Structures of MERS-CoV Macro Domain in Aqueous Solution with Dynamics: Coupling Replica Exchange Molecular Dynamics and Deep Learning at the Nano Level

Ibrahim Akbayrak [1], Burak Ulver [2], Havvanur Dervisoglu [2], Mehmet Haklidir [2], Sule Caglayan [3], Lukasz Kurgan[4], Vladimir Uversky[5], Orkun Hasekioglu [2], and Orkid Coskuner-Weber[3]

[1]Turkish-German University
[2]TUBITAK Informatics and Information Security Research Center
[3]Turkish-German University Faculty of Science
[4]Virginia Commonwealth University
[5]University of South Florida

August 3, 2020

Abstract

A novel virus, severe acute respiratory syndrome Coronavirus 2 (SARS-CoV-2), causing coronavirus disease 2019 (COVID-19) worldwide appeared in 2019. Currently, we do not have a medicament that treats the disease. One of the rea-sons for the absence of treatment is related to the scarcity of detailed scientific knowledge of the members of the Coro-naviridae family, including the Middle East Respiratory Syndrome Coronavirus (MERS-CoV). Structural studies of the MERS-CoV proteins in the current literature are extremely limited. We present here detailed characterization of the struc-tural properties of MERS-CoV macro domain in aqueous solution at the atomic level with dynamics. For this study, we conducted extensive replica exchange molecular dynamics simulations linked to a generative neural networks and we use the resulting trajectories for structural analysis. We perform structural clustering based on the radius of gyration and end-to-end distance of MERS-CoV macro domain in aqueous solution with dynamics at the atomic level. We also report and analyze the residue-level intrinsic disorder features, flexibility and secondary structure. Furthermore, we study the pro-pensities of this macro domain for protein-protein interactions and for the RNA and DNA binding. Results are in agree-ment with available nuclear magnetic resonance spectroscopy findings and present more detailed insights into the struc-tural properties of MERS CoV macro domain. Overall, this work further shows that neural networks can be used as an exploratory tool for the studies of CoV family molecular conformational space at the nano level.

Structures of MERS-CoV Macro Domain in Aqueous Solution with Dynamics: Coupling Replica Exchange Molecular Dynamics and Deep Learning at the Nano Level

Ibrahim Yagiz Akbayrak[1], Burak Ulver[2], Havvanur Dervisoğlu[2], Mehmet Haklidir[2], Sule Irem Caglayan[3], Lukasz Kurgan[4*], Vladimir N. Uversky[5,6*], Orkun Hasekioglu[2*], Orkid Coskuner-Weber[3*]

[1]Materials Sciences and Technologies, College of Sciences, Turkish-German University, Sahinkaya Caddesi, No. 106, Beykoz, Istanbul, 34820 Turkey;[2]TUBITAK, Turkish Scientific and Technological Research Council, BİLGEM, Gebze, Istanbul 41470, Turkey;[3]Molecular Biotechnology, College of Sciences, Turkish-German University, Sahinkaya Caddesi, No. 106, Beykoz, Istanbul 34820 Turkey; [4]Department of Computer Science, Virginia Commonwealth University, Richmond, VA 23284, USA;[5]Department of Molecular Medicine, USF Health Byrd Alzheimer's Research Institute, Morsani College of Medicine, University of South Florida,

Tampa, FL 33612, USA; [6]Laboratory of New Methods in Biology, Institute for Biological Instrumentation of the Russian Academy of Sciences, Federal Research Center "Pushchino Scientific Center for Biological Research of the Russian Academy of Sciences", Pushchino 142290, Russia.

Corresponding Author Emails: vuversky@health.usf.edu, lkurgan@vcu.edu, orkun.hasekioglu@tubitak.gov.tr and weber@tau.edu.tr

Keywords: MERS-CoV, REMD simulations, deep learning, bioinformatics.

Running Title: Aqueous MERS-CoV Macro Domain Structures.

ABSTRACT: A novel virus, severe acute respiratory syndrome Coronavirus 2 (SARS-CoV-2), causing coronavirus disease 2019 (COVID-19) worldwide appeared in 2019. Currently, we do not have a medicament that treats the disease. One of the rea-sons for the absence of treatment is related to the scarcity of detailed scientific knowledge of the members of the Coronaviridae family, including the Middle East Respiratory Syndrome Coronavirus (MERS-CoV). Structural studies of the MERS-CoV proteins in the current literature are extremely limited. We present here detailed characterization of the structural properties of MERS-CoV macro domain in aqueous solution at the atomic level with dynamics. For this study, we conducted extensive replica exchange molecular dynamics simulations linked to a generative neural network and we use the resulting trajectories for structural analysis. We perform structural clustering based on the radius of gyration and end-to-end distance of MERS-CoV macro domain in aqueous solution with dynamics at the atomic level. We also report and analyze the residue-level intrinsic disorder features, flexibility and secondary structure. Furthermore, we study the propensities of this macro domain for protein-protein interactions and for the RNA and DNA binding. Results are in agreement with available nuclear magnetic resonance spectroscopy findings and present more detailed insights into the structural properties of MERS CoV macro domain. Overall, this work further shows that neural networks can be used as an exploratory tool for the studies of CoV family molecular conformational space at the nano level.

INTRODUCTION

Since the first outbreak of the severe acute respiratory syndrome (SARS) in 2003, a fatal viral disease-causing pneumonia and death was first reported in Saudi Arabia in 2012. This virus was named Middle East Respiratory Syndrome Coronavirus (MERS-CoV).[1] The current SARS-CoV-2 infection, coronaviruses and coronavirus-related infection aroused the attention of the entire world. A history of the SARS-CoV outbreak justifies these high-levels of attention. By the time the global SARS-CoV outbreak was contained, the virus spread to 26 countries, infected over 8,000 people worldwide and killed almost 800. Similarly, even though MERS-CoV appeared initially in Saudi Arabia, the virus - that was new to humans – spread to several other countries in or near the Arabian Peninsula, Asia, Europe, and the United States of America.[2] The mortality of MERS was reported to be 4-fold higher than SARS-CoV.[3] In fact, at the end of 2019, there were a total of 2,494 laboratory-confirmed cases of MERS-CoV world-wide and the MERS-CoV infection was characterized by the mortality rate of 34.4%. The current version of coronavirus, namely SARS-CoV-2, is infecting and killing more people per day than SARS and MERS combined during their existence.

Despite the history of posing threats to the human health, current knowledge of coronaviruses is rather limited. It is clear that gaining insights into the structural properties of various proteins from MERS-CoV, including the conserved macro domain within the non-structural protein 3 (NSP3), can help better understanding of the*Coronaviridae* family.[4] Since the structural properties of MERS-CoV macro domain in solution with dynamics are still poorly understood, a comparison to SARS-CoV-2 macro domain in solution with dynamics cannot be provided as well.

MERS-CoV belongs to the lineage C of β-coronaviruses (β-CoVs) that includes CoVs isolated from bats and hedgehogs. CoVs use the RNA genome to encode several structural proteins, including the spike glycoprotein (S), membrane protein (M) and nucleocapsid protein (N), and various non-structural proteins (NSPs) to facilitate its fast replication processes.[5] A single large replicase gene encodes the proteins that play a role in viral replication.[4] This gene contains two open reading frames; ORF1a and ORF1b encoding the polyproteins

pp1a and pp1b, with the production of pp1b requiring a -1 ribosome frame-shift at the 3' end of ORF1a.[6] ORF1a encodes viral proteases: main protease (M^{pro}) and papain like protease (PL^{pro}). These viral proteases play a central role in the cleavage of ORF1a and ORF1b gene products in order to produce functional NSPs.[7]

The largest NSP member of the MERS-CoV genome is the ORF1a-encoded, multifunctional and multidomain protein NSP3 that serves as a major evolutionary selection target in β-CoVs.[8] NSP3 includes N-terminal acidic domain, macro domain, SARS-unique domain, PL^{pro}, nucleic acid-binding domain, marker domain (G2M), transmembrane domain, and Y-domain. The macro domain received its name based on the non-histone motif of the histone variant macroH2A, which is a crucial protein module found in eukaryotes, bacteria, and archaea. The macro-domain containing proteins and enzymes play central roles in the regulation of various cellular processes. For instance, the SARS-CoV and MERS-CoV macro domains were shown to possess poly(AD)P-ribose binding affinity, which suggested that this domain regulates cellular proteins that are important for an apoptotic way via poly(ADP)-ribosylation to mediate the host response to infection.[4]

Even though X-ray structure is available for the MERS-CoV macro domain, such structure does not capture the impact of the bulk solvent environment on protein structure and dynamics and provides a rather limited view of the underlying structural and functional residue-level characteristics. A detailed understanding of the structural properties of MERS-CoV macro domain in solution at the atomic level with dynamics linked to deep learning together residue-level characterization will provide the lacking structural information on CoVs and may be used for comparison with SARS-CoV-2 macro domain. In the long run, the information gleaned from such structural studies could help to design more efficient treatments including vaccines and small molecule drugs. Therefore, we present the characterization of the structural properties of MERS-CoV macro domain in aqueous solution at body temperature with dynamics at the atomic level via linking replica exchange molecular dynamics simulations to deep learning (generative neural networks). We combine these results with several residue-level analyses that focus on the structural flexibility, presence of intrinsically disordered regions, and functional features related to the predisposition for protein-protein and protein-nucleic acid interactions.

Proteins and long peptides, such as the NSP3 and the macro domain of the NSP3, we are investigating in this study, represent very high dimensional data with a large number of degrees of freedom. One way to compress high dimensional data and obtain a lower dimensional representation is the use of generative neural network models and auto-encoders.[9] In particular, the generative neural networks (NN) and auto-encoders have been used to encode and model large polypeptides.[10, 11] Such generative NN models are also useful for in silico drug design and drug repurposing.[32]

MATERIALS and METHODS

Many molecular simulation scenarios require ergodic sampling of conformations. Their energy landscapes may feature many minima and barriers between minima that can be difficult to cross at ambient temperatures over reachable simulation time scales. This means that the corresponding findings are confounded by the choice of initial conditions because such conditions determine the space region that is explored by a simulation.[12] On the other hand, replica exchange simulations seek to enhance the conformational sampling by running numerous independent replicas in different conditions, and periodically exchanging the coordinates of different ensembles (replicas).[12] In this study, we conduct all-atom replica exchange molecular dynamics (REMD) simulations of MERS-CoV macro domain in water between the temperatures ranging from 280 K to 320 K using 32 replicas distributed exponentially between these temperatures. We use the CHARMM36 parameters for the MERS-CoV macro domain and the explicit TIP3P model for water.[13, 14] We apply a water layer of 10Å with 11156 water molecules to solvate the macro domain using a cubic box. We perform the simulations with the GROMACS 5.1.4 package.[15] We isolated the initial structure for the MERS-CoV macro domain from the publicly available crystal structure (PDB ID: 5zu7). After solvating the macro domain in water, we first conduct equilibration simulations for 20 ns (per replica) using the canonical ensemble and then for additional 20 ns (per replica) using the isothermal-isobaric ensemble. We run REMD simulations for a total simulation time of 6.4 μs. We perform exchanges between replicas every 5 ps with a time step of 2 fs. We save trajectories every 500 steps. Following our recent studies, We use the Langevin

dynamics to maintain the temperature of each replica with a collision frequency of 2 ps^{-1}.[16] Also, following our recent studies,[17, 18] we utilize the particle mesh Ewald (PME) method to accommodate for the long-range interactions. We apply the SHAKE algorithm to constrain the bonds to hydrogen atoms and we use counterions to neutralize the charges.

We calculate the structural properties of the MERS-CoV macro domain from the structures obtained after convergence from the replica closest to physiological temperature (310 K, see Supporting Information section). Also, we used the trajectories obtained from deep learning that was linked to our REMD simulations for structural analysis. We compute the content of the secondary structure components per residue for the aqueous MERS-CoV macro domain utilizing the DSSP program both for data obtained from REMD simulations and deep learning.[19]Additionally, we determine the end-to-end distances (R_{EE}) and radius of gyration (R_{g}) of the MERS-CoV macro domain in water using all converged trajectories from REMD simulations as well as from deep learning. Based on the relationship between theR_{g} and R_{EE} values, we apply the k -means clustering method to perform vector quantization and consequently to partition the structural observations into 5 clusters. We assign each observation to the cluster with the nearest cluster centroid that serves as a prototype of the cluster.[20] This way the structural data space is partitioned into Voronoi cells and the k -means clustering minimizes within cluster variances using squared Euclidean distances. Finally, we compute the root mean square fluctuations for each residue of the MERS-CoV macro domain in water. We compare these results to findings secured by using disorder predictors, which we describe next.

In addition, we perform residue-level analysis of the intrinsic disorder predisposition of the MERS-CoV macro domain and selected functional features related to its protein and nucleic acid binding potential. We evaluate the intrinsic disorder predisposition using a set of commonly utilized and publicly available computational tools, such as PONDR® VLXT,[21]PONDR® VSL2,[22]PONDR® FIT, [23] and IUPred capable of predicting long and short disordered regions.[24, 25, 26] Residue-level predisposition of this domain to interact with proteins was evaluated with the state-of-the SCRIBER (SeleCtive pRoteIn-Binding rEsidue pRedictor) method.[27] SCRIBER is currently the most accurate method that predicts protein-binding residues (PBRs), and the only tool that eliminates the recently described issue of the cross-prediction of residues that interact with nucleic acids (RNA and DNA) as PBRs.[28] This allows us to accurately predict PBRs and maintain high specificity of our analysis by limiting contamination of the results by the cross-predictions. We also evaluate the nucleic acid binding potential of the MERS-CoV macro domain with the DRNApred predictor.[29] DRNApred is currently the only method that provides accurate results and successfully eliminates the cross-predictions.[29, 30]

Generative networks and auto-encoders can conveniently generate new conformations, replacing computationally expensive molecular dynamics computations. We use conformations produced by REMD simulations to train a NN configured as an auto-encoder, as depicted in Scheme 1. When training such a NN model, for high fidelity realizations of new conformations, it is crucial that the conformation that will determine the weights of the neurons in the NN to be very close to a plausible typical conformation. In this respect, the results of the REMD simulations are very suitable to be used for this purpose.

Auto-encoders are trained to encode (compress) high dimensional conformational data to a lower dimensional space, that we refer to as the latent space, which is a two-dimensional space in this particular study. The two-dimensional data points in the latent space generated by the encoder NN are then decoded (decompressed) again to the high dimensional space to generate new conformations. The training procedure minimizes the loss function, comprised of the average spatial distance between the residue locations in the training conformation at the input of the encoder and the conformation obtained at the output of the decoder.

The conformations obtained from REMD simulations of the macro domain constitute our training set of data. The training and test data sets consists of converged conformations, 90% of the trajectories were used for training and the remaining 10% for testing. The encoder NN consisting of decreasing number of hidden layers is then trained to project the REMD generated conformations to the 2D latent space. The second NN, the decoder, consisting of hidden layers with increasing number of neurons decode (decompress) the points in the latent space back to the original conformational space of the N atoms with 3xN spatial coordinates.

During the training, the weights in both the encoding and decoding NN layers are optimized such that the loss function is minimized. On the other hand, the latent vectors, generated in the 2D latent space through the training of the encoder do not always necessarily correspond to feasible conformations. The conformations that are physically plausible need to be eliminated. For proper selection among the latent space vectors, following[10] and, we adopted a classifier based on the random forest method.[31] The random forest model is trained to classify two separate types of data: conformations extracted from the REMD data, conformations with added random fluctuations to the nuclear coordinates. We use the atomic coordinates of the residues for training purposes. Among the configurations generated by the decoder, feasible configurations are first selected through the random forest classifier and further the configurations above a 7 Å threshold with respect to the RMSF distance are eliminated.

RESULTS and DISCUSSION

Figure 1 represents a set of the selected structures of the MERS-CoV macro domain in aqueous media that we obtained from the all-atom REMD simulations at 310 K replica and from deep learning. While presence and proportions of specific secondary structures are similar across these conformations, our results reveal a remarkable structural pliability of this protein.

The grey line in Figure 2 presents the calculated RMSF values for each residue of the MERS-CoV macro domain in aqueous media at 310 K with dynamics at the atomic level. Based on these values, we notice more significant fluctuations (higher flexibility) in the C-terminal region of the domain even with such an extensive REMD simulation. The average RMSF value for the macro domain (all residues) is 1.09 ± 0.48 Å. However, the most flexible residue are characterized by the RMSF values of up to 3.62 Å.

Figure 2A shows that some of the structural dynamic features observed in our MD simulations are correlated with the residue-level intrinsic disorder predisposition of the MERS-CoV macro domain. This is reflected in the fact that several peaks in disorder profile serve as envelopes that enclose the local RMSF peaks. However, there also some regions (e.g. residues 18-38), which are predicted as ordered but which show noticeable structural fluctuations. This indicates that part of the structural fluctuations of the MERS-CoV macro domain in aqueous medium can be rooted in the intrinsic disorder predisposition of this domain, whereas other structural fluctuations are independent of the intrinsic disorder predisposition.

We also assess propensity of this protein to interact with other proteins and nucleic acids interactions. Like for the disorder analysis, we annotate these interactions at the level of individual amino acids. Figure 2B illustrates that the MERS-CoV macro domain is expected to have several protein binding regions, such as residues 1-12, 32, 43-47, 51, 86-88, 133-134, 137-144, 147, and 162-168. The predicted likelihood of the protein-protein interactions for these residues exceeds the 0.5 threshold. Some of these protein-binding residues are located within the disordered or flexible regions; i.e., regions characterized by the predicted disorder score exceeding 0.5 or ranging from 0.2 to 0.5, respectively. Curiously, although all highly flexible residues coincided or were located in the close proximity to the protein-binding regions/residues, not all regions with the highest protein binding potential were characterized by the highest RMSF values. Moreover, our residue-level analysis did not find any DNA- or RNA-binding regions in the MERS-CoV macro domain (see Figure 2B).

Dashed purple lines denote the RMSF values obtained through the generational autoencoder NN model. We note that they follow the REMD RMSF values quite closely on a residue basis. This indicates the viability of utilizing the NN model for the purposes of generating viable conformations.

Figure 3 presents the results that we generate with the k -means clustering of the structures of MERS-CoV macro domain in water with dynamics from REMD simulations and from deep learning. We base this calculation on the radius of gyration (R_{g}) and end-to-end distance (R_{EE}) values. Based on these results, the global compactness of the MERS-CoV macro domain structures (R_{g}) found in our experiments varies only by 1.0 Å, whereas the R_{EE} values fluctuate between 15 Å and 30 Å. The average R_{g} value of the MERS-CoV macro domain is 15.27 ± 0.08 Å from our REMD simulations and this value becomes 15.11 ± 0.45 Å using deep learning. The averageR_{EE} values is 22.35 ± 1.72 Å in water from our REMD simulations. This value is 21.89 ± 2.29 Å using the trajectories from deep learning. Experimental structural studies on MERS CoV

macro domain in solution are extremely limited and therefore we could not compare these results to data generated by the experiments. However, we use a set of independently computed residue-level predictions and the secondary structure analysis based on an NMR structure to contextualize and compare with our all-atom results.

Figures 4 and 5 display the residue-level secondary structure probabilities that we predicted from the MERS-CoV macro domain in water at 310 K using the trajectories from REMD simulations and those from deep learning. Based on these calculations, we found six α-helices (Figure 4A and Figure 5A) in the MERS-CoV macro domain structures. They are located at residues Ala25-Cys31, Gly50-Ser59, Ala62-Lys74, Ser109-Met118, Pro138-Glu148, and Gln160-Leu166. Some of these helices, especially those located within the C-terminal half of the domain, were predicted to include the protein binding residues. While the six α-helices were also observed in the NMR measurements, they also annotate adjacent residues as helical, but this might be related to the buffer used in these experiments.[32] The abundance of the helical structures formed in our REMD simulations and NN in water are, in general, higher than those reported in the NMR measurements.[32]

The location of the 3_{10}-helices – displayed in Figures 4B and 5B – are formed at Ala102-Ala104 and Val108-Ser112 regions of the MERS-CoV macro domain in water at 310 K. The NMR experiments did not measure the 3_{10}-helix content for this protein. Correspondingly, we also note that only a narrow region of the sequence of this domain adopts the 3_{10}-helical structure in water. One of these 3_{10}-helices (residues Val108-Ser112), overlaps with one of the α-helices (residues Ser109-Met118).

Figures 4C and 5C summarize the calculated β-sheet content for the macro domain in water. We note prominent β-strand formation in seven regions. Namely, these regions are His11-Thr15, Val18-Leu22, Val36-Ala41, Asp81-Leu85, Asn93-Val98, Leu123-Thr126, and Arg152-Val157. The NMR measurements similarly assigned seven similar regions for β-strand formation. However, like for the helices, residues adjacent to these regions were also found to adopt β-strand structure to some extent based on these measurements. The TALOS index measurements support our findings.[32]

Finally, we calculated the turn structure content per residue (Figures 4D and 5D). This structural element was not analyzed in the NMR experiments. As per our analysis, there are ten regions that adopt turn structure. Specifically, turn structure was detected at Met4-Phe9, Glu16-Cys17, Tyr32-Ser35, Asn42-Leu45, Lys60-Gly61, Gln78-Gly80, Gly87-Lys92, Asp101-Lys105, Asn119-Pro122, and Leu128 -Gly135 of the MERS-CoV macro domain in water. Many of the protein binding residues/regions (e.g., residues 1-12, 32, 43-47, 86-88, and 133-134) overlap with the regions with turn structure.

The estimated secondary structure propensities per residue using the trajectories obtained from deep learning (Figure 5) are in excellent accord with our REMD results (Figure 4, see above).

CONCLUSION

We conduct REMD simulations linked to deep learning at the nano level for the MERS CoV macro domain in water and present here the results for the 310 K replica. We cover several structural properties including RMSF values with dynamics, secondary structure, and the k -means clustering based on radius of gyration (R_g) and end-to-end distance (R_{EE}) of the structures of MERS CoV macro domain in water with dynamics. Our findings, which rely on the RMSF values, R_{EE} values and deviations, show that some of the residues are flexible. Furthermore, the global structure is compact, not very flexible, and varies only by 1.0 Å in water (in terms of the scale of R_g fluctuations). We detected six α-helical regions and seven β-strand regions, which are in good agreement with the available NMR measurements. In addition, we notice ten regions with turn structure in the computed here conformations of MERS CoV macro domain in water with dynamics.

Based on the results of the comparison of the independently generated intrinsic disorder analysis of the MERS-CoV macro domain with the REMD and deep learning analyses, we also show that only part of the structural fluctuations of this protein in aqueous medium can be attributed the local intrinsic disorder predisposition. The other structural fluctuations are independent of the local propensity of the MERS-CoV

macro domain to the intrinsic disorder.

Our residue-level analysis provides some functional clues. Based on putative propensities for protein and nucleic acids interactions, we suggest that while the MERS-CoV macro domain appears not to show DNA- or RNA-binding potential, it contains several protein binding regions. Many of the corresponding PBRs are located within the disordered or flexible regions. Also, some PBRs overlap with the regions with the turn structure. Furthermore, some of the α-helices found in the MERS-CoV macro domain, especially located within the C-terminal half of the protein, were predicted to contain PBRs.

We studied the structures of MERS-CoV in water using generative networks and auto-encoders linked to REMD simulations. The trajectories obtained from generative networks and auto-encoders yield structural results for MERS-CoV in full agreement with our extensive special sampling simulations.

The reported here results can be used to support activities associated with the development of new MERS-CoV treatments including vaccines and drug molecules. Currently, we are studying the structural dynamics of various regions of different proteins from the CoV family ranging from SARS-CoV to SARS-CoV-2 with MERS-CoV in between.

SUPPORTING MATERIALS SECTION AVAILABLE: Time dependent RMSD values for MERS CoV macro domain in water from our REMD simulations at 310 K.

Acknowledgements: The authors acknowledge TRUBA resources because the numerical calculations reported in this study were performed at TUBITAK ULAKBIM, High Performance and Grid Computing Center (TRUBA resources).

REFERENCES

(1) Schwarz, K.; Groneberg, D. A. [Current overview on MERS-CoV].*Zentralblatt Arbeitsmedizin Arbeitsschutz Ergon.* 2015, *65*(6), 353–354. https://doi.org/10.1007/s40664-015-0062-8.

(2) Subbaram, K.; Kannan, H.; Khalil Gatasheh, M. Emerging Developments on Pathogenicity, Molecular Virulence, Epidemiology and Clinical Symptoms of Current Middle East Respiratory Syndrome Coronavirus (MERS-CoV). *Hayati J. Biosci.* 2017, *24* (2), 53–56. https://doi.org/10.1016/j.hjb.2017.08.001.

(3) Ford, N.; Vitoria, M.; Rangaraj, A.; Norris, S. L.; Calmy, A.; Doherty, M. Systematic Review of the Efficacy and Safety of Antiretroviral Drugs against SARS, MERS or COVID-19: Initial Assessment.*J. Int. AIDS Soc.* 2020, *23* (4). https://doi.org/10.1002/jia2.25489.

(4) Cho, C.-C.; Lin, M.-H.; Chuang, C.-Y.; Hsu, C.-H. Macro Domain from Middle East Respiratory Syndrome Coronavirus (MERS-CoV) Is an Efficient ADP-Ribose Binding Module: CRYSTAL STRUCTURE AND BIOCHEMICAL STUDIES.*J. Biol. Chem.* 2016, *291* (10), 4894–4902. https://doi.org/10.1074/jbc.M115.700542.

(5) Aasiyah Chafekar; Burtram Fielding. MERS-CoV: Understanding the Latest Human Coronavirus Threat. *Viruses* 2018, *10* (2), 93. https://doi.org/10.3390/v10020093.

(6) Snijder, E. J.; Bredenbeek, P. J.; Dobbe, J. C.; Thiel, V.; Ziebuhr, J.; Poon, L. L. M.; Guan, Y.; Rozanov, M.; Spaan, W. J. M.; Gorbalenya, A. E. Unique and Conserved Features of Genome and Proteome of SARS-Coronavirus, an Early Split-off from the Coronavirus Group 2 Lineage. *J. Mol. Biol.* 2003, *331* (5), 991–1004. https://doi.org/10.1016/s0022-2836(03)00865-9.

(7) Neuman, B. W.; Buchmeier, M. J. Supramolecular Architecture of the Coronavirus Particle. In *Advances in Virus Research* ; Elsevier, 2016; Vol. 96, pp 1–27. https://doi.org/10.1016/bs.aivir.2016.08.005.

(8) Forni, D.; Cagliani, R.; Mozzi, A.; Pozzoli, U.; Al-Daghri, N.; Clerici, M.; Sironi, M. Extensive Positive Selection Drives the Evolution of Nonstructural Proteins in Lineage C Betacoronaviruses.*J. Virol.* 2016, *90* (7), 3627–3639. https://doi.org/10.1128/JVI.02988-15.

(9) Hinton, G. E. Reducing the Dimensionality of Data with Neural Networks. *Science* 2006, *313* (5786), 504–507. https://doi.org/10.1126/science.1127647.

(10) Degiacomi, M. T. Coupling Molecular Dynamics and Deep Learning to Mine Protein Conformational Space. *Structure* 2019, *27* (6), 1034-1040.e3. https://doi.org/10.1016/j.str.2019.03.018.

(11) Ma, H.; Bhowmik, D.; Lee, H.; Turilli, M.; Young, M. T.; Jha, S.; Ramanathan, A. Deep Generative Model Driven Protein Folding Simulation.*ArXiv190800496 Q-Bio* 2019.

(12) Geng, H.; Chen, F.; Ye, J.; Jiang, F. Applications of Molecular Dynamics Simulation in Structure Prediction of Peptides and Proteins.*Comput. Struct. Biotechnol. J.* 2019, *17* , 1162–1170. https://doi.org/10.1016/j.csbj.2019.07.010.

(13) Huang, J.; MacKerell, A. D. CHARMM36 All-Atom Additive Protein Force Field: Validation Based on Comparison to NMR Data. *J. Comput. Chem.* 2013, *34* (25), 2135–2145. https://doi.org/10.1002/jcc.23354.

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

(15) Berendsen, H. J. C.; van der Spoel, D.; van Drunen, R. GROMACS: A Message-Passing Parallel Molecular Dynamics Implementation.*Comput. Phys. Commun.* 1995, *91* (1–3), 43–56. https://doi.org/10.1016/0010-4655(95)00042-E.

(16) Wise-Scira, O.; Aloglu, A. K.; Dunn, A.; Sakallioglu, I. T.; Coskuner, O. Structures and Free Energy Landscapes of the Wild-Type and A30P Mutant-Type α-Synuclein Proteins with Dynamics. *ACS Chem. Neurosci.* 2013, *4* (3), 486–497. https://doi.org/10.1021/cn300198q.

(17) Coskuner, O.; Wise-Scira, O. Arginine and Disordered Amyloid-β Peptide Structures: Molecular Level Insights into the Toxicity in Alzheimer's Disease. *ACS Chem. Neurosci.* 2013, *4* (12), 1549–1558. https://doi.org/10.1021/cn4001389.

(18) Coskuner, O.; Uversky, V. N. Tyrosine Regulates β-Sheet Structure Formation in Amyloid-β $_{42}$: A New Clustering Algorithm for Disordered Proteins. *J. Chem. Inf. Model.* 2017, *57*(6), 1342–1358. https://doi.org/10.1021/acs.jcim.6b00761.

(19) Kabsch, W.; Sander, C. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features.*Biopolymers* 1983, *22* (12), 2577–2637. https://doi.org/10.1002/bip.360221211.

(20) Likas, A.; Vlassis, N.; J. Verbeek, J. The Global K-Means Clustering Algorithm. *Pattern Recognit.* 2003, *36* (2), 451–461. https://doi.org/10.1016/S0031-3203(02)00060-2.

(21) Romero, P.; Obradovic, Z.; Li, X.; Garner, E. C.; Brown, C. J.; Dunker, A. K. Sequence Complexity of Disordered Protein. *Proteins*2001, *42* (1), 38–48. https://doi.org/10.1002/1097-0134(20010101)42:1<38::aid-prot50>3.0.co;2-3.

(22) Peng, K.; Radivojac, P.; Vucetic, S.; Dunker, A. K.; Obradovic, Z. Length-Dependent Prediction of Protein Intrinsic Disorder. *BMC Bioinformatics* 2006, *7* , 208. https://doi.org/10.1186/1471-2105-7-208.

(23) Xue, B.; Dunbrack, R. L.; Williams, R. W.; Dunker, A. K.; Uversky, V. N. PONDR-FIT: A Meta-Predictor of Intrinsically Disordered Amino Acids. *Biochim. Biophys. Acta* 2010, *1804* (4), 996–1010. https://doi.org/10.1016/j.bbapap.2010.01.011.

(24) Dosztányi, Z.; Csizmok, V.; Tompa, P.; Simon, I. IUPred: Web Server for the Prediction of Intrinsically Unstructured Regions of Proteins Based on Estimated Energy Content. *Bioinforma. Oxf. Engl.* 2005,*21* (16), 3433–3434. https://doi.org/10.1093/bioinformatics/bti541.

(25) Dosztányi, Z.; Csizmók, V.; Tompa, P.; Simon, I. The Pairwise Energy Content Estimated from Amino Acid Composition Discriminates between Folded and Intrinsically Unstructured Proteins. *J. Mol. Biol.* 2005, *347* (4), 827–839. https://doi.org/10.1016/j.jmb.2005.01.071.

(26) Mészáros, B.; Erdos, G.; Dosztányi, Z. IUPred2A: Context-Dependent Prediction of Protein Disorder as a Function of Redox State and Protein Binding. *Nucleic Acids Res.* 2018, *46* (W1), W329–W337. https://doi.org/10.1093/nar/gky384.

(27) Zhang, J.; Kurgan, L. SCRIBER: Accurate and Partner Type-Specific Prediction of Protein-Binding Residues from Proteins Sequences.*Bioinforma. Oxf. Engl.* 2019, *35* (14), i343–i353. https://doi.org/10.1093/bioinformatics/btz324.

(28) Zhang, J.; Kurgan, L. Review and Comparative Assessment of Sequence-Based Predictors of Protein-Binding Residues. *Brief. Bioinform.* 2018, *19* (5), 821–837. https://doi.org/10.1093/bib/bbx022.

(29) Yan, J.; Kurgan, L. DRNApred, Fast Sequence-Based Method That Accurately Predicts and Discriminates DNA- and RNA-Binding Residues.*Nucleic Acids Res.* 2017, *45* (10), e84. https://doi.org/10.1093/nar/gkx059.

(30) Su, H.; Liu, M.; Sun, S.; Peng, Z.; Yang, J. Improving the Prediction of Protein-Nucleic Acids Binding Residues via Multiple Sequence Profiles and the Consensus of Complementary Methods.*Bioinforma. Oxf. Engl.* 2019, *35* (6), 930–936. https://doi.org/10.1093/bioinformatics/bty756.

(31) Zhavoronkov, A.; Aladinskiy, V. A.; Zhebrak, A.; Zagribelnyy, B. A.; Terentiev, V. A.; Bezrukov, D.; Polykovskiy, D.; Shayakhmetov, R.; Filimonov, A.; Orekhov, P.; Yilin Yan; Popova, O.; Vanhaelen, Q.; Aliper, A.; Ivanenkov, Y. A. Potential 2019-NCoV 3C-like Protease Inhibitors Designed Using Generative Deep Learning Approaches. 2020. https://doi.org/10.13140/RG.2.2.29899.54569.

(32) Huang, Y.-P.; Cho, C.-C.; Chang, C.-F.; Hsu, C.-H. NMR Assignments of the Macro Domain from Middle East Respiratory Syndrome Coronavirus (MERS-CoV). *Biomol. NMR Assign.* 2016, *10* (2), 245–248. https://doi.org/10.1007/s12104-016-9676-9.

Figure Legends

Scheme 1. Generative auto-encoder Neural Network model. The encoder input layer consists of 3N neurons corresponding to the number of coordinates in the training data. The two hidden layers contain 300 and 50 neurons. The output layer consisting of two neurons finally encode and compress the original conformation to two real numbers in the two-dimensional latent space. The decoder in reverse order maps the points in the latent space back to the conformation space.

Figure 1. Selected structures from our simulations and deep learning representing conformations of the MERS-CoV macro domain in aqueous media.

Figure 2. Comparison of the structural flexibility of MERS-CoV macro domain in aqueous media with its intrinsic disorder predisposition (A) and propensity for protein and nucleic acid binding (B). Structural flexibility in the aqueous media is reflected in root mean square fluctuations (RMSF) of the protein backbone as a function of the MERS-CoV macro domain residue number. Intrinsic disorder predisposition was evaluated using PONDR® VLXT, PONDR® VSL2, PONDR® FIT, IUPred_short, and IUPres_long (A). Predisposition of this domain to interact with proteins and nucleic acids was evaluated by SCRIBER and DRNApred, respectively (B).

Figure 3. A) R_{g} *vs* .R_{ee} values of the MERS-CoV macro domain in solution from REMD simulations that we processed with the k means clustering. 5 k values were used and centroids are located atR_{g} = 15.24 Å, R_{ee} = 17.69 Å (Centroid1), R_{g} = 15.26 Å,R_{ee} = 24.10 Å (Centroid 2),R_{g} = 15.26 Å, R_{ee} = 21.21 Å (Centroid 3), R_{g} = 15.26 Å,R_{ee} = 22.17 Å (Centroid 4), andR_{g} = 15.26 Å, R_{ee} = 22.95 Å (Centroid 5). B) Rg vs. Ree values of the MERS-CoV macro domain in solution from deep learning that we processed with the k means clustering. 5 k values were used and centroids are located at Rg = 15.12 Å, Ree = 23.36

Å (Centroid1), Rg = 15.19 Å, Ree = 21.87 Å (Centroid 2), Rg = 14.92 Å, Ree = 19.24 Å (Centroid 3), Rg = 15.18 Å, Ree = 24.98 Å (Centroid 4), and Rg = 15.10 Å, Ree = 17.80 Å (Centroid 5).

Figure 4. Secondary structure elements and their residue-level probabilities recovered from the MERS-CoV macro domain structures calculated with dynamics at the atomic level in aqueous media from REMD simulations.

Figure 5. Secondary structure elements and their residue-level probabilities recovered from the MERS-CoV macro domain structures calculated with dynamics at the atomic level in aqueous media from deep learning.

FIGURES

Scheme 1.

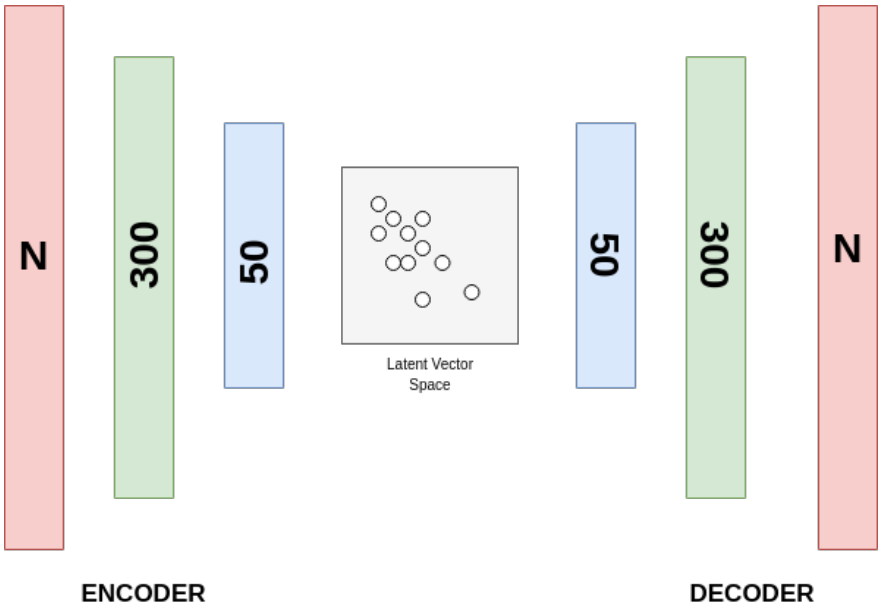


Figure 1.

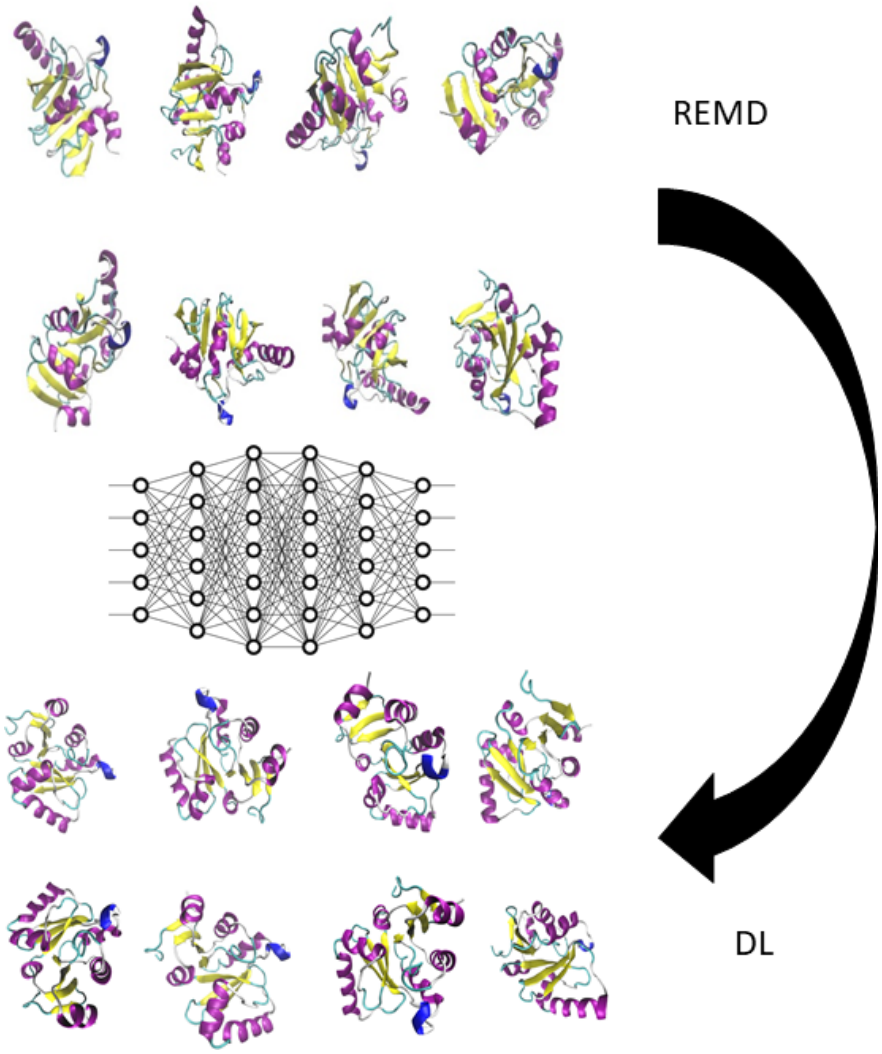


Figure 2.

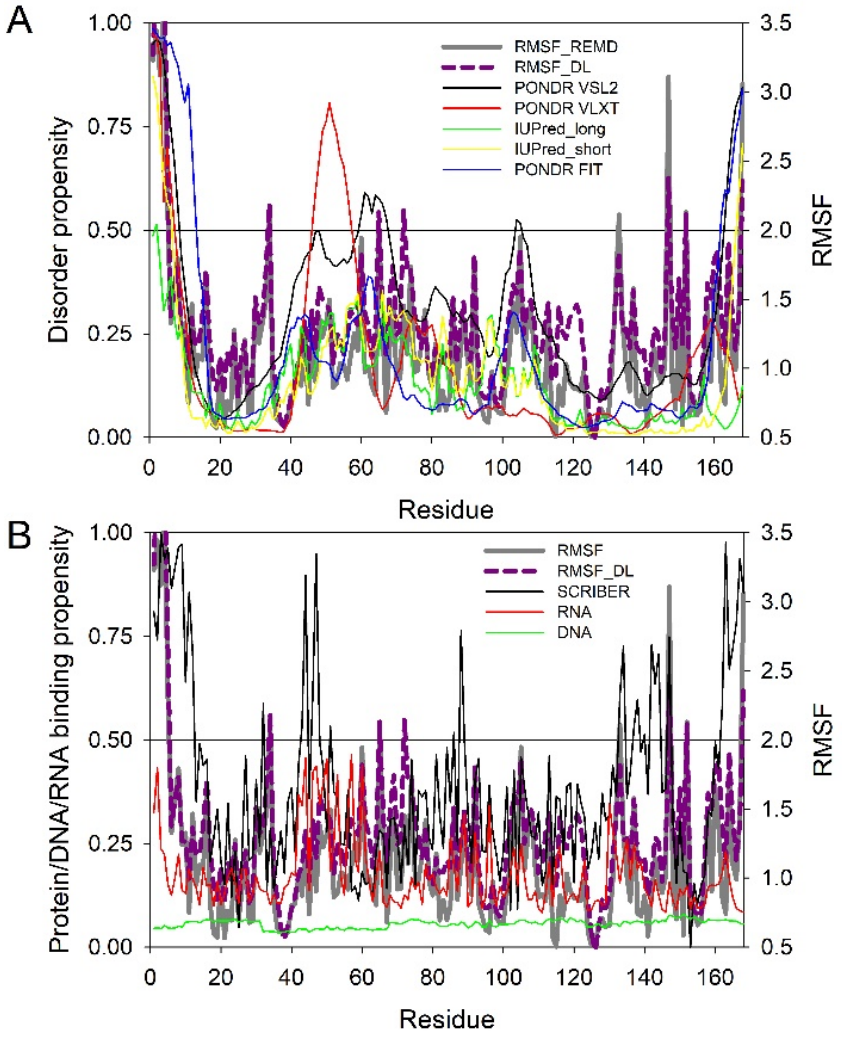


Figure 3.

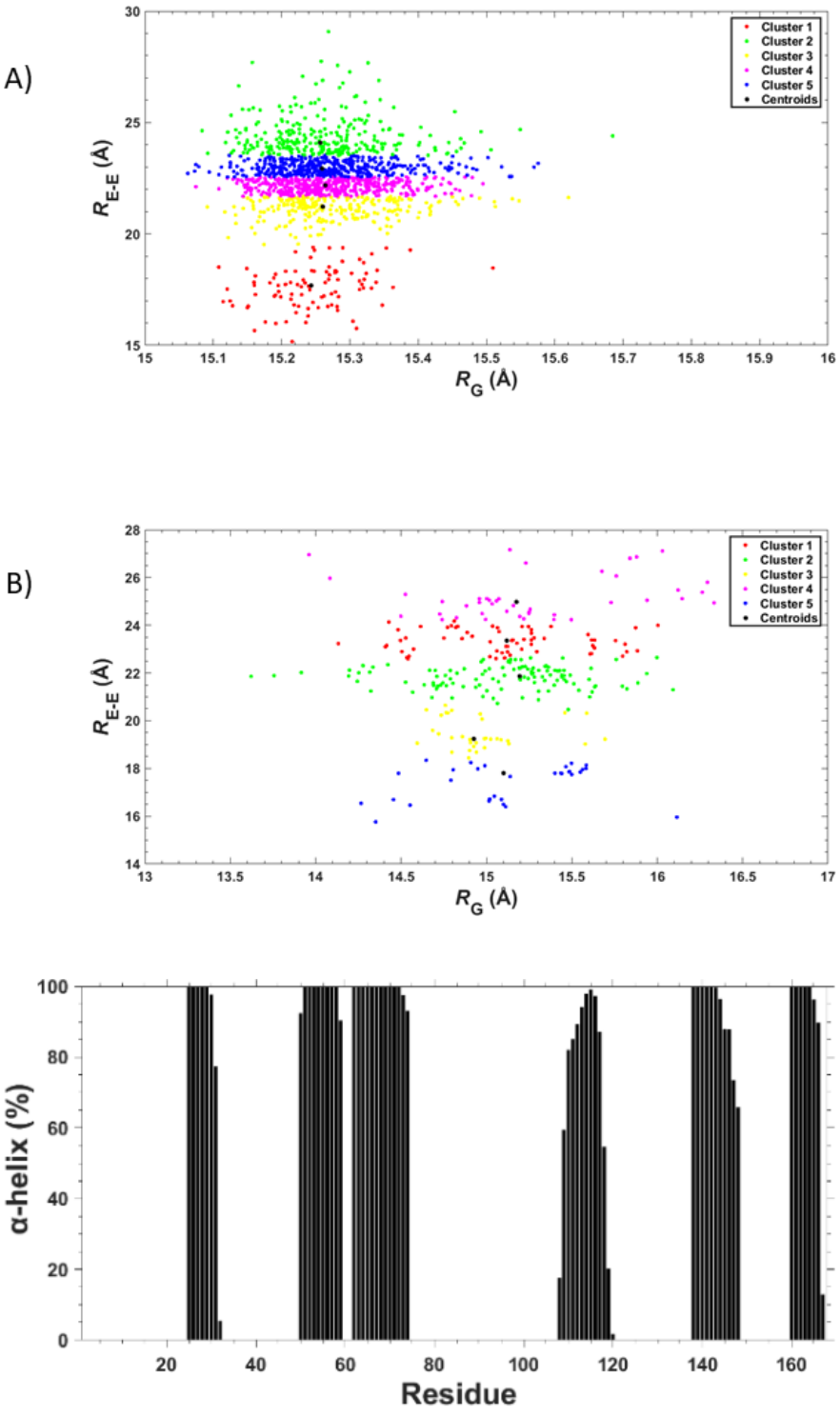
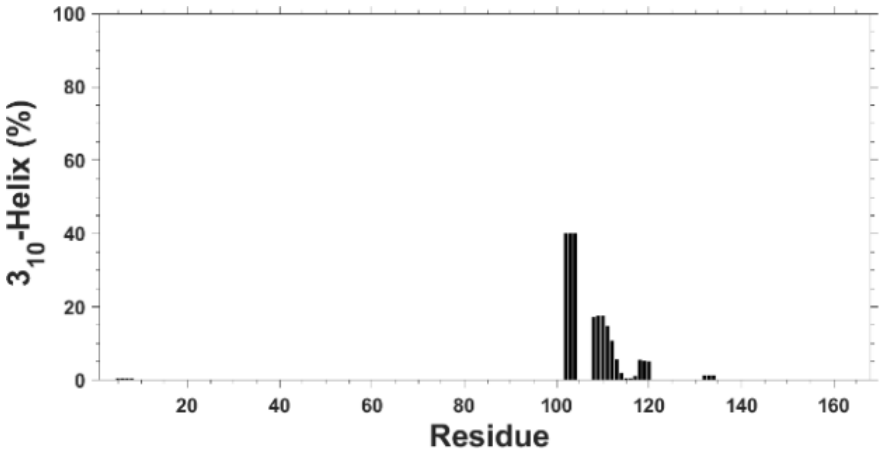
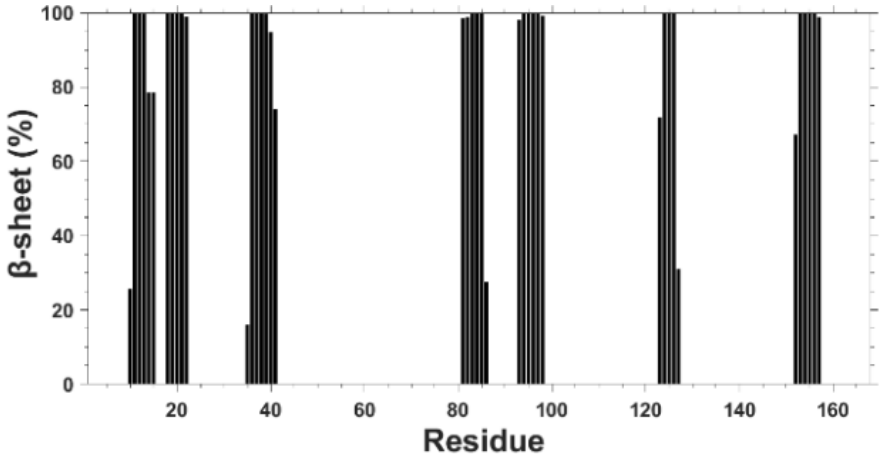


Figure 4.

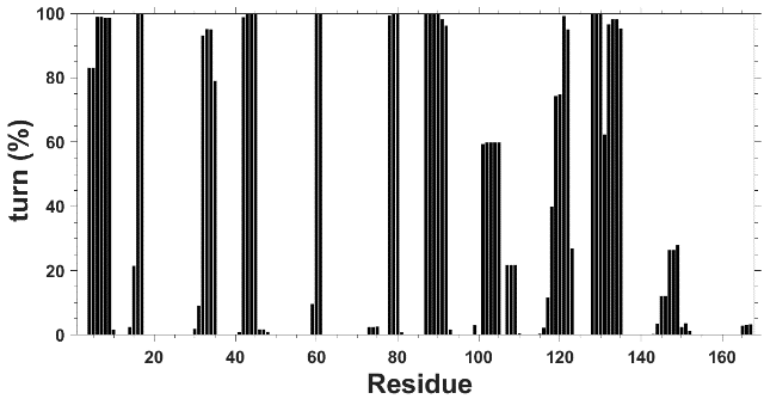
A



B

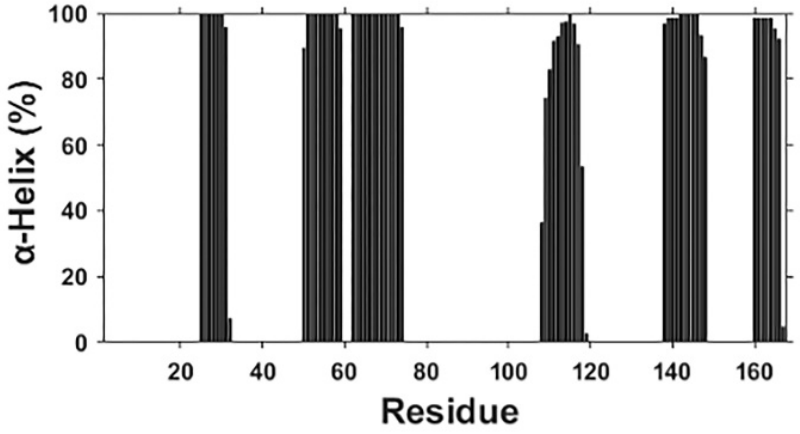


C

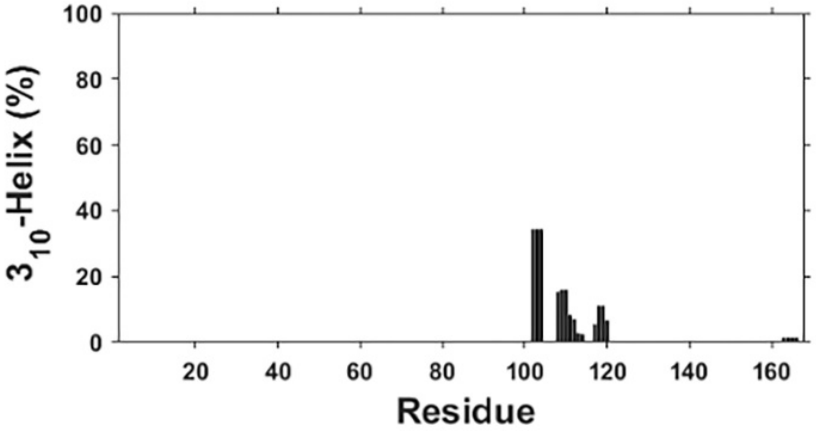


D

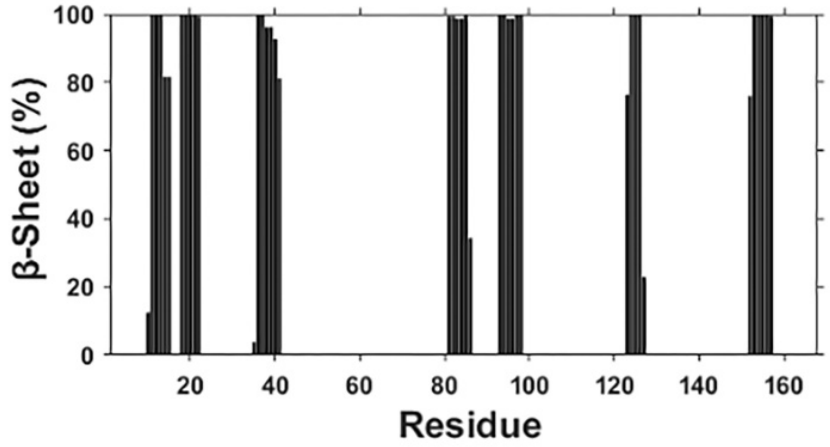
Figure 5.



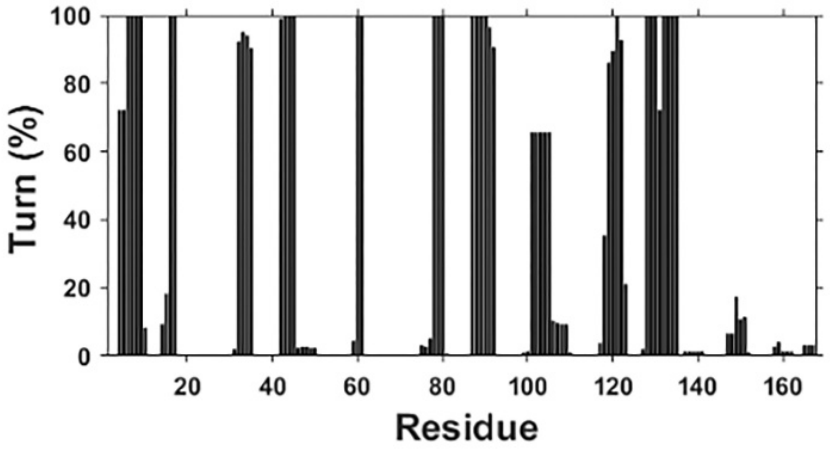
A



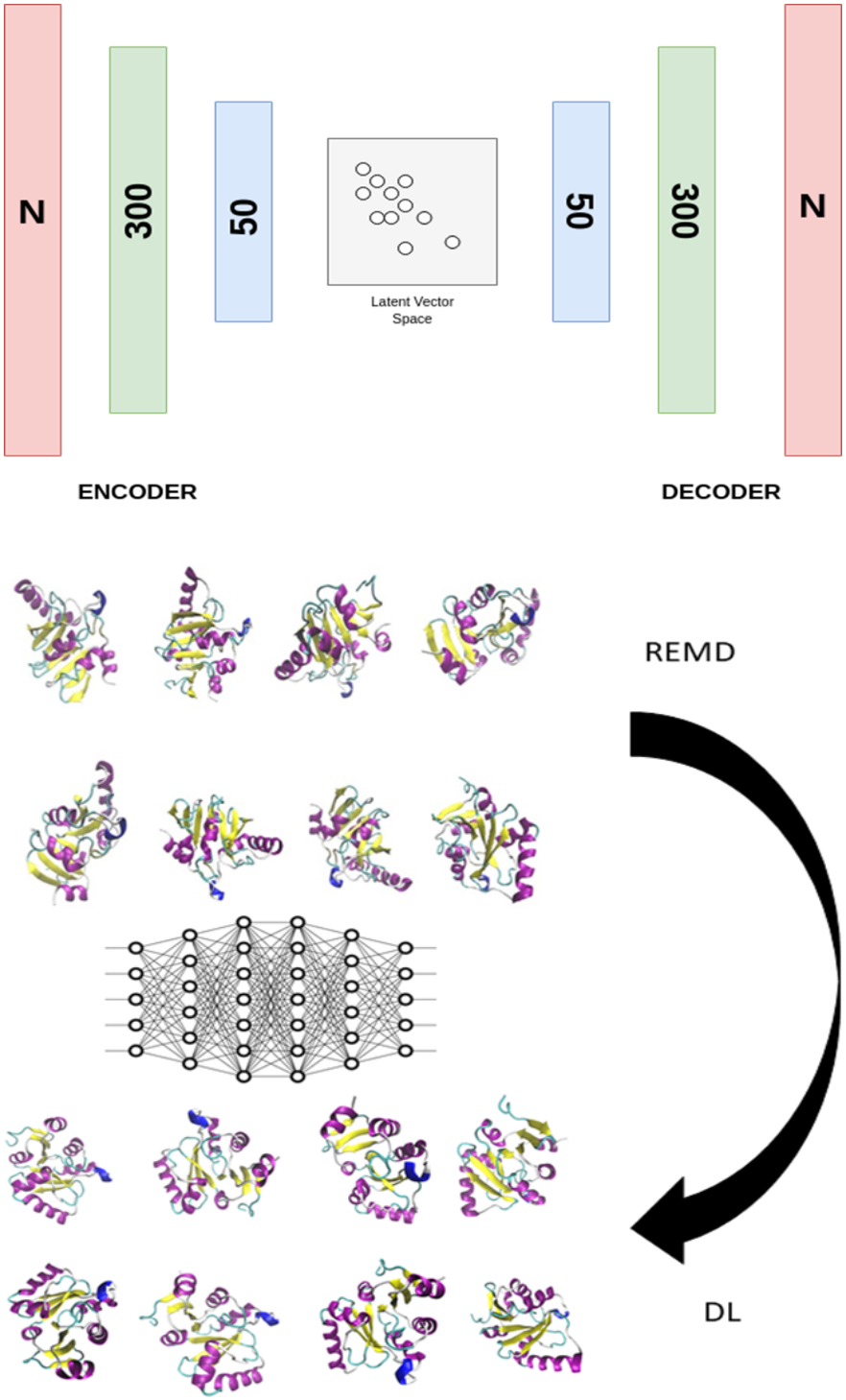
B

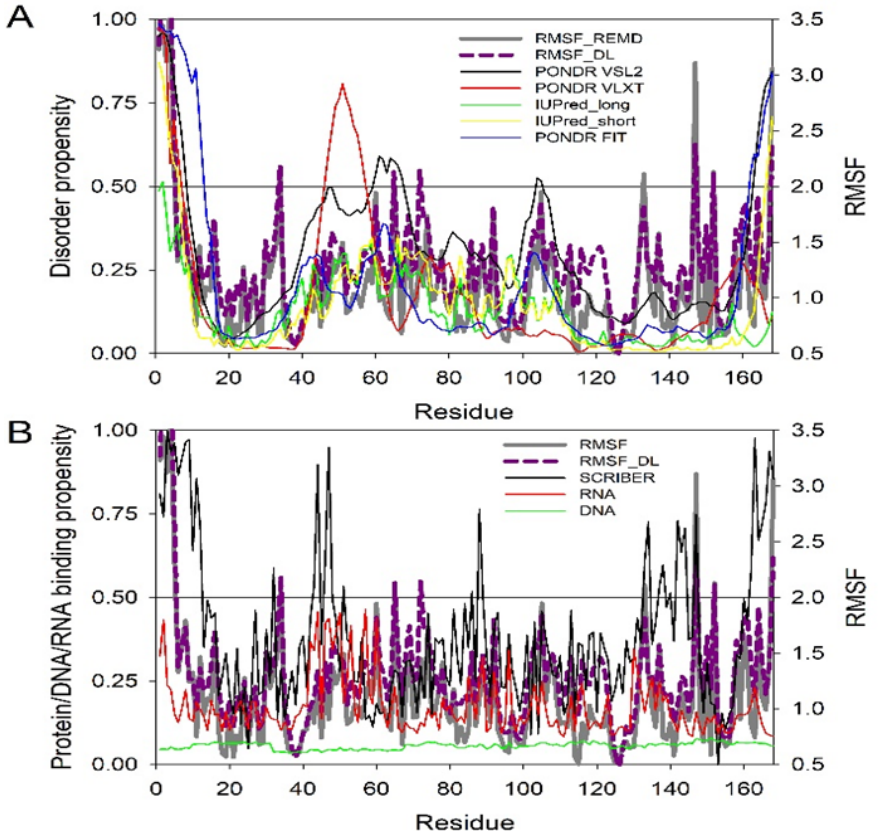


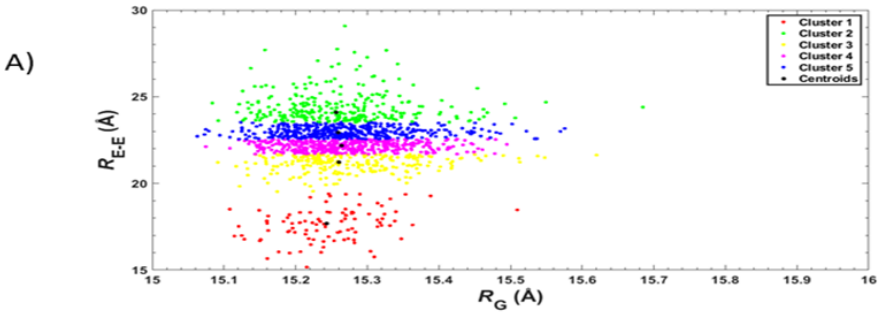
C

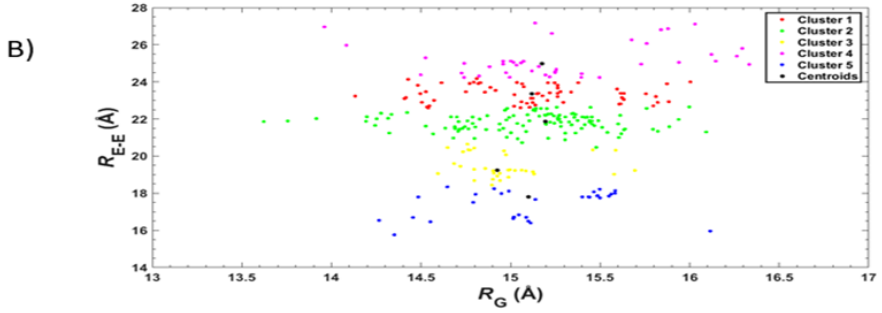


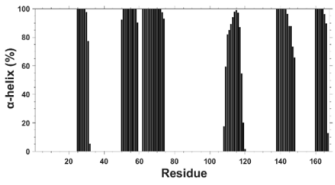
D



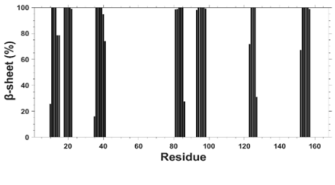




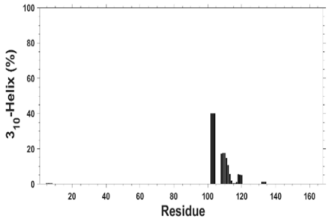




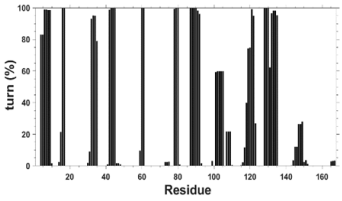
A



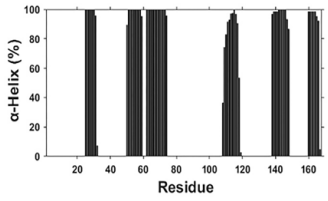
C



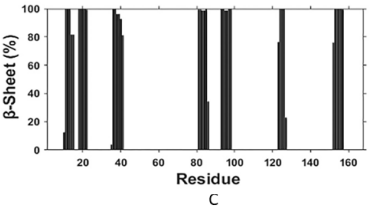
B



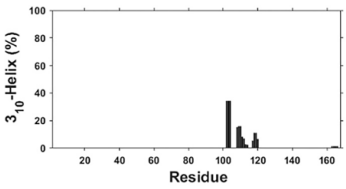
D



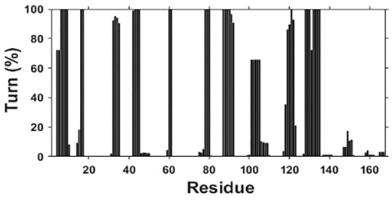
A



C



B



D