NOS Report

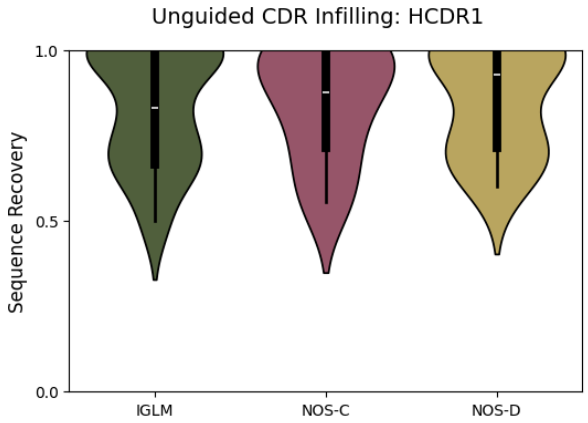
A common technique for protein design is to build generative models along with a discriminative model used for conditional sampling. The generative model samples plausible sequences while the discriminative model finds sequences with high fitness based on sampling conditionals. A challenge with generative protein models is that naive models produce outputs similar to their training data and thus are unlikely to provide valuable utility. There are two methods for protein generation: search in sequence space and search in structure space. The structure of a protein determines its function, however it is challenging and limiting to do inverse-folding because there is no guarantee that an amino acid sequence exists for a desired structure. Sequence models are faster and have access to larger datasets than structural models, thus making them more practical.

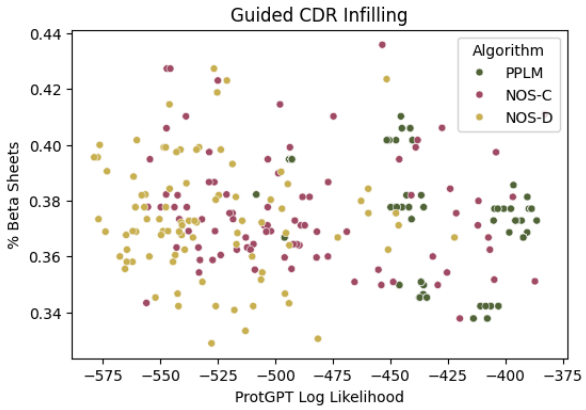
Diffusio**N** **O**ptimized **S**ampling (NOS) is a guidance method that enables discrete diffusion design directly in sequence space. It is a new method for controllable categorical diffusion that generates sequences with both high likelihood and desirable qualities by taking alternating steps between sequence corruption, guidance, and denoising in the continuous latent space of the model. For guided infilling tasks, NOS uses saliency maps to determine which positions on the sequence are most important to edit to improve the guidance objective value.

When working in sequence space, we can use categorical noise for the forward diffusion process, acting directly on the amino acid sequence. We can also use continuous noise instead by first embedding the data points with an embedding matrix, and then applying the forward diffusion process in the latent space with Gaussian noise. Thus, even though we are working in sequence space which is discrete, we can use either discrete noise (categorical) or continuous noise (Gaussian) to create the generative model. Discrete diffusion models with categorical denoising can achieve a continuous representation in the form of hidden states from an embedding matrix. Thus NOS can use gradients in both the discrete and continuous case by working in the hidden states. However, the steps taken in the hidden states of a discrete diffusion model are large between token embeddings while the steps are smooth in a continuous diffusion model.

NOS is called NOS-C for continuous diffusion and NOS-D for discrete diffusion.Training and inference was faster for NOS-D than NOS-C during experiments. The NOS methodology is made up of corrupted discriminative training that forms a modified denoising model. This model is then used to construct modified transition distributions for both the continuous and discrete case. Both NOS-C and NOS-D are trained with the bert-small transformer backbone. The modified transition distribution is sampled at each step with an update step that balances the objective with a KL term to ensure high likelihood. NOS is a form of iterative refinement which means that tokens across the entire sequence can be modified at each step because there are complex interactions between distant parts of a sequence that alter protein function.

We evaluate NOS on both guided and unguided CDR infilling for antibodies. Antibody design traditionally involves the infilling process of applying mutations to protein complementarity determining regions (CDRs). In the chart below on the left, we compare the distribution of infill sequence recovery rates during unguided infilling for NOS-C, NOS-D, and IGLM (an autoregressive antibody language model). We also train NOS-C and NOS-D for guided CDR infilling, optimizing for the beta sheets percentage of the generated proteins. In the chart below on the right, the guided models are sampled and compared to samples from PPLM, a guidance method for autoregressive models. We use hyperparameters that yield samples across the spectrum from prioritizing likelihood to prioritizing the objective. In this case, PPLM uses IGLM. The sampled infillings are also evaluated on their log likelihood, estimated with ProtGPT.





In the first chart, the NOS models are able to produce samples on-par with IGLM. This establishes that sequence diffusion models are an effective solution for protein sequence generation. In the second chart, the NOS models perform similarly to PPLM in maximizing beta sheets, though the samples tend to have a lower log likelihood. Future work should be done to verify if this difference in distributions is an issue with the algorithms or merely an artifact of hyperparameter selection. The chart shows however that NOS is able to trade off between likelihood and the objective.

It is important to highlight that the data in both graphs differ from the data presented in the original paper. The unguided distribution for NOS-C in the original paper performs noticeably better than both IGLM and NOS-D for HCDR1 while the data collected above does not show this. Also the samples of guided infilling in the original paper show an inverse relationship between log likelihood and beta sheets percentage while this is not clear in the samples shown above. Furthermore, the paper concludes that NOS offers an improvement over baselines but the above data is not able to support this. However, the reproduced data for this report confirms that NOS is an effective method for both unguided and guided protein infilling.