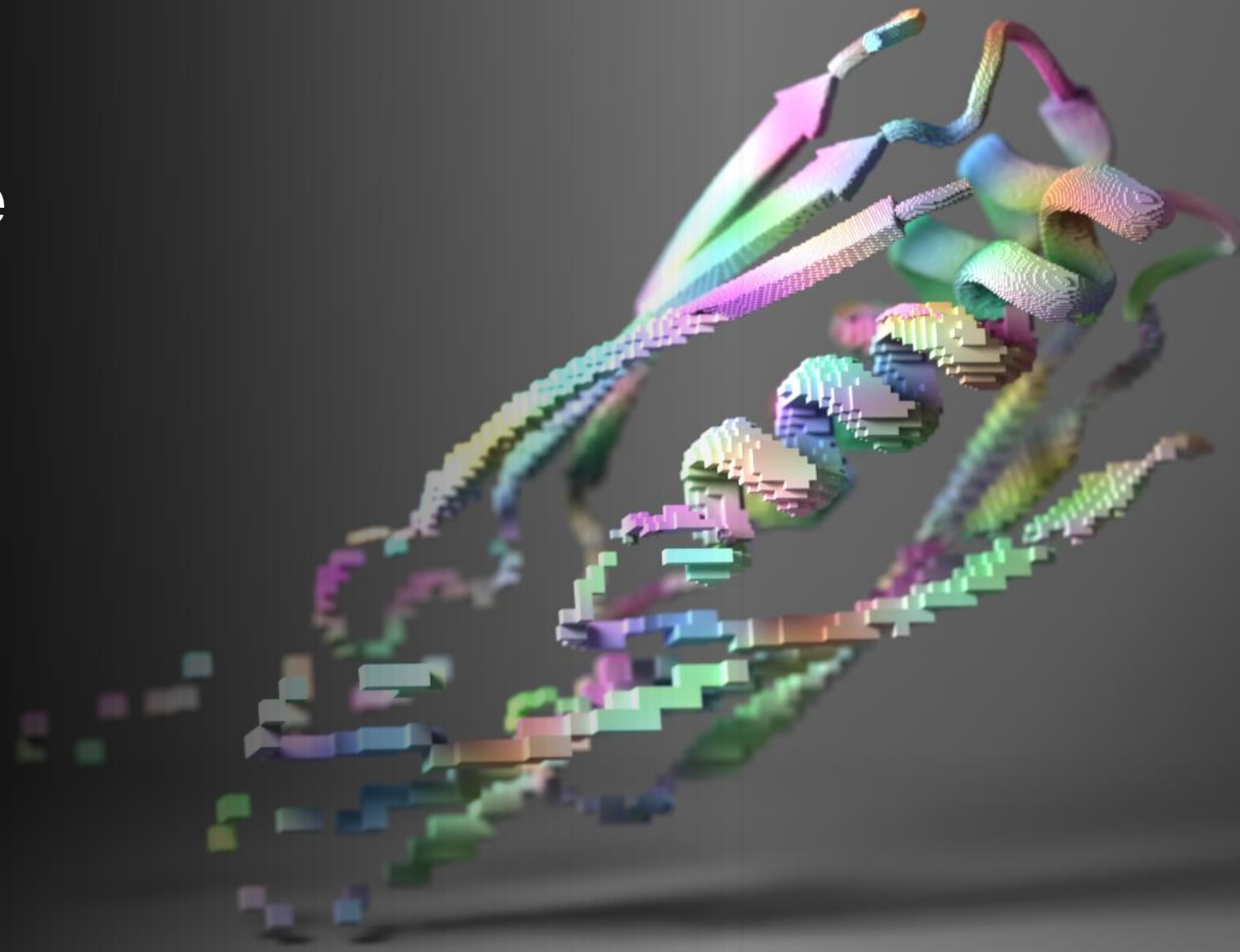
Generating Protein Conformation Ensemble Through A Transformer-based Diffusion Model

YI YANG

November 13, 2024



SMU

OUTLINE

Part I

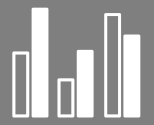
Protein structure and function, protein dynamics

Part II

How to build a transformer-based diffusion model

Part III

Computational result for applying our model to vivid protein system



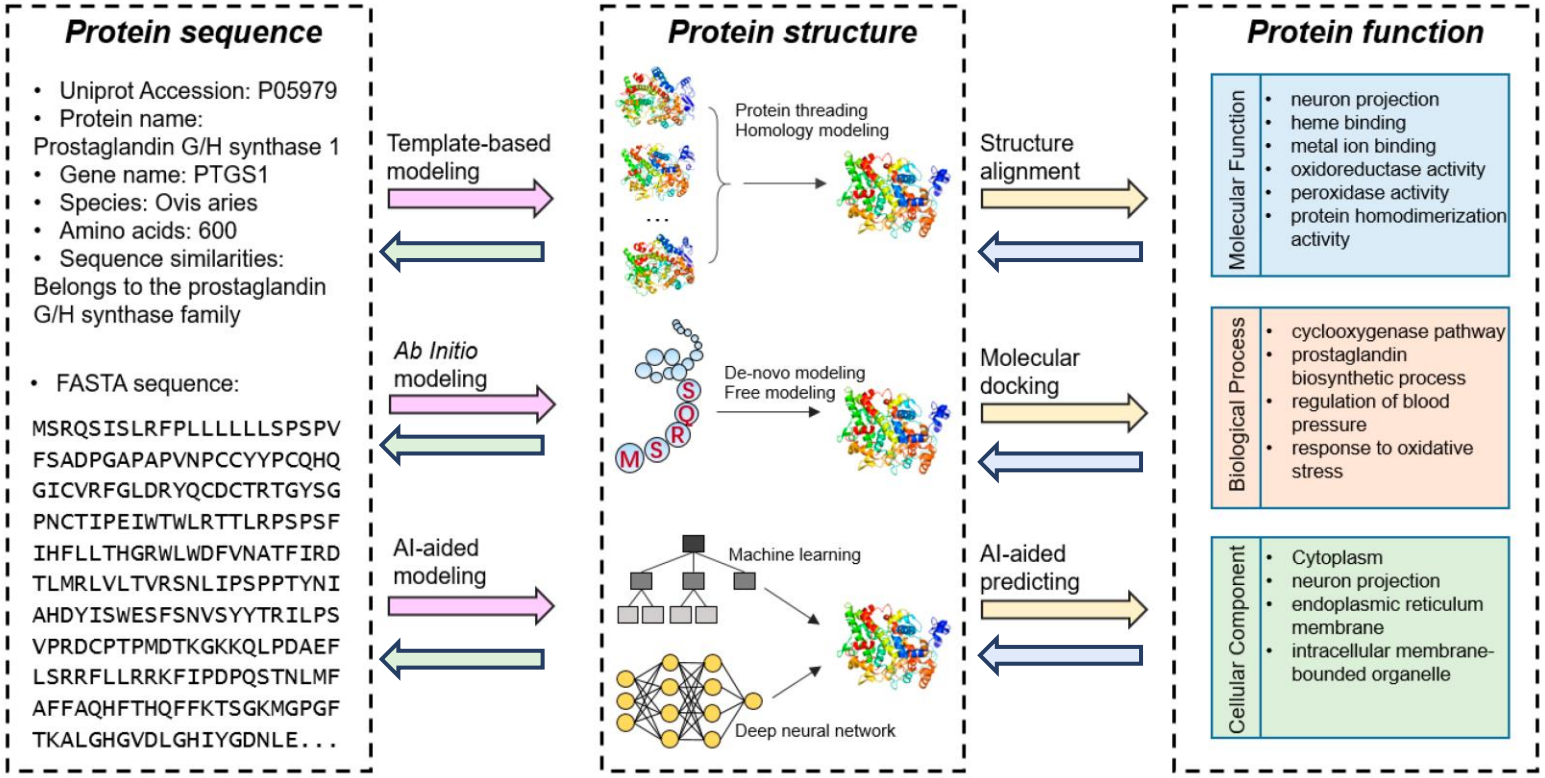
Part IV

Conclusion

Part I

The relationship between Protein structure and function

Generally, protein sequence determines structure and structure determines function.



Protein in Dynamics

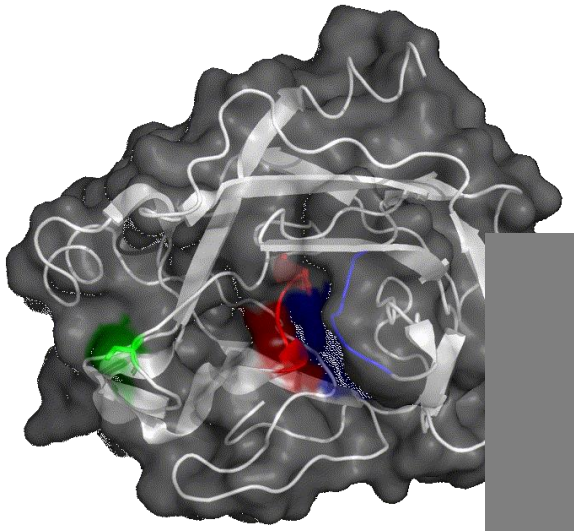
Protein Function

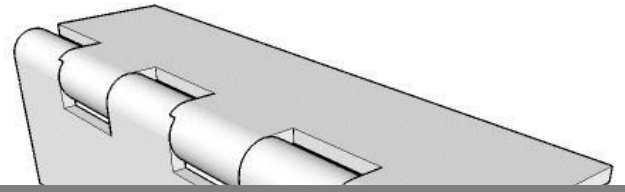
Zhang, J.; Qian, J. Advances in Computational Intelligence-Based Methods of Structure and Function Prediction of Proteins. Biomolecules 2024, 14, 1083. https://doi.org/10.3390/biom14091083

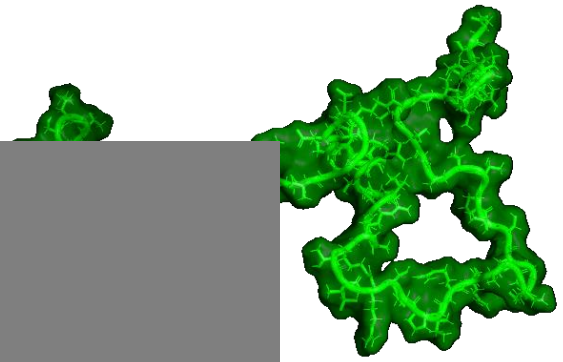
Part I

The relationship between Protein structure and function

But some proteins exhibit functional flexibility through dynamic structural changes







Obtaining protein dynamic ensemble is important !!!

- **Hinge motion** in disordered activation domain in Trypsinogen (PDB ID: 2PTN)
- Biological process: digestion

- An intrinsically disordered protein (IDP)
- A neuron-specific microtubule-associated protein that plays an important role in maintaining neuronal structure and function(MAP2C)

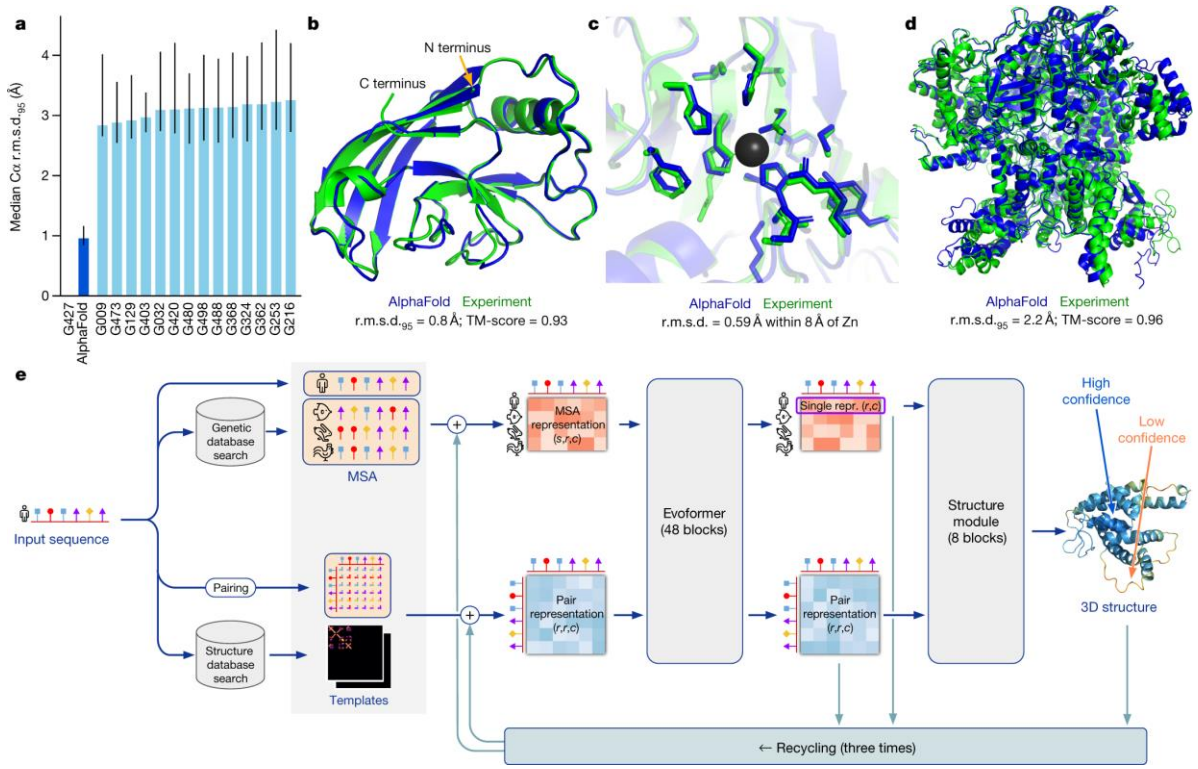
Melková K, Zapletal V, Jansen S, Nomilner E, Zachrdla M, Hritz J, Nováček J, Zweckstetter M, Jensen MR, Blackledge M, Žídek L. Functionally specific binding regions of microtubule-associated protein 2c exhibit distinct conformations and dynamics. J Biol Chem. 2018 Aug 24;293(34):13297-13309. doi: 10.1074/jbc.RA118.001769. Epub 2018 Jun 20. PMID: 29925592; PMCID: PMC6109934.

Hayward, Steven. "Structural principles governing domain motions in proteins." Proteins: Structure, Function, and Bioinformatics 36.4 (1999): 425-435.

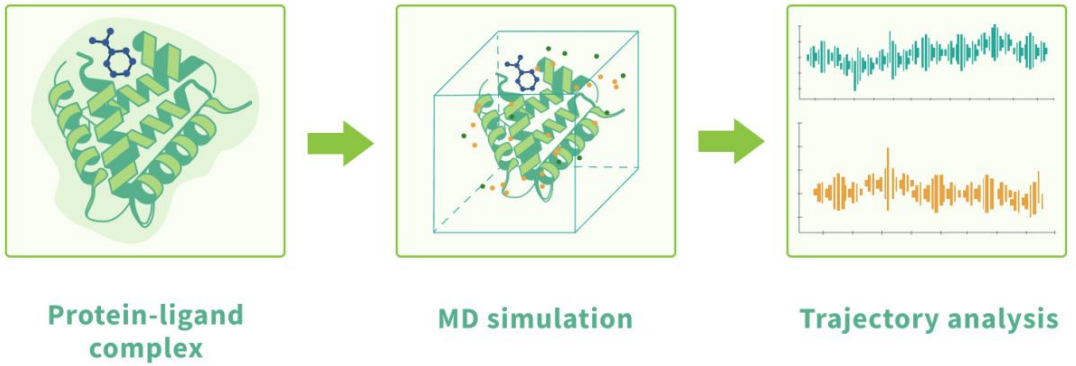
AlphaFold 2

AlpaFold2 can only give us static structures

AlphaFold 2



MD simulation



Jumper, John, et al. "Highly accurate protein structure prediction with AlphaFold." nature 596.7873 (2021): 583-589.

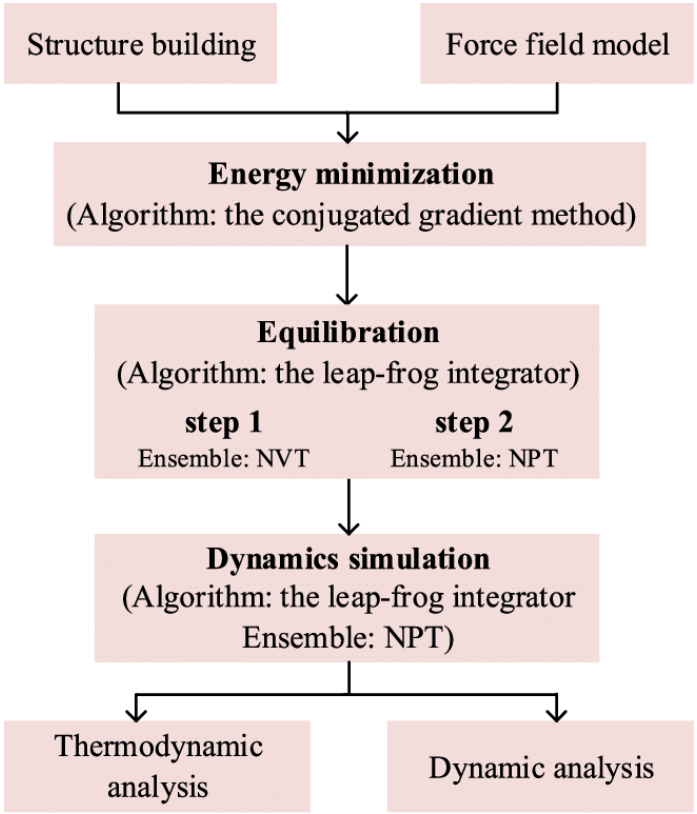
MD simulation

MD simulation:

- Newton's laws of motion
- Repeatedly update the forces on each atom by all other atoms
- Capture position and motion of every atom at every point in time

$$F = ma = -\nabla U(x)$$

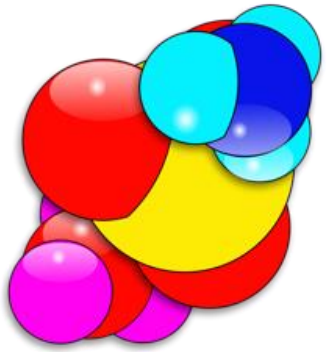
$$\frac{dx}{dt} = v \qquad \frac{dv}{dt} = a$$



Singh, S., Singh, V.K. (2020). Molecular Dynamics Simulation: Methods and Application. In: Singh, D., Tripathi, T. (eds) Frontiers in Protein Structure, Function, and Dynamics. Springer, Singapore. https://doi.org/10.1007/978-981-15-5530-5_9

MD simulation Limitation

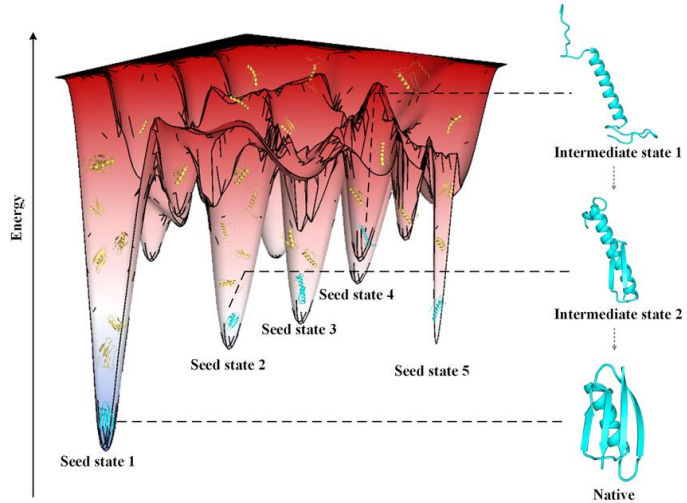
Computational expensive



Folding@home

Folding@home is a distributed computing project for simulating protein dynamics.

Low efficiency for sampling all possible conformation



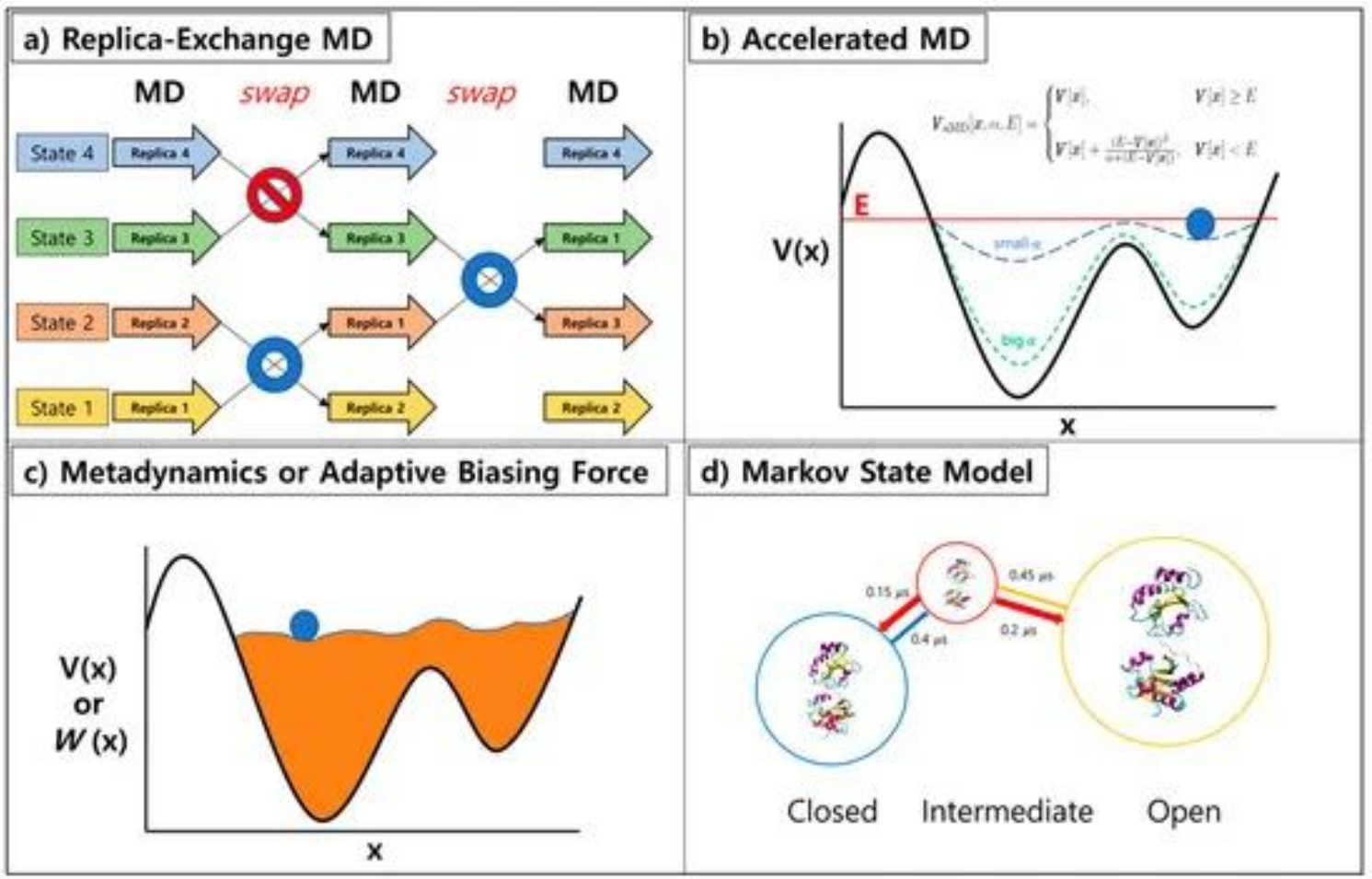
https://foldingathome.org/

Huang, Zhaohong, et al. "Pathfinder: protein folding pathway prediction based on conformational sampling." PLoS Computational Biology 19.9 (2023): e1011438.

Enhance MD simulation

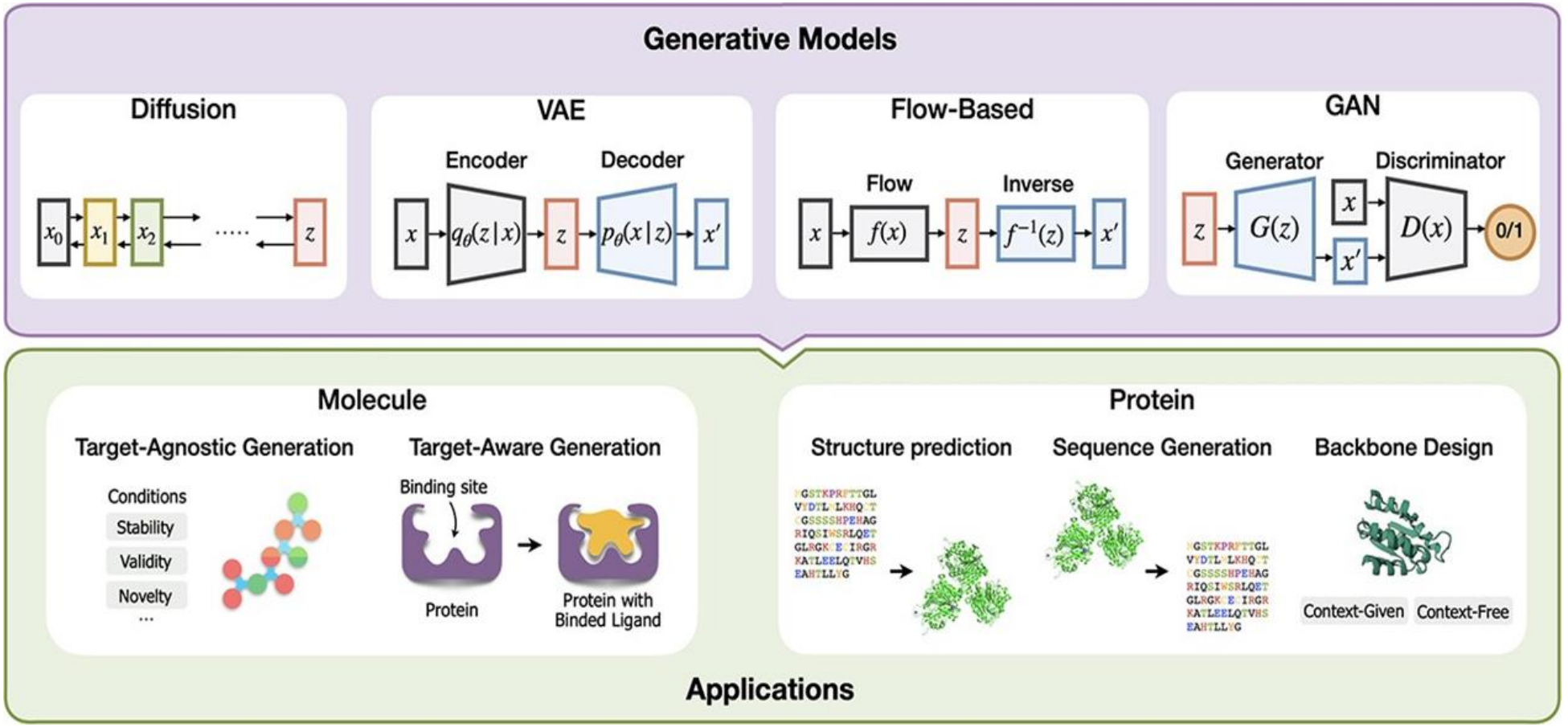
Overview of Enhanced Sampling Methods

- (a) replica-exchange MD simulation
- (b) accelerated MD
- (c) metadynamics or adaptive biasing force
- (d) Markov state model.



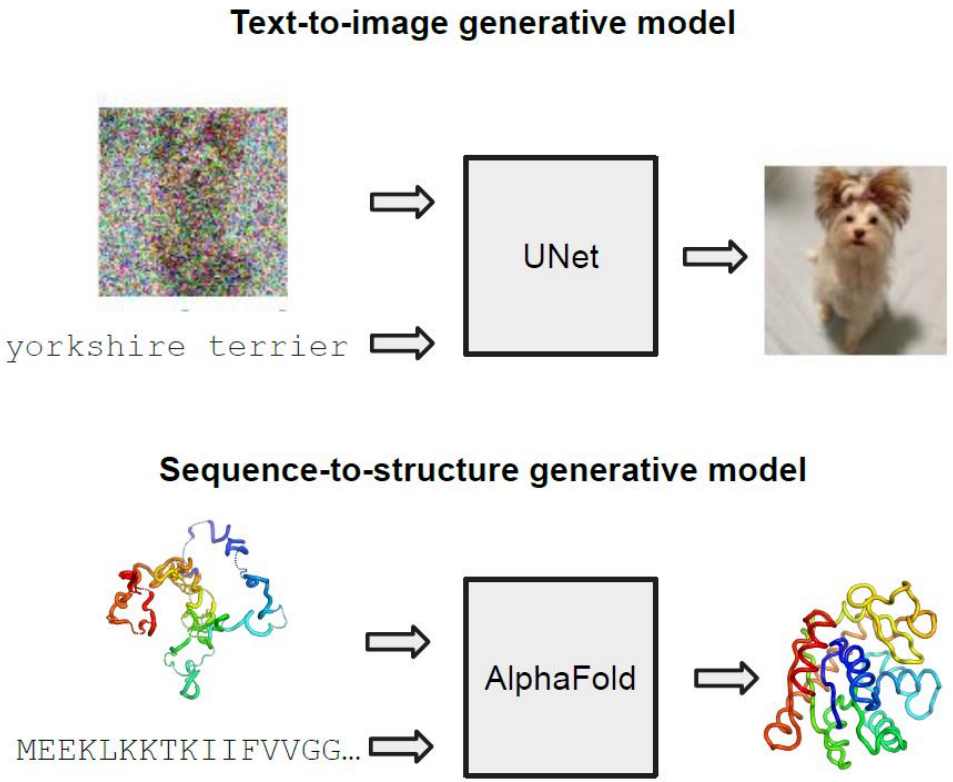
Lazim, R.; Suh, D.; Choi, S. Advances in Molecular Dynamics Simulations and Enhanced Sampling Methods for the Study of Protein Systems. Int. J. Mol. Sci. 2020, 21, 6339. https://doi.org/10.3390/ijms21176339

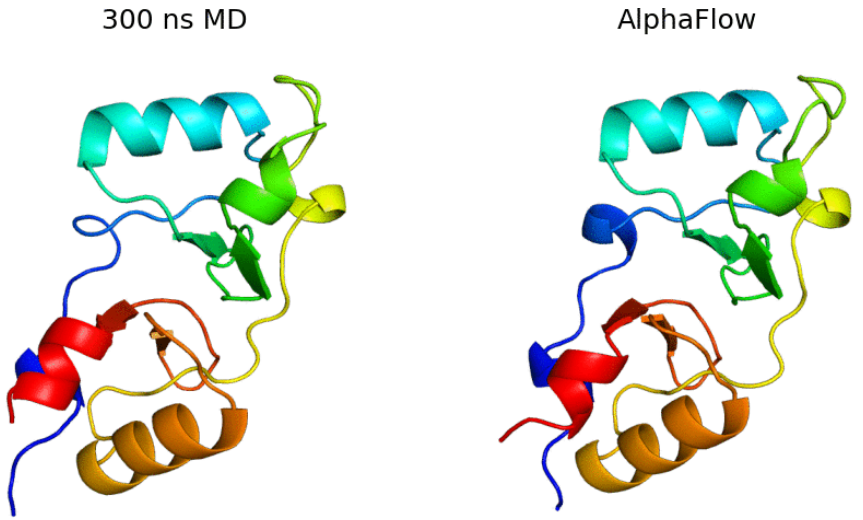
Generative AI



Tang, Xiangru, et al. "A survey of generative ai for de novo drug design: new frontiers in molecule and protein generation." Briefings in Bioinformatics 25.4 (2024): bbae338.

AlphaFlow





Jing, Bowen, Bonnie Berger, and Tommi Jaakkola. "AlphaFold meets flow matching for generating protein ensembles." arXiv preprint arXiv:2402.04845 (2024).

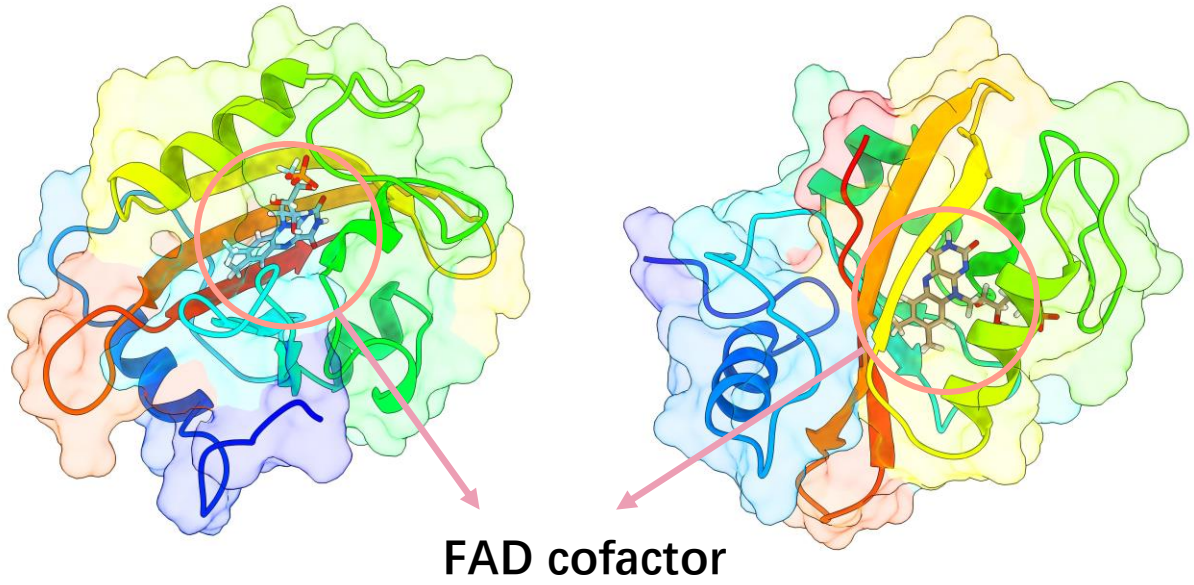
Complex Protein Systems

The presence of small molecules or cofactors that significantly influence the protein's conformational ensemble

The vivid Photoreceptor protein system

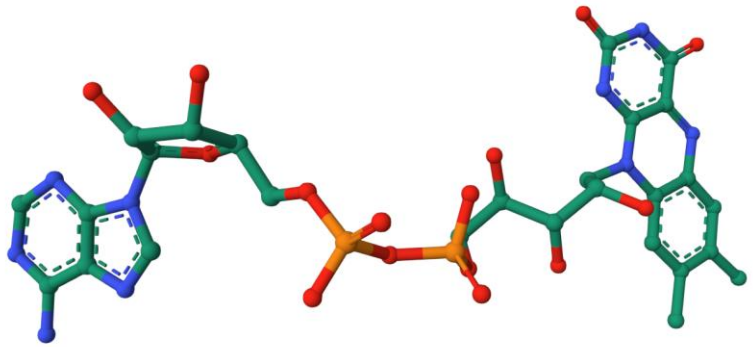
Dark State

Light State



FAD cofactor

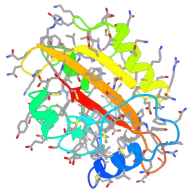
Flavin Adenine Dinucleotide (FAD)

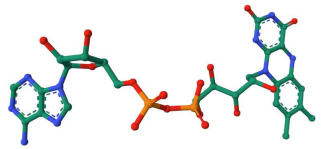


Pudasaini A, El-Arab KK, Zoltowski BD. LOV-based optogenetic devices: light-driven modules to impart photoregulated control of cellular signaling. Front Mol Biosci. 2015; 2(May):18. https://doi.org/10.3389/ fmolb.2015.00018 PMID: 25988185

Complex Protein Systems

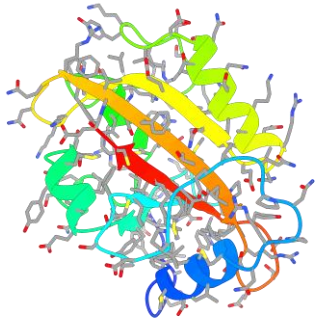
The presence of small molecules or cofactors that significantly influence the protein's conformational ensemble



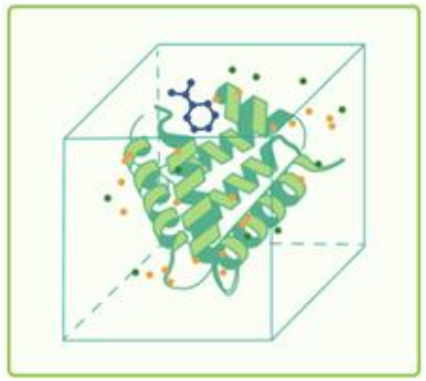


Flavin Adenine Dinucleotide (FAD)

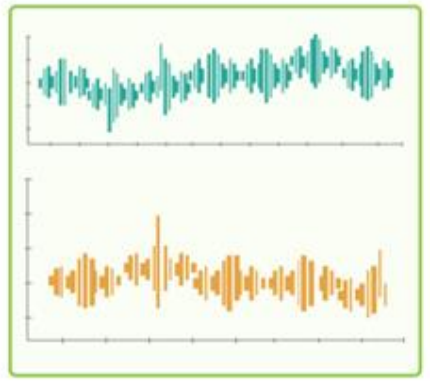
Include FAD



Not Include FAD



MD simulation



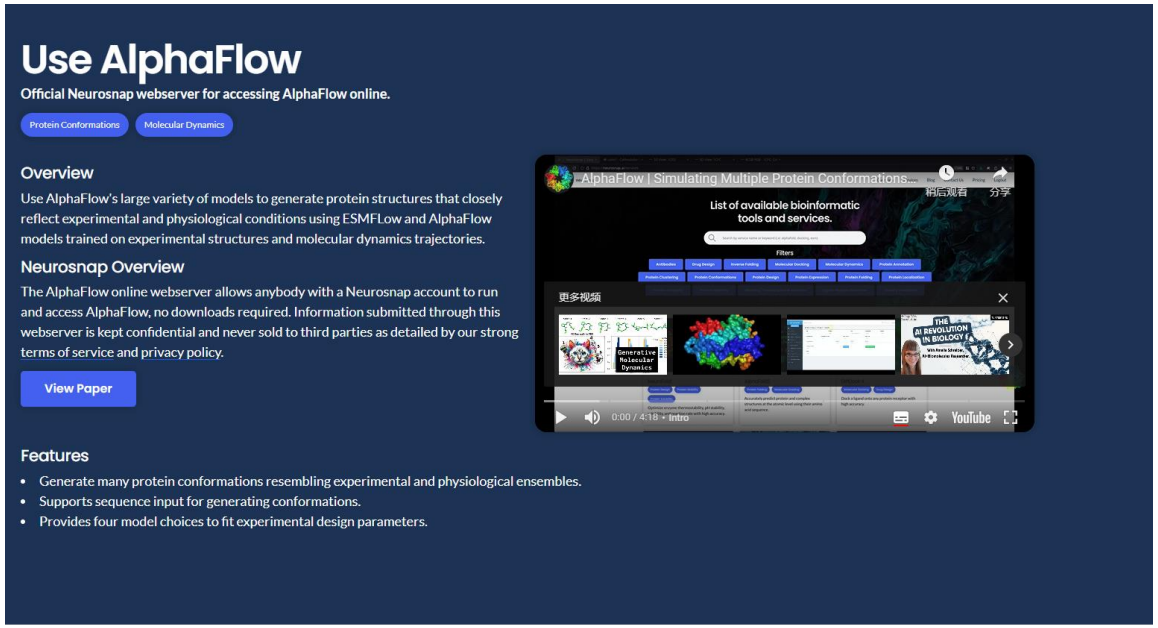
Trajectory analysis

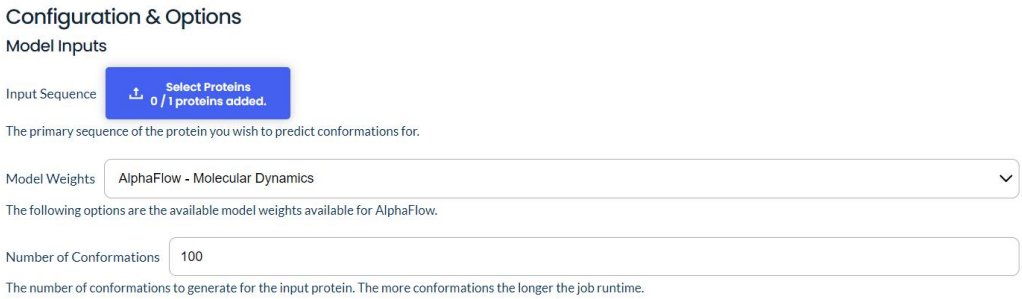
Pudasaini A, El-Arab KK, Zoltowski BD. LOV-based optogenetic devices: light-driven modules to impart photoregulated control of cellular signaling. Front Mol Biosci. 2015; 2(May):18. https://doi.org/10.3389/ fmolb.2015.00018 PMID: 25988185

Complex Protein Systems

The presence of small molecules or cofactors that significantly influence the protein's conformational ensemble

So only use the sequence as input, It is challenging to produce accurate ensembles.





https://neurosnap.ai/service/AlphaFlow

Part I

SUMMARY

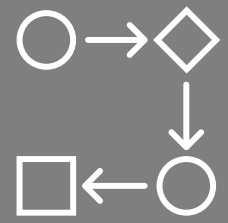
- Protein function is tied to its structure, with both static and dynamic conformations playing critical roles in defining biol ogical activity.

- MD simulation is the primary tool for exploring the dynamic prope rties of protein structures, the major limitation of MD simulations i s their requirement for extensive computational resources.

- Generative AI has opened new avenues for protein conformation ge neration, but the current models struggle to accurately generate con formations for complex protein systems.

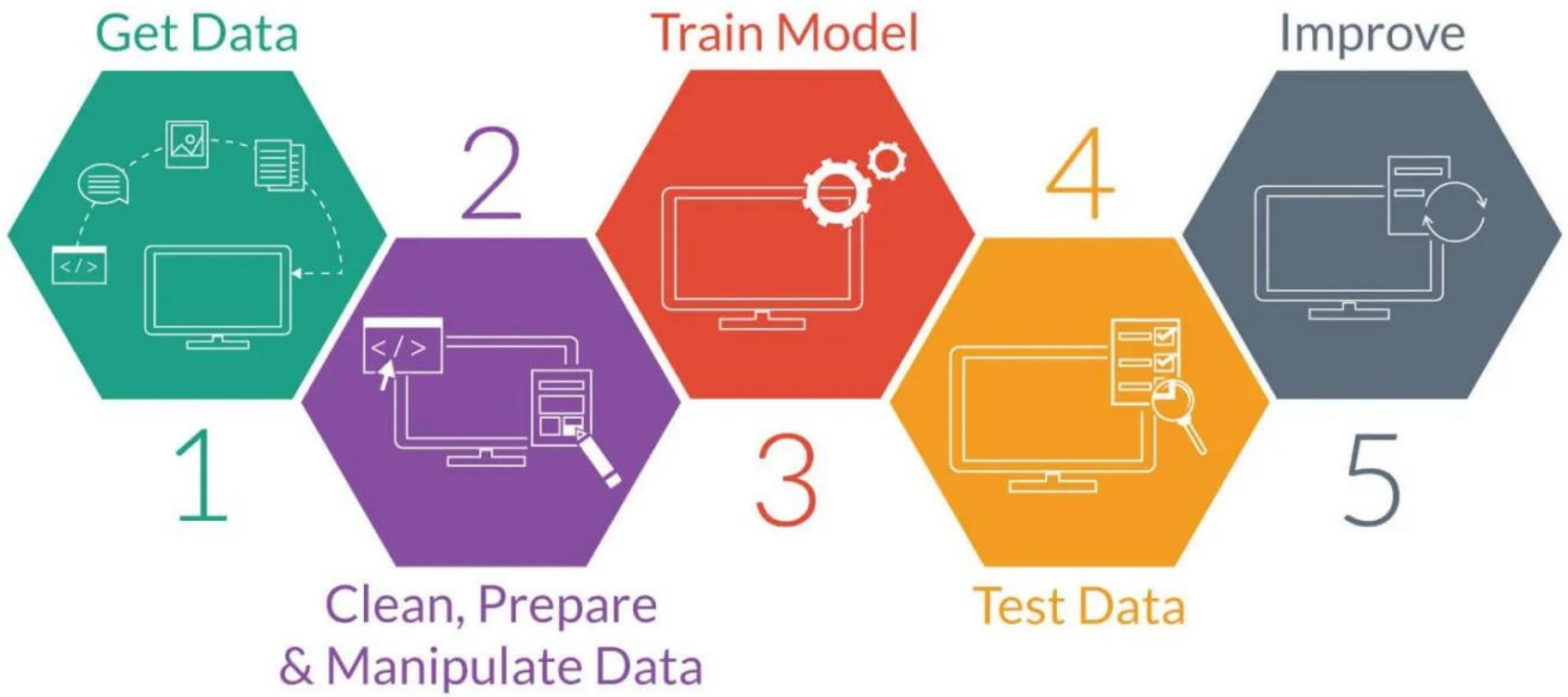
Part II

A transformer-based diffusion model to generate protein conformations



Part II

The steps of developing a deep learning model



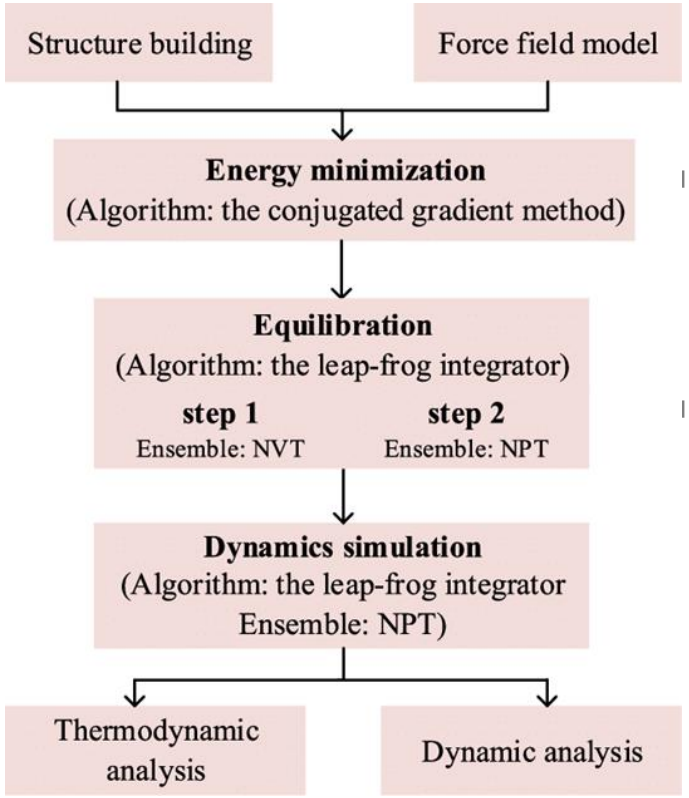
https://towardsdatascience.com/building-a-deployable-ml-classifier-in-python-46ba55e1d720

Part II

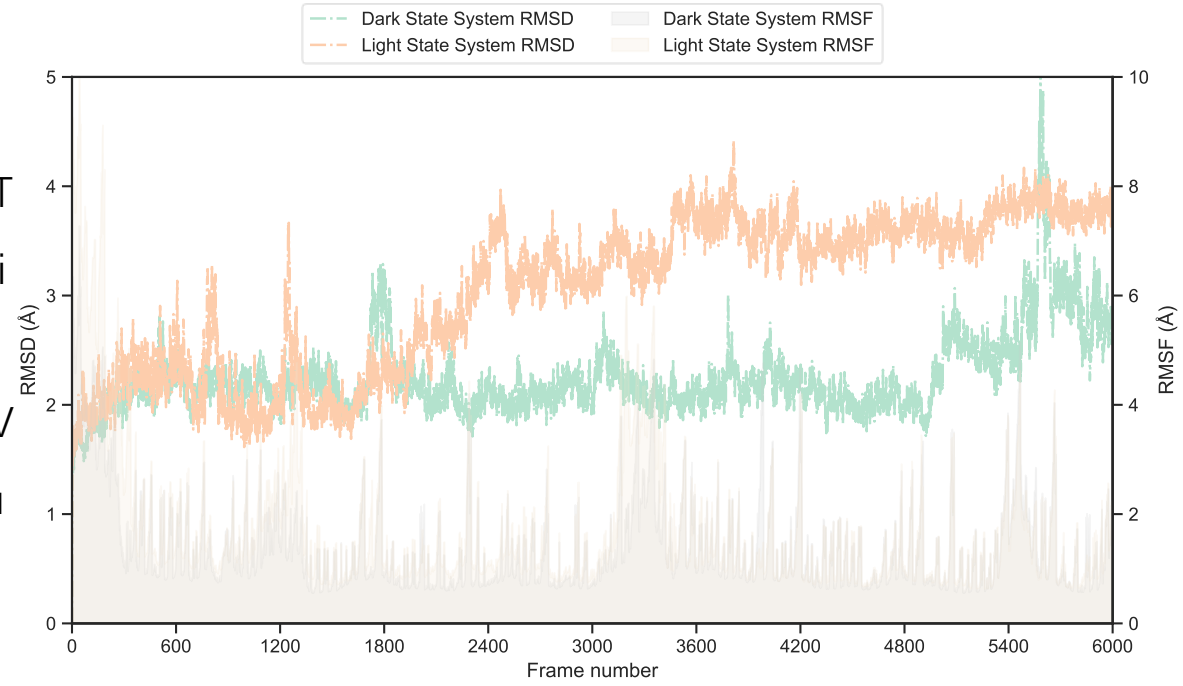
01 GET DATA

We collected a total of **6,001** distinct conformations of the VVD protein of each system.

A total of two simulation systems were constructed

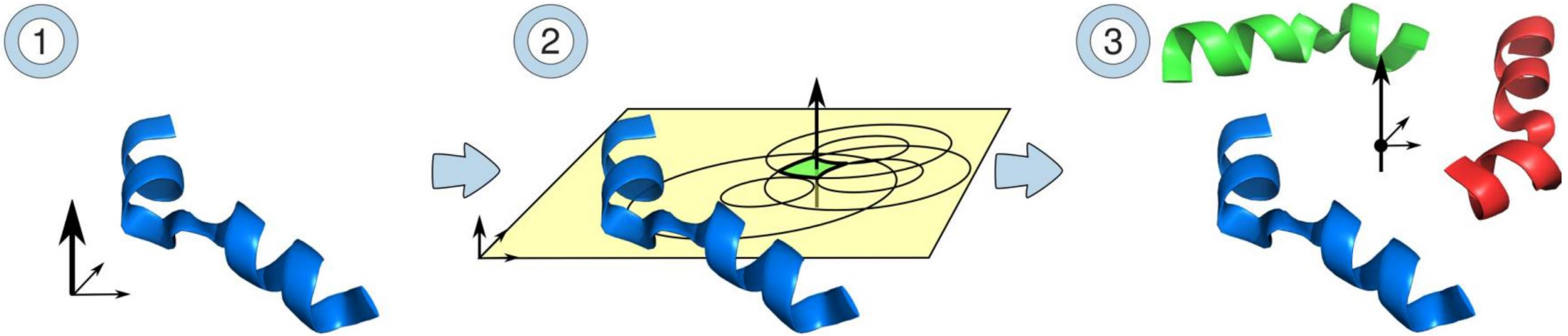


- 10 nanoseconds (ns) of NPT MD simulations for Equilibrium
- 1.1 microseconds (μs) of NVT MD simulations for production



The input feature and output feature in our model

SE(3) symmetry of protein structures



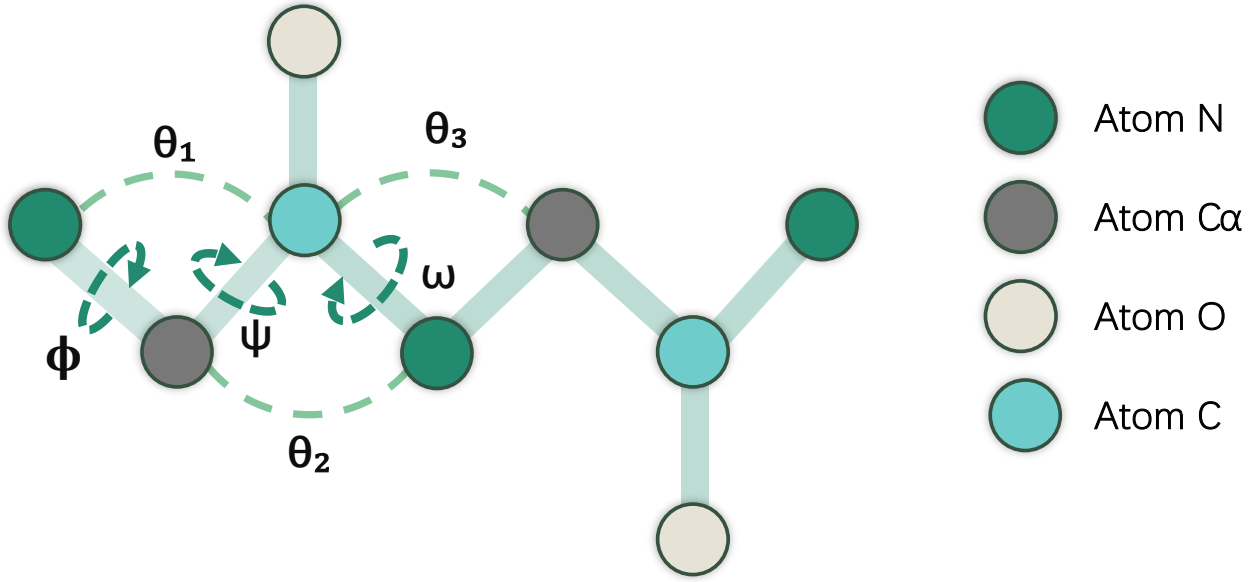
Martin, Jeffrey W., et al. "A geometric arrangement algorithm for structure determination of symmetric protein homo-oligomers from NOEs and RDCs." Journal of Computational Biology 18.11 (2011): 1507-1523.

02 Clean, Prepare, and Manipulate data

The input feature and output feature in our model

Six types of angles from the protein backbone structures during the MD simulation

- dihedral torsions ('ϕ', 'ψ', 'ω')
- bond angles ('θ₁', 'θ₂', 'θ₃')
- Data shape (N,147,6)



One of the Input and output features in our first model

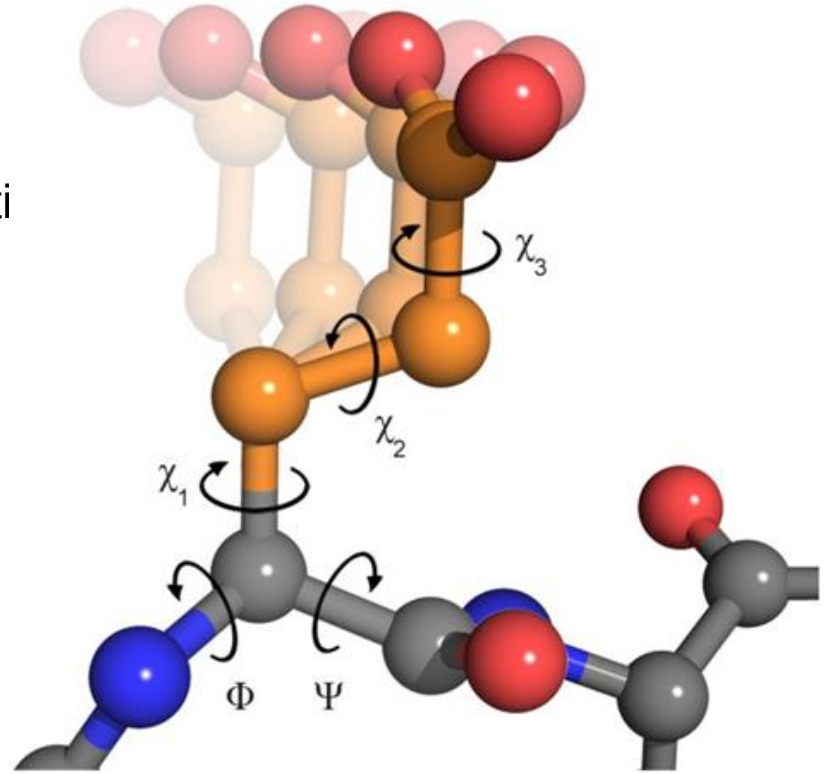
Wu, K.E.; Yang, K.K.; Van Den Berg, R.; Alamdari, S.; Zou, J.Y.; Lu, A.X.; Amini, A.P. Protein Structure Generation via Folding Diffusion. Nat Commun 2024, 15, 1059, doi:10.1038/s41467-024-45051-2.

02 Clean, Prepare, and Manipulate data

The input feature and output feature in our model

- **The dynamic nature of the structure** during MD simulation.
- **Angular Deviations**(changes) relative to the structure from the first ti of the MD conformation set.

The other of the Input and output features in **our second model**

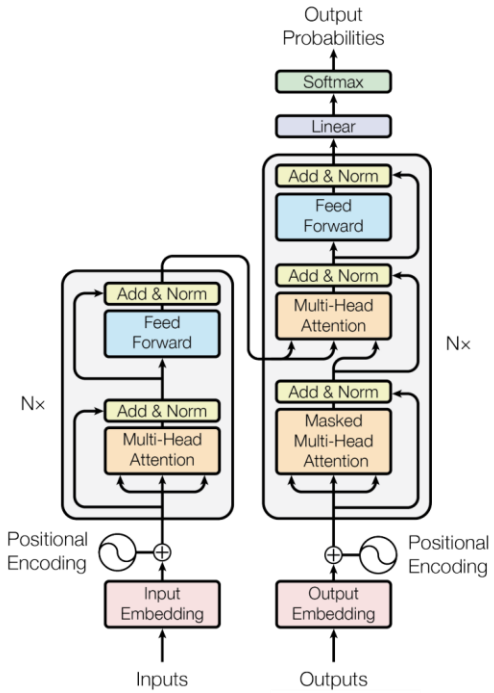


Angular Deviations

A transformer-based diffusion model

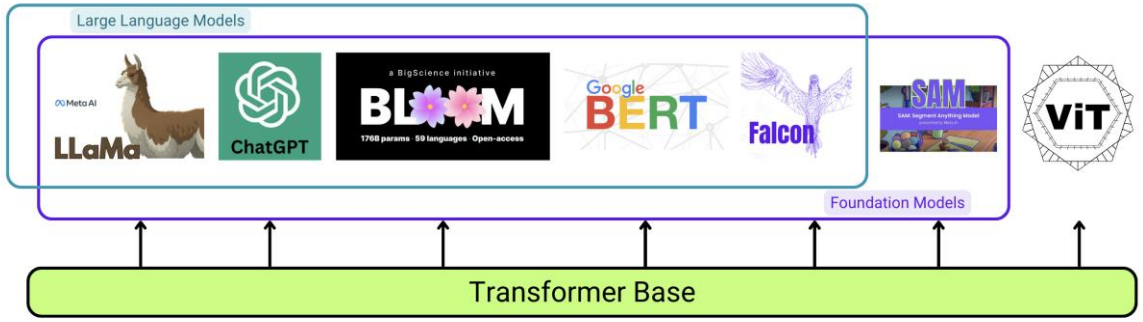
BERT

Encoder



GPT

Decoder



Hyperparameter	Value
Hidden Size	384
Number of Layers	12
Number of Attention Heads	12
Intermediate Size	768
Dropout Rate	0.1
Max Position Embeddings	147
Position Embedding Type	Relative Key
Temporal Embedding Type	Random Fourier Features

Only use the Encoder part like BERT

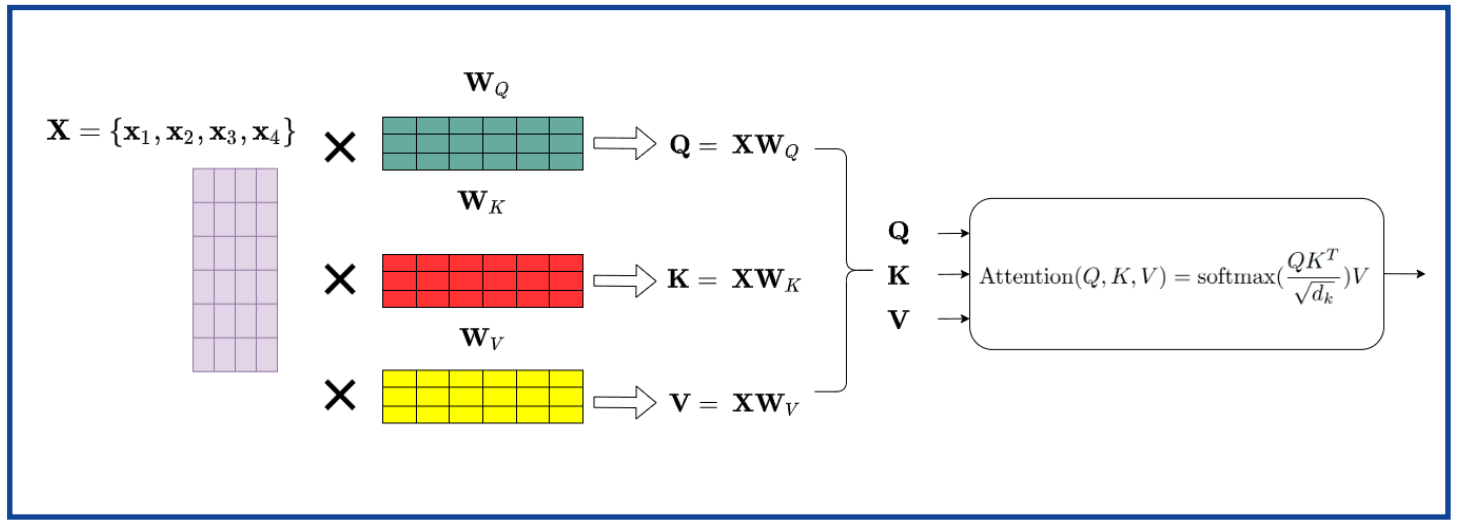
Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); Burstein, J., Doran, C., Solorio, T., Eds.; Association for Computational Linguistics: Minneapolis, Minnesota, June 2019; pp. 4171–4186.

A transformer-based diffusion model

Attention is all you need

INPUT Data Shape:
(**Batchsize, 147,6**)

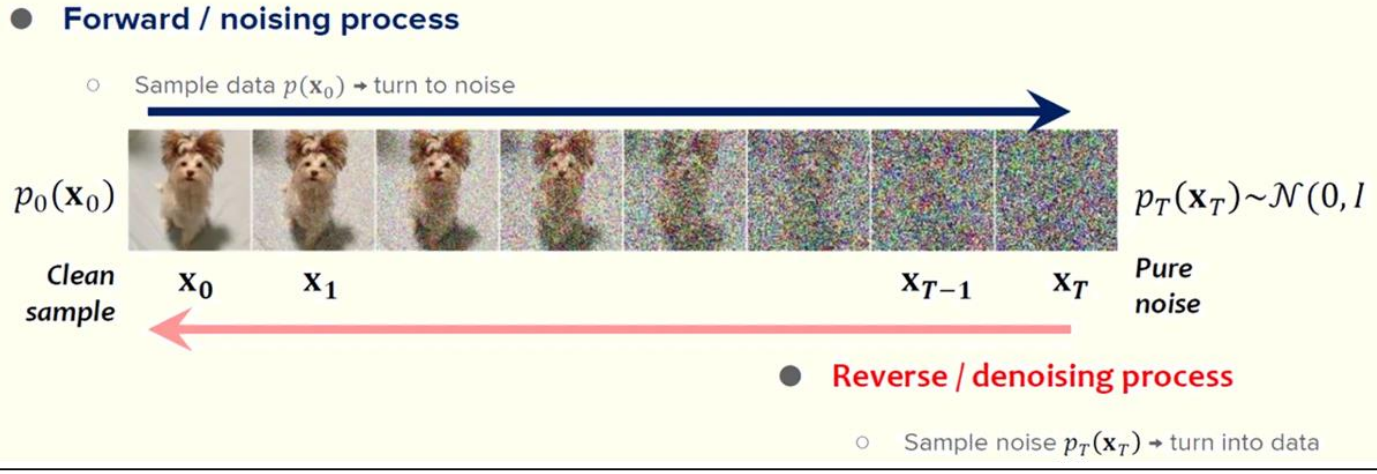
OUTPUT Data Shape:
(Batchsize, 147,6)



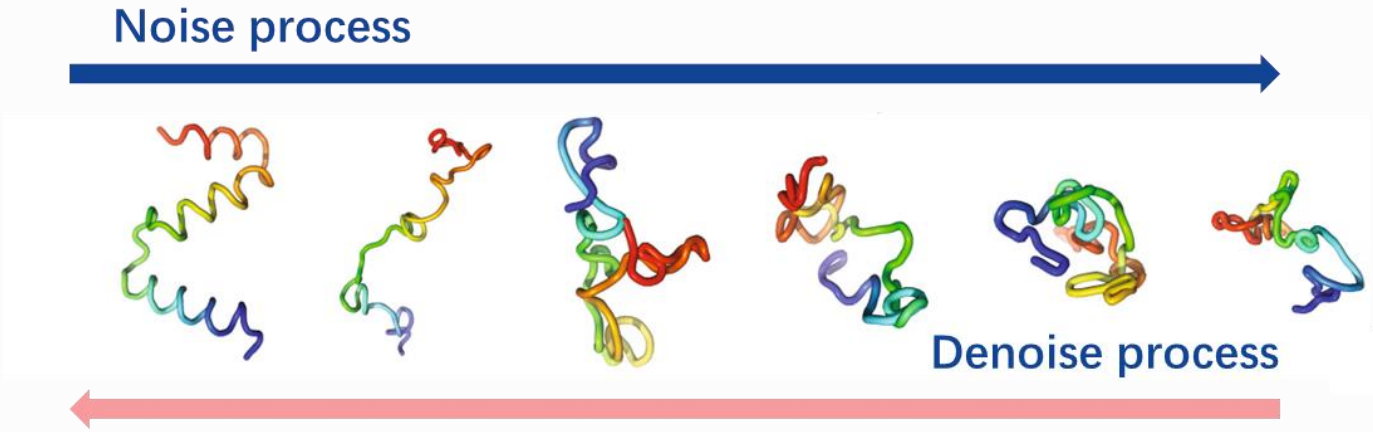
Vaswani, A. "Attention is all you need." Advances in Neural Information Processing Systems (2017).

A transformer-based diffusion model

Computer Vision



Protein Structure



Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." Advances in neural information processing systems 33 (2020): 6840-6851.

03 Model Algorithm

A transformer-based diffusion model

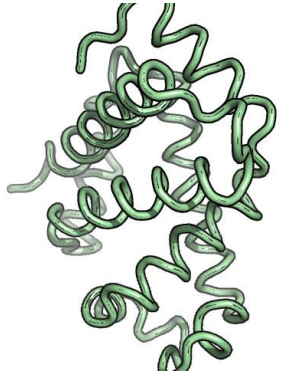
Training Algorithm

1: $x = x_{angle} - x_{ref}$ //calculate the angle deviations by subtracting the reference angle
2: do
3: $x_0 \sim q(x)$ // sample the data from training data Set q(x)
4: $t \sim U(\{1,\ldots\ldots, T\})$ //sample the time t from a uniform distribution from 1 to T
5: $\varepsilon \sim \mathcal{N}_{wrapped}(0, \mathbf{I})$ // sample the noise value from the wrapped gaussian distribution
6: gradient decent $\nabla_\theta(L_w(\varepsilon, t, x_0))$ //gradient decent on loss function
7: until reaching the converge criterion

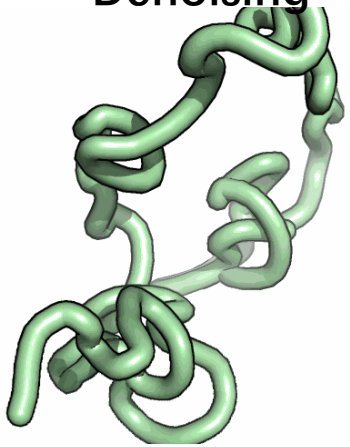
Sampling Algorithm

1: $x_t \sim \mathcal{N}_{wrapped}(0, \mathbf{I})$ //sampling from the wrapped gaussian distribution
2: for t=T,……,1 do
3: if not last step: $z \sim \mathcal{N}(0, \mathbf{I})$ else $z = 0$
4: calculated x_{t-1} by equation $x_{t-1} = \omega\left(\frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon_\theta(x_t, t))\right) + \zeta z\sigma_t$
5: end for
6: generated angle = $\omega(x_0 + x_{ref})$
7: reconstruct the Cartesian coordinate

Noising

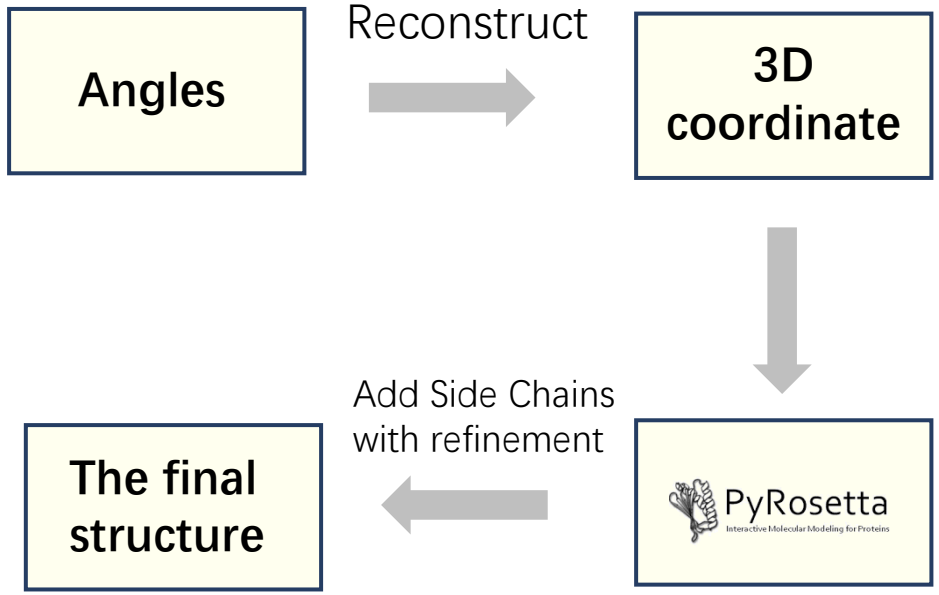


Denoising

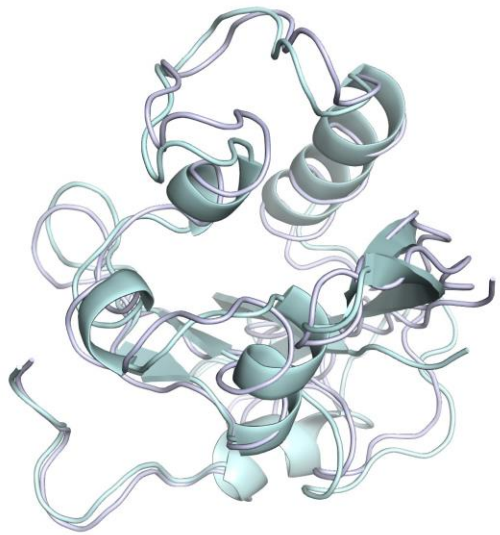


03 Add Side Chain

A transformer-based diffusion model

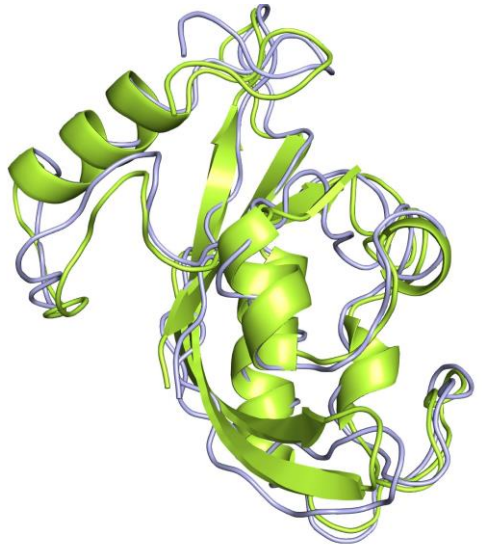


Dark State



Backbone RMSD=1.9 Å

Light State

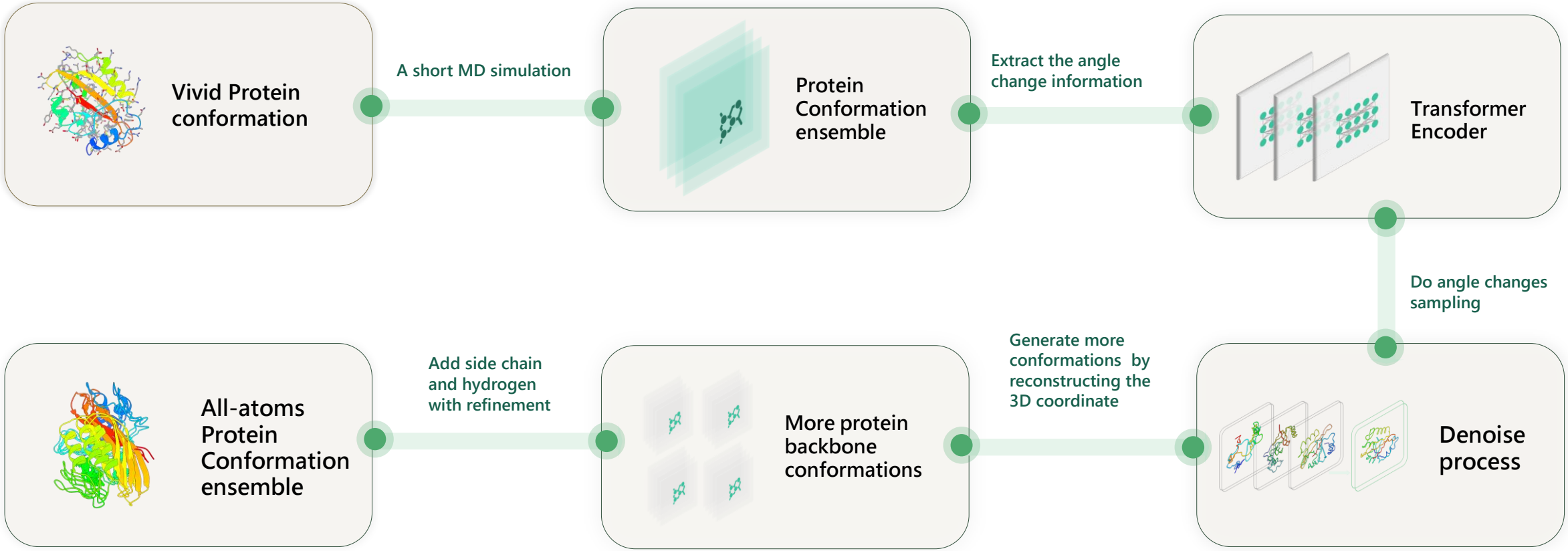


Backbone RMSD=1.8 Å

Chaudhury, S.; Lyskov, S.; Gray, J.J. PyRosetta: A Script-Based Interface for Implementing Molecular Modeling Algorithms Using Rosetta. Bioinformatics 2010, 26, 689–691, doi:10.1093/bioinformatics/btq007.

Part II

CONCLUSION

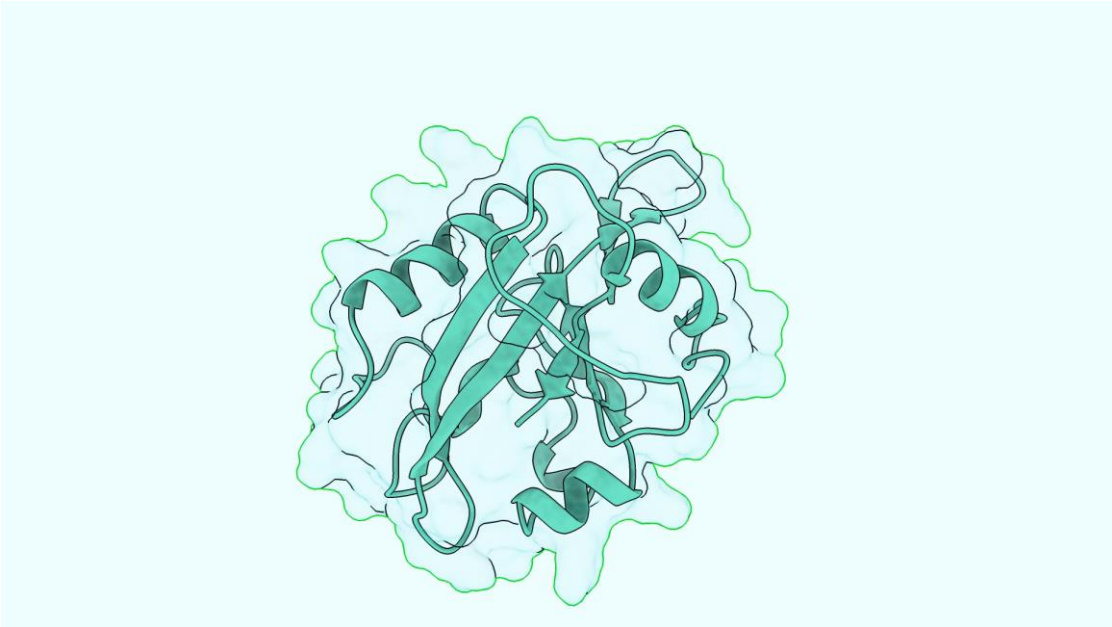


Part III

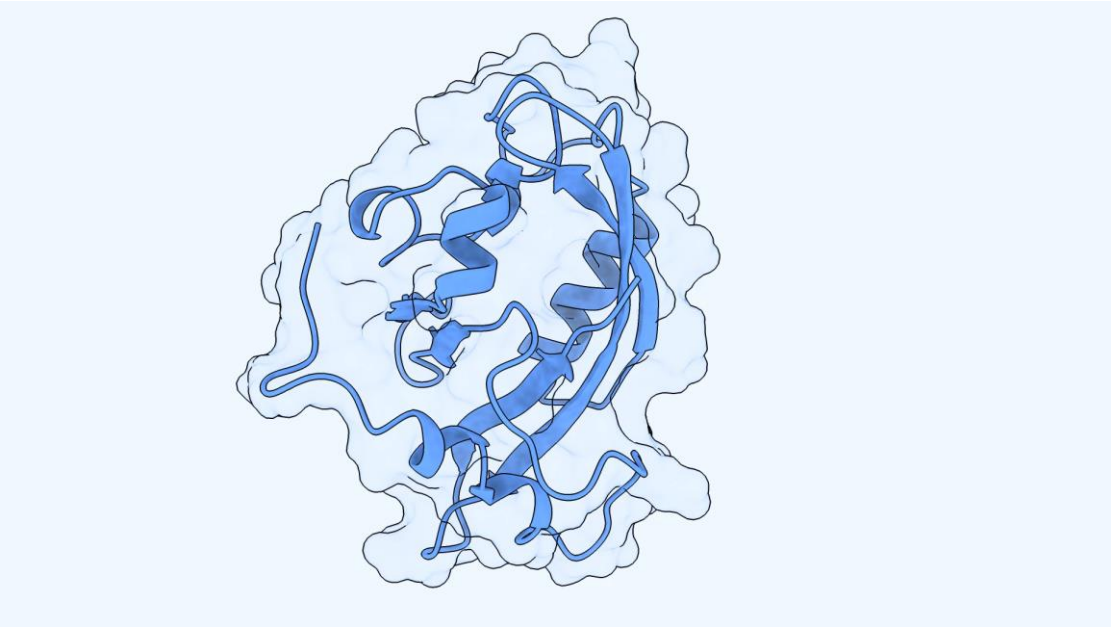
Computational Results

- Angle vs angle deviation
- Compare with MD simulation
- Compare with AlphaFlow
- Limitation

Light State



Dark State



Part III

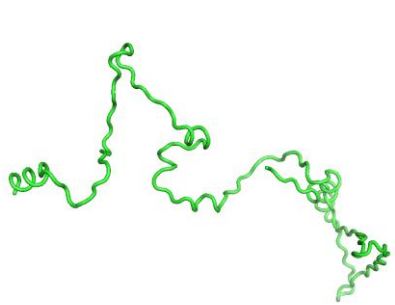
Angles vs Angle Deviation

Using Angle Deviation is Superior to Using Angles

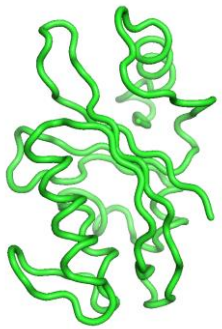
500 structures

X 10

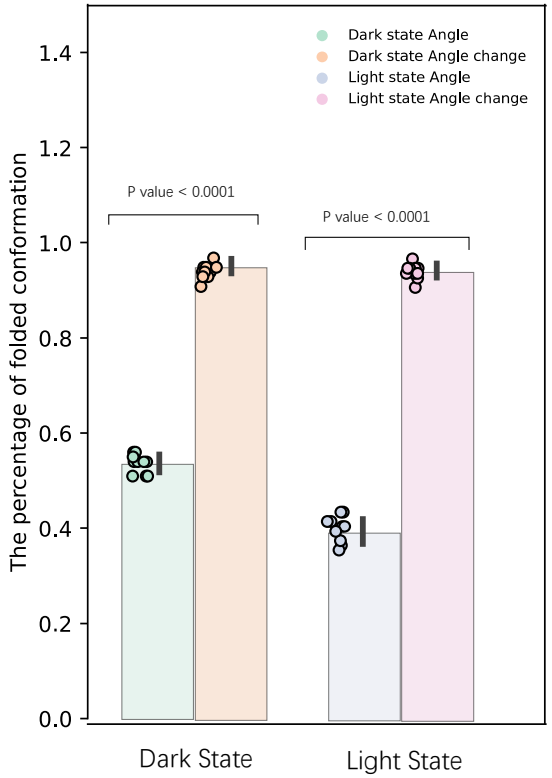
Generate 500 structures for each systems and did ten times



Unfolded Structures RMSD > 6



Folded Structures RMSD <= 6

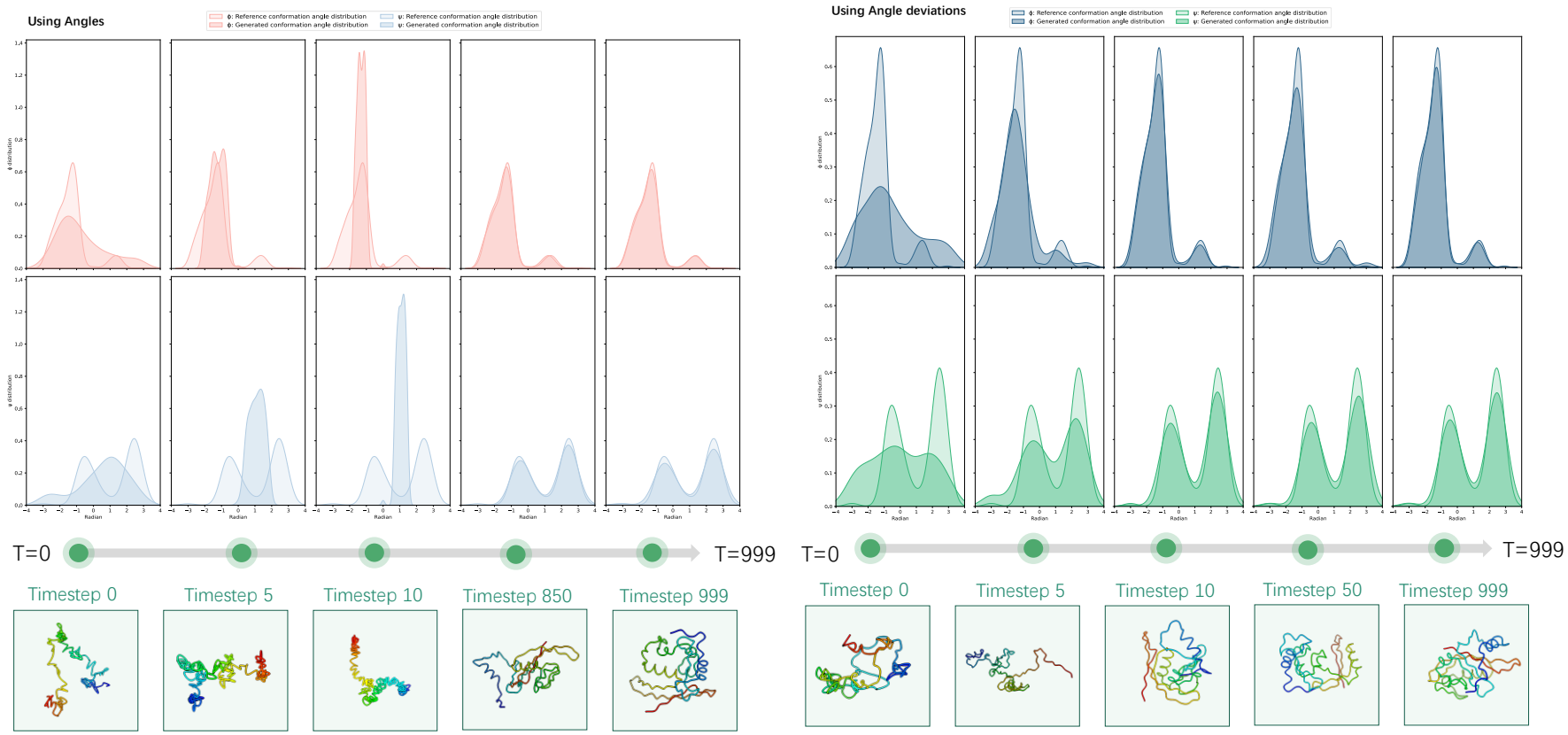


Comparison of the proportion of correctly folded structures generated using absolute angles versus angular deviation

Angles vs Angle Deviation

Using Angle Deviation is Superior to Using Angles

Denoising process

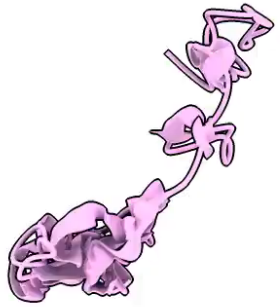


Denoising process of using angles and angle deviations for the Dark state system. Shifts in angle distributions, and conformations. (a) using angles (b) using angle deviations

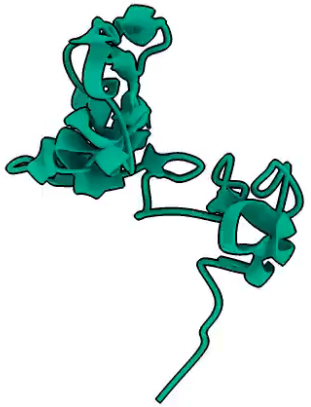
Angles vs Angle Deviation

Using Angle Deviation is Superior to Using Angles

Using Angles



Using Angle Deviations



DENOISE PROCESS OF USING ANGLES AND ANGLE DEVIATIONS

time step: 0

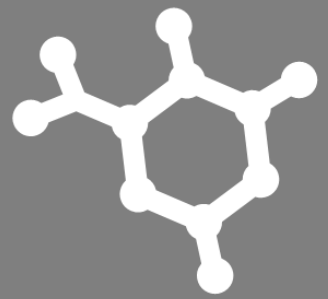
Part III

Evaluations of conformations

Using Angle Deviation

Our model

Generate 6000 conformations for each systems

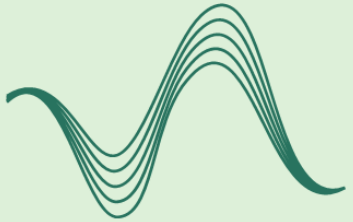


MD simulation

Collect 6001 conformations for each systems

Alpha Flow

Generate 6000 conformations for each systems only use Sequence



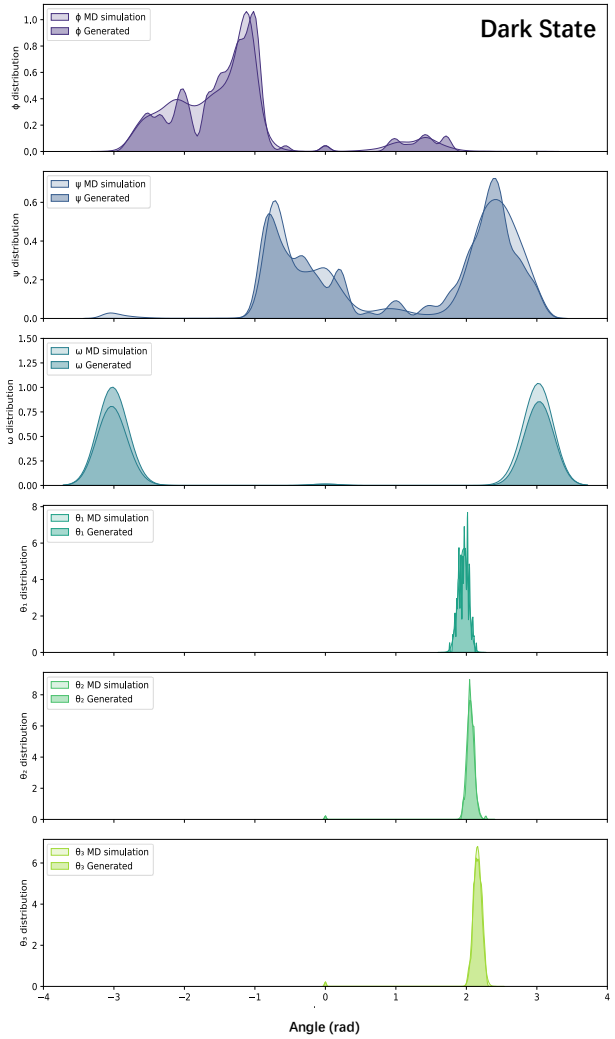
Angles Distribution

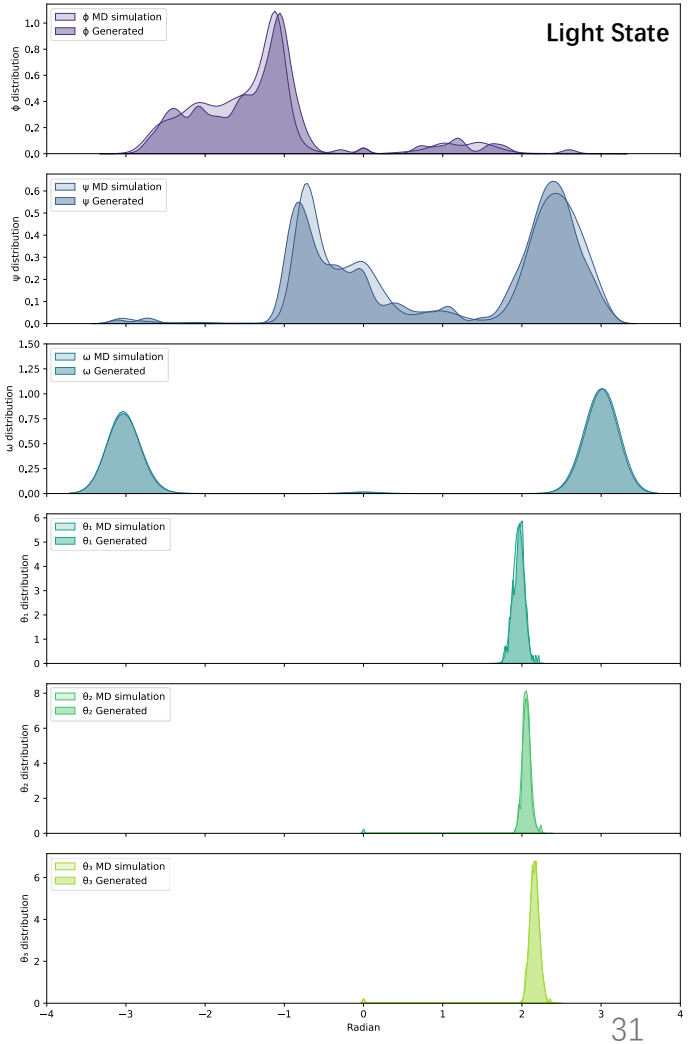
Angle distribution between the generated structures and the MD simulation.

The distributions of the six types of angles between the generated structures and the MD simulation.

(a) Dark state system (b) Light state system

Our model can successfully capture the angle distributions of the original MD simulation for both the dark state and the light state.





Part III

RMSD Distribution

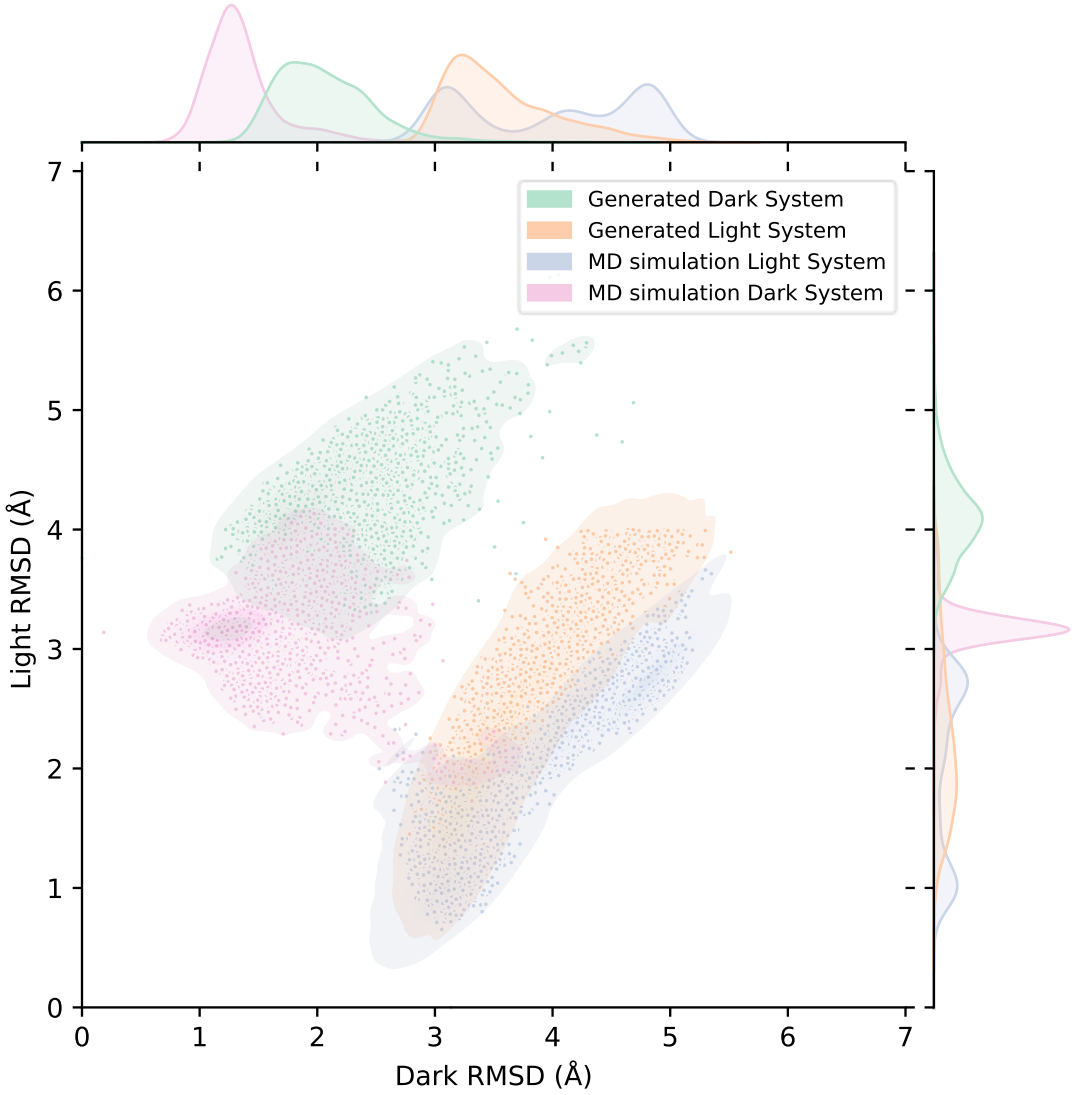
RMSD distributions of the generated structures with the training set.

RMSD distributions of the generated structures with the training set.

The x-axis represents RMSD for the dark state

while the y-axis represents RMSD for the light state.

Our model can explore a more extensive conformational space

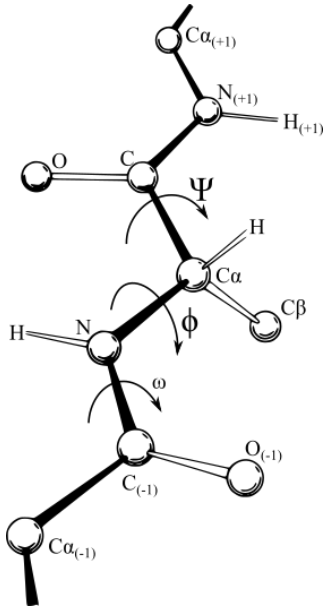


Part III

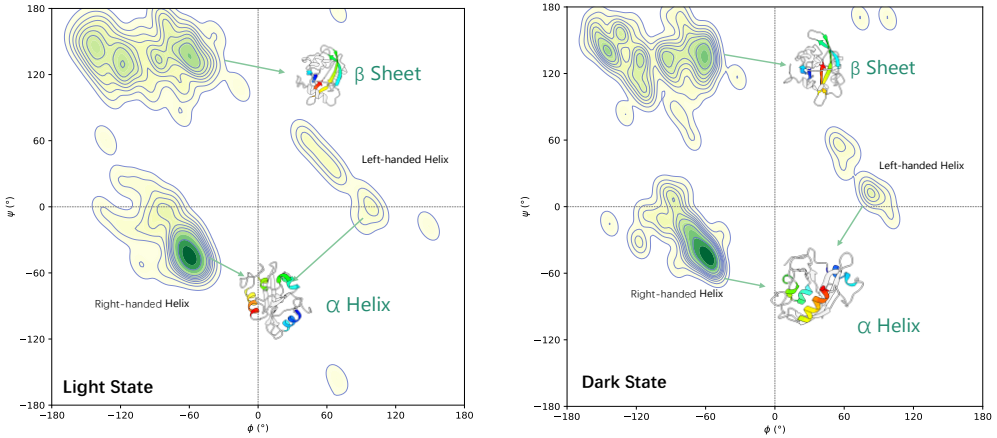
Ramachandran Plot

our generated structures possess reasonable second structures

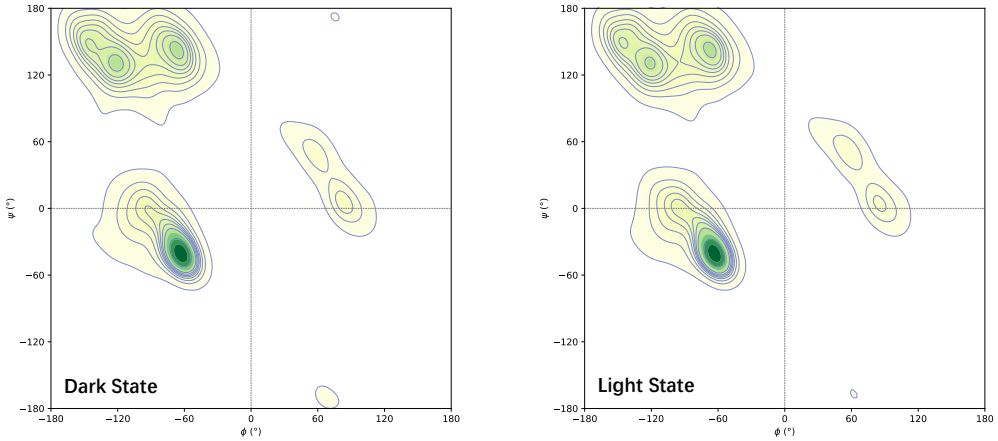
Ramachandran plot [φ,ψ] plot



The generated ensembles.



MD simulation ensembles

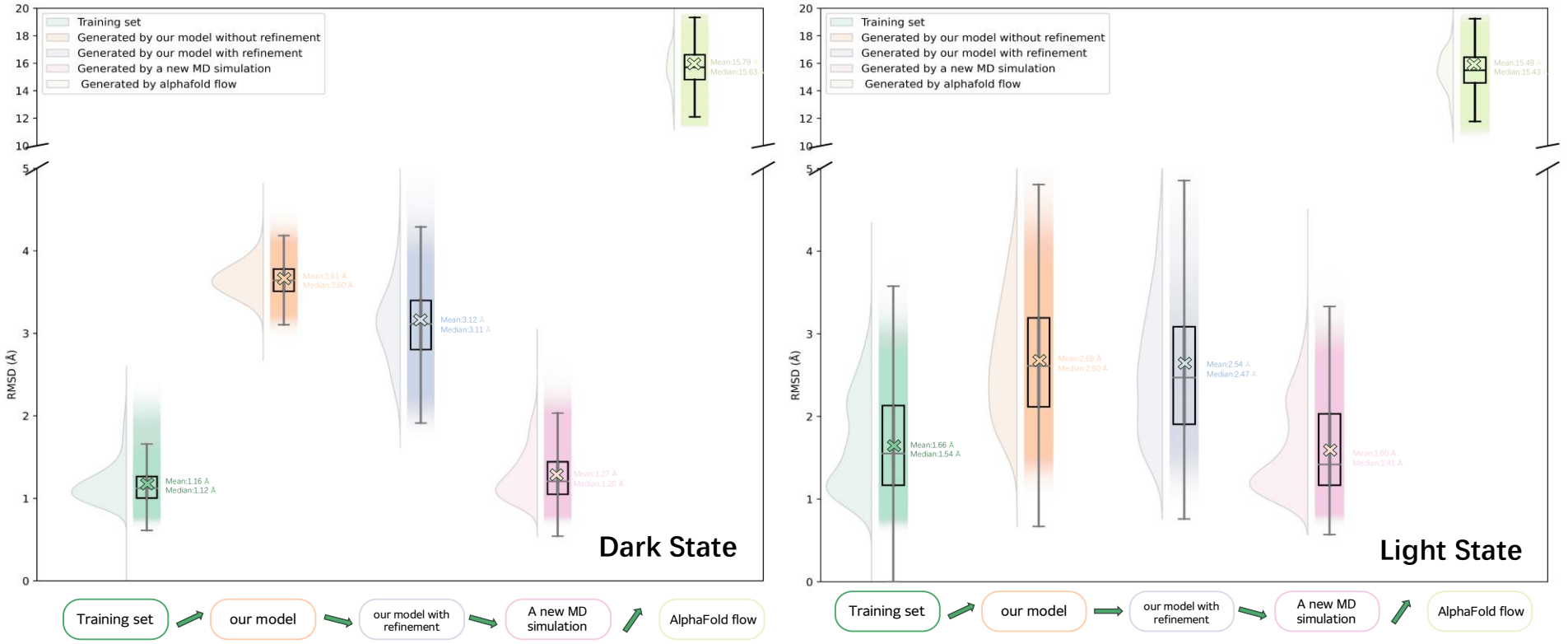


Our model can successfully capture secondary structure features of protein structures, including alpha helix and beta sheet

Part III

Pairwise RMSD Distribution

Compare with Alphaflow



our model, particularly with refinement, produces structures closer to the original MD ensemble than those generated by AlphaFold Flow,

Limitation

After refinement, there is no bond crash or break

Physical Reasonable

Bond Clash:

The distance between any two α-carbon is **less than** two times the van der Waals radius of α-carbon with an overlap tolerance: δ = 2 × 1.7 − 0.4 = **3.0 ˚A**.

5 %

Bond break

Maximum adjacent α-carbon > **4.19 ˚A**

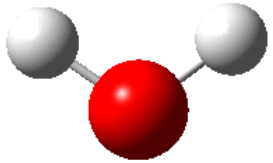
0 %

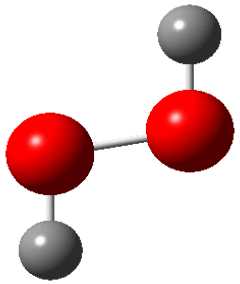
Wang, Yan, et al. "Protein conformation generation via force-guided se (3) diffusion models." arXiv preprint arXiv:2403.14088 (2024).

Future improvement

Combine force field into the model

Force fields: bonded + nonbonded





$$\sum_{bond} K_b\,(b - b_0)^2$$

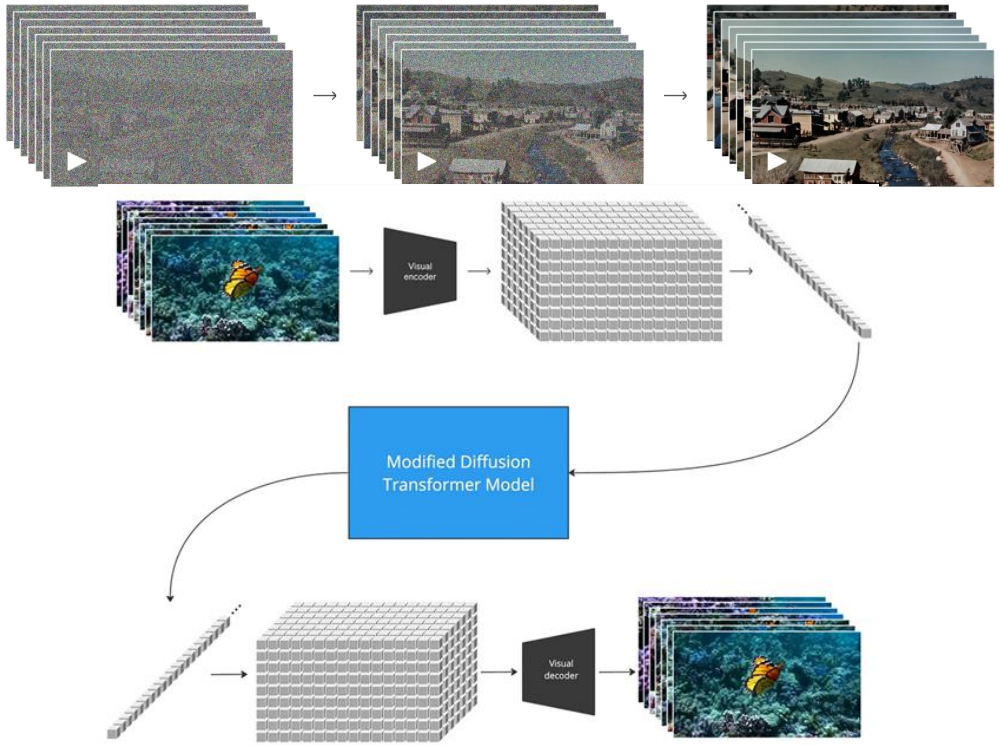
$$\sum_{angle} K_\theta\,(\theta - \theta_0)^2$$

$$\sum_{dihedral} K_\varphi\,(1 + \cos(n\varphi - \theta))^2$$

Future improvement

Video Generations

Sora: Creating video from text

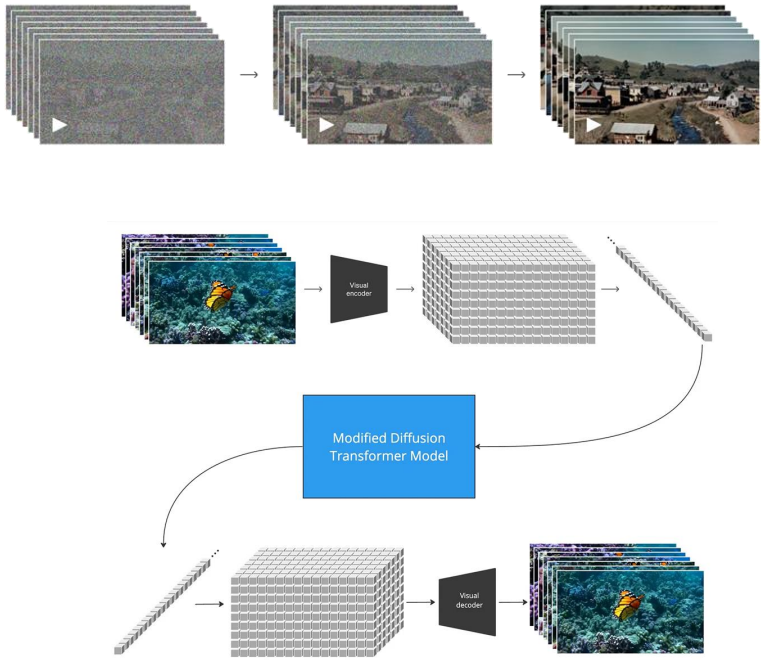


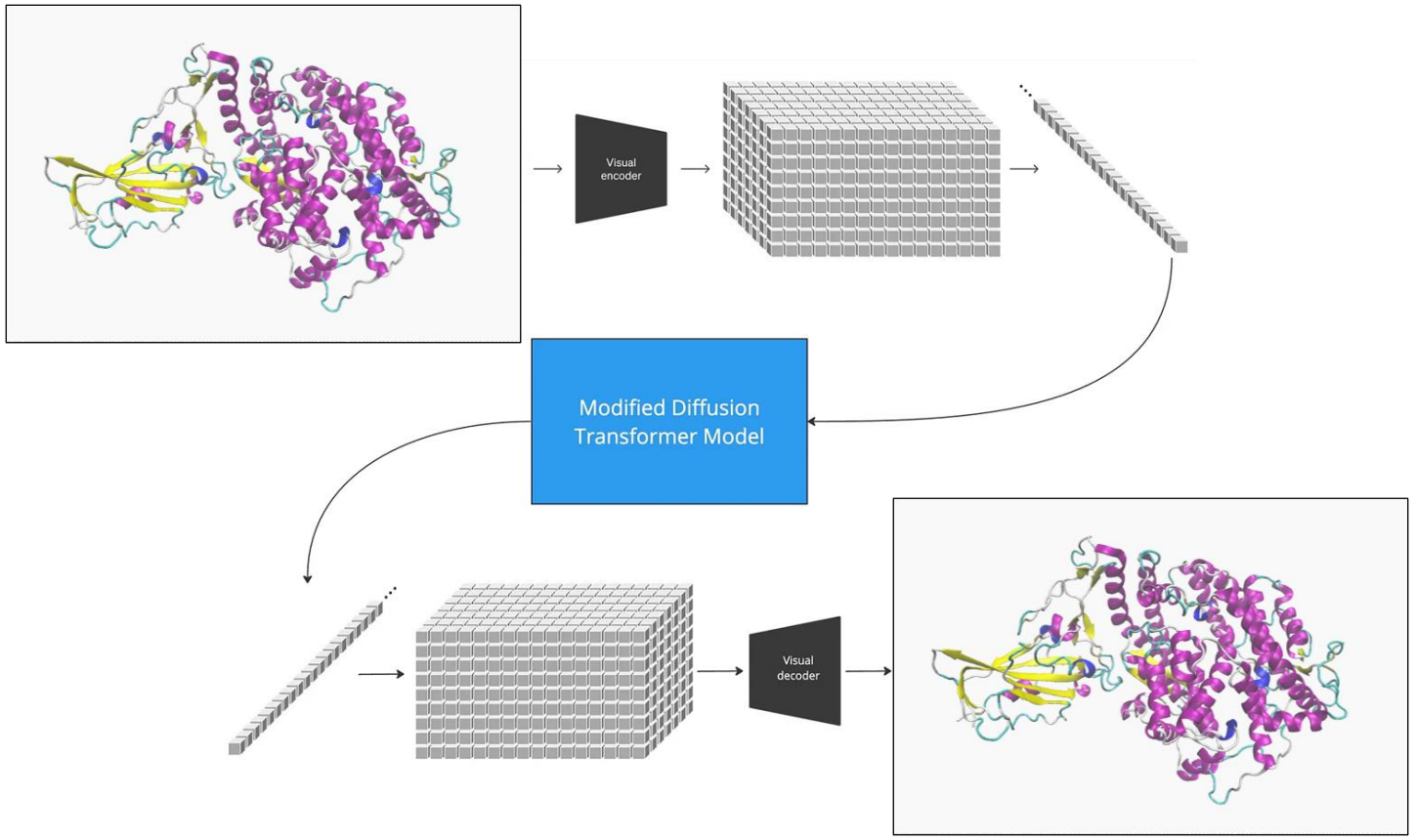
https://openai.com/index/video-generation-models-as-world-simulators/

Future improvement

Video Generations

SORA

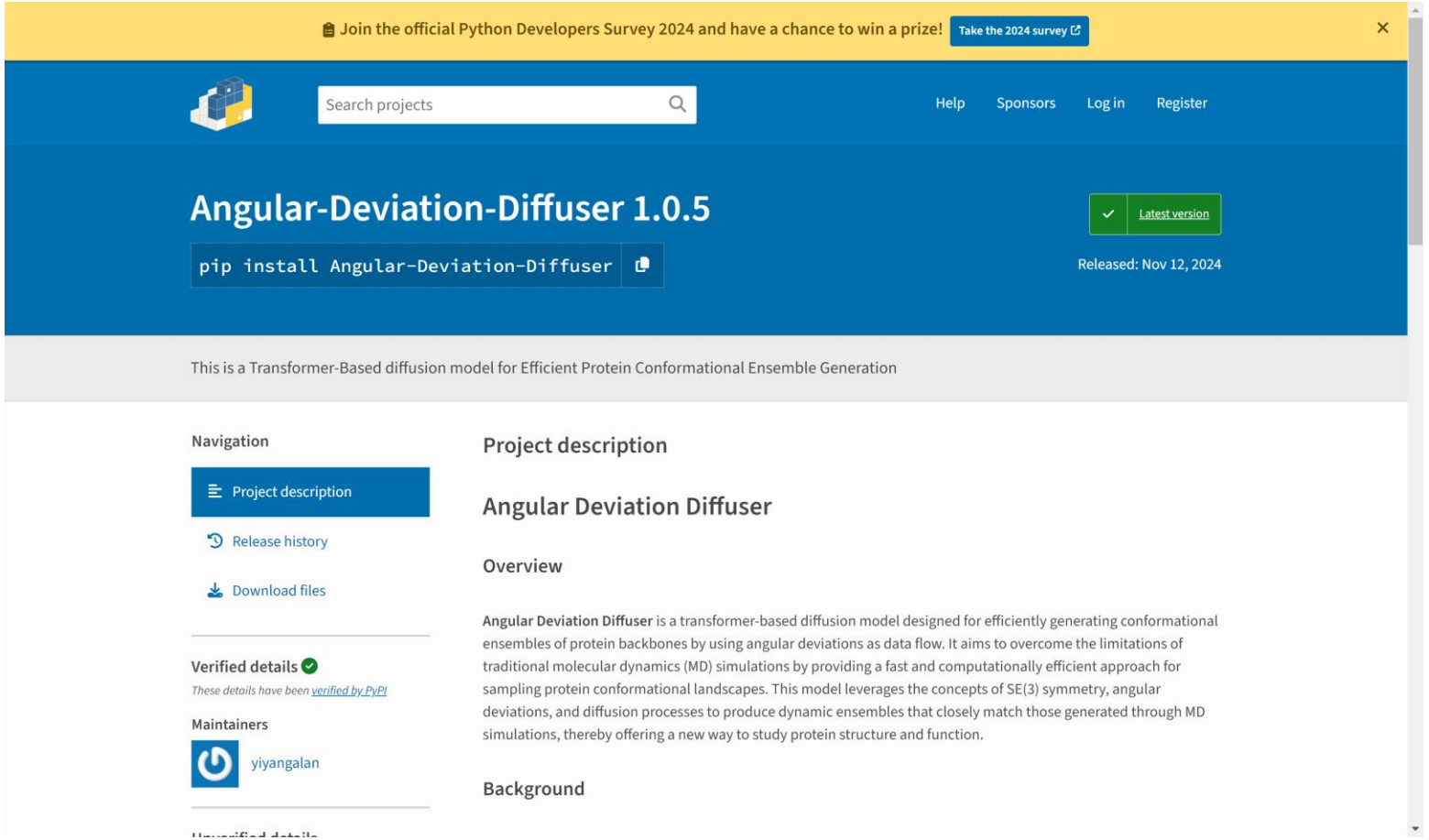




https://openai.com/index/video-generation-models-as-world-simulators/

Angular-Deviation-Diffuser

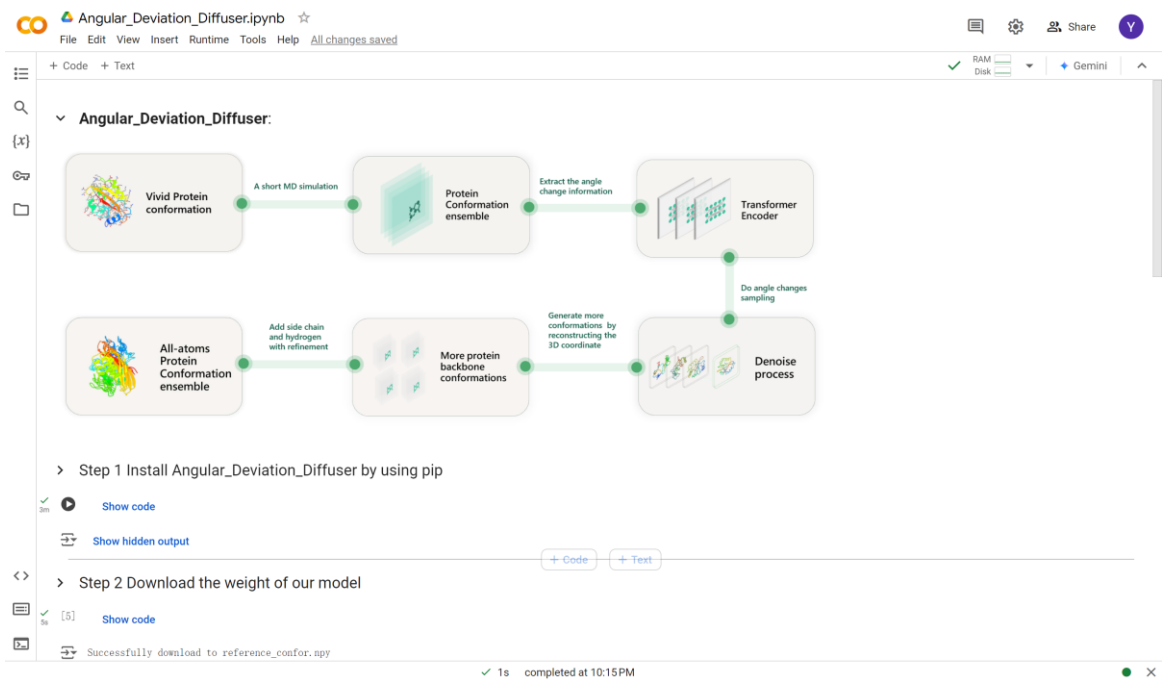
pip install Angular-Deviation-Diffuser

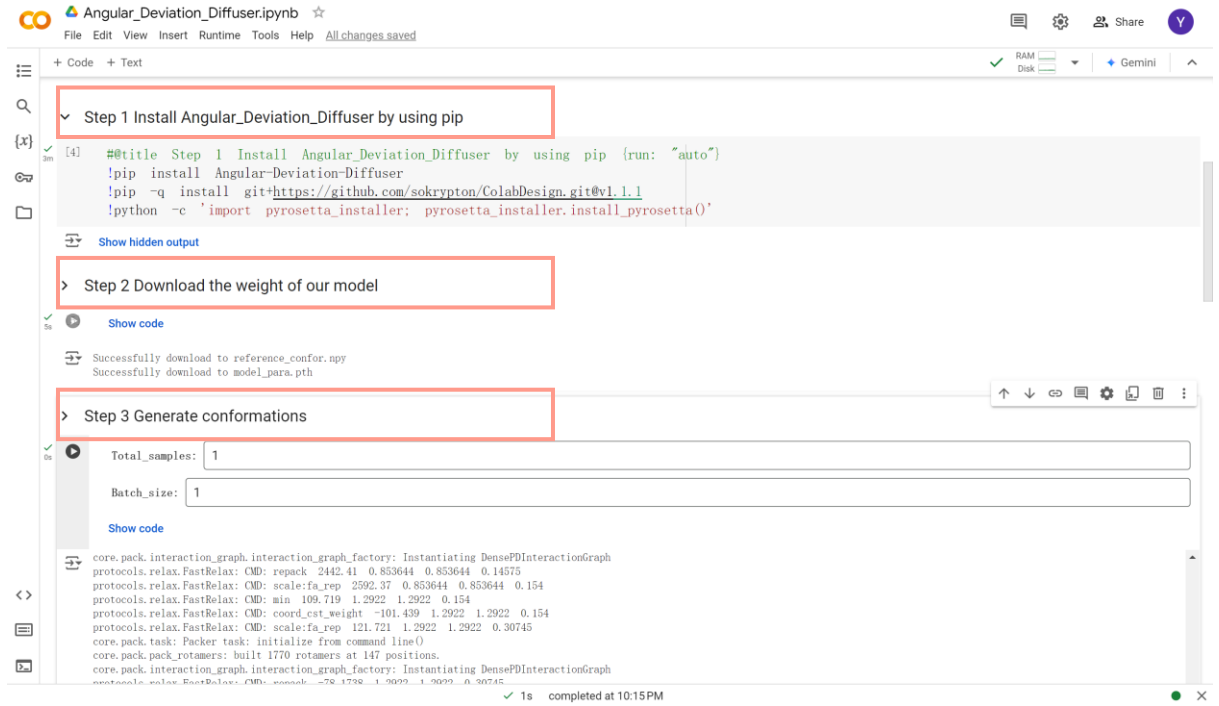


https://pypi.org/project/Angular-Deviation-Diffuser/

Part III

Angular-Deviation-Diffuser

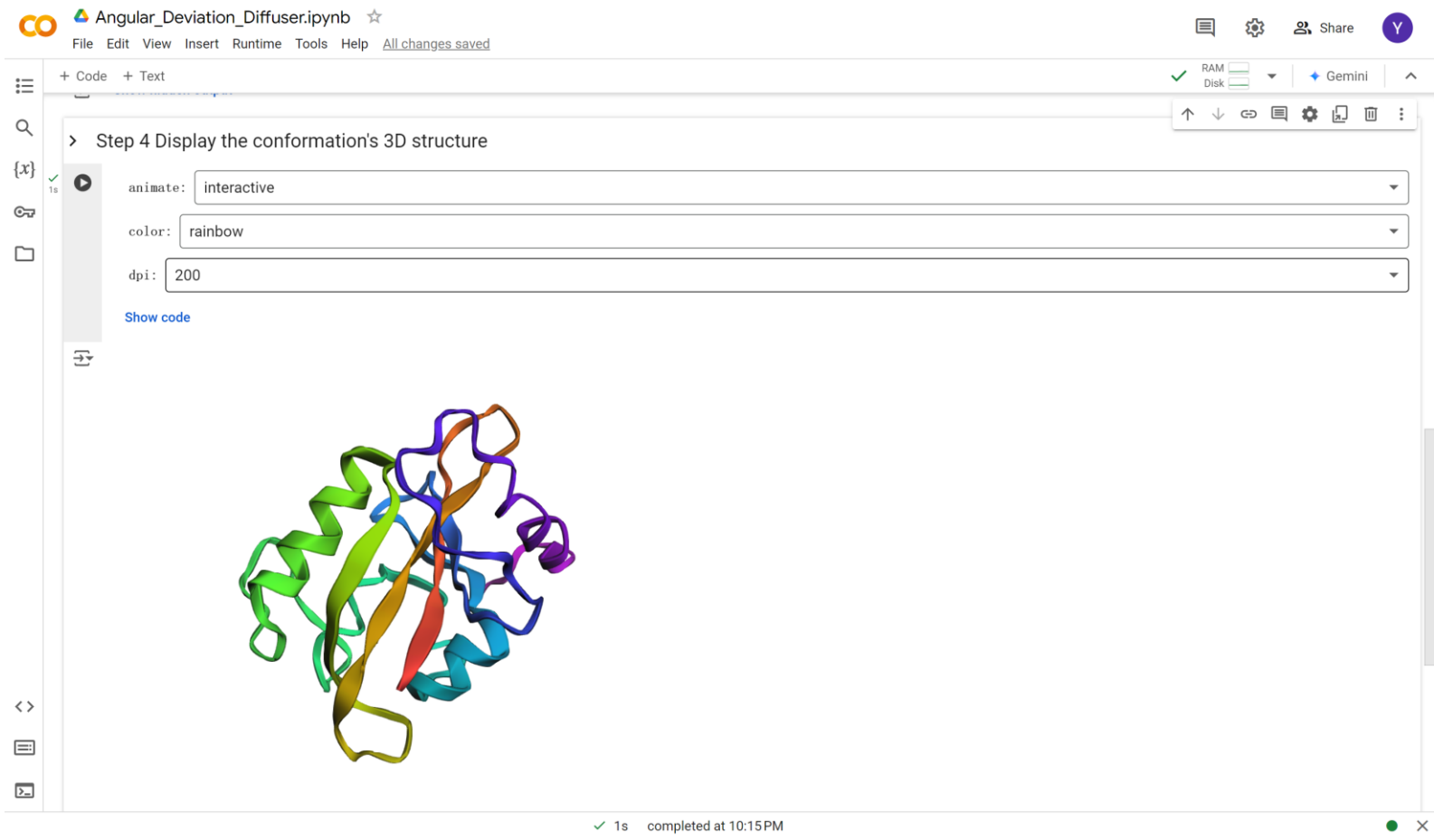




https://colab.research.google.com/drive/1paTyFVRMzD4b75DeFjYyXlMV38d1eE1I?usp=sharing

Part III

Angular-Deviation-Diffuser

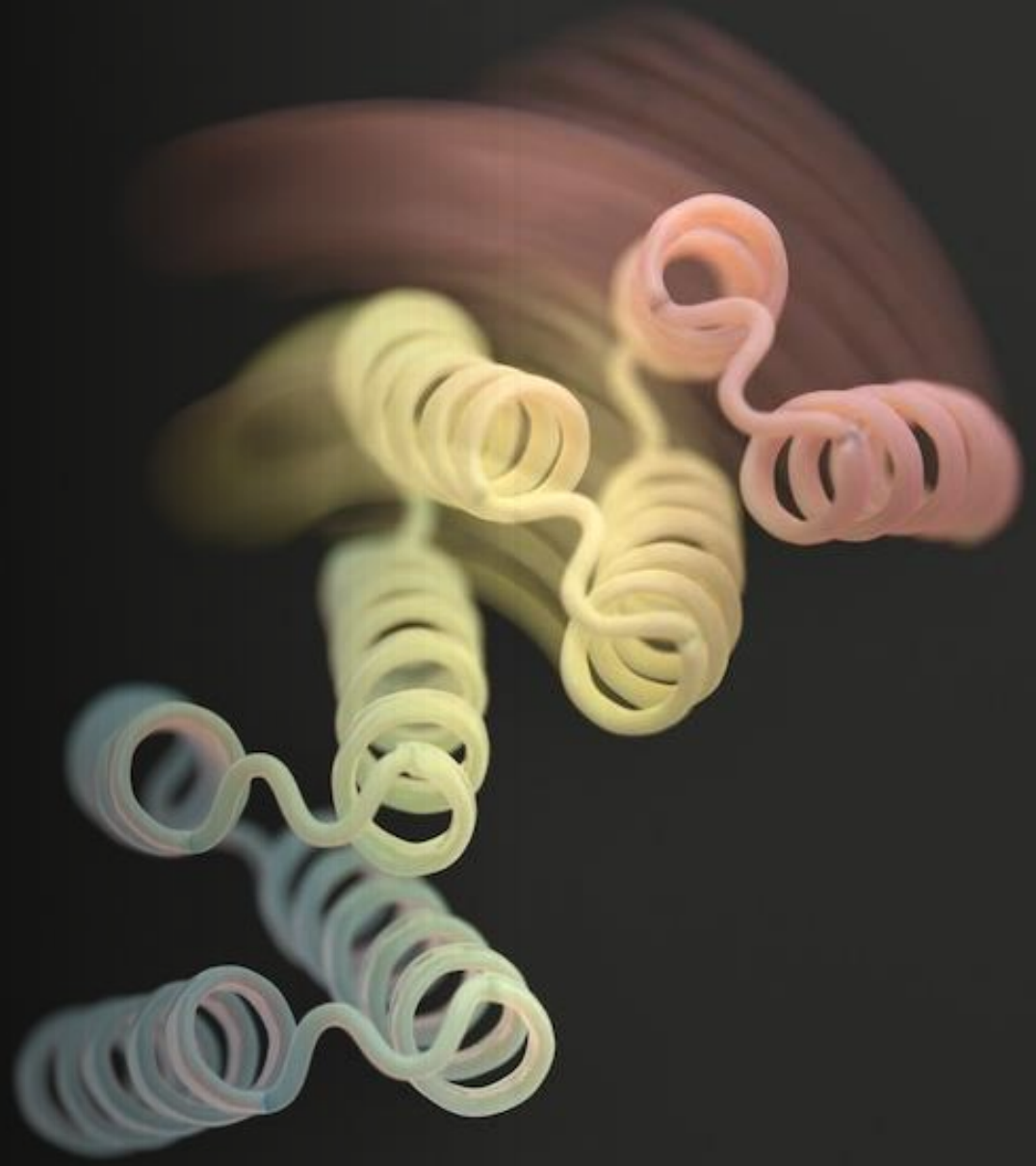


https://colab.research.google.com/drive/1paTyFVRMzD4b75DeFjYyXlMV38d1eE1I?usp=sharing

Part IV

CONCLUSION

- A transformer-based diffusion model that employs angular deviations to generate protein backbone structures with greater accuracy and efficiency.
- our model significantly outperforms traditional approaches that rely on absolute angle inputs, the use of angular deviations resulted in a more stable and reliable folding process.
- The computational experiments demonstrated that Our model, particularly when refined, has a better performance to Alpha Flow in reproducing VVD protein conformations ,also align closely with MD-generated structures.



SMU

Acknowledgements

- To all group members in Tao Group!
- To my friends and family 👨‍👩‍👦
- To HPC in SMU!
- Thanks for attending my Defense!

Final

Thank you for your attention!

Questions and suggestions?

SMU