Validating Causal Abstraction Metrics on Simulated Complex Systems

Anonymous Author(s)
Affiliation
Address
email

Abstract

A central goal of science is to produce valid explanations of complex systems: high-level causal accounts that faithfully reflect the behavior of lower-level mechanisms. Yet no consensus exists on how to measure whether a proposed high-level explanation is actually valid. We introduce a benchmark of ten complex systems spanning both discrete and continuous state spaces, as well as static and dynamical regimes, each equipped with consensual ground-truth causal explanations and invalid contrastive conditions. Within a unified causal abstraction framework, we systematically evaluate over thirty candidate metrics drawn from observational, functional, information-theoretic, and causal families. Our results show that only the latter reliably discriminates valid from invalid abstractions, and only when incorporating faithfulness testing over unmapped variables. Building on these findings, we introduce the Causal Abstraction Error (CAE), a continuous validity metric with an explicit faithfulness test, which passes all discrimination tests across every system and converges with as few as 30 sampled interventions. We offer it as a general-purpose metric for the discovery and validation of high-level explanations.

1 Introduction

A central goal of science is to produce **explanations**: not merely descriptions or predictions, but accounts of *why* phenomena occur [Hempel, 1965, Woodward, 2004]. While the philosophy of science has long debated what explanations are, a working consensus has emerged around a few core desiderata: good scientific explanations should be causally informative [Woodward, 2004, Pearl, 2009], parsimonious [Kitcher, 1962, Salmon, 1984, Batterman and Rice, 2014], and appropriately scoped to the context and level of description at which the target phenomenon is best characterized [Lombrozo, 2006, Potochnik, 2017]. Finding good explanations is particularly difficult when the system under study exhibits what Warren Weaver famously called *organized complexity* [Weaver, 1991]: systems with many interacting parts that are neither so disordered as to yield to statistical averaging, nor so simple as to admit direct analytic treatment. The appropriate high-level variables for explanation must be discovered, and the class of functions that legitimately aggregate low-level quantities into those variables is itself a subject of inquiry [Hoel et al., 2013, Potochnik, 2017].

This raises a fundamental challenge: what are useful high-level variables, and how should they be defined from lower-level quantities? Different fields studying different classes of complex systems have developed their own candidate answers, such as firing rates and population codes in systems neuroscience [Cunningham and Yu, 2014], or species abundance and trophic levels in ecology [Loreau, 2010] However, a unified cross-disciplinary methodology for evaluating high-level explanations of complex systems remains elusive. A sobering illustration of why comes from Jonas and Kording [2017], who applied the standard causal and statistical toolkit of neuroscience to a fully observable microprocessor: a relatively simple system engineered for modular, hierarchical organization. In doing so, they failed to recover meaningful high-level properties.

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

The recent success of artificial intelligence seems to offer a way forward. Modern AI systems are themselves structurally complex computational systems, yet they are fully observable and perfectly manipulable, a rare property in the natural sciences. This privileged epistemic access makes AI systems a methodological laboratory for the science of complex systems [Holtzman et al., 2025], to develop and test explanation discovery procedure in simplified settings. Yet, AI interpretability has also proven sobering, with a large tendency to produce false positive findings and explanations that are unstable and not generalizable [Méloux et al., 2025a].

A general metric for evaluating proposed high-level explanations of low-level complex systems is therefore a prerequisite to move forward: without it, the discovery and evaluation of candidate cannot be done rigorously, and progress is hard to measure or transfer across domains. Yet, no such agreed-upon, general metric exists. A growing body of work has begun addressing this gap through the lens of **causal abstraction** [Beckers and Halpern, 2019], the idea that a high-level causal model is a valid explanation of a low-level system if interventions at the high level correspond faithfully to interventions at the low level under a postulated abstraction map. This framework is conceptually appealing and promising, but existing implementations differ in important ways: in whether they treat abstraction as exact or approximate [Beckers et al., 2020], and in how they handle distributions over inputs or over interventions [Geiger et al., 2021, 2025]. Current concrete metric implementations developed in the field of AI interpretability have also been shown to suffer from the global issues of striking false positives and non-generalizable findings [Sutter et al., 2025, Méloux et al., 2025a].

In this work, we set out to systematically test, compare, and benchmark a wide range of proposed explanation metrics, observational, information-theoretic, symbolic, and causal, across a diverse collection of idealized complex systems with known, consensual high-level explanations. Rather than confining to a single domain, we deliberately span systems that vary across multiple dimensions: (i) with discrete versus continuous state spaces, (ii) with static versus temporal dynamics, (iii) with disordered versus structured complexity (in Weaver's sense). For each system, we manually construct the alignment function mapping low-level variables to high-level ones, implement all pairs within a shared conceptual formalism (causal abstraction) and a unified programmatic API supporting both observational and interventional queries. Finally, we handcraft contrastive perturbations, modifications to either the low-level or high-level model that render the proposed explanation invalid, allowing us to test whether each candidate metric correctly recognizes broken explanations as well as valid ones. This is an empirical test of the practical operationalization of existing conceptual ideas.

Our results provide strong support for the causal abstraction research program, but only for a specific distributional variants. We further introduce the Causal Abstraction Error (CAE), an error measure that incorporates a principled treatment of distributional inputs and interventions, a built-in faithfulness measurement, and produces a real-valued degree-of-violation score. The CAE passes all contrastive tests across every system class we examine and outperforms related causal abstraction variants.

2 Metrics of Explanation Validity

Following the framework of Pearl [2009] and Beckers and Halpern [2019], we model the system under study as a *structural causal model* (SCM) to implement the semantics of interventions and counterfactuals. Dyer et al. [2024] previously established that simulated complex systems belong to this class.

Definition 2.1 (Structural Causal Model) *A structural causal model is a tuple* $\mathcal{M} = \langle \mathcal{U}, \mathcal{V}, \mathcal{R}, \mathcal{F}, P_{\mathcal{U}} \rangle$*, where* $\mathcal{U}$ *is a set of* exogenous *variables with joint distribution* $P_{\mathcal{U}}$*;* $\mathcal{V}$ *is a set of* endogenous *variables with ranges* $\mathcal{R} = \{\mathcal{R}_V\}_{V \in \mathcal{V}}$*; and* $\mathcal{F} = \{f_V\}_{V \in \mathcal{V}}$ *is a set of* structural equations*, each determining* V *from its direct causes* $\mathrm{Pa}(V) \subseteq \mathcal{V}$ *and a noise term* $\mathcal{U}_V \subseteq \mathcal{U}$*.*

A *computational explanation* is a surrogate model $\mathcal{E}$ taken from some admissible class $\mathfrak{E}$, intended to account for the behavior of $\mathcal{M}$ at a relevant level of abstraction. The central question studied in this paper is: *Under what conditions is* $\mathcal{E}$ *a valid computational explanation of* $\mathcal{M}$ *under* $P_{\mathcal{U}}$*?* Radically different answers to this question have been proposed. We briefly discuss here the ones we evaluate and offer a more in-depth description in Appendix B.

Observational validity. The most elementary criterion is *observational equivalence*: $\mathcal{E}$ should reproduce the measurable outputs of $\mathcal{M}$ under the *natural* input distribution. For deterministic settings, this is operationalized via pointwise discrepancy metrics, including MSE, RMSE, L^2, and their nor-

malized variant NMSE, together with the coefficient of determination R^2 [Koza, 1994, Ljung, 1999]. Complexity-regularized variants such as AIC [Akaike, 2003], BIC [Schwarz, 1978], MDL [Mangan et al., 2017], and Mallow's C_p [Lee and Ghosh, 2009] penalize explanation complexity to prevent overfitting. When $\mathcal{M}$ is stochastic, distributional similarity is measured via measures like KL divergence, the symmetric JS divergence, or the kernel-based MMD [Gretton et al., 2012]; HSIC [Gretton et al., 2005] can further tests whether residuals of the explanation are independent of the input, probing global agreement. For dynamical systems, validity must hold over entire trajectories: we evaluate trajectory MSE, dynamic time warping (DTW; Sakoe and Chiba, 1978), temporal autocorrelation matching, and spectral analysis [Percival and Walden, 1993]. Finally, we also consider symbolic regression methods like the SINDy evaluator [Brunton et al., 2016], which scores an explanation by the residual of its governing equations on the observed state derivatives. All observational criteria share a fundamental limitation called *equifinality* [Valogianni and Padmanabhan, 2022, Collins et al., 2024]: distinct mechanisms can be observationally indistinguishable, yielding explanations that often do not generalize [Ghorbani et al., 2019, Kindermans et al., 2019, Méloux et al., 2025b].

Functional validity. Functional criteria require that $\mathcal{E}$ additionally reproduce the *input–output response profile* of $\mathcal{M}$. *Variance decomposition* via ANOVA or Sobol sensitivity indices [Sobol', 2001] decomposes output variance into per-input contributions and interactions, with validity requiring agreement between the indices of the high-level model $\mathcal{E}$ and the low-level model $\mathcal{M}$. At the local level, the infidelity metric [Yeh et al., 2019], adapted to the two-model setting, measures whether $\mathcal{E}$ and $\mathcal{M}$ agree on their per-input attribution vectors. This replaces the original single-model attribution formulation with a two-model sensitivity comparison, using $\mathcal{E}$ itself as the local linear approximation. LIME [Ribeiro et al., 2016] and SHAP [Lundberg and Lee, 2017] can be understood as proxies for this notion. Globally, relational fidelity [Collins et al., 2024] requires consistency of output differences across all input pairs, ensuring that $\mathcal{E}$ preserves the functional geometry of $\mathcal{M}$. However, functional criteria still remain agnostic to internal mechanisms.

Information-theoretic and representational validity. A third class of criteria requires that $\mathcal{E}$ reproduce the representational structure of $\mathcal{M}$. This can be measured by representational similarity analysis [Kriegeskorte et al., 2008]. Similarly, *probing accuracy* [Alain and Bengio, 2018] measures whether the variables posited by $\mathcal{E}$ are linearly decodable from the internal states of $\mathcal{M}$. The *information bottleneck* (IB) Lagrangian [Tishby et al., 2000] characterizes valid representations as those achieving an optimal compression–relevance trade-off. *Complexity shift* [Zenil et al., 2019] measures whether $\mathcal{E}$ induces comparable transformations under perturbation, assessed via Kolmogorov complexity. These criteria capture *what* information is present, but not *how* it is causally transformed [Elazar et al., 2021, Lasri et al., 2022, Teney et al., 2022].

Causal and interventional validity. The strictest criteria require $\mathcal{E}$ to correctly reproduce $\mathcal{M}$'s behavior under *active interventions*. *Symbion* [Gritti et al., 2020], focused on program explanation, scores explanations by the fraction of execution steps at which symbolic and concrete states agree, providing a tractable formal-verification proxy. The *causal sensitivity index* and the related *structural deviation* metric [Katende, 2025] measure how much the interventional consistency score between $\mathcal{E}$ and $\mathcal{M}$ changes when each parameter of $\mathcal{E}$ is zeroed-out or perturbed by a fixed fraction, respectively. Mechanistic interpretability [Olah et al., 2020] seeks to recover the sparse *circuits* [Vig et al., 2020, Meng et al., 2022] implementing a given behavior, with validity defined by output agreement under computational interventions. Finally, ϵ-machines [Crutchfield and Young, 1989, Shalizi and Crutchfield, 2001] define the canonical valid explanation as the minimal sufficient statistic for predicting $\mathcal{M}$ the next steps of a dynamical system, potentially under interventions. The *dynamic causal consistency* (DCC) criterion [Grosse-Wentrup et al., 2024] offers a practical implementation of this idea.

2.1 Zooming in on Causal Abstraction

Causal abstraction aims to provide a mathematical foundation for determining when high-level models can safely be considered as causal proxies for low-level ones [Chalupka et al., 2016, Rubenstein et al., 2017, Beckers and Halpern, 2019]. It is the central conceptual framework in this work. We explore its formal definitions and the ongoing debates regarding its precise framing.

Overview. We denote $\mathcal{M}$ as the micro- (or low-level) model and $\mathcal{E}$ as the macro- (or high-level) model. $\mathcal{E}$ is now also defined as an SCM. An abstraction between these models is defined by a

mapping $\tau : \mathcal{M} \to \mathcal{E}$ which specifies how low-level computations in $\mathcal{M}$ translate to high-level ones in $\mathcal{E}$. Zennaro [2022] reviews properties of existing definitions of valid abstractions and their evolution [Rubenstein et al., 2017, Beckers and Halpern, 2019, Beckers et al., 2020, Rischel, 2020, Rischel and Weichwald, 2021, Otsuka and Saigo, 2022]. Two key categories of properties are relevant for us: **formal structural properties** of the mapping itself and **consistency properties**, which ensure that reasoning with either the micro- or the macro-model leads to causally consistent outputs.

Structural Properties. Structural properties define constraints on the abstraction mapping τ, determining which types of abstraction maps are *permissible*. Minimal but sufficient properties to model the complex systems we consider are as follows:

1. Surjective disjoint coarse-graining of variables: The variables in the micro-model should be coarse-grained into variables of the macro-model through a surjective mapping a, called the **coarse-graining map**. This map assigns groups of micro-variables to macro-variables. If the coarse-graining is disjoint, meaning that the same micro-variables cannot compute different macro-variables, we speak of **constructive causal abstraction** [Beckers and Halpern, 2019], which is widely used [Zennaro et al., 2023a, Kekić et al., 2023, Zhu et al., 2024]. The surjectivity of a ensures that each macro-variable is derived from some micro-variables.

2. Surjective value mapping: The coarse-graining map only associates variables (structural). We must also specify how micro-states map to macro-states (contents). This is achieved via a set of **value maps** $\tau_X : a^{-1}(X) \to X$, which determines, for each macro-variable X, how its state arises from the states of its associated micro-variables. We require each value map to be surjective to guarantee that every possible macro-state can be realized from some micro-states.

Finally, the abstraction map τ consists of both a surjective (disjoint) coarse-graining map a and the set of all surjective value maps (τ_X).

Consistency Property. For a given a macro-intervention $do(X = x)$, written do_ν, multiple compatible micro-level interventions $do(a^{-1}(X) = \tau_X^{-1}(x))$ exist, since τ_X is not typically injective (multiple micro-states may realize the same macro-state). Let do_μ be one such micro-intervention. Then, the consistency property requires that there must exist a noise-variable map $\tau_u : \mathcal{U}_\mathcal{M} \to \mathcal{U}_\mathcal{E}$ such that for every compatible pair of interventions (ν, μ) and every $u \in \mathcal{U}_\mathcal{M}$, $\mathcal{E}(\mathcal{V}_\mathcal{E}|\tau_u(u), \mathrm{do}_\nu) = \tau(\mathcal{M}\,(\mathcal{V}_\mathcal{M}|u, \mathrm{do}_\mu))$. This pointwise counterfactual consistency requires that both models coincide for every exogenous noise realization. When $\mathcal{E}$ is deterministic, this equation reduces to $\mathcal{E}(\mathcal{V}_\mathcal{E} \mid \mathrm{do}_\nu) = \tau(\mathcal{M}(\mathcal{V}_\mathcal{M} \mid u, \mathrm{do}_\mu))$, which is the primary criterion implemented in this work. This typically visualized via a commuting diagram [Beckers and Halpern, 2019].

Measuring Consistency Errors. From the consistency requirement, it is possible to derive an error measure, quantifying how far a macro-model (and its associated abstraction map) is from being a valid causal abstraction of a given micro-model. A general formulation of such a measure is given by:

$$E_\tau(\nu, \mu) = D\big(\mathcal{E}(Y \mid \mathrm{do}_\nu),\ \tau_Y\big(\mathcal{M}(a^{-1}(Y) \mid \mathrm{do}_\mu)\big)\big)\,,$$

where D denotes a pointwise dissimilarity metric (e.g., MSE) in the deterministic setting, or a distributional divergence for stochastic model. The function $E_\tau(\nu, \mu)$ quantifies the degree of consistency violation for a given macro-intervention ν and one associated micro-intervention μ. Different approaches have been proposed to define a *global* consistency error: Zennaro et al. [2023a] adopts the worst-case error over (ν, μ), defining $E_\tau = \sup E_{\nu,\mu}$ and using the Jensen-Shannon divergence as the choice for D. Then, Zhu et al. [2024] and Kekić et al. [2023] employ the Kullback-Leibler divergence and compute the expectation over all single-variable interventions, facilitating sample-based approximations. A different approach is intervention interchange accuracy (IIA; Geiger et al., 2022a), a metric developed especially for neural networks. IIA fits into the consistency framework above but checks it only for output variables, not for intermediate ones, and does not require a complete abstraction map. Consequently, even perfect alignment under IIA does not guarantee that the macro-model is a causal abstraction of the micro-model in the general sense, since intermediate variables may still violate consistency [Méloux et al., 2025a, Sutter et al., 2025].

Faithfulness. The standard definition of causal abstraction does not require all low-level variables in $\mathcal{V}_\mathcal{M}$ to be used. This is natural for AI interpretability, where explanations often posit that only a small subset of a neural network is relevant to a behavior [Olah et al., 2020, Conmy et al., 2023],

and similar sparsity assumptions appear in biological systems [Rochefort et al., 2009, Meunier et al., 2010]. Yet, declaring excluded components causally irrelevant is a strong empirical claim, that must be tested. In circuit discovery, this is captured by the **faithfulness** criterion: unmapped neurons should be intervened on to verify that they do not influence the proposed circuit computation [Hanna et al., 2024]. Existing operationalizations of causal abstraction largely take the relevant variable set as modeler-provided, rather than testing whether excluded variables are genuinely inert [Zennaro et al., 2023a, Kekić et al., 2023, Zhu et al., 2024]. Enforcing faithfulness would reject explanations that incorrectly assume away causal influence from omitted components.

3 The Constructive Abstraction Error

Building upon existing causal alignment error measures [Zennaro et al., 2023a, Kekić et al., 2023, Zhu et al., 2024], we define a new metric that explicitly captures faithfulness in causal abstraction. Let Φ denote the set of *phi variables*: micro-variables that are neither mapped to any macro-variable by a nor declared as causally redundant internal variables of $\mathcal{M}$. We define two variants of CAE that differ in the direction of sampling through the consistency diagram:

$$\text{CAE}_{\downarrow}(\mathcal{M}, \mathcal{E}, \Phi, \tau) = \mathbb{E}_{\nu \sim P_I}\, \mathbb{E}_{\mu \sim P_\nu}\, D\big(\mathcal{E}(Y \mid \text{do}_\nu),\ \tau_Y\big(\mathcal{M}(a^{-1}(Y) \mid \text{do}_\mu)\big)\big)\,, \tag{1}$$

$$\text{CAE}_{\uparrow}(\mathcal{M}, \mathcal{E}, \Phi, \tau) = \mathbb{E}_{\mu \sim P_I}\, D\big(\mathcal{E}(Y \mid \text{do}_{\tau(\mu)}),\ \tau_Y\big(\mathcal{M}(a^{-1}(Y) \mid \text{do}_\mu)\big)\big)\,. \tag{2}$$

where D is a pointwise dissimilarity metric (e.g., MSE) in the deterministic setting, or a distributional divergence (JSD, KL, MMD) for stochastic models; P_I is a distribution over interventions on macro-variables extended by Φ; and for a macro-intervention $\nu = \text{do}(X = x)$, P_ν denotes the distribution over compatible micro-interventions $\mu = \text{do}(a^{-1}(X) = \mathbf{z})$ with $\mathbf{z} \sim \tau_X^{-1}(x)$. In $\text{CAE}_{\downarrow}$ (top-down), macro-interventions are grounded to compatible micro-interventions via τ_X^{-1}; in $\text{CAE}_{\uparrow}$ (bottom-up), micro-interventions are drawn directly and lifted to the macro level via τ_X. Both arguments of D lie in the macro-variable range $\mathcal{R}_Y$, avoiding ill-posedness from τ_Y being non-injective. Faithfulness is integrated in both variants: P_I includes interventions on Φ-variables, which the proposed explanation claims are causally inert. Finally, we also define non-faithful variants: $\text{CAE}_{\uparrow\text{NF}}$ and $\text{CAE}_{\downarrow\text{NF}}$.

Relation to Zennaro et al. [2023b]. That work defines four related metrics; ISIL and IC are the closest to $\text{CAE}_{\downarrow}$ and $\text{CAE}_{\uparrow}$ respectively. Beyond the faithfulness consideration, three differences are worth noting. First, their metrics aggregate by supremum; ours use expectation, which is more tractable to estimate by sampling. Second, their framework is restricted to finite variable domains, whereas CAE handles continuous value maps directly. Third, their metrics intervene on a single variable at a time; Zennaro et al. [2023b] explicitly notes that this prevents compositionality of valid abstractions from being preserved. CAE supports joint multi-node interventions, restoring this property and probing interaction terms that single-variable interventions miss.

3.1 Practical Estimation

Both variants of CAE use expectations amenable to Monte Carlo estimation. In our setting, $\mathcal{E}$ is always deterministic, so there is no $\mathcal{U}_\mathcal{E}$ to sample; the role of the exogenous noise $u \in \mathcal{U}_\mathcal{M}$ is played by the intervention values drawn from P_I, which are passed identically to both models in each Monte Carlo sample. We discuss further key design choices below.

Sampling macro-interventions ($\text{CAE}_{\downarrow}$). We draw N interventions $\{\nu^{(i)}\}_{i=1}^N \sim P_I$. To probe interaction terms, we support *multi-node interventions* by sampling a random subset of variables $\mathcal{S} \subseteq \mathcal{V}^M$ of size $|\mathcal{S}| \sim \text{Uniform}\{1, \ldots, |\mathcal{V}^M|\}$ and jointly drawing intervention values $\mathbf{x}_\mathcal{S}$, where each component x_V for $V \in \mathcal{S}$ is drawn independently and uniformly from the domain $\mathcal{R}_V$ of the corresponding macro-variable. When a geometric domain is specified for a variable, intervention values are drawn by sampling directly from that subspace.

Sampling compatible micro-interventions ($\text{CAE}_{\downarrow}$). For each macro-intervention ν on variables $\mathcal{S}$, we sample a compatible micro-intervention by applying the inverse value map to each intervened variable: for $V \in \mathcal{S}$ with intervened value x_V, we compute the preimage $\tau_V^{-1}(x_V)$ and sample a micro-state from it. In discrete settings, this corresponds to the Moore-Penrose pseudo-inverse, giving uniform probability mass to all micro-states mapping to x_V [Zennaro et al., 2023a]. For continuous

systems, where the preimage is a subspace of the micro-state space, we sample uniformly from that subspace.

Sampling micro-interventions (CAE$_\uparrow$)**.** We draw N micro-interventions $\{\mu^{(i)}\}_{i=1}^{N} \sim P_I$ directly, using the same multi-node scheme as above but operating over micro-variable domains. The corresponding macro-intervention is obtained by applying τ_X to the intervened micro-state.

Testing faithfulness. We test faithfulness by augmenting the intervention pool with a dummy Φ-sentinel alongside the real macro-variables. At each sample, Φ is included in the selected intervention targets with a configurable probability. When selected, independent noise is injected into every $\phi \in \Phi$ simultaneously before executing $\mathcal{M}$: Gaussian noise with standard deviation σ_ϕ for continuous variables, uniform $\{-1, 0, 1\}$ perturbations for integer variables, and random bit-flips for Boolean variables. If the abstracted output shifts relative to the unperturbed run, the consistency error increases proportionally. The parameter σ_ϕ controls the sensitivity of the test.

4 A Benchmark of Simulated Complex Systems

A central obstacle in evaluating explanation-validity metrics is the lack of naturally occurring systems paired with objective ground-truth explanations. Human judgment is an inadequate substitute, since users often prefer convincing explanations even when they are incomplete or misleading [Ancona et al., 2019, Herman, 2017, Petsiuk et al., 2018]. Existing benchmarks therefore typically evaluate desirable properties of explanation methods, such as sparsity, stability, localization, or causal robustness, rather than whether an explanation recovers a known valid high-level causal account [Karvonen et al., 2025, Poché et al., 2025, Liu et al., 2021, Huang et al., 2024].

We address this by constructing a benchmark of idealized complex systems whose high-level explanations are known and analytically justified at the relevant abstraction level. This enables two basic tests: a metric should assign near-zero error to correct explanations, and should detect controlled failures produced by invalid explanations. A key contribution of this work is the manual alignment of these heterogeneous systems with their high-level explanations integrated within a common framework. The benchmark spans many systems chosen to cover different abstraction types, yet all systems admit clean causal-abstraction representations and, more strongly, *constructive* causal abstractions. This provides a reusable resource and a nontrivial empirical demonstration of the breadth of the causal abstraction program. For each system, we provide the high-level model, low-level implementation, abstraction map, and manually crafted invalid contrastive scenarios. Full details are given in Appendix D; Table 1 summarizes the systems.

Table 1: **Benchmark systems.** Each row is a manually aligned ground-truth abstraction pair $(\mathcal{M}, \mathcal{E}, \tau)$. The benchmark spans discrete (D) and continuous (C) state spaces, static and dynamical regimes, and local, spatial, and global constructive maps. Rows shaded in light blue are dynamical systems. For every pair, we also provide invalid contrastive conditions by perturbing the model, the map, or the validity regime; full details are in Appendix D.

System	**Low level $\mathcal{M}$**	**High level $\mathcal{E}$**	**Constructive abstraction map τ**	
			Type	**Structure**
Gas	Lennard–Jones particles	Ideal gas / VdW law	C $\to$ C	global / aggregation
Predator–prey	Agent-based model	Lotka–Volterra ODE	D $\to$ C	global / aggregation
Heat equation 1D	Brownian particles	1D heat equation	C $\to$ C	spatial / binning
Heat equation 2D	Phonon/lattice simulation	2D heat equation	C $\to$ C	spatial / smoothing
Ising model	MD/Monte-Carlo simulator	Nearest-neighbor Ising model	C $\to$ C	global / observables
Tracr transformer	Compiled transformer states	Symbolic sort-rank program	D $\to$ D	local / decoding
Gene regulatory network	Multi-valued Boolean GRN	Binary Wg–Fz rule	D $\to$ D	local / threshold
Logic circuit	Wire-level Boolean circuit	Multi-bit gate circuit	D $\to$ D	local / functional
Transistor circuit	SPICE/MOSFET voltages	Boolean gate network	C $\to$ D	local / threshold
MOS 6502 CPU	Transistor/segment state	Logic-gate state	D $\to$ D	local / functional
MOS 6502 CPU	Logic-gate state	ISA registers and flags	D $\to$ D	local / functional
MOS 6502 CPU	Transistor/segment state	ISA registers and flags	D $\to$ D	local / functional

5 Results

We evaluate metrics along three dimensions in sequence: classification performance under high sampling, discrimination of structurally invalid abstractions in controlled experiments, and sampling efficiency as a function of model size.

Benchmarking Metrics. We begin by asking which metrics can, in principle, separate valid from invalid abstractions when sampling is not a constraint. We evaluate all metrics the asymptotic regime on every (valid, invalid) abstraction pair in the benchmark, with full results reported in Appendices A.1 and A.2. These results motivate two principled thresholds: after normalization, a metric should score at most 0.10 on a valid abstraction and at least 0.15 on an invalid one. Figure 1 uses these thresholds to assess the classification performance of each metric across the benchmark.

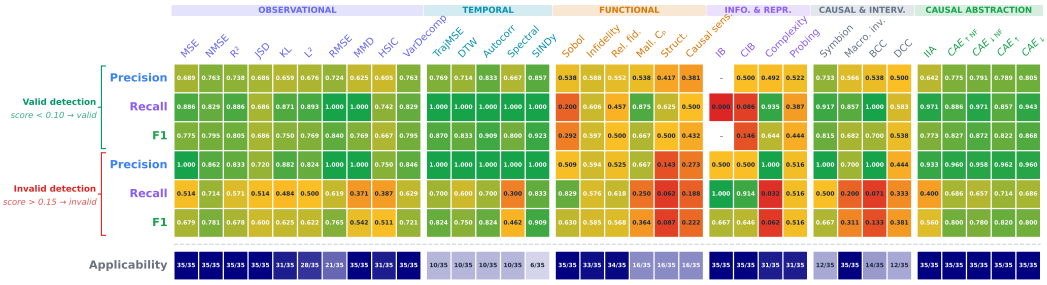

Figure 1: Precision, recall and F1-score of each metric, computed over all valid and invalid abstractions of the benchmark (higher is better). Applicability denotes the fraction of abstractions for which each metric was able to be evaluated; by design, many metrics are only applicable to a restricted class of systems. Metric normalization, implementation, and hyperparameters are reported in Appendix E.

Coverage (applicability) immediately eliminates a large subset of candidates. Temporal metrics (TrajMSE, DTW, Autocorr, Spectral, SINDy) are defined only for dynamical systems; Symbion requires finite discrete label spaces; C_p requires a linear proxy model with a well-defined noise floor; and structural deviation and the causal sensitivity index fail to produce scores on most benchmark systems. Among metrics with broad coverage, information-theoretic scores (IB and CIB Lagrangian) and several functional metrics, such as Sobol indices, structural deviation, and causal sensitivity indices, achieve mediocre to poor F1 on the valid task: their scores are sensitive to distributional details unrelated to causal consistency, so they assign non-trivial error even to valid abstractions. On the invalid task, the picture is mixed: some metrics attain reasonable precision, but at the cost of low recall. This combination of poor calibration and limited applicability immediately eliminates many candidate metrics from further consideration. Conversely, causal abstraction metrics consistently achieve among the highest F1 on both tasks and maximum coverage.

Detection of Controlled Failure Cases. Classification performance under unlimited sampling does not capture whether a metric can detect the specific structural failures that motivate causal abstraction. We therefore turn to six controlled experiments, each designed to expose a subtle violation that is observationally indistinguishable from the valid condition or manifests only under targeted interventions. The six settings are illustrated in Figure 2 and described in more detail in Appendix C. Figure 3 then reports, for each experiment and metric, whether a two-sample t-test over 100 runs reveals a statistically significant difference between the valid and invalid condition.

The results show that among the metrics that do apply, none of the observational, functional, or information-theoretic criteria achieve consistent discrimination across the six experiments. This is expected: each invalid abstraction is observationally indistinguishable from the valid one or fails only under specific interventional conditions, which are precisely the failure modes that causal abstraction was designed to detect. Within the causal abstraction family, results are more nuanced. IIA only succeeds in detecting the violations in Experiments 4 and 5. Both CAE_{NF} variants fail in Experiment 1 and 2; only the CAE variants successfully detect the violations.

Together, the benchmark results and controlled experiments narrow the candidate set to four metrics. IIA is excluded because, despite broad coverage, it fails to detect the violations in Experiments 3 and

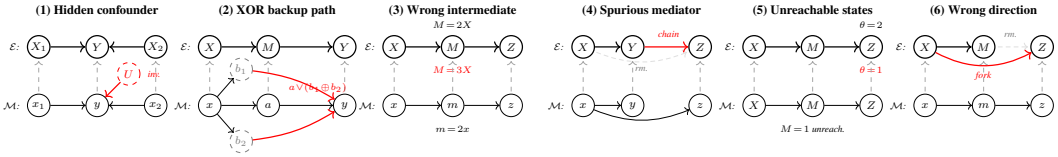

Figure 2: Causal graphs and abstraction maps for the six controlled experiments. Each panel shows $\mathcal{E}$ (top) and $\mathcal{M}$ (bottom). Gray dashed arrows are the abstraction map τ. Red marks what changes in the *invalid* condition; light gray dashed arrows mark edges present in the valid condition that are removed in the invalid condition. Dashed gray circles are ϕ variables (unmapped by τ); a dashed red circle is a ϕ variable introduced only in the invalid condition.

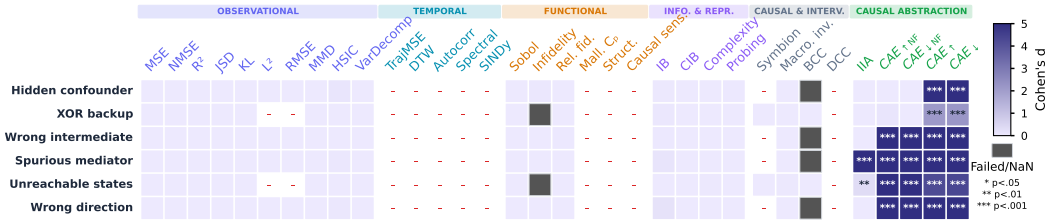

Figure 3: **Controlled failure modes.** Each cell reports whether a metric significantly separates the valid and invalid abstraction over 100 runs using a two-sample t-test. CAE is the only family that detects all six targeted invalidities.

6 where other causal abstraction metrics succeed. The remaining candidates are $\text{CAE}_{\downarrow}$ and $\text{CAE}_{\uparrow}$, which we carry forward as the primary metrics. Their faithfulness-free counterparts, $\text{CAE}_{\downarrow\text{NF}}$ and $\text{CAE}_{\uparrow\text{NF}}$, are retained as ablations to isolate the contribution of faithfulness testing.

Statistical Power. Having narrowed the candidate set to four metrics, we ask how many interventions each requires to reliably detect an invalid abstraction. Convergence results, reported in Appendix 9, show that all four metrics stabilize within 30 interventions on nearly all benchmark systems. Figure 4 reports statistical power across selected systems and controlled experiments.

Statistical power converges to nearly 100% within 30 interventions for most systems and all four metrics. Two exceptions stand out. For the gas simulation, power increases more slowly, consistent with the higher variance noted above. Experiment 2 (XOR leak) illustrates the usefulness of faithfulness testing: metrics that include it detect the invalid abstraction as the n grows, while their faithfulness-free counterparts consistently fail. The noise condition on the GRN system illustrates the value of top-down sampling: $\text{CAE}_{\downarrow}$ and $\text{CAE}_{\downarrow\text{NF}}$ can reliably detect the invalid abstraction after 25–30 interventions, while bottom-up metrics converge more slowly. For Experiment 5 (unreachable intermediate states), the picture reverses: faithfulness-aware metrics converge more slowly and exhibit lower power than plain CAE_{NF}. $\text{CAE}_{\downarrow\text{NF}}$ already probes unreachable states by sampling from the full abstract label domain; faithfulness augmentation additionally injects perturbations into unmapped Φ

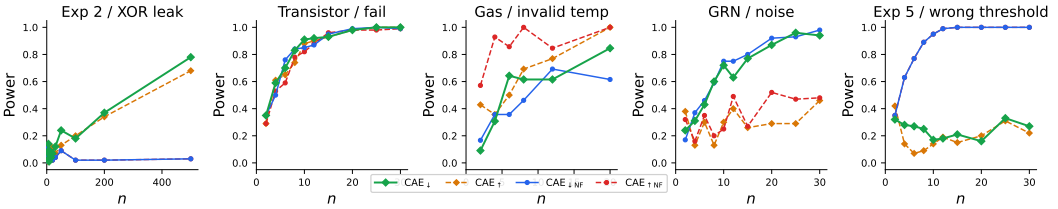

Figure 4: How quickly do metrics detect bad abstractions over good ones? For different numbers of sampled interventions n, we compute each metric 100 times on selected valid and invalid abstractions. We apply a two-sample T-test to the two sets of scores (valid and invalid) and report the fraction of runs for which the results differ significantly ($p < 0.05$). Full results are given in appendix A.

variables, which occasionally push $\mathcal{M}$ into the same micro-region as the unreachable states, inflating the valid-condition score and shrinking the discrimination gap. Faithfulness testing therefore does not miss the invalid abstraction here, but reduces specificity on the valid side.

Scaling. Power curves on fixed systems do not reveal how detection cost scales as the high-level model $\mathcal{E}$. To answer this question, we construct a controlled scaling experiment using Tracr [Lindner et al., 2023], a compiler that translates RASP programs into exact transformer weights. We compile a sort-rank program for each sequence length $\ell \in \{2, 3, 4, 5, 6, 8, 10, 12, 15\}$, yielding transformers of increasing size that compute the rank of each token in the input sequence. The valid high-level model ($\mathcal{E}$) encodes the correct formula $\mathrm{rank}_i = |\{j \neq i : \mathrm{token}_j < \mathrm{token}_i\}|$ for all positions i, while the wrong model encodes a specific incorrect hypothesis: the minimum-valued token at position 0 always receives the maximum rank rather than rank 0. This error triggers exactly when token_0 is the minimum element of the sequence, an event with probability $1/\ell$ under uniform token sampling. The expected number of interventions to observe at least one triggering event therefore grows linearly with ℓ, and analytic calculation gives $n_{95\%} \approx 3\ell$ for 95% power.

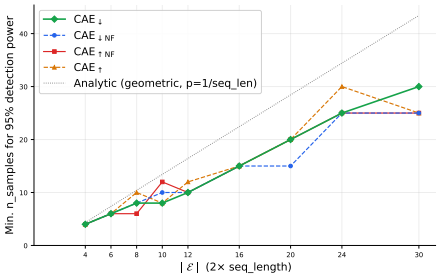

Figure 5: Minimum number of interventions required to reach 95% detection power as a function of $|\mathcal{E}|$, measured on Tracr sort-rank programs of increasing sequence length (100 runs per length). More details are provided in Appendix A.4.

Figure 5 shows the empirically measured required n for CAE and CAE_{NF} as a function of $|\mathcal{E}| = 2\ell$ (token inputs plus rank outputs). All four metrics follow the predicted linear trend closely. This confirms that the sampling budget is governed primarily by the rarity of the triggering condition in the abstract domain, not by the internal complexity of the transformer. In practice, a wrong explanation that diverges from the correct one on a fraction p of inputs requires $O(1/p)$ interventions to detect with fixed confidence: rare but systematic errors in proposed explanations remain detectable at a cost proportional to their rarity, independent of model depth or width.

6 Discussion

This paper asks when a high-level explanation of a low-level system should be trusted. Our benchmark turns this into an empirical stress test: given valid and invalid abstraction pairs, which metrics separate them? The main result is that observational agreement, distribution matching, representational similarity, compression, symbolic fit, and input-output sensitivity can be useful diagnostics, but not validity criteria. Across heterogeneous systems, causal-abstraction metrics are the most reliable. However, our results also show that operationalization matters: not all causal metrics suceed and different implementations fail in different ways. The proposed CAE metrics, which combine multi-node interventions with built-in faithfulness tests, are the only metrics that pass all tests. This makes CAE a natural objective for explanation discovery. Symbolic, gradient-based, evolutionary, and human-guided search procedures all need a criterion that rewards causally valid abstractions. In particular, future work should study adaptive intervention sampling to quickly detect invalid abstraction, CAE-based abstraction discovery, and community extensions of the benchmark. We release the benchmark and all metric implementations to support this direction.

6.1 Limitations

First, sample-level CAE assumes shared exogenous noise across abstraction levels. Our benchmark avoids this for deterministic macro-models, but stochastic micro-models are evaluated with independent stochastic realizations. Coupled stochastic abstractions remain future work. Second, the benchmark necessarily involves design choices that can be debated: systems, valid abstractions, invalid contrasts, and sampled interventions. We mitigate this by trying to span diverse domains and abstraction types, but future versions should include community-contributed systems and possibly blind evaluations. Finally, our power-scaling analysis is limited to a specific Tracr setting focused on explanation size. Yet, intervention requirements to distinguish an invalid explanation will depend on properties of the system, abstraction map, explanation size, and invalid contrast.

References

Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 2003.

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

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Gradient-based attribution methods. *Explainable AI: Interpreting, explaining and visualizing deep learning*, pages 169–191, 2019.

Khai Loong Aw, Syrielle Montariol, Badr AlKhamissi, Martin Schrimpf, and Antoine Bosselut. Instruction-tuning aligns llms to the human brain, 2024. URL https://arxiv.org/abs/2312.00575.

Robert Axtell, Robert Axelrod, Joshua M. Epstein, and Michael D. Cohen. Aligning simulation models: A case study and results. *Computational & Mathematical Organization Theory*, 1(2): 123–141, Feb 1996. ISSN 1572-9346. doi: 10.1007/BF01299065. URL https://doi.org/10.1007/BF01299065.

Robert W Batterman and Collin C Rice. Minimal model explanations. *Philosophy of Science*, 81(3): 349–376, 2014.

Sander Beckers and Joseph Y. Halpern. Abstracting causal models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):2678–2685, Jul. 2019. doi: 10.1609/aaai.v33i01.33012678. URL https://ojs.aaai.org/index.php/AAAI/article/view/4117.

Sander Beckers, Frederick Eberhardt, and Joseph Y. Halpern. Approximate causal abstractions. In Ryan P. Adams and Vibhav Gogate, editors, *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 606–615. PMLR, 22–25 Jul 2020. URL https://proceedings.mlr.press/v115/beckers20a.html.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, March 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7/.

A. Annie Steffy Beula, Geno Peter, Albert Alexander Stonier, K. Ezhil Vignesh, and Vivekananda Ganji. Behaviour analysis of modeling and model evaluating methods in system identification for a multiprocess station. *Complexity*, 2024(1):7741473, 2024. doi: https://doi.org/10.1155/2024/7741473. URL https://onlinelibrary.wiley.com/doi/abs/10.1155/2024/7741473.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences of the United States of America*, 113(15):3932–3937, April 2016. ISSN 0027-8424. doi: 10.1073/pnas.1517384113. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4839439/.

Krzysztof Chalupka, Frederick Eberhardt, and Pietro Perona. Multi-level cause-effect systems. In Arthur Gretton and Christian C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 361–369, Cadiz, Spain, 09–11 May 2016. PMLR. URL https://proceedings.mlr.press/v51/chalupka16.html.

Andrew Collins, Matthew Koehler, and Christopher Lynch. Methods that support the validation of agent-based models: An overview and discussion. *Journal of Artificial Societies and Social Simulation*, 27(1):11, 2024. ISSN 1460-7425. doi: 10.18564/jasss.5258. URL http://jasss.soc.surrey.ac.uk/27/1/11.html.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 16318–16352. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/34e1dbe95d34d7ebaf99b9bcaeb5b2be-Paper-Conference.pdf.

James P. Crutchfield and Karl Young. Inferring statistical complexity. *Phys. Rev. Lett.*, 63:105–108, Jul 1989. doi: 10.1103/PhysRevLett.63.105. URL https://link.aps.org/doi/10.1103/PhysRevLett.63.105.

RI Cukier, CM Fortuin, Kurt E Shuler, AG Petschek, and J Ho Schaibly. Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. i theory. *The Journal of chemical physics*, 59(8):3873–3878, 1973.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023. URL https://arxiv.org/abs/2309.08600.

John P Cunningham and Byron M Yu. Dimensionality reduction for large-scale neural recordings. *Nature neuroscience*, 17(11):1500–1509, 2014.

Junlan Dong and Jinghui Zhong. Recent advances in symbolic regression. *ACM Comput. Surv.*, 57 (11), June 2025. ISSN 0360-0300. doi: 10.1145/3735634. URL https://doi.org/10.1145/3735634.

Joel Dyer, Nicholas Bishop, Yorgos Felekis, Fabio Massimo Zennaro, Anisoara Calinescu, Theodoros Damoulas, and Michael Wooldridge. Interventionally consistent surrogates for complex simulation models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 21814–21841. Curran Associates, Inc., 2024. doi: 10.52202/079017-0686. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/26b8e3dc3a21fcd660d80c63b767f324-Paper-Conference.pdf.

Bradley Efron and Charles Stein. The jackknife estimate of variance. *The Annals of Statistics*, pages 586–596, 1981.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021. doi: 10.1162/tacl_a_00359. URL https://aclanthology.org/2021.tacl-1.10/.

Michael E. Fisher. Renormalization group theory: Its basis and formulation in statistical physics. *Rev. Mod. Phys.*, 70:653–681, Apr 1998. doi: 10.1103/RevModPhys.70.653. URL https://link.aps.org/doi/10.1103/RevModPhys.70.653.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal Abstractions of Neural Networks, October 2021. URL http://arxiv.org/abs/2106.02997. arXiv:2106.02997 [cs].

Atticus Geiger, Zhengxuan Wu, Karel D'Oosterlinck, Elisa Kreiss, Noah D. Goodman, Thomas Icard, and Christopher Potts. Faithful, interpretable model explanations via causal abstraction. Stanford AI Lab Blog, 2022a. URL https://ai.stanford.edu/blog/causal-abstraction/.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. Inducing causal structure for interpretable neural networks. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 7324–7338. PMLR, 17–23 Jul 2022b. URL https://proceedings.mlr.press/v162/geiger22a.html.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025. URL http://jmlr.org/papers/v26/23-0058.html.

H Georgi. Effective Field Theory. *Annual Review of Nuclear and Particle Science*, 43(1):209–252, December 1993. ISSN 0163-8998, 1545-4134. doi: 10.1146/annurev.ns.43.120193.001233. URL https://www.annualreviews.org/doi/10.1146/annurev.ns.43.120193.001233.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.751. URL https://aclanthology.org/2023.emnlp-main.751.

Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3681–3688, Jul. 2019. doi: 10.1609/aaai.v33i01.33013681. URL https://ojs.aaai.org/index.php/AAAI/article/view/4252.

M.L. Glasser. Second virial coefficient for a lennard-jones (2n-n) system in d dimensions and confined to a nanotube surface. *Physics Letters A*, 300(4):381–384, 2002. ISSN 0375-9601. doi: https://doi.org/10.1016/S0375-9601(02)00814-9. URL https://www.sciencedirect.com/science/article/pii/S0375960102008149.

Arthur Gretton, Ralf Herbrich, Alexander Smola, Olivier Bousquet, and Bernhard Schölkopf. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6(70):2075–2129, 2005. URL http://jmlr.org/papers/v6/gretton05a.html.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL http://jmlr.org/papers/v13/gretton12a.html.

Fabio Gritti, Lorenzo Fontana, Eric Gustafson, Fabio Pagani, Andrea Continella, Christopher Kruegel, and Giovanni Vigna. Symbion: Interleaving symbolic with concrete execution. In *2020 IEEE Conference on Communications and Network Security (CNS)*, pages 1–10, 2020. doi: 10.1109/CNS48642.2020.9162164.

Moritz Grosse-Wentrup, Akshey Kumar, Anja Meunier, and Manuel Zimmer. Neuro-cognitive multilevel causal modeling: A framework that bridges the explanatory gap between neuronal activity and cognition. *PLOS Computational Biology*, 20(12):1–32, 12 2024. doi: 10.1371/journal.pcbi.1012674. URL https://doi.org/10.1371/journal.pcbi.1012674.

Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms. *arXiv preprint arXiv:2403.17806*, 2024.

James V Haxby, M Ida Gobbini, Maura L Furey, Alumit Ishai, Jennifer L Schouten, and Pietro Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430, 2001.

Carl Gustav Hempel. *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. The Free Press, New York, 1965.

Bernease Herman. The promise and peril of human evaluation for model interpretability. *arXiv preprint arXiv:1711.07414*, 2017.

Erik P. Hoel, Larissa Albantakis, and Giulio Tononi. Quantifying causal emergence shows that macro can beat micro. *Proceedings of the National Academy of Sciences*, 110(49):19790–19795, 2013. doi: 10.1073/pnas.1314922110. URL https://www.pnas.org/doi/abs/10.1073/pnas.1314922110.

Ari Holtzman, Peter West, and Luke Zettlemoyer. Generative models as a complex systems science: How can we make sense of large language model behavior? *Journal of Social Computing*, 6(2): 75–94, 2025. doi: 10.23919/JSC.2025.0009. URL https://www.sciopen.com/article/10.23919/JSC.2025.0009.

Jing Huang, Zhengxuan Wu, Christopher Potts, Mor Geva, and Atticus Geiger. RAVEL: Evaluating interpretability methods on disentangling language model representations. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8669–8687, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.470. URL https://aclanthology.org/2024.acl-long.470/.

Guilherme Seidyo Imai Aldeia, Hengzhe Zhang, Geoffrey Bomarito, Miles Cranmer, Alcides Fonseca, Bogdan Burlacu, William G. La Cava, and Fabrício Olivetti de França. Call for action: towards the next generation of symbolic regression benchmark. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, GECCO '25 Companion, page 2529–2538, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400714641. doi: 10.1145/3712255.3734309. URL https://doi.org/10.1145/3712255.3734309.

Eric Jonas and Konrad Paul Kording. Could a neuroscientist understand a microprocessor? *PLoS computational biology*, 13(1):e1005268, 2017.

Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Callum McDougall, Kola Ayonrinde, Matthew Wearden, Arthur Conmy, Samuel Marks, and Neel Nanda. Saebench: A comprehensive benchmark for sparse autoencoders in language model interpretability, 2025. URL https://arxiv.org/abs/2503.09532.

Ronald Katende. Causal operator discovery in partial differential equations via counterfactual physics-informed neural networks, 2025. URL https://arxiv.org/abs/2506.20181.

Armin Kekić, Bernhard Schölkopf, and Michel Besserve. Targeted reduction of causal models. *arXiv preprint arXiv:2311.18639*, 2023.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav), 2018. URL https://arxiv.org/abs/1711.11279.

Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. *The (Un)reliability of Saliency Methods*, pages 267–280. Springer International Publishing, Cham, 2019. ISBN 978-3-030-28954-6. doi: 10.1007/978-3-030-28954-6_14. URL https://doi.org/10.1007/978-3-030-28954-6_14.

Philip Kitcher. Explanatory unification and the causal structure of the world. In Philip Kitcher and Wesley C. Salmon, editors, *Scientific Explanation*, pages 410–505. Univ of Minnesota Pr, 1962.

John R Koza. Genetic programming as a means for programming computers by natural selection. *Statistics and computing*, 4(2):87–112, 1994.

Philip A Kragel, Leonie Koban, Lisa Feldman Barrett, and Tor D Wager. Representation, pattern information, and brain signatures: from neurons to neuroimaging. *Neuron*, 99(2):257–273, 2018.

János Kramár, Tom Lieberum, Rohin Shah, and Neel Nanda. Atp*: An efficient and scalable method for localizing llm behaviour to components, 2024. URL https://arxiv.org/abs/2403.00745.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249, 2008.

Mikel Landajuela, Chak Shing Lee, Jiachen Yang, Ruben Glatt, Claudio P Santiago, Ignacio Aravena, Terrell Mundhenk, Garrett Mulcahy, and Brenden K Petersen. A unified framework for deep symbolic regression. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 33985–33998. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/dbca58f35bddc6e4003b2dd80e42f838-Paper-Conference.pdf.

Karim Lasri, Tiago Pimentel, Alessandro Lenci, Thierry Poibeau, and Ryan Cotterell. Probing for the usage of grammatical number. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8818–8831, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.603. URL https://aclanthology.org/2022.acl-long.603/.

Hyeyoung Lee and Sujit Ghosh. Performance of information criteria for spatial models. *Journal of statistical computation and simulation*, 79:93–106, 01 2009. doi: 10.1080/00949650701611143.

David Lindner, János Kramár, Sebastian Farquhar, Matthew Rahtz, Tom McGrath, and Vladimir Mikulik. Tracr: Compiled transformers as a laboratory for interpretability. *Advances in Neural Information Processing Systems*, 36:37876–37899, 2023.

Yang Liu, Sujay Khandagale, Colin White, and Willie Neiswanger. Synthetic benchmarks for scientific research in explainable machine learning, 2021. URL https://arxiv.org/abs/2106.12543.

Lennart Ljung. *System Identification*. John Wiley & Sons, Ltd, 1999. ISBN 9780471346081. doi: https://doi.org/10.1002/047134608X.W1046. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/047134608X.W1046.

Tania Lombrozo. The structure and function of explanations. *Trends in cognitive sciences*, 10(10): 464–470, 2006.

Michel Loreau. *From populations to ecosystems: Theoretical foundations for a new ecological synthesis (MPB-46)*. Princeton University Press, 2010.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

N. M. Mangan, J. N. Kutz, S. L. Brunton, and J. L. Proctor. Model selection for dynamical systems via sparse regression and information criteria. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473(2204):20170009, August 2017. ISSN 1471-2946. doi: 10.1098/rspa.2017.0009. URL http://dx.doi.org/10.1098/rspa.2017.0009.

Thomas McGrath, Matthew Rahtz, Janos Kramar, Vladimir Mikulik, and Shane Legg. The hydra effect: Emergent self-repair in language model computations, 2023. URL https://arxiv.org/abs/2307.15771.

Maxime Méloux, Giada Dirupo, François Portet, and Maxime Peyrard. The dead salmons of ai interpretability. *arXiv preprint arXiv:2512.18792*, 2025a.

Maxime Méloux, François Portet, and Maxime Peyrard. Mechanistic interpretability as statistical estimation: A variance analysis of eap-ig. *arXiv preprint arXiv:2510.00845*, 2025b.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 36, 2022.

David Meunier, Renaud Lambiotte, and Edward T Bullmore. Modular and hierarchically modular organization of brain networks. *Frontiers in neuroscience*, 4:200, 2010.

Giovanni Monea, Maxime Peyrard, Martin Josifoski, Vishrav Chaudhary, Jason Eisner, Emre Kıcıman, Hamid Palangi, Barun Patra, and Robert West. A glitch in the matrix? locating and detecting language model grounding with fakepedia, 2024. URL https://arxiv.org/abs/2312.02073.

Max D Morris. Factorial sampling plans for preliminary computational experiments. *Technometrics*, 33(2):161–174, 1991.

Kenneth A Norman, Sean M Polyn, Greg J Detre, and James V Haxby. Beyond mind-reading: multi-voxel pattern analysis of fmri data. *Trends in cognitive sciences*, 10(9):424–430, 2006.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020. doi: 10.23915/distill.00024.001. https://distill.pub/2020/circuits/zoom-in.

Jun Otsuka and Hayato Saigo. On the equivalence of causal models: A category-theoretic approach. In *Conference on Causal Learning and Reasoning*, pages 634–646. PMLR, 2022.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition, 2009. ISBN 052189560X.

Donald B Percival and Andrew T Walden. *Spectral analysis for physical applications*. cambridge university press, 1993.

Brenden K Petersen, Mikel Landajuela, T Nathan Mundhenk, Claudio P Santiago, Soo K Kim, and Joanne T Kim. Deep symbolic regression: Recovering mathematical expressions from data via risk-seeking policy gradients. *arXiv preprint arXiv:1912.04871*, 2019.

Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

Antonin Poché, Alon Jacovi, Agustin Martin Picard, Victor Boutin, and Fanny Jourdan. Consim: Measuring concept-based explanations' effectiveness with automated simulatability. *arXiv preprint arXiv:2501.05855*, 2025.

Angela Potochnik. *Idealization and the Aims of Science*. University of Chicago Press, Chicago, 2017.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3363–3377, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.295. URL https://aclanthology.org/2021.eacl-main.295.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939778. URL https://doi.org/10.1145/2939672.2939778.

Eigil F Rischel and Sebastian Weichwald. Compositional abstraction error and a category of causal models. In *Uncertainty in Artificial Intelligence*, pages 1013–1023. PMLR, 2021.

Eigil Fjeldgren Rischel. The category theory of causal models. *Master's thesis, University of Copenhagen*, 2020.

Nathalie L Rochefort, Olga Garaschuk, Ruxandra-Iulia Milos, Madoka Narushima, Nima Marandi, Bruno Pichler, Yury Kovalchuk, and Arthur Konnerth. Sparsification of neuronal activity in the visual cortex at eye-opening. *Proceedings of the National Academy of Sciences*, 106(35): 15049–15054, 2009.

Fernando E. Rosas, Bernhard C. Geiger, Andrea I Luppi, Anil K. Seth, Daniel Polani, Michael Gastpar, and Pedro A. M. Mediano. Software in the natural world: A computational approach to hierarchical emergence, 2024. URL https://arxiv.org/abs/2402.09090.

P. K. Rubenstein, S. Weichwald, S. Bongers, J. M. Mooij, D. Janzing, M. Grosse-Wentrup, and B. Schölkopf. Causal consistency of structural equation models. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence (UAI)*, page ID 11, August 2017. URL http://auai.org/uai2017/proceedings/papers/11.pdf. *equal contribution.

Maria Ryskina, Greta Tuckute, Alexander Fung, Ashley Malkin, and Evelina Fedorenko. Language models align with brain regions that represent concepts across modalities. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=2JohTFaGbW.

Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1):43–49, 1978.

Wesley C. Salmon. *Scientific Explanation and the Causal Structure of the World*. Princeton University Press, 1984. ISBN 9780691072937. URL http://www.jstor.org/stable/j.ctv173f2gh.

Andrea Saltelli, Stefano Tarantola, and KP-S Chan. A quantitative model-independent method for global sensitivity analysis of model output. *Technometrics*, 41(1):39–56, 1999.

Lucas Sánchez, Claudine Chaouiya, and Denis Thieffry. Segmenting the fly embryo: logical analysis of the role of the segment polarity cross-regulatory module. *The International Journal of Developmental Biology*, 52(8):1059–1075, 2026/03/18/13:57:44 2008. ISSN 0214-6282. doi: 10.1387/ijdb.072439ls. URL https://ijdb.ehu.eus/article/072439ls.

Michael Schmidt and Hod Lipson. Distilling free-form natural laws from experimental data. *science*, 324(5923):81–85, 2009.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, 2021.

Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.

Cosma Rohilla Shalizi and James P. Crutchfield. Computational Mechanics: Pattern and Prediction, Structure and Simplicity. *Journal of Statistical Physics*, 104(3):817–879, August 2001. ISSN 1572-9613. doi: 10.1023/A:1010388907793. URL https://doi.org/10.1023/A:1010388907793.

Francisco N. F. Q. Simoes, Mehdi Dastani, and Thijs van Ommen. The causal information bottleneck and optimal causal variable abstractions, 2025. URL https://arxiv.org/abs/2410.00535.

I.M Sobol'. Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Mathematics and Computers in Simulation*, 55(1):271–280, 2001. ISSN 0378-4754. doi: https://doi.org/10.1016/S0378-4754(00)00270-6. URL https://www.sciencedirect.com/science/article/pii/S0378475400002706. The Second IMACS Seminar on Monte Carlo Methods.

Denis Sutter, Julian Minder, Thomas Hofmann, and Tiago Pimentel. The non-linear representation dilemma: Is causal abstraction enough for mechanistic interpretability?, 2025. URL https://arxiv.org/abs/2507.08802.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery, 2023. URL https://arxiv.org/abs/2310.10348.

Wassim Tenachi, Rodrigo Ibata, and Foivos I. Diakogiannis. Deep symbolic regression for physics guided by units constraints: Toward the automated discovery of physical laws. *The Astrophysical Journal*, 959(2):99, December 2023. ISSN 1538-4357. doi: 10.3847/1538-4357/ad014c. URL http://dx.doi.org/10.3847/1538-4357/ad014c.

Damien Teney, Maxime Peyrard, and Ehsan Abbasnejad. Predicting is not understanding: Recognizing and addressing underspecification in machine learning. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIII*, page 458–476, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-20049-6. doi: 10.1007/978-3-031-20050-2_27. URL https://doi.org/10.1007/978-3-031-20050-2_27.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method, 2000. URL https://arxiv.org/abs/physics/0004057.

Silviu-Marian Udrescu and Max Tegmark. Ai feynman: a physics-inspired method for symbolic regression, 2020. URL https://arxiv.org/abs/1905.11481.

Guillermo Vallejo, Ellián Tuero-Herrero, José Carlos Núñez, and Pedro Rosário. Performance evaluation of recent information criteria for selecting multilevel models in behavioral and social sciences. *International Journal of Clinical and Health Psychology*, 14(1):48–57, 2014. ISSN 1697-2600. doi: https://doi.org/10.1016/S1697-2600(14)70036-5. URL https://www.sciencedirect.com/science/article/pii/S1697260014700365.

Konstantina Valogianni and Balaji Padmanabhan. Causal abms: Learning plausible causal models using agent-based modeling. In Thuc Duy Le, Lin Liu, Emre Kıcıman, Sofia Triantafyllou, and Huan Liu, editors, *Proceedings of The KDD'22 Workshop on Causal Discovery*, volume 185 of *Proceedings of Machine Learning Research*, pages 3–29. PMLR, 15 Aug 2022. URL https://proceedings.mlr.press/v185/valogianni22a.html.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf.

Sandra Wachter, Brent D. Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *CoRR*, abs/1711.00399, 2017. URL http://arxiv.org/abs/1711.00399.

Tor D Wager, Lauren Y Atlas, Martin A Lindquist, Mathieu Roy, Choong-Wan Woo, and Ethan Kross. An fmri-based neurologic signature of physical pain. *New England Journal of Medicine*, 368(15):1388–1397, 2013.

Warren Weaver. Science and complexity. In *Facets of systems science*, pages 449–456. Springer, 1991.

Jim Woodcock, Peter Gorm Larsen, Juan Bicarregui, and John Fitzgerald. Formal methods: Practice and experience. *ACM Comput. Surv.*, 41(4), October 2009. ISSN 0360-0300. doi: 10.1145/1592434.1592436. URL https://doi.org/10.1145/1592434.1592436.

James Woodward. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, 01 2004. ISBN 9780195155273. doi: 10.1093/0195155270.001.0001. URL https://doi.org/10.1093/0195155270.001.0001.

Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Sai Suggala, David I. Inouye, and Pradeep Ravikumar. On the (in)fidelity and sensitivity for explanations, 2019. URL https://arxiv.org/abs/1901.09392.

Rafael Yuste. Circuit neuroscience: the road ahead. *Frontiers in neuroscience*, 2:1038, 2008.

H. Zenil, N. A. Kiani, F. S Abrahão, and J. N. Tegnér. Algorithmic Information Dynamics. *Scholarpedia*, 15(7):53143, 2020. doi: 10.4249/scholarpedia.53143. revision #203921.

Hector Zenil, Narsis A. Kiani, Francesco Marabita, Yue Deng, Szabolcs Elias, Angelika Schmidt, Gordon Ball, and Jesper Tegnér. An algorithmic information calculus for causal discovery and reprogramming systems. *iScience*, 19:1160–1172, 2019. ISSN 2589-0042. doi: https://doi.org/10.1016/j.isci.2019.07.043. URL https://www.sciencedirect.com/science/article/pii/S2589004219302706.

Hector Zenil, Narsis A. Kiani, and Jesper Tegnér. *References*, page 310–321. Cambridge University Press, 2023.

Fabio Massimo Zennaro. Abstraction between structural causal models: A review of definitions and properties. *arXiv preprint arXiv:2207.08603*, 2022.

Fabio Massimo Zennaro, Máté Drávucz, Geanina Apachitei, W Dhammika Widanage, and Theodoros Damoulas. Jointly learning consistent causal abstractions over multiple interventional distributions. In *Conference on Causal Learning and Reasoning*, pages 88–121. PMLR, 2023a.

Fabio Massimo Zennaro, Paolo Turrini, and Theodoros Damoulas. Quantifying Consistency and Information Loss for Causal Abstraction Learning. In *Electronic proceedings of IJCAI 2023*, volume 5, pages 5750–5757, August 2023b. doi: 10.24963/ijcai.2023/638. URL https://www.ijcai.org/proceedings/2023/638. ISSN: 1045-0823.

Hanwei Zhang, Felipe Torres Figueroa, and Holger Hermanns. Saliency Maps Give a False Sense of Explanability to Image Classifiers: An empirical evaluation across methods and metrics. In Vu Nguyen and Hsuan-Tien Lin, editors, *Proceedings of the 16th Asian Conference on Machine Learning*, volume 260 of *Proceedings of Machine Learning Research*, pages 479–494. PMLR, 05–08 Dec 2025. URL https://proceedings.mlr.press/v260/zhang25a.html.

Yuchen Zhu, Sergio Hernan Garrido Mejia, Bernhard Schölkopf, and Michel Besserve. Unsupervised causal abstraction. In *NeurIPS 2024 Causal Representation Learning Workshop*, 2024.

A Additional results

This section contains multiple additional figures for results described in the main body of the work, as well as a new experiment on convergence rates.

A.1 Validity on Correct Abstractions

Figure 6 shows the scores produced by each baseline metric when applied to the correct abstractions in our benchmark.

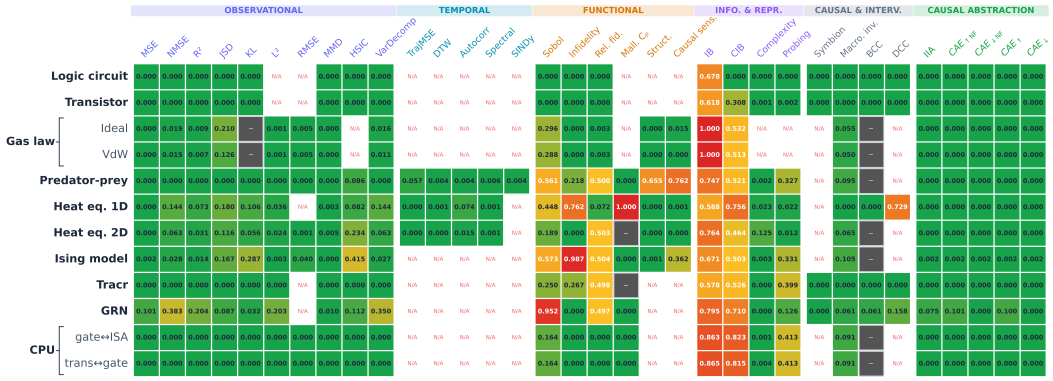

Figure 6: Error scores obtained by baseline metrics when applied to valid abstractions across benchmark systems. Scores are normalized to the $[0, 1]$ range, where 0 denotes no error and 1 maximal error. Gray cells denote computation failure or NaN results for the metric.

A.2 Discrimination Power on Invalid Abstractions

We report in Figure 7 the results of the discrimination measurement in 5, applied to all systems (including controlled experiments) and metrics.

A.3 Power Curves

Figure 8 contains the full results of the power experiments in 5, computed across all invalid abstractions for all systems.

Figure 9 shows how the score estimates produced by each causal metric stabilize for different numbers of sampled interventions.

Estimates stabilize quickly across most systems: 30 sampled interventions are generally sufficient to obtain a stable score, with the gas simulation generally requiring more samples due to the higher variance of the molecular dynamics estimator. No metric stands out as substantially slower to converge than the others. However, differences between sampling directions emerge. For the GRN with a reversed causal rule, top-down metrics ($\text{CAE}_{\downarrow}$ and $\text{CAE}_{\downarrow\text{NF}}$), which intervene directly on the abstract states that expose the mis-specification, score higher than their bottom-up counterparts, which only draw micro-states reachable from the input distribution. For the GRN noise condition, bottom-up metrics score higher than top-down ones on valid abstractions: sampling generates interventions near the boundaries of the value-map subspaces, and noisy outputs of $\mathcal{M}$ are more likely to be misclassified after abstraction. Top-down sampling stays in the interior of each subspace, reducing the noise. The separation is notably strong in Experiment 1 (hidden confounder): CAE_{NF} variants score near zero, while faithfulness-augmented variants detect the invalid abstraction (score near one). This demonstrates that faithfulness testing is critical, independently of the sampling direction.

A.4 Tracr power curves

We provide here extended results for the Tracr sort-rank scaling experiment.

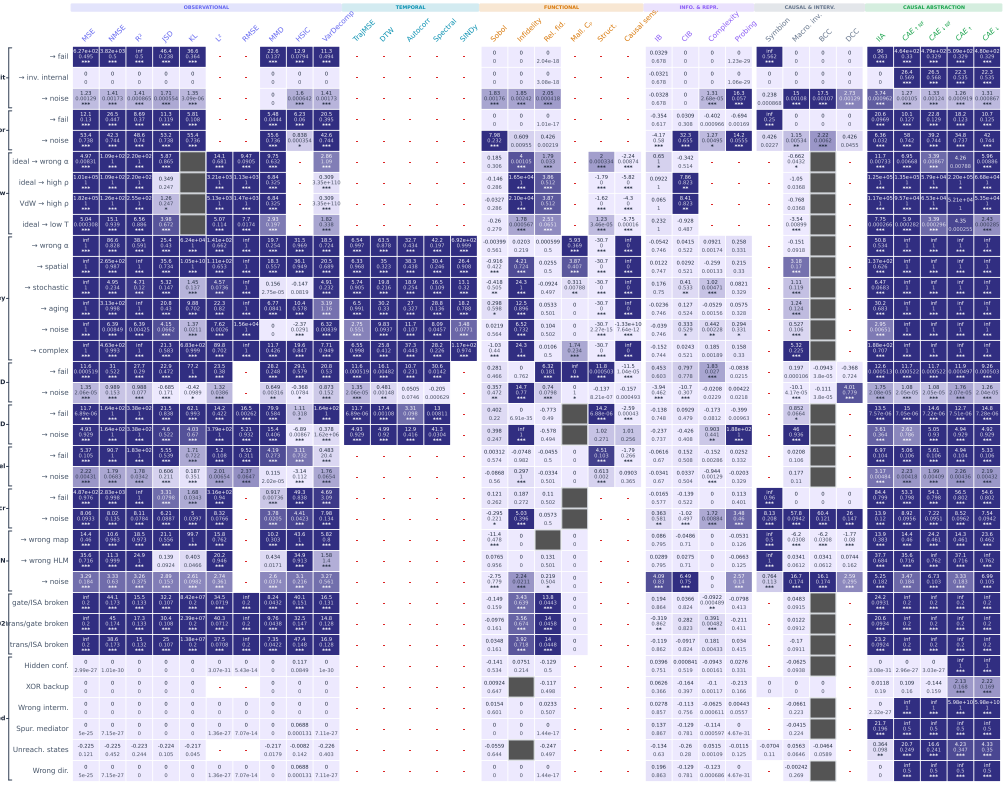

Figure 7: Discriminatory power of baseline metrics when applied to invalid abstractions compared to valid ones. For each cell, we report three values (from top to bottom): Cohen's d, computed on the scores obtained by the metric on the invalid vs. valid abstractions; the score obtained on the invalid condition; the significance level, computed on a two-sample T-test applied between the valid and invalid condition scores. Gray cells denote computation failure or NaN results for the metric. Stars represent p-values: p<0.05 (*), p<0.01 (**) and p<0.001 (***).

Figure 10 displays the full detection power curves across different sequence lengths. Figure 11 provides the minimal sample counts required to achieve a 95% detection threshold.

B Extended Background: Metrics of Explanation Validity

We survey the principal classes of metrics used to evaluate whether a computational explanation $\mathcal{E}$ is a valid account of a target system $\mathcal{M}$. Throughout, $u \sim P_{\mathcal{U}}$ denotes an exogenous input drawn from the natural input distribution, and validity is assessed relative to the observable or internal behavior of $\mathcal{M}$. We organize criteria by the type of evidence they require: observational, functional, information-theoretic, or causal.

B.1 Observational Validity

The most elementary criterion for explanation validity is *observational equivalence*: $\mathcal{E}$ is valid insofar as it reproduces the measurable outputs of $\mathcal{M}$ under the natural input distribution.

Pointwise reconstruction accuracy. In deterministic or regression settings, observational validity is operationalized as low expected point-to-point discrepancy between the outputs of $\mathcal{E}$ and $\mathcal{M}$. Standard metrics include the *mean squared error* $\mathrm{MSE} = \mathbb{E}_u[\|\mathcal{M}(u) - \mathcal{E}(u)\|^2]$, the *root mean squared error* RMSE, the ℓ_2 *distance* L_2, and the *normalized* variant NMSE, which removes the dependence on output scale. The coefficient of determination R^2 provides an interpretable measure of the fraction of variance explained. These criteria underlie classical system identification [Ljung, 1999] and general-purpose symbolic regression [Koza, 1994, Schmidt and Lipson, 2009, Petersen et al.,

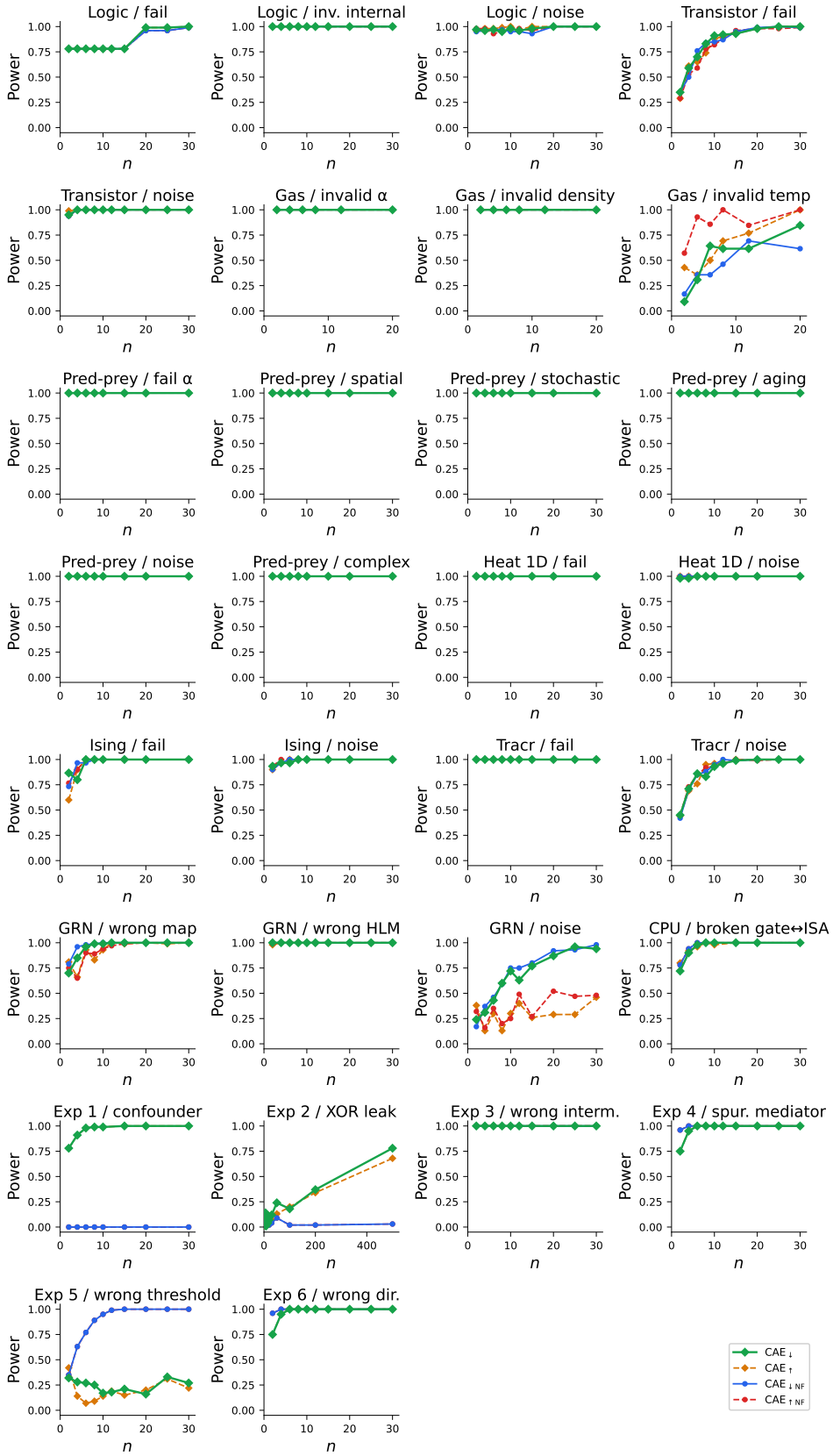

Figure 8: How quickly do metrics detect bad abstractions over good ones? For different numbers of sampled interventions n, we compute each metric 100 times on selected valid and invalid abstractions. We apply a two-sample T-test to the two sets of scores (valid and invalid) and report the fraction of runs for which the results differ significantly ($p < 0.05$).

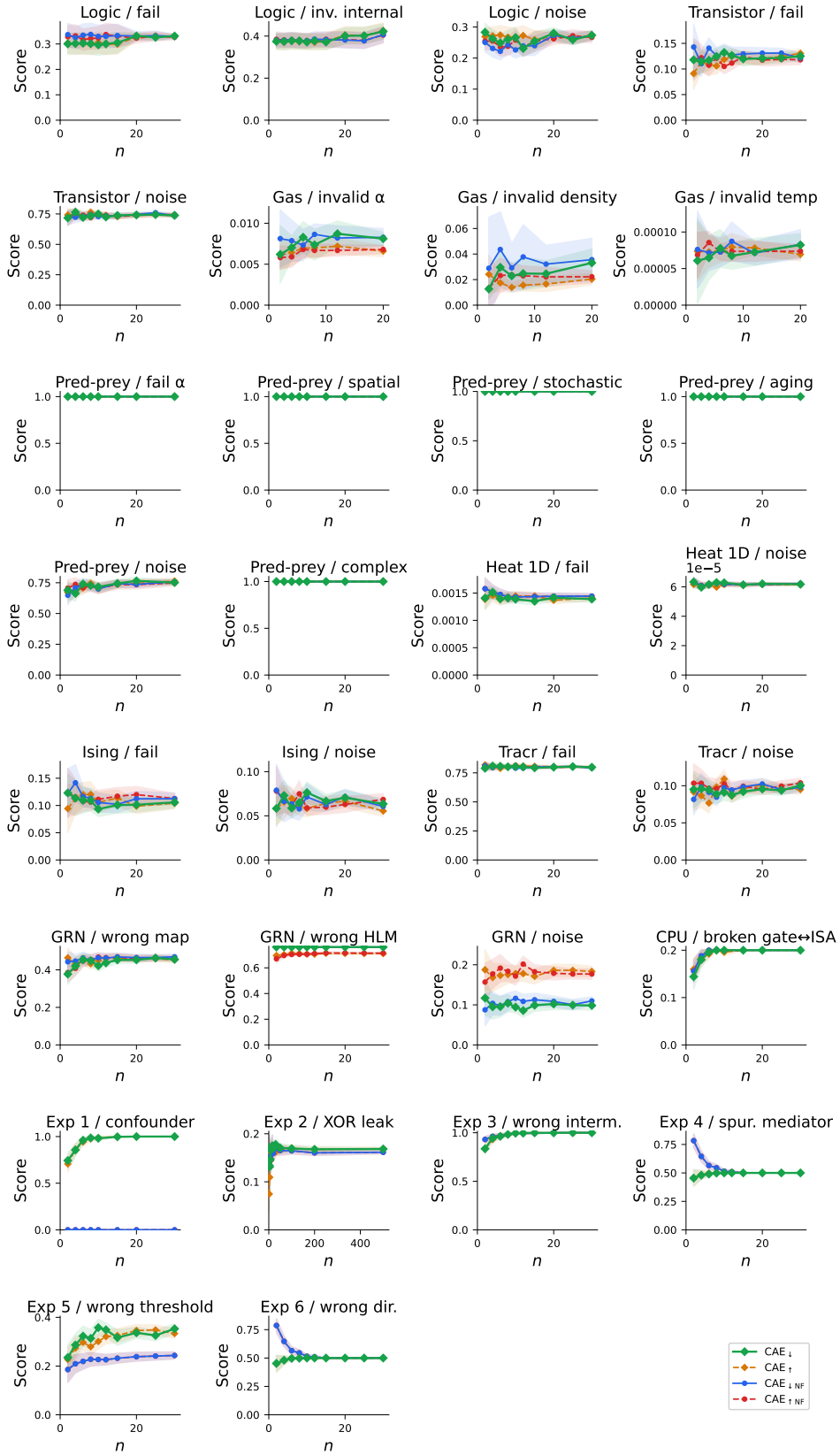

Figure 9: Convergence rate: For different numbers of sampled interventions n, we compute each metric 100 times on selected invalid abstractions.

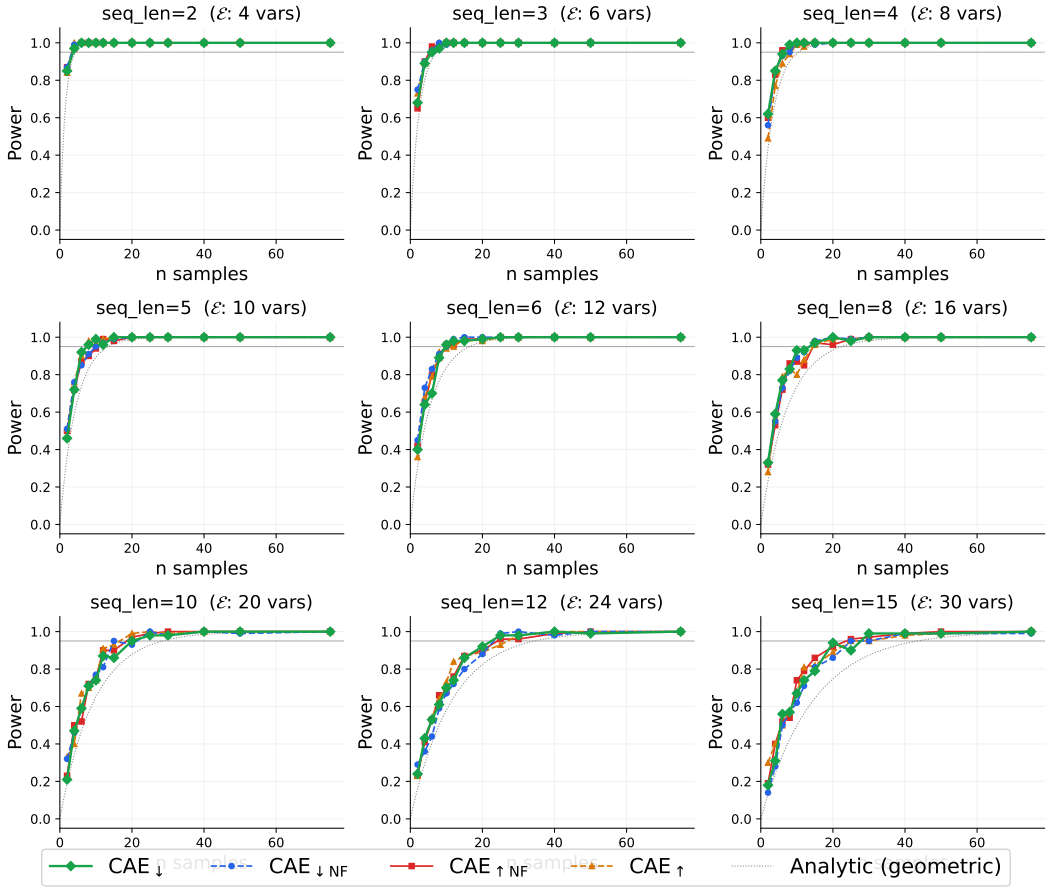

Figure 10: Detection power vs. the number of sampled interventions n across different sequence lengths for the Tracr sort-rank programs. The curves show the empirical probability of successfully detecting the invalid condition (where $\text{rank}_0 = \ell - 1$ when token_0 is the minimum).

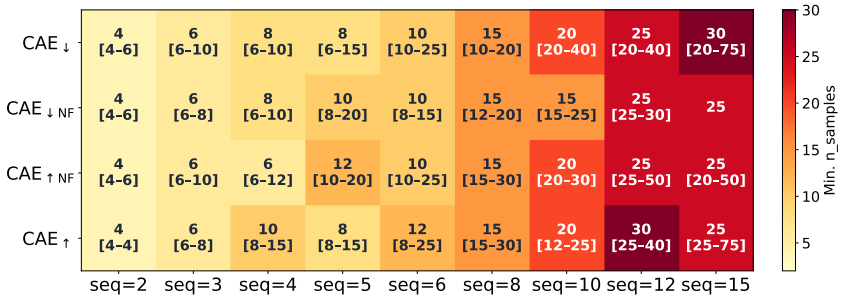

Figure 11: Minimum number of sampled interventions required to achieve 95% detection power on the Tracr sort-rank experiment, evaluated across different metrics (rows) and sequence lengths (columns). Numbers between brackets represent Wilson confidence intervals.

2019, Landajuela et al., 2022], where model selection is performed by minimizing reconstruction error on held-out data.

Complexity-regularized accuracy. Accuracy-only criteria permit arbitrarily complex explanations that overfit the observed data. Standard practice augments the reconstruction objective with a complexity penalty $\Omega(\mathcal{E})$, yielding criteria such as AIC [Akaike, 2003], BIC [Schwarz, 1978], and minimum description length [Mangan et al., 2017, Dong and Zhong, 2025]. *Mallow's* C_p statistic [Vallejo et al., 2014, Beula et al., 2024, Lee and Ghosh, 2009] provides a related bias–variance decomposition, penalizing the number of free parameters: $C_p = \mathrm{RSS}/\hat{\sigma}^2 - n + 2p$, where RSS is the residual sum of squares, p the number of parameters, and $\hat{\sigma}^2$ an estimate of noise variance. In our setting, C_p is applied to a linear proxy fitted between the predictions of $\mathcal{E}$ and the outputs of $\mathcal{M}$ so that it simultaneously penalizes residual mismatch and model complexity in a cross-model comparison: p is set to twice the number of variables in $\mathcal{E}$ (one slope and one intercept per variable), and σ^2 is estimated from the residuals of this proxy. Physics-inspired approaches embed domain-specific inductive biases, such as dimensional homogeneity, separability, known invariances, directly into the admissible hypothesis class, as in AI Feynman [Udrescu and Tegmark, 2020] and related frameworks [Tenachi et al., 2023, Imai Aldeia et al., 2025]. We also employ a *SINDy-style evaluator* [Brunton et al., 2016], which scores an explanation by how well its governing equations predict the empirical trajectory one epoch ahead. High residual indicates that the governing equations of $\mathcal{E}$ do not faithfully reproduce the system's epoch-level dynamics.

Distributional equivalence. When $\mathcal{M}$ is stochastic or the output of interest is a distributional quantity, pointwise comparison is insufficient. Observational validity is then defined by proximity between the output marginals $P_{\mathcal{E}(u)}$ and $P_{\mathcal{M}(u)}$. The *KL divergence* $D_{\mathrm{KL}}(P_{\mathcal{M}} \| P_{\mathcal{E}})$ and the symmetric *Jensen–Shannon divergence* D_{JS} measure discrepancy in a model-based sense. The *maximum mean discrepancy* $\mathrm{MMD}^2(\mathcal{M}, \mathcal{E}) = \|\mu_{P_{\mathcal{M}}} - \mu_{P_{\mathcal{E}}}\|^2_{\mathcal{H}}$, where μ_P is the kernel mean embedding of P in a reproducing kernel Hilbert space $\mathcal{H}$ [Gretton et al., 2012], provides a nonparametric alternative. The *Hilbert–Schmidt independence criterion* (HSIC) [Gretton et al., 2005] evaluates statistical dependence between residuals $\mathcal{M}(u) - \mathcal{E}(u)$ and the input u; a low HSIC indicates that errors are independent of the input, reflecting global rather than locally compensating agreement. These criteria subsume the *docking* paradigm in agent-based modeling [Axtell et al., 1996, Collins et al., 2024], where models are accepted as equivalent when they reproduce the same aggregate distributional patterns.

Temporal fidelity. For dynamical systems only, observational validity must be assessed over entire trajectories rather than marginal snapshots. *Trajectory MSE* measures the average squared deviation between state sequences $\{x_t^{\mathcal{M}}\}$ and $\{x_t^{\mathcal{E}}\}$ over a common time horizon. *Dynamic time warping* (DTW) [Sakoe and Chiba, 1978] generalizes this to temporally misaligned trajectories by computing the minimum-cost elastic alignment: $\mathrm{DTW}(\mathbf{x}^{\mathcal{M}}, \mathbf{x}^{\mathcal{E}}) = \min_\pi \sum_{(i,j)\in\pi} d(x_i^{\mathcal{M}}, x_j^{\mathcal{E}})$, where π ranges over admissible warping paths. *Temporal autocorrelation matching* assesses whether $\mathcal{E}$ reproduces the autocorrelation structure $\rho^{\mathcal{M}}(\tau) = \mathbb{E}[x_t x_{t+\tau}]$ of $\mathcal{M}$ across lags τ; failure indicates that $\mathcal{E}$ does not capture the system's memory structure. *Spectral analysis* compares the power spectral densities obtained via the Fourier transform of the respective autocorrelation functions [Percival and Walden, 1993]; agreement in dominant frequencies, harmonic structure, and spectral entropy indicates that the explanation captures the characteristic temporal modes of $\mathcal{M}$.

Limitation: equifinality. All observational criteria share a fundamental limitation: they are insensitive to the causal mechanisms generating the data. Distinct explanations may induce indistinguishable output distributions, a phenomenon termed *equifinality* [Valogianni and Padmanabhan, 2022, Collins et al., 2024]. Observationally valid models can violate the causal structure of $\mathcal{M}$ while remaining empirically accurate, leading to explanations that are unstable under distribution shift [Ghorbani et al., 2019, Kindermans et al., 2019, Zhang et al., 2025] and fragile to mechanistic perturbations [Méloux et al., 2025b,a, Katende, 2025].

B.2 Functional Validity

Functional validity strengthens observational equivalence by requiring that $\mathcal{E}$ reproduce not only the outputs of $\mathcal{M}$ but also its *input–output response profile*: how outputs vary as inputs are systematically modified. These criteria are most naturally applied to static or quasi-static input–output systems, though several generalizations to dynamical settings exist.

Variance decomposition and global sensitivity. *Variance decomposition* decomposes the total output variance $\mathrm{Var}[\mathcal{M}(u)]$ into contributions attributable to individual inputs and their interactions via the ANOVA expansion [Efron and Stein, 1981, Saltelli et al., 1999]. *Sobol sensitivity indices* [Sobol', 2001] formalize this as $S_i = \mathrm{Var}[\mathbb{E}[\mathcal{M}(u) \mid u_i]]/\mathrm{Var}[\mathcal{M}(u)]$ (first-order) and analogously for higher-order terms (interactions between input coordinates). Under this validity criterion, $\mathcal{E}$ is deemed valid if its Sobol indices $\hat{S}_i$ agree with those of $\mathcal{M}$ up to an acceptable tolerance, ensuring global functional coherence in how uncertainty in each input propagates to the output. Related global sensitivity methods including FAST [Cukier et al., 1973, Saltelli et al., 1999] and Morris screening [Morris, 1991] provide computationally cheaper approximations to the same underlying sensitivity structure.

Local attribution fidelity. A complementary notion evaluates agreement between $\mathcal{E}$ and $\mathcal{M}$ in a neighborhood of a fixed input u_0. The *infidelity* metric Yeh et al. [2019] is defined as

$$\mathrm{INFD}(\mathcal{E}, u_0) = \mathbb{E}_{\delta u}\Big[\big(\delta u^\top \phi(\mathcal{E}, u_0) - [\mathcal{M}(u_0) - \mathcal{M}(u_0 - \delta u)]\big)^2\Big],$$

where $\phi(\mathcal{E}, u_0)$ is the attribution vector of $\mathcal{E}$ at u_0 and δu is drawn from a perturbation distribution. Low infidelity indicates that the explanation's attributions accurately predict the actual output changes of $\mathcal{M}$ under local perturbations. We adapt infidelity to compare output sensitivities between $\mathcal{E}$ and $\mathcal{M}$ directly, without requiring attribution vectors. LIME Ribeiro et al. [2016] constructs a locally faithful surrogate by minimizing a locally weighted reconstruction error around u_0, and SHAP Lundberg and Lee [2017] defines attributions via Shapley values, which satisfy a consistent axiomatic decomposition of the output into per-feature contributions; both can be interpreted as proxies for local attribution fidelity.

Structural deviation and causal sensitivity index. While standard sensitivity indices quantify statistical influence under passive sampling, a complementary family of metrics evaluates the sensitivity of explanatory validity to the structural parameters of $\mathcal{E}$ itself [Katende, 2025]. The *structural deviation* metric perturbs each parameter of $\mathcal{E}$ by a small amount and measures the resulting change in alignment score, identifying which parameters are most important for the explanation's validity. The *causal sensitivity index* applies a harder test: each parameter is zeroed out in turn, and the drop in alignment score is recorded, providing a global decomposition of validity across the components of $\mathcal{E}$. Together, these metrics support model diagnosis by distinguishing robust explanatory components from those whose perturbation or removal would invalidate the abstraction.

Counterfactual and relational fidelity. *Counterfactual validity* [Wachter et al., 2017] requires that $\mathcal{E}$ reproduce the minimal perturbations δu that change the class label or output regime of $\mathcal{M}$: the decision boundaries of $\mathcal{E}$ must coincide with those of $\mathcal{M}$ in the relevant neighborhood. *Relational fidelity* [Collins et al., 2024] measures the Pearson correlation between $\Delta_{\mathcal{E}}(u, u') = \mathcal{E}(u') - \mathcal{E}(u)$ and $\Delta_{\mathcal{M}}(u, u') = \alpha(\mathcal{M}(u')) - \alpha(\mathcal{M}(u))$ over sampled pairs (u, u'), computed independently per output dimension and averaged. A score of 0 (after normalization) indicates perfect agreement in relational structure.

Limitation: structural underdetermination. Functional criteria remain agnostic to internal mechanisms. Distinct models can exhibit identical global sensitivity profiles, local attributions, and counterfactual boundaries while relying on entirely different underlying structures. Functional validity therefore does not uniquely identify the causal organization of $\mathcal{M}$, and multiple mechanistically incompatible explanations may be simultaneously admissible.

B.3 Information-Theoretic and Representational Validity

Another class of criteria requires that $\mathcal{E}$ reproduce not only the input–output behavior of $\mathcal{M}$ but also the structure of its internal representations: how information is encoded, compressed, and transmitted across the system.

Representational alignment and probing accuracy. A first approach tests whether the abstract variables posited by $\mathcal{E}$ are linearly decodable from the internal states of $\mathcal{M}$. In the case of neural networks, given a representation $h(u) \in \mathbb{R}^d$ extracted from a nominated layer of $\mathcal{M}$, *probing accuracy* measures the performance of a linear classifier or regressor trained to predict an explanatory variable from $h(u)$ [Alain and Bengio, 2018]. For real neural networks, neuroscience commonly

refers to these practices as *multivariate pattern analysis* [Haxby et al., 2001, Norman et al., 2006] or *brain signatures* [Wager et al., 2013, Kragel et al., 2018]. High probing accuracy indicates that z is explicitly represented in $\mathcal{M}$'s internal geometry [Ravichander et al., 2021, Belinkov, 2022]. For systems with continuous identity value maps, where all samples share a single abstract label, probing accuracy is trivially maximal and reflects the smoothness of the micro-model's representation rather than the discriminality of abstract categories. Supervised concept-level probing is instantiated by Concept Activation Vectors [Kim et al., 2018]; unsupervised variants based on sparse autoencoders (SAEs) [Cunningham et al., 2023, Bricken et al., 2023] recover a dictionary of monosemantic features without requiring labeled concepts. Representational similarity analysis (RSA) [Kriegeskorte et al., 2008] generalizes this by comparing representational dissimilarity matrices between $\mathcal{E}$ and $\mathcal{M}$, without committing to a specific linear structure [Schrimpf et al., 2021, Aw et al., 2024, Ryskina et al., 2025].

Information bottleneck and causal information bottleneck Lagrangians. The *information bottleneck* (IB) framework characterizes valid representations as those achieving an optimal trade-off between compression of the input and preservation of information relevant to the output [Tishby et al., 2000]. However, standard IB relies on observational mutual information, rendering it sensitive to spurious correlations in the input distribution. The *causal information bottleneck* (CIB) [Simoes et al., 2025] strengthens this by replacing observational relevance with causal relevance measured under interventions over inputs, so that $\mathcal{E}$ is assessed by its capacity to preserve causally invariant predictive information. Operationally, the IB Lagrangian is evaluated using the observational mutual information $I(\mathcal{E}(u); \mathcal{M}(u))$, whereas the CIB Lagrangian replaces this with the interventional quantity $I_c(Y \mid \mathrm{do}(T))$; this distinction is relevant when the input distribution contains spurious correlations that increase the apparent observational relevance of $\mathcal{E}$.

Complexity shift. *Algorithmic information dynamics* (AID) [Zenil et al., 2019, 2020, 2023] evaluates explanations in terms of changes in algorithmic complexity induced by interventions. The *complexity shift* metric is defined as the discrepancy between the change in Kolmogorov complexity K of the output under perturbation of $\mathcal{M}$ versus the corresponding change in $\mathcal{E}$:

$$\Delta K_{\mathcal{M}}(u, \delta u) = K(\mathcal{M}(u + \delta u)) - K(\mathcal{M}(u)), \qquad \mathrm{CS}(\mathcal{E}) = \mathbb{E}\left[\left|\Delta K_{\mathcal{M}} - \Delta K_{\mathcal{E}}\right|\right].$$

Low complexity shift indicates that $\mathcal{E}$ produces mechanistically comparable transformations under perturbation. In practice, Kolmogorov complexity is approximated via compression algorithms or block decomposition methods [Zenil et al., 2019], introducing estimation variance.

Limitation: mechanistic underdetermination. Information-theoretic criteria capture *what information* is present and how it is organized across variables, but not *how* that information is causally transformed by the system's dynamics. Distinct mechanisms can yield identical probing accuracies, identical IB optima, and identical complexity-shift profiles [Méloux et al., 2025a]. Moreover, most of these criteria, except in their explicitly causal variants, are grounded in observational regimes [Teney et al., 2022].

B.4 Causal and Interventional Validity

The strictest class of criteria requires that $\mathcal{E}$ correctly reproduce the behavior of $\mathcal{M}$ under *active interventions*. Informally, let $\mathcal{I}$ be a class of interventions on $\mathcal{M}$, and let $\mathcal{M}^{(i)}$ and $\mathcal{E}^{(i)}$ denote the respective post-intervention systems for $i \in \mathcal{I}$. Then $\mathcal{E}$ is *interventionally valid* relative to $\mathcal{I}$ if $P_{\mathcal{M}^{(i)}}(Y) \approx P_{\mathcal{E}^{(i)}}(\hat{Y}), \forall i \in \mathcal{I}$, where Y and $\hat{Y}$ are corresponding output variables under a fixed variable correspondence. Unlike observational criteria, this requirement is closed under the causal operator: agreement must hold across a family of intervention regimes.

Behavioral causal consistency and macroscopic invariance. A prerequisite for interventional validity is that the abstraction map τ must be well-defined with respect to the dynamics: distinct micro-states sharing the same abstract label must produce the same abstract output under intervention, regardless of the micro-realization. In the Neural Causal Model framework [Grosse-Wentrup et al., 2024], this is formalized as *behavioral causal consistency* (BCC): for every pair $x_1, x_2 \in \tau^{-1}(c), P(B[t] \mid \mathrm{do}(X[t] = x_1)) = P(B[t] \mid \mathrm{do}(X[t] = x_2))$. The same requirement appears independently in statistical physics under the name *macroscopic invariance*: a coarse-graining

is admissible only when all micro-configurations within a macro-state are observationally equivalent at the macro-level [Fisher, 1998]. BCC and macroscopic invariance differ in scope: macroscopic invariance tests each variable in isolation, intervening on a single variable while leaving others unspecified, whereas BCC tests each variable within a complete system state drawn from the joint distribution. Operationally, these are assessed by sampling pairs of micro-states with the same abstract label and verifying that their abstracted outputs of $\mathcal{M}$ agree under intervention.

Dynamic causal consistency. The *dynamic causal consistency* (DCC) criterion, introduced in the Neural Causal Model framework for dynamical systems [Grosse-Wentrup et al., 2024], requires that interventions on $\mathcal{E}$ correctly predict the next-step internal states of $\mathcal{M}$. Formally, DCC is satisfied if $\mathcal{E}^{(i)}(u_t) \approx h_{\mathcal{M}}(u_{t+1})$ for all interventions $i \in \mathcal{I}$ and time steps t, where $h_{\mathcal{M}}(\cdot)$ denotes the corresponding internal state of $\mathcal{M}$. DCC thus enforces step-wise causal coherence between the explanation's dynamics and those of the target system. It is a dynamical extension of behavioral causal consistency (BCC): where BCC requires that same-label micro-states produce identical behavioral outputs under intervention, DCC additionally requires that micro-states sharing the same abstract label at time t must also do so at $t + 1$ under the system dynamics. We evaluate this using exact label matching on the abstracted next state.

Constructive and formal verification. The strongest guarantees arise in settings where $\mathcal{E}$ can be analytically derived from or formally verified against $\mathcal{M}$. In physics, renormalization group methods and effective field theory [Fisher, 1998, Georgi, 1993] construct coarse-grained explanations that provably preserve interventional predictions. In computer science, formal verification [Woodcock et al., 2009] establishes that an implementation $\mathcal{M}$ satisfies a high-level specification $\mathcal{E}$ for *all* inputs, yielding universal behavioral equivalence guarantees. *Symbion* [Gritti et al., 2020] operationalizes a weaker but practically tractable version of this: it enforces consistency between concrete executions of $\mathcal{M}$ and symbolic executions of $\mathcal{E}$ at each program step, grounding abstract explanatory claims in observed concrete behavior. The validity score assigned by Symbion corresponds to the fraction of execution steps at which the symbolic and concrete states agree. In our evaluation, we implement a discrete analog suited to finite abstract domains: we exhaustively enumerate all combinations of abstract input labels, execute both models, and measure the fraction of input combinations for which they disagree on any output. In causal agent-based modeling [Valogianni and Padmanabhan, 2022], explanations are constructed to satisfy domain-theoretic structural constraints.

Mechanistic interpretability. Developed primarily in AI, mechanistic interpretability (MI) aims to reverse-engineer networks into human-interpretable algorithms [Olah et al., 2020]. MI assumes that for a given behavior of interest, a sparse subset of the network executes the relevant algorithm. This subset, often referred to as a *circuit*, is a key research target of modern AI interpretability methods Vig et al. [2020], Meng et al. [2022], Monea et al. [2024], Kramár et al. [2024], Conmy et al. [2023], Geva et al. [2023], Syed et al. [2023]. It is loosely inspired by neuroscience, which also seeks to uncover neural circuits underlying observed behaviors Yuste [2008]. A circuit is valid if it reproduces the network output under interventions on the computation, whether they are part or not of the circuit [Hanna et al., 2024]. *Intervention interchange accuracy* (IIA) [Geiger et al., 2022b] generalizes this perspective: it requires that pairwise interchange interventions on the network's internal representations and the high-level produce consistent outputs. Unlike IC and its variants, IIA evaluates interventional consistency without requiring a fully specified abstraction map over variables.

Computational mechanics and ϵ-machines. A more general perspective is provided by computational mechanics, in which valid explanations correspond to *causally closed* macro-level representations. The ϵ-machine framework [Crutchfield and Young, 1989, Shalizi and Crutchfield, 2001] defines the canonical valid explanation as the minimal sufficient statistic for predicting $\mathcal{M}$: the set of *causal states* $\mathcal{S}$ such that future behavior is conditionally independent of the past given $\mathcal{S}$. An explanation is valid under this criterion if its state space is isomorphic to the ϵ-machine of $\mathcal{M}$, ensuring that no micro-level information would improve predictive accuracy. Related perspectives interpret valid explanations as *programs implemented* by the underlying system, whose macro-level dynamics are self-contained and interventionally sufficient [Rosas et al., 2024].

Limitations While causal and interventional criteria provide the strongest notion of validity, existing formulations face practical limitations in complex, stochastic settings. Constructive approaches

require strong domain-specific assumptions. Mechanistic interpretability operates at the wrong level of abstraction, viewing low-level components as variables of the high-level explanation and, thus, suffer from causal overdetermination [McGrath et al., 2023, Méloux et al., 2025a]. Finally, computational mechanics and ϵ-machines are subsumed by the more general framework of causal abstraction.

C Controlled Experiments

We describe below the six controlled experiments from Figure 2. These highlight subtle ways in which abstraction can fail: aside from the strongest causal abstraction metrics, baseline metrics generally fail to detect the errors in these claimed abstractions.

Experiment 1: Hidden confounder. The macro-model posits that initial prey and predator populations determine the final populations (Lotka-Volterra dynamics). $\mathcal{M}$ introduces an unmapped resource variable (a phi variable) that multiplicatively scales the final prey population output by a factor of $\gamma \cdot r$, where γ is the coupling strength and r is the resource level. This creates a hidden confounder that the macro-model cannot account for. A valid metric should detect that the abstraction fails under perturbations to this resource variable. In the valid condition, the resource coupling is set to zero ($\gamma = 0$), making the resource variable inert.

Experiment 2: XOR backup path. The macro-model posits $X \to M \to Y$, with two phi variables b_1, b_2 (declared causally inert). In the valid $\mathcal{M}$, the XOR path is structurally disconnected ($y = a = x$), so perturbing b_1 and b_2 has no effect regardless of their values. In the invalid $\mathcal{M}$, Y is computed as $a \vee (b_1 \oplus b_2)$, where a is the micro-variable corresponding to M: a hidden XOR backup path to Y fires only when the phi variables are independently perturbed. This experiment specifically tests faithfulness: structural metrics that do not probe phi variables cannot detect the violation.

Experiment 3: Wrong intermediate variable. $\mathcal{M}$ implements the chain $x \to m = 2x \to z = m + 3$. The valid $\mathcal{E}$ matches this exactly. The invalid $\mathcal{E}$ posits $m = 3x$ with a compensating downstream equation ($z = \frac{2}{3}m + 3$), so that $z = 2x + 3$ in both cases. The intermediate m is wrong, but the output is identical under observation.

Experiment 4: Spurious mediator. $\mathcal{M}$ is fork-structured: $x \to y$ and $x \to z$ independently, so z does not depend on y. The invalid $\mathcal{E}$ incorrectly posits the chain $x \to y \to z$. Specifically, $y = 2x+1$ and $z = 3x + 2$ in $\mathcal{M}$; the invalid chain $\mathcal{E}$ posits $z = 1.5y + 0.5$, which reproduces $z = 3x + 2$ observationally but fails under interventions on y.

Experiment 5: Unreachable states. Both $\mathcal{E}$ and $\mathcal{M}$ share the chain $X \to M \to Z$ with $M \in \{0, 1, 2, 3\}$. THe dynamics of $\mathcal{M}$ only ever produce $M \in \{0, 2\}$ from any input $X \in \{0, 1\}$. The valid $\mathcal{E}$ uses the threshold $M \geq 2$; the invalid $\mathcal{E}$ uses $M \geq 1$, so they disagree only at $M = 1$, which is never reached from natural inputs. In micro-space, $\mathcal{M}$ uses a threshold of 1.5 to cleanly separate the $M \in \{0, 1\}$ subspaces from $\{2, 3\}$ subspaces.

Experiment 6: Wrong causal direction. $\mathcal{M}$ implements the chain $x \to y \to z$ (so z depends on y). The invalid $\mathcal{E}$ posits the fork $x \to y$, $x \to z$ (treating x as the common cause of both, ignoring $y \to z$). Specifically, $y = 2x + 1$ and $z = 1.5y + 0.5$ in $\mathcal{M}$; the invalid fork $\mathcal{E}$ posits $z = 3x + 2$, reproducing $z = 3x + 2$ observationally but failing under interventions on y.

D Benchmark Systems

For each system, we describe the high-level model $\mathcal{E}$, the low-level model $\mathcal{M}$, the abstraction map τ, the invalid conditions used to test discrimination, and key implementation details. All metrics are evaluated on a fixed number N of samples and averaged over n_{runs}. Their value, given below, were chosen to keep computational costs similar across systems.

D.1 Logic Circuit

A 2-bit ripple-carry adder evaluated at the wire level ($\mathcal{M}$) and the gate level ($\mathcal{E}$). This is the simplest discrete-to-discrete abstraction in the benchmark.

- $\mathcal{E}$: An SCM with integer-valued variables: two 2-bit operands ($\texttt{A}, \texttt{B} \in \{0, \ldots, 3\}$), a 1-bit carry-in, a 1-bit internal carry, a 2-bit sum output, and a 1-bit carry-out. Each variable is a deterministic Boolean arithmetic function of its parents.
- $\mathcal{M}$: A Boolean netlist simulator over individual wire states. Each wire carries a float value in $\{0.0, 1.0\}$, and gates are simulated by iterating until stable. Interventions force specific wire values before propagation.
- **Abstraction map:** The coarse-graining map groups pairs of 1-bit wires or single wires into variables of $\mathcal{E}$. The value map maps each integer label to the Cartesian product of unit-width subspaces: $(-0.1, 0.1)$ for bit-0 and $(0.9, 1.1)$ for bit-1. Internal gate wires are declared as internal variables and excluded from the phi set. Because every wire is either explicitly coarse-grained or declared internal, the phi set is empty.
- **Invalid conditions:**
  (i) *Fail*: $\mathcal{E}$ replaces XOR with OR in the sum logic.
  (ii) *Inverted internal*: $\mathcal{E}$ inverts `Internal_Carries` but compensates downstream, so the final output is correct, but the intermediate is wrong.
  (iii) *Noise*: Gaussian noise ($\sigma = 0.15$) is added to wire voltages.
- **Implementation details:** Metrics are evaluated over $N = 1000$ samples and averaged over 100 runs.

D.2 Transistor Circuit

A CMOS half-adder (sum + carry) simulated at the SPICE transistor level ($\mathcal{M}$) and the Boolean gate level ($\mathcal{E}$). This tests a continuous-to-discrete abstraction in which thresholding continuous voltages yields Boolean gate values.

- $\mathcal{E}$: A Boolean gate network with variables $\{a, b, \texttt{sum}, \texttt{carry}\} \in \{0, 1\}$, implementing XOR for sum and AND for carry.
- $\mathcal{M}$: A PySpice SPICE simulation of the CMOS half-adder topology. The topology uses four NAND gates and one inverter to implement the half-adder, and produces continuous node voltages for all circuit nodes (inputs, outputs, internal NAND wires, transistor junctions, and power rail). Internal nodes are declared as internal variables and excluded from the phi set.
- **Abstraction map:** The coarse-graining map associates $\{a, b, \texttt{sum}, \texttt{carry}\}$ nodes to the corresponding variables of $\mathcal{E}$. The value map maps label 0 to voltages in $(-0.5, 0.5)$ V and label 1 to $(4.5, 5.5)$ V. Voltages outside these ranges are unmapped.
- **Invalid conditions:**
  (i) *Fail*: $\mathcal{E}$ uses OR instead of XOR for sum.
  (ii) *Noise*: Gaussian noise ($\sigma = 1.5$ V) is added to circuit outputs.
- **Implementation details:** Metrics are evaluated over $N = 250$ samples and averaged over 50 runs.

D.3 Gas Simulation

An N-particle Lennard-Jones fluid ($\mathcal{M}$) evaluated against the ideal gas law or Van der Waals equation of state ($\mathcal{E}$). This is an extreme example of global aggregation: every variable of $\mathcal{E}$ aggregates over all N particles.

- $\mathcal{E}$: The ideal gas law $PV = NT$ (reduced units) or the Van der Waals equation $(P + a\rho^2)(1/\rho - b) = T$, with causal variables $\{P, V, T\}$. Unless specified, the causal grpah is reduced to the acyclic $(V, T) \to P$ forward macroscopic mapping (NVT ensemble).
- $\mathcal{M}$: An N-body molecular dynamics simulation using the 12-6 Lennard-Jones potential ($N = 128$ particles), velocity-Verlet integration, Langevin thermostat, and Berendsen barostat. Long-range tail corrections are applied analytically to the pressure virial to account for truncation of the

potential at cutoff $r_c = 3.0\sigma$ in reduced units. $\mathcal{M}$ outputs the measured P, T, V alongside particle positions and velocities. An energy minimization step (steepest descent, up to 500 iterations) is applied after lattice initialization to remove particle overlaps before dynamics begin.

- **Abstraction map:** The coarse-graining map is the identity on $\{P, V, T\}$ (these are directly computed macroscopic observables returned by the simulator: T via kinetic energy, P via the virial theorem, V = BoxLength^3). The value map is the identity.
- **Invalid conditions:**
  (i) *Wrong* α: the ideal gas $\mathcal{E}$ uses an incorrect temperature exponent.
  (ii) *High density*: particle density is raised outside the ideal gas regime.
  (iii) *Low temperature*: temperature is reduced to a regime where the ideal gas law breaks down.
- **Implementation details:** Van der Waals parameters a, b are calibrated against LJ fluid isotherms at the Boyle temperature ($T^* \approx 3.42$). Metrics are evaluated over $N = 10$ samples and averaged over 20 runs; a custom sampler initializes lattice configurations and hands equilibration to $\mathcal{M}$.

D.4 Predator-Prey Dynamics

A spatially-explicit agent-based model of predator-prey dynamics ($\mathcal{M}$) evaluated against the Lotka-Volterra ODE system ($\mathcal{E}$). The abstraction is a continuous identity mapping over total population counts: $\mathcal{M}$ outputs aggregate agent counts, which are the macro-variables.

- $\mathcal{E}$: The Lotka-Volterra equations, with parameters $\{\alpha, \beta, \delta, \gamma\}$ calibrated against the ideal (non-spatial, deterministic) ABM. The LV equations are integrated via the forward Euler method over 50 steps, matching the ABM's timestep count.
- $\mathcal{M}$: An agent-based model (ABM) of predator-prey dynamics in which individual prey and predators are tracked. Each timestep, births, predation events, and deaths are computed from per-capita probabilities. The simplest configuration (non-spatial, deterministic rates, no aging) approximates the LV equations and is used for calibration.
- **Abstraction map:** The coarse-graining map is the identity on $\{\texttt{prey_t}, \texttt{predator_t}, \texttt{final_populations}\}$. The value map is the identity.
- **Invalid conditions:** (i) *Wrong* α: prey reproduction rate in $\mathcal{E}$ is 3 times higher than in $\mathcal{M}$.
  (ii) *Spatial*: $\mathcal{M}$ uses spatial dynamics on a 20×20 grid, where agents move to and interact with neighboring cells.
  (iii) *Stochastic*: $\mathcal{M}$ uses stochastic reproduction.
  (iv) *Aging*: agents have finite lifespans.
  (v) *Noise*: Gaussian noise ($\sigma = 5$) on output populations.
  (vi) *Complex*: all three $\mathcal{M}$ extensions simultaneously active.
- **Implementation details:** Metrics are evaluated over $N = 50$ samples and averaged over 100 runs. LV parameters are calibrated via grid search.

D.5 Heat Equation (1D)

Brownian particle diffusion ($\mathcal{M}$) is abstracted to a finite-difference heat equation ($\mathcal{E}$). This tests a spatial aggregation abstraction.

- $\mathcal{E}$: A finite-difference heat equation solver on a uniform grid of K bins, parameterized by the diffusion coefficient D. The finite-difference solver uses Neumann boundary conditions (zero flux at both walls), consistent with the reflective particle boundaries in $\mathcal{M}$.
- $\mathcal{M}$: A system of $N = 1000$ Brownian particles diffusing in a 1D box $[0, L]$ with reflective boundaries over 200 time steps.
- **Abstraction map:** The value map bins particle positions into a normalized density histogram (one bin per cell in $\mathcal{E}$), and grounding samples particle positions from that histogram as a probability density function.
- **Invalid conditions:**
  (i) *Fail*: the diffusion coefficient in $\mathcal{E}$ is set to a wrong value of zero
  (ii) *Noise*: Gaussian noise ($\sigma = 0.05$) on the outputs of $\mathcal{M}$.

- **Implementation details:** Metrics are evaluated over $N = 30$ samples and averaged over 50 runs. $K = 50$ bins. The diffusion coefficient $D = 0.1$ is set to a known value matching the particle simulation's diffusion constant. Initial conditions are sampled as random Gaussian profiles (mean $\sim \mathrm{U}(0.3, 0.7)$, width $\sim \mathrm{U}(0.05, 0.15)$, grounded to particle positions and re-abstracted to obtain $\mathcal{E}$-compatible inputs.

D.6 Heat Equation (2D)

A phonon lattice model ($\mathcal{M}$) is abstracted to a 2D heat PDE ($\mathcal{E}$). This tests a spatial aggregation abstraction.

- $\mathcal{E}$: A 2D heat PDE solver with Neumann boundaries and a calibrated diffusion coefficient α.
- $\mathcal{M}$: A 16×16 phonon lattice model with harmonic springs, phonon scattering (rate $p_{\text{scatter}} = 0.2$), and $N_{\text{avg}} = 10$ simulation averages per sample.
- **Abstraction map:** The value map abstracts the kinetic energy map via Gaussian smoothing ($\sigma = 0.05 \times$ GridSize) and normalization; grounding uses a parameterized Gaussian source. The source parameters ($\texttt{source_x}, \texttt{source_y}, \texttt{source_E}$) use identity grounding.
- **Invalid conditions:**
  (i) *Fail*: the diffusion coefficient in $\mathcal{E}$ is set to a wrong value of $10\times$ the calibrated value
  (ii) *Noise*: Gaussian noise ($\sigma = 50$) on the output of $\mathcal{M}$.
- **Implementation details:** Metrics are evaluated over $N = 5$ samples and averaged over 20 runs. The diffusion coefficient α is calibrated via scalar optimization.

D.7 Ising Model

A hybrid molecular-dynamics and Monte-Carlo spin simulation ($\mathcal{M}$) evaluated against a rigid-lattice Ising model ($\mathcal{E}$). Both models operate over continuous thermodynamic variables; the internal discrete spins of $\mathcal{M}$ are not exposed as micro-variables.

- $\mathcal{E}$: A rigid $L \times L$ Ising model simulated via Metropolis-Hastings MCMC with coupling constant J. The causal variables are $\{\texttt{Temperature}, \texttt{ExternalField}, \texttt{PredictedMagnetization}\}$, all continuous.
- $\mathcal{M}$: A hybrid molecular-dynamics/Monte-Carlo simulator in which atomic positions evolve via Newtonian dynamics when vibrational temperature > 0, and spins are updated via Metropolis sweeps with distance-dependent coupling $J_{\text{eff}} = J_0 \exp(-\lambda|r - r_0|)$.
- **Abstraction map:** The coarse-graining map is the identity on all three continuous variables. The value map is the identity.
- **Invalid conditions:**
  (i) *Fail*: the coupling constant in $\mathcal{E}$ is set to $J = 2$ (double the true value).
  (ii) *Noise*: Gaussian noise ($\sigma = 0.05$) on the magnetization output of $\mathcal{M}$.
- **Implementation details:** $L = 8$. Metrics are evaluated over $N = 30$ samples and averaged over 150 runs. Sampling is performed over $[\texttt{Temperature}, \texttt{ExternalField}]$. The explicit Euler PDE solver uses adaptive sub-stepping: when the stability criterion $\alpha \Delta t / \Delta x^2 > 0.2$ is violated, the time step is subdivided accordingly.

D.8 Tracr Transformer

A transformer mechanistically compiled from a RASP sort-rank program ($\mathcal{M}$) evaluated against the symbolic program itself ($\mathcal{E}$). In this system, $\mathcal{M}$ is a neural network.

- $\mathcal{E}$: A symbolic sort-rank program over sequences of length $L = 3$: $\texttt{rank}_i = |\{j \neq i : \texttt{token}_j < \texttt{token}_i\}|$. Variables: $\texttt{token}_i \in \{1, \ldots, 5\}$ and $\texttt{rank}_i \in \{0, \ldots, 4\}$.
- $\mathcal{M}$: The Tracr-compiled JAX transformer, exposing two multi-dimensional micro-variables: $\texttt{tokens} \in \mathbb{R}^L$ (float encoding of token indices) and $\texttt{ranks} \in \mathbb{R}^L$ (float encoding of rank values).

- **Abstraction map:** The coarse-graining map slices $\texttt{token}_i \to \texttt{tokens}[i]$ and $\texttt{rank}_i \to \texttt{ranks}[i]$. The value map uses $[(v-1)-0.5, (v-1)+0.5]$ for token label v (tokens are 0-indexed internally: label $v=1 \to 0.0, v=5 \to 4.0$) and $[r-0.5, r+0.5]$ for rank label r. Tracr guarantees lossless round-trip encoding: grounding and abstraction are exact inverses for integer labels.
- **Invalid conditions:**
  (i) *Fail*: $\mathcal{E}$ uses $>$ instead of $<$ in the rank comparison (reversed ranks).
  (ii) *Noise*: Gaussian noise ($\sigma = 0.3$) on the residual stream before read-out.
- **Implementation details:** Metrics are evaluated over $N = 100$ samples and 150 runs. The sequence length is $L = 3$, the vocabulary $\{1, \ldots, 5\}$. Inputs are formed as $[\texttt{BOS}] + [\texttt{vocab_labels}]$; the BOS output position is discarded from the decoded result.

D.9 Gene Regulatory Network (GRN)

The segment polarity gene regulatory network of Sánchez et al. [2008] ($\mathcal{M}$) evaluated against a simple two-variable Boolean causal rule ($\mathcal{E}$). $\mathcal{M}$ is a synchronous multi-valued Boolean network (not a continuous ODE model).

- $\mathcal{E}$: A two-variable Boolean SCM: $\texttt{wg_src} \in \{0,1\}$ (Wg signal active?) $\to \texttt{fz_tgt} \in \{0,1\}$ (target Fz on?). Rule: $\texttt{fz_tgt} = \texttt{wg_src}$.
- **Low-level model:** One synchronous update step of the 6-cell segment polarity network. Variables are multi-valued Boolean (e.g. Wg $\in \{0,1,2\}$), and updates follow GINsim logical parameter semantics. Micro-variables exposed: $\texttt{wg_c2}$ (Wg level in cell 2, input) and $\texttt{fz_c1}, \texttt{fz_c3}, \texttt{fz_c4}, \texttt{fz_c5}$ (Fz in various cells, outputs).
- **Abstraction map:** The coarse-graining map links $[\texttt{wg_c2}] \to \texttt{wg_src}$ and $[\texttt{fz_c1}, \texttt{fz_c3}] \to \texttt{fz_tgt}$ (true neighbors of cell 2). The value map ($\texttt{WgFzValueMap}$) thresholds: $\texttt{wg_src} = 1$ iff $\texttt{wg_c2} \geq 1.9$ (effectively =2= 2 =2 for integer-valued states); $\texttt{fz_tgt} = 1$ iff any target Fz > 0.5.
- **Invalid conditions:**
  (i) *Wrong map*: the coarse-graining links $[\texttt{fz_c4}, \texttt{fz_c5}]$ (non-neighbors) instead of $[\texttt{fz_c1}, \texttt{fz_c3}]$.
  (ii) *Wrong* $\mathcal{E}$: the causal rule is reversed ($\texttt{fz_tgt} = 1 - \texttt{wg_src}$).
  (iii) *Noise*: Gaussian noise ($\sigma = 0.4$) on Fz outputs.
- **Implementation details:** Metrics are evaluated over $N = 200$ samples and averaged over 200 runs. Sampling is performed over $[\texttt{wg_src}]$.

D.10 MOS 6502 CPU

The MOS 6502 CPU, modeled at three discrete abstraction levels: the Visual6502 digital netlist, the manually extracted gate-level implementation, and a corresponding ISA-level simulation. All three levels are discrete: the transistor level uses binary switch segments, the gate level uses Boolean gate outputs, and the ISA level uses integer register and flag values (0–255). We evaluate abstractions between those three layers independently.

- **Highest-level model (ISA):** The ISA semantics of selected 6502 instructions, expressed as a causal SCM over register variables (accumulator, index registers, status flags) implemented by the ISA simulator (`ivop/fake6502`, commit b52676f).
- **Intermediate-level model (gate):** A gate-level simulator (`ivop/break6502`, commit 922af64), which implements the decoder PLA and datapath at the Boolean gate level.
- **Lowest-level model (transistor):** A transistor-level digital logic simulator (`mist64/perfect6502`, commit 6c1f1a5), extracted directly from the 6502 CPU via the Visual6502 project. This implements a segment model in which each transistor is a binary switch.
- **Abstraction maps:** All simulators share the same interface: integer register values 0–255 per variable. The coarse-graining map and value maps are both effectively identity on these integer values. All simulators mask the processor status register P with `0xCF` (clearing bits 4 and 5) before comparison, since bit 4 (the B-flag) is only meaningful during hardware interrupts and bit 5 is hardwired high; without this mask, all status-affecting instructions would appear to disagree.

- **Invalid conditions:**
  (i) *Broken gate* $\leftrightarrow$ *ISA*: accumulator bit 7 is stuck at 0 in $\mathcal{M}$.
  (ii) *Broken transistor* $\leftrightarrow$ *gate*: same fault applied at the transistor level.
- **Implementation details:** The abstraction is evaluated on a reduced set of single-instruction tests covering implied and immediate addressing modes. Metrics are evaluated over $N = 200$ samples and averaged over 100 runs (per condition).

E Baseline Metric Implementation Details

We document design choices, normalization functions, and known limitations for the baseline metrics. All metrics receive the same sampler, n samples, and output variable list as the CAE variants.

Observational metrics (MSE, RMSE, NMSE, R^2, L^2, JSD, KL, MMD, HSIC, VarDecomp). These are evaluated via a task wrapper that restricts interventions to root variables only and compares outputs at the designated output nodes.

JSD/KL: Uses adaptive histogram binning into $\lceil \log_2 n + 1 \rceil$ bins, capped at 20. KL adds Laplace smoothing ($\varepsilon = 10^{-10}$). When samples collapse to a point mass, JSD returns 0.

MMD: Uses an RBF kernel and bandwidth $\sigma = \sqrt{\text{median}(\|x_i - x_j\|^2)/2}$ from the pooled sample (median heuristic). The unbiased $\widehat{\text{MMD}}^2$ estimator is clipped to $[0, \infty)$, then passed through $\tanh$.

HSIC: Measures whether the residuals $R = \mathcal{M}(u) - \mathcal{E}(u)$ are statistically independent of the outputs of $\mathcal{E}$ (used as a proxy for the inputs u). The centered kernel alignment estimator $\widehat{\text{HSIC}} = \text{tr}(HK_XH \cdot HK_RH)/(n-1)^2$ is computed, where K_X is the RBF kernel matrix of the outputs of $\mathcal{E}$ and K_R is the RBF kernel matrix of residuals; both kernel matrices are centered by $H = I - \mathbf{1}\mathbf{1}^\top/n$. The estimator is normalized by the product of the Frobenius norms of HK_XH and HK_RH. Requires ≥ 20 samples; returns NaN below this threshold.

R^2*:* Normalized as $(1 - R^2)/2$, clipped to $[0, 1]$.

VarDecomp: Reports $\text{Var}(y - \hat{y})/\text{Var}(y)$, which is robust to constant offsets.

Temporal metrics (TrajMSE, DTW, Autocorr, Spectral, SINDy). These are implemented for the predator-prey system in a way that rolls out both $\mathcal{E}$ and $\mathcal{M}$ from shared initial conditions.

DTW: Standard dynamic programming with Euclidean local cost, normalized by the sum of sequence lengths $|s_1| + |s_2|$.

SINDy: For each sampled trajectory, we extract consecutive macro-step pairs (X_t, X_{t+1}) from the trajectory of $\mathcal{M}$ (ABM), where each step corresponds to one full model epoch. The observed epoch change is $\dot{X}_t = X_{t+1} - X_t$. The predicted change in $\mathcal{E}$ is $\hat{\dot{X}}_t = \nu(X_t) - X_t$, where $\nu(X_t)$ runs the LV equations for one full epoch starting from the actual ABM state X_t. Only steps where both species exceed 1 individual are scored. The normalized Frobenius residual $\|\dot{X} - \hat{\dot{X}}\|_F \,/\, \|\dot{X}\|_F$ is mapped through $\tanh(\cdot)$.

Sobol first-order indices: The Saltelli estimator shares the same A and B sample matrices for $\mathcal{E}$ and $\mathcal{M}$, making indices directly comparable. The metric reports mean $|S_i^{\mathcal{E}} - S_i^{\mathcal{M}}|$ over input variables.

Infidelity (two-model attribution adaptation): For each abstract input variable i, a finite-difference perturbation of size δ_i is applied to the abstract label: $\delta_i = \max(\delta_{\text{frac}} \cdot |x_i|, \delta_{\text{min}})$ for continuous labels (with $\delta_{\text{frac}} = 0.1$, $\delta_{\text{min}} = 10^{-3}$) and $\delta_i = 1$ for discrete (integer) labels. Both $\mathcal{E}$ and abstracted $\mathcal{M}$ are evaluated at the original and perturbed inputs; per-input attributions $\phi_i = (\text{output}_{\text{perturbed}} - \text{output}_{\text{original}})/\delta_i$ are computed for each model and compared by squared difference, averaged over input variables and output dimensions. The metric is normalized via $\tanh$. For the 6502 CPU, only register inputs $\{A, X, Y, S, P\}$ are perturbed; opcode and operand bytes are excluded because a perturbation of $\delta = 1$ changes the instruction identity, producing spurious attribution disagreement even for correct abstractions.

Relational fidelity: Pearson r per output dimension, averaged, mapped to $(1-r)/2 \in [0,1]$.

Mallow's C_p: $p = 2|\mathcal{V}^{\mathcal{E}}|$ (slope + intercept per variable of $\mathcal{E}$); $\hat{\sigma}^2$ is estimated as the mean variance of the outputs of $\mathcal{M}$ across output dimensions, serving as a reference noise floor. Normalized as $\tanh(\max(0, C_p - p)/N)$.

Structural deviation and causal sensitivity index: Structural deviation perturbs each parameter of $\mathcal{E}$ by 1% relative and measures $|\text{CAE}_{\uparrow\text{NF}}(\mathcal{E}_{\text{perturb}}) - \text{CAE}_{\uparrow\text{NF}}(\mathcal{E}_{\text{base}})|$. Causal sensitivity zeros each parameter instead, but measures the same quantity. Neither metric divides by the baseline $\text{CAE}_{\uparrow\text{NF}}$score.

IB vs. CIB Lagrangian: IB uses passive co-occurrence statistics for both $I(X;T)$ and $I(T;Y)$; a spurious correlate of Y in $\mathcal{M}$ can inflate $I(T;Y)$ even without causal influence. CIB replaces $I(T;Y)$ with $\text{CAE}_{\uparrow\text{NF}}(Y \mid \text{do}(T))$, estimated as $H(Y) - H_c(Y \mid \text{do}(T))$ where the conditional entropy averages over 20 label bins under direct interventions. Both Lagrangians are normalized to $[0,1]$: the IB result is shifted by $\beta H(Y)$ and divided by $H(X) + \beta H(Y)$; the CIB result is shifted by β and divided by $H(X) + \beta$, where $H(\cdot)$ denotes histogram entropy of the respective variable.

Complexity shift: For each sample, both the original input x and a Gaussian-perturbed input $x + \varepsilon$ ($\varepsilon \sim \mathcal{N}(0, 0.01^2)$) are run through both models. The complexity shift for each model is $(K(\text{output}_{x+\varepsilon}) - K(\text{output}_x))/K(\text{input})$, where $K(\cdot)$ denotes the zlib level-9 compressed byte length. The metric reports the absolute difference between the mean complexity shifts of $\mathcal{E}$ and $\mathcal{M}$. The metric is sensitive to output dimensionality and should be interpreted as a coarse ordinal measure only.

Symbion: Exhaustively enumerates all root-variable label combinations; applicable only to systems with finite discrete label spaces.

Macroscopic invariance vs. BCC: Both compare abstracted outputs of $\mathcal{M}$ for two micro-states in the same abstract cell. For discrete value maps, both metrics sample two micro-states directly from the subspace associated with each label; the difference is that macroscopic invariance iterates over individual labels, while BCC iterates over the full Cartesian product of labels and sample pairs. For continuous value maps, BCC draws a label via the sampler and calls the grounding function twice, and skips pairs where grounding is the identity, as these are trivially consistent.

DCC: Always uses hard label equality for the next-state comparison.

IIA: The interchange copies raw micro-values of $\mathcal{M}$ from the source run (not grounded preimages of the source abstract label). This tests whether the actual internal micro-state of $\mathcal{M}$ is causally exchangeable, which is a stronger condition than checking any preimage of the source label. This distinction vanishes for identity value maps.

Probing: For discrete value maps (logic circuit, Tracr, GRN), abstract labels are integers and probe predictions are rounded before scoring. For continuous identity value maps, labels are floats and the probe is a standard linear regression. The probe is trained on n_{train} samples (default 200, overridden per system) and evaluated on a held-out test set.

E.1 Computation Costs

All experiments were run on a consumer-grade laptop equipped with 16 GB of RAM and an Intel i5-13600H CPU. The computation time required to produce experimental results is estimated as such:

- Evaluation of baseline metrics on benchmark systems for valid and invalid abstraction: approx. 100 hours, of which nearly half was allocated to system 3 (gas simulation)
- Measurements of statistical power and convergence: approx. 50 hours, of which half for system 3
- Scaling laws for Tracr: Approx. 5 hours

In addition, functional tests were implemented and ran to check for correctness of the CPU implementations. There total runtime is estimated at approx. 50 hours.

Finally, preliminary experiments, including debugging and prototyping, is estimated to have required twice as much compute as the production of final results (approx. 250 hours).

F Case Study: Gas simulation (system 3)

We report additional experiments for the gas simulation system, which tests whether thermodynamic laws emerge as valid causal abstractions of a Lennard-Jones fluid simulation. Unless otherwise stated, experiments are performed at the Boyle temperature $T_{\text{Boyle}} \approx 3.418$ [Glasser, 2002] and with a low density of $\rho = 0.05$, where inter-particles interactions are rare and ideal thermodynamic behavior is expected.

We evaluate two high-level models: the ideal gas law (IGL, $PV \propto T$) and the Van der Waals equation (VdW, $(P + a/V^2)(V - b) \propto T$. These models represent successive levels of physical refinement: the IGL assumes non-interacting point particles, while VdW is a historical correction that explicitly accounts for finite particle volume and pairwise attraction. Neither is exact, but they serve as a useful pair to test whether CAE can discriminate the quality of competing valid abstractions.

For each equation, we consider three causal directions corresponding to different thermodynamic ensembles: NVT ($P \leftarrow f(V, T)$), NPT ($V \leftarrow f(P, T)$), and a non-standard inverse problem PVT ($T \leftarrow f(P, V)$), implemented via a proportional-gain temperature controller.

F.0.1 Valid Abstractions

Table 2 confirms that both models achieve near-zero CAE across all three causal directions in the dilute limit. This demonstrates that at low density, the validity of the statistical mechanical mapping matters more than the precise choice of equation of state. The VdW equation is marginally more faithful across most conditions, reflecting its tighter fit to the LJ fluid.

CAE	NVT ($P \leftarrow f(V, T)$)	NPT ($V \leftarrow f(P, T)$)	PVT ($T \leftarrow f(P, V)$)
Ideal gas law	0.0231	0.0397	0.0278
Van der Waals equation	0.0135	0.0520	0.0435

Table 2: $\text{CAE}_{\downarrow}$ for two high-level gas models across three causal directions (10 samples per evaluation).

F.0.2 Destructive Perturbations

We next systematically violated the physical assumptions required for thermodynamic laws to emerge from particle dynamics. Results are shown in Figure 12 and Table 3.

Density sweep (ρ). Increasing density from $\rho = 0.05$ to $\rho = 0.85$ in NVT mode introduces progressively stronger inter-particle interactions. The IGL error rises steadily, while VdW remains below 0.05 for significantly longer, exploiting its explicit correction terms. Both abstractions effectively fail (error ≈ 1) for $\rho > 0.45$.

Temperature sweep (T). At low temperatures ($T < 1.0$), particles cluster in attractive wells, violating the homogeneity assumptions of both laws, and both fail (error ≈ 0.8). As temperature increases, VdW recovers faster, reaching its minimum around $T \approx 2.5$. The IGL minimum falls in $[3.25, 3.75]$, consistent with the Boyle temperature where attractive and repulsive contributions cancel [Glasser, 2002]. Both models remain valid at high temperatures ($T > 4.0$, error < 0.1) where kinetic energy dominates.

Distorted exponent (α). Replacing the IGL's temperature dependence with $P \propto T^{\alpha}$ and sweeping $\alpha \in [0.5, 1.5]$ tests sensitivity to functional form. The CAE correctly identifies the physical ground truth, reaching near zero at $\alpha = 1.0$ and increasing monotonically away from it: sub-linear deviations ($\alpha < 1$) degrade error nearly linearly (reaching ≈ 0.7 at $\alpha = 0.5$), while super-linear deviations are penalized more gradually (≈ 0.42 at $\alpha = 1.5$).

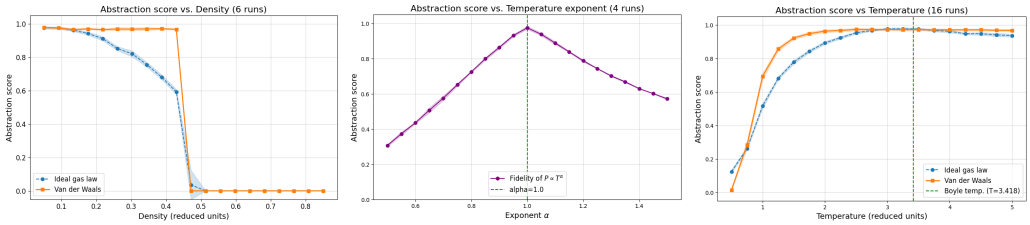

Figure 12: Average $\text{CAE}_{\downarrow}$ for IGL and VdW under (a) increasing density, (b) varying temperature exponent α, and (c) varying initial temperature T. All experiments in NVT mode, 10 samples per evaluation. Shaded areas denote $\pm 1\sigma$.

$\text{CAE}_{\downarrow}$	NVT ($P \leftarrow f(V,T)$)	NPT ($V \leftarrow f(P,T)$)	PVT ($T \leftarrow f(P,V)$)
Ideal gas law	0.5920	0.5432	0.5497
Van der Waals equation	0.5892	0.5478	0.5551

Table 3: $\text{CAE}_{\downarrow}$ for both models in non-equilibrium conditions (10 samples per evaluation).

Together, these results illustrate that $\text{CAE}_{\downarrow}$ behaves as a graded physical validity score: it correctly validates both abstractions in their regime of applicability, distinguishes the more faithful VdW model at intermediate densities and temperatures, and detects violations as soon as the underlying physical assumptions break down.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Each sentence clearly maps to a section of the article, in which the claim made in the sentence is investigated.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a short section with known limitations.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The appendices contain many details on the implementation of systems and baseline metrics; while not sufficient to ensure full reproducibility given the scope of the project, the authors commit to releasing the full library's code, including systems, baseline implementations and experimental results, upon acceptance.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The code and data are unreleased at the time of submission, but will be made available upon acceptance.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: The paper does not include typical machine learning experiments. However, baseline and implementation parameters are detailed in appendix.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The statistical power of the evaluated metrics is the subject of a dedicated subsection of the results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: A short appendix section details the estimated compute required to perform the experiments in the project.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: The main points of the Code of Ethics that apply to this work are the ones concerning reproducibility and access to research artifacts, both of which are treated in previous questions.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: The work in the paper has no direct negative societal impact; it contributes to the discussion on what is a meaningful valid explanation of a complex system, which helps advance the broad goals of science.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: There is no obvious risk of misuse for any data or models used in this work.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The work relies on external dependencies for the implementation of the MOS 6502 CPU; all three corresponding libraries are properly cited.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A].

Justification: No new assets are released at submission time.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: This research does not involve any human subject.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: This research does not involve any human subject.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [No]

Justification: The paper does not use any LLM as important, original, or non-standard component of the core methods.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.