NeurIPS 2023

ProteinNPT: Improving Protein Property Prediction & Design with Non-Parametric Transformers

Motivations

Learning fitness landscapes is critical to many tasks in biology:

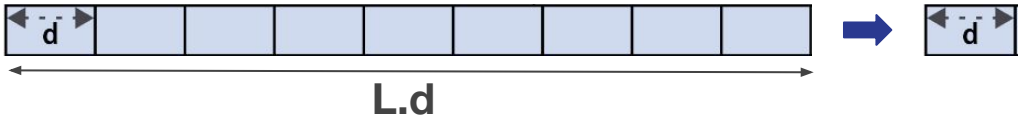
Mutation effects prediction
Effects of genetic mutations in humans

Viral evolution
Predicting which variant are likely to escape immunity

Protein engineering
Designing new biomolecules with desired properties

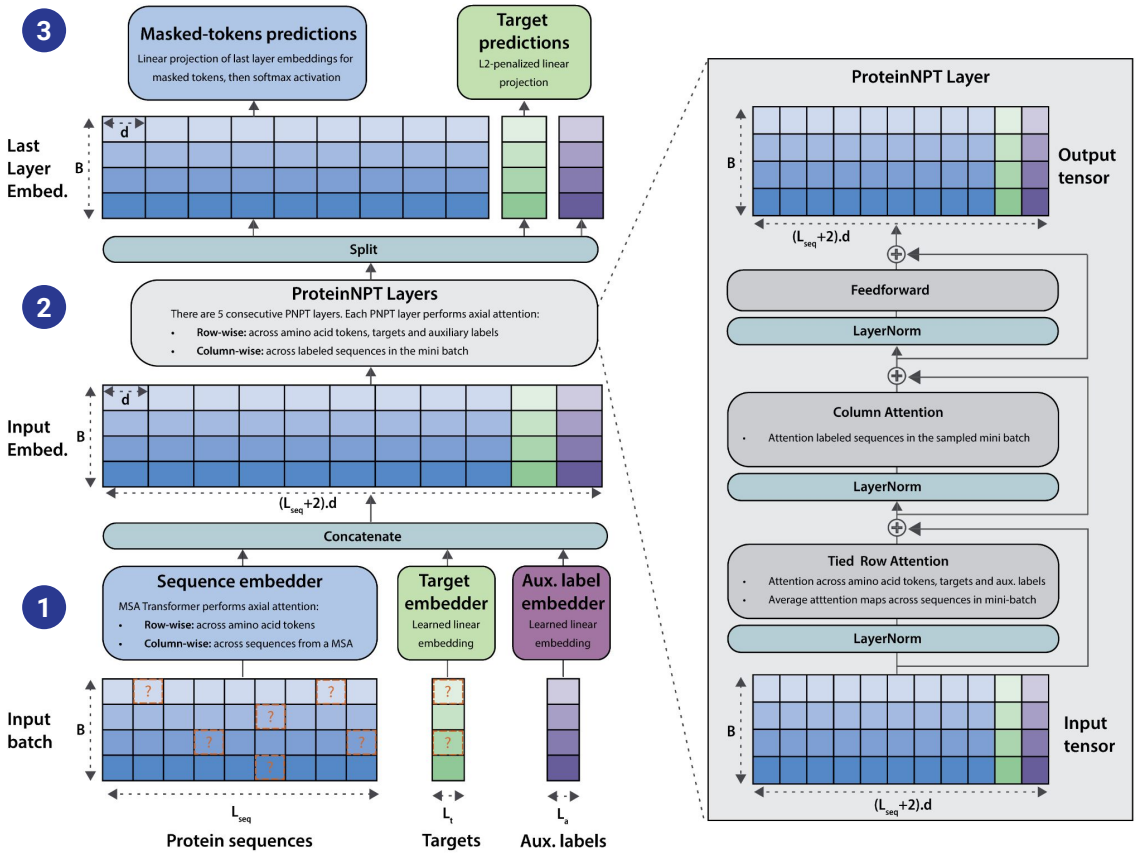
Challenges & limitations of current approaches

- The **protein space is massive** and **annotations** are **sparsely available**
- Protein language models provide **rich representation** of protein sequences. Yet, the dimensionality of the embedded sequences **is typically too large** relative to the number of available labels
- Prior approaches have relied on **limited representations** (e.g., one-hot-encodings) or **dimensionality reduction methods** (e.g., mean-pooling across sequence length[1])



1. Alley et al. Unified rational protein engineering with sequence-based deep representation learning. Nature Methods. 2019

ProteinNPT: A semi-supervised conditional pseudo-generative model for protein property prediction based on a tri-axial attention mechanism sequences



At training time

1 Batch embedding

- We embed an input batch comprised of sequences, targets & auxiliary labels
- We mask a subset of tokens and labels at random

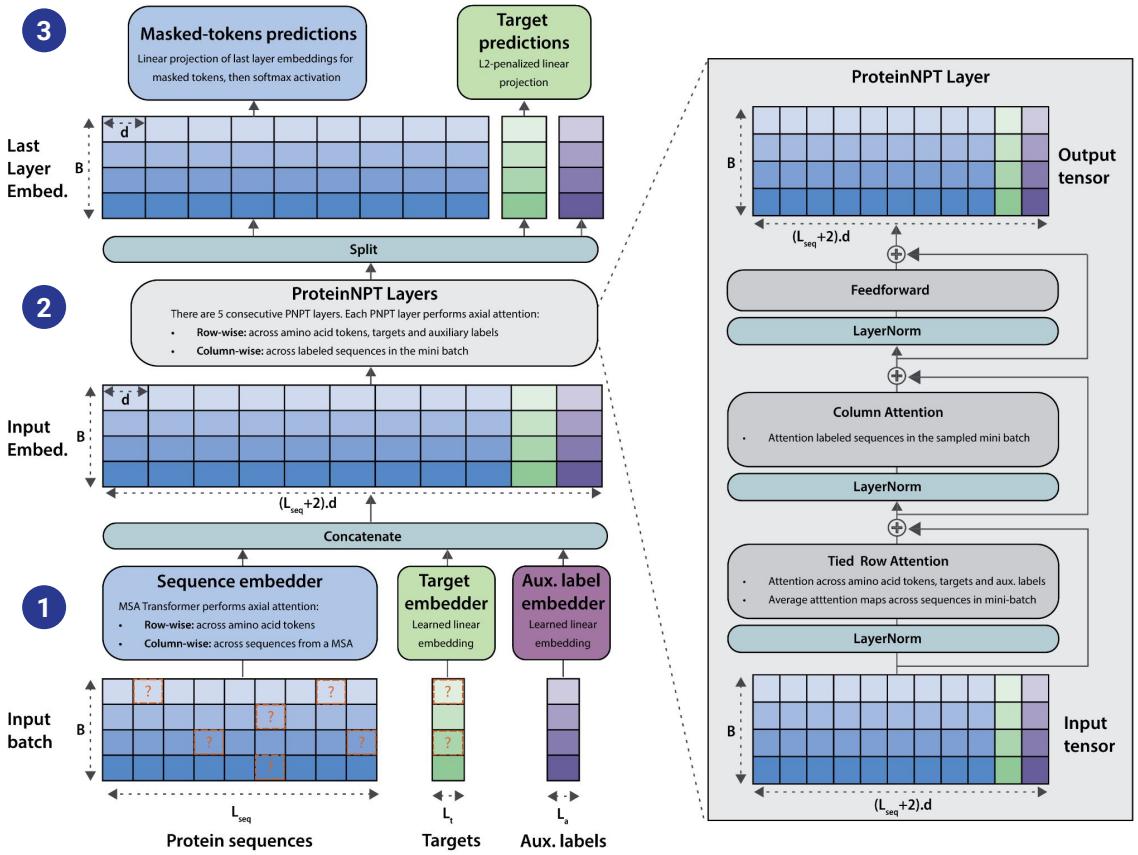
2 Axial attention

- Row attention (horizontally) across tokens and labels
- Column attention (vertically) across labeled sequences

3 Prediction loss

- Use last layer embeddings to predict masked tokens and targets

ProteinNPT: A semi-supervised conditional pseudo-generative model for protein property prediction based on a tri-axial attention mechanism sequences



At inference

1 Batch embedding

- We embed the input batch with trained embeddings
- Targets for the sequences to predict are masked but the batch also includes training sequences w/ known targets
- No sequence token is masked

2 Axial attention

- Same as during training

3 Prediction

- Predict target based on last-layer target embedding

ProteinNPT achieves SOTA performance on protein fitness prediction

Single property prediction

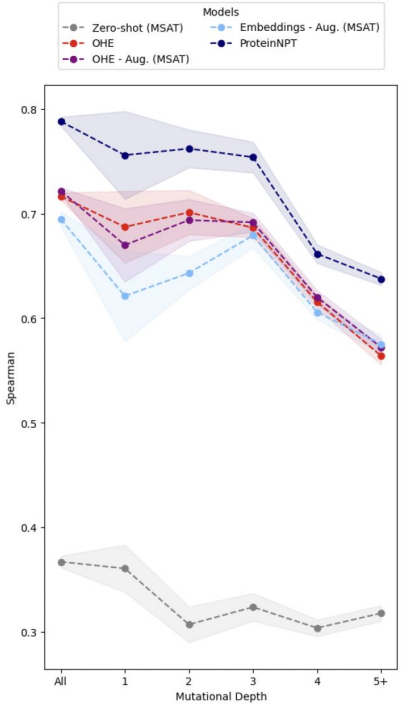
Single substitution

Model name	Spearman (↑) Contig.	Mod.	Rand.	Avg.
OHE	0.08	0.02	0.54	0.21
OHE - Aug. (DS)	0.41	0.40	0.49	0.43
OHE - Aug. (MSAT)	0.41	0.40	0.50	0.44
Embed. - Aug. (MSAT)	0.47	0.49	0.57	0.51
ProteinNPT	**0.48**	**0.51**	**0.66**	**0.55**

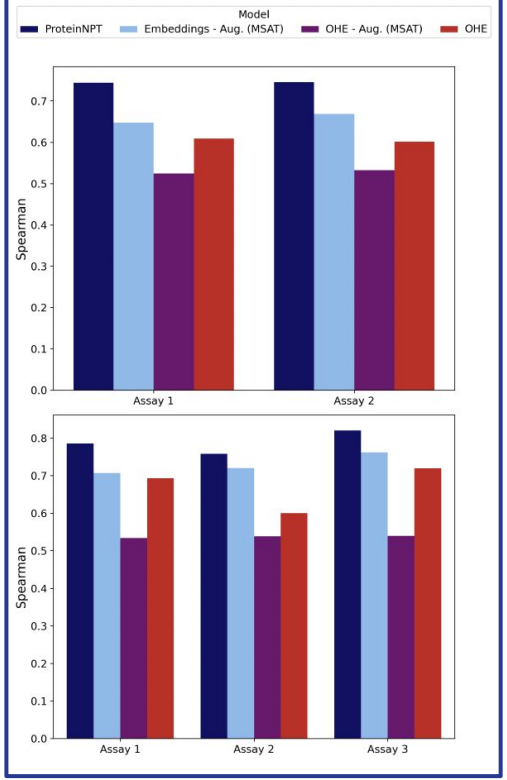
Model name	MSE (↓) Contig.	Mod.	Rand.	Avg.
OHE	1.17	1.11	0.92	1.06
OHE - Aug. (DS)	0.98	0.93	0.78	0.90
OHE - Aug. (MSAT)	0.97	0.92	0.77	0.89
Embed. - Aug. (MSAT)	**0.93**	0.85	0.67	0.82
ProteinNPT	**0.93**	**0.83**	**0.53**	**0.77**

We introduce **3 cross validation schemes** (random, modulo, contiguous) to provide stronger guarantees on ability of fitness predictors to **extrapolate across positions**

Multiple substitutions

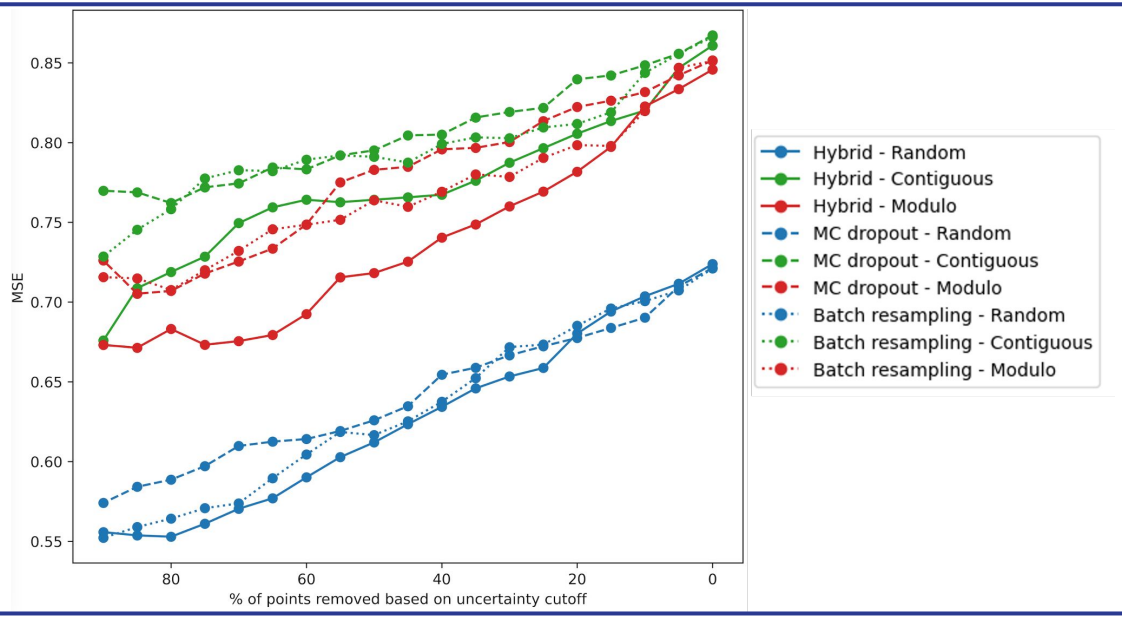


Multiple properties prediction



We implemented and tested 3 different strategies to quantify prediction uncertainty with ProteinNPT

MC dropout	Batch Resampling	Hybrid
Perform MC dropout to sample from model parameters, keeping the same set of labeled sequences across forward passes	Sample different subset of labeled sequences (with replacement) for each forward pass, with no dropout applied	Combine the MC dropout and batch resampling schemes



Uncertainty calibration curves

MSE as a function of the # of test points **excluded based on their uncertainty**:

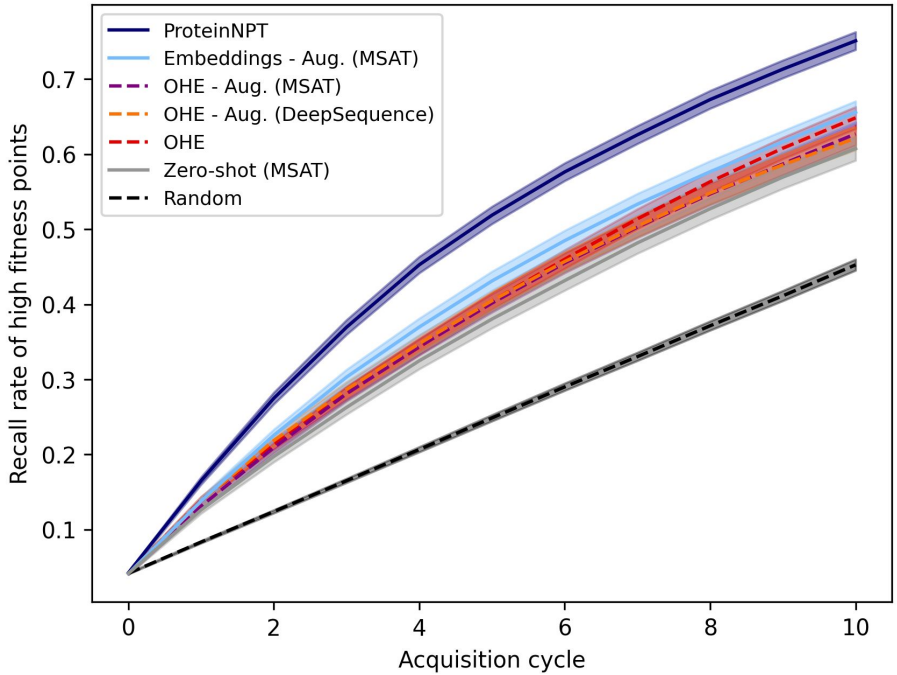
- Rightmost point → no point excluded
- Leftmost point → MSE on the subset of the 10% most confident points

In silico iterative redesign experiments demonstrate significant performance lift from ProteinNPT over prior baselines

Experiment Design

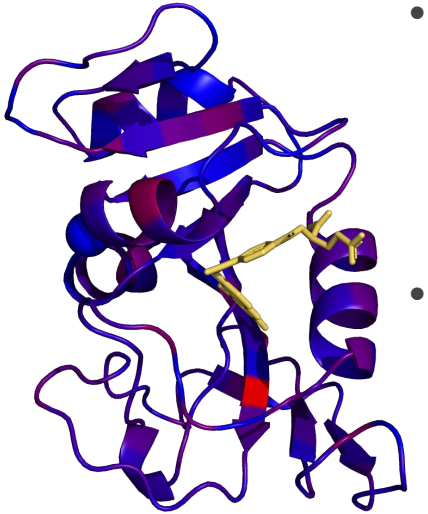
- **Goal:** Start from natural sequences and iteratively mutate sequences to design proteins with improved properties
- **In our setting:** select mutants from sequences tested in DMS assay (masking all label values)
- **Pool based optimisation:** select at each acquisition cycle which sequences to add to training pool from the unlabelled set
- **Bayesian optimization:** select points based on the **Upper Confidence Bound** acquisition function

ProteinNPT outperforms baselines at recalling the sequences with high fitness



Attention mechanisms: row-wise attention captures correlations between labels and positions; column-wise attention is critical to performance

Row-wise attention



- **Row-wise attention maps** recapitulate known dependencies between **labels and residues** -- here a substrate binding site for the DFHR protein (in red)
- It could also **help uncover unknown dependencies** between **certain positions** in sequence and the **property of interest**

Column-wise attention

- **Training ProteinNPT with column-wise attention is critical to reaching SOTA performance**

CV scheme	No column attention	With column attention
Random	0.669	**0.684**
Modulo	0.530	**0.531**
Contiguous	0.425	**0.501**
Average	0.542	**0.572**

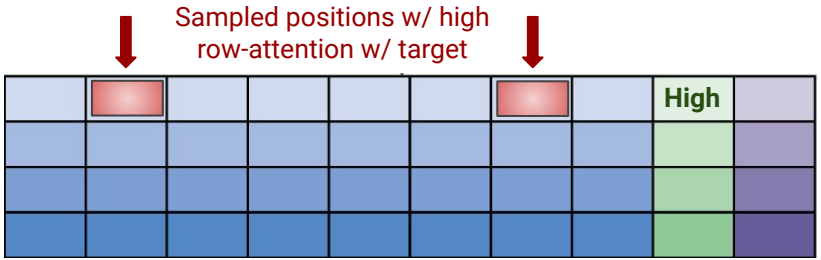
- **At inference, using as few as 100 labeled sequences** for column-wise attention **captures most of the effect**
- **No performance lift** is observed when using **more than 1k labeled sequences**

CV scheme	Nb. labelled sequences sampled at inference 0	100	200	500	1000	2000
Random	0.398	0.677	0.678	0.679	0.684	0.685
Modulo	0.299	0.533	0.531	0.531	0.531	0.531
Contiguous	0.254	0.496	0.504	0.502	0.501	0.500
Average	0.317	0.569	0.571	0.571	**0.572**	**0.572**

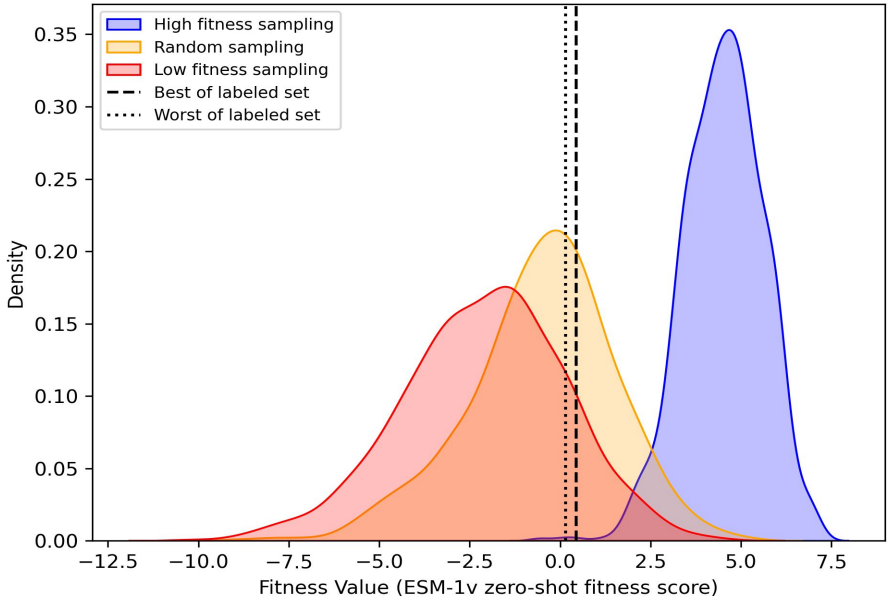
Since ProteinNPT is a conditional (pseudo-)generative model, we can sample new sequences conditioned on specific values of the properties of interest

Conditional sampling approach

1. Select sequence w/ **highest assayed property** (first batch sequence)
2. Form a **complete input batch** by drawing labeled sequences at random
3. **Sample and mask** a few positions in first sequence from subset w/ **high row-attention w/ target**
4. **Sample new amino acids** at these positions based on output **softmax from ProteinNPT**



Fitness of the proteins obtained via the ProteinNPT conditional sampling



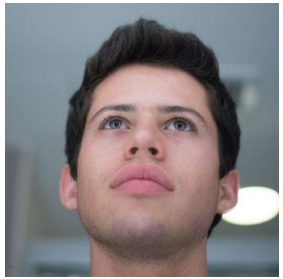
Poster - Great Hall & Hall B1+B2 #308

Come speak with us at the Conference

Pascal

Ruben

Debbie



Yarin

Thank you to our sponsors!

GSK

UKRI Engineering and Physical Sciences Research Council

NIH National Institutes of Health

The Alan Turing Institute