Presentation | 2022-12-13

Accelerated Diffusion Models for Protein Structure Generation

Haowei Lin, Shenshen Li, Yuzhe Wang
Peking University
{linhaowei, lssastar, wangyuzhe_ccme}@pku.edu.cn

Outline

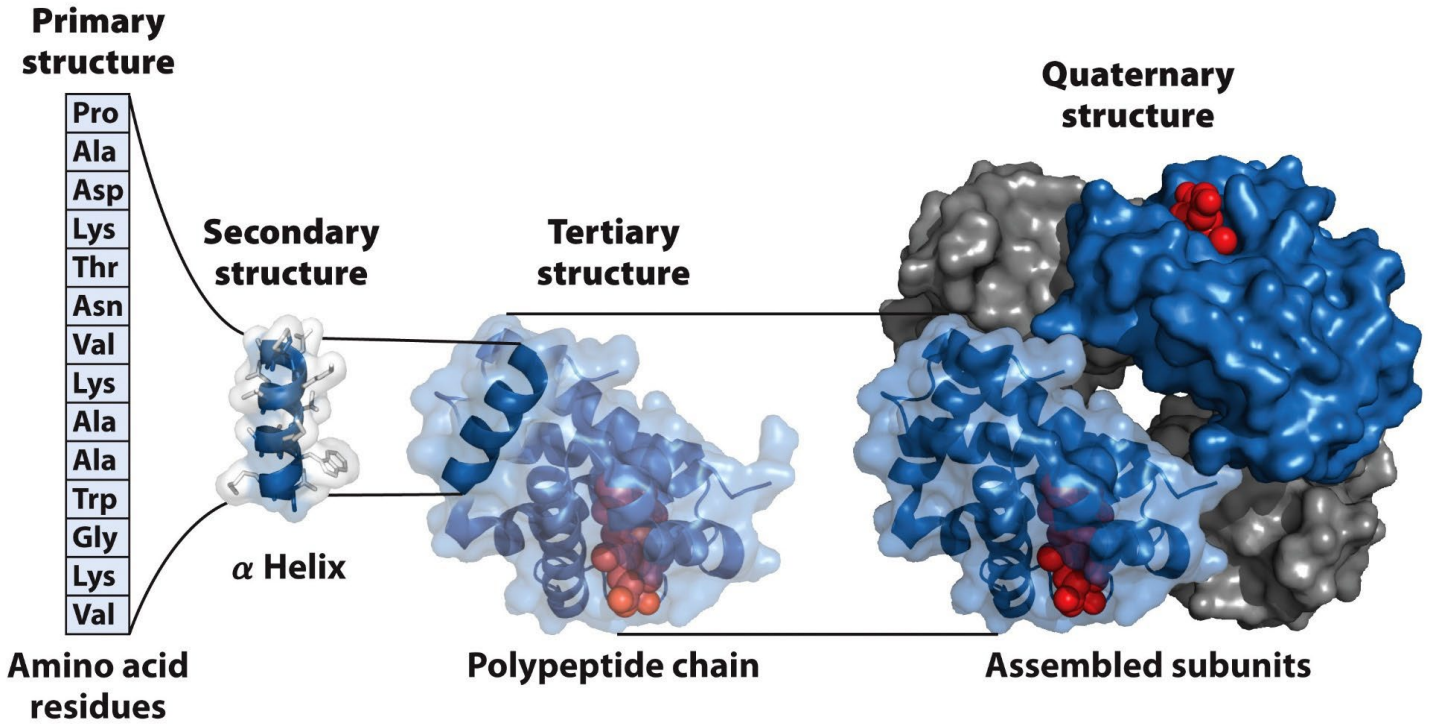
- Motivation
  - *De novo* Protein Design
  - Diffusion-based Protein Structure Generation
  - Accelerated Sampling for Diffusion Models
- Method
  - Evaluation Metric of Sample Quality
  - Representation of Protein Structures
  - Algorithm Sketch
- Experiments and Conclusion
  - Results and Analysis
  - Comparison of Various Acceleration Methods
  - Conclusion

Outline

- Motivation
  - *De novo* Protein Design
  - Diffusion-based Protein Structure Generation
  - Accelerated Sampling for Diffusion Models
- Method
  - Evaluation Metric of Sample Quality
  - Representation of Protein Structures
  - Algorithm Sketch
- Experiments and Conclusion
  - Results and Analysis
  - Comparison of Various Acceleration Methods
  - Conclusion

De novo Protein Design

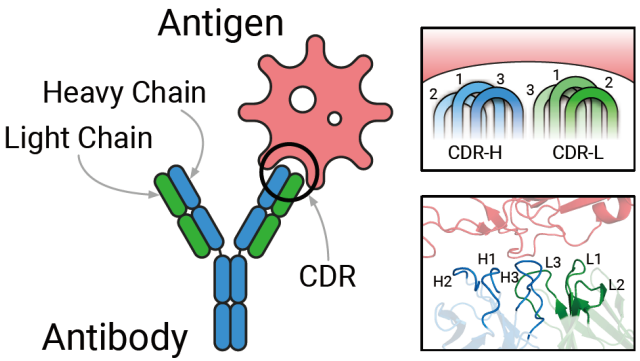
- Proteins: workhorse molecules of life
  - Protein functions: antibodies, enzymes, messengers…
  - **4 levels** of structure in proteins

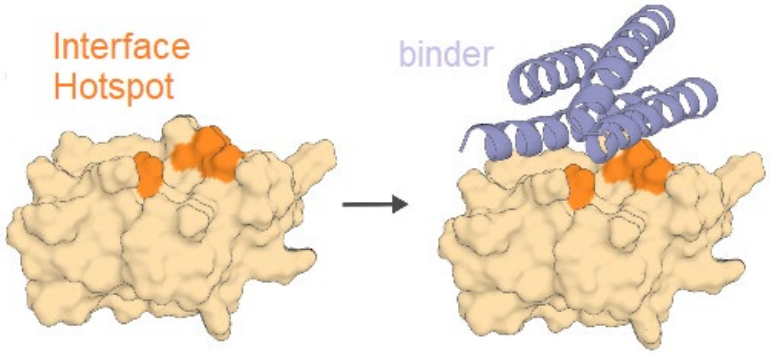


Nelson, D. L., Lehninger, A. L., & Cox, M. M. (2008). *Lehninger principles of biochemistry*. Macmillan.

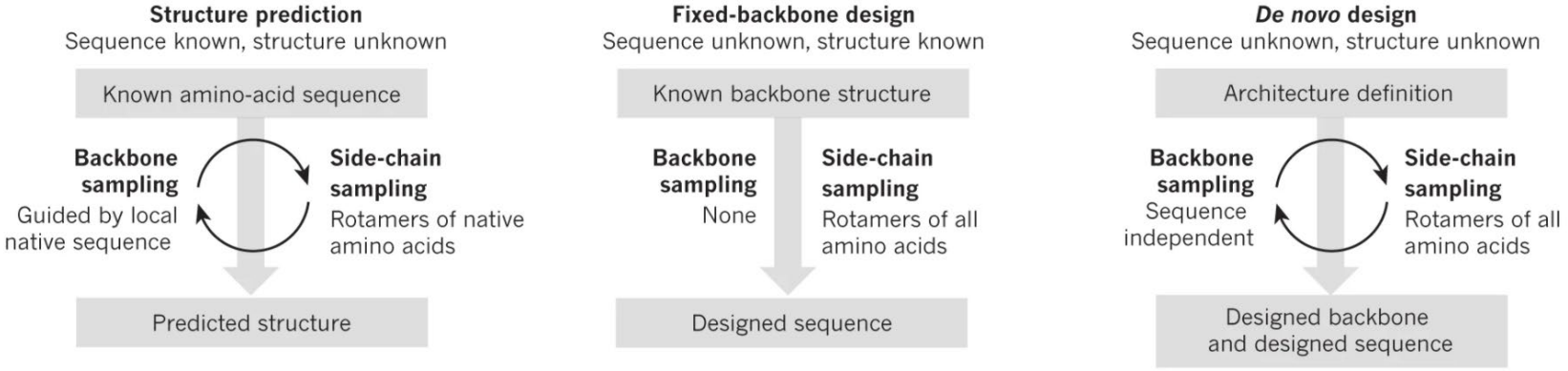
De novo Protein Design

- Protein structure generation
  - Generating novel protein structures for **antibody design** and **drug discovery**





  - Challenge of computational *de novo* protein design
    - ☑ AlphaFold2
    - ☑ Rosetta fixbb

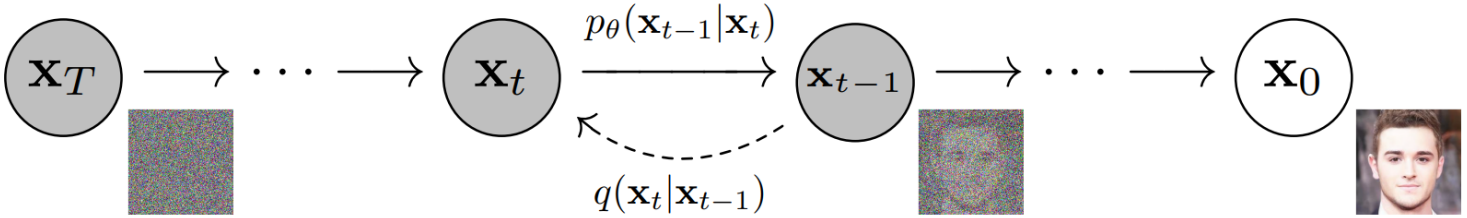


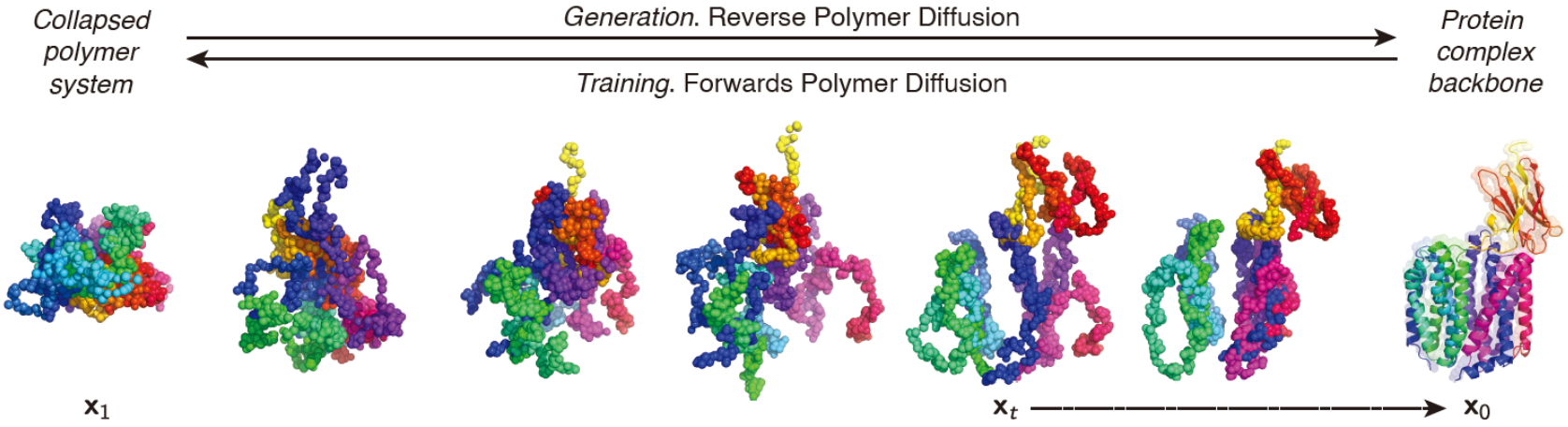
Luo, S., Su, Y., Peng, X., Wang, S., Peng, J., & Ma, J. Antigen-Specific Antibody Design and Optimization with Diffusion-Based Generative Models for Protein Structures. In *Advances in Neural Information Processing Systems*.

Huang, P. S., Boyken, S. E., & Baker, D. (2016). The coming of age of de novo protein design. *Nature*, *537*(7620), 320-327.

Diffusion-based Protein Structure Generation

- Generative diffusion of protein structures



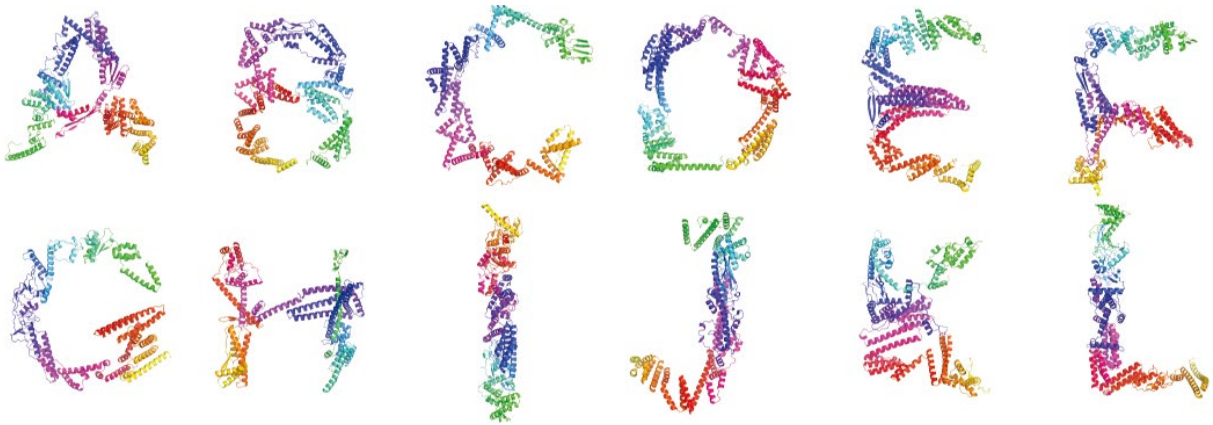


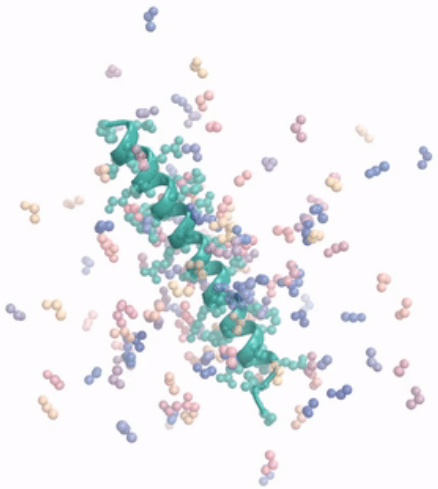
Ingraham, J., Baranov, M., Costello, Z., Frappier, V., Ismail, A., Tie, S., ... & Grigoryan, G. (2022). Illuminating protein space with a programmable generative model. *bioRxiv*.

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems, 33*, 6840-6851.

Diffusion-based Protein Structure Generation

- Generative diffusion of protein structures





- Existing problem: **slow inference**
- Project aim: apply feasible **accelerated sampling** methods to **protein structure generation**
  - Improved efficiency of antibody design and drug discovery
  - Provide convenience for model refinement

Ingraham, J., Baranov, M., Costello, Z., Frappier, V., Ismail, A., Tie, S., ... & Grigoryan, G. (2022). Illuminating protein space with a programmable generative model. *bioRxiv*.
https://www.bakerlab.org/2022/11/30/diffusion-model-for-protein-design/

Accelerated Sampling for Diffusion Models

- Accelerated sampling methods
  - Training-based: knowledge distillation, self-adaptive noise scheduling, Sample trajectory learning
  - Training-free: **DDIM**, Dynamic stepsize SDE solver, Analytic-DPM, **DPM-Solver**, **DPM-Solver++** (both used in Stable-Diffusion)
- Algorithm sketch of DPM-Solver/DPM-Solver++
  - Essence: fast ODE solver for the **probability flow ODE**

$$\frac{\mathrm{d}\boldsymbol{x}_t}{\mathrm{d}t} = f(t)\boldsymbol{x}_t - \frac{1}{2}g^2(t)\nabla_{\boldsymbol{x}} \log q_t(\boldsymbol{x}_t), \quad \boldsymbol{x}_T \sim q_T(\boldsymbol{x}_T)$$

  or equivalent the **diffusion ODE**

$$\frac{\mathrm{d}\boldsymbol{x}_t}{\mathrm{d}t} = \boldsymbol{h}_\theta(\boldsymbol{x}_t, t) := f(t)\boldsymbol{x}_t + \frac{g^2(t)}{2\sigma_t}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t), \quad \boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \tilde{\sigma}^2\boldsymbol{I})$$

  in which

$$f(t) = \frac{\mathrm{d}\log\alpha_t}{\mathrm{d}t}, \quad g^2(t) = \frac{\mathrm{d}\sigma_t^2}{\mathrm{d}t} - 2\frac{\mathrm{d}\log\alpha_t}{\mathrm{d}t}\sigma_t^2$$

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. (2020). Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456.*

Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., & Zhu, J. (2022). DPM-Solver: A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps. *arXiv preprint arXiv:2206.00927.*

Accelerated Sampling for Diffusion Models

- Accelerated sampling methods
- Algorithm sketch of DPM-Solver/DPM-Solver++
  - Decouple linear parts and nonlinear parts of the diffusion ODE by **variation of constants** formula to formulate the exact solution

$$\boldsymbol{x}_t = e^{\int_t^s f(\tau)d\tau}\boldsymbol{x}_s + \int_s^t \left(e^{\int_t^s f(\tau)d\tau}\frac{g^2(\tau)}{2\sigma_\tau}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_\tau,\tau)\right)\mathrm{d}\tau$$

  - Rewrite the solution by **change-of-variable** for $\lambda_t := \log(\alpha_t/\sigma_t)$

$$\boldsymbol{x}_t = \frac{\alpha_t}{\alpha_s}\boldsymbol{x}_s - \alpha_t\int_{\lambda_s}^{\lambda_t} e^{-\lambda}\hat{\boldsymbol{\epsilon}}_\theta(\hat{\boldsymbol{x}}_\lambda,\lambda)\mathrm{d}\lambda$$

  - Apply (*k*-1)-th order Taylor expansion to the NN term and compute the integral **analytically** using integration-by-parts to obtain **DPM-Solver-*k***

$$\boldsymbol{x}_{t_{i-1}\to t_i} = \frac{\alpha_{t_i}}{\alpha_{t_{i-1}}}\tilde{\boldsymbol{x}}_{t_{i-1}} - \alpha_{t_i}\sum_{n=0}^{k-1}\hat{\boldsymbol{\epsilon}}_\theta^{(n)}(\hat{\boldsymbol{x}}_{\lambda_{t_{i-1}}},\lambda_{t_{i-1}})\int_{\lambda_{t_{i-1}}}^{\lambda_{t_i}} e^{-\lambda}\frac{(\lambda-\lambda_{t_{i-1}})^n}{n!}\mathrm{d}\lambda + \mathcal{O}\big((\lambda_{t_i}-\lambda_{t_{i-1}})^{k+1}\big)$$

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. (2020). Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456.*

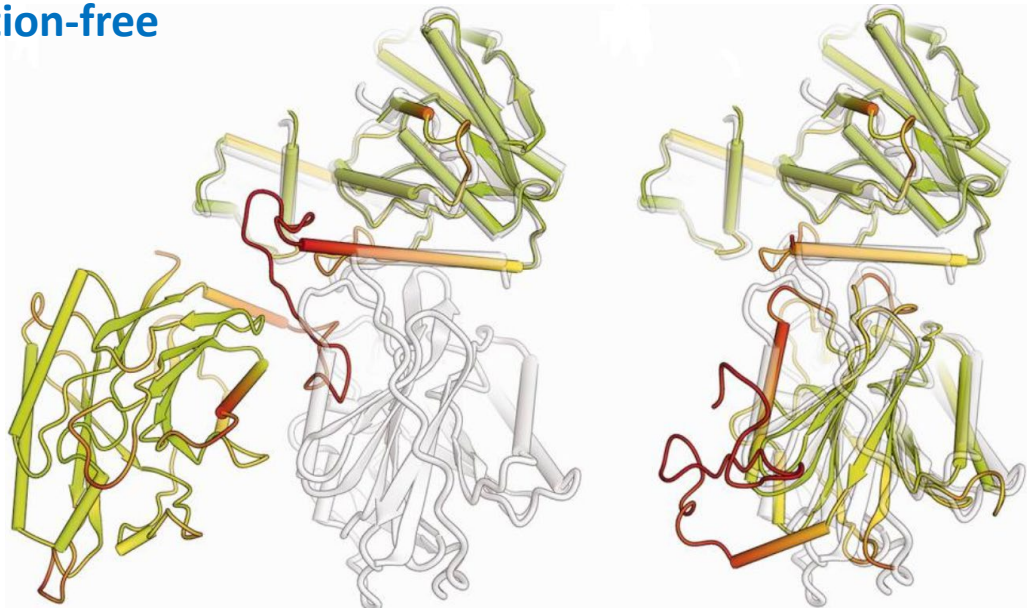
Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., & Zhu, J. (2022). DPM-Solver: A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps. *arXiv preprint arXiv:2206.00927.*

Outline

- Motivation
  - *De novo* Protein Design
  - Diffusion-based Protein Structure Generation
  - Accelerated Sampling for Diffusion Models
- Method
  - Evaluation Metric of Sample Quality
  - Representation of Protein Structures
  - Algorithm Sketch
- Experiments and Conclusion
  - Results and Analysis
  - Comparison of Various Acceleration Methods
  - Conclusion

Evaluation Metric of Sample Quality

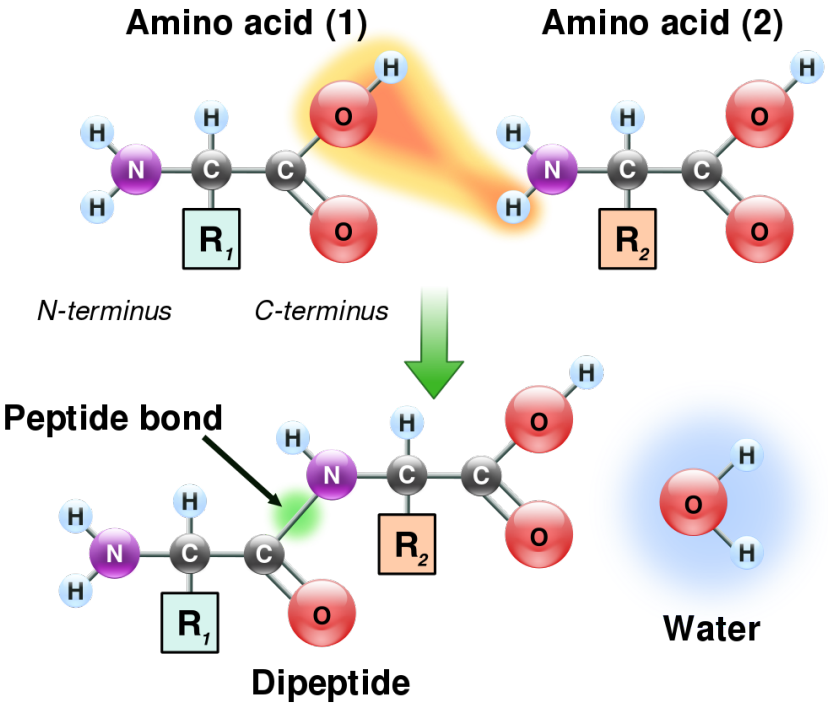
- Sample quality assessment
  - Image/text generation: perplexity, FID, inception score etc.
  - Quality assessment of generated protein structures? → structure generation of **given amino acid sequence** (whose **ground-truth** structure is known)
- Evaluation metrics: **lDDT** (**l**ocal **D**istance **D**ifference **T**est)
  - lDDT measures how well the environment in a reference structure is **reproduced** in a protein model (fraction of reproduced atom pairs, ↑)
  - **Superposition-free**

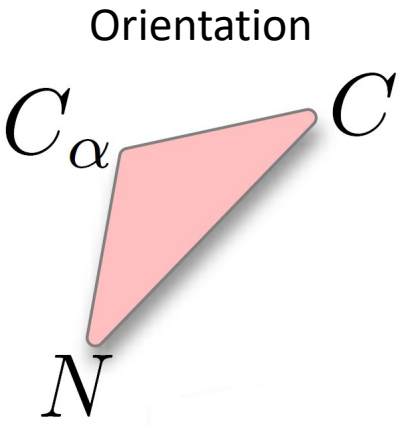


Mariani, V., Biasini, M., Barbato, A., & Schwede, T. (2013). lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 29(21), 2722-2728.

Representation of Protein Structures

- Representation of an amino acid in a protein complex
  - **Amino acid type**: $s_i \in \{\text{ACDEFGHIKLMNPQRSTVWY}\}$
  - **Cα atom coordinate**: $\boldsymbol{x}_i \in \mathbb{R}^3$
  - **Orientation**: $\boldsymbol{O}_i \in \mathrm{SO}(3)$





Luo, S., Su, Y., Peng, X., Wang, S., Peng, J., & Ma, J. Antigen-Specific Antibody Design and Optimization with Diffusion-Based Generative Models for Protein Structures. In *Advances in Neural Information Processing Systems*.
https://en.wikipedia.org/wiki/Amino_acid

Algorithm Sketch

- Diffusion for Cα coordinates
  - Forward diffusion

$$q(\boldsymbol{x}_j^t|\boldsymbol{x}_j^{t-1}) = \mathcal{N}(\boldsymbol{x}_j^t|\sqrt{1-\beta^t}\cdot\boldsymbol{x}_j^{t-1}, \beta^t\boldsymbol{I})$$

$$q(\boldsymbol{x}_j^t|\boldsymbol{x}_j^0) = \mathcal{N}(\boldsymbol{x}_j^t|\sqrt{\bar{\alpha}^0}\cdot\boldsymbol{x}_j^0, (1-\bar{\alpha}^0)\boldsymbol{I})$$

  - Generative process

$$p(\boldsymbol{x}_j^{t-1}|\boldsymbol{x}^t) = \mathcal{N}(\boldsymbol{x}_j^{t-1}|\boldsymbol{\mu}(\boldsymbol{x}^t, t), \beta^t\boldsymbol{I})$$

$$\boldsymbol{\mu}(\boldsymbol{x}^t, t) = \frac{1}{\sqrt{\alpha^t}}\left(x_j^t - \frac{\beta^t}{\sqrt{1-\bar{\alpha}^t}}G(\boldsymbol{x}^t, t)[j]\right)$$

  - Objective function

$$L^t = \mathbb{E}\left(\frac{1}{m}\sum_j ||\epsilon_j - G(\boldsymbol{x}^t, t)||^2\right)$$

Luo, S., Su, Y., Peng, X., Wang, S., Peng, J., & Ma, J. Antigen-Specific Antibody Design and Optimization with Diffusion-Based Generative Models for Protein Structures. In *Advances in Neural Information Processing Systems*

Algorithm Sketch

- Diffusion for amino acid orientations
  - Forward diffusion

$$q(\boldsymbol{O}_j^t|\boldsymbol{O}_j^0) = \mathcal{IG}_{\mathrm{SO}(3)}\left(\boldsymbol{O}_j^t|\mathrm{ScaleRot}(\sqrt{\bar{\alpha}^t}, \boldsymbol{O}_j^0), 1-\bar{\alpha}^t\right)$$

    - $\mathcal{IG}_{\mathrm{SO}(3)}$ denotes the isotropic Gaussian distribution on SO(3)

  - Generative process

$$p(\boldsymbol{O}_j^{t-1}|\boldsymbol{O}^t) = \mathcal{IG}_{\mathrm{SO}(3)}\left(\boldsymbol{O}_j^{t-1}|H(\boldsymbol{O}^t, t)[j], \beta^t\right)$$

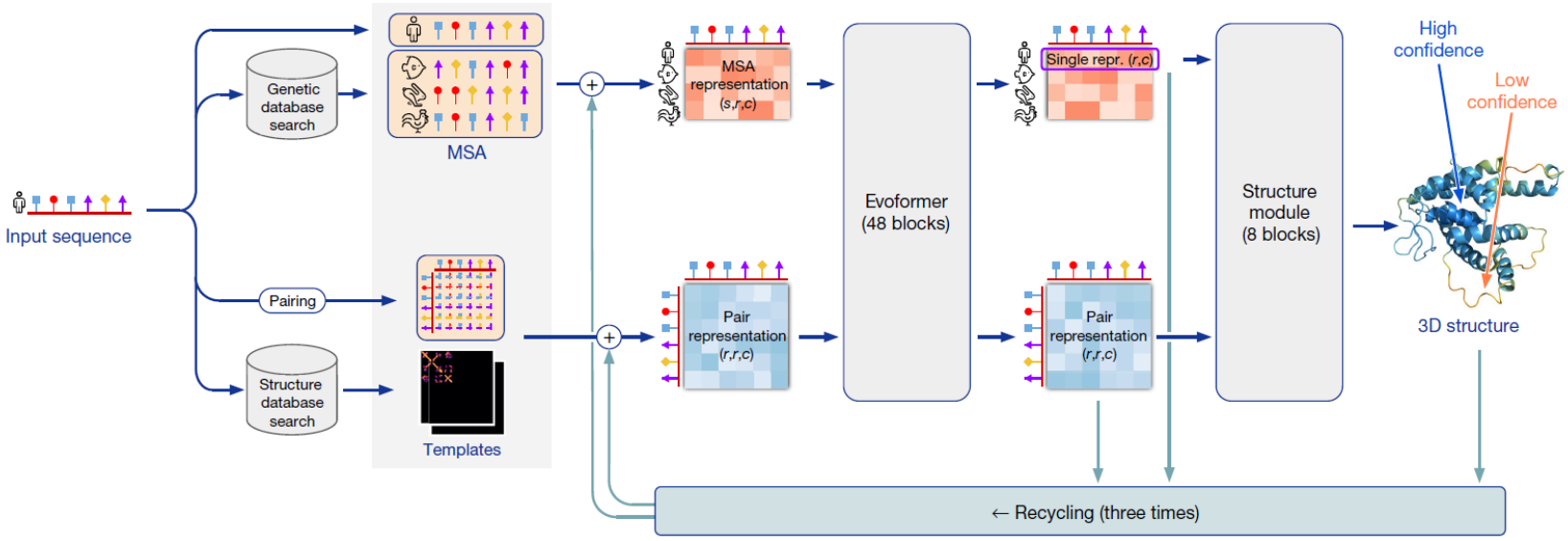
  - Objective function

$$L^t = \mathbb{E}\left(\frac{1}{m}\sum_j ||(\boldsymbol{O}_j^0)^T H(\boldsymbol{O}_j^t, t)[j] - \boldsymbol{I}||^2\right)$$

Luo, S., Su, Y., Peng, X., Wang, S., Peng, J., & Ma, J. Antigen-Specific Antibody Design and Optimization with Diffusion-Based Generative Models for Protein Structures. In *Advances in Neural Information Processing Systems*

Algorithm Sketch

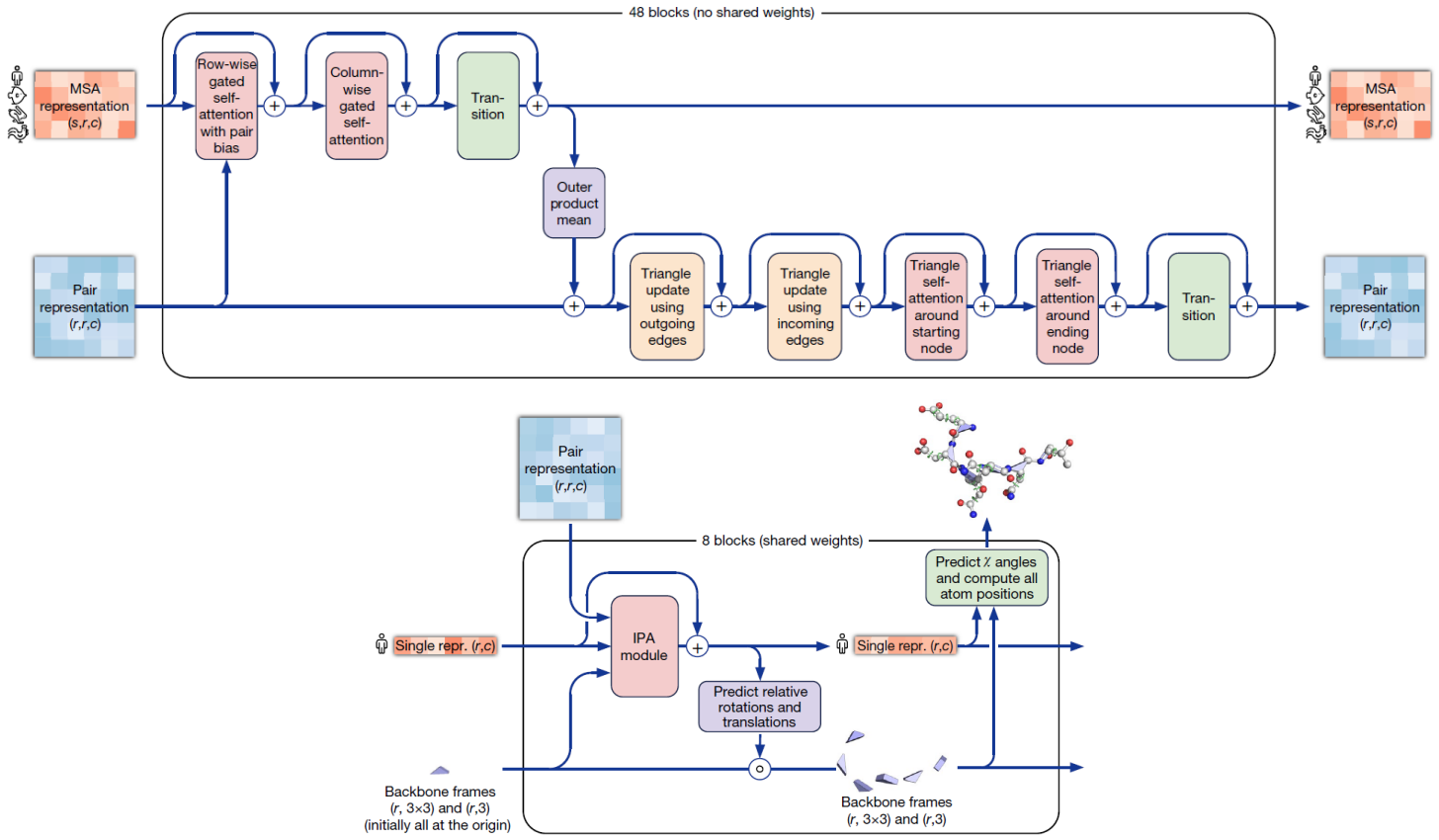
- Sequence condition
  - Latent feature of protein sequences: pretrained **AlphaFold2 Evoformer**
  - Model architecture backbone: pretrained **AlphaFold2 structure module**



Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583-589.

Algorithm Sketch

- Sequence condition
  - Latent feature of protein sequences: pretrained **AlphaFold2 Evoformer**
  - Model architecture backbone: pretrained **AlphaFold2 structure module**



Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583-589.

Outline

- Motivation
  - *De novo* Protein Design
  - Diffusion-based Protein Structure Generation
  - Accelerated Sampling for Diffusion Models
- Method
  - Evaluation Metric of Sample Quality
  - Representation of Protein Structures
  - Algorithm Sketch
- Experiments and Conclusion
  - Results and Analysis
  - Comparison of Various Acceleration Methods
  - Conclusion

Results and Analysis

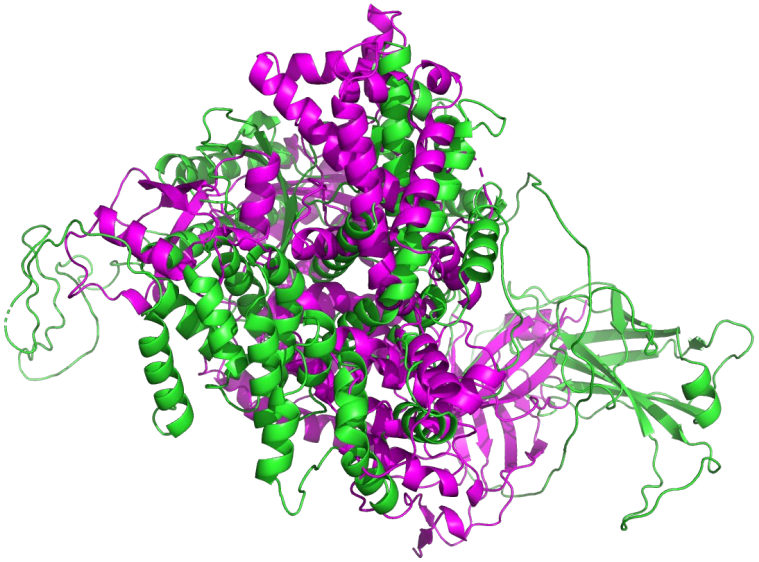
- Baseline performances of the vanilla diffusion model
  - Test set: **CAMEO dataset** which consists 146 of the most recent single-chain proteins
  - Case study: protein **PDB_ID = 7BI4_A** (on which all accelerated sampling algorithms are tested)

vanilla	1000 NFE	500 NFE	100 NFE	50 NFE	25 NFE	1 NFE
lddt(average)	0.824	0.820	0.818	0.8	0.74	0.30
lddt(7BI4_A)	0.70	0.685	0.676	0.65	0.45	0.18
time(CAMEO)	42331	21888	5356	3279	2321	1273
time(7BI4_A)	1257.3	648.2	158.7	96.9	69.8	40.1

The test results for the vanilla diffusion model

Results and Analysis

- The 3D structure of **Real** and predicted (**50 NFE** / **1000 NFE**) 7Bl4_A

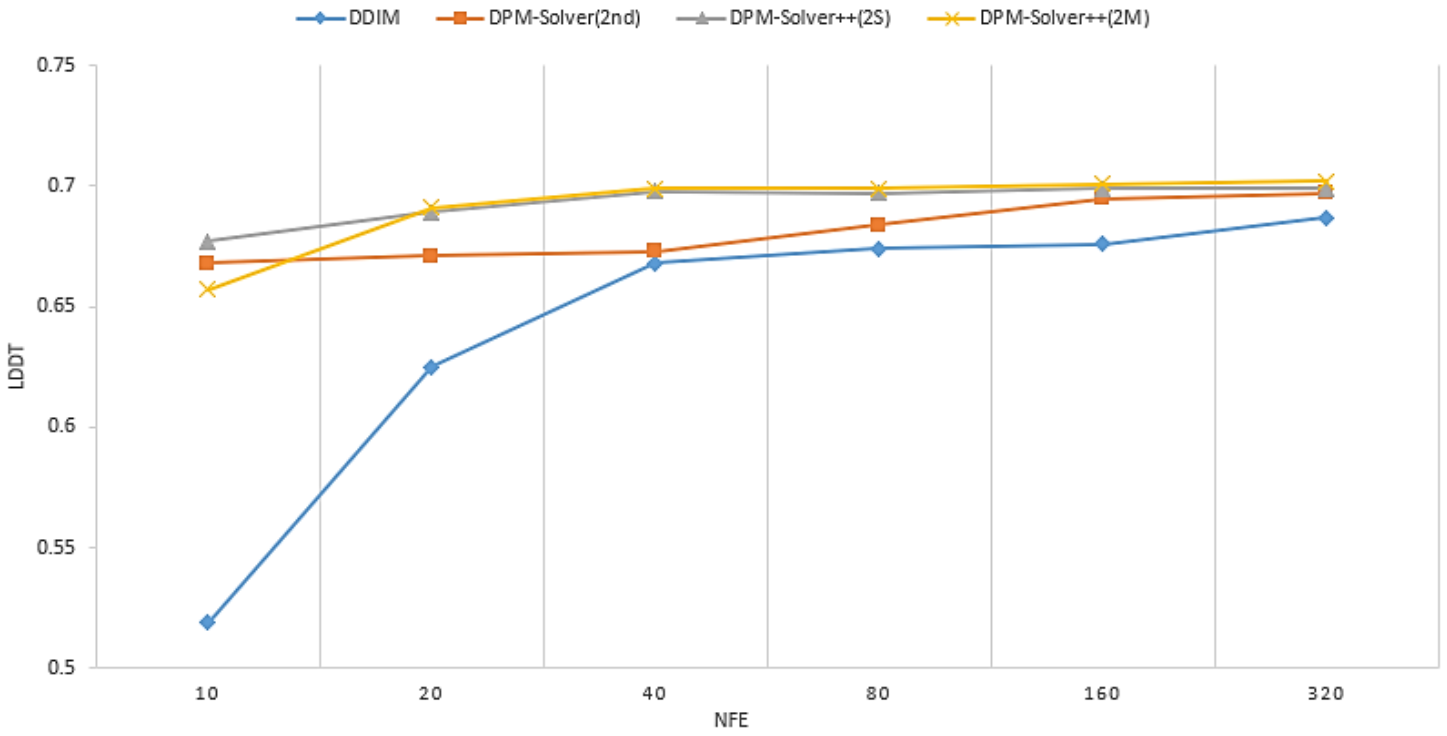


50 NFE, lddt=0.65, time=96.9 s

1000 NFE, lddt=0.70, time=1257.3 s

Comparison of Various Acceleration Methods

- The results of 7Bl4_A
  - **DPM-Solver++** converges at about 40 NFE and outperforms all other algorithms
  - DPM-Solver performs slightly worse than DPM-Solver++ but beats DDIM



Conclusion

- For 7Bl4_A:
  - The acceleration methods significantly improve the convergence speed of the model by a large margin **(1000 NFE → 40 NFE, 1200 s → 60 s)**
  - **DPM-Solver++** performs best when applied to diffusion-based protein structure generation
  - DDIM performs the worst (no convergence at 320 NFE)

Thank you

Haowei Lin, Shenshen Li, Yuzhe Wang
Peking University
{linhaowei, lssastar, wangyuzhe_ccme}@pku.edu.cn