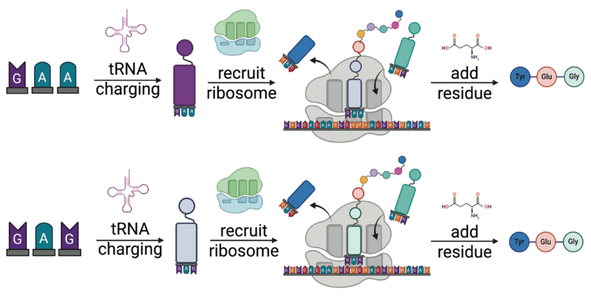
Predicting the evolutionary stability of synonymous mutations

Seh Na Mellick
Ray and Stephanie Lane Computational Biology Department
Carnegie Mellon University

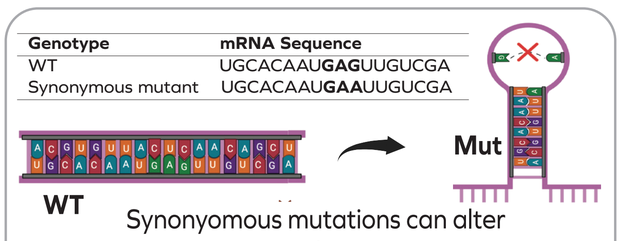
Synonymous ≠ silent

Synonymous mutations are traditionally viewed as neutral because they do not alter protein sequences. However, they can still be deleterious through effects on expression, mRNA structure, or regulation.

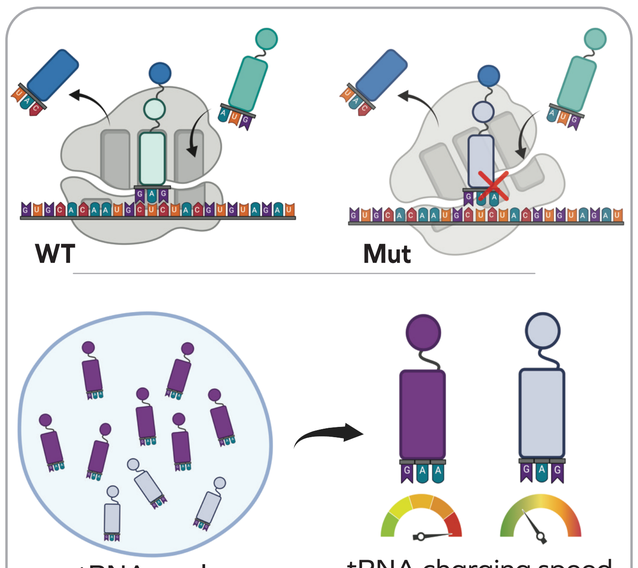


Synonymous mutations use different codons which charge the same amino acid residue.

Genotype	mRNA Sequence
WT	UGCACAAU**GAG**UUGUCGA
Synonymous mutant	UGCACAAU**GAA**UUGUCGA



Synonyomous mutations can alter mRNA secondary structure.



Rare and mismatched tRNAs slow translation

Dataset

To predict the evolutionary fitness effects of synonymous mutations, we used a dataset from Shen et al. 2022 [1] which includes:

(i) ~8,000 single SNPs over 150bp coding regions in 22 nonessential *Saccharomyces cerevisiae* genes
(ii) Relative fitness measurements for each mutation in a pooled competition assay
(iii) Wild-type and mutant codons for each mutation

Genomic, structural, and conservation features

Codon usage bias

$$CAI = (\prod_{i=1}^{L} w_i)^{1/L}$$
$$w_i = \frac{RSCU_i}{max(RSCU_{syn})}$$
L = # codons in gene
$RSCU_i$: Relative Synonymous Codon Usage of codon i

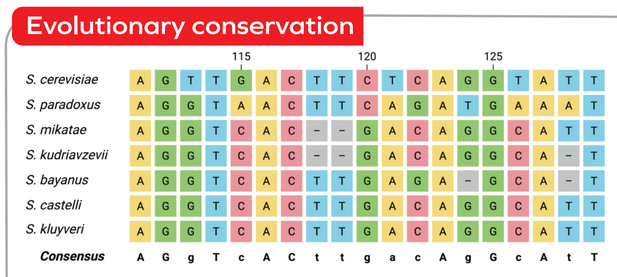
CAI [2] quantifies how well a gene's codon usage matches the preferred codons of highly expressed genes in the organism.

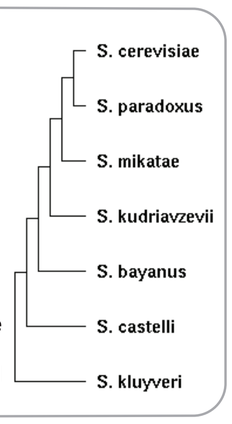
Secondary mRNA structure

$$MFE = min_{\text{all foldings}} \Delta G_{\text{structure}}$$
$\Delta G_{\text{structure}}$: sum of energy contributions from secondary structures

MFE [3] represents the predicted stability of an mRNA's secondary structure, which can affect translation efficiency and the rate of mRNA decay

Evolutionary conservation





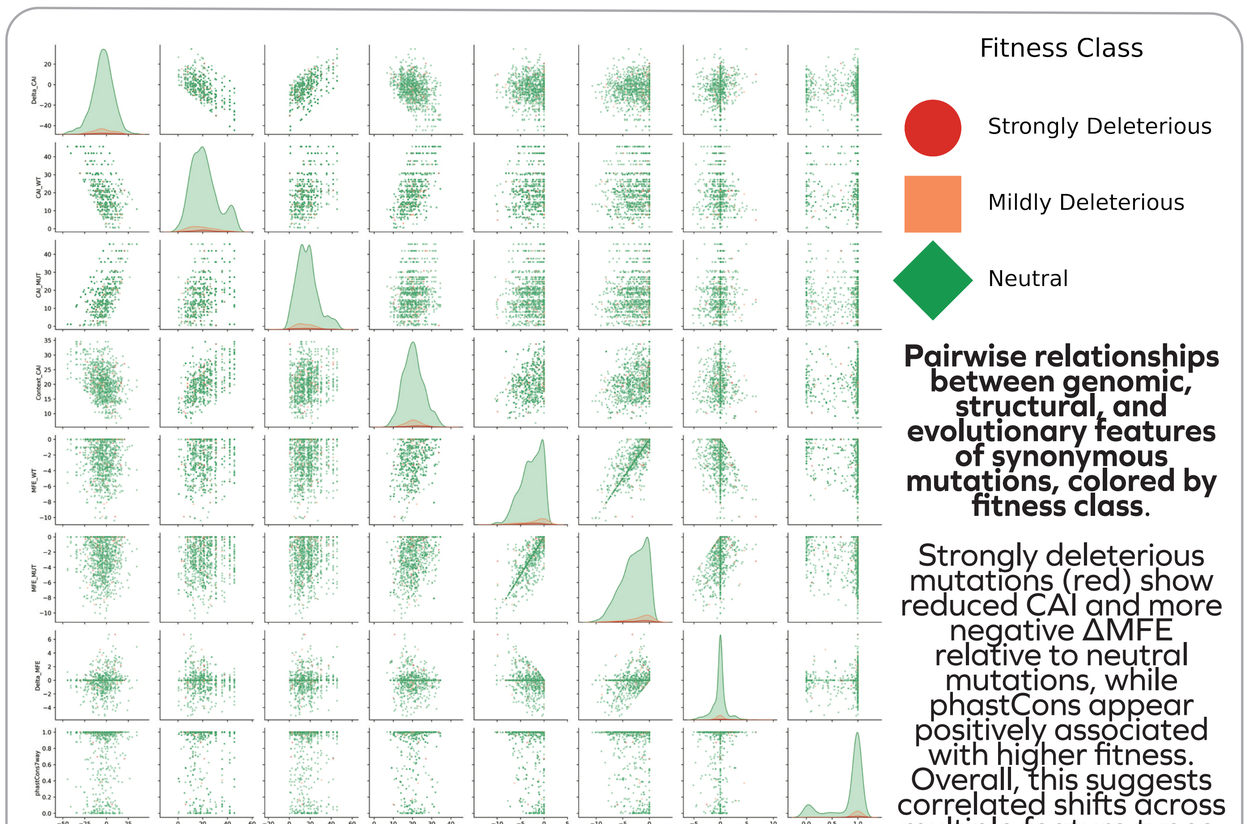
$$\text{phastCons}(x_i) = P(x_i \in \text{Conserved} \mid \text{alignment, phylogeny})$$

phastCons [4] estimate evolutionary conservation at each base in a genome based on a multiple sequence alignment of orthoogous regions across species. It uses a phylogenetic HMM to infer whether a site is conserved or not.

Objectives

We compiled a curated dataset of SNPs from *S. cerevisiae* coding regions, annotated with codon usage bias metrics, mRNA folding energies, and conservation scores. Each mutation is associated with an experimentally measured fitness value from DMS experiments [1].

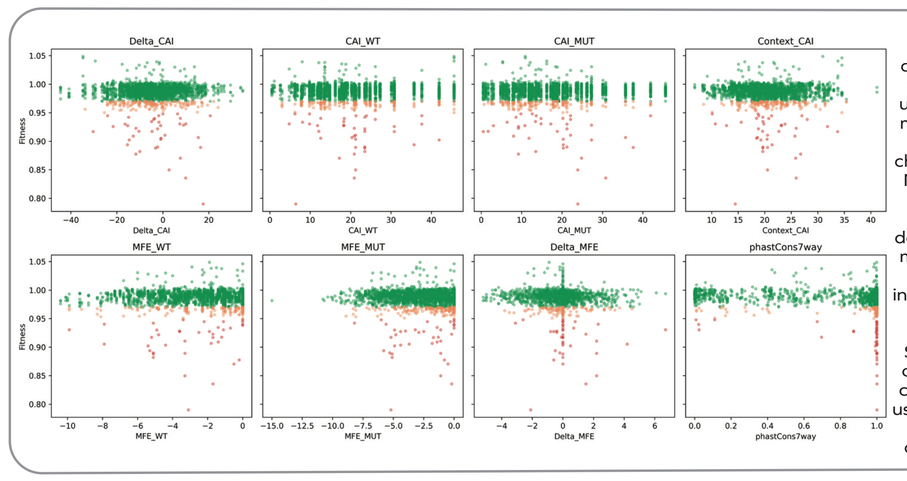
Our objective was twofold: (i) to identify molecular features which best distinguish deleterious from neutral synonymous mutations, and (ii) to classify mutations into evolutionary fitness effect categories- strongly deleterious, mildly deleterious, or neutral.



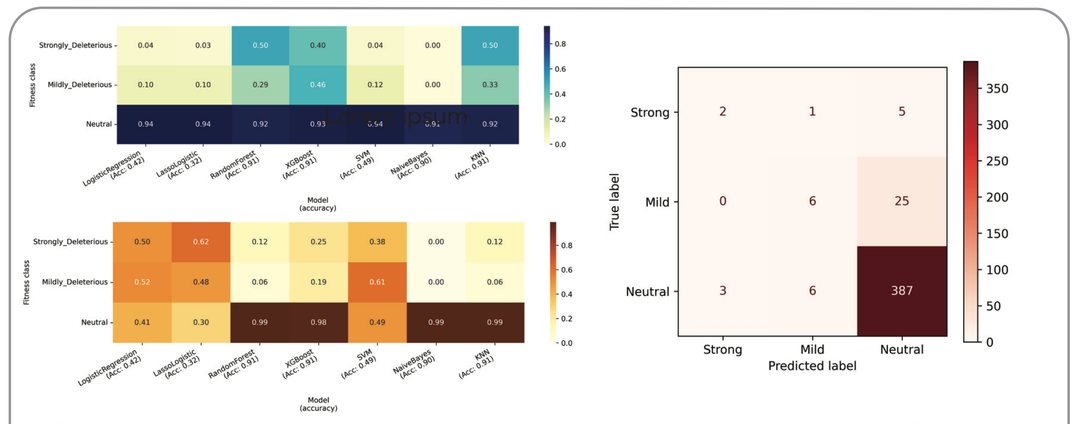
Pairwise relationships between genomic, structural, and evolutionary features of synonymous mutations, colored by fitness class.

Strongly deleterious mutations (red) show reduced CAI and more negative ΔMFE relative to neutral mutations, while phastCons appear positively associated with higher fitness. Overall, this suggests correlated shifts across multiple feature types.

Predicting fitness effects

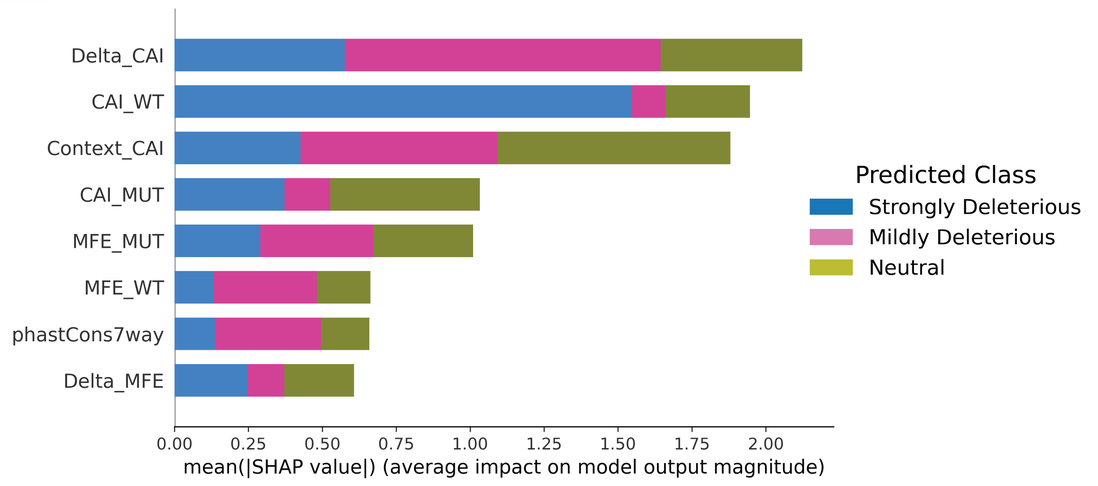


Fitness tends to decrease with less favorable codon usage (lower CAI), more destabilizing mRNA structural changes (positive Δ MFE), and higher conservation (phastCons), demonstrating that multiple molecular features independently track with fitness consequences. SMOTE synthetic oversampling and class weights were used in downstream tasks to combat class imbalances.



Classification performance across models predicting synonymous mutation fitness classes. The precision heatmap (top left) shows high performance for the Neutral class across models, with only Random Forest, XGBoost, and KNN achieving moderate precision for deleterious classes. The recall heatmap (bottom left), reveals that linear models tend to overcall deleterious mutations, while ensemble methods reliably recover Neutral cases. The confusion matrix for XGBoost (right), the top-performing model by accuracy, shows strong Neutral synonymous mutation classification but limited sensitivity to deleterious synonymous mutations.

Feature importance for classification



SHAP [5] summary plot of feature importance for XGBoost classifier of synonymous mutations. Bar length reflects each feature's average contribution to model predictions across all classes. Codon usage metrics (ΔCAI and WT CAI) dominate overall importance. Meanwhile, context-based codon usage and mRNA folding energies contribute more selectively to Mildy and Strongly Deleterious classifications of synonymous mutations.

Discussion

(i) Our work supports the hypothesis that synonymous mutations can produce substantial fitness effects through codon usage, mRNA structural stability, and evolutionary conservation, despite not altering the protein sequence. This further challenges the common assumption that synonymous mutations are functionally silent.

(ii) Our models predict neutral mutations well but struggle with predicting strongly deleterious synonymous mutations.

(iii) Future work could use additional molecular data like ribosome profiling or RNA stability data to better capture post-transcriptional effects and refine predictive resolution for borderline fitness classes.

(iv) The DMS dataset is inherently imbalanced, with most synonymous mutations being neutral and thus limiting classification performance. Future work could explore alternative sampling strategies or loss functions beyond class weighting and SMOTE.

References

1. Shen, X., et al. (2022). Synonymous mutations in representative yeast genes are mostly strongly nonneutral. *Nature, 606*(7915):725-31.
2. Puigbo, P., et al., (2008). Caical: a combined set of tools to assess codon usage adatation. *Biology Direct, 3*(38).
3. Lorenz, R., et al., (2011). Viennarna package 2.0.. *Algorithms for Molecular Biology, 6*(1):1-14.
4. Siepel, A., et al., (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Resesarch, 15*(8):1034-50.
5. Lundberg, S., and Lee, S., (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*.