h v)] \\
W^g_{trn} &= \frac{S^{HS}}{3k_B} \\
W^g_{rot} &= \frac{S^{RR}}{3k_B}
\end{aligned}
\tag{17}
$$

where $\beta = (k_B T)^{-1}$ and h is the planck constant.[105]

Protein conformational entropy

We calculated the protein conformational entropy using the program PDB2ENTROPY.[67] This program is based on the nearest-neighbor approach from probability distributions of torsion angles at a given temperature relative to uniform distributions. The entropy is computed using the Maximum Information Spanning Tree (MIST) approach.[68,69]

Supplemental Information

All supplemental figures described in the article (PDF)

Code Availability

The implementations of Machine learning protocols are publicly available in our group GitHub webpage (`https://github.com/JMLab-tifrh/Protein_Ligand_Variational_Autoencoder`)

Acknowledgements

We sincerely thank D. E. Shaw research for providing us with access to long simulation trajectories of monomeric alpha-synuclein in presence of fasudil which seeded the project. We acknowledge support of the Department of Atomic Energy, Government of India, under

Project Identification No. RTI 4007. JM acknowledges Core Research grants provided by the Department of Science and Technology (DST) of India (CRG/2023/001426).

References

(1) Maroteaux, L.; Campanelli, J. T.; Scheller, R. H. Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell. Biol.* **2015**, *16*, 18–29.

(2) Babu, M. M.; van der Lee, R.; de Groot, N. S.; Gsponer, J. Intrinsically disordered proteins: regulation and disease. *Current Opinion in Structural Biology* **2011**, *21*, 432–440.

(3) Csizmok, V.; Follis, A. V.; Kriwacki, R. W.; Forman-Kay, J. D. Dynamic Protein Interaction Networks and New Structural Paradigms in Signaling. *Chemical Reviews* **2016**, *116*, 6424–6462, PMID: 26922996.

(4) Wang, Y.; Fisher, J. C.; Mathew, R.; Ou, L.; Otieno, S.; Sublet, J.; Xiao, L.; Chen, J.; Roussel, M. F.; Kriwacki, R. W. Intrinsic disorder mediates the diverse regulatory functions of the Cdk inhibitor p21. *Nature chemical biology* **2011**, *7*, 214–221.

(5) Dyson, H. J. Making sense of intrinsically disordered proteins. *Biophysical journal* **2016**, *110*, 1013.

(6) Alberti, S.; Gladfelter, A.; Mittag, T. Considerations and challenges in studying liquid-liquid phase separation and biomolecular condensates. *Cell* **2019**, *176*, 419–434.

(7) Chong, P. A.; Forman-Kay, J. D. Liquid–liquid phase separation in cellular signaling systems. *Current opinion in structural biology* **2016**, *41*, 180–186.

(8) Shin, Y.; Brangwynne, C. P. Liquid phase condensation in cell physiology and disease. *Science* **2017**, *357*, eaaf4382.

(9) O'Flynn, B. G.; Mittag, T. The role of liquid–liquid phase separation in regulating enzyme activity. *Current opinion in cell biology* **2021**, *69*, 70–79.

(10) Uversky, V. N.; Oldfield, C. J.; Dunker, A. K. Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annual review of biophysics* **2008**, *37*, 215–246.

(11) Bronowska, A. *Thermodynamics of Ligand-Protein Interactions: Implications for Molecular Design, Thermodynamics - Interaction Studies - Solids, Liquids and Gases*; 2011.

(12) Klebe, G. Applying thermodynamic profiling in lead finding and optimization. *Nat. Rev. Drug Discov.* **2015**, *14*, 95–110.

(13) Metallo, S. J. Intrinsically disordered proteins are potential drug targets. *Current Opinion in Chemical Biology* **2010**, *14*, 481–488, Next Generation Therapeutics.

(14) Structural rationale for the coupled binding and unfolding of the c-Myc oncoprotein by small molecules. *Chemistry and Biology* **2008**, *15*, 1149–1155.

(15) Ono, K.; Li, L.; Takamura, Y.; Yoshiike, Y.; Zhu, L.; Han, H.; Mao, X.; Ikeda, T.; Takasaki, T.; Nishijo, H. Phenolic compounds prevent amyloid -protein oligomerization and synaptic dysfunction by site-specific binding. *J. Biol. Chem.* **2012**, *287*, 14631–14643.

(16) Zhu, M.; De Simone, A.; Schenk, D.; Toth, G.; Dobson, C. M.; Vendruscolo, M. Identification of small-molecule binding pockets in the soluble monomeric form of the Abeta42 peptide. *The Journal of Chemical Physics* **2013**, *139*, 035101.

(17) Attanasio, F.; Convertino, M.; Magno, A.; Caflisch, A.; Corazza, A.; Haridas, H.; Esposito, G.; Cataldo, S.; Pignataro, B.; Milardi, D.; Rizzarelli, E. Carnosine Inhibits

Abeta-42 Aggregation by Perturbing the H-Bond Network in and around the Central Hydrophobic Cluster. *ChemBioChem* **2013**, *14*, 583–592.

(18) Ehrnhoefer, D. E.; Bieschke, J.; Boeddrich, A.; Herbst, M.; Masino, L.; Lurz, R.; Engemann, S.; Pastore, A.; Wanker, E. E. EGCG redirects amyloidogenic polypeptides into unstructured, off-pathway oligomers. *Nature structural and molecular biology* **2008**, *15*, 558–566.

(19) Heller, G. T. et al. Small-molecule sequestration of amyloid as a drug discovery strategy for Alzheimers disease. *Science Advances* **2020**, *6*, eabb5924.

(20) Akoury, E.; Gajda, M.; Pickhardt, M.; Biernat, J.; Soraya, P.; Griesinger, C.; Mandelkow, E.; Zweckstetter, M. Inhibition of Tau Filament Formation by Conformational Modulation. *Journal of the American Chemical Society* **2013**, *135*, 2853–2862.

(21) others,, et al. Targeting the disordered C terminus of PTP1B with an allosteric inhibitor. *Nature chemical biology* **2014**, *10*, 558–566.

(22) Kurzbach, D.; Schwarz, T. C.; Platzer, G.; Hafler, S.; Hinderberger, D.; Konrat, R. Compensatory Adaptations of Structural Dynamics in an Intrinsically Disordered Protein Complex. *Angewandte Chemie International Edition* **2014**, *53*, 3840–3843.

(23) Biesaga, M.; Frigol-Vivas, M.; Salvatella, X. Intrinsically disordered proteins and biomolecular condensates as drug targets. *Current Opinion in Chemical Biology* **2021**, *62*, 90–100, Next Generation Therapeutics.

(24) Fields, C. R.; Bengoa-Vergniory, N.; Wade-Martins, R. Targeting Alpha-Synuclein as a Therapy for Parkinson's Disease. *Frontiers in Molecular Neuroscience* **2019**, *12*.

(25) Tah, G. et al. Targeting the Intrinsically Disordered Structural Ensemble of alpha-Synuclein by Small Molecules as a Potential Therapeutic Strategy for Parkinson's Disease. *PLOS ONE* **2014**, *9*, 1–11.

(26) Tatenhorst, L.; Eckermann, K.; Dambeck, V.; Fonseca-Ornelas, L.; Walle, H.; Lopes da Fonseca, T.; Koch, J. C.; Becker, S.; Tönges, L.; Bähr, M. Fasudil attenuates aggregation of α-synuclein in models of Parkinson's disease. *Acta neuropathologica communications* **2016**, *4*, 1–17.

(27) Cao, K.; Zhu, Y.; Hou, Z.; Liu, M.; Yang, Y.; Hu, H.; Dai, Y.; Wang, Y.; Yuan, S.; Huang, G. α-Synuclein as a Target for Metallo-Anti-Neurodegenerative Agents. *Angewandte Chemie International Edition* **2022**,

(28) Winner, B.; Jappelli, R.; Maji, S. K.; Desplats, P. A.; Boyer, L.; Aigner, S.; Hetzer, C.; Loher, T.; Vilar, M.; Campioni, S. In vivo demonstration that α-synuclein oligomers are toxic. *Proceedings of the National Academy of Sciences* **2011**, *108*, 4194–4199.

(29) Emin, D.; Zhang, Y. P.; Lobanova, E.; Miller, A.; Li, X.; Xia, Z.; Dakin, H.; Sideris, D. I.; Lam, J. Y.; Ranasinghe, R. T. Small soluble α-synuclein aggregates are the toxic species in Parkinson's disease. *Nature communications* **2022**, *13*, 1–15.

(30) Stephens, A. D.; Kölbel, J.; Moons, R.; Chung, C. W.; Ruggiero, M. T.; Mahmoudi, N.; Shmool, T. A.; McCoy, T. M.; Nietlispach, D.; Routh, A. F. Decreased Water Mobility Contributes To Increased α-Synuclein Aggregation. *Angewandte Chemie International Edition* **2022**,

(31) Ubbiali, D.; Fratini, M.; Piersimoni, L.; Ihling, C. H.; Kipping, M.; Heilmann, I.; Iacobucci, C.; Sinz, A. Direct Observation of Elongated Conformational States in α-Synuclein upon Liquid-Liquid Phase Separation. *Angewandte Chemie* **2022**, *134*, e202205726.

(32) Robustelli, P.; Ibanez-de Opakua, A.; Campbell-Bezat, C.; Giordanetto, F.; Becker, S.; Zweckstetter, M.; Pan, A. C.; Shaw, D. E. Molecular Basis of Small-Molecule Binding to alpha-Synuclein. *Journal of the American Chemical Society* **2022**, *144*, 2501–2510.

(33) Robustelli, P.; Piana, S.; Shaw, D. E. Developing a molecular dynamics force field for both folded and disordered protein states. *Proceedings of the National Academy of Sciences* **2018**, *115*, E4758–E4766.

(34) Bonomi, M.; Camilloni, C.; Cavalli, A.; Vendruscolo, M. Metainference: A Bayesian inference method for heterogeneous systems. *Science advances* **2016**, *2*, e1501177.

(35) Bottaro, S.; Lindorff-Larsen, K. Biophysical experiments and biomolecular simulations: A perfect match? *Science* **2018**, *361*, 355–360.

(36) Gomes, G.-N. W.; Krzeminski, M.; Namini, A.; Martin, E. W.; Mittag, T.; Head-Gordon, T.; Forman-Kay, J. D.; Gradinaru, C. C. Conformational ensembles of an intrinsically disordered protein consistent with NMR, SAXS, and single-molecule FRET. *Journal of the American Chemical Society* **2020**, *142*, 15697–15710.

(37) Stelzl, L. S.; Pietrek, L. M.; Holla, A.; Oroz, J.; Sikora, M.; KoÃàfinger, J.; Schuler, B.; Zweckstetter, M.; Hummer, G. Global structure of the intrinsically disordered protein tau emerges from its local structure. *Jacs Au* **2022**, *2*, 673–686.

(38) Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schutte, C.; Noe, F. Markov models of molecular kinetics: Generation and validation. *J. Chem. Phys.* **2011**, *134*, 174105.

(39) Bowman, G. R.; Beauchamp, K. A.; Boxer, G.; Pande, V. S. Progress and challenges in the automated construction of Markov state models for full protein systems. *J. Chem. Phys.* **2009**, *131*, 124101.

(40) Husic, B. E.; Pande, V. S. Markov State Models: From an Art to a Science. *Journal of the American Chemical Society* **2018**, *140*, 2386–2396, PMID: 29323881.

(41) Adhikari, S.; Mondal, J. Machine Learning Subtle Conformational Change due to

Phosphorylation in Intrinsically Disordered Proteins. *The Journal of Physical Chemistry B* **2023**,

(42) Kingma, D. P.; Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* **2013**,

(43) Rezende, D. J.; Mohamed, S.; Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. International conference on machine learning. 2014; pp 1278–1286.

(44) Hyvärinen, A. Independent component analysis: recent advances. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **2013**, *371*, 20110534.

(45) Klema, V.; Laub, A. The singular value decomposition: Its computation and some applications. *IEEE Transactions on automatic control* **1980**, *25*, 164–176.

(46) Fisher, R. A. The use of multiple measurements in taxonomic problems. *Annals of eugenics* **1936**, *7*, 179–188.

(47) Schölkopf, B.; Smola, A.; Müller, K.-R. Kernel principal component analysis. Artificial Neural Networks-ICANN 97: 7th International Conference Lausanne, Switzerland, October 8–10, 1997 Proceeedings. 2005; pp 583–588.

(48) Kruskal, J. B.; Wish, M. *Multidimensional scaling*; Sage, 1978; Vol. 11.

(49) Tenenbaum, J. B.; Silva, V. d.; Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *science* **2000**, *290*, 2319–2323.

(50) Faloutsos, C.; Lin, K.-I. FastMap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. Proceedings of the 1995 ACM SIGMOD international conference on Management of data. 1995; pp 163–174.

(51) Roweis, S. T.; Saul, L. K. Nonlinear dimensionality reduction by locally linear embedding. *science* **2000**, *290*, 2323–2326.

(52) Sumithra, V.; Surendran, S. A review of various linear and non linear dimensionality reduction techniques. *Int. J. Comput. Sci. Inf. Technol* **2015**, *6*, 2354–2360.

(53) Das, P.; Moll, M.; Stamati, H.; Kavraki, L. E.; Clementi, C. Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. *Proceedings of the National Academy of Sciences* **2006**, *103*, 9885–9890.

(54) Schwantes, C. R.; Pande, V. S. Modeling molecular kinetics with tICA and the kernel trick. *Journal of chemical theory and computation* **2015**, *11*, 600–608.

(55) Wehmeyer, C.; Noé, F. Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics. *The Journal of chemical physics* **2018**, *148*.

(56) Hinton, G. E.; Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *science* **2006**, *313*, 504–507.

(57) Scherer, M. K.; Trendelkamp-Schroer, B.; Paul, F.; Pérez-Hernández, G.; Hoffmann, M.; Plattner, N.; Wehmeyer, C.; Prinz, J.-H.; Noé, F. PyEMMA 2: A software package for estimation, validation, and analysis of Markov models. *Journal of chemical theory and computation* **2015**, *11*, 5525–5542.

(58) Hore, A.; Ziou, D. Image quality metrics: PSNR vs. SSIM. 2010 20th international conference on pattern recognition. 2010; pp 2366–2369.

(59) Heller, G. T.; Bonomi, M.; Vendruscolo, M. Structural Ensemble Modulation upon Small-Molecule Binding to Disordered Proteins. *Journal of Molecular Biology* **2018**, *430*, 2288–2292, Intrinsically Disordered Proteins.

(60) Heller, G. T.; Sormani, P.; Vendruscolo, M. Targeting disordered proteins with small molecules using entropy. *Trends Biochem. Sci.* **2015**, *40*, 491–496.

(61) Flock, T.; Weatheritt, R. J.; Latysheva, N. S.; Babu, M. M. Controlling entropy to tune the functions of intrinsically disordered regions. *Current Opinion in Structural Biology* **2014**, *26*, 62–72, New constructs and expression of proteins / Sequences and topology.

(62) Dyson, H.; Wright, P. E. Coupling of folding and binding for unstructured proteins. *Current Opinion in Structural Biology* **2002**, *12*, 54–60.

(63) Löhr, T.; Kohlhoff, K.; Heller, G. T.; Camilloni, C.; Vendruscolo, M. A Small Molecule Stabilizes the Disordered Native State of the Alzheimer's Aβ Peptide. *ACS Chemical Neuroscience* **2022**, *13*, 1738–1745.

(64) Lin, S.-T.; Blanco, M.; Goddard III, W. A. The two-phase model for calculating thermodynamic properties of liquids from molecular dynamics: Validation for the phase diagram of Lennard-Jones fluids. *The Journal of chemical physics* **2003**, *119*, 11792–11805.

(65) Lin, S.-T.; Maiti, P. K.; Goddard III, W. A. Two-phase thermodynamic model for efficient and accurate absolute entropy of water from molecular dynamics simulations. *The Journal of Physical Chemistry B* **2010**, *114*, 8191–8198.

(66) Mukherjee, S.; Schäfer, L. V. Thermodynamic forces from protein and water govern condensate formation of an intrinsically disordered protein domain. *Nature Communications* **2023**, *14*, 5892.

(67) Fogolari, F.; Maloku, O.; Dongmo Foumthuim, C. J.; Corazza, A.; Esposito, G. PDB2ENTROPY and PDB2TRENT: Conformational and Translational–Rotational Entropy from Molecular Ensembles. *Journal of Chemical Information and Modeling* **2018**, *58*, 1319–1324.

(68) King, B. M.; Tidor, B. MIST: Maximum Information Spanning Trees for dimension reduction of biological data sets. *Bioinformatics* **2009**, *25*, 1165–1172.

(69) King, B. M.; Silver, N. W.; Tidor, B. Efficient calculation of molecular configurational entropies using an information theoretic approximation. *The Journal of Physical Chemistry B* **2012**, *116*, 2891–2904.

(70) Dedmon, M. M.; Lindorff-Larsen, K.; Christodoulou, J.; Vendruscolo, M.; Dobson, C. M. Mapping long-range interactions in α-synuclein using spin-label NMR and ensemble molecular dynamics simulations. *Journal of the American Chemical Society* **2005**, *127*, 476–477.

(71) Allison, J. R.; Varnai, P.; Dobson, C. M.; Vendruscolo, M. Determination of the free energy landscape of α-synuclein using spin label nuclear magnetic resonance measurements. *Journal of the American Chemical Society* **2009**, *131*, 18314–18326.

(72) Ullman, O.; Fisher, C. K.; Stultz, C. M. Explaining the structural plasticity of α-synuclein. *Journal of the american chemical society* **2011**, *133*, 19536–19546.

(73) Esteban-Martín, S.; Silvestre-Ryan, J.; Bertoncini, C. W.; Salvatella, X. Identification of fibril-like tertiary contacts in soluble monomeric α-synuclein. *Biophysical journal* **2013**, *105*, 1192–1198.

(74) Best, R. B.; Zheng, W.; Mittal, J. Balanced protein–water interactions improve properties of disordered proteins and non-specific protein association. *Journal of chemical theory and computation* **2014**, *10*, 5113–5124.

(75) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; De Groot, B. L.; Grubmüller, H.; MacKerell, A. D. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nature methods* **2017**, *14*, 71–73.

(76) Bertoncini, C. W.; Jung, Y.-S.; Fernandez, C. O.; Hoyer, W.; Griesinger, C.; Jovin, T. M.; Zweckstetter, M. Release of long-range tertiary interactions potentiates aggregation of natively unstructured α-synuclein. *Proceedings of the National Academy of Sciences* **2005**, *102*, 1430–1435.

(77) Bernadó, P.; Blanchard, L.; Timmins, P.; Marion, D.; Ruigrok, R. W.; Blackledge, M. A structural model for unfolded proteins from residual dipolar couplings and small-angle x-ray scattering. *Proceedings of the National Academy of Sciences* **2005**, *102*, 17002–17007.

(78) Marsh, J. A.; Baker, J. M.; Tollinger, M.; Forman-Kay, J. D. Calculation of residual dipolar couplings from disordered state ensembles using local alignment. *Journal of the American Chemical Society* **2008**, *130*, 7804–7805.

(79) Bernadó, P.; Mylonas, E.; Petoukhov, M. V.; Blackledge, M.; Svergun, D. I. Structural characterization of flexible proteins using small-angle X-ray scattering. *Journal of the American Chemical Society* **2007**, *129*, 5656–5664.

(80) Ahmed, M. C.; Skaanning, L. K.; Jussupow, A.; Newcombe, E. A.; Kragelund, B. B.; Camilloni, C.; Langkilde, A. E.; Lindorff-Larsen, K. Refinement of α-synuclein ensembles against SAXS data: Comparison of force fields and methods. *Frontiers in molecular biosciences* **2021**, *8*, 654333.

(81) Brookes, D. H.; Head-Gordon, T. Experimental inferential structure determination of ensembles for intrinsically disordered proteins. *Journal of the American Chemical Society* **2016**, *138*, 4530–4538.

(82) Lincoff, J.; Krzeminski, M.; Haghighatlari, M.; Teixeira, J.; Gomes, G.-N. W.; Gradinaru, C. C.; Forman-Kay, J. D.; Head-Gordon, T. Extended Experimental Inferential Structure Determination Method for Evaluating the Structural Ensembles of Disordered Protein States. *arXiv preprint arXiv:1912.12582* **2019**,

(83) Roux, B.; Weare, J. On the statistical equivalence of restrained-ensemble simulations with the maximum entropy method. *The Journal of chemical physics* **2013**, *138*, 02B616.

(84) Boomsma, W.; Ferkinghoff-Borg, J.; Lindorff-Larsen, K. Combining experiments and simulations using the maximum entropy principle. *PLoS computational biology* **2014**, *10*, e1003406.

(85) Cavalli, A.; Camilloni, C.; Vendruscolo, M. Molecular dynamics simulations with replica-averaged structural restraints generate structural ensembles according to the maximum entropy principle. *The Journal of chemical physics* **2013**, *138*, 03B603.

(86) Robustelli, P.; Piana, S.; Shaw, D. E. Mechanism of Coupled Folding-upon-Binding of an Intrinsically Disordered Protein. *Journal of the American Chemical Society* **2020**, *142*, 11092–11101, PMID: 32323533.

(87) Lindahl, E.; Abraham, M.; Hess, B.; Spoel, v. d. D. GROMACS 2020.1 Source code. 2020.

(88) Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *The Journal of chemical physics* **2007**, *126*, 014101.

(89) Parrinello, M.; Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied physics* **1981**, *52*, 7182–7190.

(90) Páll, S.; Hess, B. A flexible algorithm for calculating pair interactions on SIMD architectures. *Computer Physics Communications* **2013**, *184*, 2641–2650.

(91) Darden, T.; York, D.; Pedersen, L. The effect of long-range electrostatic interactions in simulations of macromolecular crystals–a comparison of the ewald and truncated list methods. *J. Chem. Phys* **1993**, *99*, 10089.

(92) Hess, B.; Bekker, H.; Berendsen, H. J.; Fraaije, J. G. LINCS: a linear constraint solver for molecular simulations. *Journal of computational chemistry* **1997**, *18*, 1463–1472.

(93) Bengio, S.; Bengio, Y. Taking on the curse of dimensionality in joint distributions using neural networks. *IEEE Transactions on Neural Networks* **2000**, *11*, 550–557.

(94) Larochelle, H.; Murray, I. The neural autoregressive distribution estimator. Proceedings of the fourteenth international conference on artificial intelligence and statistics. 2011; pp 29–37.

(95) Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. International conference on learning representations. 2016.

(96) others,, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467* **2016**,

(97) Glorot, X.; Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. Proceedings of the thirteenth international conference on artificial intelligence and statistics. 2010; pp 249–256.

(98) Kingma, D. P.; Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* **2014**,

(99) Wang, Z.; Bovik, A. C.; Sheikh, H. R.; Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* **2004**, *13*, 600–612.

(100) Weinan, E.; Vanden-Eijnden, E. Towards a theory of transition paths. *Journal of statistical physics* **2006**, *123*, 503–523.

(101) Metzner, P.; Schütte, C.; Vanden-Eijnden, E. Transition path theory for Markov jump processes. *Multiscale Modeling & Simulation* **2009**, *7*, 1192–1219.

(102) Noé, F.; Schütte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R. Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proceedings of the National Academy of Sciences* **2009**, *106*, 19011–19016.

(103) Caro, M. A.; Laurila, T.; Lopez-Acevedo, O. Accurate schemes for calculation of thermodynamic properties of liquid mixtures from molecular dynamics simulations. *The Journal of chemical physics* **2016**, *145*.

(104) Caro, M. A.; Lopez-Acevedo, O.; Laurila, T. Redox potentials from ab initio molecular dynamics and explicit entropy calculations: Application to transition metals in aqueous solution. *Journal of chemical theory and computation* **2017**, *13*, 3432–3441.

(105) Huang, S.-N.; Pascal, T. A.; Goddard III, W. A.; Maiti, P. K.; Lin, S.-T. Absolute entropy and energy of carbon dioxide using the two-phase thermodynamic model. *Journal of chemical theory and computation* **2011**, *7*, 1893–1901.