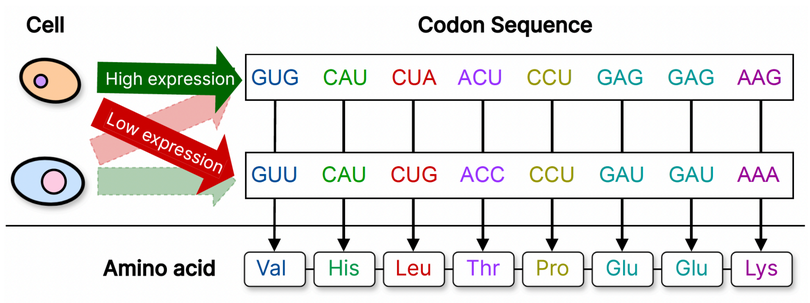
CodonMPNN for Organism Specific and Codon Optimal Inverse Folding

Hannes Stark*, Umesh Padia*, Julia Balla, Cameron Diao

Massachusetts Institute of Technology

TL;DR

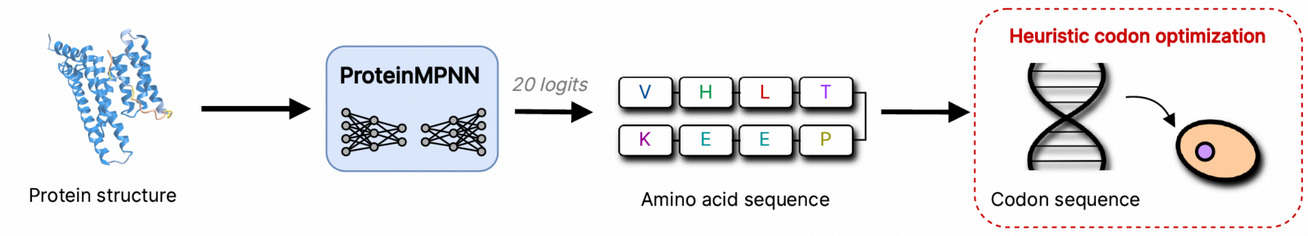
Different codon sequences can encode the same protein but interact differently within a host and have varying expression levels.



1. We propose **CodonMPNN**, which adapts ProteinMPNN [1] to generate codon sequences conditioned on a protein backbone structure and an organism label.
2. CodonMPNN retains ProteinMPNN's performance and recovers wild-type codons more frequently.
3. For the same protein sequence, CodonMPNN assigns higher likelihood to high-fitness codon sequences than low-fitness sequences.

CodonMPNN

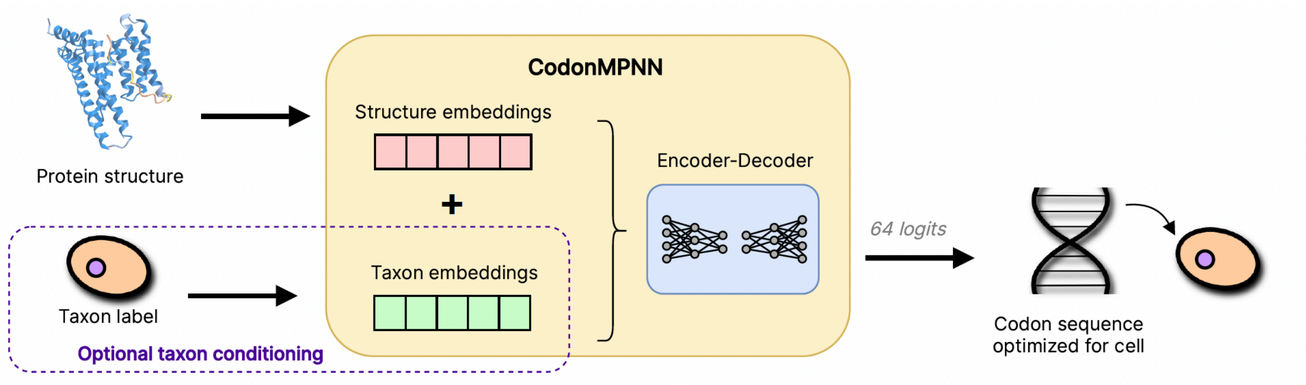
Prevailing approach: An inverse folding model (e.g., ProteinMPNN) generates an amino acid sequence that is mapped to a codon sequence via heuristic optimization.



Our approach: Let $s \in \{1,\ldots,64\}^L$ be a codon sequence and $x \in \mathbb{R}^{L\times 4\times 3}$ a protein structure of 3D coordinates with L residues and their 4 backbone atoms.

We train CodonMPNN to predict $p(s_{\sigma(i)} \mid s_{\sigma(<i)}, x; \sigma)$ for any sequence permutation σ and $i \in \{1,\ldots,64\}$.

We then sample from $p(s) = \Pi_{i=1}^{L} p(s_{\sigma(i)} \mid s_{\sigma(<i)}, x; \sigma)$.

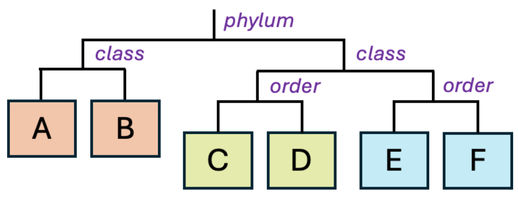


Taxon conditioning

We use the NCBI taxonomy database to group organisms into clusters with common cellular environments.

Balanced tree grouping: Recursively assign nodes to groups, keeping subtrees together without exceeding size $\lceil n/k \rceil$.

Each cluster is assigned a unique taxon label.

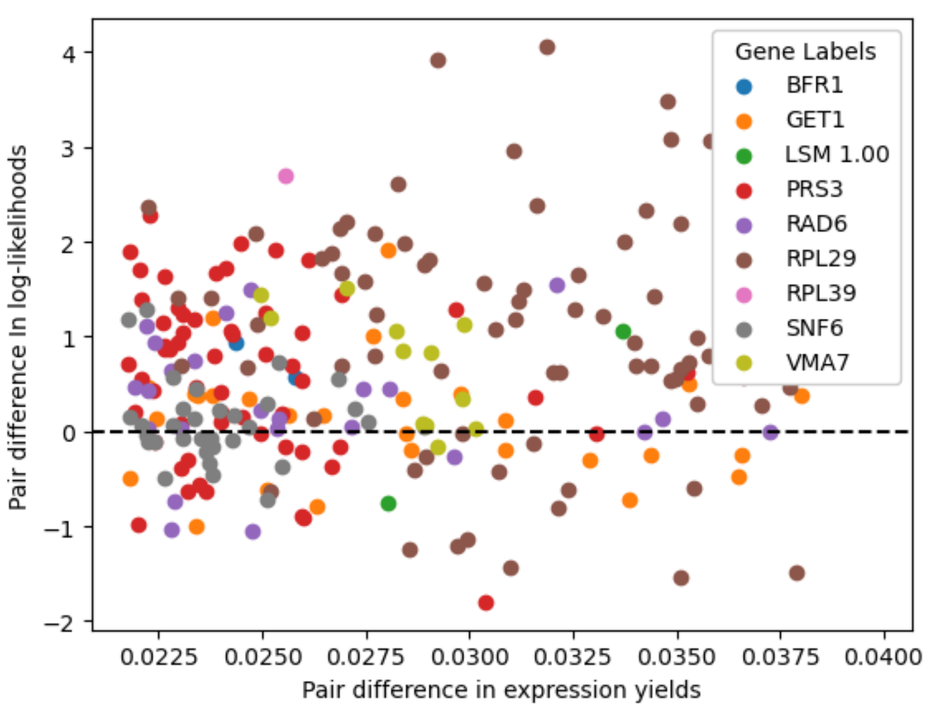


Likelihoods for synonymous coding sequences

We use yeast mutant data from [2] to evaluate CodonMPNN likelihood predictions for 250 significant synonymous mutations.

We predict protein structures for wild-type sequences with AlphaFold 2, which are used for CodonMPNN conditioning.

CodonMPNN correctly predicts higher likelihoods for the more highly expressed codon sequences in 72.4% of the cases (pairs above horizontal line).



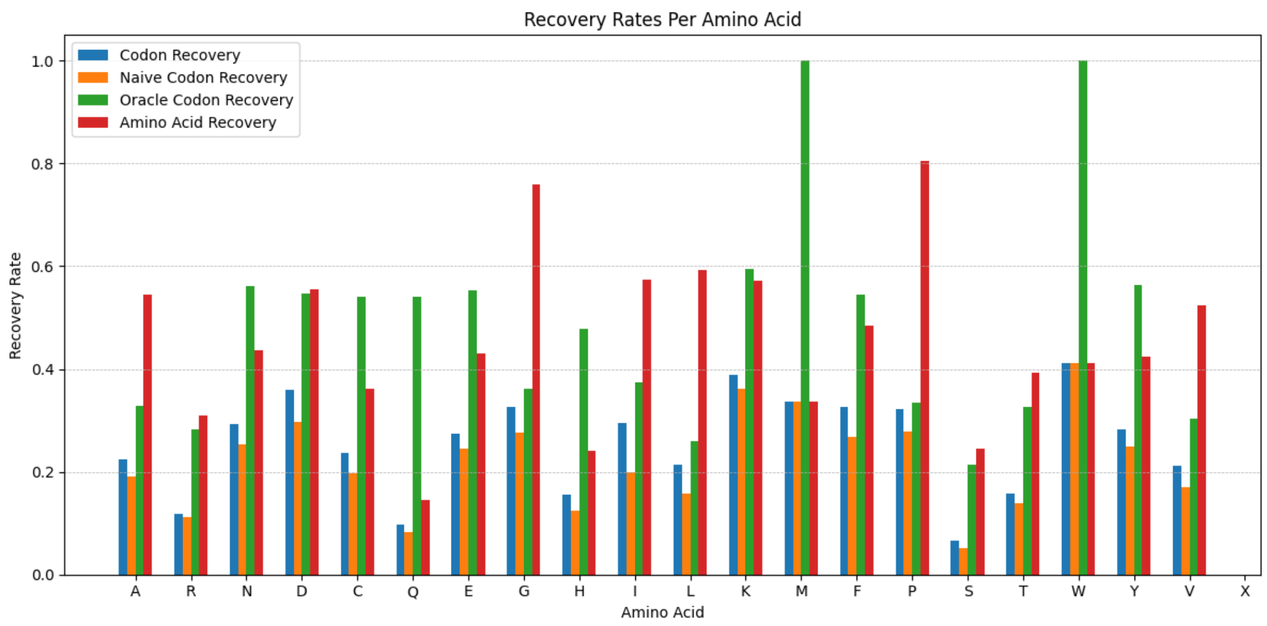
Codon recovery and designability

We train and evaluate CodonMPNN (and ProteinMPNN) on AFDB structures with pLDDT > 0.9.

	Codon %	AA %	TM
ProteinMPNN	20.5%	49.8%	0.83
ProteinMPNN-taxon	20.8%	50.3	0.86
CodonMPNN	24.8%	49.4%	0.84

CodonMPNN achieves the same AA recovery rate and designability (TM-Score) as ProteinMPNN, as well as higher codon recovery rate.

Recovery per amino acid



Codon Recovery and ***AA Recovery*** show the CodonMPNN recovery rates.
Naive Codon Recovery: Codon recovery rate obtained by translating CodonMPNN's codons to AAs by choosing their most frequent codons.
Oracle Codon Recovery: Same but for AAs.

CodonMPNN improves over choosing the most frequent codon per AA for codon recovery.

References

[1] Dauparas J, Anishchenko I, Bennett N, *et al.* Robust deep learning-based protein sequence design using ProteinMPNN. *Science*. 2022; 378(6615):49-56.
[2] Shen, X., Song, S., Li, C. *et al.* Synonymous mutations in representative yeast genes are mostly strongly non-neutral. *Nature.* 2022; 606(7915):725–731.