

The Reconstruction Gap: Boundary Noise and the Limits of Holographic Bulk Recovery in Deep Learning, with Implications for Black-Box AI Auditing

Alfredo Merlet

alfredo.merlet@gmail.com

Universidad de Sevilla <https://orcid.org/0009-0007-8558-164X>

Research Article

Keywords: AdS/CFT correspondence, holographic bulk reconstruction, deep learning, inverse problems, identifiability, underspecification, AI auditing, black-box access

Posted Date: October 15th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11365793/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Abstract

Maldacena's conjecture holds that the interior (bulk) of an anti-de Sitter spacetime is encoded on its boundary. Holographic deep learning turns this idea into a learning problem: a network whose layers are radial slices of the spacetime learns the bulk metric from boundary data alone. This literature has asked whether the bulk can be recovered, not how much of it is lost when the boundary is observed imperfectly. A Scopus relaxation funnel finds no study of reconstruction under noisy boundary data, and none linking holographic reconstruction to AI auditing. This article defines the reconstruction gap, the relative error between the learned and the true metric by radial region, and measures it in two computational experiments with a known ground truth. A factorial design of 900 training runs varies boundary noise, sample size, depth and regularisation; a second experiment trains 400 networks from different initialisations on fixed data. With clean data, networks classify 99.1% of boundary points correctly while the near-horizon error remains about 30%. One per cent boundary noise costs half a percentage point of accuracy but nearly doubles the near-horizon error. A twenty-fold increase in sample size leaves that error unchanged, and the standard smoothness regulariser worsens it above roughly 7% noise. Among networks with indistinguishable in-distribution accuracy, interiors differ by a factor of more than three and out-of-distribution accuracy by up to 42 percentage points, while a global measure of structural error does not predict which networks fail. Read as a formal analogy, not an equivalence, the results quantify why audits restricted to black-box access cannot certify a model's internal mechanism, and why white-box access helps only when it asks targeted questions.

JEL classification: C45, C63, O33

1. Introduction

Maldacena's conjecture states that a gravitational theory in an asymptotically anti-de Sitter (AdS) spacetime is equivalent to a conformal field theory living on its boundary (Maldacena 1998; Witten 1998). The correspondence is usually summarised as holography: everything that happens in the higher-dimensional interior, the *bulk*, is encoded in data on the lower-dimensional *boundary*. Reading the interior back from the boundary, a task known as bulk reconstruction, is therefore an inverse problem, and in recent years it has become a testing ground for machine learning. Hashimoto et al. (2018a) showed that a deep neural network can be built so that its layers are slices of the radial direction of AdS and its weights are the bulk metric. Trained only on boundary data, such a network lets the metric emerge as its learned weights. The idea has since been extended to holographic QCD, transport coefficients, horizon areas, Wilson loops, entanglement entropy and spacetimes that are not asymptotically AdS (Hashimoto et al. 2018b; Yan et al. 2020; Hashimoto et al. 2021; Jejjala et al. 2025; Filev 2026; Ran et al. 2026).

This literature has asked whether a network *can* recover the bulk. It has asked much less often how much of the bulk is lost when the boundary is observed imperfectly. A bibliometric screen of Scopus documents the gap. A query combining AdS/CFT or holographic duality with deep learning, neural networks or machine learning returns 33 records. Adding bulk-reconstruction terms leaves 8. Adding noise, identifiability, robustness or uncertainty leaves none (Table 1). The closest work, Filev (2026), proves an exact, structural degeneracy: equal-time entanglement entropy fixes only the spatial part of the metric. That result concerns which functions the data can ever determine. The question addressed here is different and statistical: given an identifiable

architecture, how much interior is lost to finite, noisy boundary measurements, and where along the radial direction does the loss concentrate?

The question matters beyond physics. Holographic deep learning is one of the rare settings in which the internal structure of a trained network has a known ground truth: each weight is supposed to equal a specific value of a specific function. That makes it a controlled laboratory for a problem that dominates current debates on AI governance. External auditors of AI systems are usually granted only black-box access: they can query a model and observe its outputs, but not inspect its internals (Casper et al. 2024). Whether behaviour observed at the boundary of a system certifies anything about its interior is, in that setting, an open empirical question. Machine-learning research has shown that training pipelines can return many predictors with equally good test performance but different internal solutions, a property called underspecification (D'Amour et al. 2022; see also Breiman 2001). Holographic deep learning lets us measure this divergence exactly, because the interior is known.

This article defines and measures the *reconstruction gap*: the relative error between the bulk metric a network learns from boundary data and the true metric, resolved by radial region. A full factorial computational experiment varies boundary noise, sample size, network depth and regularisation over 900 training runs. A second experiment trains 400 networks from different random initialisations on a fixed data set to measure how many structurally different interiors are observationally equivalent at the boundary, and whether structural error predicts behaviour out of distribution. The findings are summarised in Section 5 and their limits in Section 6. The connection to AI auditing is developed in Section 6 explicitly as a formal analogy, not as a claim that AI models are holographic.

2. Background and related work

2.1 Holography and bulk reconstruction

In the AdS/CFT correspondence, correlation functions of the boundary field theory are computed from fields propagating in the bulk, with the radial coordinate playing the role of an energy scale (Maldacena 1998; Witten 1998; Harlow 2018). Reconstructing bulk geometry from boundary data has been approached analytically and, more recently, through the geometry of entanglement. Tensor-network constructions such as the multiscale entanglement renormalisation ansatz (Vidal 2007) were interpreted by Swingle (2012) as discrete holographic spacetimes, an early bridge between layered computational architectures and emergent geometry. Gan and Shu (2017) made the parallel between holography and deep learning explicit, and Hashimoto (2019) reformulated the correspondence as a deep Boltzmann machine.

2.2 The AdS/DL programme

The AdS/DL construction of Hashimoto et al. (2018a) discretises the radial direction of a black-hole spacetime into network layers. A scalar field and its radial derivative are propagated from the boundary to the horizon by the bulk equation of motion, which acts as the network's forward pass. The only free weights are the values of a metric function at each layer, so a trained network is directly interpretable as a geometry. Song et al. (2021) gave simplified examples. Later work replaced the discrete layers with neural ordinary differential equations (Chen et al. 2018; Hashimoto et al. 2021), applied the method to shear viscosity (Yan et al. 2020), to strange-

metal resistivity with physics-informed networks (Ahn et al. 2025), to equations of state (Bea et al. 2024), to linear-response data from proposed tabletop experiments (Hashimoto et al. 2025) and to minimal surfaces (Hashimoto et al. 2026). Jejjala et al. (2025) learned horizon areas from boundary time series. Filev (2026) reconstructed geometry from entanglement entropy and Wilson loops, reporting maximum relative errors below 0.2% in clean settings and identifying an exact degeneracy that additional observables resolve. Ran et al. (2026) extended data-driven reconstruction beyond asymptotically AdS spacetimes. Across this literature, validation is typically performed on clean or lightly perturbed synthetic data. No study in the screened corpus maps that deviation systematically against observational noise, sample size and depth.

2.3 Black-box access and underspecification

Algorithmic auditing frameworks treat access as a design variable (Raji et al. 2020). Casper et al. (2024) argue that black-box access, in which auditors only query a system and observe its outputs, is insufficient for rigorous audits, and that white-box access to weights and activations allows substantially more scrutiny. Rudin (2019) argues that explanations fitted to black boxes cannot be fully trusted and that interpretable models should be preferred in high-stakes settings. D'Amour et al. (2022) document that pipelines validated on held-out data from the training distribution can return many distinct predictors with equivalent test performance that behave very differently once deployed. The European Union's AI Act (Regulation (EU) 2024/1689) establishes conformity-assessment and documentation obligations for high-risk systems, which makes the evidential value of different forms of access a practical regulatory question. What this literature lacks is a setting in which the interior of a trained model can be compared with a known truth, so that the information lost between behaviour and structure can be measured rather than argued. Holographic deep learning provides one.

3. Bibliometric verification of the gap

The novelty claim was verified with a sequential relaxation funnel in Scopus (TITLE-ABS-KEY field, searched on 13 October 2026). Each query adds one conceptual block to the previous one; the record counts and their substantive content are reported in Table 1. Records were assigned to queries by their content and export time, and every record of the narrower queries was read at the abstract level.

Table 1
Scopus relaxation funnel (13 October 2026)

Query	Conceptual blocks	Records	Substantive reading
Q1	(AdS/CFT OR holographic duality OR gauge/gravity duality OR holographic correspondence OR Maldacena) AND (deep learning OR neural network OR machine learning)	33	About 20 records form the AdS/DL core; the rest are supergravity vacua, Calabi–Yau volumes and two off-topic consciousness papers
Q2	Q1 AND (bulk reconstruction OR bulk metric OR inverse problem OR emergent metric)	8	All address bulk reconstruction (e.g. Hashimoto et al. 2018a, 2018b, 2021; Filev 2026; Ran et al. 2026)
Q3	Q2 AND (noise OR identifiability OR robustness OR uncertainty)	0	No systematic study of reconstruction under noisy boundary data
Q4	(AdS/CFT OR holograph* OR bulk reconstruction) AND (neural network OR artificial intelligence) AND (audit* OR interpretab* OR black-box OR governance)	87	None links holographic reconstruction to AI auditing; matches are optical holography, plus two AdS papers that use "interpretable" or "black-box ODE solver" descriptively

Note. Source: author's searches in Scopus (Elsevier). The wildcard holograph* in Q4 deliberately casts a wide net; its 87 records were screened individually.

4. Methods

4.1 Bulk model and network architecture

The design follows Hashimoto et al. (2018a). The bulk is a four-dimensional planar AdS–Schwarzschild black hole with metric $ds^2 = -f(\eta)dt^2 + d\eta^2 + g(\eta)(dx^2 + dy^2)$, where η is the radial coordinate, the horizon sits at $\eta = 0$ and the boundary at $\eta \rightarrow \infty$. A real scalar field φ with mass $m^2 = -1$ and quartic self-interaction $\lambda = 1$ obeys the radial equation of motion

$$\partial_\eta \pi + h(\eta) \pi - m^2 \varphi - \lambda \varphi^3 = 0, \pi \equiv \partial_\eta \varphi \quad (1)$$

in which the geometry enters only through the metric function $h(\eta) = \partial_\eta \log \sqrt{f g^2}$. For this background the true metric function is

$$h(\eta) = 3 \coth(3\eta) \quad (2)$$

which tends to 3 near the boundary (the AdS value) and diverges as $1/\eta$ near the horizon. Eq. (1) is discretised with an explicit Euler scheme between $\eta_{\text{ini}} = 1$ and $\eta_{\text{fin}} = 0.1$ in N layers of width $\Delta\eta = -0.9/N$:

$$\varphi_{n+1} = \varphi_n + \Delta\eta \pi_n, \pi_{n+1} = \pi_n + \Delta\eta [-h_n \pi_n + m^2 \varphi_n + \lambda \varphi_n^3] \quad (3)$$

Each layer n is a network layer whose only trainable weight is h_n ; the nonlinearity is supplied by the interaction term. The network therefore has exactly N parameters, each of which has a physical meaning and a known true value $h(\eta_n)$.

4.2 Boundary data and the learning task

A boundary observation is a pair (φ, π) at η_{ini} , drawn uniformly from $\varphi \in [0, 1.5]$ and $\pi \in [-0.2, 0.2]$. Its label is $y = 1$ if the field is regular at the horizon, operationalised as $|\pi(\eta_{\text{fin}})| < \varepsilon = 0.1$, and $y = 0$ otherwise. Labels are always generated with the true metric on a fine grid of 200 layers, independent of the network's depth, so the network never inverts the same discretisation that produced its data (avoiding the so-called inverse crime). Training sets are balanced across the two classes. The network predicts the label through

$$P(y = 1 \mid \varphi, \pi) = \sigma(s(\varepsilon - |\pi_N|)), s = 40 \quad (4)$$

and minimises the binary cross-entropy plus, when regularisation is switched on, the smoothness penalty of Hashimoto et al. (2018a):

$$L(h) = \text{BCE} + c \sum_n \eta_n^4 (h_{n+1} - h_n)^2, c = 0.01 \times N/10 \quad (5)$$

The factor $N/10$ keeps the penalty comparable across depths, since finer grids shrink each difference $h_{n+1} - h_n$.

4.3 Training

Weights are initialised at $h_n \sim \text{Normal}(3, 1)$ and optimised with Adam (Kingma and Ba 2015), learning rate 0.05, 3,000 steps and mini-batches of 500. Gradients are computed exactly by backpropagation through the discretised equation of motion (the adjoint recursion of Eq. 3), implemented in NumPy; the analytic gradient matches central finite differences with a maximum absolute discrepancy below 10^{-9} .

4.4 Boundary noise

Noise enters as measurement error on the boundary inputs: each training observation is perturbed by Gaussian noise with standard deviation equal to a fraction v of the sample standard deviation of the corresponding coordinate, while labels remain those of the clean point. This is an errors-in-variables design: the auditor of a physical system, like the auditor of an AI system, observes behaviour through an imperfect measurement channel. Test accuracy is always computed on 5,000 clean, balanced boundary points.

4.5 Outcome: the reconstruction gap

For a trained network $\hat{h}$ the reconstruction gap at layer n is the relative error

$$G_n = |\hat{h}_n - h(\eta_n)| / h(\eta_n) \quad (6)$$

averaged over three radial regions of equal width: the boundary region ($\eta \geq 0.7$), the middle region ($0.4 \leq \eta < 0.7$) and the near-horizon region ($\eta < 0.4$). The overall gap averages G_n over all layers.

4.6 Experimental designs

Experiment 1 (factorial grid). Five noise levels $v \in \{0, 1, 5, 10, 20\}\%$ $\times$ three sample sizes $n \in \{500, 2,000, 10,000\}$ $\times$ three depths $N \in \{10, 20, 40\}$ $\times$ regularisation on/off $\times$ ten seeds, giving 900 training runs. For each seed the same clean data set is shared across noise levels, so noise effects are estimated within seed. Effects

are summarised by cell means and by ordinary least squares regressions of the log regional gap on the design factors,

$$\log G_r = \beta_0 + \beta_1 v + \beta_2 \text{REG} + \beta_3 (v \times \text{REG}) + \beta_4 \log n + \delta_N + u \quad (7)$$

with heteroskedasticity-robust (HC3) standard errors; r indexes the radial region and δ_N are depth fixed effects. Standard errors clustered by seed are reported as a robustness check, bearing in mind that ten clusters is a small number.

Experiment 2 (observational equivalence). One clean training set of 10,000 points is fixed, and 100 networks are trained from different random initialisations for each combination of depth (10, 20 layers) and regularisation (on/off), giving 400 runs. Networks whose clean test accuracy lies within one percentage point of the best network in the same configuration form an *observational-equivalence class*: an auditor with boundary access only could not rank them. Within each class the dispersion of structural error is reported, together with each network's accuracy on an out-of-distribution (OOD) test set of 4,000 balanced points drawn from $\varphi \in [1.5, 2.0]$, $\pi \in [0.4, 1.2]$, a region of boundary space that the training data never visit. Spearman rank correlations test whether structural error, which only white-box access reveals, predicts OOD behaviour better than in-distribution accuracy, which black-box access reveals.

Table 2
Fixed parameters and experimental factors

Element	Setting
Bulk geometry	Planar AdS ₄ –Schwarzschild; true metric function $h(\eta) = 3 \coth 3\eta$
Scalar field	$m^2 = -1$, $\lambda = 1$; radial range $\eta \in [0.1, 1]$
Label generator	True metric on 200 layers; regular if $ \pi(\eta_{\text{fin}}) < 0.1$
Boundary domain	$\varphi \in [0, 1.5]$, $\pi \in [-0.2, 0.2]$; balanced classes
Optimiser	Adam, learning rate 0.05, 3,000 steps, batch 500, $h_n \sim \text{Normal}(3, 1)$ at start
Factor: boundary noise v	0, 1, 5, 10, 20% of each coordinate's SD (inputs only)
Factor: sample size n	500, 2,000, 10,000
Factor: depth N	10, 20, 40 layers
Factor: regularisation	Off; on with $c = 0.01 \cdot N/10$
Replications	10 seeds per cell (Experiment 1: 900 runs); 100 initialisations per configuration (Experiment 2: 400 runs)
Test sets	5,000 clean in-distribution points; 4,000 out-of-distribution points (Experiment 2)

Note. All quantities are dimensionless. Source: author's design.

Reproducibility. All code, raw run-level results (including every learned metric vector) and figure scripts are deposited in the replication package (see Data availability). The complete design runs on a single CPU core.

5. Results

5.1 Clean boundary data: accurate behaviour, partial interior

Figure 1 compares the true metric function with the metric learned by the deepest networks. Without noise, the learned metric tracks the truth closely from the boundary down to about $\eta \approx 0.45$ and then bends away from it: where the true function rises towards its horizon divergence, the learned one flattens and falls. With noise the departure starts earlier and becomes a sign reversal, so the network learns an anti-damping term exactly where the true geometry damps most strongly.

Even in the most favourable condition, clean data with regularisation, the reconstruction gap is far from uniform. Averaged over depths and sample sizes, the relative error is 1.7% in the boundary region, 6.2% in the middle region and 29.6% in the near-horizon region (Table 3). The networks nevertheless classify 99.1% of clean test points correctly, against 99.4–99.8% for the true metric at the same depths. Behaviour at the boundary is almost perfect while a substantial part of the interior is not recovered.

5.2 Boundary noise: small behavioural losses, large structural losses

Table 3 reports accuracy and regional gaps by noise level. Noise degrades both, but at very different rates. With regularisation, 1% boundary noise lowers accuracy by 0.5 percentage points and multiplies the near-horizon gap by 1.86, from 0.296 to 0.552. At 5% noise, accuracy is still 94.6% while the near-horizon error exceeds 100% of the true value (1.119). At 20% noise accuracy falls to 85.4% and the error in every region is large.

Table 3
Accuracy and reconstruction gap by boundary noise and regularisation (means over 90 runs per cell: 3 depths $\times$ 3 sample sizes $\times$ 10 seeds)

Regularisation	Noise	Test accuracy	Gap, boundary ($\eta \geq 0.7$)	Gap, middle (0.4–0.7)	Gap, near horizon ($\eta < 0.4$)	Gap, all layers
Off	0%	0.991	0.269	0.267	0.360	0.297
Off	1%	0.986	0.262	0.260	0.589	0.363
Off	5%	0.945	0.315	0.245	1.025	0.512
Off	10%	0.910	0.533	0.248	1.168	0.637
Off	20%	0.848	1.235	0.448	1.305	1.002
On	0%	0.991	0.017	0.062	0.296	0.118
On	1%	0.986	0.022	0.062	0.552	0.198
On	5%	0.946	0.099	0.092	1.119	0.411
On	10%	0.911	0.202	0.144	1.329	0.532
On	20%	0.854	0.428	0.349	1.499	0.734

Note. Gap = mean relative error $|\hat{h} - h|/h$. Accuracy is measured on 5,000 clean in-distribution boundary points; its standard deviation across runs is at most 0.009 in every cell. The true metric's accuracy is 0.994, 0.996 and 0.998 at 10, 20 and 40 layers.

The radial profile in Fig. 2 shows where the information is lost. Without regularisation the learned metric is noisy at every depth, but the error grows steeply towards the horizon. Regularisation removes most of the error near the boundary in clean conditions, down to about 1%, yet leaves the near-horizon error intact. As noise increases, the error profile rises everywhere, and with regularisation a minimum appears in the middle of the radial range, around $\eta \approx 0.65$, rather than at the boundary.

5.3 Regression estimates

Table 5 estimates Eq. (7). Each percentage point of boundary noise raises the near-horizon gap by about 5.8% without regularisation ($\exp(0.0561) - 1$) and by about 7.5% with it, while it lowers accuracy by 0.72 percentage points. Regularisation reduces the boundary-region gap by about 90% ($\exp(-2.336) - 1$) and the middle-region gap by about 78% in clean conditions, but its effect on the near-horizon gap is small (-10%) and reverses as noise grows: the interaction term (0.0161 per point) cancels it at about 6.6% noise, and above that level regularised networks reconstruct the deep interior worse than unregularised ones. In the three gap equations, the noise, regularisation and interaction coefficients are significant at the 5% level or better with HC3 errors and remain so when standard errors are clustered by seed.

Table 5

OLS estimates of Eq. (7): log reconstruction gap by radial region, and test accuracy, on the design factors (900 runs)

Term	log gap, boundary	log gap, middle	log gap, near horizon	Test accuracy
Noise (pp)	0.0809*** (0.0023)	0.0267*** (0.0021)	0.0561*** (0.0020)	-0.0072*** (0.0001)
REG	-2.3362*** (0.0513)	-1.4937*** (0.0423)	-0.1061** (0.0441)	-0.0007 (0.0006)
Noise $\times$ REG	0.0859*** (0.0044)	0.0699*** (0.0033)	0.0161*** (0.0034)	0.0003*** (0.0001)
log n	-0.0540*** (0.0149)	-0.0843*** (0.0116)	-0.0048 (0.0112)	-0.0004* (0.0002)
Depth 20	-0.0609 (0.0462)	-0.1344*** (0.0360)	0.3080*** (0.0380)	-0.0019*** (0.0006)
Depth 40	-0.0026 (0.0419)	-0.0686* (0.0359)	0.3894*** (0.0367)	-0.0031*** (0.0006)
Constant	-1.0560*** (0.1207)	-0.7918*** (0.0960)	-0.8446*** (0.0913)	0.9932*** (0.0018)
Observations	900	900	900	900
R ²	0.857	0.755	0.603	0.978

Note. HC3 robust standard errors in parentheses. Noise in percentage points; REG = 1 when regularisation is on; depth fixed effects relative to 10 layers. * $p < 0.01$, $p < 0.05$, $p < 0.1$.

5.4 Sample size and depth

Sample size has almost no effect on the reconstruction gap (Table 4, panel B). Going from 500 to 10,000 boundary points, a twenty-fold increase, leaves the near-horizon gap unchanged (coefficient on $\log n = 0.005$, $p = 0.67$) and test accuracy unchanged. In the boundary and middle regions the effect is small (-5% and -8% per log unit). This null result for the deep interior is consistent with the errors-in-variables design: noise in the inputs biases the fit rather than only adding variance, and more draws from the same noisy channel do not remove a bias.

The depth result requires care. Deeper networks reach closer to the horizon (the last layer sits at $\eta = 0.19$ with 10 layers and at $\eta = 0.12$ with 40), where the true metric is largest and hardest to learn, so part of their larger near-horizon gap is mechanical. Panel A of Table 4 therefore also reports the gap on the common support $0.19 \leq \eta < 0.4$, which all depths share. On that support, and with clean data, deeper networks still recover the interior worse (0.13 at 10 layers against 0.30 at 20 and 40 layers, with regularisation), even though their behavioural accuracy is identical (99.1%) and the true metric classifies better with more layers. With noisy data the depth effect on the common support disappears or reverses. The conclusion is limited: in clean conditions, additional capacity bought no structural fidelity, and it bought no measurable behavioural gain either.

Table 4
Reconstruction gap by depth and by sample size (means over all other design factors)

Panel / level	REG	Test accuracy	True-metric accuracy	Gap, boundary	Gap, middle	Gap, near horizon	Near horizon, common support
A. Depth 10	Off	0.938	0.994	0.514	0.277	0.732	0.732
A. Depth 10	On	0.939	0.994	0.158	0.153	0.862	0.862
A. Depth 20	Off	0.936	0.996	0.528	0.282	0.929	0.678
A. Depth 20	On	0.937	0.996	0.151	0.133	0.976	0.755
A. Depth 40	Off	0.935	0.998	0.527	0.322	1.008	0.693
A. Depth 40	On	0.936	0.998	0.152	0.139	1.038	0.739
B. n = 500	Off	0.936	–	0.597	0.330	0.879	–
B. n = 500	On	0.938	–	0.151	0.172	0.962	–
B. n = 2,000	Off	0.937	–	0.494	0.278	0.889	–
B. n = 2,000	On	0.938	–	0.150	0.130	0.954	–
B. n = 10,000	Off	0.935	–	0.478	0.273	0.900	–
B. n = 10,000	On	0.936	–	0.159	0.123	0.960	–

Note. Panel A pools noise levels, sample sizes and seeds (150 runs per row); panel B pools noise levels, depths and seeds (150 runs per row). "Common support" restricts the near-horizon gap to $0.19 \leq \eta < 0.4$, the range covered by every depth.

5.5 Behaviour does not rank structure

Figure 3 plots each run's test accuracy against its near-horizon gap. Across the full design the two are strongly related (Spearman $\rho = -0.92$), but almost entirely through the noise level, which moves both at once. Holding noise at zero, accuracy carries no information about structural quality: across the 180 clean runs, the rank correlation between test accuracy and the overall reconstruction gap is $\rho = -0.01$ ($p = 0.88$). Among networks trained on clean data, an observer who sees only boundary behaviour cannot tell which interior is closer to the truth.

5.6 Observational equivalence and out-of-distribution behaviour

Experiment 2 trains 100 networks from different random initialisations on the same clean data set, for each depth and regularisation setting. Table 6 shows that every one of the 400 networks falls within one percentage point of the best network in its configuration: the in-distribution test accuracies span only 99.0% to 99.3%. All of them are observationally equivalent to an auditor with boundary access.

Their interiors are not equivalent. Without regularisation, the overall reconstruction gap ranges from 0.12 to 0.40 among 10-layer networks and from 0.19 to 0.39 among 20-layer networks, and two networks drawn at random differ in their learned metric by about 1.05 per layer on average, roughly a third of the metric's boundary value. Regularisation compresses this dispersion by a factor of about thirty (mean pairwise difference 0.036).

Table 6

Observational-equivalence classes in Experiment 2 (one fixed clean training set of 10,000 points; 100 random initialisations per configuration)

Depth	REG	Class size	Test accuracy (min–max)	Structural gap (min / median / max)	Mean pairwise $ \Delta\hat{h} $	OOD accuracy (min / median / max)	True metric: test / OOD	$\rho(\text{gap}, \text{OOD})$	$\rho(\text{test acc.}, \text{OOD})$
10	Off	100/100	0.991–0.993	0.119 / 0.218 / 0.401	1.050	0.684 / 0.941 / 0.998	0.994 / 0.924	-0.14 (0.17)	0.06 (0.583)
10	On	100/100	0.990–0.992	0.052 / 0.068 / 0.086	0.036	0.552 / 0.687 / 0.977	0.994 / 0.924	-0.18 (0.08)	0.36 (0.000)
20	Off	100/100	0.991–0.993	0.192 / 0.272 / 0.392	1.063	0.621 / 0.950 / 0.999	0.996 / 0.966	0.06 (0.53)	-0.12 (0.242)
20	On	100/100	0.991–0.993	0.094 / 0.108 / 0.125	0.036	0.675 / 0.966 / 0.998	0.996 / 0.966	0.10 (0.30)	-0.13 (0.207)

Note. The equivalence class contains the networks whose clean test accuracy is within 1 percentage point of the best network in the configuration; here it contains all 100. OOD = out-of-distribution test set ($\varphi \in [1.5, 2.0]$, $\pi \in [0.4, 1.2]$). ρ = Spearman rank correlation with OOD accuracy (p-value in parentheses).

Out of distribution, the equivalent networks diverge sharply. Accuracy on the out-of-distribution boundary region ranges from 68% to 99.8% for unregularised 10-layer networks, and from 55% to 98% for regularised ones, against 92.4% for the true metric at that depth. Identical in-distribution behaviour therefore hides differences of up to 42 percentage points in behaviour on inputs the training data never covered. This is underspecification in the sense of D'Amour et al. (2022), measured in a setting where the interior is known.

The pre-specified test of whether structural error predicts out-of-distribution behaviour returns a null result. The rank correlation between the overall reconstruction gap and out-of-distribution accuracy is small and not significant at the 5% level in any of the four configurations (ρ between -0.18 and 0.10). In-distribution accuracy predicts out-of-distribution accuracy in only one configuration (regularised, 10 layers: $\rho = 0.36$, $p < 0.001$). An exploratory breakdown by radial region finds some significant correlations, but their sign and location change from one configuration to another, so they are not interpreted. Figure 4 shows the absence of a monotone relation.

Two observations explain why a single structural summary fails. First, several unregularised networks reach out-of-distribution accuracies well above that of the true metric discretised at the same depth (up to 99.8% against 92.4%). At finite depth the behaviourally best interior is not the true interior: the network learns an effective metric that partly compensates for the error of the discretisation itself. Second, regularisation, which brings the learned metric closer to the truth on average, also concentrates the regularised 10-layer networks at low out-of-distribution accuracy (median 69%). Fidelity in the interior that the training data probe does not transfer automatically to inputs that probe it differently.

6. Discussion

6.1 What the reconstruction gap shows about holographic deep learning

Three results bear on the AdS/DL programme itself. First, the gap is radially structured: the deep interior near the horizon is the region least determined by boundary data, and it is also the region most sensitive to measurement noise. Second, the sensitivity is strongly asymmetric. Boundary noise that costs half a percentage point of accuracy nearly doubles the near-horizon error, so a network can look excellent at the boundary while its reconstruction of the deep geometry has already failed. Third, the obvious remedies do not close the gap. Twenty times more data leaves the near-horizon error unchanged, consistent with a bias rather than a variance problem; additional layers bring no structural gain in clean conditions; and the smoothness regulariser that is standard in the programme helps near the boundary but, above about 6–7% noise, worsens the deep interior. Applications that feed real or experimental data into holographic networks, such as the proposed tabletop quantum-gravity experiments (Hashimoto et al. 2025) or condensed-matter transport data (Ahn et al. 2025), should therefore report reconstruction uncertainty by radial region and under realistic noise, not only accuracy on clean benchmarks.

6.2 A formal analogy for black-box AI auditing

The analogy proposed here is formal and limited. AI models are not holographic spacetimes, and nothing in this study implies that their internal representations obey the AdS/CFT correspondence. What the two settings share is the structure of an inverse problem: an observer with access to inputs and outputs only (the boundary) tries to make claims about the internal mechanism that produces them (the bulk). Holographic deep learning is valuable as a laboratory because it is one of the few settings where the internal mechanism has a known ground truth weight by weight.

Read in that limited sense, the results support three propositions for audit design. (i) *Behavioural evidence underdetermines structure*. Among networks trained on clean data, test accuracy carries no information about which interior is closer to the truth ($\rho = -0.01$), and 400 networks with accuracies within one percentage point of each other differ in their interiors by a factor of more than three. This is a quantitative counterpart to the argument of Casper et al. (2024) that black-box access is insufficient for rigorous audits. (ii) *Behavioural equivalence in testing does not certify behaviour in deployment*. The same equivalent networks differ by up to 42 percentage points out of distribution, which is the risk D'Amour et al. (2022) document for production pipelines and which conformity assessments for high-risk systems under the EU AI Act must confront. (iii) *White-box access is necessary but not sufficient*. Seeing the weights reveals the structural dispersion that boundary behaviour hides, but a single global summary of structural error did not predict out-of-distribution failures. Useful white-box audits must ask targeted questions about the specific parts of the mechanism that new inputs will exercise, which echoes the case of Rudin (2019) for models whose components have a meaning that can be checked.

These propositions are offered as hypotheses that the holographic laboratory makes precise, not as findings about deployed AI systems. Testing them requires interpretable architectures whose internal parameters have known targets, a condition that most deployed models do not meet.

6.3 Limitations

The study has five main limitations. (1) It uses a single bulk geometry, a single scalar field and a single learning task, the binary horizon-regularity classification of Hashimoto et al. (2018a); other observables, such as the correlators or entanglement data used by Filev (2026) and Ran et al. (2026), may be more or less robust to noise. (2) Noise is Gaussian and enters only the inputs; label noise and systematic measurement errors are not studied. (3) The networks use a discrete Euler scheme with at most 40 layers; part of the measured gap reflects the difference between the behaviourally optimal effective metric at finite depth and the continuum metric, as Experiment 2 shows, and neural-ODE formulations may separate the two more cleanly. (4) Hyperparameters (learning rate, number of steps, regularisation strength, loss steepness) were fixed a priori and not tuned per condition, so some gaps could be reduced by tuning. (5) The out-of-distribution test uses one region of boundary space, and the exploratory regional correlations were not pre-specified. All results come from synthetic data generated by known equations; this is by design, because only synthetic data provide the ground truth needed to measure the gap.

7. Conclusion

Holographic deep learning turns Maldacena's conjecture into a learning problem in which a network reads the interior of a spacetime from its boundary. This article measured how much of that interior is lost. With clean data, the boundary region is recovered to within a few per cent while the near-horizon region retains an error of about 30%, even though behaviour at the boundary is almost perfect. Boundary noise widens the gap far faster than it reduces accuracy; more data do not close it; and the standard regulariser helps near the boundary but harms the deep interior once noise exceeds about 6–7%. A second experiment showed that hundreds of networks with indistinguishable in-distribution behaviour differ substantially in their interiors and by up to 42 percentage points out of distribution, and that a global measure of structural error does not predict which of them fail.

For physics, the results argue for reporting holographic reconstructions with region-specific uncertainty under realistic noise. For AI governance, they offer a controlled, quantitative illustration of why audits restricted to input–output behaviour cannot certify the internal mechanism of a model, and why access to internals is valuable only when it is used to ask targeted questions. Future work should extend the design to neural-ODE architectures, to other holographic observables and to label noise, and should test whether the audit propositions hold in interpretable AI architectures with known internal targets.

Declarations

Competing interests.

The author declares no competing interests.

Ethics approval.

Not applicable: the study uses only computer-generated data and involves no human participants, animals or personal data.

Use of artificial intelligence.

The author used Claude (Anthropic) for assistance with code development, drafting and English-language editing. The tool was used to accelerate implementation and writing. All code, results, references and text were reviewed and verified by the author, who takes full responsibility for the content of this article.

Funding.

This research received no specific funding.

Author contributions.

A.A.M.E. conceived the study, designed the experiments, verified the literature, interpreted the results and wrote the manuscript.

Data availability.

All data generated and analysed in this study, including every learned metric vector, are openly available in Zenodo at <https://doi.org/10.5281/zenodo.23343342> under a CC BY 4.0 licence. No third-party data were used.

Code availability.

The complete code (network, training, both experiments, analysis and figures) is included in the same Zenodo record, <https://doi.org/10.5281/zenodo.23343342>.

References

1. Ahn B, Jeong H-S, Ji C-W, Kim K-Y, Yun K (2025) Deep learning-based holography for T-linear resistivity. *Phys Rev D* 112(12). <https://doi.org/10.1103/nh4n-ynlz>
2. Bea Y, Jimenez R, Mateos D, Liu S, Protopapas P, Tarancón-Álvarez P, Tejerina-Pérez P (2024) Gravitational duals from equations of state. *J High Energy Phys* 2024(7):87. [https://doi.org/10.1007/JHEP07\(2024\)087](https://doi.org/10.1007/JHEP07(2024)087)
3. Breiman L (2001) Statistical modeling: The two cultures. *Stat Sci* 16(3):199–231. <https://doi.org/10.1214/ss/1009213726>
4. Casper S, Ezell C, Siegmann C, Kolt N, Curtis TL et al (2024) Black-box access is insufficient for rigorous AI audits. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. ACM. <https://doi.org/10.1145/3630106.3659037>
5. Chen RTQ, Rubanova Y, Bettencourt J, Duvenaud D (2018) Neural ordinary differential equations. In *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*
6. D'Amour A, Heller K, Moldovan D, Adlam B, Alipanahi B et al (2022) Underspecification presents challenges for credibility in modern machine learning. *J Mach Learn Res* 23(226):1–61
7. European Union (2024) Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official J Eur Union, L series, 12 July 2024.
8. Filev VG (2026) Holographic entanglement entropy, Wilson loops, and neural networks. *J High Energy Phys* 2026(9):73. [https://doi.org/10.1007/JHEP09\(2026\)073](https://doi.org/10.1007/JHEP09(2026)073)
9. Gan W-C, Shu F-W (2017) Holography as deep learning. *Int J Mod Phys D* 26(12):1743020. <https://doi.org/10.1142/S0218271817430209>
10. Harlow D (2018) TASI lectures on the emergence of bulk physics in AdS/CFT. *Proceedings of Science, TASI2017*
11. Hashimoto K (2019) AdS/CFT correspondence as a deep Boltzmann machine. *Phys Rev D* 99(10):106017. <https://doi.org/10.1103/PhysRevD.99.106017>
12. Hashimoto K, Hu H-Y, You Y-Z (2021) Neural ordinary differential equation and holographic quantum chromodynamics. *Mach Learning: Sci Technol* 2(3):035011. <https://doi.org/10.1088/2632-2153/abe527>
13. Hashimoto K, Kyo K, Murata M, Ogiwara G, Tanahashi N (2026) Physics-informed neural network solves minimal surfaces in curved spacetime. *Mach Learning: Sci Technol* 7(1):015013. <https://doi.org/10.1088/2632-2153/ae3050>
14. Hashimoto K, Matsuo K, Murata M, Ogiwara G, Takeda D (2025) Machine-learning emergent spacetime from linear response in future tabletop quantum gravity experiments. *Mach Learning: Sci Technol* 6(1):015030. <https://doi.org/10.1088/2632-2153/adb09f>
15. Hashimoto K, Sugishita S, Tanaka A, Tomiya A (2018a) Deep learning and the AdS/CFT correspondence. *Phys Rev D* 98(4):046019. <https://doi.org/10.1103/PhysRevD.98.046019>

16. Hashimoto K, Sugishita S, Tanaka A, Tomiya A (2018b) Deep learning and holographic QCD. *Phys Rev D* 98(10):106014. <https://doi.org/10.1103/PhysRevD.98.106014>
17. Jejjala V, Mondkar S, Mukhopadhyay A, Raj R (2025) Learning holographic horizons. *Phys Rev D* 111(2):026016. <https://doi.org/10.1103/PhysRevD.111.026016>
18. Kingma DP, Ba J (2015) Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*
19. Maldacena J (1998) The large N limit of superconformal field theories and supergravity. *Advances in Theoretical and Mathematical Physics*, 2
20. Raji ID, Smart A, White RN, Mitchell M, Gebru T, Hutchinson B, Smith-Loud J, Theron D, Barnes P (2020) Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. ACM. <https://doi.org/10.1145/3351095.3372873>
21. Ran C, Wu S-F, Xian Z-Y (2026) Learning geometries beyond asymptotic AdS. *J High Energy Phys* 2026(3):31. [https://doi.org/10.1007/JHEP03\(2026\)031](https://doi.org/10.1007/JHEP03(2026)031)
22. Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* 1(5):206–215. <https://doi.org/10.1038/s42256-019-0048-x>
23. Song M, Oh MSH, Ahn Y, Kim K-Y (2021) AdS/Deep-Learning made easy: Simple examples. *Chin Phys C* 45(7):073111. <https://doi.org/10.1088/1674-1137/abfc36>
24. Swingle B (2012) Entanglement renormalization and holography. *Phys Rev D*, 86
25. Vidal G (2007) Entanglement renormalization. *Phys Rev Lett*, 99
26. Witten E (1998) Anti-de Sitter space and holography. *Advances in Theoretical and Mathematical Physics*, 2
27. Yan Y-K, Wu S-F, Ge X-H, Tian Y (2020) Deep learning black hole metrics from shear viscosity. *Phys Rev D* 102(10):101902. <https://doi.org/10.1103/PhysRevD.102.101902>

Figures

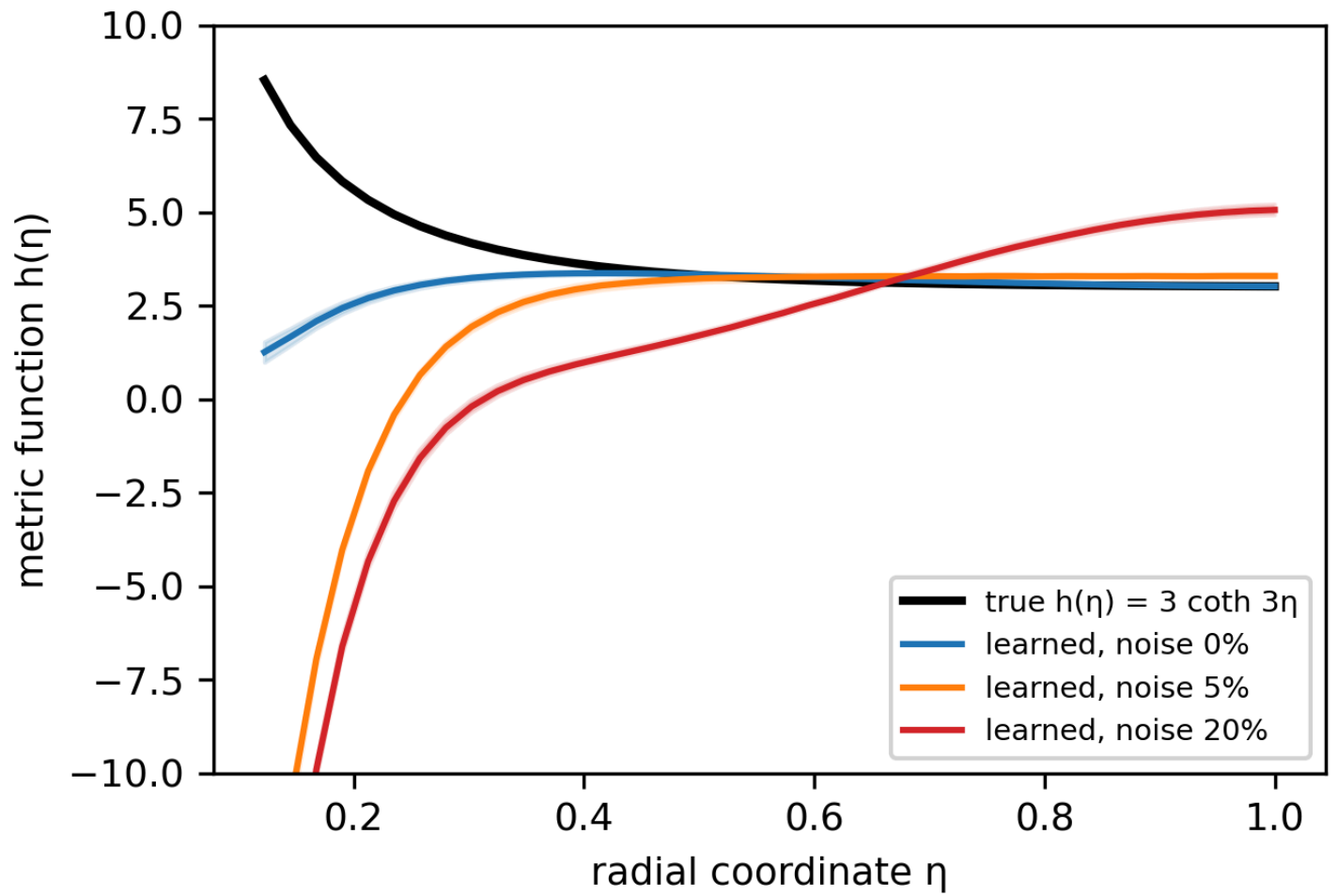


Figure 1

True metric function $h(\eta) = 3 \coth 3\eta$ (black) and learned metric $\hat{h}(\eta)$ for 40-layer networks with regularisation trained on 10,000 boundary points, at 0%, 5% and 20% boundary noise. Lines are means over ten seeds; shaded bands are ± 1 standard deviation. The horizon is at $\eta = 0$ and the boundary at large η . The vertical axis is truncated at ± 10 .

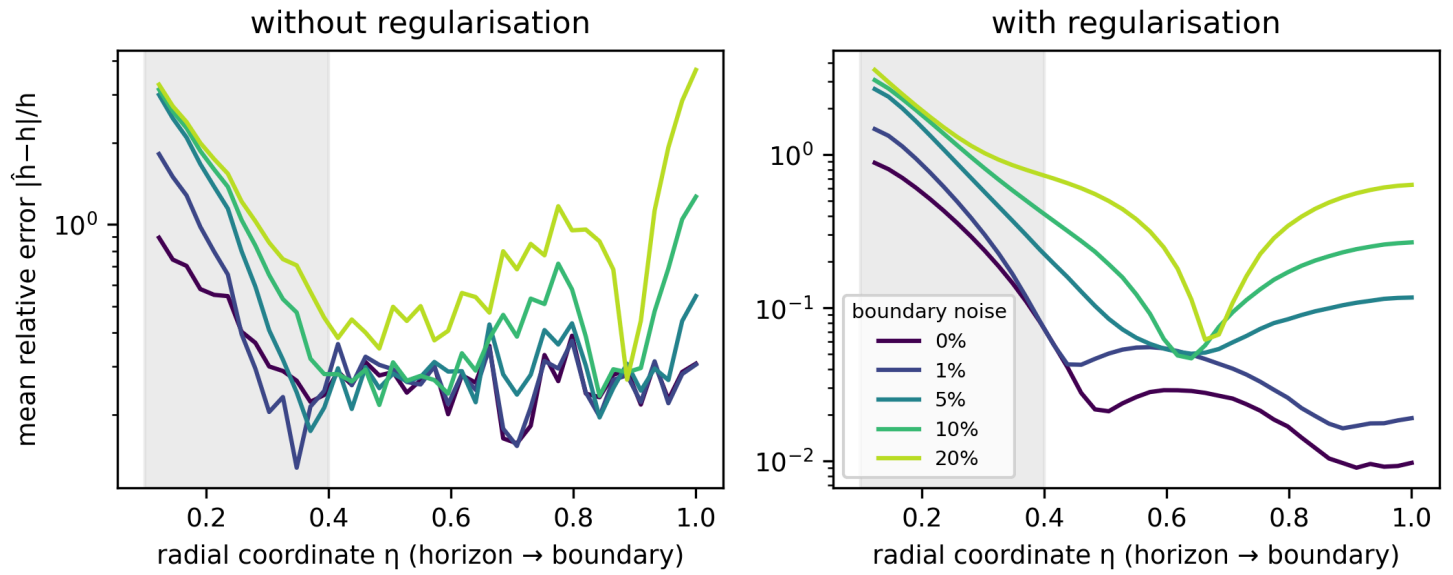


Figure 2

Radial profile of the reconstruction gap for 40-layer networks, by boundary noise level, without (left) and with (right) regularisation. Each line averages 30 runs (3 sample sizes $\times$ 10 seeds). The shaded band marks the near-horizon region ($\eta < 0.4$). Logarithmic vertical axis.

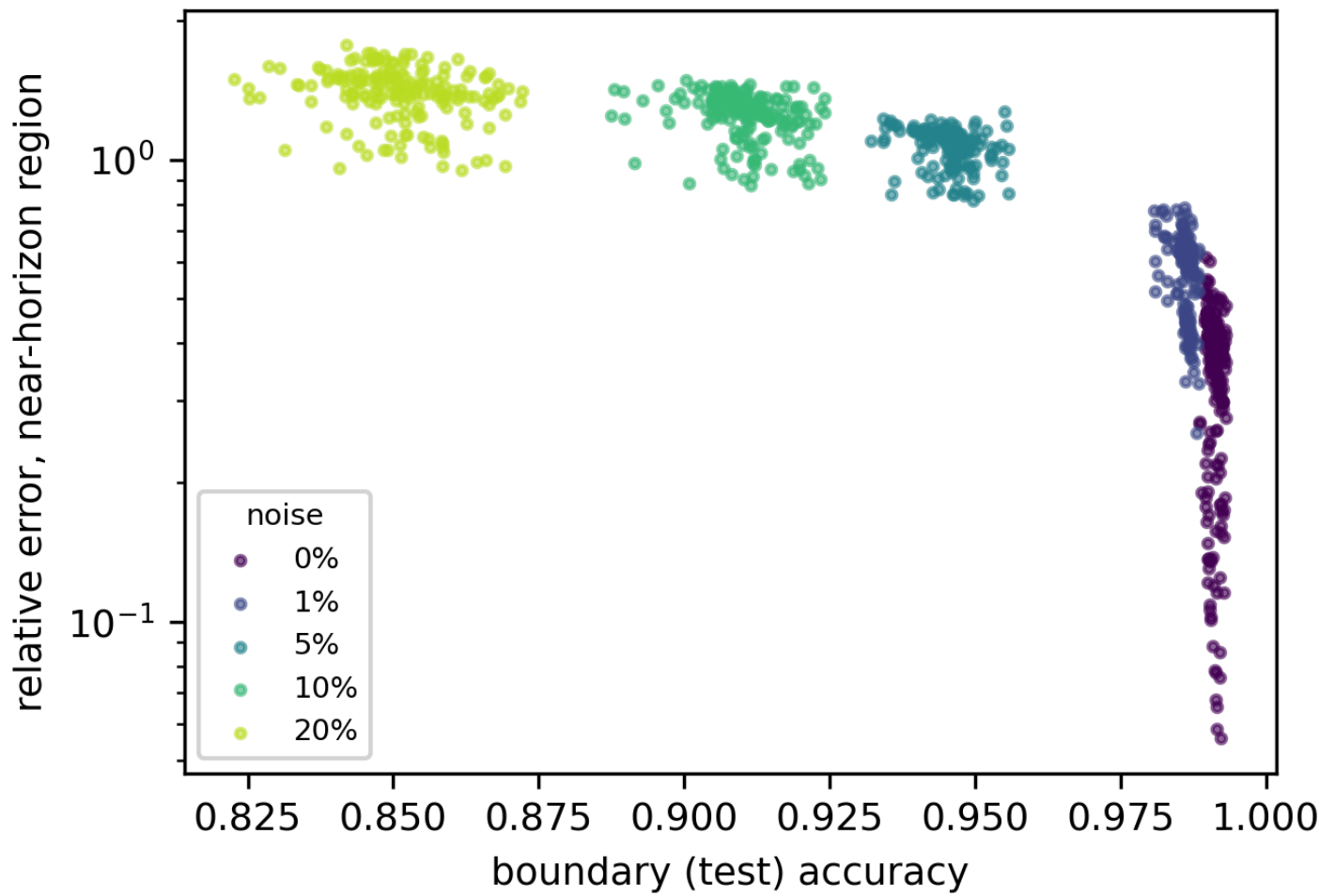


Figure 3

Test accuracy against the near-horizon reconstruction gap for all 900 runs of Experiment 1, coloured by boundary noise. Logarithmic vertical axis.

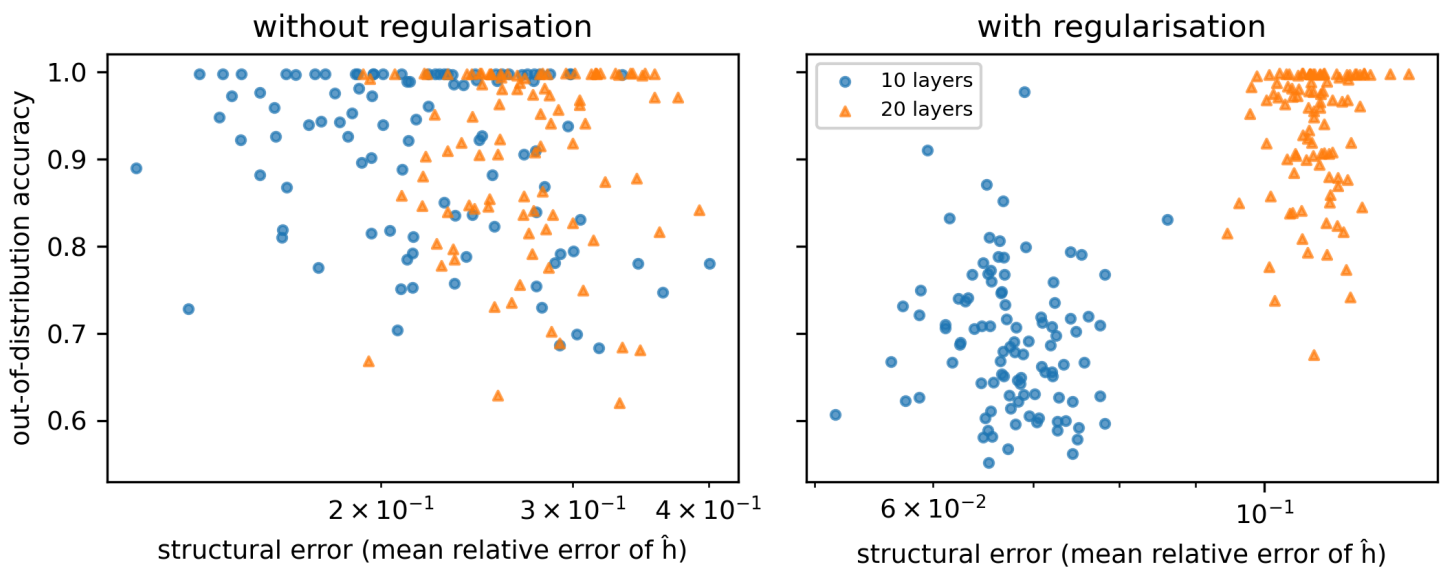


Figure 4

Structural error against out-of-distribution accuracy for the 400 observationally equivalent networks of Experiment 2, by depth, without (left) and with (right) regularisation. Logarithmic horizontal axis.