1

The planetary scope of biogenesis: the biosphere is the fourth geosphere

The origin of life was a planetary process, in which a departure from non-living states led to a new kind of order for matter and energy on this planet. To capture the role of life as a planetary subsystem we draw on the concept of geospheres from geology. Three traditional geospheres – the atmosphere, hydrosphere, and lithosphere – partition terrestrial matter into three physical states, each associated with a characteristic energetics and chemistry. The emergence of life brought the biosphere into existence as a fourth geosphere. The biosphere is an inherently dynamical state of order, which produces unique channels for energy flow through processes in carbon-based chemistry. The many similarities, and the interdependence, of biochemistry with organometallic chemistry of the lithosphere/hydrosphere interface, suggests a continuity of geochemistry with the earliest biochemistry. We will argue that dynamical phase transitions provide the appropriate conceptual frame to unify chance and necessity in the origin of life, and to express the lawfulness in the organization of the biosphere. The origin of life was a cascade of non-equilibrium phase transitions, and biochemistry at the ecosystem level was the bridge from geochemistry to cellular life and evolution. The universal core of metabolism provides a frame of reference that stabilizes higher levels of biotic organization, and makes possible the complexity and open-ended exploration of evolutionary dynamics.

1.1 A new way of being organized

The emergence of life on Earth brought with it, for the first time on this planet, a new way for matter and energy to be organized. Our goal is to understand this transition, how it happened and what it means. The question how life emerged – what sequence of stages actually occurred historically – can at present be answered only at the level of sketches and suggestions, though for some stages we believe good enough arguments can be made to guide experiments. To arrive at a sketch, however, we cannot escape making many choices of interpretation, of things known about life and its planetary context.

Life emerged in an era not accessible to us through historical reconstruction. Our claims about what happened in this era will depend on the principles we use to generalize,

simplify, and extrapolate from knowledge of modern life and a few fossilized signatures that become increasingly fragmentary and difficult to interpret on the approach to the beginning that we wish to understand. More important than any claim we can make about past events will be whether we frame the problem of emergence in terms that capture its most important abstractions.

Life appeared on Earth in a period known as the Hadean, the Earth's oldest eon, a reference to the etymology of Hades as "the unseen" [196]. Other than gross features of planetary composition, it has left no detailed signatures in the present because it is a time from which later eras preserved no memory.[1] The Hadean was, however, a time we think of as governed by laws of geophysics and geochemistry, and therefore open to understanding. (Indeed, the absence of memory makes the Hadean, more than later periods with accreted history, a simpler period to study with general laws.) The earliest stages of life were scaffolded by these geological laws, and in some respects may even have been continuous with them. Thus in geology the complement to historicity is lawfulness. What we cannot infer from preserved memories we seek to deduce by understanding the action of laws.

We argue in this book that the same is true for life. The complement to historicity in its earliest periods was not chaos, but lawfulness, albeit perhaps lawfulness of a statistical nature. Life inherited the laws of geochemistry, and grew out of geochemical precursors because some of those laws required the formation of a new state of order qualitatively unlike any of the lifeless states of matter. Life is still in large part lawful, if looked at in the right way, and some of the laws that govern the living world today are good candidates for laws that were at work during its emergence. However, modern life is also historical, to a much greater degree than modern geology, so one of the challenges to making the correct abstractions about its origin is to recognize and separate the contributions of law and of history.

The origin of life was a process of departure and a process of arrival: a departure from non-living states that we understand with natural laws, and an arrival at a new living mode of organization that is robust, persistent, and in its own respects law-like. To understand why a departure was prompted, and why the arrival has been stable, we must begin by recognizing that life is a planetary process, and that its emergence was a passage between two planetary stages.

1.1.1 Life is a planetary process

The emergence of life was a major transition in our planet's formative history, alongside the accretion of its rocky core, the deposition and eventual persistence of its oceans, and the accumulation of its atmosphere. Several facts about the timing, the planetary impacts, and the organizational nature of living systems establish the context within which any theory of origins must make sense.

[1] As is fitting, one of the five rivers in Hades from Greek mythology is Lethe, the river of forgetfulness.

Life apparently emerged early Conditions on Earth earlier than about 4 billion years ago appear to have been too hot and desiccated from asteroid infall to permit even the chemical constituents that we now associate with life to exist, much less to permit its processes to take place.[2] Yet evidence, either from explicit microfossils or from reworking of element compositions and isotope ratios that we associate with life, suggests that as early as 3.8 billion years ago, and with much greater confidence by 3.5 billion years ago, cells existed that must have possessed much of the metabolic and structural complexity that is common to all life today [444]. Given the extremely fragmentary nature of the rock record from this ancient time, the existence of the signatures we know suggests that life first became established on Earth in a geological interval that was shorter than 200 million years – possibly much shorter – an interval that was also a period of geological transition on a still young planet.

Living systems have (in some cases radically) altered planetary chemistry Living systems have always altered the chemistry of their local environments (these changes are the most reliable ancient biosignatures), and they have gone on to change global planetary chemistry. The most striking change was the filling of the atmosphere and oceans with oxygen, which changed the profiles of elements in solution in the oceans, altered continental weathering, and increased as much as threefold the diversity of minerals formed on Earth [348]. By capturing trace elements in biomass, effectively creating microenvironments for them vastly different from the surrounding physical environments, organisms also govern the concentration and distribution of metals, phosphate, and sulfur, and influence the great cycles of carbon, nitrogen, and water [236].

Living systems are ordered in many ways at many levels Whether in terms of composition, spatial configuration, or dynamics, living systems are ordered in many different ways at many scales. The chemical composition of biomass is distinguished from non-living matter by at least three major classes of synthetic innovation, in small molecules (metabolites), mesoscale molecules (cofactors), and macromolecules (lipids, polynucleotides, polypeptides, and polysaccharides). These components are organized in physical-chemical assemblies, including phase separations and gels, non-covalently bonded and geometrically interlinked molecular complexes, various compartments, cells, colonies, organisms, and ecosystems of a bewildering array of kinds. The essential heterogeneity of the many different kinds of order, and the diversity of processes that have been harnessed to generate them, is a fundamental and not merely incidental aspect of life's complexity.

Although diverse and heterogeneous, living order is also highly selected At the same time as the diversity and heterogeneity of living order creates a complex challenge of

[2] Whether asteroid infall underwent a late pulse, known as the "late heavy bombardment," sufficient to melt and desiccate the entire Earth's surface concurrently, or tapered more gradually so that sub-crustal water was only locally and intermittently removed, is a point of uncertainty, and this creates uncertainty of as much as 200–300 million years in estimates of the earliest time the Earth could have sustained organic compounds.

explanation, it is important to recognize that, within each "kind" of order, the observed ordered forms comprise a vanishingly small set within the possible arrangements of similar kind.[3] We may characterize the sparseness of observed kinds of order by saying that each ordered form is "selected" – for stability, for functionality, or by some other criteria – but in first surveying the qualitative character of life, we wish to suspend theoretical assumptions about how the selection is carried out, because (we will argue) this turns out to be a complicated question to frame properly. Whether observed forms of living order are sparse because they are uniquely specified by first principles, or because the unfolding of a historically contingent evolutionary process has only sparsely sampled its possibilities, will be fundamental to our understanding of the role of laws in biology, including life's origin.

An invariant simple foundation underlies unlimited complexity at higher levels At the core of life lies a network for the synthesis of the small organic molecules from which all biomass is derived. Remarkably, this core network of molecules and pathways is small (containing about 125 basic molecular building blocks) and very highly conserved. If viewed at the ecosystem level – meaning that, for each compound, one asks what pathways must have been traversed in the course of its synthesis, disregarding which species may have performed the reaction or what trophic exchanges may have befallen pathway intermediates along the way – the core network is also essentially universal. Some gateway reactions are strictly conserved. In areas where synthetic pathways do show variation, the network tends to be highly modular, and variations take the form of modest innovations constrained by key molecules that serve as branching points.[4] This universality made possible S. Dagley's and Donald Nicholson's assembly of a chart of intermediary metabolism [173] that generalized across organisms. Other universal features of life include its use of several essential cofactors and RNA, and some chemical aspects of bioenergetics and cellular compartmentalization. A more complex and enigmatic, but also nearly universal, feature of life is the genetic code for ribosomal protein synthesis. These higher-level universals are discussed in Chapter 5.

This small and universal foundation of life is a platform for the generation of apparently unlimited variation and complexity in higher-level forms. These range from cell architectures to species identities and capabilities, and ecological community assemblies and their coevolutionary dynamics. The contrast that is so striking is that in its invariant core elements, life is universal down to much more particular components and rules of assembly even than other broad classes of matter such as crystalline solids. Yet in its higher levels of aggregation, it appears to have open-ended scope for variation that has no counterpart in non-living states of matter.

[3] We return in Chapter 7 to a more systematic discussion of possible versus realized forms of order, and the significance of the fact that realized order has vanishingly small measure relative to possible forms.
[4] We provide a much more detailed discussion of metabolic modularity, conservation, and variations, to explain these claims, in Chapter 4.

Life as a whole has been a durable feature of Earth The presence of living systems has apparently been a constant and continuous feature of the Earth since their first appearance at least 3.8 billion years ago. The persistence or tenacity of life bears a resemblance to features that arise in the geological progression of a maturing planet, and we will see that this resemblance extends to the incremental elaboration of complexity in life's universal core as well. The perplexity in this observation is that other stable, invariant geological features result from physical processes under conditions that, at least in principle, we know how to produce in the laboratory or in computer simulations. The states of matter to which they correspond also tend to be reproducible under broadly similar conditions, so they tend to recur in broadly similar environments. In contrast, we currently have essentially no understanding of what laboratory conditions would reproduce the emergence of life. Current observation of non-Earth systems, including meteorites and other planets or moons, is also consistent with the absence of signatures we would characterize as unequivocally biotic. These observations have been interpreted by Francis Crick [168] and Jacques Monod [561] (among others) as circumstantial evidence that life is improbable or accidental.[5] While we believe this interpretation is unjustified, the observations do imply at the least that the conditions for life to emerge and persist are much more *particular* than we have come to associate with robust and persistent physical states of matter.

The persistent presence of living states on Earth, including persistence of the universal core, is more striking because higher-level systems such as cell form and catalytic capability have undergone major episodes of innovation, while still higher levels such as species identities or ecological community structures exist in a state of almost constant flux or turnover. These higher-level systems appear, at any time, to be essential to carrying out core processes, yet they have been much more variable than either the core that depends on them, or the broad characteristics of the living state by which all life shares a family resemblance. Reconciling such signatures of accident and fragility, with other signatures of robustness that we normally associate with inevitability, is one of the longstanding puzzles in understanding the nature of the living state.

1.1.2 Drawing from many streams of science

Natural languages for the origins of order can be drawn from many areas in biology, including functional and comparative studies of metabolism and cell physiology, molecular and cellular architecture, the nature of catalysis, genomic mechanisms of hereditary memory and regulation, the complex and multilevel character of individuality, the ways selection and regulation interact to produce developmental programs broadly construed, the reconstructed evolutionary history of many of these systems and their functions, and a host of regularities in ecological productivity, community assembly, and dynamics in both extant

[5] Crick's characterization was "a happy accident, indeed nearly a miracle."

and reconstructed ecosystems. We will summarize some of these, and pursue others in greater depth, in Chapters 2 through 5.

In addition to the biological sciences, we have paradigms of architecture and control from engineering, and important theories of stability from physics and (closely related) of optimal error correction from information theory. These provide potentially useful abstractions of functions performed in living systems, and in some cases strong theorems about the limits of possibility. They will be developed in Chapters 7 and 8.

All of these provide windows on the nature of life. They capture patterns in the living world that do not exist without life, and in a piecemeal fashion, they often partially characterize mechanisms by which those patterns are created and maintained. We would like them to define the problem of departure from a non-living planet that must be understood.

1.2 The organizing concept of geospheres

The problem of unifying diverse phenomena is not new with biology. Similar problems have arisen in planetary science, involving as it does a variety of chemistries, physical phases of matter, and classes and timescales of dynamics. Here a traditional coarse-grained partition of planetary systems into "geospheres" has remained useful into the modern era. The 1949 text *Geochemistry* by Kalervo Rankama and Thure G. Sahama [664] partitioned the non-living matter of Earth into three "geospheres": the atmosphere, hydrosphere, and lithosphere.[6] The concern of the authors was to provide an overview of the chemical partitioning and physical states of all matter on the planet.

The "geosphere" designations are very coarse, and to understand their use it is helpful to keep in mind the kinds of distinctions they are *not* meant to make. The geosphere partition, for the most part, does not separate regions with sharply defined boundaries; their components often interpenetrate and interact. The geospheres also do not aim at strict chemical partitions. For example, water, the primary constituent of the hydrosphere and source of its name, is present in the atmosphere, and in the lithosphere both as hydrate of minerals and as a component in trapped fluids.

Despite (and in some ways, because of) its qualitative and approximating nature, the language of geospheres is useful because it groups *multiple classes of distinctions* that are inter-related and that share the same domains of space and often similar states of matter. A coarse partition into regions and aggregate states and dynamics, by pre-empting other classifications according to specific chemical identity or sharp spatial boundaries, emphasizes

[6] These three coarsely defined geospheres are all subject to much more refined description. In modern usage, the atmosphere subdivides into an ionosphere, mesosphere, stratosphere, and troposphere, and the hydrosphere layers into epipelagic, mesopelagic, bathypelagic, abyssalpelagic, and hadalpelagic zones plus crustal and sub-crustal water. The lithosphere, used by Rankama and Sahama to refer to the totality of rocky and metallic zones on Earth, is now refined into zones that are in some respects almost as different from one another as they are from the hydrosphere. The term "lithosphere" is now used specifically to refer to the crust and rigid layer of the upper mantle, followed in depth by the plastic asthenosphere, the stiff (though still plastic) lower mantle, the liquid iron/nickel metal outer core, and the solid (also Fe/Ni metal) inner core. The refinement, however, only changes in degree but not in spirit the original function of qualitatively partitioning fine-scale structures and dynamics into useful aggregate domains.

that the system-level relations and interactions are the unifying concept for each geosphere, rather than an exclusive list of material components.

1.2.1 The three traditional geospheres

Each of the three traditional geospheres is associated with one or a few primary groups of chemical constituents, a primary phase of matter, and a characteristic class of chemical reactions.

Atmosphere Gas phase, composed of small molecules made principally from non-metals and noble gases, which exist as gases over large ranges of temperature from ~ -40 to $\sim 100\,^{\circ}\mathrm{C}$. Primary chemistry is photolytically excited gas-phase free-radical chemistry, with some ionization chemistry in the upper layers. High activation energies of excited states produce reactive compounds, which persist only at low density.

Hydrosphere Liquid phase, water solutions. Oxides of nitrogen and sulfur may be present as solutes in relatively high concentrations; the concentration of metals (particularly transition metals) depends sensitively on oxidation/reduction (or *redox*[7]) state through reactions to form insoluble compounds with non-metals. Primary chemistry is oxidation/reduction, acid/base, and hydration/dehydration chemistry. Radical intermediates have high energies and are not produced except very near the surface due to screening of light by liquid-water scattering and absorption, and they quench rapidly when formed. Acid/base and oxidation/reduction reactions may be coupled due to the high solubility of protons in water, in contrast to extremely low solubility of electrons.

Lithosphere Solid phase, dominated (outside the core) by the crystallography of silicate and sulfide minerals, with carbonates, hydroxides, and other metal oxides as lesser constituents. Much of the chemistry of the lithosphere is physical chemistry of phase transitions including melt-fractionation, dissolution, precipitation, and stoichiometric rearrangement in solid solutions. Many phase transitions involve changes in oxidation states of metals, driven by the crystallography of silicates as a function of temperature and pressure. Changes in compatibility of minor elements with temperature and pressure can be a large determinant of pH for included fluids. Although redox changes for transition metals often result from coordination changes in crystallographic contexts, they are of major importance to the chemical activity of the Earth as a whole. Mantle convection can convert heat energy, through long-range transport across temperature and pressure zones, into redox disequilibria that are too energetic to be created by thermal excitations in near-equilibrium conditions.

We return to give a more detailed characterization of some of these properties of the Earth in Chapter 3.

[7] Oxidation and reduction are introduced in Chapter 2.

1.2.2 The interfaces between geospheres

1.2.2.1 Complexity often arises at interfaces where matter is exchanged

Because the same chemicals can pass between geospheres, the interfaces between them can be concentrating centers for thermodynamic disequilibrium and the emergence of complexity. For example, volcanic outgassing, believed to be the main source of the present atmosphere, can release both methane and carbon dioxide from carbon trapped in the mantle when the Earth cooled. It also supplies hydrogen, ammonia, and hydrogen sulfide. Continental weathering is a process at the interface of the lithosphere and atmosphere also involving water, which alters minerals, replenishes trace elements (particularly Ca^{2+}) in the ocean, and plays a major role in sedimentation of carbonates and regulation of both the CO_2 partial pressure of the atmosphere and the atmospheric greenhouse. The ocean/atmosphere interface, where the cross section for absorption of sunlight energy changes drastically between two matter phases, is a primary generator of surface heat that powers evaporation and drives the global weather system. On the early Earth, it was also a boundary across which N_1 compounds could diffuse, between a region where only nitrogen oxides could survive and one in which only ammonia could survive. In the present Earth with its marine biota, the surface (photic) zone is the major zone of primary productivity. Organisms actively regulate not only light absorption and scattering, but also the viscosity of the air/water interface, controlling rates of evaporation, droplet formation, entrainment of bubbles, and thus gas exchange between the atmosphere and oceans. Finally, the lithosphere/hydrosphere interface is an extraordinarily rich zone of disequilibria in temperature, chemical potentials, geometries, and physical properties of matter, which we consider next.

1.2.2.2 The lithosphere/hydrosphere interface is particularly important to life

Of great interest to us, as we try to situate the materials and processes of life in their planetary context, will be convective currents of sub-crustal water near spreading centers and volcanos. This water is part of the lithosphere/hydrosphere interface, and is one of the most chemically active zones of terrestrial matter. Whereas local regions within the mantle, crust, or oceans generally exist very near chemical equilibrium, the interface between the hot, convected rock and surface water is constantly pushed far from equilibrium by the mismatch between the primordial reducing character of the bulk Earth and an atmosphere driven to be more oxidizing through escape processes. The mismatch at the interface is constantly replenished as a secondary effect of the dissipation of heat from fission of radioactive elements present when the planet formed. Sub-crustal convected water systems are a particularly interesting feature of the Earth, because they depend on its composition and its internal heating, and their chemical activity depends on its internal convection as well.[8] The chemical activity at the rock/water interface is closely connected to the chemical

[8] Whether tectonics in the current sense of oceanic basin subduction was a feature of the Hadean Earth is currently debated. We return to this question in Chapter 3.

activity *within* living systems, and modern-day hydrothermal vent systems host rich and ancient biota capable of exploiting this overlap.

Chemical systems tend to equilibrate to within the scale of thermal fluctuations ($k_B T \sim 0.026$ eV at room temperature) if they are not continually re-energized. The thermal activation energy of typical covalent bond modifying reactions ($\gtrsim$0.5 eV) is at least 20 times the available thermal excitation energy under conditions where liquid water exists at surface pressures. Therefore covalent bond modifying chemical activity is seen only at extremely low rates in systems that are only activated thermally. The most important physics question for a chemical origin of life within geophysics is where on Earth chemical potentials can be sufficiently insulated from one another to form large differences, but then brought together rapidly enough to drive chemical reactions rather than simply dissipating as heat.

The key to this creation of sharp chemical disequilibria is **mantle convection**. The tendency of hydrogen to escape from planetary atmospheres, leaving complementary oxidants behind, is a ready source of disequilibrium between the interior and surface of the planet. The insulating layer of the crust provides a strong barrier between these systems so that their redox potentials can move far apart. Mantle convection, resulting in volcanism and under some conditions in plate tectonics, is the force that breaks through this insulating barrier to create local disequilibria. The surface phenomenon of water circulation through heated, cracked rock – a process that is particularly efficient and active at spreading centers and faulting systems on tectonically active planets – then leads to mixing zones where disequilibria that accumulate over millions of years are brought into contact on the molecular scale.

Before the 1970s, it was believed that all life on Earth ultimately owed its existence to energy captured photosynthetically from sunlight. The discovery of hydrothermal vent systems by John Corliss and collaborators [162] using the deep submersible Alvin first revealed a diverse and thriving biota existing out of contact from sunlight, and apparently fed by minerals dissolved in vent fluids and not by detrital carbon. This life was effectively decoupled from the solar energy system except by the existence of liquid water (and, though this is now known not to be limiting, the presence of oxygen produced by photosynthesizers).

Four decades of study of microbial metabolisms and energy sources [481] have gone on to show that an enormous diversity of bacteria and archaea obtain energy from geologically produced electron donors and acceptors in both surficial and subsurface environments,[9] and that this energy is sufficient to maintain self-sufficient life and growth from one-carbon inputs, molecular nitrogen, a few inorganic salts, and trace metals. Hydrothermal systems are profuse sources of these inputs, and support life in anoxic environments that provide better models for the early oceans than oxygenated environments such as surficial hot springs. Vents were quickly proposed [161] as plausible geochemical environments for the origin of life, and since phylogenetic reconstructions increasingly suggest reductive,

[9] Some vent environments support not only microbial assemblies, but complex ecosystems of worms, mollusks, and crustaceans supported by these microbes.

thermophilic metabolisms occupied all the deepest branches of the tree of life, this proposal seems historically plausible as well as energetically feasible.

Within the abiotic matter on Earth, the chemistry at the lithosphere/hydrosphere interface most closely resembles the chemistry of life both in its general character and even in detail. Biochemistry takes place in condensed phases, meaning either aqueous solution or microenvironments created by enzymes or membranes. Gas-phase chemistry is essentially impossible, and photoionization in the strict sense (such as occurs in space) is not used. Much of the bulk of biochemistry consists of reactions that are facile in water and involve either full bonding-pair exchange (oxidations and reductions), proton exchange, or group transfers. When radicals are used, they are formed at metal centers and either hosted on metal centers (as in ferredoxins) or transferred to a limited inventory of highly evolved cofactors. Several investigators (to whose ideas we return in Chapter 6) have emphasized the similarity of biological metal centers to metal sulfide minerals that would have been present in the surface and near-surface on the Hadean Earth. Even the temperatures at which biochemistry is carried out fall within the range found in hydrothermal systems.

1.2.3 The biosphere is the fourth geosphere

The three geospheres of Section 1.2.1 subsume, though only in general terms, the domains of scientific knowledge that would apply to matter and events on a lifeless planet. The part of Earth as we know it that is not even qualitatively accounted for within the three abiotic geospheres naturally defines a fourth geosphere. This is the **biosphere**, a term coined by Vladimir Vernadsky [832] to refer to the totality of living systems and their interconnections, and approached by us as a component of Earth's matter and dynamics. The phase of matter in the biosphere is defined not only by its physical state but even more fundamentally by its necessarily non-equilibrium condition. Its chemical constitution draws from a sector of covalently bonded organometallic compounds, which are not produced by abiotic processes.[10] Its chemical process comprises the reactions that these compounds mediate and by means of which they are also produced and maintained. Its characteristic activating energy scales are the barrier- and reaction-free energies typical in reactions that make and break covalent bonds among C, H, O, N, and S atoms and phosphate groups, and dative bonds of O, N, and S to metals. Its characteristic temperature covers the range for liquid water in near-surface terrestrial (including submarine and sub-crustal) environments, $\sim 0 - 120\,^{\circ}\mathrm{C}$.

In attempting to characterize what the biosphere "is," it is important to us to recognize commonalities with the abiotic geospheres, along with all the levels of organization described within biology, but at the same time to recognize that the biosphere is more than any one of these alone. At the outermost level of abstraction, we emphasize that the order

[10] Speaking more carefully: some of the compounds are not produced at all, and others, which are produced at small rates in abiotic processes, are not produced with the selectivity, yields, or functions that they take in the biosphere, by many orders of magnitude of difference.