Why This Code? A Constrained Mapping Framework for the Evolutionary Stability of the Canonical Genetic Code

Rentian Lin[*2], Chenxiao Wang[*1], Yining Hu[*1], Chuanrui Wang[*1], Xinyu Wang[*1]

[1] Westlake University
[2] Nankai University

Abstract

The canonical genetic code is used by most known forms of life, suggesting that its structure may reflect a stable evolutionary solution. Existing theories, including frozen accident, error minimization, stereochemical recognition and coevolution, provide important explanations for its origin and conservation, but the enormous space of possible codon-to-output assignments remains difficult to examine systematically. Here, the code is formulated as a hierarchy of constrained mapping problems spanning codeword length, degeneracy composition, synonymous-block partitioning, semantic assignment and decoder implementation. Exact structural analyses identify triplets as a Pareto choice under a fixed-length full-codebook model and show that anonymous degeneracy statistics alone do not explain the canonical profile. For the fixed canonical block architecture, recurrent rule learning contracts the 20! amino-acid assignment space from 2.43×10^{18} to approximately 10^{11} admissible mappings, from which 10^{8} complete codes are sampled. The standard genetic code ranks in the best 0.1411% under an equal-weight score combining mutation robustness, encoding entropy and local GC stability, and in the best 0.2824% after accessible replacement diversity is added. A constrained decoder-level redesign further increases one-mutation amino-acid accessibility with limited robustness and GC costs. These results support the interpretation of the canonical code as a rare multi-objective compromise within a clearly defined conditional space rather than a universal single-objective optimum.

Keywords: genetic code; codon degeneracy; constrained mapping; mutation robustness; evolutionary stability; synthetic biology

1 Introduction

The standard genetic code is one of the most conserved features of living systems. With limited exceptions in mitochondria and some microbial lineages, nearly all organisms use the same triplet code to translate 64 codons into 20 amino acids and termination signals. This near universality suggests that the code was established early in evolution and subsequently maintained under strong constraints, yet the reasons why this particular codon-to-output mapping became fixed remain debated.

Crick's frozen accident hypothesis argues that, once a coding scheme was adopted by early life, later changes would have caused widespread mistranslation and were therefore strongly selected against[1]. Error-minimization theories instead emphasize the internal organization of the code. Codons that are easily interconverted by point mutation or translational error often encode amino acids with similar physicochemical properties, reducing the functional cost of errors[2,3]. Other models focus on stereochemical affinity between amino acids and cognate codons or anticodons[4], or on coevolution between the genetic code and amino acid biosynthetic pathways[5]. These views are not mutually exclusive, and each captures a plausible component of code evolution. However, none alone establishes whether the present code should be regarded as a historical contingency, an optimized solution, or a compromise shaped by multiple constraints[6].

A central difficulty is the size and structure of the possible code space. If codon assignments are treated as arbitrary mappings from 64 codons to 20 amino acids and stop signals, the number of alternatives is enormous. Moreover, many random codes are biologically implausible because they ignore triplet reading, degeneracy, stop codons, codon-block organization and other structural features of translation. This makes it necessary to compare the standard genetic code not against unconstrained random mappings, but against alternative codes generated under comparable biological and combinatorial constraints.

Here, we formulate the genetic code as a constrained mapping problem from 64 codons to 20 amino acids and one termination signal. The search space is reduced by specifying codon length, degeneracy limits, degeneracy composition, codon-block partitioning and block-to-output assignment. Candidate codes are then evaluated using metrics related to mutational robustness, GC balance and encoding diversity, with *Escherichia coli* MG1655 protein-coding sequences used as a reference biological test set[7,8]. This framework allows the standard genetic code to be examined as a structured reference point within a biologically constrained space, rather than as an isolated historical outcome.

2 Methods

2.1 Hierarchical representation of genetic-code space

The nucleotide alphabet was Σ={A,C,G,U}. A genetic code was represented as Ω=(L,R,H,P,ψ), where L is codeword length, H=(h_1,...,h_{21}) is the multiset of output degeneracies satisfying $\Sigma_i h_i = 4^L$, R=max(H), P={B_1,...,B_{21}} is an anonymous partition of Σ^L with $|B_i|=h_i$, and ψ is a bijection from the 21 anonymous blocks to 20 amino acids plus termination. Analyses were conditional and layer specific: the L analysis used only code capacity and the quaternary Hamming graph; H/R analyses used integer compositions of 4^L into 21 positive parts; P analyses used codon membership and single-nucleotide adjacency without amino-acid labels; and ψ analyses introduced amino-acid semantics only after P had been fixed. This ordering prevents downstream biological information from being used to explain an upstream structural choice. For the principal large-scale experiments, L=3, the 20 canonical sense blocks and the UAA/UAG/UGA termination block were fixed, Stop was not permuted, and ψ ranged over the 20! bijections between the sense blocks and amino-acid labels[9–16].

For the local third-position partition calculation, each of six canonical four-codon boxes containing a 2+2 synonymous split was treated independently. A third-position transition had weight κ and a transversion had weight 1. The cut weight induced by the canonical U/C | A/G split was compared exactly with the two alternative pairings. For the code-length calculation, vertices of the quaternary Hamming graph were all words in Σ^L and edges joined words differing at one position. Surjective 21-way partitions were evaluated by RatioCut; exact edge-isoperimetric profiles and subset dynamic programming were used to minimize the objective over permitted block-size compositions. These structural calculations were kept separate from the fixed-block ψ experiment reported below.

2.2 Direction-neutral recurrent-rule screen and exact candidate generation

Let W_{ij} denote the number of undirected Hamming-distance-one edges joining canonical sense blocks B_i and B_j. Edges internal to a synonymous block contribute zero, and edges incident to the fixed Stop block were excluded because the amino-acid replacement table contains no Stop state. The empirical 20×20 ProteinGym-derived replacement table Cdir was directional[17]. Because the discovery graph did not specify a mutational source distribution, the screen used the direction-neutral cost Csym(a,b)=[Cdir(a→b)+Cdir(b→a)]/2, where a and b denote amino-acid labels. The robustness cost of a complete mapping ψ was

$$J(\psi) = \sum_{i<j} W_{ij} C_{\mathrm{sym}}\big(\psi(B_i), \psi(B_j)\big) \tag{2.1}$$

In equation (2.1), J(ψ) is the direction-neutral robustness cost of mapping ψ; i and j index the 20 fixed sense blocks; W_{ij} is their codon-level adjacency multiplicity; B_i and B_j are the corresponding synonymous blocks; $\psi(B_i)$ and $\psi(B_j)$ are their assigned amino acids; and Csym is the symmetric replacement penalty. Lower J indicates that single-nucleotide-adjacent blocks encode amino acids separated by smaller empirical replacement costs.

Independent uniform permutations were generated by a vectorized Fisher–Yates shuffle. Three independent seed groups were used, each containing three rounds of 10,000,000 complete mappings. In each round, the lowest-cost 10% defined the discovery tail. For every block i and amino-acid label a, relative representation was calculated as

$$r_{ia} = 20\, P[\psi(B_i) = a \mid T] = \frac{P[T \mid \psi(B_i) = a]}{P[T]} \tag{2.2}$$

In equation (2.2), r_{ia} is the relative representation of assigning amino acid a to block B_i in the discovery tail T; P denotes empirical probability; and the factor 20 normalizes the statistic to one under uniform label assignment. A non-canonical assignment was provisionally excluded when $r_{ia}<1$. Fisher's exact test and Benjamini–Hochberg-adjusted q values were recorded diagnostically; the production setting used q≤1 and therefore relied on cross-round recurrence rather than a single-round significance cutoff. A rule had to recur in at least two of the three rounds within a seed group and in all three independent seed groups. Canonical assignments were protected, so the SGC always remained admissible. The final 186 exclusions were compiled into a 20×20 Boolean allowed-assignment matrix.

The number of admissible complete mappings was calculated exactly as the permanent of the allowed matrix using subset dynamic programming. The recurrence was

$$F(0, \varnothing) = 1; \quad F(i+1, S \cup \{a\}) \mathrel{+}= F(i, S) \text{ if } A_{ia} = 1 \text{ and } a \notin S \tag{2.3}$$

In equation (2.3), F(i,S) is the number of partial bijections assigning the first i blocks to the amino-acid-label subset S; A_{ia} is one when assignment of label a to block i is allowed and zero otherwise; ∅ is the empty set; and a∉S enforces

one-to-one assignment. The terminal value F(20,{1,...,20}) gave the exact surviving count. The same completion-count table supported unbiased generation: an integer rank was drawn uniformly from [0,F−1] and recursively unranked by subtracting the completion counts associated with each legal next label. This produced 100,000,000 complete mappings directly from the surviving order-10^{11} space without rejection sampling. The rules are empirical proposal constraints, not formal proofs that every excluded completion is biologically inviable.

2.3 Three-objective scoring: mutation robustness, encoding entropy and local GC stability

The principal landscape used three independently calculated objectives. Direction-neutral mutation robustness was the cost J defined above, with lower values favourable. To calculate encoding entropy, amino-acid frequencies p_a were obtained by pooling all accepted residues in the *Escherichia coli* K-12 MG1655 reference proteome. If d_b is the number of synonymous codons in fixed block B_b and $b\psi(a)$ is the block assigned to amino acid a, synonymous codons were assumed to be equiprobable and the expected sequence-choice capacity was[18–19]

$$H_{\mathrm{enc}}(\psi) = \sum_a p_a \log_2 d_{b_\psi(a)} \quad \text{bits per residue} \tag{3.1}$$

In equation (3.1), Henc(ψ) is the expected encoding entropy per residue under mapping ψ; a indexes amino-acid identities; p_a is the observed proteome frequency of amino acid a; $b\psi(a)$ identifies the fixed block assigned to a by ψ; and d_{bψ(a)} is that block's number of synonymous codons. Thus, an occurrence of amino acid a contributes $\log_2 d$ bits when it is assigned to a block containing d synonymous codons, and the proteome-wide score is the amino-acid-frequency-weighted mean of these contributions. Higher Henc indicates a larger expected number of synonymous nucleotide sequences capable of encoding the same reference proteome under the uniform-codon assumption; it does not measure the realized entropy of native codon usage.

Local GC stability was computed analytically rather than by sampling reverse-translated sequences. For every fixed block b, μ_b and v_b were the mean and variance of the number of G/C nucleotides among its synonymous codons under uniform use. For each overlapping 20-residue window (60 nucleotides) with amino-acid count vector x, the expected squared deviation of window GC fraction G/(3w) from 0.5 was

$$\mathbb{E}\left[\left(\frac{G}{3w} - 0.5\right)^2\right] = \frac{x^{\mathsf{T}} v + (x^{\mathsf{T}}\mu - 1.5w)^2}{9w^2}, \quad w = 20 \tag{3.2}$$

In equation (3.2), E denotes expectation over uniform synonymous-codon choices; G is the random number of G or C nucleotides in the window; w is the number of amino-acid residues per window and was fixed at 20; x is the 20-component amino-acid count vector of the window; μ and v are vectors containing, for the block assigned to each amino acid, the mean and variance of its per-codon G/C count; and the superscript T denotes vector transpose. All codon-aligned windows were included with a one-codon stride; windows were averaged equally within each eligible protein and proteins were then averaged equally. Lower GC mean-squared error was favourable. This score measures intrinsic local composition under a uniform synonymous model and does not fit candidates to native codon bias, transcription or ribonucleic acid (RNA) folding.

Raw objectives were placed on a common dimensionless scale using empirical midranks in the complete supplied 100-million-code library. For a lower-is-better objective with candidate value x, empirical loss was

$$\ell(x) = \frac{n(y < x) + 0.5\, n(y = x)}{N} \tag{3.3}$$

In equation (3.3), ℓ(x) is the loss percentile of candidate value x; y denotes values of the same objective among the N supplied candidate codes; and n(·) counts candidates satisfying the stated relation. For a higher-is-better objective, y<x was replaced by y>x. Utility was then defined as

$$U(x) = 1 - \ell(x) \tag{3.4}$$

In equation (3.4), U(x) is the empirical utility of value x, with zero worst and one best in the supplied library. The equal-weight three-objective composite was

$$Q_3 = \frac{U_{\mathrm{robustness}} + U_{\mathrm{entropy}} + U_{\mathrm{GC}}}{3} \tag{3.5}$$

In equation (3.5), Q_3 is the composite three-objective quality; Urobustness, Uentropy and UGC are the utilities of mutation robustness, encoding entropy and local GC stability, respectively. The SGC was evaluated explicitly against the same empirical distributions. For a candidate permutation π relative to the SGC permutation πSGC, the minimum label-transposition distance was

$$d(\pi, \pi_{\mathrm{SGC}}) = 20 - \mathrm{cycles}(\pi_{\mathrm{SGC}}^{-1}\pi) \tag{3.6}$$

In equation (3.6), d is the minimum number of pairwise label exchanges; π is the candidate block-label permutation; πSGC is the SGC permutation; $\pi\mathrm{SGC}^{-1}\pi$ is their relative permutation; and cycles(·) counts its disjoint cycles, including fixed points. At each integer distance d, enrichment of globally top-1% codes was

$$E_d = \frac{P(Q_3 \text{ in global top } 1\% \mid d)}{0.01} \tag{3.7}$$

In equation (3.7), E_d is the fold enrichment at distance d and P denotes the within-distance empirical probability. Only strata with at least 1,000 sampled codes were displayed, and uncertainty was reported as a two-sided 95% Wilson interval for the within-stratum binomial proportion. Equal weighting was used as a transparent descriptive scalarization; Pareto comparisons were retained as a weight-free sensitivity analysis[20–21].

2.4 Accessible replacement diversity and constrained decoder-unit redesign

Accessible replacement diversity was calculated for each fixed-block mapping as follows. A source amino acid was drawn uniformly, one of its synonymous codons was drawn uniformly, and one of the nine one-nucleotide substitutions was drawn uniformly. After global conditioning on a sense missense outcome, two independent target amino acids B_1 and B_2 were drawn from the conditional distribution for the same source. Their separation was scored with the unmodified directional ProteinGym-derived cost table:

$$D(\psi) = \mathbb{E}[C_{\mathrm{dir}}(B_1 \to B_2)] \tag{4.1}$$

In equation (4.1), D(ψ) is accessible replacement diversity under mapping ψ; E denotes expectation over the declared source-codon-mutation sampling process conditional on a sense missense event; B_1 and B_2 are independent target amino acids drawn for the same source amino acid; and Cdir($B_1 \to B_2$) is the directional ProteinGym-derived replacement cost from B_1 to B_2. Because B_1 and B_2 are identically distributed, ordered target pairs occur with equal probability in both orientations, so the expectation is insensitive to the antisymmetric component of the table even though the original directional entries were retained. Higher D indicates broader accessible outcomes in empirical replacement-cost space; it is not a measured probability of beneficial mutation. Four-objective quality was

$$Q_4 = \frac{U_{\mathrm{robustness}} + U_{\mathrm{entropy}} + U_{\mathrm{GC}} + U_{\mathrm{diversity}}}{4} \tag{4.2}$$

In equation (4.2), Q_4 is the four-objective composite and Udiversity is the empirical utility of accessible replacement diversity; the other utilities are defined in equation (3.5). Q_4 was analysed with the same empirical-rank procedure[17,22].

Fixed-block relabeling leaves the raw number of distinct SNV target blocks invariant, so target-richness optimization was performed in a larger decoder-unit space. The 61 sense codons were grouped by SGC amino-acid identity and first-two-base box. Within each group, U/C and A/G third-position pairs were treated as atomic units when both codons were present; remaining codons were singletons. This coarse wobble model yielded 32 units (29 pairs and three singletons). It is a transparent mechanistic constraint, not a reconstruction of the complete modified-tRNA repertoire of MG1655[11,13–14,23]. AUG remained a singleton fixed to Met, the three Stops remained fixed, and proposals exchanged amino-acid identities only between decoder units of equal size, thereby preserving amino-acid degeneracies and preventing an atomic unit from being split.

For a candidate decoder state, accessible richness was

$$R_{\mathrm{acc}}(\psi) = \frac{1}{61} \sum_{c \in C_{\mathrm{sense}}} |A_{\mathrm{alt}}(c; \psi)| \tag{4.3}$$

In equation (4.3), Racc(ψ) is mean accessible richness; Csense is the set of 61 sense codons; c is a source codon; Aalt(c;ψ) is the set of distinct alternative amino-acid labels reached from c through sense one-nucleotide neighbours

under decoder state ψ; and $|\cdot|$ denotes set cardinality. Directional missense cost assigned equal prior weight to each source amino acid and equal weight to all directed sense-to-sense one-base edges originating from codons assigned to that source. The 60-nucleotide GC score was recalculated from the candidate's dynamic synonymous sets using the same sufficient-statistic expression above. Feasibility required

$$\frac{J_{\mathrm{dir}}(\psi)}{J_{\mathrm{dir}}(\mathrm{SGC})} \leq 1.05; \qquad \frac{\mathrm{GC}_{\mathrm{local}}(\psi)}{\mathrm{GC}_{\mathrm{local}}(\mathrm{SGC})} \leq 1.10 \tag{4.4}$$

In equation (4.4), Jdir is the directional missense cost and GClocal is the 60-nucleotide local GC-stability cost; each ratio compares a candidate decoder state ψ with the SGC. Engineering burden was measured by the minimum number of nucleotide substitutions required to recode the MG1655 reference coding sequence (CDS) collection:

$$B_{\mathrm{edit}}(\psi) = \sum_{c \in C_{\mathrm{changed}}} n_c h_{\mathrm{min}}(c, \psi) \tag{4.5}$$

In equation (4.5), Bedit(ψ) is total recoding burden; Cchanged is the set of codon types whose amino-acid meaning changes; n_c is the genomic occurrence count of codon c in the reference CDS collection; and hmin(c,ψ) is the smallest nucleotide Hamming distance from c to any codon that retains c's original amino-acid meaning under ψ.

All decoder-unit exchanges within change radius two were enumerated exactly. For radii 4, 6, 8, 10 and 12, simulated annealing used eight independent restarts of 200,000 proposal steps, with seed 20260803. The search score maximized richness while subtracting the minimum substitution fraction and large penalties for constraint violations. A memory-bounded archive retained at most 200 unique feasible states per search stratum. The final Pareto archive was computed across feasible nonredundant designs by maximizing richness and minimizing changed decoder units, changed codon types, minimum nucleotide substitutions, directional missense cost and GC cost[20–21]. Reported designs are therefore feasible solutions found by the declared search, not certificates of global optimality.

3 Results

3.1 Genetic-code space decomposes into nested, non-interchangeable structural layers

A genetic code written directly as an arbitrary function from 64 codons to 20 amino acids and one termination signal has a nominal search space of 21^{64}. Although this expression conveys the scale of the problem, it conflates biologically distinct questions: why codewords have length three, how 64 codons are distributed among 21 outputs, why synonymous codons form blocks of their observed sizes and shapes, and why those blocks carry their present amino-acid meanings. We therefore represented a genetic code over the nucleotide alphabet Σ={A,C,G,U} as a sequence of nested variables:

$$\Omega = (L, R, H, P, \psi) \tag{1.1}$$

Here, L is the codeword length; H=(h_1,...,h_{21}) is the multiset of degeneracies assigned to the 21 outputs, with $\Sigma_i h_i = 4^L$; R=$\max_i h_i$ is the maximum degeneracy; P={B_1,...,B_{21}} is an anonymous partition of Σ^L satisfying $|B_i| = h_i$; and ψ:P→$\mathcal{A}$ assigns the anonymous blocks to 20 amino acids and a termination signal. The resulting dependency order is:

$$L \rightarrow H\ (R = \max H) \rightarrow P \rightarrow \psi \tag{1.2}$$

These variables are not independent parameters. Fixing one layer restricts the next layer to a conditional subspace: a single degeneracy profile H can support many codon partitions P, and a single partition can support many semantic assignments ψ. This decomposition separates code capacity, redundancy allocation, synonymous-block geometry and amino-acid semantics into distinct, testable problems.

The decomposition also determines which evidence is meaningful at each layer. Amino-acid properties, codon adjacency, transfer ribonucleic acid (tRNA) decoding and protein sequences are undefined at the L layer, which can therefore be evaluated only through capacity, length cost, information efficiency and the optimal mutation-exposure envelope attainable after partitioning[9,18]. At the H/R layer, codons have not yet been assigned to specific blocks, so only the composition of the 21 block sizes and the downstream partitions permitted by that composition can be compared. Entropy, the Gini coefficient or maximum degeneracy cannot distinguish different geometric arrangements of blocks with identical sizes. At the P layer, the codon members of each block are specified but amino-acid labels remain absent; single-nucleotide adjacency, translational confusion and decoder coverage can therefore be evaluated, whereas amino-acid replacement damage cannot. Amino-acid replacement costs, proteome composition and

semantic rewiring become well-defined only at the ψ layer. Thus, different structural features of the genetic code require different randomization spaces, and every result must specify both the layer being tested and the downstream structure held fixed.

At the L layer, we fixed a four-letter alphabet, 21 equally weighted outputs, a uniform codeword length and use of all 4^L words, without invoking H, P or ψ. The capacity constraint is:

$$4^L \geq 21 \tag{1.3}$$

The 16 words available at L=2 cannot cover 21 outputs, whereas L=3 is the first feasible length. This capacity bound establishes only that triplets are sufficiently short; it does not exclude the possibility that longer codewords exploit additional redundancy to obtain greater robustness. We therefore constructed the quaternary Hamming graph on Σ^L and minimized the 21-way RatioCut over all surjective 21-block partitions[9]. Exact edge-isoperimetric profiles, dynamic programming over block-size compositions and blockwise reachability audits gave the same global minimum for L=3, 4, 5 and 6:

$$\mathrm{RatioCut}_L^* = 146 \tag{1.4}$$

The corresponding optimal macro-averaged robustness was:

$$\rho_L^* = 1 - \frac{146}{63L} \tag{1.5}$$

After correction for the number of nucleotide-level mutation opportunities, the effective mutation exposure was:

$$B_L^* = L(1 - \rho_L^*) = \frac{146}{63} = 2.317460 \tag{1.6}$$

This value remained unchanged from L=3 to L=6. Longer codewords increased the conditional probability that a mutation remained within the original block, but did not reduce optimal output-changing exposure after accounting for the additional mutable nucleotide positions. At the same time, sequence-length cost increased from 1 at L=3 to 2 at L=6 when measured relative to triplets, whereas information efficiency was:

$$\eta_L = \frac{\log_2(21)}{2L} \tag{1.7}$$

Information efficiency decreased from 0.7321 to 0.3660. Under this fixed-length, full-codebook and equal-weight single-nucleotide model, L=3 therefore achieved the same optimal effective mutation exposure as longer codewords while minimizing nucleotide cost and maximizing information efficiency. Triplets consequently formed a strict Pareto choice within this model. This result applies to the fixed-length total-map branch; sparse codebooks, variable-length codes and systems that require external synchronization occupy different candidate spaces.

After fixing L=3, 64 codons must be distributed among 21 non-empty blocks. If amino-acid identities are ignored and block sizes are sorted in nondecreasing order, every possible H corresponds to a restricted integer partition of 64 into 21 positive parts[10]. Exact enumeration yielded 59,755 anonymous degeneracy profiles. Including the termination block, the standard genetic code (SGC) has the profile:

$$H_{\mathrm{SGC}} = (1^2, 2^9, 3^2, 4^5, 6^3) \tag{1.8}$$

This profile contains two one-codon blocks, nine two-codon blocks, two three-codon blocks, five four-codon blocks and three six-codon blocks, giving R_{SGC}=6. Although it lies between uniform and extremely concentrated allocations, that intermediate position did not distinguish it from the 59,755 candidates. For every profile, we calculated anonymous summary descriptors including maximum degeneracy, Shannon entropy, the Gini coefficient, the Herfindahl-Hirschman index and collision probability. Under the pilot descriptor set, 50 noncanonical profiles Pareto-dominated H_{SGC}. The canonical profile was therefore neither the most uniform nor the entropy-maximizing or least-concentrated allocation; R=6 is a coarse projection of H rather than an independent explanatory variable.

Natural code variation further exposed the information boundary of H. In Euplotes, reassignment of UGA from termination to Cys expands the Cys block from two to three codons and contracts the Stop block from three to two[24]. The two block sizes are exchanged, leaving the sorted multiset H unchanged even though both P and ψ have changed. Any statistic that depends only on H is blind to this natural recoding event. Consequently, the biological value of H

must be evaluated through the downstream partitions, decoders and semantic assignments that it can support. The exact enumeration therefore establishes a limitation rather than deriving the canonical profile from an anonymous balance principle: block-size statistics alone do not explain canonical degeneracy, and additional constraints must enter at the P and decoder-implementation layers.

At the P layer, the question changes from how many codons occupy each block to which codons belong to the same block. This analysis requires three distinct structures: the mutation graph generated by single-nucleotide substitutions, the translational-confusion graph generated by near-cognate tRNA misreading, and the decoding hypergraph in which one decoder may recognize multiple codons[11–14]. Because these structures have different edge weights and biological origins, they cannot be merged a priori into a single unweighted Hamming graph. We first constructed an exactly enumerable local P space from the six canonical family boxes with strict 2+2 splitting at the third codon position[13–14]. Each four-codon box admits three perfect matchings that divide the four third-position nucleotides into two doublets, giving:

$$3^6 = 729 \text{ combinations} \tag{1.9}$$

Let transition edges have weight κ and transversion edges have weight 1. The canonical pyrimidine/purine split U/C | A/G has cross-block cut 4 in each box, whereas either alternative pairing has cut $2\kappa+2$. When $\kappa>1$, the canonical U/C | A/G split uniquely minimizes mutation flow across synonymous blocks within this local subspace. When $\kappa=1$, all three pairings have equal cost and all 729 six-box combinations are tied. The standard doublet orientation at the third codon position is therefore not selected by the premise that the third position is intrinsically less important; its local advantage emerges from the interaction between substitution anisotropy and block geometry. An unweighted Hamming graph cannot select the canonical orientation, whereas unequal transition and transversion weights break the symmetry. Because this calculation covers only six 2+2 family boxes, it provides exact local evidence for a P-layer mechanism rather than a global optimality proof over all partitions of 64 codons.

Having separated L, H and P conceptually, the large-scale analysis fixed L=3, the 20 canonical sense-codon blocks and the positions of the three termination codons, and varied only the bijection ψ between sense blocks and amino-acid labels, matching the fixed-block randomization space used in prior error-minimization analyses[15–16]. Because the termination block was excluded from amino-acid label permutation, the conditional candidate space was exactly:

$$|\Psi| = 20! = 2{,}432{,}902{,}008{,}176{,}640{,}000 \tag{1.10}$$

Each candidate in this space is not an arbitrary 64-to-21 function. It is a complete genetic code that preserves the canonical degeneracy composition and synonymous-block geometry while reassigning the meanings of the 20 amino acids. All subsequent large-scale results on rule learning, space contraction, mutational robustness, encoding entropy, local guanine–cytosine (GC) stability and the rank of the SGC therefore answer a strictly conditional question: given that the canonical triplet-block architecture has already formed, does the present amino-acid assignment occupy a favorable region? These results do not constitute a proof over the complete space of L, H and P combinations.

3.2 A recurrent-rule screening algorithm reduces the fixed-block assignment space from 10e18 to 10e11 mappings

Conditioning on the canonical synonymous blocks reduced the original mapping problem to a precisely defined space, but the resulting 20! assignments, of order 10^{18}, remained too numerous for exhaustive downstream evaluation. We therefore developed a recurrent-rule screening and compilation algorithm that learns which local block–amino-acid assignments are consistently incompatible with the high-robustness region, converts those observations into generative constraints and counts the resulting global permutation space exactly. The purpose of the algorithm was not to identify a single optimum or to classify every mapping individually. Instead, it was designed to remove recurrently unfavourable local assignments before expensive biological scoring while preserving a broad population of complete candidate codes.

The discovery stage began with independent uniform samples from the unconstrained 20! space. Each complete mapping was scored with a direction-neutral robustness objective in which the canonical codon blocks were connected by their undirected single-nucleotide Hamming adjacency and neighbouring amino-acid labels were penalized by a symmetric replacement cost. This choice deliberately avoided assuming which amino acid was more likely to be the mutational source during proposal generation. Within each sample, the lowest-cost 10% was treated as a high-robustness discovery tail. This fraction defined a comparative training stratum rather than a biological threshold: candidates outside the tail were not declared nonfunctional, and the screen asked only which local assignments occurred less often than expected among the relatively robust codes.

For each of the 400 possible codon-block–amino-acid pairings, the algorithm compared its prevalence in the discovery tail with its prevalence under uniform assignment. A pairing was provisionally marked as disfavoured when its conditional representation in the tail fell below the uniform expectation. To prevent sampling noise from becoming a permanent constraint, rule discovery was repeated in a nested recurrence design: three independent samples formed each seed group, a provisional rule had to recur in at least two samples within a group, and the resulting group-level rule then had to recur across all three seed groups. Canonical block–amino-acid assignments were protected throughout, ensuring that the SGC remained inside the admissible space. This stabilization stage converted noisy sample-level associations into a compact set of 186 recurrent local exclusions.

The stabilized exclusions were next compiled into a 20×20 Boolean allowed-assignment matrix. A complete candidate code was admissible only if it corresponded to a perfect one-to-one assignment through this matrix. Exact subset dynamic programming over all label subsets showed that the admissible space had contracted from order 10^{18} to order 10^{11}. This reduction was not imposed as a target and was not inferred by multiplying the nominal coverage of individual rules. It emerged from the combinatorial amplification of a small number of local exclusions: forbidding a single block–label pairing simultaneously removes every full permutation that contains it, while overlapping exclusions further restrict which global bijections can be completed. The exact count therefore demonstrated that recurrent local information was sufficient to remove more than six orders of magnitude without collapsing the search to a small set of near-identical solutions.

The same dynamic-programming table was then used as a generative index. Integer ranks sampled uniformly from the exact surviving space were unranked into complete permitted mappings, allowing 10^8 candidate genetic codes to be generated directly without rejection sampling or repeated testing against the rule list. The algorithm therefore produced three linked outputs: an interpretable set of recurrent local exclusions, an exact measure of the remaining global design space and a scalable generator for downstream multi-objective experiments. All subsequent large-scale comparisons of the SGC were performed on this direction-neutral-screened library. The exclusions remain empirical proposal constraints rather than logical certificates of biological impossibility, and later ranks are conditional on the robustness-enriched library; nevertheless, the result establishes that the fixed-block assignment space can be reduced rationally, reproducibly and by several orders of magnitude before detailed sequence-level evaluation.

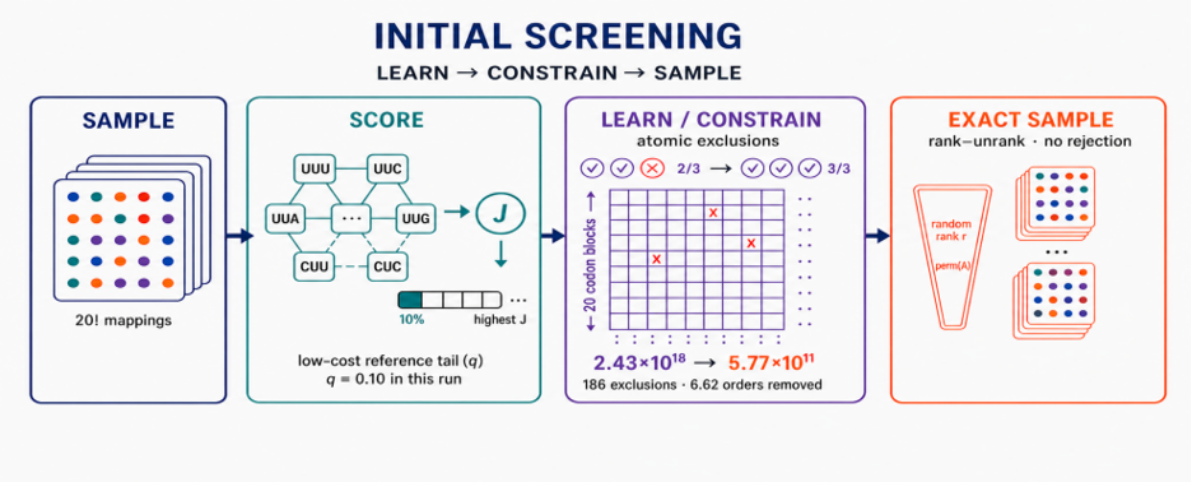

Figure 1 | Recurrent-rule screening and exact sampling of the fixed-block genetic-code space. Complete amino-acid label permutations were sampled uniformly from the 20! fixed-block mapping space and scored using the direction-neutral mutational-robustness cost J on the canonical codon-block Hamming graph. The lowest-cost 10% of each sampled batch defined a reference tail for identifying block–amino-acid assignments depleted among comparatively robust codes. Assignment-level exclusions were retained only when they recurred in at least two of three sampling rounds within each seed group and across all three independent seed groups. The resulting 186 atomic exclusions were compiled into a Boolean allowed-assignment matrix, reducing the admissible space from 2.43×10^{18} to approximately 10^{11} complete mappings. Dynamic-programming completion counts were then used to rank and unrank mappings, enabling uniform, rejection-free sampling directly from the constrained space.

3.3 The SGC occupies a rare compromise across three independently defined code-level objectives

We next examined whether the canonical amino-acid assignment ψ_{SGC} was jointly favoured by three code-level objectives that were defined without using the SGC assignment itself as a fitting target. Mutation robustness combined the codon single-nucleotide neighbourhood with an independently derived amino-acid replacement-cost matrix assembled from ProteinGym-scale mutational measurements; encoding entropy quantified the synonymous sequence capacity generated by the degeneracy assigned to each amino acid; and intrinsic local GC stability measured the

expected fluctuation of GC content around 0.5 in 60-nucleotide coding windows under uniform synonymous-codon usage[3,6,17,19,25]. Although the latter two objectives used the amino-acid composition of the MG1655 proteome, and all candidates retained the canonical synonymous-block partition, none of the three objectives directly rewarded similarity to the SGC block-to-amino-acid mapping.

The SGC showed a strongly asymmetric but complementary performance profile across the three objectives (Figure 2a). It ranked within approximately the best 0.57% of candidates for mutation robustness and the best 0.0205% for encoding entropy, whereas its local GC-stability performance was more moderate, lying within the best 22.7%. Thus, the SGC was not uniformly optimal across all individual objectives. Instead, it combined near-extreme performance in robustness and encoding capacity with a non-extreme but biologically acceptable level of local nucleotide-composition stability. This pattern is more consistent with a compromise among distinct constraints than with optimization of a single property.

To quantify this joint performance, each raw score was converted into an empirical-rank utility, with zero representing the worst and one the best value within the conditional candidate library. The three utilities were then combined with equal weights. Under this composite score, the SGC ranked 141,070th among 10^8 evaluated mappings, placing it within the best 0.1411% of the screened library (Figure 2b). Its position therefore cannot be attributed solely to exceptional performance on one objective: few candidate mappings simultaneously retained high robustness, high synonymous encoding capacity and acceptable local GC stability. Because the candidate library had already been contracted using a direction-neutral robustness screen, this percentile is conditional on the retained fixed-block mapping space rather than an unbiased estimate over all 20! assignments.

The spatial organization of high-quality codes provided an additional observation. Here, distance from the SGC was defined as the minimum number of pairwise exchanges of amino-acid labels between complete synonymous codon blocks required to transform a candidate code into the SGC. For example, exchanging the meanings assigned to the Gly and Pro blocks gives distance one, whereas rotating three amino-acid labels among three blocks requires two exchanges. This is therefore a combinatorial distance between code assignments, not a nucleotide mutation count, biochemical evolutionary distance or elapsed evolutionary time. At each distance, we calculated the enrichment of globally top-1% codes relative to the 1% background expectation. High-quality codes were enriched several-fold in the well-sampled strata closest to the SGC, and this enrichment progressively declined towards and eventually below the global expectation as assignment distance increased (Figure 2c). The result shows that high-performing mappings are preferentially concentrated around the SGC in the fixed-block assignment space. It does not, by itself, demonstrate that evolution followed these swap paths or that the SGC is the centre of a dynamical attraction basin.

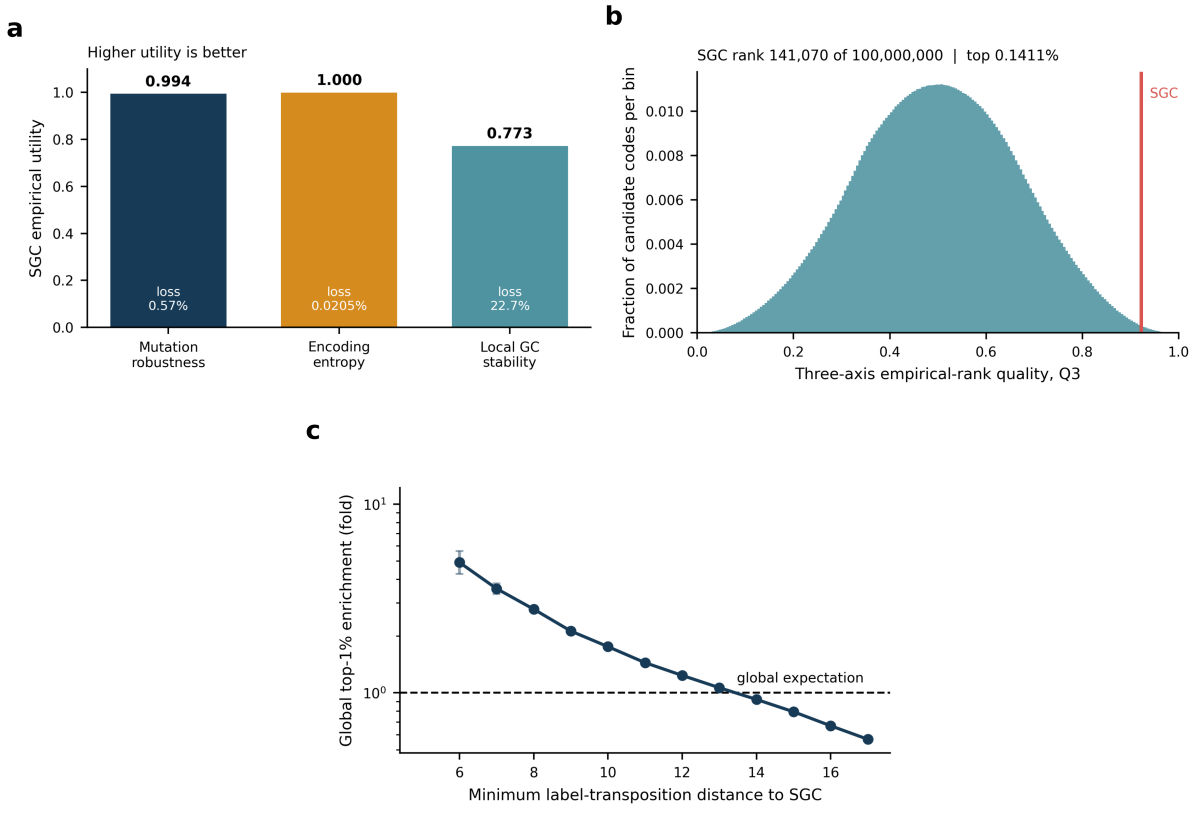

Figure 2 | The SGC occupies a rare and locally enriched region of the three-objective fixed-block landscape. (a)

Empirical utilities of the SGC for mutation robustness, encoding entropy and local GC stability; higher values indicate better performance within the direction-neutral-screened 100-million-code library. (b) Distribution of the equal-weight score Q3=(Urobustness+Uentropy+UGC)/3; the red line marks the SGC (rank 141,070; top 0.1411%). (c) Enrichment of global top-1% Q3 codes as a function of minimum block-label transposition distance from the SGC. Enrichment is the within-distance top-1% fraction divided by the 1% global expectation. Only distance strata containing at least 1,000 sampled codes are shown; error bars are binomial 95% confidence intervals.

Together, these results suggest a plausible explanation for the long-term stability of the canonical assignment. The SGC places empirically less disruptive amino-acid replacements near one another on the codon mutation graph, allocates large synonymous blocks to amino acids that contribute strongly to proteome-wide sequence demand, and avoids severe local GC-composition instability. Once such a mapping had become established, most large-scale block reassignments would be expected to sacrifice at least one of these properties. Subsequent coadaptation of tRNAs, aminoacyl-tRNA synthetases, coding sequences and translational regulation could then further increase the cost of leaving this high-performance region[6,19,25]. Our analysis therefore supports multi-objective stabilization followed by biological lock-in as a possible explanation for the present code, while not distinguishing this scenario from historical contingency or proving that the SGC is a global optimum.

3.4 Accessible replacement diversity exposes an evolvability trade-off and enables constrained decoder-level redesign

We next introduced accessible replacement diversity as a fourth objective. Unlike the three conservative objectives above, this metric did not quantify preservation of the current protein-coding system. Instead, it measured how broadly the sense-missense outcomes accessible from the same source amino acid were distributed in the empirical amino-acid replacement-cost space. It therefore represented an evolvability-like design axis without assuming that any particular accessible mutation was beneficial, extending earlier genetic-code redesign work that explicitly maximized the amino-acid outcomes accessible to single-nucleotide changes[22].

Adding this objective changed the interpretation of the SGC rather than overturning its overall high performance. The SGC retained empirical utilities of 0.994 for mutation robustness, approximately 1.000 for encoding entropy and 0.773 for local GC stability, but its accessible-replacement-diversity utility was only 0.551 (Figure 3a). Correspondingly, its equal-weight composite position shifted from the best 0.1411% under the three-objective score to the best 0.2824% under the four-objective score (Figure 3b). The SGC therefore remained an unusually strong overall solution, but it was not optimized to maximize the breadth of amino-acid outcomes exposed by a single nucleotide substitution. This contrast is consistent with a tension between preserving existing proteins and increasing the range of phenotypic variation available to mutation.

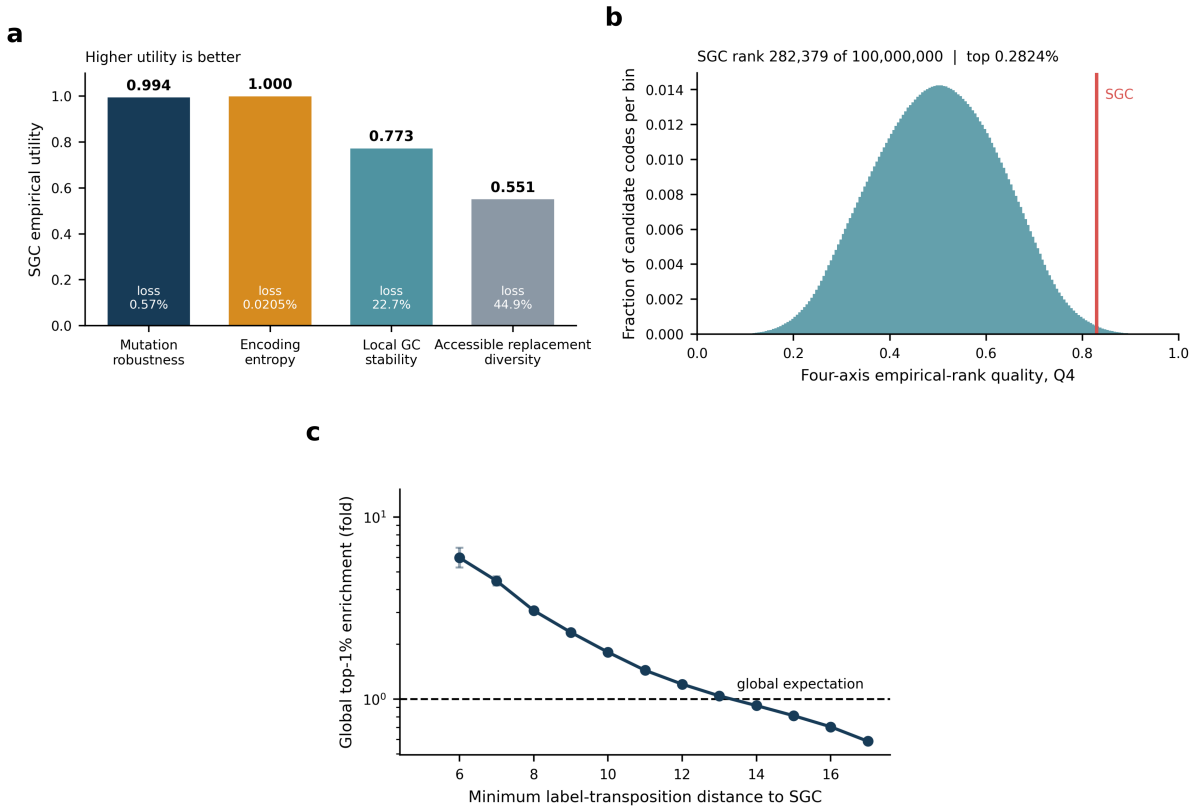

Figure 3 | Accessible replacement diversity exposes a robustness–explorability trade-off without eliminating the SGC's globally unusual composite performance. (a) SGC utilities after accessible replacement diversity is added as a fourth axis; the diversity utility is substantially lower than the three conservative utilities. (b) Distribution of the equal-weight four-axis score Q4; the red line marks the SGC (rank 282,379; top 0.2824% of the conditional 100-million-code library). (c) Enrichment of global top-1% Q4 codes by minimum block-label transposition distance from the SGC. Only strata containing at least 1,000 sampled codes are shown; error bars are binomial 95% confidence intervals and the dashed line denotes the global expectation.

Importantly, the raw number of distinct amino acids reachable by one nucleotide substitution is invariant under a bijective relabeling of the fixed canonical synonymous blocks. A genuine increase in single-nucleotide-variant (SNV) target richness therefore cannot be achieved by rearranging ψ alone. We consequently expanded the design space from block-label permutations to reassignment of mechanistically constrained decoder units. Under a transparent coarse wobble-pair model, the 61 sense codons were represented by 32 atomic decoder units—29 paired units and three single-codon units—and only equal-sized units were exchanged[11,13–14,23]. This preserved the canonical degeneracy of every amino acid, retained AUG as Met, and left all three termination codons unchanged.

We then maximized the mean number of distinct alternative amino acids reachable by one sense single-nucleotide substitution while constraining the directional missense cost to no more than 1.05 times the SGC value and the local GC-stability cost to no more than 1.10 times the SGC value. The search produced 639 feasible nonredundant designs, of which 131 were retained on the Pareto archive. A representative decoder-realizable design increased the mean accessible richness from 5.803 in the SGC to 6.115, while retaining a robustness ratio of 1.0445 and a GC-stability ratio of 1.0004. This design altered five decoder units comprising ten codon types, rather than independently reassigning arbitrary individual codons.

Table 1: Performance and constraint summary for the representative decoder-realizable design.

Measure	SGC	Designed code
Accessible amino-acid richness	5.803	6.115
Directional missense cost	1.0000	1.0445
Local GC-stability cost	1.0000	1.0004
Constraint status	reference	feasible

Table 2: Decoder-unit reassignment combinations in the representative design.

Combination	Decoder-unit change	Amino-acid change	Units
1	ACA/ACG ↔ CGA/CGG	Thr ↔ Arg	2
2	AGA/AGG → UUA/UUG → UAU/UAC → AGA/AGG	Arg → Leu → Tyr → Arg	3

The resulting design illustrates how the framework can be used in reverse. Instead of asking only why the SGC performs well, a desired phenotype can be specified—in this case, broader one-SNV amino-acid accessibility—and the least disruptive decoder-level changes satisfying explicit robustness, compositional and implementation constraints can be recovered. The output is therefore not merely a higher-scoring abstract genetic code, but a ranked set of mechanistically interpretable redesigns with quantified biological costs.

4 Discussion and Conclusion

This study reframes the canonical genetic code as a hierarchy of conditional design problems rather than a single unrestricted mapping from 64 codons to 21 outputs. That distinction is central to interpreting the results. Codeword length, degeneracy composition, synonymous-block geometry, semantic assignment and decoder implementation are not interchangeable layers, and an apparent optimum at one layer cannot explain choices made at another. Within the fixed-length, full-codebook model, triplets form a strict Pareto choice: L=3 is the shortest feasible codeword length, while longer words from L=4 to L=6 do not improve mutation exposure after the greater number of mutable nucleotide positions is accounted for. This result does not prove that every possible coding system must use triplets, but it identifies a precise set of assumptions under which triplets are jointly efficient and robust.

The degeneracy analysis gives an equally important negative result. Exact enumeration produced 59,755 anonymous block-size profiles, and 50 noncanonical profiles Pareto-dominated the standard profile under the pilot balance descriptors. Maximum degeneracy R=6 therefore does not by itself explain the canonical allocation. Natural reassignment can also change block membership and biological meaning while leaving the sorted degeneracy multiset unchanged. These observations show that degeneracy statistics are useful constraints on the search space, but are insufficient as an optimization principle unless the downstream block geometry and decoding machinery are included.

The strongest positive evidence arises after conditioning on the canonical triplet-block architecture. Recurrent local exclusions reduced the 20! semantic-assignment space from 2.43×10^{18} to approximately 10^{11} admissible mappings, enabling direct generation and evaluation of 10^8 complete codes. In this robustness-enriched library, the standard genetic code ranked in the best 0.57% for mutation robustness and the best 0.0205% for encoding entropy, while its local GC-stability rank was more moderate at the best 22.7%. The equal-weight three-objective score placed it $141{,}070^{\text{th}}$, or in the best 0.1411%. Thus, the main result is not that the canonical code uniquely optimizes one quantity. Rather, it occupies a rare region in which strong error tolerance and high synonymous encoding capacity coexist without extreme local GC instability. The enrichment of top-scoring codes near the standard assignment further suggests local structural organization, although the transposition distance used here is not an evolutionary trajectory.

Adding accessible replacement diversity exposes the cost of this conservative design. The standard code's diversity utility was 0.551, and its composite rank shifted to the best 0.2824% under four objectives. This remains unusually high, but indicates that the code is not optimized to maximize the breadth of amino-acid outcomes reachable by a single nucleotide change. The result is consistent with a trade-off between robustness and explorability: a code that protects an established proteome need not be the code that most rapidly explores new phenotypes. Decoder-level redesign made this trade-off constructive. A representative feasible design increased mean accessible richness from 5.803 to 6.115 while keeping directional missense cost at 1.0445 times and local GC cost at 1.0004 times the canonical values. Because the redesign exchanges coarse decoder units rather than arbitrary codons, it provides an interpretable engineering proposal rather than only an abstract high-scoring permutation.

Several limitations define the scope of these conclusions. The large-scale comparison fixes the canonical synonymous blocks and stop positions, so its percentiles are conditional and cannot establish global optimality over all L, H and P structures. The recurrent exclusions are empirical proposal rules learned from a robustness tail, not proofs of biological impossibility. Encoding entropy and local GC stability assume uniform synonymous-codon use; the coarse wobble model does not reconstruct the complete MG1655 tRNA pool, modified nucleosides or context-dependent mistranslation. The ProteinGym-derived replacement table aggregates heterogeneous proteins and does not capture residue-specific structural or functional effects. The present objectives also omit expression level, native codon bias, mRNA folding, transcriptional regulation, amino-acid biosynthetic cost and the cellular consequences of recoding.

Future work should therefore connect code-level screening to protein- and cell-level tests. Native MG1655 codon usage, measured mutation spectra, tRNA abundance, wobble pairing and mistranslation matrices can replace uniform assumptions. Reverse translation of a deliberately broad *E. coli* protein panel can then quantify mRNA folding, GC-window behavior, recoding burden and site-specific mutation effects. Weight sensitivity, Pareto-front analysis and held-out protein sets should be used alongside any composite score. Overall, the canonical genetic code is best interpreted here as a highly unusual, multi-objective stable compromise within a clearly defined conditional space. The same framework also identifies where alternative codes may be engineered when robustness, evolvability or implementation burden is assigned a different priority.

5 Acknowledgements

This work was developed during the 2026 PEBBLE BioFusion Workshop held at Westlake University in Hangzhou, China, from July 24 to August 4, 2026. All authors gratefully acknowledge the speakers for their inspiring lectures and discussions and the teaching assistants for their generous guidance and support throughout the workshop. We especially thank Dr. Fangzhou Xiao, the workshop organizer, whose enthusiasm, energy, dedication and generous assistance strongly supported the development of this project and enriched the workshop experience.

6 Conflict of Interest

The authors declare that they have no conflict of interest.

7 References

1. Crick, F. H. C. The origin of the genetic code. J. Mol. Biol. 38, 367–379 (1968). doi: 10.1016/0022-2836(68)90392-6.

2. Freeland, S. J. & Hurst, L. D. The genetic code is one in a million. J. Mol. Evol. 47, 238–248 (1998). doi: 10.1007/PL00006381.

3. Freeland, S. J. & Hurst, L. D. Load minimization of the genetic code: history does not explain the pattern. Proc. R. Soc. B 265, 2111–2119 (1998). doi: 10.1098/rspb.1998.0547.

4. Woese, C. R. et al. On the fundamental nature and evolution of the genetic code. Cold Spring Harb. Symp. Quant. Biol. 31, 723–736 (1966). doi: 10.1101/SQB.1966.031.01.093.

5. Wong, J. T. F. A co-evolution theory of the genetic code. Proc. Natl Acad. Sci. USA 72, 1909–1912 (1975). doi: 10.1073/pnas.72.5.1909.

6. Koonin, E. V. & Novozhilov, A. S. Origin and evolution of the genetic code: the universal enigma. IUBMB Life 61, 99–111 (2009). doi: 10.1002/iub.146.

7. Blattner, F. R. et al. The complete genome sequence of *Escherichia coli* K-12. Science 277, 1453–1462 (1997). doi: 10.1126/science.277.5331.1453.

8. Keseler, I. M. et al. EcoCyc: a comprehensive database of *Escherichia coli* biology. Nucleic Acids Res. 39, D583–D590 (2011). doi: 10.1093/nar/gkq1143.

9. Hamming, R. W. Error detecting and error correcting codes. Bell Syst. Tech. J. 29, 147–160 (1950). doi: 10.1002/j.1538-7305.1950.tb00463.x.

10. Kelleher, J. & O'Sullivan, B. Generating all partitions: a comparison of two encodings. arXiv:0909.2331 (2009; revised 2014).

11. Crick, F. H. C. Codon–anticodon pairing: the wobble hypothesis. J. Mol. Biol. 19, 548–555 (1966). doi: 10.1016/S0022-2836(66)80022-0.

12. Kramer, E. B. & Farabaugh, P. J. The frequency of translational misreading errors in *Escherichia coli* is largely determined by tRNA competition. RNA 13, 87–96 (2007). doi: 10.1261/rna.294907.

13. Grosjean, H. & Westhof, E. An integrated, structure- and energy-based view of the genetic code. Nucleic Acids Res. 44, 8020–8040 (2016). doi: 10.1093/nar/gkw608.

14. Ye, S. & Lehmann, J. Genetic code degeneracy is established by the decoding center of the ribosome. Nucleic Acids Res. 50, 4113–4126 (2022). doi: 10.1093/nar/gkac171.

15. Freeland, S. J., Knight, R. D., Landweber, L. F. & Hurst, L. D. Early fixation of an optimal genetic code. Mol. Biol. Evol. 17, 511–518 (2000). doi: 10.1093/oxfordjournals.molbev.a026331.

16. Buhrman, H., van der Gulik, P. T. S., Kelk, S. M., Koolen, W. M. & Stougie, L. Some mathematical refinements concerning error minimization in the genetic code. IEEE/ACM Trans. Comput. Biol. Bioinform. 10, 312–318 (2013). doi: 10.1109/TCBB.2011.40.

17. Notin, P. et al. ProteinGym: large-scale benchmarks for protein fitness prediction and design. Adv. Neural Inf. Process. Syst. 36 (2023).

18. Shannon, C. E. A mathematical theory of communication. Bell Syst. Tech. J. 27, 379–423, 623–656 (1948). doi: 10.1002/j.1538-7305.1948.tb01338.x.

19. Plotkin, J. B. & Kudla, G. Synonymous but not the same: the causes and consequences of codon bias. Nat. Rev. Genet. 12, 32–42 (2011). doi: 10.1038/nrg2899.

20. Emmerich, M. T. M. & Deutz, A. H. A tutorial on multiobjective optimization: fundamentals and evolutionary methods. Nat. Comput. 17, 585–609 (2018). doi: 10.1007/s11047-018-9685-y.

21. Deb, K., Pratap, A., Agarwal, S. & Meyarivan, T. A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans. Evol. Comput. 6, 182–197 (2002). doi: 10.1109/4235.996017.

22. Pines, G., Winkler, J. D., Pines, A. & Gill, R. T. Refactoring the genetic code for increased evolvability. mBio 8, e01654-17 (2017). doi: 10.1128/mBio.01654-17.

23. Alkatib, S. et al. The contributions of wobbling and superwobbling to the reading of the genetic code. PLoS Genet. 8, e1003076 (2012). doi: 10.1371/journal.pgen.1003076.

24. Meyer, F. et al. UGA is translated as cysteine in pheromone 3 of Euplotes octocarinatus. Proc. Natl Acad. Sci. USA 88, 3758–3761 (1991). doi: 10.1073/pnas.88.9.3758.

25. Novozhilov, A. S., Wolf, Y. I. & Koonin, E. V. Evolution of the genetic code: partial optimization of a random code for robustness to translation error in a rugged fitness landscape. Biol. Direct 2, 24 (2007). doi: 10.1186/1745-6150-2-24.