Poster

DiffEMA: Diffusion-based generative network for protein model local quality assessment andauto-completion

H. Khachatryan[1], C. Wild[1,2], F. von Delft[1,2]

[1]University of Oxford - Oxford, UK, [2]Diamond Light Source - Didcot, UK

hamlet.khachatryan@lincoln.ox.ac.uk, c.wild@diamond.ac.uk

X-ray crystallography is the gold standard of protein structure determination. Diamond's XChem facility has enabled high throughput collection of crystal datasets, allowing the acquisition of over 1000 structures during a single screen. That advancement brought new opportunities in fragment-based drug design, allowing the creation of a fast-acting drug design platform for emergencies (*e.g.*, the COVID Moonshot project [1]).

The main hurdle for this approach is accurately interpreting partial occupancy electron density. Previously, Pearce et al. [2] published a statistical approach to automatically identify changed states by simultaneous analysis of electron density distributions across datasets

PanDDA: Pan-Dataset Density Analysis), allowing the identification of small molecule binding events and conformational changes

To automate the interpretation of the changed states revealed by PanDDA, we developed DiffEMA, a Diffusion-based model for Experimental Model assessment and Auto-completion. Inspired by the recent development of DiffDock [3], we repurposed a diffusion-based generative model for amino acid-wise quality assessment and auto-completion (amino acid's structure generation conditioned on electron density maps). As described in DiffDock, the diffusion model acts on a submanifold corresponding to the product space of the conformation transformation groups, i.e., translation, rotation, and torsional updates. Structures are represented as heterogeneous geometric graphs formed by amino acid atoms and electron density patch nodes (masked on 1.0 sigma level). Electron density patch nodes receive one-hot encoding of intensity sigma levels as initial features. Nodes are sparsely connected based on distance cutoffs that depend on the types of nodes being linked and on the diffusion time. During the initial limited experiments, DiffEMA placed 94% of its top-1 predictions within 1.5A, allowing autocompletion of local partial occupancies with high accuracy

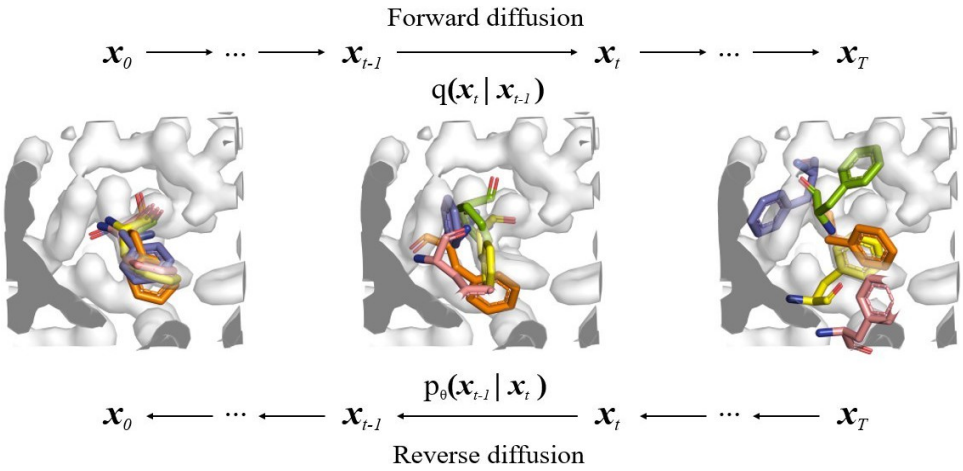

Figure 1. Overview of diffusion model in DiffEMA framework

[1] Boby, M.L., Fearon, D., Ferla, M., Filep, M., Koekemoer, L., Robinson, M.C., Chodera, J.D., Lee, A.A., London, N., Von Delft, A., Von Delft, F., et al (2023). *Science.* **382**, 6671.

[2] Pearce, N.M., Krojer, T., Bradley, A.R., Collins, P., Nowak, R.P., Talon, R., Marsden, B.D., Kelm, S., Shi, J., Deane, C.M., Von Delft, F., (2017). *Nature Communications*, **8**, 1.

[3] Corso, G., Stärk, H., Jing, B., Barzilay, R., Jaakkola, T., (2022). *arXiv,* URL https://arxiv.org/abs/2210.01776