DISCRIMINATIVE PROTEIN SEQUENCE MODELLING WITH LATENT SPACE DIFFUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce a framework for protein sequence representation learning that decomposes the task between manifold learning and distributional modelling. Specifically we present a Latent Space Diffusion architecture which combines a protein sequence autoencoder with a denoising diffusion model operating on its latent space. We obtain a one-parameter family of learned representations from the diffusion model, in addition to the autoencoder's latent representation. To address the challenge of identifying an appropriate latent space for diffusion, we propose and evaluate two autoencoder architectures: a homogeneous model forcing amino acids of the same type to be identically distributed in the latent space, and an inhomogeneous model employing a noise-based variant of masking.

1 INTRODUCTION

Proteins are an important class of biomolecules whose function, interaction, and evolutionary relationships are central to understanding cellular mechanisms and the complexity of life. While the underlying principles governing proteins and their behaviour admit explicit formulations in quantum chemistry, these are in practice too complex to model directly. The quest for simplifying representations and approximations which strike a balance between generality, accuracy and computational efficiency is a core challenge of computational biology.

Machine learning provides a powerful suite of tools for representation learning. A prominent method is an autoencoder employing an information bottleneck to learn a compressed latent representation Tishby et al. (2000). For proteins this is less applicable however, as a protein's primary sequence already provides an incredibly compact representation (each amino acid, being a categorical variable of size 20, can be represented with just 5 bits). Indeed the primary sequence completely determines a protein, and the key challenge is how to decode from this. Applications of machine learning to protein representation learning from sequence data can be roughly organised around two main threads: those such as AlphaFold which leverage multiple sequence alignments (MSAs) capturing co-evolutionary information Jumper et al. (2021); Rao et al. (2021); Truong Jr & Bepler (2023), and those such as ESM Rives et al. (2021); Lin et al. (2023); Hayes et al. (2024) utilising masked language modelling (MLM) Devlin et al. (2019); Elnaggar et al. (2021); Brandes et al. (2022).

Generative modelling is closely tied to representation learning Kingma & Welling (2014); Goodfellow et al. (2014). Indeed masked language modelling is a form of reconstructive learning, where a model is trained to restore partially corrupted input, which underlies its ability to learn rich contextually aware representations Devlin et al. (2019). For continuous spaces, Gaussian diffusion has emerged as a leading generative method due to its ability to produce diverse high-quality samples from complex distributions Sohl-Dickstein et al. (2015); Ho et al. (2020); Song et al. (2020). By learning how Gaussian noise diffuses through a data space, a diffusion model learns to approximate the score function of the data distribution, $s(x) = \nabla_x \log p_{data}(x)$. From a statistical physics perspective, expressing the distribution in Boltzmann form $p_{data}(x) \sim \exp(-E(x))$, the score function admits a natural interpretation as a distributional force: $F(x) \equiv -\nabla_x E(x) = s(x)$, which underlies a diffusion model's ability to both navigate the distribution efficiently as well as to learn a meaningful representation of the data Song et al. (2020); Vincent et al. (2008).

Studies of generative modelling on protein sequence data have primarily focused on discrete diffusion methods Gruver et al. (2024); Alamdari et al. (2023); Wang et al. (2024). These are counterparts

to Gaussian diffusion that place emphasis on the categorical nature of amino acids. Unlike the continuous case however, there is not yet an established discrete diffusion method, in part due to the challenges that categorical variables lack a natural order, and absence of continuity which means that a single change can have an abrupt effect (e.g. a single-point mutation). A notable work however is DPLM Wang et al. (2024), whose discrete diffusion method is a generalised form of masked language modelling. Indeed the authors highlight the representation learning capabilities of their generative model and demonstrate that it achieves competitive performance with ESM.

Let us question however whether a masking-based approach is the best route towards modelling protein sequence data. From a reconstructive learning perspective, it is unclear whether masking is an optimal way to represent a corrupted sequence. For instance, the unmasked amino acids are fully specified with no ambiguity, while for the masked amino acids only their ambiguity is conveyed. One can imagine instead an alternative corruption process where partial information is retained/erased, for example expected physiochemical properties or aspects of long-range dependencies. Similarly from a generative perspective, one may question the task of performing distributional modelling directly in the discrete domain, as at the level of sequence the protein landscape is far from smooth: single mutations can have abrupt consequences while compound mutations may be strongly correlated.

This motivates us to explore a switch in focus from a discrete representation of sequence space to a continuous one. This can be framed as making a distinction between two aspects of protein representation learning: manifold learning and distribution learning. The first addresses the question of how to embed protein sequences in a continuous latent space, while the second concerns the distribution of protein sequences over this latent space. Here Gaussian diffusion can be employed for the distributional modelling, and so the question then is how to learn an appropriate latent space.

Previous works adopting a latent diffusion approach for protein sequence data examined the use of the ESM embeddings Chen et al. (2024); Meshchaninov et al. (2024). These had limited success however, which can be attributed to the embeddings retaining much of the discreteness of the underlying sequence Li et al. (2023). In essence, amino acid embeddings are too robust to added noise, which obstructs the learning ability of denoising. Indeed a parallel can be made here to high-resolution images, for which latent diffusion models were originally introduced Rombach et al. (2022).

In this work we attempt to address the challenge of how to construct a latent space which facilitates the distributional modelling of proteins sequence data. To this end we propose two novel sequence autoencoder architectures: a homogeneous model forcing amino acids of the same type to be identically distributed in the latent space, and an inhomogeneous model employing a noise-based variant of masking. We train a diffusion model on their latent space, and identify how this gives rise to an additional one-parameter family of learned representations. We focus on this discriminative capability of the diffusion model, and evaluate it on a diverse set of representatation learning benchmarks.

2 Related work

The use of diffusion/denoising for protein representation learning was introduced for the structural representation Zaidi et al. (2022); Liu et al. (2022), based on a connection between the learned score function and molecular force fields. DSMBind employs SE(3) denoising score matching as an unsupervised pre-training task for binding energy prediction Jin et al. (2023). From a generative perspective, diffusion has also been mostly applied to structure Watson et al. (2023); Ingraham et al. (2023); Yim et al. (2023); Lee et al. (2023), for which the Gaussian form of diffusion can be employed. On protein structure prediction, AlphaFold 3 model Abramson et al. (2024) trains a conditioned diffusion model for the generation of its structural predictions. Application to the generation of conformational ensembles has also been explored Jing et al. (2024); Hassan et al. (2024).

Discrete diffusion applied on protein sequence data has been explored in LaMBO-2 Gruver et al. (2024), EvoDiff Alamdari et al. (2023), and DPLM Wang et al. (2024). DPLM stands out for evaluating the representation learning capabilities of their model, demonstrating it's ability to perform competitively across a range of prediction tasks. Two studies explored latent diffusion on pre-trained ESM2 embeddings Chen et al. (2024); Meshchaninov et al. (2024).

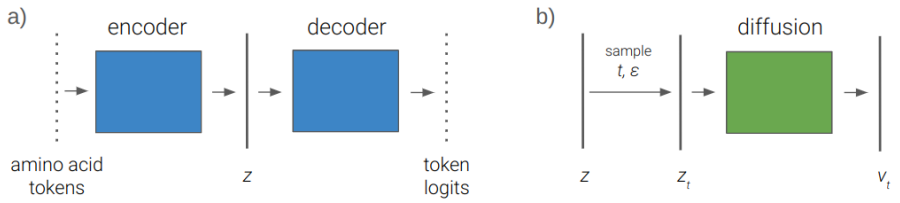


Figure 1: The LSD model is comprised of **(a)** a protein sequence autoencoder which learns a latent space z, and **(b)** a diffusion model acting on this latent space. The autoencoder is trained end-to-end by balancing a reconstruction loss, between input amino acid tokens and output token logits, against a normalization loss on the distribution of latent space embeddings. We consider two variants, LSD-TN with a non-trivial normalization loss and LSD-NM with a non-trivial reconstruction loss, as described in Sec. 3. The diffusion model learns to map noised latent embeddings $z_t = \cos(\pi t/2)z + \sin(\pi t/2)\varepsilon$ to their orthogonal complements $v_t = -\sin(\pi t/2)z + \cos(\pi t/2)\varepsilon$, which thereby provides an additional one-parameter family of sequence representation to that obtained at the latent space.

More broadly, the link between discriminative and generative modelling underlies the autoregressive approach Ferruz et al. (2022); Madani et al. (2023), the masked language modelling approach Rives et al. (2021); Lin et al. (2023); Hayes et al. (2024); Elnaggar et al. (2021); Brandes et al. (2022), as well as variational autoencoder approaches Sinai et al. (2017); Sevgen et al. (2023), to modelling protein sequence data.

Beyond the field of protein modelling, denoising autoencoders date back to the seminal work Vincent et al. (2008). Diffusion-based representation learning was advanced in Abstreiter et al. (2021). Latent diffusion models were introduced in the image domain Rombach et al. (2022), and the application of latent diffusion to discrete data has been predominantly studied in the natural language processing literature Li et al. (2022); Dieleman et al. (2022); Strudel et al. (2022); Gao et al. (2024); Ye et al. (2024); Gulrajani & Hashimoto (2024).

3 Latent Space Diffusion

We employ a Latent Space Diffusion (LSD) architecture, illustrated in Fig. 1, comprised of

- an autoencoder learning a latent manifold embedding of protein sequences,
- a diffusion model for the distributional modelling of protein sequence datasets over this learned latent space.

We adopt a transformer-based architecture for each component as shown in Appendix A. The autoencoder is composed of an encoder-decoder pair, which we set to have an equal number of layers. The encoder takes as input tokenized protein sequences, and outputs latent embeddings $z_{a,i}$ of amino acids, with a indexing amino acid position and i the coordinate of the embedding space. The decoder takes the latent embeddings $z_{a,i}$ as input, and outputs corresponding token logits. The diffusion model is a conditioned transformer and we describe its action on the latent space below in the description of the diffusion loss. A more detailed description of the architecture is provided in Appendix A.

We train the auto-encoder under competition between a reconstruction loss and a normalization loss on the latent embeddings. For the reconstruction loss we employ the standard cross-entropy between the input tokens and output token logits. The normalization loss is less straightforward. In contrast to a variational auto-encoder Kingma & Welling (2014) which maps input to distributions over the latent space for which it learns the mean and variance, we let the encoder map directly to the latent space and guide the distribution of $z_{a,i}$ over a to be normally distributed (this foregoes the generative capability of the autoencoder, which is compensated for here by the diffusion model). Specifically, given a batch of sequences we employ a univariate parametric form of the Kullback–Leibler diver-

gence

$$\mathcal{L}_N = \frac{1}{2d}\sum_{i=1}^{d}(\mu_i^2 + \sigma_i^2 - \log\sigma_i^2 - 1), \tag{1}$$

expressed in terms of the empirical mean μ_i and variance σ_i^2 of the $z_{a,i}$, with d the embedding dimension. We considered also a multivariate parametric form but found this to degrade performance, see ablation in Appendix C.

The simple combination of reconstruction loss and normalisation loss is not however sufficient to drive meaningful learning in the latent space. To achieve this, we consider two variants as follows:

- **Token Norm (LSD-TN):** here we modify the normalization loss by applying it separately to the embeddings of each amino acid type. Specifically, within each batch, we partition the latent embeddings into 20 sets, each corresponding to one of the 20 canonical amino acids, and employ a separate normalization loss for each set. While the default normalization loss allows different amino acid types to occupy distinct regions under the same normal distribution, this approach imposes a stricter constraint, creating an effective bottleneck for representation learning.
- **Noise Masking (LSD-NM):** here we modify the reconstruction loss to a variant of MLM designed for greater robustness to noise. Unlike standard MLM, where a fraction of amino acid embeddings are fully masked while the rest remain unaltered, our approach applies varying levels of corruption by inhomogeneously adding Gaussian noise to the latent embeddings. Specifically, we transform each amino acid embedding vector as

  $$z_a \to \cos(\pi t_a/2) z_a + \sin(\pi t_a/2)\varepsilon, \tag{2}$$

  where $t_a \in (0,1)$ controls the noise level and ε is an embedding vector sampled from $\mathcal{N}(0,1)$. To reflect this corruption in the reconstruction loss, we weight each embedding's contribution by the noise amplitude $\sin^2(\pi t_a/2)$, ensuring that highly corrupted embeddings dominate the training signal, while minimally corrupted ones contribute negligibly.

  We explored two sampling strategies for t_a: uniform sampling over (0,1) and sampling proportional to the signal amplitude $\cos^2(\pi t_a/2)$, as used for training the diffusion model (see below). The latter approach, which results in most amino acids being weakly noised while a few are strongly noised, performed better, and we adopt this choice in the models we present. See ablation in Appendix C.

The diffusion model is trained on the autoencoder's latent space. For this we employ a variance-preserving cosine noise schedule Nichol & Dhariwal (2021), the v-target objective Salimans & Ho (2022), and epsilon prediction loss Ho et al. (2020). Specifically, the latent embeddings z get (here uniformly) noised to

$$z_t = \cos(\pi t/2) z + \sin(\pi t/2)\varepsilon, \tag{3}$$

for $t \in (0,1)$ and $\varepsilon \in \mathcal{N}(0,1)$, and the diffusion model is trained to learn

$$v_t = -\sin(\pi t/2) z + \cos(\pi t/2)\varepsilon. \tag{4}$$

i.e. $\hat{v}_t = \texttt{Diffusion}(z_t, t)$. The epsilon prediction loss, expressed in terms of v, is weighted by the signal amplitude

$$\mathcal{L}_D = \frac{1}{2}\mathbb{E}_{t\sim(0,1),\,\varepsilon\sim\mathcal{N}(0,1)} \cos^2(\pi t/2)\|\hat{v}_t - v_t\|^2, \tag{5}$$

and we evaluate this with importance sampling. Indeed, from a representation learning perspective the increased weight for sampling t closer to 0 is intuitive, as information gets washed out with increasing t.

In this work we focus on the discriminative capability of the diffusion model. While an ultimate objective of the LSD construction is to develop also the generative capability, we take the perspective that the discriminative capability serves as a useful guide for identifying an appropriate autoencoder architecture, and so defer the more challenging generative aspect until this is established.

Through its t-dependence, the diffusion model provides a one-parameter family of learned representations. There are two subtleties to this however. The first arises from the fact that the input to

Model	Encoder / Decoder	Diffusion
S	4.7M	7.3M
M	18.9M	29.0M

Table 1: Number of parameters for the S and M versions of the LSD model. The decoder is trivialised for the MLM diffusion baseline.

the diffusion model, z_t, depends on both t and sampled ε. As ε essentially amounts to Gaussian broadening, we can treat this as regularization and employ the mean value, i.e. take

$$\bar{v}_t(z) = \texttt{Diffusion}(\cos(\pi t/2)z, t). \tag{6}$$

The second subtlety is that the diffusion model learns nothing for $t \sim 1$ as the input there is noise. This can be compensated for by switching to the score function, which from Tweedie's formula Robbins (1992) is expressed as

$$s_t(z) = -\frac{\hat{\varepsilon}_t(z)}{\sin(\pi t/2)} = -\frac{2}{\sin(\pi t)}\big(\hat{v}_t(z) + \sin(\pi t/2)z\big). \tag{7}$$

Dropping the singular prefactor, we thus take the diffusion representations as

$$\bar{v}_t(z) + \sin(\pi t/2)z. \tag{8}$$

At $t = 0$, this reduces back to $\hat{v}_t(z)$.

4 Evaluation

We present here two trained models for both LSD-TN and LSD-NM variants. We call these S and M, and provide their parameter counts in Table 1. Full model hyperparameters and training details are given in Appendix A. To establish a baseline for their performance we additionally train corresponding MLM models, along with a diffusion model on their learned embeddings, using an identical setup. (In terms of Fig. 4 of Appendix A, the decoder's transformer trunk is trivialized. Masking is applied at the input to the encoder, as opposed the input of the decoder for LSD-NM. We employ a 15% masking rate, and extract the latent embeddings after the layer norm following the transformer layers to ensure they are appropriately normalised.) All models are trained on the Uniref50 protein sequence dataset Suzek et al. (2015), with sequences of maximum length 254, and omitting sequences with unknown or non-canonical amino acids (0.5% of the dataset).

We assess the discriminative capabilities of these models across a set of a property prediction tasks assessing stability, interaction and functional characterization, which we adopt from SaProt Su et al. (2023). We conduct zero-shot evaluation, freezing the backbone and training a simple predictor on the mean of the embeddings across the sequence. We provide further information on the datasets and predictor architecture in Appendix B.

We report the performance of the models in Table 2. For the LSD-NM and LSD-TN models and the MLM diffusion baseline, we evaluate on both the latent space, at the output of the encoder, and on the $t = 0$ output of the diffusion model applied on the latent space. To further benchmark these results we additionally evaluate the predictor on two prominent protein representation learning models, ESM2 Lin et al. (2023) and the discrete diffusion model DPLM Wang et al. (2024).

We first highlight the diffusion model results, which are the primary focus of this work. We see that the LSD-TN and LSD-NM diffusion models consistently outperform the MLM diffusion models across all evaluation metrics. Comparing the between the LSD-TN and LSD-NM variants, we see that the LSD-NM performs better than LSD-TN on all but one task. The exception is the HumanPPI, on which LSD-TN-M performs notably better. This may indicate a complementarity in how the two different constructions organise correlations within their respective latent spaces.

Turning to the latent representations of the encoder we see that the situation is reversed, with the MLM results here greatly outperforming their LSD counterparts. This aligns with the well-established strength of masked language modelling for representation learning, in contrast to the LSD autoencoders which were not designed to optimise for this. Indeed, the MLM encoder performs

Models	Thermostability ↑	HumanPPI ↑	Metal Ion Binding ↑	DeepLoc ↑	
				Subcellular	Binary
	Spearman's ρ	Acc (%)	Acc (%)	Acc (%)	Acc (%)
ESM 8M	0.648	72.7	63.2	68.2	88.8
ESM 650M	0.690	81.3	66.8	77.6	91.0
DPLM 650M	0.693	76.7	69.1	78.5	90.8
MLM-S: Encoder	0.606	72.3	65.0	60.1	86.3
MLM-S: Diffusion ($t = 0$)	0.474	57.7	63.0	46.1	74.3
MLM-M: Encoder	0.613	72.3	63.5	62.4	87.3
MLM-M: Diffusion ($t = 0$)	0.543	60.6	61.7	52.2	76.2
LSD-TN-S: Encoder	0.560	58.6	64.6	53.0	76.6
LSD-TN-S: Diffusion ($t = 0$)	0.562	62.6	62.8	48.2	75.3
LSD-TN-M: Encoder	0.571	59.1	63.2	54.4	76.7
LSD-TN-M: Diffusion ($t = 0$)	0.571	65.9	62.6	52.7	76.5
LSD-NM-S: Encoder	0.553	62.6	64.1	54.6	77.6
LSD-NM-S: Diffusion ($t = 0$)	0.567	60.2	65.0	53.5	76.1
LSD-NM-M: Encoder	0.571	61.6	64.6	55.0	77.3
LSD-NM-M: Diffusion ($t = 0$)	0.581	61.1	64.7	54.2	76.8

Table 2: Zero-shot performance on protein property prediction tasks. The $t = 0$ diffusion representations are highlighted, green for the LSD models and gray for the MLM baseline. Reported scores are computed as the mean of 5 randomly initialized predictors.

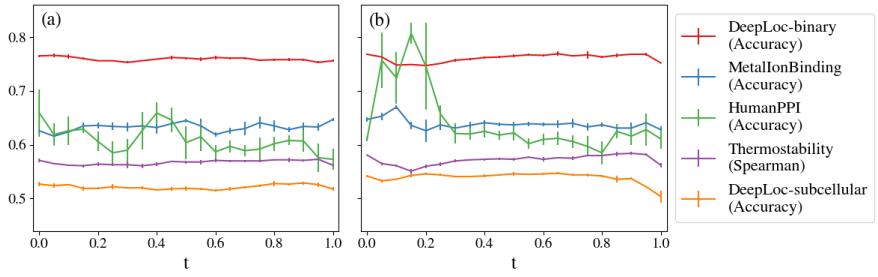


Figure 2: Evaluation of the t-dependence of the diffusion representation for the five protein property prediction tasks: **(a)** LSD-TN-M, **(b)** LSD-NM-M. The error bars are computed from the results of 5 randomly initialized predictors.

best of all the evaluated MLM, LSD-TN and LSD-NM representations, and performs significantly better than even the best LSD diffusion models. This trend is also exhibited by the ESM 8M model, which has a smaller parameter count than all the LSD models.

We also compare between the encoder and diffusion representations. For the LSD-TN model we observe a complementarity, with results consistently better for the encoder on Metal Ion Binding and DeepLoc-Subcellular, on a par for Thermostability and DeepLoc-Binary, and better for the diffusion model on HumanPPI. For LSD-NM the results are more similar between the two modules, while for the MLM model the diffusion representations all significantly score lower than the encoder's.

We now turn to the t-dependence of the diffusion representations. In Fig. 2 we evaluate the performance of the regularised score function $\bar{v}_t(z) + \sin(\pi t/2)z$ for all five tasks for the LSD-TN-M and LSD-NM-M models. For the LSD-TN model we observe that the curves are notably flat, with the exception of HumanPPI although that could reflect the greater uncertainty in that metric. For LSD-NM on the other hand, there is some variation to the curves with different trends for the different tasks. This may reflect the expectation that different correlations are captured at the different scales parameterised by t, but a definitive conclusion cannot be made. Again the HumanPPI metric stands out. We observe a consistent peak at $t = 0.15$ with value 0.807 ± 0.019, which (remarkably) is on a par with the ESM2 and DPLM 650M models.

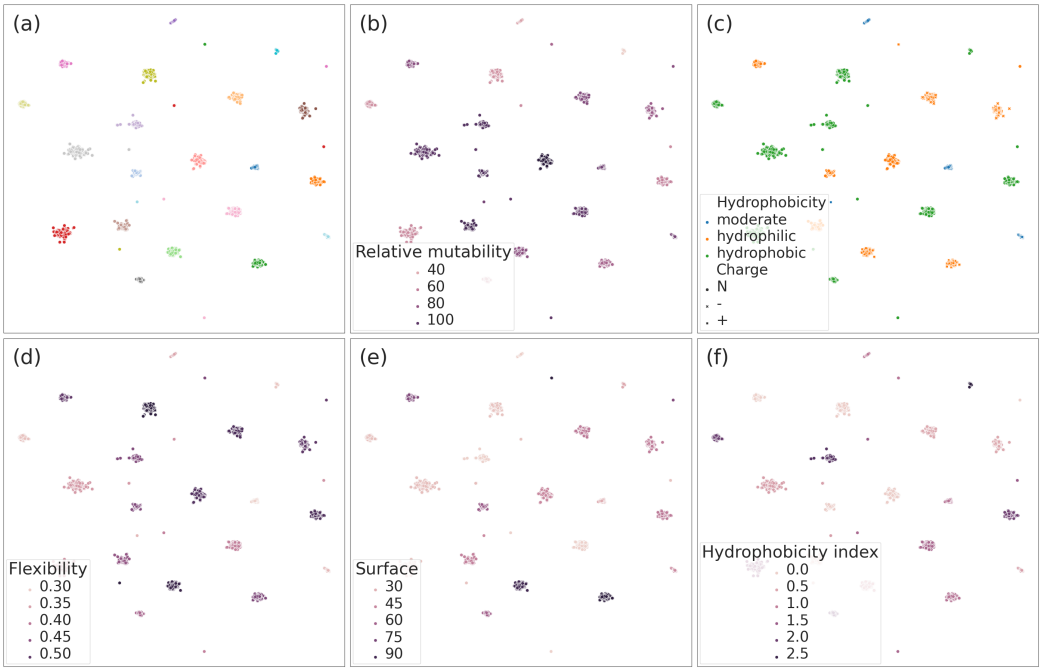


Figure 3: UMAP projections of the latent space learned by LSD-NM-M Diffusion model. **(a)** Coloured by amino acid. **(b)** Coloured by relative mutability Jones et al. (1992). **(c)** Coloured by hydrophobicity nature. **(d)** Coloured by hydrophobicity index Argos et al. (1982). **(e)** Coloured by average flexibility index Bhaskaran & Ponnuswamy (1988). **(f)** Coloured by residue accessible surface area in folded protein Chothia (1976).

4.1 Visualization

To complement the above quantitative analysis, we provide a UMAP-based visual analysis of the learned representations in Fig. 3. We focus on the best performing LSD-NM diffusion representation, and use colouring to highlight the learned biological features. For each plot, we sample 64 sequences of length 100 amino acids from UniRef50, process them through the encoder and diffusion models, and employ UMAP to project the resulting embeddings to 2D.

We also conduct an attention map analysis for the LSD-NM-S and MLM-S models in appendix D to better understand how contextual information is integrated in each of these models.

5 Discussion

The results of our evaluation highlight the key challenge in applying latent space diffusion to protein sequences: identifying an appropriate latent space. We observe that embeddings optimized for representation learning, e.g. those from the MLM baseline, result in an underperforming diffusion model. To address this, we proposed and analyzed alternative latent space learning methods designed to prioritise well-distributed embeddings. While these achieved the goal of boosting the diffusion model's performance, they ultimately fell short of matching the overall performance of token-based reconstructive learning methods like MLM, or the discrete diffusion method of DPLM.

Nevertheless the autoencoder architectures we present have interesting features that may warrant further study. To our knowledge, the Token Norm bottleneck introduced here is novel. It is particularly suitable for protein sequence data, where the 20 amino acids provide a limited vocabulary compared to the much larger token vocabularies commonly used in NLP sequence modelling. The LSD-TN model is notable for its simplicity, achieving reasonable representation learning performance despite possessing a homogeneous bottleneck. We remark also that the univariate parametric form of the Kullback-Leibler divergence normalization loss is crude, and can perhaps be improved.

Our noise masking strategy for the LSD-NM model is quite similar to the diffusion denoising. A key difference however is that for the autoencoder the noise is applied inhomogeneously, while for the diffusion model it is applied uniformly across the sequence. The former places more emphasis on locality, while the latter learns the overall data distribution, underpining its generative capability. It may be interesting to explore if the two can be effectively combined. One possibility is to let the decoder and diffusion model share the same transformer trunk.

We comment also on the one-parameter representations offered by the t-dependence of the diffusion model. As described in Sec. 3 these amount to a regularised form of the score function, $s_t(z)$. We recall from the Introduction that the score function admits an interpretation as a distributional force. Given that our models are trained on the evolutionary-scale Uniref50 dataset, we can thus offer an interpretation of $s_t(z)$ as a representation of the forces governing proteins, with the parameter t setting the scale of the latent space over which these forces are computed.

The DPLM representation included in Table 2 correspond to the $t = 0$ pass of their discrete diffusion model. It is unclear whether these can be extended non-zero t, as in this case the self-averaging property of Gaussian noise is lost, but this may be worthy of further investigation.

6 Conclusion and Outlook

We have presented a Latent Space Diffusion approach for modelling protein sequence data, with an initial focus on discriminative modelling. We highlighted the key challenge in developing this framework, which is to learn a sufficiently well-distributed latent space for the effective training of the diffusion model. To this end we proposed two novel autoencoder architectures: LSD-TN and LSD-NM. We evaluated their zero-shot predictive performance across a range of protein prediction tasks and conducted an ablation study of key design choices. We found that while the diffusion performed better than with an MLM baseline, ultimately our trained models underperformed relative to token-based reconstructive learning approaches.

Our study provides an initial exploration of the LSD approach and opens up several interesting directions for future work. The architecture itself is not settled, and further research is needed to refine the question of what constitutes a good latent space. Another aspect is the latent space's dimensionality. It has been demonstrated that the learned embeddings of ESM(Fold) can be massively compressed without significantly degrading their information content Lu et al. (2024). This motivates a compression of the latent space, which in turn can facilitate more effective distributional modelling.

There is much to be learned about the richness of the information captured by the one-parameter family of diffusion representations, and how this can be best employed for protein modelling. Looking ahead, we also highlight that LSD could serve as a pre-trained model for fine-tuning on specific tasks. In particular, freezing the autoencoder while fine-tuning the diffusion model offers a particularly natural route forward, and may help to bypass the catastrophic forgetting issues observed in masked language models Wallat et al. (2021); Schmirler et al. (2024).

Another promising avenue is the incorporation of additional modalities, particularly protein structural data Mansoor et al. (2024). In this regard the continuous nature of the LSD formulation provides an advantage over discrete token-based approaches Hayes et al. (2024); Su et al. (2023).

Finally, it would be of great interest to scale up the model and explore its generative capabilities.

Meaningfulness Statement

Proteins are a fundamental class of biomolecules whose functions, interactions, and evolutionary relationships are critical to understanding cellular mechanisms and the complexity of life. Recent breakthroughs in deep learning and high-throughput sequencing have revolutionized protein representation learning, opening new avenues for understanding biological systems. In this exploratory study at the frontier of protein sequence modelling, we explore how manifold learning and distributional modelling can be integrated to capture complementary aspects of protein representations.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pp. 1–3, 2024.

Korbinian Abstreiter, Sarthak Mittal, Stefan Bauer, Bernhard Schölkopf, and Arash Mehrjou. Diffusion-based representation learning. *arXiv preprint arXiv:2105.14257*, 2021.

Sarah Alamdari, Nitya Thakkar, Rianne van den Berg, Alex Xijie Lu, Nicolo Fusi, Ava Pardis Amini, and Kevin K Yang. Protein generation with evolutionary diffusion: sequence is all you need. *bioRxiv*, pp. 2023–09, 2023.

José Juan Almagro Armenteros, Casper Kaae Sønderby, Søren Kaae Sønderby, Henrik Nielsen, and Ole Winther. Deeploc: prediction of protein subcellular localization using deep learning. *Bioinformatics*, 33(21):3387–3395, 2017.

Patrick Argos, JK Mohana Rao, and Paul A Hargrave. Structural prediction of membrane-bound proteins. *European Journal of Biochemistry*, 128(2-3):565–575, 1982.

RPPK Bhaskaran and PK Ponnuswamy. Positional flexibilities of amino acid residues in globular proteins. *International Journal of Peptide and Protein Research*, 32(4):241–255, 1988.

Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110, 2022.

Tianlai Chen, Pranay Vure, Rishab Pulugurta, and Pranam Chatterjee. Amp-diffusion: Integrating latent diffusion with protein language models for antimicrobial peptide generation. *bioRxiv*, pp. 2024–03, 2024.

Cyrus Chothia. The nature of the accessible and buried surfaces in proteins. *Journal of molecular biology*, 105(1):1–12, 1976.

Christian Dallago, Jody Mou, Kadina E Johnston, Bruce J Wittmann, Nicholas Bhattacharya, Samuel Goldman, Ali Madani, and Kevin K Yang. Flip: Benchmark tasks in fitness landscape inference for proteins. *bioRxiv*, pp. 2021–11, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *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)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Sander Dieleman, Laurent Sartran, Arman Roshannai, Nikolay Savinov, Yaroslav Ganin, Pierre H Richemond, Arnaud Doucet, Robin Strudel, Chris Dyer, Conor Durkan, et al. Continuous diffusion for categorical data. *arXiv preprint arXiv:2211.15089*, 2022.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):7112–7127, 2021.

Noelia Ferruz, Steffen Schmidt, and Birte Höcker. ProtGPT2 is a deep unsupervised language model for protein design. *Nature communications*, 13(1):4348, 2022.

Zhujin Gao, Junliang Guo, Xu Tan, Yongxin Zhu, Fang Zhang, Jiang Bian, and Linli Xu. Empowering diffusion models on the embedding space for text generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4664–4683, June 2024. doi: 10.18653/v1/2024.naacl-long.261.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Nate Gruver, Samuel Stanton, Nathan Frey, Tim GJ Rudner, Isidro Hotzel, Julien Lafrance-Vanasse, Arvind Rajpal, Kyunghyun Cho, and Andrew G Wilson. Protein design with guided discrete diffusion. *Advances in neural information processing systems*, 36, 2024.

Ishaan Gulrajani and Tatsunori B Hashimoto. Likelihood-based diffusion language models. *Advances in Neural Information Processing Systems*, 36, 2024.

Majdi Hassan, Nikhil Shenoy, Jungyoon Lee, Hannes Stark, Stephan Thaler, and Dominique Beaini. Et-flow: Equivariant flow-matching for molecular conformer generation. *arXiv preprint arXiv:2410.22388*, 2024.

Tomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *bioRxiv*, pp. 2024–07, 2024.

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

Mingyang Hu, Fajie Yuan, Kevin Yang, Fusong Ju, Jin Su, Hui Wang, Fei Yang, and Qiuyang Ding. Exploring evolution-aware &-free protein language models as protein function predictors. *Advances in Neural Information Processing Systems*, 35:38873–38884, 2022.

John B Ingraham, Max Baranov, Zak Costello, Karl W Barber, Wujie Wang, Ahmed Ismail, Vincent Frappier, Dana M Lord, Christopher Ng-Thow-Hing, Erik R Van Vlack, et al. Illuminating protein space with a programmable generative model. *Nature*, 623(7989):1070–1078, 2023.

Wengong Jin, Caroline Uhler, and Nir Hacohen. SE (3) denoising score matching for unsupervised binding energy prediction and nanobody design. In *NeurIPS 2023 Generative AI and Biology (GenBio) Workshop*, 2023.

Bowen Jing, Bonnie Berger, and Tommi Jaakkola. Alphafold meets flow matching for generating protein ensembles. *arXiv preprint arXiv:2402.04845*, 2024.

David T Jones, William R Taylor, and Janet M Thornton. The rapid generation of mutation data matrices from protein sequences. *Bioinformatics*, 8(3):275–282, 1992.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with Alphafold. *Nature*, 596(7873):583–589, 2021.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

Jin Sub Lee, Jisun Kim, and Philip M Kim. Score-based generative modeling for de novo protein design. *Nature Computational Science*, 3(5):382–392, 2023.

Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. *Advances in Neural Information Processing Systems*, 35:4328–4343, 2022.

Yifan Li, Kun Zhou, Wayne Xin Zhao, and Ji-Rong Wen. Diffusion models for non-autoregressive text generation: A survey. *arXiv preprint arXiv:2303.06574*, 2023.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Shengchao Liu, Hongyu Guo, and Jian Tang. Molecular geometry pretraining with se (3)-invariant denoising distance matching. *arXiv preprint arXiv:2206.13602*, 2022.

Amy X Lu, Wilson Yan, Kevin K Yang, Vladimir Gligorijevic, Kyunghyun Cho, Pieter Abbeel, Richard Bonneau, and Nathan Frey. Tokenized and continuous embedding compressions of protein sequence and structure. *bioRxiv*, pp. 2024–08, 2024.

Ali Madani, Ben Krause, Eric R Greene, Subu Subramanian, Benjamin P Mohr, James M Holton, Jose Luis Olmos, Caiming Xiong, Zachary Z Sun, Richard Socher, et al. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 41(8):1099–1106, 2023.

Sanaa Mansoor, Minkyung Baek, Hahnbeom Park, Gyu Rie Lee, and David Baker. Protein ensemble generation through variational autoencoder latent space sampling. *Journal of Chemical Theory and Computation*, 20(7):2689–2695, 2024.

Viacheslav Meshchaninov, Pavel Strashnov, Andrey Shevtsov, Fedor Nikolaev, Nikita Ivanisenko, Olga Kardymon, and Dmitry Vetrov. Diffusion on language model embeddings for protein sequence generation. *arXiv preprint arXiv:2403.03726*, 2024.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pp. 8162–8171. PMLR, 2021.

Xiao-Yong Pan, Ya-Nan Zhang, and Hong-Bin Shen. Large-scale prediction of human protein-protein interactions from amino acid sequence based on latent topic features. *Journal of proteome research*, 9(10):4992–5001, 2010.

William Peebles and Saining Xie. Scalable Diffusion Models with Transformers, March 2023. URL `http://arxiv.org/abs/2212.09748`. arXiv:2212.09748 [cs].

Roshan M Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. MSA transformer. In *International Conference on Machine Learning*, pp. 8844–8856. PMLR, 2021.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

Herbert E Robbins. An empirical bayes approach to statistics. In *Breakthroughs in Statistics: Foundations and basic theory*, pp. 388–394. Springer, 1992.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

Robert Schmirler, Michael Heinzinger, and Burkhard Rost. Fine-tuning protein language models boosts predictions across diverse tasks. *Nature Communications*, 15(1):7407, 2024.

Emre Sevgen, Joshua Moller, Adrian Lange, John Parker, Sean Quigley, Jeff Mayer, Poonam Srivastava, Sitaram Gayatri, David Hosfield, Maria Korshunova, et al. Prot-vae: protein transformer variational autoencoder for functional protein design. *bioRxiv*, pp. 2023–01, 2023.

Noam Shazeer. GLU Variants Improve Transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Sam Sinai, Eric Kelsic, George M Church, and Martin A Nowak. Variational auto-encoding of protein sequences. *arXiv preprint arXiv:1712.03346*, 2017.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. PMLR, 2015.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Robin Strudel, Corentin Tallec, Florent Altché, Yilun Du, Yaroslav Ganin, Arthur Mensch, Will Grathwohl, Nikolay Savinov, Sander Dieleman, Laurent Sifre, et al. Self-conditioned embedding diffusion for text generation. *arXiv preprint arXiv:2211.04236*, 2022.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. Saprot: Protein language modeling with structure-aware vocabulary. *bioRxiv*, pp. 2023–10, 2023.

Baris E Suzek, Yuqi Wang, Hongzhan Huang, Peter B McGarvey, Cathy H Wu, and UniProt Consortium. Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932, 2015.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Timothy Truong Jr and Tristan Bepler. PoET: A generative model of protein families as sequences-of-sequences. *Advances in Neural Information Processing Systems*, 36:77379–77415, 2023.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pp. 1096–1103, 2008.

Jonas Wallat, Jaspreet Singh, and Avishek Anand. BERTnesia: Investigating the capture and forgetting of knowledge in BERT. *arXiv preprint arXiv:2106.02902*, 2021.

Xinyou Wang, Zaixiang Zheng, Fei Ye, Dongyu Xue, Shujian Huang, and Quanquan Gu. Diffusion language models are versatile protein learners. *arXiv preprint arXiv:2402.18567*, 2024.

Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with RFdiffusion. *Nature*, 620(7976):1089–1100, 2023.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. On layer normalization in the transformer architecture. In *International Conference on Machine Learning*, pp. 10524–10533. PMLR, 2020.

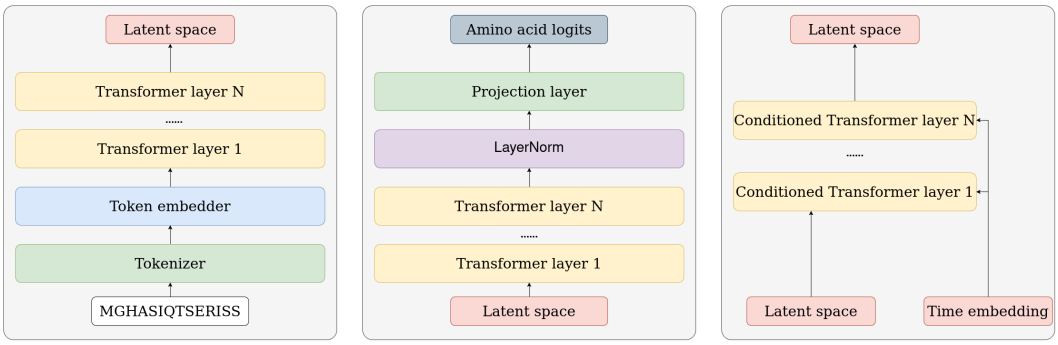
Minghao Xu, Zuobai Zhang, Jiarui Lu, Zhaocheng Zhu, Yangtian Zhang, Ma Chang, Runcheng Liu, and Jian Tang. Peer: a comprehensive and multi-task benchmark for protein sequence understanding. *Advances in Neural Information Processing Systems*, 35:35156–35173, 2022.

Jiasheng Ye, Zaixiang Zheng, Yu Bao, Lihua Qian, and Mingxuan Wang. DINOISER: Diffused Conditional Sequence Learning by Manipulating Noises, 2024. URL https://arxiv.org/abs/2302.10025.

Jason Yim, Brian L Trippe, Valentin De Bortoli, Emile Mathieu, Arnaud Doucet, Regina Barzilay, and Tommi Jaakkola. SE (3) diffusion model with application to protein backbone generation. *arXiv preprint arXiv:2302.02277*, 2023.

Sheheryar Zaidi, Michael Schaarschmidt, James Martens, Hyunjik Kim, Yee Whye Teh, Alvaro Sanchez-Gonzalez, Peter Battaglia, Razvan Pascanu, and Jonathan Godwin. Pre-training via denoising for molecular property prediction. *arXiv preprint arXiv:2206.00133*, 2022.

A ADDITIONAL MODEL DETAILS



(a) (b) (c)

Figure 4: Detailed architecture: **(a)** encoder, **(b)** decoder **(c)** diffusion model.

All three components of the model, encoder-decoder-diffusion, are based on the transformer architecture as illustrated in Fig. 4. We employ RoPE positional encoding Su et al. (2024), Pre-LN Xiong et al. (2020), SwiGLU activation functions Shazeer (2020). Sequences are padded to maximum sequence length 256, and each component has internal BOS/EOS embeddings which are learned independently and not output. For the decoder we employ a simple projection head onto the amino acid logits. Time-conditioning for the diffusion model is implemented using the adaLN-zero prescription described in DiT Peebles & Xie (2023).

The models are trained on Uniref50 with sequences limited to a maximum length of 254 (+2 for BOS/EOS embeddings), with a batch size of 512, using the AdamW optimizer with weight decay 1e-3 and learning rate 2e-5, on one A100 80GB GPU. Model hyper-parameters are provided in Table 3.

Table 3: Model details

Model name	Module	Model size	Channels	Heads	Layers	Steps
MLM-S	Encoder Diffusion	4.7M 7.3M	256	16	6	200k
MLM-M	Encoder Diffusion	18.9M 29.0M	512	16	6	100k
LSD-TN-S	Encoder/Decoder Diffusion	4.7M 7.3M	256	16	6	200k
LSD-TN-M	Encoder/Decoder Diffusion	18.9M 29.0M	512	16	6	100k
LSD-NM-S	Encoder/Decoder Diffusion	4.7M 7.3M	256	16	6	200k
LSD-NM-M	Encoder/Decoder Diffusion	18.9M 29.0M	512	16	6	100k

B EVALUATION DETAILS

In Section 4 we evaluate representation learning on a set of protein property prediction tasks which we adopt from SaProt Su et al. (2023):

- **Thermostability**: protein melting temperature T_m data from the "Human-cell" splits of the Thermostability task of the FLIP benchmark Dallago et al. (2021).

- **HumanPPI**: binary classification whether two proteins interact from HumanPPI data Pan et al. (2010) of the PEER benchmark Xu et al. (2022).

- **Metal Ion Binding**: binary classification of presence of metal ion–binding sites within a protein Hu et al. (2022).

- **DeepLoc**: Predicts subcellular localization of proteins from the DeepLoc dataset Almagro Armenteros et al. (2017).

  - **Subcellular**: multi-class classification identifying one of 10 distinct subcellular compartments.

  - **Binary**: binary classification between membrane-bound or soluble.

We perform zero-shot evaluation, training a predictor on the output of a frozen backbone. This approach differs from the SaProt pipeline, which fine-tunes the backbone. As a result the values we obtain for ESM2 are not identical to theirs; however they remain directly comparable. The same applies to DPLM, which also uses the SaProt evaluation pipeline.

The predictor is a simple 2-layer regressor or classifier head. At input we take the mean of the embeddings across the sequence, and for the HumanPPI task we concatenate the mean embeddings of the two proteins. We take the hidden layer dimension of predictor equal to the dimension of the input.

C Ablations

Model	Importance sampling	Modules	Thermostability ↑	HumanPPI ↑	Metal Ion Binding ↑	DeepLoc ↑	
						Subcellular	Binary
LSD-NM-S	✓	Encoder	0.553	62.6	64.1	54.6	77.6
		Diffusion	0.567	60.2	65.0	53.5	76.1
	Off for noise masking	Encoder	0.558	61.7	63.7	54.4	77.3
		Diffusion	0.545	64.7	63.8	52.7	75.8
	Off for decoder	Encoder	0.543	60.6	62.8	55.2	77.2
		Diffusion	0.540	68.0	59.2	52.2	76.5

Table 4: Importance sampling ablation.

Model	Loss	Modules	Thermostability ↑	HumanPPI ↑	Metal Ion Binding ↑	DeepLoc ↑	
						Subcellular	Binary
LSD-TN-S	Univariate	Encoder	0.560	58.6	64.6	53.0	76.6
		Diffusion	0.562	62.6	62.8	48.2	75.3
	Multivariate	Encoder	0.548	60.6	63.6	51.6	76.5
		Diffusion	0.528	53.2	61.1	44.9	75.3

Table 5: Normalization loss ablation: we compare the univariate parametric form of the Kullback–Leibler divergence $\frac{1}{2d}\sum_i(\mu_i^2+\sigma_i^2-\log\sigma_i^2-1)$ to its multivariate counterpart $\frac{1}{2d}(\mu^\top\mu+\Sigma-\log\det\Sigma-d)$, with Σ the covariance matrix.

D CONTEXT LEARNING ANALYSIS

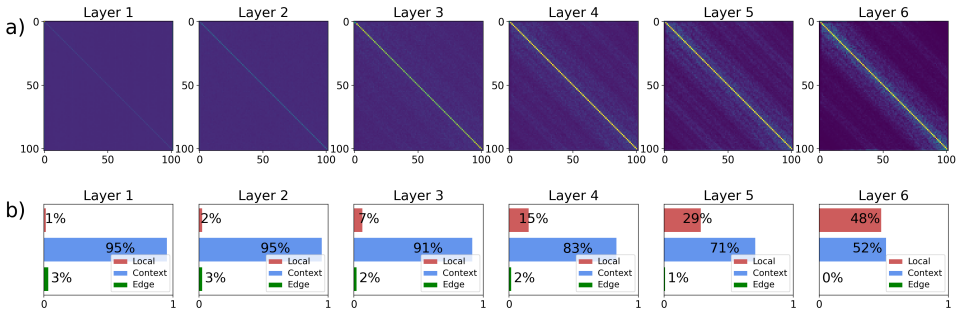


Figure 5: Attention Map Analysis for LSD-NM-S diffusion model. **a)** Average attention logits per layer, aggregated over all heads and 128 protein sequences, each consisting of 100 amino acids. **b)** Distribution of attention scores across different types: *Context* attention, *Local* attention, and edge-token attention.

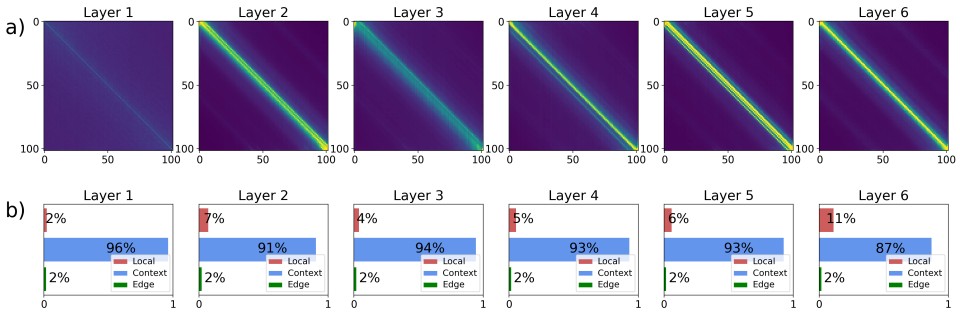


Figure 6: Attention Map Analysis for MLM-S model. **a)** Average attention logits per layer, aggregated over all heads and 128 protein sequences, each consisting of 100 amino acids. **b)** Distribution of attention scores across different types: *Context* attention, *Local* attention, and edge-token attention.

We can better understand which elements influence a given token's representation and how contextual information is integrated by analyzing the attention map of the transformer model and studying its distributions across the layers

In figures 5 and 6, we define *Context* as the sum of the attention logits that connect each position in the sequence to all other different positions. *Local* refers to the attention logits located along the diagonal of the attention weight matrix, representing how much a position attends to itself. Lastly, *Edge* corresponds to the attention logits assigned to the EOS and BOS tokens.

For the LSD-NM-S diffusion model, the early layers predominantly focus on contextual information, with *Context* attention reaching 95% and *Local* attention remaining as low as 1%. This suggests that initial layers are primarily responsible for embedding global context into each amino acid position. As the layer index increases, attention shifts to be more *Local* focused, indicating that final layers refine token embeddings based on the already incorporated contextual information. The model exhibits minimal focus on edge tokens, with attention weight not exceeding 3% and dropping to 0% in the final layer.

In contrast, the MLM-S model maintains a strong reliance on context across all layers, consistently prioritizing *Context* attention. A key difference is that attention in the MLM-S model is more short-range, with logits concentrated around nearby positions along the diagonal, whereas LSD-NM-S diffusion model distributes attention more broadly. This distinction highlights the differing information integration strategies between the two models, where LSD-NM-S diffusion model gradually

transitions from global to local representation, while MLM-S persistently relies on short-range context.