Learning the Language of Life: Chaotic Exploration and the Emergence of Biological Abstraction

Maxime MOUGET
Independent Researcher, France
Corresponding author: mougetm@gmail.com

Abstract

Understanding the fundamental principles governing living systems remains a central challenge in biology. While advances in synthetic genomics have enabled the construction of minimal cells, a significant fraction of essential genes remains functionally uncharacterized, highlighting persistent gaps in our comprehension of biological organization. In parallel, theoretical models of self-replicating systems, notably those introduced by John von Neumann, provide a formal framework for describing autonomous, information-driven entities.

In this work, we propose a conceptual framework that unifies minimal cell engineering with a guided chaotic exploration of genomic space. We introduce the notion of a "Von Neumann minimal cell" as a biological platform embodying the core components required for self-maintenance and replication: an informational substrate, an execution machinery, and a boundary-defined environment. Within this constrained architecture, we define a "chaotic exploration algorithm" in which large-scale, AI-assisted generation of genomic variants is iteratively evaluated under controlled, non-propagative conditions.

Rather than optimizing for predefined functions, this approach emphasizes exploratory diversity and emergent behavior, enabling the identification of novel genotype–phenotype relationships and potentially uncovering alternative organizational principles of living matter. By coupling stochastic variation with computationally guided selection and reverse inference, the framework aims to transform biological engineering into a process of systematic discovery.

We discuss the theoretical implications of this approach for synthetic biology, artificial life, and the study of universal constraints on living systems, while outlining safety considerations and experimental boundaries. This perspective suggests a shift from incremental modification of existing organisms toward the construction of a programmable, exploratory substrate for life-like systems.

1. Introduction

Despite major advances in molecular biology and genomics, the functional interpretation of DNA remains fundamentally incomplete. While it is well established that genomic sequences encode the instructions necessary for cellular behavior, these instructions are not accessible in a structured or semantically interpretable form. In practice, biological systems are still manipulated through empirical modification and iterative experimentation, rather than through a principled understanding of their internal logic.

This situation contrasts sharply with other domains of complex systems engineering, particularly computer science, where multiple layers of abstraction separate low-level code from high-level functional design. Concepts such as iteration, recursion, control flow, and modularity enable the transformation of raw instructions into intelligible and manipulable structures. These abstractions are not inherent to the underlying machine code, but emerge through systematic analysis, formalization, and language construction.

By comparison, DNA can be seen as a form of low-level code shaped by evolutionary processes rather than by explicit design. Although its execution produces highly organized and robust systems, the absence of a clear functional abstraction layer prevents a direct mapping between genomic structure and system-level behavior. As a result, the engineering of biological systems lacks the equivalent of a “programming language” that would allow rational design and predictable modification.

This raises a fundamental question:
can the principles of abstraction and language construction, successfully applied in computer science, be extended to biological systems?

In this work, we propose that such a transition requires a shift in methodology. Rather than attempting to directly interpret genomic sequences, we introduce a framework in which functional patterns are allowed to emerge through large-scale, structured exploration of genomic space. This approach is grounded in two key ideas: the definition of a minimal, programmable cellular substrate inspired by von Neumann architectures, and the use of a guided chaotic exploration process to generate and analyze diverse genomic configurations.

The objective is not merely to produce functional biological systems, but to identify recurring organizational patterns that may constitute the basis of a functional language of biology. By observing how variations in genomic structure affect system behavior, it becomes possible to infer higher-level constructs analogous to those found in programming languages, enabling a progressive transition from low-level manipulation to design-oriented bioengineering.

In this perspective, biological research evolves from the study of evolved systems to the systematic construction of an interpretable and potentially programmable representation of living processes. We outline a minimal experimental roadmap to ground this framework in testable conditions.

2. Conceptual Framework: The Cell as a Programmable Machine

2.1. Biological Systems as Integrated Machines

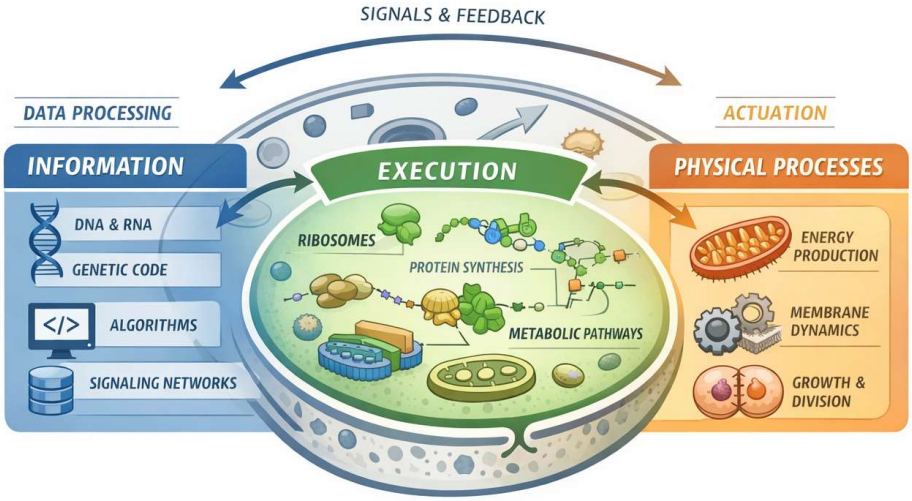


Figure 3 — *Conceptual representation of the cell as an integrated programmable system combining information, execution, and physical processes.*

Living cells can be interpreted as integrated systems combining information processing, energy transformation, material production, and environmental interaction. In this sense, a cell is not only an information-driven entity but also a **self-constructing and self-maintaining machine**. This analogy is not literal but heuristic.

The genomic sequence does not merely encode structural components; it defines the operational logic governing multiple interdependent processes, including:

- replication of the system
- construction and maintenance of internal structures
- energy acquisition and transformation
- synthesis of functional molecules
- regulation of interactions with the environment

This multi-layered functionality suggests that DNA operates as a **global specification**, encompassing both the structure and behavior of the system.

2.2. Parallel with Von Neumann Architectures

This perspective aligns with the theoretical model of self-replicating automata introduced by John von Neumann, in which a system contains:

- a symbolic description of itself
- a mechanism for interpreting that description
- a process for constructing a copy of the system

In biological systems:

- DNA corresponds to the symbolic description
- transcription and translation machinery act as interpreters
- cellular processes enable both construction and replication

However, biological systems extend beyond classical von Neumann architectures by integrating metabolism and environmental coupling as intrinsic components of their operation. This suggests that cells can be viewed as **extended computational systems**, in which execution is inseparable from physical embodiment.

2.3. DNA as a Unified Low-Level Specification

Unlike engineered systems, where software, hardware, and energy systems are often modularized, biological systems rely on a **single encoded layer**—the genome—to define:

- the construction of the system (structural components)
- its operational logic (regulation and control)
- its energy management (metabolic pathways)
- its production capabilities (biosynthesis)
- and its interaction with the environment

In this sense, DNA can be interpreted as a **low-level code that simultaneously defines architecture, execution, and resource management**, without explicit separation of concerns.

This absence of modular abstraction contributes to the difficulty of interpreting genomic sequences in functional terms.

2.4. Convergence of Biological and Engineered Systems

Modern engineering systems—particularly in computing and industrial design—are structured around principles such as:

- modularity
- control flow
- feedback regulation
- hierarchical abstraction
- resource management

These principles are also observed in biological systems, albeit in implicit and distributed forms.

Rather than implying direct imitation, this convergence suggests that **similar organizational constraints may lead to the emergence of analogous functional patterns** in both natural and engineered systems.

In particular, concepts such as:

- signal encoding and transformation
- feedback loops
- conditional activation
- iterative processes

appear in both domains, indicating the presence of underlying structural regularities.

2.5. Toward a Functional Interpretation of Genomic Systems

The absence of an explicit abstraction layer in biological systems prevents a direct mapping between genomic sequences and high-level functions. While DNA encodes the full operational logic of the cell, this logic remains distributed, entangled, and difficult to interpret.

The central hypothesis of this work is that these functional patterns do exist and can be **inferred through systematic exploration**, rather than directly decoded.

By analyzing large-scale variations of genomic configurations within a controlled framework, it becomes possible to identify recurring structures that may correspond to:

- functional modules
- control mechanisms
- interaction patterns
- higher-level organizational principles

Such patterns could constitute the basis of a **functional language of biology**, enabling a transition from low-level sequence manipulation to structured understanding and design.

3. Chaotic Exploration as a Pathway to Functional Understanding

Here, 'chaotic' refers to high-diversity stochastic exploration rather than strict mathematical chaos.

3.1. From Opaque Code to Observable Behavior

A central difficulty in molecular biology lies in the fact that genomic sequences encode functional behavior in a form that is not directly interpretable. While the mapping between genotype and phenotype is known to exist, it remains highly distributed, non-linear, and context-dependent.

As a result, DNA can be viewed as an **opaque low-level code**, whose execution produces observable system-level behaviors, but whose internal logic cannot be readily decomposed into meaningful functional units.

This limitation suggests that direct interpretation of genomic sequences may not be the most effective pathway toward understanding biological systems. Instead, an alternative approach consists in treating the genome as an executable system and focusing on the **relationship between variations in code and variations in behavior**.

3.2. Minimal Cellular Systems as a Controlled Substrate

To enable systematic analysis, the framework relies on **minimal unicellular systems** as a starting point. These systems provide:

- reduced complexity

- limited functional redundancy
- clearer mapping between genetic variation and observable effects

By constraining the biological substrate, it becomes possible to observe the impact of genomic variations with greater resolution and interpretability.

In this context, the minimal cell acts as a **controlled execution environment**, analogous to a simplified machine in which the effects of low-level instructions can be more directly observed.

3.3. Chaotic Exploration of Genomic Space

Within this controlled substrate, we introduce a process of **chaotic exploration**, in which large numbers of genomic variations are generated and evaluated.

This exploration is characterized by:

- high diversity of generated sequences
- absence of predefined functional targets
- iterative refinement guided by observed outcomes

The role of chaos is not to produce random configurations, but to ensure that exploration is sufficiently broad to reveal **non-obvious functional relationships**.

By repeatedly perturbing the genomic structure and observing the resulting system behavior, the framework generates a large dataset of **input–output mappings**, linking sequence variations to functional signatures.

3.4. Learning Functional Relationships through AI

The accumulation of these mappings enables the use of machine learning systems to infer relationships between genomic structure and system-level behavior.

Rather than attempting to decode DNA directly, the approach allows functional understanding to emerge through:

- pattern recognition across multiple observations
- identification of recurring structural motifs
- detection of dependencies and interactions
- construction of approximate models linking genotype to phenotype

This process progressively transforms raw genomic data into **structured functional knowledge**.

3.5. Emergence of Functional Abstractions

As exploration and learning progress, higher-level regularities begin to emerge. These regularities may correspond to:

- functional modules
- control mechanisms
- interaction patterns
- reusable organizational structures

Such elements can be interpreted as the building blocks of a **functional abstraction layer**, analogous to the constructs used in programming languages.

At this stage, the system transitions from:

- observing isolated effects
  to:
- identifying coherent structures within the biological system

This marks the beginning of a **semantic interpretation of genomic information**.

3.6. Toward a Global Understanding of Biological Systems

The iterative combination of chaotic exploration and functional inference ultimately enables the construction of a more global representation of biological organization.

Rather than viewing the genome as an indivisible sequence, it becomes possible to:

- decompose it into functional components
- understand interactions between these components
- identify constraints governing system behavior
- infer general principles of organization

This progression leads to a shift from low-level manipulation to **conceptual understanding**, in which biological systems can be described in terms of structured, interpretable models.

3.7. From Exploration to Design

The ultimate outcome of this process is the transition from empirical experimentation to **design-oriented bioengineering**.

Once functional abstractions are identified and validated, they can be:

- recombined
- modified
- optimized
- and deployed in a controlled manner

In this perspective, biological systems are no longer treated as opaque entities to be adjusted through trial and error, but as **systems whose logic can be progressively understood, structured, and engineered**.

4. Implications: From Genomic Complexity to Design Abstraction

4.1. The Limits of Direct Genomic Manipulation

Genomic sequences contain an extraordinary density of information, even in minimal unicellular organisms. The combinatorial complexity of nucleotide arrangements, coupled with the non-linear interactions governing gene expression and cellular behavior, renders direct manipulation of DNA inherently difficult to interpret and control.

In practice, current approaches rely on localized modifications, guided by partial knowledge of specific genes or pathways. However, this strategy does not scale to a comprehensive understanding of the system, as the genome cannot be effectively reasoned about as a flat sequence of instructions.

This situation is analogous to attempting to design complex software systems directly at the level of machine code, without access to intermediate representations or structured abstractions.

4.2. Necessity of Abstraction Layers

In computer science, the management of complexity is achieved through the introduction of multiple abstraction layers, separating low-level execution from high-level design. These layers enable:

- reduction of cognitive load
- modular reasoning
- reuse of functional components
- predictable system behavior

By contrast, biological systems lack an explicit hierarchy of abstraction accessible to human interpretation. The genome simultaneously encodes structure, control logic, and resource management, without a clear separation of concerns.

The framework proposed in this work suggests that such abstraction layers are not absent, but rather **implicit and unformalized**, and can be progressively uncovered through systematic exploration and analysis.

4.3. From Functional Patterns to Biological Representations

As chaotic exploration and AI-driven inference reveal recurring functional structures, it becomes possible to construct intermediate representations of biological systems.

These representations may take forms analogous to:

- functional modules
- regulatory motifs
- interaction networks
- hierarchical process descriptions

Such structures provide a bridge between raw genomic sequences and system-level understanding, enabling a transition toward **interpretable representations of biological function**.

At this stage, the genome is no longer treated as a monolithic sequence, but as a composition of identifiable and manipulable elements.

4.4. Toward a Biological Compilation Framework

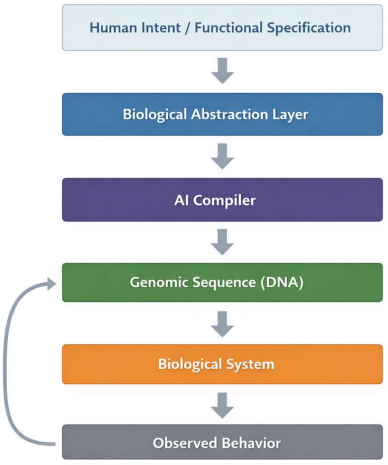


Figure 2 — Biological Compilation Framework

The emergence of functional abstraction naturally leads to the concept of a **biological compilation process**. This represents a long-term perspective rather than an immediate application.

In this paradigm:

- high-level functional descriptions define desired behaviors
- intermediate representations encode system organization
- low-level genomic sequences implement these specifications

An AI-assisted system could act as a **compiler**, translating structured biological descriptions into viable genomic configurations, while ensuring consistency with physical and biochemical constraints.

Such a system would not operate through direct encoding of DNA sequences by humans, but through a process of **guided translation**, in which:

- human-readable specifications
- learned functional patterns
- and constraint-aware generation

are combined to produce executable biological systems.

4.5. From Understanding to Design

The ultimate implication of this framework is a transition from descriptive biology to **design-oriented bioengineering**.

Once sufficient abstraction has been achieved, it becomes conceivable to:

- specify biological systems at a functional level
- compose systems from reusable elements
- predict system behavior prior to implementation
- and iteratively refine designs through compilation and testing

In this context, the role of the human shifts from low-level manipulation to high-level specification, while AI systems handle the translation into operational genomic structures.

4.6. Cognitive Constraints and the Need for Abstraction

The necessity of such a framework is not only technical, but also cognitive. The scale and complexity of genomic information exceed the limits of direct human comprehension.

Without abstraction, biological systems remain:

- opaque
- difficult to reason about
- and resistant to systematic design

By introducing structured representations and hierarchical organization, it becomes possible to bring biological complexity within the scope of human understanding, enabling meaningful interaction with systems that would otherwise remain inaccessible.

5. Limitations and Ethical Considerations

The approach proposed in this work faces two major categories of barriers: ethical considerations and technical feasibility. Research involving living systems inevitably raises concerns regarding the limits of human intervention in biological processes. However, it is important to clarify that the framework described here does not aim at organism-level

modification or eugenic applications, but rather focuses on fundamental cellular systems within controlled and minimal contexts. The objective is to understand the underlying principles of biological organization, not to direct or optimize complex organisms.

In parallel, significant technical challenges remain. The ability to encode arbitrary genomic sequences and reliably produce viable cellular systems from them is still limited. While advances in synthetic genomics have demonstrated the feasibility of constructing minimal genomes, the full design and instantiation of functional living systems from first principles remain beyond current capabilities. As such, the proposed framework should be understood as a progressive research direction, requiring incremental advances in both experimental platforms and computational modeling.

5.1. Experimental and Computational Constraints

The exploration of high-dimensional genomic spaces remains fundamentally limited by combinatorial complexity. Even within a constrained minimal cell framework, the number of possible configurations vastly exceeds what can be physically instantiated or evaluated.

As a result, the proposed framework relies critically on:

- efficient generative models
- informed sampling strategies
- scalable evaluation environments

Despite these tools, exploration will necessarily remain partial, and the inferred structure of the biological design space may be biased by the chosen constraints and evaluation criteria.

5.2. Model Dependency and Interpretation Limits

The process of functional inference is inherently dependent on the models used to interpret observed behaviors. Machine learning systems, while powerful, may introduce:

- approximation errors
- spurious correlations
- incomplete representations of underlying mechanisms

Consequently, the relationships inferred between genomic configurations and functional outcomes should be interpreted as **hypothesis-generating** rather than definitive explanations.

5.3. Biosafety and Containment Principles

The generation and evaluation of novel genomic configurations raise important biosafety considerations. To mitigate risks, the framework is conceptually restricted to **non-propagative and controlled environments**, such as:

- cell-free expression systems
- non-replicative constructs
- physically contained experimental platforms

These constraints ensure that generated systems cannot:

- replicate autonomously outside controlled conditions
- interact unpredictably with natural ecosystems

The emphasis on minimal, constrained substrates further reduces the likelihood of unintended biological complexity.

5.4. Ethical Scope and Responsible Exploration

The proposed framework aims at advancing fundamental understanding of biological systems rather than creating fully autonomous or uncontrolled forms of life. As such, its scope is aligned with existing principles of responsible research in synthetic biology.

Key ethical considerations include:

- maintaining strict containment of experimental systems
- avoiding the generation of harmful or pathogenic configurations
- ensuring transparency and traceability of generated sequences
- aligning research practices with established regulatory guidelines

This approach positions chaotic exploration not as an unrestricted manipulation of life, but as a **controlled scientific methodology for probing its underlying principles**.

5.5. Conceptual Boundaries

While the framework expands the space of biological exploration beyond traditional evolutionary pathways, it does not imply a complete understanding or mastery of living

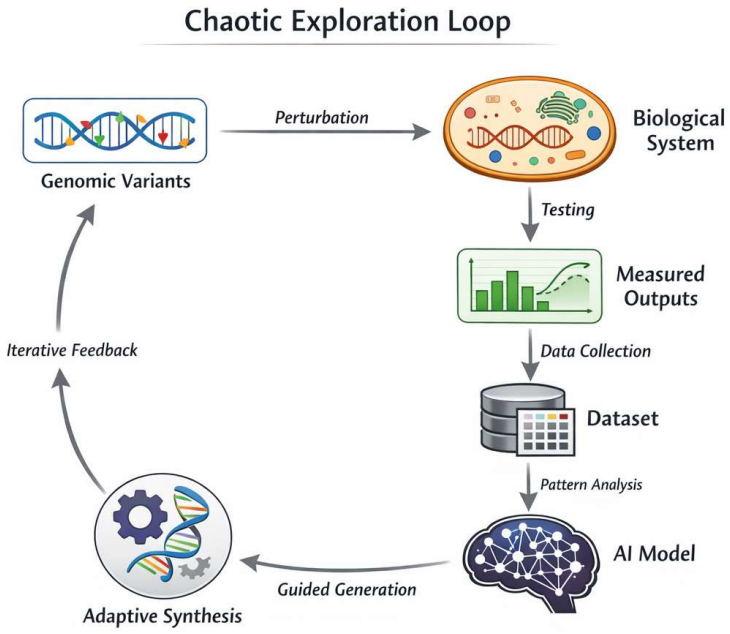
systems. The complexity of biological organization remains deeply rooted in multi-scale interactions that may not be fully captured within simplified experimental platforms.

Accordingly, the proposed approach should be viewed as a **complementary tool**, contributing to the progressive refinement of biological knowledge rather than a definitive solution to the problem of life's complexity.

6. Minimal Experimental Roadmap

Minimal Experimental Roadmap

To transition from a conceptual framework to a testable scientific approach, it is necessary to define a minimal experimental roadmap enabling the controlled exploration of genomic–functional relationships. The objective of this roadmap is not to achieve full biological design, but to establish a reproducible pipeline linking genomic variation, observable behavior, and computational inference.



6.1. Choice of Biological Substrate

The experimental framework relies on simplified and controlled biological systems, selected to balance robustness and interpretability. Two complementary classes of systems can be considered:

- robust model organisms (e.g., *Escherichia coli*), enabling large-scale perturbation and high-throughput evaluation
- minimal or reduced cellular systems (e.g., synthetic or genome-reduced bacteria), offering lower complexity and improved interpretability

Robust organisms provide tolerance to perturbations and support extensive exploration of genomic space, while minimal systems reduce redundancy and facilitate the interpretation of genotype–phenotype relationships.

The use of non-propagative or partially constrained systems is preferred when possible, in order to ensure safety and reproducibility.

6.2. Dual-Scale Exploration Strategy

A dual-scale strategy can be employed to reconcile exploratory breadth with mechanistic clarity:

- **exploratory phase**: large-scale perturbation of robust systems, enabling statistical mapping of genomic variations to functional outcomes
- **analytical phase**: targeted investigation of identified patterns within simplified or minimal systems, enabling refined interpretation

This separation allows high-diversity exploration to be conducted in resilient biological substrates, while preserving the ability to extract interpretable structure in reduced systems.

6.3. Generation of Genomic Variants

Genomic variation is introduced through structured perturbations of a baseline genome. These may include:

- random mutations (substitutions, insertions, deletions)
- recombination or rearrangement of genomic segments
- AI-guided sequence generation under biological constraints

Here, “chaotic” exploration refers to high-diversity stochastic sampling of genomic space, rather than strict mathematical chaos. The objective is not to optimize predefined functions, but to ensure sufficiently broad coverage to reveal non-obvious functional relationships.

To maintain tractability, variation can be constrained by:

- preserving essential core functions
- limiting mutation density per iteration
- restricting exploration to specific genomic regions

6.4. Observable Functional Outputs

Each generated variant must be evaluated through measurable system-level outputs. Depending on the experimental platform, these may include:

- growth dynamics or replication capacity
- metabolic activity
- gene expression profiles
- stress response signatures
- viability or stability metrics

These outputs define a mapping between genomic configuration and functional behavior, forming the empirical basis for subsequent analysis.

As an illustrative example, an initial implementation could focus on perturbations within a limited metabolic or regulatory pathway in a bacterial system, enabling localized genotype–phenotype mapping within a controlled functional scope.

6.5. High-Throughput Evaluation and Data Acquisition

Given the combinatorial nature of genomic space, the framework requires scalable evaluation methods, such as:

- automated culturing platforms
- microfluidic systems
- multiplexed sequencing and phenotyping

The objective is to generate large datasets linking:

- genomic variants (input space)
- functional signatures (output space)

This dataset constitutes the empirical foundation of the exploration process, even though coverage of the full genomic space remains inherently partial.

6.6. AI-Assisted Functional Inference

Machine learning models are used to infer relationships between genomic structure and observed behavior. This includes:

- identification of recurring sequence patterns
- detection of dependencies between genomic regions
- clustering of functional behaviors
- construction of predictive genotype–phenotype models

Rather than providing exact mechanistic explanations, these models aim to:

- approximate functional relationships
- guide further exploration
- highlight candidate structures for deeper analysis

6.7. Iterative Exploration Loop

The experimental process follows an iterative closed-loop structure:

1. generate genomic variants
2. evaluate functional outputs
3. update computational models
4. guide subsequent variant generation

This loop progressively refines both the explored region of genomic space and the inferred functional structure of the system, enabling convergence toward more informative regions of the design space.

6.8. Expected Outcomes

At early stages, the framework is expected to produce:

- coarse mappings between genomic perturbations and system behavior

- identification of sensitive regions and functional hotspots
- emergence of statistical regularities in genotype–phenotype relationships

At later stages, it may enable:

- identification of functional modules
- partial abstraction of biological processes
- construction of intermediate representations of genomic logic

---

6.9. Scope and Limitations of the Roadmap

This roadmap does not aim to immediately achieve full biological design or complete understanding of genomic systems. Instead, it provides:

- a minimal, testable instantiation of the proposed framework
- a scalable methodology for progressive exploration
- a bridge between conceptual theory and experimental practice

The approach remains constrained by:

- experimental throughput limitations
- partial observability of biological systems
- model approximation errors

However, even partial results can contribute to the progressive emergence of functional abstractions and the development of higher-level representations of biological systems.

7. Conclusion

Understanding biological systems requires more than incremental advances in molecular analysis or isolated functional discoveries. The fundamental limitation lies in the absence of a structured, interpretable representation of genomic information. DNA, as it stands, remains a low-level encoding of biological processes—rich in information, yet largely inaccessible in terms of functional semantics.

In this work, we have argued that this limitation cannot be overcome through direct interpretation alone. Instead, it requires a shift in methodology: from reading genomic

sequences to systematically exploring their functional space. By combining minimal cellular systems with chaotic exploration and AI-driven inference, it becomes possible to progressively uncover the patterns and structures that underlie biological organization.

This approach reframes biology as a problem of language and abstraction. Just as complex computational systems became tractable through the introduction of layered representations, the understanding and engineering of living systems depend on the emergence of a functional abstraction layer bridging genomic sequences and system-level behavior.

However, the scale of this challenge should not be underestimated. The combinatorial complexity of genomic space, the non-linearity of biological interactions, and the limits of human cognition make this problem fundamentally different from traditional engineering tasks. It is not a matter of incremental optimization, but of constructing entirely new frameworks for representation, exploration, and interpretation.

As such, progress in this direction will require resources commensurate with the scope of the problem. This includes:

- large-scale experimental and computational infrastructures
- integration of artificial intelligence with biological experimentation
- development of standardized minimal systems for controlled exploration
- and interdisciplinary collaboration across biology, computer science, and systems theory

The objective is not merely to advance biological knowledge, but to establish the foundations of a new paradigm in which living systems can be understood, described, and ultimately designed through structured and interpretable representations.

In this perspective, the learning of the “language of life” is not a metaphor, but a concrete scientific goal—one that requires coordinated effort, conceptual clarity, and sustained investment. Achieving this transition would mark a fundamental step toward transforming biology from an empirical science into a design discipline.

References

- **von Neumann, J.** *Theory of Self-Reproducing Automata* (University of Illinois Press, 1966).
- **Hutchison, C. A. III et al.** "Design and synthesis of a minimal bacterial genome." *Science* (2016).
- **Schreier, H. I. et al.** "Exploratory adaptation in large random networks." *Nature Communications* (2017).
- **Dewachter, L. et al.** "Deep mutational scanning of essential bacterial proteins…" *Nature Communications* (2023).
- **Hwang, Y. et al.** "Genomic language model predicts protein co-regulation…" *Nature Communications* (2024).
- **Watson, J. L. et al.** "De novo design of protein structure and function with RFdiffusion." *Nature* (2023).
- **Castle, S. D. et al.** "Engineering is evolution: a perspective on design across scale and causality." *Nature Communications* (2024).
- **Faure, A. J. et al.** "The genetic architecture of protein stability." *Nature* (2024).
- **Cagiada, M. et al.** "Discovering functionally important sites in proteins." *Nature Communications* (2023).
- **Zhang, Q. et al.** "Integrating protein language models and automatic evolution…" *Nature Communications* (2025).