AIMLAI @ ECML PKDD 2026
September 7th 2026

From Mechanistic to Compositional Interpretability

Ward Gauderis · Thomas Dooms

VUB ARTIFICIAL INTELLIGENCE RESEARCH GROUP
Universiteit Antwerpen

fwo Flanders AI Research Program

This chapter...

Interpretability Paradigms

Ward Gauderis

Compositional Interpretations

Ward Gauderis & Thomas Dooms

Architectures & Algorithms

Thomas Dooms

As AI adoption grows, regulation asks questions our interpretability tools *cannot* yet answer.

Post-hoc interpretability

"What information was used to make this particular decision?"

Behavioural faithfulness
Decisions, but no mechanisms

(Madsen, *Interpretability Needs a New Paradigm*. 2024; Sharkey, *Open Problems in Mechanistic Interpretability*. 2025)

As AI adoption grows, regulation asks questions our interpretability tools *cannot* yet answer.

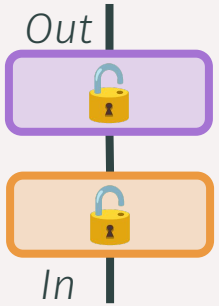
Post-hoc interpretability

"What information was used to make this particular decision?"

Behavioural faithfulness
Decisions, but no mechanisms

Intrinsic interpretability

"How do we design a model that explains all its behaviour? "



Inherent faithfulness
Guarantees, but constrained

(Madsen, *Interpretability Needs a New Paradigm*. 2024; Sharkey, *Open Problems in Mechanistic Interpretability*. 2025)

As AI adoption grows, regulation asks questions our interpretability tools *cannot* yet answer.

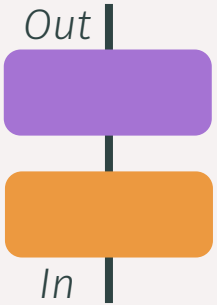
Post-hoc interpretability

"What information was used to make this particular decision?"

Behavioural faithfulness
Decisions, but no mechanisms

Mechanistic interpretability

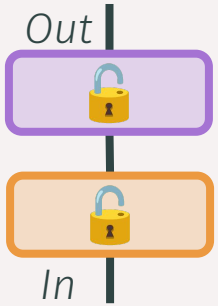
"How does the model solve this general class of problems?"



Explanatory faithfulness
Heuristic, but no foundations

Intrinsic interpretability

"How do we design a model that explains all its behaviour? "



Inherent faithfulness
Guarantees, but constrained

(Madsen, *Interpretability Needs a New Paradigm*. 2024; Sharkey, *Open Problems in Mechanistic Interpretability*. 2025)

Interpretability Paradigms

Ward Gauderis

Compositional Interpretations

Ward Gauderis & Thomas Dooms

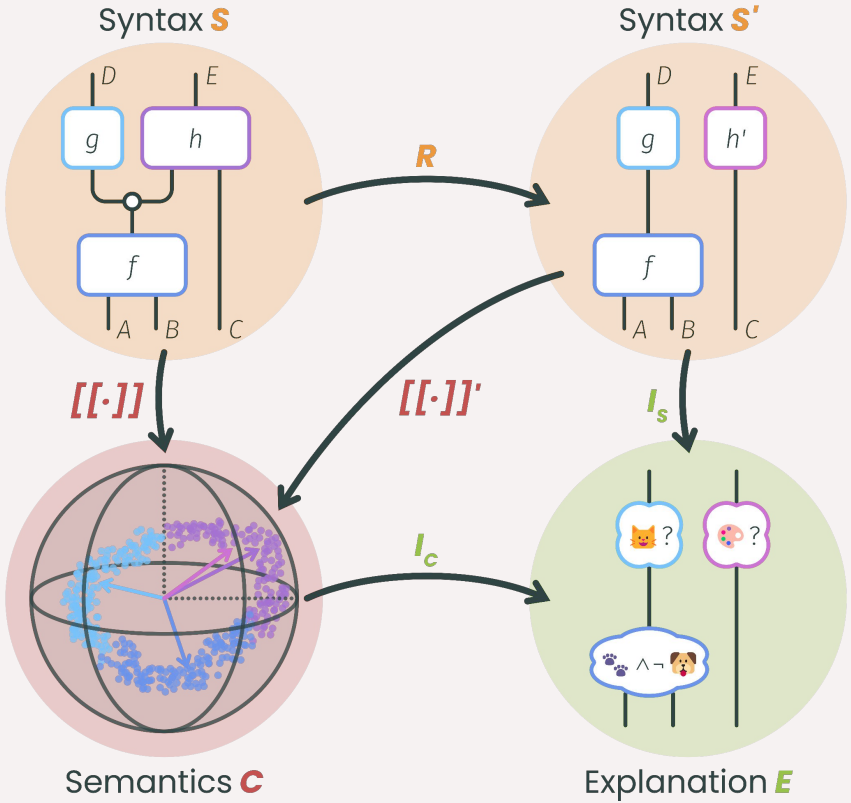
Architectures & Algorithms

Thomas Dooms

Compositional interpretability is a formal framework to evaluate, compare and improve explanations.

Two model-agnostic principles, formalised,

cast interpretability as **constrained optimisation.**



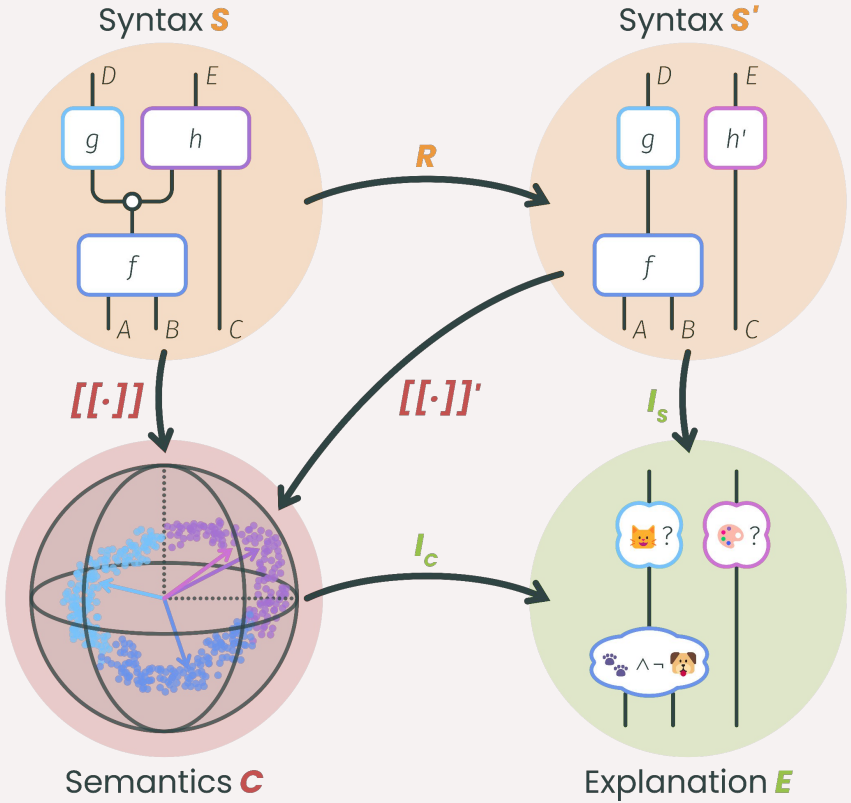
(Gauderis, Dooms, *From Mechanistic to Compositional Interpretability*. 2026)

Compositional interpretability is a formal framework to evaluate, compare and improve explanations.

Two model-agnostic principles, formalised,

- **Compositionality**
  - Category Theory

cast interpretability as **constrained optimisation.**



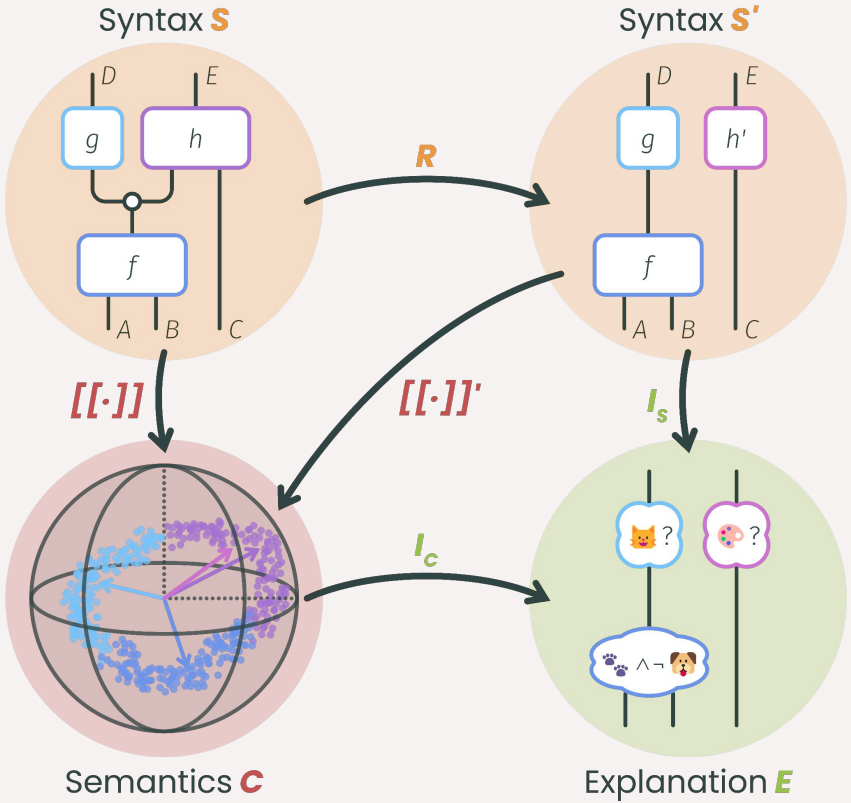
(Gauderis, Dooms, *From Mechanistic to Compositional Interpretability.* 2026)

Compositional interpretability is a formal framework to evaluate, compare and improve explanations.

Two model-agnostic principles, formalised,

- **Compositionality**
  - Category Theory
- **Occam's Razor**
  - Minimum Description Length

cast interpretability as **constrained optimisation.**

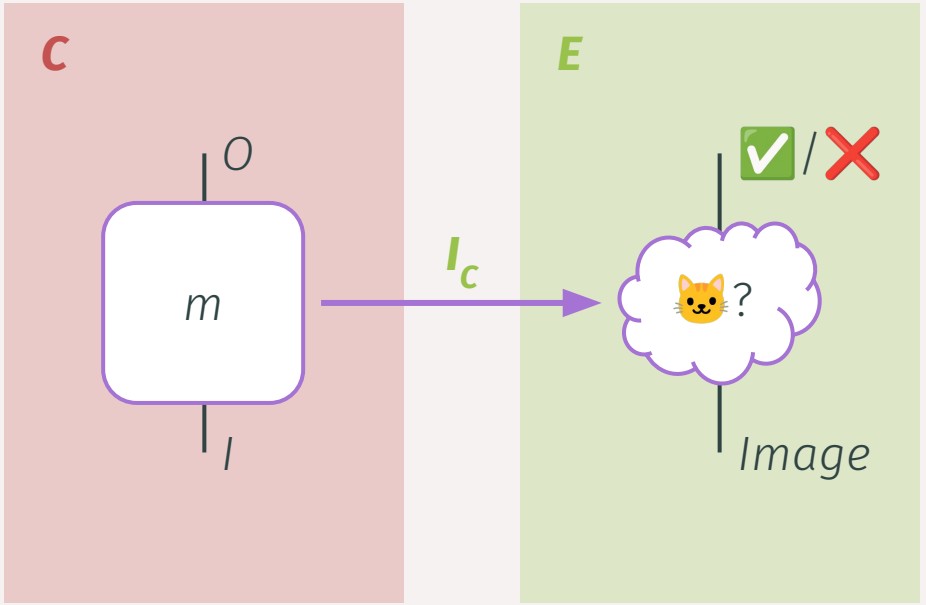


(Gauderis, Dooms, *From Mechanistic to Compositional Interpretability*. 2026)

Coming up...

1. **What are compositional interpretations?**
2. What makes explanations good?
3. How do we find better explanations?

Post-hoc interpretations only map input-output behaviour to explanations.

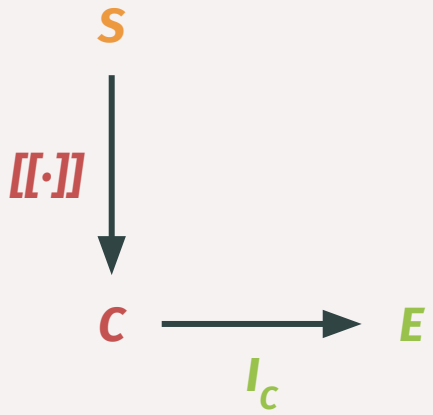


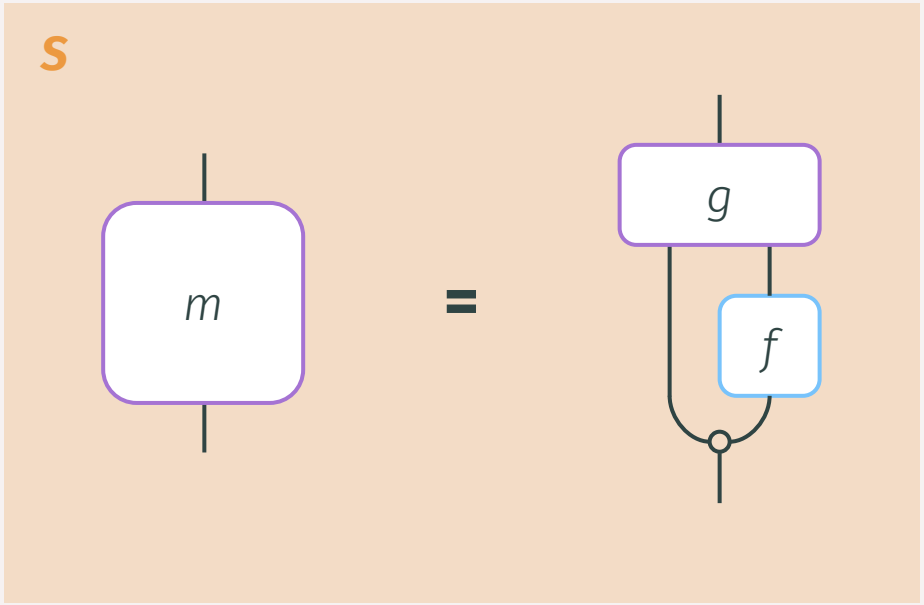
C Semantics
E Explanation

I_C Semantic interpretation

(Gauderis, Dooms, *From Mechanistic to Compositional Interpretability*. 2026)

Compositional models abstract high-level syntax from low-level semantics.



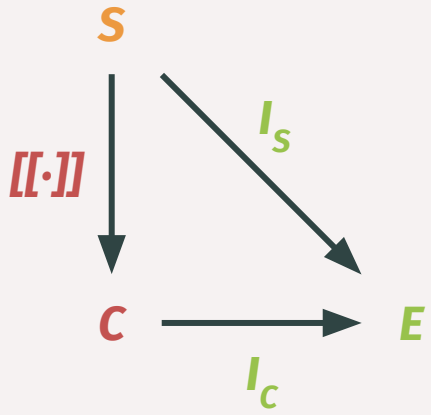


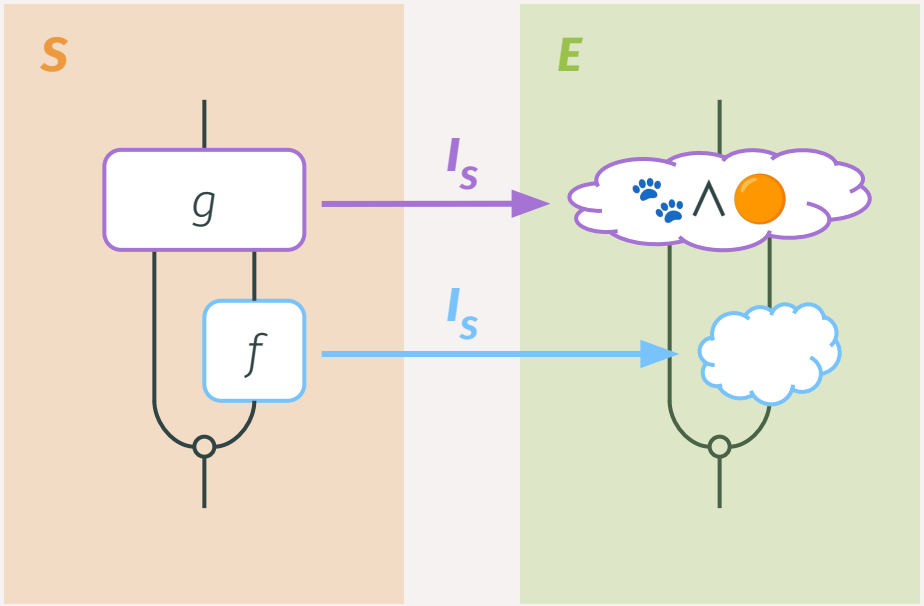
S Syntax
C Semantics
E Explanation

[[·]] Representation
I_C Semantic interpretation

(Tull, *Towards Compositional Interpretability for XAI*. 2024; Gauderis, Dooms, *From Mechanistic to Compositional Interpretability*. 2026)

Compositional interpretations mechanistically explain how a model works, step-by-step.





S Syntax
C Semantics
E Explanation

[[·]] Representation
I_C Semantic interpretation
I_S Syntactic interpretation

(Tull, *Towards Compositional Interpretability for XAI*. 2024; Gauderis, Dooms, *From Mechanistic to Compositional Interpretability*. 2026)

Category theory and string diagrams are the mathematics of compositional structure.

A **compositional model** $M = (D, G, [\![-]\!], C)$ consists of:

- G, a signature inducing $S = Free(G)$, an abstract syntactic category
- $D \in Hom\square(I, O)$, a composite morphism in S
- C, a concrete semantic category
- $[\![-]\!] : S \to C$, a representation functor

A **Category** C consists of:

- $ob(C)$, a class of objects
- $hom_c(A, B)$, a class of morphisms for $A, B \in ob(C)$
- $id_a \in hom_c(A, A)$, an identity morphism for $A \in ob(C)$
- $\circ : hom_c(B, C) \times hom_c(A, B) \to hom_c(A, C)$, an associative operator that respects identity morphisms

A **compositional interpretation** $I(I_a, I_c, I_c^*, H)$ for M into H consists of:

- $I_a : S \to H$, an abstract interpretation functor
- $I_c : C \to H$, a concrete interpretation functor s.t.

$$I_a = I_c \circ [\![-]\!]$$

- $I_c^* : H \to C$, the left adjoint to I_c ($I_c^* \dashv I_c$) defined by the natural isomorphism

$$Hom_c(I^*(h), c) \cong Hom\square(h, I_c(c))$$

A **string diagram** is a visual representation of a composite morphism in a **symmetric monoidal category** C equipped with:

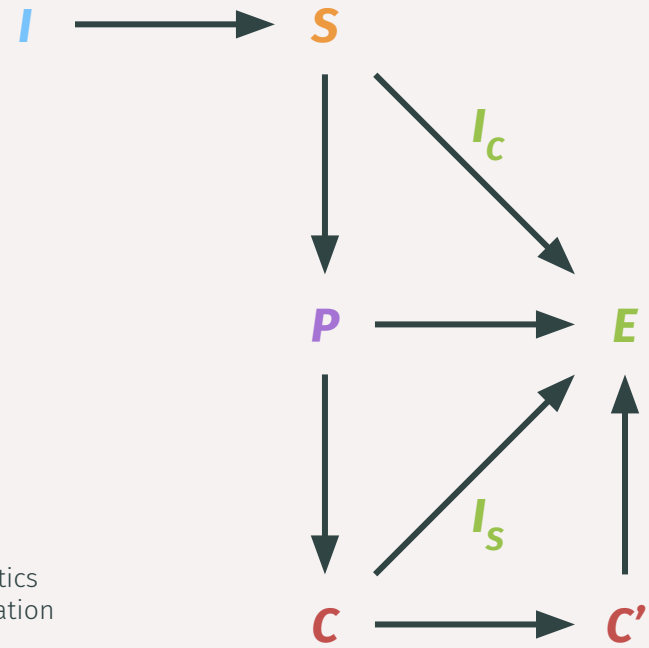
- $I \in ob(C)$, the monoidal unit object
- $\otimes : C \times C \to C$, a bifunctor for the tensor product
- Natural isomorphisms for the associator, symmetry, left and right unitors (left as an exercise to the reader)

Examples include:

- **CartSp** of Euclidean spaces and continuous functions
- **Stoch** of finite sets and probability channels
- **Quant** of finite-dimensional Hilbert spaces and CP-maps

(Tull, *Towards Compositional Interpretability for XAI*. 2024)

This language lets us formalise and structurally compare interpretability methods.



S Syntax
C Semantics
E Explanation

I_C Semantic interpretation
I_S Syntactic interpretation

I Input-dependent structure

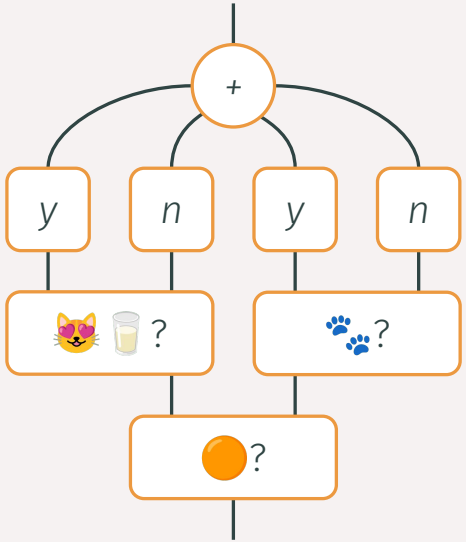
P Parameter space

C' Proxy semantics

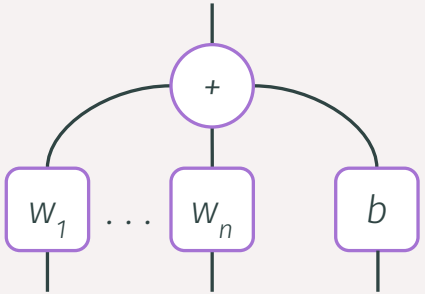
(Duneau, *Towards a Comparative Framework for Compositional AI Models*. 2025; Braun, *Open Problems in Mechanistic Interpretability*. 2025; Bushnaq, *Stochastic Parameter Decomposition*. 2025)

The objective of interpretability is to search for complete compositional interpretations.

Decision tree



Linear regression



Intrinsically interpretable models have compositional interpretations.

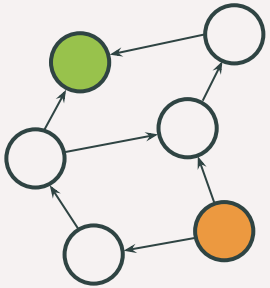
(Tull, *Towards Compositional Interpretability for XAI*. 2024; Madsen, *Interpretability Needs a New Paradigm*. 2024)

Coming up...

1. What are compositional interpretations?
2. **What makes explanations good?**
3. How do we find better explanations?

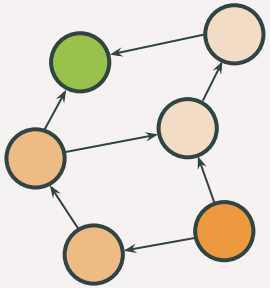
What makes a fastest path algorithm good?

Brute force



exact

Dijkstra

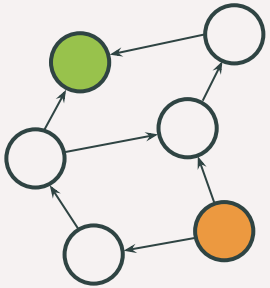


exact

(Kowalski, *Algorithm = Logic + Control*. 1979)

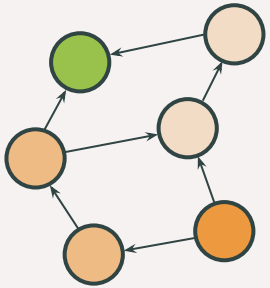
What makes a fastest path algorithm good?

Brute force



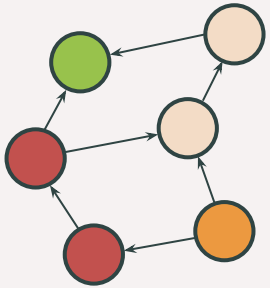
exact

Dijkstra



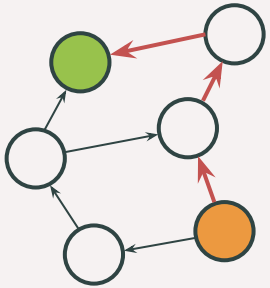
exact

Weighted A*



bounded error

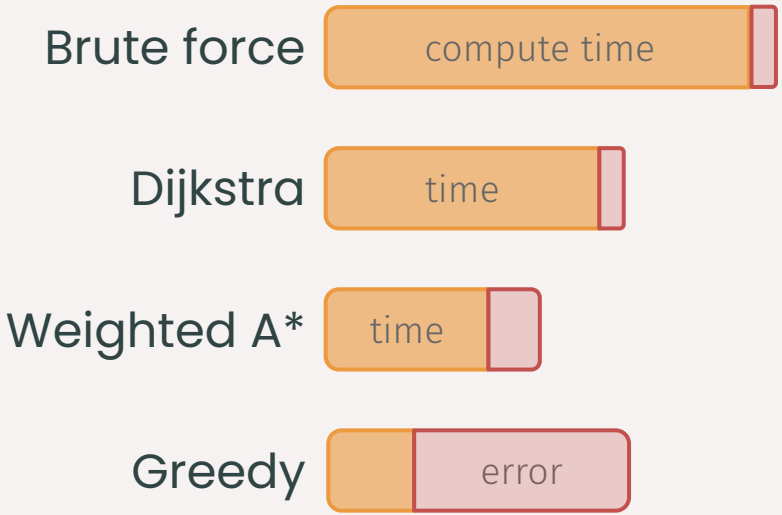
Greedy



unbounded error

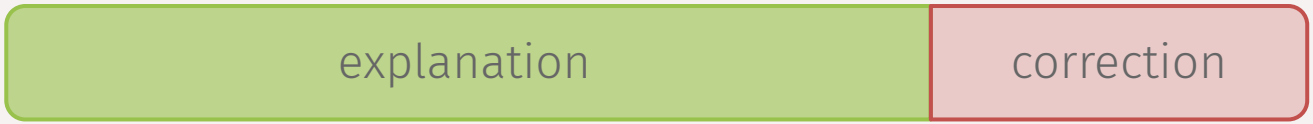
(Kowalski, *Algorithm = Logic + Control*. 1979)

We can rank an algorithm goodness based on time and error.



(Kowalski, *Algorithm = Logic + Control*. 1979)

Interpretability is a communication problem: the best explanations are the most compressed.



L(model) ***L(data | model)***

description length

$$L(data) = L(model) + L(data \mid model)$$

(Ayorinde, *A Mathematical Philosophy of Explanations in Mechanistic Interpretability*. 2025; Gauderis, Dooms, *From Mechanistic to Compositional Interpretability*. 2026)

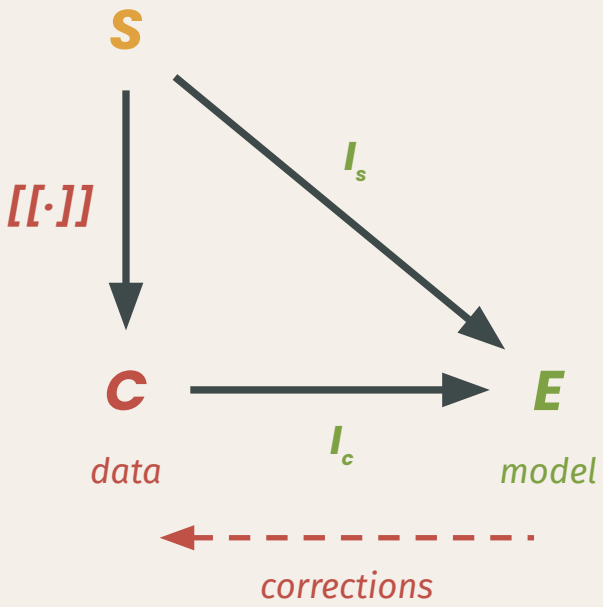
Interpretability is a communication problem: the best explanations are the most compressed.

description length

$$L(Data) = L(Model) + L(Data \mid Model)$$

(Rissanen, Modeling by Shortest Data Description. 1978)

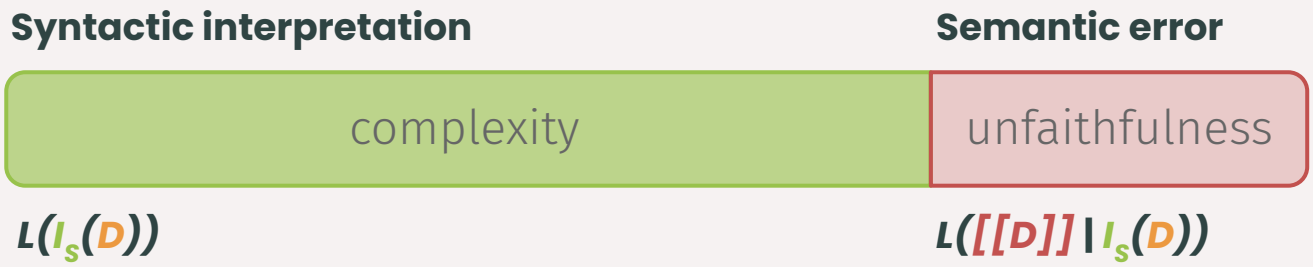
Describing the model through an explanation has two different costs.



$$L([[D]]) \leq L(I_s(D)) + L([[D]] \mid I_s(D))$$

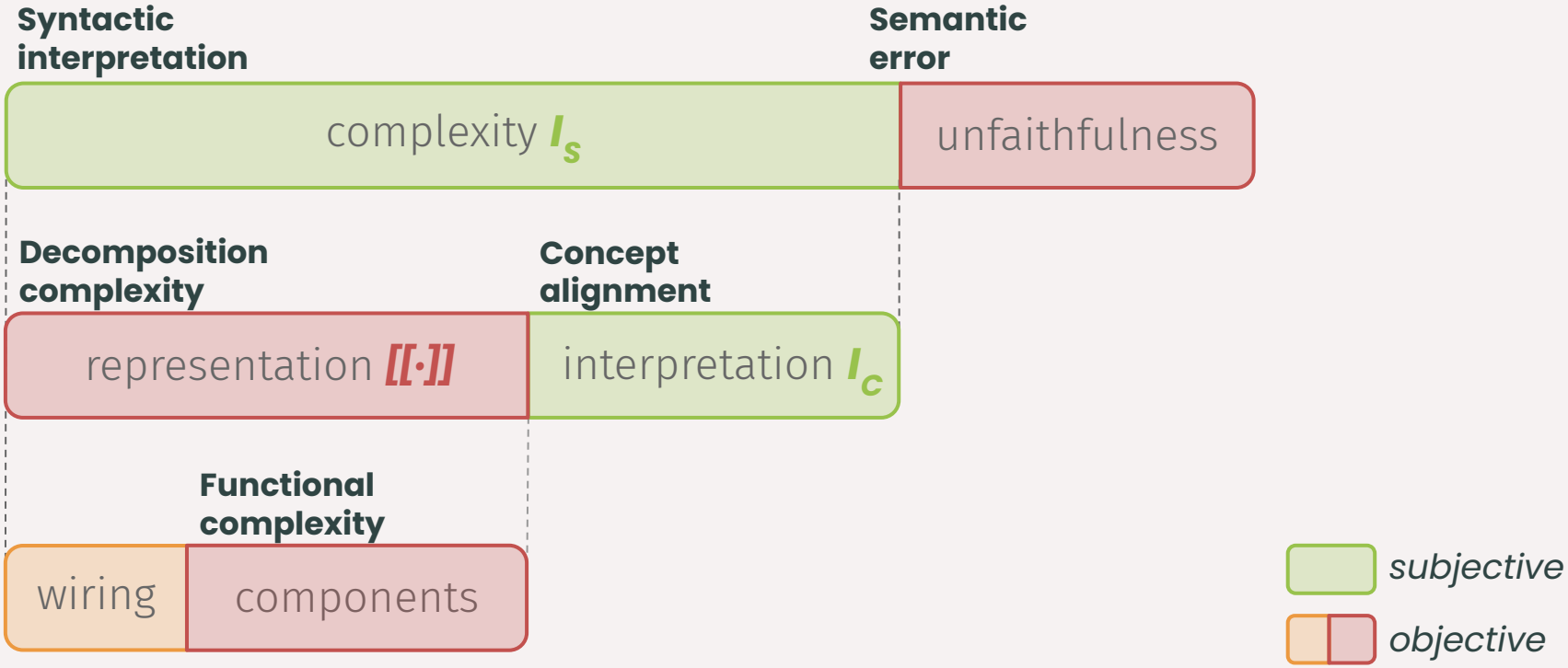
(Rissanen, Modeling by Shortest Data Description. 1978; Gauderis, Dooms, From Mechanistic to Compositional Interpretability. 2026)

For interpretations, we formalise this ranking with the notion of description length.



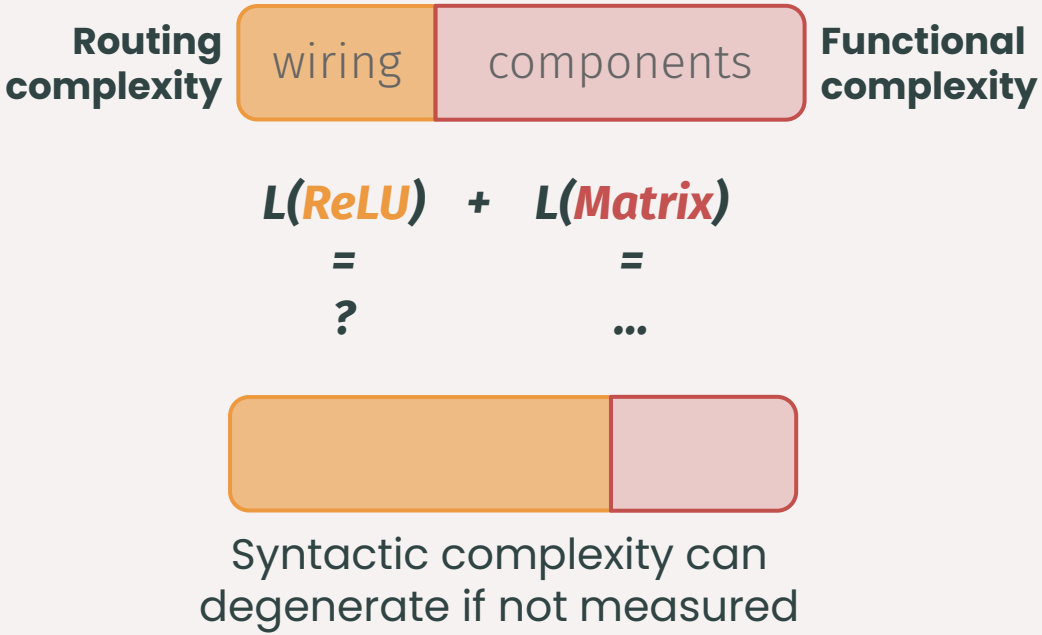
(Ayorinde, *A Mathematical Philosophy of Explanations in Mechanistic Interpretability*. 2025; Gauderis, Dooms, *From Mechanistic to Compositional Interpretability*. 2026)

Since an interpretations is subjective, we decompose into parts.



(Gauderis, Dooms, *From Mechanistic to Compositional Interpretability*. 2026)

Description length gives us a framework to assess when interpretations cheat.

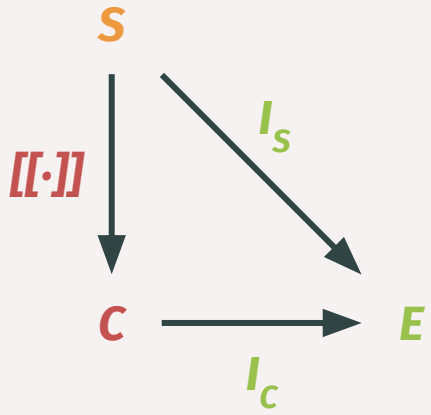


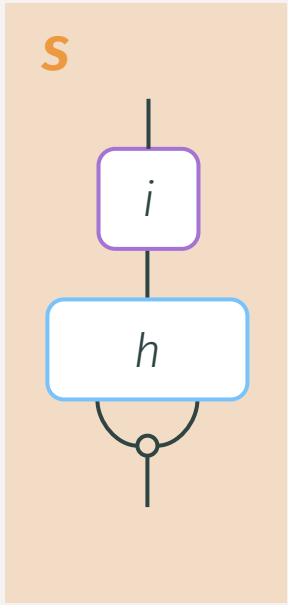
(Gauderis, Dooms, *From Mechanistic to Compositional Interpretability*. 2026)

Coming up...

1. What are compositional interpretations?
2. What makes explanations good?
3. **How do we find better explanations?**

Compressive refinements can simplify representations without changing model behaviour.





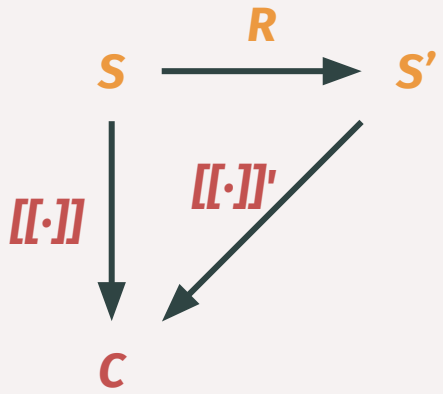
$L([[\cdot]])$

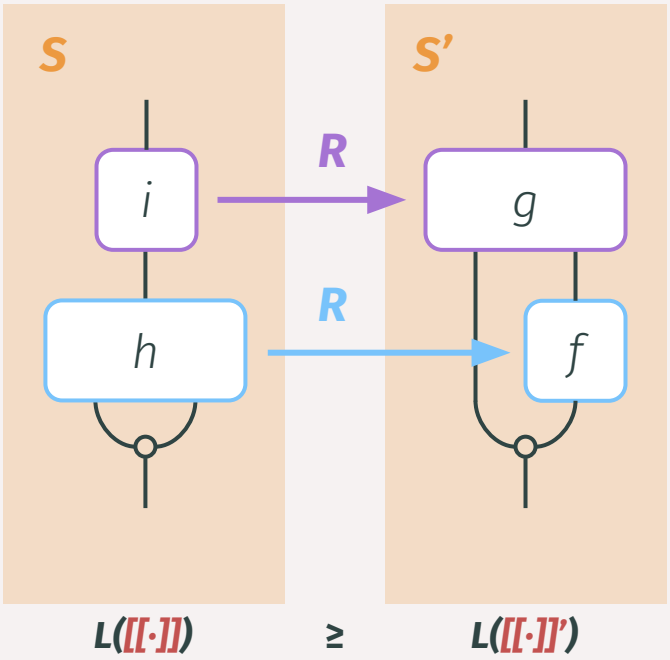
S Syntax
C Semantics
E Explanation

[[·]] Representation
I_C Semantic interpretation
I_S Syntactic interpretation

(Gauderis, Dooms, *From Mechanistic to Compositional Interpretability*. 2026)

Compressive refinements can simplify representations without changing model behaviour.



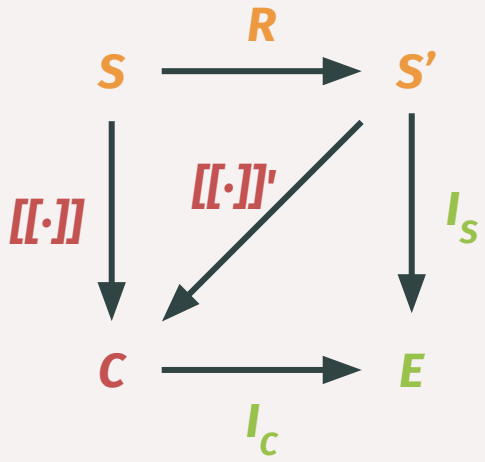


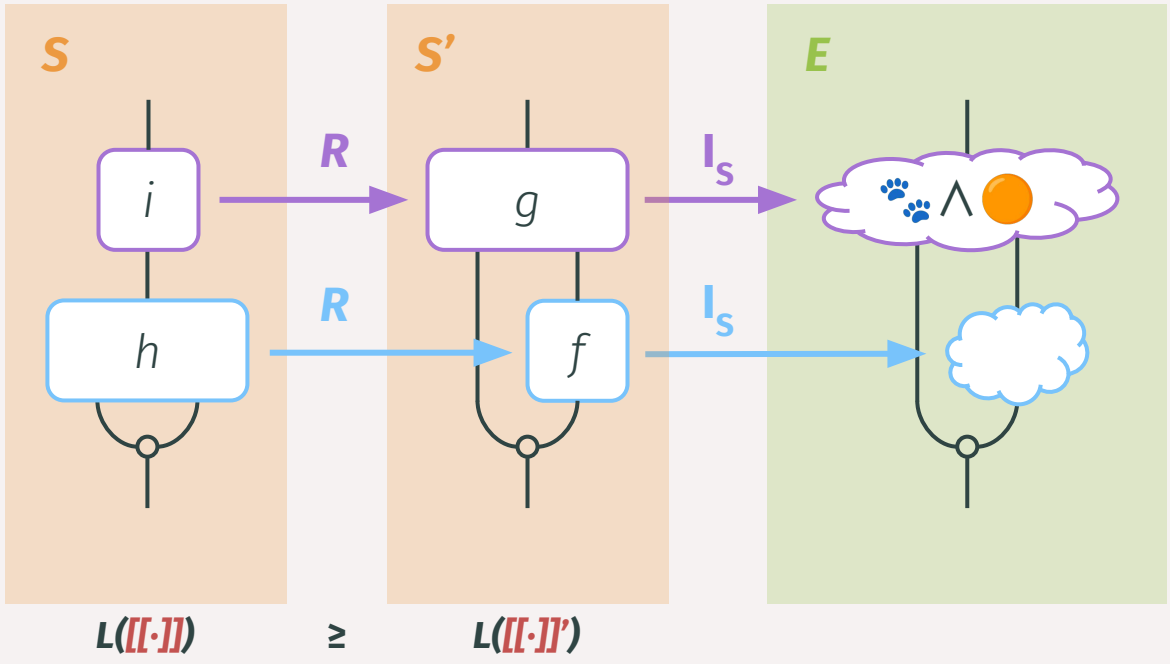
S Syntax
C Semantics
E Explanation

R Syntactic refinement
[[·]] Representation
I_C Semantic interpretation
I_S Syntactic interpretation

(Gauderis, Dooms, *From Mechanistic to Compositional Interpretability*. 2026)

Compressive refinements can simplify representations without changing model behaviour.



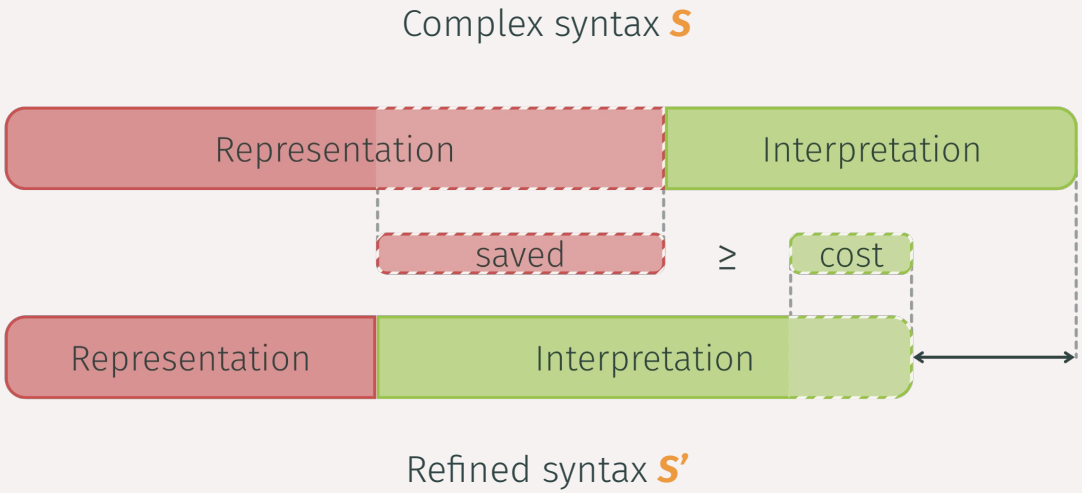


S Syntax
C Semantics
E Explanation

R Syntactic refinement
[[·]] Representation
I_C Semantic interpretation
I_S Syntactic interpretation

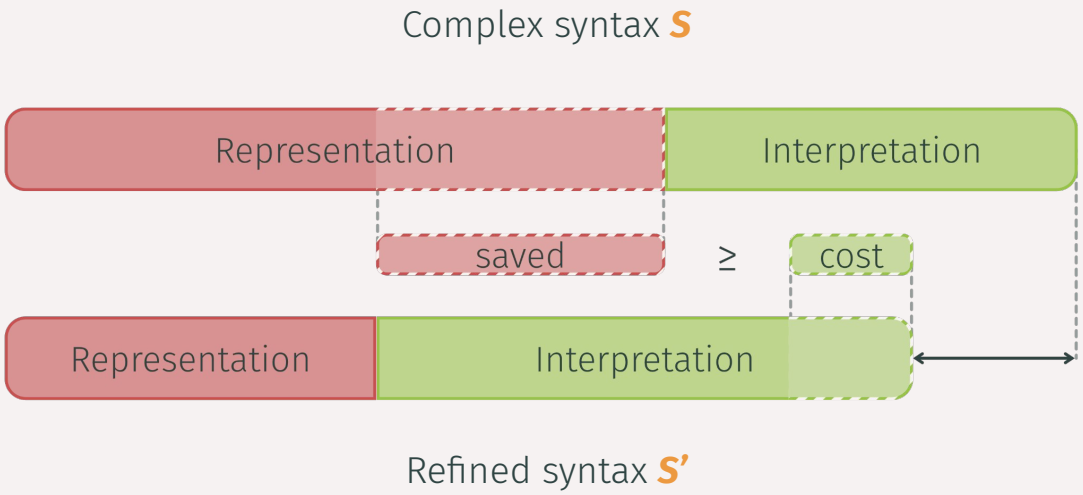
(Gauderis, Dooms, *From Mechanistic to Compositional Interpretability*. 2026)

Simpler does not always mean more interpretable… But it does under the *parsimony criterion*.



(Ayorinde, *A Mathematical Philosophy of Explanations in Mechanistic Interpretability*. 2025; Gauderis, Dooms, *From Mechanistic to Compositional Interpretability*. 2026)

Simpler does not always mean more interpretable…
But it does under the *parsimony criterion*.



Comonotonic coding suffices.

A coding distribution is a prior.
Encode *explanatory optimism* in it,
and both complexities fall together.

$$L([\![\cdot]\!]) := -\log_2 P([\![\cdot]\!])$$

(Ayorinde, *A Mathematical Philosophy of Explanations in Mechanistic Interpretability*. 2025; Gauderis, Dooms, *From Mechanistic to Compositional Interpretability*. 2026)

Make faithfulness and complexity *measurable*, and interpretability becomes *constrained optimisation*.

Search space
refinement in ***S***

Constraint
faithful in ***C***

Objective
complexity *of* ***[[·]]***

(Gauderis, Dooms, *From Mechanistic to Compositional Interpretability*. 2026)

Make faithfulness and complexity *measurable*, and interpretability becomes *constrained optimisation*.

Search space
refinement in ***S***

Constraint
faithful in ***C***

Objective
complexity *of* ***[[·]]***

Interpretability is a spectrum, not a constraint.

- Including post-hoc, mechanistic and intrinsic

(Gauderis, Dooms, *From Mechanistic to Compositional Interpretability*. 2026)

Make faithfulness and complexity *measurable*, and interpretability becomes *constrained optimisation*.

Search space
refinement in ***S***

Constraint
faithful in ***C***

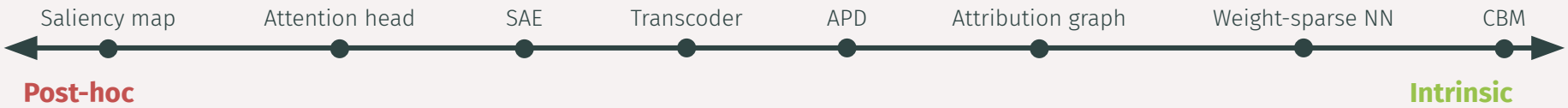
Objective
complexity *of* ***[[·]]***

Interpretability is a spectrum, not a constraint.

- Including post-hoc, mechanistic and intrinsic

Interpretability is relative to assumptions on *E*.

- To compare across ***E***s, use description length.
- Choosing ***E*** is also a question for cognitive science.



(Gauderis, Dooms, *From Mechanistic to Compositional Interpretability*. 2026)

Make faithfulness and complexity *measurable*, and interpretability becomes *constrained optimisation*.

Search space
refinement in ***S***

Constraint
faithful in ***C***

Objective
complexity *of* ***[[·]]***

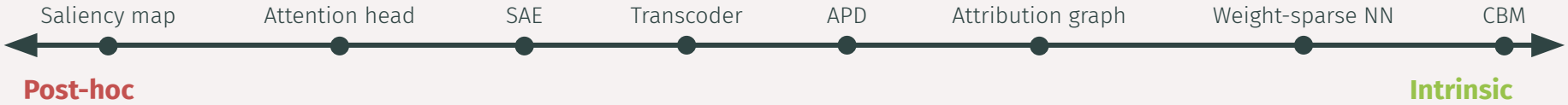
Interpretability is a spectrum, not a constraint.

- Including post-hoc, mechanistic and intrinsic

Interpretability is relative to assumptions on *E*.

- To compare across ***E***s, use description length.
- Choosing ***E*** is also a question for cognitive science.

If description length stays high, design new models.



(Gauderis, Dooms, *From Mechanistic to Compositional Interpretability*. 2026)

Interpretability Paradigms

Ward Gauderis

Compositional Interpretations

Ward Gauderis & Thomas Dooms

Architectures & Algorithms

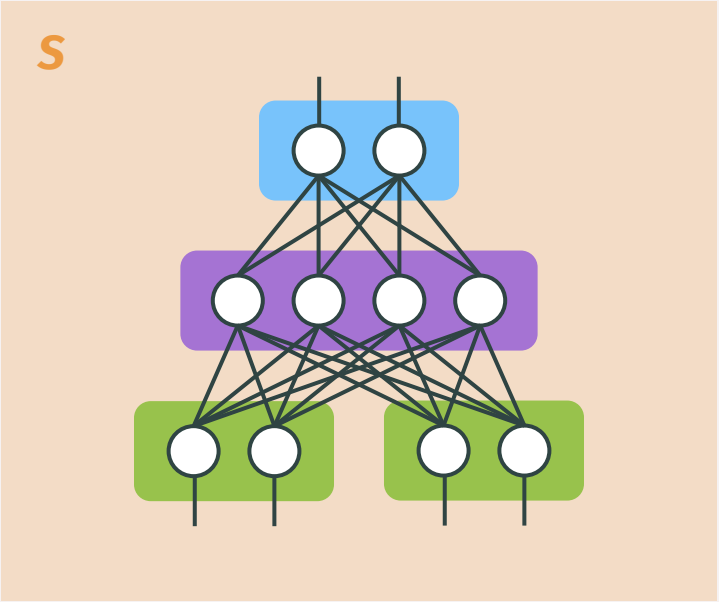
Thomas Dooms

Neural networks are hard to refine and to measure. Syntactic refinement becomes approximation.

No interesting way to rewrite structure

A **neural model** is a compositional model with
- the syntax of a **copy-discard** string diagram
- semantics in **CartSp** (continuous maps between Cartesian spaces)

A box can be any function (in theory)



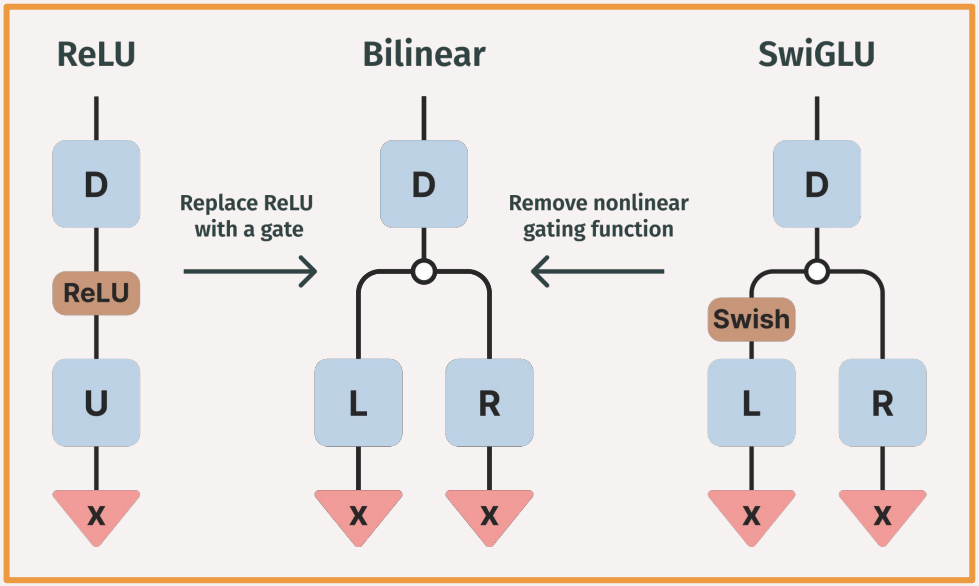
(Khatri, *Graphical einops: bridging tensor networks and computation graphs*. 2026; Méloux, *Is Mechanistic Interpretability Identifiable?*. 2025)

Tensor models are tractable to refine and measure. Their structure resembles neural architectures.

A rich framework for decompositions

A **tensor model** is a compositional model with
- the syntax of an **additive hypergraph**
- semantics in **FdHilb$\mathbb{R}$** (linear maps and finite-dimensional real Hilbert spaces)

A box is a higher-order interaction



(Pearce, *BIlinear MLPs Enable Weight-based Mechanistic Interpretability*. 2024; Dooms, Gauderis, *Compositionality Unlocks Deep Interpretable models*. 2024)

The appealing properties of compositional understanding can be scaled to state-of-the-art networks using tensor-based models.

Coming up...

1. **How do we operationalise compositional interpretability?**
2. What can we do with these models?
3. How are these used in practice?

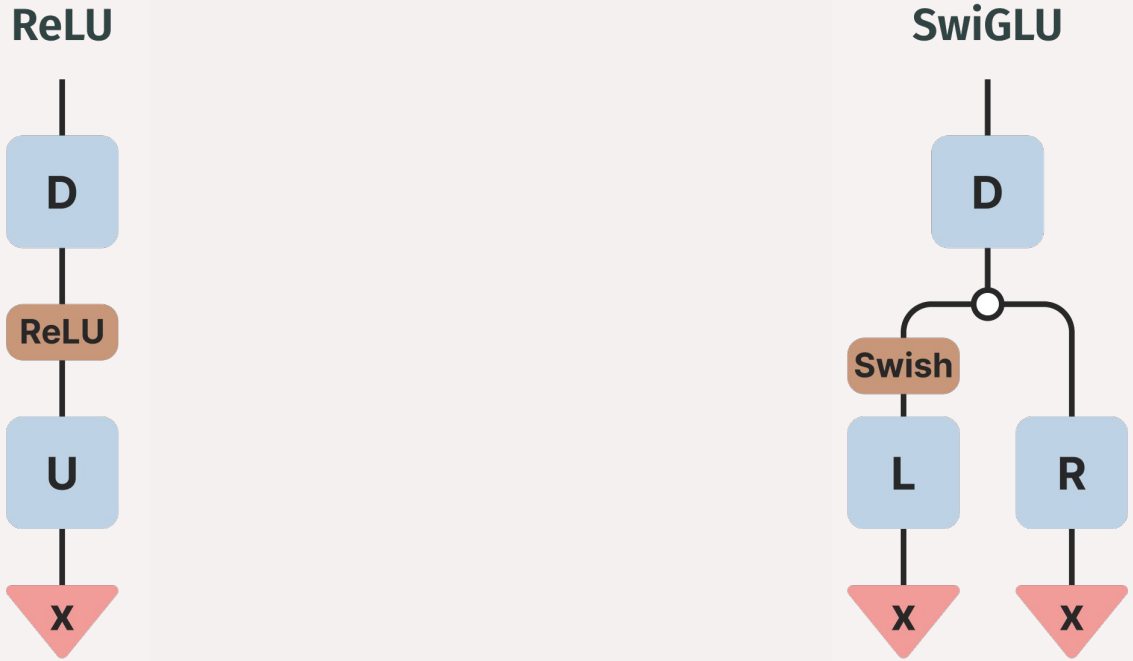
A curious table from a curious paper.

Training Steps	65,536	524,288
$\text{FFN}_{\text{ReLU}}(\textit{baseline})$	1.997 (0.005)	1.677
FFN_{GELU}	1.983 (0.005)	1.679
$\text{FFN}_{\text{Swish}}$	1.994 (0.003)	1.683
FFN_{GLU}	1.982 (0.006)	1.663
$\text{FFN}_{\text{Bilinear}}$	1.960 (0.005)	1.648
$\text{FFN}_{\text{GEGLU}}$	**1.942** (0.004)	**1.633**
$\text{FFN}_{\text{SwiGLU}}$	**1.944** (0.010)	**1.636**
$\text{FFN}_{\text{ReGLU}}$	1.953 (0.003)	1.645

"We offer no explanation as to why these architectures seem to work; we attribute their success, as all else, to divine benevolence."

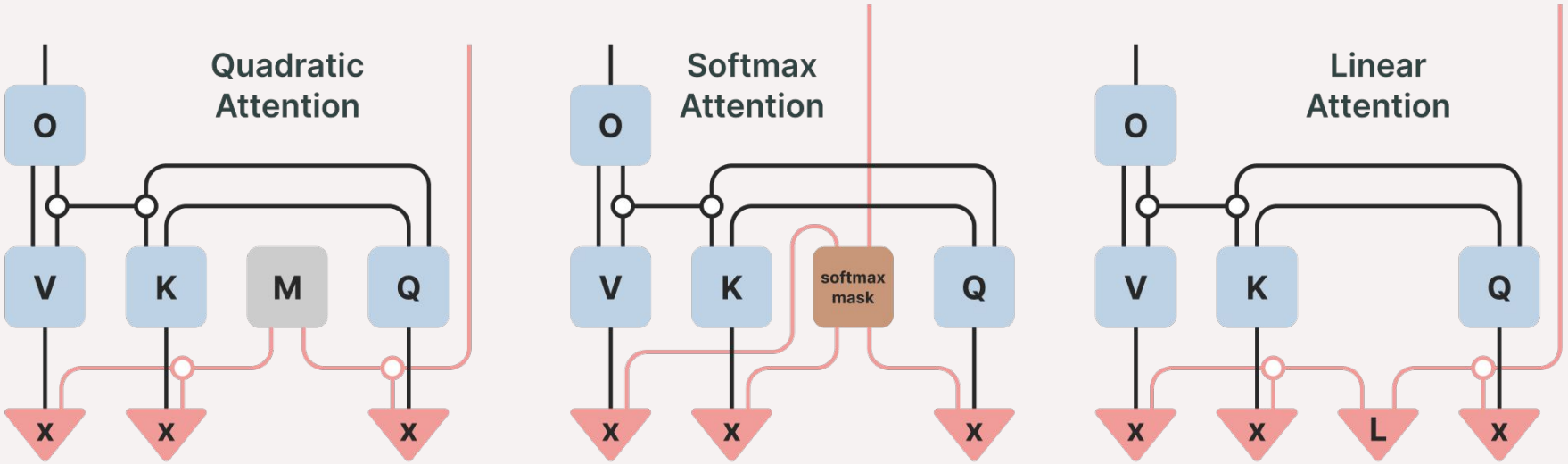
(Shazeer; 2020)

A bilinear layers is a tensor, which can be interpreted compositionally.



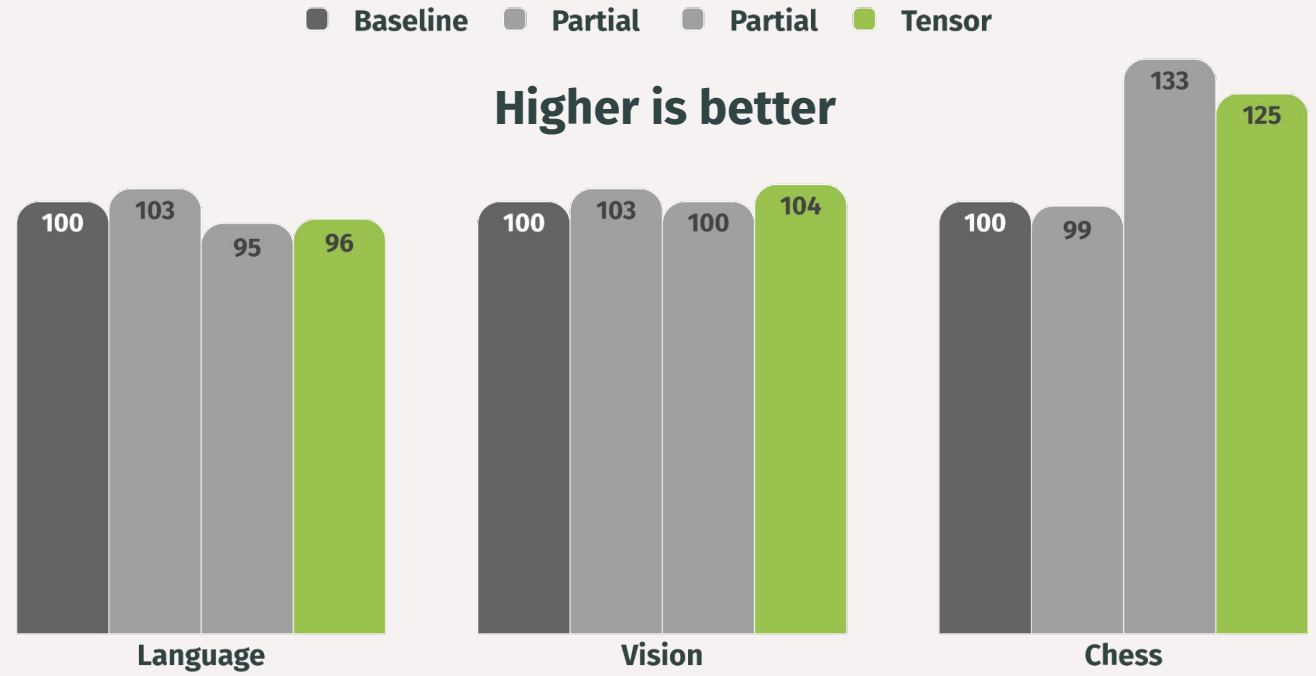
(Chrysos et al. 2024; Dooms, Gauderis et al. 2024; Hamreras et al. 2025; Gauderis, Dooms et al. forthcoming)

Attention can also be easily converted.
We can use a similar trick by making it quadratic.



(Chrysos et al. 2024; Dooms, Gauderis et al. 2024; Hamreras et al. 2025; Gauderis, Dooms et al. forthcoming)

Tensor models are performant across tasks. Training to the same loss is equally fast.



40M parameter transformers

(Shazeer et al. 2020; Pearce et al. 2024; Gauderis, Dooms et al. forthcoming)

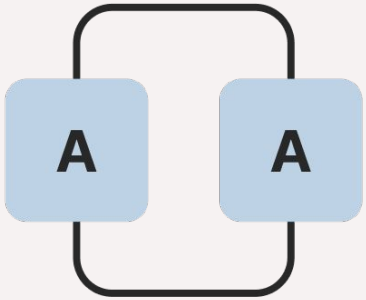
Coming up...

1. How do we operationalise compositional interpretability?
2. **What can we do with these models?**
3. How are these used in practice?

Many operations on matrices are tensor networks.

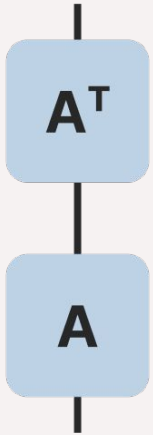
Frobenius Norm

$$\|A\|_F^2 = \sum_{ij} A_{ij} A_{ij}$$



Gram Matrix

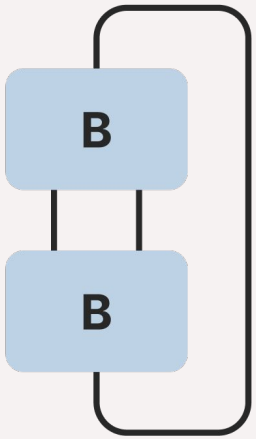
$$(A^\top A)_{jk} = \sum_i A_{ij} Aik$$



These operations intuitively generalise to tensors. This enables efficient computation of metrics.

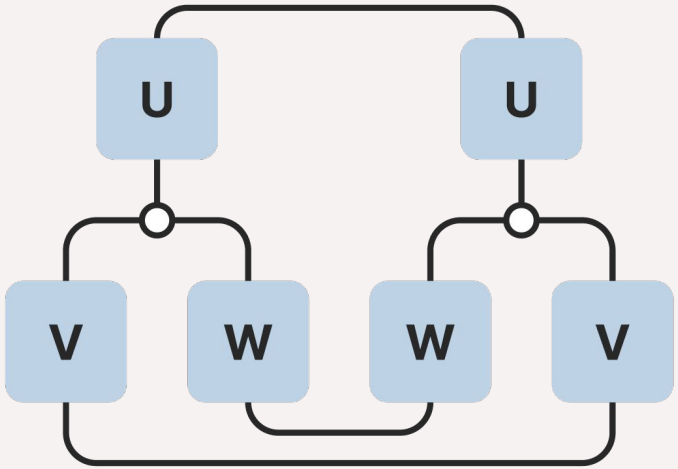
3-Tensor Norm

$$\|B\|_F^2 = \sum_{ijk} B_{ijk} B_{ijk}$$

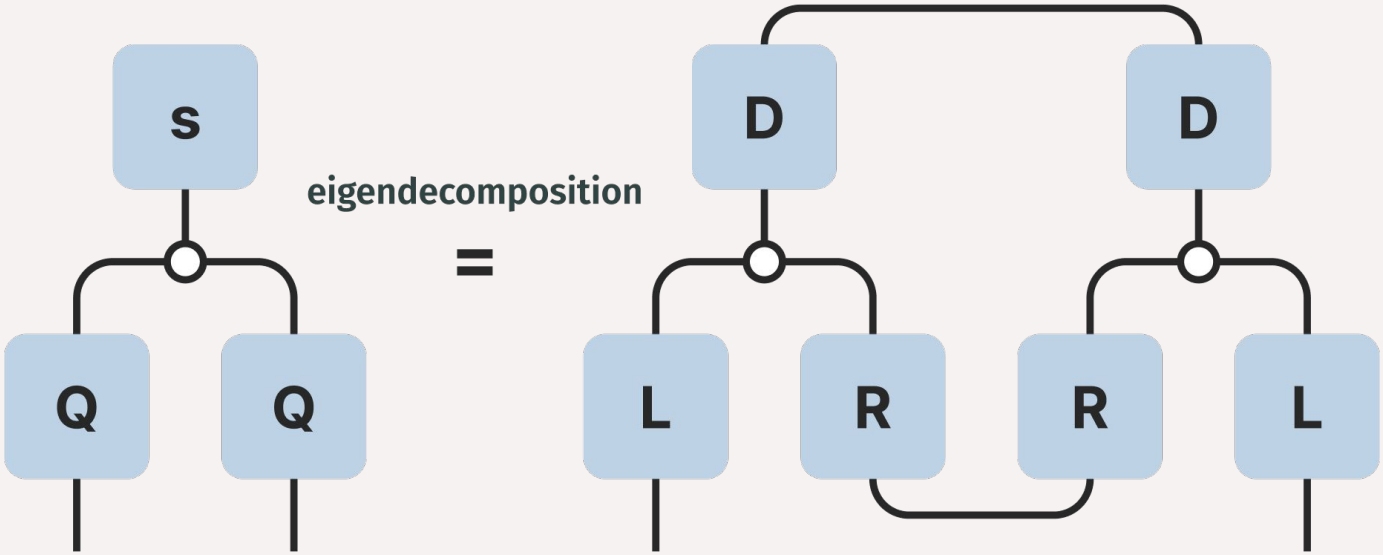


3-Tensor Norm

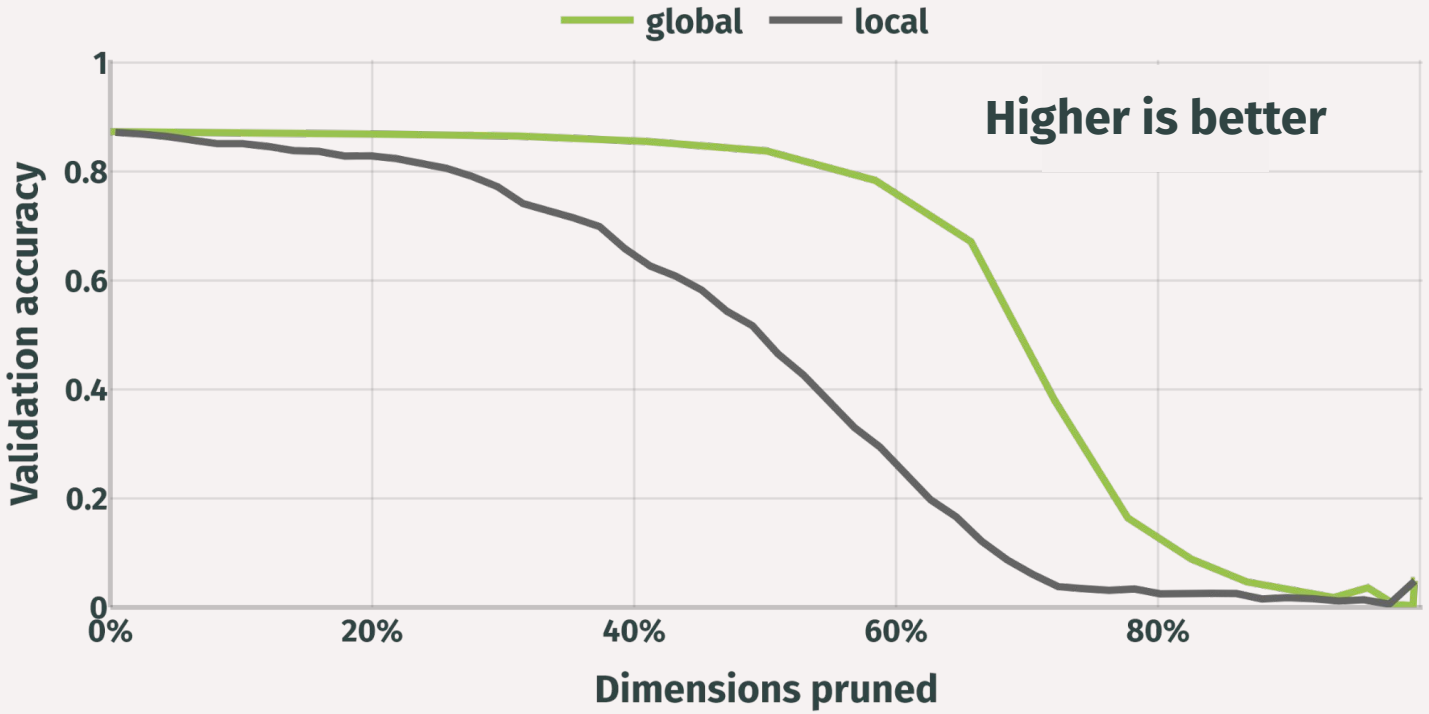
$$\| \, . \, \|_F^2 = \sum_{ijknm} U_{in} V_{jn} W_{kn} U_{im} V_{jm} W_{km}$$



This enables decomposing models analytically while keeping into account all present structure.

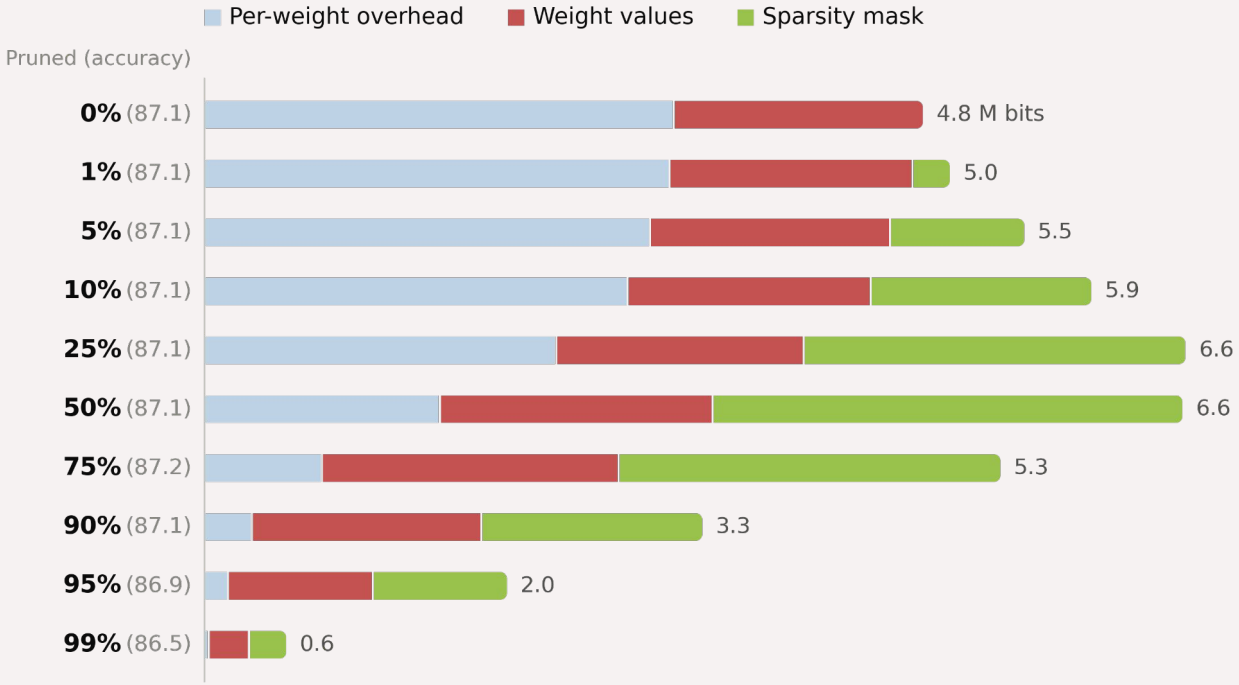


Global decompositions outperforms local versions. Models retain accuracy under strong compression.



(Grelier et al. 2018; Dooms, Gauderis et al. 2024; Gauderis, Dooms et al. forthcoming)

We can use weight sparsity to prune networks. This allows even lower description length.

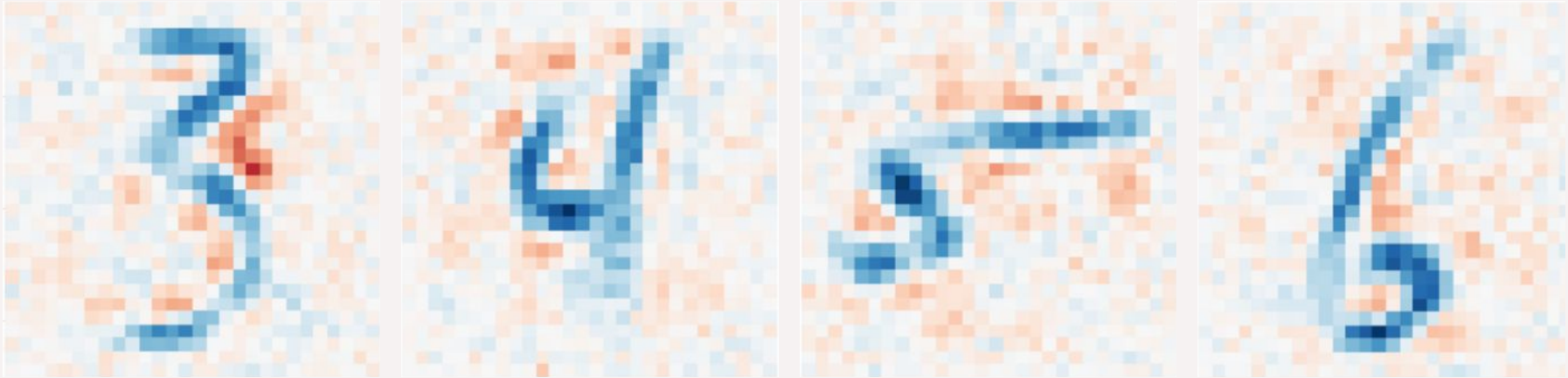


(Frankle et al. 2018; Gao et al. 2025; internal experiments)

Coming up...

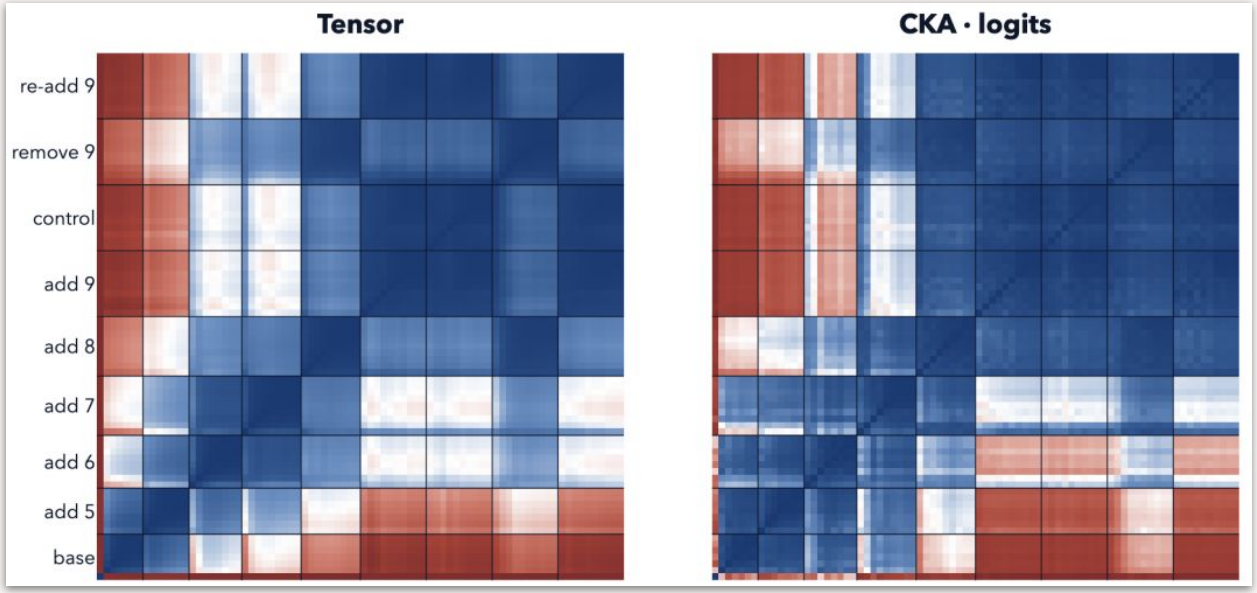
1. How do we operationalise compositional interpretability?
2. What can we do with these models?
3. **How are these used in practice?**

Rank decompositions also reveal what models learn, purely from their weights.



(Gauderis, Dooms et al. forthcoming)

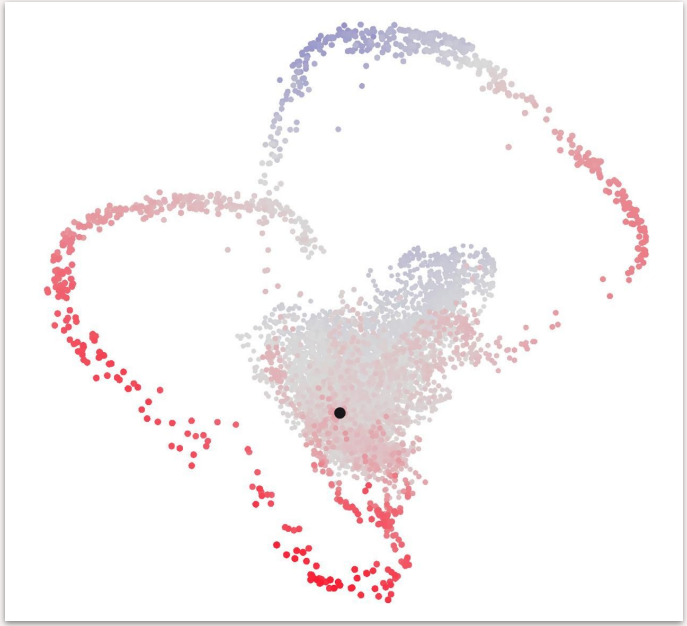
Tensors are not only useful for decomposing, but can also measure similarity across training.



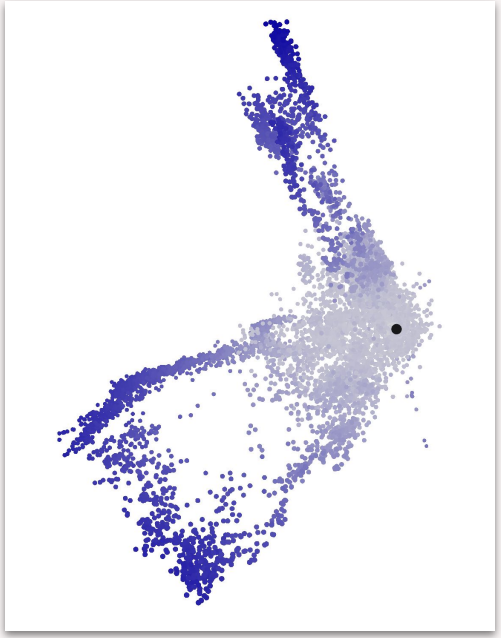
(Nissen et al.; When are two networks the same? tensor similarity for mechanistic interpretability. 2026)

Interpretability aims to find the geometry of concepts inside neural models.

Sparse composition

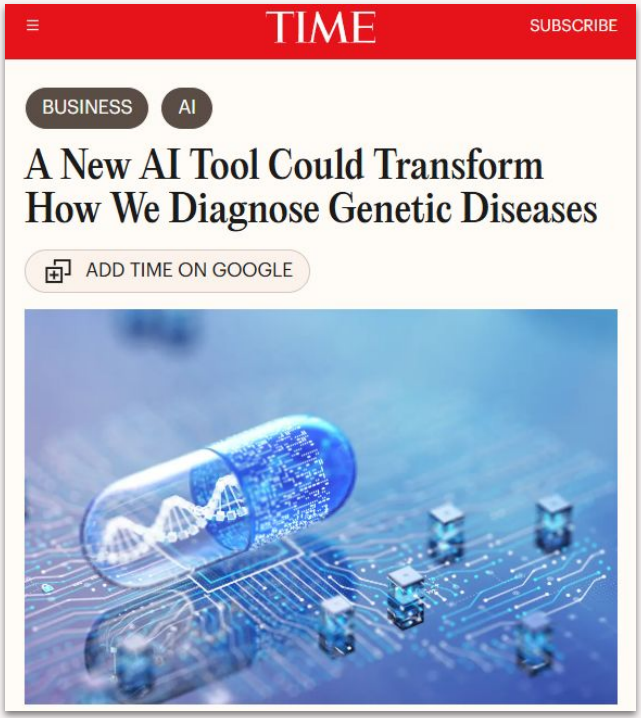


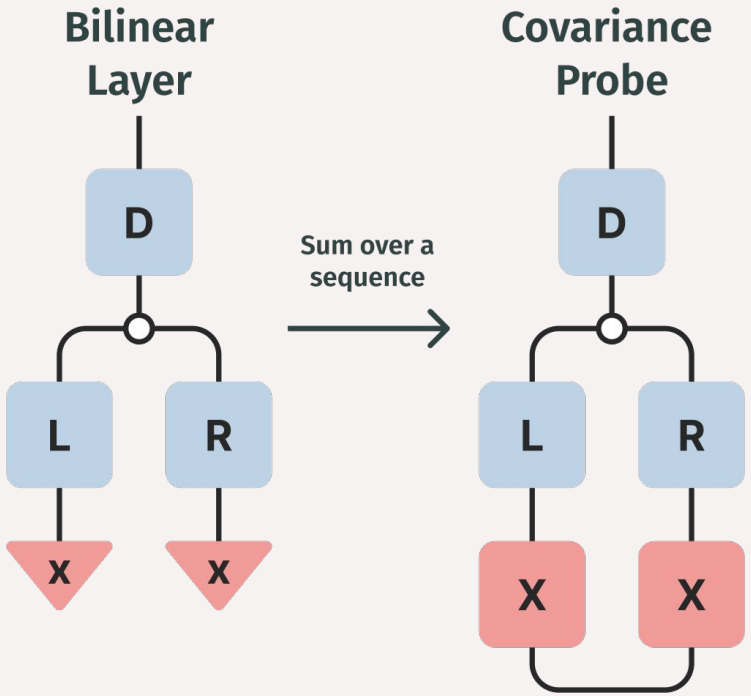
Sparse activation



(Internal experiments)

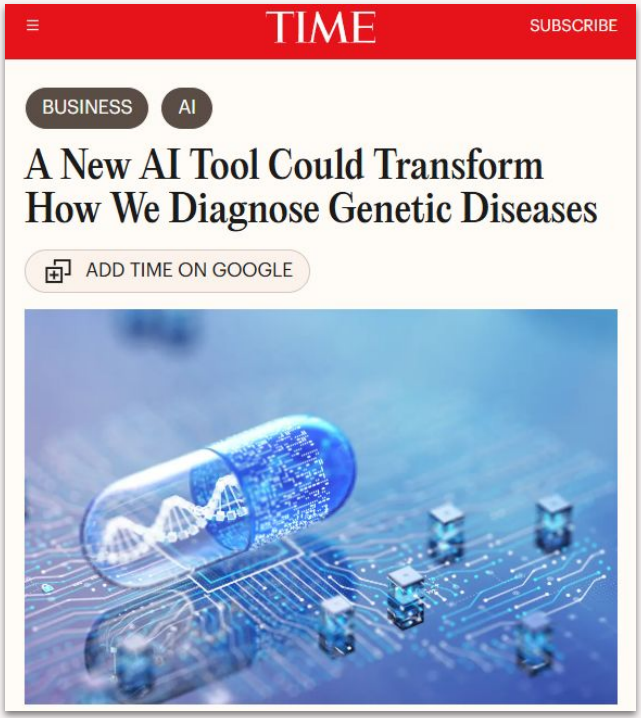
These tools are useful outside of language models, they can become drivers of scientific discovery.

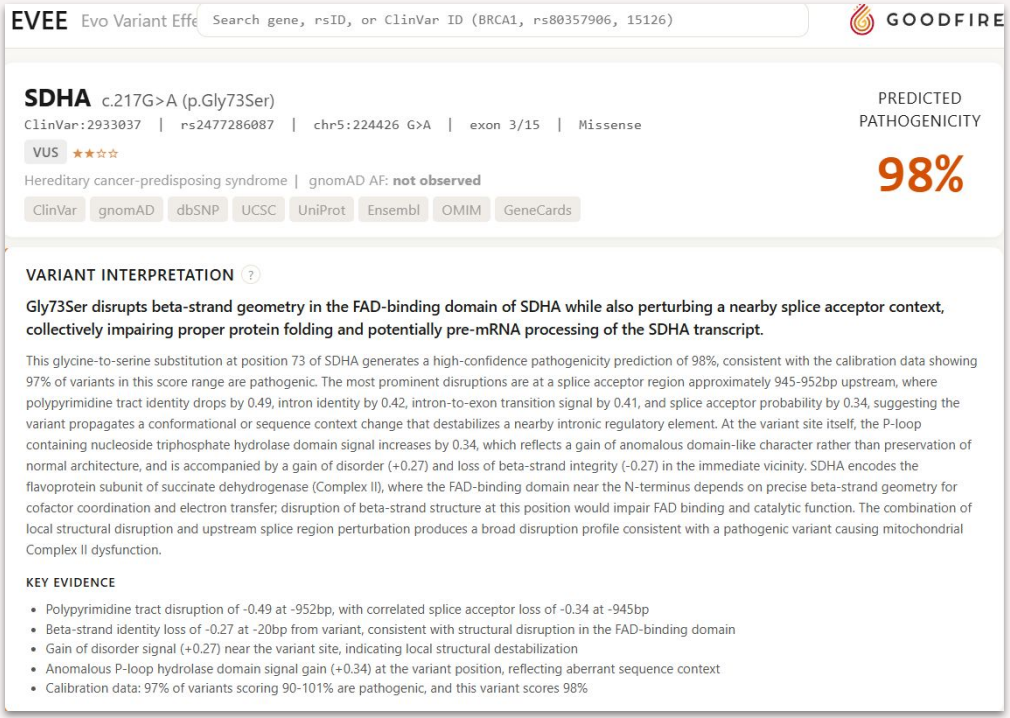




(Dooms et al. 2026; Pearce et al. 2026)

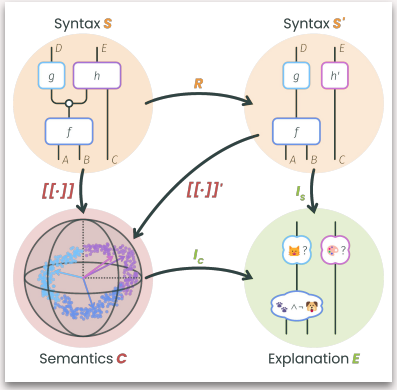
These tools are useful outside of language models, they can become drivers of scientific discovery.





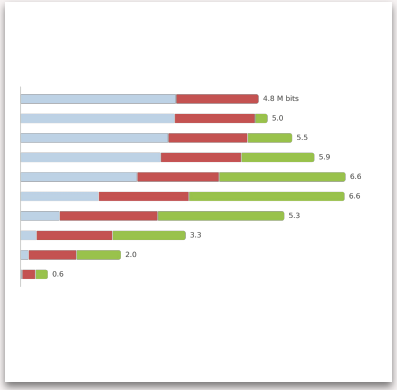
(Dooms et al. 2026; Pearce et al. 2026)

Interpretability is a constrained optimisation problem. Optimize simplicity under compositional constraints.



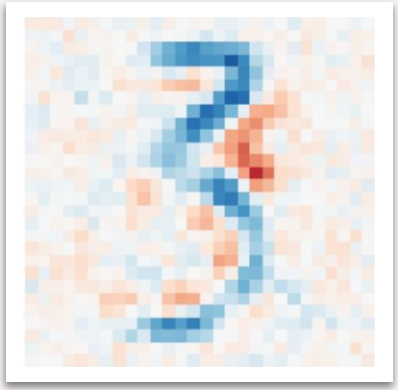
Composition

+



Compression

=



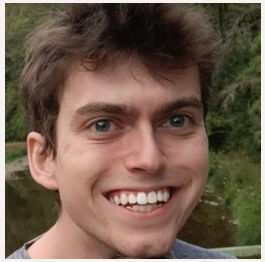
Interpretability*

*Left as exercise to the reader.

Want to know more or reach out?

Ward Gauderis

Geraint Wiggins



Thomas Dooms

Jose Oramas

VUB ARTIFICIAL INTELLIGENCE RESEARCH GROUP

Universiteit Antwerpen