How are scientific concepts birthed? Typing rules of concept formation in theoretical physics reasoning

Abstract

This work aims to formalize some of the ways scientific concepts are formed in the process of theoretical physics discovery. Since this may at first seem like a task beyond the scope of the exact sciences, we begin by presenting arguments for why scientific concept formation can be formalized. Then, we introduce type theory as a natural framework for this formalization. We formalize what we call "ways of discovering new concepts" including property preservation and concept change, as cognitive typing rules. Next, we apply these cognitive typing rules to a case study of conceptual discovery in the history of physics: Einstein's conceptual path to the relativity of time. Then, we recast what a physicist might informally call "ways of discovering new scientific concepts" as compositional typing rules built from cognitive typing rules—thus formalizing them as scientific discovery mechanisms. Lastly, we computationally model the type-theoretic reconstruction as a program synthesis task.

Introduction

The standard picture of the scientific method often emphasizes hypothesis testing. That picture leaves a prior question underspecified: how are new scientific concepts formed? One of the first concrete answers is Peirce's view that concept formation is abduction: the process of imposing constraints on admissible hypotheses to generate plausible, though fallible, explanations (Chomsky 2006). Building on Peirce, more recent work in cognitive science of science treats scientific concepts as entities shaped by the constraints of the knowledge-construction practices scientists use to generate them (Nersessian 2008; Gentner 1983). Perhaps the most detailed engagement between cognitive science of science and natural science comes from Nancy Nersessian's *cognitive-historical analysis* of theoretical physics. It treats historical records of theoretical physics practice as empirical evidence and reads them as traces of cognitive mechanisms such as analogy, mental modeling, and thought experiments (Nersessian 1987).

Since the discovery methods Nersessian reveals are systematic procedures with definite and finite steps, one may ask:

Can concept formation in theoretical physics be formalized?

To address this question, one may be inclined to explore machine learning approaches that have successfully automated other aspects of scientific discovery (Battaglia et al. 2018; Udrescu et al. 2020; Lalande et al. 2023) , with the aim of adapting these techniques to model scientific concept formation. However, these methods still face the challenge of modeling conceptual development, as they were built to solve predefined problems rather than define new ones (Battleday and Gershman 2024). This suggests that we should take a step back to clarify what has so far been left vague: what we mean by constraints in scientific concept formation.

As part of this effort, physicists might recall that when searching for new scientific concepts, they often use the word "type" or "kind" informally to describe what they are looking for. For instance, Maxwell initially conceived of the magnetic field as a "type" or "kind" of stress composed of pressures and tensions (Maxwell 1865). These *intuitive types* are not necessarily physically correct claims, but rather common-sense constraints that help narrow the space of possible concepts moving forward. Their intuitiveness stems from the fact that they are used most frequently in everyday human reasoning (e.g., "What time is it?" calls for a time, not a food).

Felix Sosa has recently proposed a natural formalization of these intuitive types as *theoretical types* which are the building blocks of *type theory* (Sosa and Ullman 2022). This makes one wonder: if intuitive types used in everyday thought—like "bread" or "dog"—are slippery yet still conceivable as theoretical types, what about those used in scientific reasoning, such as "has 3 distinct components" which are already well-defined? This brings us to the central question of this paper:

Can concept formation in theoretical physics be formalized using type theory?

Our answer is yes. We substantiate this claim through a type-theoretic formalization of scientific concept formation driven by detailed reconstructions of scientific reasoning from the history of physics. In this framework, intuitive types of scientific concepts are formalized as types and different "ways of discovering scientific concepts" as typing rules—thereby recasting them as *discovery mechanisms*. Moreover, scientific concepts are formalized both as *terms*

and as the results of *typing rules*—thus defining them by how they are discovered, in the spirit of Peirce and cognitive science of science.

Lastly, a type system can be understood as a mathematical formalization of a programming language (Cardelli 1996), where terms correspond to programs and types specify how those programs may be used. This perspective allows us to recast scientific discovery as *program synthesis*: the inference of programs from data and constraints (Rule, Tenenbaum, and Piantadosi 2020). To illustrate this idea, we will carry out type-theoretic reconstructions of a canonical example of concept formation from the history of physics, and computationally implement it using program synthesis.

Overview

This work aims to provide a type-theoretic formalization of some ways in which scientific concepts are formed in theoretical physics, based on reasoning patterns reconstructed from the history of theoretical physics. To this end, the paper undertakes the following recastings:

1. Intuitive types are formalized as theoretical types.
2. "Ways of discovering new concepts"—such as property preservation and concept change—are formalized as elementary typing rules, which we call cognitive typing rules.
3. "Ways of discovering new scientific concepts" are formalized as compositional typing rules that integrate cognitive typing rules with functional and algebraic operations. In doing so, these "ways" are recasted as discovery mechanisms.
4. Scientific concepts are formalized both as terms and as the outcomes of discovery mechanisms.

The formalism we develop is applied to a paradigmatic conceptual breakthrough in the history of physics: Einstein's special relativity. Specifically, we examine a key episode in Einstein's reasoning that paved the way to his discovery. Namely, the conceptual path that led him to relative time. Einstein's reasoning leading to the relativity of time is formalized as a discovery mechanism that reverses the roles of assumption and conclusion through a composition of property preservation and concept change typing rules with functional and algebraic operations; we call this the `assumption–conclusion switch` discovery mechanism. Lastly, we use typed-program synthesis to simulate Einstein's conceptual path to the relativity of time.

Related work

Computational scientific discovery, or AI for science, spans symbolic and deep-learning approaches, but much of it optimizes within fixed problem formulations rather than modeling the conceptual work of forming new representations and questions (Battleday and Gershman 2024). LLM-based AI scientist systems can propose hypotheses in natural language, yet their outputs are often produced by opaque procedures and lack formal representations that let concepts be precisely defined and related, leaving them vulnerable to ambiguity and meaning overload (Liu et al. 2023; Wolfram 2024). This work targets this gap.

Methods

Type theory

Type theory is a branch of mathematical logic that classifies mathematical objects (terms) by assigning them types, which constrain how terms can be used and combined (Pierce 2002). It can also be regarded as a formalization of programming languages (Pierce 2002). Here we extend type theory to model concept formation in theoretical physics reasoning by treating scientific concepts as terms and their properties (intuitive types) as types.

Terms, and types. A *term* is an expression (sequence of symbols) that carries semantic content in the system, so it can be assigned a type. A *type* classifies terms and constrains their use.

Contexts. A context Γ is a collection of assumptions under which an assertion is made.

Typing judgments. A typing judgment asserts that, under context Γ, a term t has type A:

$$\Gamma \vdash t : A. \tag{1}$$

Typing rules. Typing judgments are defined inductively as a collection of typing rules. Each rule derives a conclusion judgment from premise judgments. Schematically:

$$\frac{\text{premise}_1 \ \cdots \ \text{premise}_n}{\text{conclusion}}. \tag{2}$$

Formally:

$$\frac{\Gamma_1 \vdash t_1 : A_1 \quad \ldots \quad \Gamma_n \vdash t_n : A_n}{\Gamma \vdash t : A} \tag{3}$$

where Γ_i is a local extension of Γ specific to the i^{th} premise

Intersection type. The intersection type $A \cap B$ combines types A and B; a term of type $A \cap B$ has both types A and B.

Cognitive typing rules

Forward property preservation rule.

$$\frac{\Gamma \vdash x : A}{\Gamma \vdash f(x) : A} \tag{4}$$

Under this rule, if a concept x has property A, then applying f preserves that property.

Backward property preservation rule.

$$\frac{\Gamma \vdash f(x) : A}{\Gamma \vdash x : A} \tag{5}$$

Under this rule, if a transformed concept $f(x)$ has property A, then the source concept x must also have property A.

Concept change rule. This rule states that if two terms, x and y, are identical but carry different types under different contexts, then in the combined context, the term x is assigned the intersection type.

$$\frac{\Gamma_1 \vdash x : A \quad \Gamma_2 \vdash y : B \quad x = y}{\Gamma_1 \cup \Gamma_2 \vdash x : A \cap B} \tag{6}$$

Results

The Michelson–Morley experiment is often regarded as one of the crucial optical experiments that Einstein sought to explain in his effort to extend the principle of relativity to electrodynamics. However, according to Einstein himself, that was not the case; instead, the experiment of stellar aberration played the most critical role. In particular, his reinterpretation of Lorentz's local time, in Lorentz's stellar aberration account, as empirical time led him to the relativity of time.

Norton (Norton 2004) argues that reading Lorentz's account of stellar aberration through the lens of relativity could have led Einstein to conclude that stellar aberration provided experimental support for the physical reality of local time, and thus that it could be regarded simply as time. This line of reasoning can be viewed as a discovery mechanism that swaps assumption and conclusion, which is described in natural language and type-theoretic formalism below.

Assumption-conclusion switch discovery mechanism

Step 1. Lorentz regarded the concept of local time $t' = t - \frac{ux}{c^2}$, as an artificial mathematical coordinate without physical interpretation, yet one that was useful for explaining stellar aberration (Lorentz 1895, p. 77). This qualitative remark is formalized by assigning t' the type $\texttt{art}$ (artificial construct):

$$t' : \texttt{art} \tag{7}$$

Step 2. Now, a Lorentz transformation applied to stellar aberration yields a light waveform in the Earth's rest frame:

$$f(\omega t - ky) \;\rightarrow\; f\Big(\omega\Big(t - \frac{ux}{c^2}\Big) - ky\Big) \tag{8}$$

Here, Einstein might have wondered: what does the artificial nature of local time imply for the transformed wave? The simplest answer—and the one Einstein seems to have adopted—is that anything derived from an artificial construct is itself artificial. This reasoning can be formalized by applying the forward property preservation rule to t', which yields:

$$f\Big(\omega\Big(t - \frac{ux}{c^2}\Big) - ky\Big) : \texttt{art} \tag{9}$$

Step 3. Simplifying expression and using $k = \frac{\omega}{c}$, we have

$$f\Big(\omega t - k\frac{ux}{c} - ky\Big) : \texttt{art} \tag{10}$$

Step 4. Note that we can summarize the previous steps as follows: Under the assumption that local time is an artificial construct, the transformed waveform must also be artificial. Thus, under the context $\Gamma_1 = t' : \texttt{art}$, we have

$$\Gamma_1 \vdash f_{\text{art}} : \texttt{art} \tag{11}$$

where

$$f_{\text{art}} = f\Big(\omega t - kx\frac{u}{c} - ky\Big)$$

Step 5. On the other hand, Einstein recognized that, in the Earth's rest frame, the light wave deflected by stellar aberration $f(\omega t - kx\frac{u}{c} - ky)$ is an empirical observation that does not rely on theoretical assumptions. To formalize this reasoning, we state that, under the empty context $\Gamma_2 = \varnothing$, the deflected light wave has type $\texttt{emp}$ (empirical):

$$\Gamma_2 \vdash f_{\text{emp}} : \texttt{emp} \tag{12}$$

where

$$f_{\text{emp}} = f\Big(\omega t - k\frac{ux}{c} - ky\Big)$$

Step 6. Here, Einstein would have recognized that although f_{art} is artificial, it is also equal to f_{emp}, which is empirical. According to Norton's analysis, this suggests that, at the very least, f_{art} could not be solely artificial but must also possess an empirical aspect. Thus, within our formalism, Einstein would have concluded that f_{art} has both types: $\texttt{art}$ and $\texttt{emp}$. This reasoning process can be formalized using the concept change rule:

$$\frac{\Gamma_1 \vdash f_{\text{art}} : \texttt{art} \quad \Gamma_2 \vdash f_{\text{emp}} : \texttt{emp} \quad f_{\text{art}} = f_{\text{emp}}}{\Gamma_1 \cup \Gamma_2 \vdash f_{\text{art}} : \texttt{art} \cap \texttt{emp}} \tag{13}$$

where

$$\Gamma_1 = \{t' : \texttt{art}\} \quad \text{and} \quad \Gamma_2 = \varnothing$$

Step 7. At this point, according to Norton (Norton 2004), Einstein reasons that since his earlier conclusion is not merely artificial but also empirically true, he can now reinterpret it as an assumption. In terms of our formalism, Einstein assumes that f_{art} is of both types: $\texttt{art}$ and $\texttt{emp}$. As a result, he begins to derive the previous steps in reverse order, while cautiously maintaining that the same type ($\texttt{art} \cap \texttt{emp}$) continues to propagate backwards.

Eventually, by reasoning backwards, Einstein reaches a new version of his former assumption, now framed as a conclusion: local time t' has type $\texttt{art} \cap \texttt{emp}$. That is, local time could not be wholly artificial; it had to be, at least in part, empirically true. In other words, local time could be time itself, not merely an artifice. Einstein's reasoning process can be formalized using a backward property preservation rule as follows:

$$\frac{\Gamma_1 \cup \Gamma_2 \vdash f_{\text{art}} : \texttt{art} \cap \texttt{emp}}{\Gamma_1 \cup \Gamma_2 \vdash t' : \texttt{art} \cap \texttt{emp}} \tag{14}$$

Because local time depends on relative velocity, it was at this moment that Einstein realized time could be relative. That realization launched special relativity.

Step 8. Ultimately, Einstein's full discovery mechanism of switching Lorentz' assumption and conclusion can be formalized using the following rule:

$$\frac{\Gamma_1 \vdash t' : \texttt{art} \quad \Gamma_2 \vdash f_{\text{emp}} : \texttt{emp} \quad g(t') = f_{\text{emp}}}{\Gamma_1 \cup \Gamma_2 \vdash t' : \texttt{art} \cap \texttt{emp}} \tag{15}$$

That is, under assumptions from contexts Γ_1 and Γ_2, the concept of local time t' is assigned both types: `art` and `emp`. This rule builds compositionally upon the concept change and property preservation rules along with other simple mathematical operations.

The assumption-conclusion switch discovery mechanism can be stated in its most general form, without referring to the example's specific variable names, as follows:

$$\frac{\Gamma_1 \vdash x : A \quad \Gamma_2 \vdash y : B \quad g(x) = y}{\Gamma_1 \cup \Gamma_2 \vdash x : A \cap B} \tag{16}$$

where g denotes the function that transforms x into y.

Bridge from Type theory to Computation

To carry the type-theoretic formalism into a computational setting, we examine the three-way relationship among type theory, programming languages and scientific concepts we introduced earlier. We begin with the connection between type theory and programming languages. Specifically, we note that once a type system is equipped with fixed operational semantics (step-by-step evaluation rules), the system's terms can be viewed as programs and its theoretical types as programming-language types (Plotkin 1977).

Turning to the link between programming languages and scientific concepts, recent work in computational cognitive science holds that symbolic programs (code) offer the best formal representation of concepts because of their expressive power (Rule, Tenenbaum, and Piantadosi 2020). As noted in earlier, Sosa builds on this perspective by suggesting that types in programming languages may capture the cognitive constraints that impose structure in concepts and allow us to generate reasonable, though sometimes incorrect, answers (Sosa and Ullman 2022). Motivated by these three developments, we introduce two analogies—one theoretical and one computational—for formalizing scientific concept formation.

1. **Theoretical analogy:** Intuitive types are formalized as theoretical types; discovery mechanisms as typing rules; and scientific concepts as terms and the conclusions of typing rules.
2. **Computational analogy:** Intuitive types are formalized as types in a programming language; discovery mechanisms and scientific concepts as programs.

These analogies provide the basis for the Python implementation of the type-theoretic reconstruction of Einstein's conceptual path to the relativity of time.

Program synthesis

We recast the type-theoretic reconstruction of Einstein's route to relative time as a program-synthesis problem: the goal is to recover an eight-step sequence of admissible symbolic operations and typing rules that transforms Lorentz's initial judgment that time is artificial into Einstein's judgment that time is relative. We compare three search strategies over this space—enumerative, Bayesian, and Bayes–neural—and evaluate by success rate versus program index, averaged over $N = 150$ runs (Fig. 1). With search constraints (inductive biases) enabled (Fig. 1 (a)), Pure-Bayes performs best, Bayes–neural is next, and enumeration is worst; without these constraints (Fig. 1 (b)), Bayes–neural is markedly more sample-efficient, while Pure-Bayes and enumeration rarely succeed even with large program budgets.

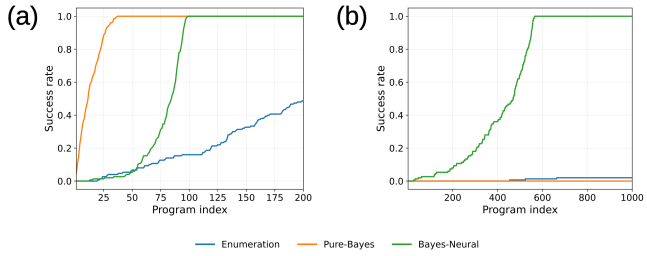

Figure 1: Success rate vs. program index for Enumeration (blue), Pure-Bayes (orange), and Bayes-Neural (green). (a) Biased: all heuristics on except "backward-only after target waveform". (b) Unbiased: all heuristics off.

Discussion and Conclusion

So how are scientific concepts birthed? This work does not offer a final answer, but argues that a central part of the answer lies in the intuitive types that constrain and guide scientific knowledge-making. Reconstructing Einstein's path to relative time makes these intuitive types explicit and yields a typed discovery mechanism—the assumption–conclusion switch. The program-synthesis results indicate that combining Bayesian guidance with neural methods can mitigate the curse of compositionality in conceptual search. Ultimately, a type-theoretic formalism of concept formation can serve both as tool AI (Aguirre 2025) and as a supportive framework for scientists. It may also inform pedagogy—not only by helping people learn existing scientific concepts, but by equipping them with the tools needed to discover and create new ones themselves.

References

Aguirre, A. 2025. Engineering the Future: What We Should Do Instead. In *Keep the Future Human: Why and How We Should Close the Gates to AGI and Superintelligence, and What We Should Build Instead*. 47–56.

Battaglia, P. W.; Hamrick, J. B.; Bapst, V.; Sanchez-Gonzalez, A.; Zambaldi, V.; Malinowski, M.; Tacchetti, A.; Raposo, D.; Santoro, A.; Faulkner, R.; Gulcehre, C.; Song, F.; Ballard, A.; Gilmer, J.; Dahl, G. E.; Vaswani, A.; Allen, K.; Nash, C.; Langston, V. J.; Dyer, C.; Heess, N.; Wierstra,

D.; Kohli, P.; Botvinick, M.; Vinyals, O.; Li, Y.; and Pascanu, R. 2018. Relational inductive biases, deep learning, and graph networks. arXiv:1806.01261 [cs.LG].

Battleday, R. M., and Gershman, S. J. 2024. Artificial intelligence for science: The easy and hard problems. arXiv:2408.14508 [cs.AI].

Cardelli, L. 1996. Type systems. In Tucker, A. B., ed., *The CRC handbook of computer science and engineering*. CRC Press. 2208–2236.

Chomsky, N. 2006. *Language and Mind*. Cambridge University Press, 3rd edition.

Gentner, D. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive Science* 7(2):155–170.

Lalande, F.; Matsubara, Y.; Chiba, N.; Taniai, T.; Igarashi, R.; and Ushiku, Y. 2023. A Transformer Model for Symbolic Regression towards Scientific Discovery. arXiv 2312.04070 [cs.LG].

Liu, A.; Wu, Z.; Michael, J.; Suhr, A.; West, P.; Koller, A.; Swayamdipta, S.; Smith, N. A.; and Choi, Y. 2023. We're afraid language models aren't modeling ambiguity. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 790–807. Association for Computational Linguistics.

Lorentz, H. A. 1895. *Attempt of a Theory of Electrical and Optical Phenomena in Moving Bodies*. E. J. Brill.

Maxwell, J. C. 1865. On Physical Lines of Force. In Niven, W. D., ed., *The Scientific Papers of James Clerk Maxwell*, volume 1. Dover Publications. 155–229.

Nersessian, N. J. 1987. A Cognitive-Historical Approach to Meaning in Scientific Theories. In *The Process of Science: Contemporary Philosophical Approaches to Understanding Scientific Practice*, volume 3 of *Science and Philosophy*. Martinus Nijhoff Publishers. 161–177.

Nersessian, N. J. 2008. *Creating scientific concepts*. MIT Press.

Norton, J. D. 2004. Einstein's investigations of Galilean covariant electrodynamics prior to 1905. *Archive for History of Exact Sciences* 59(1):45–105.

Pierce, B. C. 2002. *Types and Programming Languages*. MIT Press.

Plotkin, G. D. 1977. LCF Considered as a Programming Language. *Theoretical Computer Science* 5:223–255.

Rule, J. S.; Tenenbaum, J. B.; and Piantadosi, S. T. 2020. The Child as Hacker. *Trends in Cognitive Sciences* 24(11):900–915.

Sosa, F. A., and Ullman, T. D. 2022. Type Theory in Human-Like Learning and Inference. arXiv:2210.01634 [cs.AI].

Udrescu, S.-M.; Tan, A.; Feng, J.; Neto, O.; Wu, T.; and Tegmark, M. 2020. AI Feynman 2.0: Pareto-optimal symbolic regression exploiting graph modularity. In *Advances in Neural Information Processing Systems*, volume 33, 9140–9150.

Wolfram, S. 2024. Can AI Solve Science? Stephen Wolfram Writings.