



How To Quantify a Genetic Firewall? A Polarity-Based Metric for Genetic Code Engineering

Markus Schmidt*^[a] and Vladimir Kubyshkin*^[b]

Genetic code engineering aims to produce organisms that translate genetic information in a different way from that prescribed by the standard genetic code. This endeavor could eventually lead to genetic isolation, where an organism that operates under a different genetic code will not be able to transfer functional genes with other living species, thereby standing behind a genetic firewall. It is not clear however, how distinct the code should be, or how to measure the distance. We have developed a metric (Δ_{code}) where we assigned polarity indices ($\log D_i$) to amino acids to calculate the distances between pairs of genetic codes. We then calculated the

distance between a set of 204 genetic codes, including the 24 known distinct natural codes, 11 extreme-distance codes created computationally, nine theoretical special purpose codes from literature and 160 codes in which canonical amino acids were replaced by noncanonical chemical analogues. The metric can be used for building strategies towards creating semantically alienated organisms, and testing the strength of genetic firewalls. This metric provides the basis for a map of the genetic codes that could guide future efforts towards novel biochemical worlds, biosafety and deep barcoding applications.

1. Introduction

The wealth of natural biodiversity of an estimated 10 million species^[1] is surprisingly uniform and highly conserved on a deep biochemical and informational level. All species described so far, for example, store genetic information in just one very specific biopolymer (DNA) and the translation from RNA to 20 (22) amino acids is carried out predominantly by just one, "standard", genetic code.

The genetic code provides the rule for the correspondence between nucleic acid and protein sequences (Figure 1A). It is often considered a universal principle or language, a *lingua franca*, that unites species on the planet into a massive superorganism. Due to the genetic code universality, viruses can be transmitted between different species, as we learned from the outbreaks of viral diseases of animal origin.^[2] Another outcome of the code universality is the ability to spread fragments of genetic information between species in the course of the horizontal gene transfer. The latter is believed to be among the mechanisms behind the spread of antibiotic

resistance in pathogenic bacteria, which causes severe health threats for humanity.^[3] The scope of gene transfer, however, extends far beyond disease-causing organisms. In fact, exchange of genes-between bacteria but also between bacteria and eukaryotes-has shaped the web of life and is one of the most important factors in evolution.^[4] Uncontrolled horizontal gene transfer between released genetically modified organisms and wild-type organisms is also a biosafety concern.^[5]


Not surprisingly, researchers have been tempted by the idea of breaking the code via genetic code engineering. Eventually, a complete organism-wide (genome and proteome) genetic code engineering should yield genetically recoded organisms. This is one of the key targets of the xenobiological endeavor, which aims to create and study artificial biodiversity.^[6] One can expect to diverge species from extant natural versions by modifying the base,^[7] and backbone^[8] of the nucleic acids. Similarly, attempts to modify and enlarge the proteinogenic amino acid portfolio are well underway.^[9]


As one of the main and striking features of the biochemistry, the genetic code universality represents one of the most intriguing targets for manipulation. Nonetheless, one should note that minor variations in the genetic code do occur in nature. Currently 25 natural genetic codes have been identified,^[10] many in cell organelles. They reveal an extreme uniformity, with a mean modification of just 2.43 (min:1, max: 5) out of 64 codons;^[11] the natural code #25 was only published in 2019. Future research might discover more natural genetic codes.^[12] Nonetheless, it is already clear that the natural codes represent an infinitesimal fraction out of the astronomical sized combinatorial space. For example, 4.18×10^{84} possible genetic codes can be generated from 64 codons, 20 amino acids and at least one stop codon.^[13]


Even when acknowledging that the vast majority of potential codes in the full combinatorial space is useless,^[14] the set of "viable" codes is still much larger than the set of known

[a] Dr. M. Schmidt
Biofaction KG
Kundmannngasse 39/12, 1030 Vienna (Austria)
E-mail: schmidt@biofaction.com

[b] Dr. V. Kubyshkin
Department of Chemistry, University of Manitoba
Dysart Road 144, Winnipeg, R3T 2N2 (Canada)
E-mail: vladimir.kubyshkin@umanitoba.ca

 Supporting information for this article is available on the WWW under <https://doi.org/10.1002/cbic.202000758>

 This article is part of a Special Collection on Xenobiology. To view the complete collection, visit our homepage

 © 2020 The Authors. ChemBioChem published by Wiley-VCH GmbH. This is an open access article under the terms of the Creative Commons Attribution Non-Commercial NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

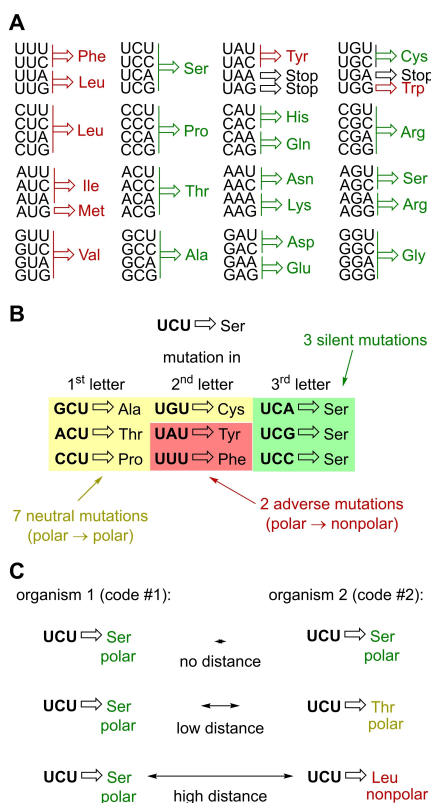


Figure 1. A) The standard genetic code. The code prescribes correspondences between the 64 coding elements in the mRNA sequence (codons) and 20 amino acids in the protein sequence. Nonpolar amino acids are highlighted in red, polar in green. B) In the natural genetic code, the codon UCU codes for serine, which is a polar amino acid. Mutations in this codon have different effects. Those that do not change the resulting amino acid are called *silent*. Those that lead to substitution with other amino acids of similar polarity are called *neutral*, here the polarity difference is low. Adverse mutations appear when the resulting amino acids have significantly different polarity. Robustness is a parameter that shows the mean square value of the polarity differences of all possible single point mutations.^[16] C) Comparison between two different genetic codes (all 64 codons) yields a distance parameter. If two codons stand for amino acids with the same polarity, the contribution to the distance is zero; the higher the difference between the polarities, the higher the contribution to overall genetic code distance.^[13]

natural codes. It has been shown that even in a massively reduced set of potential codes that maintain, for example, the same codon block distribution as in the standard code (when 64 codons of the genetic code are divided into 21 synonymous codon sets, 20 for sense codons and 1 for stop codons) will result in 2.43×10^{18} possible codes of which only one in a million (ca. 10^{12} codes) have the same or a better error robustness.^[15] By robustness we mean the error tolerance of the genes in case of single nucleotide replacements, and the effect of the resulting amino acid change. A robust genetic code is characterized by the high ability to generate similar, neutral or silent substitutions from single mutations (Figure 1B).

The study of the key features of the genetic code is of great interest for understanding the origin of biochemical mechanisms of life. At the same time, the experimental engineering of the genetic code holds a tremendous potential for creating novel biopolymers and biochemical reactions. Enhanced chem-

ical versatility of the organisms operating under novel genetic codes might be useful in medicine, industrial biotechnology or bioremediation. It might also contribute to the future biotechnological age endowed with superior biosafety features. In particular, generation of biological systems with distinct genetic codes would allow for better control of the exchange of genetic information with natural organisms through horizontal gene transfer.^[4c,5,17]

In 2009, Marlière wrote “The farther, the safer: a manifesto for securely navigating synthetic species away from the old living world”,^[17a] describing how estrangement can lead to less interaction and more safety. The isolation (or orthogonalization) strategy between natural and xeno-organism will eventually lead to a genetic firewall.^[18]

Conceptually, the metaphor of “distance” or “firewall” seems easy to parse, but the question of course is: what is the strength of the genetic firewall, and how can it be made stronger? This question points to the lack of a standardized way to measure the distance or strength of the genetic firewall. To address this weakness, a metric space for genetic code engineering was recently proposed. The idea of the metric space is to quantify the *distance*, or *firewall strength* between any two codes. The calculation was done comparing codon assignments of two codes by calculating the mean difference value of the different amino acid polarity values^[13] (Figure 1C). The underlying principle of this metric is inspired by the calculation of the mutational robustness.^[19] The metric space approach suggests repurposing of these calculations towards a metric or a scale that would allow to enumerate a distance between any two genetic codes. Thus, the metric space approach deals with the disturbance that should occur in the proteome, due to the differences in the genetic codes.

Numerical¹ values can be generated by taking known amino acid values such as *polar requirement*^[20] and *hydropathy*.^[21] The use of these values has become a common standard in the research field that calculates mutational robustness.^[19,22] Nonetheless, these values are only available for the 20 canonical amino acids, and for the wealth of the other potential amino acid substrates they are lacking. As the result, a metric based on these values would neglect a whole research branch dedicated to the integration of the artificial chemical components into the protein repertoire. It would be far more beneficial, to base a genetic code metric on a value that would readily integrate various canonical and noncanonical substrates, that are existent or even hypothetical.

Herein we propose the use of an empiric partitioning based scale (clog D_7) for parametrization of the codes. We show that this scale works as good as the former ones in predicting code similarity, yet it can readily integrate virtually any noncanonical amino acid. By using this approach, we generate a map of viable and distant genetically recoded organisms that may allow for the construction of a network of genetic firewalls. A

¹The calculation for mutational robustness compares 288 codon pairs, representing all possible changes that could be triggered by single base mutations, while the genetic code distance compares amino acids value pairs of 64 codons in two different codes. For more details, see Ref. [13].

genetic firewall is meant to massively reduce or even completely block any horizontal gene transfer between engineered organisms, and a genetic firewall metric is a way to quantify the strength of the genetic firewall. We believe that our approach will help to rationalize existing and future efforts towards deep artificial biodiversity. Our method enables us to propose a network of artificial biodiversity and map efforts towards a novel form of biocontainment. While biocontainment is typically understood as the physical isolation of organisms,^[23] in our case it is a semantic isolation, hence a form of semantic biocontainment.

2. Results and Discussion

2.1. Design of the study

We designed the study with the aim to measure the distance between any two given codes following a single principle. In the core of our approach, we assume that a gene may be transferred from an organism #1 operating under genetic code 1 to an organism #2 operating under genetic code 2. In the organism #2, the translations of this gene is likely to produce a functional protein if the amino acid assignments prescribed by the codes are similar. Conversely, translation is likely to produce a nonfunctional (nonsense) protein if the amino acid assignments are different (Figure 1C). For this reason, we needed a numerical value that would reflect the (dis)similarity between the amino acids. The simplest and first parameter to characterize the amino acids is their polarity. It is well known that the folding of globular proteins relies on the “nonpolar in – polar out” principle, whereas membrane proteins follow an opposite “polar in – nonpolar out” principle. Thus, any exchange of an amino acid residue with another one having a distinct polarity is likely to yield detrimental effects onto the protein structure. The differences in the polarities between the codon assignments in the codes 1 and 2, is what will be used to characterize the distance between the codes. Our approach does not take into account potential reading of specific functional residues such as catalytic residues in the enzymes, or backbone folding, for example, secondary structure propensities.

After we choose the physicochemical principle behind our approach, we then have to select a reliable numerical scale that reflects the polarity of amino acid side chains. We aim to operate with a numerical value that can be found for both canonical and noncanonical amino acids. After having polarity values assigned to the amino acid residues, we can calculate the distance between any two given codes numerically. We then set out to compare the standard genetic code with other possible codes, that can be generated by either 1) reshuffling of the amino acid assignments; or by 2) replacing canonical with noncanonical residues. With additional assumptions, the method will integrate genetic codes made by 3) changing the stop codon assignments; or 4) introducing empty codons that lack a corresponding tRNA match.

The method we choose to calculate the distance, or dissimilarity, between two codes is expressed according to Equation 1:

$$A_{\text{code}} = \frac{\sum_{k=1}^n |x_k - y_k|}{n} \quad (1)$$

where x_k is the (polarity) value for the k th amino acid in code x and y_k the (polarity) value for the k th amino acid in code y , while n is the total number of compared codons, so 64.

There is also another method to compare different genetic codes, which is based on the formula used to calculate mutational error robustness and which uses mean square values.^[13,19c] This formula however doesn't allow the establishment of a universal metric space.²

2.2. Choice of the polarity scale

Polar requirement and hydropathy have been previously used in the calculations of the genetic code parameters such as their robustness. As long as one uses only the 20 canonical amino acids in the genetic repertoire these values work just as well to calculate the genetic code distance between any two codes. The limitations of these values appear once we attempt to integrate amino acids beyond the set of 20, so-called non-canonical amino acids. As a matter of fact, corresponding experimental values for noncanonical amino acids are lacking. For example, the polar requirement values for 20 canonical amino acids were generated in mid 1960s using a rather primitive chromatographic approach, which is no longer in use in modern labs.^[20] Thus, in order to generate a universal genetic code metric that can integrate canonical and noncanonical amino acids, polarity scales need to be analyzed.

At first, we surveyed the common bioinformatics resource used for mapping properties of the protein sequences.^[24] We found over 60 common scales that allow to evaluate a sequence composed by the set of 20 canonical amino acid substrates. They reflect different parameters of the residues, such as bulkiness, hydrophobicity, accessibility in a protein structure, propensities to adopt secondary structures and more. Among these, we decided to analyze those scales that reflect polarity of the amino acids in one way or another. In fact, there are 31 scales that can be related to polarity, with 27 scales that can be considered non-redundant. Closer look shows that these scales can be grouped in a three large groups dependent on the physical phenomenon they are based upon:

² In this other formula the mean square (MS) value of all 64 codon differences between any two codes are calculated: $\Delta^2 = \frac{\sum_{k=1}^n (x_k - y_k)^2}{n}$ where x_k is the (polarity) value for the k th amino acid in code x and y_k the (polarity) value for the k th amino acid in code y , while n is the total number of compared codons, so 64. While the mean square values work fine for one-dimensional scales ref. [13], some genetic code triples (three Δ_{code} values between three codes) might not comply with the fourth requirement of the metric space, namely the triangle inequality. Thus, we did not use the squared but the absolute difference as a basis for our calculations.

Type 1: statistical use of a certain type of residues in known proteins. This would usually be based either on the known protein crystal structures or immunology data (antibody recognition fragments). The residues enriched in the protein cores and transmembrane domains are commonly considered more hydrophobic than those exposed to water. In this way these scales reflect the polarities of the residues.

Type 2: chromatography. Here, the oldest values are based on paper chromatography retention factors, more recently they were generated by detecting reversed-phase HPLC retention times. Chromatography is fundamentally based on multiple extraction, thereby it reflects the polarity of the substrates: amino acids, and peptides or proteins built from them.

Type 3: partitioning. These scales are based on calculating the energy of transfer of a substrate from a polar medium, which is usually water, to a nonpolar medium. The later can be water-air interface, protein interior, or organic solvent (ethanol or octanol).

The hydrophathy scale stands out from this classification. It is a so-called amalgamated scale, that was generated by averaging a number of other scales available by 1982, and assigning arbitrary values to a few residues, such as glycine and arginine.^[21] It is therefore in principle not possible to extend this scale to noncanonical substrates. The same is true for the scales that are based on structure statistics (type 1), because these statistics are simply not available for most noncanonical residues. The chromatography and partitioning based values are more convenient. These are single-phenomena based values, and can be found experimentally. We focused on the octanol/water partitioning based scale, since they can be evaluated both experimentally^[25] and can be calculated^[26] using available empirical calculations. The results of octanol/water partitioning are commonly designated as $\log P$, which is the partitioning value expressed on a logarithmic scale. The P value is a ratio between the substrate concentration in octanol (nonpolar phase) and its concentration in water (polar phase) at an equilibrium (Figure 2A). There are large databases of the partitioning data both commercially and publicly available, that allow calculation of the $\log P$ values for virtually any organic molecule. Resulting values are designated as $\text{clog} P$ values, where "c" indicates that the value is not experimental but computed. The $\text{clog} P$ calculations are simple and can be performed using a desktop or web-client on a click. The problem with the $\log P$ scale is that it does not take into account ionization of a substrate when this is transferred to water. To correct for the ionization that occurs at certain pH values, another value is used, $\log D_{\text{pH}}$, where D is distribution coefficient. For non-ionizable molecules $\log P = \log D$.

For our calculations, we choose the $\text{clog} D_7$ scale from the ChemAxon website,^[27] which is one of the most widely used public databases for calculating the partitioning values. $\text{clog} D_7$ indicates the partitioning of a substrate between octanol and water buffered at pH 7, and the value is calculated from an empiric dataset. We thus can assign any amino acid structure with a $\text{clog} D_7$ value, and use this for our scale. A few problems remained though. One of them is the impact of stereochemistry on polarity, which is not taken into account by state-of-the-art

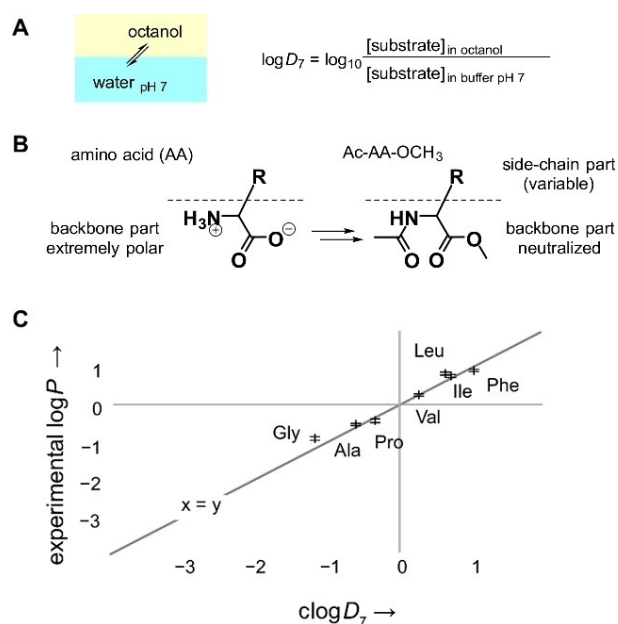


Figure 2. A) $\log D_7$ is a value that characterizes the distribution of a substrate between octanol and aqueous buffer with pH 7. Hydrophobic substrates exhibit positive values. B) Methyl esters of *N*-acetyl amino acids are the derivatives of amino acids that are used for $\text{clog} D_7$ calculations in this study. For most substrates (except glycine and proline analogues) the backbone part remains constant, with only the side chain varying. C) Correlation between the calculated $\text{clog} D_7$ values and the available experimental $\log P$ shows a good agreement between the experiment and prediction (correlation coefficient $R = 0.9824$).

$\log P/\log D$ calculation algorithms. Another problem is a large polar contribution of the backbone groups. Most of the canonical amino acids share same backbone features with only the side chains varying. The backbone is very polar, and its contribution cannot be neglected when calculating features of amino acids in peptides and proteins.^[28] Secondary structure transitions lead to differences in the backbone exposure to the medium, thereby the polarity of a particular residue can differ depending on the context of the secondary structure. For our calculations, we neglected the backbone contribution to polarity. For this reason, we choose a simple and conservative amino acid derivative: methyl esters of *N*-acetyl amino acids (Figure 2B). Calculations of the $\text{clog} D_7$ values for these molecules yield positive values (more substrate in octanol) in case of hydrophobic side chains, and negative values (more substrate in aqueous buffer) for hydrophilic side chains, when the backbone is based on an alanine structure. One should be careful though, comparing amino acids with distinct backbone features, especially proline and glycine. For example, proline has a significantly distinct backbone (secondary amino group), as well as the secondary structure propensities. Therefore, comparison of the common alanine-based amino acids and proline and its analogues cannot be accurate, and should be avoided. This is not only true for the $\text{clog} D_7$ scale, but also for other scales.

With this in mind, we assigned 20 canonical amino acids with the $\text{clog} D_7$ values that reflect the polarity of their side

chains. The found values correlate with the experimental $\log P$ fairly well (Figure 2C).^[29] We also correlated the existing hydrophobicity scales to the $\text{clog}D_7$ scale including the hydrophathy and polar requirement scales (Table 1). The largest discrepancies were found for cysteine and tyrosine residues, that can be listed as hydrophobic/buried in some scales, and hydrophilic/exposed in others. Overall, the new $\text{clog}D_7$ scale correlates to the former hydrophobicity scales in a similar way as the former hydrophobicity scales correlate between each other, and enables to distinguish the amino acid side-chains by their polarity features. Note that $\text{clog}D_7$ provides a formal accuracy of 10^{-2} with three significant digits, while the hydrophathy and polar requirement values only produce two significant digits. Nonetheless, since the values were generated by entirely different methods, we can compare values between scales but not their actual accuracy or error.

Table 1. Comparison of some traditionally used amino acid values for polar requirement and hydrophathy, and the $\text{clog}D_7$ values.^[a] Ordered by $\text{clog}D_7$ value.

Amino acid 1 letter	Amino acid 3 letters	$\text{clog}D_7$	Hydrophathy	Polar requirement
W	Trp	+1.14	-0.9	5.2
F	Phe	+1.04	2.8	5
Y	Tyr	+0.74	-1.3	5.4
I	Ile	+0.72	4.5	4.9
L	Leu	+0.64	3.8	4.9
V	Val	+0.27	4.2	5.6
M	Met	+0.04	1.9	5.3
P	Pro	-0.34	-1.6	6.6
C	Cys	-0.57	2.5	4.8
A	Ala	-0.61	1.8	7
H	His	-1.16	-3.2	8.4
G	Gly	-1.18	-0.4	7.9
T	Thr	-1.24	-0.7	6.6
S	Ser	-1.66	-0.8	7.5
Q	Gln	-1.77	-3.5	12.5
N	Asn	-2.06	-3.5	10
K	Lys	-3.56	-3.9	10.1
E	Glu	-3.74	-3.5	8.6
R	Arg	-3.94	-4.5	9.1
D	Asp	-4.14	-3.5	13

[a] Correlations: 1) hydrophathy vs. polar requirement $R = -0.786$, 2) $\text{clog}D_7$ vs. hydrophathy $R = 0.765$, 3) $\text{clog}D_7$ vs. polar requirement $R = -0.824$.

2.3. Distant genetic codes generated by reshuffling of codon assignments

After having chosen the type of the values, we carried out the distance calculations. We first set up to examine possible distant genetic codes generated by the reshuffling of the same portfolio of 20 canonical amino acids. First, an algorithm was carried out in the reduced genetic code search space that have the 20 canonical amino acids distributed in standard canonical codon blocks, leaving the three stop codons unchanged (combinatorial space = 2.43×10^{18} codes^[19c]). The genetic algorithm is identical to that described in ref. [13], which applies codon block swap operations^[30] selecting the most distant genetic codes for 50 consecutive generations, while controlling for mutational error tolerance relative to the standard code. The mutational error tolerance was calculated as the average mean square difference of all 288 possible single nucleotide mutations, with $i=1$ as the normalized value for the standard code error tolerance, with for example, 1.1 meaning that the error tolerance value is allowed to be higher than 110% of the standard code error tolerance^[22a] (Figure S1). The codes generated in this way were called X01<?_>unlimited (i unlimited), X02<?_>1<?_>1 ($i=1.1$), X03<?_>1<?_>01 ($i=1.01$) etc. Second, the restrictions on the codon blocks were lifted, meaning that the 20 canonical amino acids are distributed over all 61 sense codons, resulting in the codes called X11<?_>unlimited, X12<?_>1<?_>5, X13<?_>1<?_>25, X14<?_>1<?_>1 and X15<?_>1. Third, for code X21<?_>unlimited, only one stop codon (UAA) was conserved while the other 63 codons were available for 20 amino acids or were left empty (unassigned). Finally, for X31<?_>unlimited, 20 amino acids and one stop were randomly assigned over the 64 codons, the remaining 43 codons populated with either canonical amino acids, stop or empty codons.

The difference from the method used previously^[13] is that instead of using the values for polar requirement and hydrophathy, the $\text{clog}D_7$ values were used to quantify amino acids, and that the method used to calculate Δ_{code} was not mean square but mean absolute difference.

Figure S2 shows the evolution of distant codes following the genetic algorithms described above. Table 2 summarizes key values of the different extreme distance genetic codes. The

Table 2. Key results of most distant codes in different search spaces with increasingly relaxed robustness ratios. In general, the fewer restrictions on robustness and search space, the faster and the higher Δ_{code} values are reached. All results are based on $\text{clog}D_7$ values.

Code name	robustness, i upper limit	$\Delta_{\text{code max}}$ to Standard Code	# of codes in combinatorial space	Restrictions for combinatorial space
X01_unlimited	unlimited	2.9446	2.43×10^{18}	stop codons unchanged.
X02_1_1	1.10	2.8764		20 codon blocks unchanged
X03_1_01	1.01	2.8268		
X04_1	1.00	2.8178		
X11_unlimited	unlimited	3.1475	9.42×10^{78}	3 stop codons unchanged.
X12_1_5	1.50	3.2375		20 AA assigned to 61 sense codons
X13_1_25	1.25	3.3441		
X14_1_1	1.10	3.1365		
X15_1	1.00	2.9578		
X21_unlimited	unlimited	4.4876	7.73×10^{82}	1 stop codon unchanged.
X31_unlimited	unlimited	4.8283	2.74×10^{85}	20 AA + empty codons assigned to 63 codons Stop, empty and 20 AA assigned to 64 codons

more relaxed the error tolerance (the higher the robustness ratio, i) the higher the final distance from the standard code and the shorter it takes to reach the final distance. The search space for $i=1$ is estimated to be in the order of 10^{12} compared to 10^{18} for $i=1000$, and still the most distant code for $i=1$ reaches 95.7% of the full distance $i=\text{unlimited}$ (Table 2).

While it cannot be known for sure that the resulting codes are indeed the most distant (local maximum vs. global maximum) we can see a clear inversion—an indicator of distance—of the $\text{clog}D_7$ values in the codon assignment of most distant codes compared to the standard code (Figure 3). These most distant codes represent the strongest possible genetic firewalls vis-à-vis the standard code under different conditions, such as whether the same 20 sense codon blocks are used or not, whether the three stop codons remain unchanged or not, or whether the occurrence of empty codons that do not code for an amino acid is permitted or not. These distant codes could serve as a genetic firewall reference point to other genetic codes presented here.

2.4. The distance between codes when using different polarity scales

Before leaving the traditional amino acid scales (polar requirement and hydrophathy) behind and taking up the new $\text{clog}D_7$ scale, we compared the two scales in order to find out how this transition affects the assessment of the two sets of extreme distant and natural codes. To do so we used the set of known natural codes,^[10] the set of extreme codes identified running the genetic algorithm with the underlying polar requirement and hydrophathy scales, taken from ref.[13], and the set of extreme distant codes generated with the underlying $\text{clog}D_7$ scale (Figure 4). While in the two sets of extreme codes only changes between sense codons were performed, the situation is different in many natural codes. There are some cases where it is not trivial to calculate the distance value, namely in the case of: 1) changes between sense and stop codons, 2) ambiguous codons; and 3) unassigned codons.^[13] In natural codes there are several cases of sense to stop codon changes and also a few ambiguous codons.

Of the 24 nonstandard natural codes, 20 involve a stop to sense codon reassignment, and three harbor ambiguous codons, for example, in the karyorelict nuclear code, UGA can code for either stop or Trp. For the sense to stop codon changes we follow ref.[11]. Thus, we assigned the largest possible difference between any two canonical amino acids, in the case of $\text{clog}D_7$ this is -4.14 (Asp) and $+1.14$ (Trp), which is 5.28 and the value assigned in case of a stop to sense or sense to unassigned (empty) codon change. In the case of ambiguous codes, two codes were generated. For example, karyorelict nuclear code N19_A_27A with UGA read as Trp and hence a stop-sense codon change, and karyorelict nuclear code N19B_27B with UGA left as the stop codon. Average values were not calculated but both subcodes were used in the calculations in order to provide a lower and upper bound. Unassigned

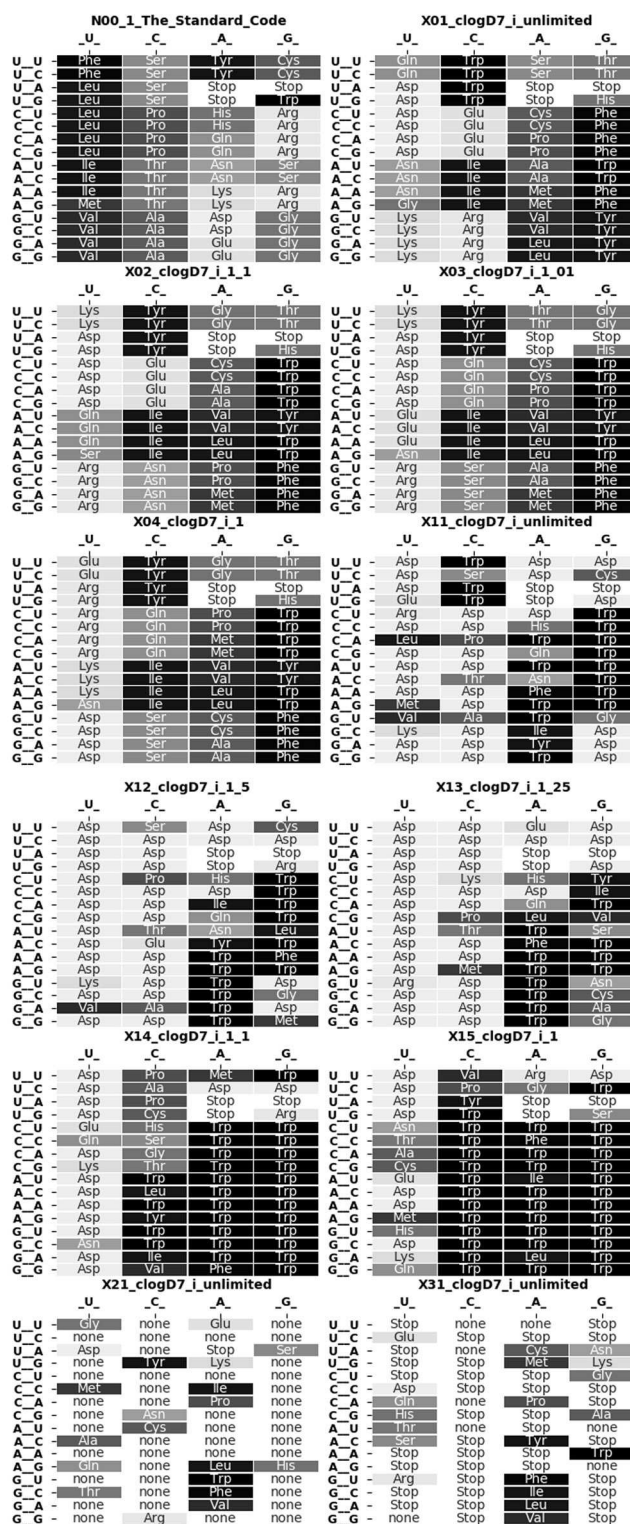


Figure 3. Tables of the standard and extreme distant codes, allowing for different error tolerance robustness ratios (see Figure S2 and Table 2 for details). The position of each codified amino acid is shaded on a gray scale representing its $\text{clog}D_7$ value (light gray: negative values, dark gray: positive values). Note the different gray shades in the four columns between the standard and extreme codes.

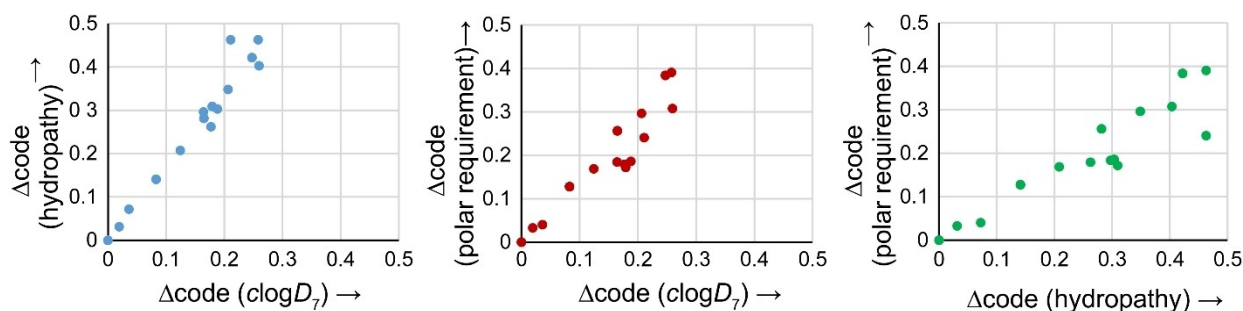


Figure 4. Comparison of the Δ_{code} values between the standard and all other natural codes using three different amino acid value sets. Correlation between $\text{clog}D_7$ and hydropathy is 0.97970, for $\text{clog}D_7$, the polar requirement is 0.93099, and polar requirement vs. hydropathy is 0.91284.

(emancipated) codons are not known in natural codes, only in engineered codes.^[31]

For all natural codes the Δ_{code} values to the standard code were calculated, using the traditional amino acid scales (polar requirement and hydropathy) and the new $\text{clog}D_7$ scale (Table 3). Results were plotted against each other and the correlation coefficients were calculated (Figure 4). The Δ_{code} values for natural codes are highly similar, with a correlation coefficient of at least 0.93099. This illustrates an excellent agreement between the analysis based on different polarity scales. Thus, the $\text{clog}D_7$ values can be utilized just as well as

others to show polarity distances, and parametrize the genetic firewalls.

2.5. Genetic codes generated by incorporation of noncanonical amino acids

Genetic code engineering provides various opportunities for integrations of novel amino acids in the translation set. These amino acids are called noncanonical, thus disobeying the genetic code “rule” (Greek *kanon*-rule, norm). In contrast, the 20 common amino acids are called canonical. The use of the

Table 3. Comparison of some traditionally used amino acid values for polar requirement and hydropathy, and the $\text{clog}D_7$ values^[a]. Ordered by $\text{clog}D_7$ values rounded to four significant digits).

Natural code name	Δ_{code} to Standard code based on		
	$\text{clog}D_7$	Hydro pathy	polar requirement
N00_1 The Standard Code ^[a]	0	0	0
N08_11 The Bacterial, Archaeal and Plant Plastid Code ^[a]	0	0	0
N20B_28B Condylostoma Nuclear Code (ambivalent) ^[a]	0	0	0
N18_26 Pachysolen tannophilus Nuclear Code	0.01953	0.03125	0.03281
N09_12 The Alternative Yeast Nuclear Code	0.03594	0.07187	0.04062
N03_4 The Mold, Protozoan, and Coelenterate Mitochondrial Code and the Mycoplasma - Spiroplasma Code ^[b]	0.0825	0.1406	0.1281
N23B_31B Blastocrithidia Nuclear Code (ambivalent) ^[b]	0.0825	0.1406	0.1281
N07_10 The Euplotid Nuclear Code	0.0825	0.1406	0.1281
N12_16 Chlorophycean Mitochondrial Code	0.0825	0.1406	0.1281
N15_23 Thraustochytrium Mitochondrial Code	0.0825	0.1406	0.1281
N17_25 Candidate Division SR1 and Gracilibacteria Code	0.0825	0.1406	0.1281
N16_24 Pterobranchia Mitochondrial Code	0.1241	0.2078	0.1687
N04_5 The Invertebrate Mitochondrial Code	0.1644	0.2969	0.1844
N05_6 The Ciliate, Dasycladacean and Hexamita Nuclear Code	0.1650	0.2812	0.2562
N14_22 Scenedesmus obliquus Mitochondrial Code	0.1650	0.2812	0.2562
N19B_27B Karyorelict Nuclear Code (ambivalent)	0.1650	0.2812	0.2562
N21_29 Mesodinium Nuclear Code	0.1650	0.2812	0.2562
N22_30 Peritrich Nuclear Code	0.1650	0.2812	0.2562
N06_9 The Echinoderm and Flatworm Mitochondrial Code	0.1772	0.2625	0.1797
N10_13 The Ascidian Mitochondrial Code	0.1794	0.3094	0.1719
N13_21 Trematode Mitochondrial Code	0.1878	0.3031	0.1859
N24_33 Cephalodiscidae Mitochondrial UAA-Tyr Code	0.2066	0.3484	0.2969
N02_3 The Yeast Mitochondrial Code	0.2106	0.4625	0.2406
N19 A_27 A Karyorelict Nuclear Code (ambivalent) ^[c]	0.2475	0.4219	0.3844
N20 A_28 A Condylostoma Nuclear Code (ambivalent) ^[c]	0.2475	0.4219	0.3844
N23 A_31 A Blastocrithidia Nuclear Code (ambivalent)	0.2475	0.4219	0.3844
N01_2 The Vertebrate Mitochondrial Code	0.2581	0.4625	0.3906
N11_14 The Alternative Flatworm Mitochondrial Code	0.2597	0.4031	0.3078

Codes marked with [a], [b], or [c] are identical.

Table 4. Typical values for the parameters used in the program: codes: name of codes; n : total number of program runs; x : number of offspring codes per generations; i : maximum allowed mutational robustness (normalized for the standard code); y : ratio of retained codes in each generation; g_{\max} : maximum number of generations. P : relative weight (function) that a certain code (based on its ranking according to Δ_{code}) is used in the next generation to generate offspring codes.

Code	n	x	i	y	g_{\max}	P
X01_unlimited			unlimited			
X02_1_1	480	400	1,1	0.1	50	
X03_1_01			1,01			
X04_1			1			
X11_unlimited			unlimited			
X12_1_5			1,5			$=0.4*((1-0.4)^{\alpha})$
X13_1_25	480	400	1,25	0.1	120	
X14_1_1			1,1			
X15_i_1			1			
X21	480	400	unlimited	0,1	120	
X31	480	400	unlimited	0,1	120	

^a with $\Delta_{\text{code_rank}}$ being the rank the new code has according to Δ_{code} to the standard code. In each generation the code with the highest Δ_{code} has rank 0, so the value for the first code to serve as parent to new codes is thus 0.4. For the second ranked code P is 0.24 for the third 0.144, fourth 0.0864, etc.

noncanonical amino acids provides an opportunity to replace any given amino acid in the code with a set of analogues with various polarity features. In our next step, we identified the noncanonical replacements to the set of canonical counterparts. The replacement structures were suggested using chemical analogy. For example, for arginine ($\text{clog}D_7$ -3.94), a potential replacement would include a number of arginine and ornithine-based structures such as citrulline, canavanine and more, that span the set of $\text{clog}D_7$ values from -4.54 to $+0.48$ (Figure 5). In a similar fashion we identified the replacements for other amino acids, except alanine and glycine. The latter two are generic structures, and these cannot be assigned any specific set of analogues. Amino acids with hydrophobic side chains, isoleucine, valine and leucine were not considered separately, but they were assigned with analogues having different number of carbon atoms in the aliphatic side chain. Proline analogues represent their own separate set of structures due to the different backbone of proline, however, in our approach they were treated in the same way as other amino acid analogue sets.

By suggesting a replacement, we aim to mimic a situation of a genetic code where a canonical amino acid is replaced with an analogue with either higher or lower polarity indices. This generates a code that is distant to the starting (standard) code by the value Δ_{code} . The higher is the value, the stronger is the corresponding genetic firewall between the codes. The structures of the amino acid analogues were selected by considering natural and artificial amino acids. We selected some candidates that we considered interesting from the experimental point of view (Table S1). However, the reader is encouraged to try out the Δ_{code} calculations for their own structure of interest by pasting the $\text{clog}D_7$ values from the web resource^[27] into Equation (1).

Figure 6 shows Δ_{code} values for some structures that are readily used in organisms as either toxic antimetabolites (e.g., canavanine) or post-translationally installed residues (e.g., phospho-serine).^[32–34] There is also a rich set of structures that have been incorporated into single protein structures using the genetic code expansion and selective pressure incorporation approaches.^[35] Especially rich is the set of analogues for aromatic amino acids. Large portfolio of aromatic amino acids originates from the fact that many experimental genetic code expansion systems are derived from the systems that incorporate tyrosine or phenylalanine in their natural environment. Finally, experiments reported proteome wide replacement of canonical amino acids. In these experiments, tryptophan was replaced with fluortryptophans or thienopyrrolylalanine.^[9a,36] The full set of the analyzed 160 amino acid replacements is listed in the supplementary information (Table S1).

By using the $\text{clog}D_7$ indices of the replacement amino acids, we calculated the corresponding Δ_{code} values. We considered that the estranged genetic codes should contain 19 amino acid and all stop assignments unchanged, but one amino acid (e.g., arginine) would be replaced with its analogues counterpart (e.g., citrulline) at all its codons. Results of these calculations are shown in Figure 6. The largest Δ_{code} to the standard genetic code was found for code R_Leu_10 with leucine replaced by miristyl-glycine ($\Delta_{\text{code}} = 0.4313$), while the shortest distance found, was code R_Met_5 with methionine replaced by homopropargylglycine ($\Delta_{\text{code}} = 0.0007813$). Between R_Leu_10 and R_Met_5 the Δ_{code} distance spans three orders of magnitude demonstrating the ability to fine tune Δ_{code} distance with noncanonical replacements.

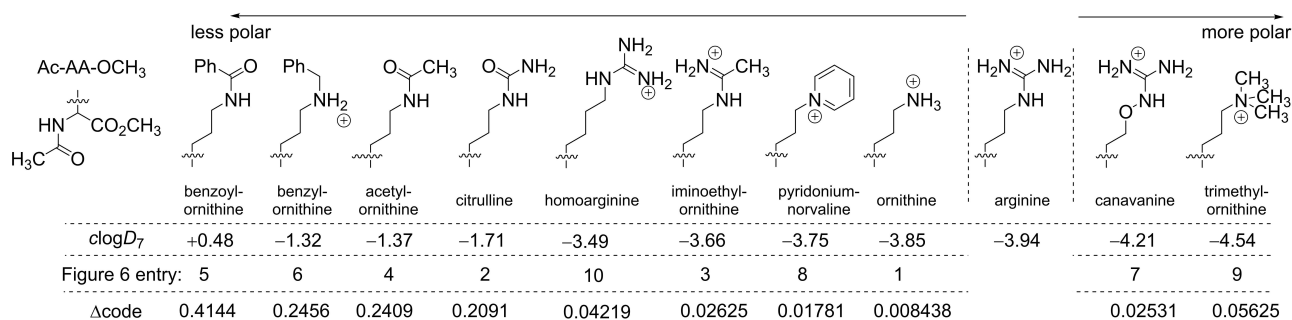


Figure 5. The set of arginine analogues sorted by their polarity values. This set of analogues was considered for replacement of arginine in the genetic code.

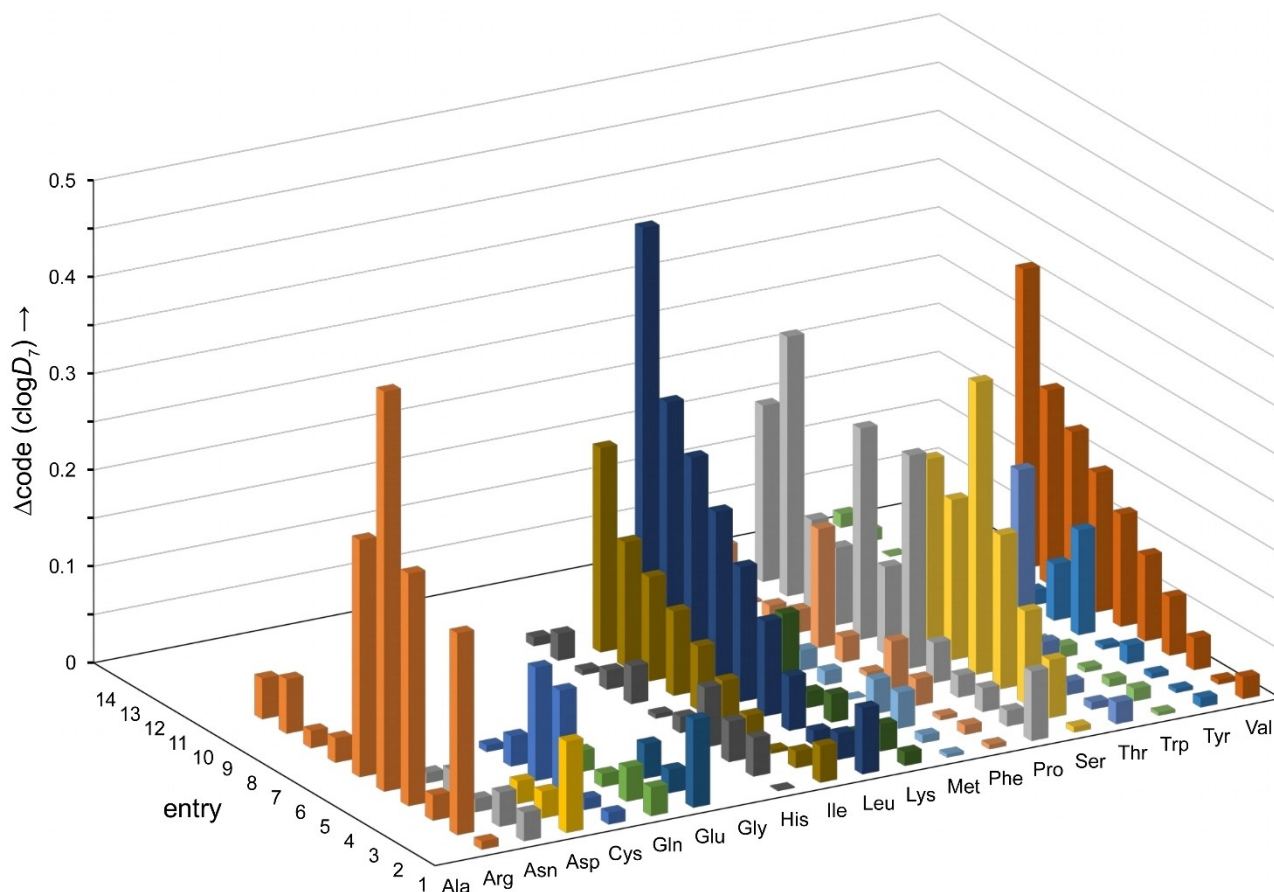


Figure 6. Clog D_7 -based Δ_{code} values between standard code and 155 replacement codes, arranged by amino acid and numbered replacement noncanonical amino acids (see Tables S1 and S2 for details). Replacement Δ_{code} distances span three orders of magnitude, from 7.8×10^{-4} (minimal value) to 4.3×10^{-1} (maximal value), whereas in natural codes it spans only one order of magnitude, from 2.0×10^{-2} to 2.6×10^{-1} , thus highlighting the opportunity to fine tune Δ_{code} values with noncanonical amino acid replacement. Note that no replacements were assigned for alanine and glycine. Only those codes that replaced amino acids throughout their respective codon blocks were included, five codes (R_AGA_1, R_Stop_1, R_Stop_2, R_M_1, R_M_2) are not included in the figure.

Analysis of the values for arginine, for example, shows that the largest Δ_{code} (0.4144) is for replacement of arginine with *N*-benzoyl ornithine, which lacks a charge, and contains a large aromatic group that effectively inverts the polarity of the residue. Very notably, large distance to the natural code can be obtained even with the maintenance of the positive charge in the side chain, as can be seen for *N*-benzyl-ornithine ($\Delta_{\text{code}} = 0.2456$). Replacement of arginine with natural metabolic amino acids such as *N*-acetyl-ornithine and citrulline generate codes with Δ_{code} 0.2409 and 0.2091 respectively. Canavanine represent an interesting example of a natural amino acid, which causes toxic effects due to its misincorporation into proteins^[37] as well as the impairment of the nitric oxide metabolism.^[38] The toxicity originates partially from promiscuity of the natural arginine-tRNA synthetase and the translation machinery, which leads to statistic incorporation of canavanine in places meant to be occupied by arginine. As this replacement is evidently toxic in nature, meaning that proteins with canavanine are dysfunctional, one can take it as an example of a natural firewall. Although, the polarity change in proteins might be not the exact mechanism of the toxicity, the change in the clog D_7 ,

reflects the chemical change upon the amino acid substitution. The Δ_{code} distance value for the arginine-to-canavanine replacement found in our study was as low as 0.02531. This result suggests that the firewall does not have to be extreme in value to readily generate dysfunctional proteomes. Even relatively low Δ_{codes} may suffice to estrange organisms operating with noncanonical amino acid substrates in their genetic repertoire.

To illustrate other replacement cases, we selected a few more noncanonical substitutes with their respective clog D_7 and Δ_{code} as presented in Figure 7. For example, methionine is often oxidized into a more polar methionine sulfoxide residue ($\Delta_{\text{code}} = 0.03719$) in natural proteins, which is a part of an oxidative damage that renders proteins dysfunctional.^[39] Homopropargylglycine (0.0007812), and azidohomoalanine (0.01453) are common methionine substitutes in proteins employed for click chemistry applications.^[40] Some other methionine analogues represent interesting targets for proteome wide replacement. For instance, norleucine (0.01187) or ethionine (0.0039) seem to produce moderate Δ_{code} values that might not be sufficient for a firewall. The phenylalanine analogues shown in Figure 7 are those that have been used in complete genome-

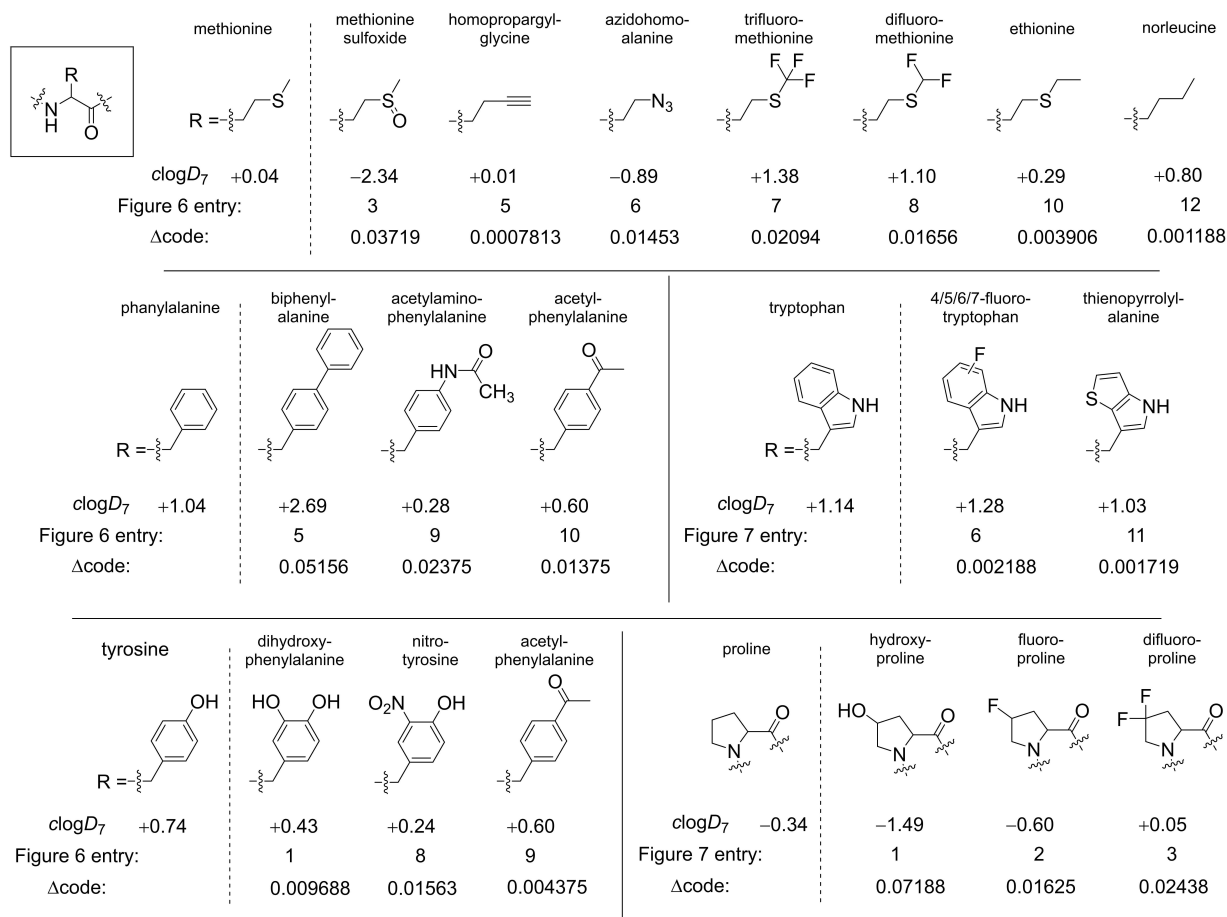


Figure 7. Selected noncanonical amino acid replacement structures from Figure 6.

wide recoded organisms.^[9b-d] Interestingly, further analysis of the Δ_{code} values shows that substitution of phenylalanine with its acetyl- (0.01375) and acetamino-containing analogues (0.02375) did not generate high Δ_{code} values, while substitution with biphenyl-alanine (0.05156) produced slightly higher value that may secure a firewall. In addition, the replacement of tyrosine with acetyl-phenylalanine (0.004375) produces a rather low Δ_{code} which is lower when the same amino acid is supposed to replace phenylalanine (0.009687). The amino acids fluorotryptophan and thienopyrrolyl-alanine have been incorporated proteome-wide in place of tryptophan in course of an adaptive laboratory evolution.^[9a,36a,c] Interestingly, the proteome wide tryptophan replacements exhibit relatively low Δ_{code} values, thus the genetic firewall strength is weak if present at all in these cases. Indeed, it has been observed that the fluorotryptophan adapted *E. coli* strains, can still grow on a tryptophan supporting media, thus a genetic firewall with such low values ($\Delta_{\text{code}} = 0.002188$) seems clearly inefficient in impeding gene functionality and horizontal gene transfer.^[9a,36a, 41] Dihydroxyphenylalanine is a common tyrosine post-translational derivative, and a constituent of the proteins in patients treated with levodopa,^[42] whereas nitro-tyrosine is a common oxidation stress marker generated from the tyrosine residues in the presence of nitric oxide.^[43] Hydroxyproline is a common

natural post-translationally^[44] generated proline substitute, which can also be incorporated translationally, so as fluoroproline.^[45] For all these cases the metric space approach allows to establish a numerical estimate for the estrangement of the genetic code containing the noncanonical substitutes. Considering evidently detrimental effects from canavanine, methionine-sulfoxide, and nitro-tyrosine, we conjecture that the lower bound for a robust genetic firewall is $\Delta_{\text{code}} 0.02$.

2.6. Mapping the complete graph of all codes

So far, we have only given one dimensional Δ_{code} distance values of natural codes, extreme codes and replacement codes relative to the standard genetic code. But of course, the Δ_{code} distance can be calculated between any two codes. With more than just one dimension it is important that the distance metric fulfills the criteria of a metric space,^{[3][13]} defined as:^[46]

³The triangle inequality, in particular, was the reason we discontinued to square the differences in hydrophobicity values, but used the absolute difference instead. Imagine that three points A, B and C lie on a straight line. When the distance from A to B is 5 and B to C is 10 the distance from A to C is 15. By squaring these values the distance A to B becomes 25, and B to C 100, which is less than the squared distance A to C, namely 225, which is a

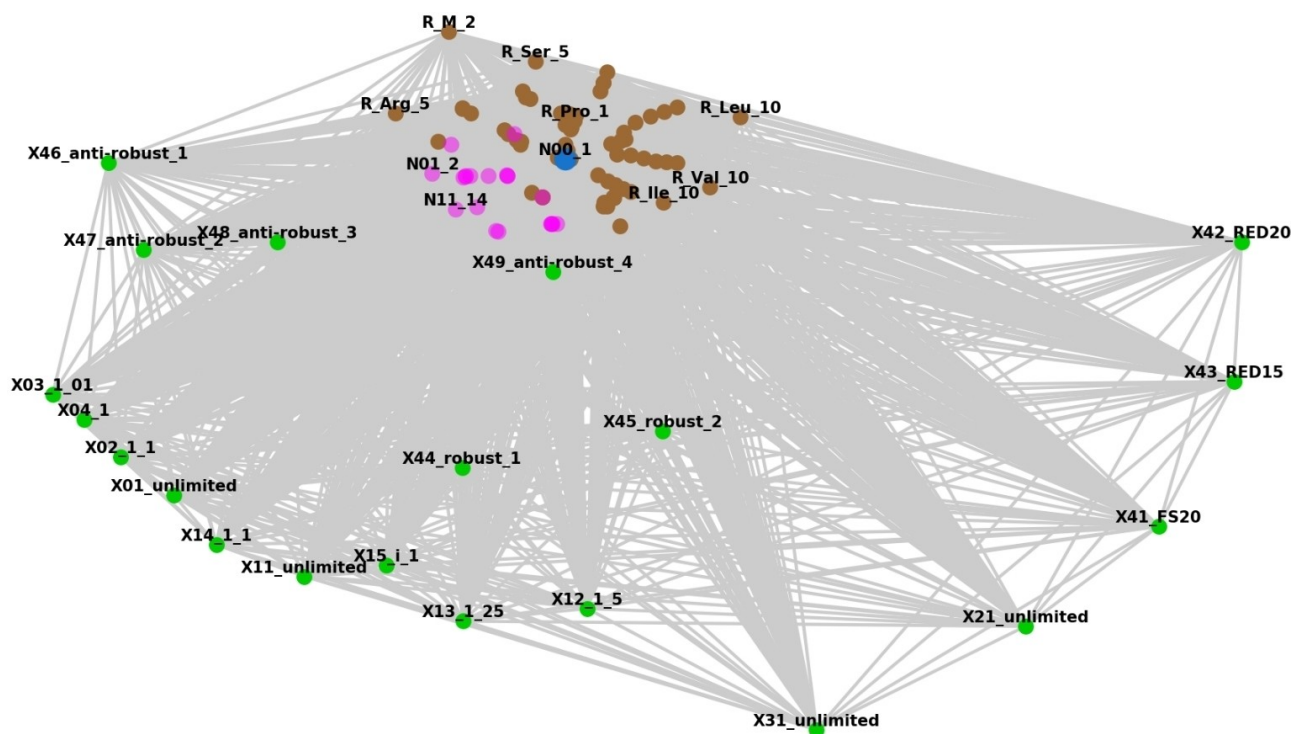


Figure 8. Two-dimensional re-representation of the complete graph of 204 distinct genetic codes based on graph drawing by force-directed placement. The farther away two codes are the more different they are. Large blue circle: standard code N00_1, pink: natural codes (N); brown: selected replacement codes (R); green: extreme distance codes (X). For better visibility, not all codes are labeled. Note that this is a 2D representation of a 203 dimensional space, which means the length of the 20706 edges (gray lines) is not exactly the same as the real values, but an approximation based on the error minimization Fruchterman and Reingold algorithm.^[59] For an interactive 3D representation of the graph, see ref. [47].

- 1) Identity: the distance from a point (code) to itself is zero
- 2) Non-negativity: the distance between two distinct points (codes) is positive (and hence never negative)
- 3) Symmetry: the distance from *A* to *B* is the same as the distance from *B* to *A*, and
- 4) Triangle inequality: the distance from *A* to *B* (directly) is less than or equal to the distance from *A* to *B* via any third point *C*.

This means that a metric space is an ordered pair (*M*, *d*) where *M* is a set and *d* is a metric, such that for any *x*, *y*, *z* ∈ *M*, the following holds:^[46]

- 1) $d(x, y) = 0 \Leftrightarrow x = y$ identity of indiscernibles
- 2) $d(x, y) = d(y, x)$ symmetry
- 3) $d(x, z) \leq d(x, y) + d(y, z)$ triangle inequality

From the above three axioms it follows that $d(x, y) \geq 0$ for any *x*, *y* ∈ *M*.

We used a set *M* of 204 discernible genetic codes (24 natural codes, 20 extreme codes and 160 replacement codes),

violation of the triangle inequality. Using a metric that is based on the absolute difference instead of the squared difference, fulfills all criteria of the metric space.

⁴ In mathematics this is known as a complete graph, which is a graph in which each pair of graph vertices (codes) is connected by an edge (Δ_{code} distance). A complete graph with *n* graph vertices has $(\text{Optn}2) = n * \frac{n-1}{2}$ edges.

summarized in Tables S1 and S2, and calculated all 20706 possible Δ_{code} distances between them⁴.

Figure 8 provides a two-dimensional graphical visualization of this 203 (*n*−1) dimensional space, giving an overview of distance relations between all codes.

Within the 20706 Δ_{code} values we found two that are in fact zero, that is between R_Phe_01 and R_Phe_06 (phenylalanine replaced by fluorophenylalanine and ethynyl-phenylalanine that have identical $\log D_7$ value of +1.19); and R_Trp_02 and R_Trp_07 (tryptophane replaced by 4-aminotryptophan and 4-azatryptophan, both with $\log D_7$ value +0.31). Different amino acids with identical $\log D_7$ values point to the limits of resolution to the Δ_{code} metric space. They also point out to the fact that parametrization of amino acids by just one parameter cannot reflect the changes in the underlying chemical structure in its entirety. In spite of this, the polarity index $\log D_7$ is very sensitive to the changes in the molecules, such as replacement of a carbon atom with a heteroatom, or changes in the constitution. For comparison, when considering the molecular weight, the chance of obtaining same values is much higher. For example, ornithine, fluoroproline, methoxinine, homothreonine all have the same molecular weight. At the same time, they have differences in polarities that allow to distinguish between the structures. See supporting Tables S3 and S4 for comparing molecular weight with $\log D_7$.

Since $\text{clog}D_7$ gives a precision of two digits behind the comma, the smallest theoretically possible Δ_{code} is as in Equation (2) and every Δ_{code} is a multiple of this.

$$\Delta_{\text{code min}} = \frac{0.01}{64} = 0.00015625 \quad (2)$$

In our set of codes, we indeed found one pair that has exactly the $\Delta_{\text{code min}}$ namely between R_Trp_05 and R_Trp_13, where tryptophane was replaced by 5-bromotryptophan and benzothiophenyl-alanine with $\text{clog}D_7$ values of +1.91 and +1.92, respectively.

The highest Δ_{code} value in our set of codes was 5.0439 between X31_unlimited and R_Leu_10 (leucine replaced by mirystyl-glycine), which is 32281 times $\Delta_{\text{code min}}$, so Δ_{code} values span a total of four orders or magnitude in our code set.

2.7. Using the Δ_{code} metrics

In mathematics a complete graph is a graph in which each pair of graph vertices (codes) is connected by an edge (Δ_{code} distance). A complete graph with n graph vertices has

$$\binom{n}{2} = n \cdot \frac{n-1}{2}$$

edges.

Right now, it is not known, how large Δ_{code} has to be in order to bring the horizontal gene transfer below a certain probability or acceptable level. We can assume, however, that

Δ_{code} values below 0.002 -as in the case of proteome wide tryptophan replacements with fluorotryptophan in adapted *E. coli* strains- should definitely be too low to guarantee a gene transfer impediment. Arginine-to-canavanine so as methionine-to-methionine sulfoxide replacement, that form dysfunctional proteins and are thus toxic to cells, yield Δ_{code} values of about 0.02-0.04 and might be a first hint of a lower bound for a genetic firewall.

With the Δ_{code} metric space, we provide a metric that is not restricted to canonical amino acids but open to noncanonical amino acids. Here, we suggest a metrological basis to design and carry out dedicated future experiments to measure horizontal gene transfer depending on Δ_{code} distance (Figure 8 in 2D and ref. [47] in 3D). We suggest that the difference between two genetic codes should reach a sufficient Δ_{code} value in order to reach a genetic isolation between corresponding organisms. We expect this to happen when one or several codons are recoded with amino acids that are distinctly different in its polarity to the original canonical one.

As an example, arginine and serine are coded by six codons each. A substitution of arginine ($\text{clog}D_7$ -3.94) with homoarginine (-3.49) has been previously accomplished at one codon.^[48] Here, the difference between the amino acid structures is only one methylene unit, and this is reflected in a rather small difference between the polarities of two amino acids. When taken at just one arginine codon ($\Delta_{\text{code}}=0.007031$), this replacement is not expected to reach the firewall threshold level. Conversely, when taken at all six codons ($\Delta_{\text{code}}=0.04218$), this replacement may produce a Δ_{code} value that may operate as a (weak) firewall (Figure 9A). In the case of serine, a replacement with very distinct amino acids, such as recoding serine ($\text{clog}D_7$ -1.66) with phospho-serine (-4.89) should yield a

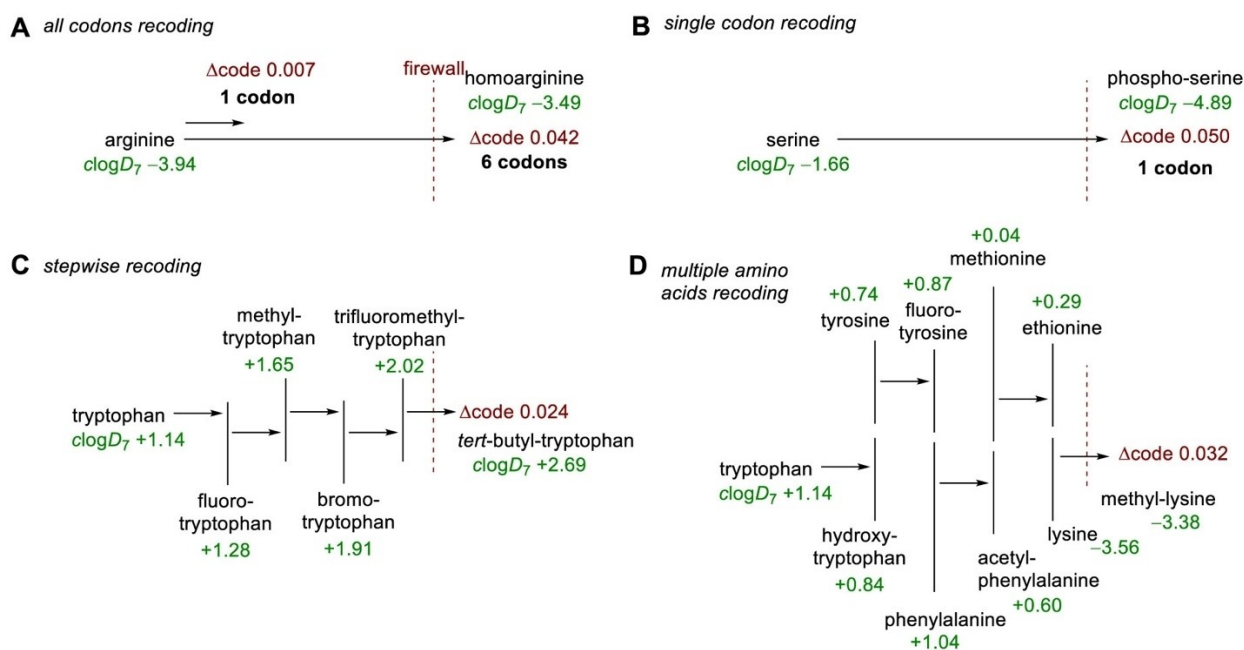


Figure 9. Examples of amino acid replacement strategies towards genetic isolation of organisms that can be experimentally scrutinized.

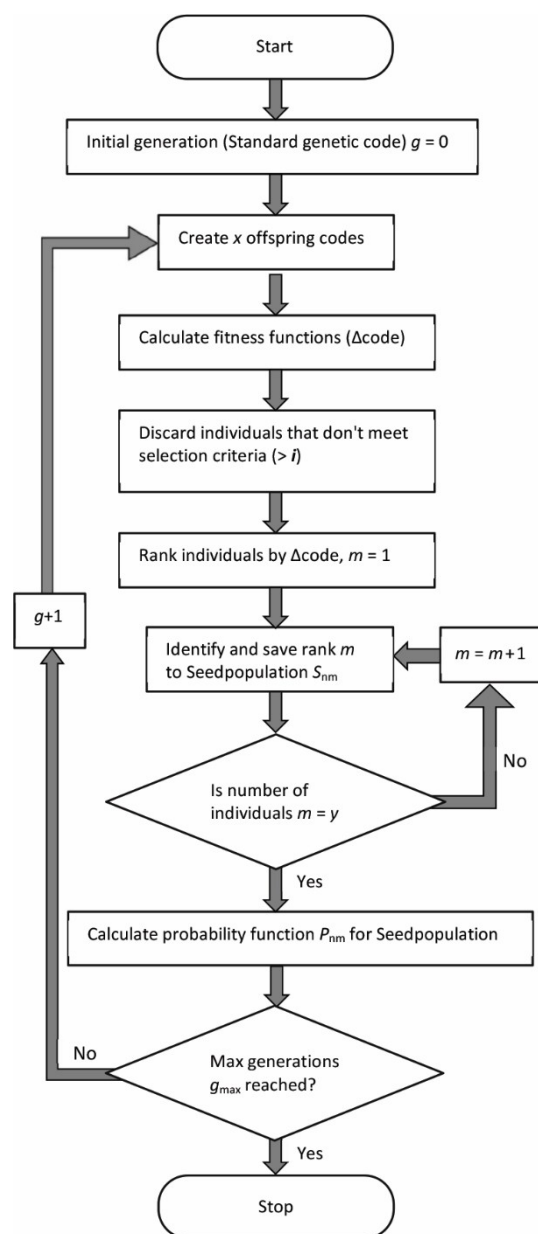


Figure 10. Overview of the program routine to identify extreme distance codes. For more details on the parameters, see Table 4. Each time the program reaches Stop, the total number of runs (n) is increased by 1.

higher Δ_{code} value even with one codon re-assignment ($\Delta_{\text{code}} = 0.05$; Figure 9B).

Another example we would like to discuss is the substitution of tryptophan with its analogues containing additional moieties. It has been mentioned that replacement of tryptophan with fluorotryptophan does not result in a large Δ_{code} due to relatively low polarity differences between the amino acid structures. However, we would like to propose to alienate the adapted strains further in a stepwise manner. Since fluorotryptophan ($\log D_7 + 1.28$) is more hydrophobic compared to tryptophan (+1.14), the next substitution should proceed towards even more hydrophobic analogues: methyltryptophan

(+1.65), bromo-tryptophan (+1.91), trifluoromethyltryptophan (+2.02), *tert*-butyl-tryptophan (+2.69), but not hydroxytryptophan, which is more hydrophilic compared to the parent structure (+0.84). While an immediate replacement of an amino acid with a distinctly different analogue might be too difficult to achieve, its stepwise replacement seems to be more practical (the same principle is applied for example in fish ladders). Eventually, *tert*-butyl-tryptophan might yield an isolation level ($\Delta_{\text{code}} 0.024$) with a significant reduction in horizontal gene transfer efficiency (Figure 9C).

Another potential option is the step wise alienation of several amino acids, one after another to generate higher and higher Δ_{code} values, and thus more alien genetic codes. As a mind experiment, we could suggest replacement of tryptophan with hydroxytryptophan, tyrosine with fluorotyrosine, phenylalanine with acetylphenylalanine, methionine with ethionine, lysine with *N*-methyl-lysine. Each of these substitutions alone generates a rather small difference in the genetic code that would probably be insufficient for genetic isolation. However, when taken together, these can actually feature the Δ_{code} value needed to reach the firewall (0.032; Figure 9D). In this way, consideration of the Δ_{code} value might help to build experimental strategies for genetic code engineering.

In the distant future, the Δ_{code} map could be very useful when it is necessary to locate existing and new codes with required minimal distances to the existing set of codes. For example, it could be required to keep a certain Δ_{code} distance to all the natural codes and engineered codes that have been or are foreseen to be released into the environment in order to impede horizontal gene transfer between them, maintaining existing genetic firewall strength of existing codes. These restricted areas around existing codes would make sure that no intermediate code “stepping stones” are established that would weaken existing genetic firewalls, like a fish ladder helps fish overcome a steep dam. For a related albeit inverse example in ecology, see “ring species”.^[49]

The map could also help to point out empty areas, genetic code voids, that could be colonized without risking the proximity of close neighbors and thus horizontal gene transfer.

Yet, another future application of the Δ_{code} metrics is making available a highly secure barcoding of synthetic biology agents in order to facilitate unique identification, tracing and if necessary, forensics. So far, the concept of genetic barcoding is based on incorporation of specific genetic “watermark” sequences into the genome that could later be sequenced to identify the strain. These barcoding sequences, however, might get lost due to mutation and selection pressure.^[50] For what we know, genetic code seems highly conserved and extremely stable over long evolutionary timescales, certainly more stable and long-lasting than a genetic sequence. Given the almost limitless amount of potential genetic codes, future advances in modifying the genetic code will open the door to use the genetic code itself as the barcode. The Δ_{code} based genetic code map could support the rational allocation of genetic codes to serve as a form of barcoding that is deeply embedded in the organisms and extremely difficult to shed off or lose, hence the term “deep barcoding”.^[51]

2.8. Generalization to other types of codes: expansion and limitations

The metric is applicable for all codes who share the same base alphabet (e.g., A, U, G, C) and codon length (e.g., 3) resulting in $4^3 = 64$ codons. Alternative approaches to genetic code engineering involve modification of: codon length, such as in quadruplet codons,^[52] the nucleic acid structure,^[8a] and different base pairs.^[53] All these examples employ a different chemical principle to build genetic isolation. The Δ_{code} metric only considers what happens on the level of the proteome, and cannot easily integrate phenomena that occur on the level of the nucleic acid and decoding principles. For example, the metric cannot be used to compare a AUGC triplet with a AUGC quadruplet code.^[52,54] However, the metric can indeed be adapted to compare all codes within the quadruplet combinatorial code space, by enlarging the number of codons from 64 to 256.

Although direct comparison of quadruplet to triplet codes is not possible, triplet to triplet codes with distinct base pairs is feasible. Take, for example, an extended genetic alphabet with three instead of two base pairs (the two canonical base pairs plus one noncanonical one) organized in triplets.^[53,55] The resulting $6^3 = 216$ codons contain the 64 canonical triplets plus 152 codons that contain one or both of the new bases. Within this larger codon set, the standard code can be notated as having the 64 canonical codons assigned as stop and sense codons, and the remaining 152 codons assigned as empty. All distance calculations performed in this manuscript can also be performed in the 6^3 codon space. All Δ_{code} values presented in this manuscript would have to be multiplied by 64/216 and all distance relations would remain the same.

Yet another situation is the modification of the nucleic acid backbone^[8a,c, 56] resulting in xeno nucleic acids (XNA), such as hexitol nucleic acid^[57] or cyclohexenyl nucleic acids.^[58] It is not possible to directly link chemical modifications in the protein structure with the structure of the XNA. Additional complications arise from the fact that the genetic information operates via two nucleic acids: a transcription messenger (DNA) and a translation messenger (RNA).

It remains to be seen how potential future organisms with XNAs will interact with DNA/RNA and other XNA based organisms. The design or directed evolution of (reverse) transcriptases that convert either between different XNAs, between XNA and DNA, or between XNA and RNA^[8a] has already started and could at one point overcome the orthogonality of the different XNA organism. In this case the different nucleic acid-based organisms could be made accessible to a unified Δ_{code} metric.

The fact that (reverse) transcriptases can possibly overcome the orthogonality embedded in nucleic acids, once again highlights the value of semantic containment, such as genetic code engineering.

3. Conclusions

Genetic code engineering provides an invaluable set of solutions that will advance both our understanding of biochemistry and biotechnology of the future. One of the most intriguing outcomes of it is that it suggests that the genetic code universality can be manipulated. Thus, we can envision a set of living species that will operate under different genetic codes: their genes are (for now) written in same nucleic acid alphabet, but the meaning behind these letters, the translation to protein sequences is different. In this work, we considered two major options for the genetic code engineering: reshuffling of the already existing amino acid assignments and introducing noncanonical amino acids (and empty codons) in the repertoire.

Here, we introduced and examined a versatile metric to quantify the distance between the different genetic codes numerically. The $\log D_7$ based Δ_{code} metric proposed here is capable of calculating the distance not only of genetic codes with the 20 canonical amino acids, but allow also for the incorporation of a vast number of noncanonical residues. A notable improvement in the calculation was the discontinuation of the mean square difference in favor of the mean absolute difference, which allows the establishment of a true metric space that can represent the correct distance relations between various natural and engineered genetic codes. The resulting value matrix may therefore become a tool to identify the position of any new code relative to other available codes.

In the core of our approach, we followed the polar/nonpolar dualistic scale that is a common classificatory for amino acids. The reason for this is that most of the amino acids structurally share the same type of backbone with only the side chains varying. This architecture principle was suggested to be called the *alanine world*.^[9f] In the frame of this world, the side chains are the functional elements that differ amino acids to one another. We used the $\log D_7$ as a value that can be easily obtained for any given structure using contemporary empirical prediction programs and databases. We then calculated the distance value Δ_{code} between the natural and engineered (existent and potentially interesting) genetic codes.

We found that the novel $\log D_7$ -based scale correlated fairly good with the previously existing scales, that are also polarity based. The natural codes displayed the Δ_{code} values in the range 0.00 to 0.26 from the standard code. For comparison, the most extreme codes that were generated by complete reshuffling of the codon assignments displayed values up to 4.828 from the standard code.

We then examined the codes that can be generated by replacement of the canonical amino acid with their chemical noncanonical analogues. The distances ranged from 0.00078 for methionine-to-homopropargyl-glycine substitution to 0.43 for leucine-to-mirystyl-glycine substitution, thus the range of values spans three orders of magnitude. The strongest determinants for the distance values of replacement codes are the polarity differences to the parent structures (in $\log D_7$ units) and the number of codons occupied by the canonical substrate. Interestingly, experiments have been published that showed a genome-wide replacement of tryptophan by analogues, how-

ever, the corresponding species can still accept tryptophan for the structures of their proteins even when completely adapted to a noncanonical replacement.^[9a,36] We found that in these cases the distance values to be very low: 0.0022 for fluoro-tryptophan and 0.0017 for thienopyrrolyl-alanine. From this, we suggest that the polarity differences introduced by these structures of the amino acids into proteome may not be sufficient to establish a genetic firewall. A lower bound for a genetic firewall is more likely to be in the range of the arginine-to-canavanine and methionine-to-methionine sulfoxide replacements, Δ_{code} of about 0.02, that proved to be toxic. For a final verdict, we propose to carry out dedicated experiments to evaluate gene transfer under different Δ_{code} distances. In this way the proposed metric space can support specifications for the design, build, test, and learn cycle^[60] for what is called semantic biocontainment.

Limitations of the study include the known problem of assigning a numerical value to sense to stop, or sense to empty codon changes, for which we provided a “patch” in substituting a value that is the largest $\log D_7$ difference between any two canonical amino acids. Future Δ_{code} dependent gene transfer and expression experiments may contribute to a more adequate substitution value. Also, the $\log D_7$ values represent only the polar nature of the sidechains, while the backbone features such as structure propensities or solvation are much harder to enumerate and take into account. The metric can always be applied to a set of genetic codes that have the same codon length, shared base pairs and the same transcription messenger chemistry (e.g., RNA). This fact points out the need for a numerical approach in parametrizing modifications of nucleic acids (XNA) and codon length (e.g., quadruplet code).

We claim that the more distant the genetic code is between two organisms, the lower are the chances that the genetic information could “leak” through horizontal gene transfer. This form of semantic isolation could be used as a genetic firewall for synthetic biology agents^[5,18a] to shield their genetic information from natural organisms and vice versa, but also between synthetic biology agents with different codes. Despite the claim that synthetic biology is the true application of engineering principles to biotechnology, right now there are surprisingly few metrics (and standards) available.^[61] The Δ_{code} metric provided here might fill the gap to allow for a rational design in genetic code engineering, to enable an intrinsic form of (deep) barcoding and to control horizontal gene flow via semantic biocontainment.

Experimental Section

Genetic algorithm for extreme distance code generation. : A genetic algorithm was programmed in Python 3.7. to evolve the standard genetic code into extreme distance codes.

The genetic algorithm aims to maximize the genetic distance between the new and the standard code while maintaining a defined upper threshold for error threshold, i . Calculations were run on a MacBookPro (OSX 10.14.6; 16 GB RAM, 2,3 GHz Intel Core i9, 8 core) with Python 3.7. run on Anaconda’s Navigator 1.19.12. Scientific Python Development Environment (SPYDER) 3.3.6.

In order to generate variations of the standard genetic code, a genetic algorithm with six key parameters was used (Figure 10 and Table 4), inspired by the Non-Dominated Sorting Genetic Algorithm.^[62] The genetic codes were explored in Python and new codes were generated starting from the standard code and using a swap operator (for codes X0+). The operator interchanged the contents of two codon sets (a codon set includes all triplets that code for the same amino acid), that is, once two codon set have been randomly selected, the amino acids codified by the two respective codon sets are swapped.^[30]

The X1+ new codes were generated starting from the standard code. Two codons were randomly chosen and the amino acid of the first codon was “copied” into the second codon, as long as this step didn’t eliminate an amino acid from the entire code. In codes X0+ and X1+ the stop codons were not changed.

For the X21 code only one codon (UAA) was reserved for the stop and kept unchanged. Codons UAG and UGA were made the 21st codon block, in addition to the 20 codon blocks from the standard code. Next the 20 amino acids and the designation empty codon was randomly assigned to the 21 codon blocks. This code was used as the starting point (and not the standard code as in X0+ and X1+) and then the same amino acid replacement strategy was applied as in codes X1+, with the exception that also empty codons could be copied to other codons.

For the X31 there was no reserved or fixed codon or codon block. 20 amino acids, a stop and a empty designation were randomly assigned to the 64 codons, the only restriction being that the codes must have all 20 amino acids and a stop codon present. Then the same acid replacement strategy was applied as in code X21.

Figure 8 was generated by using the full matrix of all 204 codes and their 20706 Δ_{code} values (edge; Table S1). All edge values were converted by using Equation (3):

$$\Delta_{\text{code spring}}_{\text{layout}} = \frac{1}{\Delta_{\text{code}}^2} \quad (3)$$

to meet the requirements of force-directed placement drawing. The fruchterman_reingold_layout from the networkx python library was used with iterations = 10^5 , threshold = 10^{-16} , dimensions = 2, scale = 5.0. The 3D graph shown at <http://markusschmidt.eu/3DG/example/GCE/index.html> is based on the 3D force graph software kit available under: <https://github.com/vasturiano/3d-force-graph>. In the 3D visualisation no conversion of edge values was needed.

Acknowledgements

M.S. received funding from the European Commission’s Horizon 2020 project MADONNA (766975) and BioRoboost (820699). V.K. acknowledges the Canadian government for funding the Canada research chair for chemical synthetic biology (grant no. 950-231971; lead by Dr. Nediljko Budisa).

Conflict of Interest

The authors declare no conflict of interest.

Keywords: amino acids · biosafety · genetic code engineering · genetic firewall · xenobiology

- [1] C. Mora, D. P. Tittensor, S. Adl, A. G. B. Simpson, B. Worm, *PLoS Biol.* **2011**, *9*, e1001127.
- [2] J. A. Fok, C. Mayer, *ChemBioChem* **2020**, *21*, 3291–3300.
- [3] a) K. L. Palmer, V. N. Kos, M. S. Gilmore, *Curr. Opin. Microbiol.* **2010**, *13*, 632–639; b) H. Heuer, H. Schmitt, K. Smalla, *Curr. Opin. Microbiol.* **2011**, *14*, 236–243.
- [4] a) S. M. Soucy, J. Huang, J. P. Gogarten, *Nat. Rev. Genet.* **2015**, *16*, 472–482; b) F. Husnik, J. P. McCutcheon, *Nat. Rev. Microbiol.* **2018**, *16*, 67–79; c) T. Y. Pang, M. J. Lercher, *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 187–192.
- [5] a) M. Schmidt, V. de Lorenzo, *FEBS Lett.* **2012**, *586*, 2199–2206; b) M. Schmidt, V. de Lorenzo, *Curr. Opin. Biotechnol.* **2016**, *38*, 90–96.
- [6] N. Budisa, V. Kubyskhin, M. Schmidt, *ChemBioChem* **2020**, *21*, 2228–2231.
- [7] P. Marlière, J. Patrouix, V. Döring, P. Herdewijn, S. Tricot, S. Cruveiller, M. Bouzon, R. Mutzel, *Angew. Chem. Int. Ed.* **2011**, *50*, 7109–7114; *Angew. Chem.* **2011**, *123*, 7247–7252.
- [8] a) V. B. Pinheiro, A. I. Taylor, C. Cozens, M. Abramov, M. Renders, S. Zhang, J. C. Chaput, J. Wengel, S.-Y. Peak-Chew, S. H. McLaughlin, P. Herdewijn, P. Holliger, *Science* **2012**, *336*, 341–344; b) S. Arangundy-Franklin, A. I. Taylor, B. T. Porebski, V. Genna, S.-Y. Peak-Chew, A. Vaisman, R. Woodgate, M. Orozco, P. Holliger, *Nat. Chem.* **2019**, *11*, 533–542; c) J. C. Chaput, P. Herdewijn, M. Hollenstein, *ChemBioChem* **2020**, *21*, 1408–1411.
- [9] a) M. G. Hoesl, S. Oehm, P. Durkin, E. Darmon, L. Peil, H. R. Aerni, J. Rappsilber, J. Rinehart, D. Leach, D. Söll, N. Budisa, *Angew. Chem. Int. Ed.* **2015**, *54*, 10030–10034; *Angew. Chem.* **2015**, *127*, 10168–10172; b) D. J. Mandell, M. J. Lajoie, M. T. Mee, R. Takeuchi, G. Kuznetsov, J. E. Norville, C. J. Gregg, B. L. Stoddard, G. M. Church, *Nature* **2015**, *518*, 55–60; c) A. J. Rovner, A. D. Haimovich, S. R. Katz, Z. Li, M. W. Grome, B. M. Gassaway, M. Amiram, J. R. Patel, R. R. Gallagher, J. Rinehart, F. J. Isaacs, *Nature* **2015**, *518*, 89–93; d) H. Xiao, F. Nasertorabi, S.-H. Choi, G. W. Han, S. A. Reed, R. C. Stevens, P. G. Schultz, *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 6961–6966; e) F. Agostini, J. S. Voller, B. Kokschi, C. G. Acevedo-Rocha, V. Kubyskhin, N. Budisa, *Angew. Chem. Int. Ed.* **2017**, *56*, 9680–9703; *Angew. Chem.* **2017**, *129*, 9810–9835; f) V. Kubyskhin, N. Budisa, *Curr. Opin. Biotechnol.* **2019**, *60*, 242–249.
- [10] A. Elzanowski, J. Ostell, *The Genetic Codes*, <https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi>, **2019**
- [11] B. P. M. Wnętrzak, P. Mackiewicz in *11th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2018)*, SCITEPRESS – Science and Technology Publications, Lda., **2018**, pp. 154–159.
- [12] Y. Li, K. M. Kocot, M. G. Tassia, J. T. Cannon, M. Bernt, K. M. Halanych, *Genome Biol. Evol.* **2019**, *11*, 29–40.
- [13] M. Schmidt, *BioSystems* **2019**, *185*, 104015.
- [14] A. A. Louis, *Stud. Hist. Philos. Biol. Biomed. Sci.* **2016**, *58*, 107–116.
- [15] S. J. Freeland, L. D. Hurst, *J. Mol. Evol.* **1998**, *47*, 238–248.
- [16] D. Haig, L. D. Hurst, *J. Mol. Evol.* **1991**, *33*, 412–417.
- [17] a) P. Marlière, *Syst. Synth. Biol.* **2009**, *3*, 77–84; b) P. Herdewijn, P. Marlière, *Chem. Biodiversity* **2009**, *6*, 791–808; c) N. Conde-Pueyo, B. Vidiella, J. Sardanyés, M. Berdugo, F. T. Maestre, V. De Lorenzo, R. Solé, *Life (Basel)* **2020**, *10*.
- [18] a) M. Schmidt, *BioEssays* **2010**, *32*, 322–331; b) M. Schmidt, in *21st Century Borders/Synthetic Biology: Focus on Responsibility and Governance*, Institute on Science for Global Policy, Tucson, **2013**, c) C. G. Acevedo-Rocha, N. Budisa, *Angew. Chem. Int. Ed.* **2011**, *50*, 6960–6962; *Angew. Chem.* **2011**, *123*, 7094–7096 d) N. Budisa, *Curr. Opin. Chem. Biol.* **2014**, *18*, 936–943; e) C. G. Acevedo-Rocha, N. Budisa, *Microb. Biotechnol.* **2016**, *9*, 666–676.
- [19] a) M. Wnętrzak, P. Błażej, D. Mackiewicz, P. Mackiewicz, *BMC Evol. Biol.* **2018**, *18*, 192; b) J. Santos, Á. Monteagudo, *BMC Bioinf.* **2017**, *18*, 195; c) D. Haig, L. D. Hurst, *J. Mol. Evol.* **1991**, *33*, 412–417.
- [20] C. R. Woese, D. H. Dugre, S. A. Dugre, M. Kondo, W. C. Saxinger, *Cold. Spring. Harb. Symp. Quant. Biol.* **1966**, *31*, 723–736.
- [21] J. Kyte, R. F. Doolittle, *J. Mol. Biol.* **1982**, *157*, 105–132.
- [22] a) G. Pines, J. D. Winkler, A. Pines, R. T. Gill, *mBio* **2017**, *8*, e01654-01617; b) L. L. de Oliveira, P. S. L. de Oliveira, R. Tinós, *BMC Bioinf.* **2015**, *16*, 52.
- [23] J. W. Lee, C. T. Y. Chan, S. Slomovic, J. J. Collins, *Nat. Chem. Biol.* **2018**, *14*, 530–537.
- [24] *Swiss Institute of Bioinformatics, "ProtScale"*, <https://web.expasy.org/protscale>, **2020**.
- [25] a) J. L. Fauchère, V. Pliska, *Eur. J. Med. Chem.* **1983**, *18*, 369–375; b) M. A. Roseman, *J. Mol. Biol.* **1988**, *200*, 513–522.
- [26] a) D. J. Abraham, A. J. Leo, *Proteins* **1987**, *2*, 130–152; b) S. D. Black, D. R. Mould, *Anal. Biochem.* **1991**, *193*, 72–82.
- [27] Chemaxon, *LogD Predictor*, <https://disco.chemaxon.com/apps/demos/logd>, **2020**.
- [28] J. R. Robalo, S. Huhmann, B. Kokschi, A. Vila Verde, *Chem* **2017**, *3*, 881–897.
- [29] V. Kubyskhin, N. Budisa, *J. Pept. Sci.* **2018**, *24*, e3076.
- [30] J. Santos, Á. Monteagudo, *BMC Bioinf.* **2011**, *12*, 56.
- [31] a) N. Ostrov, M. Landon, M. Guell, G. Kuznetsov, J. Teramoto, N. Cervantes, M. Zhou, K. Singh, M. Napolitano, M. Moosburner, E. Shrock, B. Pruitt, N. Conway, D. Goodman, C. Gardner, G. Tyree, A. Gonzales, B. Wanner, J. Norville, M. Lajoie, G. Church, *Science* **2016**, *353*, 819–822; b) J. Fredens, K. Wang, D. de la Torre, L. F. H. Funke, W. E. Robertson, Y. Christova, W. H. Schmid, D. L. Dunkelmann, V. Beránek, C. Uttamapinant, A. G. Llamazares, T. S. Elliott, J. W. Chin, *Nature* **2019**, *569*, 514–518; c) N. J. Ma, C. F. Hemez, K. W. Barber, J. Rinehart, F. J. Isaacs, *eLife* **2018**, *7*, 819–22.
- [32] J. P. Oza, H. R. Aerni, N. L. Pirman, K. W. Barber, C. M. Ter Haar, S. Rogulina, M. B. Amroffell, F. J. Isaacs, J. Rinehart, M. C. Jewett, *Nat. Commun.* **2015**, *6*, 8168.
- [33] D. T. Rogerson, A. Sachdeva, K. Wang, T. Haq, A. Kazlauskaitė, S. M. Hancock, N. Huguenin-Dezot, M. M. Muqit, A. M. Fry, R. Bayliss, J. W. Chin, *Nat. Chem. Biol.* **2015**, *11*, 496–503.
- [34] H.-S. Park, M. J. Hohn, T. Umehara, L. T. Guo, E. M. Osborne, J. Benner, C. J. Noren, J. Rinehart, D. Söll, *Science* **2011**, *333*, 1151–1154.
- [35] a) L. Wang, J. Xie, P. G. Schultz, *Annu. Rev. Biophys. Biomol. Struct.* **2006**, *35*, 225–249; b) A. Dumas, L. Lercher, C. D. Spicer, B. G. Davis, *Chem. Sci.* **2015**, *6*, 50–69.
- [36] a) J. M. Bacher, A. D. Ellington, *J. Bacteriol.* **2001**, *183*, 5414–5425; b) A. C.-S. Yu, A. K.-Y. Yim, W. K. Mat, A. H.-Y. Tong, S. Lok, H. Xue, S. K.-W. Tsui, J. T.-F. Wong, T. F. Chan, *Genome Biol. Evol.* **2014**, *6*, 629–641; c) F. Agostini, L. Sinn, D. Petras, C. J. Schipp, V. Kubyskhin, A. A. Berger, P. C. Dorrestein, J. Rappsilber, N. Budisa, B. Kokschi, *ACS Cent. Sci.* **2020**, doi: 10.1021/acscentsci.0c00679.
- [37] D. A. Thomas, G. A. Rosenthal, *Toxicol. Appl. Pharmacol.* **1987**, *91*, 395–405.
- [38] U. Krasuska, O. Andrzejczak, P. Staszek, W. Borucki, A. Gniazdowska, *Plant Physiol. Biochem.* **2016**, *103*, 84–95.
- [39] H. Weissbach, L. Resnick, N. Brot, *Biochim. Biophys. Acta* **2005**, *1703*, 203–212.
- [40] a) Y. Ma, H. Biava, R. Contestabile, N. Budisa, M. L. di Salvo, *Molecules* **2014**, *19*, 1004–1022; b) C. Kofoed, S. Riesenber, J. Šmolíková, M. Meldal, S. Schoffelen, *Bioconjugate Chem.* **2019**, *30*, 1169–1174.
- [41] A. C. Yu, A. K. Yim, W. K. Mat, A. H. Tong, S. Lok, H. Xue, S. K. Tsui, J. T. Wong, T. F. Chan, *Genome Biol. Evol.* **2014**, *6*, 629–641.
- [42] K. J. Rodgers, P. M. Hume, J. G. L. Morris, R. T. Dean, *J. Neurochem.* **2006**, *98*, 1061–1067.
- [43] N. Campolo, F. M. Issoglio, D. A. Estrin, S. Bartesaghi, R. Radi, *Essays Biochem.* **2020**, *64*, 111–133.
- [44] S. Boddapati, J. Gilmore, K. Boone, J. Bushey, J. Ross, B. Gfeller, W. McFee, R. Rao, G. Corrigan, A. Chen, H. Clarke, J. Valliere-Douglass, S. Bhargava, *PLoS One* **2020**, *15*, e0241250.
- [45] a) M. Larregola, S. Moore, N. Budisa, *Biochem. Biophys. Res. Commun.* **2012**, *421*, 646–650; b) S. A. Lieblich, K. Y. Fang, J. K. B. Cahn, J. Rawson, J. LeBon, H. T. Ku, D. A. Tirrell, *J. Am. Chem. Soc.* **2017**, *139*, 8384–8387.
- [46] E. W. Weisstein, *Metric Space*, <http://mathworld.wolfram.com/MetricSpace.html>, **2019**
- [47] M. Schmidt, *3D representation of Genetic Code Metric Space* <http://markusschmidt.eu/3DG/example/GCE/index.html> **2020**
- [48] T. Mukai, A. Yamaguchi, K. Ohtake, M. Takahashi, A. Hayashi, F. Iraha, S. Kira, T. Yanagisawa, S. Yokoyama, H. Hoshi, T. Kobayashi, K. Sakamoto, *Nucleic Acids Res.* **2015**, *43*, 8111–8122.
- [49] D. E. Irwin, J. H. Irwin, T. D. Price, *Genetica* **2001**, *112–113*, 223–243.
- [50] J. Tellechea-Luzardo, C. Winterhalter, P. Widera, J. Kozyra, V. de Lorenzo, N. Krasnogor, *ACS Synth. Biol.* **2020**, *9*, 536–545.
- [51] V. de Lorenzo, N. Krasnogor, M. Schmidt, *Nat. Biotechnol.* **2020**, *60*, 44–51.
- [52] H. Neumann, K. Wang, L. Davis, M. Garcia-Alai, J. W. Chin, *Nature* **2010**, *464*, 441–444.

- [53] a) S. A. Benner, A. M. Sismour, *Nat. Rev. Genet.* **2005**, *6*, 533–543; b) Y. Zhang, B. M. Lamb, A. W. Feldman, A. X. Zhou, T. Laverigne, L. Li, F. E. Romesberg, *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 1317–1322.
- [54] J. C. Anderson, N. Wu, S. W. Santoro, V. Lakshman, D. S. King, P. G. Schultz, *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 7566–7571.
- [55] V. T. Dien, S. E. Morris, R. J. Karadeema, F. E. Romesberg, *Curr. Opin. Chem. Biol.* **2018**, *46*, 196–202.
- [56] a) I. Anosova, E. A. Kowal, M. R. Dunn, J. C. Chaput, W. D. Van Horn, M. Egli, *Nucleic Acids Res.* **2016**, *44*, 1007–1021; b) J. C. Chaput, H. Yu, S. Zhang, *Chem. Biol.* **2012**, *19*, 1360–1371.
- [57] E. Eremeeva, A. Fikatas, L. Margamuljana, M. Abramov, D. Schols, E. Groaz, P. Herdewijn, *Nucleic Acids Res.* **2019**, *47*, 4927–4939.
- [58] V. Kempeneers, M. Renders, M. Froeyen, P. Herdewijn, *Nucleic Acids Res.* **2005**, *33*, 3828–3836.
- [59] T. M. J. Fruchtermann, E. M. Reingold, *Soft. Pract. Exp.* **1991**, *21*, 1129–1164.
- [60] E. Appleton, C. Madsen, N. Roehner, D. Densmore, *Cold Spring Harbor Perspect. Biol.* **2017**, *9*, a023978.
- [61] J. Beal, A. Goñi-Moreno, C. Myers, A. Hecht, M. d. C. de Vicente, M. Parco, M. Schmidt, K. Timmis, G. Baldwin, S. Friedrichs, P. Freemont, D. Kiga, E. Ordozgoiti, M. Rennig, L. Rios, K. Tanner, V. de Lorenzo, M. Porcar, *EMBO Rep.* **2020**, *21*, e50521.
- [62] A. Golchha, S. G. Qureshi, *Int. J. Comp. Sci. Inf. Technol.* **2015**, *6*, 252–255.

Manuscript received: November 5, 2020
Revised manuscript received: November 20, 2020
Accepted manuscript online: November 24, 2020
Version of record online: December 30, 2020