communications biology

ARTICLE

https://doi.org/10.1038/s42003-022-03562-y OPEN

Artificial intelligence guided conformational mining of intrinsically disordered proteins

Aayush Gupta[1], Souvik Dey[1], Alan Hicks[1] & Huan-Xiang Zhou[1,2]✉

Artificial intelligence recently achieved the breakthrough of predicting the three-dimensional structures of proteins. The next frontier is presented by intrinsically disordered proteins (IDPs), which, representing 30% to 50% of proteomes, readily access vast conformational space. Molecular dynamics (MD) simulations are promising in sampling IDP conformations, but only at extremely high computational cost. Here, we developed generative autoencoders that learn from short MD simulations and generate full conformational ensembles. An encoder represents IDP conformations as vectors in a reduced-dimensional latent space. The mean vector and covariance matrix of the training dataset are calculated to define a multivariate Gaussian distribution, from which vectors are sampled and fed to a decoder to generate new conformations. The ensembles of generated conformations cover those sampled by long MD simulations and are validated by small-angle X-ray scattering profile and NMR chemical shifts. This work illustrates the vast potential of artificial intelligence in conformational mining of IDPs.

[1] Department of Chemistry, University of Illinois at Chicago, Chicago, IL 60607, USA. [2] Department of Physics, University of Illinois at Chicago, Chicago, IL 60607, USA. ✉email: hzhou43@uic.edu

Artificial intelligence (AI) is gradually overshadowing traditional physics-based approaches[1,2], achieving breakthroughs in solving some of the most challenging problems in chemistry and physics. For example, a deep neural network has obtained nearly exact solutions of the electronic Schrödinger equation for small molecule[3]. Another recent breakthrough is the prediction of three-dimensional structures of proteins by neural network-based methods, Alphafold[4] and RoseTTafold[5]. With problems facing structured proteins being solved by these and other AI-based methods[6–9], a new frontier is now presented by intrinsically disordered proteins (IDPs). Instead of adopting well-defined three-dimensional structures, IDPs readily access vast conformational space. Here we report on the development of a generative AI model to mine the conformational space of IDPs.

IDPs, accounting for 30% to 50% of proteomes, perform many essential cellular functions including signaling and regulation, and are implicated in numerous human diseases[10,11]. In particular, polyglutamine expansion is associated with Huntingtin's and other diseases[12]. Amyloid-beta peptides, including Aβ40, are linked to Alzheimer's disease[13]. The cell division machinery of *Mycobacterium tuberculosis*, the causative agent of tuberculosis, contains a number of membrane proteins, including ChiZ, with disordered cytoplasmic regions[14,15]. The functional and disease mechanisms of these and other IDPs remain unclear, in large part because we lack knowledge of their conformational ensembles in various states (e.g., in isolation, in aggregation, and bound with interaction partners).

The vastness of IDPs' conformational space poses great challenges. Experimental techniques are limited to probing some aspects of the conformational space. For example, small-angle x-ray scattering (SAXS) provides information on the overall shapes and sizes of IDPs[16], whereas NMR properties, such as secondary chemical shifts, carry residue-specific information but still vastly under-represent the degrees of freedom of IDPs[17]. Molecular dynamics (MD) simulations offer an attractive approach for IDPs, with an atomic representation for each conformation, but the simulation time that can be presently achieved, which directly determines the extent of conformation sampling, is largely limited to 10 s of μs. The conformational ensembles of the 64-residue cytoplasmic disordered region of ChiZ (referred to simply as ChiZ hereafter) sampled by multiple replicate simulations, totaling 36 μs in solution and 38 μs at membrane, have been validated by SAXS and NMR data[14,15]. While we cannot answer whether 10 s of μs of simulations are really long enough, we do know that shorter simulations are insufficient. For example, Kukharenko et al.[18] have shown that the conformations of a 22-residue fragment of α-synuclein sampled in 1 μs represent only a small subset of the ensemble collected from 13 μs of "expansion" simulations. The latter are a large number (200) of short simulations (30–100 ns) started from sparsely populated regions in a two-dimensional embedded space (via sketch-map embedding). How to exhaustively cover the conformational space of IDPs without an inhibitory amount of computational time remains an open question.

For structured proteins, autoencoders have been developed to represent structures in two-dimensional latent spaces and reconstruct the structures back in Cartesian coordinates[6,8]. In another recent study[9], an autoencoder was trained to project the inter-residue distances of the ribose-binding protein into a two-dimensional latent space. The open and closed states of the protein were found to occupy separate regions in the latent space. The authors linearly interpolated points from these two states and decoded the interpolated points into inter-residue distances that represent conformations on the transition paths between the open and closed states. The inter-residue distances from interpolation were finally coupled to an all-atom model to enhance the latter's conformational sampling in MD simulations. Noé et al.[7] built Boltzmann generators, which use neural networks to represent protein structures sampled from short MD simulations as a Gaussian distribution in the latent space; points sampled from the Gaussian are transformed back as structures in Cartesian coordinates. In toy problems, the authors demonstrated that points located in different energy wells in conformational space are repacked into a dense distribution with a single peak in the latent space. These and other AI-based methods might potentially be adapted to study IDPs[19]. Several other approaches may also provide inspirations for IDPs, including variational autoencoders for dimensionality reduction of protein folding trajectories and subsequent identification of intermediate states by clustering in the latent space[20], and variational autoencoders and other neural networks for optimal selection of Markov states by training with conformations at a fixed lag time[21,22].

Here we present generative autoencoders designed to mine the conformational space of IDPs. Our design goal is to accurately sample the entire conformational space while limiting cost, which is MD conformations needed for training the autoencoders. The performance of the resulting autoencoders rivals that of expensive MD simulations and is validated by SAXS and chemical shift data. Our work opens the door to modeling IDPs in various functional states.

Results

We built autoencoders first to represent IDP conformations as vectors in a reduced-dimensional latent space (Fig. 1a). Training of the autoencoders involved reconstructing the conformations from the latent vectors and minimizing deviations from the original conformations. The training datasets consisted of conformations sampled from short MD simulations. We then modeled the latent vectors of the training datasets as multivariate Gaussian distributions (Fig. 1b). By sampling from these distributions for reconstruction, we generated the full conformational ensembles of IDPs (Fig. 1c). These generative autoencoders were built for three IDPs: polyglutamine Q15, Aβ40, and ChiZ, and were validated by their ability to cover all conformations sampled in long MD simulations and to reproduce experimentally measured properties. These IDPs contain 17 (including two capping groups), 40, and 64 residues (denoted by N_{res}).

Note that our goal is to use the smallest amount of training data – sampled from MD simulations as short as possible – to build autoencoders that will generate the most accurate full conformational ensemble of an IDP. To achieve this goal, we limit the training dataset to conformations sampled from the initial portion of the MD simulations, and use the subsequent portion only for the purpose of testing the accuracy of the autoencoders. Although increasing the training size or including conformations randomly sampled anywhere from the simulations, such as by shuffling the MD conformations before separating them into training and test sets, can potentially increase the accuracy of the autoencoders, doing so will depart from our goal.

Representation in a reduced-dimensional space. As a steppingstone to generating new conformations, we first reduced the dimensionality of the conformational space. The original conformations of the IDPs were specified by the Cartesian coordinates of heavy atoms (with truncation of some side chains). The dimension of the conformational space was thus $3N$, where N, denoting the number of heavy atoms included, was 140, 230, and 385, respectively, for Q15, Aβ40, and ChiZ. We chose the dimension (n) of the latent space for each IDP to be $0.75N_{res}$, or 13 for Q15, 30 for Aβ40, and 48 for ChiZ.

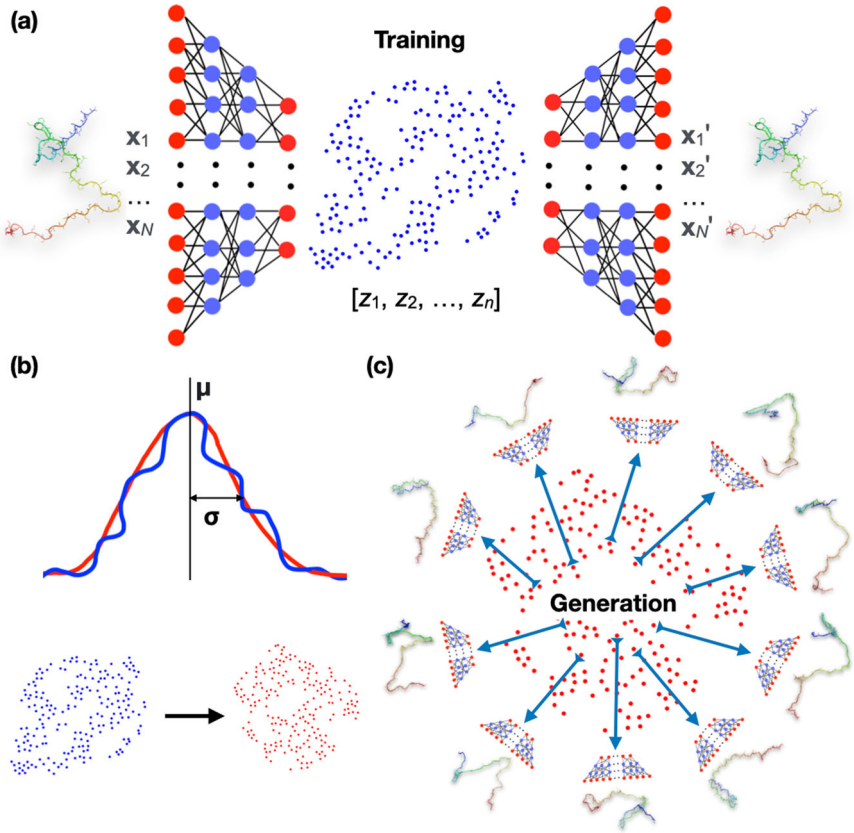

Fig. 1 Design of generative autoencoders. a Illustration of the architecture of an autoencoder. The encoder part of the autoencoder represents the conformations of an IDP, specified by the Cartesian coordinates [$\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_N$] of N heavy atoms, as n-dimensional vectors [$\boldsymbol{z}_1$, $\boldsymbol{z}_2$, ..., $\boldsymbol{z}_n$] in the latent space. The decoder then reconstructs the latent vectors back to conformations in Cartesian coordinates, [$\mathbf{x}_1'$, $\mathbf{x}_2'$, ..., $\mathbf{x}_N'$]. During training, the weights of the neural networks are tuned to minimize the deviation of the reconstructed conformations from the original ones. **b** Modeling of the distribution of the latent vectors (blue) of the training set by a multivariate Gaussian (red). The mean vector and covariance matrix of the Gaussian are those of the training latent vectors. The curves illustrate a Gaussian fit to the distribution of the training data; the scatter plots show a comparison of the training data and the Gaussian model. **c** Generation of new conformations. Vectors sampled from the multivariate Gaussian are fed to the decoder to generate new conformations. IDP structures are shown in a color spectrum with blue at the N-terminus and red at the C-terminus.

Conformations for training and testing the autoencoders came from multiple μs-long MD simulations[14,23]. We collected 95,000, 140,000, and 145,000 frames, respectively, at 10 ps intervals for Q15 and 20 ps intervals for Aβ40 and ChiZ from each replicate run; the numbers of replicate runs were 2, 4, and 12, respectively. An initial portion (e.g., 10%) of each run was taken as a training set whereas the remaining portion was the test set. The accuracy of an autoencoder was assessed by the root-mean-square deviations (RMSDs) between test conformations and their reconstructions. These RMSDs were averaged for the entire 100-fold diluted test set (comprising frames saved at 1-ns intervals for Q15 and 2-ns for Aβ40 and ChiZ). As adjacent frames in MD simulations tend to have similar three-dimensional structures, the dilution serves to reduce redundancy of the test set. The reconstruction RMSD results are shown in Fig. 2.

For Q15, the average reconstruction RMSDs are below 5 Å even when only 5% of the MD simulations (corresponding to 95 ns of simulation time) is used for training (Fig. 2a). When the training size is increased to 10% and 20%, the RMSDs stay around 4.75 Å for run1 but decrease successively from 4.96 Å to 4.73 Å and 4.43 Å for run2. This decrease in reconstruction RMSD with increasing training size is likely because run2 was started from an all α-helical conformation, which mostly melted away over time (Fig. S1a). For Q15, we chose autoencoders trained at the 10% size for generating new conformations.

For Aβ40, training with the first 10% of the MD simulations results in reconstruction RMSDs of 6.4 ± 1.3 Å (mean ± standard deviation among four MD runs) (Fig. 2b). The reconstruction RMSDs decrease to 6.0 ± 1.4 Å with a 20% training size and further to 5.4 ± 1.1 Å with a 30% training size. The higher RMSD of run2 is probably due to more compact initial conformations (Fig. S1b). For this IDP we chose a 20% training size for generating new conformations.

Reconstruction becomes more challenging as the IDP size increases. This is already apparent when Aβ40 is compared to Q15, and is much more so for ChiZ, where training with 10% of the MD simulations results in reconstruction RMSDs at 8.3 ± 1.1 Å for 10 of the 12 MD runs, and >10 Å for the other two runs (Fig. 2c). Still, the reconstruction RMSDs decrease to 7.4 ± 1.3 Å with a 20% training size and further down to 6.4 ± 1.0 Å with a 30% training size. For this larger IDP, we chose 30% training size (corresponding to 870 ns of simulation time) for generating new conformations.

To check whether the dimensions of the latent space chosen according to $0.75N_{res}$ were adequate, we trained autoencoders with a 200-dimentional latent space. The reconstruction RMSDs improve for Q15 and Aβ40, but not for ChiZ (Fig. S2). So increasing the latent-space dimension does not necessarily improve accuracy, especially for the larger, more challenging IDPs, in reconstruction (or in generating new conformations; see below).

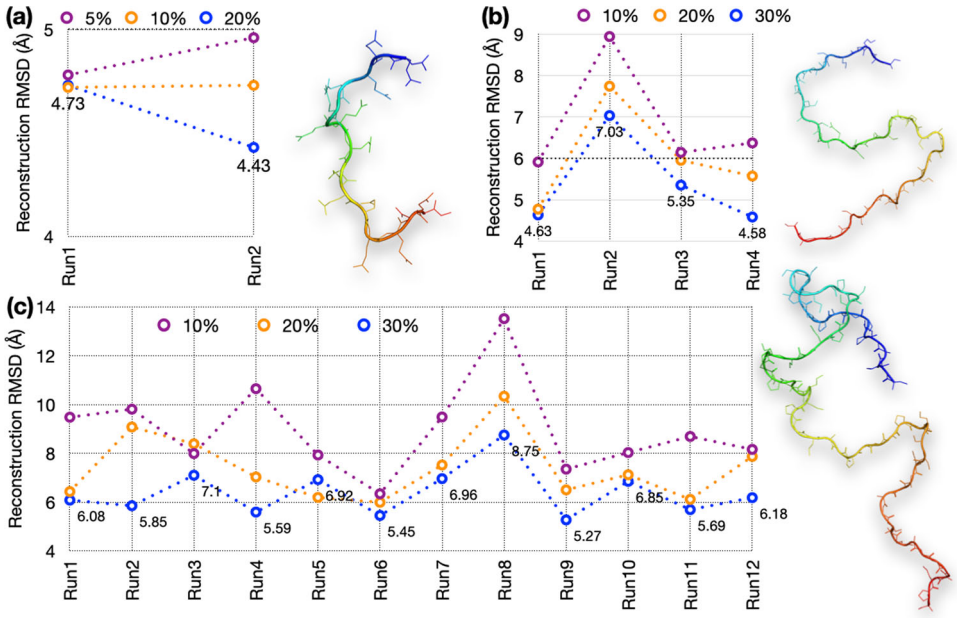

Fig. 2 Average reconstruction RMSDs at different sizes of the training sets sampled from replicate MD runs. a Q15 at 5%, 10%, and 20% training sizes from two runs. **b** Aβ40 at 10%, 20%, and 30% training sizes from four runs. **c** ChiZ at 10%, 20%, and 30% training sizes from 12 runs. A structure for each IDP is shown.

We tested autoencoders where the input was dihedral angles or distance matrices instead of Cartesian coordinates. The performance of these models in reconstruction was much worse than that with Cartesian coordinates as input (Supplementary Note 1).

Multivariate Gaussian models in latent space. The conformational ensembles of IDPs are broad and difficult to model[14]. A possible crucial benefit of representing the conformations in the latent space is that, due to the reduced dimensionality, the distribution of the latent vectors would be more compact and therefore easier to model. To assess this expectation, we calculated histograms in two-dimensional subspaces of the latent space. For each autoencoder, about one half of the encoder output values were consistently at or near zero, thereby further reducing the effective dimension of the latent space. We only calculated histograms for pairs of nonzero output neurons.

For the run1 training set of Q15, only 7 of the 13 output neurons were nonzero, resulting in 21 possible pairs. In Fig. S3, we display the histograms of 10 pairs calculated for the training (10% size) and test datasets. These histograms are indeed compact. Moreover, the counterparts of the training and test sets look very similar, with only minor differences for one or two pairs. For example, in the (9, 11) pair, the histogram of the training set is somewhat broader than the counterpart of the test set. The substantial overlap between the distributions of the training and test sets in the latent space explains the good performance of the autoencoder in reconstruction.

The autoencoder for Aβ40 (run1; 20% training size) had only 15 nonzero output neurons (out of 30). Fig. 3 displays the histograms of 8 nonzero pairs. All of these are single-peaked, and the peak positions are the same for the training and test counterparts in most cases, but with some shift for the (0, 27) pair. The high-level of overlap between the training and test sets allows for the satisfactory reconstruction of Aβ40 conformations reported above. In comparison, for the larger ChiZ, the histograms representing conformations sampled from a single MD run (run1) become irregular in shape (e.g., the (38, 39) pair) and the divergence between the training and test sets becomes prominent (e.g., the (15, 16) and (44, 47) pairs) (Fig. S4). These features exhibited by the distributions in the latent space illustrate the growing difficulty in reconstructing the conformations of larger IDPs.

The compact distributions of Q15 and Aβ40 in the latent space motivated us to model them as multivariate Gaussians. As shown in Figs. S3 and 3, the distributions of the training sets and their multivariate Gaussian models look very similar. More importantly, the multivariate Gaussian models also overlap well with the distributions of the test sets. Indeed, the overlap between the test sets and the Gaussian models is greater than that between the test sets and the corresponding training sets, as illustrated by the (9, 11) pair of Q15 and the (0, 3) pair of Aβ40. Therefore the multivariate Gaussian models seem promising for generating new conformations that are similar to those in the test sets of Q15 and Aβ40. For ChiZ, multivariate Gaussians are inadequate to model the irregular shapes of the single-run distributions in the latent space (Fig. S4).

The foregoing qualitative observations are confirmed by calculating the Kullback-Leibler (KL) divergence between the Gaussian models and the distributions of the training and test data in the latent space (Table S1). For both Q15 and Aβ40, the Gaussian models provide good representations of the training data, with KL divergence values at or below 0.1 for all the pairs shown in Figs. S3 and 3. Moreover, for all but one pair, the KL divergence values between the test data and Gaussian models are lower than those between the training data and test data. For example, for the (9, 11) pair of Q15, the KL divergence decreases from 0.10 for training vs test to 0.06 for test vs Gaussian; for the (0, 3) pair of Aβ40, the KL divergence decreases from 0.58 for training vs test to 0.38 for test vs Gaussian. On the other hand, for ChiZ, the Gaussian model provides a poor representation of the training data, with KL divergence as high as 0.48 (for the (0, 4) pair).

Autoencoder-generated conformations of Q15 and Aβ40. By sampling from a multivariate Gaussian in the latent space and using the decoder to reconstructing conformations, we turned the autoencoder into a generative model. The multivariate Gaussian

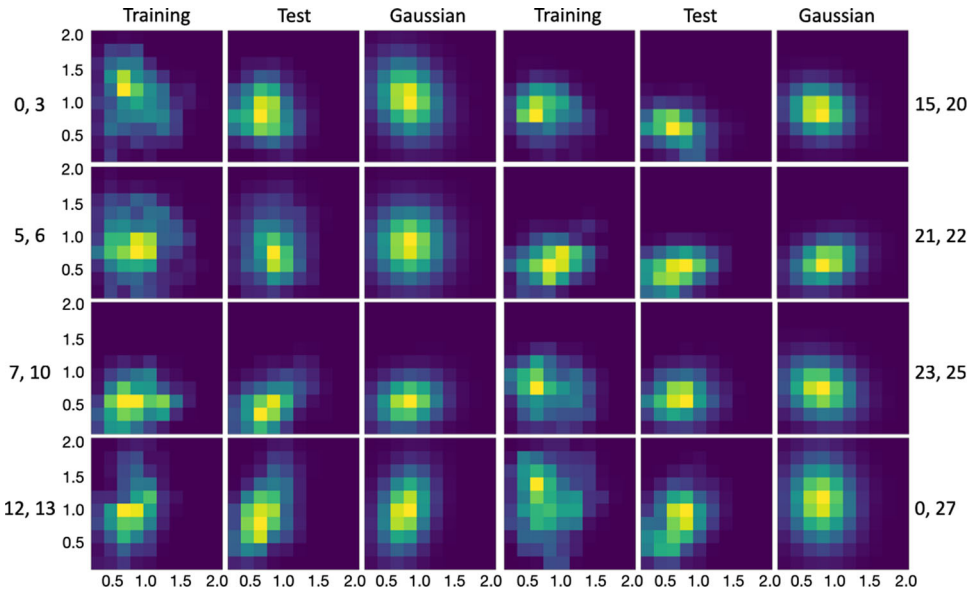

Fig. 3 Histograms of Aβ40 in the latent space, calculated from training data, test data, and multivariate Gaussian. Histograms for pairs of encoder nonzero outputs from run1 are shown as heat maps, with yellow representing pixels with the highest counts and dark blue representing pixels with 0 count.

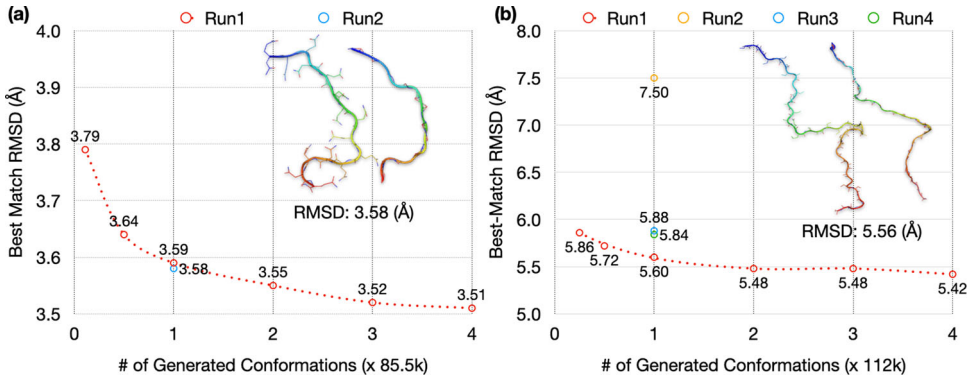

Fig. 4 Best-match RMSDs for autoencoder-generated conformations of Q15 and Aβ40. The average best-match RMSDs of 100-fold diluted test sets of (**a**) Q15 and (**b**) Aβ40, against generated sets at different sizes. The latter sizes are measured in multiples of the test size of each IDP (= 85,500 for Q15 and 112,000 for Aβ40). For run1, results are shown at sizes of the generated set ranging from the training size to 4×. For other MD runs, results are shown at 1×. In the inset of each panel, an IDP conformation and its generated best match, with an RMSD close to the average values at 1×, is compared.

was parameterized on the same dataset for training the autoencoder. For Q15, the training size was 9500 and the test size was 85,500. The size of the generated set was measured as multiples of the test size (1× = 85,500). For each conformation in the 100-fold diluted test set, we found its best match (i.e., lowest RMSD) in the generated set. We then used the average of the best-match RMSDs for the diluted test set as the measure for the accuracy of the generated set. With the generated sets at size 1×, the average best-match RMSDs of the test sets are 3.59 and 3.58 Å for MD run1 and run2, respectively. As illustrated in the inset of Fig. 4a, a test conformation and its generated best match at 3.58 Å RMSD show very similar backbone traces. Since generating new conformations by the autoencoder is extremely fast, the generated set can be easily expanded. With expanding sizes of the generated set, the average best-match RMSDs show small but systematic decreases, to 3.55 Å at 2×, 3.52 Å at 3×, and 3.51 Å at 4× for run1 (Fig. 4a). The improvement in RMSD occurs because the expanded size of the generated set yields better matches for the test conformations. Conversely, the average best-match RMSDs increase to 3.64 Å when the size of the generated set is reduced to 0.5× and further to 3.79 Å when the generated set is reduced to the same size as the training set (at 0.11×).

High accuracy is also achieved for generated conformations of Aβ40 on autoencoders trained with 20% (=28,000 conformations) of MD simulations (Fig. 4b). With the size of the generated sets at 1× (=112,000 conformations), the average best-match RMSDs of the 100-fold diluted test sets are 5.60 Å, 7.50 Å, 5.88 Å, and 5.84 Å, respectively, for MD run1 to run4. A test conformation and its generated best match at 5.56 Å RMSD show very similar backbone traces (Fig. 4b, inset). The higher average RMSD of the autoencoder for run2 in generating new conformations mirrors the poorer performance of this autoencoder in reconstruction (Fig. 2b), and can also be attributed to the overly compact conformations in the training set of this MD run (Fig. S1b). With an expansion of the generated set, the average best-match RMSD shows a slight decrease, to 5.42 Å at 4× for run1 (Fig. 4b). Conversely, the average best-match RMSD increases to 5.72 Å at 0.5× and to 5.86 Å at 0.25× (=size of the training set).

Autoencoder-generated conformations of ChiZ. We first used a similar protocol to train and test an autoencoder for ChiZ on a single MD run (run1). The training size was 30% or 43,500 and the test size was 101,500. With the generated set at size 1×

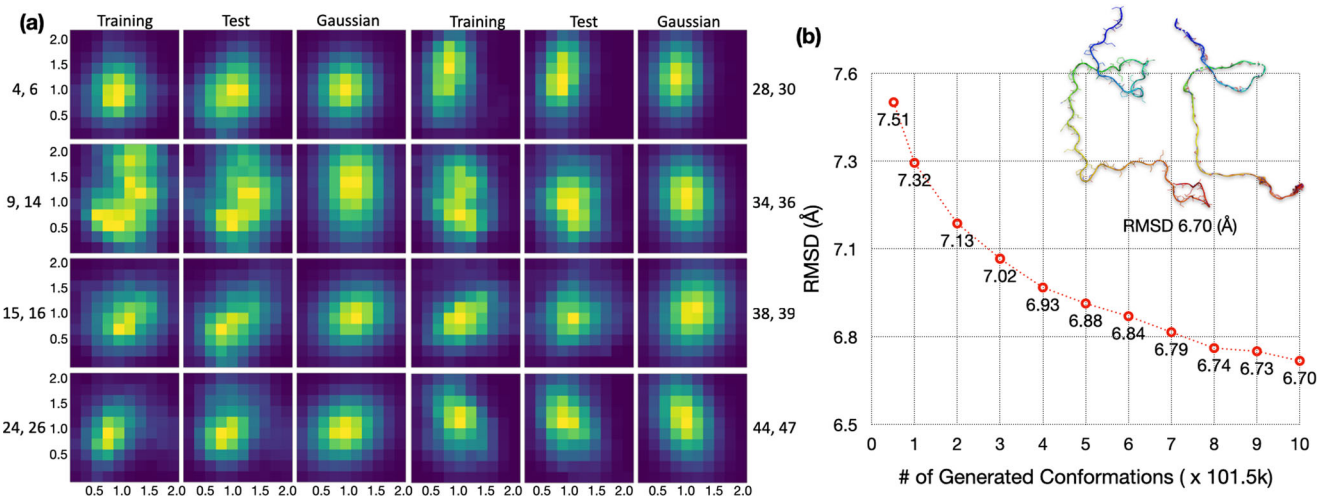

Fig. 5 Increased data overlap and prediction accuracy by combining MD runs of ChiZ. a Histograms in the latent space, shown as heat maps, with yellow representing pixels with the highest counts and dark blue representing pixels with 0 count. Histograms were calculated for pairs of nonzero elements, using 52,200, 121,800, and 101,500 vectors from the training and test sets and the multivariate Gaussian, respectively. The training and test sets were from combining conformations sampled in 12 MD runs; the multivariate Gaussian was parameterized on the combined training set. **b** The average best-match RMSDs of the 1000-fold diluted, combined test set against generated sets at different sizes. The autoencoder was trained on a 10-folded diluted, combined training set (size = 52,200) from all the 12 MD runs. The sizes of the generated sets are measured in multiples of the test size in a single MD run (=101,500), and range from 0.51× (=training size) to 10×. The inset displays an IDP conformation and its generated best match, with an RMSD of the average value at 10×.

(=101,500 conformations), the average best-match RMSD of the 100-fold diluted test set is 7.95 Å (Fig. S5a). Again the RMSD decreases slightly with expanding sizes of the generated set, but is still 7.35 Å even at size 12× (=1.2 million conformations). The high RMSD of the autoencoder trained on a single MD run is presaged by the inadequate modeling of the training data by a multivariate Gaussian in the latent space (Fig. S4 and Table S1). One idea for improving the modeling is to represent the training data in the latent space by a mixture of multiple Gaussians. We tested this idea (Supplementary Note 2). The multiple-Gaussian model indeed improves the representation of the training data, but actually does worse in predicting the test conformations. For example, with 8 Gaussians, the best-match RMSD of a generated set at size 1× increases from 7.95 Å to 8.50 Å. In essence, as the model tries to fit into the details of the training data, its ability to capture generic features shared by the test data suffers.

It is possible that a single MD run may mine a limited region in conformational space, but the regions mined by different MD runs may partially overlap and the combined mining may generate an ensemble that is densely distributed in the latent space. Indeed, when we combine the conformations from 12 MD runs for ChiZ, the histograms in the latent space for both the training set and the test set become compact and have a single peak for all but one (i.e., (9, 14)) of the nonzero pairs (Fig. 5a). The distributions of the training and test latent vectors overlap very well and are also modeled well by the multivariate Gaussian parameterized on the combined training set. The KL divergence values for training vs Gaussian, test vs Gaussian, and training vs test are all lower than 0.1 for all the pairs (Table S1); the value for training vs Gaussian is only 0.079 even for the (9, 14) pair.

The increase in overlap by combining data from multiple MD runs pointed a way to improve autoencoders. As an initial test, we pooled the generated conformations (each at size 1×) from the autoencoders of the individual MD runs. When compared with this pooled generated set (total size at 12×), the average best-match RMSD of the run1 test set is 7.04 Å (Fig. S5b), which is lower by 0.31 Å than the corresponding value when the generated set is at the same 12× size but produced solely by the run1 autoencoder (Fig. S5a). To take full advantage of the multiple MD runs of ChiZ, we used the autoencoder trained on the combined training set (a total of 52,200 conformations after a 10-fold dilution) to generate new conformations. The generated set at size 1× now gives a best-match RMSD of 7.32 Å for the 1000-fold diluted, combined test set (final size = 1218). When the generated set is expanded to a size 10×, the best-match RMSD reduces to 6.70 Å (Fig. 5b). The inset illustrates a pair of conformations, one from the test set and one from the generated set, at this RMSD.

Optimum selection of training sizes and latent-space dimensions. In Supplementary Note 3, we present additional data for the effects of varying training size and latent-space dimension on the accuracy of autoencoders in generating new conformations. In short, the selected training sizes, 10%, 20%, and 30% respectively, for Q15, Aβ40, and ChiZ, are sufficient for model convergence; additional training data do not yield appreciable gains in model accuracy, especially given that we put a premium on cost control of MD simulations. We selected $0.75N_{res}$ as the latent-space dimension. Increasing the latent-space dimension by 10-30 has little effect on model accuracy. For Q15, a very large value, 200, for the latent-space dimension actually leads to slight increases in the best-match RMSDs of generated conformations (Fig. S6, compared with Fig. 4a).

Further assessment of generated conformations. To properly benchmark the autoencoder-generated conformations, we examined the diversity of the test sets and the similarity between the training and test sets (Table S2). We calculated the RMSDs of each conformation with all others in a diluted test set. The average pairwise RMSDs are quite high even within a single MD run (run1), 6.98 Å for Q15, 11.61 Å for Aβ40, and 18.21 Å for ChiZ, showing that the conformations in each test set are very diverse. As expected, the average pairwise RMSD increases further, to 19.23 Å, for the combined and further diluted test set of ChiZ. The diversity of the test conformations again illustrates the challenge in generating conformations that are close to them.

The neighboring conformations in any MD run have relatively low RMSDs, leading to small best-match RMSDs between conformations in the test sets from single MD runs. The average best-match RMSDs in run1 are 3.71 Å for Q15, 3.83 Å for Aβ40, and 4.83 Å for ChiZ (Table S2). However, for the combined and further diluted test set of ChiZ, the average best-match RMSD increases to 8.62 Å. The latter value may be viewed as a benchmark for generated conformations to be claimed as neighbors of test conformations. Because the average best-match RMSD for the combined test set against the generated set (at size 10×) is 6.70 Å, or nearly 2 Å below the benchmark, we can claim that all the test conformations in the combined test set have neighbors in the generated set. In other words, the generated set fully covers the combined test set.

Another benchmark is given by the average best-match RMSD between a test set and the corresponding training set. For run1, values of this benchmark are 3.96 Å for Q15, 6.76 Å for Aβ40, and 10.17 Å for ChiZ (Table S2). When the comparison is against the generated sets at the sizes of the training sets (shown as the first point in Figs. 4a, b, and S5a), the average best-match RMSDs are 3.79, 5.86, and 8.16 Å, respectively, each of which is lower than the counterpart when the comparison is against the training set itself. That is, relative to the training sets, the generated sets provide better matches for the test sets. For Q15 and Aβ40, this outcome is to be expected because of the above observation that the test sets overlap better with the Gaussian models than with the training sets (Figs. S3 and 3; Table S1). For ChiZ, the combined test set from the 12 MD runs has a best-match RMSD of 8.47 Å against the combined training set, which is 1.7 Å lower than the counterpart for the comparison within run1. This decrease in best-match RMSD confirms the aforementioned increase in data overlap when multiple MD runs are combined (Figs. S4 and 5a). Moreover, the best-match RMSD of the combined test set further reduces to 7.51 Å when the generated set is of the same size as and parameterized on the combined training set (first point in Fig. 5b).

We also inspected more closely the generated conformations that best match test conformations (insets in Figs. 4a, b, and 5b). As already alluded to, test conformations and their generated best matches show overall similarities in shape and size. However, the generated conformations have considerable bond length and bond angle violations. Refinement by energy minimization restores essentially all bonds and angles to proper values (Fig. 6). The refinement results in small increases in RMSD for the best-matched test conformations, though an occasional decrease in RMSD is possible. For the pairs of conformations shown in the insets of Figs. 4a, b, and 5b, the RMSDs change from 3.58 Å to 3.45 Å, from 5.56 Å to 5.95 Å, and from 6.67 Å to 6.87 Å, respectively (Fig. 6). For the generated set of ChiZ at size 1×, the best-match RMSD increases from 7.32 Å to 7.66 Å upon conformational refinement.

Experimental validation of autoencoder-generated ChiZ conformational ensemble. To objectively assess the quality of the autoencoder-generated conformational ensemble, we calculated from it properties that can be measured experimentally. These include SAXS profile and NMR chemical shifts. In Fig. 7, we compare the experimental data for ChiZ[14] with results calculated from 12,180 conformations collected from the combined test set of the 12 MD runs, and with results calculated from 12,180 conformations generated by the autoencoder trained on the combined training set. As reported previously[14], the MD simulations reproduced both types of experimental data well: there was very good agreement for the SAXS profile over the entire q (momentum transfer) range from 0 to 0.5 $Å^{-1}$, with a mean absolute percentage error (MAPE) of 3.9%; likewise the calculated secondary chemical shifts were close to the experimental values, with a root-mean-square error (RMSE) of 0.43 ppm. The experimental SAXS profile is also reproduced well by the generated conformations, with an MAPE of 7.2%, validating the latter's sampling of the overall shape and size of ChiZ, though some deviations are seen at the high q end. For secondary chemical shifts, the RMSE increases to 0.63 ppm for the generated conformations. This RMSE is at the low end of the range of RMSEs (0.63 to 0.84 ppm) calculated on conformations from MD simulations using four other force fields[14]. Autoencoders trained on conformations from these other force fields have similar performances as the one reported above for ChiZ, demonstrating the robustness of the approach (Supplementary Note 4).

Discussion

We have developed generative autoencoders to mine the broad conformational space of IDPs. These autoencoders can not only represent IDP conformations in the latent space with high fidelity to allow for accurate reconstruction, but also generate new conformations to fill up the conformational space. The generated ensemble contains close matches for all the conformations sampled in long MD simulations, but with negligible computational time. For example, sampling 100,000 conformations (at 20 ps intervals) from MD simulations of Aβ40, even with GPU acceleration[24], takes 80 days, whereas our autoencoder generates the same number of conformations in 12 sec. In the case of ChiZ, the autoencoder-generated conformations even yielded better predictions for SAXS profile and chemical shifts than MD simulations with several force fields.

Our generative autoencoders have the flavor of variational autoencoders but are more intuitive. Rather than optimizing Gaussians in the latent space during the training process as in variational autoencoders, we only optimize reconstruction and then use the latent vectors of the training set to calculate the mean vector and covariance matrix, which are directly used to define a multivariate Gaussian for generating new conformations. We have shown that the difficulty posed by the longer sequence length of ChiZ can be overcome by training on data sampled from multiple MD runs. As the lengths of IDPs increase, the problem becomes even more challenging. One possible way to address this challenge is to break a long IDP into fragments and treat each fragment as a separate IDP. However, IDPs do form occasional long-range contacts[14,25]. The influence of long-range contacts has to be somehow taken into consideration.

The generative autoencoders designed here are for mining the conformational space of IDPs in isolation. The power of this approach demonstrated here suggests that it can be extended to study IDPs in more complex functional states, such as when bound to or associated with an interaction partner (a target protein or a membrane), or in aggregation. For example, ChiZ associated with acidic membranes has been studied by long MD simulations[15]; generative autoencoders may also be able to mine the conformational space of membrane-associated IDPs. IDPs are prone to phase separation[26], resulting in a highly concentrated phase surrounded by a dilute phase. Microsecond-long MD simulations failed to sample the equilibration between the two phases[27]. AI-based models such as generative autoencoders may open the door to solving this and other challenging conformational mining problems for IDPs.

Computational methods

Autoencoder design. We built and trained the autoencoders using the Keras package (https://keras.io/) with TensorFlow backend (https://www.tensorflow.org/) in Python 3.6[28]. The autoencoders

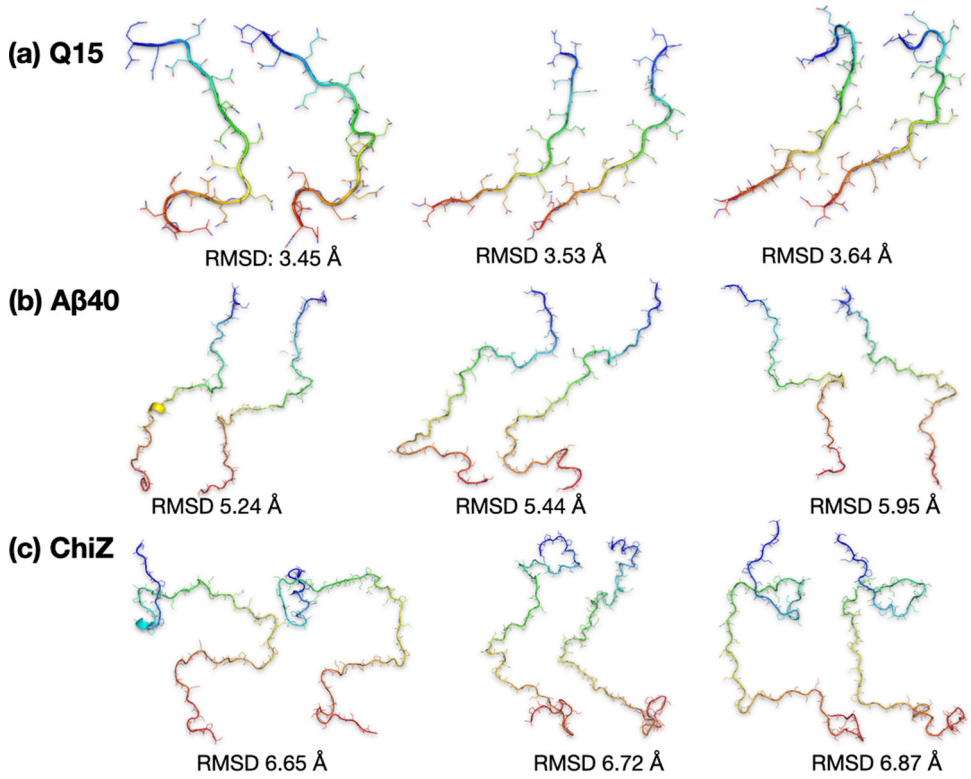

Fig. 6 Comparison of test conformations and their generated best matches after refinement. a Q15. **b** Aβ40. **c** ChiZ. Each IDP is represented by three pairs of conformations with RMSDs around the average best-match value of the diluted test set against the final generated set. In each pair, the left conformation is from the test set and the right conformation is from the generated set, after refinement. The unrefined versions of the first pair in **a**, the third pair in **b**, and the third pair in **c** are shown in Figs. 4a, b, and 5b, respectively.

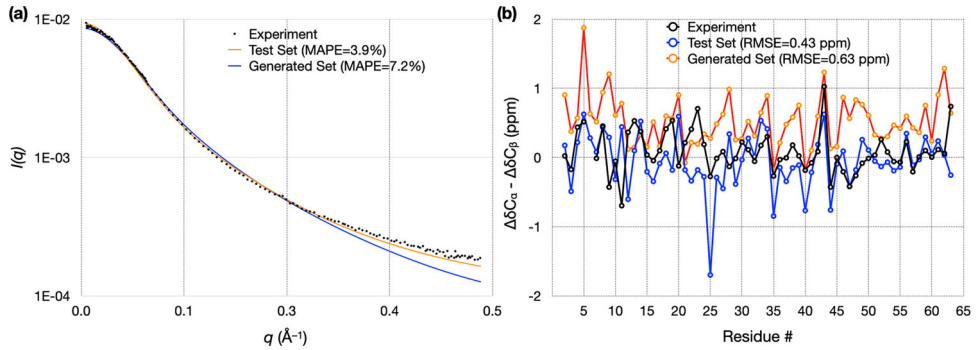

Fig. 7 Validation of autoencoder-generated conformations for ChiZ by experimental SAXS and chemical shift data. a Comparison of experimental and calculated SAXS profiles. MAPE was calculated as $<|O_i - E_i|/E_i>_{\text{all data points}} \cdot 100\%$, where E_i and O_i are experimental and predicted scattering intensities, respectively, and $<\cdots>_{\text{all data points}}$ denotes the average over all the data points. **b** Comparison of experimental and calculated secondary chemical shifts. The experimental data and the MD simulations are reported previously[14]. RMSE was calculated as the square root of $<(O_i - E_i)^2>_{\text{all data points}}$. Calculations were done on either the test set comprising 12,180 conformations sampled from 12 MD runs, or on an autoencoder-generated set comprising the same number of conformations, after refinement. The autoencoder was trained on a combined training set comprising 52,200 conformations sampled from the 12 MD runs.

consisted of an encoder and a decoder. Both the encoder and decoder had a dense neural network architecture, with two hidden layers of 300 and 50 neurons, respectively. The input, hidden, and output layers of the encoder and decoder were arranged as mirror images of each other (Fig. 1a). This arrangement was chosen based on its reduced training complexity as shown in previous reconstruction work on structured proteins[6]. All layers except for the final output layer had a rectified linear unit activation function; the final output layer had a sigmoidal activation function.

The input to the encoder consisted of the Cartesian coordinates of an IDP. Only heavy atoms (all for the backbone and selected for side chains) were included; selected side-chain atom types were CB, CG, CD, OE1, and NE2. This selection contained all the heavy atoms of polyglutamine Q15, but truncated some of the side chains in Aβ40 and ChiZ. Q15, Aβ40, and ChiZ had $N = 140$, 230, and 385 heavy atoms, respectively, for a total of $3N$ input coordinates. The loss function was the binary cross-entropy,

$$H\left(\{y_i\},\{y_i'\}\right) = \frac{1}{3N}\sum_{i=1}^{3N}\left[-y_i \ln y_i' - (1-y_i)\ln\left(1-y_i'\right)\right] \quad (1)$$

where $\{y_i\}$ denotes the $3N$ input Cartesian coordinates of the IDP after a linear transformation into the range between 0 and 1 (see below), and $\{y_i'\}$ denotes the values of the corresponding output neurons. The neural networks were trained by the Adam optimizer given its effectiveness in handling large datasets. For each autoencoder, training was done for 100 epochs using a batch size of 40. Using the mean square error as the loss function produced very similar accuracy in generating new conformation (Supplementary Note 5).

The latent space dimension and training size were tested based on reconstruction, which entailed encoding (i.e., representing the conformations as vectors in the latent space) and then decoding (i.e., constructing back full conformations from the latent vectors). The dimensions of the latent spaces for the three IDPs were finally chosen as $n = 13$, 30, and 48. Parameters of autoencoders trained on reconstruction were saved in decoder and encoder files, and the decoder was then used to generate new conformations.

Molecular dynamics simulations. Two 1 μs trajectories (100,000 frames each, saved at 10 ps intervals) for Q15, taken from Hicks and Zhou[23], were run at 298 K in GROMACS with the AMBER03ws force field[29] for protein and TIP4P2005 for water[30]. These simulations were performed using an enhanced sampling method called replica exchange with solute tempering[31,32] at constant volume and temperature, with temperature regulated by velocity rescaling[33]. The simulations were judged to be well equilibrated, as shown in particular by the agreement in the distribution of radius of gyration with simulations using a second enhanced sampling method, i.e., temperature replica exchange[34].

For ChiZ, 12 trajectories of 3 μs each (150,000 frames, saved at 20 ps intervals), taken from Hicks et al.[14], were run on GPUs using pmemd.cuda[24] in AMBER18[35] with the ff14SB force field[36] for protein and TIP4PD[37] for water. These simulations were performed at constant temperature (300 K) and pressure 1 atm), with temperature regulated by the Langevin thermostat (damping constant at 3 ps^{-1})[38] and pressure regulated by the Berendsen barostat (pressure relaxation time at 2 ps)[39]. These simulations were thoroughly validated by experimental data including SAXS, chemical shifts, and NMR relaxation properties. Additional simulations were performed using four other protein/water force field combinations, including AMBER03ws/TIP4P2005, AMBER99SB-ILDN[40]/TIP4PD, AMBER15IPQ/SPCEb[41], and CHARMM36m/TIP3Pm[42]. The protocol for ChiZ was used to run four replicate simulations of Aβ40 at 278 K (3.5 μs each; 175,000 frames saved at 20 ps intervals)[25]. Again the simulations were thoroughly validated by experimental data including chemical shifts and NMR relaxation properties.

Data preprocessing. MD trajectories in GROMACS and AMBER trajectory formats were first converted to conformations in PDB format (with solvent stripped). An initial portion of each trajectory (5000, 35000, and 5000 frames for Q15, Aβ40, and ChiZ, respectively) were removed. The remaining trajectory was split into two parts, the first (e.g., 10%) as the training dataset and the second as the test dataset.

The Biobox library in Python (https://github.com/degiacom/biobox)[6] was used to preprocess the coordinates in each dataset. All the frames were aligned to the first one according to RMSD, and shifted to have all coordinates positive. Coordinates were then scaled between 0 and 1 (via dividing by the maximum coordinate value) for using as input to the encoder. The output coordinates of the decoder were scaled back to real coordinates using the same scaling factor. The choice of the reference frame for the structural alignment before shifting and scaling the coordinates had no effect on the accuracy in generating new conformations (Supplementary Note 5).

RMSD calculation. We used a code of Ho (https://boscoh.com/protein/rmsd-root-mean-square-deviation.html) to calculate RMSDs of output conformations. A custom Python code (https://github.com/aaayushg/generative_IDPs/tree/main/RMSD) was written to find the lowest RMSD between a given test conformation against a set of generated conformations, and calculate the average of these best-match RMSDs for the test set (100-fold diluted).

Generating new conformations. The mean vector $\boldsymbol{\mu}$ with elements

$$\mu_l = <z_l>_{\text{training}} \tag{2}$$

and covariance matrix $\widetilde{\sigma}$ with elements

$$\sigma_{lm} = <(z_l - \mu_l)(z_m - \mu_m)>_{\text{training}} \tag{3}$$

were calculated from the latent vectors, $\{z_l\}$, of the training dataset; here $<\cdots>_{\text{training}}$ denotes an average over the training set. The latter two quantities in turn defined a multivariate Gaussian distribution (Fig. 1b),

$$Q(\mathbf{z}) = \frac{1}{\sqrt{(2\pi)^n \det\widetilde{\sigma}}} e^{-(\mathbf{z}-\boldsymbol{\mu})^{\mathrm{T}}\cdot\widetilde{\sigma}\cdot(\mathbf{z}-\boldsymbol{\mu})} \tag{4}$$

from which vectors were sampled and fed to the decoder to generate new conformations (Fig. 1c). In the above, det represents determinant of a matrix, and the superscript "T" signifies transpose. Sampling from multivariate Gaussians was implemented using the NumPy library (https://numpy.org/) in Python. Histograms were calculated in two-dimensional subspaces of the latent space, for qualitative comparison among the training, test, and multivariate Gaussian datasets (https://github.com/aaayushg/generative_IDPs/tree/main/Plot_histogram).

We used the Kullback-Leibler divergence

$$D_{\mathrm{KL}}(p|q) = \int\int dz_l dz_m p(z_l, z_m) \ln \frac{p(z_l, z_m)}{q(z_l, z_m)} \tag{5}$$

to quantify the difference between two distributions, $p(z_l, z_m)$ and $q(z_l, z_m)$, in a two-dimensional subspace of the latent space. $p(z_l, z_m)$ and $q(z_l, z_m)$ are proportional to the histograms but are normalized. The integral was evaluated as a summation over the two-dimensional grid over which the histograms were calculated (see, e.g., Fig. 3). For any grid point where either $p(z_l, z_m)$ or $q(z_l, z_m)$ was 0, the contribution from that grid point to $D_{\mathrm{KL}}(p|q)$ was set to 0.

Refinement of autoencoder-generated conformations. The generated conformations had considerable bond length and bond angle violations. We used a simple procedure to remedy this problem. First all the missing heavy and hydrogen atoms were added in each structure using tleap in AmberTools[35]. Then the structure was subject to 500 steps of conjugate-gradient energy minimization in vacuum using NAMD 2.13[43]. The protein force field was AMBERff14SB[36].

Calculation of SAXS profile and chemical shifts for ChiZ. The SAXS profile for each conformation was then calculated using FoXS[44] and scaled to optimize agreement with the experimental profile[14]. Chemical shifts were calculated using SHIFTX2[45] (www.shiftx2.ca). Chemical shifts for random-coil conformations calculated using POTENCI[46] were subtracted to obtain secondary Cα and Cβ chemical shifts. SAXS profiles and chemical shifts

were averaged over all the conformations in the diluted test set (12180 frames from 12 trajectories) or a generated set (of the same size). For the latter, the conformations after refinement by energy minimization were used.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Saved autoencoder models and example data are available on GitHub: https://github.com/aaayushg/generative_IDPs.

Code availability

The implementation codes and tutorials are available on GitHub: https://github.com/aaayushg/generative_IDPs.

Received: 14 December 2021; Accepted: 7 June 2022;
Published online: 20 June 2022

References

1. Fleming, N. How artificial intelligence is changing drug discovery. *Nature* **557**, S55–S57 (2018).
2. Callaway, E. 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. *Nature* **588**, 203–204 (2020).
3. Hermann, J., Schatzle, Z. & Noe, F. Deep-neural-network solution of the electronic Schrodinger equation. *Nat. Chem.* **12**, 891–897 (2020).
4. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
5. Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
6. Degiacomi, M. T. Coupling molecular dynamics and deep learning to mine protein conformational space. *Structure* **27**, 1034–1040 (2019).
7. Noé, F., Olsson, S., Köhler, J. & Wu, H. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science* **365**, eaaw1147 (2019).
8. Jin, Y., Johannissen, L. O. & Hay, S. Predicting new protein conformations from molecular dynamics simulation conformational landscapes and machine learning. *Proteins* **89**, 915–921 (2021).
9. Moritsugu, K. Multiscale enhanced sampling using machine learning. *Life (Basel)* **11**, 1076 (2021).
10. Xue, B., Dunker, A. K. & Uversky, V. N. Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. *J. Biomol. Struct. Dyn.* **30**, 137–149 (2012).
11. Uversky, V. N. Introduction to Intrinsically Disordered Proteins (IDPs). *Chem. Rev.* **114**, 6557–6560 (2014).
12. Ross, C. A. Polyglutamine pathogenesis: emergence of unifying mechanisms for Huntington's disease and related disorders. *Neuron* **35**, 819–822 (2002).
13. Sadigh-Eteghad, S. et al. Amyloid-beta: A crucial factor in Alzheimer's disease. *Med Princ. Pr.* **24**, 1–10 (2015).
14. Hicks, A., Escobar, C. A., Cross, T. A. & Zhou, H. X. Sequence-dependent correlated segments in the intrinsically disordered region of ChiZ. *Biomolecules* **10**, 946 (2020).
15. Hicks, A., Escobar, C. A., Cross, T. A. & Zhou, H. X. Fuzzy association of an intrinsically disordered protein with acidic membranes. *JACS Au* **1**, 66–78 (2021).
16. Kachala, M., Valentini, E. & Svergun, D. I. Application of SAXS for the Structural Characterization of IDPs. *Adv. Exp. Med Biol.* **870**, 261–289 (2015).
17. Jensen, M. R., Zweckstetter, M., Huang, J.-R. & Blackledge, M. Exploring free-energy landscapes of intrinsically disordered proteins at atomic resolution using NMR spectroscopy. *Chem. Rev.* **114**, 6632–6660 (2014).
18. Kukharenko, O., Sawade, K., Steuer, J. & Peter, C. Using dimensionality reduction to systematically expand conformational sampling of intrinsically disordered peptides. *J. Chem. Theory Comput* **12**, 4726–4734 (2016).
19. Ramanathan, A., Ma, H., Parvatikar, A. & Chennubhotla, C. S. Artificial intelligence techniques for integrative structural biology of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.* **66**, 216–224 (2020).
20. Bhowmik, D., Gao, S., Young, M. T. & Ramanathan, A. Deep clustering of protein folding simulations. *BMC Bioinform.* **19**, 484 (2018).
21. Hernández, C. X., Wayment-Steele, H. K., Sultan, M. M., Husic, B. E. & Pande, V. S. Variational encoding of complex dynamics. *Phys. Rev. E.* **97**, 062412 (2018).
22. Mardt, A., Pasquali, L., Wu, H. & Noé, F. VAMPnets for deep learning of molecular kinetics. *Nat. Commun.* **9**, 5 (2018).
23. Hicks, A. & Zhou, H. X. Temperature-induced collapse of a disordered peptide observed by three sampling methods in molecular dynamics simulations. *J. Chem. Phys.* **149**, 072313 (2018).
24. Salomon-Ferrer, R., Götz, A. W., Poole, D., Le Grand, S. & Walker, R. C. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *J. Chem. Theory Comput* **9**, 3878–3888 (2013).
25. Dey, S., MacAinsh, M. & Zhou, H.-X. Sequence-dependent backbone dynamics of intrinsically disordered proteins. *bioRxiv*, https://doi.org/10.1101/2022.02.11.480128 (2022).
26. Zhou, H. X., Nguemaha, V., Mazarakos, K. & Qin, S. Why do disordered and structured proteins behave differently in phase separation? *Trends Biochem Sci.* **43**, 499–516 (2018).
27. Zheng, W. et al. Molecular details of protein condensates probed by microsecond long atomistic simulations. *J. Phys. Chem. B* **124**, 11671–11679 (2020).
28. Abadi, M. et al. TensorFlow: A system for large-scale machine learning. In: *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16).* USENIX Association, Savannah, GA (2016).
29. Best, R. B., Zheng, W. & Mittal, J. Balanced protein–water interactions improve properties of disordered proteins and non-specific protein association. *J. Chem. Theory Comput.* **10**, 5113–5124 (2014).
30. Abascal, J. L. F. & Vega, C. A general purpose model for the condensed phases of water: TIP4P/2005. *J. Chem. Phys.* **123**, 234505 (2005).
31. Liu, P., Kim, B., Friesner, R. A. & Berne, B. J. Replica exchange with solute tempering: A method for sampling biological systems in explicit water. *Proc. Natl Acad. Sci.* **102**, 13749–13754 (2005).
32. Terakawa, T., Kameda, T. & Takada, S. On easy implementation of a variant of the replica exchange with solute tempering in GROMACS. *J. Comput Chem.* **32**, 1228–1234 (2011).
33. Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **126**, 014101 (2007).
34. Mitsutake, A., Sugita, Y. & Okamoto, Y. Generalized-ensemble algorithms for molecular simulations of biopolymers. *Biopolymers* **60**, 96–123 (2001).
35. Case, D. A. et al. AMBER 2018, University of California, San Francisco (2018).
36. Maier, J. A. et al. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **11**, 3696–3713 (2015).
37. Piana, S., Donchev, A. G., Robustelli, P. & Shaw, D. E. Water dispersion interactions strongly influence simulated structural properties of disordered protein states. *J. Phys. Chem. B.* **119**, 5113–5123 (2015).
38. Pastor, R. W., Brooks, B. R. & Szabo, A. An analysis of the accuracy of langevin and molecular dynamics algorithms. *Mol. Phys.* **65**, 1409–1419 (1988).
39. Berendsen, H. J. C., Postma, J. P. M., Vangunsteren, W. F., Dinola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684–3690 (1984).
40. Lindorff-Larsen, K. et al. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **78**, 1950–1958 (2010).
41. Debiec, K. T. et al. Further along the Road Less Traveled: AMBER ff15ipq, an Original Protein Force Field Built on a Self-Consistent Physical Model. *J. Chem. Theory Comput.* **12**, 3926–3947 (2016).
42. Huang, J. et al. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods.* **14**, 71–73 (2017).
43. Phillips, J. C. et al. Scalable molecular dynamics on CPU and GPU architectures with NAMD. *J. Chem. Phys.* **153**, 044130 (2020).
44. Schneidman-Duhovny, D., Hammel, M., Tainer, J. A. & Sali, A. FoXS, FoXSDock and MultiFoXS: Single-state and multi-state structural modeling of proteins and their complexes based on SAXS profiles. *Nucleic Acids Res.* **44**, W424–W429 (2016).
45. Han, B., Liu, Y., Ginzinger, S. W. & Wishart, D. S. SHIFTX2: significantly improved protein chemical shift prediction. *J. Biomol. NMR.* **50**, 43–57 (2011).
46. Nielsen, J. T. & Mulder, F. A. A. POTENCI: prediction of temperature, neighbor and pH-corrected chemical shifts for intrinsically disordered proteins. *J. Biomol. NMR.* **70**, 141–165 (2018).

Acknowledgements

This work was supported by National Institutes of Health Grant GM118091.

Author contributions

A.G. and H.X.Z designed research; A.G., S.D., A.H. conducted research; A.G., S.D., and H.X.Z. analyzed data; A.G. and H.X.Z. wrote manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-022-03562-y.

Correspondence and requests for materials should be addressed to Huan-Xiang Zhou.

Peer review information *Communications Biology* thanks Arvind Ramanathan and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Gene Chong.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.