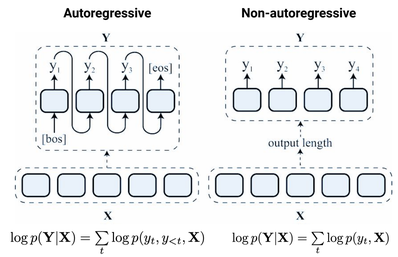
Function-guided protein design by deep manifold sampling

Vladimir Gligorijevic, Daniel Berenberg, Stephen Ra, Simon Kelow, Andy Watkins, Kyunghyun Cho, Richard Bonneau

Prescient Design
A Genentech Accelerator

Motivation

- Protein design remains challenging as it requires searching through a vast combinatorial space that is only sparsely functional [1]
- Conditional design of proteins can accelerate this search by yielding candidates that satisfy constraints
- Some generative model-based approaches, including autoregressive (AR) models, have shown promise [2, 3, 4]
- Known issues with AR models -- decoding latency [6, 8], difficulty of parallelizing inference [7, 9], and exposure bias at test-time generation [11] -- motivates a new approach

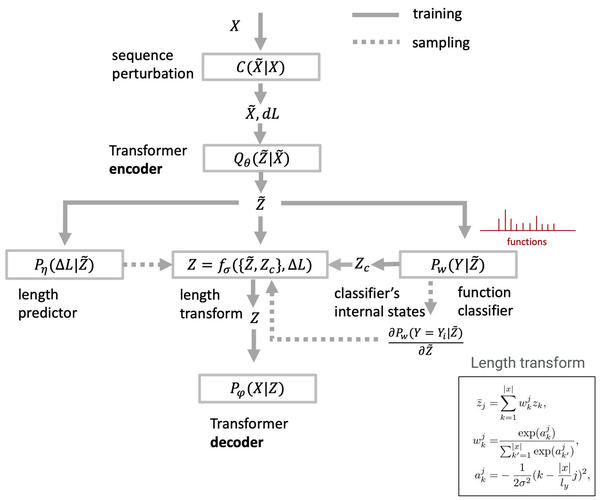


Summary

- We introduce a denoising autoencoder (DAE) [10] that learns a manifold of protein sequences from a large number of potentially unlabelled proteins in a self-supervised manner
- The DAE is combined with a function predictor that guides sampling towards sequences with higher levels of desired functions
- We present preliminary case studies below that demonstrate the effectiveness of this proposed approach, which we refer to as **deep manifold sampling**

Deep Manifold Sampler

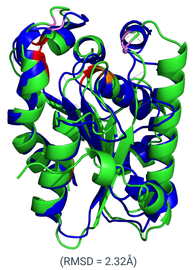
- Corruption function for modeling insertion, deletions & mutations: $C(\tilde{X}|X)$
- Guided-sampling with function predictor: $P_w(Y|\tilde{Z})$
- Length predictor: $P_\eta(\Delta L|\tilde{Z})$



- DAE estimates structure of data-generating density by denoising stochastically-corrupted training examples
- Length predictor [6, 7] outputs a categorical distribution over the length difference between original and corrupted input sequences
- Adaptive length transform [7]
- Non-autoregressive inference procedure makes changes in multiple positions of a target sequence in parallel

Experiments

(1a) Redesign of a cutinase with enhanced functions



(RMSD = 2.32Å)

Ground truth

Designed

Input sequence (PDB: 3GBS_1| Chain A | Cutinase 1 | Aspergillus oryzae)

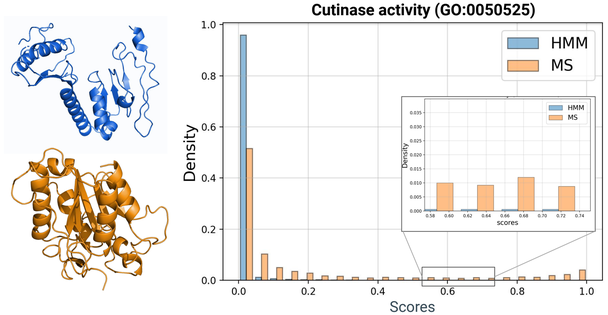
SPVDLQDRQLTGGDELRDGPCKPITFIFARASTEPGLLGISTGPAVCN
RLKLARSGDVACQGVGPRYTADLPSNALPEGTSQAAIAEAQGLFEQA
VSKCPDTQIVAGGYSQGTAVMNGAIKRLSADVQDKIKGVVLFGYTRN
AQERGQIANFPKDKVKVYCAVGDLVCLGTLIVAPPHFSYLSDTGDAS
DFLLSQLG

67% sequence identity

Designed sequence

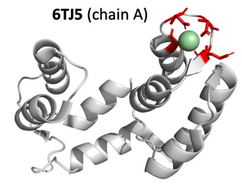
SPVDLQRRQLTGGDELRRGPCKPITFIFARAATEPPGLGISTGGAVKN
RLLKARMGGVACGVVPRRYLDLLNNLLPETSQMAIIYEIGLFQAVVK
CPDDTIVVAGVGSQGGAVMMGGIIRRLSDDQKIKRGGVLLFGYTGNQ
QERGQIAFFPTDKVKVYAAVDDLVCLGTLVVNAPPHGSYLSTTGDAS
PLLSQLG

(1b) Comparison with Hidden Markov Models (HMM)



- **(1a)** We diversify a cutinase sequence by conditioning on "*cutinase activity*" GO term and generate sequences with preserved catalytic residues and higher scores for "*cutinase activity*" (computed by DeepFRI [5])

(2) Addition of a metal-binding site



>>SMTCPPRVREAFALFDTDGDGEISGRDLVLAIRSCGVSPTPDEIKALPMSMAWPDFEAWMSKKLASYNPEEELIKSFKAFD
RSNDGTVSADELSQVMLALGELLSDEEVKAMIKEADPNGTGKIQYANFVKMLLK

Input sequence:

(remove Ca^+ binding residues)

>>SMTCPPRVREAFALFTGGISGRDLVLAIRSCGVSPTPDEIKALPMSMAWPDFEAWMSKKLASYNPEEELIKSFKAFDRSND
GTVSADELSQVMLALGELLSDEEVKAMIKEADPNGTGKIQYANFVKMLLK

Iteration 1

>>SPTCPPYVREAFALFVGGEISGRDLVLAIRSSVVSPKPDDKKALLMMSAWPPDEAAMMKLLAYYLEENLIKSFKKADRRND
GTTVAADDLSQVMALLAELLSDEEVKACIKEADLNGTGKIQYANFVKMLLK

...

Iteration 6

Designed sequence:

>>SNKAPYQLEDRALFVGGSSGRGVVLAIRSSVVSIKPDKKALLMMSAEPPDLEAAGKLLAYYGEARNLDRFYVKADRDGDGG
TTVVWDDSSQFMALLAELLMEEIKHAIKEEDLAGDTKNILYNFVKMLLK

Designed

Conditioning on "calcium ion binding" (GO:0005509)

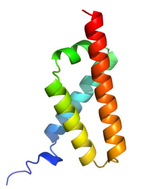
trRosetta

(3) Design of novel secondary structures

Input sequence (PDB: 4N0F | Chain B)

IQRTPKIQVYSRHPAENGKSNFLNCYVSGF
HPSDIEVDLLKNGERIEKVEHSDLSFSKDWS
FYLLYYTEFTPTEKDEYACRVNHVTLSQPKI
VKWDRDM

Redesigned by conditioning on ion transmembrane transporter activity (GO:0015075)



Designed sequence (trRosetta)

IRPPQQVRHYPENNKKMMHFYAELLKYLKKG
EEKVVAKRASFPQLAAYYSWLLEEAVTTQQPK
KFDDAALYYALLYTTVYLPAAFLVDDM

Known functions:

protein binding
identical protein binding
protein homodimerization activity
protein dimerization activity

Predicted functions:

GO term	Name	Prob	Reliability
GO:0015085	transmembrane transport	0.962	H
GO:0034220	ion transmembrane transport	0.942	H
GO:0051171	regulation of nitrogen compound metabolic process	0.926	H
GO:0006811	ion transport	0.916	H
GO:0019222	regulation of metabolic process	0.915	H
GO:1903506	regulation of nucleic acid-templated transcription	0.911	H
GO:0006810	transport	0.910	H
GO:0051252	regulation of RNA metabolic process	0.909	H

- **(1b)** Comparison of MS-designed sequences with HMM; ~1000 sequences with approx. same length as seed sequences (folded with trRosetta [12])
- **(2)** Recovery of metal-binding sites after ablation of known Ca^{2+} binding residues from a calcium-binding protein
- **(3)** Design of α-helical protein sequence by altering β-protein sequence by conditioning on "*ion transmembrane transporter activity*" function label
- Sequence folded by trRosetta [12] and function confirmed by an external function classifier (FFPred3)

NEURAL INFORMATION PROCESSING SYSTEMS

References

[1] Romero & Arnold (2009). Exploring protein fitness landscapes by directed evolution. Nature Reviews Molecular Cell Biology, 10(12)

[2] Madani *et al.* (2020). Progen: Language modeling for protein generation. Nature Reviews Molecular Cell Biology, 10(12)

[3] Shin *et al.* (2021). Protein design and variant prediction using autoregressive generative models. Nature Communications, 12(1)

[4] Brookes *et al.* (2019). Conditioning by adaptive sampling. International Conference on Machine Learning

[5] Gligorijevic *et al.* (2021). Structure-based function prediction using graph convolutional networks. Nature Communications, 12(1)

[6] Lee *et al.* (2018). Deterministic non-autoregressive neural sequence modeling by iterative refinement. arXiv:1802.06901

[7] Shu *et al.* (2020). Latent-variable non-autoregressive neural machine translation with deterministic inference using a delta prior Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34

[8] Cho (2016). Noisy parallel approximate decoding for conditional recurrent language model. arXiv:1606.03835

[9] Gu *et al.* (2018). Non-autoregressive neural machine translation. International Conference on Learning Representations.

[10] Vincent *et al.* (2008). Extracting and composing robust features with denoising autoencoders. International Conference on Machine Learning

[11] Bengio *et al.* (2015). Scheduling sampling for sequence prediction with recurrent neural networks. International Conference on Machine Learning

[12] Yang *et al.* (2020). Improved protein structure prediction using predicted interresidue orientations. Proceedings of the National Academy of Sciences, 117(3)