Position-2 Chemical Coherence as a Load-Bearing Constraint on the Genetic Code

Forrest Bishop

Independent Researcher

Abstract

The universal genetic code exhibits chemical coherence at codon position 2: U-block codons encode predominantly hydrophobic amino acids, A-block codons encode predominantly polar and charged amino acids. We present three independent categories of evidence that this organization is load-bearing—that disruption imposes measurable costs on translation accuracy, regulatory efficiency, or both. First, aminoacyl-tRNA synthetase class distribution correlates with position-2 nucleotide, ribosome geometry enforces position-2 discrimination, and cells have built functional dependencies (membrane targeting, nitrogen metabolism) on this organization. Second, natural genetic code variants preserve position-2 chemistry at high rates: seven distinct stop-to-sense amino acid targets across two codon types all preserve position-2 block identity ($p < 2.3 \times 10^{-4}$ against a random null; chemistry cannot confound these events because stop codons encode no prior amino acid); three of four sense-to-sense reassignments preserve block chemistry, with the single violation requiring ongoing translational ambiguity and extensive genomic renovation. Third, regulatory gene scaling ($R \sim N^{\gamma}$, $\gamma \approx 1.7–2.0$) creates a complexity ceiling at ~16,000 prokaryotic genes; hierarchical organization by position-2 block identity reduces this overhead. Position-2 chemical coherence is not an incidental pattern but a mechanical constraint: violations are possible but costly, and cumulative disruption is prohibitive.

Introduction

The genetic code partitions into chemical blocks defined by codon position 2. U-block codons (NUN) encode predominantly hydrophobic amino acids: phenylalanine, leucine, isoleucine, methionine, and valine. A-block codons (NAN) encode predominantly polar and charged amino acids: asparagine, aspartate, glutamate, glutamine, histidine, lysine, and tyrosine. C-block codons (NCN) encode a mix including proline, threonine, alanine, and serine, while G-block codons (NGN) encode glycine, arginine, cysteine, and tryptophan. This organization was first recognized by Rumer (1966) and quantified by Woese et al. (1966), who attributed it to error minimization. However, error minimization predicts optimization for mutational robustness but does not require the specific mechanical constraints observed in present-day translation machinery, nor does it explain why independent test cases—organisms with variant genetic codes—systematically respect this organization.

We present an alternative framework: position-2 chemical coherence is strongly constrained by the mechanical and information-processing properties of translation systems, such that violations are costly and cumulative disruption is prohibitive. We use "load-bearing" to mean operationally testable: disrupting position-2 chemical organization should impose measurable costs on translation accuracy, protein targeting, metabolic coordination, or regulatory efficiency. This framework generates specific predictions about present-day systems: aminoacyl-tRNA synthetase class distribution should correlate with position-2 nucleotide, particularly in chemically coherent blocks; ribosome architecture should enforce position-2 discrimination; cells should have built functional machinery that depends on this organization; observed code variants should respect position-2 chemistry, with violations incurring measurable cost; and translation regulatory architecture should organize by position-2 block identity.

The information-processing perspective addresses why block structure is universal. Translation systems face a fundamental scaling problem: regulatory circuits governing amino acid biosynthesis, tRNA charging, and codon usage face combinatorial pressure as proteome complexity increases. Prokaryotic regulatory genes scale as $R \sim N^{\gamma}$ where $\gamma \approx 1.7–2.0$ and N is genome size (van Nimwegen 2003; Molina & van Nimwegen 2009). This creates a ceiling at approximately 16,000 genes where regulatory overhead consumes available genomic capacity. Hierarchical organization by chemistry at position 2 provides a structural basis for block-level regulation rather than requiring exhaustive coordination of all 20 amino acids individually. This paper examines the present-day machinery—ribosome geometry, aaRS architecture, regulatory network structure—to determine whether position-2 organization is an incidental pattern or a load-bearing feature of translation systems.

Results

Mechanical Structure: Translation Machinery Correlates With and Depends on Position-2 Organization

Aminoacyl-tRNA Synthetase Fold Topology and Position-2 Nucleotide

Aminoacyl-tRNA synthetases partition into two structural classes based on catalytic domain architecture. Class I enzymes employ a Rossmann fold and approach tRNA from the minor groove, while Class II enzymes employ an antiparallel β-sheet architecture and approach from the major groove (Eriani et al. 1990). Carter & Wills (2018) demonstrated that urzyme fragments (~130 amino acids) retain substrate specificity, indicating the fold-substrate relationship is intrinsic to the class architecture.

The aaRS class distribution correlates with codon position-2 nucleotide, but the strength of this correlation varies across blocks. U-block shows a 4:1 ratio of Class I to Class II—predominantly Class I recognition. C-block shows a 0:4 ratio—exclusively Class II. A-block shows approximately 3:4 (Class I : Class II), essentially no strong preference. G-block shows approximately 3:1 to 3:2 depending on how multi-block amino acids are counted. The correlation is strongest where the chemical distinction between blocks is most extreme (U-block hydrophobic, C-block polar) and weakest where block chemistry is mixed (A-block and G-block contain amino acids spanning a broader chemical range). This pattern is consistent with a constraint that tracks chemical coherence within blocks rather than block identity per se. The mechanistic link explaining why minor groove approach (Class I) would favor hydrophobic substrates remains unclear. Additionally, some aaRSs (notably AlaRS) recognize tRNA primarily through acceptor stem elements rather than the anticodon. The aaRS class–position-2 correlation represents a strong empirical pattern consistent with the constraint hypothesis while acknowledging these mechanistic gaps and exceptions.

Table 1. Aminoacyl-tRNA synthetase class distribution by position-2 block.

Block	Pos-2	Class I	Class II	Dominant Class	Chemical Character
C	C	0	4	II (exclusive)	Small, framework
U	U	4	1	I	Hydrophobic, core
G	G	3	2	I	Structural extremes
A	A	3	4	II	Polar, charged

Class I aaRS: Arg, Cys, Glu, Gln, Ile, Leu, Met, Trp, Tyr, Val. Class II aaRS: Ala, Asn, Asp, Gly, His, Lys, Phe, Pro, Ser, Thr.

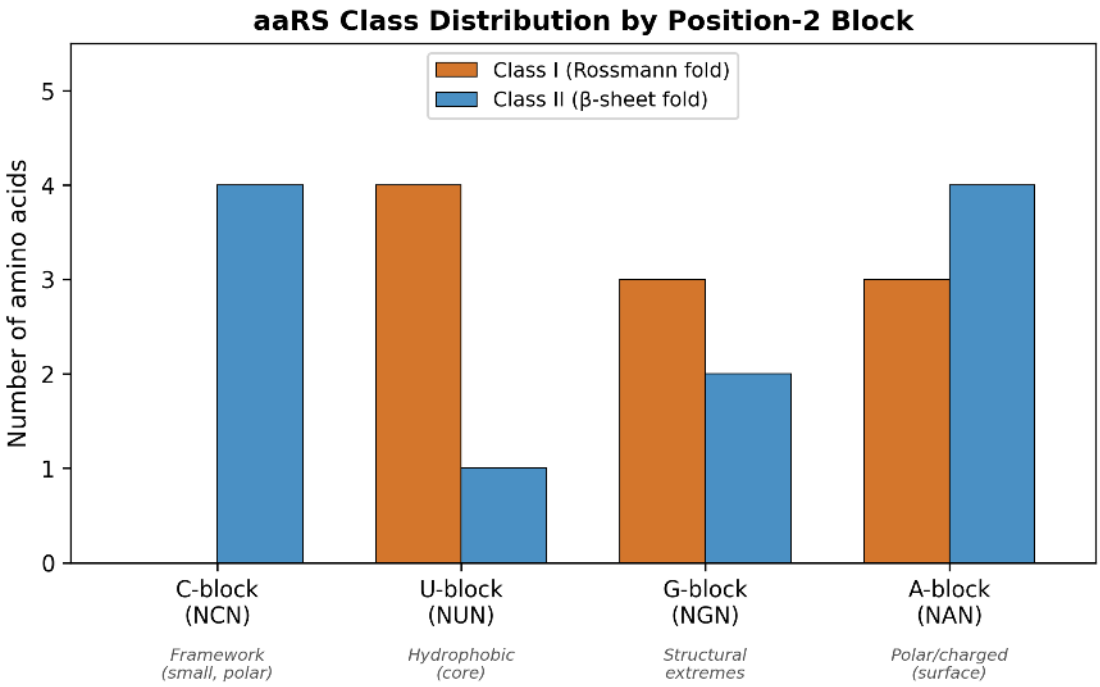

Figure 1. aaRS class distribution by position-2 block. Class I enzymes (Rossmann fold, orange) predominate in the U-block (hydrophobic amino acids) while Class II enzymes (β-sheet fold, blue) predominate in the C-block (framework amino acids) and A-block (polar/charged amino acids). The G-block shows mixed class representation, consistent with its chemically diverse composition.

Ribosome Geometry: Position 2 Versus Position 1

Ribosome geometry requires Watson-Crick pairing at codon positions 1 and 2, while position 3 permits wobble base pairing (Demeshkina et al. 2012). A-site nucleotides G530, A1492, and A1493 inspect the minor groove geometry of the first two codon-anticodon base pairs, rejecting non-Watson-Crick geometries. Position-2 mismatches trigger more robust quality control responses than position-1 mismatches (Rozov et al. 2016), indicating position 2 serves as the primary discrimination point. Structurally, the anticodon loop U-turn motif places position 35 (reading codon position 2) at the apex of the loop, where it is maximally exposed to molecular recognition by any approaching macromolecule (Quigley & Rich 1976).

Both positions 1 and 2 are geometrically constrained to Watson-Crick pairing, but they carry different types of chemical information. Position-2 variation changes amino acid hydrophobicity and charge—the chemical properties relevant to block-level regulatory organization. Position-1 variation correlates with different amino acid properties (biosynthetic family membership, molecular weight), though these correlations are weaker and less systematic than the position-2 hydrophobicity pattern. The regulatory arguments developed in this paper—subset optimization, block-level metabolic coordination, hierarchical circuit organization—depend specifically on chemical property clustering at position 2. Position 2 is the primary axis for regulatory architecture because it carries the chemical information that enables hierarchical organization. This primacy has a structural basis: as noted above, position 35 sits at the apex of the tRNA anticodon U-turn, making position-2 the primary information-bearing axis for chemical identity while position-1 carries secondary information and position-3 permits wobble degeneracy. The code uses a multi-axis system; this paper concerns the axis relevant to regulatory scaling.

Functional Dependency: Membrane Protein Targeting

Integral membrane proteins across Bacteria, Archaea, and Eukaryota exhibit position-2 U-bias in their coding sequences: 50% U content at codon position 2 compared to a 25% genome average (Prilusky & Bibi 2009). Taken alone, this compositional bias is a mathematical consequence of the code table and protein composition—membrane proteins are hydrophobic, the code assigns hydrophobic amino acids to U-block, therefore membrane protein mRNAs are U-enriched at position 2. The compositional statistic, by itself, is not independent evidence that the code table is mechanically constrained.

The non-trivial observation is that cells have built molecular machinery on top of this compositional pattern. U-rich mRNA segments encoding hydrophobic transmembrane helices interact with cold shock proteins, enabling translation-independent targeting to membranes (Benhalevy et al. 2015). This is an RNA-level recognition system that depends on position-2 U-content as a compositional signal. Disrupting U-block hydrophobic coherence would break this targeting pathway. The argument for load-bearing architecture is not the compositional bias per se but the functional dependency: molecular systems have been constructed that rely on the position-2 correlation, making it costly to disrupt.

Functional Dependency: Nitrogen Metabolism Coordination

A-block chemical coherence—where position-2 = A encodes predominantly polar and charged amino acids—enables coordinate regulation of nitrogen metabolism. Six A-block amino acids function as nitrogen donors in biosynthetic pathways: glutamate, glutamine, aspartate, asparagine, lysine, and histidine. All six are polar or charged, consistent with A-block chemical organization. All six are upregulated under nitrogen starvation and form the central nitrogen assimilation network, coordinately regulated through GS/GOGAT and transamination pathways.

The functional significance emerges under nitrogen limitation. Cells must preferentially biosynthesize amino acids that donate nitrogen to other pathways rather than amino acids that consume nitrogen without donation. Nitrogen-donor amino acids (the six A-block amino acids listed above) cluster together by position-2 identity, while nitrogen-sink amino acids occupy different blocks: arginine (consumes 4 nitrogens, position-2 = G), proline (derived from glutamate but does not donate, position-2 = C), tryptophan (consumes 2 nitrogens, position-2 = G). This organization enables block-level metabolic coordination: nitrogen-sensing pathways can coordinately affect biosynthesis and translation of chemically similar amino acids (A-block: polar/charged) that serve related metabolic functions (nitrogen donation).

The nitrogen metabolism example demonstrates functional exploitation of position-2 chemical coherence, but a circularity must be acknowledged: A-block encodes polar and charged amino acids by the structure of the standard code, and nitrogen-donor amino acids are necessarily polar or charged because nitrogen functional groups confer polarity. The clustering of nitrogen donors in A-block may therefore follow automatically from chemistry without requiring any position-2-specific explanation. The non-trivial observation is not the clustering itself but that cells have built coordinate regulatory pathways (GS/GOGAT, transamination networks, nitrogen starvation responses) that exploit this clustering for block-level metabolic coordination. Whether the clustering reflects position-2 organization or chemical necessity, the regulatory dependency on it is real: disrupting A-block chemical coherence would disrupt nitrogen metabolism coordination regardless of why the coherence exists.

Observed Code Flexibility: Natural Variants Test the Constraint Space

Natural genetic code variants provide test cases for what the position-2 constraint space permits. If position-2 chemistry were incidental—a pattern with no mechanical cost to violating—codon reassignments should show no preference for maintaining block boundaries. We surveyed documented

genetic code variants from NCBI tables and literature (2000–2024), identifying 7 distinct stop-to-sense amino acid targets and 4 sense-to-sense reassignments (see Methods for inclusion criteria). Context-dependent readthrough mechanisms (selenocysteine, pyrrolysine) were excluded.

Among documented reassignments, we identified 11 stop-to-sense reassignment events (counting independent phylogenetic origins) involving 7 distinct amino acid targets across 2 codon types, and 4 sense-to-sense reassignments. Because multiple independent origins of the same reassignment (e.g., UAR→glutamine in 5–6 lineages) test the same codon-amino acid pairing, we report statistics on distinct amino acid targets rather than total phylogenetic events to avoid pseudoreplication. Stop-to-sense reassignments show complete position-2 preservation across all 7 distinct targets (7/7, 100%): UAR codons (position-2 = A) reassign to glutamine (5–6 independent origins), lysine (1 origin), glutamic acid (1 origin), and tyrosine (1 origin)—all A-block amino acids. UGA codons (position-2 = G) reassign to tryptophan (3–4 independent origins), cysteine (1 origin), and glycine (1 origin)—all G-block amino acids. Selenocysteine (UGA, context-dependent) is excluded from formal counting as a specialized incorporation mechanism rather than a standard reassignment.

Sense-to-sense reassignments show 3 of 4 preserving position-2 chemistry (75%): AUA isoleucine→methionine preserves U-block hydrophobic character; AGA/AGG arginine→serine preserves the fact that serine already occupies G-block codons (AGY) alongside C-block codons (UCN); AAA lysine→asparagine in echinoderm and flatworm mitochondria preserves A-block polar/charged character (both position-2 = A). CUG leucine→serine violates block chemistry (U-block hydrophobic → dual-block polar).

Table 2. Natural genetic code variant reassignments and position-2 block preservation.

Type	Codon(s)	Pos-2	Source AA	Target AA	Target Block	Preserved?	Independent Origins
Stop→Sense	UAR	A	Stop	Gln	A	Yes	5–6
Stop→Sense	UAR	A	Stop	Lys	A	Yes	1
Stop→Sense	UAR	A	Stop	Glu	A	Yes	1
Stop→Sense	UAR	A	Stop	Tyr	A	Yes	1
Stop→Sense	UGA	G	Stop	Trp	G	Yes	3–4
Stop→Sense	UGA	G	Stop	Cys	G	Yes	1
Stop→Sense	UGA	G	Stop	Gly	G	Yes	1
Sense→Sense	AUA	U	Ile	Met	U	Yes	multiple
Sense→Sense	AGA/ AGG	G	Arg	Ser	G	Yes	multiple
Sense→Sense	AAA	A	Lys	Asn	A	Yes	multiple
Sense→Sense	CUG	U	Leu	Ser	C/G	**No**	1

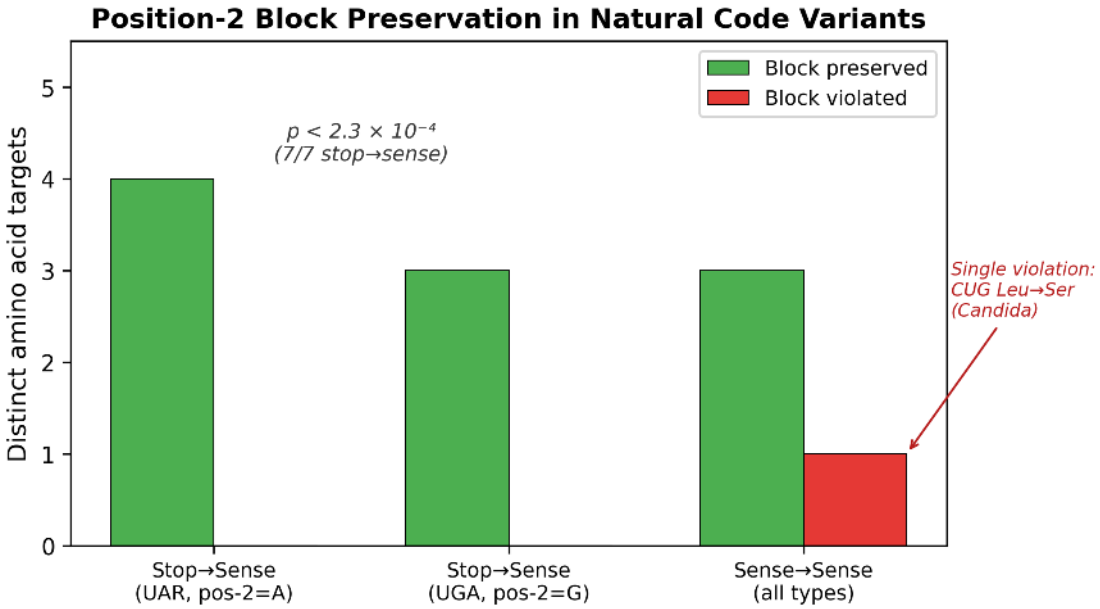

Figure 2. Position-2 block preservation in natural code variants. All seven distinct stop-to-sense amino acid targets preserve position-2 block identity ($p < 2.3 \times 10^{-4}$). Three of four sense-to-sense reassignments preserve block chemistry. The single violation (CUG Leu→Ser in Candida) is the only documented block-crossing sense-to-sense reassignment.

The overall preservation rate is 10/11 (91%) counting distinct amino acid targets (7 stop-to-sense, 4 sense-to-sense). Two null models are required because stop-to-sense and sense-to-sense reassignments face different constraints. Stop-to-sense reassignments provide the primary statistical evidence for position-2 constraint because chemistry cannot confound the analysis, while sense-to-sense reassignments provide a supportive pattern but lack independent statistical power given the small sample size.

For stop-to-sense reassignments, no prior amino acid exists, so chemical similarity cannot confound the analysis. The appropriate null model is random assignment among all 20 amino acids. Under this null, the probability of landing in the same position-2 block ranges from 0.25 (G-block, 5 of 20) to 0.35 (A-block, 7 of 20). Across 7 distinct targets (4 A-block, 3 G-block), the expected number preserving position-2 is 2.15. All 7 preserve position-2 ($p < 2.3 \times 10^{-4}$, Poisson binomial test, exact). Stop-to-sense reassignments provide the cleanest test of position-2 constraint because the confound of chemical similarity is structurally absent.

For sense-to-sense reassignments, chemical similarity between source and target amino acids contributes to position-2 preservation, because the standard code clusters chemically similar amino acids within position-2 blocks. A chemistry-controlled null—asking what fraction of chemically similar amino acids share the source's position-2 block—yields per-reassignment probabilities of 0.63 (Ile→Met, hydrophobic pool), 0.43 (Arg→Ser, polar pool), 0.43 (Lys→Asn, polar pool), and 0.00 (Leu→Ser, no polar amino acids in U-block). Under this null, the expected number preserving position-2 is 1.48 out of 4; the observed 3 out of 4 is consistent with but does not reach significance ($p = 0.11$, $n = 4$). The sense-to-sense sample is too small for independent statistical power. However, the single violation (Leu→Ser) crosses both position-2 block boundaries and chemical class boundaries simultaneously, and it is the only reassignment operating through ambiguous decoding rather than clean codon capture. The pattern indicates position-2 chemistry functions as a constraint on viable reassignments, particularly evident where identical solutions (UAR→glutamine 5–6 times, UGA→tryptophan 3–4 times) are independently discovered while preserving position-2 chemistry.

Independent evidence that reassignment targets are selected rather than random comes from proteome composition analysis. Sengupta & Higgs (2005) analyzed 24 phylogenetically independent mitochondrial codon reassignments and found that the amino acid gaining a codon was selectively favored over the amino acid losing it in each case, with proteome composition shifting to match code changes. Castresana et al. (1998) documented the mechanism in detail for echinoderm mitochondria: AAA reassignment from lysine to asparagine proceeds through a hemichordate intermediate in which the AAA codon is absent entirely, and protein lysine content tracks the code change across the phylogeny. These findings establish that reassignment targets reflect proteome requirements rather than neutral drift, and that the block-preserving pattern documented here operates within a selective framework.

Stop Codon Reassignments

UAA and UAG codons (position-2 = A) have been reassigned in multiple independent lineages. All reassignments assign these codons to A-block amino acids: UAR→glutamine in 5–6 independent origins across ciliates, green algae, and protists; UAR→lysine and glutamic acid in one origin (Oligohymenophorea); UAR→tyrosine in one origin (Mesodinium). UGA codon (position-2 = G) has been reassigned in multiple independent lineages, all to G-block amino acids: UGA→tryptophan in 3–4 independent origins; UGA→cysteine in one origin (Euplotes); UGA→glycine in bacteria. No reassignment places UAR codons into U-block amino acids despite these being common in proteomes. No reassignment places UGA into A-block amino acids despite these being essential for all cells.

The concentration of UAR reassignments on glutamine across 5–6 independent lineages may partly reflect mechanistic constraints on suppressor tRNA availability and release factor competition rather than block structure per se. However, the broader pattern—all stop codon reassignments respecting block boundaries regardless of target amino acid (glutamine, lysine, tyrosine, glutamic acid for UAR; tryptophan, cysteine, glycine for UGA)—is not readily explained by mechanism-specific funneling toward any single target.

Sense Codon Reassignments

Sense-to-sense reassignments provide stronger evidence than stop codon reassignments because they test whether the system can tolerate switching one amino acid chemistry for another. AUA codon reassignment from isoleucine to methionine in vertebrate mitochondria preserves U-block hydrophobic character: both amino acids are hydrophobic and occupy U-block. AGA/AGG codon reassignments from arginine to serine in invertebrate mitochondria preserve the fact that serine already has G-block codons (AGY): both arginine and serine are polar, and the reassignment extends serine within G-block territory. AAA codon reassignment from lysine to asparagine in echinoderm and flatworm mitochondria preserves A-block polar/charged character: both amino acids are polar, both occupy A-block (position-2 = A), and both contain amide or amine functional groups.

The Candida Exception: Cost of Block Crossing

Candida species reassign CUG from leucine (U-block, hydrophobic) to serine (a dual-block amino acid with codons in both C-block and G-block; polar). This is the only documented sense-to-sense reassignment that crosses a position-2 block boundary. The reassignment operates through ambiguous decoding rather than clean codon capture: $tRNA^{Ser}$ with CAG anticodon retains residual leucine identity, maintaining ongoing low-level translational ambiguity at CUG positions. Engineered strains tolerate leucine misincorporation rates of up to 28% at CUG positions without lethality (Santos et al. 2007), though wild-type ambiguity rates are substantially lower. That the lineage survives reflects extensive genomic renovation, not low cost: comparative genomics reveals that approximately 99% of ancestral CUG codons have been erased from Candida genomes (Butler et al. 2009), the organism has upregulated chaperone and stress-response systems, and the remaining CUG positions have been redistributed to

contexts tolerant of ambiguous decoding. Silva et al. (2007) demonstrated that reconstructing this reassignment in *Saccharomyces cerevisiae* blocked mating and locked cells in a diploid state, indicating that obligate diploidy in *C. albicans* buffers against translational errors. Notably, the CUG-to-serine reassignment occurred independently at least three times across the yeast phylogeny, with CUG reassigned to different amino acids in other fungal lineages (Krassowski et al. 2018), demonstrating that this block-crossing event is convergent rather than idiosyncratic to a single lineage. That this remains the only documented block-crossing sense-to-sense reassignment despite thousands of sequenced genomes suggests most lineages cannot pay this cost.

This single case cannot establish a general rule that all block-crossing violations are comparably costly. It establishes that this particular violation required genome-wide renovation—codon erasure, chaperone upregulation, ploidy buffering—to survive. Combined with the absence of any other documented block-crossing sense-to-sense reassignment, the pattern is consistent with block boundaries imposing a cost function: crossing is possible but expensive, and the cost explains why observed reassignments overwhelmingly preserve block chemistry.

Mitochondrial Reduction: Block Structure Under Maximal Pressure

Mitochondrial genomes represent extreme reduction: human mitochondria encode only 13 proteins, all hydrophobic respiratory chain subunits. Despite minimal genetic content, mitochondria retain translation machinery for all 20 amino acids, employing 19 aminoacyl-tRNA synthetases (nuclear-encoded, imported): 17 mitochondria-specific enzymes and 2 dual-localized between cytosol and mitochondria (GlyRS, LysRS). Under maximal pressure for consolidation, a single consolidation event occurs, and it preserves block structure.

Glutamine lacks a dedicated mitochondrial synthetase. Glutamyl-tRNA synthetase charges $tRNA^{Gln}$ with glutamate, then transamidation converts glutamate to glutamine on the tRNA (Nagao et al. 2009). Glutamate codons (GAR) and glutamine codons (CAR) both have position-2 = A. Both amino acids are polar and charged. Both differ by a single amide group and participate in the GS/GOGAT cycle.

This consolidation is consistent with block preservation, but an alternative explanation must be acknowledged: Glu→Gln transamidation is a single enzymatic step and one of the most biochemically straightforward indirect aminoacylation pathways known. It exists across bacteria, archaea, and mitochondria. The alternative consolidation—glutamine with asparagine (both polar, both contain amide groups)—would require an entirely different enzymatic pathway not known to exist in any organism. The consolidation choice may reflect enzymatic accessibility rather than block-structure preservation. The observation that bears on block structure: no cross-block consolidation exists in any mitochondrial system despite maximal selection pressure for genome reduction. If enzymatic accessibility were the sole determinant, some cross-block pathways should be comparably easy. The absence of cross-block consolidation is the data point; the Glu→Gln mechanism alone is ambiguous.

Mitochondrial Codon Usage: Subset Functionality

Mitochondria demonstrate that block structure enables subset functionality. The mitochondrial proteome consists exclusively of hydrophobic membrane proteins. Codon usage shows strong bias toward U-block codons: leucine, isoleucine, valine, phenylalanine, and methionine account for the majority of mitochondrial codon usage, while A-block codons appear at low frequency. Mitochondrial genomes encode only 22 tRNAs rather than the 32–45 typical of bacteria. Nuclear genes encoding mitochondrial translation factors number approximately 100–200, far fewer than the several hundred required for cytoplasmic translation, because regulatory complexity is reduced when the proteome is chemically homogeneous.

A genetic code where hydrophobic amino acids are distributed randomly across all position-2 values would prevent this simplification. An organism with a hydrophobic proteome would still require codons from all four blocks, could not focus its tRNA pool, could not simplify aaRS diversity, and could not implement position-based regulation. The hierarchical organization would exist in principle but provide no practical benefit because the chemical types would not be coherently organized within the hierarchy. Mitochondrial code modifications reinforce this: AUA reassigned from isoleucine to methionine (both hydrophobic, both U-block) and UGA reassigned from stop to tryptophan (hydrophobic, maintaining U-block coherence) both preserve block chemistry.

Information-Processing Architecture: Regulatory Scaling and Hierarchical Organization

Empirical Scaling Law

Prokaryotic regulatory genes scale super-linearly with genome size. Van Nimwegen (2003) first measured regulatory gene content across fully sequenced prokaryotic genomes and discovered that transcriptional regulatory genes scale faster than linearly with total gene count. Molina & van Nimwegen (2009) confirmed this across 682 prokaryotic genomes spanning 13 distinct clades: regulatory gene count R scales as $R \sim N^{\gamma}$ where N is total gene count and $\gamma \approx 1.7–2.0$, with best fit near $\gamma = 2$ indicating approximately quadratic scaling. This relationship holds across phylogenetically distant clades and organisms with radically different lifestyles, indicating that the scaling law reflects a structural property of regulatory networks.

The quantitative pattern is striking. Buchnera aphidicola, with 504 total genes, encodes only 3 transcription factors representing 0.6% of its genome. Escherichia coli, with 4,300 genes, encodes approximately 300 transcription factors representing 7% of its genome. Burkholderia species with 7,717 genes encode 801 transcription factors representing 10.4% of the genome. Across a 15-fold increase in genome size, the absolute number of transcription factors increases 267-fold, and the fraction of the genome devoted to regulation increases 17-fold.

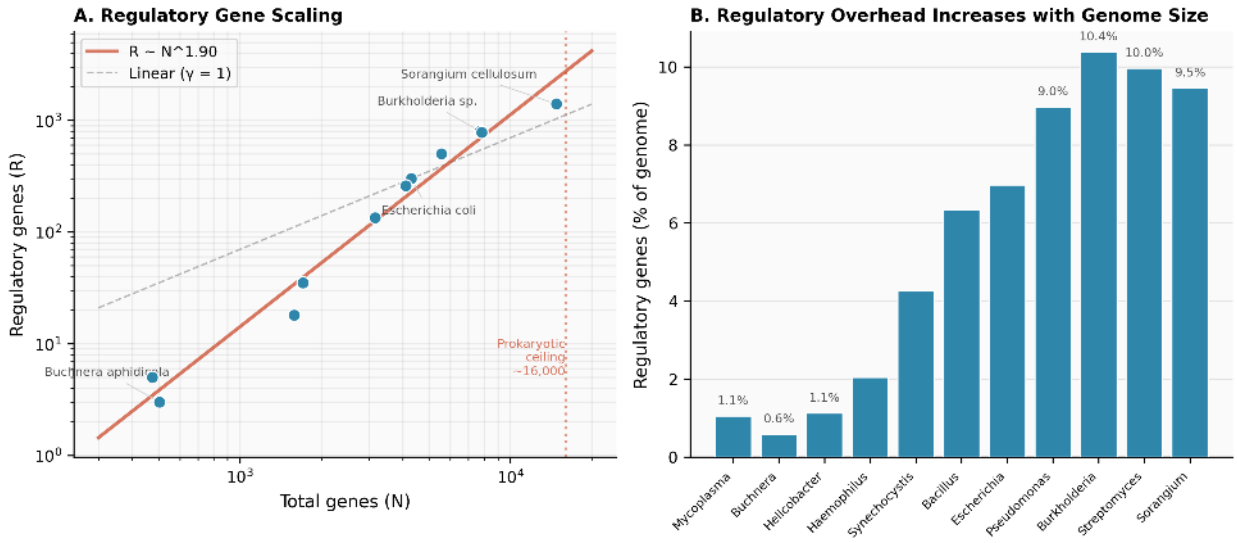

Figure 3. Super-linear scaling of regulatory genes in prokaryotes. Regulatory gene count R scales as $R \sim N^{\gamma}$ where $\gamma \approx 1.7–2.0$, creating a ceiling at approximately 16,000 genes where regulatory overhead consumes available genomic capacity. Data points from van Nimwegen (2003) and Molina & van Nimwegen (2009).

The Prokaryotic Genome Ceiling

Super-linear scaling creates a ceiling on genome complexity. If $R \sim N^2$, then adding genes eventually requires more regulatory capacity than the new genes provide in functional benefit. The ceiling occurs at approximately 15,000–20,000 genes where regulatory overhead exceeds sustainable proportions of genomic capacity (Ahnert et al. 2008). No prokaryote exceeds approximately 16,000 genes—the largest known prokaryotic genome is Sorangium cellulosum strain So0157-2 at approximately 14,800 genes. Thousands of prokaryotic genomes have been sequenced across diverse environments, and none approach the theoretical maximum.

Konstantinidis & Tiedje (2004) measured gene category scaling across 115 prokaryotic genomes. As genome size increases, regulatory genes increase as a fraction of the genome ($R^2 > 0.52$), while translation and replication genes each decrease as a fraction ($R^2 > 0.4$). Larger genomes allocate increasing fraction to regulation and decreasing fraction to core molecular processes. This pattern is incompatible with alternative ceiling hypotheses (replication time, membrane surface area, energy constraints). Regulatory overhead is the limiting factor.

Mechanistic Basis for Quadratic Scaling

The quadratic scaling reflects combinatorial pressure in control systems. When a new gene is added to a genome containing N genes, it potentially interacts regulatorily with any of the existing N genes. The space of possible regulatory interactions scales as $N(N-1)/2$, which is $O(N^2)$. If each new gene explores this interaction space with some probability p that an interaction will require dedicated regulation, then the expected number of new regulators per gene added is $\Delta R = p \times N$. Integrating gives $R \propto N^2$. This is an information-theoretic property of control systems: any system where components can interact pairwise and where those interactions require control faces this scaling pressure.

Transcriptional Versus Translational Regulation

The van Nimwegen scaling law was measured for transcriptional regulatory genes—primarily transcription factors governing gene expression. The scaling has not been independently measured for translation-specific regulation. While translation regulation likely faces analogous combinatorial pressure due to comparable interaction complexity, the quantitative scaling relationship for translation-specific regulators remains to be measured. Translation systems coordinate 20 amino acid biosynthesis pathways, 20 or more aminoacyl-tRNA synthetases, 45 or more tRNA species, and multiple quality control systems, all responding to nutritional state, growth conditions, and proteome requirements. The pairwise interaction space is comparable. However, translation regulation may employ different circuit architectures than transcriptional regulation, and the quantitative scaling may differ. The analysis that follows examines the actual regulatory architecture of E. coli's amino acid control systems to determine whether position-2 block identity correlates with real regulatory organization.

Real Regulatory Organization in E. coli

Translation regulation in E. coli operates through multiple overlapping systems organized hierarchically by chemical class and metabolic function. The stringent response provides global amino acid starvation sensing through ppGpp accumulation (Hauryliuk et al. 2015; Potrykus & Cashel 2008), affecting all amino acid biosynthesis simultaneously. Individual amino acid biosynthesis operons (trp, leu, ilv, his) respond to specific amino acid availability through attenuation and repression mechanisms (Henkin & Yanofsky 2002). Regulons coordinate chemically or metabolically related amino acids: the branched-chain amino acid regulon (ilvGMEDA) coordinates leucine, isoleucine, and valine biosynthesis (Umbarger 1978); the aromatic amino acid regulon coordinates phenylalanine, tyrosine, and tryptophan pathways (Yanofsky 2000). tRNA charging feedback operates through uncharged tRNA accumulation

triggering RelA-dependent ppGpp synthesis. This architecture organizes regulatory decisions at multiple levels rather than through exhaustive pairwise interactions between all 20 amino acids.

Position-2 Block Correlation in Observed Regulons

Natural regulatory groupings correlate with position-2 block identity. The branched-chain amino acid regulon coordinates leucine (UUR/CUN, position-2 = U), isoleucine (AUY, position-2 = U), and valine (GUN, position-2 = U)—all three are U-block amino acids, all hydrophobic, all share biosynthetic enzymes. Nitrogen metabolism coordinates glutamate, glutamine, aspartate, asparagine, lysine, and histidine—six amino acids, all A-block (position-2 = A), all polar or charged, all nitrogen donors. The aromatic amino acid regulon coordinates phenylalanine (UUY, position-2 = U), tyrosine (UAY, position-2 = A), and tryptophan (UGG, position-2 = G)—spanning three position-2 blocks. This cross-block regulon reflects shared biosynthetic pathway origin (the shikimate pathway) rather than shared chemistry; it represents regulation organized by metabolic origin rather than by chemical class. Regulatory organization is multi-hierarchical: position-2 provides one organizational axis (chemical class), biosynthetic pathways provide another (metabolic origin). The aromatic regulon demonstrates that metabolic hierarchy can override position-2 when pathway sharing is strong, consistent with position-2 providing block-level organization within a broader regulatory landscape rather than being the sole organizing principle. These groupings reflect the multi-axis nature of translation regulation: amino acids sharing a position-2 nucleotide tend to share chemistry and can be coordinated at the block level, while amino acids sharing biosynthetic pathways are coordinated at the family level regardless of block identity.

Position-2 Composition Enables Block-Level Regulatory Mechanisms

Position-2 composition enables mRNA-level regulatory mechanisms operating on chemical class. Membrane protein mRNAs show 50% U-content at position 2 (Prilusky & Bibi 2009), enabling compositional recognition for protein targeting. Nitrogen starvation triggers coordinate upregulation of A-block amino acids through metabolic sensing pathways detecting nitrogen availability. Codon usage bias in highly expressed genes reflects tRNA availability, which varies by position-2 nucleotide due to wobble base pairing rules affecting the third position but not the second. Ribosome pausing at rare codons depends on tRNA availability for specific position-2 classes. Block structure provides a mechanistic basis for regulatory coordination: decisions affecting chemical class (hydrophobic versus polar) operate through position-2 composition rather than requiring individual amino acid discrimination.

Hierarchical Organization Reduces Regulatory Complexity

The observed regulatory architecture demonstrates hierarchical organization that reduces complexity relative to flat pairwise regulation. Rather than regulating 20 amino acids through independent circuits for each, cells employ a multi-level architecture: global starvation response (the stringent response, a single circuit sensing overall amino acid availability); block-level chemical coordination (U-block for branched-chain hydrophobic amino acids, A-block for nitrogen donors); family-level metabolic coordination (operons and regulons grouping biochemically related amino acids); and individual amino acid fine-tuning (specific operons with attenuation). This hierarchy enables appropriate-level responses: nitrogen limitation affects A-block amino acids as a class rather than requiring six independent decisions for glutamate, glutamine, aspartate, asparagine, lysine, and histidine individually. Block structure by chemistry at position 2 provides the organizational scheme that enables this hierarchical architecture.

Scaling Law Implications

Prokaryotic regulatory genes scale super-linearly because regulatory circuit complexity grows faster than proteome size (van Nimwegen 2003; Molina & van Nimwegen 2009). Without hierarchical organization, adding genes requires additional regulatory machinery to prevent unwanted interactions and enable

coordinated responses. Flat regulatory architectures where every gene potentially interacts with every other gene scale as N^2, creating a ceiling where regulatory overhead consumes genomic capacity. Hierarchical organization reduces effective N by partitioning genes into groups regulated at appropriate levels. Position-2 block structure provides this partitioning for translation systems: amino acids group by chemistry, enabling block-level regulatory decisions for chemical-class coordination and family-level decisions for metabolic pathway coordination. This reduces the growth rate of regulatory complexity, delaying the point where regulatory overhead becomes limiting.

Information-Theoretic Framework

Standard information theory provides a framework for understanding why hierarchical organization reduces regulatory overhead. In a flat architecture regulating N components with pairwise interactions, the regulatory state space scales as $O(N^2)$. Hierarchical organization compresses this: if components are organized into K groups, the state space reduces to $O(K^2 + K \times M^2)$ where M is average group size. For the genetic code with $K = 4$ position-2 blocks and $M \approx 5$ amino acids per block, this represents a substantial reduction. The quantitative savings depend on biological details (how many pairwise interactions are actually regulated, what precision is required), but the scaling advantage of hierarchical over flat architecture is a general property of control systems, not specific to the genetic code.

A complementary observation concerns the compressibility of the code itself. A genetic code with random amino acid distribution requires specifying each of 64 assignments independently. The standard code can be partially compressed: specify that position-2 = U encodes hydrophobic amino acids, then specify the particular amino acid from positions 1 and 3. This compressibility directly reflects block structure. Compressible patterns are more robust to perturbation than incompressible ones, because the underlying rule persists even when individual elements change. The universality of block structure across all known genetic codes is consistent with a stable, low-complexity organizational state.

Discussion

Three Categories of Evidence Converge on Position 2 as Load-Bearing

We have presented three categories of evidence bearing on whether position-2 chemical coherence is load-bearing in translation systems. These categories are logically independent. Mechanical structure (aaRS fold topology, ribosome geometry, functional dependencies in membrane targeting and nitrogen metabolism) describes present-day machinery whose operation correlates with position-2 organization. Observed code flexibility (natural code variants, mitochondrial reduction) describes test cases that reveal what the constraint space permits and what it penalizes. Information-processing architecture (regulatory scaling, hierarchical circuit organization) describes the structural basis for why block-level organization reduces regulatory overhead.

Within each category, some observations are causally connected rather than independent. The aaRS fold distribution shapes the code table, which creates the block structure that cells then exploit for membrane targeting and nitrogen coordination. These are causally downstream of each other, not independent measurements. Similarly, code variants and mitochondrial reduction both test the flexibility of codon assignments, differing in context but probing the same underlying question. We claim three independent lines of argument, not five or more independent experiments.

Chemistry and Position-2: Inseparable by Design

An important objection holds that chemical similarity, not position-2 identity, drives reassignment targets: because the standard code clusters chemistry by position-2, preserving chemistry automatically preserves position-2, and position-2 may be along for the ride. This objection treats chemistry and position-2 as

separable variables. They are not. Position-2 identity is how chemistry is organized in the genetic code. Asking whether chemistry or position-2 constrains reassignments is analogous to asking whether weight or the structural wall constrains a building—the wall is the means by which weight is distributed. This inseparability is not a flaw in the analysis; it is the thesis. The claim is not that position-2 constrains independently of chemistry, but that the code's organization of chemistry by position-2 is load-bearing: molecular systems depend on it, and disruption is costly.

The stop-to-sense reassignments partially resolve the ambiguity. Stop codons encode no amino acid, so chemical similarity to a prior assignment cannot explain target selection. Yet all 7 distinct stop-to-sense amino acid targets respect position-2 block identity ($p < 2.3 \times 10^{-4}$). For sense-to-sense reassignments, chemistry and position-2 cannot be fully disentangled with current sample sizes ($n = 4$). The prediction is clear: as additional code variants are discovered, block-crossing sense-to-sense reassignments should remain rare relative to block-preserving ones, and should incur measurable cost where they occur.

Strongly Constrained, Not Absolutely Required

The evidence indicates that position-2 organization is strongly constrained rather than absolutely required. Candida is viable. Serine splits across blocks in the standard code. These are exceptions that survive. However, Candida's violation required erasure of 99% of ancestral CUG codons, upregulated chaperone systems, and obligate diploidy to buffer the ongoing translational ambiguity (Butler et al. 2009; Silva et al. 2007). Serine's split is forced by arithmetic constraints—it needs six codons and no single block accommodates this without splitting.

The appropriate framing is a cost function rather than a hard wall. Violations are possible but costly, and the cost should scale with the degree of violation. An organism tolerating one block-crossing reassignment with high error penalty is Candida. An organism with multiple such crossings would face cumulative cost—compounded error rates, loss of subset optimization capability, degraded block-level regulatory coordination—that is presumably prohibitive. The prediction is that organisms with more block violations should show measurably higher regulatory overhead or compensatory mechanisms. This is testable.

This framework is falsifiable. A block-crossing sense-to-sense reassignment tolerated without measurable compensatory cost — no codon erasure, no chaperone upregulation, no ploidy buffering, no elevated error rates — would contradict the constraint hypothesis. The prediction is specific: compensatory cost should scale with the degree of chemical class violation. A reassignment between chemically adjacent blocks should impose less cost than one between chemically extreme blocks. Discovery of an organism carrying multiple block-crossing reassignments without elevated regulatory overhead would falsify the load-bearing claim entirely.

The Code is Mechanically Constrained, Not Historically Frozen

Crick's frozen accident hypothesis proposed that the genetic code is universal because any change would be lethal—the code is frozen by historical contingency. The data on present-day systems contradict this. Codon assignments can change—numerous documented reassignments across multiple independent lineages demonstrate this. What cannot change cheaply is the position-2 chemical organization: reassignments preserving block boundaries succeed through clean mechanistic pathways, while reassignments violating block boundaries require costly workarounds.

The distinction is between constraint on the architecture versus constraint on all possible changes. All observed test cases—organisms with variant codes, mitochondria under reduction pressure, the Candida exception—respect the same boundary. Any future observed code variant is predicted to respect position-2 chemistry, or to incur measurable cost where it does not.

Error Minimization: Quality Metric Versus Architectural Constraint

Error minimization remains the dominant explanation for genetic code organization. It proposes that codes minimizing the functional impact of point mutations are selectively favored, producing chemical similarity clustering. Error minimization predicts that chemically similar amino acids should share similar codons. It does not predict the specific mechanical features documented here: aaRS fold topology correlating with position-2 nucleotide, cells building targeting machinery on position-2 composition, observed code variants systematically respecting block boundaries with costly violations where they do not, or super-linear regulatory scaling whose overhead is reduced by hierarchical organization. Moreover, the standard code is not particularly outstanding in terms of error minimization: Koonin & Novozhilov (2009) showed that the code sits approximately halfway along an upward trajectory from a random code toward its local optimum, achieving a minimization percentage of roughly 70–78%, with many taller peaks existing in the fitness landscape. The code occupies a mediocre local peak rather than a global optimum, consistent with error minimization being a consequence of the architecture rather than the force shaping it.

The key distinction: error minimization is a quality metric that could be satisfied through multiple architectural solutions. Any code placing chemically similar amino acids in adjacent codon space would achieve error minimization regardless of whether the similarity clustering maps to position-2 blocks specifically. Position-2 mechanical constraints describe an architectural feature of the present-day translation system—specific properties of specific molecular machines. The genetic code satisfies both error minimization and the mechanical constraints. The evidence presented here indicates that the mechanical properties of translation machinery constrain the solution space, and error minimization is a beneficial property of the resulting organization.

Implications for Code Structure

Our findings constrain what any viable genetic code must look like given the mechanical properties of contemporary translation machinery. Any code using standard ribosome-aaRS-tRNA mechanics must satisfy: aaRS class distribution correlating with substrate chemistry and position-2 nucleotide; ribosome geometry enforcing Watson-Crick pairing at position 2; sufficient chemical coherence at position 2 to support the functional dependencies (membrane targeting, metabolic coordination) that present-day systems rely on; hierarchical organization compatible with sustainable regulatory overhead.

The intersection of these constraints narrows the space of viable codes. This paper takes no position on the origin of this configuration. It establishes that the configuration is strongly constrained by the properties of present-day machinery, that violations are costly, and that the architectural features serve identifiable functions in information processing.

Conclusion

Position-2 chemical coherence is a load-bearing feature of the genetic code. Three independent categories of evidence—mechanical structure, observed code flexibility, and information-processing architecture—converge on this conclusion. Aminoacyl-tRNA synthetase geometry correlates with position-2 nucleotide, with the strongest correlation at the blocks with most extreme chemical distinction, while acknowledging mechanistic gaps in explaining why specific fold geometries favor specific chemistries. Ribosome architecture enforces position-2 Watson-Crick pairing while position 2 carries the chemical information relevant to regulatory organization. Cells have constructed functional machinery—membrane protein targeting, nitrogen metabolism coordination—that depends on this organization.

Natural code variants respect position-2 block boundaries at high rates. Among stop-to-sense reassignments, 7 distinct amino acid targets across 2 codon types all preserve position-2 block identity (p

$< 2.3 \times 10^{-4}$); because stop codons encode no prior amino acid, chemical similarity cannot confound these observations. Among sense-to-sense reassignments, 3 of 4 preserve block chemistry; the sample is too small for independent statistical power, but the single violation required genome-wide renovation including 99% codon erasure and obligate diploidy to buffer ongoing translational ambiguity. Mitochondrial genome reduction preserves block structure under maximal consolidation pressure, though the specific consolidation pathway (Glu→Gln transamidation) may reflect enzymatic accessibility as well as block preservation; the absence of any cross-block consolidation is the stronger observation.

Present-day translation regulatory architecture in E. coli demonstrates hierarchical organization that maps to position-2 block identity. Branched-chain amino acids coordinate within U-block. Nitrogen-donor amino acids coordinate within A-block. The stringent response provides global sensing across all blocks. Block structure provides the organizational scheme enabling regulatory decisions at the appropriate level. Regulatory genes scale super-linearly with genome size, creating a ceiling at approximately 16,000 genes; hierarchical organization by position-2 chemistry reduces the growth rate of regulatory complexity.

The evidence distinguishes between two classes of genetic code properties. Quality metrics—error minimization, translational efficiency, codon usage optimization—could be satisfied through multiple architectural solutions. Architectural constraints—aaRS geometry, ribosome structure, functional dependencies, regulatory scaling requirements—restrict the solution space. Position-2 organization satisfies both. The constraint analysis indicates that the mechanical properties of present-day translation machinery impose specific requirements on code organization, and the observed code satisfies them.

The code is not historically frozen. It is mechanically constrained. Codon assignments can change—numerous documented reassignments across independent lineages demonstrate this—but changes must preserve position-2 chemistry or incur substantial cost. Any future observed code variant is predicted to respect this boundary, or to exhibit compensatory mechanisms where it does not. Organisms with weakened block structure should require more regulatory machinery to achieve equivalent control. Codon usage in metabolically coordinated genes should show position-2 compositional bias. Translation regulatory circuits should organize by block identity. Each prediction tests whether block structure functions as load-bearing architecture.

Position-2 chemical coherence is not one organizational feature among many. It is the architectural feature of the genetic code that enables hierarchical regulation of translation systems. The universal code represents the mechanically constrained solution space for scalable translation using ribosome-aaRS-tRNA machinery. Alternatives violating these constraints fail not because change is impossible, but because the constraints impose escalating cost that restricts viable configurations to a narrow region of code space.

Methods

Genetic Code Variant Survey

Documented genetic code variants were surveyed from NCBI genetic code tables (https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi) and literature 2000–2024. Approximately 15–20 natural variants exist with characterized codon reassignments; of these, 11 were classified as distinct reassignment events (7 distinct stop-to-sense amino acid targets and 4 sense-to-sense reassignments). Remaining variants were excluded because they involve context-dependent readthrough (e.g., selenocysteine, pyrrolysine) rather than standard reassignment, or because the reassignment is disputed or insufficiently characterized. For each included variant: reassigned codon(s); position-2 nucleotide identity; source amino acid and target amino acid; chemical properties (hydrophobic, polar, charged); block assignment based on position 2; number of independent phylogenetic origins; mechanism

(clean reassignment versus ambiguous decoding versus context-dependent). Independent origins were identified when identical reassignments occurred in phylogenetically distinct lineages.

Genetic Code Variant Classification

Reassignments were classified as stop-to-sense (7 distinct amino acid targets) or sense-to-sense (4 identified). Stop-to-sense reassignments are distinguished because stop codons encode no amino acid, eliminating chemical similarity as an explanation for block preservation. Multiple independent phylogenetic origins of the same reassignment (e.g., UAR→glutamine in 5–6 lineages) were documented but not counted as independent data points for statistical testing, to avoid pseudoreplication. Preservation rates were calculated on distinct amino acid targets.

Null Model Analysis

Two null models were applied. For stop-to-sense reassignments (where no prior amino acid exists), a random null was used: each target amino acid was assumed equally likely among all 20 amino acids, yielding per-reassignment block-match probabilities of 0.25–0.35 depending on block size. For sense-to-sense reassignments (where chemical similarity to the source amino acid may constrain target selection), a chemistry-controlled null was used: each target was assumed drawn randomly from amino acids sharing the same broad chemical class (hydrophobic, polar, or charged) as the observed target, yielding per-reassignment block-match probabilities of 0.00–0.63. Significance was assessed using the Poisson binomial test (exact), which accounts for heterogeneous per-trial probabilities.

Aminoacyl-tRNA Synthetase Analysis

AaRS class assignments from Eriani et al. (1990) and updated classifications. For each amino acid: cognate codons identified; position-2 nucleotide(s) determined; aaRS class (I or II) identified; class distribution by position-2 block calculated. Class I aaRSs: Arg, Cys, Glu, Gln, Ile, Leu, Met, Trp, Tyr, Val. Class II aaRSs: Ala, Asn, Asp, Gly, His, Lys, Phe, Pro, Ser, Thr. Ratios calculated as (Class I count): (Class II count) for each position-2 block.

Mitochondrial Aminoacyl-tRNA Synthetase Inventory

Human mitochondrial aminoacyl-tRNA synthetase inventory compiled from Wei et al. (2019) and primary literature. For each of 20 amino acids: presence of dedicated mitochondrial aaRS; dual-localization (cytosolic + mitochondrial); indirect aminoacylation pathways; aaRS class; position-2 nucleotide of cognate codons. Consolidation events analyzed for pathway mechanism, block preservation, and metabolic relationship.

Nitrogen Metabolism Analysis

Amino acid classification by nitrogen metabolism role compiled from EcoSal Plus reviews, stringent response literature, and nitrogen starvation studies. Amino acids classified as nitrogen donors, nitrogen sinks, or neutral. Position-2 composition calculated for each category.

Regulatory Scaling Analysis

Prokaryotic regulatory scaling data from van Nimwegen (2003), Molina & van Nimwegen (2009), Konstantinidis & Tiedje (2004), and Ahnert et al. (2008). Scaling exponent γ and genome ceiling calculations from published datasets. Gene category scaling from Konstantinidis & Tiedje measurements across 115 prokaryotic genomes. E. coli regulatory architecture from EcoCyc database (Keseler et al. 2021), amino acid regulon literature, and stringent response reviews.

AI Assistance Statement

The author used Claude (Anthropic) as a research tool for literature verification, statistical analysis, and manuscript preparation. All scientific claims, interpretations, and conclusions are the author's own. The author takes full responsibility for the content.

References

Ahnert, S. E., Fink, T. M., & Zinovyev, A. (2008). Is prokaryotic complexity limited by accelerated growth in regulatory overhead? *arXiv preprint* arXiv:0808.3597.

Benhalevy, D., Biran, I., Bochkareva, E. S., Sorek, R., & Bibi, E. (2015). Evidence for a cytoplasmic pool of ribosome-free mRNAs encoding inner membrane proteins in *Escherichia coli*. *PLoS One*, 10(1), e0115139.

Butler, G., et al. (2009). Evolution of pathogenicity and sexual reproduction in eight *Candida* genomes. *Nature*, 459(7247), 657–662.

Carter, C. W., & Wills, P. R. (2018). Interdependence, reflexivity, fidelity, impedance matching, and the evolution of genetic coding. *Molecular Biology and Evolution*, 35(2), 269–286.

Castresana, J., Feldmaier-Fuchs, G., & Pääbo, S. (1998). Codon reassignment and amino acid composition in hemichordate mitochondria. *Proceedings of the National Academy of Sciences*, 95(7), 3703–3707.

Crick, F. H. (1968). The origin of the genetic code. *Journal of Molecular Biology*, 38(3), 367–379.

Demeshkina, N., Jenner, L., Westhof, E., Yusupov, M., & Yusupova, G. (2012). A new understanding of the decoding principle on the ribosome. *Nature*, 484(7393), 256–259.

Eriani, G., Delarue, M., Poch, O., Gangloff, J., & Moras, D. (1990). Partition of tRNA synthetases into two classes based on mutually exclusive sets of sequence motifs. *Nature*, 347(6289), 203–206.

Hauryliuk, V., Atkinson, G. C., Murakami, K. S., Tenson, T., & Gerdes, K. (2015). Recent functional insights into the role of (p)ppGpp in bacterial physiology. *Nature Reviews Microbiology*, 13(5), 298–309.

Henkin, T. M., & Yanofsky, C. (2002). Regulation by transcription attenuation in bacteria: how RNA provides instructions for transcription termination/antitermination decisions. *BioEssays*, 24(8), 700–707.

Keseler, I. M., et al. (2021). The EcoCyc database in 2021. *Frontiers in Microbiology*, 12, 711077.

Konstantinidis, K. T., & Tiedje, J. M. (2004). Trends between gene content and genome size in prokaryotic species with larger genomes. *Proceedings of the National Academy of Sciences*, 101(9), 3160–3165.

Koonin, E. V., & Novozhilov, A. S. (2009). Origin and evolution of the genetic code: the universal enigma. *IUBMB Life*, 61(2), 99–111.

Krassowski, T., Coughlan, A. Y., Shen, X.-X., Zhou, X., Kominek, J., Opulente, D. A., Riley, R., Grigoriev, I. V., Maheshwari, N., Shields, D. C., Kurtzman, C. P., Hittinger, C. T., Rokas, A., & Wolfe, K. H. (2018). Evolutionary instability of CUG-Leu in the genetic code of budding yeasts. *Nature Communications*, 9, 1887.

Molina, N., & van Nimwegen, E. (2009). Scaling laws in functional genome content across prokaryotic clades and lifestyles. *Trends in Genetics*, 25(6), 243–247.

Nagao, A., Suzuki, T., Katoh, T., Sakaguchi, Y., & Suzuki, T. (2009). Biogenesis of glutaminyl-mt tRNA^{Gln} in human mitochondria. *Proceedings of the National Academy of Sciences*, 106(38), 16209–16214.

Potrykus, K., & Cashel, M. (2008). (p)ppGpp: still magical? *Annual Review of Microbiology*, 62, 35–51.

Prilusky, J., & Bibi, E. (2009). Studying membrane proteins through the eyes of the genetic code revealed a strong uracil bias in their coding mRNAs. *Proceedings of the National Academy of Sciences*, 106(16), 6662–6666.

Quigley, G. J., & Rich, A. (1976). Structural domains of transfer RNA molecules. *Science*, 194(4267), 796–806.

Rozov, A., Demeshkina, N., Westhof, E., Yusupov, M., & Yusupova, G. (2016). New structural insights into translational miscoding. *Trends in Biochemical Sciences*, 41(9), 798–814.

Rumer, I. B. (1966). On codon systematization in the genetic code. *Doklady Akademii Nauk SSSR*, 167(6), 1393–1394.

Santos, M. A., Moura, G., Massey, S. E., & Tuite, M. F. (2007). Driving change: the evolution of alternative genetic codes. *Trends in Genetics*, 23(12), 630–637.

Sengupta, S., & Higgs, P. G. (2005). Pathways of genetic code evolution in mitochondrial systems. *Journal of Molecular Evolution*, 60(4), 499–507.

Silva, R. M., Paredes, J. A., Moura, G. R., Manadas, B., Lima-Costa, T., Rocha, R., Miranda, I., Gomes, A. C., Koerkamp, M. J. G., Perrot, M., Holstege, F. C. P., Boucherie, H., & Santos, M. A. S. (2007). Critical roles for a genetic code alteration in the evolution of the genus *Candida*. *EMBO Journal*, 26(21), 4555–4565.

Umbarger, H. E. (1978). Amino acid biosynthesis and its regulation. *Annual Review of Biochemistry*, 47, 533–606.

van Nimwegen, E. (2003). Scaling laws in the functional content of genomes. *Trends in Genetics*, 19(9), 479–484.

Wei, N., Zhang, Q., & Yang, X. L. (2019). Neurodegenerative Charcot-Marie-Tooth disease as a case study to decipher novel functions of aminoacyl-tRNA synthetases. *Journal of Biological Chemistry*, 294(14), 5321–5339.

Woese, C. R., Dugre, D. H., Saxinger, W. C., & Dugre, S. A. (1966). The molecular basis for the genetic code. *Proceedings of the National Academy of Sciences*, 55(4), 966–974.

Yanofsky, C. (2000). Transcription attenuation: once viewed as a novel regulatory strategy. *Journal of Bacteriology*, 182(1), 1–8.