

Concept Paper

Not peer-reviewed version

Scientific Super Intelligence: From Science Tokens to ScienceRSI

Zaixi Zhang ^{*}, Mingyang Xu, Jiaxian Yan, Zhicai Wang, Dubuke Ma, Yufeng Liu, Qizhi Pei

Posted Date: 13 October 2026

Keywords: RSI; AI4Science



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC, OpenAlex.

Copyright: This open access article is published under a [Creative Commons CC BY 4.0 license](https://creativecommons.org/licenses/by/4.0/), which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Concept Paper

Scientific Super Intelligence: From Science Tokens to ScienceRSI

Zaixi Zhang ^{1,2,*}, Mingyang Xu ¹, Jiaxian Yan ¹, Zhicai Wang ¹, Dubuke Ma ^{1,3}, Yufeng Liu ^{1,4} and Qizhi Pei ^{1,5}

¹ Scinetics Company
² Hong Kong University of Science and Technology
³ University of Michigan
⁴ University of Cambridge
⁵ Renmin University of China
* Correspondence: zaixi.zhang@scinetics.ai

Abstract

Scientific reasoning moves beyond natural language. The search for a discovery runs across sequences, structures, fields, omics, images and measurements, while language models reason mainly in text. This Perspective sets out four proposals from the Scinetics team. First, we define *science tokens*: modality-appropriate units of scientific information, spanning native discrete symbols, typed values, continuous representations and learned, decodable codes. A shared interface for these representations would let one model read, write and reason across modalities, so that a *chain of science tokens* operates on scientifically meaningful states, not on descriptions of them. Second, we define *Scientific Super Intelligence* (SSI) as a joint crossover: one frozen model that outperforms vertical specialists, such as the AlphaFold series on protein structure prediction, on their own tasks, and outperforms general large language models of the same size on broad scientific reasoning. Capacity, published training precedents, data inventories and compute place the first test of that crossover at 100–300 billion parameters and 10–30 trillion training tokens. Third, we propose *ScienceRSI* (Science Recursive Self-Improvement) as the route to that window. Following the AI4AI pattern, each larger generation is trained on data that the smaller generation before it produced: the earlier model generates candidate reasoning, screens it and keeps what passes, and in science that screening can be done by a simulation, a database or an experiment. Training this way reduces the dependence on scarce human-written scientific reasoning data and shortens each iteration. Fourth, we propose *science-token economics*, which separates metered computation from the value of evidence-labelled, reusable scientific assets.

Keywords: RSI; AI4Science

Four messages

1. **Science tokens.** Learned tokenizers turn sequences, structures, images, omics and measurements into science tokens that share one vocabulary with text. A model can then reason as a chain of science tokens, taking intermediate steps directly in scientific space.
2. **Scientific Super Intelligence.** One model that beats vertical specialists, such as the AlphaFold series, on their own tasks and beats same-size general LLMs on general scientific reasoning. We place the first test of this crossover at 100–300 billion parameters and 10–30 trillion training tokens.
3. **ScienceRSI.** Science Recursive Self-Improvement: a smaller model generates and screens the reasoning data that trains the larger model after it, with a simulator, a database, an assay or an expert deciding what survives. Each generation therefore depends less on scarce human-written scientific reasoning data, and the next one arrives sooner.
4. **Science-token economics.** Language-model services price tokens. Science-token economics prices validated, reusable scientific assets.

1. Introduction

Scientific reasoning moves beyond natural language. A structural biologist states a hypothesis in words, but tests it on a sequence, a 3D structure, an electron-density map or an assay readout, and revises it from what those objects show. Chain-of-thought prompting showed that writing out intermediate steps improves a model’s reasoning [1]. For science, those steps should be allowed to be scientific objects, with language remaining an active participant.

Specialist systems show what representations suited to scientific structure can achieve. AlphaFold and AlphaFold 3 couple biological information with geometric computation, and Evo 2 and ESM3 extend foundation modelling across genomes and proteins [2–5]. Each excels within its own representation. General language models are broad, but they reason over descriptions of objects rather than the objects themselves. The next step is one model that combines specialist depth with general scientific reasoning. We call this target *Scientific Super Intelligence* (SSI).

This Perspective sets out four proposals from the Scinetics team: what science tokens are and why modalities should be unified as science tokens (Section 2); how SSI should be defined and at what scale it could first be tested (Section 3); how a model might get there, by carrying the AI4AI pattern, in which one generation of models trains the next, into science as what we call ScienceRSI (Section 4); and how science-token economics differs from the token economics of today’s language-model services (Section 5). Molecular and biological science supply the quantitative examples.

2. Science Tokens and Reasoning in Scientific Space

2.1. What Is a Science Token?

Supplying a model with scientific inputs does not ensure that it uses them. When Fang and colleagues perturbed the biological inputs of BioReason and BioReason-Pro, aggregate performance barely changed, although task information was recoverable from those inputs; in evidence-conflict tests the models preferred the accompanying text [6]. Text annotations can thus become a shortcut. A scientific model must make sample-specific sequences, geometries and measurements consequential to its decisions, and that begins with how they are represented.

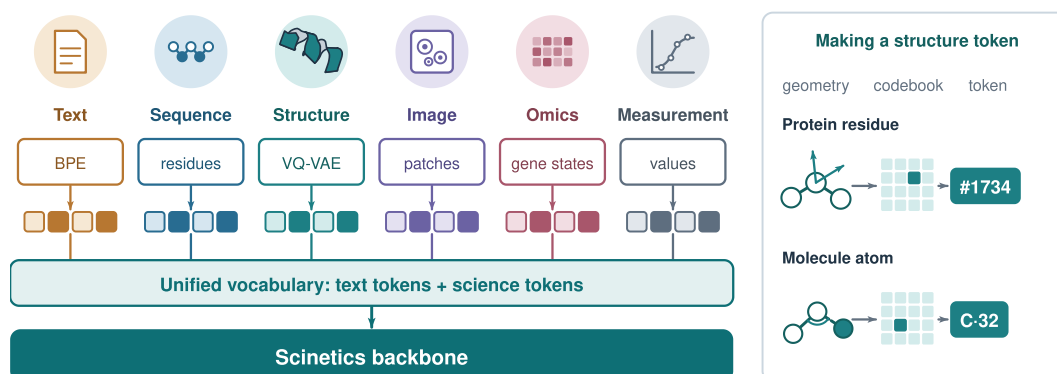
We call the representational unit a *science token* (Box 1; Figure 1a). A science token is the output of a learned, decodable tokenizer applied to one local unit of a scientific object. Recent structure tokenizers make the idea concrete. GCP-VQVAE encodes the backbone geometry around each protein residue with an SE(3)-equivariant encoder and assigns it to one of 4,096 codebook entries; its decoder recovers backbones with sub-ångström error [7]. Mol-StrucTok pairs each atom of a molecular line notation with a structure code that quantizes the atom’s local spherical coordinates into one of 256 entries [8]. Both turn continuous 3D geometry into discrete tokens that a language model can read and write, and both can be decoded back into coordinates.

Box 1. Key terms: science tokens.

Science token. A unit of a scientific object produced by a learned, decodable tokenizer: for example, the codebook index of a residue’s backbone geometry. It carries its modality and, where needed, units, conditions and provenance.
Token families. Sequence, structure, image, omics and measurement tokens. Fields sampled on grids are tokenized like images.
Modality tokenizer. An encoder, a discrete codebook or continuous embedding, and a decoder that maps tokens back to the original data.
Scientific vocabulary. Text tokens and science tokens from all modalities, embedded in one space and read by one backbone.
Chain of science tokens. A reasoning trajectory whose intermediate steps include scientific states as well as text, each consumed by the steps that follow. A chain of thought is the special case in which every step is text.
Reasoning in scientific space. Consecutive steps that transform one scientific state into another (placing a side chain, revising a geometry, editing a sequence) in the object’s own representation, where constraints can be checked rather than described.

We distinguish five families of science tokens. *Sequence tokens* are residues or nucleotides. *Structure tokens* encode the local 3D geometry of proteins, nucleic acids or small molecules. *Image tokens* encode patches of microscopy and other images; physical fields sampled on grids can be tokenized the same way. *Omics tokens* encode gene-state features of cells and tissues. *Measurement tokens* encode assay readouts and other values. Every token is typed: modality tags and metadata keep units, conditions and provenance, so a predicted structure and a measured one remain distinguishable. Text tokens and science tokens together form one *scientific vocabulary*.

a Science tokens: one vocabulary for scientific modalities



b Chain of science tokens: text and structure revising one another

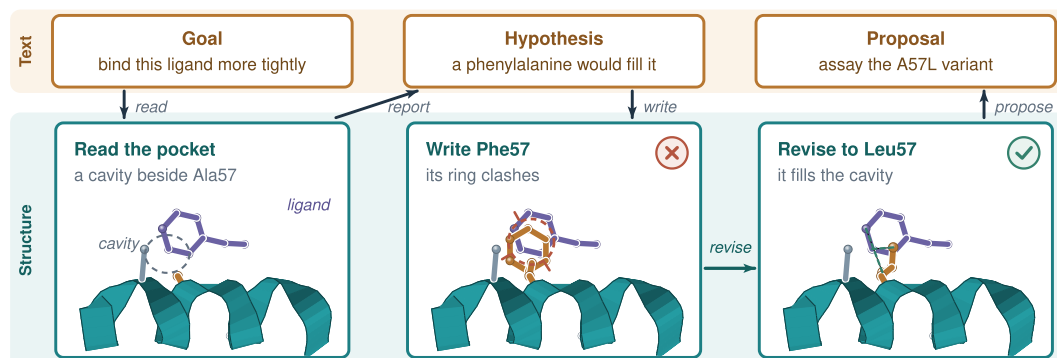


Figure 1. Science tokens and a chain of science tokens. **a**, Learned tokenizers convert each modality into tokens: text by byte-pair encoding, sequences by residue or nucleotide, structures by vector-quantized (VQ) codebooks, images by patches, omics by gene-state features and measurements by value encoders. Text and science tokens form one vocabulary that the Scinetics backbone reads and writes. Inset: a structure token is the codebook index of local geometry: a residue's backbone frame, as in GCP-VQVAE (4,096 codes), or an atom's local spherical coordinates, as in Mol-StrucTok, where atom type and structure code form one token (C-32). Decoders map tokens back to coordinates. Token indices are illustrative. **b**, An illustrative chain of science tokens in which text and structure revise one another. A goal stated in text sends the model to read the binding pocket as structure tokens, where it finds a cavity beside Ala57; the reading is reported back in text as a hypothesis; the model writes the phenylalanine side chain and a steric check shows its ring overlapping the ligand (red); it revises within structure space to leucine, which fills the cavity and makes contact without a clash (green); and only then does it return to text with a testable proposal. Arrows crossing between the lanes are the hand-offs between language and scientific space; the horizontal arrow stays inside structure space. The backbone is a helical segment drawn as a shaded ribbon cartoon, with the side chain and ligand as sticks; it is a schematic local view built from helix geometry, not an experimental structure. The trajectory is a design illustration, not a reported experimental result: every state shown is predicted or simulated, and none is measured or experimentally validated.

2.2. Why Unify Modalities as Science Tokens?

Unifying modalities as science tokens could offer four advantages, although each remains an empirical hypothesis.

A shared interface: a shared backbone could support interactions across text, sequences, structures, images, omics and measurements, potentially enabling cross-modal transfer when the training data meaningfully connect those modalities. ESM3 demonstrates joint modelling across protein sequence, structure and function, while Transfusion shows that next-token prediction and diffusion can be combined within one multimodal architecture [5,9]. These results establish relevant architectural precedents, but they do not demonstrate that a universal scientific vocabulary will improve every task or modality.

Generation within an integrated model: when an appropriate decoder is available, science tokens could allow a model to generate or revise a sequence, structure, image or other scientific state within an integrated reasoning system. This differs from tool-using agents, such as Coscientist, in which a language model plans actions and invokes external software or experimental tools [10]. It also differs from pipelines that pass outputs between separately trained specialist models, including domain-specific systems such as the AlphaFold series [2,3]. These approaches are complementary: external tools and specialist modules may still provide higher-fidelity calculations, established validation procedures or interfaces to laboratory equipment.

Fidelity and checkability: scientific information can be represented with modality-appropriate precision and associated metadata, including units, conditions, provenance and evidence status. Such representations could make intermediate states easier to evaluate using physical constraints, simulations, databases or experiments. Tokenization does not, however, guarantee lossless encoding, scientific correctness or the availability of a reliable validator.

A distinct reasoning substrate: ordinary multimodal models may receive or generate several modalities without using scientific representations as causally consequential intermediate steps. Text-only chain-of-thought expresses intermediate reasoning in natural language [1], whereas continuous latent reasoning operates through hidden neural states that need not be scientifically typed, decodable or directly interpretable [11]. A chain of science tokens instead proposes explicit scientific states that are consumed by subsequent reasoning steps and are decodable or otherwise scientifically interpretable where appropriate. This distinction is testable: matched examples can hold the text fixed while changing a relevant scientific input. The model's decision should respond to scientifically meaningful interventions while remaining stable to irrelevant transformations.

2.3. How Science Tokens Are Obtained

A science-token system would require four related but distinct stages. They should not be conflated, because an effective representation does not by itself create cross-modal alignment, a reasoning process or scientifically grounded optimization.

Modality-specific encoding or tokenization: each modality should use a representation suited to its structure. Text can use subword units, and DNA, RNA and protein sequences already consist of discrete symbols, so these modalities do not require vector quantization or a learned codebook. Images can be represented as patches, measurements as typed numerical values, and omics profiles as sparse feature sets or continuous embeddings. For continuous three-dimensional geometry, learned discrete codebooks are one option, as illustrated by GCP-VQVAE for protein backbones and Mol-StrucTok for molecular coordinates, but these are examples rather than a universal requirement [7,8]. When generation or reconstruction is required, the representation also needs an appropriate decoder and metadata for units, conditions and provenance.

Cross-modal alignment: after encoding, representations must be linked when they describe the same scientific object, experiment or observation. Alignment objectives could connect text, sequences, structures and measurements through paired prediction, reconstruction or generation. ESM3 provides a biological precedent by jointly modelling protein sequence, structure and function, whereas Transfusion shows that modality-specific objectives can be trained within one transformer architecture [5,9]. These systems demonstrate relevant design principles, but they do not establish that one alignment method will work across all scientific modalities.

Reasoning-process training: aligned representations do not ensure that a model will use them as causally consequential intermediate states. Training would therefore require interleaved trajectories in which text and scientific states are consumed by later steps, using expert demonstrations, simulations or curated experimental records. Textual chain-of-thought and COCONUT's continuous latent reasoning provide precedents for training or eliciting intermediate reasoning processes, but neither alone demonstrates reasoning with explicit, decodable scientific states [1,11].

Outcome-based or experimentally grounded optimization: when reliable evaluators are available, candidate trajectories could be assessed against physical constraints, simulations, databases or experimental measurements. Reinforcement learning is appropriate only when the reward faithfully represents the scientific objective; otherwise, optimization may amplify a proxy rather than improve scientific validity. Closed-loop robotic chemistry demonstrates that measured outcomes can guide an experiment–analysis–decision cycle, although it does not yet validate end-to-end training of the proposed science-token system [12].

Representation, alignment, reasoning-process training and outcome-based optimization are therefore separable design choices. Their integration is a testable research programme rather than an established recipe.

2.4. Chains of Science Tokens: Reasoning in Scientific Space

A chain of science tokens extends chain-of-thought reasoning by making scientific states part of the trajectory (Figure 1b). As an illustrative reasoning example rather than a reported experimental result, consider tightening a ligand's binding by editing one side chain. Text states the goal. The model reads the binding pocket as structure tokens and finds a cavity beside Ala57. Text proposes a hypothesis: a phenylalanine would fill it. The model writes the mutated side chain as structure tokens, and a steric check shows that the phenyl ring overlaps the ligand. Without leaving structure space, the model revises the choice to leucine, whose branched side chain fills the cavity and contacts the ligand without a clash. Only then does it return to text and propose A57L for an assay.

What makes this a chain rather than a pipeline is that each step consumes the state the previous step wrote. The hypothesis is conditioned on the geometry the model read, not on a textual summary of it; the steric check applies to the coordinates the model itself generated; and the revision is made in the same representation that failed, so the reason for the failure is available to it. The hand-offs run in both directions: language sets the goal and carries the interpretation, while the scientific states decide what survives.

We call such consecutive steps *reasoning in scientific space*. In a text-only chain, "a phenylalanine would fill the cavity" is a sentence that nothing checks. In a chain of science tokens it is a set of coordinates that either clashes or does not. Working in the object's own representation lets the model compare candidates and revise against constraints that language can state but cannot evaluate.

An evaluation should therefore combine intervention and invariance tests. Holding the text fixed, change a scientifically relevant input, such as residue identity or pocket geometry; the model's intermediate states and conclusion should change appropriately. Then change an irrelevant representational feature, such as a rigid rotation or an equivalent atom ordering; the reasoning should remain invariant. Every state in the trajectory should also be labelled as predicted, simulated, measured or experimentally validated. These labels must not be treated as interchangeable: a plausible prediction or successful simulation is not a measurement, and experimental validation requires a declared empirical test.

3. Scientific Super Intelligence: Definition and Scale

3.1. An Operational Definition

We define SSI as a joint crossover achieved by one fixed model checkpoint (Box 2; Figure 2a). The model must meet two conditions. On vertical scientific tasks, it must outperform the leading specialist models on their own ground, for example the AlphaFold series on protein structure prediction [2,3]. On general scientific reasoning, it must outperform general large language models with the same

number of parameters. Figure 2a sketches the idea. As scale grows, a model trained on science tokens improves faster than a general LLM and eventually crosses the specialist level; the smallest model past that point is the SSI crossover. The definition names what is surpassed. It does not claim superiority over human scientists. Because apparent emergence can depend on the scoring rule [13], continuous learning curves should accompany any claim.

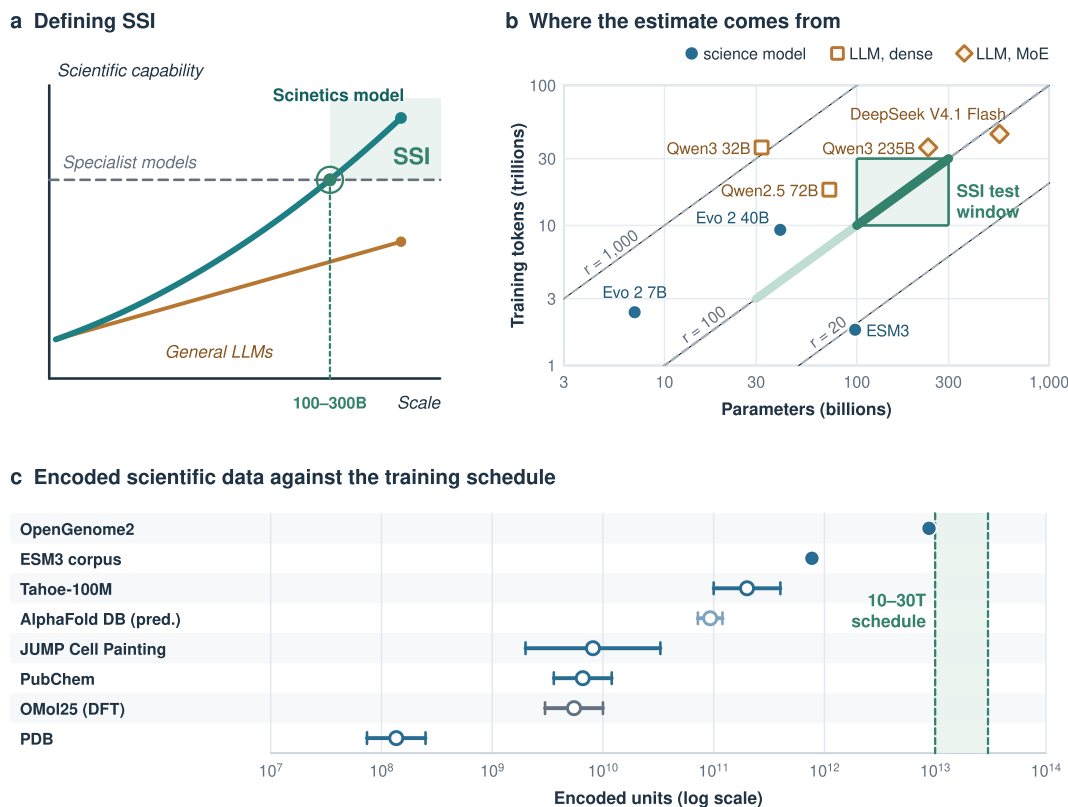


Figure 2. Scientific Super Intelligence: definition and the basis of the scale window. **a**, Schematic capability of a model trained on science tokens (teal) and of general LLMs (gold) as scale grows. The dashed line marks specialist models on their own tasks, such as the AlphaFold series on protein structure prediction. SSI begins at the crossover, where one model exceeds the specialists while staying above general LLMs of the same size; the green band marks our proposed 100–300B test window. Curves are illustrative, not measured. **b**, Published training configurations of scientific foundation models (Evo 2 7B/2.4T and 40B/9.3T; ESM3, about 98B/1.8T) and general LLMs (Qwen2.5-72B, 18T; Qwen3-32B and Qwen3-235B-A22B, 36T; DeepSeek-V4.1-Flash, 552B backbone parameters with 8–16B active, 45T). Mixture-of-experts (MoE) models are plotted at total parameters. Dashed lines show constant tokens per parameter, $D = rN$. Green: the SSI test window (100–300B parameters, 10–30T tokens at $r = 100$) and the 30–100B pilot ladder. **c**, Encoded volumes of selected scientific resources (Table 1) against the 10–30T schedule. Filled circles are reported counts; open circles and bars are assumed encoding ranges with geometric midpoints. AlphaFold DB entries are predictions and OMol25 entries are DFT calculations. Counts are neither additive nor equivalent to text tokens.

Comparisons are fixed before evaluation. The specialist panel might include a fixed AlphaFold 3 release for new structures and designated systems for molecular properties and cellular perturbation response, with external evidence such as alignments and templates matched. A unified model that calls AlphaFold as a tool cannot claim that its own structure prediction exceeds AlphaFold. Size counts the complete model, including scientific encoders and decoders; mixture-of-experts models report total and active parameters. The general-science suite spans knowledge, data interpretation, quantitative reasoning, coding and experimental planning, in the direction of GPQA and LAB-Bench [14,15]. Each improvement must exceed a preregistered practical margin with a positive lower confidence bound, without severe regression in any declared domain.

Box 2. Key terms: Scientific Super Intelligence and scale

Scientific Super Intelligence (SSI). One frozen model checkpoint that outperforms vertical specialist models on their own tasks (for example, the AlphaFold series on protein structure prediction) and outperforms general LLMs of the same size on general scientific reasoning, both by preregistered margins.

Crossover. The smallest model, along a declared training schedule, that meets both conditions.

Parameters, tokens and the ratio r . N counts every parameter, including scientific encoders and decoders; D counts processed training tokens, including repeats; $r = D/N$.

Encoded volume versus exposures. Encoded volume $T_m = n_m \ell_m$ is what a resource yields once encoded (n_m objects of mean length ℓ_m); exposures $D_{\text{train}} = \sum_m e_m n_m \ell_m$ count e_m passes. Repetition adds computation, not independent evidence.

Table 1. Auditable representation budgets for selected scientific resources.

Resource and version	Reported holdings	Counting assumption	Encoded volume
OpenGenome2 [4]	>8.8 trillion nucleotides	One token per nucleotide	$> 8.8 \times 10^{12}$
ESM3 corpus [5]	771 billion unique tokens	Author-reported corpus count	7.71×10^{11}
Tahoe-100M [16]	100 million measured cells	1,000–4,000 gene-state positions per cell	1.0×10^{11} – 4.0×10^{11}
AlphaFold DB v6 [17]	241,070,489 predictions	300–500 residues per structure, one token each	7.2×10^{10} – 1.2×10^{11}
JUMP imaging [18]	>8 million images	8 million as anchor; 256–4,096 patches per image	2.0×10^9 – 3.3×10^{10}
PubChem 2025 [19]	119 million compounds	30–100 tokens per compound	3.6×10^9 – 1.2×10^{10}
OMol25 [20]	>100 million DFT calculations	100 million as anchor; 30–100 atom-state positions	3.0×10^9 – 1.0×10^{10}
PDB, end of 2025 [21]	247,287 coordinate entries	300–1,000 residues per entry, one token each	7.4×10^7 – 2.5×10^8

Anchors refer to the cited releases and are not additive. Length ranges are explicit encoding scenarios, not measured averages; counts and lengths must be re-measured after filtering and deduplication. Continuous gene-state and atom-state positions are not equivalent to text tokens; predicted structures and DFT calculations retain distinct provenance.

3.2. A Test Window: 100–300 Billion Parameters, 10–30 Trillion Tokens

We propose a test window for the SSI crossover in molecular and biological science: **100–300 billion parameters** trained on **10–30 trillion tokens**, with 30–100-billion-parameter models providing the first scaling measurements. Four lines of evidence motivate it (Figure 2b,c).

Capacity. An SSI candidate must hold, in one checkpoint, the broad reasoning of a language model, several scientific vocabularies and the transitions between them. Single-domain foundation models already use tens to about a hundred billion parameters, with Evo 2 at 40 billion and ESM3 near 100 billion [4,5]. A model that must surpass several such specialists, and a same-size LLM, is unlikely to need less. The window therefore begins where the largest single-domain models stop and stays within the range of current open general-purpose models. Shared representations could improve parameter efficiency, whereas interference between tasks could reduce it. Neither effect has a calibrated magnitude, so the window is a hypothesis to be tested.

Training precedents. Evo 2 trained 7 and 40 billion parameters on 2.4 and 9.3 trillion nucleotide tokens [4]; ESM3 approaches 100 billion parameters with about 1.8 trillion processed positions [5]. General-purpose models now train on far larger budgets. Qwen2.5 scaled pretraining to 18 trillion tokens, including for the dense 72-billion-parameter model [22]. All Qwen3 models, from the dense 32B model to the 235B mixture-of-experts model with 22B active parameters, were trained on 36 trillion tokens [23]. DeepSeek-V4.1-Flash, a mixture-of-experts model with 552 billion backbone parameters and 8–16 billion active, was trained on 45 trillion multimodal tokens [24]. Training at 100–300 billion parameters on 10–30 trillion tokens is therefore well within demonstrated practice.

Tokens per parameter. We pair the window with $r = D/N = 100$, giving 10–30 trillion tokens. Published ratios are about 18 for ESM3, 233 for Evo 2 40B, 250 for Qwen2.5-72B and 1,125 for Qwen3-32B. We choose $r = 100$ for two reasons. It lies well above the compute-optimal ratio of about 20 for text [25], because many scientific positions carry less information than a word piece, since a nucleotide carries at most two bits. It lies below current LLM practice, because high-quality scientific data are

scarcer than web text (Section 3.3). This is a planning assumption, and Figure 2b shows its sensitivity. Within the window, $r = 20$ would require 2–6 trillion tokens and $r = 250$ would require 25–75 trillion. Conversely, 10–30 trillion tokens correspond to 500–1,500 billion parameters at $r = 20$ and 40–120 billion at $r = 250$.

Compute. With $C \approx 6ND$ [25], the schedule requires about 6×10^{24} FLOP at 100B/10T, 2.4×10^{25} at 200B/20T and 5.4×10^{25} at 300B/30T. The same approximation gives about 7.8×10^{24} FLOP for Qwen2.5-72B (72B parameters, 18T tokens). The lower end of the window therefore matches the compute of a current open dense model, and the upper end is about seven times larger. Modality encoders, long contexts and post-training add to this proxy, and experiments are budgeted separately. Mixture-of-experts models need separate accounting, because only active parameters enter the per-token compute.

3.3. Is THERE Enough Scientific Data?

Scientific archives could support the encoded-data volume this schedule assumes (Table 1; Figure 2c). OpenGenome2 holds more than 8.8 trillion curated nucleotides, and the European Nucleotide Archive reports 26 trillion nucleotides of assembled and annotated sequence [4,26]; the two overlap and cannot be added. Coverage is uneven. The Protein Data Bank held 247,287 experimental entries at the end of 2025, whereas AlphaFold DB v6 lists about 241 million predicted structures [17,21], which can supervise a model but are not independent experiments. Tahoe-100M provides 100 million single-cell measurements across about 60,000 drug–cell-line combinations and JUMP more than eight million images of perturbed cells, while PubChem and OMol25 contribute molecular records and quantum-chemical calculations [16,18–20].

Holdings become budgets only through explicit counting (Box 2). For example, 100 million cells encoded with 1,000–4,000 gene-state positions yield 100–400 billion positions. Every extra pass adds computation without independent evidence. The 10–30-trillion-token schedule therefore counts a declared mixture of scientific and language exposures, including repeated passes, not 10–30 trillion independent observations. Inventories establish volume; cross-modal alignment, intervention coverage and held-out transfer determine usefulness. New evidence matters most where archives are thinnest. High-throughput assays, simulation and automated laboratories can supply it [10,27,28]. At constant annual growth of 20%, 50% or 100%, a consistently measured resource grows about 2.5-, 7.6- or 32-fold in five years. The most valuable loop is selective: a model identifies an uncertainty, chooses a discriminating experiment and learns from the result, negative outcomes included. That loop also produces the interleaved trajectories needed to train chains of science tokens.

3.4. Making the Window Falsifiable

The crossover location should be inferred from capability curves, not from loss alone. A pilot ladder at, for example, 7, 13, 30, 70 and 100 billion parameters would vary training volume, the fraction of paired cross-modal data and inference computation. Response surfaces for specialist and general-science performance, fitted with uncertainty and tested on withheld scales and on temporal, sequence-family, scaffold and batch holdouts, then predict where the crossover lies. Runs inside the 100–300-billion-parameter window would test it prospectively: a smaller failing model and a larger passing one bracket the crossover. Success at smaller scales would move the window down; failure throughout it would point to changes in representation, evidence or objectives. Either outcome is informative, which makes the window a research programme rather than a forecast.

4. ScienceRSI: Recursive Self-Improvement for Science

4.1. From AI4AI to ScienceRSI

Recursive self-improvement (RSI) is the idea that an AI system takes part in building its own successor. Its current practical form is narrow and concrete, and usually travels under the name *AI4AI*: one generation of models produces the training signal for the next. The models sample candidate

solutions and reasoning traces, filter and curate the data that survive, supply rewards or judgments, and help evaluate the result, so that capability compounds from one generation to the next rather than only within a single training run. Surveys of this literature separate *bounded* self-refinement, in which improvement is measured against an evaluator the system does not control, from *open-ended* self-modification, in which the system also rewrites the criteria by which it improves; the loops that work in practice are the bounded ones, and their ceiling is set by how reliable the evaluator is [29].

We propose *ScienceRSI*, or Science Recursive Self-Improvement, as AI4AI carried over to scientific reasoning models (Box 3; Figure 3). Concretely, a generation is trained on data that the generation before it produced. The earlier and smaller model, together with the agent platforms built on it, takes on three jobs that in ordinary training are done by people: it generates candidate chains of science tokens, it screens them, and it labels what each surviving chain is worth as evidence. Scientific evaluators external to the model make the accept or reject decision, and the verified result trains a larger next generation, which becomes the next producer. The loop is bounded by construction: the criteria of improvement are physical constraints, simulations, databases and experiments, and they do not change because the model changed. What distinguishes ScienceRSI from AI4AI in general is therefore not the mechanism but the evaluator. In open-domain reasoning the evaluator is usually another model; in science it can be the world.

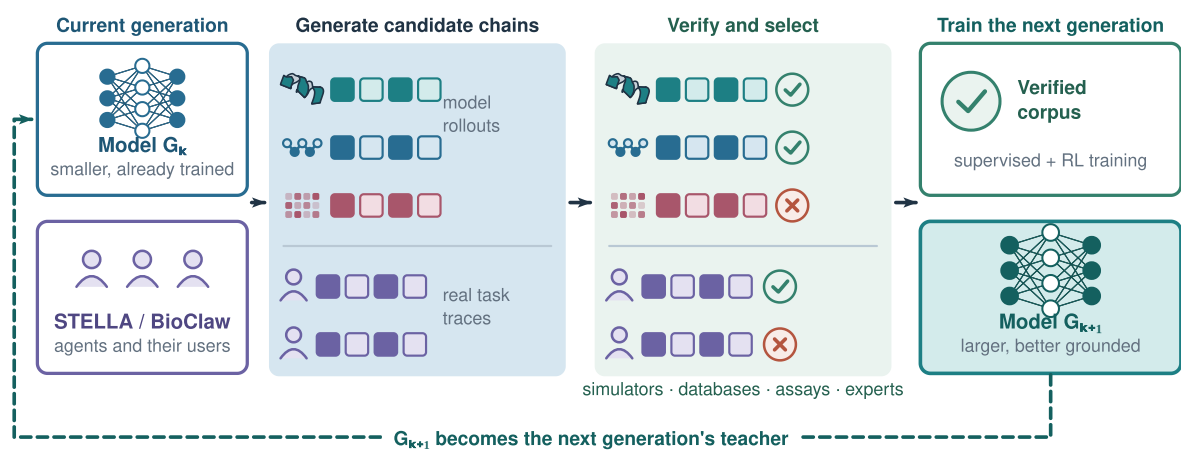


Figure 3. The ScienceRSI loop. The AI4AI pattern, in which one generation of models produces the training signal for the next, carried over to scientific reasoning. Generation G_k and the agent platforms built on it produce candidate chains of science tokens and traces from real tasks. External evaluators (simulators, databases, assays and domain experts) accept or reject each candidate and label the accepted ones with the evidence state they can support. The verified corpus trains a larger generation G_{k+1} by supervised and reinforcement learning whose reward comes from the evaluators, and G_{k+1} then becomes the producer for the next round. Rejected traces are kept as negative examples. Repeating the loop across generations is what carries a seed model towards the test window of Section 3.2.

This matters for AI4S because the data required are missing in a specific way. The binding constraint on reaching the window of Section 3.2 is not parameters or compute, both of which are within demonstrated practice, but *reasoning data*. Section 3.3 counted encoded scientific objects: genomes, structures, cells, images, measurements. A chain of science tokens needs something those archives do not contain, namely trajectories in which text and scientific states inform one another step by step, each step checked. Textual reasoning data can be mined at web scale; interleaved scientific trajectories cannot, because they were never written down. They have to be produced, and the only affordable producers are models and agents. Organizing that production as a loop across generations buys two things. It reduces how much the programme depends on scientific reasoning data written by people, which is the scarcest input and the one that cannot be bought at scale. And it shortens the

iteration: the generator improves every round, so the cost of obtaining the next verified corpus falls while the models that consume it grow.

Box 3. Key terms: ScienceRSI.

Recursive self-improvement (RSI). An AI system taking part in building its successor. Its practical form today is <i>AI4AI</i>: one generation of models generates, filters, rewards and evaluates the data that train the next generation, so capability compounds across generations. ScienceRSI. Science Recursive Self-Improvement: AI4AI applied to scientific reasoning models, with scientific evaluators external to the model deciding which generated chains of science tokens survive and what evidential status they carry. Generation. One model trained on the verified output of its predecessor. G_0 is seeded from literature, simulators and expert demonstrations. Verified trace. A reasoning trajectory retained after external screening, carrying the evidence state its evaluator supports and the identity of that evaluator. Verifier coverage. The fraction of generated claims that any available evaluator can decide. It bounds what the loop can learn. Verified yield. The fraction of generated candidates that survive screening. Its change across generations is the loop’s primary health signal. Bounded versus open-ended. ScienceRSI is bounded: the criteria of improvement are external evidence and stay fixed as the model changes. It does not propose a system that rewrites its own objectives.

4.2. The Loop

Generate. The current generation G_k produces candidate chains of science tokens across modalities: structures and their checks, sequence edits and their consequences, analyses of omics or measurement data. In parallel, agent platforms supply trajectories from real work. Our STELLA and BioClaw agents place a model in contact with real tasks, real tools and real users, so their traces carry a declared goal, an executed tool call, a dataset and sometimes an experiment, rather than a prompt invented for training [30,31]. These two sources are complementary: rollouts give coverage and volume, agent traces give grounding and task realism.

Verify and select. Candidates are screened by evaluators that are external to the model: physical and geometric constraints, simulators, database and literature lookups, executable tools, laboratory assays where they exist, and domain experts where they do not. Screening does more than filter. It labels each surviving trace with the evidence state its evaluator can support (predicted, simulated, measured or independently replicated) in the sense defined in Section 5, and records which evaluator made the decision. Rejected traces are retained as negative examples rather than discarded.

Train. The verified corpus trains generation G_{k+1} , which is larger than G_k , through supervised learning on accepted trajectories and reinforcement learning whose reward comes from the verifier rather than from a model’s own judgement. The distinction matters: a reward model trained to predict the verifier is a convenience, but it inherits the verifier’s errors and adds its own.

Promote. G_{k+1} becomes the producer for the next round, and the loop repeats at a larger scale and, if the loop works, a higher yield of verifiable traces. Agent platforms are upgraded to the new generation, so the quality of the task traces rises with the model.

4.3. Verification Is the Ceiling

The standard objection to recursive self-improvement is that a loop which grades itself degrades: model collapse and self-confirming bias accumulate, and the ceiling is whatever the evaluator can actually decide (Section 4.1). We think the objection is correct, and it is the reason AI4AI loops in open domains tend to plateau. Our claim is that science is the setting where it bites least, because scientific claims come with evaluators that are not models: a geometry either satisfies a constraint or does not, a simulation either converges or does not, a sequence either appears in a database or does not, an assay either reproduces or does not. The asymmetry between generating a candidate and checking it is what makes the loop plausible here and implausible for open-ended reasoning in general.

Three rules follow. First, the model proposes and something that is not the model decides; where no external evaluator exists, the claim is retained as unverified rather than used as supervision. Second, a trace may support training only for claims at or below the evidence state its evaluator can establish: a successful simulation does not license a claim about measured activity. Third, provenance is recorded so that later generations can be decontaminated against the benchmarks used to declare an SSI crossover; a loop that trains on its own traces must not be evaluated on tasks those traces contaminate.

4.4. How We Would Know It Is Working

The loop makes predictions that can fail. Three quantities should be reported for every generation. *Verified yield*: the fraction of generated candidates that survive screening. If it does not rise, later generations are producing more text but not more evidence. *Verifier coverage*: the fraction of generated claims that any evaluator can decide. If it is low, the loop is mostly training style, and the remedy is better evaluators rather than a larger model. *Held-out gain*: the improvement on specialist and general-science tasks that were withheld from every generation, with the contribution of self-generated data separated from that of newly acquired external data. A generation that improves none of the three ends the loop, and that outcome would be evidence against the proposal rather than a reason to add another round.

Two risks deserve explicit monitoring. Diversity can narrow as each generation trains on the previous one's accepted output, so the loop needs a continuing supply of external data and tasks; the agent platforms are one source of it precisely because their tasks are set by users rather than by the model. And optimization can target the evaluator rather than the science when the two come apart, which is the general failure mode of outcome-based training noted in Section 2.3 and is why evaluator design, not reward engineering, is the substantive work.

5. Science-Token Economics

5.1. Token Economics Today

Language-model services made computation easy to meter and to buy: input and output tokens are counted and charged at a rate, and reasoning models add the tokens they generate while thinking. This token economics rests on three implicit assumptions: the token is the unit of account, value scales with use, and verification is left to the user. For text these are workable approximations. For scientific outputs these approximations can be insufficient: a 300-residue enzyme design costs about the same number of tokens whether it is a hallucination or a working catalyst.

5.2. Two Units of Account and an Evidence Lifecycle

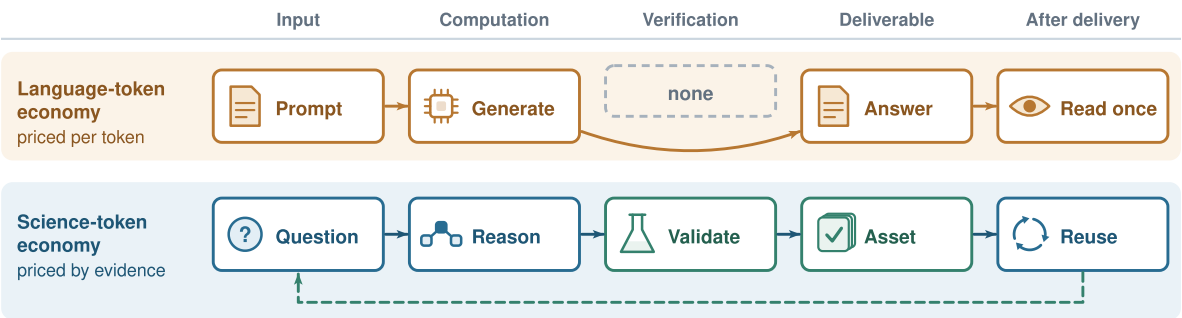
Science-token economics separates two units that should not be conflated (Box 4; Figure 4). An *elementary model token* is a unit processed during model training or inference. It meters computation and representation, but token count alone does not establish whether an output is correct, experimentally supported or reusable. A *scientific asset* is not a larger or more expensive model token. It is a persistent, versioned bundle containing one or more science-token-encoded objects together with context, units, conditions, evidence state, provenance, uncertainty, validation records and reuse rights. The same encoded molecule or enzyme motif can therefore incur the same token-level compute cost while supporting assets of different evidential status and downstream utility.

Science-token economics is therefore a proposed framework for separating computational accounting from the evaluation and exchange of reusable scientific outputs. It is not an existing market, and it assumes no universal exchange rate for evidence. The intended asset layer includes a protocol with its validation record, or a molecule linked to its synthesis, measurements and applicability limits. Assets extend the principle that scientific digital objects should be findable, accessible, interoperable and reusable [32]; tokenization makes them machine-operable, while evidence and metadata bound their claims. The proposed value of an asset lies in what another model, laboratory or development pro-

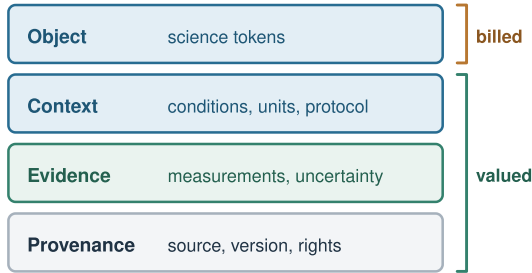
gramme can reliably do with it. Any implementation would require domain-specific validation criteria, governance and contractual arrangements, and its usefulness would need to be tested empirically.

Assets move through a lifecycle (Figure 4a, lower lane): generation by models, experts, tools or experiments; validation by database checks, computation, expert review or wet-laboratory measurement, each explicitly labelled; and curation, reuse, composition and licensing across repositories and development pipelines. Validation records, including negative results that prevent repeated failures, are themselves reusable, and each cycle seeds the next question. A predicted motif and an experimentally characterized enzyme therefore occupy different evidence states even when their encodings have the same length (Figure 4c).

a Two token economies



b Anatomy of a scientific asset



c Same tokens, different value



Figure 4. Science-token economics. **a**, In language-token economics, price follows the number of tokens and verification is left to the user. In science-token economics, reasoning produces candidates, validation assigns an evidence state, and validated assets are reused across models and laboratories; new evidence, including negative results, seeds the next question (dashed). **b**, A scientific asset bundles science tokens with context, evidence and provenance. Token count bills only the object layer; the value resides in the layers that accompany it. **c**, The same encoded enzyme motif costs the same number of tokens at every evidence state, but its scientific value rises from prediction to simulation, measurement and independent replication. All panels are schematic and illustrate a proposed framework, not a measured market.

Actors and transactions. Assets may be generated by models, individual researchers, laboratories or automated platforms. Validation is performed by declared evaluators, which may include computational services, domain experts, qualified laboratories or independent replicating groups. Repositories or institutional stewards curate the object, metadata, access conditions, version history and links to validation records. Ownership does not arise from tokenization: the relevant rights remain with the parties designated by applicable law and contracts, and should be recorded separately from scientific provenance. A rights holder or authorized repository may license access or sell a validated deliverable to research groups, companies, funders or public institutions. Reusers must comply with declared terms, preserve provenance and attach subsequent evidence as a new version rather than silently overwriting the original asset.

Evidence-state transitions. An output is *predicted* when it is generated by a model and accompanied by model provenance and uncertainty but no external evaluation. It becomes *simulated* when evaluated

using a declared computational method with versioned software, settings and acceptance criteria. It becomes *measured* when linked to a physical experiment with a protocol, controls, data, uncertainty and quality-control records. *Independently replicated* requires a repeat by a separate team, laboratory or independent sample under declared criteria. These transitions are recorded by the responsible validator rather than inferred from the token sequence; negative, inconclusive or contradictory results remain attached to the asset and may revise its status.

Box 4. Key terms: science-token economics.

Token economics. Pricing AI services by metered tokens: input and output tokens times a rate.
Science-token economics. A proposed framework that accounts separately for model computation and for reusable scientific assets; it is not an established market or pricing standard.
Elementary model token. A unit processed during training or inference. It meters computation and representation but has no evidential status of its own.
Scientific asset. A persistent, versioned bundle of science-token-encoded objects with context, evidence, provenance, validation records and reuse rights: for example, a characterized enzyme with its assay record.
Evidence state. The level of support behind an asset: predicted, simulated, measured or independently replicated.
Scientific value density. The useful, verifiable, decision-relevant information an asset delivers, relative to the full cost of obtaining and applying it.

Illustrative lifecycle: an enzyme candidate. Consider a model-generated enzyme sequence proposed to catalyse a specified reaction. At the prediction stage, the candidate is accompanied by the generating model and version, input conditions, confidence estimates and known applicability limits. A computational group may then evaluate its predicted fold, active-site geometry and stability using declared simulation methods, producing a simulated evidence record rather than proof of function. A laboratory subsequently synthesizes the sequence and measures expression, activity and specificity under documented experimental conditions. If a separate laboratory reproduces the result using an independent sample and declared acceptance criteria, the candidate acquires independently replicated status. The sequence, provenance, uncertainty, simulation settings, protocols, measurements, negative results, validation history and permitted reuse terms can then be curated as a versioned, reusable scientific asset (Figure 4c). Payment may separately cover model computation, computational evaluation, laboratory validation and independent replication; subsequent access fees or licences compensate access to the validated deliverable and its specified reuse rights. These payments concern particular services, evidence and rights, not a universal conversion from model-token count to scientific value.

No universal conversion rate. No fixed formula maps a number of model tokens to scientific value. Model tokens can meter computation, whereas the potential value of an asset depends on evidence strength, decision relevance, novelty, transferability, applicability limits, rights, and the costs of acquisition and validation. Licences, access fees, milestone payments or public funding could compensate different contributions, but these mechanisms would price particular rights or deliverables rather than convert tokens into a universal scientific currency.

Relation to existing approaches. The proposed framework combines elements of several established mechanisms but is not equivalent to any of them. Ordinary API pricing meters input and output tokens and charges for model use, but token billing does not by itself certify the correctness or downstream utility of an output [33]. Scientific-data licences specify permissions and contractual conditions for access and reuse, whereas they do not establish whether a result has been simulated, measured or independently replicated [34]. FAIR principles and repositories make digital research objects more findable, accessible, interoperable and reusable, but they do not on their own assign the evidence states proposed here [32]. Milestone-based research contracts can tie funding to negotiated technical metrics, completion criteria and deliverables, but those criteria apply within a particular agreement and do not necessarily create a portable, versioned asset for later reuse [35]. Knowledge graphs organize entities and relations for integration and query, while data marketplaces provide infrastructure through which providers and consumers can exchange data [36,37]; neither mechanism necessarily requires a transition from prediction to simulation, measurement and independent replication. Science-token

economics is proposed as a complementary asset layer in which token billing records computational cost, repositories and knowledge graphs support discovery and provenance, and licences, marketplaces or contracts govern exchange, while domain-specific evidence states, uncertainty, conditions and reuse rights remain attached to the scientific output.

5.3. From Token Volume to Value Density

This is the central difference from ordinary token economics: what should be priced is not the number of tokens but the evidence they carry (Table 2). We define *scientific value density* as the useful, verifiable, decision-relevant information an asset delivers relative to the full cost of obtaining and applying it. Its dimensions include novelty, reliability, decision relevance and reuse across tasks; we assume no universal scalar or exchange rate. One measurement that resolves a key mechanistic uncertainty can be worth more than millions of redundant encoded positions. Science-token economics would combine metered computation with licensed evidence, payment for validated deliverables and payments tied to predefined research milestones, with standard asset interfaces that make assets easy to compare, compose and exchange. It requires no cryptocurrency.

Table 2. Language-token economics and science-token economics compared.

	Language-token economics	Science-token economics
Unit of account	Model token, input and output	Validated scientific asset: object, context and evidence
What is metered	Tokens processed	Computation, simulation and experiment, reported separately
Basis of value	Volume of use	Value density: verified, decision-relevant evidence per unit cost
Verification	Outside the product; left to the user	Inside the product; evidence state labelled from predicted to replicated
Life of output	Consumed within a session	Durable, versioned, reusable and composable
Negative results	Discarded	Recorded as assets that prevent repeated failures
Settlement	Price per token	Metered compute, licensing, validated delivery, milestone payments
Principal risk	Paying for fluent but wrong output	Attribution, endpoint definition and independent verification

Computation remains a cost even when discovery becomes the unit of value: a short structural output may require extensive sampling, and experimental verification adds material, labour, quality-control and delay costs. Additional reasoning, higher-fidelity simulation and a new measurement should be compared by their expected improvement in the scientific decision relative to full cost, which is the value-density calculation applied before the work is done. Outcome-based contracts need agreed endpoints, attribution and independent verification.

Testing the framework. The value of evidence-labelled assets should be tested prospectively rather than assumed. In matched research tasks, participating teams could receive either conventional outputs or the same outputs packaged as structured assets containing evidence state, provenance, uncertainty, experimental conditions, validation records and reuse rights. Pre-registered endpoints should ask whether the structured assets: (i) increase independent replication rates and agreement between laboratories; (ii) reduce avoidable repetition of experiments that add no new decision-relevant evidence; (iii) shorten the time, number of iterations and cost required to reach predefined downstream decisions; and (iv) increase the number of candidates reaching measured or independently replicated status per unit of total expenditure. The cost calculation should include model computation, curation, expert review, laboratory work and independent validation. Label errors, curatorial overhead and false confidence should also be recorded as possible costs. Results should be reported separately by

scientific domain and evidence state because the framework assumes no universal measure of scientific value. Failure to improve these endpoints would count as evidence against the proposed framework.

6. Outlook: A Testable Route Towards Scientific Super Intelligence

A scientific vocabulary gives heterogeneous scientific objects a shared computational interface, and chains of science tokens make their transformations part of reasoning, with language guiding and being informed by scientific states. Together they define a concrete programme: train cross-modal reasoning processes and test that they use their evidence; measure the joint SSI crossover against strong, fixed baselines along a pilot ladder towards the test window of 100–300 billion parameters and 10–30 trillion tokens; and connect model development to experiments that generate informative new evidence. ScienceRSI is our proposed route along that ladder: AI4AI with the world, rather than another model, as the evaluator. It is falsifiable in the same way: if verified yield, verifier coverage and held-out gain do not improve from one generation to the next, the loop has not compounded and the proposal fails. Controlled scaling studies can narrow, revise or reject that window. An economy built around reusable, verifiable scientific assets can sustain this progress by rewarding knowledge that later models and laboratories can use.

References

1. Wei, J.; et al. Chain-of-thought prompting elicits reasoning in large language models. Preprint at <https://arxiv.org/abs/2201.11903>, 2022.
2. Jumper, J.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, 596, 583–589. <https://doi.org/10.1038/s41586-021-03819-2>.
3. Abramson, J.; et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **2024**, 630, 493–500. <https://doi.org/10.1038/s41586-024-07487-w>.
4. Brix, G.; et al. Genome modelling and design across all domains of life with Evo 2. *Nature* **2026**, 652, 1349–1361. <https://doi.org/10.1038/s41586-026-10176-5>.
5. Hayes, T.; et al. Simulating 500 million years of evolution with a language model. *Science* **2025**, 387, 850–858. <https://doi.org/10.1126/science.ads0018>.
6. Fang, A.; et al. When do biological reasoning models use their biological inputs? Preprint at <https://arxiv.org/abs/2610.00898>, 2026.
7. Pourmirzaei, M.; Morehead, A.; Esmaili, F.; Ren, J.; Pourmirzaeiolaei, M.; Xu, D. GCP-VQVAE: a geometry-complete language for protein 3D structure. Preprint at bioRxiv, doi: 10.1101/2025.10.01.679833, 2025.
8. Gao, K.; Wang, Y.; Guan, H.; Wang, Z.; Pei, Q.; Hopcroft, J.E.; He, K.; Wu, L. Tokenizing 3D molecule structure with quantized spherical coordinates. Preprint at <https://arxiv.org/abs/2412.01564>, 2024.
9. Zhou, C.; et al. Transfusion: predict the next token and diffuse images with one multi-modal model. In Proceedings of the International Conference on Learning Representations (ICLR), 2025.
10. Boiko, D.A.; MacKnight, R.; Kline, B.; Gomes, G. Autonomous chemical research with large language models. *Nature* **2023**, 624, 570–578. <https://doi.org/10.1038/s41586-023-06792-0>.
11. Hao, S.; et al. Training large language models to reason in a continuous latent space. In Proceedings of the Conference on Language Modeling (COLM), 2025.
12. Burger, B.; et al. A mobile robotic chemist. *Nature* **2020**, 583, 237–241. <https://doi.org/10.1038/s41586-020-2442-2>.
13. Schaeffer, R.; Miranda, B.; Koyejo, S. Are emergent abilities of large language models a mirage? Preprint at <https://arxiv.org/abs/2304.15004>, 2023.
14. Rein, D.; et al. GPQA: a graduate-level Google-proof Q&A benchmark. Preprint at <https://arxiv.org/abs/2311.12022>, 2023.
15. Laurent, J.M.; et al. LAB-Bench: measuring capabilities of language models for biology research. Preprint at <https://arxiv.org/abs/2407.10362>, 2024.
16. Arc Institute. Tahoe open sources Tahoe-100M as the inaugural contribution to Arc Institute’s Virtual Cell Atlas, 2025. Accessed 4 October 2026.
17. Protein Data Bank in Europe. AlphaFold Database release notes, version 6 (21 October 2025), 2025. Accessed 4 October 2026.

18. Broad Institute Imaging Platform. Cell Painting Gallery complete datasets, entry cpq0016-jump, 2026. Accessed 4 October 2026.
19. Kim, S.; et al. PubChem 2025 update. *Nucleic Acids Research* **2025**, *53*, D1516–D1525. <https://doi.org/10.1093/nar/gkae1059>.
20. Levine, D.S.; et al. The Open Molecules 2025 (OMol25) dataset, evaluations, and models. Preprint at <https://arxiv.org/abs/2505.08762>, 2025. Version 2, 2026.
21. RCSB Protein Data Bank. 2025 annual report. Archive contents as of 31 December 2025, 2026.
22. Qwen Team. Qwen2.5 technical report. Preprint at <https://arxiv.org/abs/2412.15115>, 2024.
23. Qwen Team. Qwen3 technical report. Preprint at <https://arxiv.org/abs/2505.09388>, 2025.
24. DeepSeek-AI. DeepSeek-V4.1-Flash: pushing the limits of KV cache compression. Technical report and model card, <https://huggingface.co/deepseek-ai/DeepSeek-V4.1-Flash>, 2026. Accessed 7 October 2026.
25. Hoffmann, J.; et al. Training compute-optimal large language models. Preprint at <https://arxiv.org/abs/2203.15556>, 2022.
26. Yuan, D.; et al. The European Nucleotide Archive in 2025. *Nucleic Acids Research* **2026**, *54*, D120–D127. <https://doi.org/10.1093/nar/gkaf1295>.
27. Gelman, S.; et al. Biophysics-based protein language models for protein engineering. *Nature Methods* **2025**, *22*, 1868–1879. <https://doi.org/10.1038/s41592-025-02776-2>.
28. Szymanski, N.J.; et al. An autonomous laboratory for the accelerated synthesis of inorganic materials. *Nature* **2023**, *624*, 86–91. Corrected 2026, <https://doi.org/10.1038/s41586-023-06734-w>.
29. Chen, M.; Wang, L.; Qu, B. Recursive self-improvement in AI: from bounded self-refinement to autonomous research loops. Preprint at <https://arxiv.org/abs/2607.07663>, 2026.
30. Jin, R.; Zhang, Z.; Wang, M.; et al. STELLA: self-evolving LLM agent for biomedical research. Preprint at <https://arxiv.org/abs/2507.02004>, 2025.
31. Xu, M.; Yan, J.; Feng, R.; et al. BioClaw: human–bot research collaboration ecosystems in group chats. Preprint at bioRxiv, doi: 10.64898/2026.04.11.716807, 2026.
32. Wilkinson, M.D.; et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **2016**, *3*, 160018. <https://doi.org/10.1038/sdata.2016.18>.
33. OpenAI. API pricing. <https://developers.openai.com/api/docs/pricing>, 2026. Accessed 8 October 2026.
34. Carroll, M.W. Sharing research data and intellectual property law: a primer. *PLOS Biology* **2015**, *13*, e1002235. <https://doi.org/10.1371/journal.pbio.1002235>.
35. Defense Advanced Research Projects Agency. Proposer instructions: other transactions. <https://www.darpa.mil/about/offices/contracts-management/proposer-transactions>, 2026. Accessed 8 October 2026.
36. Hogan, A.; Blomqvist, E.; Cochez, M.; et al. Knowledge graphs. *ACM Computing Surveys* **2021**, *54*, 1–37. <https://doi.org/10.1145/3447772>.
37. Stahl, F.; Schomm, F.; Vossen, G.; Vomfell, L. A classification framework for data marketplaces. *Vietnam Journal of Computer Science* **2016**, *3*, 137–143. <https://doi.org/10.1007/s40595-016-0064-2>.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.