Biophysical interactions underpin the emergence of information in the genetic code

Aaron Halpern[1], Lilly R. Bartsch[1], Kaan Ibrahim[1], Stuart A. Harrison[1], Minkoo Ahn[2], John Christodoulou[2], Nick Lane[1]*

[1] *UCL Centre for Life's Origins and Evolution (CLOE)*
Department of Genetics, Evolution and Environment
University College London

[2] *Department of Structural and Molecular Biology*
Institute of Structural and Molecular Biology (ISMB)
University College London

Correspondence: nick.lane@ucl.ac.uk

Key words: origin of life, genetic code, biophysical interactions, hydrophobicity, anticodon, molecular dynamics, NMR

Aaron Halpern ORCID: 0000-0002-0105-7354
Stuart Harrison ORCID: 0000-0002-5329-7747
Minkoo Ahn ORCID: 0000-0001-9131-7334
John Christodoulou: 0000-0002-6710-3843
Nick Lane ORCID: 0000-0002-5433-3973

Abstract

The genetic code conceals a ‘code within the codons’, which hints at biophysical interactions between amino acids and their cognate nucleotides. But research over decades has failed to corroborate systematic biophysical interactions across the code. Using molecular dynamics simulations and NMR, we have analysed interactions between the 20 standard proteinogenic amino acids and four RNA mononucleotides in three charge states. Our simulations show that 50% of amino acids bind best with their anticodonic middle base in the −1 charge state common to the backbone of RNA, while 95% of amino acids interact most strongly with at least one of their codonic or anticodonic bases. Preference for the cognate anticodonic middle base was greater than 99% of randomized assignments. We verify a selection of our results using NMR, and highlight challenges with both techniques for interrogating large numbers of weak interactions. Finally, we extend our simulations to a range of amino acids and dinucleotides, and corroborate similar preferences for cognate nucleotides. Despite some discrepancies between the predicted patterns and those observed in biology, the existence of weak stereochemical interactions means that random RNA sequences could template non-random peptides. This offers a compelling explanation for the emergence of genetic information in biology.

1. Introduction

"The whole case here rests upon the demonstration that codon-amino acid pairing interactions exist and that the codon assignments in some way reflect these interactions… all-or-none specificities are not required for such interactions to determine the form of the codon catalog, either a general form or one specified down to the very last detail. All that is required here is that a sufficient number of slight preferences be shown." Carl Woese, 1969.

The origin of the genetic code and the emergence of biological information is a notoriously elusive question. Even on its discovery, it was clear there are non-random patterns in the code [1–3]. These patterns loosely correspond to the biosynthetic precursors of the amino acids encoded [4–7], the hydrophobicity of those amino acids [8,9], and less clearly, their size [4,10], all of which point to some kind of direct biophysical interactions. Nonetheless, as suspected by Woese as early as 1969, 'all-or-none preferences' do not exist [1,11]. Woese argued that a large number of slight preferences (weak and not highly specific) could still have played a strong role in fixing the codon assignments [11]. But this prescient argument was displaced by Crick's adaptor hypothesis, which highlighted the lack of direct interactions between amino acids and either the codon or anticodon in tRNA [12,13]. Crick's position was unequivocal: since neither evidence for such interactions nor a reasonable model for them existed, the interactions themselves did not exist under any circumstances [9,14]. While others have continued to pursue the idea that direct stereochemical interactions underpin the code, even the most promising experiments and simulations over the following half century [15–22] did little to dispel the prevailing scepticism [23–25].

A second theme of Woese's early ideas on the code has also been largely neglected since the rise of the RNA world hypothesis. Woese noted that the main ideas on the optimization of the code implicitly assumed selection at the level of the cell, whether through minimizing the effects of mutations [26,27], constraining ambiguity [28–31], expanding an early amino-acid vocabulary [13], or fine-tuning weak biophysical preferences into the all-or-none assignments seen today [9,11,31]. In contrast, the RNA world hypothesis advocated selection at the level of the gene. Beginning with another foundational paper in 1968, Orgel questioned how far complexity could emerge in a peptide world versus an RNA world [32]. He concluded that only the simple copying of RNA templates could account for the emergence of natural selection, and so focused attention on the replication of RNA as a unit of selection [32]. By the early 1980s, the discovery of ribozymes gave the RNA world a sense of concrete reality [33–36]. Orgel's influential scenario held that "the very first replicators were 'naked genes' adsorbed on the surface of mineral particles"; and later on

“impermeable membrane caps were ‘invented’ by the genetic system as it became metabolically competent [37].” While these ideas do not rule out the emergence of the genetic code in cells, selection at the level of RNA is proposed to generate a set of functional RNA catalysts that sustain exponential growth in a prebiotic environment [36]. Metabolism emerged with ‘RNA cofactors’ such as NAD, and the first proteins performed the same reactions as ribozymes, but more effectively, thereby eventually displacing RNA [36,38]. Neither the code nor cells are emphasised as critical early steps.

There is no doubt that RNA played a determining role in the early translational system and the emergence of the code [35,39]. Yet the idea that RNA ‘invented metabolism’, though dominant, has not been expansively developed. Attention has instead focused on the difficult problem of RNA replication [38,40,41], and escaping the tendency to select for replication speed, often leading to parasitic collapse [42–46]. Producing stable longer chain RNAs has been an end in itself [47], on the assumption that natural selection can then drive the emergence of metabolism, ‘bit by bit’ [48]. But this scenario overlooks the complexity of metabolic pathways, the known endpoint of evolution. While cofactors probably played an important role, by no means do they catalyse every step. And if RNA encoded ribozymes or enzymes that each catalyzed individual steps, introducing them one at a time would have no benefit if the rest of the pathway was missing [49–51]. Building metabolic pathways step by step from one end or the other lowers the combinatorial odds but requires that all intermediates be stable, useful and available [52–54], which is certainly not the case for pathways such as purine synthesis. The problem is not simply combinatorial. As observed by Walker and Davies, “biological information has an additional quality which may roughly be called ‘functionality’ – or ‘contextuality’ – that sets it apart from a collection of mere bits as characterized by its Shannon information content.” [55] The simplest solution to this near-intractable problem is to assume that the core of metabolism is thermodynamically and kinetically favoured in a propitious environment [56–58], so the first RNA genes only had to enhance flux through the protometabolic network [59]. But that presumes a lot from prebiotic chemistry, to the point that Orgel memorably dismissed it as ‘an appeal to magic’ [60].

Experimental work over the last decade has now weakened Orgel’s position. Starting from CO_2 and H_2 – the autotrophic core of all life – much of intermediary metabolism is not only thermodynamically favoured but occurs spontaneously following the universally conserved biochemical pathways. Mineral catalysts [61,62] or proton gradients across inorganic barriers [63] can drive CO_2 fixation, directly generating the carboxylic-acid intermediates of the acetyl CoA pathway [64] and reverse Krebs cycle [65,66]. From these universal precursors, α-

amino acids [66–69], acetyl phosphate [70], sugars [71–74], and some nucleobases [75] have been formed via pathways that prefigure metabolism in the absence of genes and enzymes. While much still needs to be done, the concept of a spontaneous protometabolism is no longer an appeal to magic. Going further, modelling shows that catalytic positive feedbacks from nucleotide cofactors could drive flux through protometabolism, giving rise to autotrophic protocells growing from CO_2 and H_2 [59]. If RNA polymerization could occur within these replicating protocells, the emergence of the genetic code through weak biophysical interactions between amino acids and cognate bases could solve the RNA-world problem along the lines postulated by Woese [11].

Reexamining patterns in the code from the standpoint of autotrophic protometabolism is little short of revelatory. Harrison *et al*. have shown that the base at the first position of the codon corresponds to the distance from CO_2 fixation, following the universal metabolic map [59,76]. Amino acids encoded by a G at the first position of the codon are usually the closest to CO_2 fixation, followed by A, then C and U, which might suggest a purine-rich early metabolism [76]. Given that most nucleotide cofactors are derived from purines, including NAD, FAD, CoA, ATP, folates and pterins, the idea of a purine-rich early metabolism is consistent with cofactor-catalyzed positive feedbacks. When structured according to the base at the first position of the codon, there is a much stronger relationship between the hydrophobicity of the amino acid and the base at the second position of the anticodon – in other words, the correlation is stronger for earlier amino acids (closer to CO_2 fixation) than for later amino acids [76]. Finally, the patterns of redundancy across the code are far from random but are governed by rules pertaining partly to the size of the amino acid and adjoining bases [76]. These patterns predict weak biophysical interactions between amino acids and cognate bases. As noted, while there are tantalizing hints that such interactions do exist [15–22], they have not yet been systematically demonstrated across the full code [23,24]. Here, we have taken a novel approach, using molecular dynamics simulations to analyze the forces acting between atoms, which has allowed us to revisit the interactions between the 20 standard proteinogenic amino acids and four RNA mono-nucleotides in three distinct charge states. We find that half of the amino acids do bind best with their anticodonic middle base in the -1 charge state common to the backbone of RNA (which is greater than 99% of randomized assignments), while 95% of amino acids interact most strongly with at least one of their cognate codonic or anticodonic bases. We verify a selection of our results using NMR. Our results corroborate Woese's proposals from more than half a century ago, and offer a compelling framework for the emergence of genetic information in biology.

2. Methods

2.1 Molecular dynamics pipeline

Mol2 files representing the 20 standard proteinogenic L-amino acids in a zwitterionic state and with protonation states representative of pH 7 were produced in Avogadro [77]. Where used, dinucleotides were simulated in the NpN format. Three files for each of AMP, CMP, GMP, and UMP were also produced with the phosphate at either a -2, -1, or neutral charge. These files were then uploaded to CHARMM-GUI's Ligand Reader and Modeller [78,79], which output psf, crd, prm, and rtf files for the molecules. These were passed to CHARMM-GUI's Multicomponent Assembler [80] to produce the input files for the MD simulations. Each system contained 10 copies of an amino acid and 1 nucleotide in order to increase the frequency of collisions. A 40 Ångstrom periodic box was used, with water as the solvent. 150 mM $MgCl_2$ was added using Monte Carlo ion placement, for charge neutralization, resulting in a variable number of ions depending on the charge states of the monophosphate and amino acid. Default PME parameters were used. The temperature was set at 25 °C using an NVT ensemble.

The input files were then used as starting points for the simulations using NAMD 2.0 [81] and the CHARMM-36m forcefield [80]. Simulations were minimized and equilibrated for 10000 and 125000 timesteps respectively, and run for 48 hours using 2 MPI cores 100 times in parallel on UCL's myriad cluster, with randomized initial velocities in each parallel run. This produced total simulation times of approximately 1.5 microseconds.

Output trajectories were extracted in MATLAB R2020b using MDtoolbox [82]. The Euclidian distance between the nucleotide and the closest amino acid was calculated, as determined by the distance between the closest atom of each molecule. Binding was approximated by time spent within a 5 Å threshold. This value was chosen qualitatively as a single threshold covering the many varied binding modes observed between the large range of molecules. Uncertainty was calculated by treating the 100 parallel runs as individual experiments and bootstrap sampling these 100,000 times, recalculating the binding in each pseudo-repeat, and finding the range in which 95% of these values fell.

Because larger molecules are more likely to be close to each other simply as a result of their size, the volume of the molecules needed to be determined to enable comparisons. This was conducted through a Monte Carlo method in which 100,000 random coordinates within the periodic box were selected and the fraction of these points within 5 Å of the molecule was calculated. In order to account for molecular flexibility, this was repeated at 150 randomly

chosen simulation frames. The uncertainty presented is the 95% range of these volume fractions (SI Tables S1-3). The calculated volume fractions were determined to have a 96% correlation with empirical volume measurements for amino acids made by Tien *et al.,* 2013 [83]. Expected time spent within the 5 Å threshold was observed to increase linearly with increasing molecular volume fraction, so artefactual "binding" resulting from larger molecular volumes was eliminated by dividing the proximity-based binding measure by the molecular volume (SI Tables S7,S8, S12, S13, S17, S18, S22, S23). This enabled comparison between the various systems. The preferred binding nucleotide for each amino acid was determined by ranking the size adjusted proximity measure, and this was then compared to cognate nucleotide assignments in the genetic code, with a null hypothesis of uniform random preferences based on a binomial distribution. For overall elevation in preference, 200,000 randomized rank preferences of nucleotides to amino acids were generated and the sum of these ranks was calculated for each randomized run. Additional assignments for amino acids with multiple codons were included. For amino acids with multiple cognate dinucleotides in the dinucleotide simulations, randomized preferences for the cognate dinucleotides were not allowed to be identical.

2.2 Hydrophobicity trends

The influence of hydrophobicity was determined by using multiple linear regression. The proxy binding measures were compared against the volume of amino acids (before and after dividing by volume), the volume of nucleotides (again before and after dividing by volume), and hydrophobicity. Hydrophobicity was mainly determined using a composite scale where an amino acid's position was determined as the mean hydrophobicity rank of the amino acid across 43 hydrophobicity scales compiled by Trinquier and Sanejouand [8,76]. The individual scales were also compared on their own (SI Tables S28 and S29). Regressions were performed in MATLAB R2020b using the "fitlm" function.

2.3 Rings

Where mentioned, instead of calculating the distance between all atoms, only the distance between the nitrogen atoms in the NMP rings and atoms in the amino acids were calculated. This is an imperfect measure as nitrogen atoms are not homogenously laid out around the bases, but greatly simplified calculations due to irregular atom labelling produced by the software pipeline. This crude representation of the rings is flawed, but sufficient for some broad comparisons. Where considered, the volume of the rings was estimated using the same Monte Carlo method as for the full molecules, but using just the ring nitrogens.

2.4 NMR

Samples were created in HPLC Gradient Grade H_2O. Monophosphate nucleotides were added at 0.1 mM, while amino acid concentrations varied from 0.1-100 mM. All samples were in 10 mM phosphate buffer with 0.6 mM $MgCl_2$. The pH was adjusted to 7.40-7.42 using NaOH and HCl (measured with Fisher Scientific accumet AE150 meter with VWR semi-micro pH electrode) to mimic MD conditions. All chemicals were obtained from Fisher Chemicals and Sigma Aldrich. If required, samples were stored in a 4 °C fridge. The samples were transferred to 5 mm diameter borosilicate glass NMR tubes for 600 MHz frequencies with 10% (v/v) D_2O as the lock signal and 0.001 (w/v) DSS as an internal chemical shift reference. Proton (1H) spectra were recorded at 298.2K on a Bruker Avance II 600 MHz spectrometer equipped with a TXO cryogenic probe.

Peak locations and other features were determined using Topspin v.4.1.4 and then binding parameters including K_D and max shift were inferred by fitting the results to the ligand binding equation (equation 6) from Williamson (2013).The following protons were used as probes: CMP H5 and H6, AMP H2 and H8, UMP H5, and GMP H8. 500 fits were conducted using the "fit" function in MATLAB R2020b, with the following conditions: lower bounds of K_D = 0 M^{-1}, delta shift max = 0 ppm; no upper bounds; starting point for delta shift max = 0 ppm, and starting point for kD was randomly selected in the interval 0.5-1.5 M^{-1}. In order to mitigate noise in the data, one datapoint was randomly excluded in each fit. The K_D was determined as the mean of these fits, with uncertainty as the range in which 95% of the fitted constants fell. The relative preferences for NMPs were further compared by randomly selecting inferred binding constants and ranking them 10,000 times.

3. Results

3.1 Amino acids prefer cognate nucleotides

We simulated all 20 proteinogenic amino acids as zwitterions in the protonation state expected at pH 7, with mononucleotides in each of three charge states; −2, −1 and 0. Each simulation had a 10:1 ratio of amino acids to mononucleotides, as well as 150 mM $MgCl_2$. The large number of amino acids increased the frequency of collisions in the simulations, strengthening the signal from subtle differences between the weakly interacting molecules. The systems were simulated for approximately 1.5 μs timescales. The charge state for dissolved mononucleotides at pH 7 should be −2, but we also used the −1 charge state as that is more representative of what would be found in an RNA backbone (while lowering the combinatorial odds of interacting). Mononucleotides are unlikely to ever be in the neutral

state under any relevant situations, but we performed these experiments anyway to better understand the effect of charge on the interactions.

Figure 1 gives a selection of examples showing how the interactions vary between amino-acid–mononucleotide pairs, with mononucleotides in the −1 charge state. The figure demonstrates how preference for spending time at a given proximity varies. The proximity distribution is adjusted for molecular volume because larger molecules tend to be closer to one another simply due to the fact that they take up more of the simulation space. Figure 1A shows that proline spends a large proportion of simulation time bound to GMP at 1.9 Å, but is much less likely to be found at the same distance from the other three nucleotides. Another interaction mode is demonstrated at 2.5 Å, but proline behaves similarly with the four nucleotides here. There is third interaction mode at 4 Å, but for proline this is relatively indistinct.

Figure 1B shows arginine, which demonstrates more subtle and complex differences in preferences between the four nucleotides. Figure 1C shows aspartate, which also shows relatively weak binding overall, but quite dramatic preferences for GMP and UMP at 2.5 Å. Figure 1D shows phenylalanine, which most commonly interacts at 2.5 Å rather than 1.9 Å. Finally, Figure 1E shows glycine, which interacts similarly with each of the four nucleotides, and demonstrates most clearly the 4 Å binding mode. Because of the wide range and complexity of the binding interactions, an agnostic approach was taken to determining relative preferences, whereby the total amount of time the molecules spent within 5 Å of one another was calculated (by integrating the area under the peaks within 5 Å). This threshold, as demonstrated by the vertical red dotted lines in Figure 1, aims to encompass all observed binding modes while avoiding the 'free-in-solution' behaviour where the closest unbound amino acid tends to be around 5.5 Å from the mononucleotide on average.

Figure 2 shows a summary of binding preferences between the full set of 20 amino acids and the four nucleotides in the −1 charge state. Binding preferences were calculated based on the proportion of simulation time spent within the 5 Å threshold, as shown in Figure 1. This measure has been adjusted to account for amino acid and nucleotide volume, and focuses specifically on interactions with the nucleobase. Three additional pairings were included for amino acids with multiple cognate 1st and 2nd nucleotide base assignments. Figure 2A shows a significant elevation in amino acids that spend the greatest proportion of simulation time with their cognate anticodonic middle base [P = 0.0139] in the −1 charge state (which best matches the RNA backbone). This is the case for half of all amino acids

(11 of 23 cognate nucleotides). There is a corresponding significant reduction in the number of amino acids that have the least favourable interactions with their anticodonic middle base [P = 0.0492]. Figure 2B shows that 95% of amino acids bound best to at least one cognate nucleotide in either their codon or anticodon (excluding base 3) [P = 0.0243] under these conditions. Glycine was the only exception. We included the same three additional cognate codons/anticodons for the same reasons as in Figure 2A. However, no specific first choice preferences for cognate nucleotides in the −2 or neutral charge states were predicted using these measures (SI Figures 2 and 3).

Figure 2C breaks down these preferences more granularly, giving the ranked preferences of each amino acid by hydrophobicity. This reveals the prediction that hydrophilic amino acids are more likely to bind to the anticodonic middle base than more hydrophobic amino acids. We can also see that UMP is predicted to be the most commonly preferred binding partner, which matches the observation from the modern codon table that UMP is the most common anticodonic middle base, utilized by 7 of 20 amino acids, and it is never redundant [76]. This trend was repeated in the neutral state, but did not appear in the −2 state (SI Figure 3). Notably, AMP was predicted to be the least favoured binding partner; it was never ranked as first choice by any amino acids in either the −1 or neutral state (Figure 2C, SI Figure 2), though this bias was not displayed in the −2 state. Given that AMP is the most hydrophobic nucleotide, this finding seems to indicate that our molecular dynamics simulations do not model hydrophobic interactions well.

Figure 2D compares the predicted preferences of the cognate middle bases to randomized preferences. For example, in the scenario where all amino acids spent the highest proportion of simulation time within 5 Å of their anticodonic middle base nucleotide, the score for this base would be rank 1 x 20 = 20. The worst case scenario would be the cognate base being rank 4 x 20 = 80. The three nucleotides with multiple codon assignments are also included, increasing the minimum possible score by 3. Left of the red dotted line represents the best 5% of randomized assignments. Compared to random assignments, cognate anticodonic middle bases were more strongly preferred than 99% of randomized pairings. An elevated preference was also observed for codon base 1, which gave a higher rank preference than 82% of randomized assignments. Elevations in affinity to codon base 1 nucleotides were also observed in the −2 state, with rank preferences greater than 95% of randomized assignments, but no elevation was observed for anticodonic nucleotides (SI Figure 3). With nucleotides in the neutral charge state, behaviour was indistinguishable from random (SI Figure 2).

3.2 Hydrophobicity plays a role in binding

To explore how far amino-acid–nucleotide interactions were influenced by their relative hydrophobicity, we compared the interactions between the nucleobase rings against the hydrophobicity of the amino acids. We continued to use the volume-adjusted proximity measure. Due to the large variation in hydrophobicity given by different scales, we primarily utilized the mean of 43 scales collated by Trinquier and Sanejouand [8]. Figure 3 shows how the interactions were influenced by relative hydrophobicity in neutral, −1 and −2 charge states for each of the four nucleotides. The hydrophobicity rankings on the X axis of Figure 3 are the same as those in Figure 2C, with the most hydrophobic amino acids being allocated the lowest numbers (at left), and the most hydrophilic with the highest numbers (at right).

Overall, we found hydrophobicity to be a significant factor influencing binding in some cases, but not all (SI Tables S24 and S25). In the neutral (Figure 3A) and −1 states (Figure 3B), the more hydrophobic amino acids tended to bind more strongly across the board, i.e. the slight negative correlation indicates that the more hydrophobic amino acids bound best. But we expected to see the strongest inverse relationship with the most hydrophobic base (A), and that was not the case. Conversely, we expected the opposite relationship with uracil, the most hydrophilic base, but again that was not the case. GMP was the only exception in the −1 state. In this case, the positive correlation shows that the more hydrophobic amino acids bound less well than their hydrophilic counterparts. While G is sometimes considered to be a relatively hydrophobic base, the hydrophobicity of the bases is ambiguous, and we have followed Lacey *et al.* [15,85] in considering C to be more hydrophobic than G.

In the −2 state (Figure 3C), the trends were less clear and not statistically significant at the 5% threshold. However, if the entire nucleotide was considered instead of just the rings, the −2 state showed a very strong negative dependence on hydrophobicity, meaning the most hydrophilic amino acids bound most strongly to all nucleotides (SI Figure 4). A similar relative decrease in the binding of hydrophobic amino acids was observed for the other charge states when considering the whole mononucleotide, suggesting that the phosphate, especially when charged, was generally interacting with hydrophilic amino acids, whereas the rings were usually interacting with hydrophobic amino acids.

We also found that different specific hydrophobicity scales predicted different dependencies on hydrophobicity, but broadly trends were similar across scales (SI Tables 28 and 29). The strongest hydrophobicity dependence for ring interactions in the −1 state was predicted by

the Krigbaum scale [86] (based on protein geometry) and the Sweet scale [87] (derived from mutational matrices) for the whole-molecule interactions. Krigbaum also predicted the strongest hydrophobicity dependence for the neutral state, but Sweet was once again the most predictive scale in the −2 state. This further suggests a complex interplay between ring and phosphate interactions, where the −1 state may balance both dynamics in these simulations.

3.3 NMR corroborates binding interactions

In order to validate the existence of preferential binding of the sort predicted by the molecular dynamics, we attempted to measure and compare the behaviour of a small selection of amino acid-mononucleotide pairs using NMR. We chose to use phenylalanine, arginine, glycine, and aspartate, hoping to cover a variety of hydrophobicities and charges. We produced mixtures of amino acids and mononucleotides with 150 mM $MgCl_2$ in potassium phosphate buffer at pH 7. We generated a range of ratios of amino acids to mononucleotides, maintaining the mononucleotides as 0.1 mM concentration and varying amino acid concentrations from 0.1 mM up to 100 mM. Increasing the concentration of a binding ligand, in this case the amino acids, should increase the proportion of the nucleotides in a bound complex. Effective binding changes the local environment of protons near the binding site, resulting in chemical shift perturbations (CSPs) of their peaks on the NMR spectra (Figure 4). The extent of CSPs will be modulated by varying amino acid concentrations, which allows us to infer details of the binding, such as binding strength, K_D [84]. K_D measures the proportion of molecules in a complex for a given set of concentrations, where a lower K_D indicates that more molecules form a complex as a result of stronger binding.

Shifts in proton peak location characteristic of binding interactions were identified in the overwhelming majority of amino acid-mononucleotide pairs (SI Tables S30-S33). While some systems appeared to reach maximum peak shift quickly (glycine and certain proton probes for aspartate), we were unable to reach saturation while retaining consistent conditions for certain pairings of phenylalanine and arginine. This supports the prediction of diverse and distinctive interactions from the molecular dynamics. We found that many of these systems were also very sensitive, producing noisy results at low concentrations. The inability to achieve binding saturation also produced large ranges for the inferred binding constants for some pairs. The varied structures of the different nucleobases also meant that proton probes are not distributed consistently (SI Figure 1), making direct comparisons of binding challenging because not all binding sites have corresponding proton probes.

Despite these issues, the results show that two of the four tested amino acids preferentially bound their anticodonic middle base (Figure 5), as determined by comparing the inferred K_d between curve fits. These were arginine, which bound best to CMP (the cognate anticodon) in 85% of fits, and aspartate which bound best to UMP (the cognate anticodon) in 76% of fits. Aspartate also demonstrated elevated preference for AMP (the cognate codon) in 23% of fits. In contrast, phenylalanine demonstrated binding preferences for its codonic middle base instead, binding best to UMP (the cognate codon) in 60% of fits. AMP (the cognate anticodon) was actually its worst binding partner in 77% of fits. Glycine was the only amino acid that had no strong preference for either its codonic or anticodonic middle base, though GMP (the cognate codon) was its second best partner in 93% of fits. However, we note that glycine almost always interacted preferentially with AMP 8, whereas AMP 2 was typically the worst binding partner. AMP 8 is the proton next to the glycosidic nitrogen, while AMP 2 is the on the opposite side of the nucleobase (SI Figure S1), suggesting that smaller amino acids might interact with only parts of larger nucleobase rings. This sort of regional preference may be important, as ignoring AMP 8 would make GMP (the cognate codon) the preferred binding partner in 95% of fits and CMP (cognate anticodon) the best in 5%, with this preference swapped for second best partner. Conversely, AMP 8 was a poor binding target for aspartate, whereas AMP 2 displayed relatively strong binding in this case. Similar variation appears to be present for CMP 5 vs CMP 6, highlighting the wider problem of inconsistency in the proton probes between nucleotides.

Importantly, our NMR results follow similar behaviour to the molecular dynamic simulations on the microscale, wherein an elevated preference for the cognate nucleotides appears to be demonstrated. We note that the preference for the anticodonic middle base is predicted by both NMR and MD for arginine and aspartate, and that phenylalanine's preference for UMP is repeated in both cases (even though this is not the cognate anticodon). Glycine's lack of clear preference for cognate nucleotides is also predicted by both techniques. While our NMR investigations only analysed 20% of proteinogenic amino acids, we suspect that these patterns would continue to emerge over larger numbers of amino acids. Extending this experimental avenue is a goal for future research, although it will also be worth moving towards polynucleotides, which could allow for greater demonstrations of specificity and wider options for proton probes.

3.4 Dinucleotides also show affinity for their cognate amino acids:

In order to take the first steps towards more complex systems, we simulated a selection (approximately 20%) of the 320 proteinogenic amino acid and dinucleotide pairs, using the same simulation and analysis pipelines as we did for the mononucleotides. This included six amino acids (Phe, Arg, Ser, Gly, Ala and Asp) and 11 of the 16 dinucleotides, including all the codons and anticodons for the six amino acids plus a selection of other dinucleotides, including homodimers and heterodimers (SI Table S19-23). Figure 6 shows that these amino acids had weak elevations in preference for cognate dinucleotides.

We observed that amino acids had a higher preference for their codonic nucleotides than about 73% of randomized assignments when considering either whole dinucleotides or the rings in isolation (Figure 6A and B). The amino acids had a higher affinity for their cognate anticodonic nucleotides than 65% of randomized assignments when considering the whole molecule (Figure 6A), which increased to 83% when considering the ring alone (Figure 6B). The previously observed high affinity for U was repeated with the dinucleotides – in this case, the dinucleotides with the highest average affinity across the six amino acids were UU, CU, and UC in order of most to least favoured. Directionality effects were also observed for a small subset of pairings, with differences in preferences predicted for glycine with AG vs GA and with arginine and glycine for CG vs GC. This was also observed for the rings in isolation, with differences between CG and GC for glycine, arginine and phenylalanine, as well as between CU and UC for glycine and phenylalanine (SI Table S22 and S23).

4. Discussion

In this work, we have explored whether preferential stereochemical interactions between amino acids and nucleotides exist, and if so, whether the interactions match the patterns observed in the genetic code. Our results, using both MD and NMR, strongly support the hypothesis that stereochemical interactions do exist, and we find they often do match the modern codon assignments. The interactions are weak and probabilistic, but at scale they can be identified through the noise. Other notable features of the codon table, such as the frequency and non-redundance of uracil at anticodon base 2, and the patterning related to hydrophobicity, also appear to arise from these biophysical interactions. These patterns even hold true in modern codon reassignments, with 77% of known codon reassignments retaining the same middle base [88,89]. Taken together, our results corroborate Woese's prescient conjectures from more half a century ago, that the genetic code is based on a set of ancient and spontaneous interactions that have not been overwritten since the origin of life [1,11].

While this central idea is supported by both MD and NMR, some discrepancies remain. In the NMR, all the nucleotides are likely to be in the −2 charge state at pH 7 [90] but the MD simulations mainly predicted cognate preferences in the −1 state (which matters because this corresponds to the charge state in the RNA backbone). Exact rank preferences were also not replicated perfectly between the two techniques. More generally, while our results support the hypothesis that hydrophobicity shapes interactions, we have struggled to identify patterns that directly resemble those observed in the codon table. For example, being the most hydrophobic base [15,85] we expected AMP to interact most strongly with hydrophobic amino acids. UMP is the least hydrophobic base [15,85], so we expected it to interact most strongly with hydrophilic amino acids. We anticipated similar but weaker patterning for CMP and GMP, which are intermediate in their hydrophobicity. But none of these differences were recovered in our simulations. We suspect that this discrepancy may reflect limitations in the fixed-charge forcefields used in our MD simulations, which also struggle to model dynamic charges and subtle changes in electron densities [91,92]. If so, then hydrophobic effects were poorly captured in our simulations, which could explain the poor overall binding of amino acids to AMP. This interpretation is supported by Figure 2, which shows that our simulations correctly predicted the cognate anticodon middle base for most hydrophilic amino acids (to the right end of the Trinquier scale) but fared badly with the hydrophobic amino acids (to the left). The negative charge on the phosphate group might then be an overpowering factor in our simulations, as suggested by the strong negative correlation between hydrophobicity and amino-acid binding in the −2 state across all nucleotides (SI Figure 4). Because this trend reverses as charge is decreased towards zero (SI Figure 3) the charge on the phosphate seems to play a dominant role.

Another puzzle relates to the striking prediction that the anticodonic middle base and first codonic base were both preferred binding partners for cognate amino acids (Figure 2). This matches the observation that clear patterning is observed for both these bases in the codon table [76]. It also matches the complementary prediction from Figure 6, showing an elevated affinity of amino acids to the cognate anticodonic and codonic dinucleotides. These findings suggest the emergence of coding in some sort of binding pocket containing both the codon and anticodon. Nonetheless, it is still surprising that the binding affinities of amino acids to dinucleotides were not greater than those for mononucleotides. One possible explanation might be that very short polynucleotides are a known weak-spot for MD [93]. Or it could be that our results reflect a sampling bias due to the relatively limited number of amino-acid–dinucleotide pairs simulated. More interesting: if the patterns do reflect binding to a pocket, this would probably require the cognate nucleotides to be positioned opposite one another

(as in normal codon-anticodon interactions). Without this multidirectional binding, the extra complexity of dinucleotides might confound preferences rather than strengthen specificity.

If our results do indeed point to some sort of selective RNA binding pocket for amino acids, then the challenge becomes: how could this pocket evolve into to the modern translational and informational system? While much remains ambiguous, we imagine a model in which non-enzymatic chemistry and stereochemical interactions between amino acids and their cognate nucleotides could build incrementally towards the modern translational system. At issue here, once more, is Crick's adaptor hypothesis [12,13], which stresses the absence of any direct correspondence between either the codon or anticodon and amino-acid binding. On the contrary, on tRNA, the amino acid always binds to a CCA acceptor stem at one end of the molecule, while the anticodon-codon interactions take place far away at another end of the molecule. While there may be interactions between anticodons and amino acids in the ribosome [94], these have nothing to do with amino-acid loading of the tRNA by aminoacyl tRNA synthetases. The question then becomes how, physically, could the interaction between a binding pocket containing the anticodon and its cognate amino acid become separated into an interaction between a CCA acceptor stem and the cognate amino acid, and elsewhere, between the anticodon on tRNA and the codon on mRNA? We sketch a possible model in Figure 7.

We propose that translation began in autotrophically growing protocells, as outlined in the Introduction [59,76]. In this structured setting, the undirected polymerization of nucleotides and amino acids could in theory occur spontaneously, driven by nucleoside triphosphates, notably ATP [95], and catalysed by metal ions such as Mg^{2+} and amino acids such as aspartate (which is conserved in the active sites of modern RNA polymerases [96]) or lysine (which is conserved in the active site of modern RNA ligase enzymes [97]). Nucleotide polymerization in turn should form short non-templated RNA aptamers, some of which may resemble a single hairpin loop of tRNA in structure [18,98] as depicted in Figure 7A. We imagine amino acids binding to these RNA pockets by way of the weak biophysical affinities demonstrated here, giving a statistical likelihood of repeatability. This biophysical patterning can in principle explain the emergence of information in biology. Consider: if random RNA sequences bind amino acids in a non-random fashion, and this facilitates the polymerization of those amino acids into short templated peptides with non-random sequences, then biological meaning, linked with function, is introduced in the context of autotrophically growing protocells. Functions in growing protocells could include CO_2 fixation, RNA

polymerization and cofactor binding, all of which would facilitate protocell growth and heritability [59,76].

For RNA to template functional peptides, the next necessary step would be the transfer of the amino acid from its binding pocket onto the proto-tRNA acceptor stem. The universal acceptor stem is the CCA terminal of tRNA, and this is rigorously enforced in biology [99]. Curiously, CCA is an anticodon for glycine, the simplest amino acid with only an H for an R group, meaning that CCA is most likely to interact with the amino or carboxyl groups rather than the R group. This general binding affinity means that the CCA could act as a universal 'fishing rod' for all amino acids; the presence of a CCA terminus could facilitate the binding of amino acids and may have been what initially began to differentiate proto-tRNA from other sequences. In effect, any short RNA hairpin with a terminal CCA would behave like a proto-tRNA. The stacking of ATP on the terminal AMP of the CCA stem would help colocalise the ATP and amino acid, which is an essential function of amino acyl-tRNA synthetases (aaRS) as depicted in Figure 7A. Progenitors of aaRS, even very short polypeptides [18,100], have been shown to catalyse amino-acid adenylation and even tRNA acylation, by protecting the adenylate from water, and colocalising reactants. We show the adenylation of amino acid in Figure 7B, followed by transfer of the amino acid onto the CCA acceptor stem in Figure 7C, which acylates the tRNA. Thus, with no more than a short tRNA, we can picture the binding of an amino acid to a specific pocket, followed by its adenylation and acylation of the tRNA. These simple proto-tRNA molecules could eventually augment their specificity beyond the simple biophysical preferences shown here, for example through the size discrimination of amino acids, which constitutes the deep split between the two classes of aaRS [101].

Exactly how such a tRNA could facilitate amino acid polymerization is another question. As depicted in Figure 7D, we imagine that the anticodon was initially positioned at the opposite end of the tRNA hairpin loop to the CCA acceptor stem, on a flexible hinge that could twist around to interact directly with the codon on a proto-rRNA. In Figure 7D, we depict a small peptide growing from these interactions, but have not specified a mechanism. It is feasible that other short RNAs could catalyse amino-acid polymerization by colocalizing proto-tRNA and mRNA templates. Such an assemblage of RNA would be the first steps to forming the proto-ribosome, and there is some evidence of this process in the structure and sequences of modern ribosomes [102,103]. These RNA complexes presumably catalysed peptide bond formation, but later proto-ribosomes had to facilitate alignment with templates and begin to enforce reading frames from looser stereochemical roots.

This model is admittedly based on some rather large extrapolations from the literature, but provides a route to build a full translational system from the simple biophysical interactions observed here. The critical steps to test will be the non-enzymatic chemistry of nucleotide polymerization in water, the specificity of RNA pockets for amino acids, and the templated polymerization of peptides on RNA. We will also address some of the mechanistic puzzles relating to the complexification of key components, notably tRNA, which eventually moves the cognate nucleotides away from the acceptor stem. More mundane but immediate goals include exploring a wider range of experimental conditions for MD simulations. Advances in polarizable forcefields would also improve our results if they are able to simulate non-polar interactions with more sophistication. There is also plenty of scope for further work with the NMR, including improving the precision of the experimentally determined binding constants, investigating more amino acids, and progressing to polynucleotides, though the number of possible combinations when using longer RNAs will present a challenge to comprehensive exploration.

In conclusion, the results presented here support the existence of weak and probabilistic binding preferences between amino acids and nucleotides, as argued by Woese more than 50 years ago. Our results point to the origin of translation in binding pockets in RNA hairpin loops. The fact that these interactions are evident even with mononucleotides suggests that genetic information is based on spontaneous interactions built into the structure of the code from the very origins of polymerization. That these biophysical interactions still shine through the genetic code shows they form a cornerstone that has supported the dazzling complexity of life ever since.

Acknowledgements

We thank Professor Andrew Pomiankowski and Raquel Nunes Palmeira for stimulating conversations about the genetic code. This work was supported by the Biomolecular NMR Facility at UCL, and by funding from the Biotechnology and Biological Sciences Research Council to NL (BB/V003542/1) and from the Natural Environment Research Council to AH and NL (2236041).

References

(1) Woese, C. R. Order in the Genetic Code. *Proc Natl Acad Sci U S A* **1965**, *54* (1), 71–75.

(2) Eck, R. V. Genetic Code: Emergence of a Symmetrical Pattern. *Science* **1963**, *140* (3566), 477–481.

(3) Nirenberg, M. W.; Jones, O. W.; Leder, P.; Clark, B. F. C.; Sly, W. S.; Pestka, S. On the Coding of Genetic Information. *Cold Spring Harb Symp Quant Biol* **1963**, *28*, 549–557.

(4) Taylor, F. J. R.; Coates, D. The Code within the Codons. *Biosystems* **1989**, *22* (3), 177–187.

(5) Copley, S. D.; Smith, E.; Morowitz, H. J. A Mechanism for the Association of Amino Acids with Their Codons and the Origin of the Genetic Code. *Proceedings of the National Academy of Sciences of the United States of America* **2005**, *102* (12), 4442–4447.

(6) Wong, J. T. A Co-Evolution Theory of the Genetic Code. *Proceedings of the National Academy of Sciences of the United States of America* **1975**, *72* (5), 1909–1912.

(7) Di Giulio, M. An Autotrophic Origin for the Coded Amino Acids Is Concordant with the Coevolution Theory of the Genetic Code. *J Mol Evol* **2016**, *83* (3–4), 93–96.

(8) Trinquier, G.; Sanejouand, Y. H. Which Effective Property of Amino Acids Is Best Preserved by the Genetic Code? *Protein Engineering Design and Selection* **1998**, *11* (3), 153–169.

(9) Woese, C. R.; Dugre, D. H.; Saxinger, W. C.; Dugre, S. A. The Molecular Basis for the Genetic Code. *Proc Natl Acad Sci U S A* **1966**, *55* (4), 966–974.

(10) Fontecilla-Camps, J. C. The Stereochemical Basis of the Genetic Code and the (Mostly) Autotrophic Origin of Life. *Life (Basel)* **2014**, *4* (4), 1013–1025.

(11) Woese, C. Models for the Evolution of Codon Assignments. *J Mol Biol* **1969**, *43* (1), 235–240.

(12) Crick, F. H. On Protein Synthesis. *Symp Soc Exp Biol* **1958**, *12*, 138–163.

(13) Crick, F. H. The Origin of the Genetic Code. *Journal of molecular biology* **1968**, *38* (3), 367–379.

(14) Crick, F. H. C., as quoted by Hoagland, M., in The Nucleic Acids, ed. E. Chargaff and J. Davidson (New York: Academic Press, 1960), vol. 3; Crick, F. H. C., J. Griffith, and L. Orgel,
these PROCEEDINGS, 43, 416 (1957).

(15) Weber, A. L.; Lacey, J. C. Genetic Code Correlations: Amino Acids and Their Anticodon Nucleotides. *Journal of Molecular Evolution* **1978**, *11* (3), 199–210.

(16) Lacey, J. C.; Pruitt, K. M. Origin of the Genetic Code. *Nature* **1969**, *223* (5208), 799–804.

(17) Lacey, J. C.; Mullins, D. W. Experimental Studies Related to the Origin of the Genetic Code and the Process of Protein Synthesis--a Review. *Orig Life* **1983**, *13* (1), 3–42.

(18) Shimizu, M. Specific Aminoacylation of C4N Hairpin RNAs with the Cognate Aminoacyl-Adenylates in the Presence of a Dipeptide: Origin of the Genetic Code. *Journal of Biochemistry* **1995**, *117* (1), 23–26.

(19) Shimizu, M. Molecular Basis for the Genetic Code. *J Mol Evol* **1982**, *18* (5), 297–303.

(20) Yarus, M.; Widmann, J. J.; Knight, R. RNA–Amino Acid Binding: A Stereochemical Era for the Genetic Code. *Journal of Molecular Evolution* **2009**, *69* (5), 406–429.

(21) Yarus, M.; Caporaso, J. G.; Knight, R. ORIGINS OF THE GENETIC CODE: The Escaped Triplet Theory. *Annual Review of Biochemistry* **2005**, *74* (1), 179–198.

(22) Hobish, M. K.; Wickramasinghe, N. S. M. D.; Ponnamperuma, C. Direct Interaction between Amino Acids and Nucleotides as a Possible Physicochemical Basis for the Origin of the Genetic Code. *Advances in Space Research* **1995**, *15* (3), 365–382.

(23) Moghadam, S. A.; Preto, J.; Klobukowski, M.; Tuszynski, J. A. Testing Amino Acid-Codon Affinity Hypothesis Using Molecular Docking. *BioSystems* **2020**, *198*.

(24) Koonin, E. V.; Novozhilov, A. S. Origin and Evolution of the Genetic Code: The Universal Enigma. *IUBMB Life* **2009**, *61* (2), 99–111.

(25) Di Giulio, M. Arguments against the Stereochemical Theory of the Origin of the Genetic Code. *Biosystems* **2022**, *221*, 104750.

(26) Haig, D.; Hurst, L. D. A Quantitative Measure of Error Minimization in the Genetic Code. *J Mol Evol* **1991**, *33* (5), 412–417.

(27) Sonneborn, T. M. (1965). In Evolving Genes and Proteins, ed. by V. Bryson & H. Vogel, p. 377. New York: Academic Press.

(28) Barbieri, M. Evolution of the Genetic Code: The Ambiguity-Reduction Theory. *Biosystems* **2019**, *185*, 104024.

(29) Woese, C. R. On the Evolution of the Genetic Code. *Proc Natl Acad Sci U S A* **1965**, *54* (6), 1546–1552.

(30) Delarue, M. An Asymmetric Underlying Rule in the Assignment of Codons: Possible Clue to a Quick Early Evolution of the Genetic Code via Successive Binary Choices. *RNA* **2007**, *13* (2), 161–169.

(31) Woese, C. R. The Fundamental Nature of the Genetic Code: Prebiotic Interactions between Polynucleotides and Polyamino Acids or Their Derivatives. *Proceedings of the National Academy of Sciences of the United States of America* **1968**, *59* (1), 110–117.

(32) Orgel, L. E. Evolution of the Genetic Apparatus. *J Mol Biol* **1968**, *38* (3), 381–393.

(33) Kruger, K.; Grabowski, P. J.; Zaug, A. J.; Sands, J.; Gottschling, D. E.; Cech, T. R. Self-Splicing RNA: Autoexcision and Autocyclization of the Ribosomal RNA Intervening Sequence of Tetrahymena. *Cell* **1982**, *31* (1), 147–157.

(34) Guerrier-Takada, C.; Gardiner, K.; Marsh, T.; Pace, N.; Altman, S. The RNA Moiety of Ribonuclease P Is the Catalytic Subunit of the Enzyme. *Cell* **1983**, *35* (3 Pt 2), 849–857.

(35) Cech, T. R. The RNA Worlds in Context. *Cold Spring Harbor Perspectives in Biology* **2012**, *4* (7), a006742–a006742.

(36) Gilbert, W. Origin of Life: The RNA World. *Nature* **1986**, *319* (6055), 618–618.

(37) Leslie E., O. Prebiotic Chemistry and the Origin of the RNA World. *Critical Reviews in Biochemistry and Molecular Biology* **2004**, *39* (2), 99–123.

(38) Joyce, G.F. and Orgel, L.E., 1993. Prospects for understanding the origin of the RNA world. Cold Spring Harbor Monograph Series, 24, pp.1-1.

(39) Bose, T.; Fridkin, G.; Davidovich, C.; Krupkin, M.; Dinger, N.; Falkovich, A. H.; Peleg, Y.; Agmon, I.; Bashan, A.; Yonath, A. Origin of Life: Protoribosome Forms Peptide Bonds and Links RNA and Protein Dominated Worlds. *Nucleic Acids Research* **2022**, *50* (4), 1815–1828.

(40) Joyce, G. F. Evolution in an RNA World. *Cold Spring Harb Symp Quant Biol* **2009**, *74*, 17–23. https://doi.org/10.1101/sqb.2009.74.004.

(41) Orgel, L. E. Some Consequences of the RNA World Hypothesis. *Orig Life Evol Biosph* **2003**, *33* (2), 211–218.

(42) Spiegelman, S.; Haruna, I.; Holland, I. B.; Beaudreau, G.; Mills, D. The Synthesis of a Self-Propagating and Infectious Nucleic Acid with a Purified Enzyme. *Proceedings of the National Academy of Sciences of the United States of America* **1965**, *54* (3), 919–927.

(43) Szathmáry, E.; Demeter, L. Group Selection of Early Replicators and the Origin of Life. *J Theor Biol* **1987**, *128* (4), 463–486.

(44) Matsumura, S.; Kun, Á.; Ryckelynck, M.; Coldren, F.; Szilágyi, A.; Jossinet, F.; Rick, C.; Nghe, P.; Szathmáry, E.; Griffiths, A. D. Transient Compartmentalization of RNA Replicators Prevents Extinction Due to Parasites. *Science* **2016**, *354* (6317), 1293–1296.

(45) Smith, J. M. Hypercycles and the Origin of Life. *Nature* **1979**, *280* (5722), 445–446.

(46) Eigen, M. Selforganization of Matter and the Evolution of Biological Macromolecules. *Naturwissenschaften* **1971**, *58* (10), 465–523.

(47) Salditt, A.; Keil, L. M. R.; Horning, D. P.; Mast, C. B.; Joyce, G. F.; Braun, D. Thermal Habitat for RNA Amplification and Accumulation. *Phys. Rev. Lett.* **2020**, *125* (4), 048104.

(48) Joyce, G.F., 1995. The RNA world: Life before DNA and protein. Extraterrestrials: where are they, 2, pp.139-151.

(49) Ralser, M. The RNA World and the Origin of Metabolic Enzymes. *Biochem Soc Trans* **2014**, *42* (4), 985–988.

(50) Ralser, M. An Appeal to Magic? The Discovery of a Non-Enzymatic Metabolism and Its Role in the Origins of Life. *Biochem J* **2018**, *475* (16), 2577–2592.

(51) Harrison, S. A.; Lane, N. Life as a Guide to Prebiotic Nucleotide Synthesis. *Nature Communications* **2018**, *9* (1), 1–4.

(52) Lazcano, A.; Miller, S. L. On the Origin of Metabolic Pathways. *J Mol Evol* **1999**, *49* (4), 424–431.

(53) Horowitz, N. H. On the Evolution of Biochemical Syntheses. *Proc Natl Acad Sci U S A* **1945**, *31* (6), 153–157.

(54) Horowitz NH (1965) The evolution of biochemical synthesis— Retrospect and prospect. In: Bryson V, Vogel HJ (eds). Evolving genes and proteins. Academic Press, New York, pp 15–23

(55) Walker, S. I.; Davies, P. C. W. The Algorithmic Origins of Life. *Journal of The Royal Society Interface* **2013**, *10* (79), 20120869.

(56) Amend, J. P.; McCollom, T. M. Energetics of Biomolecule Synthesis on Early Earth. In *Chemical Evolution II: From the Origins of Life to Modern Society*; ACS Symposium Series; American Chemical Society, 2009; Vol. 1025, pp 63–94. https://doi.org/10.1021/bk-2009-1025.ch004.

(57) Amend, J. P.; LaRowe, D. E.; McCollom, T. M.; Shock, E. L. The Energetics of Organic Synthesis inside and Outside the Cell. *Philosophical Transactions of the Royal Society B: Biological Sciences* **2013**, *368* (1622), 20120255.

(58) Wimmer, J. L. E.; Xavier, J. C.; Vieira, A. d. N.; Pereira, D. P. H.; Leidner, J.; Sousa, F. L.; Kleinermanns, K.; Preiner, M.; Martin, W. F. Energy at Origins: Favorable Thermodynamics of Biosynthetic Reactions in the Last Universal Common Ancestor (LUCA). *Frontiers in Microbiology* **2021**, *12*.

(59) Nunes Palmeira, R.; Colnaghi, M.; Harrison, S. A.; Pomiankowski, A.; Lane, N. The Limits of Metabolic Heredity in Protocells. *Proceedings of the Royal Society B: Biological Sciences* **2022**, *289* (1986), 20221469.

(60) Orgel, L. E. Self-Organizing Biochemical Cycles. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97* (23), 12503–12507.

(61) Preiner, M.; Igarashi, K.; Muchowska, K. B.; Yu, M.; Varma, S. J.; Kleinermanns, K.; Nobu, M. K.; Kamagata, Y.; Tüysüz, H.; Moran, J.; Martin, W. F. A Hydrogen-Dependent Geochemical Analogue of Primordial Carbon and Energy Metabolism. *Nat Ecol Evol* **2020**, *4* (4), 534–542.

(62) Beyazay, T.; Belthle, K. S.; Farès, C.; Preiner, M.; Moran, J.; Martin, W. F.; Tüysüz, H. Ambient Temperature CO2 Fixation to Pyruvate and Subsequently to Citramalate over Iron and Nickel Nanoparticles. *Nat Commun* **2023**, *14* (1), 570.

(63) Hudson, R.; de Graaf, R.; Strandoo Rodin, M.; Ohno, A.; Lane, N.; McGlynn, S. E.; Yamada, Y. M. A.; Nakamura, R.; Barge, L. M.; Braun, D.; Sojo, V. CO2 Reduction Driven by a PH Gradient. *Proceedings of the National Academy of Sciences* **2020**, *117* (37), 22873–22879.

(64) Varma, S. J.; Muchowska, K. B.; Chatelain, P.; Moran, J. Native Iron Reduces CO2 to Intermediates and End-Products of the Acetyl-CoA Pathway. *Nat Ecol Evol* **2018**, *2* (6), 1019–1024.

(65) Muchowska, K. B.; Varma, S. J.; Chevallot-Beroux, E.; Lethuillier-Karl, L.; Li, G.; Moran, J. Metals Promote Sequences of the Reverse Krebs Cycle. *Nat Ecol Evol* **2017**, *1* (11), 1716–1721.

(66) Muchowska, K. B.; Varma, S. J.; Moran, J. Synthesis and Breakdown of Universal Metabolic Precursors Promoted by Iron. *Nature* **2019**, *569* (7754), 104–107.

(67) Barge, L. M.; Flores, E.; Baum, M. M.; VanderVelde, D. G.; Russell, M. J. Redox and PH Gradients Drive Amino Acid Synthesis in Iron Oxyhydroxide Mineral Systems. *Proceedings of the National Academy of Sciences* **2019**, *116* (11).

(68) Huber, C.; Wächtershäuser, G. Primordial Reductive Amination Revisited. *Tetrahedron Letters* **2003**, *44* (8), 1695–1697.

(69) Mayer, R. J.; Moran, J. Quantifying Reductive Amination in Nonenzymatic Amino Acid Synthesis. *Angewandte Chemie International Edition* **2022**, *61* (48), e202212237.

(70) Whicher, A.; Camprubi, E.; Pinna, S.; Herschy, B.; Lane, N. Acetyl Phosphate as a Primordial Energy Currency at the Origin of Life. *Origins of Life and Evolution of Biospheres* **2018**, *48* (2), 159–179.

(71) Keller, M. A.; Turchyn, A. V.; Ralser, M. Non-Enzymatic Glycolysis and Pentose Phosphate Pathway-like Reactions in a Plausible Archean Ocean. *Mol Syst Biol* **2014**, *10* (4), 725.

(72) Messner, C. B.; Driscoll, P. C.; Piedrafita, G.; De Volder, M. F. L.; Ralser, M. Nonenzymatic Gluconeogenesis-like Formation of Fructose 1,6-Bisphosphate in Ice. *Proceedings of the National Academy of Sciences* **2017**, *114* (28), 7403–7407.

(73) Piedrafita, G.; Varma, S. J.; Castro, C.; Messner, C. B.; Szyrwiel, L.; Griffin, J. L.; Ralser, M. Cysteine and Iron Accelerate the Formation of Ribose-5-Phosphate, Providing Insights into the Evolutionary Origins of the Metabolic Network Structure. *PLOS Biology* **2021**, *19* (12), e3001468.

(74) Camprubi, E.; Harrison, S. a.; Jordan, S. f.; Bonnel, J.; Pinna, S.; Lane, N. Do Soluble Phosphates Direct the Formose Reaction towards Pentose Sugars? *Astrobiology* **2022**, *22* (8), 981–991.

(75) Yi, J.; Kaur, H.; Kazöne, W.; Rauscher, S. A.; Gravillier, L.-A.; Muchowska, K. B.; Moran, J. A Nonenzymatic Analog of Pyrimidine Nucleobase Biosynthesis. *Angewandte Chemie International Edition* **2022**, *61* (23), e202117211.

(76) Harrison, S. A.; Palmeira, R. N.; Halpern, A.; Lane, N. A Biophysical Basis for the Emergence of the Genetic Code in Protocells. *Biochimica et biophysica acta. Bioenergetics* **2022**, *1863* (8), 148597.

(77) Hanwell, M. D.; Curtis, D. E.; Lonie, D. C.; Vandermeersch, T.; Zurek, E.; Hutchison, G. R. Avogadro: An Advanced Semantic Chemical Editor, Visualization, and Analysis Platform. *Journal of Cheminformatics* **2012**, *4* (1), 17.

(78) Jo, S.; Kim, T.; Iyer, V. G.; Im, W. CHARMM-GUI: A Web-Based Graphical User Interface for CHARMM. *Journal of Computational Chemistry* **2008**, *29* (11), 1859–1865.

(79) Kim, S.; Lee, J.; Jo, S.; Brooks, C. L.; Lee, H. S.; Im, W. CHARMM-GUI Ligand Reader and Modeler for CHARMM Force Field Generation of Small Molecules. *Journal of Computational Chemistry* **2017**, *38* (21), 1879–1886.

(80) Lee, J.; Cheng, X.; Swails, J. M.; Yeom, M. S.; Eastman, P. K.; Lemkul, J. A.; Wei, S.; Buckner, J.; Jeong, J. C.; Qi, Y.; Jo, S.; Pande, V. S.; Case, D. A.; Brooks, C. L.; MacKerell, A. D.; Klauda, J. B.; Im, W. CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations Using the CHARMM36 Additive Force Field. *Journal of Chemical Theory and Computation* **2016**, *12* (1), 405–413.

(81) Phillips, J. C.; Hardy, D. J.; Maia, J. D. C.; Stone, J. E.; Ribeiro, J. V.; Bernardi, R. C.; Buch, R.; Fiorin, G.; Hénin, J.; Jiang, W.; McGreevy, R.; Melo, M. C. R.; Radak, B. K.; Skeel, R. D.; Singharoy, A.; Wang, Y.; Roux, B.; Aksimentiev, A.; Luthey-Schulten, Z.; Kalé, L. V.; Schulten, K.; Chipot, C.; Tajkhorshid, E. Scalable Molecular Dynamics on CPU and GPU Architectures with NAMD. *The Journal of Chemical Physics* **2020**, *153* (4), 044130.

(82) Matsunaga, Y. *MDToolbox 1.0*. https://mdtoolbox.readthedocs.io/en/latest/introduction.html (accessed 2022-02-21).

(83) Tien, M. Z.; Meyer, A. G.; Sydykova, D. K.; Spielman, S. J.; Wilke, C. O. Maximum Allowed Solvent Accessibilites of Residues in Proteins. *PLoS ONE* **2013**, *8* (11).

(84) Williamson, M. P. Using Chemical Shift Perturbation to Characterise Ligand Binding. *Prog Nucl Magn Reson Spectrosc* **2013**, *73*, 1–16.

(85) Lacey, J. C.; Mullins, D. W.; Khaled, M. A. The Case for the Anticode. *Origins of Life* **1984**, *14* (1–4), 505–511.

(86) Krigbaum, W. R.; Komoriya, A. Local Interactions as a Structure Determinant for Protein Molecules: II. *Biochim Biophys Acta* **1979**, *576* (1), 204–248.

(87) Sweet, R. M.; Eisenberg, D. Correlation of Sequence Hydrophobicities Measures Similarity in Three-Dimensional Protein Structure. *J Mol Biol* **1983**, *171* (4), 479–488.

(88) Schoch, C. L.; Ciufo, S.; Domrachev, M.; Hotton, C. L.; Kannan, S.; Khovanskaya, R.; Leipe, D.; Mcveigh, R.; O'Neill, K.; Robbertse, B.; Sharma, S.; Soussov, V.; Sullivan, J. P.; Sun, L.; Turner, S.; Karsch-Mizrachi, I. NCBI Taxonomy: A Comprehensive Update on Curation, Resources and Tools. *Database (Oxford)* **2020**, *2020*, baaa062.

(89) Sayers, E. W.; Cavanaugh, M.; Clark, K.; Ostell, J.; Pruitt, K. D.; Karsch-Mizrachi, I. GenBank. *Nucleic Acids Res* **2019**, *47* (D1), D94–D99.

(90) The Merck Index Online. *The Merck Index Online - chemicals, drugs and biologicals*. https://www.rsc.org/merck-index (accessed 2023-02-23).

(91) Riniker, S. Fixed-Charge Atomistic Force Fields for Molecular Dynamics Simulations in the Condensed Phase: An Overview. *J. Chem. Inf. Model.* **2018**, *58* (3), 565–578.

(92) Wang, X.; Yan, J.; Zhang, H.; Xu, Z.; Zhang, J. Z. H. An Electrostatic Energy-Based Charge Model for Molecular Dynamics Simulation. *J. Chem. Phys.* **2021**, *154* (13), 134107.

(93) Condon, D. E.; Kennedy, S. D.; Mort, B. C.; Kierzek, R.; Yildirim, I.; Turner, D. H. Stacking in RNA: NMR of Four Tetramers Benchmark Molecular Dynamics. *Journal of Chemical Theory and Computation* **2015**, *11* (6), 2729–2742.

(94) Johnson, D. B. F.; Wang, L. Imprints of the Genetic Code in the Ribosome. *Proceedings of the National Academy of Sciences of the United States of America* **2010**, *107* (18), 8298–8303.

(95) Pinna, S.; Kunz, C.; Halpern, A.; Harrison, S. A.; Jordan, S. F.; Ward, J.; Werner, F.; Lane, N. A Prebiotic Basis for ATP as the Universal Energy Currency. *PLOS Biology* **2022**, *20* (10), e3001437.

(96) Sosunov, V.; Zorov, S.; Sosunova, E.; Nikolaev, A.; Zakeyeva, I.; Bass, I.; Goldfarb, A.; Nikiforov, V.; Severinov, K.; Mustaev, A. The Involvement of the Aspartate Triad of the Active Center in All Catalytic Activities of Multisubunit RNA Polymerase. *Nucleic Acids Res* **2005**, *33* (13), 4202–4211.

(97) Unciuleac, M.-C.; Goldgur, Y.; Shuman, S. Two-Metal versus One-Metal Mechanisms of Lysine Adenylylation by ATP-Dependent and NAD+-Dependent Polynucleotide Ligases. *Proc Natl Acad Sci U S A* **2017**, *114* (10), 2592–2597.

(98) Hopfield, J. J. Origin of the Genetic Code: A Testable Hypothesis Based on TRNA Structure, Sequence, and Kinetic Proofreading. *Proc. Natl. Acad. Sci. U.S.A.* **1978**, *75* (9), 4334–4338.

(99) Betat, H.; Mörl, M. The CCA-Adding Enzyme: A Central Scrutinizer in TRNA Quality Control. *Bioessays* **2015**, *37* (9), 975–982.

(100) Li, L.; Francklyn, C.; Carter, C. W.; Jr. Aminoacylating Urzymes Challenge the RNA World Hypothesis. *The Journal of biological chemistry* **2013**, *288* (37), 26856–26863.

(101) Carter, C. W.; Wills, P. R. Hierarchical Groove Discrimination by Class I and II Aminoacyl-TRNA Synthetases Reveals a Palimpsest of the Operational RNA Code in the TRNA Acceptor-Stem Bases. *Nucleic Acids Research* **2018**, *46* (18), 9667–9683.

(102) Davidovich, C.; Belousoff, M.; Bashan, A.; Yonath, A. The Evolving Ribosome: From Non-Coded Peptide Bond Formation to Sophisticated Translation Machinery. *Research in Microbiology* **2009**, *160* (7), 487–492.

(103) Farias, S. T.; Rêgo, T. G.; José, M. V. Origin and Evolution of the Peptidyl Transferase Center from Proto-TRNAs. *FEBS open bio* **2014**, *4*, 175–178.

Figures

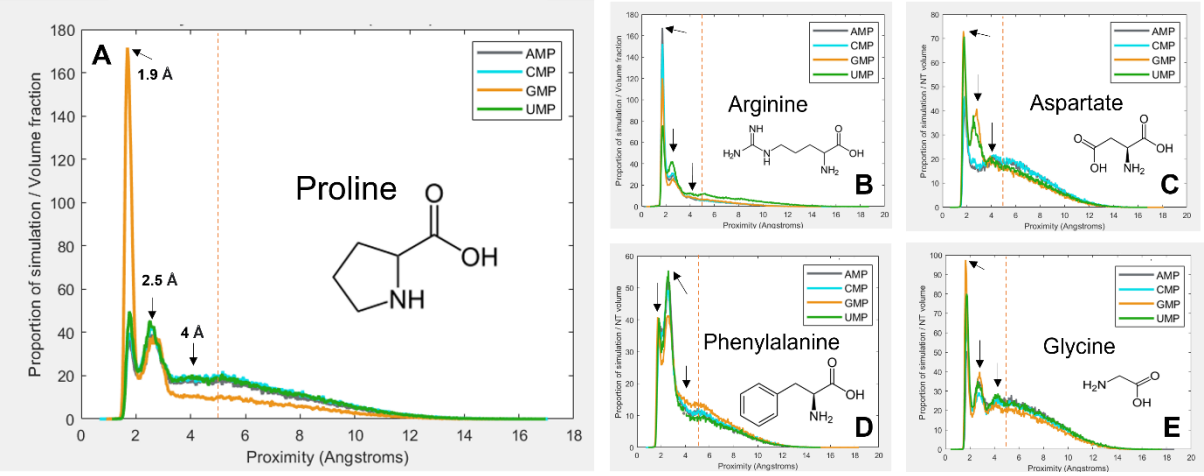

Figure 1. Volume adjusted proximity probability distributions for a selection different combinations of amino acids and nucleotides. Proximity is the closest atom of the nucleotide to the closest atom of the closest of 10 amino acids in the 40 Å periodic box. Proportion of simulation time is adjusted for the volume of the amino acid and nucleotide, so that different systems are comparable. Most interactions demonstrate multiple binding modes, at ~ 1.9 Å, 2.5 Å and 4 Å. An additional peak at around 5.5 Å is also visible, which is interpreted as the average distance of the closest amino acid when not bound. The vertical red line indicates the 5 Å threshold for binding. **(A)** proline; **(B)** arginine; **(C)** aspartate; **(D)** phenylalanine; **(E)** glycine.

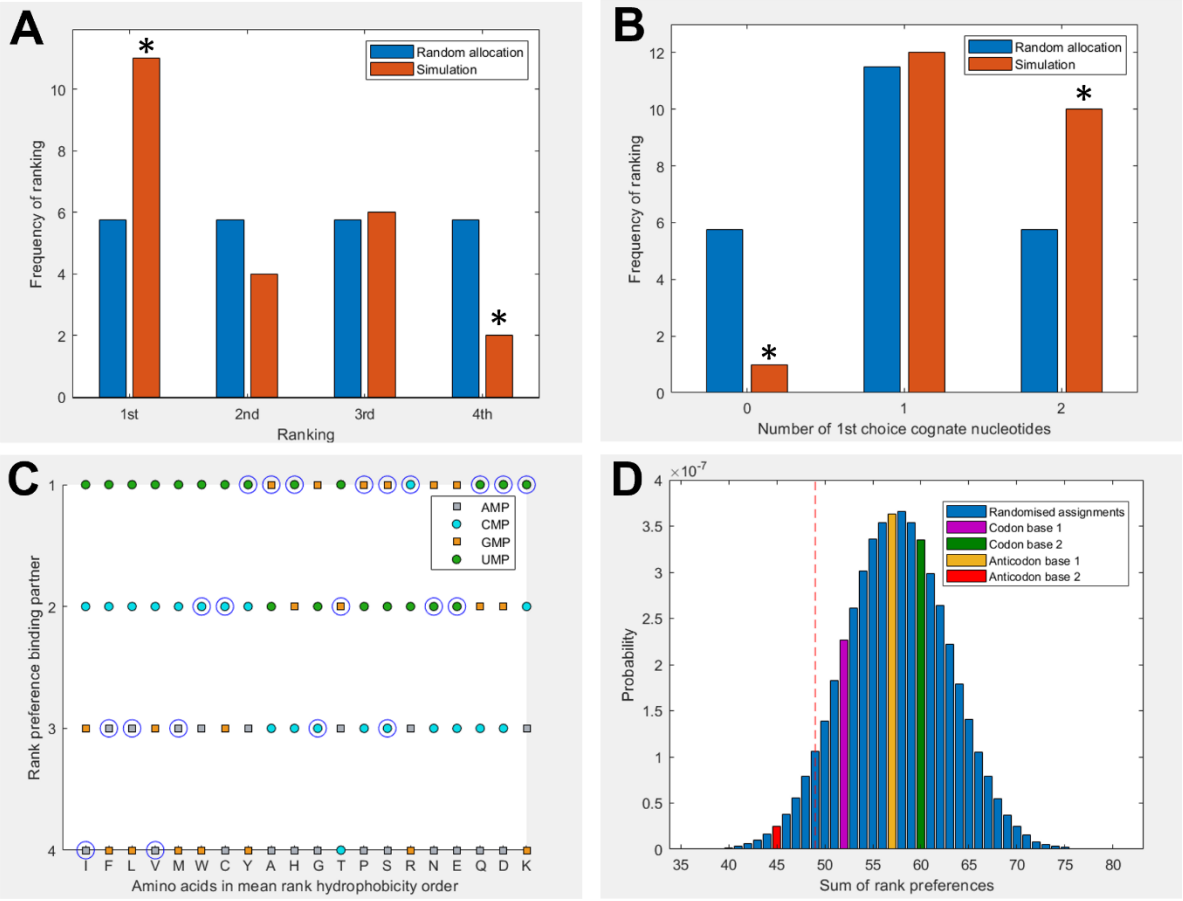

Figure 2. **(A)** Rank preference of the anticodonic middle base nucleotide for each the 20 amino acids as predicted by MD simulations. Three additional cognate assignments are included for amino acids with multiple 1st and 2nd base assignments. Random allocation shows equal chance of each amino acid interacting most strongly with any nucleotide. **(B)** Number of amino acids predicted to interact most strongly with both, one, or no nucleotides cognate at either the 1st or 2nd position of either the anticodon or codon, compared to random allocations where the probability of interacting with any given nucleotide is equal. **(C)** Rank preference of each amino acid for each nucleotide, ordered by mean hydrophobicity rank of amino acids [8,76]. The true anticodonic middle base for each amino acid is circled. Circular datapoints show pyrimidines; square datapoints are purines. Green is UMP, blue CMP, orange GMP and grey AMP. **(D)** Sum of rank preferences of the amino acids for their cognate nucleotides compared to randomized preferences. Highlighted bars are scores for: codon base 1 (purple), codon base 2 (green), anticodon base 1 (yellow), anticodon base 2 (red). All nucleotides are in the −1 charge state in all panels and interactions are given with respect to ring nitrogens.

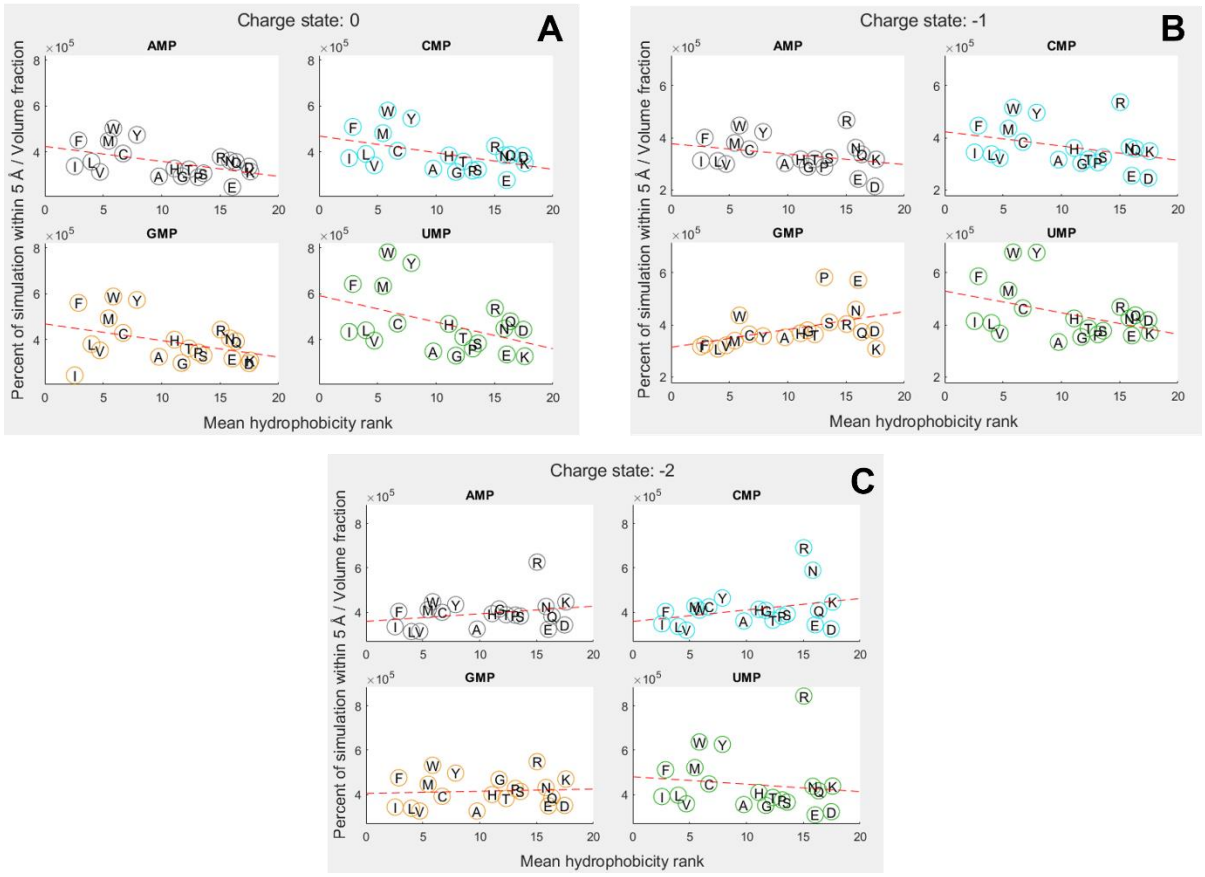

Figure 3. Proportion of simulation time for each amino acid within 5 Å of a nucleotide, adjusted for molecular volume. Amino acids are denoted by their single letter codes, and organised by mean hydrophobicity rank, calculated from Trinquier's 43 scales [8,76]. **(A)** phosphate charge state = 0; **(B)** phosphate charge state = −1. **(C)** phosphate charge state = −2. Dotted red line shows best fit from a linear regression. Distance is measured relative to nucleobase rings.

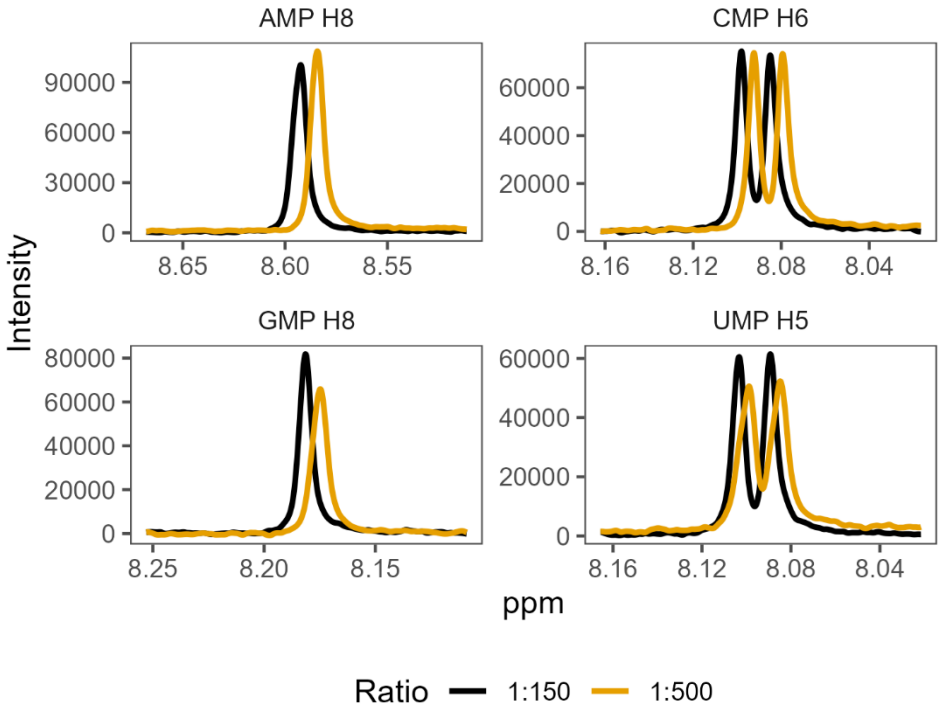

Figure 4. ^{1}H NMR spectra showing characteristic changes in proton chemical shift perturbations (or changes) as the ratio of amino acid to nucleotide increases. The panels show mixtures of phenylalanine with each of the four mononucleotides, at two representative ratios: 1:150 nucleotide to amino acid, and 1:500 nucleotide to amino acid. Numbers next to each NMP refer to the proton probes used for measurement of peak shifts (see SI Figure 1 for numbering).

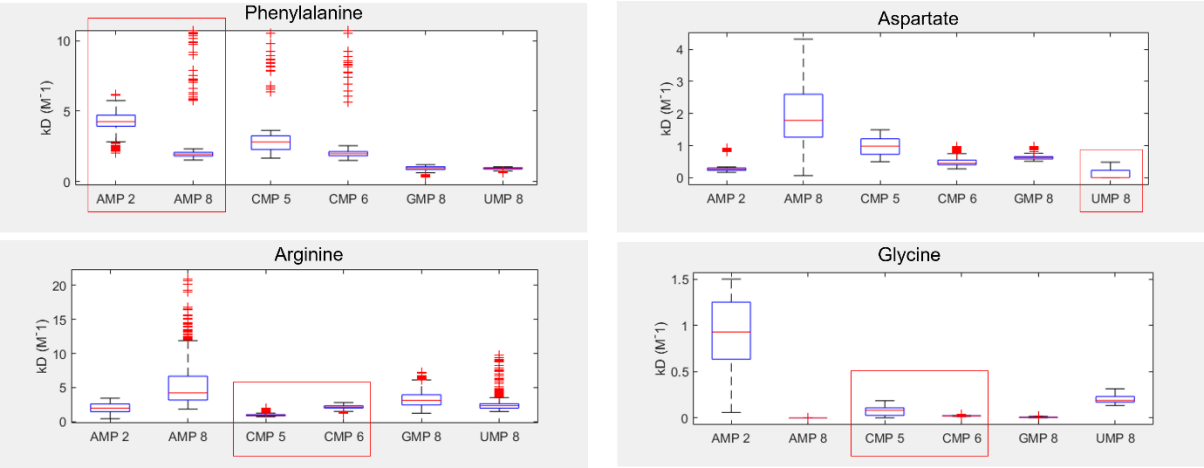

Figure 5. Inferred NMR binding constants (K_D) for phenylalanine, aspartate, arginine and glycine, with each of the four RNA bases. Lower K_D indicates stronger binding. Numbers next to each NMP refer to the proton probes used for measurement of peak shifts (see SI Figure 1 for numbering). Red boxes highlight the cognate anticodonic middle bases for each amino acid.

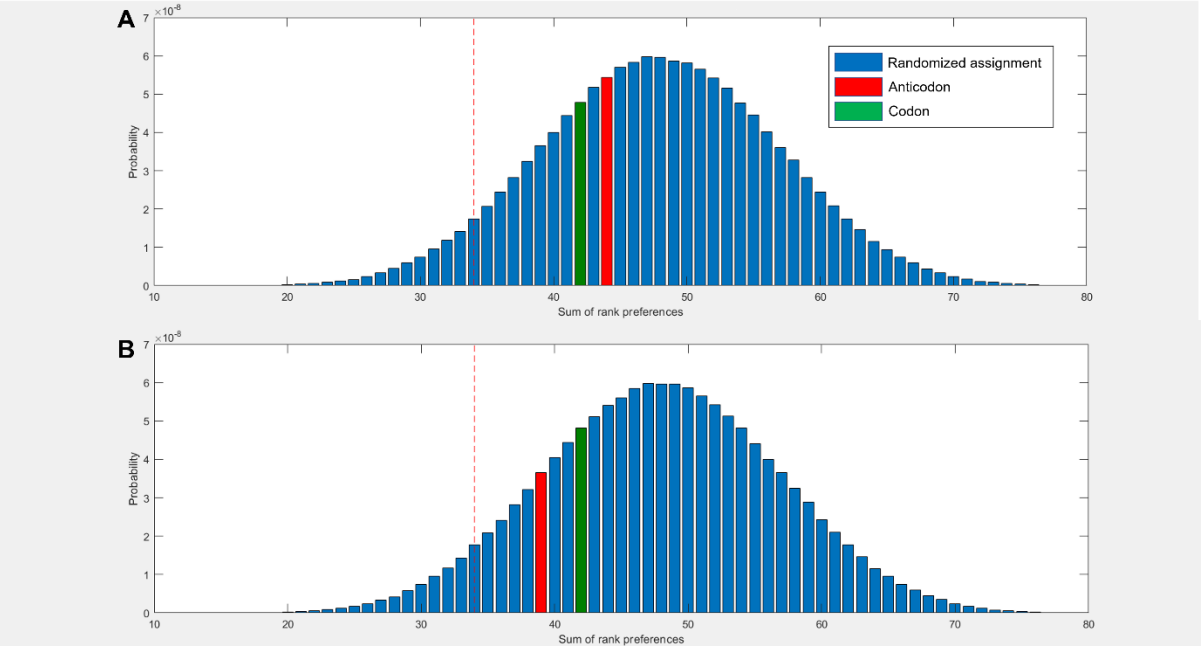

Figure 6. Sum of rank preferences for a selection of amino acids and dinucleotides with respect to proportion of simulation time spent within 5 Å, for **(A)** whole dinucleotide **(B)** ring nitrogens, adjusted for molecular volumes. If all amino acids had their cognate nucleotides as their best binding partner, this gives the lowest score. Red = bases 1 and 2, anticodonic dinucleotide. Green = bases 1 and 2, codonic dinucleotide. Left of red line is the top 5% of randomized assignments.

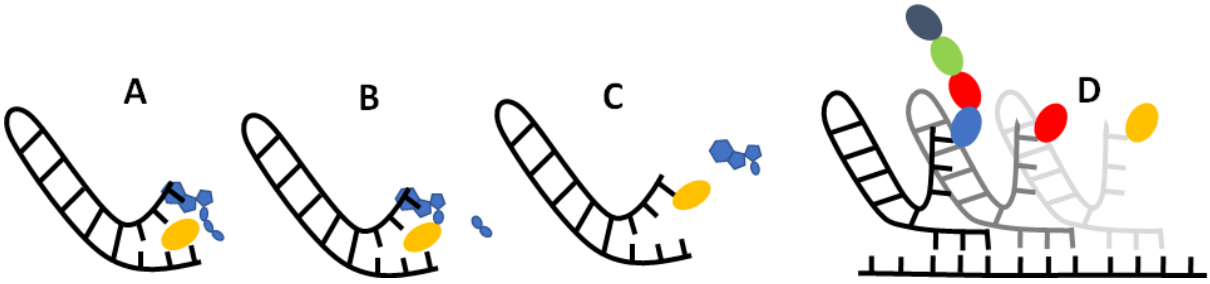

Figure *7.* Possible mechanism for amino acid binding and adenylation on a proto-tRNA. **(A)** An amino acid (yellow oblong) binds to its cognate anticodon on a short RNA with a hairpin loop. An ATP (blue) stacks onto the terminal A adjacent to the anticodon. **(B)** Nucleophilic attack of the carboxylate oxygen on the α-phosphate releases the pyrophosphate tail (blue oblongs), which adenylates the amino acid. **(C)** Transfer of the amino acid from the adenosine to the 2' ribose of the terminal A on the 'acceptor stem' results in an amino-acylated RNA. **(D)** Possible primordial mechanism of translation, where a flexible hinge of the proto-tRNA allows binding of the anticodon to a 'codon' on an adjacent proto-mRNA (where the reading frame is also determined by stereochemical interactions) enabling synthesis of short peptide sequences specified by the RNA sequence.