Title: The Origin of Genetic Heredity

Raquel Nunes Palmeira[1], Nick Lane[1], Andrew Pomiankowski[1]

[1]Department of Genetics, Evolution and Environment, University College London, UK

Raquel Nunes Palmeira 0000-0003-3873-1821

Andrew Pomiankowski 0000-0002-5171-8755

Nick Lane 0000-0002-5433-3973

Author for correspondence: nick.lane@ucl.ac.uk

Keywords: Origin of life, origin of heredity, protocells, RNA world, Frozen accident, Statistical proteins, Eigen's paradox, genetic code.

Abstract

The dominant view that genetic heredity arose in an RNA world stems from the ideas of Francis Crick and Leslie Orgel, who suggested that early nucleotide polymers could act as both phenotype and genotype and that the genetic code was shaped by a *frozen accident*. At around the same time, Carl Woese proposed a different model: that naturally occurring early translation was biased by weak stereochemical affinities between nucleotides and amino acids, producing what he called statistical proteins. This alternative was never dismissed on empirical grounds but fell out of favour as gene-level selection and the idea of an RNA world became entrenched. This review re-examines these competing models by tracing the origins of their two central divergences: whether selection first acted on naked genes or whole protocells, and whether early coding depended on stereochemical interactions. We summarise how these assumptions arose, how they have shaped theoretical models, and how they constrain explanations for the emergence of translation. Focusing first on the level of selection, we outline why models of naked-gene evolution face inherent obstacles related to information integration, whereas protocell-level models naturally favour the coexistence of sequences. We then turn to the question of stereochemical interactions in early translation. We consider the predictions of each hypothesis for the structure of the genetic code and compare these to the patterns observed today. Finally, we review the experimental evidence for amino acid-nucleotide interactions and outline the key empirical gaps. Together, these lines of evidence clarify which assumptions in classical models remain viable and point to the most informative experiments that can be done to elucidate the origin of heredity.

I. Introduction

The origin of genetic heredity is arguably the single most puzzling step in the origin of life. The genetic code and the translation machinery are extremely well conserved across life and must have been present in the last universal common ancestor (LUCA) (Koonin, 2003, Ouzounis et al., 2006, Crapitto et al., 2022). But it is hard to conceive how and in what order the many interacting parts of the genetic apparatus arose (i.e. mRNA, tRNA, the ribosome, aminoacyl tRNA synthetases, etc.). The paradox at the heart of the problem is that a system as complex as the genetic apparatus must be a result of evolution, but evolution itself requires a functional heredity system.

The conventional resolution to this paradox is the idea of an RNA world, an era of prebiotic evolution where RNA was both genotype and phenotype. RNA has an almost unique ability to act both as an information carrier and a catalyst (Orgel, 1968, Gilbert, 1986, Rich, 1962). This makes possible a gradual transition to the extant genetic system without the need for translation in early evolution. As peptides are better catalysts than RNA, any RNA strands that catalyse the synthesis of peptides should be favoured. Using this logic, the translational machinery and the genetic code could, in principle, evolve from an RNA world.

But the RNA world has its own problems. The RNA-world modelling tradition has largely been devoted to the problem of maintaining information (Eigen and Schuster, 1977, Szathmáry and Demeter, 1987, Zachar et al., 2011). Early work showed that information sequences face competitive exclusion by selfish sequences (those good at copying themselves) causing parasitic collapse, as shown in Spiegelman's classic experiment (Spiegelman, 1970). Some of these problems are solved by compartmentalisation of RNA sequences within protocells.

Group-selection at the level of the cell favours the retention of information (Szathmáry and Demeter, 1987). But the current theoretical framework fails to explain how RNA sequences could foster the evolution of translation while maintaining their own replication and increasing protocell fitness.

Another possible, albeit less popular, answer lies in the idea of rudimentary peptide synthesis through naturally occurring nucleotide-peptide interactions (Woese, 1969, Woese, 1965, Lacey et al., 1984). This could have arisen because some nucleotides (or short polynucleotides) had weak preferences for certain types of amino acids. Even if most processes in the protocell were inefficient, any nucleotides capable of templating a mixture of similar peptides could be selectively favoured in protocells. This loose but biased proto-translation could bootstrap itself up through selection to more specific and robust translation (Woese, 1965). However, until recently, the evidence for interactions between nucleotides and amino acids was sparse and based on unreliable methods (Lacey et al., 1984, Shimizu, 1987, Hobish et al., 1995). The paucity of robust evidence resulted in the idea being overlooked, and lacking rigorous theoretical consideration.

This review examines these two major hypotheses for the origin of heredity. Starting with the foundational models that gave rise to the hypotheses in the 1960s, we discuss the assumptions underlying them, as well as the historical context that consolidated ideas in the field. The models differ both in their proposed mechanisms for the evolution of the genetic code and in their assumptions about the level at which selection operates at the origin of life. Because these assumptions are not intrinsically linked, we discuss them separately. We summarise the most influential models for selection at the level of the naked gene and highlight conceptual issues that arise from this theoretical standpoint. We then summarise the evidence for stereochemical interactions that might have guided early coding, and discuss patterns present in the universal

genetic code and how they fit with the competing models. Finally, we point to gaps in the experimental and modelling data and suggest a line of enquiry that will help clarify the emergence of genetic heredity.

II. Seminal models of the origin of the genetic code

Much thinking around the origin of heredity can be traced back to debates in the literature in the 1960s. These papers were written at the dawn of molecular biology and so were among the first to approach the origin of heredity with the advantage of knowing how the system worked in modern cells. The main question at hand was the evolution of the recently discovered genetic code. Two prominent answers to this question were proposed by Leslie Orgel and Francis Crick on one hand (Crick, 1968, Orgel, 1968), and Carl Woese on the other (Woese, 1965). These hypotheses were similar in proposing rudimentary genetic codes that expanded gradually while increasing the fitness of the protocell. Both attempted to explain why the genetic code is non-random, and how heredity could evolve in the absence of a heredity system. But they differed greatly in their underlying assumptions, with very different implications for the origin of life.

(1) Crick and Orgel's evolution of adaptors and the frozen accident theory

Orgel and Crick proposed that before the advent of the genetic apparatus, nucleotide polymers were the main players of early evolution (Orgel, 1968; Crick, 1968). According to Orgel there are only two obvious alternatives for a prebiotic heredity system: "(1) life based on proteins in the absence of nucleic acids, (2) life based on nucleic acids in the absence of proteins" (Orgel, 1968). Since there was evidence that nucleotide polymers formed secondary structures but no evidence that proteins could replicate, he concluded that the first cells were most likely preceded by an era in which heredity was carried solely by nucleotide polymers. Crick reached

the same conclusion in his companion paper, referring to tRNAs as “Nature’s attempt to make RNA do the job of a protein” (Crick, 1968).

This idea resolves the paradox of the origin of a heredity system because it assumes that (1) there must have been some sort of evolution before the advent of the genetic code, and (2) any early heredity system could be composed solely of genes or proteins, rather than a mix of the two. That relinquishes the need for a translational apparatus and genetic code from the start, opening the possibility for individual nucleotide strands to evolve into proto-tRNA structures and gradually evolve translation.

Importantly, Crick and Orgel both rejected the idea that interactions between amino acids and nucleotides might have guided early translation (Crick, 1968, Orgel, 1968). Their rejection most likely stemmed from the recent vindication of Crick’s adaptor hypothesis, the idea that translation was accomplished through an intermediate molecule that independently recognised the anticodon and its cognate amino acid. This contrasted with earlier views in the 1950s, which proposed that a polynucleotide would serve as a direct template onto which amino acids could bind next to each other to form a peptide (Gamow, 1954). Crick did not accept that nucleotides could chemically “recognise” amino acids, and suggested that translation must instead be done using “adaptor molecules” as intermediates (Crick, 1958, Woese, 1970, Woese, 1967, Crick, 1955, Woese et al., 1966). These adaptors would be nucleotide strands forming structures that preferentially attached to an amino acid, while also containing a specific nucleotide sequence (later called the anticodon) that would base-pair to a template (later termed messenger RNA) (Crick, 1958, Crick, 1955). When the discovery of tRNAs (Hoagland et al., 1958) proved Crick’s adaptor hypothesis correct, the underlying assumption of no chemical recognition between the anticodons and amino acids was also indirectly supported. Later, the structure of

tRNA lent strength to this view, as the anticodon sits at the opposite end to the amino acid, and so was taken not to interact with it, (Kim et al., 1974, Robertus et al., 1974).

Building on this view of translation as adaptor mediated, Crick suggested a mechanism for how the assignments of proto-tRNAs and amino acids could evolve into the structured genetic code we have today. He proposed that the early genetic code was highly ambiguous, with all codons initially assigned to a few amino acids, and other amino acids were gradually added to the code later (Crick, 1968). For Crick, the process of early codon assignment occurred randomly and was modified through random mutations to the "activating enzyme" that assigned amino acids to certain proto-tRNAs. Crick also suggested, in accordance with Orgel's idea of "life made of nucleic acids in the absence of proteins" that "the primitive tRNA itself could be its own activating enzyme". Given this, the proposed mechanism proceeded as follows. Imagine a population of protocells with an early genetic code in which all codons (AAA through to UUU) code for only two amino acids, with half coding for AA1 (amino acid 1) and the other half coding for AA2 (Figure 1). Consider a random mutation occurs in one of the proto-tRNAs so that it now picks up AA3, which is chemically similar to AA1 but very different to AA2. If the mutated proto-tRNA had previously picked up AA1, this mutant could spread and be established as it adds to the repertoire of amino acids in peptides with limited disruption to the existing peptides of that protocell. But if the mutation occurred in a proto-tRNA that currently codes for AA2 the disruption is likely to great and very likely would reduce protocell fitness (Figure 1). The incorporation of new amino acids to the code therefore gave protocells a selective advantage in the longer term at little short-term cost. But as codon assignments became more specific, further codon reassignments to novel amino acids were less likely to be beneficial. The gain in catalytic repertoire through adding additional amino acids would diminish with the number already encoded, which could already produce chemically diverse peptides. Additional changes in assignment necessarily cascade through multiple "genes" and

disrupt fitness, to the point that any further changes to the code would be lethal. This is reflected in the almost complete stasis of evolutionary change in the genetic code across extant life. For this reason, Crick called this way of conceiving the origin of the genetic code a "frozen accident theory".

(2) Woese's statistical proteins

In contrast, Carl Woese's hypothesis posited the existence of loose discriminative interactions between "codons" and amino acids (Woese, 1969, Woese et al., 1966, Woese, 1970). He proposed that a natural genetic code consisted of groups of contiguous codons (i.e. differing in one nucleotide) roughly assigned to groups of chemically similar amino acids through weak stereochemical associations (Woese, 1970). Further evolution of the code would have been driven by increasing specificity, rather than including new amino acids to the cell's vocabulary.

Woese argued that the resolution to the paradoxical origins of heredity lay in having naturally occurring preferences between codons and amino acids. For example, more hydrophobic nucleobases would tend to associate with more hydrophobic amino acids. In this model, nucleotide templates would not be translated into specific peptides. Rather, they give rise to a non-random (but weakly specified) distribution of peptides that he called "statistical proteins" (Woese, 1965, Woese, 1970). Two protocells containing different nucleotide templates would therefore contain, on average, slightly different distributions of peptides. Evolutionary change follows because some nucleotide templates would produce statistical proteins advantageous to the protocell. A more efficient translational system could then evolve. Suppose that a mutant template increased the specificity of an amino acid assignment, so the next group of statistical proteins to be translated was less variable (i.e. had a tighter distribution around the average sequence). This would improve the fitness of the protocell if it enabled the production of more advantageous statistical proteins and fewer disadvantageous ones. Any protocell with more

tightly specified statistical proteins would be favoured by selection (Figure 2). In other words, if some statistical proteins offer a slight advantage to a protocell, new mutants that improve translation would be favoured (Woese, 1965, Woese, 1970). Improved translation also feeds back on itself, as mutations that enhance translation make their own statistical proteins more specifically.

III. Selection at the origin of life

Returning to Orgel, a somewhat secondary aspect of his hypothesis (at the time) turned out to be a watershed moment in the field. He posited that early evolution was based solely on how well single molecules replicated (Orgel, 1968). He distinguished between two phases: an earlier phase of "natural selection without function", in which evolution occurred at the level of individual replicating molecules, followed by "natural selection with function", when selection acts on the fitness effects of genes at the level of protocells. This was the first time anyone had proposed a period when selection operated at the level of the "naked gene" alone. Before Orgel's "*natural selection without function*" paper, both Crick and Woese's models for the origin of the genetic code (and other hypotheses proposed at the same time) had considered selection only at the level of a protocell (Goldberg and Wittes, 1966, Bryson and Vogel, 1965, Woese, 1969, Crick, 1968, Orgel, 1968). The idea of selection at the level of the naked gene was hugely influential, and became the most pervasive framework for the origin of heredity: the so-called RNA world (Joyce, 2009, Joyce and Orgel, 1993, Robertson and Joyce, 2012, Bernhardt, 2012, Salditt et al., 2023).

IV. Modelling the evolution of naked genes: the RNA world

(1) The RNA world

Although Orgel himself was initially agnostic about which genetic polymer (DNA or RNA) was the early protagonist of life (Orgel, 1968), his model of "*natural selection without function*" kickstarted what would later be called the RNA world hypothesis. This was bolstered by the discovery that nucleotide cofactors were almost ubiquitous in metabolism. Despite being monomers or dimers (rather than longer RNA sequences) these were suggested to be fossils of an era in life's history when nucleotides played the main role in catalysis (White, 1976). Orgel's explicit identification of selection at the level of the gene (Orgel, 1968) also chimed with the rebuttal of group selection (Hamilton, 1964a, Hamilton, 1964b, Maynard Smith, 1964) and the rise of the gene-centred view of evolution (Williams, 1966, Ågren, 2016), most popularly expounded by Dawkins' *The Selfish Gene* (1976). These ideas fundamentally changed the way the field saw evolution, and inevitably had an impact on the theories of the origin of life, providing fertile ground for the idea that the origin of life involved selection at the level of genes. In the 1980s, the discovery of the first ribozymes (RNase P, which cleaves precursor sequences to give a mature tRNA) (Cech et al., 1981, Kruger et al., 1982, Guerrier-Takada et al., 1983) cemented the hypothesis, and in 1986 Gilbert coined the name *RNA world* (Gilbert, 1986). He described the now consolidated idea of this world as "*containing only RNA molecules that serve to catalyse the synthesis of themselves*".

(2) The issue of information integration

Much of the modelling work around an RNA world has been devoted to solving the problem of how genetic information could be maintained before the evolution of the genetic apparatus. This problem, which we will refer to as "the issue of information integration" can be divided

into three distinct but interconnected problems. They relate to high copying error rates, selfish RNA strands and the competitive exclusion of sequences in the absence of compartments.

(a) The quasi-species model and Eigen's paradox

Manfred Eigen's modelling work is among the most important contributions to the framework for the evolution of "naked genes" in an RNA world. Eigen's models were initially concerned with the conservation of information given the high error rates expected of early copying before the advent of proof-reading enzymes. He proposed that, given a high copying error, information could not exist as a single sequence, but rather had to be a distribution of similar sequences centred around the fittest "master" sequence. This cluster of related sequences he termed a *quasi-species* (Eigen, 1971).

In the *quasi-species* model a particular sequence might be created by its own correct copying, but it might also be created by the incorrect copying of a similar sequence. The frequency of an RNA sequence is not only dictated by its own fitness (its ability to self-replicate) but also by the probability that it can be generated *de-novo* through incorrect copying of related sequences in the *quasi-species* (Eigen, 1971). Modelling the evolution of these distributions reveals the presence of an error-threshold: if the error rate is too great, even a sequence with high fitness cannot be maintained in the population because the probability of it being produced by copying (of itself or similar sequences) is too low, leading to the collapse of the *quasi-species* distribution (Eigen, 1971, Swetina and Schuster, 1982). The same phenomenon limits sequence length. As the number of possible configurations increases with sequence length, it becomes more difficult for selection to preserve longer sequences, even if they are very fit (Eigen, 1971, Swetina and Schuster, 1982).

This theoretical demonstration of an error threshold provides the first significant obstacle to the evolution of a genetic system from "naked genes". Given the expected error rates for non-

246 enzymatic copying, the upper limit of sequence that could be reliably maintained at high
247 frequency was estimated to be ~100 nucleotides (Eigen, 1971, Eigen and Schuster, 1977). To
248 maintain the information of longer sequences, an error-correcting enzyme must evolve. But
249 error-correcting enzymes are encoded by significantly more than 100 nucleotides (Maynard
250 Smith et al., 1983). This catch-22 scenario was later named Eigen's paradox (Szathmáry, 1989)
251 and many modelling efforts in the field to this day are dedicated to finding ways around it.

252 *(b) Spiegelman's monsters*

253 Another major obstacle to the evolution of a heredity system was uncovered in the classic
254 experiment by Sol Spiegelman (Spiegelman, 1970). In the experiment, a long strand of RNA
255 isolated from the Q_β virus, accompanied by its replicase enzyme and the necessary nucleotides,
256 was allowed to replicate in a test tube. After 74 generations the viral RNA, initially 4500
257 nucleotides long, had evolved into a 218-nucleotide strand containing only the recognition site
258 for the replicase. This was the first experimental demonstration of the evolution of a naked
259 RNA molecule (Spiegelman, 1970). When selection acts at the level of single RNA molecules,
260 it promotes the evolution of short, fast replicating strands, Spiegelman's "little monsters". Like
261 the error-threshold, selection on RNA strands tends to hinder the evolution of longer sequences
262 that might code for error-proofing polymerases needed for a heredity system.

263 *(c) Competitive exclusion of sequences*

264 One might think that the two problems demonstrated above are easily solved if the information
265 needed for an error-proofing polymerase could, instead, be contained in several short RNA
266 strands coding for different parts of the enzyme. These parts could come together to form the
267 whole enzyme after translation. This would indeed be the case if these small sequences existed
268 within a protocell, as their overall function would be selected for together (Szathmáry, 1989).
269 But because the scenario at hand here is the evolution of naked genes, we're faced with the

third obstacle to information integration. Different sequences cannot coexist in a system with limited resources. Instead, they compete until only one sequence is left. This is demonstrated indirectly in both Eigen's *quasi-species* model where variation is limited to a cloud of similar sequencers (Eigen, 1971) and Spiegelman's experiment where variation is pruned down to a sequence or sequences with the fastest replication (Spiegelman, 1970).

(3) The hypercycle

Eigen himself proposed an ingenious solution to the problem of information integration, the hypercycle (Eigen and Schuster, 1979). This is an auto-catalytic cycle where more than one "master" sequence can coexist, increasing the total amount of information that can be held in the system. In a hypercycle each sequence codes for a replicase, which specifically catalyses the copying of another sequence in the cycle. Consider a two-member hypercycle (its simplest configuration). A sequence I_1 codes for a replicase E_1 that, in turn, catalyses the copying of sequence I_2; while I_2 encodes a replicase E_2 which catalyses the copying of sequence I_1, closing the hypercycle (Figure 3). Because the two sequences enhance each other's replication, they can exist simultaneously as they have higher fitness (replication rates) than versions of I_1 and I_2 that can only self-replicate. The information integration of I_1 and I_2 in the hypercycle also escapes the problem of competitive exclusion, which is inevitable between unlinked genes that differ in their self-limited replication rates.

Although an elegant solution to the problem of information integration, the hypercycle model immediately encounters several serious conceptual and evolutionary difficulties. The first obstacle is the "growth" of an established hypercycle. Hypercyclic growth requires one of its sequences to duplicate and then diverge from its precursor without disrupting the existing cycle. This is particularly difficult because this duplicate sequence needs to have a similar replication rate to the original while it diverges, otherwise it would either replace it or become

294 replaced itself. For this reason, Eigen and Schuster argued that a hypercycle cannot grow
295 incrementally, but must "nucleate", meaning "the abundant presence of all members of the
296 hyper cycle seems to be a prerequisite for its origin". A related problem is that hypercycles are
297 vulnerable to the evolution of parasites. These sequences benefit from the hypercycle but only
298 to maximise their own replication. That eventually collapses the productivity of the system.
299 The more complex the hypercycle, the more elements are subject to parasitic invasion, which
300 also limits the existence of larger hypercycles.

301 Other issues arise from the fact that selection in hypercycles acts only at the level of individual
302 replicating sequence. One problem is that the evolution of "altruistic" functions, which benefit
303 the hypercycle as a whole, is severely constrained. Consider the case of a "selfish" mutation
304 I_1' of strand I_1 in the hypercycle shown in Figure 3, which makes it a better target to replicase
305 E_2. As I_1' will be replicated more often than I_1, these mutants will increase in frequency in the
306 population until they become the new "master sequence" (Figure 4a). But *a priori* this
307 evolutionary change has done nothing to improve the hypercycle, it simply switched one
308 version of I_1 for another. Also note this is not a selfish mutant in the sense of raising its own
309 fitness at the cost of others (*sensu* Dawkins, 1976) but is neutral as regards the hypercycle's
310 fitness. Now, consider that I_1' is an "altruistic" mutant that codes for more E_1 (or an improved
311 replicase E_1') that is better at replicating I_2. This leads to more I_2, which in turn codes for more
312 E_2. That might seem to improve the hypercycle. However, more E_2 leads to more I_1' but also
313 more I_1 and all its other mutants. Therefore, I_1' is neutral relative to I_1 so is not selectively
314 favoured (Figure 4b). Eigen and Schuster called these "phenotypic" and "genotypic" mutations
315 and wrote "*'Phenotypic' advantages, i.e., those variations which are of direct advantage to the*
316 *mutant, are immediately stabilized. On the other hand, 'genotypic' advantages, which favour*
317 *[…] only indirectly the replicative unit in which the mutation occurred, require special*
318 *separation for competitive fixation"* (Eigen and Schuster, 1979). As a consequence, even

though hypercycles do allow many sequences to coexist, they do not create a new level of selection. Hypercycles are not individuals, they are groups of individuals interacting ecologically (Maynard Smith, 1988). Selection is still at the level of the naked gene, right where Orgel put it back in the 1960s.

A related important issue is that the original formulation of the hypercycle model made the significant assumption of RNA sequences being translated adequately into replicases (Eigen and Schuster, 1979). This was especially problematic as the model set out to solve problems arising from high copying-error rates in the first place. Because translation is significantly more error prone than copying (which relies on intrinsically accurate base-pairing) this assumption was ungrounded. After the discovery of ribozymes, Eigen's successors proposed hypercycles of ribozymes (Sardanyés et al., 2017, Boerlijst and Hogeweg, 1991, Sardanyés and Solé, 2006). That solves the problem only by pushing the origin of translation further down the road. As noted above this is an "altruistic" outcome that cannot evolve from an RNA-only hypercycle. Translation requires the incorporation of new elements to an established hypercycle, invoking the "duplication and divergence" logic with the additional requirement that any translational function must feedback catalytically into the hypercycle. Worse, a hypercycle that maintains sequences encoding translational machinery is necessarily long, and so is especially vulnerable to parasitic invasion.

While the question of information integration is important, the focus on it distracted the field from the bigger question of how complexity evolves. The hypercycle model seems to create more obstacles than solutions to this bigger question. It is far from clear how a hypercycle in which individual genes connected by many specific replicases could evolve into a system with the generalist polymerases and translation seen in extant life. There is no route to the generation

342 of adaptor molecules or ribosomes that underpin translation, let alone forms of metabolism
343 which support the generation of the fundamental molecules needed to build cellular life.

344 **(4) The stochastic corrector**

345 Since many of the problems with hypercycles come from selection being at the level of the
346 gene, it might come as no surprise that answers to the problem of information integration arise
347 when genes exist within a compartment or protocell, abandoning the view of selection
348 exclusively at the level of naked gene initiated by (Orgel, 1968). Protocells have been modelled
349 as information integrators in different ways (Michod, 1983), including compartmentalising
350 hypercycles (Boerlijst and Hogeweg, 1991). But the simplest approach is to investigate whether
351 protocells alone can facilitate the coexistence of multiple RNA sequences.

352 The most notable protocell model is the Stochastic Corrector (Szathmáry and Demeter, 1987).
353 This consists of protocells each containing small number of different sequences, "replicators",
354 that copy themselves at different rates, leading to selection at the level of replicators. Selection
355 also happens at the level of the protocell. The sequences randomly segregate into daughter cells
356 on protocell division. Fitness is greater if a protocell maintains a specified ratio between the
357 types of replicators, and falls to zero if one of the replicator populations is lost. Simulations
358 show that the stochastic corrector is a reliable information integrator (Zintzaras et al., 2002). It
359 allows multiple distinct sequences to co-exist in a protocell even if they replicate at different
360 rates, increasing the information content that can be maintained.

361 Postulating selection on protocells solves the problem of maintaining diverse sequences that
362 collectively contribute to function. This occurs in the primitive condition in which sequences
363 have different replication rates (Szathmáry and Demeter, 1987). It also solves the problem of
364 parasitic takeover by selfish fast-replicating sequences (Spiegelman's monsters, Spiegelman
365 (1970)). They can spread within a protocell, but when they take over they substantially reduce

366 protocell fitness and so fall extinct. They also can be lost through stochastic segregation at
367 protocell division allowing selection at the level of the protocell to weed them out (Szathmáry
368 and Demeter, 1987). Another neglected advantage of compartmentalisation is that it removes
369 the constraints on function introduced by hypercyclic organisation. New functions can evolve
370 simply in new sequences, without being bound to exist on the complementary strand to a
371 replicase. Most importantly, altruistic functions are favoured. There are no longer any
372 theoretical obstacles to the evolution of a general polymerase, a translational system or genes
373 for metabolism.

374 ## V. Interactions between nucleotides and amino acids.

375 Whilst recognition of selection at the level of the naked gene was a key development, it should
376 not have diverted attention away from selection at the level of the protocell. The issues arising
377 from an exclusive focus on gene-level selection distracted attention from an important point of
378 divergence between the initial models: the degree to which interactions between nucleotides
379 and amino acids is key to the emergence of coding.

380 This question arose in the first place because the genetic code is remarkably non-random,
381 showing patterns in assignment of similar amino acids to related codons. As noted earlier, Crick
382 proposed that modifications to the structure of the adaptors would assign new amino acids in a
383 way that minimised disruption to existing proteins (Figure 1). As a result, newly added amino
384 acids would tend to occupy contiguous codons to chemically similar amino acids, giving rise
385 to the clustering pattern. According to Crick's model, the relationship reflects shared ancestry
386 rather than any persistent stereochemical interactions. Conversely, in Woese's model of
387 statistical proteins, amino acids are assigned to adaptors from the beginning. These assignments
388 are overlapping, and emerge from broad biases in stereochemical preferences between codons
389 and amino acids. These biases shaped the chemistry of the first statistical proteins. Assuming

that early proteins conferred some advantage to protocells, there is selective pressure for more specific assignments (Figure 2). So, this narrowing process occurs on top of the pre-existing biased landscape rather than erasing it. As specificity improves, the original interactions become sharper, so the model predicts that correlations in the chemistry of amino acids and codons should persist in the code today.

(1) The code within the codons

Given these contrasting predictions, how do they relate to the patterns associated with each position of the codon, the so-called "code within the codons" (Lacey et al., 1984, Taylor and Coates, 1989, Wong, 2005, Harrison et al., 2022, Copley et al., 2005)? The first position of the codon has long been known to group cognate amino acids by biosynthetic pathway (Wong, 2005, Tze-Fei Wong, 1981, Wong, 1975, Di Giulio and Medugno, 2000). This observation gave rise to the coevolution theory of the genetic code, which proposes that the reactions converting one amino acid into another occurred while the precursor amino acid was attached to the adaptor molecule. As a result the adaptor molecule becomes assigned to the product amino acid. The evolutionary forces acting on this assignment are not explicitly addressed in the work of the proponents of the coevolution theory. As in Crick's model, the code starts out as a reduced version of itself and expands as new amino acids are added. But as in Woese's model the resulting clustering of similar amino acids to similar codons is not attributed to shared ancestry but to chemical determinism.

If metabolism is taken to have started from CO_2 fixation (i.e. an autotrophic rather than heterotrophic origin of life (Lane et al., 2010, Moody et al., 2024, Xavier et al., 2021, Ouzounis et al., 2006, Martin and Russell, 2003, Martin and Russell, 2007, Morowitz et al., 1988, Huber and Wächtershäuser, 1997), a recent analysis shows that amino acids are not necessarily grouped by being contiguous in a biosynthetic pathway, but instead by how distant they are

414 from CO_2 fixation (Harrison et al., 2022) (Figure 5). It is notable that amino acids encoded by
415 a purine in the first position of the codon are predominantly synthesised "early" in metabolism,
416 while amino acids encoded by a pyrimidine in the first position of the codon tend to be furthest
417 away from CO_2 fixation (Figure 5).

418 This is an intriguing but as yet unexplained pattern in the genetic code. It suggests that in the
419 absence of enzymes, proto-metabolism imposed a temporal ordering in the production of amino
420 acids. Those amino acids produced "early" would have been the first to be recruited to the code
421 followed by others that emerged as biosynthetic pathways expanded. This view of the genetic
422 code starting as a reduced version of itself conforms to Crick's idea of the sequential addition
423 of amino acids to the code, in this case linked to the ease of biosynthesis or temporal appearance
424 of amino acids. But why this would entail linkage to purines in the first position of the codon,
425 followed by pyrimidines remains to be explained (Harrison et al., 2022). Neither the rationale
426 of a frozen accident nor stereochemical interactions obviously helps to understand this pattern.

427 **(2) The code within the anticodons**

428 A second, widely recognised, pattern in the genetic code relates to its second position. Amino
429 acids that share the same base in the second position of the anticodon have long been known
430 to have similar hydrophobicity (Lacey et al., 1984). But new observations show that if amino
431 acids are grouped by their first codon (as above) the pattern becomes far more revealing. For
432 "early" amino acids, those with purines at the first position of the codon, the most hydrophobic
433 amino acids associate with the most hydrophobic bases at the second position of the anticodon,
434 and *vice versa* (Harrison et al., 2022). For instance, take only the amino acids that have purines
435 in the first position of their codons (Figure 6). The most hydrophobic of these amino acids
436 (isoleucine, I) is encoded by the most hydrophobic nucleotide (A) in the second position of its
437 anticodon. Conversely the least hydrophobic amino acid, aspartate (D) is encoded by least

hydrophobic nucleotide (U). For whatever reason, during the expansion of biosynthetic pathways, the earliest amino acids (coded by a G or A at the first position of the codon) display the strongest physical associations with their cognate middle anticodons.

These findings fit with what Woese called the "*optimization of their energy of interaction*" (Woese et al., 1966). Exactly what relevant properties (or correlates) of "hydrophobicity" are crucial is yet to be identified and several suggestions remain to be fully investigated (Trinquier and Sanejouand, 1998, Caldararo and Di Giulio, 2022).

The third position of the codon is often considered to be redundant, with half of all codons being completely degenerate (Harrison et al., 2022). Yet whether the third position exhibits redundancy is itself non-random, and relates to the composition of the other positions of the codon (Lagerkvist, 1978, Harrison et al., 2022). Most relevant to the argument here, when the third position of the codon is not redundant, it serves as an indicator of the size or "complexity" of the cognate amino acid. Pyrimidines in the third position of the anticodon code for larger amino acids, while purines code for smaller amino acids (Lagerkvist, 1978, Harrison et al., 2022). As double-ringed purine bases are larger than single-ringed pyrimidines, a pyrimidine third base anticodon (C, U), could provide more space for larger amino acids to bind. This pattern might indicate a binding pocket for amino acids, which discriminates for both size and hydrophobicity, though this hypothesis lacks empirical support. But regardless of its existence, there is a critical point here: these patterns do not simply group similar amino acids with similar codons, as predicted by Crick. There is a correlation between the physical properties of amino acids and those of their cognate codons. Crick's hypothesis specifically predicts no physical interactions between the anticodon and the amino acid, hence specifically rejects such a correlation. Its existence therefore suggests an era where interactions between nucleotides and

amino acids were important in codon assignment, as predicted by Woese in the 1960s (Woese, 1969, Woese, 1967).

(3) Experimental evidence for stereochemical interactions

These patterns in the code suggest ancient stereochemical relationships, which ought to be experimentally detectable. Most work aiming to identify amino acid-nucleotide interactions was done in the 1980s and 1990s. The results included preferential interactions between anticodon dinucleoside monophosphates and their cognate amino acids, detected by UV-vis (Shimizu, 1987), specific aminoacylation of RNA hairpin loops containing the anticodon with the correct cognate amino acid in the absence of enzymes (Shimizu, 1995) and association of amino acids to their anticodons shown by NMR (Hobish et al., 1995). On the face of it, this work was largely successful in showing preferential interactions with a range of experimental methods. Yet, despite the overwhelmingly positive results, this literature is often overlooked, rarely cited, and the field largely abandoned this line of experimental work, possibly reflecting the rising popularity of the RNA world hypothesis. Perhaps there were also unpublished limitations and negative results of the work cited above that led others to abandon this experimental avenue (but we have no evidence for this).

A number of studies have revisited amino acid-nucleotide interactions, giving mixed support. One study queried amino acid binding sites in existent RNA structures such as riboswitches, aptamers and RNPs (Yarus et al., 2009). They showed that binding sites for amino acids are enriched in both codon and anticodon nucleotides more often than expected from random assignments. This connection is weak, and the most significant results were related to amino acids synthesised "late" in metabolism, which is puzzlingly the opposite of the patterns in the code itself (Figure 6). Another study using the computational technique of molecular docking

showed negative results, with no preferential interactions between amino acids and their codons or anticodons (Moghadam et al., 2020).

More recently, systematic *in silico* experiments using molecular dynamics simulations of the interactions between amino acids and single nucleotides showed that half of all canonical amino acids interact most strongly with their cognate nucleobase at the second position of the anticodon, with weaker preference for binding with their cognate nucleobase in the first codon (Halpern et al., 2023). Experimental data using NMR to detect interactions confirmed the general trend of these preferences for a selection of amino acids (Halpern et al., 2023). While these preferential interactions are weak and statistical, this probably reflects the detection limitations of the methods used. Interestingly, the correlation in hydrophobicity between amino acids and anticodon middle bases present in the code was not replicated in either the molecular dynamics or the NRM data (Halpern et al., 2023). This could simply be a reflection of the types of force field used for the simulations (which generally failed to predict hydrophobic interactions accurately), but also emphasises the need for further experiments (Halpern et al., 2023).

The evidence cited above implies the existence of preferential stereochemical interactions between the anticodon and its cognate amino acid, in accordance with Woese's model for the origin of the genetic code. But a key limitation is that it ignores the adaptor context in which early coding interactions would most plausibly have occurred. Even though the rejection of stereochemical coding interactions motivated Crick's adaptor hypothesis, if such interactions are real they must have existed in the context of an adaptor molecule. Both statistical-protein and frozen accident hypotheses rely on early RNA adaptors being able to "select" amino acids. The key difference lies in whether this selection occurs naturally from biases associated with the anticodons (Woese's statistical proteins), or evolved in the structure of the adaptor (Crick's

508 frozen accident). Experimental work demonstrates that specific aminoacylation of tRNAs
509 without enzymes is possible (Giel-Pietraszuk and Barciszewski, 2006, Krzyżaniak et al., 1994,
510 Krzyzaniak et al., 1998), but it does not resolve the source of this specificity. It is unclear
511 whether bias arises from interactions involving the anticodon itself, as indicated by studies on
512 the anticodon alone, or from other as yet unidentified structural features of the adaptor. Further,
513 studies on simpler adaptor structures (i.e. proto-tRNAs) are needed to establish the source of
514 bias (anticodon interactions or adaptor structure). Resolving this distinction is needed to
515 illuminate the decades old debate on the origin of genetic heredity.

516 ## VI. Conclusions

517 (1) Modern theories of the origin of heredity derive from an original disagreement in the
518 1960s between two opposing hypotheses, associated with Carl Woese on one side and
519 Francis Crick and Leslie Orgel on the other.
520 (2) Crick and Orgel suggested that early life was composed solely of nucleotide polymers
521 without proteins. Translation arose from this scenario through adaptor molecules
522 structurally matched to specific amino acids (Orgel, 1968, Crick, 1968). In this view the
523 genetic code started as a reduced version of itself and expanded as mutated adaptors
524 preferentially bound new amino acids. Selection acted to minimise disruption to existing
525 peptides.
526 (3) Woese, on the other hand, proposed that peptides were already present in early life
527 and were synthesised through rudimentary translation guided by loose stereochemical
528 interactions between anticodons and amino acids (Woese, 1965, Woese, 1970, Woese,
529 1967). Selection then acted to make these assignments more specific.

(4) Two key assumptions differentiate the hypotheses: whether selection happened at the level of the protocell or the naked gene; and whether preferential interactions between nucleotides and amino acids led to a coding bias in early translation.

(5) Crick and Orgel's ideas gained traction and became dominant with the rise of gene-centred view of evolution, ultimately giving rise to the RNA world hypothesis.

(6) Models stemming from Orgel's proposal of selection acting exclusively at the level of the naked gene repeatedly encountered fundamental problems related to information maintenance, including mutational decay, competitive exclusion of sequences and parasitic takeover (Eigen, 1971, Spiegelman, 1970).

(7) The hypercycle model was proposed to mitigate these issues (Eigen and Schuster, 2012). But hypercyclic organisation does not introduce a new level of selection and so remains vulnerable to parasites, while also constraining the evolution of new functions.

(8) A return to protocell-level selection, demonstrated by the stochastic corrector model, resolves the problems of integrating information and allows for the evolution of new functions.

(9) The focus on information integration distracted the field from the other key divergence between models of the origin of heredity: whether stereochemical interactions guided early translation.

(10) The genetic code is known to be non-random, with similar amino acids being encoded by similar codons. This could be explained by shared ancestry, as noted in Crick's model, or by biases from stereochemical interactions in early coding, as put forwards by Woese.

(11) A new interpretation of the patterns in the genetic code shows correlations between hydrophobicity and size of amino acid and their anticodonic nucleotides. This is only compatible with the existence of physical interactions between amino acids and anticodons (Harrison et al., 2022).

(12) Experimental evidence for the presence of these interactions is broadly positive but still sparse; the studies with the strongest and most positive results were conducted in the 1980s and 1990s but have not been replicated since (Shimizu, 1987, Shimizu, 1995, Hobish et al., 1995, Lacey et al., 1984). More recent work shows some evidence for weak preferences for interactions between amino acids and cognate nucleotides (Yarus et al., 2009, Halpern et al., 2023).

(13) Experimental work looking at whether the main source of bias in assignment of simple adaptor molecules to amino-acids stems from anticodon interactions or structural features, is still needed to settle the decades old debate.

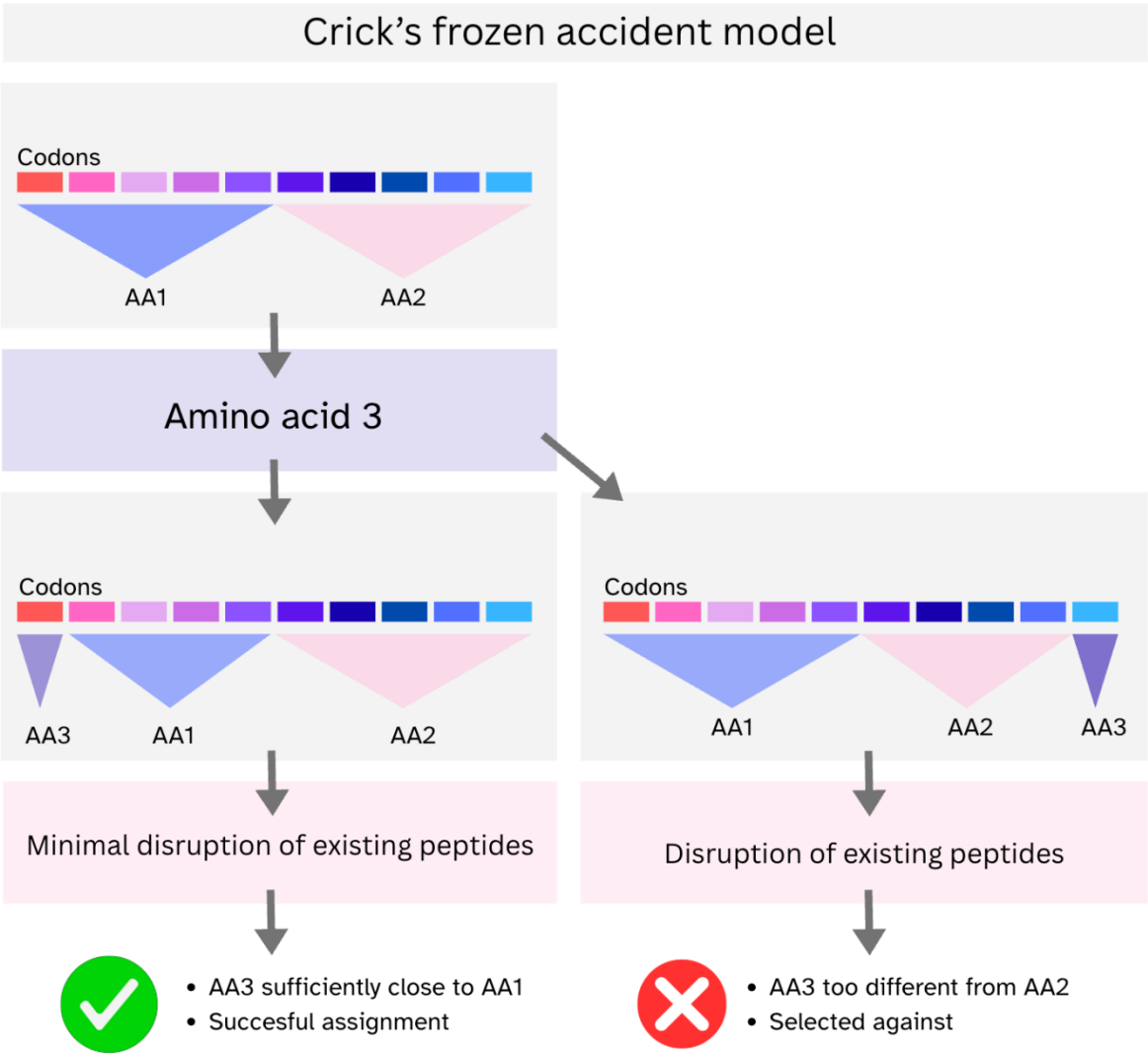
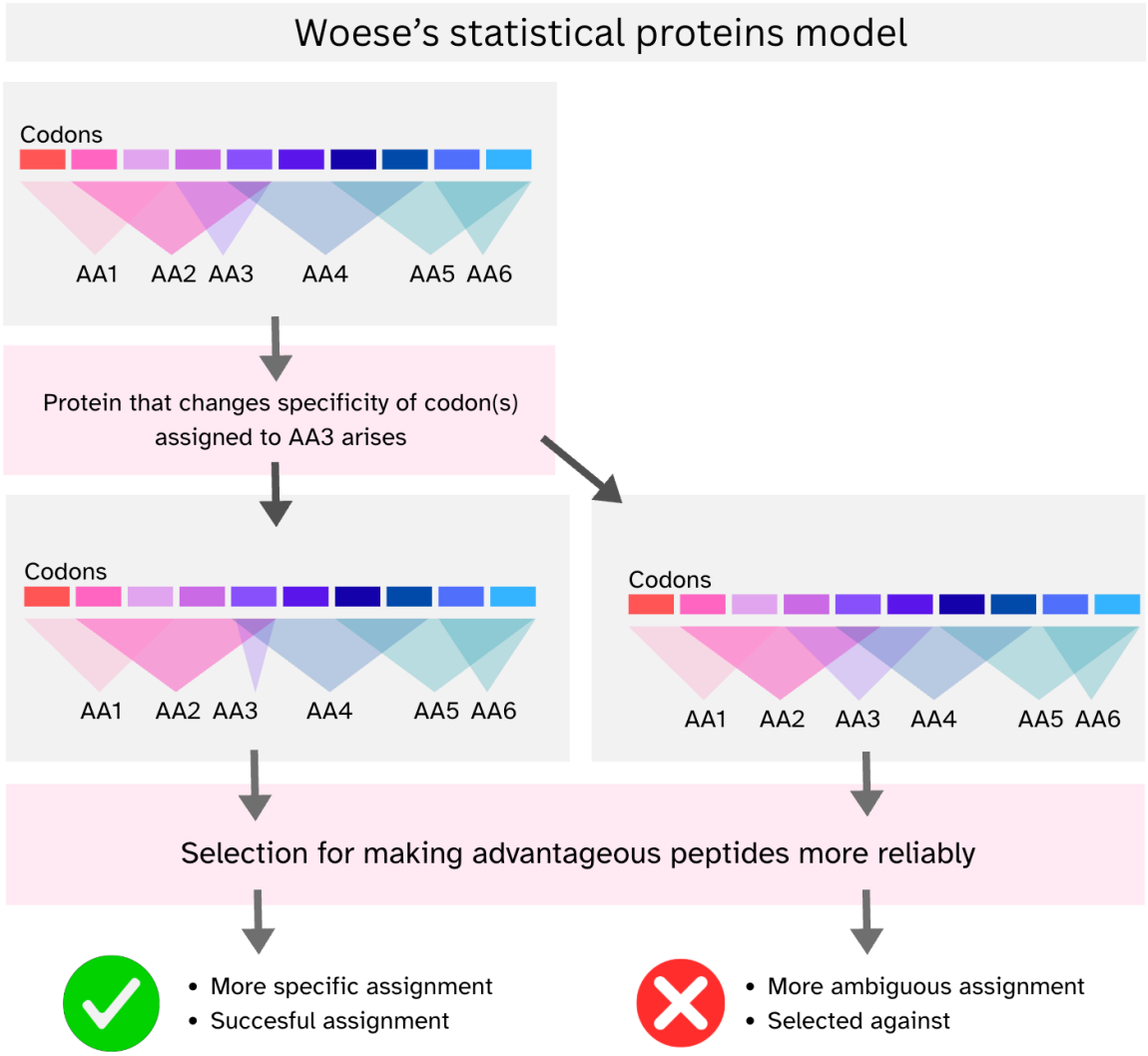


Figure 1. Crick's frozen accident model showing the introduction of a new amino acid assignment according to Crick's frozen accident model. The colours of rectangles represent different codons and of triangles representing assignments to amino acids (AA). These do not match, indicating that chemical affinity between nucleotides and amino acids is specifically precluded in this model.



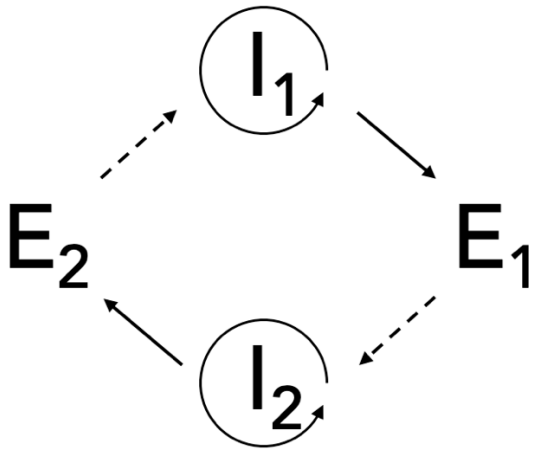
572

573 *Figure 2. Woese's statistical proteins model, showing the introduction of new amino acid*
574 *assignments. The colours of the triangles, representing assignments, and rectangles,*
575 *representing codons, deliberately match to indicate that this model predicts that codons and*
576 *amino acids would have chemical properties that enabled stereochemical interactions.*

577

578

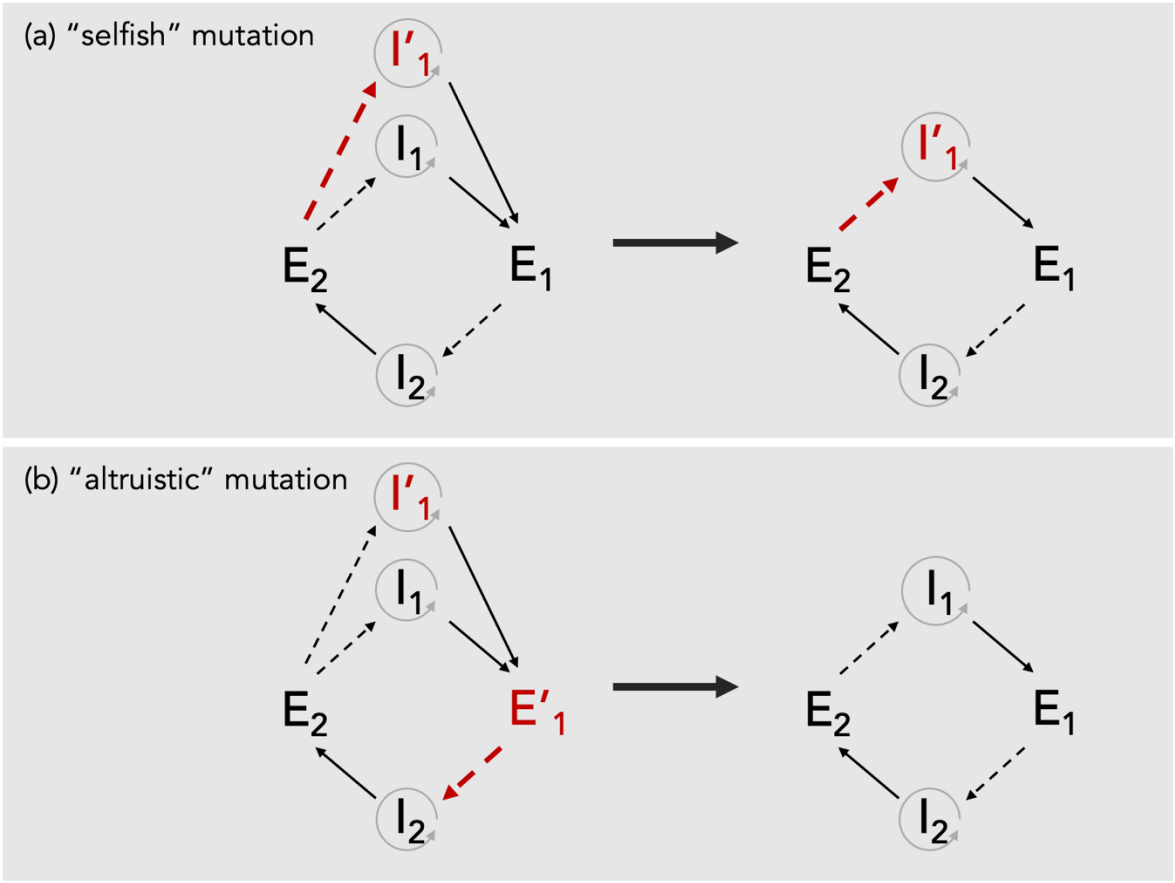
579



580

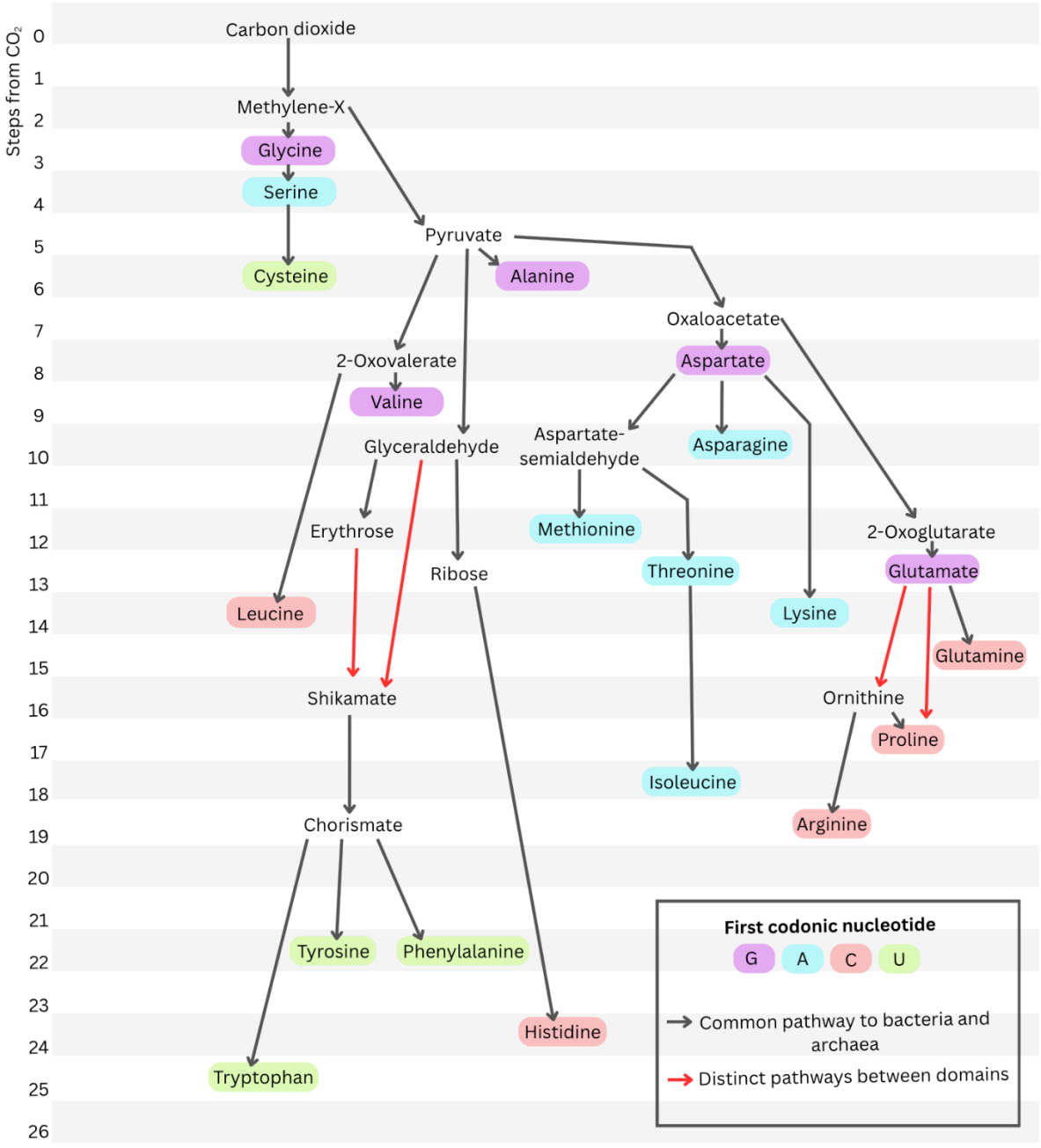
581 *Figure 3. A two-member hypercycle. I_1 and I_2 are nucleotide sequences, E_1 and E_2 are the*
582 *protein replicases encoded by I_1 and I_2, respectively. Round solid arrows show sequence self-*
583 *replication, straight solid arrows show sequences coding for a replicase, dashed arrows show*
584 *the replicase catalysing the replication of a sequence.*

585



586

587 *Figure 4. Hypercycles with (a) "genotypic" mutations that make sequences better target for*
588 *replicases, and (b) "phenotypic" or "altruistic" mutations that make replicases better at*
589 *replicating their target sequence. Solid arrows show sequences coding for a replicase enzyme,*
590 *dashed arrows show the replicase catalysing the replication of a sequence. Stronger catalysis*
591 *by replicases is indicated by red dashed lines.*



592

593 *Figure 5. The structure of protometabolism, illustrating how the autotrophic biosynthesis of*
594 *amino acids links to the genetic code. Amino acids are coloured according to the nucleotide*
595 *specified at their first codon position. Black arrows indicate synthetic routes common to both*
596 *bacteria and archaea; red arrows indicate potentially distinct pathways between the domains.*
597 *White and grey lines in the background indicate the number of reaction steps from CO_2. Note*
598 *this is a conservative map, as glycine, alanine, aspartate and glutamate are a few steps from*

CO_2 as an initial precursor, but only one step from a synthesis involving CO_2 fixation (e.g. the step from pyruvate to oxaloacetate involves fixing CO_2). This figure was adapted from Harrison et al. (2023).

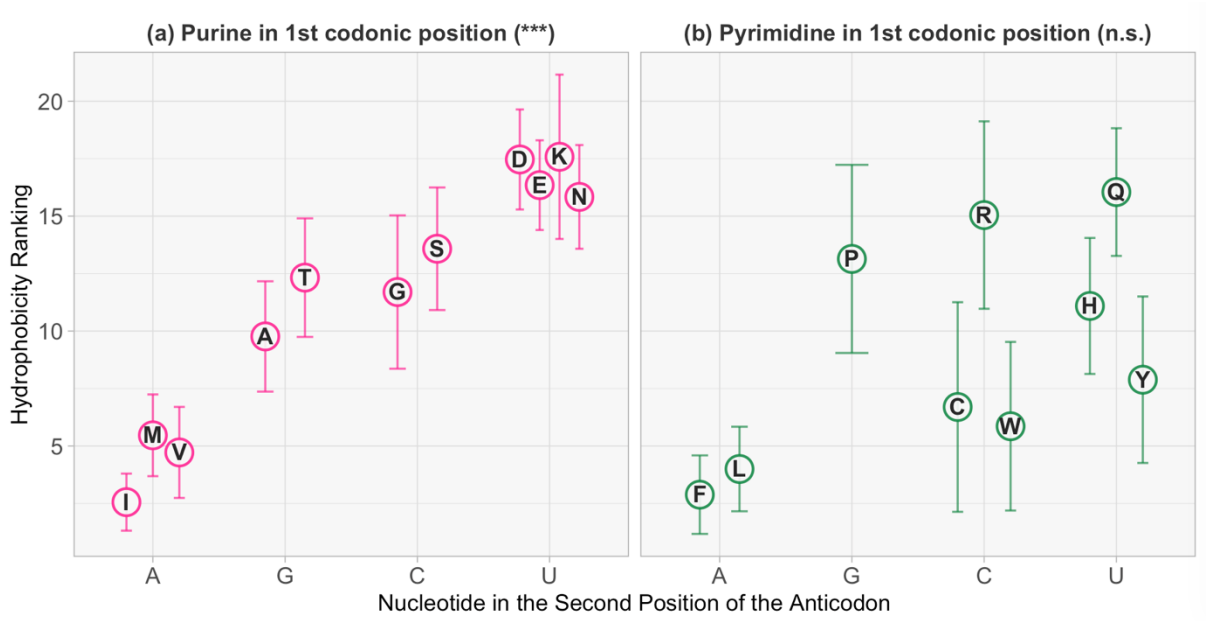


Figure 6. Plot of amino acid hydrophobicity grouped by codon structure. Circles show mean amino-acid hydrophobicity across 42 published scales. Vertical bars indicate ± 1 standard deviation, reflecting disagreement in hydrophobicity scales. Amino acids are grouped on the x-axis by the nucleotide in the second anticodon position (going from most hydrophobic, A, to most hydrophilic, U). Panels indicate first codon position (purine or pyrimidine). Letters denote the standard one-letter amino-acid codes (A, alanine; C, cysteine; D, aspartate; E, glutamate; F, phenylalanine; G, glycine; H, histidine; I, isoleucine; K, lysine; L, leucine; M, methionine; N, asparagine; P, proline; Q, glutamine; R, arginine; S, serine; T, threonine; V, valine; W, tryptophan; Y, tyrosine). Asterisks denote the significance of the effect of second-position nucleotide ordering on amino-acid hydrophobicity, tested using linear mixed-effects

615 *models that account for amino-acid identity and first-position nucleotide. P-values were*
616 *adjusted for multiple testing; *** $p < 0.001$, n.s., not significant. Data based on Harrison et*
617 *al. (2022).*

618

ÅGREN, J. A. 2016. Selfish genetic elements and the gene's-eye view of evolution. *Current Zoology,* 62**,** 659-665.

BERNHARDT, H. S. 2012. The RNA world hypothesis: the worst theory of the early evolution of life (except for all the others)a. *Biology Direct,* 7**,** 23.

BOERLIJST, M. C. & HOGEWEG, P. 1991. Spiral wave structure in pre-biotic evolution: Hypercycles stable against parasites. *Physica D: Nonlinear Phenomena,* 48**,** 17-28.

BRYSON, V. & VOGEL, H. J. 1965. Evolving Genes and Proteins. *Science,* 147**,** 68-71.

CALDARARO, F. & DI GIULIO, M. 2022. The genetic code is very close to a global optimum in a model of its origin taking into account both the partition energy of amino acids and their biosynthetic relationships. *Biosystems,* 214**,** 104613.

CECH, T. R., ZAUG, A. J. & GRABOWSKI, P. J. 1981. In vitro splicing of the ribosomal RNA precursor of Tetrahymena: involvement of a guanosine nucleotide in the excision of the intervening sequence. *Cell,* 27**,** 487-96.

COPLEY, S. D., SMITH, E. & MOROWITZ, H. J. 2005. A mechanism for the association of amino acids with their codons and the origin of the genetic code. *Proc Natl Acad Sci U S A,* 102**,** 4442-7.

CRAPITTO, A. J., CAMPBELL, A., HARRIS, A. J. & GOLDMAN, A. D. 2022. A consensus view of the proteome of the last universal common ancestor. *Ecology and Evolution,* 12**,** e8930.

CRICK, F. H. 1958. On protein synthesis. *Symp Soc Exp Biol,* 12**,** 138-63.

CRICK, F. H. 1968. The origin of the genetic code. *J Mol Biol,* 38**,** 367-79.

CRICK, F. H. C. On Degenerate Templates and the Adaptor Hypothesis: A Note for the RNA Tie Club. 1955.

DAWKINS, R. 1976. *The selfish gene*, Oxford university press.

DI GIULIO, M. & MEDUGNO, M. 2000. The robust statistical bases of the coevolution theory of genetic code origin. *J Mol Evol,* 50**,** 258-63.

EIGEN, M. 1971. Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften,* 58**,** 465-523.

EIGEN, M. & SCHUSTER, P. 1977. The hypercycle. A principle of natural self-organization. Part A: Emergence of the hypercycle. *Naturwissenschaften,* 64**,** 541-65.

EIGEN, M. & SCHUSTER, P. 1979. *The Hypercycle : A Principle of Natural Self-Organization / by M. Eigen, Peter Schuster,* Berlin, Heidelberg, Springer Berlin Heidelberg.

EIGEN, M. & SCHUSTER, P. 2012. *The Hypercycle: A Principle of Natural Self-Organization*, Springer Berlin Heidelberg.

GAMOW, G. 1954. Possible relation between deoxyribonucleic acid and protein structures. *Nature,* 173**,** 318-318.

GIEL-PIETRASZUK, M. & BARCISZEWSKI, J. 2006. Charging of tRNA with non-natural amino acids at high pressure. *The FEBS Journal,* 273**,** 3014-3023.

GILBERT, W. 1986. Origin of life: The RNA world. *Nature,* 319**,** 618-618.

GOLDBERG, A. L. & WITTES, R. E. 1966. Genetic code: Aspects of organization. *Science,* 153**,** 420-424.

GUERRIER-TAKADA, C., GARDINER, K., MARSH, T., PACE, N. & ALTMAN, S. 1983. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell,* 35**,** 849-57.

HALPERN, A., BARTSCH, L. R., IBRAHIM, K., HARRISON, S. A., AHN, M., CHRISTODOULOU, J. & LANE, N. 2023. Biophysical interactions underpin the emergence of information in the genetic code. *Life (Basel),* 13.

HAMILTON, W. D. 1964a. The genetical evolution of social behaviour. I. *Journal of Theoretical Biology,* 7**,** 1-16.

HAMILTON, W. D. 1964b. The genetical evolution of social behaviour. II. *Journal of Theoretical Biology,* 7**,** 17-52.

HARRISON, S. A., PALMEIRA, R. N., HALPERN, A. & LANE, N. 2022. A biophysical basis for the emergence of the genetic code in protocells. *Biochimica et Biophysica Acta (BBA) - Bioenergetics***,** 148597.

HOAGLAND, M. B., STEPHENSON, M. L., SCOTT, J. F., HECHT, L. I. & ZAMECNIK, P. C. 1958. A soluble ribonucleic acid intermediate in protein synthesis. *J Biol Chem,* 231**,** 241-57.

HOBISH, M. K., WICKRAMASINGHE, N. S. & PONNAMPERUMA, C. 1995. Direct interaction between amino acids and nucleotides as a possible physicochemical basis for the origin of the genetic code. *Adv Space Res,* 15**,** 365-82.

HUBER, C. & WÄCHTERSHÄUSER, G. 1997. Activated Acetic Acid by Carbon Fixation on (Fe,Ni)S Under Primordial Conditions. *Science,* 276**,** 245-247.

JOYCE, G. F. 2009. Evolution in an RNA world. *Cold Spring Harb Symp Quant Biol,* 74**,** 17-23.

JOYCE, G. F. & ORGEL, L. E. 1993. Prospects for understanding the origin of the RNA world. *Cold Spring Harbor Monograph Series,* 24**,** 1-1.

KIM, S., SUDDATH, F., QUIGLEY, G., MCPHERSON, A., SUSSMAN, J., WANG, A., SEEMAN, N. & RICH, A. 1974. Three-dimensional tertiary structure of yeast phenylalanine transfer RNA. *Science,* 185**,** 435-440.

KOONIN, E. V. 2003. Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat Rev Microbiol,* 1**,** 127-36.

KRUGER, K., GRABOWSKI, P. J., ZAUG, A. J., SANDS, J., GOTTSCHLING, D. E. & CECH, T. R. 1982. Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. *Cell,* 31**,** 147-57.

KRZYŻANIAK, A., BARCISZEWSKI, J., SAŁAŃSKI, P. & JURCZAK, J. 1994. The non-enzymatic specific aminoacylation of transfer RNA at high pressure. *International Journal of Biological Macromolecules,* 16**,** 153-158.

KRZYZANIAK, A., TWARDOWSKI, T., BARCISZEWSKI, J., SAŁANSKI, P. & JURCZAK, J. 1998. tRNA aminoacylated at high pressure is a correct substrate for protein biosynthesis. *IUBMB Life,* 45**,** 489-500.

LACEY, J. C., MULLINS, D. W. & KHALED, M. A. 1984. The case for the anticode. *Origins of life,* 14**,** 505-511.

LAGERKVIST, U. 1978. "Two out of three": an alternative method for codon reading. *Proceedings of the National Academy of Sciences,* 75**,** 1759-1762.

LANE, N., ALLEN, J. F. & MARTIN, W. 2010. How did LUCA make a living? Chemiosmosis in the origin of life. *BioEssays,* 32**,** 271-280.

MARTIN, W. & RUSSELL, M. J. 2003. On the origins of cells: a hypothesis for the evolutionary transitions from abiotic geochemistry to chemoautotrophic prokaryotes, and from prokaryotes to nucleated cells. *Philos Trans R Soc Lond B Biol Sci,* 358**,** 59-83; discussion 83-5.

MARTIN, W. & RUSSELL, M. J. 2007. On the origin of biochemistry at an alkaline hydrothermal vent. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 362**,** 1887-1926.

MAYNARD SMITH, J. 1964. Group Selection and Kin Selection. *Nature,* 201**,** 1145-1147.

MAYNARD SMITH, J. 1988. Hypercycles and the Origin of Life. *In:* SMITH, J. M. (ed.) *Did Darwin Get It Right? Essays on Games, Sex and Evolution.* Boston, MA: Springer US.

MAYNARD SMITH, J., BROOKFIELD, J. F. Y., BODMER, W. F. & KINGMAN, J. F. C. 1983. Models of evolution. *Proceedings of the Royal Society of London. Series B. Biological Sciences,* 219**,** 315-325.

MICHOD, R. E. 1983. Population Biology of the First Replicators: On the Origin of the Genotype, Phenotype and Organism. *American Zoologist,* 23**,** 5-14.

MOGHADAM, S. A., PRETO, J., KLOBUKOWSKI, M. & TUSZYNSKI, J. A. 2020. Testing amino acid-codon affinity hypothesis using molecular docking. *Biosystems,* 198**,** 104251.

MOODY, E. R. R., ÁLVAREZ-CARRETERO, S., MAHENDRARAJAH, T. A., CLARK, J. W., BETTS, H. C., DOMBROWSKI, N., SZÁNTHÓ, L. L., BOYLE, R. A., DAINES, S., CHEN, X., LANE, N., YANG, Z., SHIELDS, G. A., SZÖLLŐSI, G. J., SPANG, A., PISANI, D., WILLIAMS, T. A., LENTON, T. M. & DONOGHUE, P.

C. J. 2024. The nature of the last universal common ancestor and its impact on the early Earth system. *Nature Ecology & Evolution*.

MOROWITZ, H. J., HEINZ, B. & DEAMER, D. W. 1988. The chemical logic of a minimum protocell. *Origins of Life and Evolution of the Biosphere,* 18**,** 281-287.

ORGEL, L. E. 1968. Evolution of the genetic apparatus. *Journal of Molecular Biology,* 38**,** 381-393.

OUZOUNIS, C. A., KUNIN, V., DARZENTAS, N. & GOLDOVSKY, L. 2006. A minimal estimate for the gene content of the last universal common ancestor—exobiology from a terrestrial perspective. *Research in Microbiology,* 157**,** 57-68.

RICH, A. 1962. On the problems of evolution and biochemical information transfer. *Horizons in biochemistry Albert Szent-Györgyi dedicatory volume***,** 103-126.

ROBERTSON, M. P. & JOYCE, G. F. 2012. The origins of the RNA world. *Cold Spring Harb Perspect Biol,* 4.

ROBERTUS, J., LADNER, J. E., FINCH, J., RHODES, D., BROWN, R., CLARK, B. & KLUG, A. 1974. Structure of yeast phenylalanine tRNA at 3 Å resolution. *Nature,* 250**,** 546-551.

SALDITT, A., KARR, L., SALIBI, E., LE VAY, K., BRAUN, D. & MUTSCHLER, H. 2023. Ribozyme-mediated RNA synthesis and replication in a model Hadean microenvironment. *Nat Commun,* 14**,** 1495.

SARDANYÉS, J. & SOLÉ, R. V. 2006. Bifurcations and phase transitions in spatially extended two-member hypercycles. *Journal of Theoretical Biology,* 243**,** 468-482.

SARDANYÉS, J., TOMÁS LÁZARO, J., GUILLAMON, A. & FONTICH, E. 2017. Full analysis of small hypercycles with short-circuits in prebiotic evolution. *Physica D: Nonlinear Phenomena,* 347**,** 90-108.

SHIMIZU, M. 1987. Specific interactions of dinucleoside monophosphates with their cognate amino acids. *Journal of the Physical Society of Japan,* 56**,** 43-45.

SHIMIZU, M. 1995. Specific aminoacylation of C4N hairpin RNAs with the cognate aminoacyl-adenylates in the presence of a dipeptide: origin of the genetic code. *The Journal of Biochemistry,* 117**,** 23-26.

SPIEGELMAN, S. 1970. Extracellular evolution of replicating molecules. *The neurosciences. The Rockefeller University, New York, NY***,** 927.

SWETINA, J. & SCHUSTER, P. 1982. Self-replication with errors: A model for polvnucleotide replication22This paper is considered as part II of Model Studies on RNA replication. Part I is the Gassner and Schuster [14]. *Biophysical Chemistry,* 16**,** 329-345.

SZATHMÁRY, E. 1989. The integration of the earliest genetic information. *Trends Ecol Evol,* 4**,** 200-4.

766 SZATHMÁRY, E. & DEMETER, L. 1987. Group selection of early replicators and the
767 origin of life. *J Theor Biol,* 128**,** 463-86.

768 TAYLOR, F. J. & COATES, D. 1989. The code within the codons. *Biosystems,* 22**,** 177-87.

769 TRINQUIER, G. & SANEJOUAND, Y.-H. 1998. Which effective property of amino acids is
770 best preserved by the genetic code? *Protein engineering,* 11**,** 153-169.

771 TZE-FEI WONG, J. 1981. Coevolution of genetic code and amino acid biosynthesis. *Trends*
772 *in Biochemical Sciences,* 6**,** 33-36.

773 WHITE, H. B. 1976. Coenzymes as fossils of an earlier metabolic state. *Journal of Molecular*
774 *Evolution,* 7**,** 101-104.

775 WILLIAMS, G. C. 1966. *Adaptation and Natural Selection*

776 *A Critique of Some Current Evolutionary Thought*, Princeton University Press.

777 WOESE, C. 1969. Models for the evolution of codon assignments. *J Mol Biol,* 43**,** 235-40.

778 WOESE, C. R. 1965. On the evolution of the genetic code. *Proc Natl Acad Sci U S A,* 54**,**
779 1546-52.

780 WOESE, C. R. 1967. *The genetic code : the molecular basis for genetic expression / Carl R.*
781 *Woese,* New York, Harper and Row.

782 WOESE, C. R. 1970. The Problem of Evolving a Genetic Code. *BioScience,* 20**,** 471-485.

783 WOESE, C. R., DUGRE, D. H., SAXINGER, W. C. & DUGRE, S. A. 1966. The molecular
784 basis for the genetic code. *Proc Natl Acad Sci U S A,* 55**,** 966-74.

785 WONG, J. T.-F. 1975. A co-evolution theory of the genetic code. *Proceedings of the*
786 *National Academy of Sciences,* 72**,** 1909-1912.

787 WONG, J. T.-F. 2005. Coevolution theory of the genetic code at age thirty. *BioEssays,* 27**,**
788 416-425.

789 XAVIER, J. C., GERHARDS, R. E., WIMMER, J. L. E., BRUECKNER, J., TRIA, F. D. K.
790 & MARTIN, W. F. 2021. The metabolic network of the last bacterial common
791 ancestor. *Communications Biology,* 4**,** 413.

792 YARUS, M., WIDMANN, J. J. & KNIGHT, R. 2009. RNA–amino acid binding: a
793 stereochemical era for the genetic code. *Journal of Molecular Evolution,* 69**,** 406-429.

794 ZACHAR, I., FEDOR, A. & SZATHMÁRY, E. 2011. Two different template replicators
795 coexisting in the same protocell: stochastic simulation of an extended chemoton
796 model. *PLOS ONE,* 6**,** e21380.

797 ZINTZARAS, E., SANTOS, M. & SZATHMÁRY, E. 2002. "Living" Under the Challenge
798 of Information Decay: The Stochastic Corrector Model vs. Hypercycles. *Journal of*
799 *Theoretical Biology,* 217**,** 167-181.

800