Universality up to Causal Abstraction: Aligning SAE Feature Bases to Build a Cross-Model Motif Library

Liz Lemma Future Detective

January 18, 2026

Abstract

Mechanistic interpretability has produced compelling circuits (induction, retrieval, factual recall) but struggles to reuse findings across models because neurons and attention heads are not aligned and computations may self-repair. Building on the feature/circuit/motif framing and causal intervention tooling surveyed in mechanistic interpretability work, we formalize *universality up to abstraction*: two models share a motif not when they share the same neurons, but when there exists a causal abstraction mapping from a canonical motif into each model such that interventional effects match. We operationalize this notion by (i) learning sparse, overcomplete feature coordinates with sparse autoencoders/transcoders, (ii) aligning these feature bases across models using interventional effect signatures, and (iii) validating motif equivalence via causal abstraction/causal scrubbing style tests. We give provable guarantees in a simplified generative setting (sparse latent features mixed into model activations with bounded noise) showing alignment recovery and motif validation are possible with $\tilde{O}(\log m)$ samples per intervention and $\tilde{O}(\log m)$ interventions, and we provide matching lower bounds showing purely observational alignment is impossible under rotations. Empirically (implementation work strengthening the claim), we propose to instantiate the pipeline on compiled-transformer testbeds (ground-truth circuits) and then scale to families of open transformer checkpoints, releasing a motif library and evaluation harness that amortizes interpretability across 2026-era model churn.

Table of Contents

1. 1. Introduction and motivation: why neuron-level universality fails; why 2026 model churn (fine-tuning, RLHF, tool-use) demands universality at the right equivalence class; summary of contributions and what is proven vs empirically validated.

2. 2. Background and related work: features, superposition, SAEs/transcoders; circuits/motifs; activation/path/subspace patching; causal abstraction and causal scrubbing; prior universality claims (weak vs strong) and their limitations.

3. 3. Formal setup: models as causal graphs; intervention families; readouts Y; definition of motifs as bounded-size SCMs; definition of ε-causal equivalence; definition of universality up to abstraction; threat model for basis non-identifiability.

4. 4. Problem formulation: Cross-Model Motif Library (CMML) task; alignment subproblem (feature matching across models); motif equivalence testing subproblem (causal abstraction check); library selection objective (coverage vs size).

5. 5. Algorithm: (A) train SAE feature dictionaries per model and choose layers/streams; (B) compute interventional effect fingerprints for features; (C) align feature bases across models via assignment over fingerprints; (D) discover candidate motifs per behavior via automated circuit discovery; (E) canonicalize motifs and validate equivalence via causal abstraction tests; (F) select a compact motif library with guarantees (approximation).

6. 6. Theory I (identifiability and alignment): generative assumptions; recovery guarantees for feature alignment using interventional fingerprints; observational impossibility results and matching lower bounds; robustness to noise and feature splitting/merging.

7. 7. Theory II (motif equivalence testing): fixed-k tractability; sample/intervention complexity to accept/reject motif equivalence up to ε; hardness of the unrestricted problem; tightness discussion.

8. 8. Library objective and approximation: formulate motif library construction as set cover / minimum description; greedy selection approximation guarantee; implications for scaling across many models/behaviors.

9. 9. Experimental plan (strengthening contribution): (i) compiled transformers with known motifs; (ii) model families across seeds/sizes; (iii) base vs instruction-tuned vs RLHF/DPO; metrics: alignment accuracy, motif transfer accuracy, OOD robustness, falsification tests; ablations on fingerprint design and intervention family.

10. 10. Released artifacts: motif library schema (canonical SCM + abstraction maps + validation reports), tooling for adding new models, benchmark suite for universality up to abstraction; recommended reporting standards.

11. 11. Limitations and failure modes: non-linear/manifold features; hydra/self-repair; off-distribution interventions; partial observability of internal state; compute costs; dual-use considerations.

12. 12. Conclusion and future work: extending to multimodal and agentic tool-using systems; dynamic motifs over trajectories; adversarially robust universality testing.

1 Introduction and motivation

A recurring aspiration in mechanistic interpretability is a *universality* claim: that the internal organization responsible for a behavior is, in some stable sense, the same across models trained on similar data. The most literal form of this claim—*neuron-level* universality—is, however, poorly posed for modern transformers. Individual neurons (or MLP units, attention heads, etc.) are not identifiable objects under natural symmetries of the parameterization, and even when we fix a specific implementation, the representational degrees of freedom induced by superposition permit many inequivalent internal decompositions that realize essentially identical input–output behavior. Consequently, a claim of the form "neuron u in M_i corresponds to neuron v in M_j" is not invariant under benign reparameterizations, and tends to be fragile under retraining, fine-tuning, or modest architectural changes.

We make this critique precise at the level needed for algorithmic and statistical analysis. Even restricting to linear submodules, if an activation vector $a \in \mathbb{R}^d$ is expressed as a mixture $a = As$ of latent coordinates $s \in \mathbb{R}^m$, then for any invertible transform R on the latent space we obtain an equally valid factorization $a = (AR^{-1})(Rs)$. Observationally, and often even under limited probing, these rotated (or otherwise transformed) descriptions are indistinguishable. Thus, without additional structure, "the" features are not identifiable, and any purported universality at the granularity of coordinates is at best a convention. In practice, we see this convention drift substantially across random seeds and training runs, and even more under post-training modifications.

The year-to-year "model churn" that is standard by 2026 magnifies the inadequacy of neuron-level correspondence. Production and research models undergo frequent transitions: instruction tuning, RLHF/RLAIF, safety fine-tuning, tool-use scaffolding, retrieval augmentation, multimodal adapters, and distillation. These operations may preserve broad behaviors on a benchmark family $\mathcal{B}$ while reorganizing the internal computation. Two models may agree on a task distribution $\mathcal{D}$ yet implement the behavior using different internal routes; conversely, two models may differ in output on edge cases while sharing a common mechanistic subroutine. Any notion of "universality" that is too rigid will fail to be stable under such transformations; any notion that is too permissive will be scientifically vacuous.

We therefore seek the appropriate equivalence class: one in which mechanistic claims are (i) invariant to irrelevant reparameterizations, (ii) stable under routine post-training changes, and (iii) strong enough to support falsifiable predictions under *interventions*. Our organizing proposal is that universality should be formulated *up to causal abstraction*. Concretely, we fix a readout variable Y capturing the behavior of interest (e.g. a logit, a tool-call decision, a refusal indicator), and we demand agreement of *interventional effects* on Y rather than agreement of internal coordinates. In this framing, a mechanistic description is acceptable if it supports the same counterfactual predictions under an explicit intervention family $\mathcal{I}$, over both in-distribution inputs $\mathcal{D}$ and a designated out-of-distribution family $\mathcal{D}_{\text{OOD}}$.

This move resolves two distinct failure modes of naive universality. First, it avoids the coordinate-choice problem: if two internal descriptions are related by a benign change of basis that preserves all interventional responses relevant to Y, then they are equivalent for our purposes. Second, it prevents purely correlational "universality" from passing as mechanistic. Many internal units correlate with a behavior while being causally downstream or epiphenomenal. By requiring that candidate correspondences preserve the effect of *doing* something to the internal state (within $\mathcal{I}$), we force a causal commitment. Put differently, our basic object is not an activation but a causal relation between manipulable internal variables and the readout.

The immediate methodological implication is that alignment across models should be done using *interventional signatures* rather than raw activations. In particular, we treat learned sparse feature bases (e.g. via SAEs/transcoders) as a means of producing candidate internal variables that are more nearly identifiable than neurons, while recognizing that even these features are only defined up to permutation and sign (and, empirically, occasional splitting/merging). We then characterize a feature not by its mean activation, but by its *effect fingerprint*: the vector of changes in Y induced by interventions targeting that feature, aggregated over inputs. This yields a basis-invariant matching signal: a feature correspondence is acceptable only insofar as it preserves the causal influence profile relevant to Y.

On top of feature alignment, we introduce a second level of organization: *motifs* (bounded-size causal subgraphs) that capture reusable computational patterns such as induction, retrieval, copy-suppression, or safety refusal gating. The aspiration is not merely to say that isolated features recur, but that small causal structures recur, possibly with model-specific instantiations. Formally, we model a motif type τ by a canonical structural causal model SCM_τ with a fixed (small) number of variables, and we ask whether each model M_i admits an abstraction map that embeds SCM_τ into its internal causal graph in a way that preserves interventional behavior up to tolerance.

Our contributions are accordingly organized into three parts.

- **Algorithmic pipeline.** We specify an explicit procedure (CMML) that (a) learns sparse feature bases per model, (b) computes interventional fingerprints for a selected feature subset, (c) aligns features across models via assignment on fingerprint distances, (d) extracts bounded-size candidate motifs within each model, (e) clusters motifs across models into canonical types, and (f) validates each proposed abstraction by an interchange-style causal abstraction test using interventions from $\mathcal{I}$.

- **Provable guarantees in a simplified setting.** Under a shared latent-feature generative model with sparsity and incoherence conditions, and assuming that permitted interventions approximate do-operations on latent coordinates, we prove sample/intervention complexity bounds for recovering correct feature correspondences from fingerprints. We also prove that, for fixed motif size bound k, deciding ε-equivalence of candidate motifs up to abstraction is tractable (in the sense of polynomial dependence on ambient dimension with an exponential dependence only on k).
- **Lower bounds and computational barriers.** We show that without interventions, alignment is information-theoretically impossible in general due to rotational symmetries of the latent space. Separately, we show that removing the bounded-size restriction yields NP-hardness for motif matching and minimal library construction, with the expected logarithmic approximation barrier via Set Cover.

It is essential to separate what we *prove* from what we *propose to validate empirically.* The theorems in this work are conditional: they assume (i) a simplified generative model in which sparse latent features exist and are sufficiently separated in their causal effects, and (ii) an intervention family that can approximate feature-level do-operations while keeping activations approximately on-distribution. These assumptions are not tautological, and in practice must be checked by diagnostics (e.g. reconstruction quality and sparsity for the feature basis; stability of effects under resampling ablations; sensitivity of results to the intervention choice). Likewise, the existence of a compact motif library $\mathcal{L}$ that provides broad coverage is an empirical question; our greedy selection guarantee concerns approximation quality *given* a candidate set of validated motifs, not the completeness of that set.

Finally, we emphasize the scope of our universality claim. We do not claim that *all* internal mechanisms are shared across all models, nor that a single canonical representation exists. Rather, we claim that for practically relevant behavior families $\mathcal{B}$, there exist *small* causal motifs whose interventional semantics recur across models, and that these motifs can be identified and certified with explicit error tolerance ε and failure probability δ, provided we restrict attention to bounded-size structures and intervention-accessible internal variables. This is the strongest notion of universality that remains both stable under model churn and falsifiable by causal testing.

2 Background and related work

We summarize the prior technical ingredients that our formalism relies on: (i) feature discovery under superposition, with an emphasis on sparse autoencoders (SAEs) and related transcoders; (ii) circuit- and motif-level analyses in transformers; (iii) intervention methods used to ascribe causal responsibility to internal components; and (iv) causal abstraction frameworks that aim to make mechanistic claims invariant to representational choices. Our emphasis is on what is known to be identifiable, what is not, and which aspects of the literature implicitly depend on intervention access.

2.1 Features, superposition, and sparse dictionary learning in networks

A standard starting point is the observation that internal activations often appear to encode many "features" (directions corresponding to semantically or algorithmically coherent properties), but these features are typically *not* realized as single neurons. Rather, features may be *distributed* across many coordinates and multiple features may be *superposed* within the same neuron. In the simplest linearized picture, one writes an activation stream $a \in \mathbb{R}^d$ (e.g. residual stream at a layer, or MLP pre-activations) as a mixture

$$a \approx Dh,$$

where $D \in \mathbb{R}^{d\times m}$ is a dictionary and $h \in \mathbb{R}^m$ is a sparse code. This is precisely the sparse coding / dictionary learning model, in which identifiability (up to permutation and sign, and sometimes scaling) can be obtained under sparsity and incoherence assumptions. The relevance for mechanistic interpretability is that, when such a model is a reasonable approximation, the "right" internal variables to talk about are not individual coordinates of a but the sparse coordinates of h, which better track stable latent causes.

Sparse autoencoders (SAEs) instantiate this idea by learning an encoder–decoder pair (E, D) so that $h = E(a)$ is sparse and Dh reconstructs a with small error, typically by minimizing a reconstruction loss plus a sparsity penalty (e.g. ℓ_1 or variants) **??**. Transcoders and related architectures aim to map between activation streams (e.g. from one layer or module to another) while maintaining a sparse representation **?**. In either case, the learned features are intended to provide a *feature basis* F_i for model M_i, with the practical benefit that interventions can be applied at the level of feature activations (ablate, clamp, resample, or patch) rather than at the level of neurons.

Two limitations are central for our purposes. First, even when sparse coding assumptions are approximately satisfied, identifiability is only guaranteed under additional conditions; empirically, SAEs can exhibit splitting/merging of features across training runs, dead features, and dependence on hyperparameters and training data. Second, in the absence of strong assumptions, the learned features are not canonical objects: different dictionaries may support comparable reconstruction quality while giving different decompositions, and some of these differences may be behaviorally irrelevant. Our subsequent emphasis on *interventional fingerprints* is motivated by the need to identify feature correspondences by their causal role with respect to a readout, rather than by reconstruction or correlation alone.

2.2 Circuits, motifs, and bounded mechanistic structure

The circuits tradition in transformer interpretability studies how specific behaviors are realized by small sets of components (attention heads, MLP neurons, residual streams) and their interaction patterns; canonical examples include induction-like behavior in attention, copy mechanisms, and task-specific subroutines discovered in toy settings **??**. The operational content of a circuit claim is usually that: (i) ablating or perturbing a proposed set of components degrades a target behavior, and (ii) the information flow between those components can be demonstrated by targeted patching or attribution methods.

Our work uses the term *motif* to denote a bounded-size causal pattern that recurs across instances (models, seeds, or post-training variants). This is a generalization of a circuit: we abstract from the precise implementation details to a small directed structure with designated variables and edges, together with a family of interventions that test its functional role. The bounded-size requirement is not merely methodological; it is what permits tractable validation procedures and meaningful statistical guarantees. In practice, circuit discovery is typically performed with a mixture of manual hypotheses and automated tools (e.g. procedures that search over paths, edges, or components that most influence a measured output). However, without a bounded-size bias or a strong prior, the search space is combinatorial, and many candidate explanations can be consistent with the same observational behavior.

2.3 Intervention methods: activation, path, and subspace patching

A large fraction of mechanistic interpretability relies on *interventions* applied to internal activations. The simplest form is *activation patching*: one runs a “clean” and a “corrupted” input through the model, caches activations from one run, and then substitutes them into the other run at selected sites, observing changes in a chosen readout (often logits) **?**. Variants include patching individual layers, attention head outputs, MLP activations, or residual stream vectors. Patching provides evidence of causal mediation: if substituting a component restores performance, that component contains information relevant to the behavior under the specified input manipulation.

A refinement is *path patching* or *edge ablation*, where one intervenes not on an entire node but on a specific computational pathway (for instance, the contribution of one attention head to another module) **?**. This aims to isolate the mediating route and reduce confounding from other sources of information. Relatedly, *subspace patching* targets a linear subspace (often defined by a probe direction or by a feature dictionary) rather than the full activation vector, allowing finer-grained statements such as "the behavior depends on this subspace of the residual stream" rather than on the entire stream.

These methods naturally suggest the intervention family $\mathcal{I}$ we will formalize later. Two practical considerations are worth isolating. First, naive ablations can drive activations off-distribution, producing artifacts; resample ablations and related techniques attempt to keep interventions within the typical activation manifold by replacing components with draws from a reference distribution. Second, patching-based evidence is inherently relative to the chosen input manipulations (clean/corrupt pairs) and to the chosen readout; consequently, it is best understood as producing a *conditional* causal claim, which motivates our explicit bookkeeping of Y, $\mathcal{D}$, and $\mathcal{D}_{\text{OOD}}$.

2.4 Causal abstraction, causal scrubbing, and invariance of mechanistic claims

The causal abstraction program aims to relate a high-level causal model (capturing an algorithmic description) to a low-level causal model (capturing mechanistic implementation) via an abstraction map that preserves interventional semantics **??**. In mechanistic interpretability, this perspective is operationalized through *interchange interventions* and *causal scrubbing*: one defines a hypothesized decomposition into variables, specifies which interventions should commute with the abstraction, and then tests whether substituting internal states according to the hypothesized map preserves the relevant outputs **?**. The key point is that abstraction is not merely a compression of observables; it is a constraint on *counterfactual behavior* under interventions.

Our use of causal abstraction differs in emphasis rather than in kind. We are not primarily concerned with relating a neural network to an external symbolic program, but with relating multiple networks to a *shared canonical motif model.* This shifts the burden from explaining a behavior in one model to establishing a cross-model equivalence class of mechanisms, with explicit tolerance. In particular, we require that the abstraction be validated not only on a single distribution but also on a designated OOD family, to reduce the risk that a putative motif match is an artifact of a narrow benchmark.

2.5 Universality claims: weak notions, strong notions, and failure modes

A variety of "universality" results appear in the literature, ranging from empirical observations that certain attention heads or directions recur across seeds, to claims that specific motifs (e.g. induction-like heads) are common in language models trained on next-token prediction. These results often come in at least two forms. A *weak* universality claim asserts that some recognizable pattern exists in many models (e.g. a head with a particular attention pattern, or a direction linearly decodable as a feature), typically established by correlational or probe-based evidence and occasional ablation checks. A *strong* universality claim asserts a stable correspondence of internal variables across models, sometimes suggesting a near one-to-one mapping between units or directions.

Both forms face characteristic limitations. Weak claims can be true but scientifically underspecified: without explicit intervention semantics, it may be unclear whether the recurring pattern is causally responsible for the behavior, merely correlated with it, or downstream of the true causal factors. Strong claims, when formulated at the level of neurons or arbitrary coordinates, conflict with the non-identifiability induced by reparameterizations and superposition; when formulated at the level of learned features, they inherit the non-canonicity of the learned basis and can fail under feature splitting/merging or under post-training changes.

These limitations motivate two design choices in our framework. First, we treat correspondences as hypotheses to be tested by *interventional effect profiles* with respect to a readout Y, rather than by activation similarity. Second, we elevate the unit of universality from isolated features to bounded-size motifs whose semantics are defined by a canonical causal model and validated by intervention. The next section introduces the formal machinery needed to state these claims precisely.

3 Formal setup

We now fix notation and state the objects that will be manipulated throughout. Our goal is to phrase cross-model mechanistic claims in a way that is (i) explicit about intervention semantics, (ii) invariant to superficial reparameterizations, and (iii) compatible with bounded-size validation procedures.

3.1 Models as structural causal graphs

For each $i \in \{1, \ldots, r\}$, let M_i denote a transformer with parameters θ_i. We view M_i as a (possibly stochastic) mapping from an input sequence x to a readout Y of interest. The readout Y may be a scalar (e.g. a refusal indicator), a discrete action (e.g. tool selection), or a vector of logits $Y \in \mathbb{R}^{|V|}$ for a vocabulary V. Randomness may enter through explicit sampling, dropout-like noise (if present), or through resampling-based interventions.

We associate to M_i an *internal causal graph* G_i. Concretely, fix a finite set of internal variables $V(G_i)$ corresponding to activation vectors at chosen sites (e.g. residual stream at each layer and position, attention head outputs, MLP activations). Directed edges correspond to direct functional dependence in the forward computation. For a fixed input x, the forward pass induces assignments

$$v = f_v\big(\mathrm{Pa}(v), x; \theta_i\big), \qquad v \in V(G_i),$$

where $\mathrm{Pa}(v)$ denotes the parent variables of v in G_i. Together with a readout function g_i producing Y from a subset of internal nodes and x, this yields a structural causal model (SCM) in the standard sense: internal nodes are endogenous variables, x (and any explicit noise variables) are exogenous, and the equations are induced by the computation graph. We emphasize that G_i is not intended to be uniquely defined; rather, it is a bookkeeping device for intervention targets, and it may be coarsened or refined depending on the chosen granularity of analysis.

3.2 Inputs, distributions, and readouts

We fix a *behavior family* $\mathcal{B}$, which determines a collection of tasks, templates, or benchmarks that instantiate inputs. We write $x \sim \mathcal{D}$ for in-distribution inputs and $x \sim \mathcal{D}_{\mathrm{OOD}}$ for a specified out-of-distribution family. The union $\mathcal{D} \cup \mathcal{D}_{\mathrm{OOD}}$ is the domain on which we will demand agreement of mechanistic claims.

A *readout* Y is any measurable function of the model execution that we intend to explain or constrain. When Y is vector-valued (logits), we will compare models via a chosen discrepancy functional, e.g. expected logit differences on designated coordinates or total variation distance between induced next-token distributions. The definitions below are stated for general Y; the choice of discrepancy will be made explicit when needed.

3.3 Intervention families

Let $\mathcal{I}$ be a family of allowed interventions. An intervention $I \in \mathcal{I}$ specifies (i) a target set of internal variables in $V(G_i)$ and (ii) a rule for modifying those variables during the forward pass. We write M_i^I for the intervened model. Typical examples include:

1. *Node ablation*: replace an activation vector v by 0 or by its mean.

2. *Resample ablation*: replace v by a sample from a reference distribution conditioned on layer/position, intended to keep activations on-manifold.

3. *Subspace interventions*: decompose $v = v_{\parallel} + v_{\perp}$ relative to a chosen subspace, then ablate/patch only $v_{\parallel}$.

4. *Patching*: substitute cached activations from a counterfactual run.

5. *Path/edge interventions*: modify only a designated contribution along an edge (e.g. one head's contribution to the residual stream).

We will require interventions to be *well-posed* (they define a measurable mapping from (x, θ_i) and internal randomness to outputs) and, when claims depend on approximate equivalence, to be chosen so that the intervened activations remain approximately on-distribution (resample/patching variants are intended to satisfy this).

3.4 Feature bases and the threat model of non-identifiability

To express motif variables in a basis that is more stable than raw neurons, we assume that for each M_i we learn a feature basis F_i via an SAE or transcoder. Formally, for a chosen activation stream $a \in \mathbb{R}^d$ we obtain a dictionary $D_i \in \mathbb{R}^{d \times m}$ and an encoder E_i producing sparse codes $h_i = E_i(a) \in \mathbb{R}^m$ with $a \approx D_i h_i$. The coordinates $(h_i)_j$ are our candidate *features.*

Our threat model is that, absent additional assumptions, such a basis is not canonical. Even when reconstruction is good, multiple dictionaries may exist that induce comparable reconstruction error but differ by permutations, sign flips, scalings, or more general rotations within subspaces of correlated features. Moreover, feature splitting/merging can occur across SAE training runs or across models with different training histories. Consequently, any definition of "the same feature across models" that relies only on geometric similarity (e.g. cosine similarity between dictionary vectors) is vulnerable to non-identifiability. The purpose of our subsequent formalism is to tie feature identity to *interventional role* with respect to a readout Y, thereby breaking symmetries that are invisible observationally.

3.5 Motifs as bounded-size SCMs

A *motif type* τ is specified by a canonical SCM, denoted SCM_τ, with endogenous variables $Z_1, \dots, Z_k$ for some k bounded by a fixed constant. The SCM includes:

- a directed graph on $\{Z_1, \dots, Z_k\}$ specifying parent sets $\mathrm{Pa}_\tau(Z_j)$;
- structural equations $Z_j = \varphi_j(\mathrm{Pa}_\tau(Z_j), U_j)$, with exogenous noise U_j (which may be degenerate);
- a designated *interface* to the readout, e.g. a function $Y = \psi_\tau(Z_1, \dots, Z_k, U_Y)$ or, more generally, a specification of which interventions on the Z_j should induce which qualitative changes in Y.

The bounded-size condition $k = O(1)$ is not a modeling convenience but a computational constraint: it is what makes it possible to enumerate and validate candidate abstractions with finite intervention budgets.

3.6 ε-causal equivalence and abstraction maps

We require a notion of equivalence that is explicitly interventional. Fix a model M_i, a motif type τ, and a candidate *abstraction map* α_i that assigns each canonical variable Z_j to a concrete internal target in M_i. In practice, α_i will often map Z_j to an SAE feature (a coordinate of h_i) at a particular layer/position, possibly together with a linear readout/subspace identifying how interventions on Z_j are implemented in the underlying activation stream.

Given α_i, each intervention $\mathrm{do}(Z_j \leftarrow z)$ in SCM_τ induces an intervention $I \in \mathcal{I}$ on M_i (e.g. clamping, ablating, or resampling the corresponding feature coordinate). Let $P^{\tau}_{\mathrm{do}(\cdot)}(Y \mid x)$ denote the interventional distribution over Y in the canonical motif SCM, and let $P^{i,\alpha_i}_{I}(Y \mid x)$ denote the interventional distribution over Y in M_i under the corresponding intervention.

Definition 3.1 (ε-causal agreement)**.** Fix a discrepancy $d(\cdot,\cdot)$ between distributions on Y (e.g. total variation, or expected ℓ_2 distance of logits). We say that (M_i, α_i) *ε-agrees* with SCM_τ on $\mathcal{D}'$ (where $\mathcal{D}' \subseteq \mathcal{D} \cup \mathcal{D}_{\mathrm{OOD}}$) if for all $x \in \mathrm{supp}(\mathcal{D}')$ and for all allowed interventions in the motif intervention set,

$$d\Big(P^{i,\alpha_i}_{I}(Y \mid x),\ P^{\tau}_{\mathrm{do}(\cdot)}(Y \mid x)\Big) \le \varepsilon.$$

Definition 3.2 (ε-causal equivalence of motifs across models)**.** Two instances (M_i, α_i) and (M_j, α_j) implementing the same motif type τ are *ε-causally equivalent* on $\mathcal{D}'$ if both ε-agree with SCM_τ on $\mathcal{D}'$ under the same canonical intervention set. Equivalently, their induced interventional distributions on Y match up to 2ε by the triangle inequality.

This definition separates the representational question (what internal variables realize Z_j?) from the semantic question (what counterfactual constraints does τ impose on Y?). In particular, α_i is not required to be an isomorphism between graphs; it is required to preserve the relevant interventional semantics up to tolerance.

3.7 Universality up to abstraction

We can now state the cross-model notion we aim to certify.

Definition 3.3 (Universality up to abstraction)**.** Fix a motif type τ, an intervention family $\mathcal{I}$, a readout Y, and distributions $\mathcal{D}, \mathcal{D}_{\mathrm{OOD}}$. We say that τ is *universal up to ε-abstraction* across the model family $\{M_1, \ldots, M_r\}$ if for each i there exist (i) a feature basis F_i (e.g. an SAE dictionary on a chosen stream) and (ii) an abstraction map α_i from SCM_τ into G_i (typically via features in F_i) such that (M_i, α_i) ε-agrees with SCM_τ on $\mathcal{D} \cup \mathcal{D}_{\mathrm{OOD}}$.

The quantification over F_i reflects the fact that different models may require different learned feature bases for the motif variables to be sparse and intervention-accessible. The dependence on $\mathcal{D}_{\text{OOD}}$ is explicit: we are not merely fitting a motif to a benchmark, but demanding that its interventional semantics persist under a designated shift family.

3.8 Interventional effect fingerprints (preview)

Although the algorithmic details appear in the next section, we record one object that will reoccur: for a feature f (an SAE coordinate) in model M_i, an intervention $I \in \mathcal{I}$, and an input x, let $\Delta_i(f; I, x)$ denote the induced change in Y when intervening on f using I at input x. Aggregating these effects over a collection of interventions yields a vector-valued summary, which we will use as a basis-invariant signature of causal role. This is the mechanism by which we will mitigate the non-identifiability described above: rotated or permuted feature bases that are observationally indistinguishable need not preserve interventional effect profiles on Y.

With these definitions in place, we can pose a concrete computational task: given multiple models, interventions, and data distributions, we seek to align features across models, cluster bounded motifs into canonical types, and select a compact library that maximizes validated coverage. This is the Cross-Model Motif Library problem formulated next.

4 Problem formulation: the Cross-Model Motif Library (CMML) task

We now formalize the computational problem implicit in the preceding definitions. The input consists of a family of models $\{M_1, \ldots, M_r\}$, a behavior family $\mathcal{B}$ that induces input distributions $\mathcal{D}$ and $\mathcal{D}_{\text{OOD}}$, a readout Y, and an allowed intervention family $\mathcal{I}$. The output we seek is a compact library of canonical motifs together with per-model abstraction maps that certify ε-causal agreement on $\mathcal{D} \cup \mathcal{D}_{\text{OOD}}$. Algorithmically, this decomposes into three interacting subproblems: (i) aligning feature bases across models, (ii) testing whether candidate motifs are ε-equivalent up to abstraction, and (iii) selecting a small set of motifs covering as many validated model–behavior instances as possible.

4.1 The CMML task

Fix a motif size budget k (treated as a constant in our fixed-k results). For each model M_i we may choose a feature basis F_i (e.g. via an SAE/transcoder on a specified activation stream) and a candidate set of features $S_i \subseteq [m]$ on which we are willing to spend intervention budget. A *candidate motif instance* in model M_i is then a bounded-size object

$$\mu = (V_\mu, E_\mu, \text{loc}_\mu),$$

where $V_\mu \subseteq S_i$ with $|V_\mu| \leq k$ is a set of feature indices, E_μ is a directed edge set encoding an hypothesized causal/functional dependency among these features (possibly annotated by layer/position), and loc_μ records where in the computation graph the features are realized. We deliberately keep this representation abstract: different circuit discovery procedures instantiate different notions of edges E_μ (e.g. attention-mediated paths versus MLP composition), but for CMML we require only that edges can be targeted (directly or indirectly) by interventions from $\mathcal{I}$ when performing validation.

A *motif type* τ is represented canonically as an SCM SCM_τ over variables $Z_1, \ldots, Z_{k_\tau}$ with $k_\tau \leq k$. An *instance* of τ in model M_i is specified by an abstraction map $\alpha_{i,\tau}$ that assigns each Z_j to a concrete intervention target in G_i (typically a feature coordinate in F_i at a designated site) together with the intervention semantics needed to implement $\mathrm{do}(Z_j \leftarrow z)$ via some $I \in \mathcal{I}$.

CMML output. The CMML problem asks us to produce:

1. a library $\mathcal{L} = \{(\tau, \mathsf{SCM}_\tau)\}$ of canonical motif types;
2. for each i and each $\tau \in \mathcal{L}$ judged present in M_i, an abstraction map $\alpha_{i,\tau}$;
3. a validation report certifying ε-causal agreement of $(M_i, \alpha_{i,\tau})$ with SCM_τ on $\mathcal{D} \cup \mathcal{D}_{\mathrm{OOD}}$ (up to failure probability δ arising from finite sampling).

We emphasize that $\mathcal{L}$ is not a taxonomy of all internal phenomena; it is a parsimonious set of causal templates whose interventional signatures are reused across models.

4.2 Alignment subproblem: matching features across models

Because feature bases F_i are only identifiable up to reparameterization, we treat feature identity across models as an *interventionally defined* notion. Concretely, for each model M_i and feature $f \in S_i$, we define an *interventional effect fingerprint* by aggregating the effect on the readout Y across a collection of interventions and inputs. Fix a list of T interventions $(I_t)_{t=1}^T \subseteq \mathcal{I}$ intended to act "locally" on the target feature (e.g. ablation, resample ablation, or subspace patching restricted to that feature). We define

$$\Phi_i(f) \;=\; \Big(\mathbb{E}_{x \sim \mathcal{D}}\big[\Delta_i(f; I_t, x)\big] \Big)_{t=1}^{T}, \tag{1}$$

optionally concatenated with an analogous vector over $x \sim \mathcal{D}_{\mathrm{OOD}}$. Here $\Delta_i(f; I, x)$ is a chosen effect functional (e.g. a logit difference on designated coordinates, or a scalar change in refusal probability) computed as the discrepancy between $M_i(x)$ and $M_i^I(x)$.

The *feature alignment problem* between models M_i and M_j is then a minimum-cost matching between subsets of S_i and S_j using a metric on fingerprints. Writing $d_\Phi(\cdot,\cdot)$ for a distance on $\mathbb{R}^T$ (e.g. ℓ_2, cosine distance, or a robust M-estimator), we seek a partial bijection π_{ij} minimizing

$$\min_{\pi\in\Pi} \sum_{f\in\mathrm{dom}(\pi)} d_\Phi\big(\Phi_i(f), \Phi_j(\pi(f))\big), \tag{2}$$

where Π ranges over matchings of prescribed cardinality (or, more generally, over matchings with dummy nodes to allow unaligned features). In settings where sign or scale flips are expected (e.g. due to dictionary conventions), we may allow d_Φ to minimize over these symmetries, or augment π_{ij} to record sign/scale parameters.

This alignment is the only place where we attempt to establish cross-model correspondences at the level of individual features. Subsequent motif canonicalization may use these correspondences to propose that two motifs drawn from different models instantiate the same motif type; the validity of that proposal is then checked interventionaly, not assumed from the alignment alone.

4.3 Motif equivalence testing as causal abstraction checking

Given a candidate motif type τ (represented by SCM_τ) and a candidate abstraction map α into M_i, the central decision problem is whether (M_i, α) ε-agrees with SCM_τ on $\mathcal{D}\cup\mathcal{D}_{\mathrm{OOD}}$ under the motif's canonical intervention set. Since the quantification "for all x" and "for all interventions" is intractable in full generality, we phrase validation in terms of a *finite* test suite derived from $\mathcal{I}$, with explicit sample size N per test and failure probability δ.

Formally, fix a finite intervention set $\mathcal{I}_\tau \subseteq \mathcal{I}$ sufficient to characterize the motif semantics of τ (e.g. single-node ablations, pairwise interchange interventions, and a small number of compositional interventions on size-k subsets). For each $I \in \mathcal{I}_\tau$ and each distribution $\mathcal{D}' \in \{\mathcal{D}, \mathcal{D}_{\mathrm{OOD}}\}$, we estimate

$$\widehat{d}_{i,\alpha}(I;\mathcal{D}') \;=\; \frac{1}{N}\sum_{n=1}^{N} d\Big(\widehat{P}_I^{i,\alpha}(Y \mid x_n), \widehat{P}_I^{\tau}(Y \mid x_n)\Big), \qquad x_n \sim \mathcal{D}'.$$

We accept (M_i, α) as an instance of τ if $\widehat{d}_{i,\alpha}(I;\mathcal{D}') \le \varepsilon$ for all $I \in \mathcal{I}_\tau$ and both $\mathcal{D}'$, with slack chosen so that concentration bounds imply a true discrepancy at most ε with probability at least $1-\delta$. Importantly, this validation step treats the model as a causal system, not merely a pattern recognizer: passing the test entails that the counterfactual constraints encoded by SCM_τ are respected (up to tolerance) when realized through α.

We also require a *well-posedness* condition on interventions used in validation: interventions should keep internal activations approximately on-distribution, so that discrepancies reflect causal mismatch rather than pathological off-manifold behavior. This motivates using resample ablations or patching-based interventions as the default elements of $\mathcal{I}_\tau$ when available.

4.4 Library selection: coverage versus parsimony

After alignment and motif validation we obtain a collection $\mathcal{C}$ of *validated motif clusters*. Each cluster $c \in \mathcal{C}$ corresponds to a proposed motif type $\tau(c)$ together with a set of validated instances across models. Let

$$U \;=\; \{(i,b) : i \in [r],\, b \in \mathcal{B} \text{ is a target behavior instance}\}$$

denote the universe of model–behavior instances we seek to explain (one may equivalently take U to be pairs (i,τ) if motif presence is defined independently of $\mathcal{B}$). Each validated cluster c covers a subset $U(c) \subseteq U$, consisting of those instances for which there exists an abstraction map $\alpha_{i,\tau(c)}$ passing the ε-agreement test.

The *library selection problem* is to choose a subcollection $\mathcal{L} \subseteq \{\tau(c) : c \in \mathcal{C}\}$ that achieves high coverage while remaining small. In its strictest form, we require full coverage of a designated target set $U_\star \subseteq U$ and minimize $|\mathcal{L}|$, yielding a set cover instance:

$$\min_{\mathcal{L}} |\mathcal{L}| \quad \text{s.t.} \quad U_\star \subseteq \bigcup_{\tau \in \mathcal{L}} U(\tau).$$

More generally, we may trade off coverage and parsimony by optimizing

$$\max_{\mathcal{L}} \; \left|\cup_{\tau \in \mathcal{L}} U(\tau)\right| \;-\; \lambda |\mathcal{L}|,$$

or by imposing a hard budget $|\mathcal{L}| \leq B$ and maximizing covered instances. The key point is that motif discovery alone does not solve CMML: without an explicit selection objective, one may produce an unwieldy and redundant library that fits idiosyncrasies of particular models. CMML therefore treats library construction as an optimization step layered atop interventional validation.

In summary, CMML is a structured pipeline of decision and optimization problems whose primitives are (i) interventional fingerprints for aligning feature identities across models, (ii) fixed-k causal abstraction tests for certifying motif equivalence, and (iii) a set-cover-style selection objective for producing a compact explanatory library. The next section presents an explicit algorithm realizing this decomposition under the intervention and sampling budgets parameterized by T and N.

5 Algorithm: constructing a Cross-Model Motif Library

We now describe an explicit procedure realizing the CMML decomposition stated in Section 4. The algorithm has six stages: (A) learn sparse feature coordinates within each model, (B) measure interventional effect fingerprints for a selected subset of features, (C) align feature identities across models by solving a fingerprint matching problem, (D) extract bounded-size candidate motifs associated with behaviors of interest, (E) canonicalize and validate motifs by causal abstraction tests, and (F) select a compact library by a set-cover objective. Throughout, we treat k as a small constant and make the intervention and sampling budgets explicit via parameters T and N.

(A) Per-model sparse features via SAE/transcoder. For each model M_i we first fix an *activation stream* to be featurized (e.g. MLP pre-activations, attention outputs, or residual stream) and a set of layers and token positions over which we will collect training data. This choice is part of the algorithmic interface: we require only that the stream admits interventions from $\mathcal{I}$ that are sufficiently local to approximate do-operations on individual feature coordinates. We then train a sparse autoencoder (or a transcoder) to obtain a dictionary $D_i \in \mathbb{R}^{d\times m}$ and encoder E_i producing codes $h_i \in \mathbb{R}^m$. The only properties we use downstream are (i) that the codes are sparse (so individual coordinates are plausibly interpretable intervention targets) and (ii) that the learned basis is stable enough that interventions on a coordinate can be implemented without systematically driving activations off-manifold. In practice we therefore prefer objectives and regularizers that encourage sparsity and decoder incoherence, and we evaluate reconstruction quality to avoid degenerate dictionaries.

Having learned F_i, we select a manageable subset $S_i \subseteq [m]$ of candidate features. Since intervention budget scales linearly in $|S_i|$, we typically combine multiple filters: activation frequency (to remove rarely used directions), magnitude/variance (to remove near-noise coordinates), and behavioral relevance (to prioritize features implicated by quick screening interventions on Y). This stage produces a per-model feature set on which we will compute fingerprints and search for motifs.

(B) Interventional effect fingerprints. Fix an intervention list $(I_t)_{t=1}^T \subseteq \mathcal{I}$ intended to probe the causal role of a feature coordinate while remaining approximately on-distribution. For each i and each $f \in S_i$ we estimate the fingerprint $\Phi_i(f)$ from (1) by sampling $x \sim \mathcal{D}$ (and optionally $x \sim \mathcal{D}_{\mathrm{OOD}}$) and computing the effect $\Delta_i(f; I_t, x)$ on the readout Y. We emphasize two design constraints. First, the effect functional should be sensitive to the behavior under study (e.g. class-logit differences for classification behaviors, tool-call indicators for agentic behaviors, refusal probability for safety behaviors). Second, the interventions should be "well-posed" in the sense discussed earlier: we prefer resample ablation, subspace patching, or interchange-style interventions that preserve the marginal distribution of the remaining activation space, so that measured discrepancies reflect a feature's causal contribution rather than arbitrary distribution shift.

We treat fingerprint estimation as a statistical task: for each (i, f, t) we obtain an empirical mean over N samples and retain (when useful) an empirical variance estimate. These variance estimates will later serve as confidence weights in the alignment objective and as a mechanism for automatically discarding unstable fingerprints. When OOD robustness is required, we concatenate fingerprints over $\mathcal{D}$ and $\mathcal{D}_{\mathrm{OOD}}$ (or maintain them separately and enforce agreement in both regimes), thereby preventing alignments that match only in-distribution idiosyncrasies.

(C) Cross-model feature alignment by assignment. Given fingerprints $\{\Phi_i(f) : f \in S_i\}$ and $\{\Phi_j(g) : g \in S_j\}$, we align features by solving a minimum-cost bipartite matching problem as in (2). The cost between f and g is given by a distance $d_\Phi(\Phi_i(f), \Phi_j(g))$, optionally re-optimized over sign/scale symmetries if the SAE convention admits them. When fingerprint variances are available, we may use a Mahalanobis-style distance or a robust loss that downweights noisy coordinates. To accommodate features that have no reliable counterpart (due to training differences, feature splitting/merging, or insufficient probing power), we allow dummy nodes so that the solution returns a partial bijection π_{ij} together with an "unmatched" set.

Since this alignment is computed pairwise, we optionally impose a global consistency step across multiple models: for example, we may choose a reference model M_{i_0} and align all others to it, or we may compute a cycle-consistent alignment by solving a synchronization problem over permutations. In either case, the alignment is used only as a proposal mechanism for motif canonicalization; it does not itself certify causal equivalence.

(D) Candidate motif discovery under a size bound. For each behavior template $b \in \mathcal{B}$ and each model M_i, we use a circuit discovery routine restricted to SAE features to produce bounded-size candidate motifs μ with $|V_\mu| \leq k$. The discovery procedure is modular: it may be implemented by greedy attribution over features, by path patching restricted to a small set of feature coordinates, or by searching for small subgraphs whose interventions substantially move Y. The crucial restriction is that the output must be representable as a directed graph on at most k nodes with intervention targets in G_i (as recorded by loc_μ), so that subsequent equivalence testing remains tractable in the fixed-k regime. We also store sufficient metadata (layer/position annotations and intervention semantics) to replay the motif under validation interventions.

(E) Canonicalization and causal-abstraction validation. We next cluster discovered motifs across models into candidate motif types. Canonicalization uses two sources of structure: the aligned feature identities (via π_{ij}) and the internal motif graph structure (edge patterns, layer orderings, and any annotations). Concretely, we map each motif instance into a canonical representation by relabeling nodes according to aligned feature IDs and by applying a fixed ordering convention on layers/positions; motifs with similar canonical representations are grouped into a cluster c. For each cluster we propose a canonical SCM $\mathsf{SCM}_{\tau(c)}$ whose variables correspond to the motif nodes and whose edges follow the shared structure observed in the cluster; we also propose abstraction maps $\alpha_{i,\tau(c)}$ into each participating model.

Canonicalization is only a hypothesis generator. Acceptance is determined by an interventional validation step: for each (i, τ) we construct a finite test suite $\mathcal{I}_\tau \subseteq \mathcal{I}$ of single-node and multi-node interventions that probe the motif's defining counterfactual constraints (including interchange interventions when appropriate). We then estimate discrepancies between the model-under-abstraction and the canonical SCM prediction on both $\mathcal{D}$ and $\mathcal{D}_{\mathrm{OOD}}$, accepting the instance only if all estimated discrepancies fall below ε with sampling slack chosen to ensure overall failure probability at most δ. This step is the point at which motif equivalence is certified: observational similarity or aligned fingerprints alone are insufficient.

(F) Library selection with approximation guarantees. After validation we obtain a set $\mathcal{C}$ of motif clusters, each covering a subset of target instances U (or $U_\star$). We then select a library $\mathcal{L}$ by solving the induced set cover (or budgeted maximum coverage) problem. Because exact optimization is intractable in general, we use the greedy algorithm that iteratively adds the motif type providing the largest marginal increase in coverage (or the best marginal gain per unit cost if motifs have heterogeneous validation costs). The output consists of the selected canonical motif SCMs, the validated abstraction maps $\alpha_{i,\tau}$ for each covered instance, and the associated validation reports. In Section 6 we will state the conditions under which this procedure admits formal guarantees: interventionally defined fingerprints enable reliable alignments; fixed-k validation is tractable with polynomial query complexity; and greedy selection achieves the standard harmonic-number approximation ratio for set cover.

6 Theory I: identifiability and alignment from interventional fingerprints

We isolate the portion of the CMML pipeline whose role is to assign *stable identities* to internal feature coordinates across models. The central difficulty is that a learned feature basis F_i is not, by itself, canonically labeled: even when two models compute the same abstract function, their internal representations may be related by permutations, sign flips, or (in the absence of additional structure) more general rotations. Our approach is to use *interventional effect fingerprints* as signatures that are invariant to such relabelings and that concentrate with finitely many interventions and samples.

Generative assumptions and identifiability target. We work in a simplified latent-feature setting sufficient to state recovery guarantees. For each model M_i and a fixed activation stream of dimension d, we posit latent coordinates $s \in \mathbb{R}^m$ (shared across i) and a mixing matrix $A_i \in \mathbb{R}^{d\times m}$ such that

$$a_i \;=\; A_i s + \eta, \tag{3}$$

where s is ρ-sparse (or approximately sparse), η is mean-zero sub-Gaussian noise, and A_i satisfies standard incoherence conditions. In this regime, sparse dictionary learning is identifiable up to permutation and coordinate-wise sign (and, depending on normalization, scale). Concretely, if D_i denotes an SAE decoder trained on samples of a_i, then under these assumptions we expect

$$D_i \approx A_i P_i \Sigma_i, \tag{4}$$

where P_i is a permutation and Σ_i is diagonal with entries in $\{\pm 1\}$ (and possibly positive scales). Thus a coordinate index $f \in [m]$ in model i is *not* directly comparable to a coordinate $g \in [m]$ in model j without recovering the relative permutation $P_j^{-1} P_i$ (and sign/scale).

The remaining ingredient is causal: we assume the intervention family $\mathcal{I}$ contains operations that, to a good approximation, implement $\mathrm{do}(s_\ell \leftarrow \tilde{s}_\ell)$ for a single latent coordinate ℓ, realized in practice by intervening on the corresponding SAE coordinate. This is an approximation statement: interventions must be sufficiently local and sufficiently on-manifold that the measured change in the readout Y reflects the causal contribution of the targeted coordinate rather than an arbitrary distribution shift.

Fingerprints and concentration. Fix a list of interventions $(I_t)_{t=1}^T \subseteq \mathcal{I}$. For each model i and feature $f \in S_i$ we define the population fingerprint

$$\Phi_i(f) \;:=\; \big(\mathbb{E}_{x\sim\mathcal{D}}[\Delta_i(f; I_t, x)]\big)_{t=1}^T, \tag{5}$$

and its empirical estimate from N samples per intervention,

$$\widehat{\Phi}_i(f) \;:=\; \left(\frac{1}{N} \sum_{n=1}^{N} \Delta_i(f; I_t, x_t^{(n)}) \right)_{t=1}^{T}, \qquad x_t^{(n)} \overset{\text{iid}}{\sim} \mathcal{D}. \tag{6}$$

Under bounded-variance (or sub-Gaussian) assumptions on $\Delta_i(f; I_t, x)$, standard concentration yields

$$\|\widehat{\Phi}_i(f) - \Phi_i(f)\|_2 \;\le\; O\left(\sqrt{\frac{T \log(1/\delta)}{N}} \right) \tag{7}$$

uniformly over a finite candidate set S_i by a union bound, or more sharply using empirical Bernstein bounds when variances are estimated. This motivates the scaling $N = \tilde{O}(\varepsilon^{-2} \log(|S_i|/\delta))$ when we seek ℓ_2-accuracy ε per fingerprint.

Alignment as minimum-cost matching and a margin condition. For a pair (i, j), we align features by solving a bipartite assignment problem minimizing a total cost

$$\min_{\pi} \sum_{f \in S_i} d_\Phi\left(\widehat{\Phi}_i(f), \widehat{\Phi}_j(\pi(f)) \right), \tag{8}$$

where d_Φ is a chosen metric (e.g. ℓ_2 or a variance-weighted norm), and π may be partial via dummy nodes. The analysis hinges on a separation (margin) assumption: for identifiable features f, there exists a unique counterpart $g^\star$ such that

$$d_\Phi(\Phi_i(f), \Phi_j(g^\star)) + \gamma \;\le\; d_\Phi(\Phi_i(f), \Phi_j(g)) \quad \text{for all } g \neq g^\star \tag{9}$$

for some margin $\gamma > 0$. When γ dominates the fingerprint estimation error, the optimal assignment is forced to select the true correspondence.

Formally, in the latent-feature model above with interventions that act as approximate do-operations on single latent coordinates, one obtains the following recovery statement (stated earlier as Theorem 1): with $T = \tilde{O}(\log m)$ appropriately chosen interventions and $N = \tilde{O}(\gamma^{-2} \log(m/\delta))$ samples per intervention, the Hungarian assignment recovers the correct permutation/sign on all features satisfying the margin condition, with probability at least $1 - \delta$. The proof proceeds by (i) identifiability of the sparse dictionary up to permutation/sign, which reduces the alignment problem to identifying that permutation, and (ii) uniform concentration of $\widehat{\Phi}$ around Φ, which ensures the empirical cost matrix preserves the argmin structure implied by the margin.

Observational impossibility and lower bounds. The use of interventions is not merely algorithmically convenient; it is information-theoretically necessary in general. In particular, if one observes only the joint distribution of activations and outputs under $x \sim \mathcal{D}$ (with no interventions), then there exist model pairs whose internal feature spaces are related by an orthogonal rotation that preserves all observable statistics while destroying any notion of coordinate-wise correspondence. In the latent model, if s is isotropic on its support and $A_j = A_i R$ for an orthogonal R acting within an identifiable subspace, then a_j and a_i can induce the same distribution even though the individual coordinates of s have been mixed. Under such constructions, any estimator attempting to match coordinates across models from observational data alone has expected accuracy at most $1/m + o(1)$ on the rotated subset (Theorem 2). The proof is a standard indistinguishability argument (e.g. Le Cam or Fano): two hypotheses (two different rotations) generate identical observations, so no test can reliably distinguish them, and therefore no alignment procedure can do better than chance on the affected coordinates.

These lower bounds justify why we define fingerprints in terms of *causal effects* on Y: interventions break the rotational symmetry by selecting privileged directions through their effect on the readout.

Robustness to noise and approximate interventions. Two deviations from the idealized setting are unavoidable: fingerprints are noisy estimates, and interventions are only approximately equal to do-operations on latent coordinates. Both can be incorporated as perturbations.

First, if the true fingerprints are perturbed by additive noise $\xi_{i,f}$ with $\|\xi_{i,f}\|_2 \leq \beta$, then the margin condition degrades from γ to $\gamma - 2\beta$ by the triangle inequality; hence exact recovery persists as long as $\beta < \gamma/2$. Since β can be taken as the empirical estimation error, this reproduces the scaling $N = \tilde{O}(\gamma^{-2} \log(m/\delta))$.

Second, suppose the intervention I_t on SAE coordinate f implements a mixture of latent do-operations with a small leakage term, e.g.

$$\mathrm{do}(s \leftarrow s + \epsilon e_\ell + \zeta), \qquad \|\zeta\|_2 \leq \lambda, \tag{10}$$

for some λ capturing off-target effects and off-manifold drift. If the causal effect of ζ on Y is Lipschitz (or otherwise bounded) in magnitude by $L\|\zeta\|_2$, then fingerprints incur a systematic bias of size at most $L\lambda$. Again, recovery reduces to requiring an effective margin γ that dominates both statistical error and intervention bias. This is the point at which "on-distribution" intervention design becomes theoretically relevant: resample ablation and related procedures can be viewed as methods for reducing λ by preserving the marginal distribution of non-target coordinates.

Feature splitting, merging, and partial matchings. Across independently trained models, it is common for a semantic factor to be represented by multiple coordinates in one model (splitting) or for multiple factors to be entangled into one coordinate (merging). In such cases a bijective alignment is not well-posed. Our assignment formulation accommodates this by allowing dummy nodes (unmatched features), but it is useful to articulate what can and cannot be recovered.

A minimal model of splitting is: in model i a latent coordinate s_ℓ exists, while in model j the same effect on Y is carried by two coordinates $s_{\ell,1}, s_{\ell,2}$ whose interventions each produce approximately half the effect, so that no single coordinate in j matches s_ℓ under the fingerprint distance. Then any one-to-one matcher must either (a) leave s_ℓ unmatched, or (b) commit an error by matching it to an imperfect counterpart. By introducing a rejection option (dummy nodes) and a calibrated threshold on d_Φ, we can guarantee a *partial recovery* statement: all pairs with distance below threshold are correct, and all ambiguous cases are rejected. Formally, if true correspondences satisfy $d_\Phi(\Phi_i(f), \Phi_j(g^\star)) \leq \eta$ while all non-correspondences satisfy $d_\Phi(\Phi_i(f), \Phi_j(g)) \geq \eta + \gamma$, then choosing a threshold in $(\eta, \eta + \gamma)$ yields a precision guarantee (no false matches) at the cost of recall (some features rejected). The same logic applies to merging: a merged coordinate lacks a unique counterpart and should be rejected rather than forcibly matched.

This robustness mechanism also interacts favorably with downstream motif discovery. Since motifs are extracted as bounded-size subgraphs, it is typically preferable to retain only high-confidence aligned features; missing nodes can be re-introduced later via motif-level validation, where multi-node interventions can detect whether a split representation collectively realizes the same causal role.

In summary, the identifiability story is as follows: sparse coding provides a representation that is unique up to simple symmetries; interventions define fingerprints that break those symmetries by referencing causal effects on Y; and matching under a margin condition yields finite-sample alignment guarantees, while observational data alone admits impossibility results. The remaining theoretical question is how to *validate* that aligned subgraphs correspond to the same canonical causal motif, which we address next.

7 Theory II: motif equivalence testing and fixed-k tractability

We now formalize the validation step in which we decide whether a bounded-size subgraph extracted from a model implements a given canonical motif SCM_τ *up to causal abstraction*, and we bound the interventions and samples needed to accept/reject this hypothesis with error probability at most δ. The key observation is that once we restrict attention to motifs of size at most k, equivalence testing reduces to a finite family of interventional constraints whose cardinality depends only on k (and mild structural parameters such as indegree), yielding a tractable procedure. Conversely, if k is unrestricted, the same problem subsumes standard NP-hard graph matching tasks.

Motifs as bounded causal subgraphs. Fix a model M_i with learned features F_i and internal causal graph G_i. A *candidate motif instance* is a directed subgraph

$$\mathcal{M}_i \;=\; (V_i, E_i), \qquad V_i \subseteq F_i, \quad |V_i| \le k,$$

together with a designated set of *intervenable* nodes V_i and a readout Y. In practice, $\mathcal{M}_i$ is produced by automated circuit discovery methods constrained to SAE features and bounded size, and E_i encodes a hypothesized causal dependency (e.g. as suggested by path patching scores). For theory, we assume that for each $v \in V_i$ we may apply interventions from $\mathcal{I}$ localized to that feature (e.g. resample ablation, clamping, or subspace patching), and that each intervention yields a well-defined interventional distribution over Y under inputs $x \sim \mathcal{D}$ (and similarly for $\mathcal{D}_{\mathrm{OOD}}$).

A canonical motif type τ is represented by a structural causal model SCM_τ with variables $Z_1, \dots, Z_{k_\tau}$ for $k_\tau \le k$. We assume SCM_τ comes with a specified set of allowed interventions on its variables (mirroring $\mathcal{I}$ at the abstract level), and with a prediction for the induced effect on Y as a function of those interventions and the input distribution. This covers both deterministic motifs (where Y is a deterministic functional of Z) and stochastic motifs (where Y has an induced conditional distribution).

Equivalence up to abstraction as a finite test family. An *abstraction map* $\alpha_{i,\tau}$ assigns each abstract variable Z_j to an internal feature $v_j \in V_i$ (possibly together with sign/scale and a layer index when the same feature appears at multiple sites), and interprets abstract interventions $\mathrm{do}(Z_j \leftarrow \cdot)$ as concrete interventions $I \in \mathcal{I}$ applied to v_j. Given $\alpha_{i,\tau}$, we obtain an induced family of interventional distributions on the readout,

$$\mathcal{P}_{i,\alpha} \;:=\; \big\{\, \mathcal{L}\big(Y \,\big|\, \mathrm{do}_\alpha(u),\, x \sim \mathcal{D}\big) \;:\; u \in \mathcal{U}_\tau \,\big\},$$

where $\mathcal{U}_\tau$ indexes a chosen set of abstract interventions (single-node, multi-node, and/or interchange interventions), and $\mathrm{do}_\alpha(u)$ denotes the corresponding concrete intervention pattern on M_i. Analogously, SCM_τ induces a family $\mathcal{P}_\tau$ over Y under the same intervention index set $\mathcal{U}_\tau$.

We say that $\mathcal{M}_i$ *ε-implements* SCM_τ under α on distribution $\mathcal{D}' \in \{\mathcal{D}, \mathcal{D}_{\mathrm{OOD}}\}$ if

$$\sup_{u \in \mathcal{U}_\tau} d\big(\mathcal{L}\big(Y \mid \mathrm{do}_\alpha(u),\, x \sim \mathcal{D}'\big),\; \mathcal{L}\big(Y \mid \mathrm{do}(u),\, x \sim \mathcal{D}'\big)\big) \;\le\; \varepsilon, \tag{11}$$

for a fixed discrepancy d (e.g. total variation when Y is discrete, or an absolute difference of expected logit/readout when Y is real-valued). The validation task is to decide whether there exists an α such that (11) holds simultaneously for $\mathcal{D}$ and $\mathcal{D}_{\text{OOD}}$.

The crucial point is that for fixed k we may choose $\mathcal{U}_\tau$ to be finite and of size poly(k) while still pinning down the causal role of each node in the motif. Concretely, we may include: (i) single-node interventions on each Z_j at two or more reference values (e.g. ablate vs. resample); (ii) selected pairwise interventions $\mathrm{do}(Z_j \leftarrow a, Z_\ell \leftarrow b)$ to detect interaction terms; and (iii) interchange interventions that swap representations between two inputs while holding other variables fixed, which operationalize functional dependencies without requiring a dense grid over values. Under bounded indegree Δ, the number of such constraints needed to rule out non-isomorphic causal graphs on k nodes is bounded by a function $f(k, \Delta)$, which we treat as constant in fixed-k analysis.

Algorithmic structure of the equivalence test. Given $\mathcal{M}_i = (V_i, E_i)$ and SCM_τ, we enumerate candidate abstraction maps α from $\{Z_1, \dots, Z_{k_\tau}\}$ into V_i. Without additional structure this is $O(k!)$ maps, and with bounded indegree and typed nodes (e.g. "retrieval head", "induction feature") it can be substantially smaller; in any case it is $f(k, \Delta)$ for fixed k. For each candidate α, we estimate the discrepancy in (11) using $T := |\mathcal{U}_\tau|$ interventions and N input samples per intervention:

$$\widehat{\mu}_{i,\alpha}(u) \;=\; \frac{1}{N}\sum_{n=1}^{N} \psi\Big(Y_{i,\alpha,u}^{(n)}\Big), \qquad \widehat{\mu}_\tau(u) \;=\; \frac{1}{N}\sum_{n=1}^{N} \psi\Big(Y_{\tau,u}^{(n)}\Big),$$

where ψ is either the identity (for real Y) or an indicator/test function used to approximate a distributional discrepancy, and $Y_{i,\alpha,u}^{(n)}$ denotes the model readout under intervention u and input $x^{(n)} \sim \mathcal{D}'$. We accept α if

$$\max_{u \in \mathcal{U}_\tau} |\widehat{\mu}_{i,\alpha}(u) - \widehat{\mu}_\tau(u)| \;\le\; \varepsilon_{\text{test}},$$

and we accept that $\mathcal{M}_i$ realizes τ if any α passes (on both $\mathcal{D}$ and $\mathcal{D}_{\text{OOD}}$). The dependence on m enters only through the upstream motif extraction; conditional on having a k-node candidate, validation is independent of m up to the cost of forward passes.

Intervention and sample complexity. Assume that for each fixed $u \in \mathcal{U}_\tau$, the scalar statistic $\psi(Y)$ is sub-Gaussian with parameter σ^2 under both the model and SCM_τ, uniformly over inputs and any stochasticity induced by resampling interventions. Then, for any fixed α, Hoeffding- or Bernstein-type bounds imply

$$\mathbb{P}\Big(\max_{u \in \mathcal{U}_\tau} |\widehat{\mu}_{i,\alpha}(u) - \mu_{i,\alpha}(u)| > t\Big) \;\le\; 2T \exp\Big(-c\,\frac{Nt^2}{\sigma^2}\Big),$$

for a universal constant $c > 0$. Taking a union bound over the $f(k, \Delta)$ candidate maps α and over the two distributions $\mathcal{D}, \mathcal{D}_{\mathrm{OOD}}$, it suffices to choose

$$N \;=\; \tilde{O}\left(\frac{\sigma^2}{\varepsilon^2}\log\frac{Tf(k,\Delta)}{\delta}\right) \tag{12}$$

so that all empirical estimates are simultaneously within $O(\varepsilon)$ of their population counterparts with probability at least $1 - \delta$. This yields the fixed-k tractability claim: with $T = \mathrm{poly}(k, \log(1/\delta))$ interventions (determined by the test family $\mathcal{U}_\tau$) and N as in (12), we can accept/reject ε-equivalence using $O(TN)$ model queries per (i, τ) validation attempt, up to the multiplicative enumeration factor $f(k, \Delta)$. This is the content of the fixed-k result stated abstractly as Theorem 3.

We emphasize that ε^{-2} scaling in (12) is unavoidable for mean-estimation-based tests under sub-Gaussian noise, and the $\log(1/\delta)$ dependence is similarly tight up to constants. What fixed-k buys is that the number of tested constraints T and the number of candidate abstractions $f(k, \Delta)$ do not grow with the ambient feature count m, so validation remains stable as models scale provided we can extract bounded motifs.

Approximate interventions and the effective tolerance. As in the alignment setting, concrete interventions may deviate from ideal abstract do-operations. If each mapped intervention incurs a bounded bias $b(u)$ in the relevant statistic, so that $|\mu_{i,\alpha}(u) - \mu^\star_{i,\alpha}(u)| \le b(u)$ where $\mu^\star$ denotes the counterfactual quantity under perfect interventions, then the test can only certify equivalence up to an *effective* tolerance $\varepsilon + \max_u b(u)$. Operationally, one sets $\varepsilon_{\mathrm{test}}$ to account for both estimation error and a measured (or conservatively bounded) intervention bias term, and one treats failures localized to specific u as evidence that the intervention family $\mathcal{I}$ is insufficiently on-manifold for that motif instance.

Hardness without bounded size and tightness of the restriction. If we remove the restriction $|V_i| \le k$ and permit arbitrary motif graphs, then motif matching and equivalence testing inherit the computational hardness of classical subgraph problems. In particular, even if feature alignment is given, deciding whether a graph encoding of one motif embeds into another (or whether two models share a common motif of size at least K) subsumes Subgraph Isomorphism and Maximum Common Subgraph, yielding NP-hardness (as summarized by Theorem 4). Moreover, any algorithm that attempts to search over large motifs faces an exponential blowup in the number of candidate abstractions and constraints required to distinguish non-equivalent graphs; thus the fixed-k regime is not merely a convenient assumption but the natural boundary at which we can obtain both provable guarantees and practical scaling.

This concludes the validation component: with aligned features and bounded candidate motifs, we can certify motif instances by checking a finite interventional constraint family with controlled sample complexity. We next address how to select a compact motif library across many validated instances.

8 Theory III: library objective and approximation via set cover

Having reduced validation of a *fixed* motif hypothesis to a finite interventional test family, we now address the *cross-model* objective: from a potentially large collection of validated motif hypotheses, we wish to produce a compact library $\mathcal{L}$ that explains as many model–behavior instances as possible, with explicit approximation guarantees and a clear scaling law in the number of models and behaviors.

Instances to be explained and candidate motif clusters. Let $\mathcal{B}$ be a behavior family (tasks, templates, or benchmark items) equipped with a readout Y and intervention family $\mathcal{I}$. For each model M_i and each behavior $b \in \mathcal{B}$, we run the upstream pipeline (feature learning, candidate motif extraction, canonicalization, and validation) and obtain a set of *validated explanations* in the form of motif clusters (canonical motif types) that ε-implement their corresponding canonical SCMs under some abstraction map. We define the universe of explainable instances as

$$U \;:=\; \big\{(i,b)\;:\; i\in[r],\; b\in\mathcal{B},\; \exists \text{ a validated motif cluster explaining } (i,b)\big\}.$$

Let $\mathcal{C}$ denote the set of all candidate motif clusters produced by canonicalization. Each cluster $c \in \mathcal{C}$ is associated with a proposed canonical motif $(\tau(c), \mathsf{SCM}_{\tau(c)})$ and a collection of abstraction maps $\{\alpha_{i,b,c}\}$ for the instances it explains. We write the *coverage set* of cluster c as

$$S_c \;:=\; \big\{(i,b)\in U\;:\; \text{cluster } c \text{ passes the causal abstraction test for } (M_i,b) \text{ on } \mathcal{D}\cup\mathcal{D}_{\mathrm{OOD}}\big\}.$$

By construction, S_c depends on the tolerance ε, the allowed intervention family $\mathcal{I}$, and the chosen validation test family $\mathcal{U}_{\tau(c)}$; however, at this stage these are fixed, and we treat $\{S_c\}_{c\in\mathcal{C}}$ as known.

Set cover formulation (unweighted and weighted). The most basic library objective is to explain all explainable instances using the smallest number of motifs. This yields the (unweighted) set cover problem:

$$\min_{\mathcal{L}\subseteq\mathcal{C}} |\mathcal{L}| \quad \text{s.t.} \quad \bigcup_{c\in\mathcal{L}} S_c \;=\; U. \tag{13}$$

In many applications we may prefer a weighted objective capturing that some motifs are intrinsically more complex (larger k, richer intervention semantics, or additional parameters). Assign a nonnegative cost w_c to each cluster c and consider the weighted set cover:

$$\min_{\mathcal{L}\subseteq\mathcal{C}} \sum_{c\in\mathcal{L}} w_c \quad \text{s.t.} \quad \bigcup_{c\in\mathcal{L}} S_c \;=\; U. \tag{14}$$

Both (13) and (14) are NP-hard in general (cf. Theorem 4), and thus one should not expect exact polynomial-time solutions without additional structure. We therefore adopt a greedy approximation algorithm with the standard logarithmic guarantee (Theorem 5).

MDL interpretation and explicit cost choices. The set cover objective admits a minimum description length (MDL) reading that is useful for selecting w_c and for handling partial coverage. Consider a two-part code: (i) describe the selected library $\mathcal{L}$ and its canonical SCMs, and (ii) describe, for each instance (i, b), which motif in $\mathcal{L}$ is used (together with the validated abstraction map identifier) or, if uncovered, describe a residual "exception" model. A minimal library then corresponds to a short description of the family of behaviors across models. Concretely, one may set

$$w_c \;\approx\; \underbrace{\mathrm{DL}(\mathsf{SCM}_{\tau(c)})}_{\text{canonical motif complexity}} \;+\; \underbrace{\mathrm{DL}(\mathcal{U}_{\tau(c)})}_{\text{intervention semantics}} \;+\; \underbrace{\lambda\, k_{\tau(c)}}_{\text{size penalty}},$$

for a tunable $\lambda > 0$. The MDL view also suggests relaxing the hard constraint $\cup_{c\in\mathcal{L}} S_c = U$ to a penalized objective when coverage is imperfect due to limited interventions or overly strict ε:

$$\min_{\mathcal{L}\subseteq\mathcal{C}} \sum_{c\in\mathcal{L}} w_c \;+\; \beta\,\big|U \setminus \cup_{c\in\mathcal{L}} S_c\big|,$$

where β encodes how costly it is to leave an instance unexplained relative to adding a new motif. The hard-constraint set cover is the limit $\beta \to \infty$.

Greedy selection and approximation guarantee. For the unweighted case, the greedy algorithm iteratively selects the motif cluster covering the largest number of currently uncovered instances. For the weighted case, it selects the cluster maximizing the ratio $|S_c \cap U_{\text{rem}}|/w_c$, where U_{rem} denotes uncovered instances at the current iteration. Formally, starting from $\mathcal{L}_0 = \emptyset$ and $U_{\text{rem}}^{(0)} = U$, define

$$c_t \in \arg\max_{c\in\mathcal{C}} \frac{|S_c \cap U_{\text{rem}}^{(t-1)}|}{w_c}, \qquad \mathcal{L}_t := \mathcal{L}_{t-1} \cup \{c_t\}, \qquad U_{\text{rem}}^{(t)} := U_{\text{rem}}^{(t-1)} \setminus S_{c_t},$$

until $U_{\text{rem}}^{(t)} = \emptyset$. Theorem 5 yields that the resulting library has total weight at most $H(|U|)$ times the optimal (where $H(n)$ is the n-th harmonic number), and in particular at most $(1 + \ln |U|)$ times optimal. As is standard, this approximation factor is essentially best possible in polynomial time under conventional complexity assumptions.

Implications for scaling in r and $|\mathcal{B}|$. The relevance of the greedy guarantee is that the approximation factor depends only on $|U|$, not on the number of candidate motifs $|\mathcal{C}|$ nor on the ambient feature dimension m. Since $|U| \le r|\mathcal{B}|$, the worst-case approximation scales as $O(\log(r|\mathcal{B}|))$. Thus, as we increase the number of models (additional seeds, sizes, or post-training variants) and enlarge the behavior family, the degradation in the optimality guarantee is only logarithmic.

Equally important, once the coverage sets $\{S_c\}$ are computed, the greedy selection stage is computationally light relative to upstream validation: it requires maintaining uncovered counts and selecting maxima, which can be implemented in time $O\big(\sum_{c\in\mathcal{C}} |S_c|\big)$ up to data-structure overhead. In typical regimes, the dominant cost remains the model-query budget used to certify each membership $(i, b) \in S_c$, not the combinatorial selection itself.

Library stability and incremental updates. The set cover/MDL formulation also clarifies how to update the library when new models or behaviors are added. Suppose we append a new model M_{r+1} or new behavior b_{new}. We need only validate candidate motifs on the new instances to update the coverage sets S_c, and then rerun greedy (or continue greedy from the existing $\mathcal{L}$ by covering newly added universe elements). While greedy need not produce identical solutions under incremental growth, the logarithmic approximation guarantee continues to hold for the enlarged universe, and in practice the library changes only when genuinely novel motifs appear that cover instances previously uncovered by existing motifs.

Interpretation: universality as small set cover. Finally, the library viewpoint provides an operational measure of "universality up to abstraction." If a small number of canonical motifs suffice to cover a large fraction of U across diverse $\mathcal{D}_{\text{OOD}}$ settings, then the behavior family admits a compact causal basis stable across the model family. Conversely, if covering U requires a library whose size grows nearly linearly in $|U|$, then either (i) the behavior family is not decomposable into bounded motifs under the intervention family $\mathcal{I}$, (ii) the extracted candidates are too noisy to canonicalize reliably, or (iii) the models implement genuinely heterogeneous mechanisms. In this sense, the set cover objective provides not merely a selection heuristic but a quantitative diagnostic: library size (or MDL cost) summarizes the extent to which our motif abstraction captures cross-model commonality under interventions.

9 Experimental plan: empirical validation of cross-model motif libraries

We outline an experimental program designed to (i) verify the alignment and motif-equivalence claims in settings with known ground truth, (ii) measure scaling and robustness across realistic model families, and (iii) stress-test the causal nature of our validation via falsification and ablation studies. Throughout, we fix a readout variable Y appropriate to the behavior (e.g. next-token logits on a designated position, a discrete tool-call, or a refusal indicator), a behavior family $\mathcal{B}$, an allowed intervention family $\mathcal{I}$, and in-distribution / out-of-distribution input families $\mathcal{D} \cup \mathcal{D}_{\mathrm{OOD}}$.

(I) Compiled-transformer testbeds with known motifs. We first evaluate on "compiled" transformers whose internal computation implements a known algorithmic motif with an explicit, human-specified causal graph (e.g. induction heads, associative retrieval, parity checks, multi-step copying). Concretely, we consider synthetic sequence tasks where a small circuit is sufficient and can be embedded into a transformer with controlled superposition and noise. This setting supports a ground-truth SCM_τ and an intended abstraction map $\alpha_i^\star$ from canonical variables $Z_1, \ldots, Z_k$ to internal nodes.

The primary goal is to test whether our pipeline recovers (a) the correct feature correspondence (up to permutation/sign/scale) across independently trained instances, and (b) the correct motif type τ and causal structure up to the equivalence induced by the validation test family. We measure:

- *Alignment accuracy:* given ground-truth correspondences among latent variables (or designated internal modules), we score π_{ij} by top-1 matching accuracy and by a soft score based on rank of the true match under fingerprint distance.

- *Motif recovery:* whether the discovered motif cluster is isomorphic (under aligned features) to the planted motif, and whether the learned canonical SCM_τ matches the planted intervention response table on a held-out intervention set.

- *Statistical efficiency:* empirical dependence on the number of interventions T and samples per intervention N, including the observed probability of correct recovery as a function of (T, N) for fixed tolerance ε.

Because $\mathcal{I}$ can be implemented exactly in these testbeds (true do-operations on known latent coordinates), we also quantify the gap between ideal interventions and their SAE-feature approximations by comparing fingerprints computed in latent space versus activation space.

(II) Model families across seeds and sizes. We next consider realistic model families $\{M_i\}_{i=1}^{r}$ produced by varying (a) random seed, (b) model width/depth, and (c) training data scale, while holding architecture class fixed. For each family we train a feature basis F_i (SAE/transcoder) on a chosen activation stream (e.g. MLP residual stream at a subset of layers) and compute interventional fingerprints $\Phi_i(f)$ for a candidate feature set S_i. We emphasize two evaluation regimes:

1. *Within-size, across-seed:* tests stability of recovered correspondences under minimal distribution shift, isolating the effect of non-identifiability and superposition.

2. *Across-size:* tests whether motif types and their canonical causal constraints persist under scaling, allowing that abstraction maps $\alpha_{i,\tau}$ may involve different layers/heads/features as width changes.

Here we do not have ground truth, so we use internal consistency metrics: cycle-consistency of alignments ($\pi_{ij} \circ \pi_{jk} \approx \pi_{ik}$ on high-confidence features), agreement of motif clusters across subsets of models, and predictive validity of motifs via transfer tests (below).

(III) Base versus instruction-tuned versus preference-optimized (RLHF/DPO). A central claim is that motifs can be universal up to abstraction even when post-training substantially changes behavior. We therefore construct triplets of related models: a base pre-trained M_{base}, an instruction-tuned M_{IT}, and a preference-optimized M_{pref} (RLHF or DPO), ideally sharing a common pre-training initialization. For each behavior $b \in \mathcal{B}$ we ask:

- whether a motif validated in M_{base} remains present (possibly with different α) in M_{IT} and M_{pref};
- whether novel motifs appear that explain post-training-specific behaviors (e.g. refusal mechanisms, policy shaping, tool-use routing);
- whether alignments based on fingerprints computed on neutral $\mathcal{D}$ transfer to domains where the post-training signal dominates (selected subsets of $\mathcal{D}_{\text{OOD}}$).

We treat these as paired comparisons: we hold $\mathcal{I}$ fixed and report motif coverage changes and changes in validation margins (empirical effect mismatches relative to ε).

Metrics: alignment, transfer, robustness, and certification. We report four families of metrics, each computed with explicit confidence intervals over both intervention sampling and input sampling.

1. *Feature alignment metrics.* For each pair (i, j), we report (a) matching cost under the optimal assignment, (b) fraction of features aligned above a margin threshold (a calibrated “high-confidence” subset), and (c) stability of π_{ij} under resampling of $\mathcal{D}$ and under moderate changes in $\mathcal{I}$.

2. *Motif transfer accuracy.* Given a motif cluster c learned primarily from a subset of models $I_{\text{train}} \subseteq [r]$, we test on held-out models I_{test} by instantiating the proposed $\mathsf{SCM}_{\tau(c)}$ with abstraction maps $\alpha_{i,\tau}$ derived from aligned features, and measuring agreement of interventional effect predictions on Y. This yields a direct measure of whether canonical motifs generalize across the family rather than merely re-fitting per model.

3. *OOD robustness.* For each validated (i, τ) we evaluate the same interventional constraints on $\mathcal{D}_{\text{OOD}}$, reporting the distribution of effect mismatches and the fraction of motifs whose certification holds uniformly (or with bounded degradation) under OOD shift. We distinguish semantic OOD (new topics/templates) from syntactic OOD (longer contexts, altered token distributions), since different motifs are expected to fail differently.

4. *Falsification and negative controls.* We report the false-accept rate of the causal abstraction test under deliberately incorrect hypotheses: mismatched motif types, random abstraction maps, and “shuffled” alignments produced by permuting fingerprints. A valid certification procedure should reject these at rates consistent with the target failure probability δ, up to estimation error.

Falsification tests targeted to causality (not correlation). To ensure that success is not driven by incidental correlations in activations, we include interventions that specifically break correlational explanations:

- *Counterfactual interchange:* swap the activations of candidate motif variables across inputs x, x' matched on superficial statistics but differing in the motif-relevant property (e.g. presence/absence of an induction trigger), and test whether Y changes according to SCM_τ.

- *Path-constraint checks:* intervene on purported parent variables while holding downstream activations fixed via patching; reject motifs that only reproduce marginal effects but fail conditional (path-specific) constraints.

- *Anticausal baselines:* define a “reverse” motif graph with edges reversed (or randomized) and show that it cannot pass the interventional constraints despite matching observational correlations.

These tests are reported alongside standard effect-size summaries so that acceptance corresponds to genuinely causal invariants under $\mathcal{I}$.

Ablations on fingerprint design and intervention family. We perform controlled ablations to identify which components are necessary for reliable cross-model alignment.

1. *Fingerprint statistics.* We compare fingerprints based on (a) mean logit differences at selected positions, (b) KL divergence between output distributions, (c) discrete readout changes (e.g. tool-call flips), and (d) multi-position summaries. We also ablate whether fingerprints are computed on $\mathcal{D}$ alone or on a mixture $\mathcal{D} \cup \mathcal{D}_{\mathrm{OOD}}$.

2. *Intervention operators.* Within $\mathcal{I}$, we compare zero-ablation, resample ablation, mean replacement, subspace projection, and sparse feature editing (setting a coordinate in h_i). We quantify (i) on-distributionness of intervened activations (e.g. via reconstruction error or density proxies) and (ii) downstream stability of the alignment.

3. *Budget scaling.* We sweep T and N to empirically map the tradeoff between cost and accuracy, and to verify that observed sample complexity tracks the qualitative dependence predicted by the separation-margin story (in particular, the emergence of a stable high-confidence aligned subset).

As a final robustness check, we vary the SAE dictionary size m and sparsity regularization to test whether motifs are stable across different feature granularities, and we report when increased m merely refines abstraction maps versus when it changes the inferred motif types.

Deliverables of the experimental section. The empirical outcome is a set of quantitative curves and certification tables: alignment quality across families, motif transfer and OOD validity, and calibrated rejection rates on falsification tests. The experimental narrative is thereby forced into explicit, testable statements: which motifs are shared, under what interventions, at what tolerance ε, and with what failure probability δ as estimated from repeated trials.

(X) Released artifacts: library schema, tooling, benchmarks, and reporting standards. To make the claims of universality up to abstraction auditable and extensible, we will release a set of artifacts that render each accepted motif instance (i, τ) as a *reproducible object*: one should be able to (a) instantiate the canonical SCM_τ, (b) apply the corresponding abstraction map $\alpha_{i,\tau}$ to a concrete model M_i, (c) re-run the agreed-upon intervention suite from $\mathcal{I}$ on fresh samples from $\mathcal{D} \cup \mathcal{D}_{\mathrm{OOD}}$, and (d) recover the validation decision (accept/reject) within statistical tolerance. We therefore treat the motif library not as a narrative description but as a typed schema with explicit interfaces for interventions, effect measurements, and uncertainty.

Motif library schema. We will publish $\mathcal{L} = \{(\tau, \mathsf{SCM}_\tau)\}$ in a machine-readable format that separates *canonical* causal content from *model-specific* instantiations. Each motif type entry includes:

- *Canonical variables and graph:* a node list $(Z_1, \ldots, Z_k)$, directed edges, and (when available) a parametric or tabular specification of structural relations, including which variables are designated as intervene-able under the intended $\mathcal{I}$.
- *Readout interface:* a declaration of the readout Y (e.g. logits at a named position, a categorical action, or a scalar score) and an allowed family of effect summaries (e.g. expected logit difference, KL divergence, or probability shift).
- *Interventional constraints:* the finite set of inequalities/equalities used by the causal abstraction test (e.g. interchange constraints, path-specific constraints), stated as estimable functionals of the interventional distribution of Y. This is the portion of the schema that is intended to be stable under re-implementations.
- *Reference intervention suite:* a recommended family of interventions $\{I_t\}_{t=1}^{T_v} \subset \mathcal{I}$ sufficient to validate the motif for typical choices of ε, δ, together with any admissibility checks (e.g. resample-ablation requirements to keep activations on-distribution).

We emphasize that SCM_τ need not be fully identified as a set of structural equations; it suffices to encode the interventional invariants that define the equivalence class we certify. Accordingly, the schema is designed to admit both *equation-based* motifs (where a compact parametric form exists) and *constraint-based* motifs (where the canonical object is a list of interventional tests and predicted qualitative effects).

Abstraction maps as first-class objects. For each validated (i, τ), we will release $\alpha_{i,\tau}$ as a structured mapping from canonical variables to internal nodes/features of G_i (typically SAE features in F_i), with enough detail to execute the test suite without ambiguity. Concretely, each map entry specifies:

- *Target location:* model identifier and version hash, layer/module coordinates, activation stream, and the feature index (or a small set thereof) in the learned basis F_i. Where applicable, we include sign/scale conventions induced by dictionary normalization so that aligned features can be compared across models without hidden gauge choices.

- *Realization type:* whether the canonical variable is realized as a single feature, a sparse linear combination in code space h_i, or a subspace (e.g. an SAE feature group). When the map is non-injective (one canonical variable realized by multiple internal degrees of freedom), we record the aggregation rule used by the intervention operator.
- *Intervention operator binding:* an explicit recipe translating a canonical do-operation on Z_j into a concrete intervention $I \in \mathcal{I}$ on the chosen internal representation (e.g. resample-ablate the corresponding feature coordinate; project out a subspace; patch from a matched donor input). This includes any matching criteria for donor selection in interchange interventions.

This makes the abstraction map more than an alignment hint: it is an executable specification of how to *realize* the canonical motif within M_i. Where multiple abstraction maps pass validation, we record them as alternative realizations with their respective validation margins, rather than committing to a single "true" placement.

Validation reports as certificates with uncertainty. Each accepted instance (i, τ) will ship with a validation report that functions as a certificate relative to declared tolerances (ε, δ). The report includes: (a) the exact intervention set used for validation (including random seeds where sampling is involved), (b) the sample sizes N and any stratification over $\mathcal{D}$ versus $\mathcal{D}_{\text{OOD}}$, (c) the empirical effect mismatch statistics for each constraint (point estimates and confidence intervals), and (d) the acceptance criterion used to aggregate constraints into a single decision (e.g. max-norm mismatch, FDR-controlled multiple testing, or a composite test statistic). In addition, we include negative-control results produced by the same harness (e.g. randomized α or shuffled alignments) so that a consumer can sanity-check calibration of false acceptance rates under the declared δ.

Tooling for adding new models and re-validating motifs. We will release a reference implementation of the end-to-end pipeline sufficient for third parties to (i) train a feature basis F_i on a new model M_i, (ii) compute fingerprints $\Phi_i(f)$ and align to an existing library via π_{ij}-style assignments, (iii) propose candidate abstraction maps for known motifs, and (iv) run the validation harness to either accept an existing motif or reject it with diagnostic traces. The tooling will expose a minimal interface:

$$\texttt{forward}(x) \to (\text{activations}, Y), \qquad \texttt{intervene}(\text{activations}, I) \to \text{activations}', \qquad \texttt{readout}(\text{activation}$$

together with adapters for common transformer libraries. We will provide caching and batched execution utilities, since the practical cost is dominated by repeated forward passes under interventions. The intent is that adding a new model requires only an implementation of these hooks plus a declaration of the activation stream on which F_i is trained; all other steps (fingerprints, alignments, motif tests) are then standardized.

Benchmark suite for universality up to abstraction. We will package a benchmark suite that operationalizes the task "given $\mathcal{L}$, certify which motifs hold in a new model." The suite includes (a) a behavior family $\mathcal{B}$ specified as prompt templates with labeled readout positions and scoring functions, (b) an explicit split into $\mathcal{D}$ and multiple $\mathcal{D}_{\mathrm{OOD}}$ families (semantic shift, syntactic/length shift, and adversarially constructed perturbations), and (c) reference implementations of interventions in $\mathcal{I}$ together with admissibility checks. For compiled-transformer settings, we will also include generators with planted motifs and the corresponding ground-truth motif annotations, enabling direct measurement of alignment/motif recovery in a controlled environment. The benchmark will report not only pass/fail but also *coverage* (how many (i, τ) are validated), *margins* (distance to the ε threshold), and *cost* (effective T, N, and wall-clock).

Recommended reporting standards. To ensure comparability across studies, we will recommend that any reported motif universality claim include, at minimum:

- model provenance (architecture, training regime, checkpoint hash) and the exact definition of Y;
- SAE/transcoder configuration (activation stream, dictionary size m, sparsity regularization, reconstruction metrics) and the candidate feature selection procedure for S_i;
- intervention family $\mathcal{I}$ and the concrete operators used (including on-distributionness diagnostics for intervened activations);
- the tolerances ε and target failure probability δ, with the actual T_v and N_v used for validation and an explicit multiple-testing policy if applicable;
- a table of validated (i, τ) with margins on both $\mathcal{D}$ and each $\mathcal{D}_{\mathrm{OOD}}$, plus negative-control rejection rates under the same harness.

When alignments π_{ij} are used to transfer motif hypotheses across models, we recommend reporting cycle-consistency statistics on the high-confidence subset and sensitivity of conclusions to re-training F_i under modest hyperparameter changes. These standards are intentionally mechanistic: they force the claim “motif τ is universal up to abstraction” to be grounded in executable objects (maps, interventions, tests) rather than in qualitative similarity of circuits.

11. Limitations and failure modes. Our notion of “universality up to abstraction” is intentionally operational: we certify agreement of interventional effects on a declared readout Y under a declared intervention family $\mathcal{I}$ and input families $\mathcal{D} \cup \mathcal{D}_{\mathrm{OOD}}$, up to tolerance ε and failure probability δ. This choice yields auditable claims, but also fixes what can go wrong. In particular, a negative result may reflect a mismatch between the motif and the model, or merely a mismatch between the motif and the *available* interventions; conversely, a positive result may certify only that the motif reproduces the tested effect constraints, not that it captures all causal structure relevant to $\mathcal{B}$. We therefore enumerate limitations that arise when the assumptions implicit in our alignment, extraction, and validation steps are violated.

Non-linear or manifold features beyond sparse linear codes. Our alignment guarantees (e.g. Theorem 1) are stated in a simplified setting where internal activations admit a sparse linear representation $a_i = A_i s + \eta$ and where SAE features approximate do-interventions on coordinates of s. Empirically, many internal variables are better modeled as lying on curved manifolds, with behavior depending on directions that are only locally linear, context-dependent, or representationally entangled across layers. In such cases, a global dictionary $D_i \in \mathbb{R}^{d\times m}$ with sparse codes h_i may fail to produce a stable basis F_i, and the induced intervention operator (e.g. resample-ablation of a coordinate) may not approximate any well-defined do-operation in the underlying computation. A concrete failure mode is that two features with similar reconstruction properties yield sharply different $\Delta_i(f; I, x)$ depending on x, destroying the separation margin γ required for reliable matching. While one may attempt to address this by (i) learning multiple dictionaries conditioned on activation regimes, (ii) fitting non-linear feature maps, or (iii) using tangent-space interventions (projecting onto local principal directions), these extensions weaken identifiability and complicate the meaning of feature-level causal claims.

Redundancy, hydra effects, and self-repair under intervention. Transformers frequently exhibit redundant representations and alternative pathways implementing the same function. Under an intervention I that ablates or perturbs a feature, downstream computation may “route around” the perturbation by recruiting correlated features, a phenomenon sometimes described as hydra behavior or self-repair. In our framework this manifests as an intervention-dependent abstraction: a map $\alpha_{i,\tau}$ that passes validation for one subset of $\mathcal{I}$ may fail for another, even when both subsets are ostensibly aimed at the same canonical do-operation. The failure is not merely statistical; it reflects that the intervention changes the mechanism rather than setting a variable in an invariant SCM. This is especially acute when interventions are large in magnitude, when they break layernorm statistics, or when they implicitly induce compensatory attention patterns. Mitigation typically requires restricting to "gentle" interventions (e.g. resample-ablation with matched donors, low-rank subspace patching) and explicitly checking invariance across multiple operators intended to represent the same do-operation. Nonetheless, in highly redundant systems, a motif may be present but not *isolatable*: no intervention in $\mathcal{I}$ cleanly severs the relevant causal path without activating alternatives.

Off-distribution interventions and the problem of admissibility. All causal conclusions here are conditional on interventions being admissible in the sense of keeping internal activations approximately on-distribution. When an intervention produces an activation pattern outside the training support, the resulting output shift in Y may reflect pathological extrapolation rather than the counterfactual effect of toggling a meaningful internal variable. This can cause both false positives (a spurious but consistent effect pattern that matches SCM_τ by accident) and false negatives (true motifs obscured by intervention-induced noise). Our recommended resample-ablation and donor-matched patching operators reduce but do not eliminate this risk, since matching in a limited activation stream may fail to control for other latent state. A principled admissibility theory would specify, for each $I \in \mathcal{I}$, a constraint such as

$$d\big(\mathrm{Law}(a_i \mid x), \mathrm{Law}(I(a_i) \mid x)\big) \le \kappa$$

for an appropriate distance d and tolerance κ, and would propagate this to bounds on causal-effect estimation error. We do not yet have such a theory in general, so validation should be read as an empirical certificate under a chosen admissibility diagnostic rather than as an unconditional causal guarantee.

Partial observability of internal state and limited intervention access. Even in a white-box setting, we typically intervene on a selected activation stream (e.g. residual stream features) and treat other internal quantities as fixed or endogenous. However, the effective internal state relevant to Y includes attention scores, key/value caches, normalization statistics, and sometimes stochastic components (sampling, dropout during training-time analysis, or tool-use randomness in agentic settings). If the true causal parents of the canonical variables Z_j are partly unobserved or un-intervenable, then abstraction maps $\alpha_{i,\tau}$ may be non-identifiable: multiple distinct internal structures can satisfy the same finite set of tested interventional constraints. This is not merely an inconvenience; it bounds what we can claim. Our certificates establish that the *tested* interventional distributions of Y are consistent with SCM_τ under $\alpha_{i,\tau}$, but they need not pin down a unique internal causal story. In practice, we expect to report equivalence classes of realizations (multiple passing maps) and to treat failure to find a passing map as inconclusive when intervention access is too narrow.

Statistical and multiple-testing failure modes in validation. Validation aggregates many estimated effects into a single accept/reject decision. When the number of tested constraints is large, naive thresholds can silently inflate false acceptance. Conversely, conservative corrections can yield false rejections, especially under heterogeneous variance across $\mathcal{D}$ and $\mathcal{D}_{\mathrm{OOD}}$. Although Theorem 3 gives sample requirements $N = \tilde{O}(\varepsilon^{-2}\log(1/\delta))$ for individual constraints under concentration assumptions, the constants can be large in practice, and model outputs can be heavy-tailed under OOD prompts. Moreover, the choice of effect summary (e.g. expected logit shift versus KL divergence) can change sensitivity. A robust implementation therefore requires explicit multiple-testing policy (e.g. FDR control) and negative controls (randomized α, shuffled π_{ij}) to estimate calibration. These measures reduce, but do not eliminate, the risk that a library grows by accumulating borderline motifs that pass due to correlated test statistics.

Compute and scaling constraints. The pipeline is compute-intensive: training SAEs per model, computing fingerprints for $|S_i|$ features across T interventions with N samples, performing pairwise (or hub-based) alignments, and then validating candidate motifs with additional $T_v N_v$ queries. The complexity terms $O(r \cdot |S| \cdot TN)$ for fingerprints and $O(r^2 \cdot |S|^3)$ for exact assignment become prohibitive when r and $|S|$ are large, even before motif extraction. Approximate matching, locality-sensitive hashing on fingerprints, and restricting to a small high-confidence subset can help, but may systematically bias the library toward frequent or easy-to-intervene-on features, reducing coverage of rarer but behaviorally important motifs. Furthermore, the fixed-k tractability story does not remove the practical cost of candidate generation: without strong pruning heuristics, the search space over $|S|^k$ remains large even for modest k. Thus our approach is likely to be most reliable when applied to carefully chosen behaviors $\mathcal{B}$ and narrow activation streams where interventions are cheap and interpretable.

Dual-use and release-risk considerations. A motif library $\mathcal{L}$ with executable abstraction maps $\alpha_{i,\tau}$ can lower the cost of targeted model editing, behavioral steering, or bypassing safety mechanisms. The same intervention harness that enables scientific reproducibility can, if misused, enable systematic discovery of control levers for undesirable behaviors (e.g. eliciting disallowed content, manipulating tool calls, or amplifying persuasion). This risk is accentuated when motifs correspond to high-level control signals (refusal, deception, goal persistence) rather than benign subroutines (copying, retrieval). We therefore view unrestricted release of model-specific maps and intervention recipes as a policy decision rather than an automatic consequence of scientific publication. Practical mitigations include releasing (i) canonical motifs SCM_τ and validation statistics without fine-grained model coordinates, (ii) coarse-grained or rate-limited tooling interfaces, (iii) red-teamed subsets of motifs under controlled access, and (iv) audit logs for intervention execution in shared environments. Any such restriction, however, reduces external verifiability; this tradeoff is intrinsic and should be made explicit in claims of universality.

Summary. In short, our certificates are strongest when (i) feature bases approximate identifiable causal variables, (ii) interventions are admissible and do not trigger self-repair, (iii) the relevant internal state is sufficiently observable, and (iv) the tested constraints are appropriately calibrated under finite samples. Where these conditions fail, the framework remains useful as a disciplined empirical protocol, but its conclusions should be read as conditional statements relative to $(\mathcal{I}, Y, \mathcal{D} \cup \mathcal{D}_{\mathrm{OOD}}, \varepsilon, \delta)$, not as absolute mechanistic identifications.

12. Conclusion and future work. We have formulated an operational notion of *universality up to abstraction* for mechanistic motifs across a family of transformer models, and we have given a concrete pipeline (CMML) that (i) learns sparse feature bases, (ii) aligns features across models using interventional fingerprints, (iii) proposes candidate motifs of bounded size k, (iv) canonicalizes these into a library $\mathcal{L}$ of motif types τ with associated canonical SCM_τ, and (v) certifies ε-agreement of interventional effects on a declared readout Y over $\mathcal{D} \cup \mathcal{D}_{\mathrm{OOD}}$. The main value of this framing is that it turns informal claims of "shared circuits" into a finite family of testable interventional constraints, together with explicit dependence on $(\mathcal{I}, Y, \varepsilon, \delta)$. Theorems 1–5 delineate a tractable regime (identifiable sparse features; bounded motif size) and clarify where interventions are provably necessary and where computational hardness is unavoidable. We view the present work as a base layer upon which several extensions—multimodality, agentic tool use, temporal motifs, and adversarial robustness—can be developed without abandoning the core requirement of auditable causal tests.

Multimodal models: shared motifs across heterogeneous streams. A direct next step is to extend the feature/alignment/validation story to multimodal transformers whose internal state mixes text, vision, audio, or other modalities. In such settings, a motif τ may be realized by variables that are not confined to a single activation stream, and an abstraction map $\alpha_{i,\tau}$ may necessarily span multiple layers and modalities. Technically, this suggests replacing a single dictionary $D_i \in \mathbb{R}^{d \times m}$ by a *typed* collection $\{D_i^{(\text{text})}, D_i^{(\text{vision})}, \ldots\}$ together with cross-modal linking constraints that specify which feature types are allowed to serve as images of a given canonical variable Z_j. One can then define fingerprints $\Phi_i(f)$ with interventions that are modality-appropriate (e.g. patching a visual token subspace, ablation of a cross-attention head, or resample-ablation in the residual stream) while preserving the same downstream readout Y. An open theoretical question is whether analogues of Theorem 1 can be proved when the latent sparse code s has block structure and the mixing maps A_i are modality-specific and partially shared; a plausible route is to impose structured incoherence assumptions that prevent arbitrary rotations across modalities while still allowing within-modality symmetries. Empirically, we expect that many canonical motifs (retrieval, copy-suppression, refusal gating) will admit both unimodal and multimodal realizations, and a library $\mathcal{L}$ that explicitly records modality-typed variables may clarify when "the same behavior" is implemented by genuinely shared subroutines versus modality-contingent special cases.

Agentic and tool-using systems: motifs over interaction graphs. When Y denotes an action (tool call, parameter choice, or refusal) produced by an agentic model interacting with an environment, the relevant causal graph is no longer confined to a single forward pass. Instead, the system includes external state (tool outputs, memory buffers, scratchpads) and the model's own action-conditioned future inputs. We can still remain within the present framework by treating the entire closed-loop computation as an extended causal graph G_i^{loop} with additional nodes for environment variables and tool I/O. The intervention family $\mathcal{I}$ must then include operations such as (i) patching internal activations at specific time steps, (ii) intervening on tool outputs (counterfactual tool responses), and (iii) freezing or replacing memory states. A motif type τ in this setting may describe, for example, a canonical SCM_τ in which a latent goal variable influences both a plan variable and a termination decision, with observed tool calls mediating information acquisition. Validation would require interchange interventions that hold the environment fixed while swapping internal variables according to $\alpha_{i,\tau}$, in a manner analogous to interchange intervention tests but now indexed by time and conditioned on the history. A key methodological issue is to ensure that the test suite distinguishes mechanistic motifs from policy-level invariances; in particular, we will want tests that separate "the agent chooses tool t because it inferred fact q" from "the agent chooses tool t whenever the prompt matches a template," even if both patterns produce similar observational behavior.

Dynamic motifs over trajectories: causal templates for computation-in-time. Even without external tools, many behaviors of interest (multi-step reasoning, iterative refinement, self-correction) are naturally temporal. We therefore expect the appropriate canonical object to be a *dynamic* SCM or causal state-space motif. One minimal formalization is to index canonical variables by time $t \in \{1, \ldots, T_x\}$, writing $Z_j^{(t)}$ and allowing edges across time steps. A dynamic motif of bounded *per-step* size k can then be represented by structural equations

$$Z^{(t+1)} = g_\tau\big(Z^{(t)}, U^{(t)}\big), \qquad Y^{(t)} = h_\tau\big(Z^{(t)}\big),$$

where $U^{(t)}$ are exogenous noises and g_τ, h_τ are canonical mechanisms. An abstraction map $\alpha_{i,\tau}$ now maps each $Z_j^{(t)}$ to a time-localized internal node (or feature) in G_i, such as an SAE feature at layer $\ell(t)$ and token position $p(t)$. The intervention family must support temporally targeted operations, and fingerprints must be defined over interventions that perturb a variable at a given time and measure its downstream effect on future readouts. The main conceptual benefit is that we can state universality claims not merely about static subcircuits, but about canonical *update laws* that recur across positions and layers. The main technical obstacle is combinatorial: even if each time slice has size k, the unrolled motif has size kT_x. One promising direction is to validate *local* constraints that are sufficient to certify a repeating template (e.g. time-homogeneous mechanisms) rather than validating the entire unrolled graph.

Adversarially robust universality testing: worst-case prompts and interventions. Our present guarantees are phrased in terms of average-case estimation over $\mathcal{D}$ (and a chosen $\mathcal{D}_{\mathrm{OOD}}$). For safety-relevant behaviors, we often require stronger statements: the motif should predict interventional effects not only on typical inputs, but also under adversarially chosen prompts, contexts, or tool outputs. We can express such a requirement by replacing expected discrepancies with a supremum over an adversary class $\mathcal{A}$:

$$\sup_{x \in \mathcal{A}} \sup_{I \in \mathcal{I}_\tau} d\big(\mathrm{Law}(Y \mid \mathrm{do}_\alpha(I), x)\,,\ \mathrm{Law}(Y_\tau \mid \mathrm{do}(I), x)\big) \le \varepsilon,$$

where $\mathcal{I}_\tau$ denotes canonical interventions on SCM_τ and $\mathrm{do}_\alpha(I)$ denotes the corresponding realized intervention via α. Achieving such a bound would require (i) an explicit adversary model (prompt-space constraints, norm bounds in embedding space, or discrete search procedures), and (ii) a testing protocol that either covers $\mathcal{A}$ by a finite ε-net or uses adaptive search to find counterexamples. This suggests a synthesis between causal abstraction testing and adversarial evaluation: we can treat the validation step as a falsification game in which the tester searches for (x, I) that maximizes mismatch. Theoretical work here would aim to relate the complexity of this search to capacity measures of the motif and to regularity properties of the model's response under admissible interventions.

Compositional libraries, hierarchy, and minimality. The library objective in CMML is a set-cover criterion over validated instances (i, τ), which already implies an implicit notion of parsimony. A further goal is to construct libraries that are compositional: complex behaviors should be explained by wiring together simpler motifs with explicitly represented interfaces. Concretely, we can treat motifs as typed components with input/output variables, and define a composition operator that produces a larger SCM by identifying interface variables (or by adding coupling equations). The resulting representation would allow us to express, for example, that a tool-use policy motif composes a retrieval motif with an action-selection motif. On the algorithmic side, compositionality suggests searching for motifs that not only explain Y directly but also serve as reusable subcomponents across behaviors $\mathcal{B}$. On the theoretical side, it raises a minimality question: under what conditions is a canonical decomposition identifiable from interventional fingerprints, and how does bounded-size tractability interact with hierarchical reuse? While unrestricted structure learning remains hard, we expect that constraining the motif grammar (bounded arity, typed interfaces, limited depth) may yield tractable special cases analogous to the fixed-k regime of Theorem 3.

Closing remark. We emphasize that each proposed extension retains the same core discipline: universality claims should be phrased as equivalence of interventional effects on declared readouts under declared intervention families, with explicit tolerances and failure probabilities. The central open problem is to broaden the class of systems and behaviors for which we can construct such certificates without relying on fragile assumptions, while maintaining computational and statistical feasibility at scale.