From Mechanistic to Compositional Interpretability

Ward Gauderis [*1] **Thomas Dooms** [*2] **Steven T. Homer** [2] **Kola Ayonrinde** [3] **Geraint A. Wiggins** [1]

Abstract

Mechanistic interpretability aims to explain neural model behaviour by reverse-engineering learned computational structure into human-understandable components. Without a formal framework, however, mechanistic explanations cannot be objectively verified, compared, or composed. We introduce *compositional interpretability*, a category-theoretic framework grounded in the principles of compositionality and minimum description length. *Compositional interpretations* are pairs of syntactic and semantic mappings that must commute to enforce consistency between a model's decomposition and its observed behaviour. We deconstruct explanation quality into measures of *faithfulness* and *complexity* to cast interpretability as a constrained optimisation problem, and introduce *compressive refinement* to systematically restructure models into simpler parts without altering their function. Finally, we derive a parsimony criterion under which syntactic compression theoretically guarantees more concise, human-aligned explanations. Our framework situates prominent mechanistic methods as subclasses of refinement, and clarifies why their compressibility heuristics tend to align with human interpretability. Our work provides a measurable, optimisable blueprint for automating the discovery and evaluation of mechanistic explanations.

1. Introduction

Neural models derive their capabilities from emergent complexity rather than explicit design. While they learn sophisticated and meaningful functions, their behaviour is difficult to understand because their internal representations are opaque to humans (Ayonrinde & Jaburi, 2025). Yet interpreting them could yield new scientific insights (Brixi et al., 2025; Nainani et al., 2025; Pearce et al., 2026) and give us greater control over their behaviour (Templeton et al., 2024). Doing so requires a measurable definition of interpretation quality that can be tested and optimised (Barbiero et al., 2025; Madsen et al., 2024). Without it, the resulting explanations cannot systematically be compared or composed.

Interpretability approaches can be seen as spanning a spectrum from intrinsically interpretable models to post-hoc explanations. Post-hoc methods such as saliency and attribution (Binder et al., 2016) explain individual predictions through surface-level heuristics, but provide little insight into the model's internal mechanisms (Ayonrinde & Jaburi, 2025; Adebayo et al., 2020; Kim et al., 2018). Conversely, intrinsically interpretable models aim to make the structure of the computation transparent by design but require changes to architecture or training (Madsen et al., 2024). Mechanistic interpretability sits between these endpoints by seeking to relate observed behaviour to internal features (Alain & Bengio, 2018) or to mechanisms discovered through compressibility heuristics such as sparsity (Ayonrinde et al., 2024b; Bricken et al., 2023; Gao et al., 2024). While the goal is sound, current mechanistic methods approximate faithfulness rather than enforce it structurally.

The standard reverse-engineering view constructs interpretations in three stages: decomposition, description, and validation (Sharkey et al., 2025). These correspond to breaking a model into simpler functional parts, assigning them explanatory hypotheses, and testing their correctness, respectively. Decomposition fixes the structural foundation, or syntax, on which the latter two stages operate. If that basis is non-compositional, then description and validation can be locally correct while remaining globally uninformative. With a flawed syntax, even perfectly described parts will not add up to a complete explanation of global behaviour.

To avoid the above shortcomings, we argue that interpretations should be **compositional**, made of individually describable parts whose composition equals the whole; **faithful**, approximating the model's internal computation rather than merely its input-output behaviour; and **concise**, having low description length for a human interpreter. Compositionality enables bidirectional tracing between local parts and global behaviour, while faithfulness prevents explanations from becoming just-so stories. Together, these criteria cast interpretability as constrained optimisation: among all

[*]Equal contribution [1]Vrije Universiteit Brussel, Brussels, Belgium [2]Independent [3]UK AI Security Institute, London, United Kingdom. Correspondence to: Ward Gauderis <ward.gauderis@vub.be>, Thomas Dooms <doomsthomas@gmail.com>.

Accepted to the 2nd Workshop on Compositional Learning at ICML 2026, Seoul, South Korea. Copyright 2026 by the author(s).

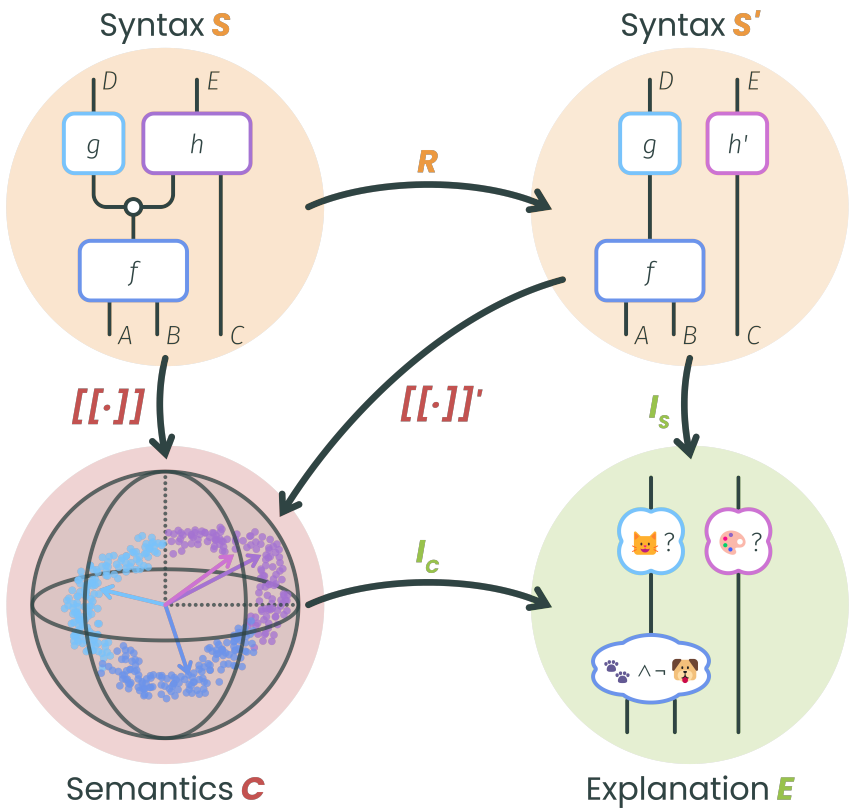

Figure 1. A commutative diagram illustrating compositional interpretability through compressive refinement for a model that classifies animals and their colour. In the original decomposition, a string diagram in S, the mechanisms appear structurally entangled even though their learned representations $[[\cdot]]$ in C are not. Through a compressive refinement R, a new model decomposition S' is discovered that clearly separates the two mechanisms, while the new implementation $[[\cdot]]'$ remains faithful to the model's input-output behaviour in C. Based on this new syntax, a compositional interpretation I_S can assign much simpler explanations in E to the disentangled mechanisms. Because all paths commute, the resulting explanation coincides with the post-hoc observed behaviour I_C of the original model.

faithful compositional explanations, prefer the most concise.

This paper formalises interpretability through the notion of a *compositional interpretation* (Section 2), which requires grounding a model in both its structure (syntax) and behaviour (semantics). We characterise the quality of such an interpretation by its faithfulness and complexity (Section 3), and upper-bound complexity by a tractable measure that can be directly optimised. The optimisation takes the form of *compressive refinement* (Section 4), restructuring the model into simpler parts while preserving its semantics. We show how compressive refinement aligns interpretability under the principles of explanatory optimism and compositional sparsity. Finally, we situate compositional interpretability relative to other prominent formalisms (Section 5). A summary of the formal notation and colour-coding is provided in Appendix E.

2. Defining Compositional Interpretations

The objective of interpreting neural models is rarely specified, leading to ambiguity about what constitutes a valid explanation. We build on an *explanatory view* of interpretability, which holds that explaining model behaviour requires examining internal mechanisms, not just input-output pairs. This perspective distinguishes between two forms of faithfulness (Ayonrinde & Jaburi, 2025). *Behavioural faithfulness* means that the explanation and the model produce the same outputs. Post-hoc interpretability methods typically aim for this, describing what a model computes by generating explanations that correlate with its data-local surface behaviour, thereby answering questions like "Why did the model make this particular decision?" In contrast, *explanatory faithfulness* means the internal steps of the step-by-step explanation match the model's own mechanisms, not just its input-output behaviour. Mechanistic interpretations pursue this by examining how a model operates and elucidating the mechanisms by which it reaches conclusions. The mechanistic approach addresses broader questions such as "How did the model solve this general class of problems?", aiming to explain the model's generalisation capabilities rather than just individual decisions (Sharkey et al., 2025).

Definition 2.1 (Explanatory faithfulness)**.** An explanation E is *explanatorily faithful* to a model M over a data distribution $\mathcal{D}$, to the extent that intermediate activations s_i at each layer i that are given by an algorithmic explanation E

closely match the intermediate activations x_i of the model M for input data in $\mathcal{D}$ (Ayonrinde & Jaburi, 2025).

Two aspects deserve emphasis. First, explanation quality depends on the interpreter's goals and abilities. Second, faithfulness is relative to a data distribution and may be approximate. This motivates a non-binary notion of faithfulness, which we develop in Section 4. For now, we discuss the structural implications of explanatory faithfulness.

Compositionality Neural networks expose a limitation of naive reductionism: despite complete knowledge of their low-level components, their emergent behaviour remains opaque, posing a fundamental challenge for interpretability (Ayonrinde & Jaburi, 2025). Achieving explanatory faithfulness—matching an explanation's structure to a model's computational mechanisms—requires decomposing the model into meaningful components. While a model's layers offer a natural starting point for such a decomposition (Definition 2.1), there is no guarantee that these architectural boundaries align with intelligible atoms of computation. Indeed, because multi-layer computations can be flattened or vice versa (Lin et al., 2017; Danhofer et al., 2025), architectural depth is an implementation choice. Coupled with the discovery of cross-layer circuits (Olsson et al., 2022; Lindsey et al., 2024), this suggests the fundamental units of computation are functional, not architectural. These observations motivate the need for a more general, functionally grounded principle for explanatory faithfulness: *compositionality*, commonly attributed to Frege (Coecke, 2021).

Definition 2.2 (Compositionality)**.** The meaning of the whole can be derived from the meanings of its parts and how these parts are structured together.

Compositionality describes the relationship between a system's global properties and those of its components (Puca et al., 2023; Montague, 1970). For interpretability, this enables a divide-and-conquer approach: a model's behaviour should emerge from the local understanding of its parts (composition), while behaviour must also be traceable back to the contributions of those parts (decomposition) (Puca et al., 2023). However, compositionality is not theory-neutral; it requires a formal theory of both syntax (the parts and their structure) and semantics (their meaning) (Hupkes et al., 2020). The central challenge, therefore, is to identify a formal syntax and semantics for a model that is both faithful to its mechanisms and useful for human understanding.

Category theory provides the mathematical language to formalise this distinction between syntax and semantics and to reason about compositional structures (Fong & Spivak, 2018; Dudzik et al., 2022). Compositional methods in AI are largely inspired by fields like quantum mechanics, which shift the focus of study from isolated components to their structured interactions (Coecke & Kissinger, 2018). These methods combine explicit structure with distributed representations, helping overcome the curse of dimensionality in model generalisation and interpretation (Tull et al., 2024; Coecke, 2021; Hitzler & Sarker, 2021). The application of category theory in AI has already produced unifying frameworks for organising insights across the field (Shiebler et al., 2021; Dudzik et al., 2022), whose utility has been demonstrated in model design (Gavranović et al., 2024; Khatri et al., 2024; Rodatz et al., 2025) and conceptual, causal and generative modelling (Tull et al., 2023b; Gauderis & Wiggins, 2023; Lorenz & Tull, 2023; Tull et al., 2023a).

Compositional interpretability *String diagrams* (Coecke & Kissinger, 2018) provide the syntactic skeleton needed to separate a model's wiring from its computational behaviour. They describe the flow of information as pure structure (see Figure 1 and Appendix A). By endowing this bare syntax with a semantic representation, we form a *compositional model* (Tull et al., 2024).

Definition 2.3 (Compositional model)**.** A *compositional model* $M = (\boldsymbol{D}, \boldsymbol{S}, \boldsymbol{C}, [[\,\cdot\,]])$ consists of:

- a string diagram $\boldsymbol{D} : \boldsymbol{I} \to \boldsymbol{O}$ in the monoidal category $\boldsymbol{S}$, representing the *syntax* $\boldsymbol{S}$ of M;
- a monoidal category $\boldsymbol{C}$, providing the *semantics* of M;
- a monoidal *representation* functor $[[\,\cdot\,]] : \boldsymbol{S} \to \boldsymbol{C}$, which assigns a semantic instantiation in $\boldsymbol{C}$ to each syntactic component in $\boldsymbol{S}$, preserving the diagrammatic structure.

The diagram $\boldsymbol{D}$ is built from a finite collection of components. Its *signature* $\boldsymbol{S}_{\boldsymbol{D}}$ consists of the set of all individual wire types (objects $\mathrm{ob}(\boldsymbol{S}_{\boldsymbol{D}})$) and boxes (generators $\mathrm{mor}(\boldsymbol{S}_{\boldsymbol{D}})$) occurring in $\boldsymbol{D}$[1]. The image of $\boldsymbol{S}_{\boldsymbol{D}}$ under $[[\,\cdot\,]]$ is denoted as $\boldsymbol{C}_{\boldsymbol{D}}$, the minimal set of semantic elements required to implement the representation of the diagram $\boldsymbol{D}$.

Within this framework, the *syntax* $\boldsymbol{S}$, described by the string diagram $\boldsymbol{D}$, defines the model's abstract internal structure, such as neurons, layers, or attention heads in neural models. The diagram provides a high-level structure that can be inspected or reasoned about. The *semantics* $\boldsymbol{C}$ provides the model's concrete instantiation or implementation, enabling practical computation and evaluation. The representation $[[\,\cdot\,]] : \boldsymbol{S} \to \boldsymbol{C}$ links syntax to semantics, assigning (learned) functions, distributions, or matrices to these syntactic components (Figure 1). For interpretability, this distinction between syntax and semantics offers a useful lens for decomposing model behaviour into understandable parts. It elucidates how global model behaviour arises from learned semantic representations composed by syntax. Both features and mechanisms (Olah et al., 2020) can be viewed as syntactic components, differing primarily in the degree

[1]The syntactic category $\boldsymbol{S} = \mathrm{Free}(\boldsymbol{S}_{\boldsymbol{D}})$ is the free monoidal category generated by $\boldsymbol{S}_{\boldsymbol{D}}$, comprising all valid diagrams formed from $\mathrm{ob}(\boldsymbol{S}_{\boldsymbol{D}})$ and $\mathrm{mor}(\boldsymbol{S}_{\boldsymbol{D}})$. The signature $\boldsymbol{S}_{\boldsymbol{D}}$ may also admit equations $\mathrm{eq}(\boldsymbol{S}_{\boldsymbol{D}})$ specifying equivalence between diagrams.

to which they can further be decomposed into subdiagrams. We can understand how different syntactic constructions are (approximately, within known bounds) equivalent by relating them to their concrete semantic instantiations.

While any model can be framed compositionally, this is only useful when oriented toward a specific goal (Duneau, 2025). To this end, we identify *compositional interpretations* (Tull et al., 2024): human-understandable explanations tied to both a model's syntax and semantics.

Definition 2.4 (Compositional Interpretation)**.** Given a compositional model $M = (\boldsymbol{D}, \boldsymbol{S}, \boldsymbol{C}, [[\cdot]])$ and a signature of human-interpretable terms $\boldsymbol{H}$, a *compositional interpretation* consists of two compatible mappings:

- A syntax-based, *compositional interpretation* $\boldsymbol{I_S} : \boldsymbol{S} \to \boldsymbol{H}$, which provides *syntactic grounding* by assigning explanations to the structural components of M.
- A semantics-based, *post-hoc interpretation* $\boldsymbol{I_C} : \boldsymbol{C} \to \boldsymbol{H}$, which provides *semantic grounding* by producing explanations based on the behaviour implemented by M.

For consistent groundings, this diagram must commute[2]:

$$\begin{tikzcd} S \arrow[d, "{[[\cdot]]}"'] \arrow[dr, "I_S"] & \\ C \arrow[r, "I_C"'] & H \end{tikzcd}$$

The commutativity condition, $\boldsymbol{I_S} = \boldsymbol{I_C} \circ [[\cdot]]$, is the cornerstone of this definition. It enforces consistency between two forms of interpretation (Figure 1). The compositional interpretation $\boldsymbol{I_S}$ ties the explanation to the model's structure, while the post-hoc interpretation $\boldsymbol{I_C}$ ties it to the model's input-output behaviour. Without this condition, interpretations risk becoming disconnected from the mechanisms they claim to describe. An interpretation defined solely in terms of semantics, without factoring through a valid syntax, remains causally disconnected from the model's mechanisms: it describes *what* the model does, not *how* it is done. Consequently, components with distinct implementations might be given identical explanations despite structural differences. Conversely, an interpretation that is defined solely in terms of syntax labels internal components without verification against their semantics: these labels may not reflect their actual function. Components with identical semantics might be given different, misleading explanations based on superficial structural distinctions. By requiring syntax and semantics to agree, a compositional interpretation connects mechanism to behaviour.

[2]Categorically, both mappings would ideally be complete monoidal functors. In practice, these constraints are relaxed to partial maps to account for (as yet) unexplained components.

In practice, interpretations are partial, meaning they may not account for every component of a model. A *compositionally-interpretable model* admits a complete compositional interpretation $\boldsymbol{I_S}$, where every syntactic component receives a human-understandable explanation. Tull et al. (2024) show that this definition generalises the existing notion of intrinsically interpretable models, such as decision trees or linear models, which are considered transparent precisely because their explicit structure permits direct and complete interpretation . The central objective of compositional interpretability is therefore *the search for compositional interpretations that are as complete as possible*.

3. Evaluating Compositional Interpretations

Compositionality is only a *structural constraint*: it defines the space of valid explanations but does not distinguish the good from the bad. When offered multiple valid choices, which explanation should be preferred? We take inspiration from Kowalski's (1979) dictum

$$\text{Algorithm} = \textbf{Logic} + \textbf{Control}.$$

In logic programming, the logic component determines the meaning of the algorithm, whereas the control component only affects its implementation. By separating the two, Kowalski argued that computer programs can be improved by freely optimising control for computational efficiency, while the algorithm's logic must be held constant. Compositional interpretability mirrors this separation as[3]

$$\text{Model} = \textbf{Semantics} + \textbf{Syntax}.$$

We treat the semantics as the fixed logic—the input-output behaviour we aim to understand—and the syntax as the variable control structure that decomposes it. Our objective differs from Kowalski's only in the target metric: rather than optimise syntax for efficiency, we optimise for intelligibility.

Our constrained optimisation view of interpretability is formalised through the *Minimum Description Length* (MDL) principle (Grünwald, 2007). In model selection, MDL identifies the best model M for a dataset $\mathcal{D}$ as the one that minimises the total description length $L(\cdot)$ in bits:

$$L(\mathcal{D}) = L(M) + L(\mathcal{D} \mid M).$$

The total description cost is the length of the model $L(M)$ plus the length of the data encoded, given that model $L(\mathcal{D} \mid M)$. This two-part code operationalises Occam's razor: all else being equal, prefer the more concise explanation (Grünwald, 2007). We apply this principle at a meta-level: interpretability is framed as a constrained communication problem where the goal is to communicate the model's semantics $[[\boldsymbol{D}]]$ as efficiently as possible in

[3]This also parallels Marr's (1982) computational and algorithmic levels of analysis, respectively.

a human-comprehensible language $\boldsymbol{H}$ (Ayonrinde et al., 2024a; Braun et al., 2025). We treat the model's fixed semantics $[[\boldsymbol{D}]]$ as the data and the compositional interpretation $\boldsymbol{I_S}(\boldsymbol{D})$ as the hypothesis. The description length of the semantics then decomposes as:

$$L([[\boldsymbol{D}]]) = \underbrace{L(\boldsymbol{I_S}(\boldsymbol{D}))}_{\text{Complexity}} + \underbrace{L([[\boldsymbol{D}]] \mid \boldsymbol{I_S}(\boldsymbol{D}))}_{\text{Faithfulness}}$$

The first term measures the *complexity* of the explanation. The second quantifies *faithfulness* in the information lost by the explanation. It is the cost of correcting the explanation to match the model's behaviour. Formulating interpretability through MDL presents distinct challenges compared to standard machine learning: the hypothesis class potentially spans all human-interpretable statements $\boldsymbol{H}$, making the complexity of an explanation subjective and dependent on the interpreter's goals, priors, and computational capabilities. However, unlike inferring a model from noisy empirical data, we have access to the system's closed-form description. Knowing the ground truth without aleatoric uncertainty allows us to separate the MDL terms. Rather than minimising the sum as a soft trade-off, we treat it as a rate-distortion problem and minimise the explanation's complexity subject to a strict lower bound on faithfulness.

We thus refine the central criterion of compositional interpretability as *the search for complete compositional interpretations that yield the least complex explanations while remaining faithful up to a user-specified threshold*. In the remainder of this section, we formalise these measures for compositional interpretations so their optimisation can be automated. Evaluating them naively in the human-interpretable category $\boldsymbol{H}$ would require subjective human involvement. To bypass this, we instead ground them directly in the model's syntax $\boldsymbol{S}$ and semantics $\boldsymbol{C}$.

Faithfulness The *faithfulness measure* $L([[\boldsymbol{D}]] \mid \boldsymbol{I_S}(\boldsymbol{D}))$ quantifies the information needed to correct an explanation. Rate-distortion theory operationalises this by bounding a semantic distortion metric d (e.g., L_2 norm or KL-divergence) between the model's true semantics and those reconstructed from the explanation. However, restricting this bound to the global input-output behaviour $[[\boldsymbol{D}]]$ ensures only behavioural faithfulness, lacking the mechanistic insight required by Definition 2.1. While compositional interpretations are structurally well-defined (Definition 2.4), they are not inherently informative: describing a component merely as "a mathematical function" is valid but trivial. To ensure explanations capture specific structural mechanisms, we impose the stricter measure of *compositional faithfulness*: the interpretation must be detailed enough to reconstruct the semantics of *each* component that it explains.

Definition 3.1 (Compositional faithfulness)**.** Given a compositional model $M = (\boldsymbol{D}, \boldsymbol{S}, \boldsymbol{C}, [[\,\cdot\,]])$, a compositional interpretation $\boldsymbol{I_S} : \boldsymbol{S} \to \boldsymbol{H}$ is *compositionally faithful* if there exists a *reconstruction* $\boldsymbol{I_C^*} : \boldsymbol{H} \to \boldsymbol{C}$ such that the following diagram commutes:

$$\begin{tikzcd} S \arrow[d, "{[[\,\cdot\,]]}"'] \arrow[dr, "I_S"] & \\ C \arrow[r, shift right, "I_C"'] & H \arrow[l, shift right, "I_C^*"'] \end{tikzcd}$$

This constraint is relaxed up to a tolerance $\epsilon \in \mathbb{R}$ to reconcile faithfulness and complexity[4]. Given a metric d on boxes (parallel morphisms) in $\boldsymbol{C}$, we require that $\forall \boldsymbol{f} \in \text{mor}(\boldsymbol{S})$:

$$d([[\,\boldsymbol{f}\,]], \boldsymbol{I_C^*}(\boldsymbol{I_S}(\boldsymbol{f}))) \leq \epsilon$$

This condition subsumes explanatory and behavioural faithfulness: reconstructing individual component mechanisms necessarily recovers global behaviour upon composition. Consequently, the interpreter can understand the global system by combining local component explanations according to the diagrammatic structure. The metric d defines the semantic reconstruction error relevant to the interpreter, making faithfulness relative rather than absolute. By relaxing d to a premetric over specific subdistributions, we ensure faithfulness where it matters while permitting abstraction elsewhere. For example, evaluating d only over naturally occurring or task-relevant data captures observational alignment, whereas evaluating it over interventional data ensures the explanations are causally faithful (Joshi et al., 2026).

Complexity The *complexity measure* $L(\boldsymbol{I_S}(\boldsymbol{D}))$ of an explanation is inherently more elusive than faithfulness since it depends on the interpreter's subjective frame of reference. The lack of a universal metric has posed a significant obstacle to automated evaluation of interpretations. However, the dual nature of compositional interpretations offers a tractable surrogate. Just as we operationalised faithfulness in the semantics $\boldsymbol{C}$ instead of $\boldsymbol{H}$, we bound complexity by the structural derivation of the explanation rather than by the explanation itself. Relying on the factorisation $\boldsymbol{I_S} = \boldsymbol{I_C} \circ [[\,\cdot\,]]$ and functoriality of $[[\,\cdot\,]]$, we define the *compositional description length*: the cost of describing the model's compositional representation plus the complexity of interpreting its components' semantics.

Definition 3.2 (Compositional description length)**.** Let $M = (\boldsymbol{D}, \boldsymbol{S}, \boldsymbol{C}, [[\,\cdot\,]])$ be a compositional model and $\boldsymbol{I_S} : \boldsymbol{S} \to \boldsymbol{H}$ a compositional interpretation for M. The *compositional description length* $L(M, \boldsymbol{I_S})$ of its explanation $\boldsymbol{I_S}(\boldsymbol{D})$ is defined as the sum of its representation and

[4] Categorically, relaxing strict inversion to an adjunction $\boldsymbol{I_C^*} \dashv \boldsymbol{I_C}$ formalises lossy abstraction: $\boldsymbol{I_C}$ discards details while $\boldsymbol{I_C^*}$ reconstructs the canonical implementation consistent with the explanation. However, our analysis does not rely on such assumptions.

interpretation complexities.

$$L(M, \boldsymbol{I_S}) := L^{\text{rep}}(M) + L^{\text{int}}(M, \boldsymbol{I_C})$$

Where *representation complexity* $L^{\text{rep}}(M)$ is the syntax-based cost of describing the semantics of the individual components in $\boldsymbol{S_D}$ and their wiring in $\boldsymbol{D}$ separately.

$$L^{\text{rep}}(M) := L(\boldsymbol{D}) + \sum_{\boldsymbol{f} \in \text{mor}(\boldsymbol{S_D})} L([[\,\boldsymbol{f}\,]])$$

And *interpretation complexity* $L^{\text{int}}(M, \boldsymbol{I_C})$ is the cost of describing the post-hoc interpretation method for the individual semantic elements in $\boldsymbol{C_D}$.

$$L^{\text{int}}(M, \boldsymbol{I_C}) := L(\boldsymbol{I_C}|_{\boldsymbol{C_D}})$$

The decomposition splits the description length of an explanation into two distinct parts: an objective structural cost and a subjective alignment cost. The *representation complexity* $L^{\text{rep}}(M)$ depends solely on the model. It captures the trade-off between the granularity of the decomposition $\boldsymbol{D}$ and the semantic complexity of the atomic units $[[\boldsymbol{f}]]$. This makes it a tractable target for automated optimisation (Appendix D). Conversely, the *interpretation complexity* $L^{\text{int}}(M, \boldsymbol{I_C})$ isolates the subjective difficulty of aligning the semantics of the atomic units in $\boldsymbol{D}$ with human concepts. This cost is inherently variable, scaling with the conceptual gap (Ayonrinde, 2025) between the components' semantics relative to the interpreter's mental model (Sutter et al., 2025). Optimising $L(M, \boldsymbol{I_S})$ therefore balances syntactic parsimony with semantic simplicity: we seek a decomposition that is concisely described yet composed of parts that are naturally aligned with the interpreter. As $L(M, \boldsymbol{I_S})$ represents the information cost of the compositional derivation, it upper-bounds the theoretical minimal description length $L(\boldsymbol{I_S}(\boldsymbol{D}))$ of the explanation itself.

Theorem 3.3 (Description length bound)**.** *Compositional description length $L(M, \boldsymbol{I_S})$ is an upper bound on the minimal description length $L(\boldsymbol{I_S}(\boldsymbol{D}))$ of a model's explanation.*

$$L(\boldsymbol{I_S}(\boldsymbol{D})) \leq L(M, \boldsymbol{I_S})$$

The proof is deferred to Appendix B.

4. Improving Compositional Interpretations

Since $L(M, \boldsymbol{I_S})$ corresponds to the length of a realisable encoding, it serves as a tractable surrogate to optimise $L(\boldsymbol{I_S}(\boldsymbol{D}))$. Optimising this bound formalises the natural workflow in mechanistic interpretability: identifying a suitable decomposition, interpreting its components, and verifying their descriptions causally (Sharkey et al., 2025). Since structurally modifying a model is more tractable than solving the subjective alignment problem, recent research has predominantly been successful in reducing representation complexity $L^{\text{rep}}(M)$ (Gao et al., 2025; Dunefsky et al., 2024; Bushnaq et al., 2025), while interpretation complexity $L^{\text{int}}(M, \boldsymbol{I_C})$ remains less explored (Sutter et al., 2025; Voita & Titov, 2020; Oldfield et al., 2025). In this work, we align with this strategy, focusing on the first step: finding decompositions that minimise representational cost. We formalise this process as *syntactic refinement* (Tull et al., 2024): compressing the model's representation by optimising the syntax while keeping the semantics fixed.

Definition 4.1 (Syntactic refinement)**.** A *syntactic refinement* of a compositional model $M = (\boldsymbol{D}, \boldsymbol{S}, \boldsymbol{C}, [[\,\cdot\,]])$ into a semantically equivalent model $M' = (\boldsymbol{D}', \boldsymbol{S}', \boldsymbol{C}, [[\,\cdot\,]]')$ is a functor $\boldsymbol{R} : \boldsymbol{S} \to \boldsymbol{S}'$ such that $\boldsymbol{D}' = \boldsymbol{R}(\boldsymbol{D})$[5].

For unchanged semantics $\boldsymbol{C}$, this diagram must commute[6]:

$$\begin{tikzcd} S \arrow[r, "R"] \arrow[d, "{[[\,\cdot\,]]}"'] & S' \arrow[dl, "{[[\,\cdot\,]]'}"] \\ C & \end{tikzcd}$$

If $L^{\text{rep}}(M') \leq L^{\text{rep}}(M)$, the refinement reduces the representation complexity and is *compressive*.

Compressive refinement $\boldsymbol{R}$ operationalises the redistribution of information from semantic atoms to syntactic structure, subject to $[[\,\cdot\,]] = [[\,\cdot\,]]' \circ \boldsymbol{R}$. By unpacking a black box with high $L([[\boldsymbol{D}]])$ into a diagram of simpler sub-components (increased $L(\boldsymbol{D})$ but lower $\sum_{\boldsymbol{f} \in \text{mor}(\boldsymbol{S_D})} L([[\boldsymbol{f}]])$), the model's internal regularities become explicit (Figure 1). An efficient refinement seeks the optimal granularity where the reduction in component complexity outweighs the cost of describing the additional wiring (Appendix D).

This gives a formal rationale for the search for "atoms of computation" (Olah et al., 2020): primitive components that are algebraically simple and structurally explicit. However, mathematical simplicity does not guarantee human understanding. As captured by the interpretation complexity term in Definition 3.2, syntactic compression alone is insufficient. A refinement might simplify syntax at the cost of introducing components that are harder to interpret. For instance, a syntactic refinement could decrease representation complexity $L^{\text{rep}}(M)$ by replacing linearly alignable components with nonlinear ones (Elhage et al., 2022), thereby increasing interpretation complexity $L^{\text{int}}(M, \boldsymbol{I_C})$. Consequently, a compressive refinement improves the explanation only if the reduction in representation complexity outweighs any

[5]The dual functor $\boldsymbol{R}' : \boldsymbol{S}' \to \boldsymbol{S}$ yields *syntactic abstraction*. By functor composition, one can selectively refine and abstract representations while preserving a trace to the original model.

[6]In practice, this constraint is relaxed up to a tolerance $\epsilon \in \mathbb{R}$ to reconcile (global) faithfulness and description length. Given a metric d (or generalisation thereof) on $\text{Hom}_C([[\boldsymbol{I}]], [[\boldsymbol{O}]])$, we require $d([[\,\boldsymbol{D}\,]], [[\,\boldsymbol{D}'\,]]') \leq \epsilon$.

increase in interpretation complexity. We formalise this necessary condition via the *parsimony criterion*, a decision rule that follows from Theorem 3.3.

Proposition 4.2 (Parsimony criterion for refinement)**.** *Let M' be a compressive refinement of M. Let $\boldsymbol{I_S}$ and $\boldsymbol{I'_S}$ be the compositional interpretations of M and M' respectively, with the same semantic grounding $\boldsymbol{I_C} : \boldsymbol{C} \to \boldsymbol{H}$. The refined model M' yields a more parsimonious explanation in terms of compositional description length*

$$L(M', \boldsymbol{I'_S}) \leq L(M, \boldsymbol{I_S})$$

if and only if the gained representation simplicity outweighs the interpretation complexity:

$$\underbrace{L^{rep}(M) - L^{rep}(M')}_{\textit{Decrease in representation complexity}} \geq \underbrace{L^{int}(M', \boldsymbol{I_C}) - L^{int}(M, \boldsymbol{I_C})}_{\textit{Increase in interpretation complexity}}$$

The parsimony criterion underscores that a compressive refinement is not an end in itself but a means to balance two tensions in finding concise compositional explanations:

1. *Representation vs. interpretation complexity:* There is an external risk that optimising the objective representation complexity $L^{\text{rep}}(M)$ comes at the cost of the subjective interpretation complexity $L^{\text{int}}(M, \boldsymbol{I_C})$. A concise description is not necessarily easy for a human to interpret.
2. *Wiring vs. component complexity:* Internal to representation complexity, there is a tension between the wiring complexity $L(\boldsymbol{D})$ and the semantic complexity of the individual components $\sum_{f \in \text{mor}(\boldsymbol{S_D})} L([[\boldsymbol{f}]])$. Decomposing components reduces their individual cost but increases the diagram size $\boldsymbol{D}$. An exponentially large diagram undermines behavioural faithfulness, as the interpreter cannot tractably compose all interactions.

While these trade-offs suggest that naively optimising $L^{\text{rep}}(M)$ alone could potentially lead to uninterpretable model decompositions, practical evidence indicates that compression often yields interpretable results (Gao et al., 2025; Dunefsky et al., 2024; Braun et al., 2025; Hinton & Salakhutdinov, 2006). Under compressive refinements, the two pressures described above are often aligned rather than adversarial. We argue this can be made deliberate, addressing the first tension through a comonotonic coding distribution and the second through compositional sparsity.

Representation vs. interpretation complexity To resolve the first tension between representation and interpretation complexity, we draw on the intuition that the regularities a model exploits to solve a task often correspond to concepts humans use to understand that task. This intuition is captured by the *Weak Principle of Explanatory Optimism*, a conjecture usually left implicit in interpretability (Ayonrinde & Jaburi, 2025).

Conjecture 4.3 (Weak principle of explanatory optimism)**.** Most important behaviour of an intelligence with implicit explanatory knowledge within some explanatory complexity class is human-understandable (Ayonrinde & Jaburi, 2025).

Since structure is fundamentally linked to compressibility, compressive refinement acts as a filter for these regularities. Lossy compression that retains only the most important model behaviour (measured by d in Definition 4.1) strips away redundancy to isolate the core mechanisms. The conjecture implies that if the model performs well, these isolated mechanisms are human-understandable. To ensure alignment to human concepts, we encode explanatory optimism in the coding distribution that defines representation complexity, $L^{\text{rep}}(M) := -\log_2 P^{\text{rep}}(M)$ (Grünwald, 2007). By choosing a $P^{\text{rep}}(M)$ that encodes our prior *belief* in explanatory optimism, assigning higher probability to representations we expect to be interpretable (such as those exhibiting sparsity, low dimensionality, or hierarchy), $L^{\text{rep}}(M)$ measures how well a model satisfies this prior, aligning representation complexity with interpretation complexity.

Corollary 4.4 (Comonotonic coding)**.** *Any compressive refinement satisfies the parsimony criterion (Proposition 4.2) if its coding distribution $P^{rep}(M)$ is* comonotonic *with the induced interpretation distribution:*

$$P^{rep}(M) \leq P^{rep}(M') \implies 2^{-L^{int}(M, \boldsymbol{I_C})} \leq 2^{-L^{int}(M', \boldsymbol{I_C})}$$

Consequently, the closer our chosen coding distribution captures actual interpretation complexity, the more reliably a compressive refinement improves the full compositional description length[7]. By explicitly adopting a comonotonic coding distribution, compressive refinement is guaranteed to yield more concise explanations.

Crucially, none of our formal results assume explanatory optimism to be true; a comonotonic code only makes compressibility explicitly track interpretability. When this belief fails for a given model, so does compression with a comonotonic code: if the components' semantics admit no simpler interpretation, the absence of a successful compressive refinement indicates exactly this.

Wiring vs. component complexity To address the second tension between wiring and component complexity, ensuring that the string diagram $\boldsymbol{D}$ remains manageable through syntax refinement, we invoke a result from approximation theory regarding the nature of the functions neural networks can approximate. *Compositional sparsity* ensures that the

[7]Readers reluctant to assert explanatory optimism as a literal belief can instead adopt a *luckiness function* within the Normalised Maximum Likelihood framework (Grünwald & Roos, 2019). This replaces the subjective Bayesian *belief* with a pragmatic *hope* for interpretability; if the assumption is wrong, the code is merely suboptimal, but worst-case guarantees remain intact.

search for "atoms of computation" is feasible. The graph size $L(D)$ is expected to grow only polynomially because the semantics approximated by the model admit a sparse decomposition.

Theorem 4.5 (Compositional sparsity)**.** *Deep (neural) networks of polynomial size efficiently approximate compositionally sparse functions. A function is* compositionally sparse *if it can be decomposed into polynomially many constituent functions, each depending on a small, constant number of variables. This class includes all efficiently Turing-computable functions (Danhofer et al., 2025).*

Together, the operationalised principles of explanatory optimism and compositional sparsity justify minimising representation complexity as a tractable surrogate for the full description length. Compressive refinement distils the model into a sparse syntax with concise semantics that are aligned with human concepts. Therefore, compressive refinement yields better compositionally interpretable models.

5. Related Work

Mechanistic interpretability Existing methods use various syntactic refinements $R : S \to S'$ (Definition 4.1), differing in how tightly S' must track the given S. Early work often used the identity refinement $R = \mathrm{id}_S$, treating architectural units such as neurons as the atoms of analysis (Olah et al., 2017). Contemporary research is systematically moving to stronger decomposition classes, ranging from piecewise substitution (Bricken et al., 2023), to relying on architectural invariants (Lindsey et al., 2024), or disregarding the original syntax altogether (Gao et al., 2025). Compositional interpretability makes this balance explicit and provides a parsimony criterion for decomposition selection. A more detailed comparison can be found in Appendix C.

Measurable interpretability To bridge the traditional intrinsic and post-hoc interpretability paradigms, Madsen et al. (2024) propose *inherently faithfulness-measurable models*. Rather than imposing inflexible constraints that directly generate explanations, such models allow efficient evaluation of explanation faithfulness by design. When a measurable target metric exists, one can search directly for metric-optimal explanations rather than rely on heuristic proxies like sparsity or low-rankness (Zhou & Shah, 2023). Compositional interpretability extends this principle to mechanistic explanations, but recognises that any metric of faithfulness must be balanced against a measure of complexity. Only when both criteria are efficiently measurable can the discovery of interpretable decompositions be cast as a standard optimisation procedure.

Symmetry-formalised interpretability Barbiero et al. (2026) propose defining actionable interpretability through invariance under four symmetries. These strict constraints naturally guide the design of intrinsically interpretable models (Barbiero et al., 2025), but are difficult to apply to mechanistic or post-hoc interpretations of black-box models, where such symmetries are not guaranteed to hold. Compositional interpretability instead builds on two fundamental principles: compositionality and MDL for explanation grounding and selection, respectively. These two principles recover analogous desiderata: *inference equivariance* is implied by the functoriality of compositional interpretations, *information invariance* by minimisation of representation complexity via compressive refinement, *concept-closure invariance* by minimisation of interpretation complexity, and *structural invariance* by separation of syntax from semantics. Replacing symmetries with an MDL objective yields a spectrum of compositionally interpretable models that also includes non-interpretable models. In this spectrum, actionable interpretability is reframed from an architectural constraint into a continuous optimisation objective.

Causal abstraction Geiger et al. (2025) frame mechanistic interpretability as causal alignment between low-level model decompositions and high-level explanations. While rigorous for top-down hypothesis testing, this approach assumes the high-level explanation and the decomposition search space are known a priori (Geiger et al., 2024). Moreover, Sutter et al. (2025) note that without assuming linear representations, this alignment becomes unidentifiable: any arbitrary high-level explanation can be mapped onto any low-level decomposition. Compositional interpretability focuses instead on the antecedent step of structural decomposition. It instead searches bottom-up for simple decompositions from which candidate algorithms can be extracted. Its formulation is directly compatible with causal abstraction and inference (Lorenz & Tull, 2023; 2026): our definition of syntactic refinement aligns with the stronger notion of *strict component-level abstraction*, and our formulation of compositional faithfulness can incorporate causal validity through the semantic distortion measure d. As such, compressive refinement supplies the candidate decompositions and algorithms required for distributed alignment search.

6. Conclusion

Summary This paper establishes a formal foundation for mechanistic interpretability based on the first principles of compositionality and minimum description length. We argue that interpretability methods should explicitly aim to make models *compositionally interpretable*, and that reducing an explanation's description length provides a quantifiable objective to improve its quality. We show how structural measures of *faithfulness* and *complexity* can serve as tractable surrogates for minimising compositional description length. Finally, we demonstrate that *compressive refinement* generalises current mechanistic approaches to neural network decomposition and supply a parsimony criterion

under which minimising representation complexity is guaranteed to produce better compositional explanations.

Future work We discuss three directions to translate these theoretical foundations into practical tools for interpretation.

- We focus on decomposition; the description and validation stages of the reverse-engineering workflow (Sharkey et al., 2025) require separate treatment, and integrating compressive refinement with causal inference (Joshi et al., 2026) is a natural next step.
- Automated compressive refinement requires faithfulness and complexity to be efficiently measurable and optimisable. Tensor networks could meet these requirements, and early results suggest they can scale these objectives tractably (Dooms et al., 2024; Hamreras et al., 2025).
- While allowing subjective priors, standard complexity assumes unbounded computational capacity, poorly modelling humans. Epiplexity (Finzi et al., 2026) proposes a bounded-observer alternative to capture extractable structure better.

Acknowledgements

This research received funding from the Research Foundation Flanders (FWO) grant 11A8R26N (WG) and the Flanders AI Research Program (TD).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. Sanity checks for saliency maps, 2020. URL https://arxiv.org/abs/1810.03292.

Alain, G. and Bengio, Y. Understanding intermediate layers using linear classifier probes, 2018. URL https://arxiv.org/abs/1610.01644.

Ayonrinde, K. Position: Interpretability is a Bidirectional Communication Problem. In *ICLR 2025 Workshop on Bidirectional Human-AI Alignment*, April 2025.

Ayonrinde, K. and Jaburi, L. A Mathematical Philosophy of Explanations in Mechanistic Interpretability – The Strange Science Part I.i, May 2025.

Ayonrinde, K., Pearce, M. T., and Sharkey, L. Interpretability as compression: Reconsidering sae explanations of neural activations with mdl-saes, 2024a. URL https://arxiv.org/abs/2410.11179.

Ayonrinde, K., Pearce, M. T., and Sharkey, L. Interpretability as Compression: Reconsidering SAE Explanations of Neural Activations with MDL-SAEs, October 2024b.

Barbiero, P., Zarlenga, M. E., Termine, A., Jamnik, M., and Marra, G. Foundations of Interpretable Models, August 2025.

Barbiero, P., Zarlenga, M. E., Giannini, F., Termine, A., Bonchi, F., Jamnik, M., and Marra, G. Actionable Interpretability Must Be Defined in Terms of Symmetries, January 2026.

Binder, A., Montavon, G., Bach, S., Müller, K.-R., and Samek, W. Layer-wise relevance propagation for neural networks with local renormalization layers, 2016. URL https://arxiv.org/abs/1604.00825.

Braun, D., Bushnaq, L., Heimersheim, S., Mendel, J., and Sharkey, L. Interpretability in parameter space: Minimizing mechanistic description length with attribution-based parameter decomposition, 2025. URL https://arxiv.org/abs/2501.14926.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., McLean, B., Burke, J. E., Hume, T., Carter, S., Henighan, T., and Olah, C. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Brixi, G., Durrant, M. G., Ku, J., Poli, M., Brockman, G., Chang, D., Gonzalez, G. A., King, S. H., Li, D. B., Merchant, A. T., Naghipourfar, M., Nguyen, E., Ricci-Tam, C., Romero, D. W., Sun, G., Taghibakshi, A., Vorontsov, A., Yang, B., Deng, M., Gorton, L., Nguyen, N., Wang, N. K., Adams, E., Baccus, S. A., Dillmann, S., Ermon, S., Guo, D., Ilango, R., Janik, K., Lu, A. X., Mehta, R., Mofrad, M. R., Ng, M. Y., Pannu, J., Re, C., Schmok, J. C., St. John, J., Sullivan, J., Zhu, K., Zynda, G., Balsam, D., Collison, P., Costa, A. B., Hernandez-Boussard, T., Ho, E., Liu, M.-Y., McGrath, T., Powell, K., Burke, D. P., Goodarzi, H., Hsu, P. D., and Hie, B. Genome modeling and design across all domains of life with evo 2. *bioRxiv*, 2025. doi: 10.1101/2025.02.18.638918. URL https://www.biorxiv.org/content/early/2025/02/21/2025.02.18.638918.

Bushnaq, L., Braun, D., and Sharkey, L. Stochastic Parameter Decomposition, September 2025.

Coecke, B. Compositionality as we see it, everywhere around us, October 2021.

Coecke, B. and Kissinger, A. Picturing Quantum Processes. In Chapman, P., Stapleton, G., Moktefi, A., Perez-Kriz, S., and Bellucci, F. (eds.), *Diagrammatic Representation and Inference*, Lecture Notes in Computer Science, pp. 28–31, Cham, 2018. Springer International Publishing. ISBN 978-3-319-91376-6. doi: 10.1007/978-3-319-91376-6_6.

Danhofer, D. A., D'Ascenzo, D., Dubach, R., and Poggio, T. Position: A Theory of Deep Learning Must Include Compositional Sparsity, July 2025.

Dooms, T., Gauderis, W., Wiggins, G., and Mogrovejo, J. A. O. Compositionality Unlocks Deep Interpretable Models. In *Connecting Low-Rank Representations in AI: At the 39th Annual AAAI Conference on Artificial Intelligence*, November 2024.

Dudzik, A., Gavranović, B., João Guilherme Araújo, Petar Veličković, and Pim de Haan. Categories for AI, October 2022.

Duneau, T. Towards a Comparative Framework for Compositional AI Models, June 2025.

Dunefsky, J., Chlenski, P., and Nanda, N. Transcoders Find Interpretable LLM Feature Circuits, November 2024.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., and Olah, C. Toy models of superposition, 2022. URL https://arxiv.org/abs/2209.10652.

Finzi, M., Qiu, S., Jiang, Y., Izmailov, P., Kolter, J. Z., and Wilson, A. G. From Entropy to Epiplexity: Rethinking Information for Computationally Bounded Intelligence, March 2026. URL http://arxiv.org/abs/2601.03220. arXiv:2601.03220 [cs].

Fong, B. and Spivak, D. I. Seven Sketches in Compositionality: An Invitation to Applied Category Theory, October 2018.

Gao, L., la Tour, T. D., Tillman, H., Goh, G., Troll, R., Radford, A., Sutskever, I., Leike, J., and Wu, J. Scaling and evaluating sparse autoencoders, 2024. URL https://arxiv.org/abs/2406.04093.

Gao, L., Rajaram, A., Coxon, J., Govande, S. V., Baker, B., and Mossing, D. Weight-sparse transformers have interpretable circuits, November 2025.

Gauderis, W. and Wiggins, G. *Quantum Theory in Knowledge Representation: A Novel Approach to Reasoning with a Quantum Model of Concepts*. PhD thesis, Vrije Universiteit Brussel, Brussels, Belgium, August 2023.

Gavranović, B., Lessard, P., Dudzik, A., von Glehn, T., Araújo, J. G. M., and Veličković, P. Categorical Deep Learning: An Algebraic Theory of Architectures, February 2024.

Geiger, A., Wu, Z., Potts, C., Icard, T., and Goodman, N. D. Finding Alignments Between Interpretable Causal Variables and Distributed Neural Representations, February 2024.

Geiger, A., Ibeling, D., Zur, A., Chaudhary, M., Chauhan, S., Huang, J., Arora, A., Wu, Z., Goodman, N., Potts, C., and Icard, T. Causal Abstraction: A Theoretical Foundation for Mechanistic Interpretability, May 2025.

Grünwald, P. D. *The Minimum Description Length Principle*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2007. ISBN 978-0-262-07281-6. Foreword by Jorma Rissanen.

Grünwald, P. and Roos, T. Minimum Description Length Revisited. *International Journal of Mathematics for Industry*, 11(01):1930001, December 2019. ISSN 2661-3352, 2661-3344. doi: 10.1142/S2661335219300018. URL http://arxiv.org/abs/1908.08484. arXiv:1908.08484 [stat].

Hamreras, S., Singh, S., and Orús, R. Tensorization is a powerful but underexplored tool for compression and interpretability of neural networks, May 2025.

Hinton, G. E. and Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006. doi: 10.1126/science.1127647. URL https://www.science.org/doi/abs/10.1126/science.1127647.

Hitzler, P. and Sarker, M. K. (eds.). *Neuro-Symbolic Artificial Intelligence: The State of the Art*, volume 342 of *Frontiers in Artificial Intelligence and Applications*. IOS Press, December 2021. ISBN 978-1-64368-244-0 978-1-64368-245-7. doi: 10.3233/FAIA342.

Hupkes, D., Dankers, V., Mul, M., and Bruni, E. Compositionality decomposed: How do neural networks generalise?, February 2020.

Joshi, S., Mueller, A., Klindt, D., Brendel, W., Reizinger, P., and Sridhar, D. Causality is Key for Interpretability Claims to Generalise, February 2026. URL http://arxiv.org/abs/2602.16698. arXiv:2602.16698 [cs].

Khatri, N., Laakkonen, T., Liu, J., and Wang-Maścianica, V. On the Anatomy of Attention, July 2024.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., and Sayres, R. Interpretability beyond feature attribution: Quantitative testing with concept activation

vectors (tcav), 2018. URL https://arxiv.org/abs/1711.11279.

Kowalski, R. Algorithm = logic + control. *Commun. ACM*, 22(7):424–436, July 1979. ISSN 0001-0782. doi: 10.1145/359131.359136.

Lin, H. W., Tegmark, M., and Rolnick, D. Why does deep and cheap learning work so well? *J Stat Phys*, 168(6): 1223–1247, September 2017. ISSN 0022-4715, 1572-9613. doi: 10.1007/s10955-017-1836-5.

Lindsey, J., Templeton, A., Marcus, J., Conerly, T., Batson, J., and Olah, C. Sparse crosscoders for cross-layer features and model diffing. https://transformer-circuits.pub/2024/crosscoders/index.html, October 2024. Transformer Circuits (Anthropic Interpretability Team). Accessed: 2025-09-26.

Lorenz, R. and Tull, S. Causal models in string diagrams, April 2023.

Lorenz, R. and Tull, S. Causal and Compositional Abstraction, February 2026.

Madsen, A., Lakkaraju, H., Reddy, S., and Chandar, S. Interpretability Needs a New Paradigm, November 2024.

Marr, D. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Henry Holt and Co., Inc., New York, NY, USA, 1982. ISBN 0716715678.

Montague, R. Universal grammar. *Theoria*, 36(3):373–398, 1970. doi: 10.1111/j.1755-2567.1970.tb00434.x.

Nainani, J., Reimer, B. M., Watts, C., Jensen, D., and Green, A. G. Mechanistic evidence that motif-gated domain recognition drives contact prediction in protein language models. *bioRxiv*, 2025. doi: 10.1101/2025.08.22.671739. URL https://www.biorxiv.org/content/early/2025/08/28/2025.08.22.671739.

Olah, C., Mordvintsev, A., and Schubert, L. Feature visualization. *Distill*, 2017. doi: 10.23915/distill.00007. https://distill.pub/2017/feature-visualization.

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., and Carter, S. Zoom In: An Introduction to Circuits. *Distill*, 5(3):e00024.001, March 2020. ISSN 2476-0757. doi: 10.23915/distill.00024.001.

Oldfield, J., Torr, P., Patras, I., Bibi, A., and Barez, F. Beyond Linear Probes: Dynamic Safety Monitoring for Language Models, September 2025.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Johnston, S., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. In-context learning and induction heads, 2022. URL https://arxiv.org/abs/2209.11895.

Pearce, M. T., Dooms, T., Yamamoto, R., Meehl, J., Molnar, C., Bissell, M., Hazra, D., Fang, C., Nguyen, N., Anderson, M., Osborne, C., Duffy, P., Toomey, B., Klee, E., Myasoedova, E., Ryu, A. J., Ayanian, S., Korfiatis, P., Redlon, M., Jain, A., Balsam, D., and Wang, N. K. Evee: Interpretable variant effect prediction from genomic foundation model embeddings. *bioRxiv*, 2026. doi: 10.64898/2026.04.10.717844. URL https://www.biorxiv.org/content/early/2026/04/16/2026.04.10.717844.

Puca, C., Hadzihasanovic, A., Genovese, F., and Coecke, B. Obstructions to Compositionality. *Electron. Proc. Theor. Comput. Sci.*, 397:226–245, December 2023. ISSN 2075-2180. doi: 10.4204/EPTCS.397.14.

Rodatz, B., Fan, I., Laakkonen, T., Ortega, N. J., Hoffmann, T., and Wang-Mascianica, V. A Pattern Language for Machine Learning Tasks, May 2025.

Sharkey, L., Chughtai, B., Batson, J., Lindsey, J., Wu, J., Bushnaq, L., Goldowsky-Dill, N., Heimersheim, S., Ortega, A., Bloom, J., Biderman, S., Garriga-Alonso, A., Conmy, A., Nanda, N., Rumbelow, J., Wattenberg, M., Schoots, N., Miller, J., Michaud, E. J., Casper, S., Tegmark, M., Saunders, W., Bau, D., Todd, E., Geiger, A., Geva, M., Hoogland, J., Murfet, D., and McGrath, T. Open Problems in Mechanistic Interpretability, January 2025.

Shiebler, D., Gavranović, B., and Wilson, P. Category Theory in Machine Learning, June 2021.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014. URL https://arxiv.org/abs/1312.6034.

Sutter, D., Minder, J., Hofmann, T., and Pimentel, T. The Non-Linear Representation Dilemma: Is Causal Abstraction Enough for Mechanistic Interpretability?, November 2025.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Freeman, C. D., Sumers, T. R., Rees, E., Batson, J., Jermyn, A., Carter, S., Olah, C., and Henighan, T. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL https:

//transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Tull, S., Kleiner, J., and Smithe, T. S. C. Active Inference in String Diagrams: A Categorical Account of Predictive Processing and Free Energy, August 2023a.

Tull, S., Shaikh, R. A., Zemljic, S. S., and Clark, S. From Conceptual Spaces to Quantum Concepts: Formalising and Learning Structured Conceptual Models, November 2023b.

Tull, S., Lorenz, R., Clark, S., Khan, I., and Coecke, B. Towards Compositional Interpretability for XAI, June 2024.

Voita, E. and Titov, I. Information-Theoretic Probing with Minimum Description Length. In Webber, B., Cohn, T., He, Y., and Liu, Y. (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 183–196, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.14.

Zhou, Y. and Shah, J. The Solvability of Interpretability Evaluation Metrics, February 2023.

A. String diagrams

String diagrams (Coecke & Kissinger, 2018) are a category-theoretic tool for visually describing compositional structures and formally reasoning about them. They place the composition of processes, representing the flow of information, at the centre of analysis. By linking topology to algebra, intuitive visual manipulations directly correspond to formal algebraic proofs of behavioural equivalence, with the mathematical details hidden from view. Diagrams can be manipulated via equational rules, summarised by the principle: "Only connectivity matters" (Coecke & Kissinger, 2018).

Definition A.1 (String Diagram)**.** A *string diagram* D is a graphical representation of a composite process (formally a morphism $D : I \to O$ in a monoidal category S) characterised by:

- *Wires* representing types (objects $A, B, C, \ldots \in \mathrm{Obj}(S)$);
- *Boxes* representing processes between types (morphisms $M : A \to B \in \mathrm{Hom}_S(A, B)$).

Diagrams are composed via:

- *Sequential composition:* Connecting output wires of one box to matching input wires of another vertically (composition of morphisms $M : A \to B$ and $N : B \to C$ gives $N \circ M : A \to C$);
- *Parallel composition:* Placing boxes side-by-side horizontally (monoidal product of morphisms $M : A \to C$ and $N : B \to D$ gives $M \otimes N : A \otimes B \to C \otimes D$)

String diagrams further allow defining *states* as boxes with no inputs, *effects* as boxes with no outputs, and *scalars* as boxes with neither inputs nor outputs. Figure 1 gives examples of string diagrams; comprehensive introductions are available elsewhere (Coecke & Kissinger, 2018; Shiebler et al., 2021; Tull et al., 2024).

B. Proof of the description length bound

Proof. Proof of Theorem 3.3 By Definition 2.4

$$L(I_S(D)) = L(I_C([[\, D \,]]))$$

The description length of a mapping's output is bounded by the sum of the description length of the mapping and its input (information non-increase). Therefore

$$L(I_C([[\, D \,]])) \leq L([[\, D \,]]) + L(I_C|_{C_D})$$

Because, by Definition 2.3, $[[\,\cdot\,]]$ is a structure-preserving (monoidal) functor, the description length of the semantic object $[[\, D \,]]$ is bounded by the sum of the description lengths of its constituent parts and their arrangement (information subadditivity)

$$L([[\, D \,]]) \leq L(D) + \sum_{f \in \mathrm{mor}(S_D)} L([[\, f \,]])$$

Combining these two inequalities yields

$$L(I_S(D)) \leq L(D) + \sum_{f \in \mathrm{mor}(S_D)} L([[\, f \,]]) + L(I_C|_{C_D}) = L^{\mathrm{rep}}(M) + L^{\mathrm{int}}(M, I_C)$$

The right-hand side matches the definition of $L(M, I_S)$ in Definition 3.2, proving that $L(I_S(D)) \leq L(M, I_S)$. □

C. Comparison of interpretability methods

Mechanistic interpretability methods trade off the scope of refinement against traceability to the original model: a broader refinement space can expose more natural atoms of computation, but at the cost of losing that trace. Table 1 lists common methods along this spectrum, with a representative example for each category.

D. Worked example of representation complexity

To illustrate the representation complexity $L^{\mathrm{rep}}(M)$, we compute it for the linear animal-and-colour classifier of Figure 1, encoding each parameter with 8 bits and each box or wire with 16 bits. The encoding is naive and not comonotonic (Corollary 4.4), and the specific constants matter less than the direction of the inequality. The model maps 64 input features to logits over 4 classes (two animals, two colours), split across three wires of 8, 24, and 32 features.

Level	Constraint on R	Limitation on I_S
L0. Post-hoc saliency maps (Simonyan et al., 2014)	R is not constructed.	No syntax-based interpretation I_S is created; two models that implement the same function by different mechanisms receive the same explanation.
L1. Architectural attention heads (Olsson et al., 2022)	$R = \mathrm{id}_S$.	The atoms of analysis are fixed by the architecture. Any functional unit that crosses architectural boundaries, or lives inside one, cannot be captured.
L2. Piecewise transcoders (Dunefsky et al., 2024)	$d([[\, f \,]], [[\, R(f) \,]]') \leq \epsilon$ for each component f of D.	Each component is refined in isolation. Regularities that span or combine components (cross-layer superposition) are structurally excluded — refining them would break the per-component match.
L3. Global attribution graphs (Lindsey et al., 2024)	$d([[\, D \,]], [[\, R(D) \,]]') \leq \epsilon$ only.	Individual components of S need not correspond to anything in S'. Localising a claim to a specific component requires an additional invariance in C (e.g. linearity of the residual stream).
L4. Intrinsic weight-sparse transformers (Gao et al., 2025)	R is not constrained.	R is trivial as an interpretation of a pretrained source model; the framework analyses M' directly rather than explaining a given M.

Table 1. Interpretability methods as subclasses of refinement. Reading down column 2, the constraint on R weakens from per-component match, to whole-diagram match, to trivial match against a collapsed source; reading down column 3, what each method cannot see grows accordingly. L4 fits only as a limiting case: the framework analyses M' directly, not as an interpretation of any pretrained model.

Naive decomposition A naive decomposition treats the classifier as a single dense map $[[\, W \,]] \in \mathbb{R}^{4\times 64}$: one box with an input and an output wire (3 elements).

$$L^{\text{rep}} = \underbrace{3 \cdot 16}_{L(D)} + \underbrace{4 \cdot 64 \cdot 8}_{\sum_f L([[\, f \,]])} = 48 + 2048 = 2096 \text{ bits}$$

Entangled decomposition S The original syntax of Figure 1 routes the first two input wires through a bottleneck $[[f]] \in \mathbb{R}^{4\times(8+24)}$ that feeds both an animal head $[[g]] \in \mathbb{R}^{2\times 4}$ and a colour head $[[h]] \in \mathbb{R}^{2\times(4+32)}$ (10 elements). Both heads read $[[f]]$, so they are entangled.

$$L^{\text{rep}}(M) = \underbrace{10 \cdot 16}_{L(D)} + \underbrace{(4 \cdot (8+24) + 2 \cdot 4 + 2 \cdot (4+32)) \cdot 8}_{\sum_f L([[\, f \,]])} = 160 + 1664 = 1824 \text{ bits}$$

Refined decomposition S' The compressive refinement R replaces $[[h]]$ by $[[h']] \in \mathbb{R}^{2\times 32}$, reading only its own input wire, while $[[f]]$ and $[[g]]$ are unchanged (9 elements). This severs the wire carrying the hidden representation to the colour head, disentangling the two paths.

$$L^{\text{rep}}(M') = \underbrace{9 \cdot 16}_{L(D')} + \underbrace{(4 \cdot (8+24) + 2 \cdot 4 + 2 \cdot 32) \cdot 8}_{\sum_f L([[\, f \,]]')} = 144 + 1600 = 1744 \text{ bits}$$

Representation complexity decreases at each step, $2096 > 1824 > 1744$, so the refinement is compressive. It also disentangles the colour head from the animal pathway, which does not raise interpretation complexity, so by the parsimony criterion (Proposition 4.2) the refined model gives a more concise compositional explanation.

E. Notation

The paper uses colour to distinguish three roles: syntax, semantics, and explanations. A symbol's colour indicates the specific category an object or morphism inhabits, or the target category of a functor. Uncoloured symbols fall outside these categories.

Compositional models

- $M = (D, S, C, [[\,\cdot\,]])$ — a compositional model (Definition 2.3).

- $\boldsymbol{S}$ — syntactic monoidal category; $\boldsymbol{S_D} = (\mathrm{ob}(\boldsymbol{S_D}), \mathrm{mor}(\boldsymbol{S_D}))$ its signature (objects and generators appearing in $\boldsymbol{D}$).
- $\boldsymbol{C}$ — semantic monoidal category; $\boldsymbol{C_D}$ the image of $\boldsymbol{S_D}$ under the representation.
- $\boldsymbol{D} : \boldsymbol{I} \to \boldsymbol{O}$ — string diagram over $\boldsymbol{S}$; $\boldsymbol{f} \in \mathrm{mor}(\boldsymbol{S_D})$ a generator of $\boldsymbol{D}$.
- $[[\,\cdot\,]] : \boldsymbol{S} \to \boldsymbol{C}$ — monoidal representation functor; $[[\,\boldsymbol{f}\,]]$ the semantic instantiation of the generator $\boldsymbol{f}$; $[[\,\boldsymbol{D}\,]]$ the semantics of the full diagram.
- $M' = (\boldsymbol{D'}, \boldsymbol{S'}, \boldsymbol{C}, [[\,\cdot\,]]')$ — a refined compositional model sharing semantics $\boldsymbol{C}$ with M (Definition 4.1).

Interpretations

- $\boldsymbol{H}$ — signature of human-interpretable terms.
- $\boldsymbol{I_S} : \boldsymbol{S} \to \boldsymbol{H}$ — compositional interpretation (syntactic grounding).
- $\boldsymbol{I_C} : \boldsymbol{C} \to \boldsymbol{H}$ — post-hoc interpretation (semantic grounding); $\boldsymbol{I_C}|_{\boldsymbol{C_D}}$ its restriction to the semantic components appearing in $\boldsymbol{D}$.
- $\boldsymbol{I_C^*} : \boldsymbol{H} \to \boldsymbol{C}$ — reconstruction map witnessing compositional faithfulness (Definition 3.1).
- Commutativity $\boldsymbol{I_S} = \boldsymbol{I_C} \circ [[\,\cdot\,]]$ links the two groundings.

Refinements

- $\boldsymbol{R} : \boldsymbol{S} \to \boldsymbol{S'}$ — syntactic refinement; a functor with $\boldsymbol{D'} = \boldsymbol{R}(\boldsymbol{D})$ and $[[\,\cdot\,]] = [[\,\cdot\,]]' \circ \boldsymbol{R}$.
- $\boldsymbol{R} = \mathrm{id}_{\boldsymbol{S}}$ — the identity refinement (L1 in Table 1).
- A refinement is *compressive* when $L^{\mathrm{rep}}(M') \leq L^{\mathrm{rep}}(M)$.

Description lengths

- $L(\,\cdot\,)$ — description length (in bits).
- $L^{\mathrm{rep}}(M) = L(\boldsymbol{D}) + \sum_{\boldsymbol{f} \in \mathrm{mor}(\boldsymbol{S_D})} L([[\,\boldsymbol{f}\,]])$ — representation complexity; upper-bounds $L([[\,\boldsymbol{D}\,]])$ (Appendix B).
- $L^{\mathrm{int}}(M, \boldsymbol{I_C}) = L(\boldsymbol{I_C}|_{\boldsymbol{C_D}})$ — interpretation complexity.
- $L(M, \boldsymbol{I_S}) = L^{\mathrm{rep}}(M) + L^{\mathrm{int}}(M, \boldsymbol{I_C})$ — compositional description length; upper-bounds $L(\boldsymbol{I_S}(\boldsymbol{D}))$ (Theorem 3.3).

Faithfulness

- d — semantic distortion metric on parallel morphisms in $\boldsymbol{C}$ (or a generalisation thereof).
- $\epsilon \in \mathbb{R}$ — tolerance for relaxed equalities.
- $d([[\,\boldsymbol{D}\,]], [[\,\boldsymbol{R}(\boldsymbol{D})\,]]') \leq \epsilon$ — behavioural faithfulness of a refinement $\boldsymbol{R}$.
- $d([[\,\boldsymbol{f}\,]], \boldsymbol{I_C^*}(\boldsymbol{I_S}(\boldsymbol{f}))) \leq \epsilon$ for every $\boldsymbol{f} \in \mathrm{mor}(\boldsymbol{S})$ — compositional faithfulness of an explanation (Definition 3.1).