ACS Chemical Neuroscience

pubs.acs.org/chemneuro

Research Article

A Small Molecule Stabilizes the Disordered Native State of the Alzheimer's Aβ Peptide

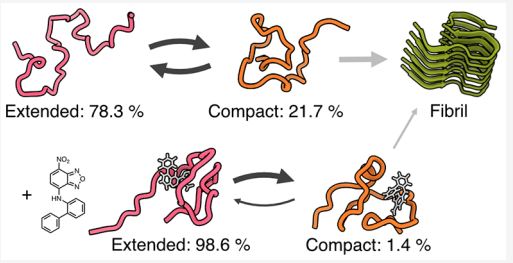
Thomas Löhr, Kai Kohlhoff, Gabriella T. Heller, Carlo Camilloni, and Michele Vendruscolo*

Cite This: ACS Chem. Neurosci. 2022, 13, 1738−1745

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The stabilization of native states of proteins is a powerful drug discovery strategy. It is still unclear, however, whether this approach can be applied to intrinsically disordered proteins. Here, we report a small molecule that stabilizes the native state of the Aβ42 peptide, an intrinsically disordered protein fragment associated with Alzheimer's disease. We show that this stabilization takes place by a disordered binding mechanism, in which both the small molecule and the Aβ42 peptide remain disordered. This disordered binding mechanism involves enthalpically favorable local π-stacking interactions coupled with entropically advantageous global effects. These results indicate that small molecules can stabilize disordered proteins in their native states through transient non-specific interactions that provide enthalpic gain while simultaneously increasing the conformational entropy of the proteins.



KEYWORDS: *small molecule, Alzheimer's disease, Aβ42 peptide, native state*

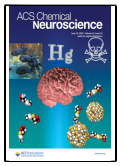
INTRODUCTION

Drug development for Alzheimer's disease has been a tremendous challenge in the past decades.[1] This condition is characterized by the formation of protein aggregates, such as fibrillar forms of the amyloid-β42 peptide (Aβ42).[2,3] This protein fragment is intrinsically disordered, i.e., it does not form a single stable folded structure as a monomer, but instead exists in a dynamic equilibrium of states with transient local structure and fast transitions.[4−12] Many drug development efforts focused on aggregation-prone proteins such as Aβ42 attempt to target the already-formed fibril and/or the structurally elusive oligomeric species.[13−15] Other attempts aimed to identify small molecules capable of stabilizing monomeric Aβ42 into a well-structured conformation[16−18] or generally interfering with the interaction of disordered proteins to structured partners by binding to their interfacing regions.[19] Since the most populated state of disordered proteins is conformationally highly heterogeneous, it has also been suggested that it may be more convenient to identify small molecules capable of stabilizing this disordered state.[20,21] The idea is that since the free energy landscape of disordered proteins is "inverted" when compared with the funnel concept of folded proteins, with the disordered state as the free energy minimum and the ordered states exhibiting relatively high free energies,[22] small molecules stabilizing this minimum would be easier to develop, as they would not have to restructure the topology of the free energy landscape itself.

Independent from the strategy pursued, however, it is extremely challenging to characterize the binding mode of small molecules to disordered protein on an atomistic level. While some experimental methods such as nuclear magnetic resonance spectroscopy can provide quantitative information, it is often not sufficient to clearly understand the interactions and kinetics underlying the binding.[20]

Molecular dynamics is one of the tools that can provide the necessary spatial and temporal resolution to study the interaction between disordered proteins and small molecules.[20] Together with Bayesian restraints from experimental data, molecular dynamics simulations have been used to characterize the thermodynamics of these binding modes in the case of the oncoprotein c-Myc[23] and Aβ42.[5] In the former study, urea was used as a control molecule to assess the sequence specificity of the drug. In the latter case of Aβ42, we studied the interaction with the small molecule 10074-G5 and showed that it was able to inhibit Aβ42 aggregation by binding the disordered monomeric form of the peptide. The interaction was characterized both experimentally, using various biophysical techniques, and computationally, using restrained molecular dynamics simulations with enhanced sampling, yielding thermodynamic information. While in both systems, the binding mode was found to be highly dynamic, a quantitative study of the kinetics was not possible due to the use of time-dependent restraints and biases applied during the simulation.

Received: February 19, 2022
Accepted: May 4, 2022
Published: June 1, 2022



https://doi.org/10.1021/acschemneuro.2c00116
ACS Chem. Neurosci. 2022, 13, 1738−1745

The microscopic kinetics in the form of contact lifetimes and autocorrelations have, to the best of our knowledge, never been calculated for this kind of interaction and could be especially instructive to fully understand the origin of entropic and enthalpic stabilization in these extremely dynamic binding events (Figure 1).[21]

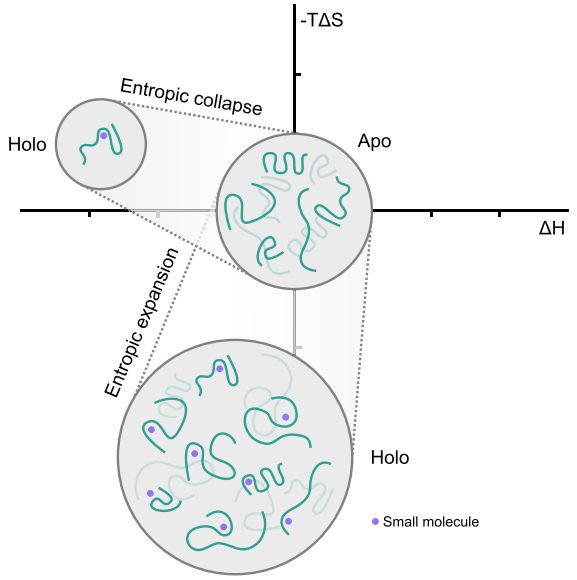


Figure 1. Illustration of two different native state stabilization mechanisms of disordered proteins. The interaction with a small molecule can result in a reduction or increase of conformational space of the protein, thus resulting in a positive or negative entropic contribution to the binding free energy. A loss of entropic native state stabilization will often be compensated for with a stronger enthalpic binding affinity, while an increase in entropy often requires more dynamic and thus weaker binding.

A quantitative study of the kinetics of these interactions may allow a more targeted approach to the design of both drugs and better experiments to probe their binding modalities. Because we can view each transient interaction of the small molecule with a residue as a binding or unbinding event, detailed knowledge of the contact lifetimes could act as a basis for a rational drug design strategy. However, even with atomistic computational approaches, gaining insight into the kinetics, i.e., transition rates, relaxation constants, autocorrelations, and state lifetimes, can be challenging. This is because in contrast to folded systems, the definition of states for disordered proteins is not always clear: due to the generally shallow free energy landscape state, transitions may be fast but not always distinct.[6] New developments in the theory of dynamical systems now allow an optimal state decomposition and transition operator to be learned using deep neural networks, for example, using the VAMPNet framework.[24,25] To acquire kinetic information for a system, one would traditionally use a Markov state model:[26,27] One first finds a suitable low-dimensional embedding of the system coordinates, followed by using a clustering algorithm to define microstates. Transitions between these can then be counted to build up statistics and thus construct a transition matrix. This matrix can then be coarse-grained to obtain macroscopic kinetics.[28,29]

Koopman operator[30,31] based models present a generalization of Markov models and have provided a basis for new method developments. The VAMPNet approach combines the steps of dimensionality reduction and clustering into a single function that can be approximated by a neural network and also yields a probabilistic state assignment in lieu of a discrete one.[24,25] Probabilistic state assignments are inherently well suited to disordered proteins, as the typically shallow free energy basins and low barriers can be encoded with some ambiguity. While hidden Markov models also allow for probabilistic state assignments, VAMPNet simplifies the model construction process, as the hyperparameter search over various dimensionality reduction and clustering techniques is replaced by a simplified search over neural network parameters, also allowing a more accurate model due to the use of a single arbitrarily non-linear function compared to two steps that are heavily restricted in terms of search space. This constrained VAMPNet approach was recently utilized by us to determine the kinetic ensemble of the disordered Aβ42 monomer.[6]

Here, we use this technique to build kinetic ensembles of Aβ42 with 10074-G5 and urea as a control molecule to expand on our previous thermodynamic ensembles.[5] We compare the transition rates, lifetimes, and state populations with the previous kinetic ensemble of the Aβ42 monomer[6] and further characterize the atomic-level protein−small molecule interactions.

RESULTS

Molecular Dynamics Simulations and Soft Markov State Models. We performed two explicit-solvent molecular dynamics simulations of Aβ42 with one molecule of urea and one molecule of 10074-G5, respectively. Both simulations were performed in multiple rounds of 1024 trajectories on the Google Compute Engine as described previously.[6] As before, we used a soft Markov state model approach using the constrained VAMPNet framework[24] to construct kinetic ensembles. The major advantages of this method, compared to regular discrete Markov state models, are the soft state definitions and the use of a single function mapping directly from arbitrary system coordinates to a state assignment probability, allowing for more optimal models. To aid our analysis, we added our previous simulation of Aβ42 with no additional molecules to our dataset. We refer to it as the *apo* ensemble.[6] We compared all ensembles using a decomposition into two states. In addition to being easier to interpret, this approach allows for a direct comparison of the slowest timescales in contrast to higher state-count models.

Computational and Experimental Validation. Constructing a kinetic ensemble using the constrained VAMPNet approach requires choosing the number of states and the model lag time. The latter is a critical parameter that needs to be chosen such that the model can accurately resolve both long and short timescales. This can be done by plotting the dependence of the slowest relaxation timescales on the lag time (Figure S1). A stricter measure is the Chapman−Kolmogorov test, comparing multiple applications of the Koopman operator estimated at a certain lag time τ with a Koopman operator estimated at a multiple of this lag time $n\tau$ (Figure S2).[32] To evaluate sampling convergence, we visualized the dependence of the mean relaxation timescales on the number of trajectories used to evaluate these timescales (Figure S8). With sufficient sampling of kinetics, we would expect the global timescales to be unchanged within error. Experimental validation was performed by comparing back-calculated chemical shifts to ones from experiments. Because the small molecule 10074-G5

https://doi.org/10.1021/acschemneuro.2c00116
ACS Chem. Neurosci. 2022, 13, 1738−1745

only has minor effects on the chemical shifts of Aβ42[23] and below the prediction error of the model[33] used to back-calculate the chemical shifts, we compared our calculated values to the experiment without the small molecule (Figure S3). We also computed the distribution of back-calculated chemical shifts over the full ensembles for each residue and atom type (Figures S4−S6).

10074-G5 Has Minor Impact on Ensemble-Averaged Structural Properties of Aβ42. To evaluate the influence of 10074-G5 and urea on the structural conformations of Aβ42, we calculated state-averaged contact maps and secondary structure content for each state of all ensembles (Figure S7a−c). In all cases, we find a state decomposition into a more extended state with few inter-residue contacts, and a slightly more compact form with a higher number of local backbone interactions. We will refer to these as the compact and extended states, respectively. The addition of a small molecule has little effect on the formation of contacts and other structural motifs. This finding is consistent with our recent experimental thermodynamic and kinetic characterization of this interaction, and the absence of strong chemical shift perturbations in the holo ensemble.[5]

10074-G5 and Urea Decelerate the Formation of More Compact States. Compared to the previously published kinetic ensemble of the apo form of Aβ42, the kinetic ensembles in the presence of both urea and 10074-G5 show a deceleration of more compact state formation (Figure 2). Notably, the transition from the more compact form to the more extended state is unaffected. This change is also mirrored in the state populations, which exhibit a strong shift toward the extended state. We note that even though there are strong changes in the state populations, the ensemble-averaged contact maps are very similar (Figure S7a−c). This is likely due to the high sensitivity of the VAMPNet method to minor changes in free energy barrier regions. These will have a significant effect on the kinetics and thus state populations but not on the ensemble averaged structure due to the relatively low thermodynamic weight.[34] While the lifetimes of the extended states increase, the ones for the more compact form are unchanged within model error. We can thus conclude that within our model, the small molecule has a strong effect on the contact formation rates but no influence on the contact dissociation rates.

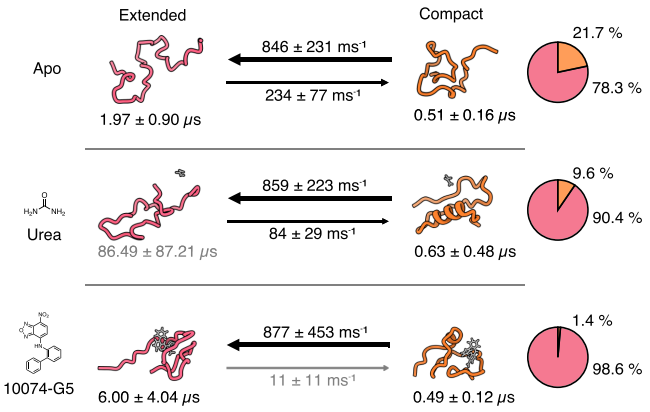


Figure 2. Impact of small molecules on the state transition rates, state lifetimes, and populations. The arrows indicate the mean state transition rates, the number below the representative structures is the mean state lifetime, and the pie charts show the mean state populations. Errors are the standard deviations of the bootstrap sample of the mean over all 20 models.

Small Molecules Shift the System to More Entropically Stable States by Short-Lived Local Interactions. To evaluate the impact of 10074-G5 on the conformational space of Aβ42, we calculated the Ramachandran and state entropy for all ensembles, as well as the autocorrelation of side-chain χ_1 dihedral angles (Figure 3). The Ramachandran entropy can indicate relative flexibility of the backbone, thus revealing potential regions of dynamic changes as a result of interactions between the peptide and small molecule.[5] Resolving this change in the entropy over residues (Figure 3a) indicates strong increases in the relatively hydrophobic C-terminal region of Aβ42. This entropy increase is confirmed globally by the sum of the entropies over all residues (Figure 3b). As an alternative metric, we also calculated the entropy in the state assignments (Figure 3c), this can be thought of as indicating the overall ambiguity in the state definition. Again, we find a

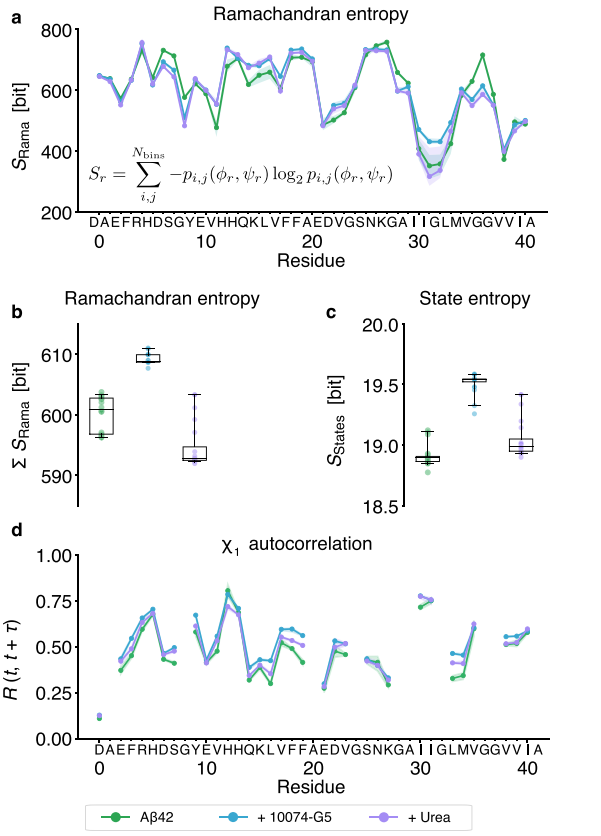


Figure 3. Effect of small molecules on conformational and state entropy of Aβ42, showing that 10074-G5 increases the conformational entropy of the peptide. (a) Ramachandran entropy, i.e., information entropy over the distribution of φ and ψ backbone dihedral angle conformations, using 100 bins. (b) Sum of the Ramachandran entropies over all residues for all ensembles. (c) State entropy, i.e., the population-weighted mean of the information entropy of each set of state assignments. More ambiguity in the state assignments leads to a correspondingly higher state entropy. (d) Autocorrelation of all sidechain χ_1 dihedral angles with a lag time of τ = 5 ns. Shaded areas in panels (a) and (d) indicate the 95th percentiles of the bootstrap sample of the mean over all 20 models. Whiskers and boxes in panels (b) and (c) indicate the 95th percentiles and quartiles of the bootstrap sample of the mean over all 20 models, respectively.

https://doi.org/10.1021/acschemneuro.2c00116
ACS Chem. Neurosci. 2022, 13, 1738−1745

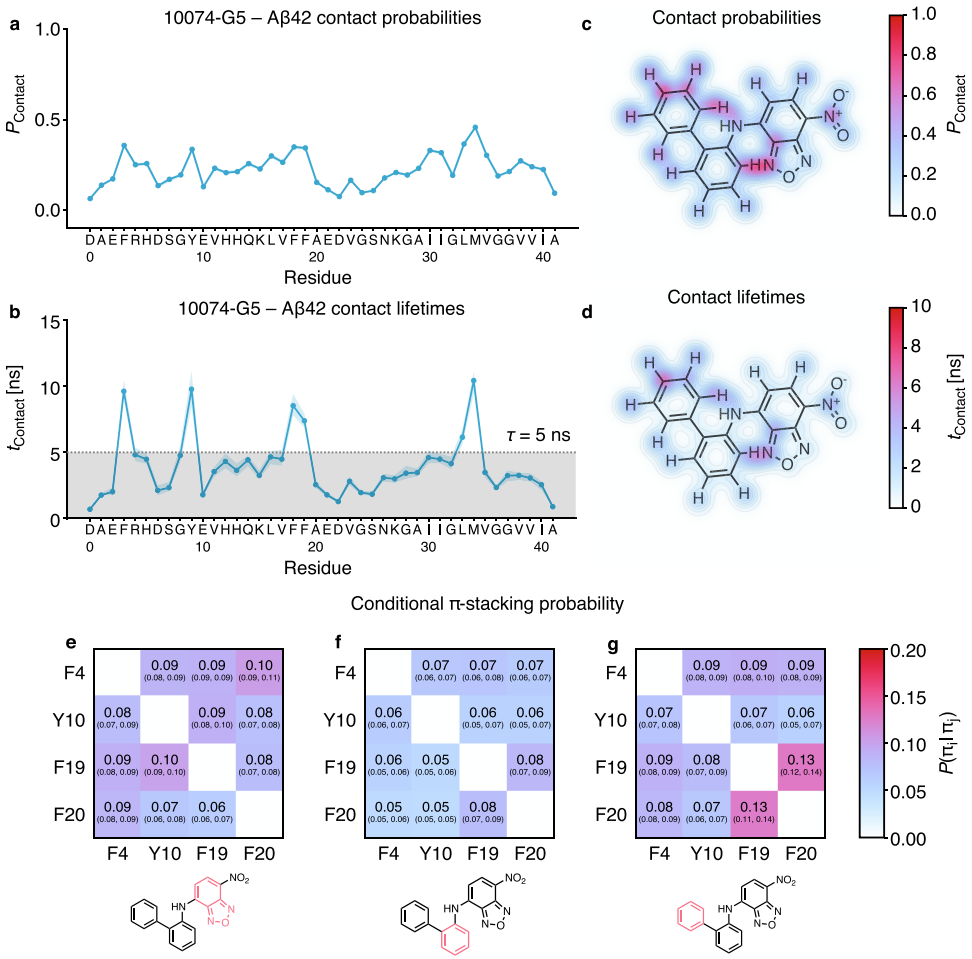


Figure 4. Residue- and atomic-level interactions of 10074-G5 with Aβ42 showing regions on the small molecule responsible for binding. (a) Contact probabilities of 10074-G5 and AB42 with a cutoff of 0.45 nm. (b) Lifetimes of these contacts, estimated using a Markov state model for each contact formation with a lag time of τ = 5 ns, indicated with gray shading. Colored shaded areas in panels (a) and (b) indicate the 95th percentiles of the bootstrap sample of the mean over all 20 models. (c,d) Contact probabilities and lifetimes of each atom of 10074-G5 with any residue of AB42. (e–g) Conditional probability of forming a π-stacking interaction, given an existing π-stacking interaction for all aromatic groups in the small molecule. Tuples indicate the 95th percentiles of the bootstrap sample of the mean over all 20 models.

relatively strong increase in the conformational entropy of Aβ42 for the ensemble with 10074-G5 and only minor increases for urea. These results are in agreement with our previous observations from simulations of the equilibrium ensembles in that the presence of 10074-G5 increases the conformations available to Aβ42, via the entropic expansion mechanism.[5,21]

To better understand the impact of the small molecule on local kinetics, we calculated the autocorrelation of the sidechain χ_1 dihedral angles (Figure 3). We see an increase in the autocorrelation, specifically for aromatic residues and M35, indicating a slowing of side chain rotations. This suggests that despite an increase in the backbone entropy, the peptide is able to visit many locally stable states, resulting in local enthalpic stabilization.

Interactions of 10074-G5 with Aβ42 Are Dominated by π-Stacking and Other Electrostatic Effects. To better understand the origin of the global and local effects of 10074-G5 on the ensemble, we analyzed the interactions on a residue and atomistic level (Figure 4). While the probability of forming a contact between the small molecule and a residue shows certain mild preferences (Figure 4a), these become more evident when looking at the lifetimes of these contacts (Figure 4b). Here, the longest contacts are formed by π-stacking with certain aromatic residues (F4, Y10, F19, F20) and by interactions with M35. This result also explains the reduction in side-chain rotations for these residues (Figure 3d). On an atomistic level, the π-interactions exhibit some anisotropy (Figure S9). The importance of the nitro- and benzofurazan fragments is also highlighted. Finally, we also investigated the conditionality of π-interactions, i.e., if we see an interaction between the molecule and residue i, what is the probability of also observing an interaction with residue j (Figure 4e–g)? The probabilities here are uniformly low but indicate a slight preference (13%) for a triple π-stack involving the terminal aromatic ring of 10074-G5 and residues F19 and F20 of Aβ42. The importance of π-stacking interactions was also noted in a computational study on the interactions of small molecules with α-synuclein.[35]

These results indicate that this disordered binding mechanism operates on two levels whereby local enthalpically favorable interactions coupled with global entropically advantageous effects. The local interactions are predominantly of an electrostatic nature and result in a reduction of side-chain rotations on specific residues. At the same time, these interactions also allow the exploration of more backbone

https://doi.org/10.1021/acschemneuro.2c00116
ACS Chem. Neurosci. 2022, 13, 1738–1745

conformations, thus resulting in a net entropy increase for Aβ42. This influence expands into the global kinetics of the system, significantly slowing the formation of local structure.

DISCUSSION

The results outlined above present a possible example of the previously proposed entropic expansion mechanism for the binding of small molecules to disordered proteins.[21,36] This mechanism is distinct from the entropic collapse and folding-on-binding mechanisms.[37,38] The concept of disordered binding is difficult to probe, as the tools suitable to detecting small changes in the conformational ensemble of disordered proteins are still in their infancy.[20] Nuclear magnetic resonance experiments can provide information, but it should usually be interpreted in a structural framework, necessitating molecular simulations with ensemble-averaged restraints,[39] or re-weighting approaches.[40] This constraint causes issues whenever we are also interested in kinetics, as by enhancing the sampling, we modify the natural dynamics of the system. Nevertheless, an approach to incorporate ensemble-averaged experimental measurements into Koopman models has recently been proposed.[41] Neither is it generally possible to use enhanced sampling methods to study kinetics without having some *a priori* knowledge of the system states. A framework allowing the incorporation of experimental data into a kinetic model and also allowing the use of enhanced sampling methods such as metadynamics,[42] without prior knowledge of states, would make the study of these systems easier and more accurate.

As we have shown, a kinetic model is crucial to fully explain the nature of these binding interactions. This is in part due to the ability to use the slowest timescales of the system to reliably define metastable states, something that is notoriously difficult for disordered proteins without access to the time dimension. This clustering alone is already sensitive enough to reveal differences between systems that are nearly invisible when comparing ensemble-averaged results and more conventional clustering methods.[5] Increases in local autocorrelation and global state transitions might be seen as indicators of both local enthalpic stabilization and global entropic expansion. The former result hints at the possibility of designing small molecules that exhibit high specificity, as the global entropic stabilization effect may be due to transient, local, enthalpically favorable interactions.[21] The two level global entropy−local enthalpy effect becomes especially visible when looking at the timescales: The slowest state transitions of the protein are on the order of microseconds, while the local, enthalpically favorable π-interaction lifetimes are no longer than tens of nanoseconds. Tuning these contact lifetimes and keeping them in a specific range could be argued to be essential, as stronger enthalpic interactions may have the effect of reducing the entropic contribution to the binding free energy of the small molecule. The entropic binding mechanism thus requires delicate balancing of the enthalpic and entropic terms to design active molecules. The lifetimes are specifically useful in this context as frequent but short-lived contacts may not have an effect on global state transitions, whereas too-long lived contacts could cause an entropic collapse and corresponding loss of binding affinity. Information on the contact probabilities can thus be argued to be insufficient to fully explain the binding mechanism. On the other hand, the simultaneous tuning of the binding affinity for multiple different contacts across the protein sequence carries a greater risk of loss of specificity and subsequent off-target effects. Striking an appropriate balance to achieve high specificity and affinity for this kind of native state stabilization thus presents a major challenge.

The observed binding mechanism also identifies π-stacking interactions as a major driving force. Similar effects have been observed for the binding of another small molecule, fasudil, and α-synuclein, which is also intrinsically disordered.[35] We note that while that study proposed a "shuttling model" mechanism to explain the diffusion of the small molecule on the α-synuclein surface, here, we demonstrate the stabilization of a native state of a disordered protein by a disordered binding mechanism. The $\pi-\pi$ stacking phenomenon also plays a major role in liquid−liquid phase separation,[43] suggesting a possible link between the effect of these small molecules and the hypothesized state of some proteins in a crowded environment. For molecular simulations, the force field may present a barrier in studying $\pi-\pi$ interactions in more detail. This is because these interactions are not explicitly part of the potential, but only approximated with a combination of electrostatic and hydrophobic terms.[44] Polarizable force fields may offer a computationally affordable alternative that could more accurately model this type of binding.[45]

Looking forward, it may become possible to pursue a drug discovery strategy for disordered proteins based on the stabilization of their native states through the disordered binding mechanism that we have described here. This strategy would extend an approach to disordered proteins that has already proven successful for folded proteins[46] and would have the advantage of maintaining the proteins in their native functional states.

METHODS

Details of the Simulations. All simulations were performed on the Google Compute Engine with `n1-highcpu-8` preemptible virtual machine instances, equipped with eight Intel Haswell CPU cores and 7.2 GB of RAM. Molecular dynamics simulations were performed with GROMACS 2018. 1,[47] with 1024 starting structures sampled from the previously performed apo simulation[6] using the Koopman model weights. Each conformation was placed in the center of a rhombic dodecahedron box with a volume of 358 nm^3, and the corresponding small molecule was placed in the corner of the box. The force field parameters for urea were taken from the CHARMM22* force field[48] and the ones for 10074-G5 were computed using the Force Field Toolkit (FFTK)[49] and Gaussian 09,[50] as described previously.[5] The systems were then solvated using between 11,698 (11,707) and 11,740 (11,749) water molecules. Both systems were minimized using the steepest descent method to a maximum target force of 1000 kJ/mol/nm. Both systems were subsequently equilibrated, first over a time range of 500 ps in the NVT ensemble using the Bussi thermostat[51] and then over another 500 ps in the NPT ensemble using Berendsen pressure coupling.[52] In both equilibrations, position restraints were placed on all heavy atoms. All production simulations were performed using 2 fs time steps in the NVT ensemble using the CHARMM22*[48] force field and TIP3P water model[53] at 278 K and LINCS constraints[54] on all bonds. Electrostatic interactions were modeled using the Particle-Mesh-Ewald approach[55] with a short-range cutoff of 1.2 nm. All simulations used periodic boundary conditions. We again used the fluctuation-amplification of specific traits (FAST) approach[56] to adaptive sampling, with clustering performed through time-lagged independent component analysis (TICA)[57,58] using a lag time of 5 ns and C distances fed to the *k*-means clustering algorithm tearthurKmeansAdvantagesCareful2007 to yield 128 clusters. 1024 new structures were then sampled from these clusters based on maximizing the deviation to the mean C distance matrix for each cluster and maximizing the sampling of the existing clusters, using a balance

https://doi.org/10.1021/acschemneuro.2c00116
ACS Chem. Neurosci. 2022, 13, 1738−1745

parameter of α = 1.0, with all amino acids weighted equally. This approach was performed once for each ensemble; however, we also chose to perform 32 additional long-trajectory simulations for the 10074-G5 ensemble, yielding a total of 2,079 trajectories for the latter, and 2,048 trajectories for the urea ensemble. The total simulated times were 306 and 279 μs for the 10074-G5 and urea ensembles, respectively. The shortest and longest trajectories for 10074-G5 (urea) were 21 (24) ns and 1134 (196) ns. All trajectories were subsampled to 250 ps time steps for further analysis.[59]

Details of the Neural Networks. State decomposition and kinetic model construction was performed using the constrained VAMPNet approach,[24,25] using the same method described previously.[6] We again chose flattened inter-residue nearest-neighbor heavy-atom distance matrices as inputs, resulting in 780 input dimensions. We used the self-normalizing neural network architecture[60] with scaled-exponential linear units, normal LeCun weight initialization[61] and alpha dropout. We chose an output dimension of 2, thus yielding a soft two-state assignment. The datasets were prepared by first creating a test dataset by randomly sampling 10% of the frames. In the case of 10074-G5, we excluded all frames in which the closest distance between the small molecule and peptide was higher than 0.5 nm. We then created 20 randomized 9:1 train-validation splits to allow a model error estimate. Training was performed by using three trials for each train-validation split and picking the best-performing model based on the VAMP2 score[31] of the test set. We implemented the model using Keras 2.2.4[62] with the Tensorflow 2.1.0[63] backend. We chose the following model hyperparameters based on two successive coarser and finer grid searches: A network lag time of 5 ns, a layer width of 512 nodes, a depth of 2 layers, an L2 regularization strength of 10^{-7}, and a dropout of 0.05. Training was performed in 10,000 frame pairs using the Adam minimizer[64] with a learning rate of 0.05, β_2 = 0.99, and epsilon of 10^{-4}, and an early stopping criterion of a minimum validation score improvement of 10^{-3} over the last five epochs. For the constrained part of the model, we reduced the learning rate by a factor of 0.02. We used a single Google Compute Engine instance with 12 Intel Haswell cores, 78 GB of RAM, and an NVidia Tesla V100 GPU.

Details of the Kinetic Analysis. After training, VAMPNet yields a state assignment vector $\chi(\mathbf{x}_t)$ for each frame $\mathbf{x}_t$ of the ensemble. Based on this vector, we can calculate state averages $\langle A_i \rangle$ for any observable $A(\mathbf{x}_t)$:

$$\langle A_i \rangle = \left(\sum_{t=1}^{T} \chi_i(\mathbf{x}_t) \right)^{-1} \sum_{t=1}^{T} \chi_i(\mathbf{x}_t) A(\mathbf{x}_t) \tag{1}$$

Here, i is the corresponding state and the sum runs over all time steps. To calculate an ensemble average $\langle A \rangle$, one first calculates a weight w_t for each frame using the model equilibrium distribution π:

$$w_t = \frac{\langle \chi(\mathbf{x}_t) | \pi \rangle}{\sum_{t=1}^{T} \langle \chi(\mathbf{x}_t) | \pi \rangle} \tag{2}$$

which leads to the ensemble average

$$\langle A \rangle = \sum_{t=1}^{T} w_t A(\mathbf{x}_t) \tag{3}$$

Because each trained model will classify the states in an arbitrary order, we need to sort the state assignment vectors based on state similarity. We did this by comparing the state-averaged contact maps using root-mean-square deviation as a metric and grouping states based on the lowest value. Any deviations are thus accounted for in the overall model error.

Model Validation. The Koopman matrix $\mathbf{K}(\tau)$ is given directly by the neural network model, along with the equilibrium distribution π. We validated our models using the Chapman–Kolmogorov test:

$$\mathbf{K}(n\tau) \approx \mathbf{K}^n(\tau) \tag{4}$$

where τ is the model lag time and $n\tau$ is a low integer-multiple of the lag time. The model should therefore behave the same way whether we estimate it at a longer lag time or repeatedly apply the transfer operator. We first estimate a suitable lag time τ by plotting the relaxation timescales over the chosen lag time. The lag time τ should be chosen to be as small as possible, but large enough to not have any impact on the longer relaxation timescales, which represent the slowest motions of the system. The temporal resolution of the model is thus given by this lag time. The relaxation timescales t_i are calculated from the eigenvalues λ_i of the Koopman matrix $\mathbf{K}(\tau)$ as follows:

$$t_i = \frac{-\tau}{\log|\lambda_i|} \tag{5}$$

We can similarly compute the state lifetimes $\bar{t}_i$from the diagonal elements of the Koopman matrix $\mathbf{K}(\tau)_{ii}$ using:

$$\bar{t}_i = \frac{-\tau}{\log \mathbf{K}(\tau)_{ii}} \tag{6}$$

Experimental Validation. We back-calculated the nuclear magnetic resonance chemical shifts using the CamShift algorithm[33] as implemented in PLUMED 2.4.1.[65,66] We again used the same ensemble averaging procedure described above.

Errors. Errors are calculated over all trained neural network models. To obtain a more meaningful estimate, we only consider frames that were part of the bootstrap training sample of the corresponding model, i.e., one of the 20 models described above. The reported averages are the mean, and the errors the 95th percentiles over all 20 models, unless reported otherwise.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acschemneuro.2c00116.

Analysis code and notebooks are available from https://github.com/vendruscolo-lab/ab42-g5-ensemble and https://zenodo.org/record/5659241. Subsampled trajectory and intermediate data, as well as the trained neural network weights and analysis notebooks, are available from https://zenodo.org/record/5659241.

■ AUTHOR INFORMATION

Corresponding Author

Michele Vendruscolo − *Department of Chemistry, University of Cambridge, CB2 1EW Cambridge, UK;* orcid.org/0000-0002-3616-1610; Email: mv245@cam.ac.uk

Authors

Thomas Löhr − *Department of Chemistry, University of Cambridge, CB2 1EW Cambridge, UK*

Kai Kohlhoff − *Google Research, Mountain View, California 94043, United States*

Gabriella T. Heller − *Department of Chemistry, University of Cambridge, CB2 1EW Cambridge, UK; Department of Structural and Molecular Biology, University College London, WC1E 6BT London, UK*

Carlo Camilloni − *Dipartimento di Bioscienze, Università degli Studi di Milano, 20133 Milano, Italy;* orcid.org/0000-0002-9923-8590

Complete contact information is available at: https://pubs.acs.org/10.1021/acschemneuro.2c00116

Author Contributions

T.L., K.K., G.T.H., C.C., and M.V. designed research; T.L., K.K., and G.T.H. performed research; T.L., K.K., G.T.H., and C.C. analyzed data; T.L., K.K., G.T.H., C.C., and M.V. wrote the paper.

https://doi.org/10.1021/acschemneuro.2c00116
ACS Chem. Neurosci. 2022, 13, 1738−1745

Funding

G.T.H. is supported by the Rosalind Franklin Research Fellowship at Newnham College, Cambridge and the Schmidt Science Fellowship.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We would like to thank Google and the Google Accelerated Science team for providing access to the Google Cloud Platform for simulations and analysis.

■ REFERENCES

(1) Cummings, J.; Lee, G.; Zhong, K.; Fonseca, J.; Taghva, K. *Alzheimer's & Dementia: Transl. Res. Clin. Inter.* **2021**, *7*, No. e12179.
(2) Hampel, H.; Hardy, J.; Blennow, K.; Chen, C.; Perry, G.; Kim, S. H.; Villemagne, V. L.; Aisen, P.; Vendruscolo, M.; Iwatsubo, T.; Masters, C. L.; Cho, M.; Lannfelt, L.; Cummings, J. L.; Vergallo, A. *Mol. Psychiatry* **2021**, 1−23.
(3) Hardy, J. A.; Higgins, G. A. *Science* **1992**, *256*, 184−185.
(4) Ball, K. A.; Phillips, A. H.; Nerenberg, P. S.; Fawzi, N. L.; Wemmer, D. E.; Head-Gordon, T. *Biochemistry* **2011**, *50*, 7612−7628.
(5) Heller, G. T.; et al. *Sci. Adv.* **2020**, *6*, No. eabb5924.
(6) Löhr, T.; Kohlhoff, K.; Heller, G. T.; Camilloni, C.; Vendruscolo, M. *Nat. Comput. Sci.* **2021**, *1*, 71−78.
(7) Meng, F.; Bellaiche, M. M.; Kim, J.-Y.; Zerze, G. H.; Best, R. B.; Chung, H. S. *Biophys. J.* **2018**, *114*, 870−884.
(8) Nasica-Labouze, J.; et al. *Chem. Rev.* **2015**, *115*, 3518−3563.
(9) Paul, A.; Samantray, S.; Anteghini, M.; Khaled, M.; Strodel, B. *Chem. Sci.* **2021**, *12*, 6652−6669.
(10) Roche, J.; Shen, Y.; Lee, J. H.; Ying, J.; Bax, A. *Biochemistry* **2016**, *55*, 762−775.
(11) Rosenman, D. J.; Connors, C. R.; Chen, W.; Wang, C.; García, A. E. *J. Mol. Biol.* **2013**, *425*, 3338−3359.
(12) Sgourakis, N. G.; Merced-Serrano, M.; Boutsidis, C.; Drineas, P.; Du, Z.; Wang, C.; Garcia, A. E. *J. Mol. Biol.* **2011**, *405*, 570−583.
(13) Habchi, J.; Arosio, P.; Perni, M.; Costa, A. R.; Yagi-Utsumi, M.; Joshi, P.; Chia, S.; Cohen, S. I. A.; Müller, M. B. D.; Linse, S.; Nollen, E. A. A.; Dobson, C. M.; Knowles, T. P. J.; Vendruscolo, M. *Sci. Adv.* **2016**, *2*, No. e1501244.
(14) Habchi, J.; Chia, S.; Limbocker, R.; Mannini, B.; Ahn, M.; Perni, M.; Hansson, O.; Arosio, P.; Kumita, J. R.; Challa, P. K.; Cohen, S. I. A.; Linse, S.; Dobson, C. M.; Knowles, T. P. J.; Vendruscolo, M. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, E200−E208.
(15) Lieblein, T.; Zangl, R.; Martin, J.; Hoffmann, J.; Hutchison, M. J.; Stark, T.; Stirnal, E.; Schrader, T.; Schwalbe, H.; Morgner, N. *eLife* **2020**, *9*, No. e59306.
(16) Sinha, S.; et al. *J. Am. Chem. Soc.* **2011**, *133*, 16958−16969.
(17) Zhu, M.; De Simone, A.; Schenk, D.; Toth, G.; Dobson, C. M.; Vendruscolo, M. *The Journal of Chemical Physics* **2013**, *139*, No. 035101.
(18) Tóth, G.; et al. *PLoS One* **2014**, *9*, No. e87133.
(19) Xu, Y.; Shi, J.; Yamamoto, N.; Moss, J. A.; Vogt, P. K.; Janda, K. D. *Bioorg. Med. Chem.* **2006**, *14*, 2660−2673.
(20) Heller, G. T.; Aprile, F. A.; Vendruscolo, M. *Cell. Mol. Life Sci.* **2017**, *74*, 3225−3243.
(21) Heller, G. T.; Sormanni, P.; Vendruscolo, M. *Trends Biochem. Sci.* **2015**, *40*, 491−496.
(22) Granata, D.; Baftizadeh, F.; Habchi, J.; Galvagnion, C.; De Simone, A.; Camilloni, C.; Laio, A.; Vendruscolo, M. *Sci. Rep.* **2015**, *5*, 15449.
(23) Heller, G. T.; Aprile, F. A.; Bonomi, M.; Camilloni, C.; De Simone, A.; Vendruscolo, M. *J. Mol. Biol.* **2017**, *429*, 2772−2779.
(24) Mardt, A.; Pasquali, L.; Noé, F.; Wu, H. Deep Learning Markov and Koopman Models with Physical Constraints. Proceedings of The First Mathematical and Scientific Machine Learning Conference. 2020; pp 451−475.
(25) Mardt, A.; Pasquali, L.; Wu, H.; Noé, F. *Nat. Commun.* **2018**, *9*, 5.
(26) Chodera, J. D.; Noé, F. *Curr. Opin. Struct. Biol.* **2014**, *25*, 135−144.
(27) Husic, B. E.; Pande, V. S. *J. Am. Chem. Soc.* **2018**, *140*, 2386−2396.
(28) Noé, F.; Wu, H.; Prinz, J.-H.; Plattner, N. *J. Chem. Phys.* **2013**, *139*, 184114.
(29) Rabiner, L. R.; Juang, B. H. *IEEE ASSp Magazine* **1986**, 13.
(30) Klus, S.; Nüske, F.; Koltai, P.; Wu, H.; Kevrekidis, I.; Schütte, C.; Noé, F. *J. Nonlinear Sci.* **2018**, *28*, 985−1010.
(31) Wu, H.; Noé, F. *J. Nonlinear Sci.* **2020**, *30*, 23−66.
(32) Bowman, G. R.; Pande, V. S.; Noé, F., Eds. *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation; Advances in Experimental Medicine and Biology*; Springer: Netherlands, Dordrecht, 2014; Vol. *797*.
(33) Kohlhoff, K. J.; Robustelli, P.; Cavalli, A.; Salvatella, X.; Vendruscolo, M. *J. Am. Chem. Soc.* **2009**, *131*, 13894−13895.
(34) Suárez, E.; Wiewiora, R. P.; Wehmeyer, C.; Noé, F.; Chodera, J. D.; Zuckerman, D. M. *J. Chem. Theory Comput.* **2021**, *17*, 3119−3133.
(35) Robustelli, P.; Ibanez-de-Opakua, A.; Campbell-Bezat, C.; Giordanetto, F.; Becker, S.; Zweckstetter, M.; Pan, A. C.; Shaw, D. E. *Molecular Basis of Small-Molecule Binding to α-Synuclein*. 2021.
(36) Heller, G. T.; Bonomi, M.; Vendruscolo, M. *J. Mol. Biol.* **2018**, *430*, 2288−2292.
(37) Lee, H.; Mok, K. H.; Muhandiram, R.; Park, K.-H.; Suk, J.-E.; Kim, D.-H.; Chang, J.; Sung, Y. C.; Choi, K. Y.; Han, K.-H. *J. Biol. Chem.* **2000**, *275*, 29426−29432.
(38) Robustelli, P.; Piana, S.; Shaw, D. E. *J. Am. Chem. Soc.* **2020**, *142*, 11092−11101.
(39) Bonomi, M.; Heller, G. T.; Camilloni, C.; Vendruscolo, M. *Curr. Opin. Struct. Biol.* **2017**, *42*, 106−116.
(40) Rangan, R.; Bonomi, M.; Heller, G. T.; Cesari, A.; Bussi, G.; Vendruscolo, M. *J. Chem. Theory Comput.* **2018**, *14*, 6632−6641.
(41) Mardt, A.; Noé, F. *J. Chem. Phys.* **2021**, *155*, 214106.
(42) Laio, A.; Parrinello, M. *Proc. Natl. Acad. Sci.* **2002**, *99*, 12562−12566.
(43) Vernon, R. M.; Chong, P. A.; Tsang, B.; Kim, T. H.; Bah, A.; Farber, P.; Lin, H.; Forman-Kay, J. D. *eLife* **2018**, *7*, No. e31486.
(44) Paton, R. S.; Goodman, J. M. *J. Chem. Inf. Model.* **2009**, *49*, 944−955.
(45) Baker, C. M. *WIREs Comput. Mol. Sci.* **2015**, *5*, 241−254.
(46) Burton, A.; Castaño, A.; Bruno, M.; Riley, S.; Schumacher, J.; Sultan, M. B.; Tai, S. S.; Judge, D. P.; Patel, J. K.; Kelly, J. W. *DDDT* **2021**, *15*, 1225−1243.
(47) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. *SoftwareX* **2015**, *1-2*, 19−25.
(48) MacKerell, A. D.; et al. *J. Phys. Chem. B* **1998**, *102*, 3586−3616.
(49) Mayne, C. G.; Saam, J.; Schulten, K.; Tajkhorshid, E.; Gumbart, J. C. *J. Comput. Chem.* **2013**, *34*, 2757−2770.
(50) Frisch, M. J. et al. Gaussian 16 Revision A.03. 2016.
(51) Bussi, G.; Donadio, D.; Parrinello, M. *The Journal of Chemical Physics* **2007**, *126*, No. 014101.
(52) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *The Journal of Chemical Physics* **1984**, *81*, 3684−3690.
(53) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *The Journal of Chemical Physics* **1983**, *79*, 926−935.
(54) Hess, B. *J. Chem. Theory Comput.* **2008**, *4*, 116−122.
(55) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577−8593.
(56) Zimmerman, M. I.; Bowman, G. R. *J. Chem. Theory Comput.* **2015**, *11*, 5747−5757.
(57) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. *J. Chem. Phys.* **2013**, *139*, No. 015102.
(58) Schwantes, C. R.; Pande, V. S. *J. Chem. Theory Comput.* **2013**, *9*, 2000−2009.
(59) Arthur, D.; Vassilvitskii, S. K-Means++: The Advantages of Careful Seeding. Proceedings of the Eighteenth Annual ACM-SIAM

Symposium on Discrete Algorithms. Philadelphia, PA, USA, 2007; pp 1027−1035.
(60) Klambauer, G.; Unterthiner, T.; Mayr, A.; Hochreiter, S. In *Advances in Neural Information Processing Systems 30*; Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; Garnett, R. Eds.; Curran Associates, Inc., 2017; pp. 971−980.
(61) LeCun, Y. A.; Bottou, L.; Orr, G. B.; Müller, K.-R. In *Neural Networks: Tricks of the Trade: Second Edition*; Montavon, G.; Orr, G. B.; Müller, K. R. Eds.; Lecture Notes in Computer Science; Springer: Berlin, Heidelberg, 2012; pp. 9−48.
(62) Chollet, F. 2015,
(63) Abadi, M. et al. Tensor Flow: A System for Large-Scale Machine Learning. 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16). 2016; pp 265−283.
(64) Kingma, D. P.; Ba, J. *arXiv:1412.6980 [cs]* 2014,
(65) PLUMED consortium. *Nat. Methods* **2019**, *16*, 670−673.
(66) Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. *Comput. Phys. Commun.* **2014**, *185*, 604−613.

https://doi.org/10.1021/acschemneuro.2c00116
ACS Chem. Neurosci. 2022, 13, 1738−1745