

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

Reviewed Preprint

v2 • November 8, 2024

Revised by authors

Reviewed Preprint

v1 • June 3, 2024

Sneha Menon, Subinoy Adhikari, Jagannath Mondal 

Tata Institute of Fundamental Research, Hyderabad, India

 https://en.wikipedia.org/wiki/Open_access
 Copyright information

eLife Assessment

This study describes the application of machine learning and Markov state models to characterize the binding mechanism of alpha-Synuclein to the small molecule Fasudil. The results suggest that entropic expansion can explain such binding. However, the simulations and analyses in their present form are **inadequate**.

<https://doi.org/10.7554/eLife.97709.2.sa3>

Abstract

The mis-folding and aggregation of intrinsically disordered proteins (IDPs) such as α -synuclein (α S) underlie the pathogenesis of various neurodegenerative disorders. However, targeting α S 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 α S 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 α S or gives rise to new conformational states of α S, 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 α S 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}

Considering the broad range of functionality of IDPs, their dysregulation 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 PTP1B²¹ and osteopontin²² have been successfully targeted by small molecules.²³

The IDP of this article's interest, alpha-synuclein (α S) has remained an important drug-target^{24–27} owing to its long-rooted association with neurodegenerative diseases. Often, the early aggregation stages of α S into small oligomers is implicated in the causation of Parkinson's disease^{28–31} and a potential therapeutic strategy has been the stabilization of α S in soluble monomeric form. In particular, the small molecule fasudil (focus of the present investigation) has been shown to interact with α S and retard its aggregation.²⁶ The anti-aggregative effect of fasudil on α S assembly was tested both in vivo and in vitro and it was found to significantly reduce α S 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 coworkers³² had produced an unprecedentedly long-time scale (1500 μ s) all-atom molecular dynamics (MD) simulation trajectory of small molecule drug fasudil with monomeric α S, 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 α S ensemble in neat water and that in the presence of fasudil³² (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 α S 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, nonlinear 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-learned 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 α S ensemble, we harnessed deep-learning based β -Variational AutoEncoder (β -VAE) and Markov State Models (MSM) to dissect α S ensemble in water and in aqueous fasudil solution. A projection of the atomistically simulated conformation ensemble of α S 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, α S, manifested as a change in the backbone entropy of the fasudil binding hotspots.

Results and discussion

1. Modulation of free energy landscape of α S in presence of fasudil

Simulations of α S 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 α S. In our study, in order to compare the α S-fasudil ensemble with an apo ensemble at same salt concentration, we spawned multiple all-atom MD simulations of α S in neat water and 50 mM NaCl from multiple starting conformations to generate a cumulative ensemble of $\sim 62 \mu$ s.

In a clear departure from the classical view of ligand binding to a folded globular protein, the visual change in α S 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 α S (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 C_α pairwise distances, as this feature largely incorporates the conformational space of α S. 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.

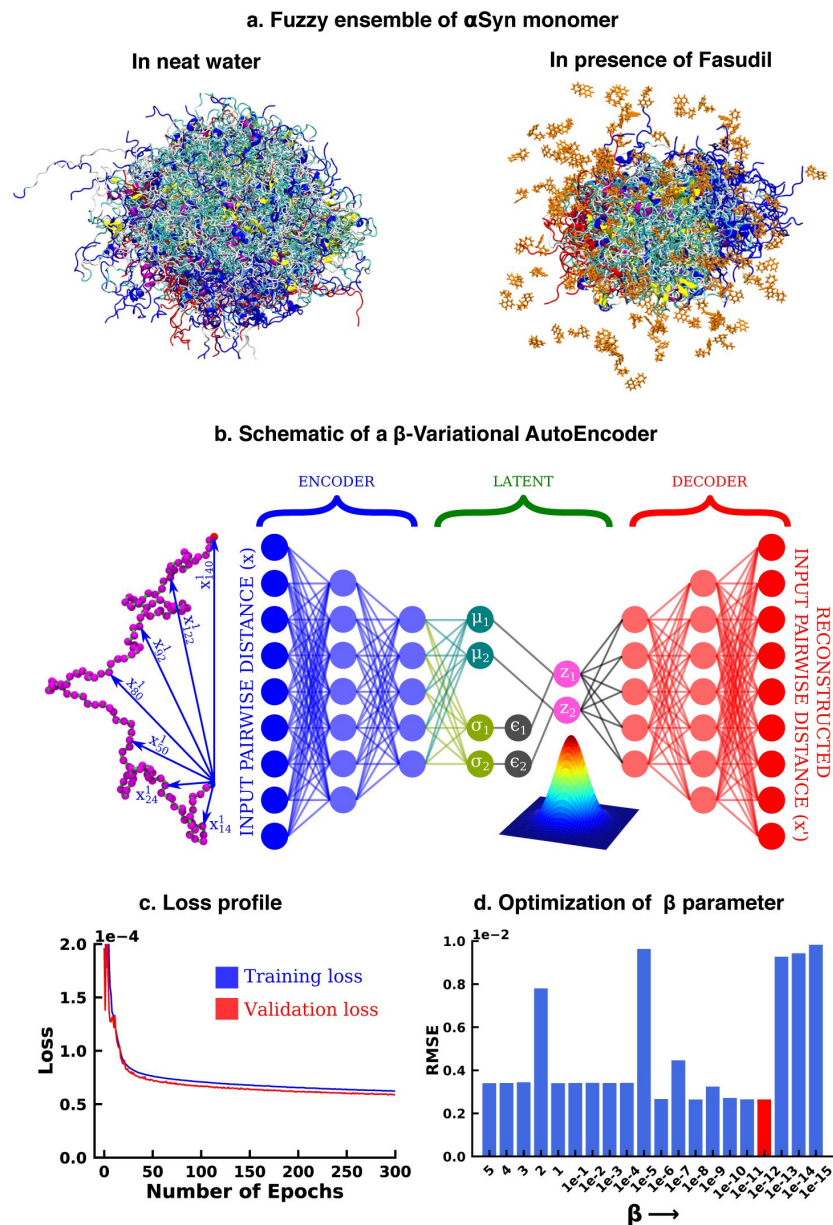


Figure 1

(a) Conformational ensemble view of α S and α S-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

We started our investigation by searching for an optimized representation of high-dimensional and heterogeneous ensemble of monomeric α S 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 analysis⁴⁶ 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 embedding⁵¹ etc.⁵² have outperformed such linear transformation methods and have been used for free energy calculations⁵³ 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 α S.

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 distance 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,

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

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 α S 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 α S-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 α S and α S-Fasudil simulations.

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 α S, 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 α S, 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 α S conformations. Subsequent projection of the apo state on the low-dimensional subspace also brought out a spatially heterogeneous landscape.

However, a close comparison of free energy landscapes indicates that a set of local minima representing conformations of α S are less clustered and appear in small patches in the apo state. Moreover, a relative comparison of the two free energy landscapes suggests that α S 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 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 α S (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 α S 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 PyEMMA⁵⁷. 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 α S in water with three kinetically separated states (see **Figure 3d** for relative population). On the other hand, a similar analysis of α S 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 α S ensemble in water whereas a six-state model was built for the α S-fasudil ensemble. Hereafter, we will refer to the macrostates created in presence of fasudil prefixed with

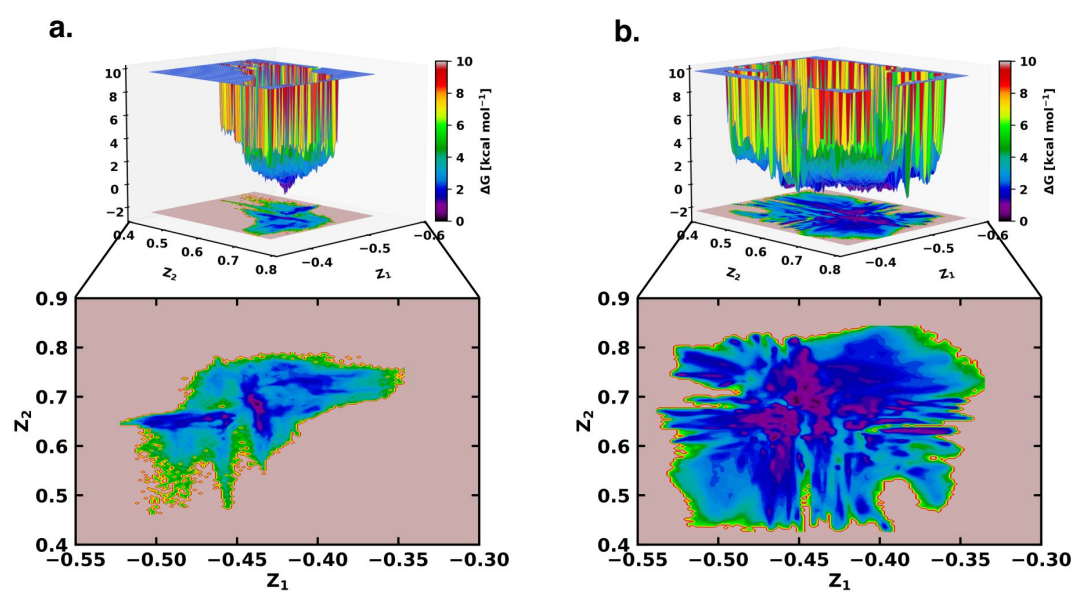


Figure 2

The free energy landscape of (a) α S and (b) α S-Fasudil system as determined from the latent dimensions of β -VAE

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** [↗](#)).

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 α S-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 α S led to sampling or generation of additional conformational states than the states populated in the apo α S ensemble.

A pertinent question arises: Are the appearances of newer states in α S in presence of fasudil a result of the presence of longer trajectory in fasudil solution (1500 μ s) than that of the ensemble populated in neat water (62 μ s)? We verified this by building an MSM for 60 μ s of data of fasudil system using 10 μ s to 70 μ s segment of the trajectory (see Figure S10). We find that the MSM built using lesser data (and same amount of data in the α S simulation) also indicated the presence of six states of α S 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 α S ensemble and found to be in good agreement (see Figure S11). Additionally, we selected a chunk spanning from 966 μ s to 1026 μ s 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 α S in fasudil solution relative to that in water is a direct outcome of the interaction of α S with the small molecule.

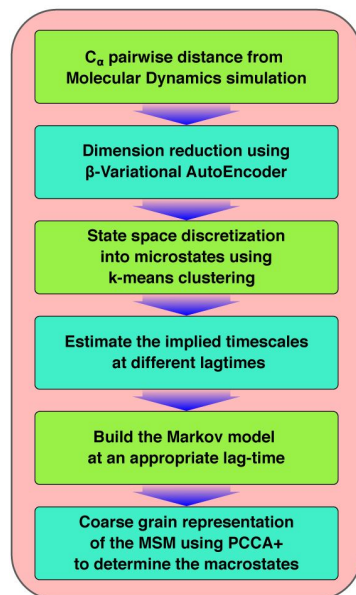
Another aspect that we evaluated was in the choice of the number of latent dimensions of β -VAE. Even though an increase in the number of latent dimensions may make the model more accurate, this can also result in overfitting. The model can simply memorize the pattern in the data instead of generalizing 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 α S monomer in presence of fasudil

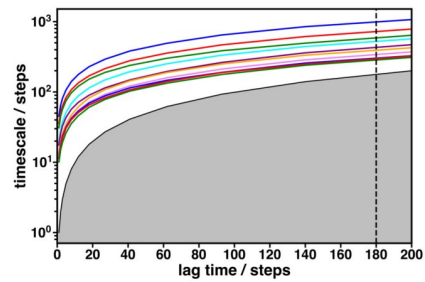
We characterised the residue-wise intramonomer contact maps of the metastable states to identify the differences in the interaction patterns in these states that can be attributed to interactions of fasudil with the monomer. The average inter-residue contact probability maps for each of the metastable 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

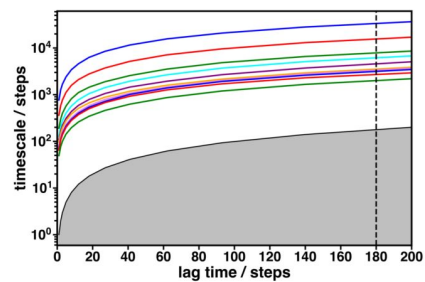
a. Protocol to construct MSM



b. α Syn



c. α Syn-Fasudil



d. MSM derived macrostate populations

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

Figure 3

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

Intramonomer contact maps

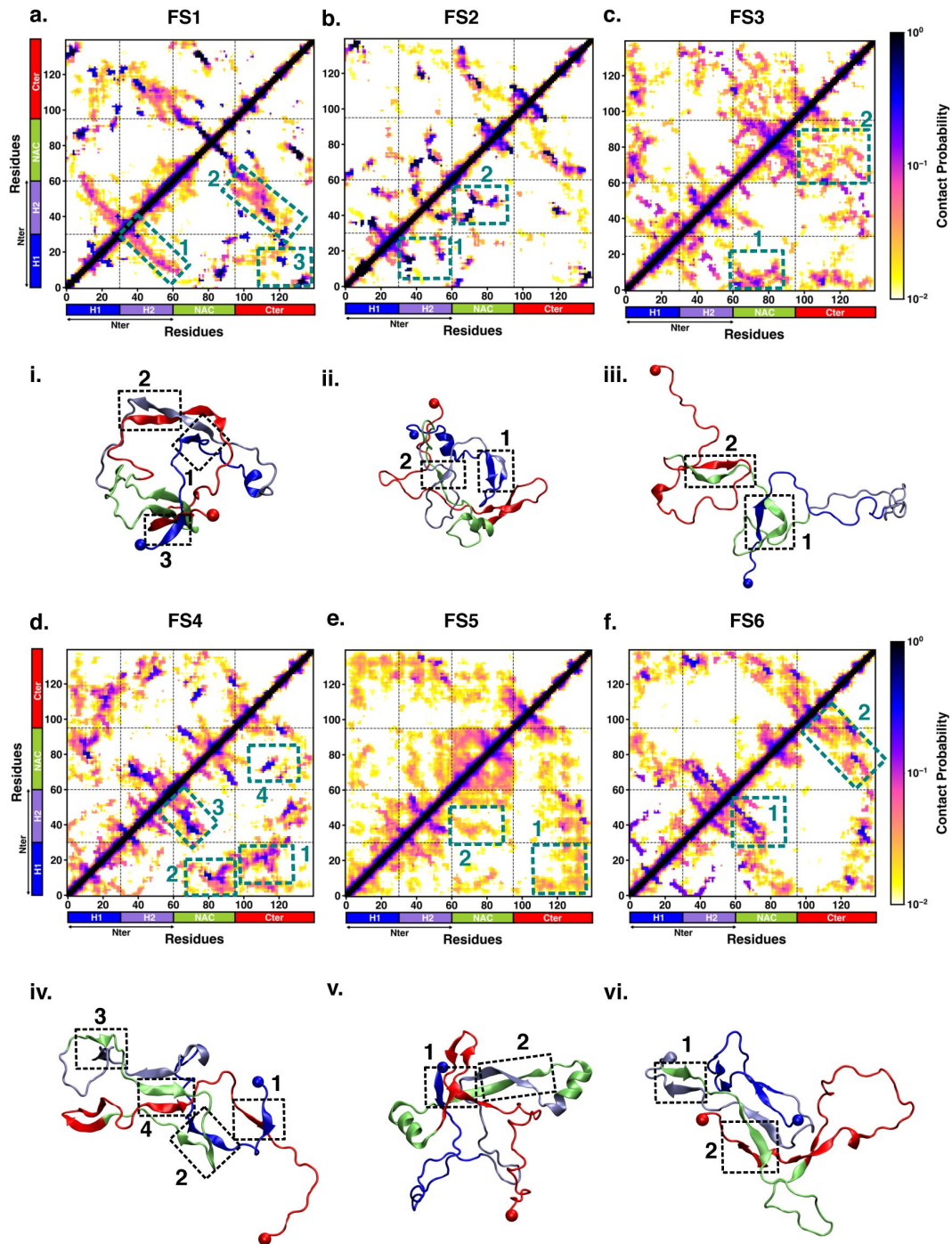


Figure 4

Intra-peptide 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 each other. 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 α S 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.

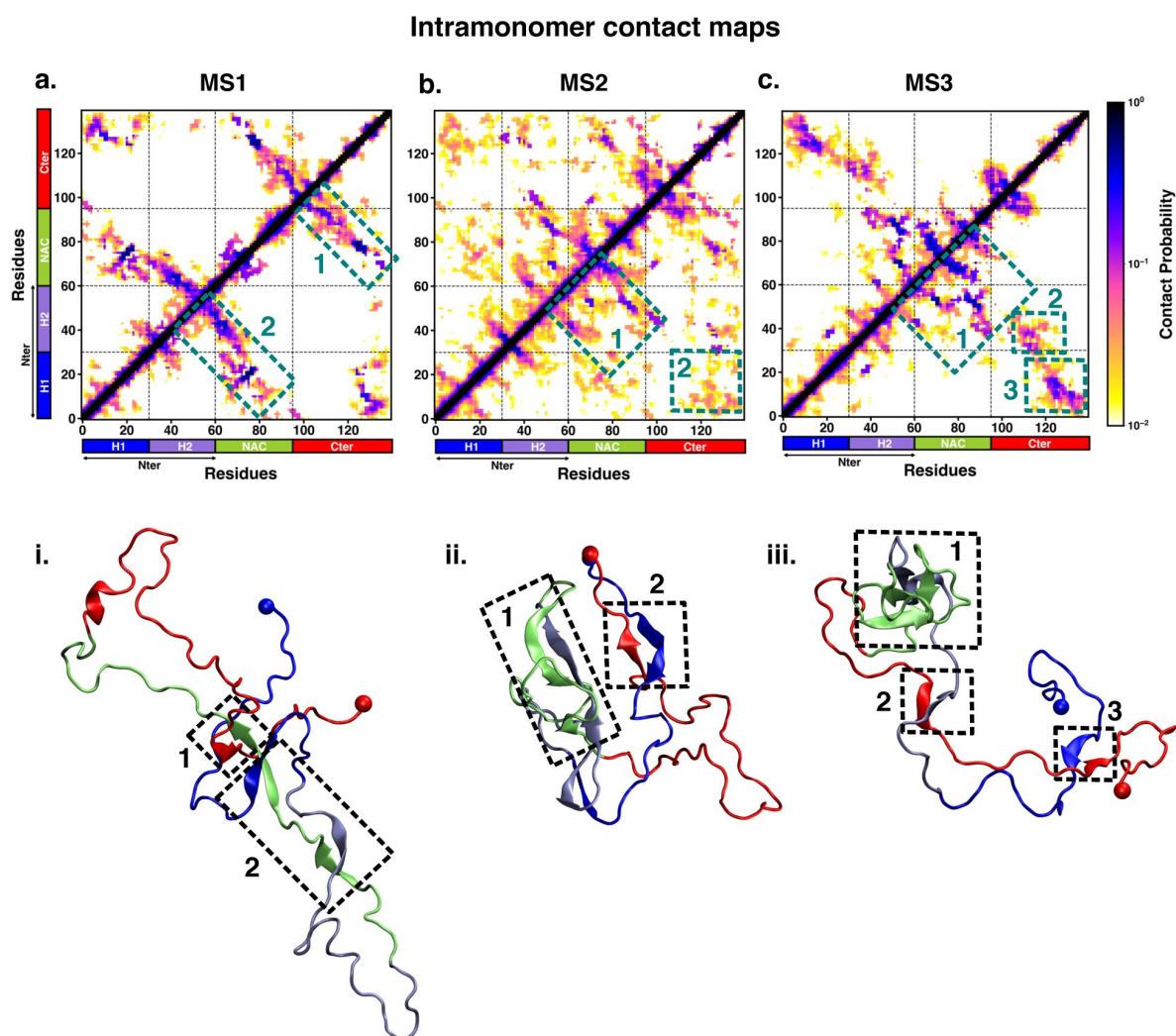


Figure 5

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 each other. 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 α S 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.

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 ~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 α S.

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

To ascertain the distinctions among these contact maps for the α S and α S-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 α S and α S-Fasudil macrostates as inputs. The model was initially trained using the six contact maps of α S-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)

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 α S and α S-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 corrupting 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 α S and α S-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 α S. Accordingly, we calculated the contact probabilities between fasudil and each α S 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 α S-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 α S 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- α S interaction via the C-terminal region of the protein while it weakly interacts with the entire α S 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 amphipathic 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).

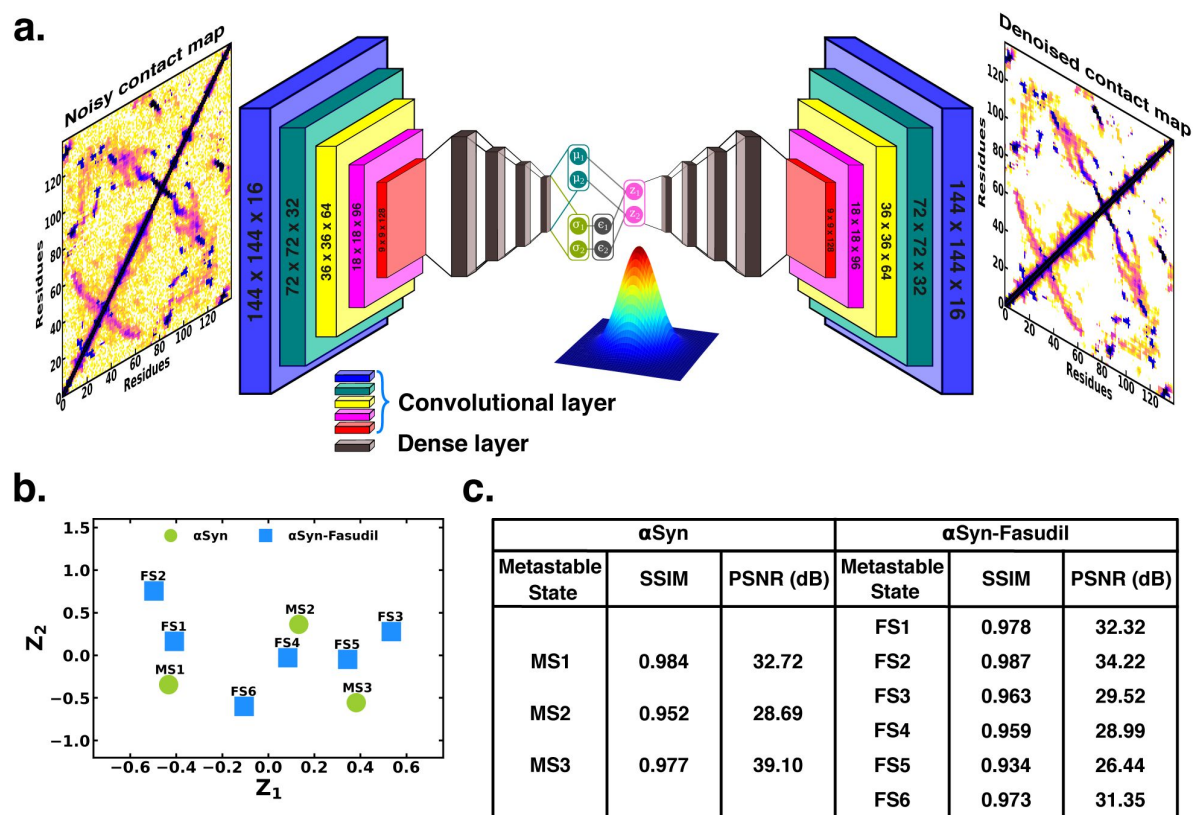


Figure 6

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

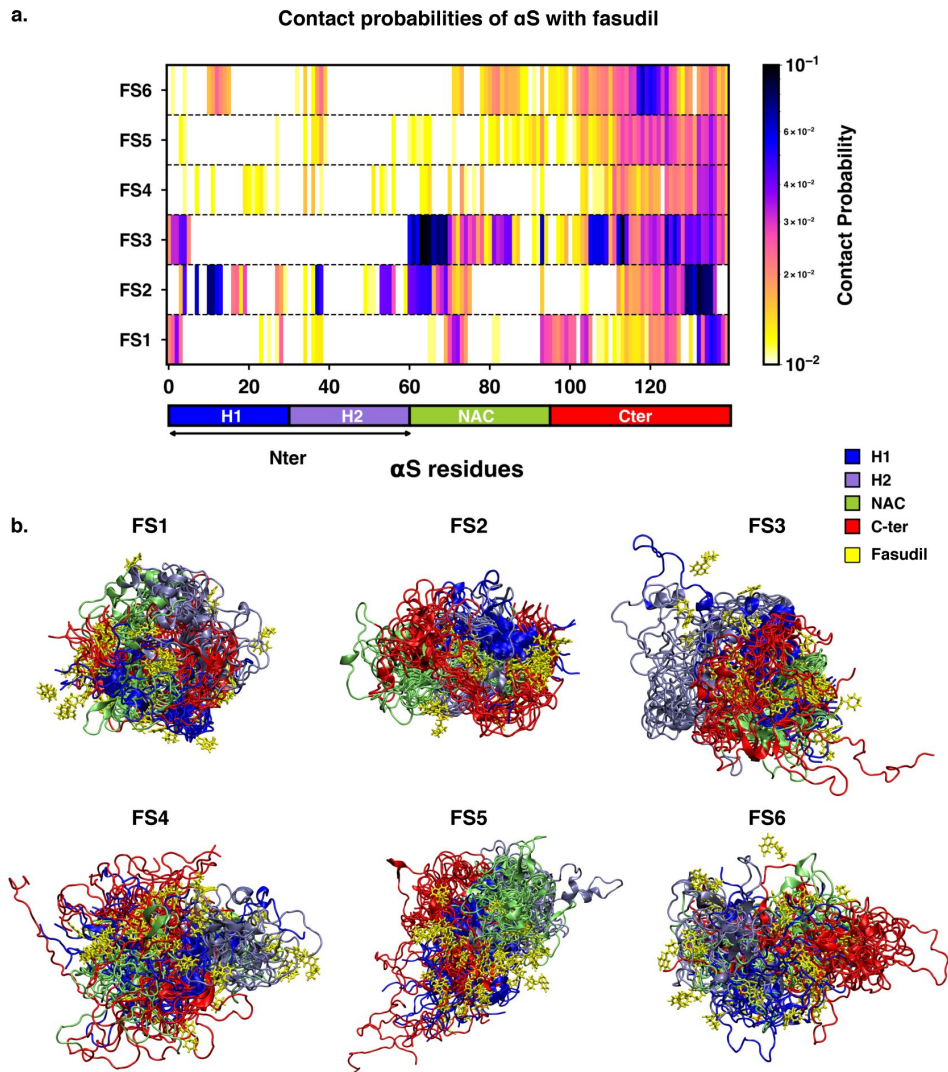


Figure 7

(a) Contact probability map of residue-wise interaction of fasudil with α S. (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.

Robustelli et al.³² have reported that in the bound state, fasudil interactions with α S 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.

6. Entropic Signatures of Small Molecule Binding

The overall analysis of the conformational ensemble of α S in presence of fasudil demonstrates that small-molecule binding substantially modulates the ensemble of α S. Our results indicate that the interaction of fasudil with α S residues governs the structural features of the protein. Estimation of the metastable states of α S 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 α S 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 α S 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 α S, 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 $55.5 \text{ J mol}^{-1} \text{ K}^{-1}$, which is close to the reported value of $54.4 \text{ J mol}^{-1} \text{ K}^{-1}$ from free energy perturbation calculation.⁶⁶ However, the molar entropy is marginally reduced in all the macrostates of both α S and α S-fasudil systems, with values ranging between $55.01 \text{ J mol}^{-1} \text{ K}^{-1}$ to $55.25 \text{ J mol}^{-1} \text{ K}^{-1}$, 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 α S 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 program⁶⁷ 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 $-75 \text{ J mol}^{-1} \text{ K}^{-1}$, the entropy of fasudil-bound states (FS1 to FS6) vary from -70 to $-86 \text{ J mol}^{-1} \text{ K}^{-1}$, indicating that binding of fasudil to α S 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 $-70.8 \text{ J mol}^{-1} \text{ K}^{-1}$ and $-72.3 \text{ J mol}^{-1} \text{ K}^{-1}$, respectively. These states are also the most populated metastable states. MS1 and MS3 states have entropy values of $-73.9 \text{ J mol}^{-1} \text{ K}^{-1}$ and $-74.5 \text{ J mol}^{-1} \text{ K}^{-1}$, 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 $-85.3 \text{ J mol}^{-1} \text{ K}^{-1}$, followed by FS2 having entropy of $-83.1 \text{ J mol}^{-1} \text{ K}^{-1}$. FS4 is also relatively more ordered than FS5 with entropy of $-76.2 \text{ J mol}^{-1} \text{ K}^{-1}$.

FS3 and FS6 are closely disordered as FS5 with entropy values of $-73.2 \text{ J mol}^{-1} \text{ K}^{-1}$ and $-72.5 \text{ J mol}^{-1} \text{ K}^{-1}$, respectively. These observations of the total extent of disorderliness of the metastable states clearly indicate that the interactions of fasudil with α S 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 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 α S 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.

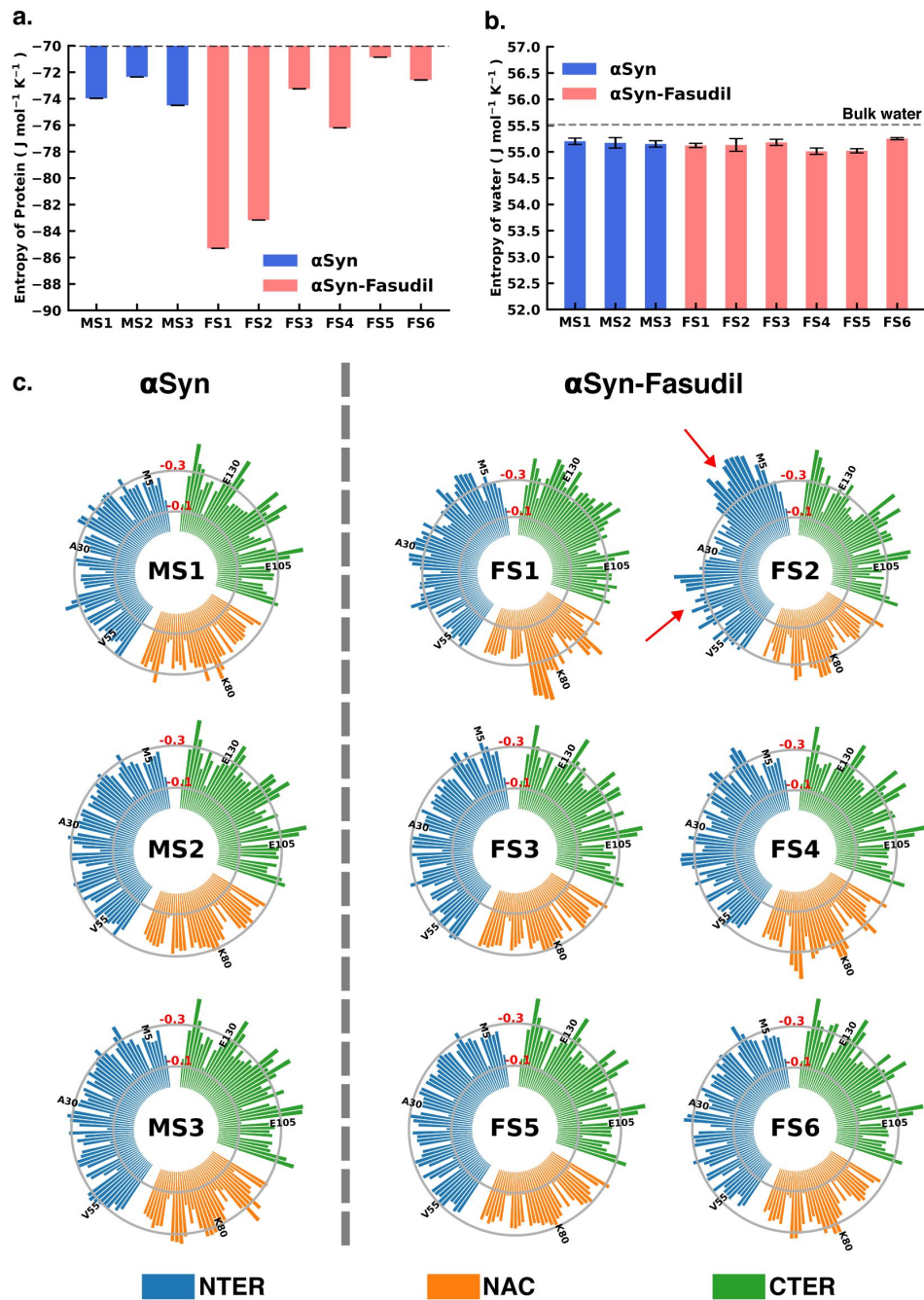


Figure 8

(a) Total Protein entropy of the metastable states of αS and $\alpha\text{S-Fasudil}$ calculated from the torsion angles, relative to a fully flexible chain (b) Total water entropy in αS and $\alpha\text{S-Fasudil}$ system and (c) Residue-wise backbone entropy within the αS and $\alpha\text{S-Fasudil}$ states.

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 α S indicate that in the presence and absence of fasudil, the transition to the most populated state is preferred (see [Figure 9](#)). For α S-Fasudil system, the transition to the most populated states takes place over timescales ranging from 74 μ s and 766.5 μ s, whereas for the apo state it occurs between 29.8 μ s and 37.4 μ s. 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 α S and α S-Fasudil system is the most favoured over all other states in the 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 α S and α S-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 α S and α S-Fasudil system. However, as compared to α S in its apo state, the apparent increase (~ 2 -20 times) in the time to transition to the major state in α S-

Fasudil system, suggests that fasudil may potentially trap the protein conformations in the intermediate states, thereby slowing down α S 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 α S both in vivo and in vitro. The ensemble view of long time scale atomic simulations of fasudil binding to α S monomer displays a fuzzy ensemble of α S 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. $C\alpha$ -pairwise distance into a two-dimensional latent space. We built a Markov State Model using this data to delineate the metastable conformational states of α S and identify the differences that arise from small-molecule interactions. Comparison of the metastable states of α S 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](#)).

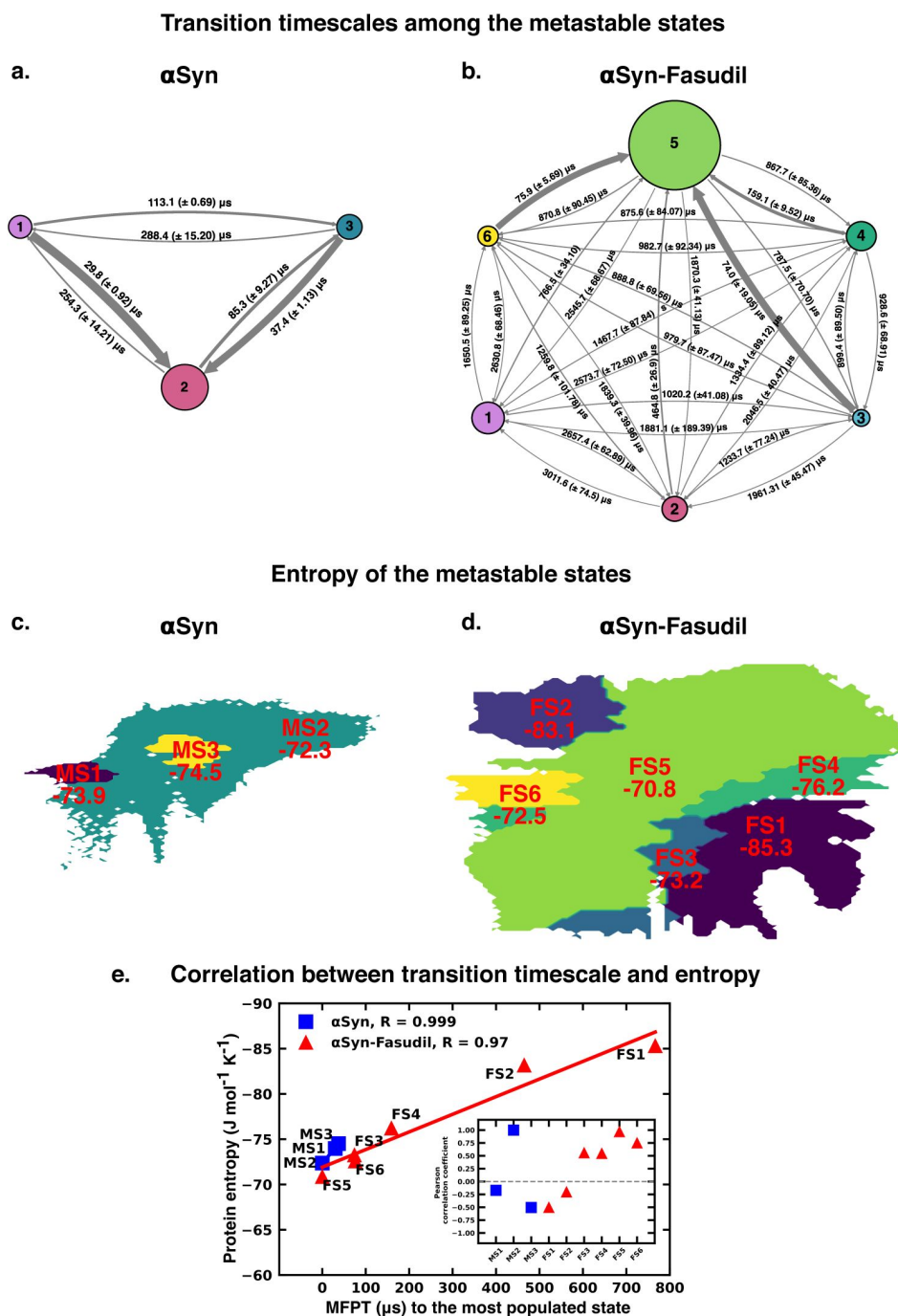


Figure 9

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

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 α S 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 α S monomer ensemble reveals the altering effects of small-molecule interaction on α S 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 μ s long all-atom MD simulation trajectory of α S 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 forcefield (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 α S ensemble in water, we performed multiple simulations of α S 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 reported³³ multi- μ s α S 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 model³³ 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 thermostat⁸⁸ with a relaxation time of 0.1 ps and a pressure of 1 bar using the Parrinello-Rahman barostat⁸⁹ with a time constant of 2 ps for coupling. The Verlet cutoff scheme⁹⁰ with 1.0 nm cutoff was applied for Lennard-Jones interactions and shortrange electrostatic interactions. Long-range electrostatic interactions were computed using the Particle-Mesh Ewald (PME) summation method⁹¹ 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 μ s, amounting to a cumulative length of $\sim 62 \mu$ s.

We have compared the ensemble properties of the α S 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 α S simulations we performed from different starting structures have effectively sampled the phase space as the single long simulation of the α S-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,

$$p(x) = \int p(x|z)p(z)dz \quad (2)$$

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

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

Upon expanding and rearranging, we have

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

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 [2](#) and is given as,

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

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

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

which modifies the loss function to

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

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 [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

$$z = \mu + \sigma \odot \epsilon \quad (8)$$

where E 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 α S protein in presence of fasudil trajectory as our input. We have built our model using Tensorflow backend⁹⁶ [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 [97](#) method in all the layers except for the output layer where we have used *sigmoid* activation. Adaptive Moment Estimation (Adam⁹⁸ [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 [99](#)).

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, 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 $\beta = 1 \times 10^{-12}$. 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 α S and α S-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

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (9)$$

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. 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

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

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

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

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 $C_1 = (k_1L)^2$,

$C_2 = (k_2L)^2$ 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.

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

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

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

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

where M is the height (number of rows) of the image, N is the width (number of columns) of the image and O_{ij} and 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 α S and α S-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,

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

where m_l is the mass of atom l , v^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

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

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 ,

$$S_k = k_B \left[\int_0^\infty I_k^s(v) W_k^s(v) dv + \int_0^\infty I_k^g(v) W_k^g dv \right] \quad (16)$$

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

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

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}

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 (2015) **Intrinsically disordered proteins in cellular signalling and regulation** *Nat. Rev. Mol. Cell. Biol* **16**:18–29
- (2) Babu M. M., van der Lee R., de Groot N. S., Gsponer J (2011) **Intrinsically disordered proteins: regulation and disease** *Current Opinion in Structural Biology* **21**:432–440
- (3) Csizmok V., Follis A. V., Kriwacki R. W., Forman-Kay J. D (2016) **Dynamic Protein Interaction Networks and New Structural Paradigms in Signaling** *Chemical Reviews* **116**:6424–6462
- (4) Wang Y., Fisher J. C., Mathew R., Ou L., Otieno S., Sublet J., Xiao L., Chen J., Roussel M. F., Kriwacki R. W (2011) **Intrinsic disorder mediates the diverse regulatory functions of the Cdk inhibitor p21** *Nature chemical biology* **7**:214–221
- (5) Dyson H. J (2016) **Making sense of intrinsically disordered proteins** *Biophysical journal* **110**
- (6) Alberti S., Gladfelter A., Mittag T (2019) **Considerations and challenges in studying liquid-liquid phase separation and biomolecular condensates** *Cell* **176**:419–434
- (7) Chong P. A., Forman-Kay J. D (2016) **Liquid-liquid phase separation in cellular signaling systems** *Current opinion in structural biology* **41**:180–186
- (8) Shin Y., Brangwynne C. P (2017) **Liquid phase condensation in cell physiology and disease** *Science* **357**
- (9) O’Flynn B. G., Mittag T (2021) **The role of liquid-liquid phase separation in regulating enzyme activity** *Current opinion in cell biology* **69**:70–79
- (10) Uversky V. N., Oldfield C. J., Dunker A. K (2008) **Intrinsically disordered proteins in human diseases: introducing the D2 concept** *Annual review of biophysics* **37**:215–246
- (11) Bronowska A., Moreno Piraján JC (2011) **Thermodynamics of Ligand-Protein Interactions: Implications for Molecular Design** *Thermodynamics - Interaction Studies - Solids, Liquids and Gases* IntechOpen
- (12) Klebe G (2015) **Applying thermodynamic profiling in lead finding and optimization** *Nat. Rev. Drug Discov* **14**:95–110
- (13) Metallo S. J (2010) **Intrinsically disordered proteins are potential drug targets** *Current Opinion in Chemical Biology* **14**:481–488
- (14) Follis A.V., Hammoudeh D.I., Wang H., Prochownik E.V., Metallo S.J. (2008) **Structural rationale for the coupled binding and unfolding of the c-Myc oncoprotein by small molecules** *Chemistry and Biology* **15**:1149–1155
- (15) Ono K., Li L., Takamura Y., Yoshiike Y., Zhu L., Han H., Mao X., Ikeda T., Takasaki T., Nishijo H (2012) **Phenolic compounds prevent amyloid -protein oligomerization and synaptic dysfunction by site-specific binding** *J. Biol. Chem* **287**:14631–14643

- (16) Zhu M., De Simone A., Schenk D., Toth G., Dobson C. M., Vendruscolo M (2013) **Identification of small-molecule binding pockets in the soluble monomeric form of the Abeta42 peptide** *The Journal of Chemical Physics* **139**
- (17) Attanasio F. *et al.* (2013) **Carnosine Inhibits Abeta-42 Aggregation by Perturbing the H-Bond Network in and around the Central Hydrophobic Cluster** *ChemBioChem* **14**:583–592
- (18) Ehrnhoefer D. E., Bieschke J., Boeddrich A., Herbst M., Masino L., Lurz R., Engemann S., Pastore A., Wanker E. E (2008) **EGCG redirects amyloidogenic polypeptides into unstructured, off-pathway oligomers** *Nature structural and molecular biology* **15**:558–566
- (19) Heller G. T., *et al.* (2020) **Small-molecule sequestration of amyloid as a drug discovery strategy for Alzheimers disease** *Science Advances* **6**
- (20) Akoury E., Gajda M., Pickhardt M., Biernat J., Soraya P., Griesinger C., Mandelkow E., Zweckstetter M (2013) **Inhibition of Tau Filament Formation by Conformational Modulation** *Journal of the American Chemical Society* **135**:2853–2862
- (21) others,,,, *et al.* (2014) **Targeting the disordered C terminus of PTP1B with an allosteric inhibitor** *Nature chemical biology* **10**:558–566
- (22) Kurzbach D., Schwarz T. C., Platzer G., Hafler S., Hinderberger D., Konrat R (2014) **Compensatory Adaptations of Structural Dynamics in an Intrinsically Disordered Protein Complex** *Angewandte Chemie International Edition* **53**:3840–3843
- (23) Biesaga M., Frigol-Vivas M., Salvatella X (2021) **Intrinsically disordered proteins and biomolecular condensates as drug targets** *Current Opinion in Chemical Biology* **62**:90–100
- (24) Fields C. R., Bengoa-Vergniory N., Wade-Martins R (2019) **Targeting Alpha-Synuclein as a Therapy for Parkinson's Disease** *Frontiers in Molecular Neuroscience* **12**
- (25) Tah G., *et al.* (2014) **Targeting the Intrinsically Disordered Structural Ensemble of alpha-Synuclein by Small Molecules as a Potential Therapeutic Strategy for Parkinson's Disease** *PLOS ONE* **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. (2016) **Fasudil attenuates aggregation of α -synuclein in models of Parkinson's disease** *Acta neuropathologica communications* **4**:1–17
- (27) Cao K., Zhu Y., Hou Z., Liu M., Yang Y., Hu H., Dai Y., Wang Y., Yuan S., Huang G (2022) **α -Synuclein as a Target for Metallo-Anti-Neurodegenerative Agents** *Angewandte Chemie International Edition*
- (28) Winner B., Jappelli R., Maji S. K., Desplats P. A., Boyer L., Aigner S., Hetzer C., Loher T., Vilar M., Campioni S (2011) **In vivo demonstration that α -synuclein oligomers are toxic** *Proceedings of the National Academy of Sciences* **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 (2022) **Small soluble α -synuclein aggregates are the toxic species in Parkinson's disease** *Nature communications* **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 (2022) **Decreased Water Mobility Contributes To Increased α -Synuclein Aggregation** *Angewandte Chemie International Edition*

- (31) Ubbiali D., Fratini M., Piersimoni L., Ihling C. H., Kipping M., Heilmann I., Iacobucci C., Sinz A (2022) **Direct Observation of Elongated Conformational States in α -Synuclein upon Liquid-Liquid Phase Separation** *Angewandte Chemie* **134**
- (32) Robustelli P., Ibanez-de Opakua A., Campbell-Bezat C., Giordanetto F., Becker S., Zweckstetter M., Pan A. C., Shaw D. E (2022) **Molecular Basis of Small-Molecule Binding to α -Synuclein** *Journal of the American Chemical Society* **144**:2501–2510
- (33) Robustelli P., Piana S., Shaw D. E (2018) **Developing a molecular dynamics force field for both folded and disordered protein states** *Proceedings of the National Academy of Sciences* **115**:E4758–E4766
- (34) Bonomi M., Camilloni C., Cavalli A., Vendruscolo M (2016) **Metainference: A Bayesian inference method for heterogeneous systems** *Science advances* **2**
- (35) Bottaro S., Lindorff-Larsen K (2018) **Biophysical experiments and biomolecular simulations: A perfect match?** *Science* **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 (2020) **Conformational ensembles of an intrinsically disordered protein consistent with NMR, SAXS, and single-molecule FRET** *Journal of the American Chemical Society* **142**:15697–15710
- (37) Stelzl L. S., Pietrek L. M., Holla A., Oroz J., Sikora M., KoÅàfinger J., Schuler B., Zweckstetter M., Hummer G (2022) **Global structure of the intrinsically disordered protein tau emerges from its local structure** *Jacs Au* **2**:673–686
- (38) Prinz J.-H., Wu H., Sarich M., Keller B., Senne M., Held M., Chodera J. D., Schutte C., Noe F (2011) **Markov models of molecular kinetics: Generation and validation** *J. Chem. Phys* **134**
- (39) Bowman G. R., Beauchamp K. A., Boxer G., Pande V. S (2009) **Progress and challenges in the automated construction of Markov state models for full protein systems** *J. Chem. Phys* **131**
- (40) Husic B. E., Pande V. S (2018) **Markov State Models: From an Art to a Science** *Journal of the American Chemical Society* **140**:2386–2396
- (41) Adhikari S., Mondal J (2023) **Machine Learning Subtle Conformational Change due to Phosphorylation in Intrinsically Disordered Proteins** *The Journal of Physical Chemistry B*
- (42) Kingma D. P., Welling M (2013) **Auto-encoding variational bayes** *arXiv*
- (43) Rezende D. J., Mohamed S., Wierstra D (2014) **Stochastic backpropagation and approximate inference in deep generative models** *International conference on machine learning* :1278–1286
- (44) Hyvärinen A (2013) **Independent component analysis: recent advances. Philosophical Transactions of the Royal Society A: Mathematical Physical and Engineering Sciences** **371**
- (45) Klema V., Laub A (1980) **The singular value decomposition: Its computation and some applications** *IEEE Transactions on automatic control* **25**:164–176
- (46) Fisher R. A (1936) **The use of multiple measurements in taxonomic problems** *Annals of eugenics* **7**:179–188

- (47) Schölkopf B., Smola A., Müller K.-R (2005) **Kernel principal component analysis** *Artificial Neural Networks-ICANN 97: 7th International Conference* :583–588
- (48) Kruskal J. B., Wish M. (1978) **Multidimensional scaling** Sage
- (49) Tenenbaum J. B., Silva V. d., Langford J. C (2000) **A global geometric framework for nonlinear dimensionality reduction** *science* **290**:2319–2323
- (50) Faloutsos C., Lin K.-I (1995) **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* :163–174
- (51) Roweis S. T., Saul L. K (2000) **Nonlinear dimensionality reduction by locally linear embedding** *science* **290**:2323–2326
- (52) Sumithra V., Surendran S (2015) **A review of various linear and non linear dimensionality reduction techniques** *Int. J. Comput. Sci. Inf. Technol* **6**:2354–2360
- (53) Das P., Moll M., Stamati H., Kavraki L. E., Clementi C (2006) **Low-dimensional, freeenergy landscapes of protein-folding reactions by nonlinear dimensionality reduction** *Proceedings of the National Academy of Sciences* **103**:9885–9890
- (54) Schwantes C. R., Pande V. S (2015) **Modeling molecular kinetics with tICA and the kernel trick** *Journal of chemical theory and computation* **11**:600–608
- (55) Wehmeyer C., Noé F (2018) **Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics** *The Journal of chemical physics* **148**
- (56) Hinton G. E., Salakhutdinov R. R (2006) **Reducing the dimensionality of data with neural networks** *science* **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 (2015) **PyEMMA 2: A software package for estimation, validation, and analysis of Markov models** *Journal of chemical theory and computation* **11**:5525–5542
- (58) Hore A., Ziou D (2010) **Image quality metrics: PSNR vs. SSIM** *2010 20th international conference on pattern recognition* :2366–2369
- (59) Heller G. T., Bonomi M., Vendruscolo M (2018) **Structural Ensemble Modulation upon Small-Molecule Binding to Disordered Proteins** *Journal of Molecular Biology* **430**:2288–2292
- (60) Heller G. T., Sormani P., Vendruscolo M (2015) **Targeting disordered proteins with small molecules using entropy** *Trends Biochem. Sci* **40**:491–496
- (61) Flock T., Weatheritt R. J., Latysheva N. S., Babu M. M (2014) **Controlling entropy to tune the functions of intrinsically disordered regions** *Current Opinion in Structural Biology* **26**:62–72
- (62) Dyson H., Wright P. E (2002) **Coupling of folding and binding for unstructured proteins** *Current Opinion in Structural Biology* **12**:54–60
- (63) Löhr T., Kohlhoff K., Heller G. T., Camilloni C., Vendruscolo M (2022) **A Small Molecule Stabilizes the Disordered Native State of the Alzheimer's A β Peptide** *ACS Chemical Neuroscience* **13**:1738–1745

- (64) Lin S.-T., Blanco M., Goddard W. A (2003) **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* **119**:11792–11805
- (65) Lin S.-T., Maiti P. K., Goddard W. A (2010) **Two-phase thermodynamic model for efficient and accurate absolute entropy of water from molecular dynamics simulations** *The Journal of Physical Chemistry B* **114**:8191–8198
- (66) Mukherjee S., Schäfer L. V (2023) **Thermodynamic forces from protein and water govern condensate formation of an intrinsically disordered protein domain** *Nature Communications* **14**
- (67) Fogolari F., Maloku O., Dongmo Fomthum C. J., Corazza A., Esposito G (2018) **PDB2ENTROPY and PDB2TRENT: Conformational and Translational–Rotational Entropy from Molecular Ensembles** *Journal of Chemical Information and Modeling* **58**:1319–1324
- (68) King B. M., Tidor B (2009) **MIST: Maximum Information Spanning Trees for dimension reduction of biological data sets** *Bioinformatics* **25**:1165–1172
- (69) King B. M., Silver N. W., Tidor B (2012) **Efficient calculation of molecular configurational entropies using an information theoretic approximation** *The Journal of Physical Chemistry B* **116**:2891–2904
- (70) Dedmon M. M., Lindorff-Larsen K., Christodoulou J., Vendruscolo M., Dobson C. M (2005) **Mapping long-range interactions in α -synuclein using spin-label NMR and ensemble molecular dynamics simulations** *Journal of the American Chemical Society* **127**:476–477
- (71) Allison J. R., Varnai P., Dobson C. M., Vendruscolo M (2009) **Determination of the free energy landscape of α -synuclein using spin label nuclear magnetic resonance measurements** *Journal of the American Chemical Society* **131**:18314–18326
- (72) Ullman O., Fisher C. K., Stultz C. M (2011) **Explaining the structural plasticity of α -synuclein** *Journal of the american chemical society* **133**:19536–19546
- (73) Esteban-Martín S., Silvestre-Ryan J., Bertoncini C. W., Salvatella X (2013) **Identification of fibril-like tertiary contacts in soluble monomeric α -synuclein** *Biophysical journal* **105**:1192–1198
- (74) Best R. B., Zheng W., Mittal J (2014) **Balanced protein–water interactions improve properties of disordered proteins and non-specific protein association** *Journal of chemical theory and computation* **10**:5113–5124
- (75) Huang J., Rauscher S., Nawrocki G., Ran T., Feig M., De Groot B. L., Grubmüller H., MacKerell A. D (2017) **CHARMM36m: an improved force field for folded and intrinsically disordered proteins** *Nature methods* **14**:71–73
- (76) Bertoncini C. W., Jung Y.-S., Fernandez C. O., Hoyer W., Griesinger C., Jovin T. M., Zweckstetter M (2005) **Release of long-range tertiary interactions potentiates aggregation of natively unstructured α -synuclein** *Proceedings of the National Academy of Sciences* **102**:1430–1435
- (77) Bernadó P., Blanchard L., Timmins P., Marion D., Ruigrok R. W., Blackledge M (2005) **A structural model for unfolded proteins from residual dipolar couplings and small-angle x-ray scattering** *Proceedings of the National Academy of Sciences* **102**:17002–17007

- (78) Marsh J. A., Baker J. M., Tollinger M., Forman-Kay J. D (2008) **Calculation of residual dipolar couplings from disordered state ensembles using local alignment** *Journal of the American Chemical Society* **130**:7804–7805
- (79) Bernadó P., Mylonas E., Petoukhov M. V., Blackledge M., Svergun D. I (2007) **Structural characterization of flexible proteins using small-angle X-ray scattering** *Journal of the American Chemical Society* **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 (2021) **Refinement of α -synuclein ensembles against SAXS data: Comparison of force fields and methods** *Frontiers in molecular biosciences* **8**
- (81) Brookes D. H., Head-Gordon T (2016) **Experimental inferential structure determination of ensembles for intrinsically disordered proteins** *Journal of the American Chemical Society* **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 (2019) **Extended Experimental Inferential Structure Determination Method for Evaluating the Structural Ensembles of Disordered Protein States** *arXiv*
- (83) Roux B., Weare J (2013) **On the statistical equivalence of restrained-ensemble simulations with the maximum entropy method** *The Journal of chemical physics* **138**
- (84) Boomsma W., Ferkinghoff-Borg J., Lindorff-Larsen K (2014) **Combining experiments and simulations using the maximum entropy principle** *PLoS computational biology* **10**
- (85) Cavalli A., Camilloni C., Vendruscolo M (2013) **Molecular dynamics simulations with replica-averaged structural restraints generate structural ensembles according to the maximum entropy principle** *The Journal of chemical physics* **138**
- (86) Robustelli P., Piana S., Shaw D. E (2020) **Mechanism of Coupled Folding-upon-Binding of an Intrinsically Disordered Protein** *Journal of the American Chemical Society* **142**:11092–11101
- (87) Lindahl E., Abraham M., Hess B., Spoel v. d. D (2020) **GROMACS 2020 Source code**
- (88) Bussi G., Donadio D., Parrinello M (2007) **Canonical sampling through velocity rescaling** *The Journal of chemical physics* **126**
- (89) Parrinello M., Rahman A (1981) **Polymorphic transitions in single crystals: A new molecular dynamics method** *Journal of Applied physics* **52**:7182–7190
- (90) Páll S., Hess B (2013) **A flexible algorithm for calculating pair interactions on SIMD architectures** *Computer Physics Communications* **184**:2641–2650
- (91) Darden T., York D., Pedersen L (1993) **The effect of long-range electrostatic interactions in simulations of macromolecular crystals—a comparison of the ewald and truncated list methods** *J. Chem. Phys* **99**
- (92) Hess B., Bekker H., Berendsen H. J., Fraaije J. G (1997) **LINCS: a linear constraint solver for molecular simulations** *Journal of computational chemistry* **18**:1463–1472
- (93) Bengio S., Bengio Y (2000) **Taking on the curse of dimensionality in joint distributions using neural networks** *IEEE Transactions on Neural Networks* **11**:550–557

- (94) Larochelle H., Murray I (2011) **The neural autoregressive distribution estimator** *Proceedings of the fourteenth international conference on artificial intelligence and statistics* :29–37
- (95) Higgins I., Matthey L., Pal A., Burgess C., Glorot X., Botvinick M., Mohamed S., Lerchner A (2016) **beta-vae: Learning basic visual concepts with a constrained variational framework** *International conference on learning representations*
- (96) Abadi M., et al. (2016) **Tensorflow: Large-scale machine learning on heterogeneous distributed systems** *arXiv*
- (97) Glorot X., Bengio Y (2010) **Understanding the difficulty of training deep feedforward neural networks** *Proceedings of the thirteenth international conference on artificial intelligence and statistics* :249–256
- (98) Kingma D. P., Ba J (2014) **Adam: A method for stochastic optimization** *arXiv*
- (99) Wang Z., Bovik A. C., Sheikh H. R., Simoncelli E. P (2004) **Image quality assessment: from error visibility to structural similarity** *IEEE transactions on image processing* **13**:600–612
- (100) Weinan E., Vanden-Eijnden E (2006) **Towards a theory of transition paths** *Journal of statistical physics* **123**:503–523
- (101) Metzner P., Schütte C., Vanden-Eijnden E (2009) **Transition path theory for Markov jump processes** *Multiscale Modeling & Simulation* **7**:1192–1219
- (102) Noé F., Schütte C., Vanden-Eijnden E., Reich L., Weikl T. R (2009) **Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations** *Proceedings of the National Academy of Sciences* **106**:19011–19016
- (103) Caro M. A., Laurila T., Lopez-Acevedo O (2016) **Accurate schemes for calculation of thermodynamic properties of liquid mixtures from molecular dynamics simulations** *The Journal of chemical physics* **145**
- (104) Caro M. A., Lopez-Acevedo O., Laurila T (2017) **Redox potentials from ab initio molecular dynamics and explicit entropy calculations: Application to transition metals in aqueous solution** *Journal of chemical theory and computation* **13**:3432–3441
- (105) Huang S.-N., Pascal T. A., Goddard W. A., Maiti P. K., Lin S.-T (2011) **Absolute entropy and energy of carbon dioxide using the two-phase thermodynamic model** *Journal of chemical theory and computation* **7**:1893–1901

Author information

Sneha Menon¹

Tata Institute of Fundamental Research, Hyderabad, India

¹These authors contributed equally to this work

Subinoy Adhikari¹

Tata Institute of Fundamental Research, Hyderabad, India

¹These authors contributed equally to this work

Jagannath Mondal

Tata Institute of Fundamental Research, Hyderabad, India

For correspondence: jmondal@tifrh.res.in

Editors

Reviewing Editor

Rosana Colleparado

University of Cambridge, Cambridge, United Kingdom

Senior Editor

Qiang Cui

Boston University, Boston, United States of America

Reviewer #2 (Public Review):

The manuscript by Menon et al describes a set of simulations of alpha-Synuclein (aSYN) and analyses of these and previous simulations in the presence of a small molecule.

Comments on latest version:

I have read the authors' response to my comments as well as to the other reviewers. Summarizing briefly, I don't think they provide substantial answer to the questions/comments by me or reviewer 3, and generally do not quantify the results/effects data. I still remain unconvinced about the analyses and conclusions. Rather than rewriting another set of comments, I think it will be more useful for all (authors and readers) simply to be able to see the entire set of reviews and responses together with the paper.

<https://doi.org/10.7554/eLife.97709.2.sa2>

Reviewer #3 (Public Review):

In this manuscript Menon, Adhikari, and Mondal analyze explicit solvent molecular dynamics (MD) computer simulations of the intrinsically disordered protein (IDP) alpha-synuclein in the presence and absence of a small molecule ligand, Fasudil, previously demonstrated to bind alpha-synuclein by NMR spectroscopy without inducing folding into more ordered structures. In order to provide insight into the binding mechanism of Fasudil the authors analyze an unbiased 1500us MD simulation of alpha-synuclein in the presence of Fasudil previously reported by Robustelli et.al. (Journal of the American Chemical Society, 144(6), pp.2501-2510). The authors compare this simulation to a very different set of apo simulations: 23 separate 1-4us simulations of alpha-synuclein seeded from different apo conformations taken from another previously reported by Robustelli et. al. (PNAS, 115 (21), E4758-E4766), for a total of ~62us.

To analyze the conformational space of alpha-synuclein - the authors employ a variational auto-encoder (VAE) to reduce the dimensionality of Ca-Ca pairwise distances to 2 dimensions,

and use the latent space projection of the VAE to build Markov state Models. The authors utilize k-means clustering to cluster the sampled states of alpha-synuclein in each condition into 180 microstates on the VAE latent space. They then coarse grain these 180 microstates into a 3-macrostate model for apo alpha-synuclein and a 6-macrostate model for alpha-synuclein in the presence of fasudil using the PCCA+ coarse graining method. Few details are provided to explain the hyperparameters used for PCCA+ coarse graining and the rationale for selecting the final number of macrostates.

The authors analyze the properties of each of the alpha-synuclein macrostates from their final MSMs - examining intramolecular contacts, secondary structure propensities, and in the case of alpha-synuclein:Fasudil holo simulations - the contact probabilities between Fasudil and alpha-synuclein residues.

The authors utilize an additional variational autoencoder (a denoising convolutional VAE) to compare denoised contact maps of each macrostate, and project onto an additional latent space. The authors conclude that their apo and holo simulations are sampling distinct regions of the conformational space of alpha-synuclein projected on the denoising convolutional VAE latent space.

Finally, the authors calculate water entropy and protein conformational entropy for each microstate. To facilitate water entropy calculations - the author's take a single structure from each macrostate - and ran a 20ps simulation at a finer timestep (4 femtoseconds) using a previously published method (DoSPT), which computes thermodynamic properties of water from MD simulations using autocorrelation functions of water velocities. The authors report that water entropy calculated from these individual 20ps simulations is very similar.

For each macrostate the authors compute protein conformational entropy using a previously published Maximum Information Spanning tree approach based on torsion angle distributions - and observe that the estimated protein conformational entropy is substantially more negative for the macrostates of the holo ensemble.

The authors calculate mean first passage times from their Markov state models and report a strong correlation between the protein conformational entropy of each state and the mean first passage time from each state to the highest populated state.

As the authors observe the conformational entropy estimated from macrostates of the holo alpha-synuclein:Fasudil is greater than those estimated from macrostates of the apo holo alpha-synuclein macrostates - they suggest that the driving force of Fasudil binding is an increase in the conformational entropy of alpha-synuclein. No consideration/quantification of the enthalpy of alpha-synuclein Fasudil binding is presented.

Strengths:

The author's utilize MD simulations run with an appropriate force field for IDPs (a99SB-disp and a99SB-disp water (Robustelli et. al, PNAS, 115 (21), E4758-E4766) - which has previously been used to perform MD simulations of alpha-synuclein that have been validated with extensive NMR data.

The contact probability between Fasudil and each alpha-synuclein residue observed in the previously performed 1500us MD simulation of alpha-synuclein in the presence of Fasudil (Robustelli et. al., Journal of the American Chemical Society, 144(6), pp.2501-2510) was previously found to be in good agreement with experimental NMR chemical shift perturbations upon Fasudil binding - suggesting that this simulation is a reasonable choice for understanding IDP:small molecule interactions.

Comments on the latest version:

While the authors have provided additional information in the updated manuscript, none of the additional analyses address the fundamental flaws of the manuscript.

The additional analyses do not convincingly demonstrate that these two extremely different simulation datasets (1500 microsecond unbiased MD for a-synuclein + fasudil, 23 separate 1-4 microsecond simulations of apo a-synuclein) are directly comparable for the purposes of building MSMs.

The additional analyses do not demonstrate that there are sufficient conformational transitions among kinetically metastable states observed in 23 separate 1-4 microsecond simulations of apo a-synuclein to build a valid MSM, or that the latent space of the VAE is kinetically meaningful.

If one is interested in modeling the kinetics and thermodynamics of transitions between a set of conformational states, and they run a small number of MD simulations that are too short to see conformational transitions between conformational states - any kinetics and thermodynamics modeled by an MSM will be inherently meaningless. This is likely to be the case with the apo a-synuclein dataset analyzed in this investigation.

Simulations of 1-4 microseconds are almost certainly far too short to see a meaningful sampling of conformational transitions of a highly entangled 140-residue IDP beyond a very local relaxation of the starting structures, and the authors provide no analyses to suggest otherwise.

Without convincingly demonstrating reasonable statistics of conformational changes from the very small apo simulation dataset analyzed here, it seems highly likely the apparent validity of the apo MSM results from learning a VAE latent space that groups structurally and kinetically distinct conformations into similar states, creating the spurious appearance of transitions between states. As such, the kinetics and thermodynamics of the resulting MSM are likely to be relatively meaningless, and comparisons with an MSM for a-synuclein in the presence of fasudil are likely to be meaningless.

In its present form, this study provides an example of how the use of black-box machine learning methods to analyze molecular simulations can lead to obtaining misleading results (such as the appearance of a valid MSM) - when more basic analyses are omitted.

<https://doi.org/10.7554/eLife.97709.2.sa1>

Author response:

The following is the authors' response to the current reviews.

Public Reviews:

Reviewer #2 (Public Review):

I have read the authors' response to my comments as well as to the other reviewers. Summarizing briefly, I don't think they provide substantial answer to the questions/comments by me or reviewer 3, and generally do not quantify the results/effects data. I still remain unconvinced about the analyses and conclusions. Rather than rewriting another set of comments, I think it will be more useful for all (authors and readers) simply to be able to see the entire set of reviews and responses together with the paper.

The authors disagree with the views of referees. The authors have provided point-wise precise responses to each of the previous comments. The authors find that the referee has not been able to engage with the responses and accompanying analysis that were provided while communicating the previous response.

The following extensive analyses were performed by the authors while submitting our revision of round 2 of peer-review to address the comments of reviewer 2 and reviewer 3 that were raised by them on the previous versions:

- (1) We calculated the distribution of multiple metrics for both the apo and holo simulations, including their secondary structure composition, and demonstrated the robustness of our findings.
 - (2) We analyzed smaller 60 μ s chunks from two parts of the 1.5 ms trajectory and showed how, in combination with the Markov state modeling (MSM) approach, these chunks effectively capture equilibrium properties.
 - (3) We thoroughly investigated the choice of starting structures, examining parameters such as Rg, RMSD, secondary structure, and SASA, in response to Referee 3's concerns about the objectivity of our dimension reduction approach.
 - (4) We conducted multiple analyses using VAMP-scores and justified the use of a Variational Autoencoder (VAE) over tICA.
 - (5) We had extensively verified the choice of hyperparameters used in constructing the MSM.
 - (6) To alleviate referee concerns, we had retrained a VAE with four latent dimensions and used it to build an MSM, ensuring the robustness of our approach.
- However, we find that Referee has not considered these additional analysis in response to his/her comments on the manuscript.

Since referee 2 also draws comments from Referee 3, it is worth noting that some of the comments from Referee 2 and Referee 3 in Round 1 were mutually contradictory. In particular, Referee 3's suggestion in Round 1 to use the same initial configuration for simulations of intrinsically disordered proteins (IDPs) in both apo and ligand-bound forms contradicts the fundamental principle that IDPs should not possess structural bias. This recommendation also directly conflicts with Referee 2's request for greater diversity in starting structures. Our manuscript provided robust evidence that our initial configurations are indeed diverse, with one configuration coincidentally matching that used in the ligand-bound simulations. Despite this, we addressed both sets of concerns in our Round 2 revisions. Unfortunately, it seems that these efforts were overlooked in the subsequent round of review.

Referee 2's suggestion in previous round of review comments to mix both holo and apo simulation trajectories for MSM construction is conceptually wrong and indicates a lack of understanding of transition matrix building in this field. Nevertheless, we addressed these comments by performing additional analyses and demonstrating the robustness of our current MSM.

Reviewer #3 (Public Review):

Summary:

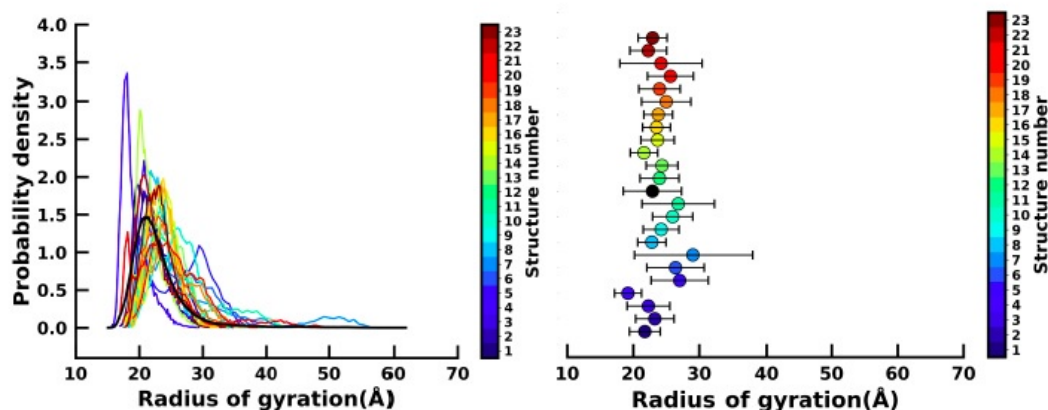
While the authors have provided additional information in the updated manuscript, none of the additional analyses address the fundamental flaws of the manuscript.

The additional analyses do not convincingly demonstrate that these two extremely different simulation datasets (1500 microsecond unbiased MD for α -synuclein + fasudil, 23 separate 1-4 microsecond simulations of apo α -synuclein) are directly comparable for the purposes of building MSMs.

The 23 unbiased 1-4 microsecond simulations of apo α S totals to ~ 60 us.

Author response image 1.

Left figure : Distribution of the radius of gyration (Rg) of the 23 apo simulation (as shown in the colourbar) and holo simulation (black). Right figure : Mean and standard deviation (as error bar) of the Rg of the 23 apo (colourbar) and holo simulations (black).



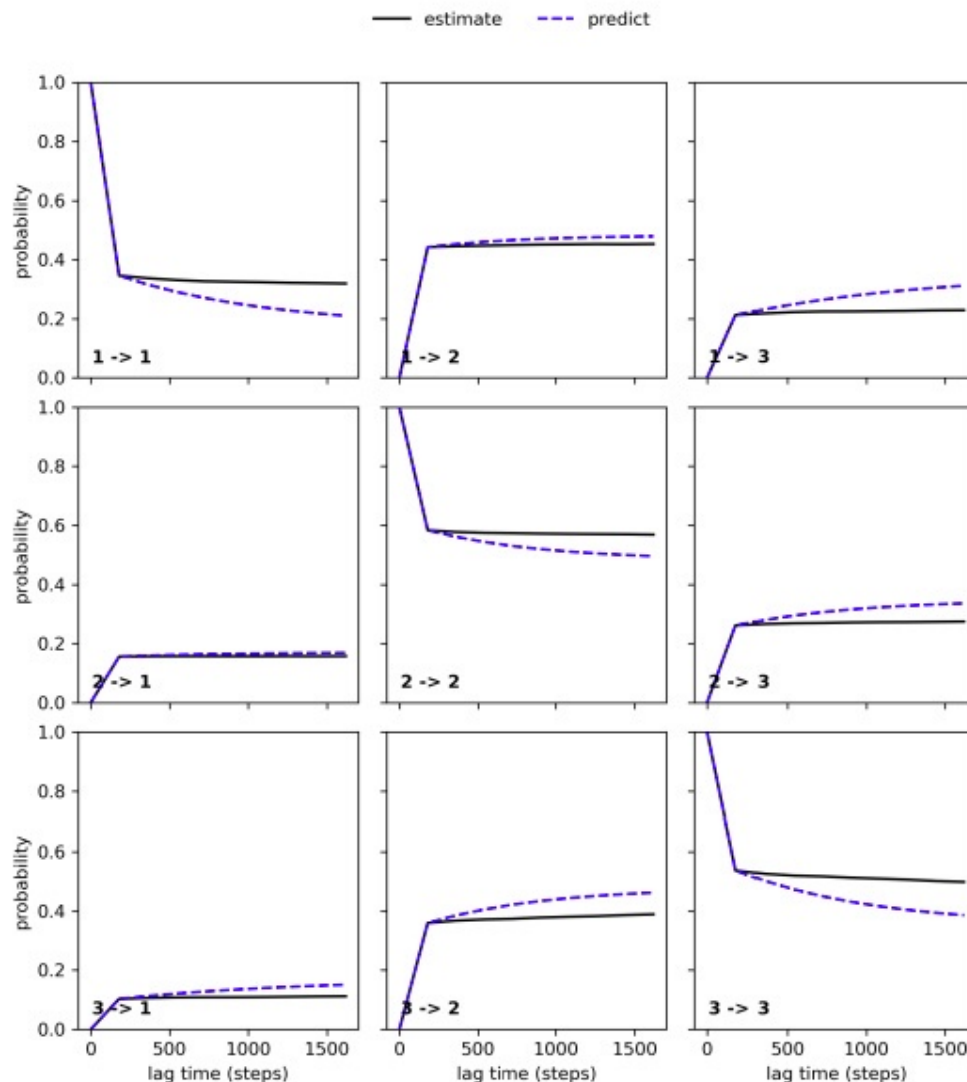
We have plotted the distribution of the Radius of gyration ((Rg) for the 23 apo simulation (colour bar) and the holo simulation (black) as shown in the left figure and also compared the mean and standard deviations of the Rg values (right figure). We find that our apo simulations span the entire space of Rg as is spanned by the holo simulation. We have also measured the mean and standard deviations (SD) (horizontal error bar) of the apo and holo simulations. The fact that the apo simulations have mean and SDs comparable to those of the holo ensemble suggests that the majority of the apo simulations are sampling similar conformational space as those observed in the ligand-bound holo form and hence can be used for building the MSM.

The additional analyses do not demonstrate that there are sufficient conformational transitions among kinetically metastable states observed in 23 separate 1-4 microsecond simulations of apo α -synuclein to build a valid MSM, or that the latent space of the VAE is kinetically meaningful.

We have performed the Chapman-Kolmogorov test to compare observed and predicted transition probabilities over increasing lag times and found good agreement between these probabilities, thereby suggesting that transitions between states are well-sampled for both the apo (Author response image 2) and holo simulation (Figure S9).

Author response image 2.

The Chapman-Kolmogorov test performed for the three state Markov State Model of the α S ensemble.



As for the latent space of VAE, we have compared the VAMP2 score and compared with tICA. VAE has a higher VAMP2 score as compared to tICA thereby indicating its efficacy in capturing slower mode for both apo and holo simulation (Fig. S7 and S8).

If one is interested in modeling the kinetics and thermodynamics of transitions between a set of conformational states, and they run a small number of MD simulations that are too short to see conformational transitions between conformational states - any kinetics and thermodynamics modeled by an MSM will be inherently meaningless. This is likely to be the case with the apo asynuclein dataset analyzed in this investigation.

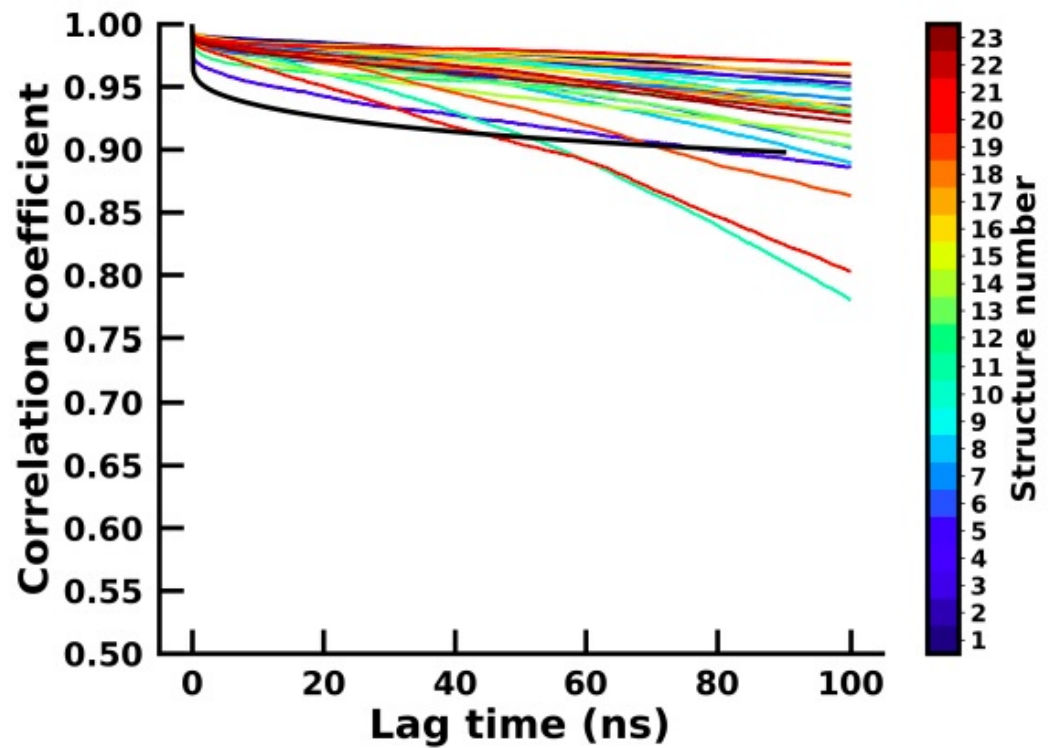
We disagree with the referee's view. The referee does not seem to understand the point of building Markov state models via short-time scale trajectories. The distribution of R_g of all the 23 apo simulations spans the entire R_g space sampled by the holo simulation, thereby suggesting that multiple short simulations can sample structures of varying sizes as sampled from the 1.5 ms holo simulation (see Author response image 1).

Simulations of 1-4 microseconds are almost certainly far too short to see a meaningful sampling of conformational transitions of a highly entangled 140-residue IDP beyond a

very local relaxation of the starting structures, and the authors provide no analyses to suggest otherwise.

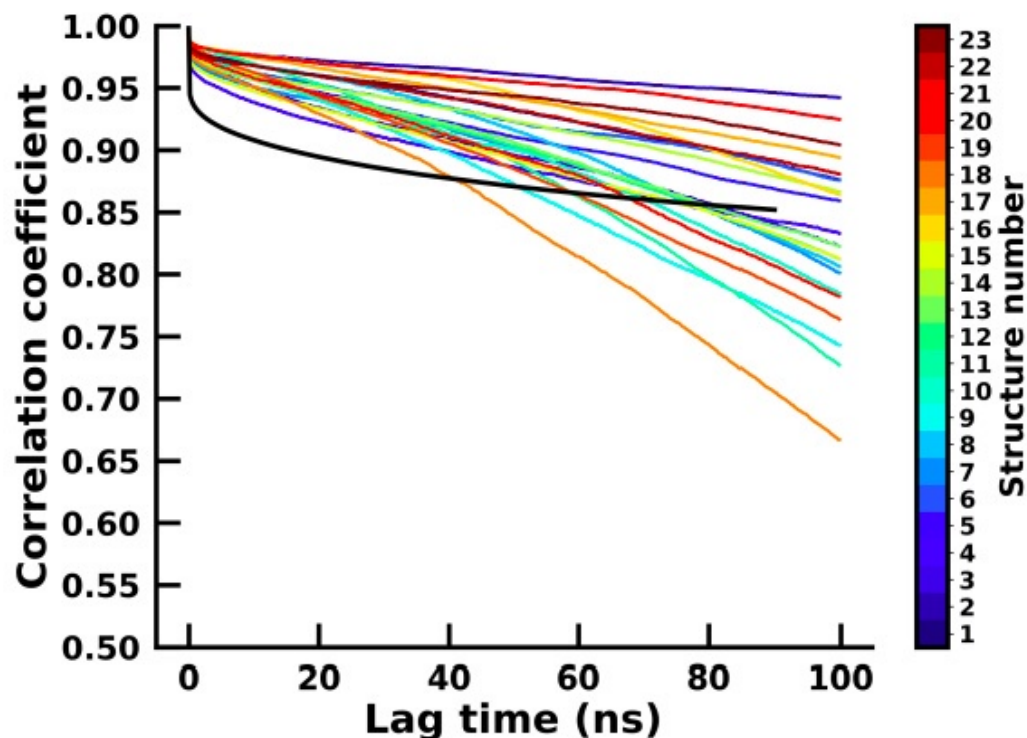
Author response image 3.

Autocorrelation of the first principal component of the backbone dihedral for the apo (colourbar) and holo (black) simulation.



Author response image 4.

Autocorrelation of the second principal component of the backbone dihedral for the apo (colourbar) and holo (black) simulation.



In order to assess the 23 short simulations in capturing meaningful kinetics and thermodynamics, we have computed the backbone dihedrals which were then reduced to two principal components for both the 23 apo and holo simulations. We then calculated the autocorrelation time for each of the components and for each of the apo and holo simulations which are plotted in Author response image 3 and Author response image 4 respectively.

The autocorrelation for the holo and most of the apo simulation is similar, thereby suggesting that there is sufficient sampling of conformational transitions between conformational states in the apo simulations and are therefore able to represent the structural changes of the system similarly to the long simulation.

Without convincingly demonstrating reasonable statistics of conformational changes from the very small apo simulation dataset analyzed here, it seems highly likely the apparent validity of the apo MSM results from learning a VAE latent space that groups structurally and kinetically distinct conformations into similar states, creating the spurious appearance of transitions between states. As such, the kinetics and thermodynamics of the resulting MSM are likely to be relatively meaningless, and comparisons with an MSM for α -synuclein in the presence of fasudil are likely to be meaningless.

We have shown above that the short simulations are able to capture the structural changes in the long simulation. In addition we have compared the VAMP2 score of the apo and holo simulation with tICA and found out that VAE is superior in capturing long timescale dynamics, for both apo and holo simulation (Fig. S7 and S8).

In its present form, this study provides an example of how the use of black-box machine learning methods to analyze molecular simulations can lead to obtaining misleading results (such as the appearance of a valid MSM) - when more basic analyses are omitted.

The authors disagree with the referee's viewpoint on our manuscript. We find that the majority of the contents of the referee's comments are cursory and lack objectivity.

The referee's loose reference on Machine learning as a *black box* lacks basic knowledge to comprehend artificial deep neural network's long-proven ability to objectively deduce optimal set of lower-dimensional representation of conformational subspace of complex biomacromolecule. The referee's views on the manuscript ignore the extensive optimization of hyper-parameters that were carried out by the authors in developing the suitable framework of beta-variational autoencoder for deducing optimal latent space representation of complex and fuzzy conformational landscape of an IDP such as alpha-synuclein. We had thoroughly investigated the choice of starting structures, examining parameters such as Rg, RMSD, secondary structure, and SASA, in response to Referee 3's concerns about the objectivity of our dimension reduction approach. However, we find that referee 3 has ignored the analysis provided to justify our choice.

Referee 3's advocacy for linear dimensional reduction techniques overlooks the necessity and generality of non-linear approaches, as enabled by artificial deep neural network frameworks, demonstrated in the present manuscript. Nevertheless, our manuscript includes evidence demonstrating the optimality of our current reduced dimensions through varied dimensional analyses. Our extensive analysis, based on the VAMP-2 score, supports the sufficiency of the present dimensions compared to other linear reduction methods.

The referee's views that developing Markov state models (MSM) of apo form of the alphasynuclein using multiple number of 1-4 microsecond long simulation length is misleading, suggests referee's lack of knowledge on the fundamental purpose and motivation for the usage of MSM, which is, to derive long-time scale equilibrium properties from significantly short-length adaptively sampled trajectories. The referee has overlooked the extensive analysis that the authors had provided while demonstrating that the Markov state models developed from short length simulation trajectories of alpha-synuclein can statistically replicate the properties derived from very long trajectories.

The following is the authors' response to the original reviews.

The following extensive analyses were performed to address the reviewer comments:

- (1) We have calculated the distribution of radius of gyration (Rg), end-to-end distance (Ree), solvent accessible surface area (SASA) of the apo and holo simulations and also their secondary structure composition.
- (2) We have performed a similar analysis for the smaller 60 μ s chunk from two parts of the 1.5 ms trajectory.
- (3) The choice of starting structures have been thoroughly investigated in terms of Rg, RMSD, secondary structure and SASA.
- (4) We have justified the use of VAE over tICA.
- (5) We have verified the choice of hyperparameters that were used to build the MSM.
- (6) We have retrained a VAE with four latent dimensions and used it to build MSM.
- (7) As per recommendation of the referee 1, we have updated the title of the manuscript by introducing 'expansion' phrase.

The manuscript has been accordingly revised by updating it with additional analysis.

Public Reviews:

Reviewer #1 (Public Review):

Summary:

This is a well-conducted study about the mechanism of binding of a small molecule (fasudil) to a disordered protein (alpha-synuclein). Since this type of interaction has puzzled researchers for the last two decades, the results presented are welcome as they offer relevant insight into the physical principles underlying this interaction.

Strengths:

The results show convincingly that the mechanism of entropic expansion can explain the previously reported binding of fasudil to alpha-synuclein. In this context, the analysis of the changes in the entropy of the protein and of water is highly relevant. The combination use of machine learning for dimensional reduction and of Markov State Models could become a general procedure for the analysis of other systems where a compound binds a disordered protein.

Weaknesses:

It would be important to underscore the computational nature of the results, since the experimental evidence that fasudil binds alpha-synuclein is not entirely clear, at least to my knowledge.

The experimental evidence of binding of fasudil to α -synuclein and potentially preventing its aggregation is reported in the paper “Fasudil attenuates aggregation of α -synuclein in models of Parkinson’s disease. Tatenhorst et al. Acta Neuropathologica Communications (2016) 4:39 DOI 10.1186/s40478-016-0310-y”. In this work, solution state 15N-1H HSQC NMR experiments were performed of α -synuclein in increasing amounts of fasudil which led to large chemical shift perturbation of Y133 and Y136 residues. Additionally single and double mutant synT-Y133A and synT-Y136A (tyrosine is replaced with alanine), when treated with fasudil, had no significant effect as evident from immunochemistry, thereby indicating that α -synuclein aggregation can be inhibited by the interaction of C-terminal tyrosines with fasudil. These two analyses point to binding specific binding sites of fasudil to α -synuclein.

In our work, we have built a MSM using the latent dimension of a deep learning method called VAE, to address how fasudil interacts with α -synuclein. An analysis of the macrostates as obtained from MSM, gives insights into how fasudil interacts with α -synuclein, in terms of transition probabilities among the states, thereby predicting which states are most favorable for binding.

Reviewer #2 (Public Review):

The manuscript by Menon et al describes a set of simulations of alpha-Synuclein (aSYN) and analyses of these and previous simulations in the presence of a small molecule.

While I agree with the authors that the questions addressed are interesting, I am not sure how much we learn from the present simulations and analyses. In parts, the manuscript reads more like an attempt to apply a whole range of tools rather than with a goal of answering any specific questions.

In this manuscript, we have employed a variational bayesian method, VAE, that uses variational inference to approximate the distribution of latent variable. Unlike conventional linear dimension reduction methods such as tICA (as provided in the SI), this method has

been found to be better (higher VAMP2 score) in capturing slow modes and thereby facilitate the study of long-time dynamics. Markov State Model was built on this lower dimension space which indicated the presence of three and six states for the apo and holo simulations respectively. The exclusivity of the states was justified by determining the backbone contact map and further mapping these states using a denoising CNN-VAE. The increase in the number of states in the presence of the small molecule was justified by calculating the entropy of the macrostates. The entropic contribution from water remained similar across all states, while for the protein in the holo ensemble, entropy was significantly modulated (either increased or decreased) compared to the apo state. In contrast, the entropy of the apo states showed much less modulation. This proves that an increase in the number of states is primarily an entropic effect caused by the small molecule. Finally we have compared the mean first passage time (MFPT) of other states to the most populated state, which reveals a strong correlation between transition time and the system's entropy for both apo and holo ensemble. However, the transition times (to the most populated state) are much lower for the holo ensemble, thereby suggesting that fasudil may potentially trap the protein conformations in the intermediate states, thereby slowing down α S in exploring the large conformational space and eventually slow down aggregation.

There's a lot going on in this paper, and I am not sure it is useful for the authors, readers or me to spell out all of my comments in detail. But here are at least some points that I found confusing/etc

Major concerns

p. 5 and elsewhere:

I lack a serious discussion of convergence and the statistics of the differences between the two sets of simulations. On p. 5 it is described how the authors ran multiple simulations of the ligandfree system for a total of 62 μ s; that is about 25 times less than for the ligand system. I acknowledge that running 1.5 ms is unfeasible, but at a bare minimum the authors should discuss and analyse the consequences for the relatively small amount of sampling. Here it is important to say that while 62 μ s may sound like a lot it is probably not enough to sample the relevant properties of a 140-residue long disordered protein.

As to referee 2's original comment on 'a lot going on in the manuscript', we believe that the complexity of the project demanded that this work needs to be dealt with an extensive analysis and objective machine learning approaches, instead of routine collective variable or traditional linear dimensional reduction techniques. This is what has been accomplished in this manuscript. For someone to get the gist of the work, the last paragraph of the introduction and first paragraph of conclusion provides a summary of the overall finding and investigation in the manuscript. First, a VAE-based machine learning approach demonstrates the modulation of free energy landscape of alpha-synuclein in presence of fasudil. Next, Markov State Model elucidates distinct binding competing states of alpha-synuclein in presence of the small-molecule drug. Then the MSM-derived metastable states of alpha-synuclein monomer are structurally characterized in presence of fasudil. Next we mapped the macrostates in apo and bound-state ensembles using denoising convolutional variational autoencoder, to ensure that these are mutually distinct. Next we show that fasudil exhibits conformation-dependent interactions with individual metastable states. Finally the investigation quantitatively brings out entropic signatures of small molecule binding.

We thank the reviewer for the question. For the apo simulations, we performed 1-4 μ s long simulations with 23 different starting structures and the ensemble amounted to an ensemble of ~62 μ s. In the Supplementary figures, we show analyses of how the starting structures used for apo simulations compare with the structure used to run the holo simulations as well as

comparison of the apo and holo ensembles in terms of structures features as Rg, Re, solvent accessible surface area (SASA) and secondary structure properties. This is updated in the manuscript on page 3,31- 33 and figures S1-S6, S25-S30.

Also, regarding the choice of starting structures, we chose multiple distinct conformations from a previous simulation of alpha synuclein monomer, reported in *Robustelli et. al, PNAS, 115 (21), E4758-E4766*. The Rg of the starting structures represent the entire distribution of Rg of the holo ensemble; from compact, intermediate to extended states. Importantly, the Rg distribution of the apo and holo ensembles are highly comparable and overlapping, indicating that the apo simulations, although of short timescale, have sampled the phase space locally around each starting conformation and thus covered the protein phase space as in the holo simulation. Similarly, other structural properties such as SASA, Re and secondary structure are comparable for the two ensembles. These analyses show that the local sampling across a variety of starting conformations has ensured sufficient sampling of the IDP phase space. This is updated in the manuscript on page 33-34 and figure S1, S25-S30.

p. 7:

The authors make it sound like a bad thing than some methods are deterministic. Why is that the case? What kind of uncertainty in the data do they mean? One can certainly have deterministic methods and still deal with uncertainty. Again, this seems like a somewhat ad hoc argument for the choice of the method used.

We appreciate the reviewer's comment. In this work, we have used a single VAE model to map the simulation of α S in its apo state and in the presence of fasudil, into two dimensions. If we had used an autoencoder, which is a deterministic model, we would have to train two independent models; one for the apo-state and one for fasudil. It would then be questionable to compare the two dimensions obtained from two different autoencoders as the model parameters are not shared.

VAE gives us this flexibility by not mapping it to a single point, but to a distribution, thereby encouraging it to learn more generalizable representation. The uncertainty is not in the data; but mapping a conformation (of the fasudil simulation) to a distribution would provide a new point for a similar structure (from the apo simulation).

p. 8:

The authors should make it clear (i) what the reconstruction loss and KL is calculated over and (ii) what the RMSD is calculated over.

(i) The reconstruction loss is calculated between the reconstructed and original pairwise distances, whereas the KL loss is calculated between the approximated posterior distribution and the prior distribution (for VAE it is a standard normal distribution)

(ii) The RMSE is the root mean square error between the original data and the reconstructed data.

(i) is updated on page 34 and (ii) is updated in the revised manuscript on page 8.

p. 9/figure 1:

The authors select a beta value that may be the minimum, but then is just below a big jump in the cross-validation error. Why does the error jump so much and isn't it slightly dangerous to pick a value close to such a large jump.

In this work, RMSE has been chosen as a metric to select the best VAE model. To do so, the β parameter (weighting factor for the KL loss) was varied. The β value was chosen as this had the minimum value.

This is updated on page 8.

p. 10:

Why was a 2-dimensional representation used in the VAE? What evidence do the authors have that the representation is meaningful? The authors state "The free energy landscape represents a large number of spatially close local minima representative of energetically competitive conformations inherent in aS" but they do not say what they mean by "spatially close". In the original space? If so, where is the evidence.

We thank the reviewer for the question. 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.

The reconstruction loss (which is the mean squared error between the input and the reconstructed data) is of the order of 10^{-4} . Also, the MSM built on the latent space of VAE is able to identify states that are distinct for both apo and holo simulations, which ensures that the latent space representation is meaningful.

We have also trained a model with 4 neurons in the latent space and built an MSM. The implied timescales indicate the presence of six states which is consistent with the model with two latent dimensions.

This is updated in the manuscript on page 13 and figure S14-S15.

No, not spatially close in the original space, but in the reduced two dimensional latent space.

p. 10:

It is not clear from the text whether the VAEs are the same for both aSYN and aSYN-Fasudil. I assume they are. Given that the Fasudil dataset is 25x larger, presumably the VAE is mostly driven by that system. Is the VAE an equally good representation of both systems?

Yes, the same model is used for both aSYN and aSYN-Fasudil ensemble.

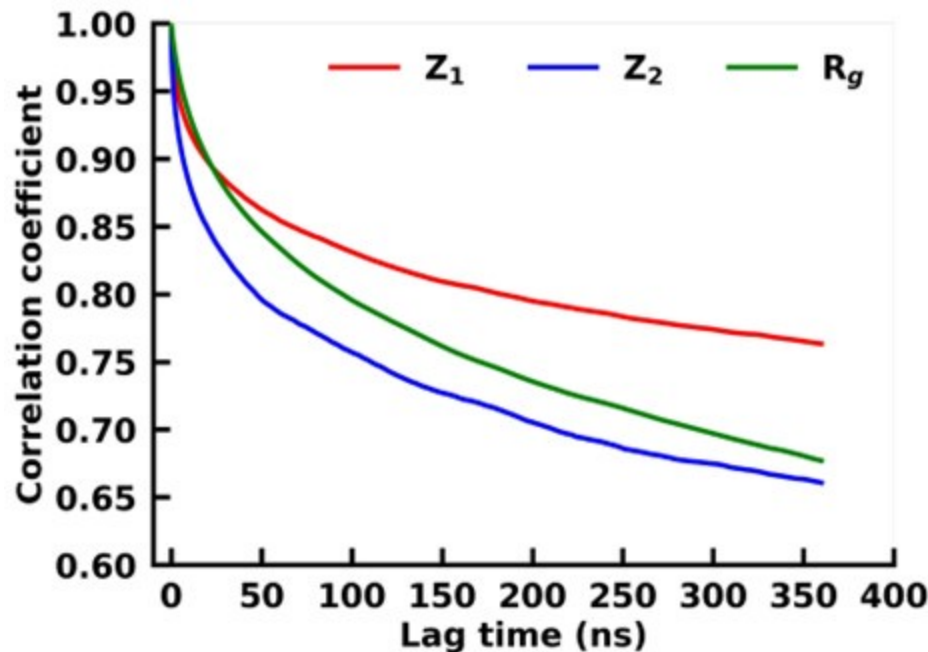
The states obtained from the MSM of the aSyn ensemble are distinct when their Ca contact maps are analyzed. So we think it is a good representation for this system.

p. 10/11:

Do the authors have any evidence that the latent space representation preserves relevant kinetic properties? This is a key point because the entire analysis is built on this. The choice of using z1 and z2 to build the MSM seems somewhat ad hoc. What does the autocorrelation functions of Z1 and Z2 look like? Are they related to dynamics of some key structural properties like Rg or transient helical structure.

Autocorrelation of z1 and z2 of the latent space of VAE and the radius of gyration for asyn-fasudil simulation.

Author response image 5.



We find that z_1 of VAE has a much slower decay as compared to R_g . This indicates that it is much better in capturing long-time-scale dynamics as compared to R_g .

p. 11:

What's the argument for not building an MSM with states shared for aSYN +- Fasudil?

We have built two different markov state models for two aSYN simulation in its apo state and in the presence of ligand. Mixing the two latent spaces to build one MSM would give incorrect transition timescales among the states as these are independent simulations.

p. 12:

Fig. 3b/c show quite clearly that the implied timescales are not converged at the chosen lag time (incidentally, it would have been useful with showing the timescales in physical time). The CK test is stated to be validated with "reasonable accuracy", though it is unclear what that means.

We have mentioned the physical timescales in the main manuscript (Page no. 38), which is 36 and 32 ns for apo and holo simulations, respectively. We used "reasonable accuracy" in the context of the Chapman-Kolmogorov test. We note that for the ligand simulations, the estimated and predicted models are in excellent agreement as compared to some of the transitions in the apo state. This good agreement implies that the model has reached Markovianity and the timescales have converged.

The CK test is updated in the manuscript on page 12.

p. 12:

In Fig. 3d, what are the authors bootstrapping over? What are the errors if the authors analyse sampling noise (e.g. bootstrap over simulation blocks)?

For bootstrapping, we randomly deleted a part of the simulation (simulation block) and rebuilt the MSM with this reduced dataset. We repeated this 10 times and reported the average value of the population and the transition timescales over the 10 iterations.

p. 13:

I appreciate that the authors build an MSM using only a subset of the fasudil simulations. Here, it would be important that this analysis includes the entire workflow so that the VAE is also rebuilt from scratch. Is that the case?

The VAE model was trained over data points of the ligand simulation sampled at every 9 ns starting from time $t=0$, for the entire 1.5 ms. We did not train it for the subset of the fasudil simulation, but rather used the trained VAE model to get the latent space of the 60 μ s of the fasudil simulation to build the MSM. Additionally, we have compared the distributions of R_g for this simulation block with the apo ensemble and found good agreement among them.

R_g distribution is updated in the manuscript on page 13 and see figure S10-S11.

p. 18:

I don't understand the goal of building the CVAE and DCVAE. Am I correct that the authors are building a complex ML model using only 3/6 input images? What is the goal of this analysis. As it stands, it reads a bit like simply wanting to apply some ML method to the data. Incidentally, the table in Fig. 6C is somewhat intransparent.

We appreciate the reviewer's valid question. The ensemble averaged contact map of the macrostates of aSyn in apo state and in the presence of ligand posed us a challenge in finding contacts that are exclusive to each state. Since VAEs are excellent in finding patterns, we employed a convolutional VAE (typically used for images). However, owing to the few number of contact maps, the model overfitted and to prevent this, we added noise to the data. A visual inspection of the ensemble averaged contact map, especially for IDPs is difficult and this lower dimensional space will give us a preliminary idea of how each macrostate is different from every other. The table in Fig. 6C provides scores for the denoised contact maps (SSIM and PSNR scores). An SSIM score above 0.9 and PSNR score between 20-48 indicates that the reconstruction of the contact map is of good quality.

p. 22:

"Our results indicate that the interaction of fasudil with aS residues governs the structural features of the protein."

What results indicate this?

By building a Markov State Model and comparing them across the apo and holo ensembles, we showed the interaction of fasudil with aSyn leads to the population of more states (than apo). In these states, we observe that fasudil interacts with aSyn in different regions as shown by the protein-ligand contact map as shown in figure 7. Also, the contact maps and the extent of secondary structure of the six states are distinct across the states. The location and extent of the helix and sheet-like character in the ensemble of the six macrostates as shown in figure S16-S17. Based on these observations, we state that the interaction of the small molecule favors the population of new aSyn states that are distinct in their structural features.

p. 23:

The authors should add some (realistic) errors to the entropy values quoted. Fig. 8 have some error bars, though they seem unrealistically small. Also, is the water value quoted from the same force field and conditions as for the simulations?

The error values are the standard deviations that are provided by the PDB2ENTROPY package. Yes, the water value is from the same force field and conditions for the simulations are the same as reported in the section “Entropy of water”

p. 23:

Has PDB2ENTROPY been validated for use with disordered proteins?

Yes, it has been used in the following paper studying liquid-liquid phase separation of an IDP.

This paper has also been cited in the manuscript (reference 66).

“Thermodynamic forces from protein and water govern condensate formation of an intrinsically disordered protein domain” by Saumyak Mukherjee & Lars V. Schäfer, *Nature Communications* volume 14, Article number: 5892 (2023) <https://doi.org/10.1038/s41467-023-41586-y>

p. 23/24:

It would be useful to compare (i) the free energies of the states (from their populations), (ii) the entropies (as calculated) and (iii) the enthalpies (as calculated e.g. as the average force field energy). Do they match up?

Our analysis stems from previous studies where enthalpy driven drug design has not led to significant advances in drug design, particularly for IDPs. In the presence of the drug/ligand, the protein may be able to explore a larger conformational space and hence an increase in the number of states accessible by the protein, which we found by building Markov State Model using the latent space of VAE. The entropy of the protein is calculated based on the torsional degrees of freedom relative to the random distribution (the protein with the most random configuration).

p. 31:

It is unclear which previous simulation the new aSYN simulations were launched from. What is the size of the box used?

The starting conformations for the new aSYN simulations were randomly chosen from a previously reported 73 μ s simulation in Robustelli et. al. (PNAS, 115 (21), E4758-E4766).

Box size for the 23 simulation has been added to the supplemental information in Table S1.

Reviewer #3 (Public Review):

Summary:

In this manuscript Menon, Adhikari, and Mondal analyze explicit solvent molecular dynamics (MD) computer simulations of the intrinsically disordered protein (IDP) alpha-synuclein in the presence and absence of a small molecule ligand, Fasudil, previously demonstrated to bind alpha-synuclein by NMR spectroscopy without inducing folding into more ordered structures. In order to provide insight into the binding mechanism of Fasudil the authors analyze an unbiased 1500us MD simulation of alpha-synuclein in the presence of Fasudil previously reported by Robustelli et.al. (Journal of the American

Chemical Society, 144(6), pp.2501-2510). The authors compare this simulation to a very different set of apo simulations: 23 separate 1-4us simulations of alphasynuclein seeded from different apo conformations taken from another previously reported by Robustelli et. al. (PNAS, 115 (21), E4758-E4766), for a total of ~62us.

To analyze the conformational space of alpha-synuclein - the authors employ a variational autoencoder (VAE) to reduce the dimensionality of Ca-Ca pairwise distances to 2 dimensions, and use the latent space projection of the VAE to build Markov state Models. The authors utilize kmeans clustering to cluster the sampled states of alpha-synuclein in each condition into 180 microstates on the VAE latent space. They then coarse grain these 180 microstates into a 3macrostate model for apo alpha-synuclein and a 6-macrostate model for alpha-synuclein in the presence of fasudil using the PCCA+ course graining method. Few details are provided to explain the hyperparameters used for PCCA+ coarse graining and the rationale for selecting the final number of macrostates.

The authors analyze the properties of each of the alpha-synuclein macrostates from their final MSMs - examining intramolecular contacts, secondary structure propensities, and in the case of alpha-synuclein:Fasudil holo simulations - the contact probabilities between Fasudil and alphasynuclein residues.

The authors utilize an additional variational autoencoder (a denoising convolutional VAE) to compare denoised contact maps of each macrostate, and project onto an additional latent space. The authors conclude that their apo and holo simulations are sampling distinct regions of the conformational space of alpha-synuclein projected on the denoising convolutional VAE latent space.

Finally, the authors calculate water entropy and protein conformational entropy for each microstate. To facilitate water entropy calculations - the author's take a single structure from each macrostate - and ran a 20ps simulation at a finer timestep (4 femtoseconds) using a previously published method (DoSPT), which computes thermodynamic properties of water from MD simulations using autocorrelation functions of water velocities. The authors report that water entropy calculated from these individual 20ps simulations is very similar.

For each macrostate the authors compute protein conformational entropy using a previously published Maximum Information Spanning tree approach based on torsion angle distributions - and observe that the estimated protein conformational entropy is substantially more negative for the macrostates of the holo ensemble.

The authors calculate mean first passage times from their Markov state models and report a strong correlation between the protein conformational entropy of each state and the mean first passage time from each state to the highest populated state.

As the authors observe the conformational entropy estimated from macrostates of the holo alphasynuclein:Fasudil is greater than those estimated from macrostates of the apo holo alphasynuclein macrostates - they suggest that the driving force of Fasudil binding is an increase in the conformational entropy of alpha-synuclein. No consideration/quantification of the enthalpy of alpha-synuclein Fasudil binding is presented.

Strengths:

The author's utilize MD simulations run with an appropriate force field for IDPs (a99SB-disp and a99SB-disp water (Robustelli et. al, PNAS, 115 (21), E4758-E4766) - which has previously been used to perform MD simulations of alpha-synuclein that have been validated with extensive NMR data.

The contact probability between Fasudil and each alpha-synuclein residue observed in the previously performed 1500us MD simulation of alpha-synuclein in the presence of Fasudil (Robustelli et. al., Journal of the American Chemical Society, 144(6), pp.2501-2510) was previously found to be in good agreement with experimental NMR chemical shift perturbations upon Fasudil binding - suggesting that this simulation is a reasonable choice for understanding IDP:small molecule interactions.

Weaknesses:

Major Weakness 1: Simulations of apo alpha-synuclein and holo simulations of alpha-synuclein and fasudil are not comparable.

The most robust way to determine how presence of Fasudil affects the conformational ensemble of alpha-synuclein conclusions is to run apo and holo simulations of the same length from the same starting structures using the same simulation parameters.

The 23 1-4 us independent simulations of apo alpha-synuclein and the long unbiased 1500us alpha-synuclein in the presence of fasudil are not directly comparable. The starting structures of simulations used to build a Markov state model to describe apo alpha-synuclein were taken from a previously reported 73us MD simulation of alpha-synuclein run with the a99SB-disp force field and water model) with 100mM NaCl, (Robustelli et. al, PNAS, 115 (21), E4758-E4766). As the holo simulation of alpha-synuclein and Fasudil was run in 50mM NaCl, snapshots from the original apo alpha-synuclein simulation were resolvated with 50mM NaCl - and new simulations were run.

No justification is offered for how starting structures were selected. We have no sense of the conformational variability of the starting structures selected and no sense of how these conformations compare to the alpha-synuclein conformations sampled in the holo simulation in terms of standard structural descriptors such as tertiary contacts, secondary structure, radius of gyration (Rg), solvent exposed surface area etc. (we only see a comparison of projections on an uninterpretable non-linear latent-space and average contact maps). Additionally, 1-4 us is a relatively short timescale for a simulation of a 140 residue IDP- and one is unlikely to see substantial evolution for many structural properties of interest (ie. secondary structure, radius of gyration, tertiary contacts) in simulations this short. Without any information about the conformational space sample in the 23 apo simulations (aside from a projection on an uninterpretable latent space)- we have no way to determine if we observe transitions between distinct states in these short simulations, and therefore if it is possible the construct a meaningful MSM from these simulations.

If the structures used for apo simulations are on average more compact or contain more tertiary contacts - then it is unsurprising that in short independent simulations they sample a smaller region of conformational space. Similarly, if the starting structures have similar dimensions - but we only observe extremely local sampling around starting structures in apo simulations in the short simulation times - it would also not be surprising that we sample a smaller amount of conformational space. By only presenting comparisons of conformational states on an uninformative VAE latent space - it is not possible for a reader to ask simple questions about how the conformational ensembles compare.

It is noted that the authors attempt to address questions about sampling by building an MSM of single contiguous 60us portion of the holo simulation of alpha-synuclein and Fasudil - noting that:

"the MSM built using lesser data (and same amount of data as in water) also indicated the presence of six states of alphaS in presence of fasudil, as was observed in the MSM of

the full trajectory. Together, this exercise invalidates the sampling argument and suggests that the increase in the number of metastable macrostates of alphaS in fasudil solution relative to that in water is a direct outcome of the interaction of alphaS with the small molecule."

However, the authors present no data to support this assertion - and readers have no sense of how the conformational space sampled in this portion of the trajectory compares to the conformational space sampled in the independent apo simulations or the full holo simulation. As the analyzed 60us portion of the holo trajectory may have no overlap with conformational space sampled in the independent apo simulations - it is unclear if this control provides any information. There is no quantification of the conformational entropy of the 6 states obtained from this portion of the holo trajectory or the full conformational space sampled. No information is presented to determine if we observe similar states in the shorter portion of the holo trajectory. Furthermore - as the authors provide almost no justification for the criteria used to select of the final number of macrostates for any of the MSMs reported in this work- and the number of macrostates is effectively a free parameter in the PCCA+ method, arriving at an MSM with 6 macrostates does not convey any information about the conformational entropy of alpha-synuclein in the presence or absence of ligands. Indeed - the implied timescale plot for 60us holo MSM (Figure S2) - shows that at least 10 processes are resolved in the 120 microstate model - and there is no information to provided explaining/justifying how a final 6-macrostate model was determined. The authors also do not project the conformations sampled in this sub- trajectory onto the latent space of the final VAE.

One certainly expects that an MSM built with 1/20th of the simulation data should have substantial differences from an MSM built from the full trajectory - so failing additional information and hyperparameter justification - one wonders if the emergence of a 6-state model could be the direct result of hardcoded VAE and MSM construction hyperparameter choices.

Required Controls For Supporting the Conclusions of the Study: The authors should initiate apo and holo simulations from the same starting structures - using the same simulation software and parameters. This could be done by adding a Fasudil ligand to the apo structures - or by removing the Fasudil ligand from a subset of holo structures. This would enable them to make apples-to-apples comparisons about the effect of Fasudil on alpha-synuclein conformational space.

Failing to add direct apples-to-apples comparisons, which would be required to truly support the studies conclusions, the authors should at least compare the conformational space sampled in the independent apo simulations and holo simulations using standard interpretable IDP order parameters (ie. Rg, end-to-end distance, secondary structure order parameters) and/or principal components from PCA or tICA obtained from the holo simulation. The authors should quantify the number of transitions observed between conformational states in their apo simulations. The authors could also perform more appropriate holo controls, without additional calculations, by taking batches of a similar number of short 1-4us segments of simulations used to compute the apo MSMs and examining how the parameters/macrostates of the holo MSMs vary with the input with random selections.

In case of IDPs, one should not bias the simulation by starting from identical structures, as IDP does not have a defined structure and the starting configuration has little significance. It is the microenvironment that matters most. As for the choice of simulation software and parameters, we have used the same force field that was used in the holo simulation at the same temperature and same salt concentration. We have performed multiple independent simulations that have varying structural signatures such as Rg, SASA and secondary structure

content. In fact, the starting structure for apo simulations covered the entire span of the R_g distribution of holo simulation, including the starting structure of the holo simulation. The simulations are unbiased w.r.t the starting structure. Although the fasudil simulation was run for 1.5 ms, we should also understand that it is difficult to run a millisecond range of simulation in reasonable time from a single starting structure. It is exactly for this reason that we start with different structures so that we do not bias ourselves and sample every possible conformation.

We have updated the manuscript on page 33-34 and figure S1, S25-S30.

Considering the computational expense for simulating 1.5 ms timescale of a 140-residue IDP, we generated an ensemble from multiple short runs amounting to $\sim 60 \mu\text{s}$. The premise of this investigation is a widely popular method, Markov State Models (MSMs) that can be used to estimate long timescale kinetics and stationary populations of metastable states built from ensembles of short simulations. We have also demonstrated that comparable to the apo data, when we build an MSM for asyn-fasudil (holo) using $60 \mu\text{s}$ simulation block, the implied timescales (ITS) plot shows identical number of metastable states as for the 1.5 ms data.

An intrinsically disordered protein (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 chose randomly 23 different conformations from the $73 \mu\text{s}$ long simulation of α -synuclein monomer reported in *Robustelli et. al, PNAS, 115 (21), E4758-E4766*. Based on the reviewer's comment on providing a justification for choice of the starting structures for apo simulations, we provide a compilation of figures below showing comparison of standard conformational properties of the chosen initial structures for apo simulations with the starting structure of the long holo simulation; we have also provided comparative analyses of the apo ($\sim 60 \mu\text{s}$) and holo ensemble (1.5 ms) properties.

Figure S1 compares the R_g of the apo and holo ensembles of $\sim 60 \mu\text{s}$ and 1.5 ms, respectively. The distributions are majorly overlapping, indicating that the apo ensemble is comparable to the holo ensemble, in terms of the extent of compaction of the conformations. In Figure 1, we have also marked the R_g values corresponding to the starting structures used to seed the apo simulations. It is evident that the 23 starting conformations chosen represent the whole range of the R_g space that is sampled in the holo ensemble. Therefore, while the apo simulations are relatively short (1-4 μs), the local sampling of these multiple starting conformations of variable compaction (R_g) ensures that the phase space is efficiently sampled and the resulting ensemble is comparable to the holo ensemble. Furthermore, the implementation of MSM on such an ensemble can be efficiently used to identify metastable states and the long timescale transitions happening between them

Another property that is proportional to R_g is the end-to-end distance of the protein conformations. Figure S2 shows that the distribution of this property in the apo and holo ensembles are highly similar.

Figure S3 depicts another fundamental structural descriptor i.e. solvent accessible surface area (SASA) that indicates the extent of folding and the exposure of the residues. The apo ensemble only shows a minimal shift in the distribution towards higher SASA values. The distributions of the two ensembles largely overlap.

In Figure S25, we have provided the root mean square deviation (RMSD) of the starting structures used in the apo simulations with the structure used to start the long simulation with fasudil. The RMSD values range from 1.6 to 3 nm, indicating that the starting structures

used are highly variable. This is justifiable for IDPs since they are not identified by a single, fixed structure, but rather by an array of different conformations.

Figures S26-S28 show the fraction of the secondary structure elements i.e. helix, beta and coil in the starting structures of apo and holo simulations. All the conformations are mostly disordered in nature with the greatest extent of coil content. The helix content ranges from 3-10 % while sheet content varies from 3-15 % in the initial simulation structures.

Figures S4-s6 represent the residue-wise percentage of secondary structure elements (helix, beta and coil) in the apo and holo ensembles. It is evident that the extent of secondary structure is comparable in the two ensembles.

The above analyses comparing distributions of several structural features clearly indicate that the apo simulations we performed from different starting structures have effectively sampled the phase space as the single long simulation of the holo system.

We have discussed the above in the manuscript: Computational Methods section, Page 33-34.

The above VAMP score analyses (Figures S7 and S8) has been now presented in the manuscript: Results and Discussion (Page 8)

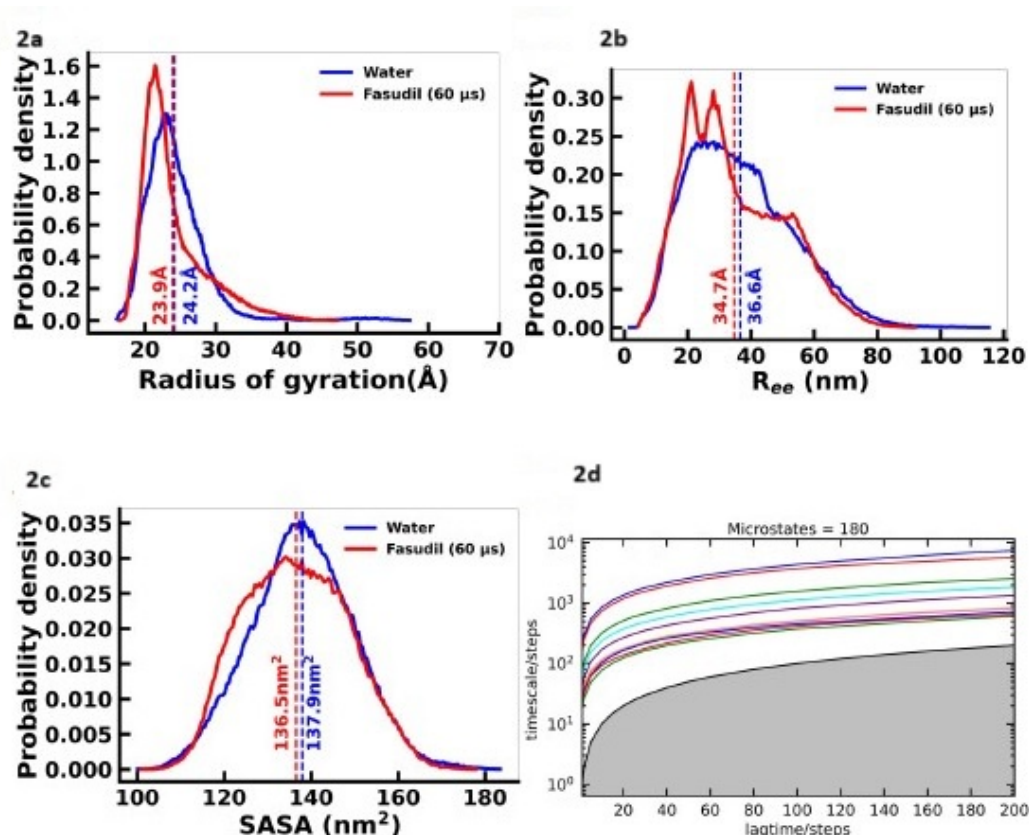
Building the MSM

While building the MSM, we iteratively varied the hyperparameters to build a reasonable model. In this process, we explored different values of the number of clusters, maximum number of iterations, tolerance, stride, metric, seed, chunk size and initialization methods. There is no possible way to perform an optimization on the choice of the above hyperparameters using gradient descent methods, as no convergence would be guaranteed. The parameters were tuned carefully so that we get the best possible implied timescales of the system. The quality of the MSM was further validated using the Chapman-Kolmogorov (CK) test on a state-by-state basis i.e by considering the transitions between each pair of the metastable states. In addition, we have built the contact maps to show that the states are mutually exclusive. This is also justified by the latent space of denoising convolutional variational autoencoders.

We have compared the conformational space in the independent apo and holo simulations for Rg, Ree, SASA and secondary structure. As for PCA/TICA, we have computed the VAMP-2 score for TICA and found out to be low as compared to VAE. In fact, neural networks have been shown previously as a better dimension reduction technique due to its non-linearity over linear methods such as PCA or TICA.

Author response image 6.

Distribution of (a) Rg, (b) Ree, (c) SASA and of the apo ensemble and a 60 μ s slice of the holo simulation trajectory. (d) ITS plot of the 60 μ s chunk.



First, someone familiar with MSM should understand that the basic philosophy of MSM is not the requirement of long simulation trajectories, which would defeat the purpose of its usage. Rather as motivated by Noe and coworkers in seminal PNAS (vol. 106, page 9011, year 2009) paper, MSM plays an important role in inferring long-time scale equilibrium properties by using significantly short-length scale non-equilibrium trajectories.

Considering the difference in the size of the ensembles in the apo and holo simulations, we verified how different is the MSM built using 60 μ s slice of the data from the 1.5 ms holo simulation in terms of the number of metastable states identified by the model. For this, we considered 60 μ s data beginning from 966 μ s - 1026 μ s. First, we compared the gross structural properties of these datasets. Author response image 6a-c compares the distributions of Rg, R_{ee} and SASA. The distributions show that the apo and holo simulations are very similar with respect to these standard properties of protein conformations.

We built the MSM for this 60 μ s data of the holo ensemble from the reduced data obtained from the same VAE model. We would like to clarify that the hyperparameters of the model are not hardcoded but rather carefully fine-tuned to obtain a good model that performs good kinetic discretization of the underlying macrostates. The implied timescale plot of this new MSM shows distinct timescales corresponding to six macrostates. This led us to conclude that the six-state model is robust despite the differences in the ensemble size. The implied timescale is shown in Author response image 6d.

The above analyses in Author response image 6 are presented in Results and Discussion, Page 13.

Major Weakness 2: There is little justification of how the hyperparameters MSMs were selected. It is unclear if the results of the study depend on arbitrary hyperparameter selections such as the final number of macrostates in each model.

It is unclear what criteria were used to determine the appropriate number of microstates and macrostates for each MSM. Most importantly - as all analyses of water entropy and conformational entropy are restricted to the final macrostates - the criteria used to select the final number of macrostates with the PCCA+ are extremely important to the results of the conclusions of the study. From examining the ITS plots in Figure 3 - it seems both MSMs show the same number of resolved processes (at least 11) - suggesting that a 10-state model could be appropriate for both systems. If one were to simply select a large number of macrostates for the 20x longer holo simulation - do these states converge to the same conformational entropy as the states seen in the short apo simulations? Is there some MSM quality metric used to determine what number of macrostates is more appropriate?

Required Controls For Supporting the Conclusions of the Study: The authors should specify the criteria used to determine the appropriate number of microstates and macrostates for their MSMs and present controls that demonstrate that the conformational entropies calculated for their final states are not simply a function of the ratio of the number macrostates chosen to represent very disparate amounts of conformational sampling.

VAMP-2 score was used to determine the number of microstates. We have calculated the VAMP2 score by varying the number of microstates, ranging from 10 to 220. We find that the VAMP-2 score has saturated at a higher number of microstates for both apo and holo simulations.

The number of macrostates were determined by the gap between the lines of the Implied timescales plot followed by a CK test (shown in figure S1). Since we plotted the first 10 slowest timescales, the implied timescales show 10 timescales and this is not an indicator of the number of macrostates. The macrostates are separated by distinct gaps in the timescales and do not merge as seen beyond 5 timescales in the plot. The timescales, when leveled off and distinct, indicate that the system has well defined metastable states and the MSM is accurate in identifying the macrostates. We find this to be three and six for the apo and holo simulations from the corresponding implied timescales.

The above is discussed in Computational Methods, Page 37-38.

Major Weakness 3: The use of variational autoencoders (VAEs) obscures insights into the underlying conformational ensembles of apo and holo alpha-synuclein rather than providing new ones

No rationale is offered for the selection of the VAE architecture or hyperparameters used to reduce the dimensionality of alpha-synuclein conformational space.

It is not clear the VAEs employed in this study are providing any new insight into the conformational ensembles and binding mechanisms of Fasudil to alpha-synuclein, or if the underlying latent space of the VAEs are more informative or kinetically meaningful than standard linear dimensionality reduction techniques like PCA and tICA. The initial VAE is used to reduce the dimensionality of alpha-synuclein conformational ensembles to 2 degrees of freedom - but it is unclear if this projection is structurally or kinetically meaningful. It is not clear why the authors choice to use a 2-dimeinsional projection instead of a higher number of dimensions to build their MSMs. Can they produce a more

kinetically and structurally meaningful model using a higher dimensional VAE latent space?

Additionally - it is not clear what insights are provided by the Denoising Convolutional Variational Autoencoder. The authors appear to be noising-and-denoising the contact maps of each macrostate, and then projecting the denoised values onto a new latent space - and commenting that they are different. Does this provide additional insight that looking at the contact maps in Figures 4&5 does not? Is this more informative than examining the distribution of the Radii of gyration or the secondary structure propensities of each ensemble? It is not clear what insight this analysis adds to the manuscript.

Suggested controls to improve the study: The authors should project interpretable IDP structural descriptors (ie. secondary structure, radius of gyration, secondary structure content, # of intramolecular contacts, # of intermolecular contacts between alpha-synuclein and Fasudil) onto this latent space to illustrate if any of these properties are meaningful separated by the VAE projection. The authors should compare these projections, and MSMs built from these projections, to projections and MSMs built from projections using standard linear dimensionality projection techniques like PCA and tICA.

We have already pointed out the IDP structural parameters for the first question.

In case of VAE, the latent space captures the underlying pattern of the higher dimensional data. A non-linear projection using VAE has shown to have a higher VAMP-2 score over linear dimension reduction methods such as tICA. The latent space of VAE was then used to build the MSM, in order to get the macrostates and also the transition timescales among them. We can project the data onto a higher dimension, but the goal is to reduce it to lower dimensions where it will be easier to interpret. Higher number dimensions would also risk overfitting; and the model, instead of learning the pattern, it may simply memorize the data. The training and validation loss curve from VAE has reached the order of 10^{-4} thereby indicating good reconstruction of the original data.

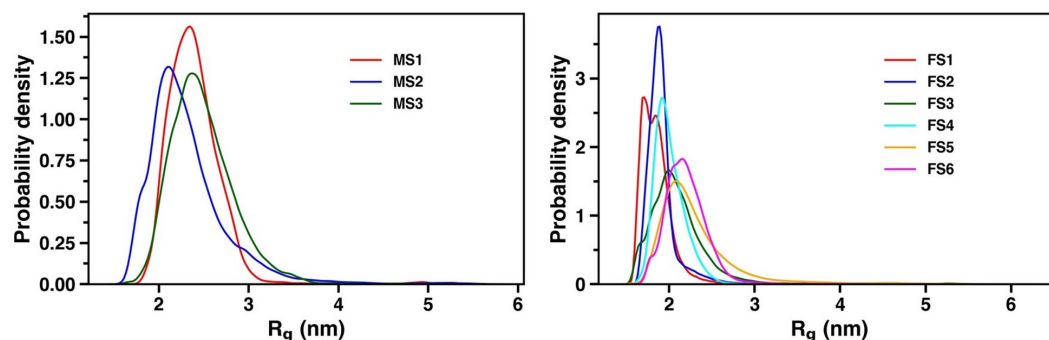
As for dimension reduction using tICA, the VAMP-2 score confirms that our VAE model performs better than tICA. This manuscript uses deep neural networks to understand the structural and kinetic process of IDP and small molecule interaction. Dimension reduction using tICA would give different reaction coordinates and MSM built using the projected data of tICA will not be one-to-one comparable with that obtained from VAE.

We had to perform noising, as we had only 9 contact maps. This led to overfitting of the CVAE model. To overcome this problem, we have introduced white noise to our data, so as to prevent the model from overfitting. The objective of the DCVAE model was to see how distinct these contact maps are based on their locations on a lower dimensional space. A visual inspection of the ensemble averaged contact map, especially for IDPs is much more difficult as compared to folded proteins. So, even before computing the Rg, Re, SASA or secondary structure, this lower dimensional space will give us a preliminary idea of how each macrostate is different from every other.

As for the distribution of Rg, we have plotted it in Author response image 7. The residue-wise percentage secondary structure is plotted in figure S4-S6 for the holo and apo simulation respectively.

Author response image 7.

Distribution of radius of gyration for the three and six macrostates in the apo and holo simulation respectively.



As for training a model with a higher number of latent dimensions, we have retrained a VAE model with four dimensions in the latent space. The loss was of the order of 10^{-4} . We built a MSM with the appropriate number of microstates and found the presence of six macrostates as evident from the ITS plot as shown in Figure S14 and S15.

This data is presented in Results and Discussion, Page 13

Major Weakness 4: The MSMs produced in this study have large discrepancies with MSMs previously produced on the same dataset by the same authors that are not discussed.

Previously - two of the authors of this manuscript (Menon and Mondal) authored a preprint titled "Small molecule modulates α -synuclein conformation and its oligomerization via Entropy Expansion" (<https://www.biorxiv.org/content/10.1101/2022.10.20.513005v1.full>) that analyzed the same 1500us holo simulation of alpha-synuclein binding Fasudil. In this study - they utilized the variational approach to Markov processes (VAMP) to build an MSM using a 1D order parameter as input (the radius of gyration), first discretizing the conformational space into 300 microstates before similarly building a 6 macrostate model. From examining the contact maps and secondary structure propensities of the holo MSMs from the current study and the previous study- some of the macrostates appear similar, however there appear to be orders of magnitude differences in the timescales of conformational transitions between the two models. The timescales of conformational transitions in the previous MSM are on the order of 10s of microseconds, while the timescales of transitions in this manuscript are 100s-1000s microseconds. In the previous manuscript, a 3 state MSM is built from an apo α -synuclein obtained from a continuous 73ms unbiased MD simulation of alpha-synuclein run at a different salt concentration (100mM) and an additional 33 ms of shorter simulations. The apo MSM from the previous study similarly reports very fast timescales of transitions between apo states (on the order ~ 1 ms) - while the MSM reported in the current study (Figure 9) are on the order of 10s-100s of microseconds).

These discrepancies raise further concerns that the properties of the MSMs built on these systems are extremely sensitive to the chosen projection methods and MSM modeling choices and hyperparameters, and that neither model may be an accurate description of the true underlying dynamics

Suggestions to improve the study: The authors should discuss the discrepancies with the MSMs reported in their previous studies.

In the previous preprint, the radius of gyration was used as the collective variable to build the MSM. In this manuscript, we have used a much more general collective variable, reduced pairwise distance using VAE. Firstly, the collective variables used to build the model in the two works are different. Secondly, for the 73 μ s apo simulation in the previous manuscript,

the salt concentration used was 100 mM, but in this work, we have used a salt concentration of 50 mM, same as the salt concentration used in the holo simulations. Since the two simulation conditions are different with respect to salt concentration, the conformational space sampled in these conditions will be different and this will be reflected in the nature/features of the metastable states and the associated transition kinetics. Thirdly, the lag time at which the MSM was built was 3.6 ns in the previous manuscript, whereas, in this work we have used 32 ns. This is already off by a factor of 10. So the order of timescales have also changed. Thus, changes in the collective variable and change in the lag time at which the system reaches Markovianity is different. Hence, the timescales of transition among the macrostates are also different. Because of these differences, it would not be correct to compare the results that we would get from the two investigations.

Recommendations for the authors:

Reviewer #1 (Recommendations For The Authors):

To highlight the role of the entropic expansion mechanism, I would suggest modifying the title to capture this result, for example: "An Integrated Machine Learning Approach Delineates an Entropic Expansion Mechanism for the Binding of a Small Molecule to α -Synuclein".

We have changed the title as suggested by the reviewer.

To my knowledge the binding of fasudil to alpha-synuclein has been shown in the simulations by Robustelli et al (JACS 2022), but the experimental evidence is less clear cut. If an experimental binding affinity and the effect on alpha-synuclein aggregation have been measured, they should be reported.

Reviewer #2 (Recommendations For The Authors):

We thank the reviewer for the careful evaluation of our manuscript and providing comments and questions that we have attempted to address and incorporate.

Minor

Abstract:

In "which is able to statistically distinguish fuzzy ensemble", what does the word "statistically" mean in this context? Do the authors present evidence that the two ensembles are statistically different, and if so in what ways?

We have analyzed the apo and holo ensembles of aSyn using the framework of Markov State Models, which provides the stationary populations of the states that the model identifies. For this reason, we have used 'which is able to statistically distinguish fuzzy ensemble' as we compare and contrast the metastable states that we resolve using MSM. The MSM provides metastable states which are identified through statistical analysis of the transitions between states (transition probability matrix). We characterize their structural features to distinguish them which gives a meaningful interpretation of the fuzzy ensemble.

Abstract:

What does "entropic ordering" mean?

We thank the reviewer for pointing this out. Here, we mean that the presence of the small molecule only affects the protein backbone entropy while the entropy of water is not affected in the simulations with fasudil. We will rewrite this more clearly in the abstract.

The changed sentence is as follows:

“A thermodynamic analysis indicates that small-molecule modulates the structural repertoire of α S by tuning protein backbone entropy, however the entropy of the water remains unperturbed.”

Abstract:

What does "offering insights into entropic modulation" mean?

In this investigation, we first discretized the ensemble of a small-molecule binding/interacting with a disordered α Syn into the underlying metastable states, followed by characterisation of these identified states. As small molecule interactions can affect the overall entropy of the IDP, we estimated the said effect of fasudil binding on α Syn. We find that small molecule binding effect is manifested in the protein backbone entropy and the solvent entropy is not affected. Through this work, we highlight these insights into the modulatory effect that fasudil brings about in the entropy of the system (*entropic modulation*).

p. 3/4:

When the authors write "However, a routine comparison of monomeric α S ensemble... ensemble" it is unclear whether they are referring to previous work (they only cite a paper with simulations of "apo" α Syn, and if so which. Do they mean Ref 32? Also, the word "routine" sounds odd in this context.

We thank the author for pointing this out. We compared the ensemble properties (such as the distributions of the radius of gyration, end-to-end distance, solvent accessible surface area, secondary structure properties) of α -synuclein monomer that we generated in neat water and the ensemble of α -synuclein in the presence of the small molecule fasudil that is reported in Robustelli et.al. (Journal of the American Chemical Society, 144(6), pp.2501-2510). We have now modified this sentence in the main manuscript as follows: (Page no 3)

“However, comparison of the global and local structural features of the α S 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.”

p. 4:

Regarding "Integrative approaches are therefore gaining importance in IDP studies", these kinds of integrative approaches have been used for 20 years for studies of IDPs (with increasing sophistication and success), so I think "gaining" is somewhat of a stretch.

We thank the reviewer for this comment. We agree with the reviewer and have now changed this sentence as follows:

“Integrative approaches have been exploited in studying IDPs as well as small-molecule binding to IDPs.”

p. 5:

What does "large scale" mean in "This study showed no large-scale differences between the bound and unbound states of α S"? Do the authors mean substantially/significantly different, or differences on a large (length) scale?

Here, we refer to the study of small molecule (fasudil) binding study to α -synuclein reported in Robustelli et.al. (Journal of the American Chemical Society, 144(6), pp.2501-2510). In this study, the authors report no substantial (“large scale”) differences in the conformational ensembles of α -synuclein in the bound and unbound states of fasudil such as the backbone conformation distributions.

p. 6:

The authors write "In a clear departure from the classical view of ligand binding to a folded globular protein, the visual change in α S ensemble due to the presence of small molecule is not so strikingly apparent." I don't understand this. Normally, there is very little difference between apo and holo protein structures for folded proteins, so I don't understand the "in a clear departure" part. This seems like a strawman. Of course, for folded proteins one can generally see the ligand bound, but here the authors are talking about the protein.

In case of folded proteins, the overall tertiary structure of the protein remains mostly the same upon binding of the ligand. Structural changes are localized in nature and primarily around the binding site. However, in case of α Syn, binding of fasudil is transient and not as strong as seen for folded proteins. “Clear departure” refers to the fact that for α Syn, binding of fasudil is more subtle and dispersed across the ensemble of conformations rather than localized changes as in case of folded proteins.

p. 6:

I don't think the term "data-agnostic" makes sense since these methods are based on data and also make some assumptions about how the data can/should be used.

We have replaced this term with “model-agnostic”.

p. 16:

How are contacts defined; please add to caption.

A contact is considered if the Ca atoms of two residues are within a distance of 8 Å of each other. We have updated the caption with this information in Figures 4 and 5.

p. 20:

What do the authors mean by "non-specific interactions" in this context?

The interactions of fasudil are predominantly with the negatively charged residues in the C-terminal region of α Syn via charge-charge and π -stacking interactions (Robustelli et.al. (Journal of the American Chemical Society, 144(6), pp.2501-2510)).

In addition, in some metastable states that we identify, we also observe transient interactions with residues in the hydrophobic NAC region and N-terminal region. We refer to these transient interactions as “non-specific” interactions.

p. 27:

Are the axes of Fig. 9c/d z1 and z2?

Yes. The axes are z1 and z2

Smaller than minor

Abstract:

Rephrase "In particular, the presence of fasudil in milieu"

We have rephrased the sentence as follows:

"In particular, the presence of fasudil in the solvent..."

p. 4:

What does the word "potentially" do in "ensemble of conformations potentially sampled"?

Here, by potentially, we mean the various conformations that the protein can adopt, subject to the environmental conditions.

p. 10:

"we trained a large array of inter-residue pairwise distances"

The distances were not trained; please reformulate

We have corrected this sentence as follows:

"We trained a VAE model using a large array of inter-residue pairwise distances."

p. 13:

N/C-terminal -> terminus (or in the C-terminal region)

We have made the changes in the manuscript at the required places.

p. 20:

Precedent -> previous (?)

We have made the change in the manuscript.

p. 30:

As far as I understand, Anton does not use GPUs and does not run Desmond.

We thank the reviewer for providing this information. We referred to the original paper of the α syn-fasudil simulations (Robustelli et.al. (Journal of the American Chemical Society, 144(6), pp.2501-2510)). The authors have performed equilibration with GPU/Desmond and used Anton for production runs. We have modified this sentence as:

We have modified this sentence as:

"A 1500 μ s long all-atom MD simulation trajectory of α S 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." on page 31

References :

(1) Schütte C, Fischer A, Huisinga W, Deuffhard P (1999) A direct approach to conformational dynamics based on hybrid monte carlo. *J Comput Phys* 151:146–168

(2) Chodera JD, Swope WC, Pitera JW, Dill KA (2006) Long-time protein folding dynamics from short-time molecular dynamics simulations. *Multiscale Model Simul* 5:1214–1226.

<https://doi.org/10.7554/eLife.97709.2.sa0>