RELIABLE NEURON IDENTIFICATION: A STATISTICAL LEARNING FRAMEWORK FOR NEURON EXPLANATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

A central goal of mechanistic interpretability is to reverse-engineer deep neural networks into human-understandable mechanisms. A foundational step in this pipeline is neuron identification, which maps internal neural representations to human-understandable concepts. However, existing methods for neuron identification largely operate as empirical heuristics, making it unclear when the resulting neuron explanations can be trusted. To bridge this critical gap, we establish the first theoretical foundation for neuron identification by uncovering a fundamental correspondence to statistical learning theory: for a given neuron, selecting the concept that best explains its activation behavior is equivalent to Empirical Risk Minimization (ERM) over a concept hypothesis class. This ERM formulation unifies popular identification methods (e.g., Network Dissection, CLIP-Dissect) under a common objective and casts explanation reliability as a classical generalization problem. Leveraging this connection, we derive PAC-style fidelity bounds that characterize how explanation reliability depends on probing sample size and concept class complexity, and reveal a critical limitation of commonly used metrics: AUROC- and recall-based identification can become substantially less reliable for rare concepts. Finally, to capture uncertainty in neuron explanations, we introduce Bootstrap Explanation (BE), a principled procedure that produces concept prediction sets, with a provable coverage guarantee for its idealized variant. By connecting interpretability with statistical learning theory, our work provides a principled foundation for reliable neuron identification.

1 INTRODUCTION

With the rapid development and application of deep neural networks, their lack of interpretability raises growing concerns (Samek et al., 2017; Zhang et al., 2021). A popular strategy to open the black box is to analyze internal representations at the level of individual neurons and associate them with human-interpretable concepts. This process, known as *neuron identification* in mechanistic interpretability, yields *neuron explanations* (Bau et al., 2017; Oikarinen & Weng, 2023). Over the past few years, a growing body of neuron identification methods has emerged to map concepts to neurons, ranging from using curated concept datasets (Bau et al., 2017) to leveraging multimodal models (Oikarinen & Weng, 2023) for automated explanations (Oikarinen & Weng, 2023; Huang et al., 2023; Gurnee et al., 2023; Mu & Andreas, 2020; La Rosa et al., 2023; Zimmermann et al., 2023; Bykov et al., 2023; Kopf et al., 2024; Shaham et al., 2024; Srinivas et al., 2025).

Despite this rapid empirical progress, our theoretical understanding of neuron identification remains limited. While recent work (Oikarinen et al., 2025) unifies the evaluations of neuron explanations within a single mathematical framework to enable fair comparisons, deeper theoretical foundations regarding their reliability and generalization are still lacking. This gap undermines the trustworthiness of neuron explanations in high-stakes domains. For example, consider a chest X-ray model that attributes its pneumonia prediction to a neuron believed to represent lung opacity, when in fact the neuron responds to hospital-specific artifacts. Such low-fidelity explanations could mislead clinicians, lead to harmful treatment decisions, and ultimately erode trust.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

These concerns motivate a central question for the reliability of neuron explanations: *Can we provide formal guarantees that a neuron explanation produced by a neuron identification algorithm is reliable?* We address this question at two levels:

1. **Fidelity**[1] **bounds for neuron explanations.** *Does the identified concept achieve high fidelity, quantified by the intrinsic similarity between concept and neuron representation?*
2. **Concept prediction sets.** *Building on these bounds, can we construct a set of candidate concepts that contains the optimal concept with high probability?*

To enable rigorous analysis for reliable neuron explanations, we formulate the neuron identification problem as an Empirical Risk Minimization (ERM) problem and establish the first formal bridge between neuron identification and statistical learning theory. Our key insight is a conceptual duality swap (we call this "structural correspondence" throughout the paper): by treating the neuron as fixed and the concept itself as the hypothesis to be learned, selecting a neuron concept becomes ERM over a candidate concept set. This formulation unifies a broad family of existing identification methods as ERM instances under different similarity metrics and enables a common theoretical foundation to analyze how explanations generalize beyond probing data. Leveraging this framework, we adapt classical generalization tools to the metric-specific objectives of neuron identification, deriving finite-sample fidelity guarantees that account for probing-set size, candidate-set size, and concept prevalence. Furthermore, this framework naturally motivates concept prediction sets that formalize uncertainty across competing explanations. We summarize our contributions below:

1. A unifying statistical learning framework for neuron identification (Sec. 2). Prior work largely treats neuron identification as a heuristic matching problem (Bau et al., 2017; Oikarinen & Weng, 2023; Huang et al., 2023; Gurnee et al., 2023; Mu & Andreas, 2020; La Rosa et al., 2023; Zimmermann et al., 2023; Bykov et al., 2023; Kopf et al., 2024; Shaham et al., 2024; Srinivas et al., 2025). We establish the first ERM formulation of neuron identification, casting the probing dataset as the training dataset, candidate concepts as the hypothesis class, and the similarity metric as defining the optimization objective. This formulation unifies existing identification methods under a single theoretical umbrella and enables rigorous analysis of explanation reliability.

2. Finite-sample fidelity guarantees for neuron explanations (Sec. 3). Leveraging our ERM framework, we derive PAC-style fidelity bounds for neuron identification. Our analysis characterizes how probing-set size, candidate concept-set size, concept prevalence, and metric choice jointly affect the fidelity guarantees for selected explanations. These theoretical results are verified by simulation studies and yield practical guidelines for metric selection, including how concept rarity affects the guarantees for AUROC and recall.

3. Concept prediction sets via Bootstrap Explanation (Sec. 4). Building on our ERM framework, we also introduce a principled approach to quantify explanation uncertainty. We propose Bootstrap Explanation (BE), which aggregates concept selections across bootstrap-resampled probing datasets into a concept prediction set. We prove coverage guarantees for an idealized independent-sampling scheme and show empirically on synthetic and Broden benchmarks that BE's bootstrap approximation substantially improves coverage over standard single-concept assignments, particularly when probing data are scarce.

2 Neuron Identification as an ERM Problem

In this section, we formalize neuron identification as a similarity-maximization problem and show that it can be equivalently formulated as Empirical Risk Minimization (ERM). Although we use the term "neuron" for simplicity, our framework applies more broadly to any functional unit whose activation can be compared with concept labels through a similarity metric, including directions in representation space, Sparse Autoencoder features (Cunningham et al., 2023), TCAV directions (Kim et al., 2018), and linear probes (Alain & Bengio, 2016). More discussion on related works can be found in Sec. B.

We begin by defining the two core objects: *neuron representations* and *concepts*.

[1]Throughout this paper, fidelity means population similarity under a specified metric and probing distribution.

Neuron representation $f(x) : \mathcal{X} \to \mathbb{R}$: A neuron representation is a function mapping an input $x \in \mathcal{X}$ to an activation value. Here, $\mathcal{X}$ denotes the input space (e.g., images[2]). For example, a neuron in an MLP maps the input to a scalar value. For general neural networks, the output may not be a single real number, e.g., for convolutional neural networks (CNNs) $f(x)$ is a 2-D feature map. For simplicity, existing works often apply pooling (avg, max) to aggregate the feature into a single real value. Methods that use binary metrics further map the representation to $\tilde{f}(x) = \mathbb{1}[f(x) > \tau_f]$, where τ_f is the threshold specified by the identification method.

Concept label $c(x)$: In the literature of neuron identification (Bau et al., 2017; Oikarinen & Weng, 2023), a concept is usually defined as a human-understandable text description, such as "cat" or "shiny blue feather." Although intuitive, this definition is not a formal mathematical one. In this work, we define a concept as a function: a concept $c(x) : \mathcal{X} \to [0, 1]$ is a function that takes images as input and outputs the probability of the concept. This definition is consistent with previous works: for example, Bau et al. (2017); Bykov et al. (2023) use human annotations that output 1 if the concept is present, and 0 otherwise; Oikarinen & Weng (2024) use SigLIP (Zhai et al., 2023) to automatically estimate the probability that concept c appears. For binary metrics, we analogously write $\tilde{c}(x) = \mathbb{1}[c(x) > \tau_c]$, with τ_c determined by the annotation convention or base method.

Similarity function $\mathsf{sim}(f, c)$. Given a neuron representation and a concept, existing methods quantify their agreement using different measures, e.g., IoU (Bau et al., 2017), WPMI (Oikarinen & Weng, 2023), AUROC (Bykov et al., 2023), F1-score (Gurnee et al., 2023), among others. We observe that all of these can be unified under a single abstraction: a *similarity function* $\mathsf{sim}(f, c)$, which is a functional measuring the agreement between $f(x)$ and $c(x)$ at the population level. We call $\mathsf{sim}(f, c)$ the **fidelity** of concept c with respect to neuron f, as it quantifies how well c captures f's true behavior. This abstraction is broadly unifying: common criteria such as "the neuron's activation can classify concept c" reduce to specific choices of sim, including accuracy (Koh et al., 2020), recall (Zhou et al., 2014), AUROC (Kopf et al., 2024), and F1-score (Huang et al., 2023). Fidelity is therefore relative to the chosen similarity metric and data distribution.

With this notation, the neuron identification problem becomes a maximization over the concept set C:

$$c^*(x) = \arg\max_{c(x) \in C} \mathsf{sim}(f(x), c(x)) \tag{1}$$

Empirical approximation. In practice, the population-level $\mathsf{sim}(f, c)$ is inaccessible. Instead, existing works approximate it using realizations $\{f(x_i), c(x_i)\}$ on a probing dataset D_{probe}, where each x_i is sampled i.i.d. from the underlying data distribution. We illustrate with two common binary metrics, accuracy and IoU, which use $\tilde{f}, \tilde{c} : \mathcal{X} \to \{0, 1\}$. The same convention applies to the other binary metrics in Table 1; continuous metrics instead operate directly on the real-valued functions supplied by the base method.

Accuracy. The population-level similarity is $\mathsf{sim}_{\text{acc}}(f, c) = \mathbb{P}(\tilde{f}(x) = \tilde{c}(x))$, and its empirical estimate is:

$$\hat{\mathsf{sim}}_{\text{acc}}(f, c; D_{\text{probe}}) = \frac{1}{|D_{\text{probe}}|} \sum_{i=1}^{|D_{\text{probe}}|} \mathbb{1}(\tilde{f}(x_i) = \tilde{c}(x_i)). \tag{2}$$

Intersection over Union (IoU). Another widely used metric (Bau et al., 2017) is defined as $\mathsf{sim}_{\text{IoU}}(f, c) = \frac{\mathbb{P}(\tilde{f}(x)=1 \wedge \tilde{c}(x)=1)}{\mathbb{P}(\tilde{f}(x)=1 \vee \tilde{c}(x)=1)}$, with empirical estimate:

$$\hat{\mathsf{sim}}_{\text{IoU}}(f, c; D_{\text{probe}}) = \frac{\sum_{i=1}^{|D_{\text{probe}}|} \mathbb{1}(\tilde{f}(x_i) = 1 \wedge \tilde{c}(x_i) = 1)}{\sum_{i=1}^{|D_{\text{probe}}|} \mathbb{1}(\tilde{f}(x_i) = 1 \vee \tilde{c}(x_i) = 1)}. \tag{3}$$

Other similarity functions (e.g., AUROC, correlation) follow the same pattern. Replacing sim with its empirical estimate $\hat{\mathsf{sim}}$ in Eq. 1, the neuron identification problem becomes:

(Neuron identification) $$\hat{c} = \arg\max_{c \in C} \hat{\mathsf{sim}}(f, c; D_{\text{probe}}). \tag{4}$$

[2]The input modality could also be audio (Wu et al., 2024) or text (Huang et al., 2023; Gurnee et al., 2023). In this work, we focus on the vision domain. Depending on the granularity of concept labels and methods, the input can also refer to whole images or individual pixels.

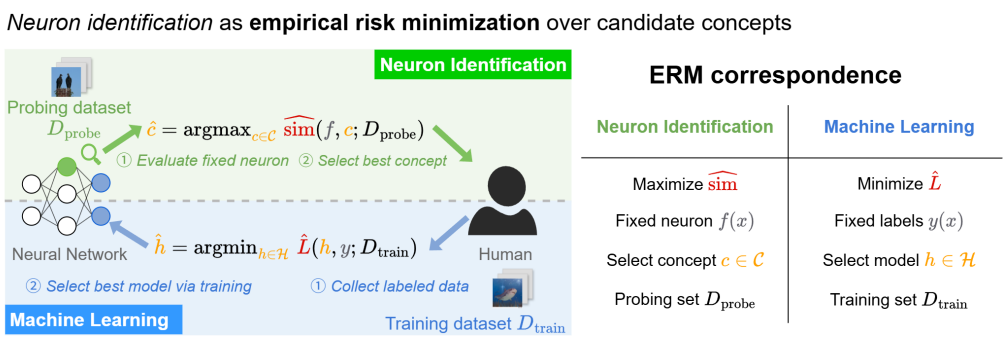


Figure 1: Analogous relationship between **neuron identification** and **machine learning**. Neuron identification searches for a concept matching a neuron, while machine learning searches for a model matching human labels. The roles of model and concept are exchanged, revealing a structural correspondence between neuron identification and the learning process.

Eq. 4 shows that the probing dataset D_{probe} plays a critical role in this approximation, yet a rigorous analysis of its effect is still lacking. We address this gap in Sec. 3 and 4.

2.1 Formal Correspondence between Neuron Identification and Empirical Risk Minimization

From the formulation in Eq. 4, we observe that the neuron identification problem closely parallels the supervised learning problem. Given a standard classification task with input x, corresponding class label $y(x)$, and a neural network model $h(x) \in H$, where H denotes the hypothesis space containing all possible neural network models, the problem can be formalized as minimizing the loss L,

$$h^* = \arg\min_{h \in H} L(h(x), y(x)), \tag{5}$$

which is typically approximated by minimizing the empirical loss $\hat{L}$ on a training dataset D_{train}, known as Empirical Risk Minimization (ERM):

(Machine Learning) $$\hat{h} = \arg\min_{h \in H} \hat{L}(h, y; D_{\text{train}}). \tag{6}$$

Comparing Eq. 6 and Eq. 4, we see that these two problems share a similar structure: both are optimization problems with objectives of similar form. The left panel of Fig. 1 compares the procedures of these two domains, while the right panel lists their detailed correspondences. As illustrated in Fig. 1, neuron identification and machine learning exhibit a structural correspondence in which the roles of the known and unknown quantities are exchanged: in machine learning, we search for a neural network $h(x)$ that approximates a target human concept $y(x)$ (e.g., ImageNet classes), whereas neuron identification searches for a concept $c(x)$ that best matches a given neuron representation $f(x)$. This parallel reveals that the neuron identification formulation in Eq. 4 is an instance of ERM: just as ERM selects a hypothesis $\hat{h} \in H$ by minimizing an empirical loss on training data, neuron identification selects a concept $\hat{c} \in C$ by maximizing an empirical similarity on a probing dataset.

What the ERM framework enables. Through this structural correspondence, questions about the reliability of a neuron explanation can now be formally cast as a generalization problem in statistical learning theory. This bridge enables two key advances in the following sections: (i) deriving generalization bounds for ERM that provide fidelity guarantees for identified concepts (Sec. 3), and (ii) introducing bootstrap aggregation over ERM solutions to construct concept prediction sets that quantify uncertainty in neuron explanations (Sec. 4).

3 PAC-Style Fidelity Bounds for Neuron Explanations

In this section, we address the first question: *Does the identified concept achieve high fidelity, quantified by the intrinsic similarity between concept and neuron representation?* Building on the

ERM correspondence established in Sec. 2.1, we cast explanation reliability as a generalization problem in statistical learning theory and derive PAC-style fidelity bounds for neuron explanations that characterize when an identified concept can be reliably trusted. We first derive the fidelity bounds in Sec. 3.1, and then instantiate them for popular similarity metrics in Sec. 3.2.

3.1 Derivation of Fidelity Bounds

Using the framework introduced in Sec. 2, the first question reduces to asking whether the identified concept truly achieves high population-level similarity $\mathsf{sim}(f, c)$ (i.e., high fidelity) to the neuron representation. From the structural correspondence we discover in Sec. 2, the population-level similarity corresponds to population risk, which can be bounded with generalization theory. Thus, we define the generalization gap as $g(D_{\text{probe}}, C, f) \triangleq \sup_{c \in C}[\hat{\mathsf{sim}}(f, c; D_{\text{probe}}) - \mathsf{sim}(f, c)]$, which measures the maximum difference between the estimated and true similarity over all concepts. Bounding this gap yields two results: (1) the identified concept $\hat{c}$ is approximately optimal (Cor. 3.3), and (2) a single-trial error probability that quantifies the chance of selecting a non-optimal concept (Cor. 3.4). The latter serves as the foundation for the prediction sets developed in Sec. 4.

We start with the simplest and most common setting where the concept set is finite. We show that the gap $g(D_{\text{probe}}, C, f)$ can be bounded in Thm. 3.2 under two mild assumptions: (i) the concept set C is finite, and (ii) the probing dataset D_{probe} is sampled i.i.d. These conditions are met by most existing neuron identification methods (e.g., Bau et al., 2017; Oikarinen & Weng, 2023; Bykov et al., 2023).

Definition 3.1. The convergence rate of a similarity function is a function $r(f, D_{\text{probe}}, \delta)$ satisfying $\mathbb{P}\big[|\hat{\mathsf{sim}}(f, c; D_{\text{probe}}) - \mathsf{sim}(f, c)| \geq r(f, D_{\text{probe}}, \delta)\big] \leq \delta$, where D_{probe} is randomly sampled from the underlying data distribution D with a fixed size $|D_{\text{probe}}|$, and the probability is taken over all possible sampled probing datasets.

Theorem 3.2. *With probability at least* $1 - \delta$*,*

$$\sup_{c \in C} |\hat{\mathit{sim}}(f, c; D_{\text{probe}}) - \mathit{sim}(f, c)| \leq r(f, D_{\text{probe}}, \frac{\delta}{|C|}), \tag{7}$$

where $r(f, D_{\text{probe}}, \delta)$ *is the convergence rate defined above. In Eq. 7, the confidence parameter* δ *is adjusted using a union bound, replacing* δ *with* $\frac{\delta}{|C|}$*.*

Corollary 3.3. *With probability at least* $1 - \delta$*,*

$$\mathit{sim}(f, \hat{c}) \geq \mathit{sim}(f, c^*) - 2r(f, D_{\text{probe}}, \frac{\delta}{|C|}), \tag{8}$$

where $\hat{c}$ *is the concept selected by Eq. 4 and* $c^* = \arg\max_{c \in C}[\mathit{sim}(f, c)]$ *is the optimal concept.*

Discussion. We defer the proof of Thm. 3.2 to Sec. A.1 due to the page limit. The problem-specific content of Thm. 3.2 and Cor. 3.3 enters through the convergence rate r, which depends on the similarity metric. Thm. 3.2 provides a PAC-style bound on the generalization gap over all concepts, where the convergence rate $r(f, D_{\text{probe}}, \delta)$ characterizes how fast the empirical similarity $\hat{\mathsf{sim}}$ converges to the true similarity sim, and $|C|$, which is independent of the similarity function, controls the error probability δ. In Sec. 3.2 below, we derive convergence rates r for several popular similarity functions, showing that many commonly used estimators have a rate $r(f, D_{\text{probe}}, \delta) = \mathcal{O}(\sqrt{\frac{-\log \delta}{|D_{\text{probe}}|}})$. Cor. 3.3 then shows that the identified concept $\hat{c}$ is *approximately optimal*, within a gap determined by r and the concept set size $|C|$. In addition, a direct consequence of Cor. 3.3 is a bound on the probability that a single identification trial selects the optimal concept exactly, as shown in Cor. 3.4 below.

Corollary 3.4 (Single-trial error probability)**.** *Suppose the optimal concept* c^* *satisfies* $\mathit{sim}(f, c^*) > \mathit{sim}(f, c) + \Delta$ *for all* $c \in C, c \neq c^*$*, where* $\Delta > 0$*. Let the error probability* $p \in [0, 1]$ *be implicitly defined such that* $r(f, D_{\text{probe}}, \frac{p}{|C|}) = \frac{\Delta}{2}$*. Then, the probability that the identified concept* $\hat{c}$ *exactly recovers the optimal concept is lower bounded by:*

$$\mathbb{P}(\hat{c} = c^*) \geq 1 - p. \tag{9}$$

This direct consequence is derived by setting $\delta = p$ in Cor. 3.3, which guarantees $\mathsf{sim}(f, \hat{c}) \geq \mathsf{sim}(f, c^*) - \Delta > \mathsf{sim}(f, c), \forall c \neq c^*$. Intuitively, the error probability p decreases as the probing dataset grows or the similarity gap Δ increases. While this guarantee may be loose when p is moderate, Sec. 4 shows that aggregating repeated identification trials into a prediction set strengthens it, and proposes a bootstrap procedure that approximates this aggregation in practice.

Extension to linear concept spaces. The framework also lets us adopt the standard treatment of infinite hypothesis classes. The above theorems can be extended to linear combinations of m base concepts (Oikarinen & Weng, 2024), yielding an infinite concept space. In this case, we show that an analogous fidelity bound still holds, with the concept set size $|C|$ replaced by a covering number that grows exponentially in m (see Sec. J for details).

sim Metric	$\mathsf{sim}(f, c)$	$\hat{\mathsf{sim}}(f, c)$	$r(f, D_{\text{probe}}, \delta)$
Accuracy (Oikarinen et al., 2025)[3]	$\mathbb{P}(\tilde{f}(x) = \tilde{c}(x))$	$\frac{\sum_{x \in D_{\text{probe}}} \mathbf{1}(\tilde{f}(x) = \tilde{c}(x))}{\lvert D_{\text{probe}} \rvert}$	$\sqrt{\frac{\log(\frac{2}{\delta})}{2\lvert D_{\text{probe}} \rvert}}$
AUROC (Bykov et al., 2023)	$\mathbb{P}(f(x) < f(y) \mid \tilde{c}(x) = 0, \tilde{c}(y) = 1)$	$\frac{\sum_{\{x \mid c(x)=0\}} \sum_{\{y \mid c(y)=1\}} \mathbf{1}[f(x) < f(y)]}{\lvert\{x \mid c(x)=0\}\rvert \lvert\{x \mid c(x)=1\}\rvert}$	$\sqrt{\frac{\log(\frac{2}{\delta})}{2\rho(1-\rho)\lvert D_{\text{probe}} \rvert}}$
IoU (Bau et al., 2017)	$\frac{W_{11}}{W_{01}+W_{11}+W_{10}}$	$\frac{F_{11}}{F_{01}+F_{11}+F_{10}}$	$\sqrt{\frac{\log(\frac{2}{\delta})}{2(F_{11}+F_{10}+F_{01})\lvert D_{\text{probe}} \rvert}}$
Recall (Srinivas et al., 2025)	$\frac{W_{11}}{W_{01}+W_{11}}$	$\frac{F_{11}}{F_{01}+F_{11}}$	$\sqrt{\frac{\log(\frac{2}{\delta})}{2(F_{11}+F_{01})\lvert D_{\text{probe}} \rvert}}$
Precision (Zhou et al., 2014)	$\frac{W_{11}}{W_{10}+W_{11}}$	$\frac{F_{11}}{F_{10}+F_{11}}$	$\sqrt{\frac{\log(\frac{2}{\delta})}{2(F_{11}+F_{10})\lvert D_{\text{probe}} \rvert}}$

Table 1: Similarity metrics $\mathsf{sim}(f, c)$, their estimators $\hat{\mathsf{sim}}(f, c)$, and corresponding convergence rates $r(f, D_{\text{probe}}, \delta)$. Different similarity metrics correspond to different neuron identification methods (e.g., AUROC for INVERT (Bykov et al., 2023)). Accuracy, IoU, recall, and precision use the base-method-specific binarizations $\tilde{f}$ and $\tilde{c}$. For these binary metrics, $W_{ij} = \mathbb{P}(\tilde{f}(x) = i, \tilde{c}(x) = j)$ and $F_{ij} = |\{\tilde{f}(x) = i, \tilde{c}(x) = j \mid x \in D_{\text{probe}}\}|/|D_{\text{probe}}|$. AUROC retains real-valued neuron activations and uses binary concept labels; ρ is the concept frequency in D_{probe}.

3.2 Convergence Rates for Popular Similarity Metrics

The convergence rate r is the key factor controlling the generalization gap $g(D_{\text{probe}}, C, f)$. Table 1 summarizes the derived r for several commonly used similarity metrics in existing neuron identification methods. For accuracy, the empirical estimate converges to the population similarity at a standard $\mathcal{O}(1/\sqrt{|D_{\text{probe}}|})$ rate via Hoeffding's inequality (Shalev-Shwartz & Ben-David, 2014). However, bounds for more complex metrics reveal an important dependence on concept properties. For instance, the convergence rate of AUROC additionally depends on the concept frequency $\rho = \mathbb{P}(c(x) = 1)$ (Agarwal et al., 2004). When ρ is small, r_{AUROC} deteriorates significantly, suggesting that AUROC-based neuron identification is unreliable for rare concepts (see Fig. E.3). Similarly, metrics such as recall, precision, and IoU can be analyzed as conditional versions of accuracy whose effective sample size is determined by the number of positive examples rather than $|D_{\text{probe}}|$ (details in Sec. C). We also extend these derivations to continuous metrics (e.g., correlation, WPMI, AUPRC) in Sec. G, where empirical studies confirm the same $\mathcal{O}(1/\sqrt{|D_{\text{probe}}|})$ scaling behavior.

Simulation studies. We validate these predictions on synthetic data with known population similarity (Sec. E). Accuracy converges fastest, whereas AUROC and recall degrade sharply for a rare concept (frequency 0.1%), matching their dependence on the concept frequency in Table 1. The gap $\mathsf{sim}(f, c^*) - \mathsf{sim}(f, \hat{c})$ decays at the $\mathcal{O}(1/\sqrt{|D_{\text{probe}}|})$ rate predicted by Thm. 3.2 and Cor. 3.3 for all metrics.

[3]Accuracy was originally used to evaluate explanations, but it can also be used to generate them.

Practical guidelines. Our derived bounds yield concrete, actionable guidance for interpretability practitioners. For example, when using the accuracy metric with a concept set of size $|C| = 1{,}000$ and desired confidence level $1 - \delta = 95\%$, Cor. 3.3 guarantees the selected concept is within 2ϵ of the optimal concept provided $|D_{\text{probe}}| \geq \log(2|C|/\delta) \,/\, (2\epsilon^2)$. Setting $\epsilon = 0.05$ gives $|D_{\text{probe}}| \approx 2{,}100$, a highly practical sample complexity for standard vision datasets. For metric selection, accuracy serves as the most stable default; AUROC should be avoided when the target concept is rare (frequency $< 1\%$), as its convergence rate degrades sharply. When using IoU, recall, or precision, the effective sample size is determined by the number of positive examples rather than the total dataset size, so users should ensure sufficient positive samples for the concept of interest.

4 Concept Prediction Sets via Bootstrap Explanation

In this section, we address the second question: *Building on the fidelity bounds, can we construct a set of candidate concepts that contains the optimal concept with high probability?* Building on our ERM framework in Sec. 2, we now quantify explanation uncertainty by constructing a concept prediction set, instead of assigning a single concept to the neuron. Specifically, we adopt bootstrap ensembles (Breiman, 1996) to propose Bootstrap Explanation (BE), which aggregates concept selections across resampled probing datasets in Sec. 4.1. We then prove a coverage guarantee for an idealized version of BE that uses independent probing sets in Sec. 4.2, and empirically evaluate the coverage of BE itself on Broden and synthetic data in Sec. 4.3.

4.1 Constructing Bootstrap Explanation

Bootstrap ensemble (Breiman, 1996) is a technique that aggregates multiple models, each trained on a resampled version of the original dataset. The final prediction is typically determined by majority voting, with confidence estimated from the proportion of models supporting the prediction (Lakshminarayanan et al., 2017). Our ERM framework makes this idea naturally applicable to neuron identification: because the probing dataset plays the role of the training set, resampling it induces multiple concept assignments for the same neuron, allowing us to quantify uncertainty in its explanation rather than returning only a single concept. Based on this idea, Bootstrap Explanation consists of four main steps:

1. **Collect bootstrap datasets:** Sample K datasets $\{D_i\}_{i=1}^{K}$ independently by randomly selecting samples from the probing dataset D_{probe} with replacement.
2. **Run neuron identification:** Apply the neuron identification algorithm to each bootstrap dataset D_i and record the predicted concept $\hat{c}_i$.
3. **Aggregate predictions:** After K runs, count the frequency of each concept: $k_c = \sum_{i=1}^{K} \mathbf{1}(\hat{c}_i = c)$.
4. **Construct prediction set:** Sort concepts by decreasing frequency and add them to the prediction set S until the cumulative count reaches a threshold t (e.g., $t = 0.95K$): $S = \{c_{r_1}, c_{r_2}, \ldots, c_{r_l}\}, \quad \text{where } l = \min\{j : \sum_{i=1}^{j} k_{r_i} \geq t\}$.

Fig. 2 summarizes the pipeline, and pseudocode is provided in Alg. 1, Sec. D. The threshold t controls the trade-off between coverage and set size: a higher t yields a larger set with stronger coverage, while a lower t produces a smaller, more selective set. When several concepts have similar scores (e.g., for a polysemantic neuron), different bootstrap samples can select different concepts, exposing ambiguity hidden by a single top-1 explanation.

4.2 An Idealized Coverage Guarantee

We now formalize the idealized setting that BE approximates. Consider the *ideal* version of Alg. 1, in which each of the K identification trials uses a fresh probing set of size $|D_{\text{probe}}|$ drawn i.i.d. from the data distribution instead of a bootstrap resample of D_{probe}. The result below extends the single-trial error probability p from Cor. 3.4 to this setting.

Theorem 4.1 (Idealized coverage)**.** *Let c^* be the optimal concept for a given neuron and suppose $\mathsf{sim}(f, c^*) \geq \mathsf{sim}(f, c) + \Delta, \forall c \in C, c \neq c^*$, where $\Delta > 0$ is a positive constant. Let $S \subseteq C$ be the*

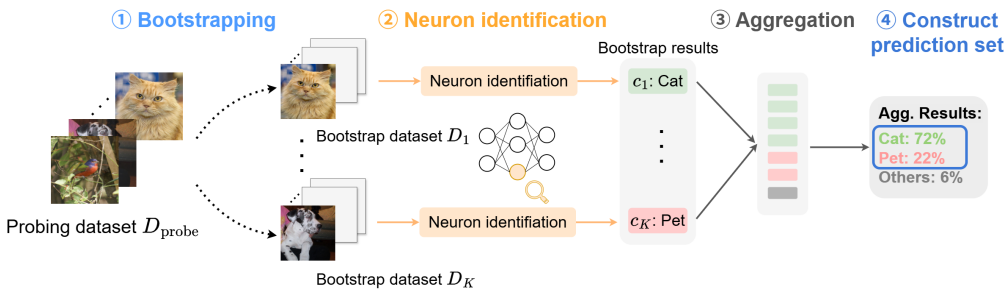


Figure 2: Illustration of Bootstrap Explanation (BE). Multiple probing datasets are generated via bootstrapping. Then, the neuron identification algorithm is applied to each dataset, and the resulting concepts are aggregated to estimate the probability of each concept.

prediction set constructed by the ideal version of Alg. 1 with K trials and threshold t, where each trial uses an independent probing set of size $|D_{\text{probe}}|$ drawn i.i.d. from the data distribution. Then

$$\mathbb{P}(c^* \in S) \geq \sum_{i=K-\lceil t \rceil+1}^{K} \binom{K}{i} (1-p)^i p^{K-i}, \tag{10}$$

where p is the single-trial error probability from Cor. 3.4.

We provide the proof of Thm. 4.1 in Sec. A.2.

Discussion. Thm. 4.1 is an extension of the single-trial result in Cor. 3.4: it reduces to Cor. 3.4 when $K = 1$, which means the single-trial output is taken as the prediction set. While Cor. 3.4 bounds the probability of a single correct identification, Thm. 4.1 shows that aggregating K independent trials into a prediction set yields a coverage probability that grows as a binomial CDF in K, which can be made arbitrarily close to 1 even when the per-trial error p is moderate. This is the key amplification: a weak single-trial guarantee is boosted into a strong set-level guarantee. The theory relies on independent trials, which cost K fresh probing sets, making it expensive to apply. The bootstrap trials of BE approximate this process, enabling prediction sets without collecting new data.

4.3 Empirical Coverage of Bootstrap Explanation

We evaluate BE on Broden (Bau et al., 2017) and in controlled synthetic settings, focusing on concept coverage and prediction-set size.

Broden dataset evaluation. We split Broden into disjoint probe and test sets and apply CLIP-Dissect to ResNet-50 layer-4 neurons with either soft-WPMI or cosine similarity. For each metric, the top-ranked concept on the test set serves as a proxy for the population-optimal concept, and we compare BE with single-concept assignment on probe sets of increasing size. Setup details are given in Sec. D.1. From Fig. 3(a, b), we can see:

1. Single-concept assignment is unreliable on real data: even with all probe images, the probe-selected concept differs from the test-set concept for about half of the neurons under soft-WPMI.
2. Single-concept coverage rises steadily with probe size under both metrics, consistent with the finite-sample behavior predicted by Thm. 3.2.
3. BE achieves higher coverage than single-concept assignment at every probe size, and the gain is largest when probe data are scarce, which is where a single assignment is least reliable.

Raising the threshold t increases coverage but also enlarges the prediction sets (Fig. D.1 in Sec. D.1).

Controlled synthetic data evaluation. We next verify these trends against a known population-optimal concept, using $|C| = 1{,}000$ candidates, an optimal agreement probability of $s^* = 0.9$, a margin of $\Delta = 0.05$ over all other concepts, and 2,000 repetitions. We vary the probe size with $K = 100$ and $t = 95$. From Fig. 3(c, d), we can see:

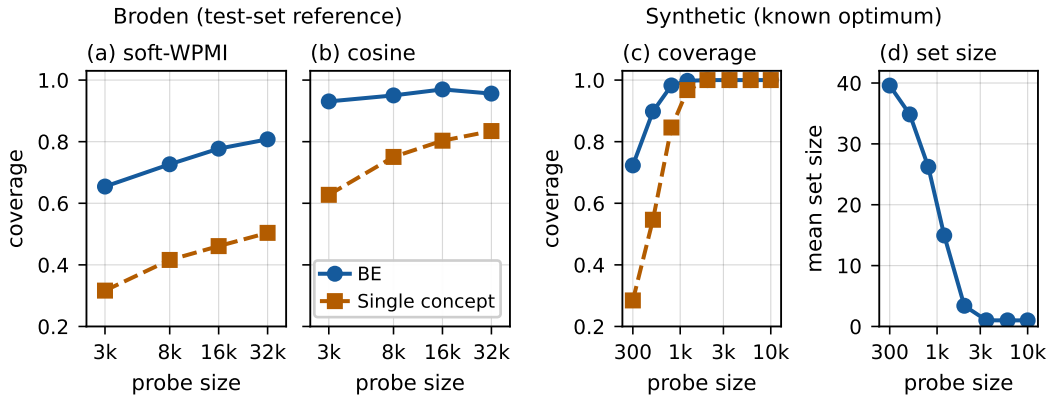


Figure 3: Coverage of BE and single-concept assignment versus probe size. (a, b) Broden with CLIP-Dissect under soft-WPMI and cosine similarity: coverage is the fraction of neurons whose prediction set contains the top-ranked test-set concept, and the probe sets are nested subsets with 10–100% of the 31,652 probe images. (c, d) Controlled synthetic experiment: coverage of the population-optimal concept over 2,000 repetitions, and mean prediction-set size. BE uses $K = 100$ and $t = 95$; single-concept assignment returns a singleton set. Probe size is on a logarithmic scale.

1. BE substantially improves on single-concept assignment when the probe set is small, and both approach full coverage as data grow, consistent with the Broden results.
2. The prediction set shrinks to a single concept as the probe size grows: BE expresses uncertainty through set size when data are limited and reduces to a single assignment once the evidence is sufficient.

We include more empirical experiments in Sec. D.

Qualitative examples. We also apply BE to NetDissect and CLIP-Dissect on all 2,048 layer-4 neurons of an ImageNet-trained ResNet-50, with Broden as the probing dataset (Sec. I.1). NetDissect returns concrete concepts that are stable across bootstrap samples, whereas CLIP-Dissect returns more diverse and abstract concepts and sometimes captures ones missed by NetDissect. The computational overhead of BE is discussed in Sec. F.

5 Conclusion

In this work, we establish a principled theoretical framework for assessing the reliability of neuron explanations by uncovering a structural correspondence between neuron identification and Empirical Risk Minimization (ERM). By formally casting explanation reliability as a generalization problem, our framework enables two key advances. First, we derive PAC-style fidelity bounds that characterize how explanation reliability depends on probing-set size, concept prevalence, and hypothesis class complexity, theoretically uncovering critical failure modes of widely used similarity metrics for rare concepts. Second, we introduce Bootstrap Explanation (BE), which moves beyond point-estimate assignments to construct concept prediction sets that quantify explanation uncertainty, together with a theoretical coverage guarantee for its idealized variant. Experiments on synthetic benchmarks and the Broden dataset corroborate our theoretical predictions and demonstrate the practical utility of BE prediction sets. Broadly, this bridge provides a principled foundation for reliable neuron identification and opens a new avenue toward trustworthy mechanistic interpretability.

AI USE STATEMENT

In this work, we used generative AI tools to assist in writing proofs (the extension to linear concept spaces in Sec. J), implement experiment code, provide feedback on experimental design, and help interpret results. We have not used generative AI tools to generate synthetic datasets, develop the conceptual framework, or formulate the main theoretical claims, and translation, dataset cleaning, and qualitative data analysis are not applicable to this work. Additionally, we used generative AI tools to edit the paper for readability and to check grammar and typos. We have reviewed all AI-assisted work: all proofs were checked by the authors, and AI-assisted code was verified and tested by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work develops statistical tools for assessing the reliability of neuron explanations. It does not involve human subjects or the collection of new data; all experiments use publicly available models and datasets (ImageNet and Broden). We believe that more reliable interpretability tools can help practitioners audit neural networks, and we are not aware of direct negative societal consequences beyond the general risk that explanations may be over-trusted, which our fidelity bounds and concept prediction sets are designed to mitigate.

REPRODUCIBILITY STATEMENT

Complete statements and proofs of our theoretical results are provided in Sections A.1 to A.3, together with the derivations of the convergence rates in Sec. C and G and the extension to linear concept spaces in Sec. J. The Bootstrap Explanation procedure is specified in Alg. 1. Details of the simulation studies are given in Sec. E, and the experimental setup for BE, including backbones, concept sets, probing data, and hyperparameters, is described in Sec. D and 4.3.

REFERENCES

Shivani Agarwal, Thore Graepel, Ralf Herbrich, and Dan Roth. A large deviation bound for the area under the roc curve. *Advances in Neural Information Processing Systems*, 17, 2004.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Alaa Anani, Tobias Lorenz, Bernt Schiele, Mario Fritz, and Jonas Fischer. Certified circuits: Stability guarantees for mechanistic circuits. In *International Conference on Machine Learning*, 2026.

Nicholas Bai, Rahul Ajay Iyer, Tuomas Oikarinen, Akshay Kulkarni, and Tsui-Wei Weng. Interpreting neurons in deep vision networks with language models. In *TMLR*, 2025.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6541–6549, 2017.

Leo Breiman. Bagging predictors. *Machine learning*, 24:123–140, 1996.

Kirill Bykov, Laura Kopf, Shinichi Nakajima, Marius Kloft, and Marina Höhne. Labeling neural representations with inverse recognition. *Advances in Neural Information Processing Systems*, 36: 24804–24828, 2023.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. Finding neurons in a haystack: Case studies with sparse probing. *arXiv preprint arXiv:2305.01610*, 2023.

Itamar Hadad, Guy Katz, and Shahaf Bassan. Formal mechanistic interpretability: Automated circuit discovery with provable guarantees. In *International Conference on Learning Representations*, 2026.

Jing Huang, Atticus Geiger, Karel D'Oosterlinck, Zhengxuan Wu, and Christopher Potts. Rigorously assessing natural language explanations of neurons. *arXiv preprint arXiv:2309.10312*, 2023.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pp. 2668–2677. PMLR, 2018.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pp. 5338–5348. PMLR, 2020.

Laura Kopf, Philine L Bommer, Anna Hedström, Sebastian Lapuschkin, Marina Höhne, and Kirill Bykov. Cosy: Evaluating textual explanations of neurons. *Advances in Neural Information Processing Systems*, 37:34656–34685, 2024.

Biagio La Rosa, Leilani Gilpin, and Roberto Capobianco. Towards a fuller understanding of neurons with clustered compositional explanations. *Advances in Neural Information Processing Systems*, 36:70333–70354, 2023.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11976–11986, 2022.

Jesse Mu and Jacob Andreas. Compositional explanations of neurons. *Advances in Neural Information Processing Systems*, 33:17153–17163, 2020.

Tuomas Oikarinen and Tsui-Wei Weng. Clip-dissect: Automatic description of neuron representations in deep vision networks. In *International Conference on Learning Representations*, 2023.

Tuomas Oikarinen and Tsui-Wei Weng. Linear explanations for individual neurons. In *Forty-first International Conference on Machine Learning*, 2024.

Tuomas Oikarinen, Ge Yan, and Tsui-Wei Weng. Evaluating neuron explanations: A unified framework with sanity checks. In *International Conference on Machine Learning*, 2025.

Nils Palumbo, Ravi Mangal, Zifan Wang, Saranya Vijayakumar, Corina S. Pasareanu, and Somesh Jha. Validating mechanistic interpretations: An axiomatic approach. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 47509–47544. PMLR, 2025.

Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*, 2017.

Tamar Rott Shaham, Sarah Schwettmann, Franklin Wang, Achyuta Rajaram, Evan Hernandez, Jacob Andreas, and Antonio Torralba. A multimodal automated interpretability agent. In *Forty-first International Conference on Machine Learning*, 2024.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Anvita A Srinivas, Tuomas Oikarinen, Divyansh Srivastava, Wei-Hung Weng, and Tsui-Wei Weng. Sand: Enhancing open-set neuron descriptions through spatial awareness. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2993–3002. IEEE, 2025.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Tung-Yu Wu, Yu-Xiang Lin, and Tsui-Wei Weng. And: audio network dissection for interpreting deep acoustic models. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 53656–53680, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

Yu Zhang, Peter Tiňo, Aleš Leonardis, and Ke Tang. A survey on neural network interpretability. *IEEE transactions on emerging topics in computational intelligence*, 5(5):726–742, 2021.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns. *arXiv preprint arXiv:1412.6856*, 2014.

Roland S Zimmermann, Thomas Klein, and Wieland Brendel. Scale alone does not improve mechanistic interpretability in vision models. *Advances in Neural Information Processing Systems*, 36:57876–57907, 2023.

APPENDIX CONTENTS

A Proofs of Theoretical Results **14**

B Related Work **17**

C Details on recall, precision, and IoU convergence rate derivation **18**

D Experimental Details for Bootstrap Explanation **19**

E Simulation Studies for the Fidelity Bounds **22**

F Computational cost of Bootstrap Explanation (BE) **25**

G Additional fidelity experiments for continuous metrics **27**

H Study on sample size (Thm. 3.2) **29**

I Additional Bootstrap Explanation (BE) experiments **30**

J Extension of Fidelity Bounds to Linear Concept Spaces **36**

A Proofs of Theoretical Results

A.1 Proofs of Thm. 3.2 and Cor. 3.3

Proof of Thm. 3.2. For each $c_i \in C$, from the definition of convergence rate, we have

$$\mathbb{P}\left[\left|\hat{\mathsf{sim}}(f, c_i; D_{\text{probe}}) - \mathsf{sim}(f, c_i)\right| \geq r(f, D_{\text{probe}}, \frac{\delta}{|C|})\right] \leq \frac{\delta}{|C|}. \tag{11}$$

Thus, with union bound, we have

$$\begin{aligned}
&\mathbb{P}\left\{\sup_{c\in C}|\hat{\mathsf{sim}}(f, c; D_{\text{probe}}) - \mathsf{sim}(f, c)| \leq r(f, D_{\text{probe}}, \frac{\delta}{|C|})\right\} \\
=&1 - \mathbb{P}\left\{\cup_{c\in C}|\hat{\mathsf{sim}}(f, c; D_{\text{probe}}) - \mathsf{sim}(f, c)| > r(f, D_{\text{probe}}, \frac{\delta}{|C|})\right\} \\
\geq&1 - \sum_{c\in C}\mathbb{P}\left\{|\hat{\mathsf{sim}}(f, c; D_{\text{probe}}) - \mathsf{sim}(f, c)| > r(f, D_{\text{probe}}, \frac{\delta}{|C|})\right\} \\
\geq&1 - |C|\frac{\delta}{|C|} \\
=&1 - \delta
\end{aligned} \tag{12}$$

□

Proof of Cor. 3.3. From the definition,

$$\hat{c} = \arg\max_{c\in C} \hat{\mathsf{sim}}(f, c). \tag{13}$$

Thus, $\hat{\mathsf{sim}}(f, \hat{c}) \geq \hat{\mathsf{sim}}(f, c^*)$. From Thm. 3.2, with probability at least $1 - \delta$, we have

$$\mathsf{sim}(f, \hat{c}) \geq \hat{\mathsf{sim}}(f, \hat{c}) - r(f, D_{\text{probe}}, \frac{\delta}{|C|}) \tag{14}$$

and

$$\mathsf{sim}(f, c^*) \leq \hat{\mathsf{sim}}(f, c^*) + r(f, D_{\text{probe}}, \frac{\delta}{|C|}). \tag{15}$$

Therefore, we have

$$\mathsf{sim}(f, \hat{c}) \geq \hat{\mathsf{sim}}(f, \hat{c}) - r(f, D_{\text{probe}}, \frac{\delta}{|C|}) \tag{16}$$

$$\geq \hat{\mathsf{sim}}(f, c^*) - r(f, D_{\text{probe}}, \frac{\delta}{|C|}) \tag{17}$$

$$\geq \mathsf{sim}(f, c^*) - 2r(f, D_{\text{probe}}, \frac{\delta}{|C|}) \tag{18}$$

with probability at least $1 - \delta$. □

A.2 Proof of Thm. 4.1

In this section, we prove Thm. 4.1: Theorem 4.1. *Let c^* be the optimal concept for a given neuron and suppose $\mathsf{sim}(f, c^*) \geq \mathsf{sim}(f, c) + \Delta, \forall c \in C, c \neq c^*$, where $\Delta > 0$. Let $S \subseteq C$ be the prediction set constructed by the ideal version of Alg. 1 with K trials and threshold t, where each trial uses an independent probing set of size $|D_{\text{probe}}|$ drawn i.i.d. from the data distribution. Then,*

$$\mathbb{P}(c^* \in S) \geq \sum_{i=K-\lceil t\rceil+1}^{K} \binom{K}{i} (1-p)^i p^{K-i}, \tag{19}$$

where p is the single-trial error probability defined implicitly by the equation $r(f, D_{\text{probe}}, \frac{p}{|C|}) = \frac{\Delta}{2}$.

Proof. We start the proof by estimating the single-trial error rate.

Lemma A.1. *Let p be defined implicitly by the equation*

$$r(f, D_{\text{probe}}, \frac{p}{|C|}) = \frac{\Delta}{2}, \tag{20}$$

where $r(\cdot)$ is the convergence rate in Thm. 3.2. Then,

$$\mathbb{P}(\hat{c} = c^*) \geq 1 - p \tag{21}$$

Remark A.2. Thm. A.1 can be easily derived from Thm. 3.2: with probability at least $1 - p$, $\sup_{c \in C} |\hat{\mathsf{sim}}(f, c; D_{\text{probe}}) - \mathsf{sim}(f, c)| \leq \frac{\Delta}{2}$, thus

$$\begin{aligned} \hat{\mathsf{sim}}(f, c^*; D_{\text{probe}}) &\geq \mathsf{sim}(f, c^*) - \frac{\Delta}{2} \\ &\geq \mathsf{sim}(f, c) + \frac{\Delta}{2} \quad \text{(margin } \Delta\text{)} \\ &\geq \hat{\mathsf{sim}}(f, c; D_{\text{probe}}). \end{aligned} \tag{22}$$

Previously, we showed that for many similarity metrics (AUROC, accuracy, IoU, etc.), $r(f, D_{\text{probe}}, \delta) = \mathcal{O}(\sqrt{\frac{-\log \delta}{|D_{\text{probe}}|}})$, i.e., $r(f, D_{\text{probe}}, \delta) \leq Q(\sqrt{\frac{-\log \delta}{|D_{\text{probe}}|}})$ for some constant $Q > 0$. In this case, we can plug in $\delta = \frac{p}{|C|}$ and get

$$\frac{\Delta}{2} = r(f, D_{\text{probe}}, \frac{p}{|C|}) \leq Q\sqrt{\frac{-\log \frac{p}{|C|}}{|D_{\text{probe}}|}}, \tag{23}$$

which gives $p \leq |C| e^{-\frac{\Delta^2}{4Q^2}|D_{\text{probe}}|}$. This shows that when the probing dataset size $|D_{\text{probe}}|$ and the gap Δ between the optimal concept and other concepts become larger, the error probability p can be reduced.

In the idealized setting, the K trials use independent probing sets $D_1, \ldots, D_K$, each drawn i.i.d. from the data distribution with $|D_i| = |D_{\text{probe}}|$. The selected concepts $\hat{c}_i = \texttt{Identify}(C, f, D_i)$ are therefore independent, and each equals c^* with probability at least $1 - p$ by Thm. A.1. Then, we have the following theorem.

Theorem A.3. *Let $k^* = \sum_{i=1}^{K} \mathbf{1}[\hat{c}_i = c^*]$ denote the number of times the target concept is selected during K experiments. Then,*

$$\mathbb{P}(k^* \geq m) \geq \sum_{i=m}^{K} \binom{K}{i} (1-p)^i p^{K-i} \tag{24}$$

Remark A.4. This holds for every integer $m \in \{1, \ldots, K\}$ and can be derived from Thm. A.1, the independence of the K trials, and the binomial distribution CDF.

Let $k(S) = \sum_{i=1}^{K} \mathbf{1}[\hat{c}_i \in S]$. If $c^* \notin S$, all k^* trials that select c^* fall outside S, so $k^* \leq K - k(S)$. Alg. 1 stops only once $k(S) \geq t$, and $k(S)$ is an integer, so $k(S) \geq \lceil t \rceil$. Using Thm. A.3 with $m = K - \lceil t \rceil + 1$, we can derive:

$$\begin{aligned} \mathbb{P}(c^* \notin S) &\leq \mathbb{P}(k^* \leq K - k(S)) \\ &\leq \mathbb{P}(k^* \leq K - \lceil t \rceil) \\ &= 1 - \mathbb{P}(k^* \geq K - \lceil t \rceil + 1) \\ &\leq 1 - \sum_{i=K-\lceil t \rceil + 1}^{K} \binom{K}{i} (1-p)^i p^{K-i} \end{aligned} \tag{25}$$

Thus,

$$\mathbb{P}(c^* \in S) \geq \sum_{i=K-\lceil t \rceil + 1}^{K} \binom{K}{i} (1-p)^i p^{K-i}, \tag{26}$$

which finishes the proof. □

A.3 EXCLUSION GUARANTEE FOR IRRELEVANT CONCEPTS

The coverage guarantee in Thm. 4.1 ensures that the optimal concept enters S, but does not prevent irrelevant concepts from contaminating the set. The following theorem addresses this complementary concern in the same idealized setting, where the K trials use independent probing sets.

Theorem A.5 (Exclusion guarantee)**.** *Let $c \in C$ satisfy $\mathsf{sim}(f, c^*) - \mathsf{sim}(f, c) \geq \Delta$ for some $\Delta > 0$, and let p be the single-trial error probability from Cor. 3.4. Suppose the K trials use independent probing sets of size $|D_{\text{probe}}|$ drawn i.i.d. from the data distribution, as in Thm. 4.1. Then, for any frequency threshold $m \geq 1$:*

$$\mathbb{P}(k_c \geq m) \leq \sum_{i=m}^{K} \binom{K}{i} p^i (1-p)^{K-i}, \tag{27}$$

where $k_c = \sum_{i=1}^{K} \mathbf{1}[\hat{c}_i = c]$ is the number of trials that select concept c.

Proof. The proof proceeds in two steps.

Step 1: Single-trial error bound. Let $c \in C$ satisfy $\mathsf{sim}(f, c^*) - \mathsf{sim}(f, c) \geq \Delta$. From Thm. A.1, we have $\mathbb{P}(\hat{c} = c^*) \geq 1 - p$. Since $\hat{c} = c$ requires $\hat{\mathsf{sim}}(f, c; D_{\text{probe}}) \geq \hat{\mathsf{sim}}(f, c^*; D_{\text{probe}})$, and by Thm. 3.2 with probability at least $1 - p$ the PAC-style bound $\sup_{c' \in C} |\hat{\mathsf{sim}}(f, c'; D_{\text{probe}}) - \mathsf{sim}(f, c')| \leq \frac{\Delta}{2}$ holds, we have under this event:

$$\begin{aligned} \hat{\mathsf{sim}}(f, c^*; D_{\text{probe}}) &\geq \mathsf{sim}(f, c^*) - \frac{\Delta}{2} \\ &\geq \mathsf{sim}(f, c) + \frac{\Delta}{2} \\ &\geq \hat{\mathsf{sim}}(f, c; D_{\text{probe}}). \end{aligned} \tag{28}$$

Therefore, $\hat{c} = c$ can only occur when the PAC-style bound fails, giving $\mathbb{P}(\hat{c} = c) \leq p$.

Step 2: Binomial tail bound over K trials. Let $k_c = \sum_{i=1}^{K} \mathbf{1}[\hat{c}_i = c]$. Since in the idealized setting the K trials use independent probing sets and $\mathbb{P}(\hat{c}_i = c) \leq p$ for each trial, k_c is stochastically dominated by $\mathrm{Binom}(K, p)$. Therefore, for any threshold $m \geq 1$:

$$\mathbb{P}(k_c \geq m) \leq \mathbb{P}(\mathrm{Binom}(K, p) \geq m) = \sum_{i=m}^{K} \binom{K}{i} p^i (1-p)^{K-i}. \tag{29}$$

□

Remark A.6. When $p < m/K$, i.e., the expected count Kp falls below the threshold m, the bound decreases exponentially in K by standard Chernoff arguments. This provides a practical guideline: given a desired exclusion confidence, one can compute the minimum gap Δ or sample size $|D_{\text{probe}}|$ required to ensure irrelevant concepts do not enter the prediction set.

Remark A.7. Together with Thm. 4.1, this yields a two-sided characterization of the prediction set S in the idealized setting: the optimal concept is included (coverage), while concepts with a similarity gap of at least Δ are excluded (exclusion). For polysemantic neurons, this means all concepts appearing in S with substantial frequency must have similarity close to the best concept. The prediction set thus naturally separates the multiple concepts a neuron responds to from irrelevant noise, without requiring explicit knowledge of the number of desired concepts.

B Related Work

B.1 Neuron identification

The goal of neuron identification is to find a human-interpretable concept that describes the behavior and functionality of a specific neuron. A variety of methods have been proposed for neuron identification. Network Dissection (Bau et al., 2017) is a pioneering work with the idea of comparing neuron activations with ground-truth concept masks. Subsequent work explored extensions such as compositional explanations (Mu & Andreas, 2020), automated labeling with CLIP (Oikarinen & Weng, 2023), and multimodal summarization (Bai et al., 2025). More recent approaches expand the concept space to linear combinations (Oikarinen & Weng, 2024). While these advances provide useful empirical tools, in this work we aim to fill the gap by providing a principled theoretical foundation for neuron identification.

B.2 Principled framework for neuron identification

To unify the rapidly growing neuron identification methods, Oikarinen et al. (2025) design a framework that summarizes most neuron identification algorithms into three major components: neuron representation, concept activations, and similarity metrics. Additionally, two meta-tests are proposed to compare similarity metrics. While this work provides a good starting point, it does not analyze how the finite probing set affects the identified concept. Our framework is complementary: by casting identification as ERM over the candidate concept set, it lets generalization and resampling results from statistical learning theory apply to neuron identification directly.

B.3 Formal validation and certification in mechanistic interpretability

Recent work provides guarantees for objectives complementary to finite-sample concept assignment. Palumbo et al. (2025) axiomatize and validate an entire proposed mechanistic interpretation, while Hadad et al. (2026) certify robustness, patching behavior, and minimality of multi-component circuits and Anani et al. (2026) certify circuit stability under bounded edits to a concept dataset. We instead analyze the assignment of a concept from a fixed candidate set to a single unit, with uncertainty arising from finite probing data. Our guarantees concern population similarity under the chosen metric and distribution; they neither validate a complete interpretation nor certify a circuit or causal mechanism.

C Details on recall, precision, and IoU convergence rate derivation

In the main text, we mention the key idea of deriving the convergence rate r for recall, precision, and IoU: they can be regarded as special cases of accuracy where data are limited to a subgroup. For recall:

$$\begin{aligned}\mathsf{sim}_{\text{recall}}(f, c) &= \frac{\mathbb{P}(f(x) = 1, c(x) = 1)}{\mathbb{P}(c(x) = 1)} \\ &= \mathbb{P}(f(x) = c(x) \mid c(x) = 1).\end{aligned} \tag{30}$$

Therefore, we can regard the calculation of recall as a rejection sampling process: the samples satisfying $c(x) = 1$ are kept and the others are rejected. Then, accuracy is calculated on the remaining samples. Thus, the convergence rate can be calculated by inserting the effective sample size $|\{c(x) = 1 \mid x \in D_{\text{probe}}\}|$ into the accuracy convergence rate:

$$r_{\text{recall}} = \sqrt{\frac{\log(\frac{2}{\delta})}{2|\{c(x) = 1 \mid x \in D_{\text{probe}}\}|}}. \tag{31}$$

For precision and IoU, the derivation is similar.

D EXPERIMENTAL DETAILS FOR BOOTSTRAP EXPLANATION

This appendix describes the setup of the quantitative evaluation in Sec. 4.3 and reports its full numerical results. Sec. D.1 covers the Broden evaluation (Fig. 3(a, b) and Fig. D.1), Sec. D.2 covers the controlled coverage experiment (Fig. 3(c, d)), and Sec. D.3 studies whether BE recovers a secondary concept with slightly lower similarity.

Throughout, BE follows Alg. 1. Each of the K bootstrap samples is drawn from D_{probe} with replacement and has the same size as D_{probe}, and the prediction set S collects the most frequently selected concepts until their counts reach the threshold t. Coverage is the fraction of neurons (on Broden) or repetitions (in the controlled experiments) whose prediction set contains the reference concept. Single-concept assignment, called the single trial in the controlled experiments, applies the identification procedure once to the full probing set and returns a singleton set.

Algorithm 1 BE: Generating a concept prediction set for a target neuron

Require: Concept set C, probing dataset D_{probe}, target neuron f, neuron identification procedure $\texttt{Identify}(C, f, D_{\text{probe}})$, bootstrap sample count K, bootstrap count threshold t
Ensure: Prediction set S of candidate concepts
for $i = 1$ to K **do**
  Sample dataset D_i from D_{probe} with replacement (same size as D_{probe})
  Calculate $\hat{c}_i = \texttt{Identify}(C, f, D_i)$
end for
For each concept $c_j \in C$, count the number of its appearances: $k_j = \sum_{i=1}^{K} \mathbf{1}[\hat{c}_i = c_j]$
Sort concepts by frequency, $k_{r_1} \geq k_{r_2} \geq \cdots \geq k_{r_s}$, where s is the number of different concepts generated during bootstrapping
Initialize $S \leftarrow \emptyset$, $j \leftarrow 1$, cur_count $\leftarrow 0$
while cur_count $< t$ **do**
  Add c_{r_j} to S: $S \leftarrow S \cup \{c_{r_j}\}$
  cur_count $\leftarrow$ cur_count $+ k_{r_j}$
  Update $j \leftarrow j + 1$
end while

D.1 BRODEN EVALUATION

We use all 63,305 RGB images in Broden and exclude the segmentation masks. The images are randomly split (seed 42) into 31,652 probe images and 31,653 test images. We apply CLIP-Dissect with the ViT-B/16 CLIP encoder to all 2,048 layer-4 neurons of an ImageNet-trained ResNet-50, using the 20,000-word concept set and either soft-WPMI or cosine similarity; neuron activations are averaged over spatial positions. For each metric, the concept ranked first on the full test set serves as the reference concept. For Fig. 3(a, b), the smaller probe sets contain the first 10%, 25%, and 50% of a random permutation of the probe split, so they are nested, while the test set and the reference concepts stay fixed. BE uses $K = 100$ and $t = 95$ at every probe size. Fig. D.1 varies t with $K = 100$ and all probe images.

Effect of the threshold. Fig. D.1 shows the coverage and mean set size for different thresholds t/K. Raising the threshold t increases coverage but also enlarges the prediction sets.

D.2 CONTROLLED SYNTHETIC EXPERIMENT

Setup. We simulate a binary neuron and use accuracy as the similarity metric, so the population similarity of every concept is known. For each probe example and candidate concept c_j, the agreement indicator $\mathbf{1}[c_j(x) = f(x)]$ is drawn independently from a Bernoulli distribution with mean a_j. The optimal concept has $a_j = s^* = 0.9$, and the other 999 concepts have $a_j = s^* - \Delta = 0.85$. The empirical similarity of a concept on a sample is the mean of its agreement indicators, and the identification procedure returns the concept with the highest empirical similarity. Each of the 2,000 repetitions draws a new probing set and applies both the single trial and BE to it. Fig. 3(c, d) varies $|D_{\text{probe}}|$ from 300 to 10,000 with $K = 100$ and $t = 95$. We additionally vary K from 1 to 200 with $t = 0.95K$, and t/K from 0.5 to 0.99 with $K = 100$, both at $|D_{\text{probe}}| = 1{,}500$.

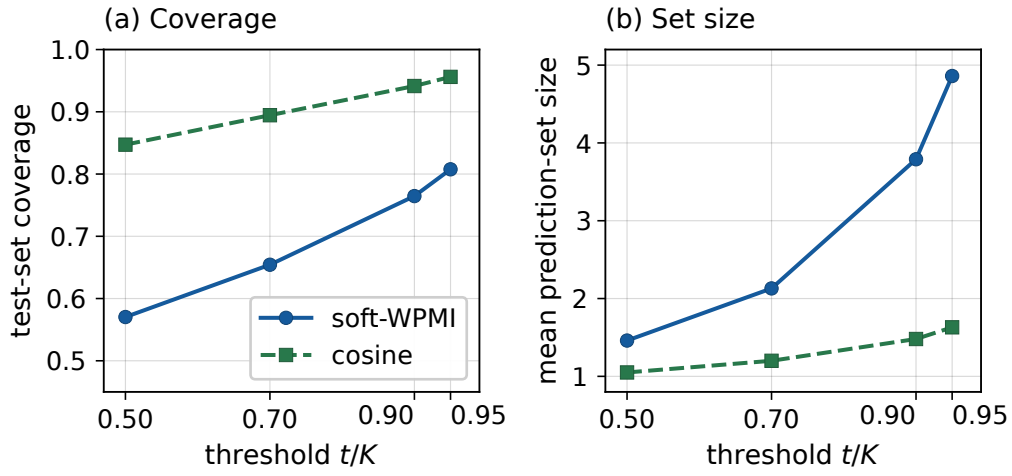


Figure D.1: Effect of the threshold t/K on (a) test-set coverage and (b) mean prediction-set size, using the full Broden probe set and $K = 100$. Coverage is the fraction of neurons whose prediction set contains the top-ranked concept on the test set.

Results. Table D.1 shows that coverage saturates around $K = 10$, whereas the set size keeps growing with K. Table D.2 shows that, at this probe size, coverage is close to one for every threshold, so raising t mainly enlarges the prediction set, in line with the Broden results in Fig. D.1.

Table D.1: Coverage and mean prediction-set size in the controlled experiment as the number of bootstrap trials K varies, with $|D_{\text{probe}}| = 1{,}500$ and $t = 0.95K$.

K	t	Coverage	Mean $\lvert S\rvert$
1	.9	.8075	1.00
2	1.9	.9265	1.30
5	4.8	.9855	1.86
10	9.5	.9950	2.69
20	19.0	.9975	3.41
50	47.5	1.0000	6.25
100	95.0	.9995	8.81
150	142.5	.9995	11.23
200	190.0	1.0000	12.57

Table D.2: Coverage and mean prediction-set size in the controlled experiment as the threshold t varies, with $|D_{\text{probe}}| = 1{,}500$ and $K = 100$.

t/K	Coverage	Mean $\lvert S\rvert$
.50	.9980	1.23
.60	.9990	1.51
.70	.9995	2.07
.80	.9995	3.34
.90	.9995	6.22
.95	.9995	8.81
.99	1.0000	11.77

D.3 Recovering a secondary concept

Setup. To simulate the polysemantic neuron setting, we use the generative model of Sec. D.2 with a main concept at $s^* = 0.9$, a secondary concept at $s^* - \delta$, and 998 distractors at $s^* - \Delta$ with $\Delta = 0.05$. We vary $\delta/\Delta \in \{0.1, 0.2, 0.4\}$; $\delta/\Delta = 1$ serves as a monosemantic reference, in which the secondary concept is indistinguishable from the distractors. BE uses $K = 100$ and $t = 95$, and results are averaged over 2,000 repetitions. The relevant set R contains the main and secondary

concepts for $\delta < \Delta$ and only the main concept for the monosemantic case. Purity is the fraction of concepts in S that belong to R, averaged over repetitions.

Results. From Fig. D.2, we can see:

1. Distractors are excluded from the prediction set once the probe set resolves the margin Δ, which happens between 2,000 and 5,000 probe images for every δ.
2. Beyond this point, BE retains the secondary concept over a range of probe sizes that widens as δ decreases. Within this range, the prediction set consists of the main and secondary concepts only (Table D.3).
3. With more probe data, the gap δ is also resolved and the prediction set reduces to the main concept. In the monosemantic reference, the concept at the position of the secondary concept is almost never included.

BE therefore reports the concepts whose similarity cannot be distinguished from that of the best concept at the given probe size. A secondary concept with a small gap appears in a small set free of distractors, whereas a large set dominated by distractors indicates that the probe set is too small to resolve the margin of the main concept.

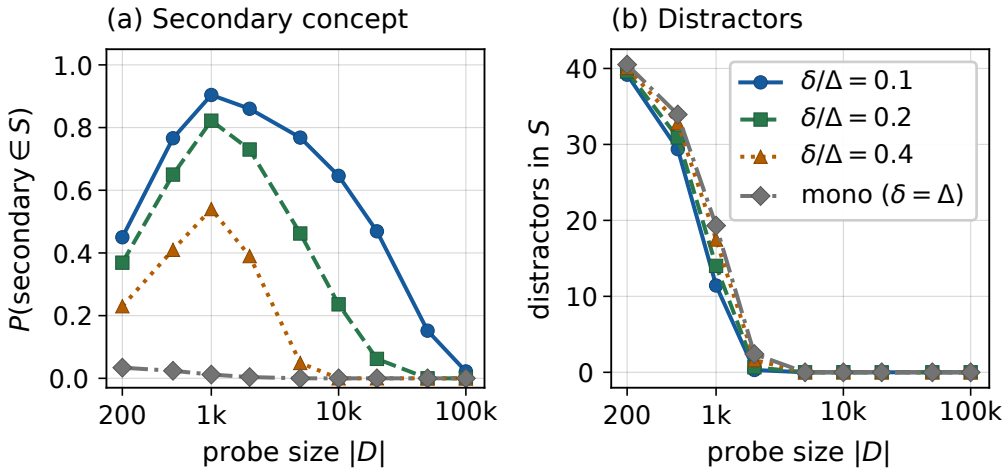


Figure D.2: Recovery of a secondary concept whose similarity is lower than that of the main concept by δ, with distractors at margin $\Delta = 0.05$, $K = 100$, and $t = 95$. (a) Probability that the secondary concept is in S. (b) Mean number of distractors in S. For the monosemantic reference ($\delta = \Delta$), the secondary concept is itself a distractor.

Table D.3: Exact recovery ($S = R$) and mean purity for the settings in Sec. D.3, with $K = 100$ and $t = 95$. The relevant set R contains the main and secondary concepts for $\delta < \Delta$ and only the main concept for the reference. $S = R$ is the fraction of repetitions in which the prediction set equals R, and purity is the fraction of concepts in S that belong to R, averaged over repetitions.

	$\delta/\Delta = 0.1$		$\delta/\Delta = 0.2$		$\delta/\Delta = 0.4$		Reference	
$\lvert D_{\text{probe}}\rvert$	$S = R$	Purity	$S = R$	Purity	$S = R$	Purity	$S = R$	Purity
200	.000	.026	.000	.024	.000	.020	.000	.015
500	.002	.071	.002	.061	.001	.049	.004	.036
1,000	.120	.330	.070	.267	.016	.196	.074	.154
2,000	.768	.956	.584	.914	.202	.816	.660	.740
5,000	.762	1.000	.462	1.000	.049	1.000	1.000	1.000
10,000	.645	1.000	.236	1.000	.000	1.000	1.000	1.000
20,000	.468	1.000	.062	1.000	.000	1.000	1.000	1.000
50,000	.152	1.000	.000	1.000	.000	1.000	1.000	1.000
100,000	.022	1.000	.000	1.000	.000	1.000	1.000	1.000

E SIMULATION STUDIES FOR THE FIDELITY BOUNDS

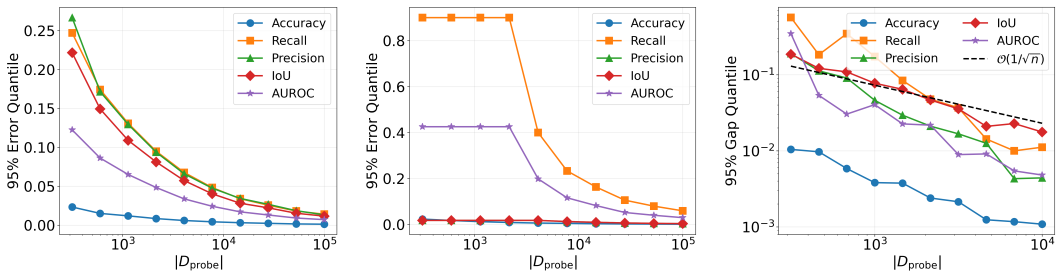
To verify the theoretical bounds developed in Sec. 3.1 and systematically compare the behavior of different similarity metrics, we conduct simulations on a synthetic dataset. This controlled environment provides ground-truth similarity values and allows us to simulate a variety of settings. Specifically, we use a binary concept $c(x) \in \{0, 1\}$ for simplicity. Neuron activations $f(x)$ are binarized by setting the top-5% activations to 1 and the remaining to 0. The joint distribution of f, c is controlled by the probability matrix M: $M_{ij} = \mathbb{P}(f(x) = i, c(x) = j),\ i, j \in \{0, 1\}$. We conduct two experiments: (1) a **single-concept study** to compare convergence rates and (2) a **multi-concept simulation** to verify Thm. 3.2. Additionally, we provide a real-data study on the ImageNet validation set in Sec. G.

Experiment 1: Convergence rate r. In Thm. 3.2, the key factor that controls the generalization gap $g(D_{\text{probe}}, C, f)$ is the convergence rate r. To investigate this, we generate synthetic data and compare different similarity functions. For the concept, we study the following two settings: (1) a regular concept with frequency 5% and (2) a rare concept with frequency only 0.1%. We simulate with $N_{\text{exp}} = 1000$ randomly sampled datasets and plot how the 95% quantile of error ($|\hat{\mathsf{sim}}(f, c; D_{\text{probe}}) - \mathsf{sim}(f, c)|$) changes with the number of samples. Details of the procedure are given in Sec. E.1. From the simulation results in Fig. E.1, we can see that

1. Accuracy has the fastest convergence in both cases. For the regular concept, IoU, recall, and precision are similar. AUROC converges faster than them.
2. For the rare concept, the convergence pattern differs: AUROC and recall are much worse than precision and IoU. This matches our analysis in Sec. 3.2, where we showed that AUROC converges much more slowly when the concept frequency is low.

Experiment 2: Gap simulation. We further verify Thm. 3.2 by generating $|C| = 1{,}000$ synthetic concepts with log-uniformly distributed frequencies and measuring the gap ($\mathsf{sim}(f, c^*) - \mathsf{sim}(f, \hat{c})$) over $N_{\text{exp}} = 1{,}000$ trials (full procedure in Sec. E.1). In Fig. E.1c, we observe that: (1) All curves have a similar slope to the reference $\mathcal{O}(\sqrt{1/n})$ curve, suggesting an asymptotic convergence rate of $\mathcal{O}(\sqrt{1/n})$, which is consistent with our theoretical analysis. (2) For the constant term, accuracy has the fastest convergence and AUROC is second. This matches our simulation of r in **Experiment 1, Setting 1 (Regular concept)**, supporting our conclusion.

In summary, these simulation experiments empirically validate the correctness of our theory in Sec. 3.1 and show its potential to help users choose a similarity metric under different settings.



(a) Regular concept (5%) (b) Rare concept (0.1%) (c) Generalization gap

Figure E.1: Simulation results for fidelity bounds. (a, b) 95% quantile of estimation error for 5 similarity metrics: accuracy converges fastest; AUROC degrades for rare concepts. (c) Generalization gap (Thm. 3.2) versus probing dataset size, confirming the $\mathcal{O}(1/\sqrt{n})$ rate.

E.1 SIMULATION DETAILS

Experiment 1. The procedure of Experiment 1 can be described as follows:

1. **Generate simulation data.** We first generate the simulation data following the settings described above. We generate paired binary variables representing ground-truth concept

activations and neuron responses i.i.d. by directly sampling from the probability distribution specified by the probability matrix M. For the two settings, the probability matrices M are:

- **Setting 1: regular concept.** $M =$

$$\begin{array}{c} \\ f=0 \\ f=1 \end{array} \begin{array}{c} \begin{array}{cc} c=0 & c=1 \end{array} \\ \begin{pmatrix} 0.93 & 0.02 \\ 0.02 & 0.03 \end{pmatrix} \end{array}$$

- **Setting 2: rare concept.** $M =$

$$\begin{array}{c} \\ f=0 \\ f=1 \end{array} \begin{array}{c} \begin{array}{cc} c=0 & c=1 \end{array} \\ \begin{pmatrix} 0.9499 & 0.0001 \\ 0.0491 & 0.0009 \end{pmatrix} \end{array}$$

  This simulates a rare concept (frequency 0.001), which often occurs when the concept is fine-grained.

2. **Calculate ground-truth metrics.** Given the probability matrix M, the ground-truth (population-level) value of each metric is computed directly from the joint probabilities $W_{ij} = M_{ij}$:
   - Accuracy: $\mathsf{sim}_{\text{acc}} = W_{11} + W_{00}$.
   - Recall: $\mathsf{sim}_{\text{recall}} = W_{11}/(W_{11} + W_{01})$.
   - Precision: $\mathsf{sim}_{\text{precision}} = W_{11}/(W_{11} + W_{10})$.
   - IoU: $\mathsf{sim}_{\text{IoU}} = W_{11}/(W_{11} + W_{10} + W_{01})$.
   - AUROC: $\mathsf{sim}_{\text{AUROC}} = \frac{1}{2}\left(\frac{W_{11}}{W_{11}+W_{01}} + \frac{W_{00}}{W_{00}+W_{10}}\right)$.

   These values serve as the ground truth against which the empirical estimates are compared.
3. **Simulation and error computation.** For each probing dataset size $N_{\text{sample}} \in \{316, 599, \ldots, 100{,}000\}$ (10 values log-spaced between $10^{2.5}$ and 10^5), we repeat the following for $N_{\text{exp}} = 1000$ iterations:
   (a) Sample N_{sample} data points i.i.d. from the joint distribution specified by M using a multinomial draw.
   (b) Compute the empirical estimate $\hat{\mathsf{sim}}$ of each metric on the sampled data (e.g., for IoU: $\hat{\mathsf{sim}}_{\text{IoU}} = \frac{\sum_i \mathbf{1}[f(x_i)=1 \wedge c(x_i)=1]}{\sum_i \mathbf{1}[f(x_i)=1 \vee c(x_i)=1]}$).
   (c) Compute the absolute error: $\text{err}_i = |\hat{\mathsf{sim}} - \mathsf{sim}|$, where sim is the ground-truth value from Step 2.

   After all iterations, we report the 0.95-quantile of $\{\text{err}_i\}_{i=1}^{N_{\text{exp}}}$ as the empirical convergence rate at confidence level $1 - \delta = 0.95$, i.e., we estimate r such that $\mathbb{P}(|\hat{\mathsf{sim}} - \mathsf{sim}| \le r) \ge 0.95$.
4. **Theoretical comparison.** We also plot the theoretical convergence rate r from Table 1 at $\delta = 0.05$ for each metric. For example, $r_{\text{acc}} = \sqrt{\frac{\log(2/\delta)}{2N_{\text{sample}}}}$ and $r_{\text{IoU}} = \sqrt{\frac{\log(2/\delta)}{2(W_{11}+W_{10}+W_{01})N_{\text{sample}}}}$. The theoretical curves serve as upper bounds on the empirical 95% quantile, confirming the validity of our bounds. The theoretical r for both settings is shown in Fig. E.2.

Experiment 2. The procedure of Experiment 2 (gap simulation) is as follows:

1. **Generate neuron representation.** The binarized neuron representation $f(x)$ is generated by setting the top-5% of activations to 1 and the rest to 0, i.e., $M_{10} + M_{11} = 0.05$.
2. **Generate concepts.** We generate $|C| = 1000$ concepts as the candidate set. For each concept c_i, we generate its frequency $\mathbb{P}(c_i(x) = 1) = M_{01} + M_{11}$ from a log-uniform distribution in $(10^{-4}, 10^{-1})$. Then, we sample $M_{11} = \mathbb{P}(f(x) = 1, c_i(x) = 1)$ uniformly from $(0, \min[\mathbb{P}(f(x) = 1), \mathbb{P}(c_i(x) = 1)])$ to ensure validity. The remaining entries of M are inferred from the concept frequency and activation binarization. Given the probabilities, we compute $\mathbb{P}(c_i(x) \mid f(x))$ and sample $c_i(x)$ accordingly.
3. **Simulation.** We repeat the above steps $N_{\text{exp}} = 1000$ times. For each trial, we compute the similarity of the top-ranked concept $\hat{c}$ and the ground-truth maximum similarity, then record the gap $\max_{c \in C} \mathsf{sim}(f, c) - \mathsf{sim}(f, \hat{c})$. The 95% quantile over all trials approximates the bound at $1 - \delta = 95\%$.

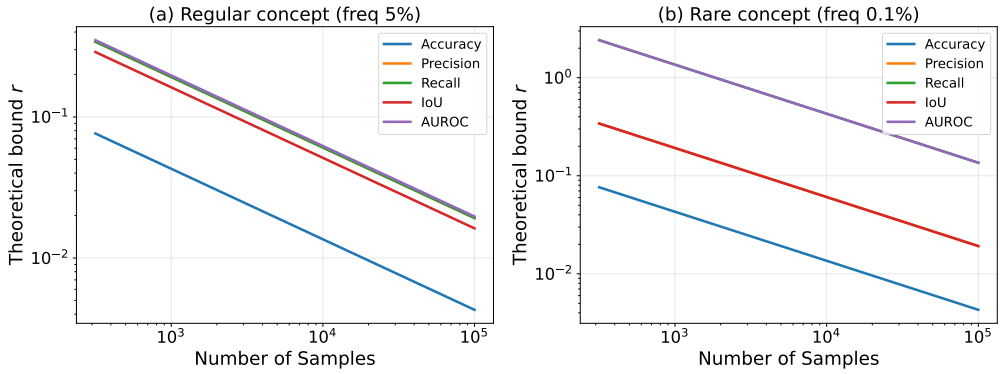


Figure E.2: Theoretical convergence rate r for five similarity metrics under both settings ($\delta = 0.05$). (a) Regular concept (frequency 5%): accuracy converges fastest; all other metrics have larger r due to smaller effective sample sizes. (b) Rare concept (frequency 0.1%): AUROC and recall degrade significantly because their effective sample sizes depend on the concept frequency ρ, which is very small in this setting.

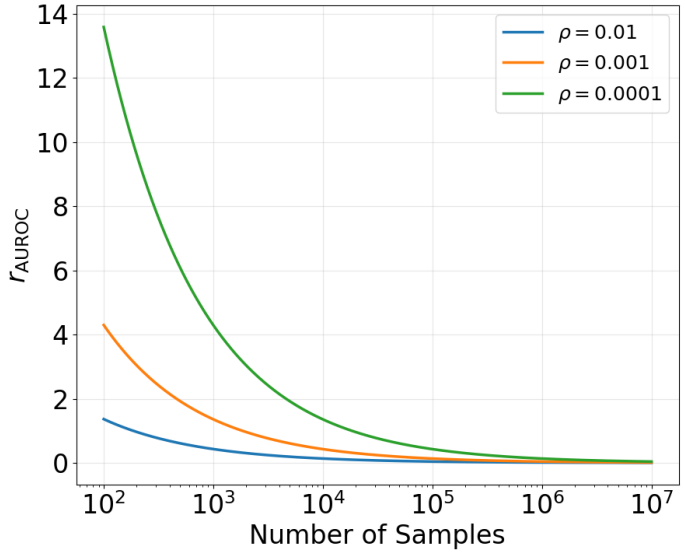


Figure E.3: Convergence rate r_{AUROC} with respect to probing dataset size $|D_{\text{probe}}|$ under different concept frequencies ρ. When both ρ and $|D_{\text{probe}}|$ are small, r_{AUROC} blows up, indicating that AUROC-based identification is unreliable for rare concepts.

F COMPUTATIONAL COST OF BOOTSTRAP EXPLANATION (BE)

In this section, we study the computational cost of BE. Theoretically, since the BE method requires K rounds of bootstrapping and running the original neuron identification algorithm, the running time should scale linearly with K. In practice, however, the BE time cost is not simply K times that of the original algorithm, as a significant portion of computation can often be shared among bootstrap samples, saving considerable time. Take CLIP-Dissect as an example: The CLIP-Dissect algorithm can be roughly divided into three steps:

1. Collect features from the target model and CLIP model.
2. Calculate the similarity matrix between target features and CLIP features.
3. Select the final concept with the highest similarity.

In the implementation of BE, the first step can be shared among bootstrap samples: we pre-compute the features of the target model and CLIP on the whole probing dataset. For each bootstrap sample, we only need to fetch the corresponding features, calculate similarity, and select the highest-similarity concept. With this optimization, we run experiments on a ConvNeXt-base model (Liu et al., 2022).[4] The profiling results are shown in Fig. F.1. We can see that although we run $K = 100$ bootstrap samples, the time overhead is only about 30% and the memory overhead is about 15%. Fig. F.2 shows how runtime changes with K, illustrating that our runtime is not as high as K times that of the original algorithm. Further, to understand the impact of concept set size, we compare results on

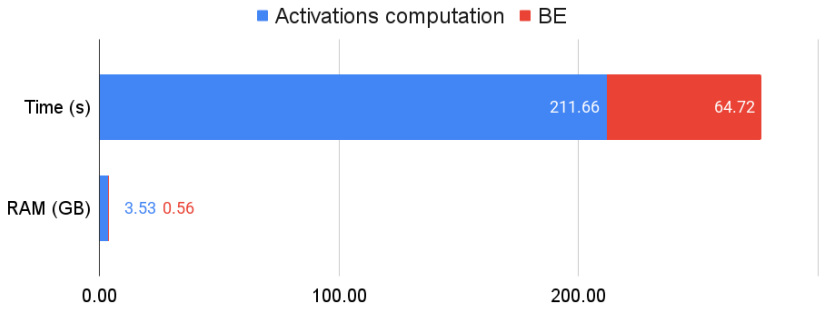


Figure F.1: Profiling result of CLIP-Dissect on a ConvNeXt-base model. $K = 100$ bootstrap samples are used.

two concept sets: the Broden concept set with 1,198 concepts and 20k English words with 20,000

[4]Experiments are done with an NVIDIA RTX A5000 GPU and an AMD Ryzen Threadripper PRO 3975WX CPU.

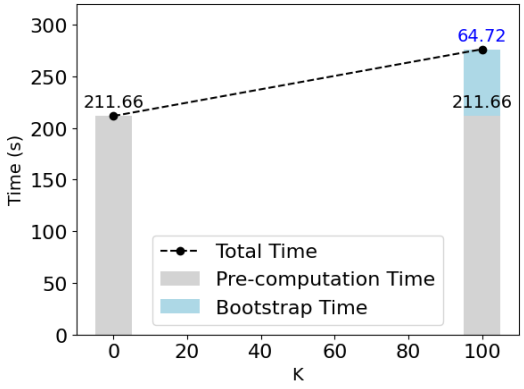


Figure F.2: Runtime vs. K for CLIP-Dissect on a ConvNeXt-base model.

concepts. The results are shown in Table F.1. From the results, we can see that bootstrapping time increases significantly with a larger concept set. However, we argue that the major cause is that CLIP-Dissect uses soft-WPMI, which is much slower with a larger concept set. The bootstrap wrapper itself does not introduce higher overhead.

Stage	Broden ($\|C\| = 1{,}198$)	20k ($\|C\| = 20{,}000$)
Activation computation	211.7	277.6
Bootstrapping	64.7	348.7

Table F.1: Runtime (s) comparison with two concept sets: Broden and 20k English words.

G ADDITIONAL FIDELITY EXPERIMENTS FOR CONTINUOUS METRICS

In Sec. 3.2, we mainly study bounds for metrics with binarized neuron representations. In this section, we conduct experiments on several metrics for continuous neuron representations: AUROC (Bykov et al., 2023), AUPRC, correlation (Oikarinen & Weng, 2024), WPMI (Oikarinen & Weng, 2023), and MAD (Kopf et al., 2024). AUROC has been defined in the main text. To calculate AUPRC, we first sort $f(x_i)$ from largest to smallest: $f(x_{(1)}) \geq f(x_{(2)}) \geq \cdots \geq f(x_{(n)})$. Thus, the precision at the kth threshold can be defined as

$$\mathrm{Prec}(k) = \frac{1}{k}\sum_{i=1}^{k} c_{(i)}. \tag{32}$$

The recall can be defined as

$$\mathrm{Rec}(k) := \frac{1}{\sum_{i=1}^{n} c_i}\sum_{i=1}^{k} c_{(i)}. \tag{33}$$

The AUPRC can be calculated as:

$$\hat{\mathsf{sim}}_{\mathrm{AUPRC}}(f, c) := \sum_{k|c_{(k)}=1} \frac{\mathrm{Prec}(k)}{\sum_{i=1}^{n} c_i}. \tag{34}$$

For MAD,

$$\mathsf{sim}_{\mathrm{MAD}}(f, c) = \mathbb{E}_{x|c(x)=1} f(x) - \mathbb{E}_{x|c(x)=0} f(x) \tag{35}$$

For WPMI, we take the definition in Oikarinen et al. (2025):

$$\mathsf{sim}_{\mathrm{WPMI}} = \mathbb{E}_{f(x)=1}[\log(c(x)) - \lambda \log[\mathbb{E}(c(x))]]. \tag{36}$$

Here, we take $\lambda = 1$. For correlation:

$$\mathsf{sim}_{\mathrm{corr}}(f, c) := \frac{\mathbb{E}\big[(f(x) - \mu_f)\,(c(x) - \mu_c)\big]}{\sqrt{\sigma_f^2\,\sigma_c^2}}, \tag{37}$$

where $\mu_f, \mu_c, \sigma_f, \sigma_c$ are the mean and standard deviation of f and c, respectively. For these metrics, a closed-form expression for the convergence rate is challenging to derive. Thus, we conduct empirical studies with synthetic data and real ImageNet validation data.

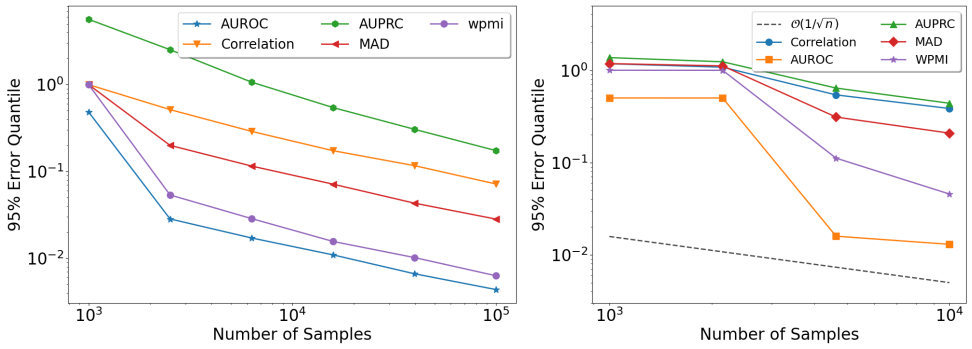
G.1 SYNTHETIC EXPERIMENT

Data generation For the synthetic dataset, we construct a simple conditional Gaussian data-generating process to obtain pairs of binary concept activations and continuous neuron representations. For each sample $i \in \{1, \ldots, n\}$ we first randomly draw a binary concept label $c_i \in \{0, 1\}$, where $\mathbb{P}(c_i = 1) = p$ is a hyperparameter. Conditioned on c_i, we then sample a one-dimensional neuron representation $z_i \in \mathbb{R}$ from a Gaussian distribution whose mean and variance depend on the concept state:

$$z_i \mid c_i = \begin{cases} \mathcal{N}(\mu_{\text{pos}}, \sigma_{\text{pos}}^2), & \text{if } c_i = 1, \\ \mathcal{N}(\mu_{\text{neg}}, \sigma_{\text{neg}}^2), & \text{if } c_i = 0, \end{cases}$$

where $(\mu_{\text{pos}}, \sigma_{\text{pos}})$ and $(\mu_{\text{neg}}, \sigma_{\text{neg}})$ are hyperparameters controlling, respectively, the distribution of the neuron activation when the concept is present or absent. In practice, we generate n i.i.d. samples $\{(c_i, z_i)\}_{i=1}^{n}$ according to the above process. In the experiment, we take $p = 0.002$, $\mu_{\text{pos}} = 1$, $\mu_{\text{neg}} = 0$, $\sigma_{\text{pos}} = 0.2$, $\sigma_{\text{neg}} = 0.5$.

Experiments We approximate the ground-truth values of the metrics with 10^6 examples, then estimate the convergence rate $r(\cdot)$ by calculating the 95% quantile of error. Since the scale of MAD is significantly different from other metrics, we compute the relative error instead. The results are shown in Fig. G.1a. From the results, we see that all metrics also follow an asymptotic $\mathcal{O}(1/\sqrt{N})$ convergence rate.



(a) Synthetic dataset

(b) ImageNet validation dataset

Figure G.1: 95% error quantile on the synthetic dataset and the ImageNet validation set.

G.2 Real data experiment

Furthermore, we conduct an experiment on the ImageNet validation set with a ResNet-50 model. We choose the final-layer neurons, and the concepts are the class labels they predict. The results are shown in Fig. G.1b. Unlike the synthetic dataset, the AUROC metric performs significantly better with more samples. We hypothesize that the reason is that these neurons are specific to classifying these concepts; thus, their ability to distinguish concepts (what AUROC measures) is especially strong. In addition, similar to the synthetic data, most metrics show an asymptotic $\mathcal{O}(1/\sqrt{N})$ convergence rate.

H STUDY ON SAMPLE SIZE (THM. 3.2)

Thm. 3.2 suggests that more probing samples can reduce the generalization gap, implying better fidelity. To evaluate this, we conduct a qualitative study using CLIP-Dissect to study ResNet-50 layer-4 neurons on the ImageNet validation set with probing dataset sizes of $|D_{\text{probe}}| = 100, 1000, 10000, 50000$. We show some examples in Fig. H.1. From the examples, we can see that with more probing data the identified concept describes the neurons more accurately, supporting our theory.

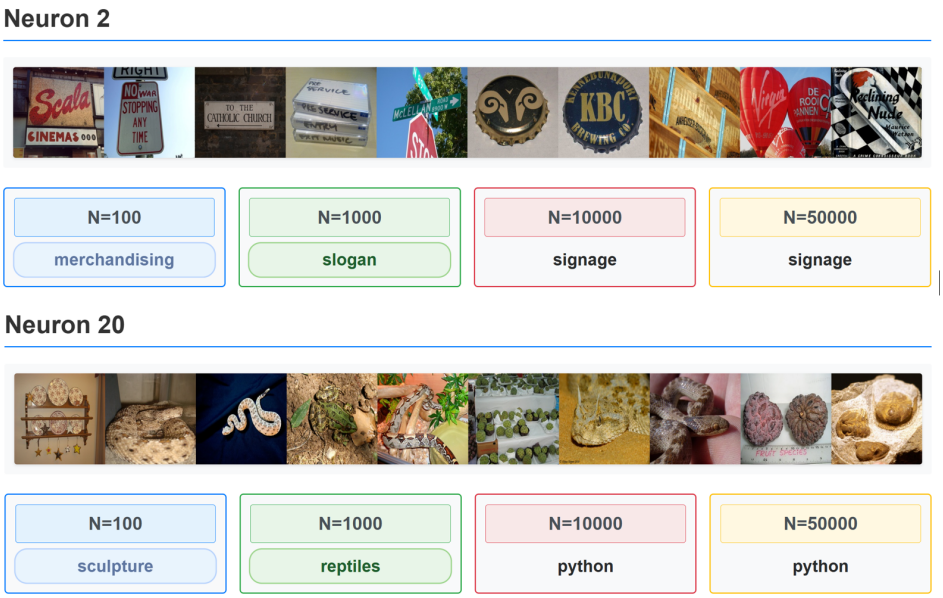


Figure H.1: CLIP-Dissect explanations with different numbers of probing samples.

I Additional Bootstrap Explanation (BE) experiments

I.1 Qualitative examples

We apply our BE method to two base methods: CLIP-Dissect (Oikarinen & Weng, 2023) and NetDissect (Bau et al., 2017). For each method, we adopt its own concept set and use Broden (Bau et al., 2017) as the probing dataset. We use a ResNet-50 model trained on the ImageNet dataset (Deng et al., 2009), evaluate all the layer-4 neurons (2048 neurons), run $K = 100$ bootstrap samples and choose the bootstrap count threshold $t = 0.95K = 95$ in Alg. 1. The results are shown in Fig. I.1. We present more results with more backbones (ViT-B/16 with 768 neurons and ConvNeXt-base with 1024 neurons) in Sec. I.2, and we discuss the computation overhead of BE in Sec. F.

From the results, we can observe interesting differences between these two methods: (i) CLIP-Dissect prefers more abstract concepts. For example, it gives concepts like "fostering" and "bibliographic." NetDissect, in contrast, tends to identify concrete concepts. (ii) In general, CLIP-Dissect provides more diverse concepts and sometimes captures ones missed by NetDissect (e.g., "birding" for Neuron 89). NetDissect is more stable across different bootstrap samples. A potential reason is that NetDissect utilizes localization information, which improves stability.

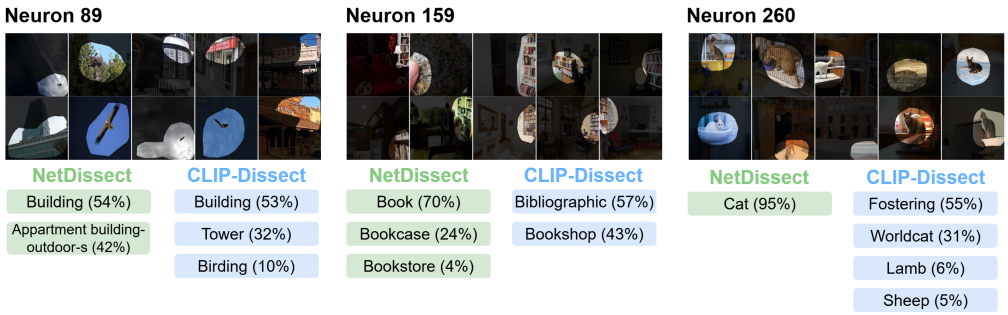


Figure I.1: Results of applying BE to NetDissect and CLIP-Dissect on ResNet-50 neurons. NetDissect shows more stable, concrete concepts. CLIP-Dissect outputs are more diverse and abstract. More examples are given in Sec. I.2.

I.2 Additional examples for BE

In Figures I.2 to I.5, we present additional results of applying BE to NetDissect and CLIP-Dissect. To enable a better comparison, we use the Broden (Bau et al., 2017) dataset as the probing dataset and the corresponding concepts as the concept set for both methods.

I.3 Experiments with different architectures

To understand the performance of NetDissect and CLIP-Dissect with different architectures, we conduct additional experiments with ConvNeXt-base (Liu et al., 2022) and ViT-B/16 (Dosovitskiy et al., 2020). We show example images in Figures I.6 and I.7. We compare (1) the average frequency of the top-1 concept and (2) the average size of the concept set that covers 90% of all runs. Since NetDissect does not support Vision Transformer features, we map the embedding back to the corresponding input patch to form a 2-D feature map. The results are shown in Tables I.1 and I.2. From the results, we can see that CLIP-Dissect provides more diverse concepts except on ConvNeXt-base. NetDissect results vary considerably across backbones: it is most stable for ViT-B/16 and most diverse for ConvNeXt-base.

Method	ResNet-50	ViT-B/16	ConvNeXt-base
CLIP-Dissect	66.9%	51.2%	53.7%
NetDissect	79.2%	91.3%	27.8%

Table I.1: Comparison of top-1 concept frequency of CLIP-Dissect and NetDissect on three vision backbones.

Method	ResNet-50	ViT-B/16	ConvNeXt-base
CLIP-Dissect	3.11	5.93	5.61
NetDissect	2.08	1.34	6.99

Table I.2: Comparison of 90% coverage concept set size of CLIP-Dissect and NetDissect on three vision backbones.

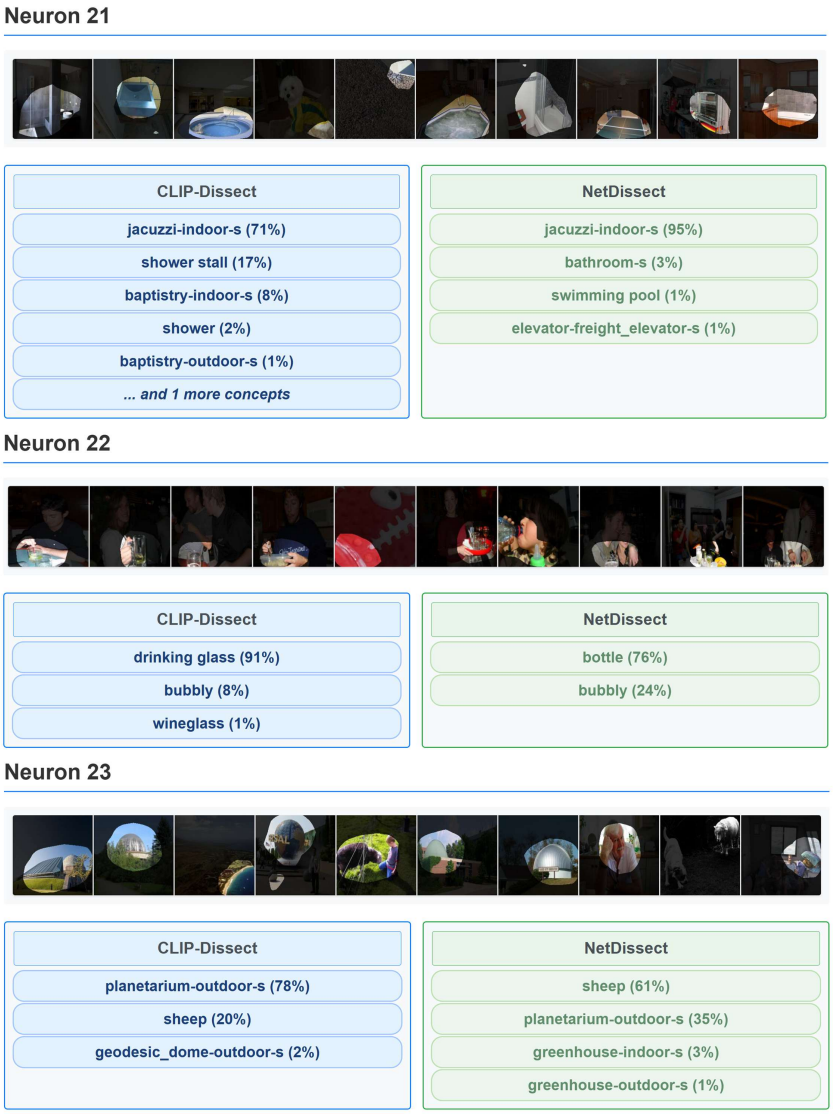


Figure I.2: Additional results of applying BE to NetDissect and CLIP-Dissect on ResNet-50 neurons.

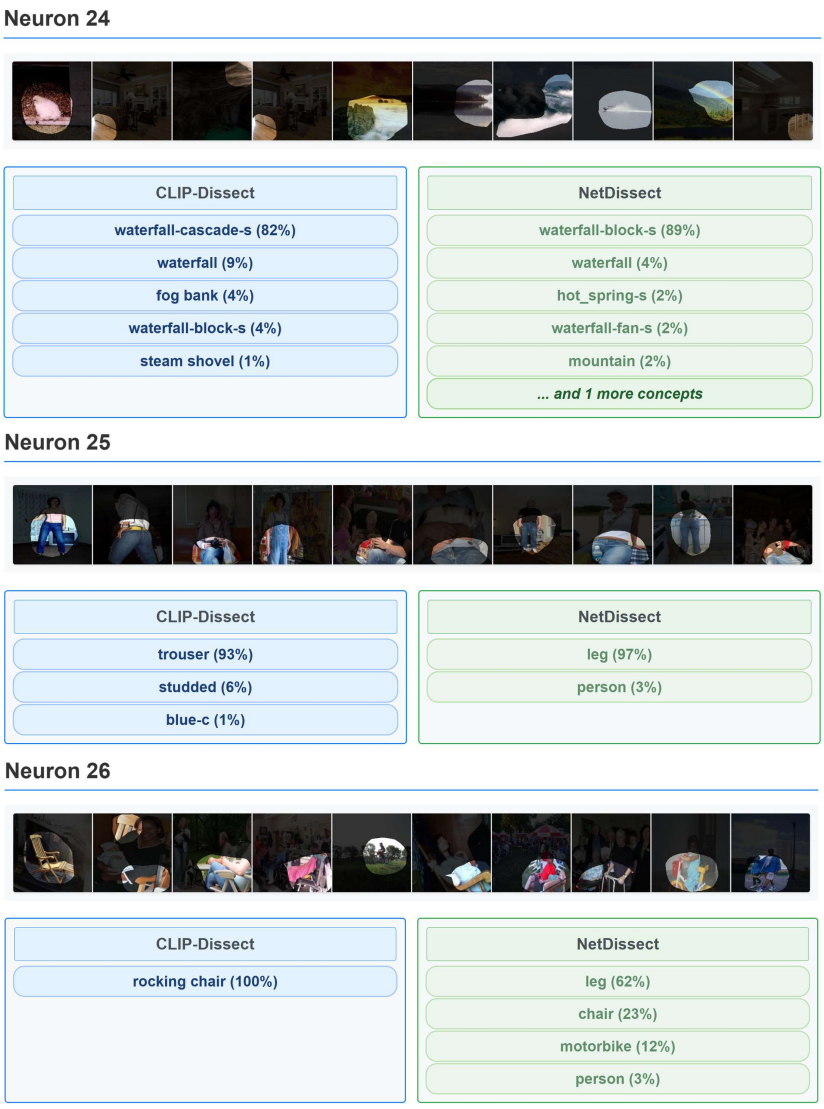


Figure I.3: Additional results of applying BE to NetDissect and CLIP-Dissect on ResNet-50 neurons.

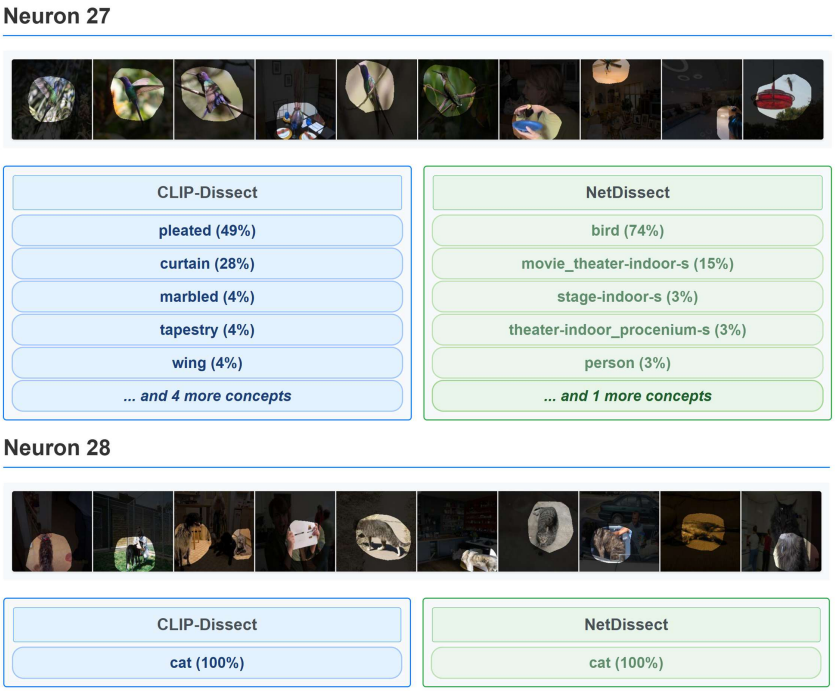


Figure I.4: Additional results of applying BE to NetDissect and CLIP-Dissect on ResNet-50 neurons.

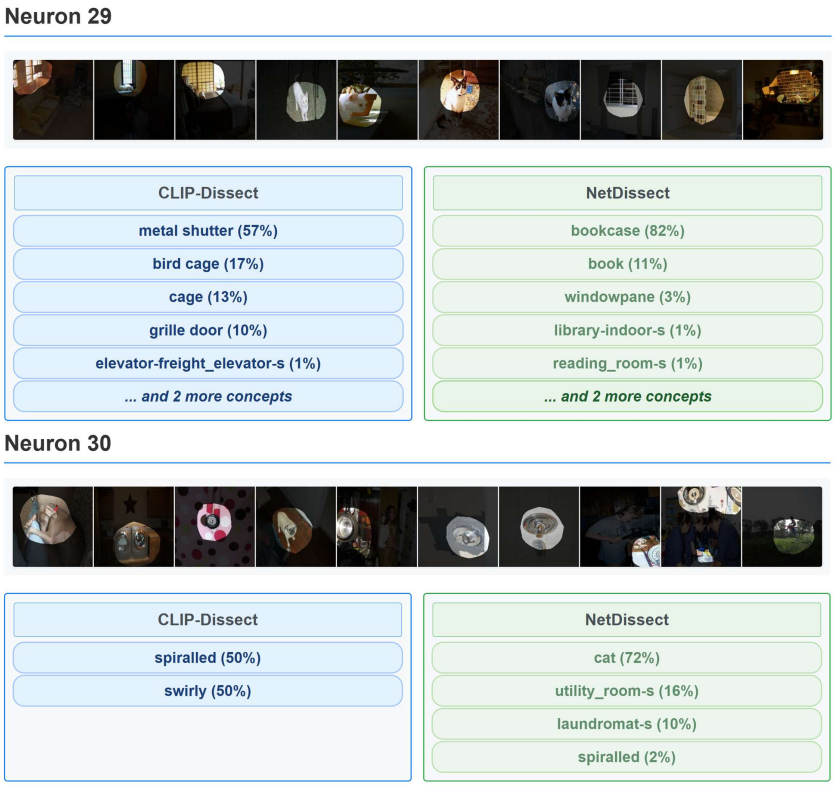


Figure I.5: Additional results of applying BE to NetDissect and CLIP-Dissect on ResNet-50 neurons.

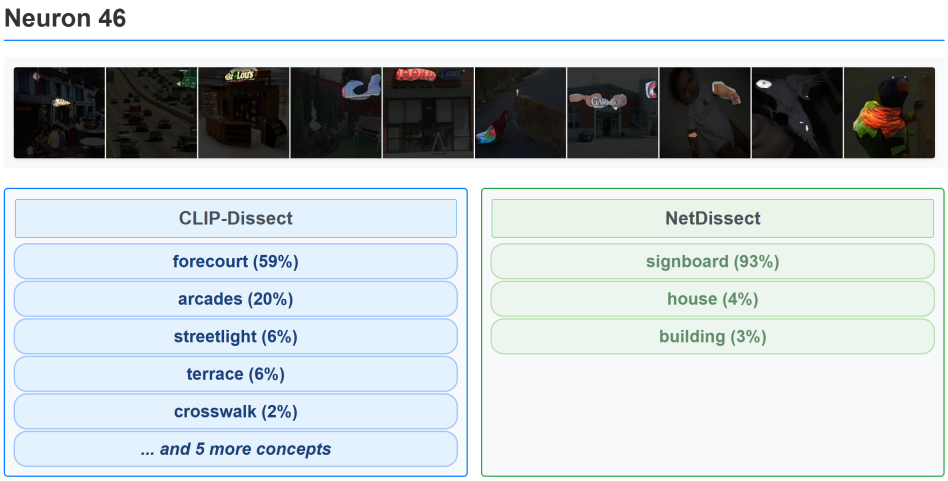


Figure I.6: BE example on ViT-B/16, last encoder layer.

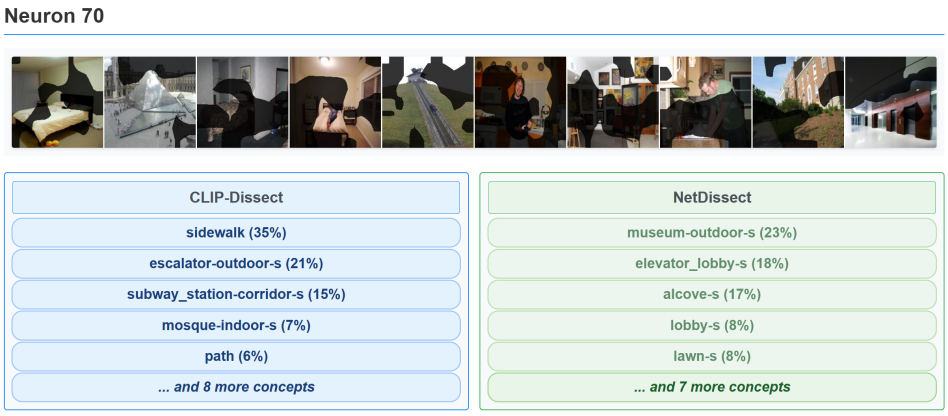


Figure I.7: BE example on ConvNeXt-base, last layer.

J Extension of Fidelity Bounds to Linear Concept Spaces

As noted in the main text, Thm. 3.2 assumes a finite concept set C and uses a union bound over $|C|$ elements. When neurons are described as linear combinations of m base concepts (Oikarinen & Weng, 2024), the concept space C_{lin} becomes infinite, and the union bound no longer applies directly. In this section, we show how to replace $|C|$ with the *covering number* of the coefficient set, yielding a bound that grows exponentially in m but remains finite. We require a Lipschitz regularity condition on the similarity function, which holds naturally for continuous metrics (e.g., correlation, MAD, AUROC).

J.1 Setup

Let $C_{\text{base}} = \{c_1, c_2, \ldots, c_m\}$ be a finite set of m base concept functions, where each $c_j : \mathcal{X} \to [0, 1]$. We define the *linear concept space* as

$$C_{\text{lin}} = \left\{ c_\alpha(x) = \sum_{j=1}^{m} \alpha_j \, c_j(x) \;\middle|\; \alpha = (\alpha_1, \ldots, \alpha_m) \in \mathcal{A} \right\}, \tag{38}$$

where $\mathcal{A} \subseteq \mathbb{R}^m$ is a compact coefficient set. A natural choice is the **bounded ball** $\mathcal{A} = \{\alpha \in \mathbb{R}^m : \|\alpha\|_2 \leq R\}$, which arises when the linear combination is learned with ℓ_2 regularization, as in ridge-penalized linear probes.

J.2 Assumptions

We require the following regularity condition on the similarity function.

Definition J.1 (Lipschitz similarity)**.** We say the similarity function $\mathsf{sim}(f, \cdot)$ is L-Lipschitz with respect to the coefficient vector if there exists a constant $L > 0$ such that for all $\alpha, \beta \in \mathcal{A}$,

$$|\mathsf{sim}(f, c_\alpha) - \mathsf{sim}(f, c_\beta)| \leq L \, \|\alpha - \beta\|_2. \tag{39}$$

We further assume the same condition holds for the empirical similarity: for all $\alpha, \beta \in \mathcal{A}$ and all probing datasets D_{probe},

$$\left| \hat{\mathsf{sim}}(f, c_\alpha; D_{\text{probe}}) - \hat{\mathsf{sim}}(f, c_\beta; D_{\text{probe}}) \right| \leq L \, \|\alpha - \beta\|_2. \tag{40}$$

Remark J.2. The Lipschitz condition holds naturally for continuous similarity metrics studied in Sec. G. For instance, for correlation: $\mathsf{sim}_{\text{corr}}(f, c_\alpha) = \text{Cov}(f, c_\alpha)/(\sigma_f \, \sigma_{c_\alpha})$ and $\text{Cov}(f, c_\alpha) = \sum_j \alpha_j \text{Cov}(f, c_j)$, so the correlation is differentiable in α and hence locally Lipschitz. On a compact $\mathcal{A}$ with σ_{c_α} bounded away from zero, it is globally Lipschitz. We defer a full derivation of the Lipschitz constant for other metrics to future work and focus here on the structural extension of the bound.

J.3 Covering number bound

The key idea is to discretize $\mathcal{A}$ into a finite ε-cover, apply Thm. 3.2 to that finite set, and control the approximation error via the Lipschitz condition.

Definition J.3 (Covering number (Vershynin, 2018, Chapter 4))**.** For a set $\mathcal{A} \subseteq \mathbb{R}^m$ and $\varepsilon > 0$, the *ε-covering number* $\mathcal{N}(\varepsilon, \mathcal{A}, \|\cdot\|_2)$ is the minimum number of ℓ_2-balls of radius ε needed to cover $\mathcal{A}$.

Theorem J.4 (Fidelity bound for linear concept spaces)**.** *Let C_{lin} be the linear concept space defined in Eq. 38, where $\mathcal{A} \subseteq \mathbb{R}^m$ is compact. Suppose the similarity function satisfies the Lipschitz conditions in Thm. J.1. Then, for any $\varepsilon > 0$, with probability at least $1 - \delta$:*

$$\sup_{\alpha \in \mathcal{A}} \left| \hat{\mathsf{sim}}(f, c_\alpha; D_{\text{probe}}) - \mathsf{sim}(f, c_\alpha) \right| \leq r\left(f, D_{\text{probe}}, \frac{\delta}{\mathcal{N}(\varepsilon, \mathcal{A}, \|\cdot\|_2)} \right) + 2L\varepsilon, \tag{41}$$

where r is the convergence rate from Thm. 3.2 and L is the Lipschitz constant.

Proof. Let $\mathcal{A}_\varepsilon = \{\alpha^{(1)}, \alpha^{(2)}, \ldots, \alpha^{(N_\varepsilon)}\}$ be a minimal ε-cover of $\mathcal{A}$, where $N_\varepsilon = \mathcal{N}(\varepsilon, \mathcal{A}, \|\cdot\|_2)$. That is, for every $\alpha \in \mathcal{A}$, there exists some $\alpha^{(k)} \in \mathcal{A}_\varepsilon$ with $\|\alpha - \alpha^{(k)}\|_2 \leq \varepsilon$.

Step 1: Uniform bound on the finite cover. Since $\mathcal{A}_\varepsilon$ is a finite set of size N_ε, we can apply Thm. 3.2 with $|C| = N_\varepsilon$. With probability at least $1 - \delta$:

$$\sup_{k \in [N_\varepsilon]} \left| \hat{\mathsf{sim}}(f, c_{\alpha^{(k)}}; D_{\text{probe}}) - \mathsf{sim}(f, c_{\alpha^{(k)}}) \right| \leq r\left(f, D_{\text{probe}}, \frac{\delta}{N_\varepsilon}\right). \tag{42}$$

Step 2: Approximation error from the cover. For any $\alpha \in \mathcal{A}$, let $\alpha^{(k)}$ be its nearest neighbor in $\mathcal{A}_\varepsilon$, so $\|\alpha - \alpha^{(k)}\|_2 \leq \varepsilon$. By the triangle inequality:

$$\begin{aligned} &\left| \hat{\mathsf{sim}}(f, c_\alpha; D_{\text{probe}}) - \mathsf{sim}(f, c_\alpha) \right| \\ \leq\ &\left| \hat{\mathsf{sim}}(f, c_\alpha; D_{\text{probe}}) - \hat{\mathsf{sim}}(f, c_{\alpha^{(k)}}; D_{\text{probe}}) \right| \\ &+ \left| \hat{\mathsf{sim}}(f, c_{\alpha^{(k)}}; D_{\text{probe}}) - \mathsf{sim}(f, c_{\alpha^{(k)}}) \right| \\ &+ \left| \mathsf{sim}(f, c_{\alpha^{(k)}}) - \mathsf{sim}(f, c_\alpha) \right|. \end{aligned} \tag{43}$$

By the Lipschitz conditions in Equations (39) and (40), the first and third terms are each bounded by $L\varepsilon$. The middle term is bounded by $r(f, D_{\text{probe}}, \delta/N_\varepsilon)$ from Eq. 42. Combining:

$$\left| \hat{\mathsf{sim}}(f, c_\alpha; D_{\text{probe}}) - \mathsf{sim}(f, c_\alpha) \right| \leq r\left(f, D_{\text{probe}}, \frac{\delta}{N_\varepsilon}\right) + 2L\varepsilon. \tag{44}$$

Step 3: Take supremum. Since $\alpha \in \mathcal{A}$ was arbitrary, taking the supremum over $\mathcal{A}$ yields Eq. 41. □

J.4 Explicit bound for bounded coefficient sets

Standard covering number estimates from geometric analysis give an explicit bound for $\mathcal{A}$.

Corollary J.5 (Bounded ball)**.** *If $\mathcal{A} = \{\alpha \in \mathbb{R}^m : \|\alpha\|_2 \leq R\}$, then $\mathcal{N}(\varepsilon, \mathcal{A}, \|\cdot\|_2) \leq \left(\frac{2R}{\varepsilon} + 1\right)^m$ (Vershynin, 2018, Corollary 4.2.13). Substituting into Eq. 41 and using the typical rate $r(f, D_{\text{probe}}, \delta) = Q\sqrt{\frac{-\log \delta}{|D_{\text{probe}}|}}$:*

$$\sup_{\alpha \in \mathcal{A}} \left| \hat{\mathit{sim}}(f, c_\alpha; D_{\text{probe}}) - \mathit{sim}(f, c_\alpha) \right| \leq Q\sqrt{\frac{m \log\left(\frac{2R}{\varepsilon} + 1\right) + \log \frac{1}{\delta}}{|D_{\text{probe}}|}} + 2L\varepsilon. \tag{45}$$

In particular, choosing $\varepsilon^* = \frac{Q\sqrt{m}}{2L\sqrt{|D_{\text{probe}}|}}$ and substituting into Eq. 45:

- *The approximation error term becomes $2L\varepsilon^* = \frac{Q\sqrt{m}}{\sqrt{|D_{\text{probe}}|}}$.*
- *The covering number evaluates to $\log \mathcal{N}(\varepsilon^*, \mathcal{A}, \|\cdot\|_2) \leq m \log\left(\frac{2R}{\varepsilon^*} + 1\right) = m \log\left(\frac{4LR\sqrt{|D_{\text{probe}}|}}{Q\sqrt{m}} + 1\right)$.*

The estimation error term thus becomes

$$Q\sqrt{\frac{m \log\left(\frac{4LR\sqrt{|D_{\text{probe}}|}}{Q\sqrt{m}} + 1\right) + \log \frac{1}{\delta}}{|D_{\text{probe}}|}}. \tag{46}$$

We simplify the logarithmic factor. For $|D_{\text{probe}}|$ large enough that $\frac{4LR\sqrt{|D_{\text{probe}}|}}{Q\sqrt{m}} \geq 1$ (which holds in any practical regime), we have

$$\begin{aligned} \log\left(\frac{4LR\sqrt{|D_{\text{probe}}|}}{Q\sqrt{m}} + 1\right) &\leq \log\left(\frac{8LR\sqrt{|D_{\text{probe}}|}}{Q\sqrt{m}}\right) \\ &= \log\left(\frac{LR\sqrt{|D_{\text{probe}}|}}{\sqrt{m}}\right) + \log\left(\frac{8}{Q}\right), \end{aligned} \tag{47}$$

where we used $a + 1 \leq 2a$ for $a \geq 1$ in the first inequality, and then split the logarithm. Since $\log(8/Q)$ is a constant independent of $|D_{\text{probe}}|$, the term $m \cdot \log(8/Q)$ can be absorbed into the $\mathcal{O}(\cdot)$ notation. Therefore, the estimation error in Eq. 46 is

$$\mathcal{O}\left(Q\sqrt{\frac{m\log(LR|D_{\text{probe}}|^{1/2}/m^{1/2}) + \log(1/\delta)}{|D_{\text{probe}}|}}\right). \tag{48}$$

Combining with the approximation error term $\frac{Q\sqrt{m}}{\sqrt{|D_{\text{probe}}|}}$, the total bound at ε^* is

$$\mathcal{O}\left(\sqrt{\frac{m\log(LR|D_{\text{probe}}|^{1/2}/m^{1/2}) + \log(1/\delta)}{|D_{\text{probe}}|}}\right) + \frac{Q\sqrt{m}}{\sqrt{|D_{\text{probe}}|}}. \tag{49}$$

The second term is $\mathcal{O}(\sqrt{m/|D_{\text{probe}}|})$, which is dominated by the first term (due to the extra $\sqrt{\log(\cdot)}$ factor inside the square root). Thus, the overall rate is

$$\mathcal{O}\left(\sqrt{\frac{m\log(LR|D_{\text{probe}}|^{1/2}/m^{1/2}) + \log(1/\delta)}{|D_{\text{probe}}|}}\right). \tag{50}$$

Comparison with the finite case. Comparing Eq. 50 with the finite-concept bound in Thm. 3.2, the structural change is that $\log|C|$ (finite vocabulary size) is replaced by $\mathcal{O}(m\log|D_{\text{probe}}|)$, where m is the intrinsic dimensionality of the linear concept space. The covering number $(2R/\varepsilon + 1)^m$ is exponential in m, but it enters the bound only inside a logarithm, so the $\mathcal{O}(1/\sqrt{|D_{\text{probe}}|})$ convergence rate is preserved. For a small number of base concepts (e.g., $m = 5$–20), the bound remains tight and practical.

J.5 Extension of the approximate optimality guarantee

The analogue of Cor. 3.3 also extends to the linear concept space.

Corollary J.6. *Under the conditions of Thm. J.4, let $c_{\hat{\alpha}} = \arg\max_{\alpha\in\mathcal{A}} \hat{\mathsf{sim}}(f, c_\alpha; D_{\text{probe}})$ be the concept selected by empirical similarity maximization and $c_{\alpha^*} = \arg\max_{\alpha\in\mathcal{A}} \mathsf{sim}(f, c_\alpha)$ be the population-optimal concept. Then, with probability at least $1 - \delta$:*

$$\mathsf{sim}(f, c_{\hat{\alpha}}) \geq \mathsf{sim}(f, c_{\alpha^*}) - 2r\left(f, D_{\text{probe}}, \frac{\delta}{\mathcal{N}(\varepsilon, \mathcal{A}, \|\cdot\|_2)}\right) - 4L\varepsilon. \tag{51}$$

Proof. Follows the same argument as Cor. 3.3: $\hat{\mathsf{sim}}(f, c_{\hat{\alpha}}) \geq \hat{\mathsf{sim}}(f, c_{\alpha^*})$ by definition of $\hat{\alpha}$, and each term deviates from its population counterpart by at most $r(f, D_{\text{probe}}, \delta/N_\varepsilon) + 2L\varepsilon$ from Thm. J.4. □

J.6 Discussion

Thm. J.4 directly applies to the linear explanation framework of Oikarinen & Weng (2024). To use the bound in practice, one needs (i) the Lipschitz constant L of the chosen similarity metric and (ii) the geometry of the coefficient set $\mathcal{A}$ (which determines the covering number). For ℓ_2-regularized probes (bounded ball), these quantities are readily available from the covering number estimate in Cor. J.5.