Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Triplet architecture enables deep error-minimization of the genetic code

**Hannah E. Ribbeck**
McNeese State University
**Janak Paudyal**
McNeese State University
**Boggavarapu Kiran**
kiran@mcneese.edu

McNeese State University

**Research Article**

**Keywords:** Genetic code, Error minimization, Codon length, Hamming graph, Dirichlet energy, Wobble position

**Posted Date:** September 3rd, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-10491657/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# Abstract

The standard genetic code (SGC) assigns amino acids to trinucleotide codons in a pattern that minimizes the phenotypic impact of translation errors. Whether this structure reflects selection on amino acid assignments or an intrinsic property of the triplet architecture has not been tested. We quantify error-minimization using two graph-theoretic metrics on the codon Hamming graph, Dirichlet energy and receiver noise distortion, scored against 8 PCA-derived physicochemical properties. The SGC outperforms all 1,000,000 random codes with matched degeneracy on both metrics ($p < 10^{-6}$). To separate the contribution of codon length from alphabet size, we compared doublet and triplet codes in a 2 × 2 factorial design. A 10-amino-acid doublet projection of the SGC falls within the tail of its null distribution ($z = -3.8$; best random doublet $z = -4.0$), whereas a 10-amino-acid triplet code with SGC degeneracy falls beyond its null ($z = -14.8$; no random code in $10^6$ matched it). Triplet architecture therefore crosses an optimization threshold that doublet architecture cannot, independently of alphabet size. The architectural affordance is synonymy at the wobble position: position 3 absorbs 69 to 77% of single-nucleotide errors as synonymous substitutions, a geometric degree of freedom that doublet codes cannot provide.


# Introduction

Single-nucleotide errors during translation occur at rates of $10^{-4}$ to $10^{-3}$ per codon (Drummond and Wilke 2008; Kramer and Farabaugh 2007; Landerer et al. 2024). Such errors cause the ribosome to insert the wrong amino acid, and the phenotypic cost depends on which amino acid replaces which. A valine-to-isoleucine substitution is nearly silent; a glycine-to-tryptophan change is often catastrophic. The standard genetic code (SGC) mitigates this cost: codons that differ by a single nucleotide tend to encode chemically similar amino acids (Woese 1965; Sonneborn 1965), so the most frequent translation errors tend to produce the least damaging substitutions.

Whether this pattern reflects natural selection or an accidental by-product has been debated for decades. Haig and Hurst (1991) introduced a Monte Carlo randomization of amino acid assignments as a test: if the SGC minimizes errors more effectively than most random alternatives with identical degeneracy, natural selection is implicated. Freeland and Hurst (1998) estimated that the SGC outperforms ~ 99.99% of such random codes on a polar requirement scale, and subsequent work by various groups confirmed this high level of optimization using different property measures (Gilis et al. 2001; Goodarzi et al. 2004; Itzkovitz and Alon 2007). However, Novozhilov, Wolf, and Koonin (2007) demonstrated that even modest local optimization of a random starting code can achieve robustness comparable to the SGC. They concluded that the code represents partial optimization within a rugged fitness space, a combination of natural selection and “frozen accident.” Their later work emphasized that the apparent optimization level depends on which amino acid property is measured, and that no single principled criterion exists for choosing among alternatives (Koonin and Novozhilov 2009).

Two problems have persisted. The first is property selection bias: each study chooses one or a few physicochemical scales, and the resulting optimization percentile varies with the chosen property (Koonin and Novozhilov 2009; Archetti 2004). The second is that all prior analyses operate within a single codon length. It is not clear whether the triplet architecture itself contributes to the code's capacity for optimization. Triplet codes ($4^3 = 64$) encode 20 amino acids with substantial synonymy, compared to doublet codes ($4^2 = 16$ codons), which can encode at most 15 amino acids plus one stop signal. What this synonymy contributes to error-minimization has not been quantified.

Here we address both problems. Information-theoretic treatments of the genetic code have a substantial prior literature. Tlusty (2007, 2008, 2010) formulated the codon-amino acid mapping as a noisy communication channel and showed that the graph-Laplacian of the codon error-graph governs the smooth modes corresponding to amino acid assignments. Radványi and Kun (2021) applied rate-distortion theory to quantify the mutational robustness of codon usage across organisms. We build on this framework by applying two specific metrics, Dirichlet energy (the standard graph-Laplacian quadratic form) and receiver noise distortion (an expected-distortion measure in the sense of Tlusty's error-load), and adding three elements absent from prior work: a degeneracy-preserving empirical null generated by $10^6$ Monte Carlo shuffles of amino acid labels, scoring against 8 orthogonal amino acid properties derived by principal component analysis (PCA) of published physicochemical data, and a $2 \times 2$ factorial design that separates the contribution of codon length from alphabet size. PCA eliminates researcher degrees of freedom in property selection; the two metrics provide a concordance check; the factorial design isolates the architectural contribution. Applying this machinery across codon lengths, we find that triplet architecture enables a qualitatively deeper level of error-minimization than doublet architecture, independent of alphabet size. The architectural affordance is synonymy at the wobble position.

## Results

# The SGC is error-minimized beyond the reach of random codes

We represented the 61 sense triplet codons as nodes in a Hamming graph, connecting codons that differ by a single nucleotide substitution (Fig. 1a). This graph is the natural geometry of translation errors: two codons are neighbors if a single misreading event can convert one into the other, so an edge represents a single possible error. Each node carries an 8-dimensional property vector derived from PCA of amino acid physicochemical descriptors (see Materials and Methods). On this graph, we computed two metrics that capture different consequences of single-nucleotide errors.

Dirichlet energy sums the squared amino-acid property differences across every edge of the graph. A low Dirichlet energy means that codons adjacent in the graph tend to encode amino acids with similar physicochemical properties, so the code is chemically smooth in the sense that single-letter mutations tend to produce chemically conservative substitutions. Receiver noise distortion is the expected amino-

acid property change when a single decoding error occurs, assuming each of a codon's nine mutational neighbors is equally likely to be misread. A low noise distortion value means the average translation error produces a small chemical perturbation. The two metrics formalize complementary aspects of the same intuition. Dirichlet energy is a property of the whole code considered at once (a structural measure of how neighbors compare throughout the graph); noise distortion is the per-error cost averaged over codons (an operational measure of the damage a typical mistranslation event inflicts).

We generated 1,000,000 random codes by permuting amino acid labels among codon blocks while preserving the SGC's degeneracy structure. The SGC scored lower than all 1,000,000 random codes on both metrics (Fig. 2a,b). We report the SGC's position relative to the null as a z-score, the number of standard deviations separating the SGC's metric value from the mean of the random-code distribution. A z of zero means the SGC is average; increasingly negative z means the SGC is increasingly better (lower energy, lower distortion) than the typical random code. No random code matched or beat the SGC on Dirichlet energy ($z = -14.6$) or noise distortion ($z = -16.6$). The two metrics are highly correlated across random codes ($r = 0.96$; Fig. 2c), confirming that they capture overlapping but not identical features of code error-tolerance.

This result strengthens the Freeland and Hurst (1998) finding by approximately two orders of magnitude: $p < 10^{-6}$ versus their $\sim 10^{-4}$. The improvement traces to two sources: the use of 8 PCA-derived properties rather than a single researcher-selected scale, and the Hamming-graph formulation, which accounts for all single-nucleotide error pathways rather than a subset.

## Doublet projections of the SGC are only mildly optimized

Does this optimization extend to shorter codon lengths? The SGC's block structure already clusters chemically related amino acids at the doublet level: positions 1 and 2 of triplet codons define amino acid families (e.g., all four GCN codons encode alanine). To test how much optimization this doublet-level structure carries, we projected the SGC onto a doublet code.

We selected 10 representative amino acids from the 10 reduced-alphabet groups of Murphy et al. (2000), choosing for each group the amino acid whose 8-PC property vector was closest to the group centroid. These 10 amino acids were assigned to 16 doublet codons following the SGC's position 1–2 structure. This doublet code is a controlled comparator derived from the SGC, not a reconstruction of any historical code. We then ran 1,000,000 Monte Carlo shuffles with matched degeneracy.

The doublet projection scored $z = -3.8$ on noise distortion and $z = -3.3$ on Dirichlet energy (Fig. 3a). These values fall in the top 0.01% of random doublet codes, but they remain within the tail of the null distribution. The best random doublet code achieved $z = -4.0$. The SGC's doublet projection is therefore consistent with a well-sampled random code: optimized, but not beyond what chance produces in a million trials. This contrast with the triplet result (where no random code came close) raised an immediate question: does the depth of achievable optimization depend on codon length?

## Triplet architecture, not alphabet size, enables deep optimization

The gap between z = − 3.8 (doublet) and z = − 16.6 (triplet) could reflect either the difference in codon length or the difference in alphabet size (10 vs. 20 amino acids). A larger alphabet creates a richer property space with more potential for optimization simply because there are more distinct entities to arrange. To disentangle these effects, we constructed a third condition: the same 10 amino acids assigned to triplet codons.

We projected the SGC's triplet structure onto the 10-amino-acid reduced alphabet, mapping each of the 20 amino acids to its reduced-alphabet representative, and ran 1,000,000 shuffles with matched degeneracy. The 10-amino-acid triplet code achieved z = − 14.8 (noise distortion) and z = − 13.3 (Dirichlet energy). Zero of 1,000,000 random codes matched or beat it on either metric (Fig. 3b).

Table 1 reports the raw z-scores for each condition. The differences between conditions (Δz) are informative as directional indicators but are not strictly commensurable across architectures: the null distributions at n = 2 and n = 3 have different variance structure because they operate over code spaces of different size and degeneracy. We therefore interpret the results in terms of whether each condition crosses the null envelope rather than in terms of z-score subtraction. On this reading: at n = 2 the SGC doublet projection falls inside the null tail (z = − 3.8; best random doublet z = − 4.0, within one unit of the SGC), so the doublet projection is not distinguishable from a well-sampled random code. At n = 3, by contrast, both the 10-AA and 20-AA conditions lie beyond the reach of $10^6$ random permutations. Triplet architecture therefore crosses an optimization threshold that doublet architecture does not, whether the alphabet contains 10 or 20 amino acids.

Table 1
Decomposition of error-minimization z-scores across architecture (n = 2 vs. n = 3) and alphabet (10 vs. 20 amino acids). At n = 2 the SGC projection lies within its null tail; at n = 3 both alphabet conditions lie beyond their null. Difference rows (C − A, B − C) are reported as directional indicators only. We do not report their ratio: the two effects are computed against null distributions with different variance structure and are not strictly commensurable, so the conclusion rests on which conditions cross the null envelope rather than on arithmetic among z-scores (see Discussion).

| Condition | Amino acids | Codon length | Noise z | Dirichlet z | D–E corr |
|---|---|---|---|---|---|
| A | 10 | 2 | −3.80 | −3.32 | 0.932 |
| B | 20 | 3 | −16.56 | −14.60 | 0.959 |
| C | 10 | 3 | −14.81 | −13.35 | 0.958 |
| Architecture (C − A) | — | — | −11.01 | −10.03 | — |
| Alphabet (B − C) | — | — | −1.75 | −1.25 | — |

Both metrics converge on the same qualitative conclusion. A triplet code with SGC degeneracy can be optimized beyond the reach of matched-degeneracy random sampling; a doublet projection of the same code cannot. The threshold is set by codon length, not by the size of the amino acid alphabet: the 10-AA

and 20-AA triplet conditions both lie beyond the null, while neither doublet condition does. The result is robust to the choice of representative amino acids: alternative selection strategies (farthest-from-centroid, mean over three random draws) shift the 10-AA triplet z-score by at most 0.56 units on noise distortion and 0.92 units on Dirichlet energy, small compared to the SGC's displacement from the null of approximately 15 standard deviations (table S1). The conclusion is also robust to the choice of metric convention: substituting the squared graph Laplacian f⊺L²f for f⊺Lf shifts the z-scores by ≤ 1.1 units across all conditions, and a Boltzmann-weighted noise-distortion variant with wobble softening strengthens rather than weakens the SGC's z-score (see Materials and Methods, Sensitivity analysis).

## The wobble position is the architectural mechanism

What structural feature of triplets accounts for this difference? The answer is synonymy concentrated at position 3.

In the SGC, 69% of single-nucleotide substitutions at position 3 produce a synonymous codon, leaving the amino acid unchanged and the property distortion at zero. At positions 1 and 2, the synonymous fractions are 5% and 10%, respectively. The weighted mean across all three positions is 28%. In the 10-amino-acid triplet code, where each reduced class commands approximately 6 codons, position-3 synonymy rises to 77% (Fig. 3b).

The doublet code has no comparable structure. Position 1 absorbs 33.3% of substitutions as synonymous and position 2 only 8.3%, so neither approaches the 68.9% that position 3 supplies in the triplet code. Both positions carry amino-acid-distinguishing information, and no single position is dedicated to error absorption.

This asymmetry has a direct consequence for optimization geometry. In a triplet code, positions 1 and 2 determine the amino acid family while position 3 absorbs errors as synonymous changes. The wobble position acts as a noise-absorbing dimension in the Hamming graph: a structural degree of freedom that converts potential errors into silent substitutions (Fig. 4). This affordance does not, by itself, produce the SGC's observed optimization depth. The null codes in our Monte Carlo preserve the SGC's degeneracy structure, so every null shares the same wobble-position synonymy as the SGC, yet none matched it. The remaining depth comes from the specific amino acid assignments at positions 1–2, consistent with the pattern identified by Freeland and Hurst (1998). What the architecture supplies is the geometric room in which this deep optimization becomes possible; the specific assignments fill that room.

Doublet codes cannot sustain a dedicated synonymy position. With only two positions and at most 15 amino acids to encode, both positions must carry discriminating information. This constraint caps the achievable optimization: the best doublet code cannot approach the error-minimization of a triplet code, not because of poor amino acid assignment, but because doublets lack the structural feature that makes deep optimization geometrically possible.

# Discussion

The SGC's error-minimizing properties have been recognized since Woese (1965), quantified by Haig and Hurst (1991) and Freeland and Hurst (1998), and placed in evolutionary context by Novozhilov et al. (2007) and Koonin and Novozhilov (2009). Our results extend this work in two directions.

First, the SGC's optimization is more extreme than previously estimated: $p < 10^{-6}$ on two concordant graph-theoretic metrics scored against PCA-derived properties. The PCA approach resolves the property-selection problem identified by Koonin and Novozhilov (2009). By extracting orthogonal components from the full physicochemical dataset, it eliminates the researcher degree of freedom that has made prior optimization estimates difficult to compare across studies.

Second, the factorial decomposition reveals that error-minimization is an architecture-level property, not merely an assignment-level one. Novozhilov et al. (2007) showed that starting from random codes and applying local optimization produces codes with robustness to errors comparable to the SGC, evidence that the SGC's optimization level is achievable through modest evolutionary tinkering. Our result adds an architectural dimension to this picture. The accessible optimization depth depends on codon length. At $n = 2$, the best random code in a million trials ($z = -4.0$) lies within one unit of the SGC projection ($z = -3.8$), so the doublet regime admits little optimization headroom. At $n = 3$, the SGC occupies a region of code space ($z = -16.6$) that random sampling cannot reach even in $10^6$ draws. The local optimization that Novozhilov et al. described can produce deep error-minimization, but only within the triplet architecture. Doublet codes offer a much shallower optimization valley, because many nearby alternatives perform similarly and selection has little geometric purchase.

The mechanism is architectural rather than assignment-specific. Position-3 synonymy supplies the geometric room in which deep optimization becomes possible: errors along the position-3 axis are silent, freeing positions 1 and 2 to encode amino acid identity in an optimized arrangement. Because every Monte Carlo null preserves SGC degeneracy, the null codes inherit this same affordance, yet none matched the SGC on either metric. The architecture does not, on its own, produce the observed depth; it makes the depth reachable, and the specific amino acid assignments at positions 1–2 then occupy the available room. Doublet codes, with both positions carrying discriminating information, cannot generate this directional structure at all, so deep optimization is geometrically inaccessible regardless of which amino acids are assigned.

This finding reframes the standard explanation for triplet codon length. The conventional argument is arithmetic: 16 doublet codons cannot encode 20 amino acids plus stop signals, so triplets are the minimum sufficient length. Our results identify a second, structural consideration. A triplet code provides a dedicated synonymy position that a doublet code cannot. Within our analysis, this is the position whose absence restricts the doublet projection to a shallow optimization valley. Whether a doublet code with a smaller alphabet could nonetheless be deeply optimized is not addressed here, but the architectural difference we identify is present regardless of alphabet size.

The present analysis addresses the lower bound on codon length. Whether codes longer than triplets would provide additional optimization depth is a separate question: adding synonymy positions adds geometric room but also imposes genome-size and frameshift costs that scale with codon length, and whether those costs are recovered by fidelity gains is not determined by the present analysis. Our result is therefore restricted to the doublet–triplet comparison, where the threshold-crossing asymmetry is clear.

The 10-amino-acid reduced alphabet derives from the Murphy et al. (2000) clustering, and the representative amino acid for each group was chosen by proximity to the group centroid. Alternative choices (farthest-from-centroid, mean over three random draws) shift the z-score by at most 0.56 units on noise distortion and 0.92 units on Dirichlet energy (table S1), small compared to the SGC's displacement from the null, so the result is robust to this decision, but the reduced alphabet is nonetheless an approximation. The doublet code table is a projection of the SGC's position 1–2 structure onto 16 codons, not a historically reconstructed primordial code. The factorial comparison shows that this particular projection has a shallower optimization valley than its triplet counterpart; it does not prove that all conceivable doublet codes must. We restricted the factorial to n = 2 and n = 3 rather than extending to n = 4 for two reasons. First, the natural n = 4 baseline is a degenerate embedding of the SGC in which position 4 is silent by construction ($AA(b_1b_2b_3b_4) := AA_SGC(b_1b_2b_3)$), so any test of its optimization depth is circular: position-4 substitutions are synonymous because we defined them to be. Second, a symmetric 2 × 2 factorial requires both axes to have paired conditions, and no biological 20-AA code at n = 4 exists to pair with the 10-AA case. The analysis measures error-minimization geometry but does not model the evolutionary process by which optimization was achieved. Three scenarios are consistent: architecture creates the optimization valley and selection fills it; architecture creates the valley and neutral block structure plus frozen history fills it; or naming-capacity constraints select for triplets and optimization follows as a secondary consequence. We do not distinguish among these.

The metrics themselves are not novel. Dirichlet energy $f^{\top}Lf$ is the standard graph-signal-processing measure of code smoothness, and a rate-distortion-theoretic noise distortion D on the codon error-graph was introduced by Tlusty (2007, 2008, 2010) and used in a different empirical context by Radványi and Kun (2021). Our contribution sits elsewhere: a degeneracy-preserving empirical null generated by $10^6$ Monte Carlo shuffles of amino acid labels (converting what had been analytic phase-transition arguments into exact z-scores and p-values), a 2 × 2 factorial design isolating codon-length from alphabet-size effects, and a PCA-compressed 8-dimensional property vector distilled from 20 descriptors that removes researcher degrees of freedom in selecting which amino acid property to weigh. We report D under uniform Hamming-1 weighting because this is the minimal-assumption specialization of Tlusty's construction; a Boltzmann-weighted version with wobble softening sharpens the SGC's z-score rather than weakening it (Materials and Methods, Sensitivity analysis), so the choice does not affect any conclusion drawn here.

The genetic code has been characterized as a partially optimized frozen accident (Koonin and Novozhilov 2009). Our results specify the architecture within which that optimization operates. The

wobble position of the triplet code supplies the geometric room in which deep error-minimization becomes reachable; the specific amino acid assignments at positions 1 and 2 then fill that room. Doublet architecture does not provide comparable room, and so the depth observed in the SGC is not accessible to a doublet code regardless of its amino acid assignments.

## Conclusion

Error-minimization in the genetic code is an architecture-level property, not only an assignment-level one. On two concordant graph-theoretic metrics scored against PCA-derived amino acid properties, the SGC outperforms all $10^6$ random codes with matched degeneracy ($p < 10^{-6}$ on both Dirichlet energy and noise distortion).

A 2 × 2 factorial decomposition separates the contributions of codon length and alphabet size. At fixed 10-amino-acid alphabet, triplet architecture crosses an optimization threshold that doublet architecture cannot reach: at n = 2, the best random code in a million trials barely exceeds the SGC projection; at n = 3, the SGC lies in a region of code space that random sampling does not reach at all. Expanding the alphabet from 10 to 20 amino acids at fixed triplet length shifts the optimization only modestly. The architectural shift is qualitative; the alphabet shift is not.

The mechanism is position-3 synonymy. The wobble position absorbs single-nucleotide errors as synonymous substitutions and leaves positions 1 and 2 free to encode amino acid identity in an optimized arrangement. Doublet codes cannot sustain a dedicated synonymy position, and deep optimization is therefore geometrically inaccessible to them regardless of assignment. The triplet code is not merely a minimum-sufficient length for encoding 20 amino acids; it is the minimum length at which deep error-minimization is structurally possible.

## Materials and methods

# Amino acid property space

Amino acid physicochemical properties were obtained from the AAindex database (Kawashima and Kanehisa 2000). We selected 20 descriptors spanning hydrophobicity, volume, charge, polarity, flexibility, and secondary structure propensity (table S2). Properties were standardized (mean 0, variance 1) across the 20 canonical amino acids. Principal component analysis extracted 8 orthogonal components explaining 97% of total variance. Each amino acid is represented by its 8-dimensional PC score vector.

# Codon Hamming graph

For a code of length n, the Hamming graph H(4, n) has $4^n$ nodes (codons) with edges connecting codons that differ at exactly one position. Biologically, this graph encodes the one-step error geometry of translation: two codons share an edge if and only if a single nucleotide misreading event can convert one into the other. The degree of each node in the n = 3 graph is 9, corresponding to the 3 alternative

bases at each of the 3 codon positions. Each sense codon inherits the 8-dimensional property vector of its assigned amino acid. Stop codons are excluded, so the effective graph at n = 3 has 61 nodes.

# Dirichlet energy

For a property vector function f on the Hamming graph, the Dirichlet energy is $E = \Sigma_{(u,v) \in \text{edges}} ||f(u) - f(v)||^2$, where the sum runs over all edges connecting sense codons. In plain terms, for every pair of codons that are one nucleotide apart, we compute the chemical distance between the amino acids they encode, square it, and sum across all such pairs. A low Dirichlet energy means that codons that are neighbors in the graph tend to encode chemically similar amino acids. This is the graph-theoretic version of saying the code is “chemically smooth”: single-letter mutations tend to produce conservative substitutions. E is the standard graph-Laplacian quadratic form $f^{\top}Lf$ on the codon error-graph; Tlusty (2007, 2010) uses the weighted variant with misreading probabilities as edge weights, and our expression is the unweighted (uniform Hamming-1) specialization.

# Noise distortion

Noise distortion measures the expected property change from a uniform random single-nucleotide error: $D = (1/|C|) \times \Sigma_{c \in C} (1/|N(c)|) \times \Sigma_{c' \in N(c)} ||f(c) - f(c')||$, where C is the set of sense codons, N(c) is the set of sense codons reachable from c by a single-nucleotide substitution, and ||.|| is the Euclidean norm in 8-PC space. In words, for each codon we compute the average chemical distance to its mutational neighbors, and then average this quantity over all codons. This is the expected amino-acid property change per decoding error, under the assumption that any of a codon’s mutational neighbors is equally likely to be the misread outcome. A low noise distortion value therefore means the typical translation error produces a small chemical perturbation. D is a specialization of the rate-distortion-theoretic distortion measure used for the codon channel by Tlusty (2007, 2008, 2010) and by Radványi and Kun (2021); our version uses a uniform codon prior, unit Hamming-1 edge weights, and an 8-dimensional PCA-derived property distance, whereas Tlusty parameterizes edge weights by misreading probabilities and Radványi and Kun use empirical codon usage and a Kimura two-parameter mutation model. The two metrics are complementary: Dirichlet energy is a global structural quantity (the sum of squared edge weights across the whole graph), while noise distortion is an expected per-error cost (the edge-weight average, codon by codon, unsquared). Both use the same graph and the same amino-acid property vectors; they differ in how they aggregate over edges and whether they weight large property jumps by squaring them. These conventions are not our invention: D inherits the unsquared rate-distortion distance convention of Tlusty (2007, 2010, Eq. 1) and Radványi and Kun (2021, Eq. 1), while E is the standard Dirichlet energy $f^{\top}Lf$ of graph signal processing (Chung 1997; Shuman et al. 2013). Reporting both is deliberate: a linear pathway cost (D) and a quadratic smoothness (E) are correlated but not equivalent, and the SGC’s displacement under both — rather than under either alone — is the relevant claim.

# Null model

Random codes were generated by permuting amino acid labels among codon blocks while preserving the degeneracy structure of the reference code. For the SGC at n = 3, this permutes amino acid identities among the 20 codon families {UUU/C (Phe), UUA/G (Leu1), CUU/C/A/G (Leu2), …}, maintaining the number and size of each family. Both metrics were computed for each permutation. One million permutations were generated for each condition. The one-sided p-value is the fraction of random codes with metric value ≤ the reference code.

# Cross-length analysis

Three conditions were compared. Condition A (10 AA, n = 2): a doublet code derived from the SGC's position 1–2 structure. The 20 amino acids were mapped to 10 reduced-alphabet groups following Murphy et al. (2000). For each group, the representative amino acid was selected as the one with minimum Euclidean distance to the group centroid in 8-PC space. The 10 representatives were assigned to 16 doublet codons with the SGC's position 1–2 degeneracy structure. Condition B (20 AA, n = 3): the standard genetic code. Condition C (10 AA, n = 3): the SGC's triplet structure with the same 10 reduced-alphabet amino acids, where each of the 20 amino acids was replaced by its group representative. For all three conditions, 1,000,000 null permutations were generated with matched degeneracy, and both metrics were computed. The architecture effect is z_C − z_A; the alphabet effect is z_B − z_C.

# Sensitivity analysis

Three representative-selection strategies were tested for Condition C: (i) the amino acid closest to the group centroid (used in the main text), (ii) the amino acid farthest from the group centroid, and (iii) a random member of each group, averaged over three independent seeds. The z-score ranges across these strategies were 0.56 units on noise distortion and 0.92 units on Dirichlet energy (table S1), small compared to the SGC's displacement from the null of approximately 15 standard deviations.

We also tested two alternative metric conventions that a reviewer might prefer. First, the squared graph Laplacian f⍁L²f, which Tlusty (2007, Eq. 3) invokes in his linear-stability analysis of eigenmodes near the coding transition, gives z-scores within 1.1 units of the standard Dirichlet energy f⍁Lf at $10^5$ null samples for all three conditions (A: −3.31 under both; C: −13.30 versus − 12.24; B: −14.59 versus − 13.52). The null-distribution correlation between the two forms is r = 0.987 for Condition B and r = 0.987 for Condition C, reflecting the relatively flat spectrum of the Hamming-1 codon graph, and any qualitative conclusion survives the substitution. Second, the main-text D assumes uniform weighting of Hamming-1 substitutions; a Boltzmann-weighted variant (implemented in src/receiver/thermo_noise.py) with position-specific penalties K = 3.0 for positions 1 and 2, wobble strength 0.4 for position 3, and RT = 1.0 gives z-scores of − 24.89 for the SGC (Condition B), − 20.12 for the 10-AA triplet (Condition C), and − 3.79 for the doublet projection (Condition A, unchanged because n = 2 has no distinct wobble position at which the softmax concentrates) at $10^5$ null samples. The Boltzmann-weighted result is 5–8 z-units stronger on triplets than the uniform version reported in the main text; the weighted model redirects probability mass onto position-3 errors, which are ~ 69% synonymous in the SGC, and random codes do not share this advantage. We report the uniform version as the conservative, minimal-assumption

baseline: the conclusion does not depend on a particular choice of misreading model or wobble parameter, and a reviewer preferring a mechanistic weighting would compute stronger evidence for SGC optimization, not weaker. Full numerical output: results/preempt_metrics.json.

# Synonymy computation

Position-specific synonymy S_k was computed as the fraction of all possible single-nucleotide substitutions at position k that produce a synonymous codon (same amino acid or same reduced-alphabet group), averaged over all sense codons.

## Declarations

## Funding

This work received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

## Conflict of interest

The authors declare no competing interests.

## Author contributions

B.K. conceived the study and supervised the project. H.E.R. designed and ran the Monte Carlo simulations, implemented the analysis pipeline, and produced the figures. J.P. contributed to manuscript drafting and provided cross-disciplinary review. B.K. wrote the manuscript with contributions from H.E.R. and J.P. All authors approved the final version.

## Data availability

All analysis outputs supporting this manuscript are contained in the accompanying repository, archived at [Zenodo DOI — CONFIRM: deposit repository snapshot and insert DOI before publication]. The AAindex descriptors used for the principal component analysis are publicly available at https://www.genome.jp/aaindex/.

## Code availability

All Monte Carlo simulation code and figure-generation scripts are available in the accompanying repository, archived at the same Zenodo DOI as above [CONFIRM].

## Acknowledgements

The authors thank Department of Chemistry and Physics colleagues at McNeese State University for discussion. Computation was performed on local research servers.

## References

1. Archetti M (2004) Codon usage bias and mutation constraints reduce the level of error minimization of the genetic code. J Mol Evol 59:258–266. https://doi.org/10.1007/s00239-004-2620-0
2. Chung FRK (1997) Spectral graph theory. CBMS Regional Conference Series in Mathematics, no. 92. American Mathematical Society, Providence, RI
3. Drummond DA, Wilke CO (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134:341–352. https://doi.org/10.1016/j.cell.2008.05.042
4. Freeland SJ, Hurst LD (1998) The genetic code is one in a million. J Mol Evol 47:238–248. https://doi.org/10.1007/PL00006381
5. Gilis D, Massar S, Cerf NJ, Rooman M (2001) Optimality of the genetic code with respect to protein stability and amino-acid frequencies. Genome Biol 2. https://doi.org/10.1186/gb-2001-2-11-research0049. :research0049.1-research0049.12
6. Goodarzi H, Nejad HA, Torabi N (2004) On the optimality of the genetic code, with the consideration of termination codons. BioSystems 77:163–173. https://doi.org/10.1016/j.biosystems.2004.05.031
7. Haig D, Hurst LD (1991) A quantitative measure of error minimization in the genetic code. J Mol Evol 33:412–417. https://doi.org/10.1007/BF02103132
8. Itzkovitz S, Alon U (2007) The genetic code is nearly optimal for allowing additional information within protein-coding sequences. Genome Res 17:405–412. https://doi.org/10.1101/gr.5987307
9. Kawashima S, Kanehisa M (2000) AAindex: amino acid index database. Nucleic Acids Res 28:374. https://doi.org/10.1093/nar/28.1.374
10. Koonin EV, Novozhilov AS (2009) Origin and evolution of the genetic code: the universal enigma. IUBMB Life 61:99–111. https://doi.org/10.1002/iub.146
11. Kramer EB, Farabaugh PJ (2007) The frequency of translational misreading errors in E. coli is largely determined by tRNA competition. RNA 13:87–96. https://doi.org/10.1261/rna.294907
12. Landerer C, Poehls J, Toth-Petroczy A (2024) Fitness effects of phenotypic mutations at proteome-scale reveal optimality of translation machinery. Mol Biol Evol 41:msae048. https://doi.org/10.1093/molbev/msae048
13. Murphy LR, Wallqvist A, Levy RM (2000) Simplified amino acid alphabets for protein fold recognition and implications for folding. Protein Eng 13:149–152. https://doi.org/10.1093/protein/13.3.149
14. Novozhilov AS, Wolf YI, Koonin EV (2007) 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. https://doi.org/10.1186/1745-6150-2-24
15. Novozhilov AS, Koonin EV (2009) Exceptional error minimization in putative primordial genetic codes. Biol Direct 4:44. https://doi.org/10.1186/1745-6150-4-44

16. Radványi Á, Kun Á (2021) Phylogenetic analysis of mutational robustness based on codon usage supports that the standard genetic code does not prefer extreme environments. Sci Rep 11:10963. https://doi.org/10.1038/s41598-021-90440-y
17. Shuman DI, Narang SK, Frossard P, Ortega A, Vandergheynst P (2013) The emerging field of signal processing on graphs: extending high-dimensional data analysis to networks and other irregular domains. IEEE Signal Process Mag 30:83–98. https://doi.org/10.1109/MSP.2012.2235192
18. Sonneborn TM (1965) Degeneracy of the genetic code: extent, nature, and genetic implications. In: Bryson V, Vogel HJ (eds) Evolving genes and proteins. Academic, New York, pp 377–397
19. Tlusty T (2007) A model for the emergence of the genetic code as a transition in a noisy information channel. J Theor Biol 249:331–342. https://doi.org/10.1016/j.jtbi.2007.07.029
20. Tlusty T (2008) Rate-distortion scenario for the emergence and evolution of noisy molecular codes. Phys Rev Lett 100:048101. https://doi.org/10.1103/PhysRevLett.100.048101
21. Tlusty T (2010) A colorful origin for the genetic code: information theory, statistical mechanics and the emergence of molecular codes. Phys Life Rev 7:362–376. https://doi.org/10.1016/j.plrev.2010.06.002
22. Woese CR (1965) On the evolution of the genetic code. Proc Natl Acad Sci USA 54:1546–1552. https://doi.org/10.1073/pnas.54.6.1546

# Figures

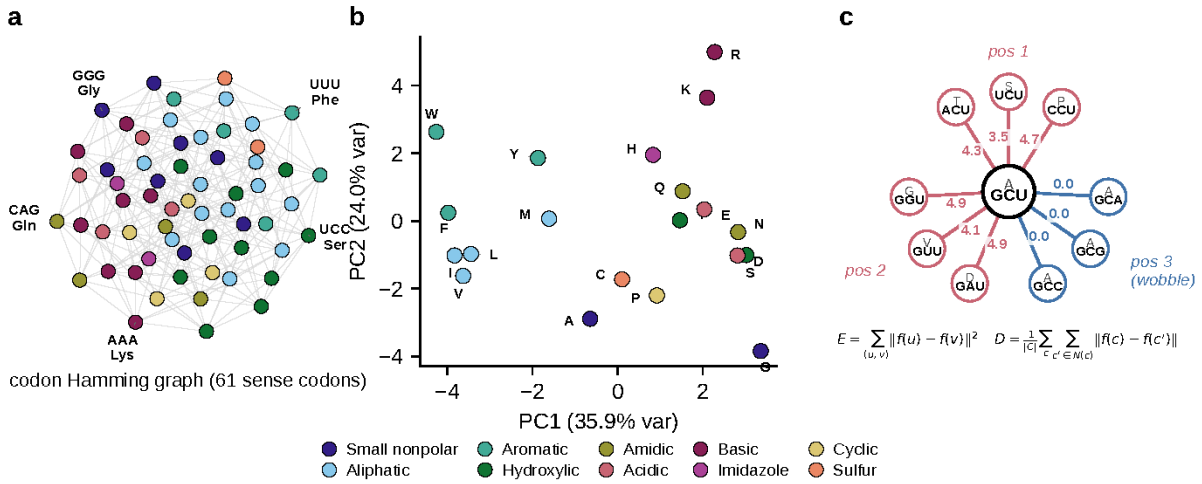



**Figure 1.** Framework. (a) Codon Hamming graph for n = 3: all 61 sense codons, with edges connecting codons that differ by a single nucleotide substitution (264 edges; average degree 8.6). Nodes are colored by amino acid chemical class (10-class RAA10 grouping, Murphy 2000); five peripheral anchor codons are labeled for orientation. (b) 20 amino acids in PC1–PC2 space of the 22-property descriptor set (after correlation-pruning and z-scoring; PC1 = 35.9 % of variance, PC2 = 24.0 %; 97.1 % total in the retained eight PCs used downstream). Hydrophobic residues (Phe, Trp, Tyr, Ile, Leu, Val, Met) lie at negative PC1; charged residues (Lys, Arg, Asp, Glu) occupy positive PC1. Class color scheme matches panel (a). (c) Hamming-1 neighborhood of the Ala codon GCU with edge-wise property distances ‖Δf‖ computed in the 8-PC space: all three position-3 neighbors (GCA, GCC, GCG) are synonymous with ‖Δf‖ = 0 (blue edges); the six position-1 and position-2 neighbors are non-synonymous with ‖Δf‖ ranging from 3.5 (UCU → Ser) to 4.9 (GGU → Gly, GAU → Asp) (red edges). Dirichlet energy E sums squared ‖Δf‖ over all graph edges; noise distortion D averages ‖Δf‖ over non-synonymous neighbors per codon.

## Figure 1

See image above for figure legend.

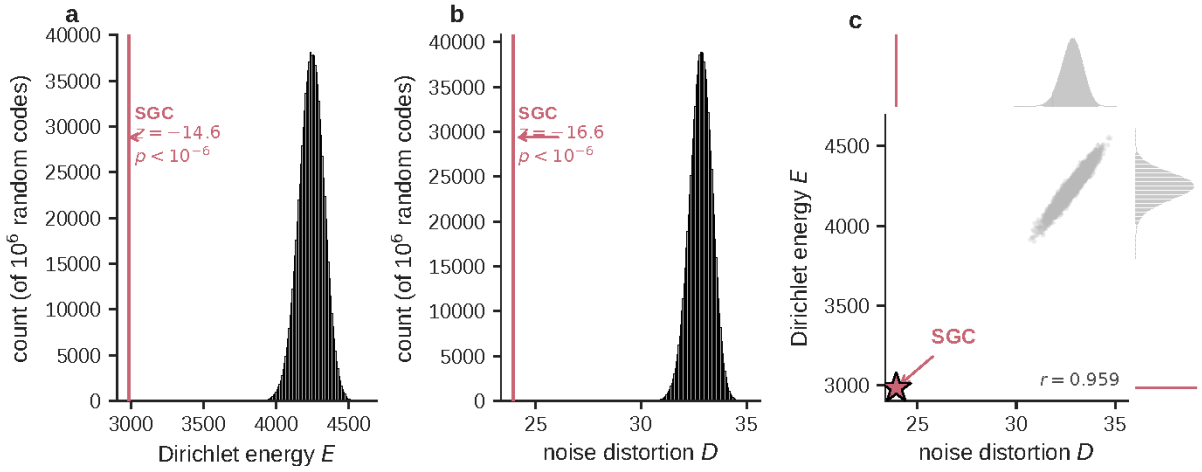



**Figure 2.** The Standard Genetic Code (SGC) is beyond the null under both metrics. (a) Null distribution of Dirichlet energy E across $10^6$ degeneracy- preserving random codes (grey). The SGC (red line) falls at z = −14.6 ($p < 10^{-6}$; 0 / $10^6$ random codes reach the SGC's energy). The null is generated by permuting amino acid labels across the 61 sense codons while preserving per-amino-acid codon counts, fixing stops, and pinning AUG → Met. (b) Same null-hypothesis test for noise distortion D: SGC at z = −16.6 ($p < 10^{-6}$; 0 / $10^6$). (c) Joint scatter of (D, E) for a 10 000-point random subsample from the null distribution (grey points). The SGC (red star) lies outside and below the null cloud in both metrics simultaneously. The two metrics are strongly but imperfectly correlated across the null (r = 0.959), so each provides partially independent evidence; neither suggests the SGC's optimization is an artefact of the other. Marginal histograms along the top (D) and right (E) axes reproduce panels (a) and (b) for direct comparison.

## Figure 2

See image above for figure legend.

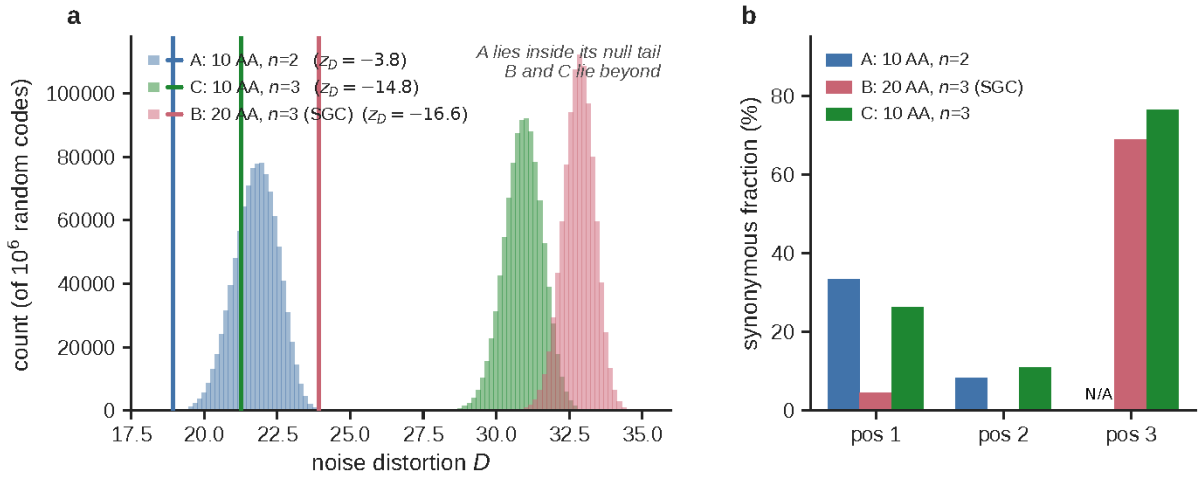



**Figure 3.** Triplet architecture enables deep optimization; reduced alphabet size does not. (a) Null distributions of noise distortion D across $10^6$ degeneracy-preserving random codes for three conditions, on a shared x-axis: Condition A (blue; 10 amino acids, n = 2 doublet code obtained by majority- vote projection of the SGC onto the first two codon positions under the RAA10 reduced alphabet), Condition C (green; 10 amino acids, n = 3 triplet with RAA10 applied to the sense codons), and Condition B (red; 20 amino acids, n = 3 — the SGC itself). Vertical lines mark the observed value for the corresponding code. The doublet SGC projection (A) sits well within its null tail (z_D = −3.8); the triplet conditions both lie beyond their nulls (z_D = −14.8 for C, z_D = −16.6 for B). The B and C null distributions substantially overlap, indicating the architectural shift from n = 2 to n = 3 is the dominant determinant of where the observed codes sit relative to their randomized backgrounds. (b) Position-specific synonymous fraction for each condition: the fraction of single-nucleotide substitutions at each codon position that preserve the coded amino acid (or RAA class). In the triplet codes (B and C), position 3 absorbs a majority of substitutions as synonymous (68.9 % for SGC, 76.5 % for RAA10 triplet), creating an error- buffering corridor. The doublet code has no dedicated synonymy position (33.3 % at pos 1, 8.3 % at pos 2; pos 3 is absent). N/A = not applicable (condition A has no third position). Raw z-scores are reported; no inter- condition z-score differences (Δz) are computed because z-scores against different null models are not directly comparable.

## Figure 3

See image above for figure legend.

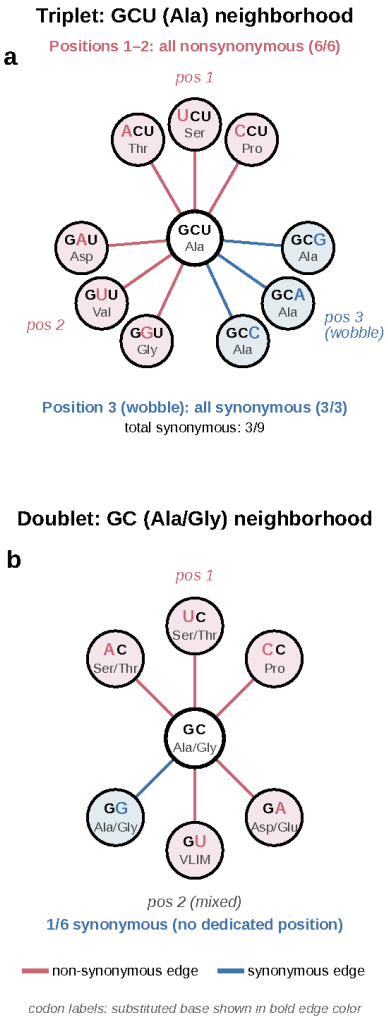



**Figure 4.** Wobble-position synonymy is the mechanism by which triplet architecture buffers single-nucleotide errors. (a) Hamming-1 neighborhood of GCU (Ala) in the SGC. Each neighbor differs from GCU by exactly one base (the substituted base is shown in bold in the edge color). The three position-3 neighbors (GCA, GCC, GCG) all encode Ala and are therefore synonymous with GCU (blue edges). All six position-1 and position-2 neighbors encode different amino acids (red edges). The wobble position converts all three position-3 substitutions into silent changes; positions 1 and 2 convert none, for a neighborhood synonymy rate of 3/9 (33 %). (b) Hamming-1 neighborhood of GC (Ala/Gly), the doublet-projection analogue of GCU under the RAA10 alphabet. The doublet code has six neighbors (three at each of its two positions). Only GG (which the majority-vote projection also assigns to Ala/Gly) is synonymous; the other five neighbors cross class boundaries. Synonymy rate 1/6 (17 %), with no position providing dedicated wobble tolerance. This explains the mechanism behind the architecture effect in Figure 3: the third codon position in triplet codes supplies a systematic error-absorbing channel that has no analogue in doublet codes, regardless of alphabet size.

## Figure 4

See image above for figure legend.

# Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementarymaterialsJME.pdf