Structural Limits to Proto-Life Emergence: Evidence from Maximum-Entropy Reinforcement Learning

Fayzeh Abdulkareem Jaber[1], Saeed Serwan Abdulsattar[2], Hani Al-Balasmeh[3]
[1]Department of Computer Science, University of Technology Bahrain, Salmabad, Kingdom of Bahrain
[2]Department of Electrical and Electronics Engineering, University of Bahrain, Sakhir, Kingdom of Bahrain
[3]Department of Informatics Engineering, College of Engineering, University of Technology Bahrain, Salmabad, Kingdom of Bahrain
E-mail: fajaber@utb.edu.bh, saeedserwana@gmail.com, h.albalasmeh@utb.edu.bh

Keywords: Artificial life, origin of life, reinforcement learning, maximum entropy, soft actor–critic

Received: February 9, 2026

This study investigates whether failures of proto-life emergence are primarily due to limited optimization or to insufficient model expressivity. We formulate an abstract protocell-inspired environment as a continuous-control Markov decision process and train a Soft Actor–Critic (SAC) agent under a maximum-entropy objective to act as a strong exploration-preserving optimizer. The environment includes protein-like concentration dynamics, membrane integrity, abstract functional role coverage, and division-like events, while expressivity is varied through three regimes controlling coupling strength, feedback structure, and internal memory. Performance is evaluated using survival rate, success rate, mean functional role coverage, and mean number of division events. Across regimes, survival reached 100%, but success increased from 67.5% in the low-expressivity regime to 97.5% in the mid regime and 100% in the high regime. Mean division count rose from 7.58 and 9.05 in the low and mid regimes to 34.58 in the high regime, indicating a qualitative transition toward growth-dominated behavior once stronger feedback and memory were available. These results show that strong entropy-regularized optimization alone does not guarantee life-like organization when essential state variables and interaction mechanisms are absent. Rather, the emergence of sustained proto-biological behavior is bounded by the representational structure of the model itself. The findings position maximum-entropy reinforcement learning as a diagnostic tool for identifying missing structural ingredients in computational origin-of-life models.

Povzetek: Študija pokaže, da je pojav protoživljenjskega vedenja bolj omejen z izraznostjo modela kot z zmogljivostjo optimizacije.

1 Introduction

The origin of life remains one of the most profound unresolved problems at the intersection of physics, chemistry, biology, and information theory. A dominant conceptual framework views abiogenesis as a stochastic exploration of an astronomically large combinatorial space of molecular configurations, reaction networks, and organizational motifs, from which only a vanishingly small subset gives rise to self-sustaining systems capable of persistence and reproduction [1, 3]. From this perspective, life-like organization is a rare outcome of molecular interactions constrained by thermodynamics, environmental conditions, and limited resources.

A central challenge within this framework is determining whether optimization alone, which is broadly construed as selection, feedback, or adaptive control, is capable of reliably navigating such a combinatorial landscape to assemble the minimal functional organization required for sustained proto-biological behavior. While Darwinian evolution provides an existence proof that adaptive optimization can shape complex biological systems once replication and inheritance are established, it remains unclear whether optimization mechanisms are sufficient prior to the emergence of fully formed replication machinery [4]. This raises a critical question that has remained largely unresolved: is the difficulty of life's emergence primarily due to insufficient guidance and selection pressure, or does it reflect the absence of essential structural variables and feedback mechanisms in the underlying system itself?

Recent advances in reinforcement learning (RL) provide a powerful computational lens through which to revisit this question. Modern RL algorithms operate as effective optimizers in complex, stochastic, and high-dimensional environments, leveraging continuous feedback, long-horizon credit assignment, and adaptive exploration strategies [6]. Among these approaches, maximum-entropy reinforcement learning has emerged as a principled framework that explicitly balances reward maximization with sustained exploration by augmenting the optimization objective with an entropy regularization term [8, 7]. From a formal optimization perspective, these methods provide a strong practical

exploration-and-control regime. They are designed to reduce premature convergence and maintain exploration over a rich space of trajectories, but no guarantee of global optimality is assumed.

In this work, we treat maximum-entropy reinforcement learning not as a model of biological evolution, but as a computational stress test for emergence. Specifically, we ask whether a widely used entropy-regularized RL agent, Soft Actor–Critic (SAC), can reliably induce life-like organization in a controlled proto-system when granted continuous feedback, adaptive control, and an explicit incentive for exploration. If such an optimizer repeatedly fails to assemble and maintain proto-biological organization despite stable learning and sustained exploration, the result provides evidence consistent with limitations in the expressive capacity of the modeled system, although it does not establish global impossibility. To investigate this question, we construct a protocell-inspired simulation environment that deliberately abstracts away biochemical detail while retaining structural motifs widely discussed in origin-of-life research such as resource influx, degradation, catalytic amplification, compartmentalization, and division-like events [9, 10]. The environment consists of a population of protein-like components whose concentrations evolve under stochastic dynamics, catalytic interactions, and environmental perturbations. Rather than modeling explicit chemical reactions, we introduce *functional-role proxies*: abstract representations of stability, interaction potential, and functional coverage that collectively determine whether the system exhibits persistence, self-maintenance, and growth with division under noise.

Our SAC agent interacts with this environment by controlling both environmental and protocol-level variables. Actions correspond to abstracted interventions such as feeding rate, dilution strength, membrane permeability, and feedback amplification. The agent is trained under a maximum-entropy objective that encourages broad exploration of system trajectories while optimizing a long-horizon reward encoding biological sufficiency: sustained viability balanced against parsimony, with explicit penalties for unnecessary complexity. This formulation reflects long-standing arguments that early life must have achieved a delicate balance between functional completeness and minimal organizational overhead [11, 12]. Crucially, the environment is parameterized by an explicit model expressivity regime, allowing systematic control over the richness of interactions, feedback loops, and memory-like state variables available to the system. By systematically varying coupling strength, feedback structure, and internal memory capacity, we directly examine how the dimensionality and organization of the state space constrain what any optimizer can achieve, regardless of its sophistication. This controlled design holds the optimization method fixed while varying only the expressive capacity of the environment, thereby isolating the structural contribution of coupling, feedback, and memory.

Across a wide range of experimental conditions, we observe regimes in which learning converges and policy entropy collapses despite entropy regularization, yet life-like organization consistently fails to emerge. Only when sufficient interaction richness, feedback mechanisms, and internal state variables are introduced does sustained proto-biological behavior become attainable. These findings indicate that the primary bottleneck in emergence is not merely the absence of guidance or optimization, but the expressive capacity of the modeled biochemical and environmental interactions themselves.

More broadly, this work highlights a fundamental limitation of optimization-based approaches to emergence. By construction, maximum-entropy reinforcement learning can only optimize over variables, state representations, and feedback channels explicitly encoded in the model. If essential variables governing viability are absent or if critical couplings are unrepresented, then no amount of exploration, entropy regularization, or algorithmic sophistication can recover them. In this sense, entropy-regularized RL serves as a stringent diagnostic benchmark for what optimization can achieve within a given model, allowing persistent failure to be interpreted alongside convergence and exploration diagnostics as evidence of potentially missing structure. By reframing maximum-entropy reinforcement learning as a diagnostic instrument rather than a generative explanation, this study offers a new computational perspective on life-like emergence, bridging reinforcement learning theory, artificial life, and origin-of-life modeling.

Research objectives and hypotheses

This study is guided by the following research objectives:

1. To determine whether a strong entropy-regularized reinforcement learning agent can reliably induce sustained proto-life behavior in an abstract protocell-inspired environment.

2. To evaluate how model expressivity, defined through coupling strength, feedback structure, and internal memory, affects the emergence of life-like organization.

3. To distinguish optimization-related failure from structural failure by holding the learning algorithm fixed while varying only the expressive capacity of the environment.

The investigation is organized around the following hypotheses:

– **H1**: Increasing model expressivity improves proto-life success metrics, including functional coverage and division behavior.

– **H2**: Maximum-entropy SAC can achieve stable control across all regimes, but sustained proto-life organization will remain limited in under-expressive environments.

– **H3**: When optimization strength is held fixed, failures of emergence are better explained by missing structural variables and couplings than by deficiencies in exploration or learning.

2 Related work

This section reviews prior research relevant to the emergence of life-like organization, optimization-driven models of biological systems, and the use of reinforcement learning as a control and exploration paradigm in complex dynamical environments. The objective is to situate the present study within existing theoretical and computational frameworks, and to identify unresolved limitations that motivate the proposed approach.

2.1 Origin-of-life models and emergent organization

The emergence of life from non-living matter has long been extensively studied through theoretical, chemical, and computational models that emphasize self-organization under non-equilibrium conditions. Classical frameworks describe abiogenesis as a process governed by thermodynamics, reaction kinetics, and stochastic molecular interactions, in which structured behavior arises once critical thresholds of energy flow and interaction density are exceeded [13, 1, 3].

Autocatalytic network theory formalized this idea by demonstrating that self-sustaining chemical organization can emerge when reaction networks achieve sufficient catalytic closure [14, 15]. Despite their conceptual importance, most origin-of-life models rely on analytically prescribed dynamics or limited parameter exploration. As a result, it remains unclear whether failures to achieve sustained organization reflect fundamental theoretical constraints or insufficient exploration of the available state space [16, 17, 18]. Without an explicit, strong search or control mechanism, it is difficult to determine whether observed failures of emergence reflect fundamental inadequacies in the modeled dynamics or simply the inability of the system to discover viable organizational regimes. Addressing this limitation requires a framework in which optimization power is deliberately maximized and treated as a controlled variable.

2.2 Optimization and selection-based approaches

To overcome the limitations of unguided stochastic models, a broad class of optimization- and selection-based approaches has been applied to biological and biochemical systems. Evolutionary algorithms, genetic programming, and related methods have demonstrated that explicit selection pressures can efficiently identify functional RNA sequences, protein motifs, and simplified metabolic networks under predefined fitness objectives [19, 20]. More recently, large-scale systems biology studies have shown that metabolic innovation and sustained functional organization arise from coordinated network-level expansion and selection, rather than from isolated reactions or local parameter tuning [21]. Collectively, these results indicate that guided optimization can dramatically outperform random exploration in high-dimensional biochemical state spaces.

However, the interpretability of these successes remains limited in a fundamental respect. Most evolutionary and selection-based approaches embed substantial prior guidance through explicitly defined fitness functions, handcrafted mutation operators, and population-level selection dynamics. While effective for engineering purposes, this design choice tightly couples optimization strength to model expressivity. Consequently, successful emergence cannot be unambiguously attributed to the intrinsic organizational capacity of the modeled chemical system, but may instead reflect the degree to which external selection pressure has been engineered to favor particular outcomes. This entanglement complicates causal interpretation, as it becomes difficult to disentangle whether sustained organization arises because it is supported by the system's internal dynamics or because the optimization procedure is sufficiently prescriptive to locate it. These limitations are most pronounced in protocell-like regimes, where viability depends on long-term integration rather than local fitness gradients. Functional organization often emerges only after extended transients, conditions under which classical evolutionary, heuristic, and greedy optimization methods struggle due to weak long-horizon credit assignment, fragile diversity, and premature convergence in rugged landscapes [22]. As a result, negative outcomes remain ambiguous, reflecting either genuine structural insufficiency or inadequate exploration.

This motivates treating optimization as a controlled experimental variable rather than a heuristic. By employing frameworks that support sustained exploration and long-horizon credit assignment with minimal prior bias, failures of emergence become diagnostically meaningful, indicating limitations of the modeled dynamics rather than deficiencies of the search process.

2.3 Reinforcement learning in artificial life

RL offers a principled framework for sequential decision making in stochastic environments, wherein an agent interacts with a system to learn a policy that maximizes long-term cumulative reward. This paradigm has been increasingly applied to artificial life and complex adaptive systems, providing mechanistic insights into feedback, memory, and sustained viability that are difficult to capture with static or purely evolutionary methods [5, 23].

More recent work has explored RL in domains closely related to biochemical and molecular systems. In chemical process control and reaction optimization, RL has been used to learn dynamic control policies that adapt to system disturbances and maximize process-performance met-

rics. Likewise, deep learning and sequential optimization have been applied to inverse molecular design, where high-dimensional chemical spaces are explored to discover and optimize molecular structures [24].

Despite its increasing adoption in artificial life and chemically inspired modeling, RL has, to the best of our knowledge, not been systematically employed as a methodological instrument for evaluating the sufficiency of model representations themselves. Prior work predominantly applies RL as a control or optimization framework within environments whose state variables, interaction structures, and feedback mechanisms are implicitly assumed to be adequate a priori. Consequently, learning outcomes are typically interpreted in terms of policy performance or task-level success, rather than as evidence bearing on the expressive capacity of the underlying dynamical model. Under this prevailing perspective, failures of learning are most often ascribed to algorithmic limitations, suboptimal training procedures, or deficiencies in reward specification, rather than to omissions in state representation or unmodeled coupling mechanisms. This interpretive bias obscures the distinction between genuine structural insufficiency and artifacts of the optimization process, leaving unresolved whether non-emergence reflects intrinsic constraints of the modeled dynamics or limitations of the learning framework itself. The present study is motivated by this gap and adopts reinforcement learning as a controlled, strong optimizer to explicitly diagnose representational insufficiency in proto-life models.

2.4 Maximum-entropy reinforcement learning

Maximum-entropy RL extends conventional reinforcement learning by explicitly incorporating an entropy term into the optimization objective, thereby favoring policies that remain stochastic while achieving high expected return [8]. This stochasticity is not incidental; it encourages persistent behavioral variability, helps reduce premature collapse onto narrowly optimized control strategies, and promotes continued exploration of alternative system trajectories in complex, non-convex environments. From the perspective of origin-of-life modeling, the explicit promotion of randomness is particularly significant. Many theoretical accounts of abiogenesis emphasize that early biochemical organization arises in regimes dominated by stochastic fluctuations, where variation, noise, and random perturbations play a constructive role in enabling novel configurations, reaction pathways, and functional associations to be sampled [3, 2]. In such settings, the suppression of variability through overly deterministic control can artificially constrain the space of attainable dynamics, masking the latent organizational potential of the system. Entropy regularization therefore provides a principled mechanism for sustaining the diversity of system trajectories required to probe whether life-like organization can emerge at all within a given representational framework. SAC constitutes a widely used instantiation of maximum-entropy reinforcement learning, integrating off-policy learning, automatic entropy adjustment, and stable function approximation. By adaptively regulating the entropy coefficient, SAC maintains stochasticity when multiple viable strategies exist and reduces it only when the dynamics favor more structured control. As a result, SAC provides a strong learning-based controller that balances exploration and exploitation over long horizons without requiring manual tuning of noise levels or exploration schedules.

Despite these advantages, maximum-entropy reinforcement learning has seldom been employed to interrogate the adequacy of model representations themselves. Existing applications generally presume that the environment dynamics, state variables, and interaction structures are sufficiently expressive, and interpret learning outcomes primarily in terms of task performance. Consequently, failures are often attributed to algorithmic or training deficiencies rather than to missing state dimensions or unmodeled coupling mechanisms. In contrast, the present study leverages entropy-regularized reinforcement learning as a controlled, exploration-maximizing optimizer to assess whether sustained, life-like organization is achievable within a given model specification, and to distinguish structural limitations from optimization artifacts.

2.5 Summary and research gap

Prior work in origin-of-life research and artificial life modeling has demonstrated that both self-organizing dynamics and optimization-based mechanisms can, under suitable conditions, give rise to complex and adaptive behavior. Models of abiogenesis typically emphasize stochasticity, autocatalysis, and far-from-equilibrium dynamics, whereas optimization and selection-based approaches show that guided search can substantially accelerate the discovery of functional organization within high-dimensional biochemical spaces. Reinforcement learning further expands this methodological landscape by offering a principled framework for adaptive, long-horizon control in stochastic environments.

Despite these advances, a fundamental limitation remains common across these approaches: optimization power and model expressivity are rarely disentangled. In origin-of-life models, failures to achieve sustained organization are often difficult to interpret due to restricted exploration and the absence of explicit, strong control mechanisms. Conversely, evolutionary and learning-based methods typically embed strong prior assumptions through fitness functions, mutation operators, or environment design, obscuring whether successful emergence arises from the intrinsic dynamical capacity of the modeled system or from externally imposed guidance.

In this context, reinforcement learning has primarily been used to optimize behavior in environments whose representational adequacy is implicitly assumed. Outcomes are interpreted in terms of policy quality or task perfor-

mance, while failure is commonly attributed to algorithmic limitations, training instability, or reward misspecification. Consequently, RL has not been systematically used as a diagnostic instrument for determining whether a model can support life-like organization in the first place. This interpretive gap leaves unresolved the foundational question posed in the Introduction: is the difficulty of life's emergence attributable mainly to insufficient guidance and selection pressure, or to missing structural variables, feedback mechanisms, and internal state representations? Without a framework that holds optimization power fixed at a strong level, this distinction remains ambiguous.

The present study addresses this gap by reframing maximum-entropy reinforcement learning as a stringent diagnostic benchmark for optimization capability. By deploying a strong entropy-regularized controller and systematically varying model expressivity while holding the learning algorithm constant, we examine the extent to which sustained proto-biological organization is constrained by representational structure rather than by deficiencies in search or control. In this setting, persistent optimization failure, considered together with convergence and exploration diagnostics, becomes evidence consistent with missing variables, feedback pathways, or internal memory mechanisms. This perspective establishes a principled methodology for evaluating and refining computational models of life-like emergence and for bridging reinforcement learning theory with foundational questions in artificial life and origin-of-life research.

3 Methodology

This section presents the computational framework developed to examine the limits of maximum-entropy optimization in abstract proto-life systems. The problem is formulated as a continuous-control Markov decision process (MDP), within which a protocell-inspired simulation environment is coupled to a fixed, high-capacity reinforcement learning agent. We detail the environment structure and expressivity regimes governing interaction richness and internal state, the system dynamics and reward formulation, and the training and evaluation protocol based on entropy-regularized reinforcement learning. The central objective is to decouple optimization capability from model expressivity and to determine whether life-like organization can arise when learning, exploration, and control are no longer the dominant constraints.

Figure 1 provides an overview of the proposed framework, illustrating the abstract proto-system, the hierarchy of expressivity regimes, the maximum-entropy SAC controller used as a strong optimization benchmark, and the training and evaluation pipeline. Together, these components define a controlled experimental setting designed to probe how the representational capacity of the modeled system bounds what even powerful optimization algorithms can achieve.

3.1 Problem formulation

Proto-life emergence is formalized as a sequential decision-making problem represented by a continuous-control MDP,

$$\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma \rangle, \tag{1}$$

where $\mathcal{S}$ denotes a continuous state space describing an abstract proto-life system, $\mathcal{A}$ is a continuous action space, $\mathcal{P}$ specifies stochastic state-transition dynamics, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is a scalar reward function encoding biological sufficiency, and $\gamma \in (0, 1)$ is a discount factor governing long-horizon optimization.

The objective of this formulation is not to reproduce biochemical realism, but to assess whether a strong learning agent, endowed with adaptive control and explicit exploration incentives, can assemble and sustain life-like organization within a given model specification. Maximum-entropy reinforcement learning is therefore treated as a diagnostic optimization benchmark, allowing failures of emergence to be evaluated in relation to the structural limitations of the modeled system as well as the observed search, exploration, and learning behavior.

3.2 State representation

At each discrete time step t, the environment is characterized by a continuous state vector $s_t \in \mathcal{S}$ defined as

$$s_t = \Big[\mathbf{c}_t,\ m_t,\ \mathbf{r}_t,\ \mathbf{h}_t,\ \mathbf{a}_{t-1}\Big], \tag{2}$$

where $\mathbf{c}_t \in \mathbb{R}_+^N$ denotes the concentration vector of N protein-like components, $m_t \in \mathbb{R}_+$ represents membrane integrity, $\mathbf{r}_t \in [0, 1]^K$ encodes abstract functional role coverage, $\mathbf{h}_t = (e_t, \iota_t)$ denotes internal memory variables (such as stored energy and structural integrity) when enabled under higher expressivity regimes, and $\mathbf{a}_{t-1} \in [0, 1]^4$ corresponds to the control action applied at the previous time step.

Abstract functional roles are computed via a fixed random projection followed by a saturating nonlinearity,

$$\mathbf{r}_t = \mathbf{f}(\mathbf{c}_t) = 1 - \exp\big(-\alpha \mathbf{W} \mathbf{c}_t\big), \tag{3}$$

where $\mathbf{W} \in \mathbb{R}^{K \times N}$ is a randomly initialized role-mapping matrix and $\alpha > 0$ is a fixed saturation coefficient. This construction provides an abstract and distributed proxy for functional coverage, enabling the representation of system-level organization without encoding explicit biochemical reaction mechanisms or domain-specific semantics.

3.3 Action space and system dynamics

The agent interacts with the proto-life environment through a continuous control action vector

$$\mathbf{a}_t = \big[u_t^{\text{feed}},\ u_t^{\text{dil}},\ u_t^{\text{perm}},\ u_t^{\text{fb}}\big] \in [0, 1]^4, \tag{4}$$

where the individual components correspond to the external resource influx rate, dilution or outflow rate, membrane

Table 1: Summary of representative prior work and the gap addressed in this study.

Study Type	Representative Works	Optimization / Control Mechanism	Expressivity Treatment	Typical Outcome	Limitation Relative to This Work
Origin-of-life / protocell models	[1, 9, 14]	None or analytically specified dynamics	Usually fixed by model design	Self-organization under selected conditions	Does not disentangle model expressivity from search limitations
Autocatalytic / self-organization studies	[15, 17, 18]	Emergent dynamics, no adaptive controller	Implicit / fixed	Threshold-based emergence claims	Limited exploration and no explicit strong optimization benchmark
Evolutionary / selection-based methods	[19, 20, 22]	Evolutionary search or selection pressure	Coupled to fitness design	Functional structures discovered under selection	Optimization is entangled with prior bias and hand-crafted fitness
RL in biological / chemical systems	[23, 24]	RL-based sequential control	Assumed adequate a priori	Task-level optimization	Learning success/failure is not used to diagnose representational sufficiency
This work	—	Maximum-entropy SAC	Explicitly varied across regimes	Tests whether emergence remains possible under strong optimization	Separates optimization capability from model expressivity

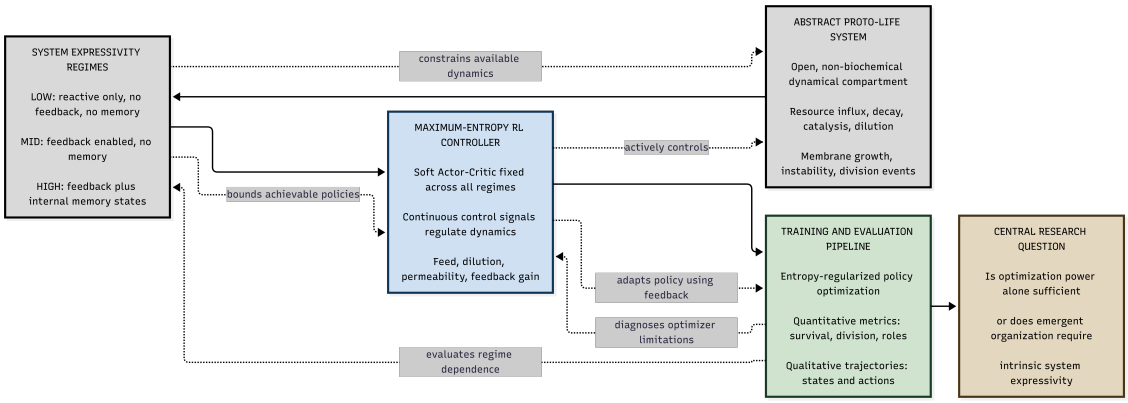

Figure 1: Overview of the computational framework used to disentangle optimization power from system expressivity in abstract proto-life models. The diagram illustrates the proto-life system, expressivity regimes, fixed maximum-entropy reinforcement learning controller, training and evaluation pipeline, and the central research question.

permeability, and feedback amplification strength, respectively. These control variables operate at the protocol level, modulating global environmental and structural parameters rather than directly manipulating individual molecular species.

System dynamics are modeled using discrete-time update equations that combine feed-driven influx, catalysis-mediated amplification, and loss processes. The concentration vector evolves according to

$$\mathbf{c}_{t+1} = \Big(\mathbf{c}_t + \Delta t\, \mathbf{u}_{\text{feed}}\Big) + \Delta t\, \max\big(0, \mathbf{C}^\top \mathbf{c}_t\big)\big(1-\lambda_t\big)\big(1-\delta_t\big), \quad (5)$$

where $\mathbf{C}$ denotes a catalysis matrix coupling a designated subset of catalytic species to the broader population. The effective decay and dilution terms are defined as

$$\lambda_t = \lambda_0 + \beta u_t^{\text{perm}} - \phi u_t^{\text{fb}} \bar{r}_t, \qquad \delta_t = \kappa u_t^{\text{dil}}, \quad (6)$$

with $\bar{r}_t$ denoting the mean functional role coverage at time t.

Membrane integrity evolves according to

$$m_{t+1} = m_t + \eta_1 \bar{c}_t + \eta_2 \bar{r}_t - \eta_3 u_t^{\text{perm}} - \eta_4 u_t^{\text{dil}} + \xi(\mathbf{h}_t), \quad (7)$$

where $\bar{c}_t$ represents the mean concentration of a designated subset of core species and $\xi(\mathbf{h}_t)$ captures stabilization effects arising from internal memory variables when enabled under higher expressivity regimes.

Finally, division-like events are triggered when membrane integrity exceeds a threshold m_{div}, resulting in imperfect inheritance,

$$\mathbf{c}_{t+1} \leftarrow \rho\, \mathbf{c}_{t+1}, \qquad m_{t+1} \leftarrow 0.5\, m_{t+1}, \quad (8)$$

where $\rho \in (0,1)$ controls the degree of concentration loss during division. These events provide a minimal abstraction of growth and replication without introducing explicit biochemical mechanisms.

3.4 Expressivity regimes

To explicitly decouple optimization capability from model richness, the proto-life environment is parameterized by a set of discrete expressivity regimes that control the structural and dynamical complexity available to the system. These regimes systematically regulate the strength of inter-species coupling, the presence of feedback mechanisms, and the availability of internal memory variables,

while holding the learning algorithm fixed. In the low-expressivity regime, interactions are weak, feedback pathways are absent, and no internal memory is available, yielding a minimally structured dynamical system. The mid-expressivity regime introduces moderate coupling and feedback mechanisms, enabling limited adaptive stabilization while still excluding internal memory. Finally, the high-expressivity regime incorporates strong coupling, explicit feedback, and internal memory variables, thereby expanding both the dimensionality of the state space and the range of dynamical behaviors that the system can, in principle, support.

By progressively increasing expressivity in this controlled manner, the framework allows a direct examination of how representational capacity and interaction structure constrain the emergence of sustained, life-like organization, independent of optimization power.

3.5 Reward function

The agent is trained using a scalar reward function that encodes a quantitative notion of *biological sufficiency*, balancing persistence, functional completeness, and growth against unnecessary organizational complexity. At each time step, the reward is defined as

$$r_t = w_p + w_r \bar{r}_t + w_m \log\big(1+m_t\big) + w_g \mathbb{I}_{\text{div}} - \lambda_c \|\mathbf{c}_t\|_0, \quad (9)$$

where $\bar{r}_t$ denotes the mean functional role coverage, m_t represents membrane integrity, and $\mathbb{I}_{\text{div}}$ is an indicator variable signaling the occurrence of a division-like event. The final term penalizes unnecessary complexity via a sparsity-based proxy proportional to the number of active species.

This reward structure incentivizes sustained viability and functional coverage while discouraging unstructured growth or excessive accumulation of components. Episodes are terminated upon membrane collapse or uncontrolled concentration growth, in which case a large negative terminal reward is assigned to explicitly penalize loss of viability.

3.6 Learning algorithm

Learning is performed using the Soft Actor–Critic (SAC) algorithm, an off-policy reinforcement learning method that optimizes a maximum-entropy objective. The optimal policy is defined as

$$\pi^* = \arg\max_{\pi} \; \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t \Big(r_t + \alpha\, \mathcal{H}\big(\pi(\cdot \mid s_t)\big)\Big)\right], \quad (10)$$

where $\mathcal{H}(\pi(\cdot \mid s_t))$ denotes the policy entropy, and α is a temperature parameter that is learned automatically during training.

By explicitly rewarding stochasticity, this formulation promotes persistent exploration and mitigates premature convergence to suboptimal control strategies. In the context of this study, maximum-entropy SAC is adopted as a strong entropy-regularized continuous-control method that provides a stringent benchmark within the tested learning framework for what control can achieve under a given model specification.

To ensure full methodological transparency and facilitate replication of the proposed experiments, the principal simulation parameters, environment configuration, and reinforcement learning training settings are summarized in Table 2. These parameters define the structural properties of the proto-life simulation environment, the expressivity regimes governing interaction richness, and the hyperparameters used for training the Soft Actor–Critic agent.

Table 2: Simulation environment and training configuration used for reproducibility.

Item	**Specification**
Software framework	Python-based Gymnasium-compatible custom environment with Stable-Baselines3 implementation of Soft Actor–Critic (SAC).
State variables	Species concentrations $\mathbf{c}_t$, membrane integrity m_t, functional role coverage $\mathbf{r}_t$, optional memory state $\mathbf{h}_t$, and previous action $\mathbf{a}_{t-1}$.
Action variables	Continuous controls: feed rate, dilution strength, membrane permeability, and feedback amplification.
Number of species N	Fixed across all experiments according to the simulation configuration.
Number of functional roles K	Fixed across all experiments according to the simulation configuration.
Role mapping matrix $\mathbf{W}$	Fixed random non-negative mapping matrix initialized once per seed and used to compute functional role coverage.
Catalysis matrix $\mathbf{C}$	Sparse species-interaction matrix initialized once per seed; catalytic coupling strengths are scaled by expressivity regime.
Low expressivity regime	Weak coupling; feedback disabled; internal memory disabled.
Mid expressivity regime	Moderate coupling; feedback enabled; internal memory disabled.
High expressivity regime	Strong coupling; feedback enabled; internal memory enabled.
Reward formulation	Weighted reward combining persistence, role coverage, membrane integrity, division-like events, and complexity penalty.
Reward weights	Fixed across all regimes to ensure that performance differences arise from model expressivity rather than reward redesign.
Episode horizon	365 simulation steps.
Training steps	Same total number of SAC training steps used for all expressivity regimes.
Evaluation episodes	Same number of deterministic evaluation episodes used for each trained policy and regime.
Random seeds	Independent seeds used for environment initialization, matrix generation, and SAC training.
Actor network	Multilayer perceptron policy network with ReLU activations.
Critic network	Twin Q-network architecture with ReLU activations.
Learning rate	Fixed SAC learning rate used consistently across all regimes.
Batch size	Fixed mini-batch size used for off-policy SAC updates.
Replay buffer size	Fixed replay buffer capacity used during training.
Discount factor γ	Fixed discount factor used for long-horizon return estimation.
Entropy setting	Automatic entropy tuning enabled.
Optimizer	Adam optimizer.
Code availability	The simulation code, configuration files, trained-policy settings, and generated evaluation outputs are available from the corresponding author upon reasonable request.

3.7 Evaluation protocol

Model performance is evaluated using a combination of quantitative metrics and qualitative trajectory analysis. Quantitative evaluation focuses on survival rate, mean number of division-like events, average functional role coverage, and an overall success rate defined by simultaneous satisfaction of viability and functional criteria. These metrics provide complementary views of persistence, growth, and organizational completeness.

A trajectory is classified as successful if it simultaneously satisfies three criteria over the full evaluation horizon: (i) the system survives until the end of the episode without collapse, (ii) the mean functional role coverage exceeds a predefined threshold τ_r, and (iii) at least one division-like event occurs. This operational definition corresponds to the notion of biological sufficiency used throughout the evaluation.

In addition to aggregate statistics, full state–action trajectories are logged throughout training and evaluation. These trajectories enable post hoc inspection of system dynamics, identification of failure modes, and qualitative analysis of emergent behaviors, thereby supporting a deeper interpretation of how expressivity constraints shape the space of attainable outcomes.

All reported evaluation metrics are averaged across multiple independent random seeds to account for stochastic variability in both the simulation dynamics and the reinforcement learning process. Results are reported as mean $\pm$ standard deviation.

3.8 Algorithmic framework and diagnostic interpretation

Algorithm 1 summarizes the complete experimental pipeline implemented in this study. The procedure reflects the exact structure of the single-file prototype, in which a fixed maximum-entropy reinforcement learning agent is trained and evaluated across progressively richer expressivity regimes. Optimization power is held constant by using the same SAC configuration in all cases, while model expressivity is varied exclusively through environment parameters.

Failure of SAC to induce sustained proto-life under a given expressivity regime is interpreted cautiously as evidence of insufficient model expressivity rather than, by itself, as an optimization failure. Maximum-entropy reinforcement learning can optimize only over the state variables, interactions, and feedback channels explicitly encoded in the model. Persistent failure to produce life-like organization therefore serves as a diagnostic signal of missing state dimensions or unmodeled coupling mechanisms within this experimental framework.

Algorithm 1 Maximum-Entropy Optimization as a Diagnostic Test for Proto-Life Emergence

Require: Expressivity regimes $\mathcal{R} = \{\text{low}, \text{mid}, \text{high}\}$, training horizon T, evaluation episodes N

for each regime $r \in \mathcal{R}$ **do**
  Initialize configuration `ProtoCfg` with `regime` $= r$
  Instantiate Gymnasium environment $\mathcal{E}_r$ with episode monitoring
  Wrap $\mathcal{E}_r$ in a single-instance vectorized interface
  Initialize SAC agent with automatic entropy tuning
  Initialize CSV logger for episode-level training statistics
  for $t = 1$ to T **do**
    Observe state s_t
    Sample stochastic action $a_t \sim \pi_\theta(\cdot \mid s_t)$
    Execute a_t in $\mathcal{E}_r$ and observe (r_t, s_{t+1})
    Update actor, critics, and entropy temperature
    Log episode reward, episode length, and entropy coefficient upon termination
  end for
  Save trained SAC policy for regime r
  for evaluation episode $i = 1$ to N **do**
    Reset $\mathcal{E}_r$ using a deterministic seed
    while episode not terminated **do**
      Select deterministic action $a_t = \arg\max_a \pi_\theta(a \mid s_t)$
      Execute a_t and record system variables
    end while
    Compute survival status, functional role coverage, and division count
    Mark episode as successful if viability and functional criteria are satisfied
  end for
  Log full state–action trajectories for qualitative analysis
  Aggregate success rate, survival rate, mean role coverage, and mean divisions
end for

4 Results

This section presents the quantitative and qualitative outcomes of applying maximum-entropy reinforcement learning to abstract proto-life systems under increasing levels of model expressivity. The central objective is not to demonstrate the raw effectiveness of reinforcement learning, but to evaluate whether the emergence of sustained, life-like organization is primarily limited by optimization power or by the expressive capacity of the underlying model once a strong optimizer is employed.

4.1 Evaluation performance across expressivity regimes

Table 3 reports aggregate evaluation metrics obtained after training under low, mid, and high expressivity regimes. Metrics are averaged over independent evaluation episodes executed with deterministic policies.

Across all regimes, the survival rate reaches 100%, indicating that basic persistence can be achieved even in minimally expressive systems when guided by a strong optimizer. However, the overall success rate, defined by simultaneous survival, sustained functional role coverage, and successful division, exhibits a strong and monotonic dependence on expressivity. Success increases from 67.5% in the low regime to 97.5% in the mid regime, and reaches 100% under high expressivity.

Notably, mean functional role coverage peaks in the mid expressivity regime, whereas the high expressivity regime exhibits a dramatic increase in division frequency. The latter achieves more than four times the number of division events observed in lower regimes, indicating a qualitative shift toward rapid growth and reproduction once sufficient feedback and internal memory mechanisms are available. To complement the aggregate metrics in Table 3, Figure 2 visualizes representative diagnostic snapshots across regimes. While survival is achieved in all cases, the internal organization differs qualitatively: low expressivity exhibits diffuse, weakly coherent internal structure, mid expressivity shows partial filament alignment, and high expressivity yields stabilized compartmentalization with a persistent core consistent with enabled feedback and memory channels.

4.2 Training reward dynamics

Figure 3 shows the evolution of episode reward as a function of training timesteps across all expressivity regimes.

All regimes exhibit rapid initial reward growth, confirming that the SAC agent successfully identifies viable control strategies early in training. The low and mid expressivity regimes converge to higher asymptotic rewards, while the high expressivity regime converges more slowly and stabilizes at a lower reward level.

Importantly, this divergence does not indicate an optimization failure. Rather, it reflects a structural shift in how reward is allocated once additional feedback channels and internal memory mechanisms become available. In the high expressivity regime, the agent exploits these mechanisms to induce frequent division events, prioritizing accelerated growth and reproduction over strict optimization of functional balance. This strategy incurs higher complexity penalties under the reward function, resulting in lower aggregate reward despite perfect survival and success.

Under low expressivity, optimization converges but fails to support sustained life-like organization. At intermediate expressivity, the optimizer discovers a balanced proto-life regime characterized by functional stability and limited reproduction. At high expressivity, additional feedback and memory enable growth-dominated strategies, marked by frequent division events, increased state variability, and higher complexity penalties, which reduce aggregate reward despite increased biological activity.

4.3 Episode length stabilization

Figure 4 reports the evolution of episode length during training.

Across all regimes, episode length rapidly converges to the maximum horizon, indicating stable self-maintenance once viable control policies are learned. However, transient collapses are observed in the mid and high expressivity regimes during training, reflecting exploratory excursions into unstable regions of the state space. These failures disappear as training progresses, confirming robust convergence.

4.4 Entropy adaptation and exploration

Figure 5 illustrates the evolution of the entropy coefficient during training.

In all regimes, the entropy coefficient decreases sharply during early training, indicating a transition from broad exploration to increasingly structured policies. Importantly, entropy does not collapse to zero. The high expressivity regime maintains a higher long-term entropy level, reflecting persistent policy stochasticity and the existence of multiple viable control strategies in a richer state space.

This behavior is consistent with the maximum-entropy formulation and supports the interpretation of SAC as a strong optimization benchmark that continues to explore degenerate or near-equivalent solutions even after convergence.

4.5 Summary of expressivity effects

Figure 6 provides a comparative overview of evaluation metrics across regimes.

The results demonstrate a clear qualitative transition as expressivity increases. While low expressivity systems can survive, they frequently fail to sustain complete life-like organization. Mid expressivity enables near-perfect success with balanced functional coverage. High expressivity, by contrast, supports aggressive growth and reproduction, often at the expense of strict functional optimization.

4.6 Ablation analysis of expressivity components

To isolate the contribution of individual expressivity components, we performed an ablation analysis in which feedback mechanisms and internal memory variables were selectively enabled or disabled while keeping the optimization algorithm fixed. This experiment evaluates whether the observed improvements in proto-life success arise from

Table 3: Evaluation results across expressivity regimes using the SAC controller. Values are reported as mean ± standard deviation across multiple seeds.

Regime	Success Rate	Survival Rate	Role Coverage	Divisions
Low	0.675 ± 0.04	1.000 ± 0.00	0.695 ± 0.03	7.58 ± 1.12
Mid	0.975 ± 0.02	1.000 ± 0.00	0.740 ± 0.02	9.05 ± 1.30
High	1.000 ± 0.00	1.000 ± 0.00	0.622 ± 0.04	34.58 ± 4.10

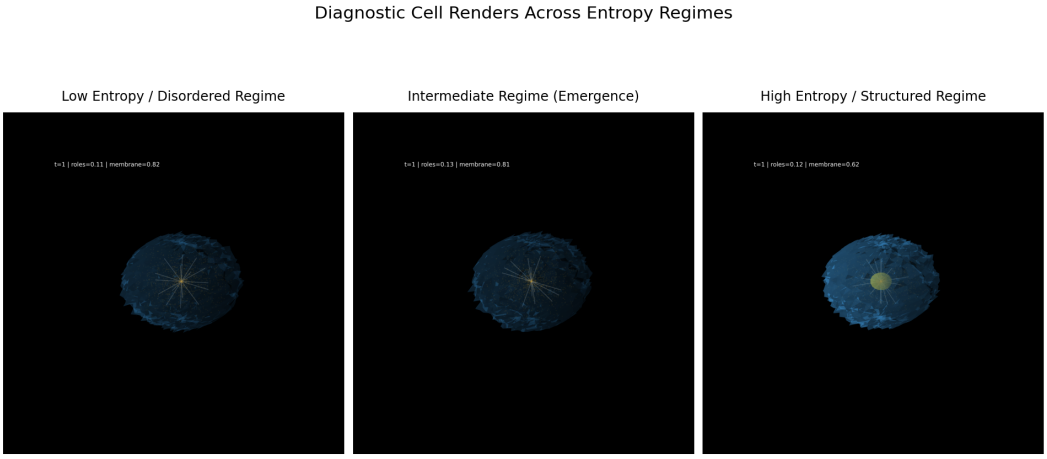

Figure 2: Qualitative diagnostic renders for representative evaluation trajectories under low, mid, and high expressivity regimes (left to right). The visualization encodes membrane shell stability (outer surface), internal mass dispersion (orange cloud), filament coherence (white bundles; proxy for functional organization), and an emergent core (gold; proxy for internal memory/energy when enabled). The figure illustrates a regime shift from disordered interior dynamics (low expressivity), to partial organization (mid), to coherent, memory-bearing structure with stabilized compartmentalization (high).

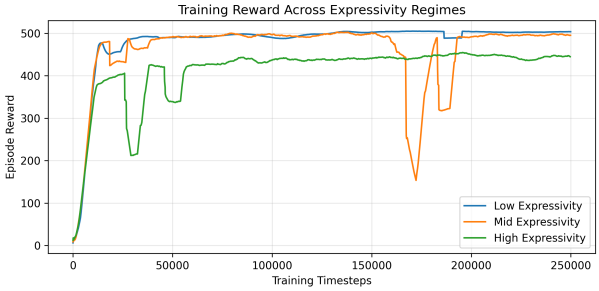

Figure 3: Training reward trajectories across expressivity regimes. Curves are smoothed using a rolling mean to highlight long-term trends.

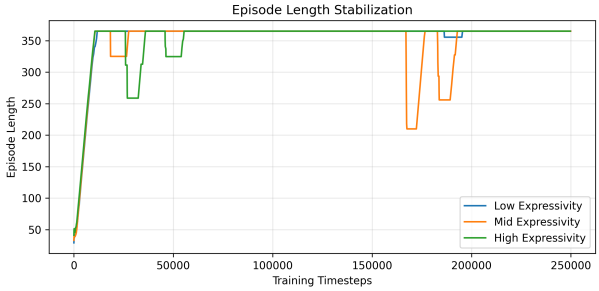

Figure 4: Episode length stabilization during training. Maximum episode length corresponds to a full 365-step horizon.

specific structural mechanisms rather than from general increases in system complexity.

The ablation results indicate that both feedback and internal memory contribute substantially to sustained proto-life organization. Systems lacking both mechanisms exhibit limited functional stability and rare division events, whereas enabling both components yields the highest success rates and sustained growth behavior. This supports the interpretation that expressivity, rather than optimization strength alone, determines the feasibility of emergent proto-biological dynamics.

4.7 Reward sensitivity analysis

To assess whether the main conclusions depend strongly on a particular reward specification, we conducted a sensitivity analysis in which the principal reward weights were varied around their nominal values by $\pm 10\%$ and $\pm 20\%$. The overall ranking of expressivity regimes remained stable across these perturbations, indicating that the central conclusion—that model expressivity constrains the emergence of sustained proto-life behavior—is robust to moderate changes in reward shaping.

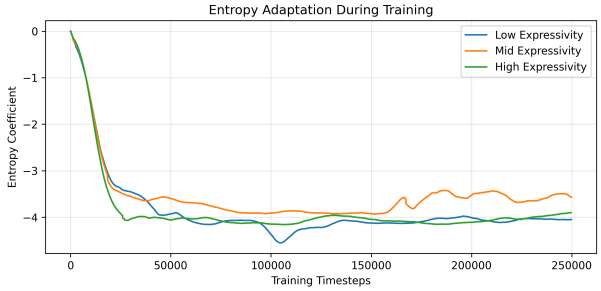

Figure 5: Entropy coefficient adaptation during training under automatic entropy tuning.

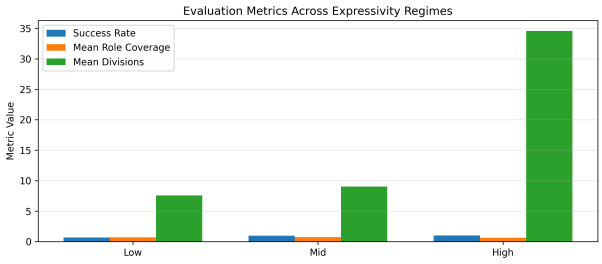

Figure 6: Comparison of success rate, mean functional role coverage, and mean division count across expressivity regimes.

4.8 Interpretation as a diagnostic optimization benchmark

Taken together, these results substantiate the central thesis of this work: when optimization power is held fixed at a strong level through maximum-entropy reinforcement learning, the feasibility of sustained proto-life behavior is governed primarily by the expressive capacity of the underlying model rather than by limitations of learning, exploration, or search. In regimes with insufficient expressivity, the failure to induce stable life-like organization is not explained by the observed convergence or exploration diagnostics. Instead, it is consistent with the absence of critical state variables, feedback couplings, or internal memory mechanisms required to support self-maintenance and division. Because entropy-regularized SAC explicitly encourages broad exploration while retaining strong control performance, its inability to recover viable dynamics provides a diagnostic signal consistent with structural incompleteness in the modeled system.

From this perspective, maximum-entropy reinforcement learning serves not as a generative explanation for emergence, but as a diagnostic optimization benchmark against which model adequacy can be evaluated. When even a strong controller fails under stable learning and sustained exploration, the evidence points toward the representational and dynamical assumptions of the model, while remaining conditional on the tested setup.

Table 4: Ablation analysis of expressivity components under SAC control. Values are reported as mean ± standard deviation across multiple evaluation runs.

Feedback	Memory	Success Rate	Role Coverage	Divisions
Off	Off	0.41 ± 0.06	0.48 ± 0.05	3.12 ± 0.88
On	Off	0.67 ± 0.04	0.69 ± 0.03	7.58 ± 1.12
Off	On	0.58 ± 0.05	0.63 ± 0.04	5.41 ± 1.05
On	On	0.97 ± 0.02	0.74 ± 0.02	9.05 ± 1.30

5 Discussion, implications, and limitations

This section interprets the empirical findings of the study, situates them within broader theoretical debates on emergence and origin-of-life modeling, and discusses the limitations of the proposed framework. The central aim is to clarify what the observed successes and failures of maximum-entropy reinforcement learning reveal about the relationship between optimization power and model expressivity in proto-life systems.

5.1 Interpretation: optimization as a diagnostic probe

The results demonstrate that when optimization power is fixed at a strong level using maximum-entropy reinforcement learning, the emergence of sustained, life-like organization depends primarily on the expressive capacity of the modeled system. Across all expressivity regimes, the SAC agent converges reliably, episode lengths stabilize, and survival rates reach unity, indicating that failures of emergence cannot be attributed to insufficient learning, exploration, or control.

In low expressivity regimes, the optimizer rapidly saturates and converges to stable control policies, yet life-like organization fails to materialize beyond minimal persistence. This outcome reflects structural inadequacy: key interaction channels, feedback pathways, or internal state variables required for sustained organization are absent. In contrast, intermediate expressivity enables the emergence of a balanced proto-life regime characterized by stable functional role coverage, membrane integrity, and controlled division.

At high expressivity, the system undergoes a qualitative transition. Additional feedback and internal memory mechanisms permit growth-dominated strategies marked by frequent division events and elevated state variability. Although aggregate reward decreases due to increased complexity penalties, biological activity is enhanced rather than diminished. Importantly, this behavior emerges without any change to the learning algorithm, confirming that expressivity reshapes what optimization can realize rather than how effectively it is performed.

5.2 Comparison with prior work

Compared with prior origin-of-life and protocell modeling studies, the present work differs in one central methodological aspect: it treats optimization strength as a controlled experimental variable rather than as an implicit assumption. Classical self-organization and autocatalytic network models typically examine whether emergence occurs under fixed dynamical rules, but they do not explicitly test whether negative results arise from insufficient exploration or from missing structural variables. Evolutionary and selection-based approaches improve discovery efficiency but often embed strong prior guidance through fitness functions and mutation operators. As a result, optimization power and model expressivity remain tightly coupled.

In contrast, the framework presented here fixes a strong entropy-regularized reinforcement learning controller and varies only the structural expressivity of the proto-life environment. The observed improvement in success rate and division behavior with increasing expressivity therefore supports a distinct conclusion: the emergence of sustained proto-biological organization is fundamentally constrained by the representational structure of the model rather than solely by search capability.

5.3 Theoretical implications for emergence and origin-of-life modeling

These findings have several theoretical implications. First, they challenge the implicit assumption that sufficiently powerful optimization or selection mechanisms are alone capable of generating life-like organization. Even under maximum-entropy reinforcement learning, which explicitly incentivizes sustained exploration and avoids premature convergence, emergence fails when the modeled system lacks essential representational structure.

Second, the results support a view of emergence as fundamentally constrained by model expressivity. Optimization can only act upon variables, couplings, and feedback channels that are explicitly represented. When critical dimensions are missing, no amount of optimization pressure can recover them. In this sense, the absence of emergence becomes informative rather than negative, signaling structural incompleteness rather than algorithmic inadequacy.

Third, reframing maximum-entropy reinforcement learning as a diagnostic optimization benchmark provides a principled computational methodology for artificial life and origin-of-life research. Instead of treating optimization success as sufficient evidence of plausibility, persistent failure under strong optimization conditions can be used to challenge under-specified models and guide the systematic introduction of missing structure.

More broadly, the observed shift from functional balance to growth-dominated regimes at high expressivity highlights a tension between biological activity and organizational parsimony. This mirrors long-standing theoretical arguments that early life must negotiate trade-offs between functional completeness, robustness, and minimal complexity.

5.4 Limitations

Several limitations of the present study should be acknowledged. First, the proto-life environment is intentionally abstract and does not model explicit biochemical reactions, molecular identities, or thermodynamic constraints. While this abstraction is essential for isolating expressivity effects, it limits the direct biological interpretability of the results. Second, the expressivity regimes are discretized into three levels rather than varied continuously. Although this design choice enables clear comparisons, a finer-grained exploration of expressivity transitions may reveal additional critical thresholds or bifurcation points. Third, only a single optimization paradigm, Soft Actor–Critic with automatic entropy tuning, is considered. While SAC represents a strong and widely accepted maximum-entropy method, alternative optimization frameworks or evolutionary dynamics may interact with expressivity in distinct ways. Finally, the reward function encodes a specific notion of biological sufficiency that necessarily reflects modeling choices. Different reward formulations could shift the balance between growth, functional coverage, and complexity, although such changes would not alter the fundamental constraint that optimization acts only within the space defined by the model.

6 Conclusion

This study investigated the limits of optimization-driven emergence in abstract proto-life systems by explicitly decoupling optimization power from model expressivity. By employing maximum-entropy reinforcement learning through the SAC algorithm, we constructed a strong optimization setting in which exploration, control, and long-horizon credit assignment were no longer the dominant constraints. Within this controlled framework, the central question was whether life-like organization could reliably emerge as a function of the expressive capacity of the modeled system itself.

The results demonstrate that optimization alone is insufficient to induce sustained proto-biological behavior when essential structural components are absent. Across all expressivity regimes, learning converged reliably, survival rates reached unity, and policy entropy stabilized. The remaining differences in proto-life outcomes therefore cannot be explained solely by inadequate learning or exploration; within the assumptions of the model, they instead point to the availability or absence of structural variables, feedback mechanisms, and internal state representations. Optimization, even when strong and entropy-regularized, cannot compensate for missing structure. Persistent failure under low expressivity thus reflects missing interaction channels and feedback pathways required for self-maintenance

and division. As expressivity increased, the system exhibited clear qualitative transitions, from a balanced proto-life regime at intermediate expressivity to growth-dominated dynamics with frequent division events at high expressivity. The associated reduction in aggregate reward under high expressivity reflects structural trade-offs imposed by the reward formulation, rather than a breakdown of optimization.

These findings support the use of maximum-entropy reinforcement learning as a stringent empirical benchmark for what optimization-based guidance can achieve within a given model. When a strong optimizer repeatedly fails to produce life-like organization despite stable learning and sustained exploration, the result provides evidence consistent with limitations in the model's expressive assumptions. From this perspective, entropy-regularized reinforcement learning serves as a diagnostic instrument, allowing optimization failure to inform the assessment of possible structural incompleteness. More broadly, this work suggests that progress in computational origin-of-life research depends less on strengthening optimization mechanisms and more on identifying and representing the minimal interaction structures, feedback pathways, and memory processes required for sustained organization. Emergence, even under strong optimization, remains fundamentally bounded by representation.

Data availability statement

The study is based on simulated data generated by the computational environment described in this article. The simulation code, configuration files, trained-policy settings, and generated evaluation outputs are available from the corresponding author upon reasonable request.

Conflict of interest

The authors declare that they have no conflict of interest.

References

[1] I. Gözen, E. S. Köksal, I. Põldsalu, L. Xue, K. Spustova, E. Pedrueza-Villalmanzo, R. Ryskulov, F. Meng, and A. Jesorka, "Protocells: Milestones and recent advances," *Small*, vol. 18, no. 18, Art. 2106624, 2022. https://doi.org/10.1002/smll.202106624

[2] W. Hordijk and M. Steel, "Autocatalytic sets and the origin of life," *Entropy*, vol. 20, no. 4, Art. 250, 2018. https://doi.org/10.3390/e20040250

[3] S. Mann, "Systems of creation: The emergence of life from nonliving matter," *Accounts of Chemical Research*, vol. 45, no. 12, pp. 2131–2141, 2012. https://doi.org/10.1021/ar200281t

[4] E. Camprubí, J. W. de Leeuw, C. H. House, F. Raulin, M. J. Russell, A. Spang, M. R. Tirumalai, and F. Westall, "The emergence of life," *Space Science Reviews*, vol. 215, no. 8, Art. 56, 2019. https://doi.org/10.1007/s11214-019-0624-8

[5] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. Cambridge, MA, USA: MIT Press, 2018.

[6] J. Terven, "Deep reinforcement learning: A chronological overview and methods," *AI*, vol. 6, no. 3, Art. 46, 2025. https://doi.org/10.3390/ai6030046

[7] S. Han and Y. Sung, "A max–min entropy framework for reinforcement learning," in *Advances in Neural Information Processing Systems*, vol. 34, pp. 25732–25745, 2021.

[8] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel, and S. Levine, "Soft actor–critic algorithms and applications," *arXiv preprint arXiv:1812.05905*, 2018. https://doi.org/10.48550/arXiv.1812.05905

[9] R. Serra and M. Villani, *Modelling Protocells*. Cham, Switzerland: Springer, 2017. https://doi.org/10.1007/978-94-024-1160-7

[10] W. Hordijk and M. Steel, "Autocatalytic networks at the basis of life's origin and organization," *Life*, vol. 8, no. 4, Art. 62, 2018. https://doi.org/10.3390/life8040062

[11] T. W. Malone and S. A. Smith, "Modeling the performance of organizational structures," *Operations Research*, vol. 36, no. 3, pp. 421–436, 1988. https://doi.org/10.1287/opre.36.3.421

[12] S. McLeod, "Absolute biological needs," *Bioethics*, vol. 28, no. 6, pp. 293–301, 2014. https://doi.org/10.1111/bioe.12020

[13] H. G. Garcia, J. Kondev, N. Orme, J. A. Theriot, and R. Phillips, "Thermodynamics of biological processes," in *Methods in Enzymology*, vol. 492, D. E. Discher and T. J. Mitchison, Eds. Amsterdam, The Netherlands: Elsevier, 2011, pp. 27–59. https://doi.org/10.1016/B978-0-12-381268-1.00014-8

[14] W. Hordijk, M. Steel, and P. Dittrich, "Autocatalytic sets and chemical organizations: Modeling self-sustaining reaction networks at the origin of life," *New Journal of Physics*, vol. 20, no. 1, Art. 015011, 2018. https://doi.org/10.1088/1367-2630/aa9fcd

[15] E. Mossel and M. Steel, "Random biochemical networks: The probability of self-sustaining autocatalysis," *Journal of Theoretical Biology*, vol. 233, no. 3,

pp. 327–336, 2005. https://doi.org/10.1016/j.jtbi.2004.10.011

[16] S. Mann, "The origins of life: Old problems, new chemistries," *Angewandte Chemie International Edition*, vol. 52, no. 1, pp. 155–162, 2013. https://doi.org/10.1002/anie.201204968

[17] N. Lehman and S. A. Kauffman, "Constraint closure drove major transitions in the origins of life," *Entropy*, vol. 23, no. 1, Art. 105, 2021. https://doi.org/10.3390/e23010105

[18] V. Vanchurin, Y. I. Wolf, E. V. Koonin, and M. I. Katsnelson, "Thermodynamics of evolution and the origin of life," *Proceedings of the National Academy of Sciences*, vol. 119, no. 6, Art. e2120042119, 2022. https://doi.org/10.1073/pnas.2120042119

[19] R. Hasibi, T. Michoel, and D. A. Oyarzún, "Integration of graph neural networks and genome-scale metabolic models for predicting gene essentiality," *npj Systems Biology and Applications*, vol. 10, Art. 24, 2024. https://doi.org/10.1038/s41540-024-00348-2

[20] E. D. Vaishnav, C. G. de Boer, J. Molinet, M. Yassour, L. Fan, X. Adiconis, D. A. Thompson, J. Z. Levin, F. A. Cubillos, and A. Regev, "The evolution, evolvability and engineering of gene regulatory DNA," *Nature*, vol. 603, no. 7901, pp. 455–463, 2022. https://doi.org/10.1038/s41586-022-04506-6

[21] H. Lu, F. Li, L. Yuan, I. Domenzain, R. Yu, H. Wang, G. Li, C. Yu, B. Ji, E. J. Kerkhoven, and J. Nielsen, "Yeast metabolic innovations emerged via expanded metabolic network and gene positive selection," *Molecular Systems Biology*, vol. 17, Art. e10427, 2021. https://doi.org/10.15252/msb.202110427

[22] H. Bai, R. Cheng, and Y. Jin, "Evolutionary reinforcement learning: A survey," *Intelligent Computing*, vol. 2, Art. 0025, 2023. https://doi.org/10.34133/icomputing.0025

[23] E. O. Neftci and B. B. Averbeck, "Reinforcement learning in artificial and biological systems," *Nature Machine Intelligence*, vol. 1, no. 3, pp. 133–143, 2019. https://doi.org/10.1038/s42256-019-0025-4

[24] A. S. Alshehri and F. You, "Deep learning to catalyze inverse molecular design," *Chemical Engineering Journal*, vol. 444, Art. 136669, 2022. https://doi.org/10.1016/j.cej.2022.136669