FlowBack: A Flow-matching Approach for Generative Backmapping of Biomolecules

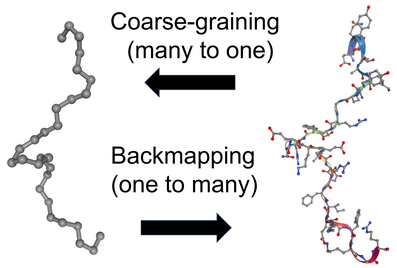
THE UNIVERSITY OF CHICAGO
PRITZKER SCHOOL OF MOLECULAR ENGINEERING

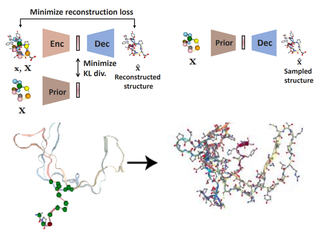
Mike S. Jones[1], Smayan Khanna[1], and Andrew L. Ferguson[1]

1 Pritzker School of Molecular Engineering, University of Chicago, Chicago, IL 60637, USA

Objective

- Generate an **ensemble** of all-atom (AA) structures consistent with a single coarse-grained (CG) trace
- Develop **generalized** approach applicable to various biomolecules and CG models
- Improve on and compare against existing methods:





GenZProt[1]
- VAE-based, internal coordinates
- Fast but limited diversity

DiAMoNDBack[2]
- Diffusion-based, autoregressive
- Good diversity but slow

Model Training

Sample

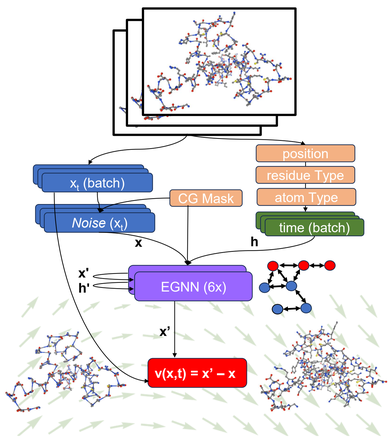
$$x_1 \sim q_1(x_1)$$
$$x_0 \sim q_0(x_0|x_1, \sigma_p, M)$$
$$t \sim \mathcal{U}(0,1)$$

Interpolate and noise

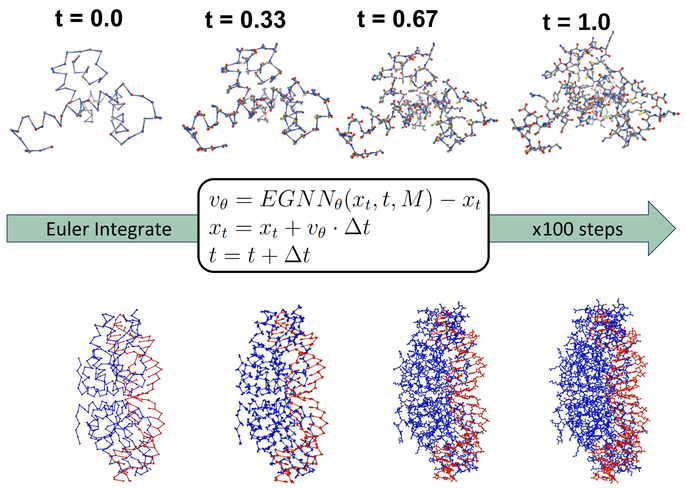
$$\mu_t \leftarrow t x_1 + (1-t)x_0$$
$$x_t \sim \mathcal{N}(\mu_t, \sigma_t^2 \mathbf{I})$$
$$x_t[M] \leftarrow \mu_t[M]$$

Regress against vector field

$$v_\theta \leftarrow EGNN_\theta(x_t, t, M) - x_t$$
$$u_t \leftarrow x_1 - x_0$$
$$\mathcal{L}_{\mathrm{CFM}} \leftarrow \|v_\theta - u_t\|^2$$



Model Inference



Training Data

- ~65k structures from PDB up to 1000 residues in length[5]
- 1577 protein-DNA complexes from DNAproDB database[6]
- mmSeqs clustering to build 50 sequence protein-DNA validation set

RCSB PDB PROTEIN DATA BANK

DNAproDB

- Proteins are coarse-grained such that only Cα atoms are retained
- DNA reduced to 3-site-per-nucleotide (3SPN) representation, beads places at sugar, base, and phosphate centers of mass

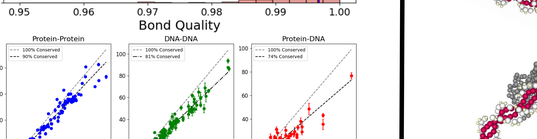
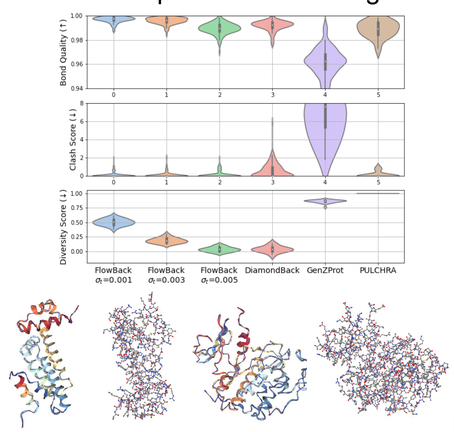


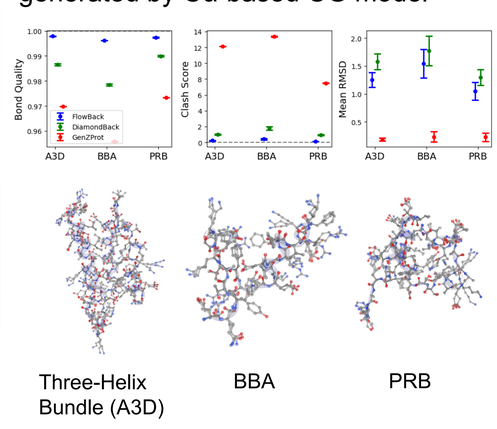
Image Adapted from[7]

Test Systems

24 protein structures from CASP structure prediction challenge[5]



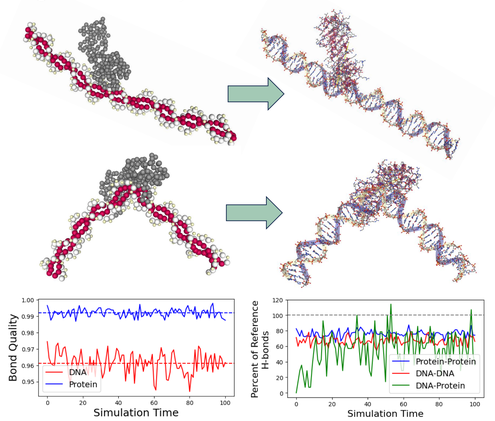
Fast-folding protein trajectories generated by Cα-based CG model[8]



Three-Helix Bundle (A3D) | BBA | PRB

50 Protein-DNA complexes from DNAproDB test set[6]

TATA-Binding protein-DNA trajectory generated by AICG + 3SPN.2 Model[9]

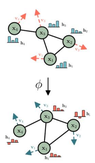


Flow-matching

Generative approach to learn an ordinary differential equation that transforms one distribution into another[3]

$$\mathscr{L} = \mathbb{E}_{t,q(z),p_t(x|z)}\left[\|v_\theta(t,x) - u_t(x|z)\|^2\right]$$

$$\mathscr{L} = \mathbb{E}_{t,q_0(x_0),q_1(x_1),p_t(x|x_0,x_1)}\left[\|v_\theta(t,x) - u_t(x|x_0,x_1)\|^2\right]$$

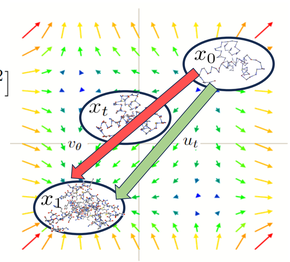


Equivariant Graph Neural Network[4]

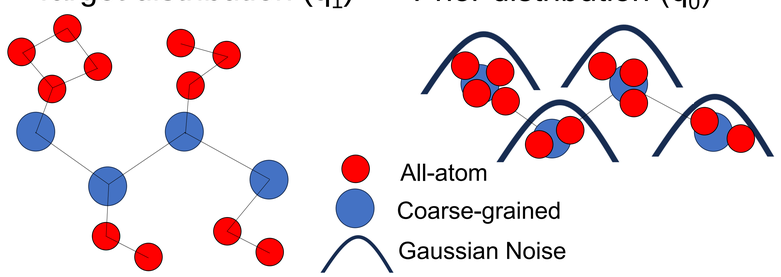
$$v_\theta(t,x) = EGNN_\theta(t,x) - x$$

Conditional vector field

$$u_t = x_1 - x_0$$

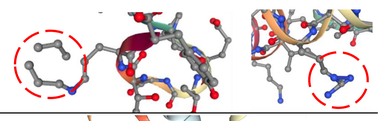


Target distribution (q_1) | Prior distribution (q_0)

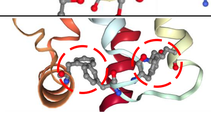


Evaluation Metrics

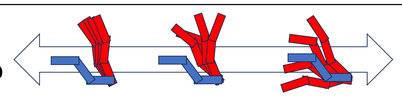
Bond Score ⬆
Percent of bonds within 10% of reference



Clash Score ⬇
Percent of residues within 1.2 Å of any other residue



Diversity Score ⬇
Similarity of reference to generate distribution



$$DIV = 1 - \frac{RMSD_{gen}}{RMSD_{ref}}$$

References

1. Yang, Soojung et al. Chemically transferable generative backmapping of coarse-grained proteins. *arXiv* preprint arXiv:2303.01569 (2023).
2. Jones et al. DiAMoNDBack: Diffusion-Denoising Autoregressive Model for Non-Deterministic Backmapping of Cα Protein Traces. *J. Chem. Theory Comput.* 2023, 19, 21, 7908–7923
3. Lipman et al. Flow matching for generative modeling. *arXiv* preprint arXiv:2210.02747 (2022)
4. Satorras et al. E (n) equivariant graph neural networks. *International conference on machine learning. PMLR*, 2021.
5. King et al. SidechainNet: An all-atom protein structure dataset for machine learning *Proteins* 2021, 89, 11, 1489-1496
6. Sagendorf et al. DNAproDB: an expanded database and web-based tool for structural analysis of DNA–protein complexes. *Nucleic Acids Research*, 2020, 48, D277–D287
7. Tan et al. Implementation of residue-level coarse-grained models in GENESIS for large-scale molecular dynamics simulations. *PLoS Comput Biol* 2022, 18(4): e1009578
8. Majewski et al. Machine learning coarse-grained potentials of protein thermodynamics. *Nat Commun* 2023, 14, 5739
9. Tan et al. Dynamic and Structural Modeling of the Specificity in Protein–DNA Interactions Guided by Binding Assay and Structure Data. *J. Chem. Theory Comput.* 2018, 14, 7, 3877–3889