FINITE PROBES SUFFICE: IDENTIFIABILITY AND UNIVERSALITY FOR WEIGHT-SPACE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning properties of neural networks has recently attracted growing interest, with existing approaches operating either directly on network parameters or through probe-based representations of network behavior. While probing methods have shown strong empirical performance, their theoretical foundations remain limited. In this work, we study when finite probe-based representations are sufficient for learning neural functionals. We establish general identification and universality results for probing, and show that using intermediate hidden representations can provide significantly more informative representations than relying only on final outputs. Motivated by these results, we introduce HIDDENPROBE, a simple architecture for learning from hidden probe responses. Across a range of neural functional benchmarks, including both MLPs and Transformers, HIDDENPROBE consistently improves over existing probing methods and achieves state-of-the-art performance.

1 INTRODUCTION

In recent years, there has been increasing interest in the topic of learning neural functionals (also known as weight space learning Han et al. (2026)). A neural functional is a mapping ψ taking a neural network f_θ to a vector $\psi(\theta) \in \mathbb{R}^m$. For example, one could consider a neural functional predicting the test accuracy of a given neural network. Another example is where f_θ is an INR representing an image, and then standard image tasks like image classification become neural functionals.

A popular approach to learning neural functionals $\psi(f_\theta)$ is via *weight space methods*, which think of the functionals as a function of the parameters $\tilde{\psi}(\theta) = \psi(f_\theta)$, and then apply neural architectures directly to the parameter weights to learn $\tilde{\psi}$. Several recent works have adopted this approach with some success Navon et al. (2023); Zhou et al. (2023b); Kofinas et al. (2024); Zhou et al. (2023a). However, weight-space learning must contend with the fact that many distinct parameter vectors can represent the same function. Accounting for these parameter symmetries is often important for performance, but typically requires symmetry-aware architectures that are more complex to design and implement.

Probing provides an alternative way to process neural models that avoids the parameterization ambiguities of weight-space methods. Rather than representing f_θ through its parameters θ, probing represents it through its evaluations on a set of inputs, called probes, $p_1, \ldots, p_N$. The learning problem is then to find a function $\hat{\psi}\big(f_\theta(p_1), \ldots, f_\theta(p_N)\big)$ that approximates the desired neural functional $\psi(f_\theta)$. Probe-based methods, such as Kahana et al. (2025), have achieved strong empirical performance on several benchmarks, but currently lack a theoretical justification. In particular, while weight space methods like Navon et al. (2023) can provably approximate all neural functionals Dayan et al. (2026), analogous universality guarantees are not known for probe-based methods. Moreover, whereas the parameters θ fully specify f_θ, a finite probe set reveals only the values of f_θ at finitely many inputs and therefore appears, a priori, to provide only partial information about the realized function. This raises two closely related questions:

Main Questions:

1. Can a function f_θ be uniquely reconstructed from a finite number of probes? If so, how many probes are needed?

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.

2. Can any neural functional $\Lambda(f_\theta)$ be approximated by a probe-based model?

Main Results: As we shall see, our answer to both questions will be positive. Under the assumption that f_θ is analytic, we show that the function f_θ can be uniquely recovered from $2q+1$ probes, where q denotes the dimension of the parameter space. As a result, we will show that any neural functions $\Lambda(f_\theta)$ can be approximated by a probe-based model with $2q+1$ probes.

In addition, we will discuss probe based methods which represent f_θ not only through the final probe value $f_\theta(p)$, but also through the value of all intermediate layers on the probes. For both MLP and self-attention, we will show that this representation requires fewer probes to obtain the same guarantees. For fully connected neural networks, this brings down the sufficient probe count from $O(q)$ to $O(d_{\max})$, where $d_{\max}$ is the maximal width, and for self-attention blocks this brings down the count to $O(d_{in}h)$, with d_{in} and h being the input dimension and number of heads, respectively. Motivated by this we suggest a specialized model to learn properties of neural networks with such probes, which we name HIDDENPROBE (see Figure 1). Empirically, we will show that HIDDENPROBE attains state of the art results on several standard neural functional datasets.

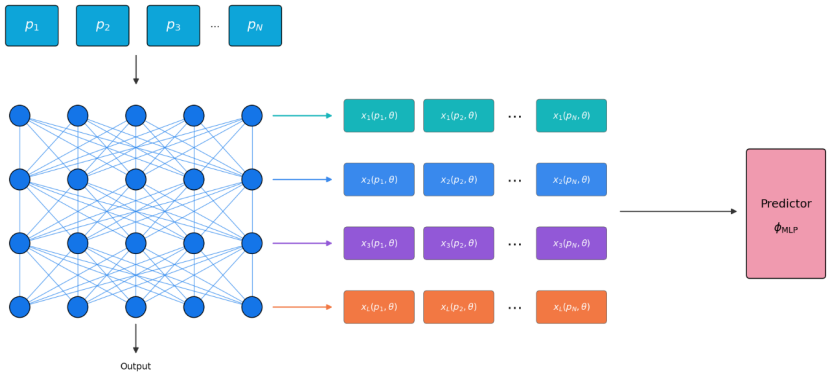


Figure 1: **Overview of HiddenProbe.** A set of learnable probes $p_1, \ldots, p_N$ is evaluated on a target neural network f_θ, producing intermediate activations $x_i(p_j, \theta)$ and final outputs $x_L(p_j, \theta)$, which are processed by an invariant model and an MLP to predict a property of the target network.

1.1 PREVIOUS WORK

Weight-space learning. Weight-space learning studies models that process the parameters of other neural networks directly; we refer to Han et al. (2026) for a recent survey. A central challenge in this setting is that different parameterizations can represent the same function, motivating architectures that explicitly respect weight-space symmetries. DWSNets (Navon et al., 2023) and Neural Functional Networks (Zhou et al., 2023a) construct linear layers equivariant to hidden-neuron permutations, while Neural Functional Transformers (Zhou et al., 2023b) incorporate the same symmetry into attention-based architectures. More recent graph-based approaches, including Graph Metanetworks (Lim et al., 2024) and Neural Graphs (Kofinas et al., 2024), represent networks as computational graphs, enabling weight-space processing across a broader range of architectures. Beyond permutations, subsequent methods incorporate additional symmetries, such as ScaleGMN (Kalogeropoulos et al., 2024) and Monomial-NFN (Tran et al., 2024), which account for scaling and sign symmetries induced by particular activation functions.

Identification and learning from probes Prior work in weight-space learning has explored combining network weights with responses to input probes to predict network properties (Herrmann et al., 2024; Kofinas et al., 2024; Horwitz et al., 2025). Complementary approaches rely solely on final outputs evaluated at learned probes (Kahana et al., 2025). Beyond weight-space learning, interpolation and identification of neural-network from input–output observations has also been studied in learning theory (Vardi et al., 2022). More recently, Bhattamishra et al. (2026); Kim & Kim (2026) have investigated recovering the weights of frozen transformers from query outputs. Beyond final outputs, hidden activations provide additional information about a network's internal representations. Prior work has used these activations to study representation properties (Alain & Bengio, 2017; Cotler et al., 2023) and, more recently, to investigate model behavior through mechanistic interpretability (Gurnee

et al., 2023). In natural language processing, probing classifiers similarly predict linguistic properties from a model's hidden representations to study the information they encode (Belinkov, 2022). Within weight-space learning, Heo et al. (2026) similarly probe selected hidden layers to predict network properties.

To the best of our knowledge, prior work has not systematically studied property prediction from complete activation traces across all hidden layers, collected by evaluating a network on input probes. We address this gap by studying when neural functions can be identified from the activations induced by input queries, and how universal approximators can be designed to predict network properties from final outputs and hidden activation traces.

1.2 Preliminaries

Consider a parametric family $f : \mathcal{X} \times \Theta \to \mathbb{R}^m$, where $\mathcal{X} \subseteq \mathbb{R}^d$ is the input domain and $\Theta \subseteq \mathbb{R}^q$ is the parameter domain. Every fixed parameter $\theta \in \Theta$ defines a function $f_\theta(x) = f(x;\theta)$. We use $f(\cdot;\theta)$ and f_θ interchangeably. We denote the collection of all such functions by $\mathcal{F}_\Theta$.

Our motivating application is the problem of learning neural functionals[1] $\Psi : \mathcal{F}_\Theta \to \mathbb{R}^m$, and our focus is on analyzing methods which rely on probing. We will now define these methods, and focus on two variants: final-layer probing and hidden-layer probing.

Final-layer probing Final-layer probing represents a function f_θ by considering its output on a number of input points, called probes. Formally, given an N tuple of points $\mathcal{P} := (p_1, \cdots, p_N) \in \mathcal{X}^N$, the model f_θ is mapped to an $m \times N$ dimensional matrix

$$\Psi_{\text{final}}(\mathcal{P};\theta) := (f_\theta(p_1), \cdots, f_\theta(p_N)) \in \mathbb{R}^{m\times N}.$$

This matrix is processed to learn properties of the neural network, e.g., by applying an MLP $\phi_{\text{MLP}} \circ \Psi_{\text{final}}(\mathcal{P};\theta)$ to the output, to obtain a neural model for learning the functional $\Lambda(f_\theta)$. A representative example of this approach is the ProbeGen method of Kahana et al. (2025), which uses this approach precisely with a carefully designed method for learning the probes.

Hidden-Layer Probing Heo et al. (2026); Horwitz et al. (2025), In applications, the parametric functions f_θ are typically constructed by a composition of simpler functions (e.g. layers), that is $f_\theta = f_{L,\theta} \circ \ldots \circ f_{2,\theta} \circ f_{1,\theta}$. When this compositional structure is present, hidden layer probing relates to the practice of retaining the output of all intermediate computations on the probes, and not only the final output. Formally, for every input $p \in \mathcal{X}$ and parameter $\theta \in \Theta$ we define

$$x_0(p,\theta) = p,\ x_1(p,\theta) = f_{1,\theta}\left(x_0(p,\theta)\right), \ldots, x_L(p,\theta) = f_{L,\theta}\left(x_{L-1}(p,\theta)\right).$$

The hidden-probe representation of f_θ is defined via the mapping

$$\Psi_{\text{all}}(\theta,\mathcal{P}) := (x_i(p_j,\theta), 0 \le i \le L,\ 1 \le j \le N)$$

1.3 Overview

In the remainder of the paper, we will prove that analytic neural networks of any kind can be identified by a number of final-layer probes proportional to the network's parameters (Section 2), and will show, for both MLPs and transformers, that the required number of probes can be greatly reduced when allowing Hidden-Layer probing (Sections 3-4). We then describe HiddenProbe, a method for learning using hidden probes (Section 5), prove the universality of probe-based methods (Section 6), and finally demonstrate the effectiveness of HiddenProbe on neural functional learning (Section 7).

2 Identifying Analytic Neural Networks from Probe evaluations

Our first result is that final-layer probe evaluations are enough to identify *analytic* parametric functions, and the number of probes is proportional to the dimension of the parameter space. This

[1] Our focus is on neural functionals which depend only on the function f_θ and not on the specific parameterization θ. In the terminology of Dayan et al. (2026) these are called function-space functionals

theorem applies to all analytic parametric families including MLPs and Transformers, provided that the activations used in these models are analytic. Examples of such activations are tanh, sigmoid, SiLU and sine activations.

Theorem 1 (Unique reconstruction from finite samples)**.** *Let $f : \mathbb{R}^d \times \mathbb{R}^q \to \mathbb{R}^m$ be an analytic function. Then, for almost any choice of probes $p_1, \dots, p_{2q+1} \in \mathbb{R}^d$, we have that for all $\theta_1, \theta_2 \in \Theta$,*

$$f(p_j, \theta_1) = f(p_j, \theta_2), \forall j = 1, \dots, 2q+1 \iff f(\cdot, \theta_1) \equiv f(\cdot, \theta_2) \text{ on } \mathbb{R}^d$$

Proof Idea. Intuitively we can interpret this result as follows: although the family of functions $\{f_\theta | \quad \theta \in \mathbb{R}^q\}$ contains infinitely many functions, it has a finite dimensional parameterization in $\mathbb{R}^q$, and so it is reasonable that a finite number of $\sim q$ evaluations can separate the realized functions. This is similar to the well-known result that a degree n polynomial can be identified by $n+1$ samples. However, this rule of thumb is not always true. For example, for ReLU networks, no finite number of probes suffices to uniquely identify the function (see Subsection 3.2). The proof that such a result does hold for analytic functions is based on the Finite Witness Theorem by Amir et al. (2023). □

In Section 6 we will show that this result implies universal approximation: probe based methods with $2q+1$ probes can approximate all neural functionals. This result is substantial, since probe based methods typically use a small number of probes, while apriori it would seem that universal approximation would require an exhaustive sampling of the input domain. At the same time, the number of probes suggested by the theorem is still much larger than what is used in practice. To alleviate this disadvantage, we will next focus on hidden probes, where function identification is possible with a much lower probe cardinality.

3 PROBING MLPS

In this section we specialize to the case where $f(x; \theta)$ is an MLP. The main message of this section is that hidden probes substantially reduce the number of probes necessary to identify an MLP.

We begin our discussion by setting our notation for the MLP model class. An MLP $f(p; \theta)$ is determined by a choice of activation σ and a width vector $\mathbf{D} := (d_0, d_1, \cdots, d_{L-1}, d_L)$. An MLP $f(p; \theta)$ is then defined via

$$x_0 := p,\ x_i := \sigma\left(W_i x_{i-1} + b_i\right),\ \forall i \in [L], \tag{1}$$

Here, $W_i \in \mathbb{R}^{d_i \times d_{i-1}}$ and $b_i \in \mathbb{R}^{d_i}$, for $i \in [L]$, are the weights and biases, which together define the parameters $\theta := [W_i, b_i]_{i \in [L]}$. The intermediate features x_i depend on p, θ and so we denote them by $x_i(p, \theta)$. The final function is defined by the last feature $f(p, \theta) = x_L(p, \theta)$. We denote the total dimension of the parameter space by $q_{\mathbf{D}}$.

3.1 HIDDEN PROBES PROVIDE USEFUL INFORMATION:

For MLPs with analytic activations, Theorem 1 implies that a finite number of probes, proportional to $q_{\mathbf{D}}$, is sufficient. For hidden probes, the number of necessary probes depends only on the maximal width:

Theorem 2 (Reconstruction from hidden layer activations)**.** *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be an analytic, injective, non-polynomial activation. Let $f(p, \theta)$ denote an MLP with activation σ and widths $\mathbf{D} := (d_0, d_1, \cdots, d_{L-1}, d_L)$. Assume that $N > \max_i d_i$. Then for almost every $\mathcal{P} \in \mathcal{X}^N$, we have that for almost every θ_1, θ_2,*

$$\Psi_{all}(\theta_1, \mathcal{P}) = \Psi_{all}(\theta_2, \mathcal{P}) \implies \theta_1 = \theta_2, \implies f(\cdot, \theta_1) = f(\cdot, \theta_2).$$

In this result we add the assumption that the activation is non-polynomial and injective (this is the case for e.g., tanh and sigmoid, but not for sine activation). Moreover, identifiability is not on the whole domain, but rather it holds for parameters θ_1, θ_2 outside an exceptional Lebesgue-null set $\mathcal{E}$. The proof in fact shows a stronger statement: the set $\mathcal{E}$ is a zero set of an analytic function, and hence is of strictly lower dimension than the parameter space, and has empty interior. In general, whenever we talk about "almost everywhere" in this paper, this stronger notion applies.

Proof Idea. The idea of the proof is based on a straightforward algorithm described in detail in in Algorithm 1 of Appendix B.1. Intuitively, the parameters W_i, b_i of the i-th layer satisfy the linear equations

$$\sigma^{-1}\left(x_i(p_j, \theta)\right) = W_i x_{i-1}(p_j, \theta) + b_i, \quad j = 1, \ldots, N$$

The proof shows that for generic probes and parameters, these linear equations have a unique solution. □

Tightness of results in Appendix B.2 we give examples of analytic MLPs that cannot be identified with $d_{\max}$ hidden probes (exactly matching the bound of Theorem 2) For final-layer probes, we construct, for every $N \leq 2q_D - 2$, a nonempty open set of probe tuples that fail to separate all realized functions. (Theorem 1 gives a sufficient bound of $2q_{\mathbf{D}} + 1$ generic probes).

3.2 Identifying ReLU Networks from Probe Evaluations

Our results on identifiability of an MLP from a finite number of probes require analytic activations, and thus do not apply to the common ReLU activation. In fact, such a result is not possible for ReLU activations:

Proposition 3 (Finite probes do not identify ReLU networks)**.** *For any finite set of probes $p_1, \ldots, p_N \in [0, 1]$, there exist two distinct ReLU networks, each with three hidden neurons, that agree on all the probes. Consequently, no finite set of probes can uniquely identify all ReLU networks.*

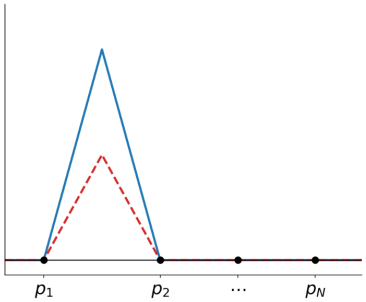


Figure 2: For any finite set of probes, a nonzero ReLU 'hat' function can be constructed whose support avoids all probes.

Proof Idea. For given probes $p_1, \ldots, p_N$, it is possible to construct two different hat functions which are zero on the probes: for example, the functions plotted in blue and orange in Figure 2. Full proof is in Appendix B.3. □

Since exact identification of ReLU networks from finitely many probes is impossible, we instead study approximate identification in $L^2(\mu)$. We show that, with probability at least $1 - \delta$ over the sampled probes, any network that agrees with the target network on those probes is within ϵ of the target in $L^2(\mu)$. This is similar in spirit to results from PAC learning. The following theorem gives a semi-formal description of these results, which are described and proved in more detail in Theorem 12 of Appendix B.3.

Theorem 4 (semi-formal-PAC identification of ReLU networks)**.** *Consider the family of* ReLU *MLPs with widths $\mathbf{D} := (d_0, d_1, \cdots, d_{L-1}, d_L)$ and q parameters. Let μ be a compactly supported probability distribution on $\mathbb{R}^{d_0}$, and assume that the parameter set is compact. Fix $\epsilon, \delta \in (0, 1)$ and a parameter vector θ_0. Then there exist constants $C_1(\epsilon, \delta)$, $C_2(\epsilon, \delta) > 0$ such that*

- *For $N_1 = C_1 \cdot Lq \ln(q)$ final layer samples, with probability of at least $1 - \delta$ on the samples, any θ which agrees with θ_0 on the samples will satisfy $\|f(\cdot; \theta) - f(\cdot; \theta_0)\|_{L^2(\mu)} \leq \epsilon$.*
- *When hidden layer samples are used, only $N_2 = C_2 \cdot q \ln(q)$ samples are sufficient.*

As in the analytic activation case, this result suggests that hidden activations can reduce the number of probes necessary to (almost) identify a ReLU network, especially for deeper ReLU nets. An experimental verification of this is given in the Appendix D.1.

4 Identifying Multi-Head Attention from Probes

Multi-head attention (MHA) forms the backbone of modern transformer architectures. In this section we discuss the identification of multiple-layer multihead attention (MHA) blocks from probes. Additional components typically used in transformers, like FFNs and LayerNorm, are omitted for mathematical tractability.

We begin with some basic definitions (details in Appendix B.4). We consider d dimensional inputs and h attention heads. An MHA layer has parameters $\theta := \{(\mathbf{Q}_a, \mathbf{K}_a, \mathbf{V}_a, \mathbf{O}_a)\}_{a=1}^{h}$, For an input probe $P \in \mathbb{R}^{n\times d}$ whose rows are the token vectors $p_1^\top, \ldots, p_n^\top$, the layer computes

$$F_n(P;\theta) = \sum_{a=1}^{h} \mathrm{softmax}(P\mathbf{M}_a P^\top)\, P\mathbf{W}_a,\ \mathbf{M}_a := \frac{\mathbf{Q}_a\mathbf{K}_a^\top}{\sqrt{k}},\ \mathbf{W}_a := \mathbf{V}_a\mathbf{O}_a. \tag{2}$$

The softmax acts rowwise: $\left[\mathrm{softmax}(P\mathbf{M}_a P^\top)\right]_{ij} = \frac{\exp(p_i^\top \mathbf{M}_a p_j)}{\sum_{s=1}^{n}\exp(p_i^\top \mathbf{M}_a p_s)}$. A probe specifies P and returns the complete matrix $F_n(P;\theta) \in \mathbb{R}^{n\times d}$. We denote by $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_L)$ the parameters of an L-layer MHA stack and let $\Theta = \prod_{\ell=1}^{L}\Theta_\ell$ denote the parameter space of the L-layer MHA architecture, with Θ_ℓ being the parameter space of the ℓ-th layer, let $q := \dim(\Theta)$.Writing $F_n^{(\ell)}$ for the ℓ-th layer map, we define

$$x_{0,n}(P;\boldsymbol{\theta}) = P,\ x_{\ell,n}(P;\boldsymbol{\theta}) = F_n^{(\ell)}\left(x_{\ell-1,n}(P;\boldsymbol{\theta});\theta_\ell\right),\ \ell = 1,\ldots,L. \tag{3}$$

Thus $x_{L,n}$ is the stack's final output. For a collection of probes $\mathcal{P} = (P^{(i)})_{i=1}^{N}$, where $P^{(i)} \in \mathbb{R}^{n_i\times d}$. As usual, we define the output-only and hidden-access probe maps for transformers by $\Psi_{\text{final}}(\boldsymbol{\theta}, \mathcal{P})$ and $\Psi_{\text{all}}(\boldsymbol{\theta}, \mathcal{P})$, respectively.

Since the softmax attention is analytic, we can apply a variant of Theorem 1, by using the Finite Witness Theorem for the effective manifold dimension of Θ to give function-level identification from probes:

Corollary 1 (Generic probes)**.** *Fix a token length $n \geq 1$. For almost every tuple of $1 + 2q$ probes in $\mathbb{R}^{n\times d}$ and all $\theta, \theta' \in \Theta$, $\Psi_{final}(\theta) = \Psi_{final}(\theta') \iff x_{L,n}(\cdot;\theta) \equiv x_{L,n}(\cdot;\theta')$.*

Note that all of these parameter counts are $O(d^2)$ per layer for the standard multi-head attention blocks. But softmax attention has more structure than analyticity, and we can query probes of varying token lengths to an MHA block, which lets us identify the function using only $O(dh)$ probes:

Theorem 5 (Structured probing for softmax attention)**.** *Let $h_{\max} := \max_\ell h_\ell$ and $N := d(2h_{\max} + 1) + 1$. There exists an open set of probe tuples $\mathcal{P} \in \mathcal{X}^N$ of token length at most $\max\{3, 1 + d/h_{\max}\}$, such that for almost every target $\theta_0 \in \Theta$ we have:*

$$\Psi_{all}(\theta;\mathcal{P}) = \Psi_{all}(\theta_0;\mathcal{P}) \implies x_{\ell,n}(\cdot;\theta) \equiv x_{\ell,n}(\cdot;\theta_0),\ \forall\,\theta \in \Theta, \forall\,\ell \in [L], \forall\, n \in \mathbb{N},$$

Proof Idea. Theorem 5 gives a specific way of choosing the probes for identifying Multilayer MHA blocks with hidden access, which we call RECOVERMHA. We elaborate on the single layer single head case of RECOVERMHA (see also Algorithm 2 in the appendix). In this case, the goal is to reconstruct two matrices: the matrices $\mathbf{M}_0$ and $\mathbf{W}_0$ from (2). The key trick is to do this in first stages: we first use single-token input: For such input the output of the softmax is always the scalar $1 \in \mathbb{R}^{1\times 1}$, and so $\mathbf{M}_0$ is eliminated, and one-token probes provide linear equations that can be used to determine $\mathbf{W}_0$. Once $\mathbf{W}$ is known, we query a single probe consisting of affinely independent tokens, which uniquely recovers the attention weights for a full rank target value matrix. We then recover $\mathbf{M}_0$ from log-ratios by utilizing the softmax structure. Algorithm 2 implements these steps using $d + 1$ probes. The extension of these results to multiple heads and layers is discussed in Appendix B.4. □

5 HIDDENPROBE: LEARNING USING HIDDEN NEURAL ACTIVATIONS

Our analysis shows that using hidden probes can greatly enhance the identifiability of neural networks from probes. Accordingly, we design a method for learning from hidden probes which we name HIDDENPROBE. We have two different instantiations of this model, one for MLPs and one for transformers. The exact details are discussed in Appendix A. However, the basic idea is simple and identical. We use learned probes, and utilize all hidden layers. The probes of each hidden layer are processed a permutation-invariant aggregator, and then the results are aggregated using a global MLP to predict the target property.

We design HIDDENPROBE for MLPs so that it respects the permutation invariance of the MLP weight space. Namely, one can apply an independent permutation to the neurons of each hidden layer,

and obtain a functionally equivalent MLP with permuted weight matrices. We attain permutation invariance by using permutation invariant Set Transformer Lee et al. (2019).

In general, an interesting advantage of (hidden) probe methods over weight space methods is that weight space symmetries become much simpler. Indeed, while in the hidden layer representation the permutations to each hidden layer are independent, in the weight space representation these symmetries are entangled, and designing permutation invariant models for this representation can be cumbersome Navon et al. (2023); Sverdlov et al. (2025). For transformers, the weight space approach needs to deal with $GL(n)$ symmetries Tran et al. (2025) which probe based methods avoid since they depend on the functional behavior of MHA and not on the exact parameterization.

6 UNIVERSALITY OF PROBE-BASED LEARNING

In this section, we use our identification results from previous sections to prove universality of probe based models with a fixed number of probes. These results are based on the standard observation that separation implies universality Chen et al. (2019); Dym & Gortler (2025). For our theorem, we will consider analytic functions $f : \mathbb{R}^d \times \mathbb{R}^q \to \mathbb{R}^m$ as in Theorem 1, but we will assume that we are only interested in a compact input domain $\mathcal{X} \subseteq \mathbb{R}^d$ with non-empty interior, and a compact parameter domain $\Theta \subseteq \mathbb{R}^q$. Accordingly, we are interested in the function space $\mathcal{F}_\Theta(\mathcal{X}, \mathbb{R}^m) := \{f_\theta : \mathcal{X} \to \mathbb{R}^m\}$ which is a subset of the set of all continuous functions from $\mathcal{X}$ to $\mathbb{R}^m$, endowed with the infinity norm $\|f_\theta\|_\infty = \max_{x\in\mathcal{X}} \|f(x,\theta)\|_2$. A functional $\Lambda : \mathcal{F}_\Theta(\mathcal{X}, \mathbb{R}^m) \to \mathbb{R}$ is continuous, if it is continuous with respect to the distance induced by this norm.

Our first result uses Theorem 1 to show that for general analytic models with q parameters, $2q + 1$ final-layer probes are sufficient for universality:

Theorem 6 (Universality of probe-based learning)**.** *With the assumptions listed above, let $\Lambda : \mathcal{F}_\Theta(\mathcal{X}, \mathbb{R}^m) \to \mathbb{R}^k$ be a continuous functional. Then there exists $p_1, \ldots, p_{2q+1} \in \mathcal{X}$, such that $\forall\, \epsilon > 0$, there exists an MLP ϕ_{MLP} such that $\sup_{\theta\in\Theta} ||\Lambda(f_\theta) - \phi_{\mathrm{MLP}}\left(f_\theta(p_1), \ldots, f_\theta(p_{2q+1})\right)|| < \epsilon$*

When using hidden probes, we can obtain universality using our model HIDDENPROBE with a significantly smaller number of probes, due to our identifiability results:

Theorem 7 (Universality of HIDDENPROBE)**.** *With our previous notation, assume that*

1. *$f(x;\theta)$ is an MLP architecture with analytic, injective, non-polynomial activation, and maximal width d_{max}, and denote $N = d_{max} + 1$. Or*
2. *$f(x;\theta)$ is a pure softmax MHA stack with d dimensional inputs and at most H heads per layer. Set $N = d(2H + 1) + 1$.*

Then, if $K \subseteq \Theta$ is a compact subspace of the parameters that does not intersect an appropriate Lebesgue-null set E, there exist N probes, and a hidden probe model $\mathcal{H}_{\text{HIDDENPROBE}}$, such that $\sup_{\theta\in K} \|\Lambda(f_\theta) - H_{\text{HIDDENPROBE}}(\theta)\|_2 < \epsilon$.

We restate formal versions of the above theorems and prove them in Appendix B.5, together with guarantees on operator learning from probes.

7 EXPERIMENTS

We evaluate HIDDENPROBE for weight space lerning tasks and compare with other prominent baselines. To understand the influence of multi-layer probes, we put a special emphasis on comparison with ProbeGen, which is a successful method relying only on final layer probes. We will show that HIDDENPROBE typically outperforms ProbeGen and other weight space and probe-based models on weight space learning baselines.

INR classification In this setting, each input network is an MLP trained as an implicit neural representation (INR) of a single image, and the goal is to predict the class of the represented image directly from the network. We evaluate on the MNIST, Fashion-MNIST, CIFAR-10 (singke and augmented views). The INR networks across datasets contain two hidden layers.

In Table 1, we compare HIDDENPROBE against the baselines: ScaleGMN (Kalogeropoulos et al., 2024), NG-T and Neural Graphs (Kofinas et al., 2024), NFT (Zhou et al., 2023b), MAGEP-NFN (Vo et al., 2025), and ProbeGen (Kahana et al., 2025). A complete comparison with all available methods is provided in Table 4 in the Appendix.

Method	MNIST	FMNIST	CIFAR-10	CIFAR-10 Aug
ScaleGMN	0.966 ± 0.002	0.808 ± 0.001	0.388 ± 0.001	0.570 ± 0.005
Neural Graphs (128 probes)	0.976 ± 0.001	0.745 ± 0.008	–	–
NG-T	0.924 ± 0.003	0.727 ± 0.006	–	–
MAGEP-NFN	–	–	0.3718 ± 0.003	–
NFT	–	–	–	0.634 ± 0.000
ProbeGen	0.984 ± 0.001	0.877 ± 0.003	0.573 ± 0.007	0.563 ± 0.003
HiddenProbe	0.986 ± 0.001	0.886 ± 0.005	0.580 ± 0.002	0.606 ± 0.002

Table 1: Classification accuracy (↑) on INR weight-space benchmarks. Uncertainties denote the standard error over five seeds. Baseline numbers are copied from the respective papers

As observed, **(1)** HIDDENPROBE achieves the best performance on MNIST, Fashion-MNIST, and CIFAR-10, and the second-best performance on augmented CIFAR-10. **(2)** It outperforms ProbeGen everywhere, highlighting the importance of having hidden layer information. **(3)** Probegen comes second in $3/4$ datasets, proving the efficacy of probe-based learning in general

Test accuracy prediction We next evaluate HIDDENPROBE on the model accuracy prediction task. In this setting, the input is a CNN trained for image classification, and the goal is to predict its test accuracy directly from the network. We evaluate on the MNIST, Fashion-MNIST, SVHN, CIFAR-10-GS, and CIFAR-10-WP benchmarks, using Kendall's correltion(τ) as the evaluation metric. In Table 2, we compare against a representative set of strong baselines: NFN (Zhou et al., 2023a), ScaleGMN (Kalogeropoulos et al., 2024), NG-T and Neural Graphs (Kofinas et al., 2024), DNG-Encoder (Wu et al., 2026), and ProbeGen (Kahana et al., 2025). A complete comparison with all available methods is provided in Table 5 in the Appendix.

The networks in the MNIST, Fashion-MNIST, SVHN, and CIFAR-10-GS benchmarks contain three hidden layers, whereas CIFAR-10-WP contains heterogeneous architectures with up to four hidden layers. HIDDENPROBE achieves the best performance on all five benchmarks and outperforms ProbeGen in every case.

Method	MNIST	FMNIST	SVHN	CIFAR-WP	CIFAR-GS
NFN	0.942 ± 0.001	0.935 ± 0.000	0.931 ± 0.005	–	0.934 ± 0.001
ScaleGMN	–	–	–	–	0.941 ± 0.000
NG-T	–	–	0.872 ± 0.001	0.817 ± 0.007	0.935 ± 0.000
DNG-Encoder	–	–	0.867 ± 0.002	0.874 ± 0.002	0.936 ± 0.000
Neural Graphs	–	–	–	0.885 ± 0.005	0.938 ± 0.000
ProbeGen	0.953 ± 0.002	0.948 ± 0.005	0.876 ± 0.005	0.932 ± 0.006	0.957 ± 0.001
HiddenProbe	0.966 ± 0.004	0.957 ± 0.002	0.957 ± 0.003	0.940 ± 0.006	0.965 ± 0.000

Table 2: Test-accuracy prediction performance measured by Kendall's τ (↑). Uncertainties denote standard error over five seeds. Baseline numbers are copid from the respective papers.

As observed: **(1)** HIDDENPROBE is the best across all datasets, and probegen comes second in $4/5$ datasets. **(2)** In Cifar-WP where the model architecture varies, probe-based models significantly outperform weight space models since they are more robust to architectural changes.

Transformer Accuracy Prediction. We evaluate HIDDENPROBE on test-accuracy prediction for Transformers trained on MNIST and AGNews. Performance for the fuill dataset is shown in Table 3.Following the evaluation protocol of Tran et al. (2026), we additionally report results across different accuracy thresholds in Appendix D.4 in tables 6,7. As observed, **(1)**HIDDENPROBE achieves state-of-the-art performance on both benchmarks, beating Transformer-NFN models with lesser parameter count **(2)** Even output-only ProbeGen is competitive with the benchmarks.

Varying performance with the number of probes In the first part of the experiments, we showed that using the intermediate layers adds significant empirical improvement. Now we want to see how the performance of the probe-based models, namely ProbeGen and HIDDENPROBE, changes as we vary the number of probes. We compare HIDDENPROBE with ProbeGen using $N \in \{16, 32, 64, 128, 256\}$ probes. We take seed 0 results for all plots. Results are in Figure 3.

Method	MNIST-Transformers	AGNews-Transformers
MLP	0.866 ± 0.002	0.879 ± 0.006
STATNN (Unterthiner et al., 2020)	0.881 ± 0.001	0.841 ± 0.002
XGBoost (Chen & Guestrin, 2016)	0.860 ± 0.002	0.859 ± 0.001
LightGBM (Ke et al., 2017)	0.858 ± 0.002	0.835 ± 0.001
Random Forest (Breiman, 2001)	0.772 ± 0.002	0.774 ± 0.003
Transformer-NFN (Tran et al., 2025)	0.905 ± 0.002	0.910 ± 0.001
Transformer-NFN large (Tran et al., 2026)	0.907 ± 0.001	0.913 ± 0.001
Transformer-NFN Quasi (Tran et al., 2026)	0.911 ± 0.001	0.914 ± 0.001
ProbeGen	0.898 ± 0.001	0.885 ± 0.001
HiddenProbe	0.920 ± 0.003	0.916 ± 0.002

Table 3: Test accuracy prediction on the MNIST-Transformers and AGNews-Transformers benchmarks, evaluated on the full datasets. Performance is measured by Kendall's τ, reported as mean $\pm$ standard error over five seeds. The best and second-best results are highlighted.

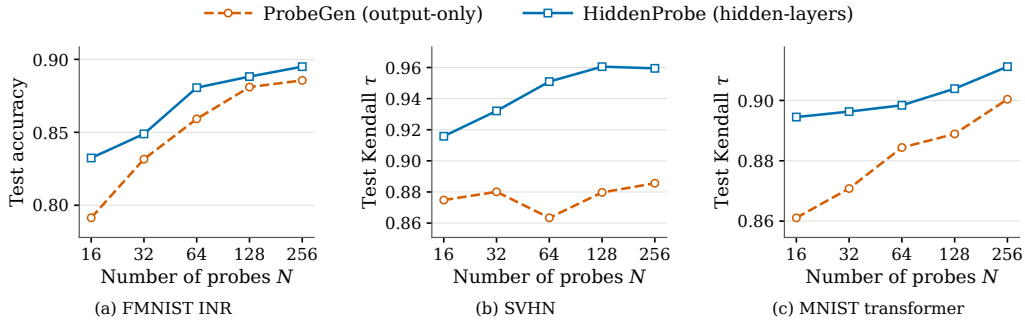


Figure 3: Performance of HiddenProbe and ProbeGen as the number of probes N varies. FMNIST INR classification is evaluated by test accuracy, while regression tasks are evaluated by test Kendall's τ. Higher is better in all panels.

We make the following observations: **(1)** HIDDENPROBE generally improves as the probe count increases. Such a trend is also observed for ProbeGen but to a weaker extent **(2)** HIDDENPROBE is comparable to, or better than, ProbeGen at a smaller ($0.5\times$ or less) probe count. This supports the efficiency of including hidden layer information in probe-based learning.

Additional Information In all our experiments, HIDDENPROBE uses the same number of parameters as ProbeGen (see D.5 for a complete parameter count comparison). A runtime comparison of HIDDENPROBE, ProbeGen, and other DWS models is provided in Section D.6. Synthetic experiments evaluating model reconstruction from finite probes are presented in Appendix D.1.

8 CONCLUSION

In this work, we have done a rigorous theoretical analysis of function-level identification of neural networks from probes. We have shown that one can use probe-based methods to identify a neural network from its outputs and also design universal approximators to learn properties of the neural net from probes. Across architectures like MLPs and Multi head attention blocks, we have shown that this identification can be done with much less number of probes if hidden layer information is incorporated, and we have accordingly proposed HIDDENPROBE, which learns properties of neural nets from full hidden traces. Our experiments demonstrate consistent improvements of HIDDENPROBE over ProbeGen and strong weight-space baselines across diverse tasks, achieving state-of-the-art performance in several. We also demonstrated that incorporating hidden-layer information enables our models to match or surpass ProbeGen using substantially fewer probes, often requiring only half as many. Together, these results establish hidden-layer probing as an effective approach to learning neural functionals across different architectures and tasks. An interesting direction of future work could be designing such methods for larger architectures like language models or diffusion kernels.

AI USE STATEMENT

In this work, we used ChatGPT to assist in refining the exposition of mathematical proofs and to check mathematical arguments. We also used ChatGPT for language editing and improving the readability of the manuscript. All AI-assisted mathematical arguments, claims, and edits were manually reviewed and verified by the authors. We have also used ChatGPT and Claude for AI-assisted coding but all the data and codes were manually checked and verified. Generative AI tools were not used to generate experimental results, numerical results, or data reported in the paper. We take responsibility for the final content of this work, including any text, claims, and mathematical exposition and code produced with the aid of generative AI.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations (ICLR), Workshop Track*, 2017. URL https://openreview.net/forum?id=HJ4-rAVtl.

Tal Amir, Steven J. Gortler, Ilai Avni, Ravina Ravina, and Nadav Dym. Neural injective functions for multisets, measures and graphs via a finite witness theorem. In *Advances in Neural Information Processing Systems*, volume 36, pp. 42516–42551, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/84b686f7cc7b7751e9aaac0da74f755a-Abstract-Conference.html.

Bruno Andreis, Soro Bedionita, Philip H. S. Torr, and Sung Ju Hwang. Set-based neural network encoding without weight tying. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Peter L. Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019. URL https://jmlr.org/papers/v20/17-612.html.

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

Satwik Bhattamishra, Kulin Shah, Michael Hahn, and Varun Kanade. Provably learning attention with queries. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2601.16873.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the vapnik-chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, 1989. doi: 10.1145/76359.76371.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016. doi: 10.1145/2939672.2939785.

Zhengdao Chen, Soledad Villar, Lei Chen, and Joan Bruna. On the equivalence between graph isomorphism testing and function approximation with gnns. *Advances in neural information processing systems*, 32, 2019.

Jordan Cotler, Kai Sheng Tai, Felipe Hernández, Blake Elias, and David Sussillo. Analyzing populations of neural networks via dynamical model embedding, 2023. URL https://arxiv.org/abs/2302.14078.

Adir Dayan, Yam Eitan, and Haggai Maron. On the expressive power of permutation-equivariant weight-space networks. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=gQXkbQSZWS.

Nadav Dym and Steven J. Gortler. Low-dimensional invariant embeddings for universal geometric learning. *Foundations of Computational Mathematics*, 25(2):375–415, 2025. doi: 10.1007/s10208-024-09641-2.

Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. Finding neurons in a haystack: Case studies with sparse probing. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=JYs1R9IMJr.

Xiaolong Han, Zehong Wang, Bo Zhao, Binchi Zhang, Jundong Li, Damian Borth, Rose Yu, Haggai Maron, Yanfang Ye, Lu Yin, and Ferrante Neri. A survey of weight space learning: Understanding, representation, and generation, 2026. URL https://arxiv.org/abs/2603.10090.

Boris Hanin and David Rolnick. Complexity of linear regions in deep networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2596–2604. PMLR, 2019. URL https://proceedings.mlr.press/v97/hanin19a.html.

David Haussler and Emo Welzl. $\epsilon-nets and simplex range queries. Discrete \ \& \ Computational \ Geometry, 2: 127 - -151, 1987. doi: .$

Eunwoo Heo, Kyeongkook Seo, and Jaejun Yoo. What linear probes miss: Multi-view probing for weight-space learning, 2026. URL https://arxiv.org/abs/2605.23410.

Vincent Herrmann, Francesco Faccio, and Jürgen Schmidhuber. Learning useful representations of recurrent neural network weight matrices. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 18205–18227. PMLR, 2024. URL https://proceedings.mlr.press/v235/herrmann24a.html.

Eliahu Horwitz, Bar Cavia, Jonathan Kahana, and Yedid Hoshen. Learning on model weights using tree experts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20468–20478, June 2025.

Jonathan Kahana, Eliahu Horwitz, Imri Shuval, and Yedid Hoshen. Deep linear probe generators for weight space learning. In *The Thirteenth International Conference on Learning Representations*, pp. 96360–96377, 2025. URL https://openreview.net/forum?id=XoYdD3m0mv.

Ioannis Kalogeropoulos, Giorgos Bouritsas, and Yannis Panagakis. Scale equivariant graph metanetworks. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30, pp. 3146–3154, 2017.

Sunyeop Kim and Insung Kim. Provably learning multi-head attention with queries, 2026. URL https://arxiv.org/abs/2608.03294.

Miltiadis Kofinas, Boris Knyazev, Yan Zhang, Yunlu Chen, Gertjan J. Burghouts, Efstratios Gavves, Cees G. M. Snoek, and David W. Zhang. Graph neural networks for learning equivariant representations of neural networks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/c592fc7e6207f82560ed45fece8d6937-Abstract-Conference.html.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning*, pp. 3744–3753. PMLR, 2019.

Moshe Leshno, V. Ya. Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6:861–867, 1993. URL https://api.semanticscholar.org/CorpusID:206089312.

Derek Lim, Haggai Maron, Marc T. Law, Jonathan Lorraine, and James Lucas. Graph metanetworks for processing diverse neural architectures. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ijK5hyxs0n.

Aviv Navon, Aviv Shamsian, Idan Achituve, Ethan Fetaya, Gal Chechik, and Haggai Maron. Equivariant architectures for learning in deep weight spaces. In *International Conference on Machine Learning*, pp. 25790–25816. PMLR, 2023.

Lajos Rónyai, László Babai, and Murali K. Ganapathy. On the number of zero-patterns of a sequence of polynomials. *Journal of the American Mathematical Society*, 14(3):717–735, 2001. 10.1090/S0894-0347-01-00367-8.

Yonatan Sverdlov, Ido Springer, and Nadav Dym. Revisiting multi-permutation equivariance through the lens of irreducible representations. In *The Thirteenth International Conference on Learning Representations*, pp. 8505–8527, 2025. URL https://openreview.net/forum?id=4v4nmYWzBa.

Viet-Hoang Tran, Thieu N. Vo, Tho Tran-Huu, An T. Nguyen, and Tan M. Nguyen. Monomial matrix group equivariant neural functional networks. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Viet-Hoang Tran, Thieu N. Vo, An Nguyen The, Tho Tran Huu, Minh-Khoi Nguyen-Nhat, Thanh Tran, Duy-Tung Pham, and Tan Minh Nguyen. Equivariant neural functional networks for transformers. In *International Conference on Learning Representations*, 2025.

Viet-Hoang Tran, An Nguyen, Benoît Guérand, Thieu N. Vo, and Tan M. Nguyen. Quasi-equivariant metanetworks. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=XMiDpi2mWY.

Thomas Unterthiner, Daniel Keysers, Sylvain Gelly, Olivier Bousquet, and Ilya Tolstikhin. Predicting neural network accuracy from weights. *arXiv preprint arXiv:2002.11448*, 2020.

Gal Vardi, Gilad Yehudai, and Ohad Shamir. On the optimal memorization power of relu neural networks. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=MkTPtnjeYTV.

Thieu Vo, Hoang V. Tran, Tho Tran Huu, An Nguyen The, Thanh Tran, Minh-Khoi Nguyen-Nhat, Duy-Tung Pham, and Tan Minh Nguyen. Equivariant polynomial functional networks. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 61689–61744. PMLR, 2025.

Alex J. Wilkie. O-minimal structures. In *Séminaire Bourbaki, Volume 2007/2008*, number 326 in Astérisque, pp. 131–142. Société mathématique de France, 2009. 10.24033/ast.842. URL https://www.numdam.org/item/AST_2009__326__131_0/.

Di Wu, Huan Liu, Zhixiang Chi, Yuanhao Yu, Konstantinos N. Plataniotis, and Yang Wang. Dynamic neural graph encoding of inference processes in deep weight space. *Transactions on Machine Learning Research*, 2026.

Allan Zhou, Kaien Yang, Kaylee Burns, Adriano Cardace, Yiding Jiang, Samuel Sokota, J. Zico Kolter, and Chelsea Finn. Permutation equivariant neural functionals. In *Advances in Neural Information Processing Systems*, volume 36, 2023a.

Allan Zhou, Kaien Yang, Yiding Jiang, Kaylee Burns, Winnie Xu, Samuel Sokota, J. Zico Kolter, and Chelsea Finn. Neural functional transformers. In *Advances in Neural Information Processing Systems*, volume 36, 2023b.

A HIDDENPROBE: PROBE BASED LEARNING USING HIDDEN NEURAL ACTIVATIONS

In this appendix we describe in full detail HIDDENPROBE, our probe-based architecture for learning from hidden-layer activations.

A.1 HIDDENPROBE: A PROBE-BASED MODEL FOR LEARNING ON WEIGHT SPACES

We begin by discussing the permutation symmetries of hidden activations induced by weight-space symmetries, which motivate the design of HIDDENPROBE.

Symmetries in weight spaces and hidden Layers Consider a depth-L MLP. Let $P_0 = P_L = I$, and let $P_1, \dots, P_{L-1}$ be arbitrary permutation matrices. The transformation

$$W_i \mapsto P_{i+1} W_i P_i^\top,\ b_i \mapsto P_{i+1} b_i,\ 0 \le i < L,$$

permutes the neurons within the hidden layers while leaving the realized function unchanged. As shown in prior work Navon et al. (2023), architectures that explicitly respect these symmetries can substantially outperform models that ignore them. We therefore design HIDDENPROBE to be invariant to hidden-neuron permutations.

Under the above transformation, the representation at hidden layer i transforms as $x_i(p;\theta) \mapsto P_i x_i(p;\theta)$, $i \in [L-1]$ Thus, each hidden layer is equivariant to its corresponding permutation P_i. In contrast, the input and final output are unchanged, since $P_0 = P_L = I$.

We now describe how this symmetry acts on hidden probe responses. Consider a tuple of probes $\mathcal{P} = (p_1, \dots, p_N)$. For hidden layer i, define the response vector of neuron k across all probes as $v^{i,k} := (x_{i,k}(p_1;\theta), \dots, x_{i,k}(p_N;\theta)) \in \mathbb{R}^N$. Since x_i is equivariant to P_i, applying a hidden-neuron permutation simply permutes the collection $\left\{v^{i,k} : k \in [d_i]\right\}$. Therefore, each hidden layer is naturally represented as an unordered set of d_i vectors in $\mathbb{R}^N$.

Overview of HIDDENPROBE Motivated by this observation, HIDDENPROBE processes each hidden layer using a permutation-invariant set encoder. Specifically, we encode every hidden layer i as $r_\phi^i := T_\phi^i\left(\left\{v^{i,k} : k \in [d_i]\right\}\right)$, where T_ϕ^i is a layer-wise set encoder. In our implementation, T_ϕ^i is an invariant Set Transformer Lee et al. (2019), making r_ϕ^i invariant to permutations of the neurons in layer i.

The resulting hidden-layer representations are concatenated in layer order, together with the final-layer probe outputs:

$$\Phi^{\text{hidden}}_{\mathcal{P},\phi}(\theta) = \left[r_\phi^1; \dots; r_\phi^{L-1}; f(p_1;\theta); \dots; f(p_N;\theta)\right].$$

A final predictor ρ_ϕ maps this representation to the target: $\widehat{y}(\theta) = \rho_\phi(\Phi^{\text{hidden}}_{\mathcal{P},\phi}(\theta))$. The probes $p_1, \dots, p_N$, the layer-wise encoders T_ϕ^i (allowed to be different across layers but can have shared parameters to reduce overall parameter count), and the predictor ρ_ϕ are trained jointly. The learned probes are shared across all target networks.

A.2 HIDDENPROBE FOR TRANSFORMERS

Learning a mixed-length probe bank. Here we introduce HIDDENPROBE, a probe-based model for learning on softmax-attention based transformers. Our reconstruction arguments from Theorem 5 and Algorithm 2 use different sequence lengths: singleton probes recover the value matrix, while multi-token probes recover the attention weights. This suggests that probes of different lengths can provide complementary information about a target. We retain this design principle, but learn a shared set of probes of varying sequence length rather than imposing the exact configurations.

Design of HIDDENPROBE predictor For each probe $P_j \in \mathcal{P}$, we record the outputs of the transformer blocks, the final normalized token states, and the decoder output:

$$\mathcal{T}_\theta = \left(X_1(P_j;\theta), \cdots, X_L(P_j;\theta),\ X_{\text{norm}}(P_j;\theta),\ f(P_j;\theta)\right)_{P_j \in \mathcal{P}}.$$

For each observed layer and channel, we concatenate its responses across all probes and token positions, using a fixed order. A shared MLP encodes each resulting channel trajectory into one feature vector. These channel features and final logit features are passed to a transformer followed by an MLP to give the output: $\widehat{y}_\theta = R_\phi(\mathcal{T}_\theta)$.

Incorporating weight space symmetries Head permutations and the transformations $Q_a \mapsto Q_a R_a$, $K_a \mapsto K_a R_a^{-\top}$ and $V_a \mapsto V_a S_a$, $O_a \mapsto S_a^{-1} O_a$, for invertible R_a, S_a, leave the attention output unchanged. Consequently, they leave every recorded block response and classifier output unchanged, so HIDDENPROBE is invariant to these symmetries by design. Handling such symmetries remains a key challenge in works like Tran et al. (2025), which we get for free.

B PROOF OF TECHNICAL RESULTS

Theorem 1 (Unique reconstruction from finite samples)**.** *Let $f : \mathbb{R}^d \times \mathbb{R}^q \to \mathbb{R}^m$ be an analytic function. Then, for almost any choice of probes $p_1, \ldots, p_{2q+1} \in \mathbb{R}^d$, we have that for all $\theta_1, \theta_2 \in \Theta$,*

$$f(p_j, \theta_1) = f(p_j, \theta_2), \forall j = 1, \ldots, 2q+1 \iff f(\cdot, \theta_1) \equiv f(\cdot, \theta_2) \text{ on } \mathbb{R}^d$$

Proof. The proof is based on the Finite Witness Theorem (Theorem A.2 in Amir et al. (2023)), which states that, for an analytic $G : \mathbb{R}^a \times \mathbb{R}^b \to \mathbb{R}$, and Lebesgue almost every $w_1, \ldots, w_{a+1} \in \mathbb{R}^b$,

$$\{y \in \mathbb{R}^a \,|\, G(y; w_i) = 0, \forall i = 1, \ldots, a+1\} = \{y \in \mathbb{R}^a \,|\, G(y; w) = 0, \forall w \in \mathbb{R}^b\}$$

To prove our theorem, we define the real analytic function $G(\theta_1, \theta_2; p) := \|f(p; \theta_1) - f(p; \theta_2)\|_2^2$. Since the pair (θ_1, θ_2) resides in $\mathbb{R}^{2q}$, the Finite Witness Theorem implies that for $2q+1$ random samples $p_1, \ldots, p_{2q+1}$, we will have that for all θ_1, θ_2, $f(p_i; \theta_1) - f(p_i; \theta_2) = 0$ for all $i = 1, \ldots, 2q+1$, if and only if $f(p; \theta_1) - f(p; \theta_2) = 0$ for all $p \in \mathbb{R}^d$. □

B.1 IDENTIFICATION OF ANALYTIC MLPS FROM HIDDEN PROBES

We borrow notations from Section 3.1 and Algorithm 1, but occasionally them slightly. Let $L \in \mathbb{N}$, $L \geq 1$, and let $d_0, \ldots, d_L$ be positive integers. Let $\mathcal{X} \subseteq \mathbb{R}^{d_0}$ be nonempty, open, and connected. For each $i \in \{1, \ldots, L\}$, let $\Theta_i \subseteq \mathbb{R}^{p_i}$ be nonempty, open, and connected, and let $f_i : \mathbb{R}^{d_{i-1}} \times \Theta_i \to \mathbb{R}^{d_i}$ be jointly real analytic. Define the following notations:

$$\Theta := \prod_{i=1}^{L} \Theta_i,\ \theta := (\theta_1, \ldots, \theta_L) \in \Theta,\ q := \sum_{i=1}^{L} q_i,\ q_{\max} := \max_{1 \leq i \leq L} q_i.$$

The full model is the composition of these layer maps:

$$x_0(u; \theta) := u,\ x_i(u; \theta) := f_i\left(x_{i-1}(u; \theta), \theta_i\right),\ \forall i \in [L],\ f(u; \theta) := x_L(u; \theta). \tag{4}$$

Given a tuple of probes $\mathcal{P} = (p_1, \ldots, p_N) \in \mathcal{X}^N$, define

$$\begin{aligned} &\mathbf{X}_i^\theta(\mathcal{P}) := (x_i(p_1; \theta) \cdots x_i(p_N; \theta)) \in \mathbb{R}^{d_i \times N},\ i = 0, \ldots, L. \\ &\overline{\mathbf{X}}_i^\theta = \begin{pmatrix} \mathbf{X}_i^\theta \\ \mathbf{1}_N^\top \end{pmatrix} \end{aligned} \tag{5}$$

Thus, full intermediate observation records the sampled states at every layer, including the final output.

Identification of Arbitrary Analytic Nets from Hidden Layers: Theorem 1 shows that $2q+1$ final-layer probes suffice to distinguish any two functionally distinct models in an analytic family. Here, we show that for analytic sequential models of the form in (4), access to intermediate-layer observations reduces the sufficient probe count to $q + q_{\max} + 1$.

Theorem 8 (Identification from intermediate observations)**.** *Under the preceding assumptions, suppose $N \geq q + q_{\max} + 1$. Then, for Lebesgue-almost every $\mathcal{P} \in \mathcal{X}^N$, simultaneously for every $\theta, \theta' \in \Theta$, one has*

$$\mathbf{X}_i^\theta(\mathcal{P}) = \mathbf{X}_i^{\theta'}(\mathcal{P})\, \forall\, i \in \{0, \cdots, L\} \implies f(u; \theta) = f(u; \theta'),\ \forall\, u \in \mathcal{X}$$

Proof. For $i \in [L]$, we use the following notations:

$$q_{<i} := \sum_{k=1}^{i-1} q_k,\ \Theta_{<i} := \prod_{k=1}^{i-1} \Theta_k,\ Q_i := q_{<i} + 2q_i.$$

For $i = 1$, set $q_{<1} = 0$ and $\Theta_{<1} = \mathbb{R}^0$. The state $x_{i-1}(u;\theta)$ depends only on $(\theta_1, \cdots, \theta_{i-1})$. For $\eta = (\eta_1, \cdots, \eta_{i-1}) \in \Theta_{<i}$, we therefore write $x_{i-1}(u;\eta)$ for the state generated by these blocks and for $i = 1$, this means $x_0(u;\eta) = u$.

We define $\Omega_i := \Theta_{<i} \times \Theta_i \times \Theta_i$, and $H_i(u;\eta,\alpha,\beta) := \|f_i\left(x_{i-1}(u;\eta),\alpha\right) - f_i\left(x_{i-1}(u;\eta),\beta\right)\|_2^2$. By composition, H_i is jointly real analytic on $\mathcal{X} \times \Omega_i$, and Ω_i is an open connected subset of $\mathbb{R}^{Q_i}$. Moreover, we have

$$Q_i + 1 = q_{<i} + 2q_i + 1 \le q + q_i + 1 \le q + q_{\max} + 1 \le N.$$

The Finite Witness Theorem (Amir et al., 2023, Theorem A.2) provides a Lebesgue-null set $E_i \subseteq \mathcal{X}^{Q_i+1}$ such that, for every $(p_1, \cdots, p_{Q_i+1}) \notin E_i$ and every $(\eta,\alpha,\beta) \in \Omega_i$,

$$H_i(p_n;\eta,\alpha,\beta) = 0,\ \forall\, n \in [Q_i+1] \implies H_i(u;\eta,\alpha,\beta) = 0,\ \forall\, u \in \mathcal{X}.$$

We define $B_i := E_i \times \mathcal{X}^{N-Q_i-1}$, $B := \bigcup_{i=1}^{L} B_i$, where $\mathcal{X}^0$ denotes a singleton. Each B_i, and hence B, is Lebesgue-null. For every $\mathcal{P} = (p_1, \cdots, p_N) \notin B$, the preceding implication holds at every layer, simultaneously for all $(\eta,\alpha,\beta) \in \Omega_i$. In particular, B does not depend on the pair of parameters being compared.

Fix $\mathcal{P} \notin B$ and $\theta = (\theta_1, \cdots, \theta_L)$, $\theta' = (\theta'_1, \cdots, \theta'_L)$ satisfying $\mathbf{X}_i^{\theta}(\mathcal{P}) = \mathbf{X}_i^{\theta'}(\mathcal{P})$, $i = 0, \cdots, L$. We prove by induction that $x_i(\cdot;\theta) \equiv x_i(\cdot;\theta')$ on $\mathcal{X}$. For $i = 0$, both maps are the identity.

Suppose the assertion holds for $i-1$, and set $\eta = (\theta_1, \cdots, \theta_{i-1})$. Then, for every $u \in \mathcal{X}$, we have

$$x_{i-1}(u;\eta) = x_{i-1}(u;\theta) = x_{i-1}(u;\theta'), \text{ and } H_i(u;\eta,\theta_i,\theta'_i) = \|x_i(u;\theta) - x_i(u;\theta')\|_2^2.$$

Equality at the probes $\mathcal{P}$ therefore gives $H_i(p_n;\eta,\theta_i,\theta'_i) = 0$, $n = 1, \cdots, N$. Since $\mathcal{P} \notin B_i$, the finite witness theorem implies $H_i(u;\eta,\theta_i,\theta'_i) = 0$ for every $u \in \mathcal{X}$. Thus $x_i(\cdot;\theta) \equiv x_i(\cdot;\theta')$, completing the induction. □

When the activation is also injective and non polynomial: We now show that if the activation is analytic, injective, and non-polynomial, the sufficient number of probes can be substantially reduced. We prove a stronger version of Theorem 2 for that. We first prove a couple of lemmas that will be useful in proving the result.

Lemma 1. *Let $y_1, \ldots, y_K \in \mathbb{R}^q$ be pairwise distinct. Let $\sigma : \mathbb{R} \to \mathbb{R}$ be real analytic and non-polynomial. Then*

$$\operatorname{span}\left\{\left(\sigma(w^\top y_1 + b), \ldots, \sigma(w^\top y_K + b)\right) : (w,b) \in \mathbb{R}^q \times \mathbb{R}\right\} = \mathbb{R}^K.$$

Consequently, for Lebesgue-a.e. $(w_1,b_1), \ldots, (w_{K-1},b_{K-1}) \in (\mathbb{R}^q \times \mathbb{R})^{K-1}$, the matrix

$$M = \begin{pmatrix} 1 & \cdots & 1 \\ \sigma(w_1^\top y_1 + b_1) & \cdots & \sigma(w_1^\top y_K + b_1) \\ \vdots & & \vdots \\ \sigma(w_{K-1}^\top y_1 + b_{K-1}) & \cdots & \sigma(w_{K-1}^\top y_K + b_{K-1}) \end{pmatrix}$$

is invertible.

Proof. We consider the non-trivial case $K \ge 2$. Choose $v \in \mathbb{R}^q$ such that the numbers $t_n := v^\top y_n$, $n \in [K]$, are pairwise distinct. Such a choice is possible because the bad set is the finite union of proper hyperplanes $\bigcup_{n<n'} \{v \in \mathbb{R}^q : v^\top(y_n - y_{n'}) = 0\}$, each of which is proper since $y_n \ne y_{n'}$. We first prove the span claim. Suppose, for contradiction, that the span is proper. Then there exists a nonzero vector $c = (c_1, \ldots, c_K) \in \mathbb{R}^K$ such that

$$\begin{aligned} &\sum_{n=1}^{K} c_n \sigma(w^\top y_n + b) = 0\, \forall\, (w,b) \in \mathbb{R}^q \times \mathbb{R} \\ \implies &\sum_{n=1}^{K} c_n \sigma(b + a t_n) = 0\, \forall\, a,b \in \mathbb{R}, \text{ by taking } w = av. \end{aligned} \tag{6}$$

Since σ is real analytic and non-polynomial, none of the derivatives $\sigma^{(k)}$, $k = 0, \ldots, K-1$, is identically zero. Indeed, if $\sigma^{(k)} \equiv 0$ for some k, then σ would be a polynomial of degree at most $k-1$. Therefore, for each $k = 0, \ldots, K-1$, the zero set of $\sigma^{(k)}$ is a proper discrete subset of $\mathbb{R}$. Hence the finite union $\bigcup_{k=0}^{K-1} \{b \in \mathbb{R} : \sigma^{(k)}(b) = 0\}$ does not cover $\mathbb{R}$. Choose $b_0 \in \mathbb{R}$ outside this union. Then we have:

$$\sigma^{(k)}(b_0) \neq 0,\ k = 0, \ldots, K-1.$$

Now set $b = b_0$ in equation 6. Since σ is analytic at b_0, for all sufficiently small a we have:

$$0 = \sum_{n=1}^{K} c_n \sigma(b_0 + a t_n) = \sum_{n=1}^{K} c_n \sum_{k=0}^{\infty} \frac{\sigma^{(k)}(b_0)}{k!} a^k t_n^k$$

$$\implies 0 = \sum_{k=0}^{\infty} \frac{\sigma^{(k)}(b_0)}{k!} \left(\sum_{n=1}^{K} c_n t_n^k \right) a^k.$$

Since the convergent power series is identically 0 in a neighbourhood around the origin, every coefficient vanishes, and since $\sigma^{(k)}(b_0) \neq 0$ for $k \leq K-1$ we get $\sum_{n=1}^{K} c_n t_n^k = 0$ for $k = 0, \ldots, K-1$. Hence c lies in the kernel of the Vandermonde matrix

$$V = \begin{pmatrix} 1 & \cdots & 1 \\ t_1 & \cdots & t_K \\ \vdots & & \vdots \\ t_1^{K-1} & \cdots & t_K^{K-1} \end{pmatrix}.$$

But the t_n are pairwise distinct, so we have:

$$\det V = \prod_{1 \leq n < n' \leq K} (t_{n'} - t_n) \neq 0.$$

Therefore $c = 0$, contradicting the choice of c. This proves the span claim. Now we define

$$v(w, b) := \big(\sigma(w^\top y_1 + b), \ldots, \sigma(w^\top y_K + b)\big) \in \mathbb{R}^K.$$

We first show that there exists some $(K-1)$-tuple of parameters such that the matrix M is full rank. We start with the one-dimensional subspace $V_0 := \operatorname{span}\{\mathbf{1}_K\}$. If every vector $v(w, b)$ belonged to V_0, then the span of all ridge vectors would be contained in V_0, but these vectors span $\mathbb{R}^K$. Hence there exists (w_1, b_1) such that $v(w_1, b_1) \notin V_0$. Set $V_1 := \operatorname{span}\{\mathbf{1}_K, v(w_1, b_1)\}$. Continuing inductively (as long as $\dim V_j < K$ some ridge vector lies outside V_j), after $K-1$ steps we obtain parameters $(w_1, b_1), \ldots, (w_{K-1}, b_{K-1})$ such that the K vectors

$$\mathbf{1}_K,\ v(w_1, b_1), \ldots, v(w_{K-1}, b_{K-1})$$

are linearly independent in $\mathbb{R}^K$. Since there are K such vectors, they form a basis of $\mathbb{R}^K$. Equivalently, the corresponding matrix M has nonzero determinant for at least one choice of parameters. Now we define

$$\Delta((w_1, b_1), \ldots, (w_{K-1}, b_{K-1})) := \det M.$$

Since σ is real analytic, every entry of M is a real-analytic function of the parameters (w_1, b_1), …, (w_{K-1}, b_{K-1}), and therefore Δ is real analytic on the connected open set $(\mathbb{R}^q \times \mathbb{R})^{K-1}$. The preceding construction shows that Δ is not identically zero. Hence its zero set is the zero set of a nonzero real-analytic function on a connected open set, so it has Hausdorff dimension strictly smaller than $(K-1)(q+1)$, implying zero Lebesgue measure. Therefore M is invertible for Lebesgue-a.e. choice of the ridge parameters. □

Now, for the following lemma, we define another notation for MLPs with width sequence $\mathbf{D} := (d = d_0, d_1, \cdots, d_{L-1}, d_L = m)$. For probes $\mathcal{P} \in \mathcal{X}^N$ and $\theta \in \Theta_{\mathbf{D}}$, we recall the definitions $\overline{\mathbf{X}}_i^\theta(\mathcal{P}) := \begin{pmatrix} \mathbf{X}_i^\theta(\mathcal{P}) \\ \mathbf{1}_N^\top \end{pmatrix} \in \mathbb{R}^{(d_i+1) \times N}$, where $\mathbf{X}_i^\theta(\mathcal{P})$ is as defined in equation 5. We show that, for sufficiently large $N := |\mathcal{P}|$, the matrix $\overline{\mathbf{X}}_i^\theta(\mathcal{P})$ is almost always full rank.

Lemma 2. *Assume that $\sigma : \mathbb{R} \to \mathbb{R}$ is real analytic, injective, and non-polynomial. Let $N \geq \max_{0\leq i\leq L-1}(d_i+1)$, and define*

$$E := E_1 \cup E_2, \text{ where :}$$
$$E_1 := \left\{ \mathcal{P} = (p_1, \ldots, p_N) \in \mathcal{X}^N : \operatorname{rank} \begin{pmatrix} p_1 & \cdots & p_N \\ 1 & \cdots & 1 \end{pmatrix} < d+1 \right\},$$
$$E_2 := \bigcup_{1\leq n<n'\leq N} \{\mathcal{P} \in \mathcal{X}^N : p_n = p_{n'}\}.$$

Then $\dim_{\mathrm{H}}(E) < Nd$, and for every probe tuple $\mathcal{P} \in \mathcal{X}^N \setminus E$ there exists a Lebesgue-null set $\mathcal{M}_{\mathcal{P}} \subseteq \Theta_{\mathbf{D}}$ such that

$$\operatorname{rank} \overline{\mathbf{X}}_i^{\theta_0}(\mathcal{P}) = d_i + 1, \, \forall\, i \in \{0, \cdots, L-1\},$$

for every $\theta_0 \in \Theta_{\mathbf{D}} \setminus \mathcal{M}_{\mathcal{P}}$.

Proof. Since $\mathcal{X}$ has nonempty interior, choose $x_0 \in \mathcal{X}$ and $\epsilon > 0$ such that $B(x_0, \epsilon) \subseteq \mathcal{X}$. Let $e_1, \ldots, e_d$ be the standard basis of $\mathbb{R}^d$. Then the determinant

$$P(p_1, \ldots, p_{d+1}) := \det \begin{pmatrix} p_1 & \cdots & p_{d+1} \\ 1 & \cdots & 1 \end{pmatrix}$$

is nonzero for the choice of points $\{x_0, x_0 + \frac{\epsilon}{2}e_1, \cdots, x_0 + \frac{\epsilon}{2}e_d\}$: subtracting the first column from each of the others gives the value $\pm(\epsilon/2)^d$. Hence P is not the zero polynomial. If a probe tuple belongs to E_1, then every $(d+1)\times(d+1)$ minor of the matrix in the definition of E_1 vanishes, so $P(p_1, \ldots, p_{d+1}) = 0$. Thus $E_1 \subseteq \{P = 0\} \times \mathcal{X}^{N-d-1}$. Since P is a nonzero polynomial, $\{P = 0\}$ has Hausdorff dimension strictly smaller than $(d+1)d$. Therefore $\dim_{\mathrm{H}}(E_1) < Nd$. Moreover, each diagonal set $\{p_n = p_{n'}\}$ has Hausdorff dimension $(N-1)d < Nd$. Since E is a finite union of such sets, we have $\dim_{\mathrm{H}}(E) < Nd$.

Fix a probe tuple $\mathcal{P} = (p_1, \ldots, p_N) \in \mathcal{X}^N \setminus E$. Then the probes are pairwise distinct and $\operatorname{rank} \overline{\mathbf{X}}_0(\mathcal{P}) = d+1$. Note that $\overline{\mathbf{X}}_i^{\theta}(\mathcal{P})$ and the sampled states $x_i(p_n; \theta)$ depend on $\theta = (B_0, \ldots, B_{L-1})$ only through the blocks $B_0, \ldots, B_{i-1}$. We shall prove inductively that, for Lebesgue-a.e. choice of the blocks $B_0, \ldots, B_{i-1}$, the following two properties hold:

$$\operatorname{rank} \overline{\mathbf{X}}_i^{\theta}(\mathcal{P}) = d_i + 1, \tag{R_i}$$
$$x_i(p_1; \theta), \ldots, x_i(p_N; \theta) \text{ are pairwise distinct.} \tag{D_i}$$

For $i = 0$, property (R_0) holds by the definition of E_1, and (D_0) holds because the probes $p_1, \ldots, p_N$ are pairwise distinct; neither involves any block. Now fix $0 \leq i \leq L-2$ and blocks $B_0, \ldots, B_{i-1}$ for which (D_i) holds. Set

$$y_n := x_i(p_n; \theta) \in \mathbb{R}^{d_i},\ n = 1, \ldots, N.$$

By (D_i), the vectors $y_1, \ldots, y_N$ are pairwise distinct. We first prove (R_{i+1}). Let $K := d_{i+1} + 1$. Since $i + 1 \leq L - 1$, we have $K \leq N$, so the first K vectors $y_1, \ldots, y_K$ are pairwise distinct. Write $(w_j^\top, b_j)$, $j \in [d_{i+1}]$, for the rows of B_i and define

$$M_{i+1} := \begin{pmatrix} 1 & \cdots & 1 \\ \sigma\left(w_1^\top y_1 + b_1\right) & \cdots & \sigma\left(w_1^\top y_K + b_1\right) \\ \vdots & & \vdots \\ \sigma\left(w_{d_{i+1}}^\top y_1 + b_{d_{i+1}}\right) & \cdots & \sigma\left(w_{d_{i+1}}^\top y_K + b_{d_{i+1}}\right) \end{pmatrix}.$$

Up to moving the row of ones to the bottom, M_{i+1} consists of the first K columns of $\overline{\mathbf{X}}_{i+1}^{\theta}(\mathcal{P})$, i.e. it is a $(d_{i+1}+1)\times(d_{i+1}+1)$ minor of $\overline{\mathbf{X}}_{i+1}^{\theta}(\mathcal{P})$. Applying Lemma 1 in $\mathbb{R}^{d_i}$ to the K points $y_1, \ldots, y_K$ gives that, for Lebesgue-a.e. choice of the rows $(w_1^\top, b_1), \ldots, (w_{d_{i+1}}^\top, b_{d_{i+1}})$ of B_i, the matrix M_{i+1} is invertible. Therefore, for Lebesgue-a.e. block B_i, we have $\operatorname{rank} \overline{\mathbf{X}}_{i+1}^{\theta}(\mathcal{P}) = d_{i+1} + 1$. Thus (R_{i+1}) holds for Lebesgue-a.e. block B_i, conditional on the previous blocks satisfying (D_i). We

next prove (D_{i+1}). Fix $n \neq n'$. Since $y_n \neq y_{n'}$, if $x_{i+1}(p_n;\theta) = x_{i+1}(p_{n'};\theta)$, then, by injectivity of σ applied coordinatewise, we get:

$$W_i y_n + b_i = W_i y_{n'} + b_i.$$

Equivalently, $W_i(y_n - y_{n'}) = 0$. Since $y_n - y_{n'} \neq 0$, this is a proper linear condition on the entries of W_i. Hence, for fixed previous blocks, it defines a Lebesgue-null subset of the space of blocks B_i. Taking the finite union over all pairs $n \neq n'$, we get that (D_{i+1}) holds for Lebesgue-a.e. block B_i. It remains to pass from a fixed choice of the previous blocks to all parameters. Write $q_k := d_{k+1}(d_k+1)$ for the size of block B_k and, for $-1 \leq i \leq L-2$, let $\mathcal{B}_i \subseteq \mathbb{R}^{q_0+\cdots+q_i}$ be the set of prefixes $(B_0,\ldots,B_i)$ for which (R_k) or (D_k) fails for some $k \leq i+1$; thus $\mathcal{B}_{-1} = \emptyset$. Each $\mathcal{B}_i$ is a Borel set: the failure of (R_k) is the vanishing of all $(d_k+1)\times(d_k+1)$ minors of $\overline{\mathbf{X}}_k^{\theta}(\mathcal{P})$, finitely many real-analytic functions of the prefix, and the failure of (D_k) is a finite union of equalities between real-analytic functions. If $(B_0,\ldots,B_i) \in \mathcal{B}_i$, then either $(B_0,\ldots,B_{i-1}) \in \mathcal{B}_{i-1}$, or this prefix satisfies (D_i) and B_i belongs to the set $F(B_0,\ldots,B_{i-1})$ of blocks for which (R_{i+1}) or (D_{i+1}) fails, which we have just shown to be Lebesgue-null. Hence, by Tonelli's theorem,

$$\lambda(\mathcal{B}_i) \leq \lambda\left(\mathcal{B}_{i-1}\times\mathbb{R}^{q_i}\right) + \int_{(B_0,\ldots,B_{i-1})\notin\mathcal{B}_{i-1}} \lambda\left(F(B_0,\ldots,B_{i-1})\right)\, d(B_0,\ldots,B_{i-1}) = 0,$$

where the first term vanishes because $\mathcal{B}_{i-1}$ is null by the induction hypothesis and the second because its integrand vanishes. Iterating this over $i = 0,\ldots,L-2$, every $\mathcal{B}_i$ is Lebesgue-null. Finally, since (R_i) only involves the blocks $B_0,\ldots,B_{i-1}$, the set of parameters $\theta \in \mathbb{R}^{q_{\mathbf{D}}}$ for which (R_i) fails for some $i \in \{0,\cdots,L-1\}$ is contained in $\mathcal{B}_{L-2}\times\mathbb{R}^{q_{L-1}}$, which is null; its intersection with $\Theta_{\mathbf{D}}$ is the required $\mathcal{M}_{\mathcal{P}}$. This completes the proof. □

We now prove Theorem 9, which is a stronger version of Theorem 2.

Theorem 9 (Reconstruction from hidden layer activations)**.** *Let $\mathcal{X} \subseteq \mathbb{R}^{d_0}$ be compact with nonempty interior and $\sigma : \mathbb{R} \to \mathbb{R}$ be real analytic, injective, and non-polynomial. Consider fully connected MLPs with biases, widths $\mathbf{D} = (d_0,\ldots,d_L)$, and raw parameter space $\Theta_{\mathbf{D}} = \mathbb{R}^{q_{\mathbf{D}}}$, where $q_{\mathbf{D}} = \sum_{i=0}^{L-1} d_{i+1}(d_i+1)$. Suppose $N \geq 1 + \max_{0\leq i<L} d_i$. For Lebesgue-almost every probe tuple $\mathcal{P} \in \mathcal{X}^N$, there exists a nonzero real-analytic function $\Delta_{\mathcal{P}} : \mathbb{R}^{q_{\mathbf{D}}} \to \mathbb{R}$ such that $\mathcal{M}_{\mathcal{P}} = \{\theta : \Delta_{\mathcal{P}}(\theta) = 0\}$ and we have*

$$\Psi_{all}(\theta,\mathcal{P}) = \Psi_{all}(\theta_0,\mathcal{P}) \iff \theta = \theta_0,\ \forall\,\theta_0 \in \Theta_{\mathbf{D}} \setminus \mathcal{M}_{\mathcal{P}},\ \forall\,\theta \in \Theta_{\mathbf{D}}$$

The exceptional set $\mathcal{M}_{\mathcal{P}}$ is closed and satisfies $\dim_{\mathrm{H}} \mathcal{M}_{\mathcal{P}} \leq q_{\mathbf{D}} - 1$; hence it is Lebesgue-null and nowhere dense.

Proof. Let $E \subseteq \mathcal{X}^N$ be the exceptional probe set in Lemma 2, applied with the given N. Then E is Lebesgue-null. Fix $\mathcal{P} \in \mathcal{X}^N \setminus E$ and define

$$\Delta_{\mathcal{P}}(\theta) := \prod_{i=0}^{L-1} \det\left(\overline{\mathbf{X}}_i^{\theta}(\mathcal{P})\overline{\mathbf{X}}_i^{\theta}(\mathcal{P})^{\top}\right),\ \mathcal{M}_{\mathcal{P}} := \left\{\theta \in \mathbb{R}^{q_{\mathbf{D}}} : \Delta_{\mathcal{P}}(\theta) = 0\right\}.$$

Each entry of $\overline{\mathbf{X}}_i^{\theta}(\mathcal{P})$ is real analytic in θ, so $\Delta_{\mathcal{P}}$ is real analytic on $\mathbb{R}^{q_{\mathbf{D}}}$.

For any real matrix A, the identity $v^{\top}AA^{\top}v = \|A^{\top}v\|_2^2$ shows that $AA^{\top}$ is positive definite exactly when A has full row rank. Consequently,

$$\Delta_{\mathcal{P}}(\theta) > 0 \iff \operatorname{rank}\overline{\mathbf{X}}_i^{\theta}(\mathcal{P}) = d_i + 1 \text{ for every } i = 0,\ldots,L-1.$$

Lemma 2 gives these ranks simultaneously for Lebesgue-almost every $\theta \in \mathbb{R}^{q_{\mathbf{D}}}$. In particular, they hold at some θ_*, and hence $\Delta_{\mathcal{P}}(\theta_*) > 0$. Thus $\Delta_{\mathcal{P}}$ is not identically zero.

By continuity, $\mathcal{M}_{\mathcal{P}}$ is closed and $\dim_{\mathrm{H}} \mathcal{M}_{\mathcal{P}} \leq q_{\mathbf{D}} - 1$ since it forms the zero-set of an analytic function. Therefore $\mathcal{M}_{\mathcal{P}}$ is Lebesgue-null. It has empty interior, since every nonempty open subset of $\mathbb{R}^{q_{\mathbf{D}}}$ has positive Lebesgue measure. Being closed with empty interior, it is nowhere dense.

Fix $\theta_0 = (B_0^0,\ldots,B_{L-1}^0) \in \Theta_{\mathbf{D}} \setminus \mathcal{M}_{\mathcal{P}}$. By the definition of $\Delta_{\mathcal{P}}$,

$$\operatorname{rank}\overline{\mathbf{X}}_i^{\theta_0}(\mathcal{P}) = d_i + 1,\ i = 0,\ldots,L-1. \tag{7}$$

If $\theta = \theta_0$ then trivially $\Psi_{\text{all}}(\theta, \mathcal{P}) = \Psi_{\text{all}}(\theta_0, \mathcal{P})$.

Conversely, let $\theta = (B_0, \ldots, B_{L-1}) \in \Theta_{\mathbf{D}}$ satisfy $\Psi_{\text{all}}(\theta, \mathcal{P}) = \Psi_{\text{all}}(\theta_0, \mathcal{P})$. We show that $B_i = B_i^0$ layer by layer. Fix $i \in \{0, \cdots, L-1\}$. By definition, $\mathbf{X}_i^{\theta}(\mathcal{P}) = \mathbf{X}_i^{\theta_0}(\mathcal{P})$ and $\mathbf{X}_{i+1}^{\theta}(\mathcal{P}) = \mathbf{X}_{i+1}^{\theta_0}(\mathcal{P})$,. The first equality gives $\overline{\mathbf{X}}_i^{\theta}(\mathcal{P}) = \overline{\mathbf{X}}_i^{\theta_0}(\mathcal{P})$, and the second equality reads $\sigma\left(B_i\, \overline{\mathbf{X}}_i^{\theta_0}(\mathcal{P})\right) = \sigma\left(B_i^0\, \overline{\mathbf{X}}_i^{\theta_0}(\mathcal{P})\right)$ entrywise.. Since σ is injective and applied entrywise, this implies $B_i \overline{\mathbf{X}}_i^{\theta_0}(\mathcal{P}) = B_i^0 \overline{\mathbf{X}}_i^{\theta_0}(\mathcal{P})$. Equivalently, we have

$$\begin{pmatrix} W_i - W_i^0 & b_i - b_i^0 \end{pmatrix} \overline{\mathbf{X}}_i^{\theta_0}(\mathcal{P}) = 0.$$

By equation 7, $\overline{\mathbf{X}}_i^{\theta_0}(\mathcal{P})$ has full row rank, so the $(d_i + 1) \times (d_i + 1)$ matrix $\overline{\mathbf{X}}_i^{\theta_0}(\mathcal{P})\, \overline{\mathbf{X}}_i^{\theta_0}(\mathcal{P})^\top$ is invertible and

$$\overline{\mathbf{X}}_i^{\theta_0}(\mathcal{P})^+ := \overline{\mathbf{X}}_i^{\theta_0}(\mathcal{P})^\top \left(\overline{\mathbf{X}}_i^{\theta_0}(\mathcal{P})\, \overline{\mathbf{X}}_i^{\theta_0}(\mathcal{P})^\top\right)^{-1}$$

is a right inverse of $\overline{\mathbf{X}}_i^{\theta_0}(\mathcal{P})$. Multiplying the previous display by it on the right gives $W_i = W_i^0$ and $b_i = b_i^0$, i.e. $B_i = B_i^0$. Since i was arbitrary, $\theta = \theta_0$. Hence $\Psi_{\text{all}}(\theta, \mathcal{P}) = \Psi_{\text{all}}(\theta_0, \mathcal{P}) \iff \theta = \theta_0$, and in particular $\Psi_{\text{all}}(\cdot, \mathcal{P})$ is injective on $\Theta_{\mathbf{D}} \setminus \mathcal{M}_{\mathcal{P}}$. □

Algorithm to recover weights from Ψ_{all} for generic targets Our proof of Theorem 9 gives us a way to actually reconstruct the weights rather than just the functions for generic targets. We formally describe the algorithm as follows:

Algorithm 1 Weights from hidden activations

Require: $\mathcal{P} = (p_1, \cdots, p_N)$,
$\Psi_{\text{all}}(\theta_0, \mathcal{P}) = (\mathbf{X}_1, \cdots, \mathbf{X}_L)$

$\mathbf{X}_0 \leftarrow (p_1, \cdots, p_N)$
for $i = 0, \cdots, L-1$ **do**
  $\overline{\mathbf{X}}_i \leftarrow \begin{pmatrix} \mathbf{X}_i \\ \mathbf{1}_N^\top \end{pmatrix}$
  $\mathbf{Z}_{i+1} \leftarrow \sigma^{-1}(\mathbf{X}_{i+1})$ ▷ entrywise
  $\widehat{B}_i \leftarrow \mathbf{Z}_{i+1}\, \overline{\mathbf{X}}_i^\dagger$ ▷ pseudoinverse
  $(\widehat{W}_i \;\; \widehat{b}_i) \leftarrow \widehat{B}_i$
end for
return $\widehat{\theta} := (\widehat{B}_0, \cdots, \widehat{B}_{L-1})$

B.2 Lower bounds on probe counts

For analytic MLPs, we have established upper bounds on the number of probes required for unique function identification, using either final-layer or hidden-layer probes. We now establish nearly matching lower bounds for shallow networks. Since these lower bounds hold for one-hidden-layer networks of arbitrary width, they also provide necessary lower bounds for the general family of MLPs with analytic activations (or analytic, injective, non-polynomial activations when observing Ψ_{all}). The formal results are stated in Theorem 10.

We now prove several lemmas needed to establish the tightness results in Theorem 10.

Lemma 3. *Let $K \geq 1$ and let T be a nonzero real trigonometric polynomial of the form $T(t) = A_0 + \sum_{k=1}^{K} (A_k \cos(kt) + B_k \sin(kt))$, not identically 0. Then T has at most $2K$ distinct zeros in $[0, 2\pi)$. Moreover, suppose another nonzero sum H of the same form has $2K$ distinct zeros in $[0, 2\pi)$. If T vanishes at all those points, then $T = \lambda H$ for some real number λ.*

Proof. Write $T(t) = \sum_{k=-K}^{K} c_k e^{ikt}$ and define $P_T(z) := \sum_{k=-K}^{K} c_k z^{k+K}$. The polynomial P_T is nonzero, has degree at most $2K$, and satisfies $P_T(e^{it}) = e^{iKt} T(t)$. Distinct zeros modulo 2π give distinct roots of P_T, proving the first assertion. For the second assertion, P_H has degree $2K$ and $2K$

distinct roots, all shared by P_T. Hence $P_T = \lambda P_H$. Therefore $T = \lambda H$, with λ real because both functions are real-valued and $H \not\equiv 0$. □

Lemma 4. *Let $K \geq 1$ and let $H(t) = \sum_{k=1}^{K} \gamma_k \sin(kt) \not\equiv 0$ have $2K$ distinct zeros $r_1, \ldots, r_{2K} \in [0, 2\pi)$. We define:*

$$\mathcal{Z} := (r_1, \ldots, r_{2K}, r_1 + 2\pi, \ldots, r_{2K} + 2\pi),$$
$$\mathcal{V}_K := \left\{ Q(t) + tR(t) : \begin{array}{l} Q \in \operatorname{span}\{1, \sin(kt), \cos(kt) : k \in [K]\}, \\ R \in \operatorname{span}\{\cos(kt) : k \in [K]\} \end{array} \right\}.$$

Writing $\mathcal{Z} = (z_1, \ldots, z_{4K})$, the set of sampled vectors $\{(S(z_1), \ldots, S(z_{4K})) : S \in \mathcal{V}_K\}$ is a $3K$-dimensional linear subspace of $\mathbb{R}^{4K}$. Moreover, a function $S \in \mathcal{V}_K$ satisfies $S(z_n) = 0$ for every $n = 1, \ldots, 4K$ if and only if there exists $\lambda \in \mathbb{R}$ such that $S(t) = \lambda H(t)$ for every $t \in \mathbb{R}$.

Proof. Every $S \in \mathcal{V}_K$ can be written as $S(t) = Q(t) + tR(t)$, $Q(t) = a_0 + \sum_{k=1}^{K}, (a_k \sin(kt) + b_k \cos(kt))$, $R(t) = \sum_{k=1}^{K} c_k \cos(kt)$.

We first verify that these coefficients are uniquely determined by S. It suffices to show that $S(t) = 0$ for every t forces all coefficients to be zero. Since Q and R are 2π-periodic, subtracting the identity at t from the identity at $t+2\pi$ gives $0 = S(t+2\pi) - S(t) = 2\pi R(t)$. Thus $R = 0$, and then $Q = 0$. For each $\ell = 1, \ldots, K$, the elementary sine and cosine integration identities give $\int_0^{2\pi} R(t)\cos(\ell t)\, dt = \pi c_\ell$, $\int_0^{2\pi} Q(t)\, dt = 2\pi a_0$, $\int_0^{2\pi} Q(t)\sin(\ell t)\, dt = \pi a_\ell$, $\int_0^{2\pi} Q(t)\cos(\ell t)\, dt = \pi b_\ell$., which uniquely identifies the coefficients, implying linear independence.

We next characterize the functions whose sampled values are all zero. Suppose $S(z_n) = 0$ for every $n = 1, \ldots, 4K$. For each $j = 1, \ldots, 2K$, periodicity gives $0 = S(r_j) = Q(r_j) + r_j R(r_j)$, $0 = S(r_j + 2\pi) = Q(r_j) + (r_j + 2\pi)R(r_j)$. Subtracting these yields $R(r_j) = 0$, $j = 1, \ldots, 2K$. Since R is a finite cosine sum using frequencies at most K, Lemma 3 implies $R = \mu H$ for some $\mu \in \mathbb{R}$; if $R = 0$, take $\mu = 0$.

For every real t, cosine symmetry and sine antisymmetry give $R(-t) = R(t)$, $H(-t) = -H(t)$. Therefore we get: $\mu H(t) = R(t) = R(-t) = \mu H(-t) = -\mu H(t)$. It follows that $2\mu H(t) = 0$ for every t. Because H is not identically zero, $\mu = 0$, and hence $R = 0$.

We now have $S = Q$ and $Q(r_j) = 0$ for all j. A second application of Lemma 3 gives $S(t) = Q(t) = \lambda H(t) \,\forall\, t \in \mathbb{R}$ for some $\lambda \in \mathbb{R}$, including $\lambda = 0$ when $Q = 0$. Conversely, every function $S = \lambda H$ vanishes at all r_j and $r_j + 2\pi$, since H is 2π-periodic. This proves the stated equivalence.

It remains to determine the dimension of the set of sampled vectors. For any function ϕ, write $v_\phi := (\phi(z_1), \ldots, \phi(z_{4K})) \in \mathbb{R}^{4K}$. Choose $k_* \in \{1, \ldots, K\}$ such that $\gamma_{k_*} \neq 0$, which is possible because H is not identically zero. Consider the following collection of $3K$ functions:

$$\mathcal{B} := \{1\} \cup \{\sin(kt) : 1 \leq k \leq K,\ k \neq k_*\} \cup \{\cos(kt) : 1 \leq k \leq K\} \cup \{t\cos(kt) : 1 \leq k \leq K\}.$$

We show that their sampled vectors $\{v_\phi : \phi \in \mathcal{B}\}$ are linearly independent and that every sampled vector from $\mathcal{V}_K$ is a linear combination of them.

For independence, suppose real coefficients d_ϕ satisfy $\sum_{\phi \in \mathcal{B}} d_\phi v_\phi = 0$. Define $S(t) = \sum_{\phi \in \mathcal{B}} d_\phi \phi(t)$. Then $S \in \mathcal{V}_K$ and $S(z_n) = 0$ for every n. The equivalence already proved implies $S = \lambda H$.

The coefficient of $\sin(k_* t)$ in the displayed expression for S is zero, because this function was omitted from $\mathcal{B}$. Its coefficient in λH is $\lambda\gamma_{k_*}$. Uniqueness of the coefficient representation therefore gives $0 = \lambda\gamma_{k_*}$. Since $\gamma_{k_*} \neq 0$, we obtain $\lambda = 0$ and hence $S = 0$. The linear independence established at the beginning of the proof now forces $d_\phi = 0$ for every $\phi \in \mathcal{B}$. Thus the $3K$ sampled vectors are linearly independent.

Finally, because H vanishes at every point of $\mathcal{Z}$, $0 = v_H = \sum_{k=1}^{K} \gamma_k v_{\sin(k\,\cdot)}$. Solving for the sampled vector of the omitted sine function gives $v_{\sin(k_*\,\cdot)} = -\sum_{\substack{k=1 \\ k \neq k_*}}^{K} \frac{\gamma_k}{\gamma_{k_*}} v_{\sin(k\,\cdot)}$. Every function in $\mathcal{V}_K$ is a linear combination of the functions in $\mathcal{B}$ and this one omitted sine function. The last identity therefore shows that every sampled vector from $\mathcal{V}_K$ is a linear combination of $\{v_\phi : \phi \in \mathcal{B}\}$.

Conversely, every linear combination of these sampled vectors is obtained by sampling the corresponding linear combination of functions in $\mathcal{B}$, which belongs to $\mathcal{V}_K$. Hence the set of sampled vectors is exactly the linear subspace generated by these $3K$ linearly independent vectors. Its dimension is therefore $3K$. □

Lemma 5. *Let $\mathcal{X} \subseteq \mathbb{R}$ be a nonempty open interval. If $w_j \neq 0$ for all $j \in [m]$ and $b_j/w_j \neq b_k/w_k$ whenever $j \neq k$, then $1, \tanh(w_1 x + b_1), \ldots, \tanh(w_m x + b_m)$ are linearly independent on $\mathcal{X}$.*

Proof. Suppose $C + \sum_{j=1}^{m} A_j \tanh(w_j x + b_j) = 0$ on $\mathcal{X}$. Extend the left-hand side to the meromorphic function

$$\Phi(z) := C + \sum_{j=1}^{m} A_j \tanh(w_j z + b_j).$$

Since $\tanh = \sinh / \cosh$ and $\cosh' = \sinh$, the poles of its jth summand occur at

$$z_{j,k} = \frac{i\pi(k + 1/2) - b_j}{w_j},\ k \in \mathbb{Z},$$

and have residue A_j/w_j. Their real part is $-b_j/w_j$, so no pole is shared by two distinct summands. The union of the pole sets is closed and discrete, and its complement in $\mathbb{C}$ is connected: a segment between any two points can be detoured around the finitely many poles it meets. By the Identity Theorem of complex analysis, Φ is zero on this complement because it vanishes on $\mathcal{X}$. Its residue at every pole therefore vanishes. Thus $A_j/w_j = 0$ for every j, giving $A_j = 0$ and then $C = 0$. □

Theorem 10 (Tightness of probe complexities)**.** *Let $f(u, \theta)$ denote a shallow universal MLP, with analytic activation σ, and layers of dimension $\mathbf{D} = (1, m, 1)$ with $m \geq 1$. Thus, the total number of parameters is $q = 3m + 1$ and the maximal dimension of any layer is $d_{\max} = m$. In each of the following cases, there exists a compact parameter space $\Theta_{\mathbf{D}} \subseteq \mathbb{R}^q$ with nonempty interior for which the stated conclusion holds. We have the following:*

1. *There exists a real-analytic activation σ (namely, $\sigma = \sin$) for which the following holds: if $N \leq 2q - 2 = 6m$ then there exists a non-empty open interval $\mathcal{U}_N \subseteq \mathbb{R}^N$ such that for all $\mathcal{P} \in \mathcal{U}_N$*

$$\exists\, \theta_1, \theta_2 \in \Theta_{\mathbf{D}},\ \text{such that } \Psi_{\text{final}}(\theta_1, \mathcal{P}) = \Psi_{\text{final}}(\theta_2, \mathcal{P}) \text{ and } f(\cdot; \theta_1) \not\equiv f(\cdot; \theta_2)\ .$$

   Thus, a generic shared probe of length $2q - 2$ cannot identify $f(\cdot; \theta)$ from the input-output evaluations.

2. *There exists a real-analytic, non-polynomial and injective activation σ (namely, $\sigma = \tanh$) for which the following holds: $N \leq d_{\max} := m$. Then, for every $\mathcal{P} \in \mathcal{X}^N$,*

$$\exists\, \theta_1, \theta_2 \in \Theta_{\mathbf{D}},\ \text{such that } \Psi_{\text{all}}(\theta_1, \mathcal{P}) = \Psi_{\text{all}}(\theta_2, \mathcal{P}) \text{ and } f(\cdot; \theta_1) \not\equiv f(\cdot; \theta_2)\ .$$

   Thus, probes of length $N \leq d_{\max}$ cannot identify $f(\cdot; \theta)$ from the hidden neuron evaluations.

Proof. We start with two observations:

- To show that the lower bound of $2q - 1$ is necessary for the family of all MLPs with analytic activations, it is sufficient to come up with one analytic activation for which the $2q - 2$ probes don't suffice. We use $\sigma = \sin$ in part 1 and $\sigma = \tanh$ in part 2.
- We show the necessary lower bounds $2q-1$ or $d_{\max}+1$ using the layer widths $\mathbf{D} = (1, m, 1)$, but we show the tightness for any $m \in \mathbb{N}$. Now by taking $m \to \infty$ we can conclude the general lower bound of $2q - 1$ is tight.

With these arguments, we now prove the two constructive counter examples that establish the lower bounds.

We start with proving the first part. First we consider an interval containing $[0, 4\pi]$. We set $K := 2m$, $\epsilon := 1/(100m)$, and choose base parameters θ_1^0, θ_2^0 whose affine pre-activations are

$$g_1(t) := g(t; \theta_1^0) = \epsilon \sum_{k=1}^{m} \sin(kt),$$
$$g_2(t) := g(t; \theta_2^0) = \epsilon \sum_{k=m+1}^{K-1} \sin(kt) + \sin(Kt),$$
$$H(t) := g_1(t) - g_2(t).$$

Both networks have exactly m hidden neurons, positive distinct slopes, nonzero readout coefficients, and zero biases.

For $t_j = (j + 1/2)\pi/K$, $j = 0, \ldots, 2K - 1$, one has

$$H(t_j) = -(-1)^j + e_j,\ |e_j| \le (K - 1)\epsilon < 1/50.$$

The signs alternate at these points, including across the final interval to $t_0 + 2\pi$. The intermediate value theorem and Lemma 3 therefore give exactly $2K$ distinct zeros modulo 2π. We choose representatives $r_1, \ldots, r_{2K} \in [0, 2\pi)$ and form the $4K$-tuple $\mathcal{Z}$ of Lemma 4.

We consider the derivative, with respect to both parameter vectors, of the difference between the pre-activations. At the base pair it ranges over the functions $Q(t) + tR(t)$, $Q + tR \in \mathcal{V}_K$. This can be seen by considering the fact that, for a neuron with base slope k, base bias zero, and nonzero base readout coefficient a, the three derivatives are:

$$\partial_a \left(a \sin(wt + b)\right) = \sin(kt),\ \partial_b \left(a \sin(wt + b)\right) = a \cos(kt),\ \partial_w \left(a \sin(wt + b)\right) = at \cos(kt).$$

The two networks together contain each slope $k \in [K]$ exactly once; the difference of their output-bias variations supplies the constant function. Thus every coefficient of Q and R can be varied independently. By Lemma 4, the Jacobian of the pre-activation difference on $\mathcal{Z}$ has rank $3K = 2q - 2$. We choose $3K$ independent rows and let $\mathcal{P}_0$ be their probe locations.

We put $N_0 := 2q - 2$ and define

$$\mathcal{E}(\theta_1, \theta_2, \mathcal{P}) := \left(g(p_n; \theta_1) - g(p_n; \theta_2)\right)_{n=1}^{N_0}.$$

Then $\mathcal{E}(\theta_1^0, \theta_2^0, \mathcal{P}_0) = 0$, and its parameter Jacobian has an invertible $N_0 \times N_0$ minor. We take the remaining $2q - N_0 = 2$ parameter coordinates at their base values. The implicit function theorem gives parameter pairs $\theta_1(\mathcal{P}), \theta_2(\mathcal{P})$, continuous on a nonempty open neighborhood $\mathcal{U}_{N_0}$ of $\mathcal{P}_0$, such that $\mathcal{E}(\theta_1(\mathcal{P}), \theta_2(\mathcal{P}), \mathcal{P}) = 0\, \forall\, \mathcal{P} \in \mathcal{U}_{N_0}$. Consequently their final sine outputs agree at every probe.

Choose $t_* \in (0, 4\pi)$ with $H(t_*) \neq 0$. At the base pair, we have $|g_1(t)| \le m\epsilon < \pi/2$, $|g_2(t)| \le 1 + (m - 1)\epsilon < \pi/2$. The sine function is injective on $(-\pi/2, \pi/2)$, so $f(t_*; \theta_1^0) \neq f(t_*; \theta_2^0)$. We shrink the neighbourhood $\mathcal{U}_{N_0}$ so that this inequality persists, the probe entries remain distinct, and both parameter vectors lie in compact boxes around their base values with positive distinct slopes and nonzero readout coefficients. Their union is a compact set $\mathcal{K}$ with nonempty interior containing every constructed pair. Thus the final functions are different for every tuple in this open set.

For a general interval $\mathcal{X}$, choose $[a, b] \subset \mathcal{X}$, $a < b$, and $t = \alpha x + \beta$, $\alpha > 0$, mapping $[a, b]$ onto $[0, 4\pi]$. Replacing each hidden row (w_j, b_j) by $(\alpha w_j, \beta w_j + b_j)$ is an invertible linear change of its parameters. It transports the construction, its open probe set, and the compact parameter set into $\mathcal{X}$.

For $N < N_0$, let $\mathcal{U}_N$ be the projection of $\mathcal{U}_{N_0}$ onto the first N probe coordinates. This set is nonempty and open. Every tuple in $\mathcal{U}_N$ extends to a tuple in $\mathcal{U}_{N_0}$, whose colliding pair also agrees on the shorter tuple and belongs to the same $\mathcal{K}$. Since a nonempty open set has positive Lebesgue measure, generic shared separation is impossible for $N \le 2q - 2$.

Now we prove the second part. Fix any $\theta_0 = (w, b, a, c) \in \mathrm{int}(\Theta_{\mathbf{D}})$ such that $w_j \neq 0$ for every j and $b_j w_k - b_k w_j \neq 0$ whenever $j < k$. These conditions hold for Lebesgue-almost every interior parameter, because their failure is a finite union of zero sets of nonzero polynomials.

Fix $1 \le N \le m$ and an arbitrary probe tuple $\mathcal{P} \in \mathcal{X}^N$. The matrix

$$A := \overline{\mathbf{X}}_1^{\theta_0}(\mathcal{P}) = \begin{pmatrix} h_1(p_1;\theta_0) & \cdots & h_1(p_N;\theta_0) \\ \vdots & & \vdots \\ h_m(p_1;\theta_0) & \cdots & h_m(p_N;\theta_0) \\ 1 & \cdots & 1 \end{pmatrix} \in \mathbb{R}^{(m+1)\times N}$$

has more rows than columns. Hence there exists $\delta \in \mathbb{R}^{m+1} \setminus \{0\}$ with $\delta^\top A = 0$. Define $\theta(t)$ by keeping w, b fixed and replacing the last affine block by

$$B_1(t) := B_1^0 + t\delta^\top, \qquad B_1^0 = (a^\top \; c).$$

The hidden functions are unchanged, and

$$g(x;\theta(t)) - g(x;\theta_0) = t\left(\delta_{m+1} + \sum_{j=1}^{m} \delta_j h_j(x;\theta_0)\right).$$

Since $\delta^\top A = 0$, the right-hand side vanishes at every probe. Therefore the sampled final outputs also agree, and

$$\Psi_{\text{all}}(\theta(t), \mathcal{P}) = \Psi_{\text{all}}(\theta_0, \mathcal{P}).$$

It remains to show that the realized functions are different. By Lemma 5, the functions $1, h_1(\cdot;\theta_0), \ldots, h_m(\cdot;\theta_0)$ are linearly independent on $\mathcal{X}$. Since $\delta \neq 0$, the function

$$r(x) := \delta_{m+1} + \sum_{j=1}^{m} \delta_j h_j(x;\theta_0)$$

is not identically zero on $\mathcal{X}$. Consequently, for every $t \neq 0$, the output preactivations $g(\cdot;\theta(t))$ and $g(\cdot;\theta_0)$ differ somewhere on $\mathcal{X}$. Injectivity of the outer $\tanh$ therefore gives

$$f(\cdot;\theta(t)) \not\equiv f(\cdot;\theta_0) \quad \text{on } \mathcal{X}.$$

Finally, $\theta_0 \in \operatorname{int}(\Theta_{\mathbf{D}})$ and $\theta(t) \to \theta_0$. Thus all sufficiently small nonzero t give parameters in $\Theta_{\mathbf{D}}$ and in any prescribed neighborhood of θ_0. This proves that, for every probe tuple with $N \le m$, almost every interior target admits arbitrarily close, functionally distinct competitors with identical complete probe observations.

For this shallow $\tanh$ family, $d_{\max} = m$ hidden probes fail to identify almost every interior target, whereas the upper bound guarantees identification of almost every target from almost every tuple of $d_{\max} + 1$ probes. Thus the hidden-probe threshold is sharp in the worst case over the class of analytic, injective, non-polynomial activations.

For the $\sin$ family, the nonempty open set of failing probe tuples at $N \le 2q - 2$ rules out a general almost-everywhere-in-probes identification guarantee at those cardinalities. Hence the sufficient bound $2q + 1$ is within two probes of the lower bound $2q - 1$ for such a guarantee. We do not claim that these bounds are tight for every individual activation. □

B.3 ReLU Network Identification from Probes

We begin with a counterexample showing that ReLU networks cannot be exactly reconstructed from finitely many probes.

Proposition 11 (Finite probes do not identify ReLU networks)**.** *For any finite set of probes $p_1, \ldots, p_N \in [0,1]$, there exist two distinct ReLU networks, each with three hidden neurons, that agree on all the probes. Consequently, no finite set of probes can uniquely identify all ReLU networks.*

Proof. Consider hat functions of the form

$$g_{c,\epsilon}(x) := \mathrm{ReLU}(x - c + \epsilon) - 2\mathrm{ReLU}(x - c) + \mathrm{ReLU}(x - c - \epsilon) = \max\{\epsilon - |x - c|, 0\}.$$

For any c, ϵ, such a function is a ReLU network with three hidden neurons, and it is supported in $[c - \epsilon, c + \epsilon]$. Given any probes $p_1, \cdots, p_N$, choose $c \in (0,1)$ off the probes and $0 < \epsilon_1 < \epsilon_2$ below the distance from c to the probes and to the boundary; then g_{ϵ_1} and g_{ϵ_2} vanish at every probe and differ at c. Thus no finite set of probes, however chosen, can completely identify ReLU networks. □

The probabilistic setting of identification: Here we describe the probabilistic setting of function identification in ReLU Nets. Let μ be a Borel probability measure on $\mathcal{X}$. The probes are drawn from μ, $\mathcal{P} = (p_1, \cdots, p_N)$ with $p_1, \cdots, p_N$ i.i.d. μ. We also consider the same measure to quantify agreement between functions. For two parameters $\theta_0, \theta \in \mathbb{R}^{q_D}$ we define the disagreement sets:

$$\begin{aligned} \Delta_\mu(\theta, \theta_0) &:= \mu\{x \in \mathcal{X} : f(x;\theta) \neq f(x;\theta_0)\} \\ \Delta_\mu^{\text{all}}(\theta, \theta_0) &:= \mu\{x \in \mathcal{X} : x_i(x;\theta) \neq x_i(x;\theta_0) \text{ for some } i \in [L]\}, \end{aligned} \tag{8}$$

as the measure of the inputs at which the two networks compute different outputs. For uniform μ these are fractions of the volume of $\mathcal{X}$. Given thresholds ϵ, δ, we consider guarantees of the form: for every target θ_0, with probability at least $1 - \delta$ over the probes, every network $f(\cdot, \theta)$ agreeing on those probes also agrees with $f(\cdot, \theta_0)$ outside a set of measure at most ϵ. If we moreover assume that $\sup_{\theta\in\Theta, x\in\mathcal{X}} \|f(x;\theta)\| \leq B$ for some constant B, then we would like to show that one can reconstruct f_θ up to any ϵ in $L^s(\mu)$ norm. Note that such boundedness assumption is automatic for the hypothesis class of ReLU networks if we assume the input space $\mathcal{X}$ and the parameter space Θ to be compact.

Theorem 4 presents a semi-formal version of our results for ReLU networks, while Theorem 12 provides the formal statement. Intuitively, this theorem says the following. Since we cannot identify the function f_θ exactly everywhere from the probes, we want to identify f_θ from $\Psi_{\text{all}}(\theta, \mathcal{P})$ or $\Psi_{\text{final}}(\theta, \mathcal{P})$ at a large region of the input space $\mathcal{X}$, namely at a set of measure at least $1 - \epsilon$ (where $\mu(\mathcal{X}) = 1$) by using N many probes. Also assuming that the probes are sampled independently at random under the same probability measure μ, we want the above property to hold with high probability (at least $1 - \delta$). Thus, we want the target function f_θ and the reconstructed function $f_{\hat{\theta}}$ to be approximately equal with high probability. This is similar in spirit (but unrelated) to the PAC learning framework used in classical learning theory. Now, we prove Theorem 12 on ReLU networks, which is a formal version of Theorem 4. For that we first define some notations and prove a couple of helper lemmas.

Proof of the formal version of Theorem 4: We first state and prove some bounds on counting polynomials and the random-sampling bound used in the proof. The zero-pattern estimate below is a specialization of Rónyai et al. (2001, Theorem 1.1). We still prove it for completeness.

Lemma 6 (Counting polynomial tests)**.** *Let $g_1, \cdots, g_J$ be real polynomials of degree at most r in P variables, where $P, r, J \geq 1$, and we define $D := \binom{P+r}{r} - 1$. The numbers of distinct vectors $(\mathbf{1}[g_j(v) > 0])_{j=1}^J$, $(\mathbf{1}[g_j(v) \neq 0])_{j=1}^J$, $v \in \mathbb{R}^P$, are at most $\sum_{k=0}^D \binom{J}{k}$ and $\binom{P+rJ}{P}$, respectively. If $J \geq D$, the first bound is at most $(eJ/D)^D$.*

Proof. We first bound the number of positive vectors. For this, we choose one parameter for each realized vector. Subtract from every g_j a positive number smaller than all its positive values at the chosen parameters. If none of these values is positive, any positive number works. This preserves every chosen binary vector and makes all tested values nonzero.

List the D nonconstant monomials of degree at most r. Each shifted polynomial is an affine function of this list. It remains to count strict sign vectors of J affine functions on $\mathbb{R}^D$. A nonempty set with fixed strict signs is an intersection of open halfspaces and is convex. Adding one zero hyperplane splits such a set only when the hyperplane meets it. The intersections of the split sets with this hyperplane have distinct strict sign vectors for the preceding functions restricted to $\mathbb{R}^{D-1}$. A repeated hyperplane causes no split, and constant functions add no choices. Thus the maximum count $a(J, D)$ satisfies

$$a(J, D) \leq a(J-1, D) + a(J-1, D-1).$$

Starting from $a(0, D) = a(J, 0) = 1$, induction gives $a(J, D) \leq \sum_{k=0}^D \binom{J}{k}$. For $J \geq D$, we set $t = D/J$ and we conclude by the following bounding:

$$t^D \sum_{k=0}^{D} \binom{J}{k} \leq \sum_{k=0}^{D} \binom{J}{k} t^k \leq (1+t)^J \leq e^D.$$

Now we bound the number of non-zero vectors. Let $\mathcal{S}$ contain the realized sets $\{j : g_j(v) \neq 0\}$. We choose a realizing parameter v_S for each $S \in \mathcal{S}$, and define $G_S := \prod_{j\in S} g_j$, with empty product 1.

Then $G_S(v_T) \neq 0$ exactly when $S \subseteq T$. Ordering the sets by nondecreasing cardinality makes the matrix $(G_S(v_T))_{S,T\in\mathcal{S}}$ upper triangular with nonzero diagonal. The polynomials G_S are therefore linearly independent. Each has degree at most rJ, and the space of such polynomials has $\binom{P+rJ}{P}$ monomials as a basis. This proves the second bound. □

For a class $\mathcal{C}$ of subsets of $\mathcal{X}$, write $\Pi_{\mathcal{C}}(M)$ for the largest number of distinct vectors $(\mathbf{1}[p_a \in C])_{a=1}^{M}$, as $C \in \mathcal{C}$ varies, on any M fixed inputs. This is the growth function of $\mathcal{C}$ from VC theory. Moreover, the following lemma is inspired from the classical double-sampling argument for random ϵ-nets from (Haussler & Welzl, 1987; Blumer et al., 1989).

We also use the following consequence of the polynomial sign bound of Bartlett et al. (2019, Lemma 17).

Lemma 7 (Counting polynomial inequalities)**.** *Let $g_1, \cdots, g_J$ be real polynomials of degree at most r in P variables, where $r \geq 1$ and $J \geq P \geq 1$. Then*

$$\left|\left\{(\mathbf{1}[g_j(v) > 0])_{j=1}^{J} : v \in \mathbb{R}^P\right\}\right| \leq \left(\frac{4erJ}{P}\right)^P.$$

Proof. Choose one parameter for each realized binary vector. For each j, choose $\eta_j > 0$ smaller than every positive value of g_j at these finitely many parameters; if there is no positive value, choose any $\eta_j > 0$. Replacing g_j by $g_j - \eta_j$ preserves the chosen binary vectors and makes every tested value nonzero. The polynomial sign bound therefore gives at most

$$2\left(\frac{2erJ}{P}\right)^P \leq \left(\frac{4erJ}{P}\right)^P$$

such vectors. □

Lemma 8 (Random samples detect large disagreement sets)**.** *Let $K, \mathcal{X}$ be compact metric spaces, let μ be a Borel probability measure on $\mathcal{X}$, and let $h : K \times \mathcal{X} \to [0, \infty)$ be continuous. Set $C_\theta = \{u : h(\theta, u) > 0\}$ and $\mathcal{C} = \{C_\theta : \theta \in K\}$. Suppose that $D, Q \geq 1$ are integers and*

$$\Pi_{\mathcal{C}}(M) \leq \left(\frac{eMQ}{D}\right)^D, \; M \geq D.$$

For $\epsilon, \delta \in (0, 1)$, if

$$N \geq \frac{6}{\epsilon}\left(D \log \frac{12eQ}{\epsilon} + \log \frac{2}{\delta}\right),$$

then, with probability at least $1 - \delta$, every $C \in \mathcal{C}$ missed by N i.i.d. samples from μ has $\mu(C) \leq \epsilon$.

Proof. Let E be the event that the sample misses some C with $\mu(C) > \epsilon$. Draw an independent second sample of size N. Let E' be the event that some $C \in \mathcal{C}$ contains no first-sample point and at least $N\epsilon/2$ second-sample points.

These events are measurable. To see this, we observe that the map $\theta \mapsto \mu(C_\theta)$ is lower semicontinuous by Fatou's lemma. In $K \times \mathcal{X}^N$ or $K \times \mathcal{X}^{2N}$, the no-hit conditions are closed, whereas the conditions $\mu(C_\theta) > \epsilon$ and the required number of second-sample hits are open. The defining sets are therefore intersections of a closed set with an open set in a compact metric space. Such sets are countable unions of compact sets, as are their coordinate projections onto the samples. Thus E, E' are Borel.

Suppose the first sample belongs to E, and choose a witnessing set of measure $q > \epsilon$. Its second-sample hit count Z has mean Nq and variance $Nq(1-q)$. If $N\epsilon \geq 8$, by Chebyshev's inequality we get

$$\mathbb{P}\{Z < N\epsilon/2\} \leq \frac{4(1-q)}{Nq} < \frac{1}{2}.$$

Consequently we have $\mathbb{P}(E) \leq 2\mathbb{P}(E')$.

Now we pair the two samples and independently swap the members of each pair with probability $1/2$. Their joint distribution is unchanged. On the fixed $2N$ points there are at most $\Pi_{\mathcal{C}}(2N)$ membership

vectors. For a fixed vector to witness E', all its marked points must be assigned to the second sample. If a pair has two marked points this is impossible; otherwise, with k marked points, its probability is 2^{-k}. Since $k \geq N\epsilon/2$, a union bound gives $\mathbb{P}(E) \leq 2\Pi_{\mathcal{C}}(2N)2^{-N\epsilon/2}$.

It remains to check the stated threshold. Put $A = D\log(12eQ/\epsilon) + \log(2/\delta)$ and $N_0 = 6A/\epsilon$. Then $N_0 > D$, $N_0\epsilon > 8$, and

$$D\log\frac{2eN_0Q}{D} + \log\frac{2}{\delta} = A + D\log\frac{A}{D} \leq 2A = \frac{N_0\epsilon}{3}.$$

The function $N\epsilon - 3D\log(2eNQ/D) - 3\log(2/\delta)$ has derivative $\epsilon - 3D/N \geq 0$ for $N \geq N_0$. Hence, for every such N,

$$D\log\frac{2eNQ}{D} + \log\frac{2}{\delta} \leq \frac{N\epsilon}{3}.$$

Applying the growth bound at $M = 2N$ and using $3\log 2/2 > 1$, we conclude that

$$\mathbb{P}(E) \leq 2\exp\left(D\log\frac{2eNQ}{D} - \frac{\log 2}{2}N\epsilon\right) \leq \delta.$$

□

We shall now state and prove Theorem 12, which is the formal version of Theorem 4. Throughout this section, $\mathcal{X} \subseteq \mathbb{R}^d$ is compact, $L \geq 2$, $\sigma = \mathrm{ReLU}$ in layers $1, \cdots, L-1$ and the output layer is linear. We use the following notations for ReLU nets, similar to what we used for analytic nets:

$$\begin{aligned} x_0(u;\theta) &:= u, \\ \overline{x}_i(u;\theta) &:= (x_i(u;\theta);1) \in \mathbb{R}^{d_i+1},\ 0 \leq i < L, \\ z_i(u;\theta) &:= B_i\overline{x}_{i-1}(u;\theta) \in \mathbb{R}^{d_i},\ i \in [L], \\ x_i(u;\theta) &:= \mathrm{ReLU}(z_i(u;\theta)),\ i \in [L-1], \\ x_L(u;\theta) &:= f(u;\theta) := z_L(u;\theta), \\ H &:= \sum_{i=1}^{L-1} d_i,\ \ell := H + m. \end{aligned} \tag{9}$$

We denote by $\beta_{ij}^\top$ the j-th row of B_i. We call the rows (i,j) with $i \leq L-1$ the hidden rows. H is thus the total number of hidden neurons.

Theorem 12 (PAC identification of ReLU networks)**.** *Let $\mathcal{X} \subset \mathbb{R}^d$ be compact and μ be a Borel probability measure on $\mathcal{X}$. Consider the family of* ReLU *MLPs $\mathcal{F}_{\Theta_{\mathbf{D}}}$ of ℓ neurons with widths $\mathbf{D} := (d, d_1, \cdots, d_{L-1}, m)$ with compact parameter space $\Theta_{\mathbf{D}} \subseteq \mathbb{R}^{q_{\mathbf{D}}}$. We consider any $\epsilon, \delta \in (0,1)$ and integers:*

$$N_1 \geq \tfrac{6}{\epsilon}\left(q_{\mathbf{D}}L\log\tfrac{48eL(\ell-m+1)}{\epsilon} + \log\tfrac{2}{\delta}\right),\ N_2 \geq \tfrac{6}{\epsilon}\left(q_{\mathbf{D}}\log\tfrac{24e\ell}{\epsilon} + \log\tfrac{2}{\delta}\right).$$

Let $\mathcal{P}_1 = (p_1, \cdots, p_{N_1}) \in \mathcal{X}^{N_1}, \mathcal{P}_2 = (p_1, \cdots, p_{N_2}) \in \mathcal{X}^{N_2}$ be probes with p_i i.i.d. μ. We fix the target $\theta_0 \in \Theta_{\mathbf{D}}$. Then we have:

(i) $\mathbb{P}_{\mathcal{P}_1\sim\mu^{N_1}}\left\{\forall\theta \in \Theta_{\mathbf{D}},\ \Psi_{final}(\theta,\mathcal{P}_1) = \Psi_{final}(\theta_0,\mathcal{P}_1) \implies \Delta_\mu(\theta,\theta_0) \leq \epsilon\right\} \geq 1-\delta$, i.e, with probability at least $1-\delta$, any parameter $\theta \in \Theta_{\mathbf{D}}$ having the same final output on the probe tuple $\mathcal{P}_1$ as the target θ_0 has disagreement set of measure at most ϵ.

(ii) $\mathbb{P}_{\mathcal{P}_2\sim\mu^{N_2}}\left\{\forall\theta \in \Theta_{\mathbf{D}},\ \Psi_{all}(\theta,\mathcal{P}_2) = \Psi_{all}(\theta_0,\mathcal{P}_2) \implies \Delta_\mu^{\mathrm{all}}(\theta,\theta_0) \leq \epsilon\right\} \geq 1-\delta$ i.e, with probability at least $1-\delta$, any parameter $\theta \in \Theta_{\mathbf{D}}$ having the same hidden activations on the probe tuple $\mathcal{P}_2$ as θ_0 can disagree on any hidden layer neuron on an input set of measure at most ϵ.

Moreover, let $B > 0$ be such that $\sup_{\theta\in\Theta_{\mathbf{D}},\ x\in\mathcal{X}}\|f(x;\theta)\|_2 \leq B$. Then we also have (a) $\mathbb{P}_{\mathcal{P}_1\sim\mu^{N_1}}\left\{\forall\theta \in \Theta_{\mathbf{D}},\ \Psi_{final}(\theta,\mathcal{P}_1) = \Psi_{final}(\theta_0,\mathcal{P}_1) \implies \|f_\theta - f_{\theta_0}\|_{L^2(\mu)} \leq 2B\sqrt{\epsilon}\right\} \geq 1-\delta$ and (b) $\mathbb{P}_{\mathcal{P}_2\sim\mu^{N_2}}\left\{\forall\theta \in \Theta_{\mathbf{D}},\ \Psi_{all}(\theta,\mathcal{P}_2) = \Psi_{all}(\theta_0,\mathcal{P}_2) \implies \|f_\theta - f_{\theta_0}\|_{L^2(\mu)} \leq 2B\sqrt{\epsilon}\right\} \geq 1-\delta$

Proof. Write $p = q_{\mathbf{D}}$, and we recall the definition of $x_i, \overline{x}_i, z_i, H$ and the details of the ReLU net structure from equation 9. We use the following notations:

$$P_i := \sum_{k=1}^{i} d_k(d_{k-1}+1),\ i \in [L],\ D_{\text{final}} := \sum_{i=1}^{L} P_i,$$
$$Q_{\text{final}} := 4\sum_{i=1}^{L-1} i\, d_i + 4L,\ D_{\text{all}} := p,\ Q_{\text{all}} := 2\ell.$$

Here P_i is the number of parameters in the first i layers, and $P_L = p$. We fix a parameter θ_0 and define

$$C_\theta^{\text{final}} = \{u : f(u;\theta) \neq f(u;\theta_0)\},\ C_\theta^{\text{all}} = \{u : x_i(u;\theta) \neq x_i(u;\theta_0) \text{ for some } i \in [L]\}.$$

Let $\mathcal{C}^\star$ consist of these sets as $\theta \in \Theta_{\mathbf{D}}$ varies. We shall show that, for $\star \in \{\text{final}, \text{all}\}$, we have $\Pi_{\mathcal{C}^\star}(M) \leq \left(\frac{eMQ_\star}{D_\star}\right)^{D_\star}$, $M \geq D_\star$, where $\Pi_{\mathcal{C}}(M)$ denotes the largest number of distinct vectors $(\mathbf{1}[p_a \in C])_{a=1}^M$, by varying $C \in \mathcal{C}$ on any M fixed inputs.

Fix $u \in \mathcal{X}$. By (9), equality with the target at every layer is equivalent to the following 3 conditions: (i) $\beta_{ij}^\top \overline{x}_{i-1}(u;\theta_0) = x_{i,j}(u;\theta_0)$, $(i < L,\ x_{i,j}(u;\theta_0) > 0)$, (ii) $\beta_{ij}^\top \overline{x}_{i-1}(u;\theta_0) \leq 0$, $(i < L,\ x_{i,j}(u;\theta_0) = 0)$ and (iii) $\beta_{Lj}^\top \overline{x}_{L-1}(u;\theta_0) = f_j(u;\theta_0)$, $(j \in [m])$, where the first two conditions range over all hidden rows. Necessity follows directly from the network equations. Conversely, starting from the common input u, these conditions imply equality of successive states by induction.

For fixed θ_0 and u, these are affine conditions in θ. Each equality $g = 0$ is determined by the two tests $g > 0$ and $-g > 0$, and each inequality $g \leq 0$ by the test $g > 0$. Thus at most 2ℓ affine tests per input suffice. On M inputs, Lemma 6 with $P = p$ and $r = 1$ gives $\Pi_{\mathcal{C}^{\text{all}}}(M) \leq \left(\frac{2eM\ell}{p}\right)^p$, $M \geq p$.

Now consider final-output observations. We count the hidden activation patterns successively over the layers, as in Bartlett et al. (2019, Section 4). Fix $M \geq D_{\text{final}}$ inputs $p_1, \cdots, p_M$. For $i = 0, \cdots, L-1$, let $\mathcal{S}_i$ be the partition of $\Theta_{\mathbf{D}}$ obtained by grouping parameters according to the bits

$$\mathbf{1}[z_{k,j}(p_a;\theta) > 0],\ k \in [i], \quad j \in [d_k], \quad a \in [M].$$

Only nonempty parts are retained, and $\mathcal{S}_0 = \{\Theta_{\mathbf{D}}\}$.

Fix $i \in [L-1]$ and $S \in \mathcal{S}_{i-1}$. On S, every preceding ReLU equals multiplication by its fixed activation bit at each sampled input. Consequently, each $z_{i,j}(p_a;\theta)$ agrees on S with a polynomial of degree at most i in the first P_i parameters. Indeed, first-layer preactivations are affine, and each subsequent affine layer increases the degree by at most one.

There are Md_i such polynomials. Since $Md_i \geq M \geq D_{\text{final}} \geq P_i$, Lemma 7 shows that S is divided into at most $(4eiMd_i/P_i)^{P_i}$ parts of $\mathcal{S}_i$. Hence

$$|\mathcal{S}_i| \leq |\mathcal{S}_{i-1}| \left(\frac{4eiMd_i}{P_i}\right)^{P_i},$$
$$|\mathcal{S}_{L-1}| \leq \prod_{i=1}^{L-1} \left(\frac{4eiMd_i}{P_i}\right)^{P_i}.$$

Zero preactivations are assigned bit zero, so these partitions also include parameters at which some preactivation vanishes.

Fix $S \in \mathcal{S}_{L-1}$. For each $a \in [M]$ and $j \in [m]$, let $f_{a,j}^S(\theta)$ be the polynomial obtained by replacing every hidden ReLU at p_a by multiplication by its bit on S. It has degree at most L and equals $f_j(p_a;\theta)$ for $\theta \in S$. Define

$$R_{a,S}(\theta) := \sum_{j=1}^{m} \left(f_{a,j}^S(\theta) - f_j(p_a;\theta_0)\right)^2.$$

This polynomial has degree at most $2L$, and, for $\theta \in S$, $R_{a,S}(\theta) \neq 0 \iff f(p_a;\theta) \neq f(p_a;\theta_0)$. The second part of Lemma 6 therefore bounds the number of disagreement vectors on S by

$$\binom{p+2LM}{p} \leq \left(\frac{e(p+2LM)}{p}\right)^p \leq \left(\frac{4eLM}{p}\right)^p,$$

where we used $M \geq p$.

Summing over $S \in \mathcal{S}_{L-1}$ and applying the weighted arithmetic–geometric mean inequality, with weights $P_1/D_{\text{final}}, \cdots, P_{L-1}/D_{\text{final}}, p/D_{\text{final}}$, gives

$$\begin{aligned}\Pi_{\mathcal{C}^{\text{final}}}(M) &\leq \left[\prod_{i=1}^{L-1}\left(\frac{4eiMd_i}{P_i}\right)^{P_i}\right]\left(\frac{4eLM}{p}\right)^p \\ &\leq \left(\frac{eM\left(4\sum_{i=1}^{L-1} i\, d_i + 4L\right)}{\sum_{i=1}^{L-1} P_i + p}\right)^{\sum_{i=1}^{L-1} P_i + p} \\ &= \left(\frac{eMQ_{\text{final}}}{D_{\text{final}}}\right)^{D_{\text{final}}}.\end{aligned}$$

Moreover, we have $D_{\text{final}} \leq pL$, $Q_{\text{final}} \leq 4L(H+1)$.

We now apply Lemma 8. Set $K = \Theta_{\mathbf{D}}$ and define:

$$h_{\text{final}}(\theta, u) := \|f(u;\theta) - f(u;\theta_0)\|_2,\ h_{\text{all}}(\theta, u) := \max_{i\in[L]} \|x_i(u;\theta) - x_i(u;\theta_0)\|_2.$$

Both functions are continuous, and their positive sets are C_θ^{final} and C_θ^{all}, respectively.

For part (i), the preceding bounds give

$$\frac{6}{\epsilon}\left(D_{\text{final}} \log \frac{12eQ_{\text{final}}}{\epsilon} + \log\frac{2}{\delta}\right) \leq \frac{6}{\epsilon}\left(pL \log \frac{48eL(H+1)}{\epsilon} + \log\frac{2}{\delta}\right) \leq N_1.$$

Lemma 8 therefore supplies an event $\mathcal{G}_1 \subseteq \mathcal{X}^{N_1}$ of probability at least $1-\delta$ on which every $C \in \mathcal{C}^{\text{final}}$ missed by the probes has measure at most ϵ. For every $\mathcal{P}_1 \in \mathcal{G}_1$ and every $\theta \in \Theta_{\mathbf{D}}$,

$$\Psi_{\text{final}}(\theta, \mathcal{P}_1) = \Psi_{\text{final}}(\theta_0, \mathcal{P}_1) \implies C_\theta^{\text{final}} \cap \{p_1, \cdots, p_{N_1}\} = \varnothing \implies \Delta_\mu(\theta, \theta_0) \leq \epsilon.$$

This proves (i).

For part (ii), apply the same lemma with $D = D_{\text{all}} = p$ and $Q = Q_{\text{all}} = 2\ell$. The assumed bound on N_2 is its required threshold. Thus, on an event $\mathcal{G}_2 \subseteq \mathcal{X}^{N_2}$ of probability at least $1-\delta$, every $\theta \in \Theta_{\mathbf{D}}$ satisfies

$$\Psi_{\text{all}}(\theta, \mathcal{P}_2) = \Psi_{\text{all}}(\theta_0, \mathcal{P}_2) \implies C_\theta^{\text{all}} \cap \{p_1, \cdots, p_{N_2}\} = \varnothing \implies \Delta_\mu^{\text{all}}(\theta, \theta_0) \leq \epsilon.$$

Since $C_\theta^{\text{final}} \subseteq C_\theta^{\text{all}}$, the final-output disagreement is also at most ϵ. This proves (ii).

Finally, if the outputs are bounded by B, then, for every $1 \leq s < \infty$, we have:

$$\|f_\theta - f_{\theta_0}\|_{L^s(\mu)}^s \leq (2B)^s \mu(C_\theta^{\text{final}}) \leq (2B)^s \epsilon.$$

Taking $s = 2$ gives the claimed $2B\sqrt{\epsilon}$ bound. □

B.4 Function-level identification of multi-head attention

We use the architecture, parameter spaces, and probe maps defined in equations 2–(3). In particular, Θ is the raw parameter space and $D = \dim \Theta$. The *canonical representation* of a layer is obtained by merging heads with identical score matrices $\mathbf{M}_a$, summing their value matrices $\mathbf{W}_a$, and removing groups whose summed value matrix is zero.

We start the section with a brief discussion on related works and then prove our results.

Algorithm 2 RecoverMHA for $h = L = 1$

Require: oracle $F_n(\cdot;\theta_0)$; $\operatorname{rank} \mathbf{W}_0 = d$
Ensure: effective matrices $(\mathbf{M}_0, \mathbf{W}_0)$
1: query $y_i \leftarrow F_1(e_i^\top;\theta_0)$, $i = 1, \ldots, d$
2: $\mathbf{W} \leftarrow [y_1; \ldots; y_d]$ ▷ Value matrix
3: $P \leftarrow [0^\top; e_1^\top; \ldots; e_d^\top]$
4: query $Y \leftarrow F_{d+1}(P;\theta_0)$
5: $Z \leftarrow P\mathbf{W}$
6: $A \leftarrow [Y\ \mathbf{1}_{d+1}][Z\ \mathbf{1}_{d+1}]^{-1}$ ▷ Attn. wts.
7: $\mathbf{M}_{ij} \leftarrow \log(A_{i+1,j+1}/A_{i+1,1})$, $i, j \in [d]$
8: **return** $(\mathbf{M}, \mathbf{W})$

Relation to query-based attention recovery. Query based recovery of attention blocks has been studied recently and we acknowledge the progress of those works and discuss what we contribute. Bhattamishra et al. (2026) study recovery of single-head attention from scalar-output queries, while Kim & Kim (2026) recover canonical multi-head representations.

Our three-token separation argument is a matrix-output adaptation of the canonical-identifiability result (Kim & Kim, 2026, Proposition 3.1).

However, our main focus is generic identification of multilayer MHA stacks from a shared, nonadaptive probe probes with hidden access. We also consider different observation models: our queries return complete tokenwise layer outputs for Ψ_{all}, whereas their attention oracles return a scalar at the final token only.

Softmax attention setup: Now we setup some detailed notation, some of which overlap with the same from main text, but in more details.

We begin with some definitions. We consider d dimensional inputs and h attention heads, with $h \mid d$. Set $r = d/h$, the value dimension of each head, and let k be the query/key dimension with $1 \le k \le d$. A bias-free, unmasked MHA layer has parameters $\theta := \{(\mathbf{Q}_a, \mathbf{K}_a, \mathbf{V}_a, \mathbf{O}_a)\}_{a=1}^{h}$, $\mathbf{Q}_a, \mathbf{K}_a \in \mathbb{R}^{d\times k}$, $\mathbf{V}_a \in \mathbb{R}^{d\times r}$, $\mathbf{O}_a \in \mathbb{R}^{r\times d}$. For an input probe $P \in \mathbb{R}^{n\times d}$ whose rows are the token vectors $p_1^\top, \dots, p_n^\top$, the layer computes

$$F_n(P;\theta) = \sum_{a=1}^{h} \mathrm{softmax}(P\mathbf{M}_a P^\top)\, P\mathbf{W}_a,\ \mathbf{M}_a := \frac{\mathbf{Q}_a\mathbf{K}_a^\top}{\sqrt{k}},\ \mathbf{W}_a := \mathbf{V}_a\mathbf{O}_a.$$

The softmax acts rowwise: $\big[\mathrm{softmax}(P\mathbf{M}_a P^\top)\big]_{ij} = \frac{\exp(p_i^\top \mathbf{M}_a p_j)}{\sum_{s=1}^{n}\exp(p_i^\top \mathbf{M}_a p_s)}$. The matrices $\mathbf{M}_a, \mathbf{W}_a \in \mathbb{R}^{d\times d}$ thus have ranks at most k and r, respectively. A probe specifies P and returns the complete matrix $F_n(P;\theta) \in \mathbb{R}^{n\times d}$. An L-layer MHA stack has parameters $\boldsymbol{\theta} = (\theta_1, \dots, \theta_L)$ and we consider the case where the token dimension d is preserved across layers. Layer ℓ has h_ℓ heads, query/key dimension k_ℓ, and value dimension $r_\ell = d/h_\ell$. Our proofs only need manifold doms of parameters.

Let $\Theta = \prod_{\ell=1}^{L} \Theta_\ell$ denote the parameter space of the L-layer MHA architecture, with Θ_ℓ being the parameter space of the ℓ-th layer. Let $q_\ell := h_\ell\left(k_\ell(2d-k_\ell) + r_\ell(2d-r_\ell)\right)$, $q := \sum_{\ell=1}^{L} q_\ell$. Then $q_\ell = \dim(\Theta_\ell)$ since the set of $d\times d$ matrices of rank at most k is an irreducible algebraic variety of dimension $k(2d-k)$. Thus, $\dim(\Theta_\ell) = q_\ell$, $\dim(\Theta) = q$.

Proof of Stronger version of Theorem 5 We use two elementary facts: a discrete subset of a Euclidean space is countable, and a nonzero analytic function on a connected open subset of $\mathbb{R}^m$ has a zero set of Hausdorff dimension at most $m-1$.

Lemma 9 (Three-token separation)**.** *Two MHA layers agree on every three-token input if and only if they agree at every sequence length.*

Proof. Define $\rho : \mathbb{R} \to \mathbb{R} : \rho(t) = e^t/(2+e^t)$. It is not hard to see that the functions $\rho(gt)$ for distinct "rates" $g \in \mathbb{R}$ are linearly independent.

Now, suppose the three-token maps agree. Equal-token inputs give equal value sums. List their distinct score matrices as B_j, and let Δ_j be the difference between the associated value sums. The repeated-token row of $(q+tv, q, q)$ then gives $\sum_j \rho(tq^\top B_j v)\, v^\top \Delta_j = 0$. For generic (q, v) the rates are distinct, since $q^\top(B_j - B_{j'})v$ is a nonzero polynomial for $j \ne j'$. Linear independence gives $v^\top \Delta_j = 0$, hence $\Delta_j = 0$. The converse, at every length, follows by merging the summands. □

We make a note that the above is the matrix-output counterpart of Kim & Kim (2026, Proposition 3.1).

Lemma 10 (Countably many candidates from a fixed probe batch)**.** *Let $H \ge 2$, $H \mid d$, and $s = d/H$. Consider probe tuples containing $d(2H+1)$ probes: d singletons, $d(H+1)$ triples, and $d(H-1)$ sequences of length $s+1$. Let $\mathbf{P}_0$ be the set of all such probes.*

Fix an MHA layer with $h \le H$ heads, $h \mid d$, query/key dimension $1 \le k \le d$, and value dimension $r = d/h$. Its effective parameter space is

$$\Theta_{\text{eff}} := \left\{\theta_{\text{eff}} = ((\mathbf{M}_a, \mathbf{W}_a))_{a=1}^{h} : \operatorname{rank} \mathbf{M}_a \le k,\ \operatorname{rank} \mathbf{W}_a \le r\right\},$$

with

$$q := \dim \Theta_{\text{eff}} = h\big[k(2d-k) + r(2d-r)\big].$$

Then, for almost all tuples $(\mathcal{P}, \theta) \in \mathbf{P}_0 \times \Theta_{\text{eff}} \setminus \mathcal{N}$ (with $\dim_{\mathrm{H}} \mathcal{N} \le \dim \mathbf{P}_0 + q - 1$) and almost all $\theta \in \theta_{\text{eff}}$ the set $\{\theta' \in \Theta_{\text{eff}} : \Psi_{\textit{final}}(\theta', \mathcal{P}) = \Psi_{\textit{final}}(\theta, \mathcal{P})\}$ is at most countable.

Proof. We first choose parameters and probes for which no nonzero first-order change in the parameters leaves all outputs unchanged. We then show that this property holds almost everywhere and implies the claimed countability.

Let $V_{ij} = \alpha_i^{j-1}$, $0 < \alpha_1 < \cdots < \alpha_d$, and let P_a keep the ath consecutive coordinate block of size r and set the other coordinates to zero. Choose $\mathbf{W}_a = VP_a$, $\mathbf{M}_a = c_a M_*$, $M_* = e_1 \mathbf{1}^\top + \sum_{j=2}^{k} e_j e_j^\top$, where $0 < c_1 < \cdots < c_h$. Partition $[d]$ into H blocks J of size s, and set $u_j = V_{:,j}$ for $j \le d - s$. Use the probes e_i, $(e_1 + t_\nu e_i, e_1, e_1)$, $(u_j, (u_j + e_b)_{b \in J})$, for $i \in [d]$, $\nu \in [H+1]$, and $j \in [d-s]$, stacking the listed tokens as rows. Their counts are exactly those specified by $\mathbf{P}_0$.

We will repeatedly use the following elementary fact: a nonzero polynomial of degree below q cannot vanish at q distinct points. Consequently, the matrix $(\alpha_i^{j-1})_{i \in I,\, j \in [q]}$ is invertible whenever $|I| = q$.

We start with showing that the singleton and triple outputs constrain the value matrices. Write $\rho(t) = e^t/(2 + e^t)$. Singletons determine $S = \sum_a \mathbf{W}_a$. For a fixed direction e_i, subtract $e_1^\top S$ from a repeated output row of the triple and divide by t_ν. The result is

$$Y_\nu = \sum_a \rho(t_\nu g_a) b_a^\top,\ g_a = e_1^\top \mathbf{M}_a e_i = c_a,\ b_a = \mathbf{W}_a^\top e_i.$$

We define:

$$C_{\nu a} = \rho(t_\nu c_a),\ B = (b_a^\top)_{a=1}^{h},\ (D_a)_\nu = t_\nu \rho'(t_\nu c_a).$$

Each row of B is nonzero and occupies a different coordinate block, so its rows are independent.

Choose the positive scales so that $\operatorname{rank}[C\ D_a] = h+1$ for every a. Here is a real-variable justification that such scales exist. For $c > 0$, as $t \to \infty$, we get $\rho(ct) - 1 = -2e^{-ct} + O(e^{-2ct})$, $t\rho'(ct) = 2te^{-ct} + O(te^{-2ct})$. Suppose $\sum_b \beta_b \rho(c_b t) + \gamma t \rho'(c_a t) = 0$. Taking the limit gives $\sum_b \beta_b = 0$. After subtracting this constant, consider the smallest rate with a nonzero coefficient. Dividing by its exponential leaves a nonzero constant or a nonzero linear function of t, plus a term tending to zero. This is impossible. Thus the $h+1$ functions are independent. Their evaluation matrices therefore have full rank for almost every choice of the scales, simultaneously for all a.

Now write $\delta\mathbf{M}_a, \delta\mathbf{W}_a$ for first-order parameter changes, and suppose every observed output has zero first-order change. Then $\delta S = 0$, $C\delta B + \sum_a D_a\, \delta g_a\, b_a^\top = 0$. Taking row combinations that cancel C, and multiplying by a right inverse of B, gives $\delta g_a = 0$ for every a, because D_a is not a linear combination of the columns of C. Then $C\delta B = 0$, so $\delta B = 0$. Doing this for every i gives $\delta\mathbf{W}_a = 0$.

Now we show that the remaining outputs constrain the score matrices corresponding to the softmax attention. Since $\mathbf{W}_b P_a = \mathbf{1}_{a=b}\mathbf{W}_a$, multiplying the output on the right by P_a isolates head a. For a block input Z, we therefore obtain $\delta A_a\, Z\mathbf{W}_a = 0$, $A_a = \operatorname{softmax}(Z\mathbf{M}_a Z^\top)$. Every s selected rows of $\mathbf{W}_a$ are independent. Indeed, after removing nonzero row factors, their first s nonzero columns form the polynomial-evaluation matrix described above. Hence the rows of $Z\mathbf{W}_a$ are affinely independent. Since also $\delta A_a \mathbf{1} = 0$, it follows that $\delta A_a = 0$.

Differentiating the attention log-ratios gives

$$u_j^\top \delta\mathbf{M}_a e_b = 0,\ e_{b'}^\top \delta\mathbf{M}_a e_b = 0\ (j \le d - s,\ b, b' \in J).$$

For each J, the vectors $(u_j)_{j \le d-s}$ and $(e_b)_{b \in J}$ span $\mathbb{R}^d$: deleting rows J from the first $d - s$ columns of V leaves an invertible polynomial-evaluation matrix. Thus $\delta\mathbf{M}_a e_b = 0$ for every b and every a. We have proved that all first-order parameter changes vanish.

Now we use this example to conclude the result for almost every parameter and probe tuple. A rank-t matrix can locally be written as

$$\begin{pmatrix} U_{11} & U_{12} \\ U_{21} & U_{21}U_{11}^{-1}U_{12} \end{pmatrix}, \det U_{11} \neq 0.$$

Its three displayed blocks provide $t(2d-t)$ independent coordinates. Thus the effective parameters with ranks exactly k and r have q independent coordinates; smaller ranks require fewer coordinates.

The previous discussion shows that the matrix of first derivatives of the outputs has rank q at our chosen parameters and probes. The same is true when differentiating with respect to the raw factors: full-rank factors allow every first-order change in the independent matrix coordinates. Some square submatrix of these derivatives therefore has a nonzero determinant. This determinant is analytic on the connected raw parameter and probe space, so its zero set has measure zero.

Passing to the independent matrix coordinates preserves this almost-everywhere conclusion. We note that full-rank factorizations are locally described by these coordinates together with an arbitrary invertible change of basis in the factors. Fubini applied to these extra coordinates gives the assertion.

Now, finally we bound the exceptional set. Record both the probes and their outputs:

$$\mathcal{G}(\mathcal{P}, \theta) = \big(\mathcal{P}, \Psi_{\text{final}}(\theta, \mathcal{P})\big),\ m = \dim \mathbf{P}_0 + q.$$

Let $\mathcal{B}$ be the union of the pairs for which a parameter matrix has smaller rank than allowed and the remaining pairs where the matrix of first derivatives of $\mathcal{G}$ has rank below m.

The smaller-rank matrices require fewer independent coordinates. For matrices of the full allowed ranks, the derivative condition fails on proper analytic zero sets, by the preceding construction. Hence $\dim_{\mathrm{H}} \mathcal{B} \leq m-1$. Here we use the elementary fact that the zero set of a nonzero analytic function on an open subset of $\mathbb{R}^m$ has Hausdorff dimension at most $m-1$.

We must also exclude parameters whose observations match those of a point in $\mathcal{B}$. Define $\mathcal{N} := \mathcal{G}^{-1}\big(\mathcal{G}(\mathcal{B})\big)$. Smooth maps do not increase Hausdorff dimension. Therefore we have: $\dim_{\mathrm{H}} \mathcal{G}(\mathcal{B}) \leq m-1$. Outside $\mathcal{B}$, select m coordinates of $\mathcal{G}$ whose derivative matrix is invertible. The inverse function theorem gives a smooth local inverse. Applying this inverse to the corresponding coordinates of $\mathcal{G}(\mathcal{B})$ gives $\dim_{\mathrm{H}}(\mathcal{N} \setminus \mathcal{B}) \leq \dim_{\mathrm{H}} \mathcal{G}(\mathcal{B}) \leq m-1$. A countable collection of these neighborhoods suffices. Since $\mathcal{B} \subseteq \mathcal{N}$, we conclude that $\dim_{\mathrm{H}} \mathcal{N} \leq m-1$.

The set $\mathcal{B}$ is closed: having the full allowed matrix ranks and an invertible selected derivative matrix are open conditions. Thus $\mathcal{B}$ is a countable union of compact sets. Its continuous image $\mathcal{G}(\mathcal{B})$ is likewise such a union, so $\mathcal{N}$ is measurable.

If $(\mathcal{P}, \theta) \notin \mathcal{N}$, every parameter with the same observations lies outside $\mathcal{B}$. The inverse function theorem therefore isolates each matching parameter from all other matches. There are at most countably many such isolated points. Finally, Fubini in the independent matrix coordinates gives the almost-everywhere conclusion in the stated order. □

Lemma 11 (Prefixes are generically locally invertible)**.** *For each fixed length n and prefix depth ℓ, the input Jacobian of $x_{\ell,n}$ is nonsingular almost everywhere in the raw prefix parameters and input. Consequently, for almost every prefix, its image contains a nonempty open set.*

Proof. It suffices to construct one nonsingular Jacobian. In each head take $\mathbf{M}_a = \varepsilon e_1 e_1^\top$ and let $\mathbf{W}_a$ be coordinate-block projections summing to I_d. Writing $X = (z, X')$, the layer is

$$(z, X') \longmapsto (f_\varepsilon(z), A_\varepsilon(z)X'),\ A_\varepsilon = \mathrm{softmax}(\varepsilon z z^\top),\ f_\varepsilon = A_\varepsilon z.$$

Choose distinct z_i. For $\varepsilon > 0$, the matrix $K_{ij} = e^{\varepsilon z_i z_j}$ is positive definite, because

$$c^\top K c = \sum_{m \geq 0} \frac{\varepsilon^m}{m!} \left(\sum_i c_i z_i^m \right)^2 > 0\ (c \neq 0)$$

by the Vandermonde determinant. Thus A_ε is invertible. For $n \geq 2$, put $w = z - \bar{z}\mathbf{1}$ and $v = \|w\|^2/n > 0$. Expansion at $\varepsilon = 0$ gives

$$Df_\varepsilon = \frac{\mathbf{1}\mathbf{1}^\top}{n} + \varepsilon \left(vI + \frac{2zw^\top}{n} \right) + O(\varepsilon^2).$$

Its first-order block on $\mathbf{1}^\perp$ is $vI + 2ww^\top/n \succ 0$, so Df_ε is invertible for sufficiently small $\varepsilon > 0$. The full input Jacobian is block triangular with invertible diagonal blocks Df_ε and $d-1$ copies of A_ε.

The output first coordinates remain distinct: the function $q \mapsto \sum_j z_j e^{\varepsilon q z_j} / \sum_j e^{\varepsilon q z_j}$ has derivative ε times a positive weighted variance. We can therefore repeat the construction through each layer. For $n = 1$ the witness is the identity. Each prefix determinant is consequently a nonzero analytic function. Its zero set is null; Fubini and the inverse function theorem give the conclusion. □

The geometric strengthening. The following standard result is the only additional external ingredient used to obtain the open-probes and closed-exception claims. We state it explicitly rather than treating those claims as consequences of analyticity.

Theorem 13 (Real-exponential geometry; external input)**.** *Call a set* definable *if it can be described by a finite formula using real constants, arithmetic, exponentials, equalities, inequalities and quantifiers over real variables. Such sets have integer Hausdorff dimension, satisfy*

$$\dim_{\mathrm{H}} \overline{A} = \dim_{\mathrm{H}} A,\ A \subseteq \mathbb{R}^m \text{ is null } \iff \dim_{\mathrm{H}} A < m,$$

and, for a definable family $B \subseteq \mathbb{R}^s \times \mathbb{R}^D$, each set $\{u : \dim_{\mathrm{H}} B_u = j\}$ is definable. Here $\dim \varnothing = -\infty$.

This is a consequence of Wilkie's o-minimality theorem for $\mathbb{R}_{\exp}$ and its dimension theory (Wilkie, 2009); we do not reprove that foundational theorem.

Theorem 14 (Structured probing for softmax attention)**.** *Define $H := \max_\ell h_\ell$ and $N = d(2H+1)+1$. There exist lengths $n_j \le \max\{3, 1 + d/H\}$ and an open, dense, full-measure set $\mathcal{U} \subseteq \mathcal{S} = \prod_{j=1}^N \mathbb{R}^{n_j \times d}$ such that every $\mathcal{P} \in \mathcal{U}$ admits a closed exceptional set $E_\mathcal{P} \subseteq \Theta$, with $\dim_{\mathrm{H}} E_\mathcal{P} \le D-1$, satisfying*

$$\Psi_{all}(\boldsymbol{\theta}; \mathcal{P}) = \Psi_{all}(\boldsymbol{\theta}_0; \mathcal{P}) \ \to\ x_{\ell,n}(\cdot; \boldsymbol{\theta}) \equiv x_{\ell,n}(\cdot; \boldsymbol{\theta}_0)\ (\ell \in [L],\ n \ge 1)$$

for every $\boldsymbol{\theta}_0 \notin E_\mathcal{P}$ and every $\boldsymbol{\theta} \in \Theta$. The probe set $\mathcal{P}$ is fixed independently of the target.

Proof. We first state and prove an equivalent result. That with some more argument proves Theorem 14.

Proposition 15 (Fixed-probes identification)**.** *Put $H = \max_\ell h_\ell$ and $N = d(2H+1)+1$. There exist lengths $n_j \le \max\{3, 1+d/H\}$ such that, for almost every probe tuples $\mathcal{P} \in \mathcal{S} = \prod_{j=1}^N \mathbb{R}^{n_j \times d}$ and almost every target $\boldsymbol{\theta}_0 \in \Theta$, every competitor with the same complete hidden transcript has the same canonical representation in every layer. Consequently,*

$$\Psi_{all}(\boldsymbol{\theta}; \mathcal{P}) = \Psi_{all}(\boldsymbol{\theta}_0; \mathcal{P}) \ \to\ x_{\ell,n}(\cdot; \boldsymbol{\theta}) \equiv x_{\ell,n}(\cdot; \boldsymbol{\theta}_0)\ (\ell \in [L],\ n \ge 1).$$

Proof. Assume first $H \ge 2$. Take the length profile of Lemma 10 and append one triple. Thus

$$N = d + d(H+1) + d(H-1) + 1 = d(2H+1) + 1.$$

We first use the hidden transcript to restrict each layer. Let $G_\ell(\zeta; \boldsymbol{\theta}_{<\ell})$ be the first batch after the target prefix, and let π_ℓ map raw factors to effective parameters. The map

$$(\zeta, \boldsymbol{\theta}) \longmapsto \big(G_\ell(\zeta; \boldsymbol{\theta}_{<\ell}), \pi_\ell(\theta_\ell)\big)$$

is a submersion almost everywhere: its probes derivative is invertible by Lemma 11, and its independent current-layer factor map is a submersion on the full-rank stratum. Pulling back the exceptional set in Lemma 10 therefore gives a null set. For almost every first batch and target, each layer has only countably many consistent effective candidates. This includes every matching competitor: its hidden transcript fixes the same layer inputs as those observed from the target.

Then we use the last triple to separate these candidates. Fix such a first batch and target, with open prefix images at length three. For a candidate η at layer ℓ not canonically equivalent to the target, the discrepancy

$$X \longmapsto F_3^{(\ell)}(x_{\ell-1,3}(X; \boldsymbol{\theta}_0); \eta) - x_{\ell,3}(X; \boldsymbol{\theta}_0)$$

is nonzero and analytic. Otherwise the standalone layer maps would agree on an open set, hence everywhere, contradicting Lemma 9. Almost every last triple avoids the zero sets of all these countably many discrepancies. Every matching stack then has the target's canonical layers and agrees at every length.

If $H = 1$, instead use d singletons, one $(d+1)$-token input, and $2d$ additional triples. The singleton states satisfy $C_\ell = C_{\ell-1}\mathbf{W}_\ell$, so generically they determine every $\mathbf{W}_\ell$. By Lemma 11, the prefix image cannot lie in $\det[\mathbf{1}\ Z] = 0$. Analyticity therefore makes the propagated $(d+1)$-token input Z affinely independent at every depth, almost everywhere. From its layer output Y recover

$$A = [\mathbf{1}\ \ Y\mathbf{W}_\ell^{-1}][\mathbf{1}\ \ Z]^{-1}.$$

The equations $\log(A_{ij}/A_{i1}) = z_i^\top \mathbf{M}_\ell(z_j - z_1)$ determine $\mathbf{M}_\ell$: the differences $z_j - z_1$, $j = 2, \dots, d+1$, form a basis and the tokens span $\mathbb{R}^d$.

Finally we show that the failure set is measurable. At each layer, give the competitor's heads positive signs and the target's heads negative signs. Canonical agreement means that these signed heads can be partitioned into groups with equal score matrices and zero signed value sums. There are finitely many partitions, each imposing polynomial equalities. Canonical agreement is therefore closed. The incidence set of matching transcripts but canonical disagreement is closed intersected with open, hence a countable union of compact sets. Its projection

$$B = \{(\mathcal{P}, \boldsymbol{\theta}_0) : \exists \boldsymbol{\theta}, \Psi_{\text{all}}(\boldsymbol{\theta}; \mathcal{P}) = \Psi_{\text{all}}(\boldsymbol{\theta}_0; \mathcal{P}), \boldsymbol{\theta} \not\sim_{\text{can}} \boldsymbol{\theta}_0\}$$

is likewise a countable union of compact sets and is measurable. The preceding argument and Fubini show that B is null. A second application of Fubini gives the stated quantifier order. □

Take the length profile of Proposition 15 and its null failure set B. The set B is definable: finite hidden transcripts involve arithmetic and exponentials, and canonical agreement is given by the finite polynomial description above.

By Fubini, almost every fibre $B_\mathcal{P}$ is null. Therefore $A = \{\mathcal{P} \in \mathcal{S} : \dim_{\mathrm{H}} B_\mathcal{P} = D\}$ is null and definable by Theorem 13. We define

$$\mathcal{U} = \mathcal{S} \setminus \overline{A},\ E_\mathcal{P} = \overline{B_\mathcal{P}}\ (\mathcal{P} \in \mathcal{U}).$$

The same theorem makes $\mathcal{U}$ open, dense and full measure, and $E_\mathcal{P}$ closed with Hausdorff dimension at most $D - 1$. Outside $E_\mathcal{P}$ there is no canonically different competitor with the same transcript. Canonical agreement implies equality of all layer maps, and thus of all composed functions at every length. □

Multilayer MHA reconstruction from a fixed set of probes The observations already supply each layer's training inputs: they are the preceding *target* states, not states generated by the reconstructed prefix. Each layer can consequently be fitted independently. Define its cached-data loss by

$$\mathcal{L}_\ell(\theta_\ell) = \sum_{j=1}^{N} \left\| F_{n_j}^{(\ell)}(Z_\ell^{(j)}; \theta_\ell) - Y_\ell^{(j)} \right\|_F^2. \tag{10}$$

Optimization is over the specified raw factors, so the prescribed head counts and rank bounds are maintained automatically.

Algorithm 3 Fixed-probe set multilayer reconstruction

Require: Architecture; probe set $\mathcal{P} = (X^{(j)})_{j=1}^N \in \mathcal{U}$; complete hidden access to a target $\boldsymbol{\theta}_0 \notin E_\mathcal{P}$
Ensure: A stack with the target's composed hidden functions at every length

1: **for** $j = 1, \dots, N$ **do**
2: Query $X^{(j)}$ once; record $H_0^{(j)} \leftarrow X^{(j)}$ and all returned states $H_1^{(j)}, \dots, H_L^{(j)}$
3: **end for**
4: **for** $\ell = 1, \dots, L$ **do** ▷ These fits may run in parallel
5: Set $Z_\ell^{(j)} \leftarrow H_{\ell-1}^{(j)}$ and $Y_\ell^{(j)} \leftarrow H_\ell^{(j)}$ for every j
6: Find any $\widehat{\theta}_\ell$ in the layer's raw parameter space with $\mathcal{L}_\ell(\widehat{\theta}_\ell) = 0$ ▷ Fit nonlinear solver
7: **end for**
8: **return** $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \dots, \widehat{\theta}_L)$

Correctness and scope. Each feasibility problem has a solution, the true layer. Induction over ℓ shows that the returned stack reproduces every recorded state. Theorem 14 therefore gives the all-length conclusion. In numerical implementations it is replaced by layerwise least-squares optimization. It should be mentioned that neither its global convergence nor an OOD error bound from a nonzero residual follows from the theorem. The theorem provides a guarantee, the implementation requires implementing a solver or independently fitting the layers. No head alignment *between layers* is required.

The special case of one layer: The above statement in Theorem 14 discusses identification of generic targets from probes in the multilayer case. However, we show in the following lemma that for a single layer, we can *exactly identify every target* at a function level.

Lemma 12 (One layer: every target from fixed probes)**.** *Let $L = 1$ and $N = 4dh + 1$. There exist an integer $n_* = n_*(d, h)$ and a fixed probe set $\mathcal{P} \in (\mathbb{R}^{n_* \times d})^N$ such that, simultaneously for all $\theta_1, \theta_2 \in \Theta$, we have:*

$$\Psi_{\textit{final}}(\theta_1; \mathcal{P}) = \Psi_{\textit{final}}(\theta_2; \mathcal{P}) \iff [F_n(\cdot; \theta_1) \equiv F_n(\cdot; \theta_2) \textit{ for every } n \geq 1] .$$

Proof. Put $K = (d-1)(2h-1) + 1$ and $q_t = (1, t, \ldots, t^{d-1})^\top$ for $t = 1, \ldots, K$, and let $\mathcal{D}$ contain the distinct vectors among $0, e_1, \ldots, e_d, q_1, \ldots, q_K$. For positive integer multiplicities $m = (m_x)_{x \in \mathcal{D}}$, let $X(m)$ contain m_x copies of each x. At a copy of x_0, output coordinate c equals

$$R(m; \xi) = \sum_{a=1}^{h} \frac{\sum_{x \in \mathcal{D}} m_x \tau_a^x x^\top w_a}{\sum_{x \in \mathcal{D}} m_x \tau_a^x}, \; \tau_a = \exp(\mathbf{M}_a^\top x_0), \; w_a = \mathbf{W}_a e_c, \tag{11}$$

where $\tau^x = \prod_i \tau_i^{x_i}$ and $\xi = ((\tau_a, w_a))_a \in \Xi = ((0, \infty)^d \times \mathbb{R}^d)^h$.

Clearing the positive denominators of $R(m; \xi) - R(m; \xi')$ gives a polynomial $P(m; \xi, \xi')$, since every $x \in \mathcal{D}$ has nonnegative integer coordinates. The Finite Witness Theorem (Amir et al., 2023), applied to the $4dh$-dimensional pair (ξ, ξ'), implies that almost every choice of $N = 4dh + 1$ generic positive multiplicity vectors detects every nonzero discrepancy.

The bad witness set is semialgebraic: writing $P = \sum_\alpha c_\alpha(\xi, \xi') m^\alpha$, it is the projection of the set defined by $\sum_\alpha c_\alpha^2 > 0$ and $P(m^{(j)}; \xi, \xi') = 0$ for every j. Its nullity implies that its closure has empty interior. We choose a good rational tuple. Since $R(\lambda m; \xi) = R(m; \xi)$, normalize each vector to sum one and clear all denominators by a common integer. This gives positive integer multiplicities of a common sum n_*. The resulting probe set works for every row and coordinate, because their parameters all belong to Ξ.

Suppose two layers agree on this set of probes. Let $B_1, \ldots, B_J$, $J \leq 2h$, be their distinct score matrices, and let Δ_j be the difference of their summed value matrices at B_j. For each j, some q_t satisfies $q_t^\top B_j \neq q_t^\top B_l$ for all $l \neq j$: any d moment-curve vectors are independent, so each proper subspace $\{q : q^\top (B_j - B_l) = 0\}$ contains at most $d - 1$ of them, fewer than K altogether.

Fix this $x_0 = q_t$ and a coordinate c. The witness property gives

$$\sum_{l=1}^{J} \frac{A_l(m)}{D_l(m)} \equiv 0, \; D_l = \sum_x m_x e^{x_0^\top B_l x}, \; A_l = \sum_x m_x e^{x_0^\top B_l x} x^\top \Delta_l e_c.$$

Each denominator has coefficient one at m_0, and its coefficients at m_{e_i} determine $x_0^\top B_l$. After merging equal denominators, D_j is a distinct linear factor. Clearing denominators shows that this prime factor divides A_j, so $A_j = \lambda D_j$. The coefficient at m_0 gives $\lambda = 0$; the coefficients at m_{e_i} then give $(\Delta_j)_{ic} = 0$. Thus all $\Delta_j = 0$, proving functional agreement at every length. □

B.5 Learning Functionals and Operators from Probes

We first prove a common approximation lemma. We then give formal versions of Theorems 6 and 7, in that order. All function spaces below carry the uniform norm. Matrix inputs, outputs, and observations are identified with vectors by listing their entries in a fixed order. All MLP readouts have biases, an affine output layer, and a fixed continuous, non-polynomial hidden activation τ.

Separation and approximation. For a compact parameter set K and a jointly continuous family $f : \mathcal{X} \times K \to \mathbb{R}^m$ on compact $\mathcal{X}$, write

$$\mathcal{F}_K := \{f_\theta = f(\cdot;\theta) : \theta \in K\}.$$

A map $R : K \to \mathbb{R}^s$ separates the represented functions if

$$R(\theta) = R(\theta') \implies f_\theta = f_{\theta'},\ \theta, \theta' \in K.$$

Note that equality of functions need not imply equality of R, because R acts on weights.

Lemma 13 (Separation implies universal approximation)**.** *Let $\mathcal{X} \subseteq \mathbb{R}^d$ and $K \subseteq \mathbb{R}^q$ be compact, let $f : \mathcal{X} \times K \to \mathbb{R}^m$ be continuous, and let $R : K \to \mathbb{R}^s$ be continuous and separate the represented functions. Then $\mathcal{F}_K$ is compact and there is a continuous map*

$$g : R(K) \to \mathcal{F}_K,\ g(R(\theta)) = f_\theta.$$

For every continuous $\Lambda : \mathcal{F}_K \to \mathbb{R}^k$, every compact $\mathcal{Y} \subseteq \mathbb{R}^a$, every continuous $\mathcal{T} : \mathcal{F}_K \to C(\mathcal{Y}, \mathbb{R}^k)$, and every $\epsilon > 0$, there are MLPs $\phi : \mathbb{R}^s \to \mathbb{R}^k$ and $\psi : \mathbb{R}^{s+a} \to \mathbb{R}^k$ such that

$$\sup_{\theta \in K} \|\Lambda(f_\theta) - \phi(R(\theta))\|_2 < \epsilon,\ \sup_{\theta \in K} \sup_{y \in \mathcal{Y}} \|\mathcal{T}(f_\theta)(y) - \psi(R(\theta), y)\|_2 < \epsilon.$$

Proof. Uniform continuity of f on $\mathcal{X} \times K$ implies continuity of the realization map $\Phi : K \to C(\mathcal{X}, \mathbb{R}^m)$, $\Phi(\theta) = f_\theta$, in the uniform norm. Thus $\mathcal{F}_K = \Phi(K)$ is compact. Put $S = R(K)$ and define $g(R(\theta)) = f_\theta$. Separation makes g well defined. For every closed $C \subseteq \mathcal{F}_K$,

$$g^{-1}(C) = R\left(\Phi^{-1}(C)\right)$$

is compact and therefore closed in S. Hence g is continuous.

The map $\Lambda \circ g$ is continuous on the compact set S. The universal approximation theorem Leshno et al. (1993) gives an MLP ϕ such that $\sup_{s \in S} \|\Lambda(g(s)) - \phi(s)\|_2 < \epsilon$. Substituting $s = R(\theta)$ proves the functional conclusion. The compact-set, vector-valued form follows by applying the Tietze extension theorem and the scalar approximation theorem coordinatewise.

For the operator result, suppose $G(s, y) := \mathcal{T}(g(s))(y)$, $(s, y) \in S \times \mathcal{Y}$. If $(s_j, y_j) \to (s, y)$, then we have

$$\begin{aligned}\|G(s_j, y_j) - G(s, y)\|_2 \le\ & \|\mathcal{T}(g(s_j)) - \mathcal{T}(g(s))\|_\infty \\ & + \|\mathcal{T}(g(s))(y_j) - \mathcal{T}(g(s))(y)\|_2 \to 0.\end{aligned}$$

Thus G is continuous on the compact set $S \times \mathcal{Y}$. We thus apply the same approximation theorem to G, and then substitute $s = R(\theta)$. □

Final-output observations. The following Theorem 16 is the formal version of Theorem 6.

Theorem 16 (Approximation from final-output probes)**.** *Let $f : \mathbb{R}^d \times \mathbb{R}^q \to \mathbb{R}^m$ be real analytic, let $\mathcal{X} \subseteq \mathbb{R}^d$ be compact with nonempty interior, and let $\Theta \subseteq \mathbb{R}^q$ be compact. For every $N \ge 2q + 1$, and for almost every $\mathcal{P} = (p_1, \cdots, p_N) \in \mathcal{X}^N$, the map*

$$T_\mathcal{P} : \mathcal{F}_\Theta \to \mathbb{R}^{mN},\ T_\mathcal{P}(g) = (g(p_1), \cdots, g(p_N)),$$

is one-to-one and has a continuous inverse on its image.

Then, for every continuous $\Lambda : \mathcal{F}_\Theta \to \mathbb{R}^k$, every compact $\mathcal{Y} \subseteq \mathbb{R}^a$, every continuous $\mathcal{T} : \mathcal{F}_\Theta \to C(\mathcal{Y}, \mathbb{R}^k)$, and every $\epsilon > 0$, there are MLPs $\phi : \mathbb{R}^{mN} \to \mathbb{R}^k$ and $\psi : \mathbb{R}^{mN+a} \to \mathbb{R}^k$ satisfying

$$\sup_{\theta \in \Theta} \|\Lambda(f_\theta) - \phi(\Psi_{\textit{final}}(\theta, \mathcal{P}))\|_2 < \epsilon,\ \sup_{\theta \in \Theta} \sup_{y \in \mathcal{Y}} \|\mathcal{T}(f_\theta)(y) - \psi(\Psi_{\textit{final}}(\theta, \mathcal{P}), y)\|_2 < \epsilon.$$

Proof. Since f is globally analytic, we may apply Theorem 1 on sufficiently large input and parameter balls containing $\mathcal{X}$ and Θ in their interiors. Restricting the parameters to Θ and intersecting the exceptional probe set with $\mathcal{X}^{2q+1}$ gives a Lebesgue-null $\mathcal{N}_0 \subseteq \mathcal{X}^{2q+1}$ outside which final-output observations separate all represented functions. For $N > 2q + 1$, take $\mathcal{N} = \mathcal{N}_0 \times \mathcal{X}^{N-2q-1}$; for $N = 2q + 1$, take $\mathcal{N} = \mathcal{N}_0$. Fix $\mathcal{P} \notin \mathcal{N}$. The map $R(\theta) := \Psi_{\text{final}}(\theta, \mathcal{P})$ is continuous and separates functions on Θ. Lemma 13, with $K = \Theta$, gives both approximation conclusions and a continuous map $g : R(\Theta) \to \mathcal{F}_\Theta$. Separation makes $T_\mathcal{P}$ one-to-one, and $g(T_\mathcal{P}(f_\theta)) = f_\theta$, so $g = T_\mathcal{P}^{-1}$. Also,

$$\|T_\mathcal{P}(f) - T_\mathcal{P}(f')\|_2 \le \sqrt{N}\|f - f'\|_\infty,\ f, f' \in \mathcal{F}_\Theta.$$

The exceptional set comes solely from the identification theorem. □

Hidden-layer observations. The following Theorem 17 is the formal version of Theorem 7. We retain the complete final probe responses in their coordinate order, but discard the ordering of the hidden neurons or channels. For MLPs and MHA stacks, respectively, define

$$v^{i,k}(\theta) := \left(x_{i,k}(p_j;\theta)\right)_{j\in[N]} \in \mathbb{R}^N,\ i \in [L-1],\ k \in [d_i],$$
$$v^{i,k}(\theta) := \left([x_{i,n_j}(P_j;\theta)]_{a,k}\right)_{j\in[N],\, a\in[n_j]} \in \mathbb{R}^{\sum_j n_j},\ i \in [L-1],\ k \in [d].$$

The coordinates within each vector retain the probe order and, for attention, the token order. Let $S_i(\theta)$ be the multiset of these vectors, including repetitions Thus $S_i(\theta)$ contains r_i vectors in $\mathbb{R}^t$, where $r_i = d_i$, $t = N$ for MLPs and $r_i = d$, $t = \sum_j n_j$ for MHA.

For a multiset S of r vectors in $\mathbb{R}^t$, we define

$$\mathcal{M}_r(S) := \left(\sum_{v\in S} v^\alpha\right)_{\alpha\in\mathbb{N}_0^t,\, 1\le|\alpha|\le r},\ \mathbb{N}_0 := \{0,1,\cdots\},$$
$$|\alpha| := \sum_{b=1}^t \alpha_b,\ v^\alpha := \prod_{b=1}^t v_b^{\alpha_b},$$
$$R_{\mathcal{P}}(\theta) := \left[\mathcal{M}_{r_1}(S_1(\theta)); \cdots; \mathcal{M}_{r_{L-1}}(S_{L-1}(\theta)); \Psi_{\text{final}}(\theta,\mathcal{P})\right].$$

Every occurrence of a repeated vector contributes to the sum.

For the MLP variant of HIDDENPROBE, we use layerwise Set Transformers followed by a final MLP, as described in Section A. For the pure-MHA specialization of HIDDENPROBE, we consider the shared trajectory encoder and layerwise pooling given by

$$\widehat{R}_{\mathcal{P},\varphi}(\theta) := \left[\left(\sum_{k=1}^d \varphi(v^{i,k}(\theta))\right)_{i=1}^{L-1}; \Psi_{\text{final}}(\theta,\mathcal{P})\right],$$
$$\mathcal{H}_{\mathcal{P},\phi}(\theta) := \rho\left(\widehat{R}_{\mathcal{P},\varphi}(\theta)\right),\ \phi := (\varphi,\rho).$$

Here φ and ρ are MLPs satisfying the preceding activation and bias assumptions, with unrestricted hidden widths and encoder output dimension. The same φ is used for every hidden channel and layer. The final responses $\Psi_{\text{final}}(\cdot,\mathcal{P})$ retain their coordinate order. For operators, the prediction MLP additionally receives the query input.

Theorem 17 (Approximation from hidden-layer probes)**.** *Consider either of the following settings.*

1. MLPs. *Consider the MLP family of Theorem 2. Its widths are $\mathbf{D} = (d_0,\cdots,d_L)$ and its activation is real analytic, injective, and non-polynomial. Let $\mathcal{X} \subseteq \mathbb{R}^{d_0}$ be compact with nonempty interior, let $N := 1 + \max_{0\le i<L} d_i$, and let $\mathcal{U} \subseteq \mathcal{X}^N$ consist of pairwise distinct tuples whose one-padded input matrix has rank d_0+1. This set has full measure by Lemma 2. Set $f_\theta = x_L(\cdot;\theta)|_{\mathcal{X}}$.*

2. Softmax attention stacks. *Let the family satisfy the assumptions of Theorem 14, and define $H := \max_{\ell\in[L]} h_\ell$, $N := d(2H+1)+1$. Let $n_1,\cdots,n_N$ and $\mathcal{U} \subseteq \prod_{j=1}^N \mathbb{R}^{n_j\times d}$ be the length profile and full-measure good-bank set supplied by that theorem. For an evaluation length $n \ge 1$ and compact $\mathcal{X}_n \subseteq \mathbb{R}^{n\times d}$, set $f_\theta = x_{L,n}(\cdot;\theta)|_{\mathcal{X}_n}$.*

Let $\Theta = \mathbb{R}^D$ be the respective full raw parameter space. For every $\mathcal{P} \in \mathcal{U}$, there is a closed set $E_{\mathcal{P}} \subseteq \Theta$ with $\dim_{\mathrm{H}} E_{\mathcal{P}} \le D-1$ such that the following holds.

For every compact $K \subseteq \Theta \setminus E_{\mathcal{P}}$, the representation $R_{\mathcal{P}}$ defined above admits a continuous map $g_{\mathcal{P}} : R_{\mathcal{P}}(K) \to \mathcal{F}_K$ with $g_{\mathcal{P}}(R_{\mathcal{P}}(\theta)) = f_\theta$, $\theta \in K$. For every continuous $\Lambda : \mathcal{F}_K \to \mathbb{R}^k$, every compact $\mathcal{Y} \subseteq \mathbb{R}^a$, every continuous $\mathcal{T} : \mathcal{F}_K \to C(\mathcal{Y},\mathbb{R}^k)$, and every $\epsilon > 0$, there are models from the respective HIDDENPROBE *or* HIDDENPROBE *family described above such that*

$$\sup_{\theta\in K} \|\Lambda(f_\theta) - \mathcal{H}_{\mathcal{P},\phi}(\theta)\|_2 < \epsilon,\ \sup_{\theta\in K}\sup_{y\in\mathcal{Y}} \left\|\mathcal{T}(f_\theta)(y) - \mathcal{H}^{\mathrm{op}}_{\mathcal{P},\psi}(\theta,y)\right\|_2 < \epsilon.$$

Proof. Fix $\mathcal{P} \in \mathcal{U}$. For MLPs, we define

$$E_{\mathcal{P}} := \left\{ \theta \in \Theta : \prod_{i=0}^{L-1} \det\left(\overline{\mathbf{X}}_i^{\theta}(\mathcal{P}) \left(\overline{\mathbf{X}}_i^{\theta}(\mathcal{P})\right)^{\top} \right) = 0 \right\}.$$

By Lemma 2, the displayed product is a nonzero real-analytic function of θ. Its zero set is closed and has Hausdorff dimension at most $D-1$. For $\theta_0 \notin E_{\mathcal{P}}$, equality of the ordered observations with any raw competitor gives, by injectivity, $(B_i - B_i^0)\overline{\mathbf{X}}_i^{\theta_0}(\mathcal{P}) = 0$, $0 \le i < L$. Full row rank gives $B_i = B_i^0$ for every i. For an affine output layer, its equality follows directly without inverting an activation. For attention, use $E_{\mathcal{P}}$ from Theorem 14. That theorem also identifies every target outside $E_{\mathcal{P}}$ against every raw competitor, at every sequence length.

We first show that $\mathcal{M}_r$ determines an r-element multiset. If $\mathcal{M}_r(S) = \mathcal{M}_r(S')$, then for every $a \in \mathbb{R}^t$ and $b \in [r]$, expansion of $(a^\top v)^b$ gives $\sum_{v\in S}(a^\top v)^b = \sum_{v\in S'}(a^\top v)^b$. For real numbers $c_1, \cdots, c_r$, their power sums $p_b = \sum_j c_j^b$ determine the coefficients of $\prod_j (z - c_j)$: writing e_b for the sum of all products of b distinct c_j and $e_0 = 1$, Newton's identities give

$$be_b = \sum_{j=1}^{b} (-1)^{j-1} e_{b-j} p_j,\ b \in [r],\ \prod_{j=1}^{r}(z - c_j) = \sum_{b=0}^{r} (-1)^b e_b z^{r-b}.$$

Consequently the projected multisets agree, including multiplicities. Choose a outside the finitely many hyperplanes $\{a : a^\top(v - v') = 0\}$ indexed by distinct vectors in $S \cup S'$. This projection is one-to-one on that union, so $S = S'$.

Now fix compact $K \subseteq \Theta \setminus E_{\mathcal{P}}$ and $\theta, \theta_0 \in K$ with $R_{\mathcal{P}}(\theta) = R_{\mathcal{P}}(\theta_0)$. The preceding argument gives equality of the hidden trajectory multisets. Thus one permutation per hidden layer aligns the competitor's responses at all probes. For MLPs, choose these permutation matrices P_i, put $P_0 = P_L = I$, and transform the competing parameter by

$$W_i' = P_{i+1} W_i P_i^\top,\ b_i' = P_{i+1} b_i,\ 0 \le i < L.$$

Induction gives $x_i(u; \theta') = P_i x_i(u; \theta)$. Hence $\Psi_{\text{all}}(\theta', \mathcal{P}) = \Psi_{\text{all}}(\theta_0, \mathcal{P})$ while $f_{\theta'} = f_\theta$.

For attention, states are row matrices. Choose permutations Π_i that align their columns. Let $\Pi_0 = \Pi_L = I$, and we define

$$Q'_{ia} = \Pi_{i-1}^\top Q_{ia},\ K'_{ia} = \Pi_{i-1}^\top K_{ia},\ V'_{ia} = \Pi_{i-1}^\top V_{ia},\ O'_{ia} = O_{ia}\Pi_i.$$

Then $\mathbf{M}'_{ia} = \Pi_{i-1}^\top \mathbf{M}_{ia} \Pi_{i-1}$ and $\mathbf{W}'_{ia} = \Pi_{i-1}^\top \mathbf{W}_{ia} \Pi_i$, so we get:

$$(X\Pi_{i-1})\mathbf{M}'_{ia}(X\Pi_{i-1})^\top = X\mathbf{M}_{ia}X^\top,\ (X\Pi_{i-1})\mathbf{W}'_{ia} = X\mathbf{W}_{ia}\Pi_i.$$

The attention scores are unchanged, and induction gives us $x_{i,n}(X; \theta') = x_{i,n}(X; \theta)\Pi_i$. Again, $\Psi_{\text{all}}(\theta', \mathcal{P}) = \Psi_{\text{all}}(\theta_0, \mathcal{P})$ and $f_{\theta'} = f_\theta$. The all-competitor identification established above implies $f_\theta = f_{\theta_0}$ in either case. It is unnecessary that θ' belong to K.

The map $R_{\mathcal{P}}$ is continuous, since its coordinates are polynomials in finitely many continuous observations. The realization is jointly continuous on the chosen compact input domain and K. Lemma 13 therefore gives $g_{\mathcal{P}}$ and continuous functional and operator readouts on $R_{\mathcal{P}}(K)$ and $R_{\mathcal{P}}(K) \times \mathcal{Y}$. It remains to realize their approximations by the specified architectures.

For HIDDENPROBE, Set Transformer universality Lee et al. (2019), with sufficient capacity and output dimension, approximates each layer's continuous invariant map $\mathcal{M}_{r_i}$. The final MLP approximates the resulting continuous readout.

For HIDDENPROBE, the case $L = 1$ follows directly from Lemma 13, since there are no hidden trajectories. Suppose $L \ge 2$, and define

$$\nu(v) := (v^\alpha)_{\alpha \in \mathbb{N}_0^t,\, 1 \le |\alpha| \le d},\quad t := \sum_{j=1}^{N} n_j,\ C := \left\{ v^{i,k}(\theta) : \theta \in K,\ i \in [L-1],\ k \in [d] \right\},$$

$$\mathcal{S} := R_{\mathcal{P}}(K),\ \mathcal{S}^1 := \{z : \operatorname{dist}(z, \mathcal{S}) \le 1\}.$$

Both C and $\mathcal{S}^1$ are compact. By Lemma 13, there are MLPs ρ, ψ satisfying

$$\sup_{\theta\in K} \|\Lambda(f_\theta) - \rho(R_{\mathcal{P}}(\theta))\|_2 < \epsilon/2,\ \sup_{\theta\in K}\sup_{y\in\mathcal{Y}} \|\mathcal{T}(f_\theta)(y) - \psi(R_{\mathcal{P}}(\theta), y)\|_2 < \epsilon/2.$$

Uniform continuity on $\mathcal{S}^1$ and $\mathcal{S}^1\times\mathcal{Y}$ gives $\delta \in (0,1)$ such that, whenever $z, z' \in \mathcal{S}^1$ and $\|z-z'\|_2 < \delta$, we have:

$$\|\rho(z) - \rho(z')\|_2 < \epsilon/2,\ \sup_{y\in\mathcal{Y}} \|\psi(z,y) - \psi(z',y)\|_2 < \epsilon/2.$$

Choose an MLP φ with $\sup_{v\in C} \|\varphi(v) - \nu(v)\|_2 < \frac{\delta}{1+d(L-1)}$. Since $\mathcal{M}_d(S_i) = \sum_{k=1}^d \nu(v^{i,k})$, we have:

$$\begin{aligned}\sup_{\theta\in K} \|\widehat{R}_{\mathcal{P},\varphi}(\theta) - R_{\mathcal{P}}(\theta)\|_2 &\le \sum_{i=1}^{L-1}\sum_{k=1}^{d} \sup_{\theta\in K} \|\varphi(v^{i,k}(\theta)) - \nu(v^{i,k}(\theta))\|_2 \\ &< \delta.\end{aligned}$$

In particular, $\widehat{R}_{\mathcal{P},\varphi}(K) \subseteq \mathcal{S}^1$. Thus we get:

$$\begin{aligned}&\sup_{\theta\in K} \|\Lambda(f_\theta) - \rho(\widehat{R}_{\mathcal{P},\varphi}(\theta))\|_2 \\ &\le \sup_{\theta\in K} \|\Lambda(f_\theta) - \rho(R_{\mathcal{P}}(\theta))\|_2 + \sup_{\theta\in K} \|\rho(R_{\mathcal{P}}(\theta)) - \rho(\widehat{R}_{\mathcal{P},\varphi}(\theta))\|_2 < \epsilon.\end{aligned}$$

Replacing $\rho(z)$ by $\psi(z,y)$ and taking the supremum over $y \in \mathcal{Y}$ gives the operator result. The independence assertions follow from the identification theorems. □

C Architectural details of HiddenProbe and HiddenProbe

We describe the architectures and training procedures of HiddenProbe for MLPs and HiddenProbe for Transformers in separate subsections.

C.1 HiddenProbe for MLPs

We describe the architectures used for INR classification and Model Zoo regression tasks.

C.1.1 INR Classification

Target networks and learned probes. For INR classification, each target is a SIREN with sine activations. We use $N = 128$ learned coordinate probes. Each probe has a trainable code $\zeta_q \in \mathbb{R}^{32}$, initialized from $\mathcal{N}(0, I_{32})$. Two shared affine maps generate the coordinates:

$$u_q = \tanh\big(A_2(A_1\zeta_q + b_1) + b_2\big) \in (-1,1)^2, \qquad A_1 \in \mathbb{R}^{32\times 32},\quad A_2 \in \mathbb{R}^{2\times 32}.$$

There is no activation between the affine maps. The same ordered probe bank is presented to every target; generation is not conditioned on the target's weights or responses.

Observed responses and neuron trajectories. For INR classification, during one forward pass, we record the two hidden layer responses $X_i^\theta \in \mathbb{R}^{N\times 32}$, $i = 1, 2$, and final outputs $O^\theta \in \mathbb{R}^{N\times 3}$. For each hidden neuron, we collect its responses across the complete bank:

$$v_x^{i,c}(\theta) = \big([X_i^\theta]_{q,c}\big)_{n=1}^{N} \in \mathbb{R}^{128}, \qquad i \in \{1,2\},\quad c \in [32].$$

We design features by concatenating the activations with global features: mean, standard deviation, maximum, minimum. Denoting these statistics by $s(v)$, the resulting feature vector is $\chi(v) = [v; s(v)]$. No normalization is applied before the token projection. Probe order is retained within every trajectory.

Neuron tokens and joint readout. A shared affine map projects every neuron feature to $\mathbb{R}^{d_e}$, and a learned embedding e_i identifies its hidden layer:

$$t^{\theta}_{i,c} = A\chi\big(v^{i,c}_x(\theta)\big) + a + e_i,$$
$$t^{\theta}_{\text{out}} = A_{\text{out}}\,\text{vec}(O^{\theta}) + a_{\text{out}}.$$

One joint Transformer processes all tokens. We use three blocks: each block has eight attention heads, feedforward, ReLU, dropout 0.1 and LayerNorm before its attention and feedforward sublayers. There is no final LayerNorm after the last block. An MLP maps the final CLS token to class logits.

There are no neuron-index embeddings or sequence positional encodings. At evaluation, when dropout is disabled, permuting the neurons within a hidden layer only permutes their tokens and leaves the prediction unchanged. Probe order, in contrast, is retained in both the trajectories and final-output vector.

Relation to the theory. The neuron tokens preserve correspondence across probes, as in the theoretical trajectory representation. The implementation processes the layers jointly rather than using separate layerwise set encoders; layer embeddings retain their identities. It also observes preactivations, from which the theoretical postactivation responses can be computed. Sine is analytic but not injective, so the parameter-identification theorem stated for injective activations does not apply directly to these targets. The learned probes and finite-width token projections are not assumed to satisfy exact separation conditions. The universality results concern capacity-growing architecture families under their stated assumptions, not every fixed-width encoder or trained parameter choice.

C.1.2 HiddenProbe for CNN-zoo accuracy regression

Target networks and learned probes The regression targets are image-classification CNNs. The grayscale zoos use three convolutional layers with channels $1 \to 16 \to 16 \to 16$, kernel size 3 and stride 2, followed by global average pooling and a ten-output affine classifier. The Wild Park zoo contains targets with different depths, channel counts, convolutional settings and activations.

Here, we use $N = 128$ learned image probes. Each probe is generated from a trainable code in $\mathbb{R}^{32\times1\times1}$ and has the appropriate image dimensions. A final $\tanh$ bounds the image values to $(-1, 1)$. The same learned probe bank is used for every target network. The output and hidden-response branches described below share this probe bank and the same target forward pass. We follow a protocol similar to that of Kahana et al. (2025) for learning probes in classification and regression tasks.

Hidden aggregation and Prediction The regressor stacks the probes' output at each channel and then pools the channels within each layer. Like in the case of INR classification, we also use additional global statistics like channelwise means, maxima, standard deviation etc over probes to design extra features to aggregate all the information from hidden layers.

One self-attention block mixes the layer vectors using a learned relative-layer distance bias, followed by a residual pre-LayerNorm feedforward block with GELU. A final attention pool over layers, funal output and a projection give $z_H \in \mathbb{R}^{d_z}$. The attention projections in this aggregation have inner dimension 16; large hidden-branch projections use factorized weights to reduce the number of parameters, and match the parameter count with that of ProbeGen.

The predictor is an MLP that taken in the hidden combined representation and gives a scalar that predicts the accuracy.

Relation to the theory The regressor follows the shared-probe principle and uses hidden responses without additional target-network queries. Its channel pooling respects channel permutations. It is an empirical adaptation of hidden probing in the context of CNN Model Zoo.

C.2 HiddenProbe for Transformers

Details on the learned probes. For MNIST, we use $N = 256$ probes divided into two families. The first contains 192 learned sequences, each consisting of 17 token vectors in $\mathbb{R}^{32}$. The remaining 64 probes are learned 28×28 images.

Each encoder probe is parameterized by independently learned 32-dimensional token codes. Each image is parameterized by 49 learned 16-dimensional codes, corresponding to its 4×4 patches. Within each route, a shared residual two-layer MLP transforms these codes: $G(z) = z + W_2\,\mathrm{GELU}(W_1 z + b_1) + b_2$. The transformed image patches are assembled into a 28×28 image.

The same learned bank is presented to every target: generation is not conditioned on the target's weights or its responses.

Observed responses and channel trajectories. For MNIST and AGNews datasets, the Transformers have two blocks. For every probe, we record the outputs of the target's two transformer blocks, its final normalized token states, and its classifier logits. Denote the three hidden layers by X_1, X_2, and X_{norm}. Each hidden layer has 32 channels, and each probe produces a 10-dimensional logit vector.

For each layer $i \in \{1, 2, \mathrm{norm}\}$ and channel $c \in [32]$, we concatenate that channel's responses across all probes and token positions:

$$v^{i,c}(\theta) = \left([X_i(P_j;\theta)]_{t,c}\right)_{j\in[N],\, t\in[n_j]} \in \mathbb{R}^T,$$

The concatenation uses a fixed order: encoder probes first, then image probes, with token positions kept in order within each probe. Consequently, each channel is represented by its responses to the complete bank.

Encoding and combining channel features. One shared trajectory encoder E is applied independently to every channel in all three blocks: $u^{i,c} = E(v^{i,c}) \in \mathbb{R}^{192}$. The encoder consists of Linear, GELU, Linear, and LayerNorm. It produces 32 feature vectors for each view; it does not pool across channels.

We then combine each block-2 feature with the final-normalized feature of the same channel: $\widetilde{u}^{2,c} = \mathrm{Fuse}\left([u^{2,c}; u^{\mathrm{norm},c}]\right) \in \mathbb{R}^{192}$. This is concatenated with the 32 block-1 features, giving 64 hidden-response tokens in total.

Joint readout and pooling. Each probe's classifier logits are projected from $\mathbb{R}^{10}$ to $\mathbb{R}^{192}$. Together with a learned CLS token, the readout input contains $1 + 32 + 32 + 256 = 321$ tokens of dimension 192. Two pre-Layernorm Transformer blocks process these tokens jointly, a final LayerNorm is applied to their outputs.

Write the resulting tokens as $z_{\mathrm{CLS}}, z_1, \ldots, z_{320} \in \mathbb{R}^{192}$. We form

$$r_\theta = \left[z_{\mathrm{CLS}};\, \frac{1}{320}\sum_{a=1}^{320} z_a;\, \max_{a\in[320]} z_a \right] \in \mathbb{R}^{576},$$

where the maximum is coordinatewise. Thus, mean and maximum pooling act jointly over all non-CLS tokens, including the logit tokens, rather than separately within each layer.

A prediction head with dimensions $576 \to 192 \to 1$ and a GELU activation maps r_θ to standardized accuracy. The predictor receives only the recorded responses and fixed semantic identifiers, not target weights, target identities, or generated probe coordinates.

Relation to the theory. The theoretical constructions motivate a shared bank with different sequence lengths and the use of intermediate responses. The experiments apply these principles to complete transformer classifiers containing feedforward and normalization layers. They do not assume that the learned probes satisfy the exact reconstruction conditions or identify the target function. Likewise, the universality results concern capacity-growing architecture families, not a guarantee for this particular finite-width configuration or its learned parameters.

D Experimental details

D.1 Function-level reconstruction on synthetic neural networks

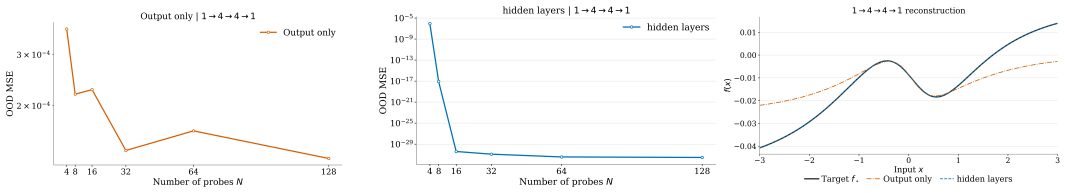
To illustrate the benefits of hidden activations, we study function-level reconstruction of frozen neural networks from a limited number of input probes. We compare *output-only regression*, which observes

only the final network outputs, with *intermediate-layer regression*, which independently fits each layer using observed teacher activations as its inputs and targets. Our goal is to examine how well each approach reconstructs the target function, both between the observed probes and outside the training interval.

Common Setup We consider function-level reconstruction of MLPs with different activations and stacks of MHA blocks, using either final outputs alone or complete hidden-layer observations. Let $\boldsymbol{\theta} := (\theta_1, \ldots, \theta_L)$ denote the parameters of a sequential model $f_{\boldsymbol{\theta}} = f_{\theta_L} \circ \cdots \circ f_{\theta_1}$. Given a set of probes $\mathcal{P} = (p_1, \ldots, p_N)$, we observe either the final outputs $\Psi_{\text{final}}(\boldsymbol{\theta}, \mathcal{P})$ or the complete hidden-layer observations $\Psi_{\text{all}}(\boldsymbol{\theta}, \mathcal{P})$.

For the output-only case we do MSE regression to interpolate on the probes. If $f_{\hat{\boldsymbol{\theta}}}$ is the "student" network trying to reconstruct the "teacher" network, we minimize the following losses: (i) For learning from Ψ_{final}, we minimize $\min_{\hat{\boldsymbol{\theta}}} \sum_{i=1}^{N} \|f_{\hat{\boldsymbol{\theta}}}(p_i) - f_{\boldsymbol{\theta}}(p_i)\|^2$ (ii) For learning from Ψ_{all}, we fit sum of MSE on each of the L layers, given by $\sum_{\ell} \sum_{i=1}^{N} \|f_{\hat{\theta}_\ell}(X_i^{\ell-1}) - X_i^{\ell}\|^2$, where $X_i^{\ell} := f_{\theta_\ell} \circ \cdots \circ f_{\theta_1}(p_i)$, i.e $\{X_i^{\ell}\}$ are the observed outputs from the ℓ-th layer. Here we fit each hidden layer independently but simultaneously using MSE regression.

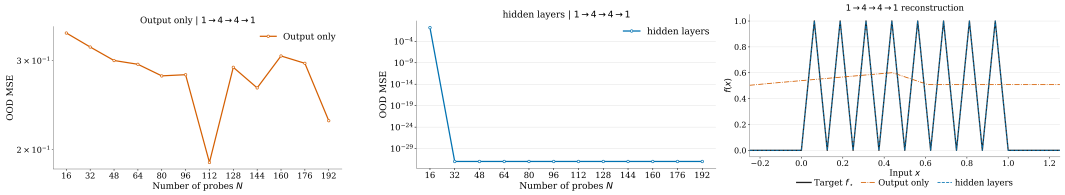
Analytic MLP: We used a neural network with $\tanh$ activation of widths $(1, 4, 4, 1)$, the training interval was $[-1, 1]$, and the evaluation interval is $[-3, 3]$. As shown in Figure 4, access to the hidden activations enables essentially exact reconstruction once sufficiently many observations are available, whereas reconstruction from final outputs alone improves much more gradually and retains a substantially larger error even as the number of observed samples increases.



(a) OOD MSE from Ψ_{final} (b) OOD MSE from Ψ_{all} (c) Reconctruction from Ψ_{all}, Ψ_{final}

Figure 4: Reconstruction of Analytic Net from Ψ_{final}, Ψ_{all}.

ReLU MLP: As in the analytic activation case, Theorem 4 suggests that hidden activations can reduce the number of probes necessary to (almost) identify a ReLU network. We take a target ReLU network with 2 hidden layers, width 4 and weights that represent a member of the "sawtooth" function class, having a large number of breakpoints. As discussed in Hanin & Rolnick (2019), such a large number of breakpoints is atypical. This makes the target a challenging example for reconstruction from final outputs alone using MSE regression from random initialization. This matches the observation: we can't reconstruct the target with just the final output. But with hidden activations, we can reconstruct the target almost exactly with far fewer probes. The same is also reflected in the out-of-distribution MSE plots



(a) OOD MSE from Ψ_{final} (b) OOD MSE from Ψ_{all} (c) Reconctruction from Ψ_{all}, Ψ_{final}

Figure 5: Reconstruction of ReLU Net from Ψ_{final}, Ψ_{all}.

Multi Head Attention Stack Here, we want to reconstruct an MHA stack. We randomly initialize a two-layer multi-head attention network with input width $d = 4$, two heads per layer and query/key

and value widths $k = r = 2$ (bias-free softmax attention, without residual connections, normalization or feed-forward blocks). We sample four single-token inputs followed by generic three-token inputs, and a budget of N probes uses the first N of them, for N from 5 to 40. We fit the "student" MHA block jointly using Ψ_{final}, and each layer independently using Ψ_{all}, both using MSE regression as described above.

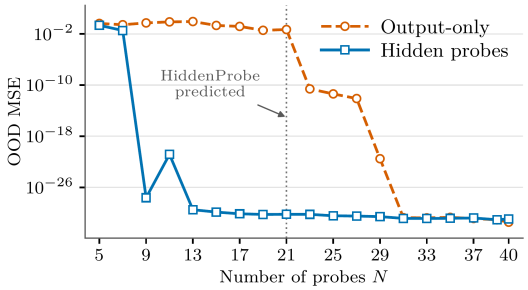


Figure 6: Reconstruction of MHA Stack from Ψ_{final}, Ψ_{all}.

As shown in Figure 6, access to the hidden activations enables essentially exact reconstruction from as few as nine probes, below the $N = 21$ probes that Theorem 5 predicts, whereas reconstruction from final outputs alone becomes exact only at about 31 probes.

Together, these three experiments suggest that hidden activations enable more accurate function reconstruction with fewer probes than final outputs alone. This observation is consistent with our empirical results on neural functional learning, where HIDDENPROBE, which uses hidden activations, outperforms ProbeGen, which relies only on final outputs.

D.2 Full INR Classification Comparison

Table 4 provides an extended comparison on the INR classification benchmarks, including both recent weight-space methods and several earlier baselines omitted from the main table for brevity. The MLP, MLP with permutation augmentation, MLP with alignment, INR2Vec, Transformer, and DWS baselines are taken from the DWSNets evaluation (Navon et al., 2023); these results were originally reported on the MNIST and Fashion-MNIST INR benchmarks. For methods evaluated on CIFAR-10 and augmented CIFAR-10, we additionally use the comparison reported by ScaleGMN (Kalogeropoulos et al., 2024), which includes DWS, NFN, NG-GNN, ScaleGMN, and ScaleGMN-B on all four INR benchmarks. We further include results from Neural Graphs (Kofinas et al., 2024), NFT (Zhou et al., 2023b), Monomial-NFN (Tran et al., 2024), Quasi-Equivariant Monomial-NFN (Tran et al., 2026), MAGEP-NFN (Vo et al., 2025), and ProbeGen (Kahana et al., 2025). Results unavailable in the corresponding source are marked with "–".

D.3 Full Test-Accuracy Prediction Comparison

Table 5 provides an extended comparison on the neural-network test-accuracy prediction benchmarks. We include a broad range of approaches for learning from neural-network weights. MLP is a simple baseline operating directly on flattened network parameters, while StatNN (Unterthiner et al., 2020) predicts performance from summary statistics of the weights and biases. SNE (Andreis et al., 2024) represents neural networks using a hierarchical set-based encoder. We additionally include the permutation-equivariant NFN variants, NFN_{NP} and NFN_{HNP} (Zhou et al., 2023a), as well as the scale- and permutation-aware ScaleGMN and ScaleGMN-B architectures (Kalogeropoulos et al., 2024).

We further compare with graph-based weight-space models. NG-GNN and NG-T (Kofinas et al., 2024) represent the input network as a graph and process it using a graph neural network or relational Transformer, respectively. DNG-Encoder (Wu et al., 2026) extends this approach by constructing a dynamic neural graph that models inference-time information flow. We also report the probe-enhanced Neural Graphs variants with 64 and 128 probes, as well as vanilla probing with 64 and 128 probes, following the evaluation of ProbeGen (Kahana et al., 2025). Finally, we include NFT (Zhou

Table 4: Classification accuracy (↑) on INR weight-space benchmarks.

Method	MNIST	FMNIST	CIFAR-10	CIFAR-10 Aug
StatNN	0.398 ± 0.001	0.418 ± 0.002	–	–
DWS	0.857 ± 0.006	0.671 ± 0.003	0.3445 ± 0.004	0.4127 ± 0.001
NFN	0.791 ± 0.008	0.689 ± 0.006	0.3341 ± 0.001	0.4660 ± 0.001
ScaleGMN	0.966 ± 0.002	0.808 ± 0.001	0.3882 ± 0.001	0.5695 ± 0.005
NG-GNN	0.914 ± 0.006	0.680 ± 0.002	0.3604 ± 0.004	0.4570 ± 0.002
NG-T	0.924 ± 0.003	0.727 ± 0.006	–	–
NFT	–	–	–	0.6340 ± 0.000
Neural Graphs(128 probes)	0.976 ± 0.001	0.745 ± 0.008	-	-
Monomial-NFN	–	–	0.3423 ± 0.003	–
Monomial-NFN Quasi	–	–	0.3532 ± 0.005	–
MAGEP-NFN	–	–	0.3718 ± 0.003	–
ProbeGen	0.984 ± 0.001	0.877 ± 0.003	0.573 ± 0.007	0.563 ± 0.003
HiddenProbe	0.986 ± 0.0006	0.886 ± 0.005	0.580 ± 0.002	0.606 ± 0.002

et al., 2023b), ProbeGen (Kahana et al., 2025), and our HiddenProbe model. Results that were not reported for a particular benchmark are marked with "–".

Method	MNIST	FMNIST	SVHN	CIFAR-WP	CIFAR-GS
MLP	0.878 ± 0.001	0.874 ± 0.001	0.809 ± 0.003	–	0.880 ± 0.000
StatNN	0.926 ± 0.000	0.915 ± 0.000	0.843 ± 0.000	0.719 ± 0.010	0.914 ± 0.000
SNE	0.941 ± 0.000	0.928 ± 0.001	0.858 ± 0.003	–	0.927 ± 0.000
NFN_{NP}	0.937 ± 0.000	0.922 ± 0.001	0.856 ± 0.001	–	0.922 ± 0.001
NFN_{HNP}	0.942 ± 0.001	0.935 ± 0.000	0.931 ± 0.005	–	0.934 ± 0.001
NFT	–	–	0.858 ± 0.000	–	0.926 ± 0.001
ScaleGMN	–	–	–	–	0.941 ± 0.006
ScaleGMN-B	–	–	–	–	0.941 ± 0.000
NG-GNN	–	–	0.863 ± 0.002	0.804 ± 0.009	0.930 ± 0.001
NG-T	–	–	0.872 ± 0.001	0.817 ± 0.007	0.935 ± 0.000
DNG-Encoder	–	–	0.867 ± 0.002	0.8743 ± 0.0021	0.936 ± 0.000
Neural Graphs (64 probes)	–	–	–	0.888 ± 0.009	0.938 ± 0.001
Neural Graphs (128 probes)	–	–	–	0.885 ± 0.005	0.938 ± 0.000
Vanilla Probing (64 probes)	–	–	–	0.885 ± 0.008	0.933 ± 0.001
Vanilla Probing (128 probes)	–	–	–	0.889 ± 0.008	0.936 ± 0.001
ProbeGen	0.955 ± 0.000	0.945 ± 0.004	0.873 ± 0.005	0.932 ± 0.006	0.957 ± 0.001
HiddenProbe	0.966 ± 0.004	0.957 ± 0.002	0.957 ± 0.003	0.940 ± 0.006	0.965 ± 0.000

Table 5: Full comparison of test-accuracy prediction performance measured by Kendall's τ (↑). Results that were not reported for a particular benchmark are denoted by "–".

D.4 Transformer Accuracy Prediction Across Thresholds

Following the evaluation protocol of Tran et al. (2026), we additionally evaluate HiddenProbe on subsets of networks obtained by applying accuracy thresholds of 20%, 40%, 60%, and 80%. Tables 6 and 7 report the results for MNIST-Transformers and AGNews-Transformers, respectively. HiddenProbe consistently outperforms the previous state-of-the-art method, Transformer-NFN Quasi, achieving the highest Kendall's τ across all evaluated thresholds on both benchmarks.

Accuracy threshold	No threshold	20%	40%	60%	80%
MLP	0.866 ± 0.002	0.873 ± 0.001	0.874 ± 0.003	0.874 ± 0.006	0.873 ± 0.007
STATNN (Unterthiner et al., 2020)	0.881 ± 0.001	0.872 ± 0.001	0.868 ± 0.001	0.860 ± 0.001	0.856 ± 0.001
XGBoost (Chen & Guestrin, 2016)	0.860 ± 0.002	0.839 ± 0.004	0.869 ± 0.003	0.846 ± 0.001	0.884 ± 0.001
LightGBM (Ke et al., 2017)	0.858 ± 0.002	0.835 ± 0.001	0.847 ± 0.001	0.822 ± 0.001	0.830 ± 0.001
Random Forest (Breiman, 2001)	0.772 ± 0.002	0.758 ± 0.004	0.769 ± 0.001	0.752 ± 0.001	0.759 ± 0.001
Transformer-NFN (Tran et al., 2025)	0.905 ± 0.002	0.899 ± 0.001	0.895 ± 0.001	0.895 ± 0.002	0.888 ± 0.002
Transformer-NFN large (Tran et al., 2026)	0.907 ± 0.001	0.904 ± 0.002	0.897 ± 0.002	0.897 ± 0.002	0.890 ± 0.001
Transformer-NFN Quasi (Tran et al., 2026)	0.911 ± 0.001	0.905 ± 0.001	0.898 ± 0.002	0.897 ± 0.001	0.892 ± 0.001
ProbeGen	0.898 ± 0.001	0.885 ± 0.001	0.873 ± 0.001	0.872 ± 0.001	0.863 ± 0.001
HiddenProbe	0.920 ± 0.003	0.910 ± 0.002	0.904 ± 0.001	0.900 ± 0.002	0.898 ± 0.002

Table 6: Performance measured by Kendall's τ on the MNIST-Transformers dataset. Uncertainties indicate the standard error over 5 seeds.

Accuracy threshold	No threshold	20%	40%	60%	80%
MLP	0.879 ± 0.006	0.875 ± 0.001	0.841 ± 0.012	0.842 ± 0.001	0.862 ± 0.006
STATNN (Unterthiner et al., 2020)	0.841 ± 0.002	0.839 ± 0.003	0.812 ± 0.003	0.813 ± 0.001	0.812 ± 0.001
XGBoost (Chen & Guestrin, 2016)	0.859 ± 0.001	0.852 ± 0.002	0.872 ± 0.002	0.874 ± 0.001	0.872 ± 0.001
LightGBM (Ke et al., 2017)	0.835 ± 0.001	0.845 ± 0.001	0.837 ± 0.001	0.835 ± 0.001	0.820 ± 0.001
Random Forest (Breiman, 2001)	0.774 ± 0.003	0.801 ± 0.001	0.797 ± 0.001	0.798 ± 0.002	0.773 ± 0.001
Transformer-NFN (Tran et al., 2025)	0.910 ± 0.001	0.908 ± 0.001	0.897 ± 0.001	0.896 ± 0.001	0.890 ± 0.001
Transformer-NFN large (Tran et al., 2026)	0.913 ± 0.001	0.910 ± 0.002	0.898 ± 0.002	0.898 ± 0.001	0.893 ± 0.002
Transformer-NFN Quasi (Tran et al., 2026)	0.914 ± 0.001	0.913 ± 0.002	0.901 ± 0.001	0.903 ± 0.002	0.896 ± 0.001
ProbeGen	0.885 ± 0.001	0.883 ± 0.002	0.860 ± 0.002	0.857 ± 0.001	0.854 ± 0.001
HiddenProbe	0.916 ± 0.002	0.918 ± 0.002	0.906 ± 0.002	0.904 ± 0.002	0.901 ± 0.002

Table 7: Performance measured by Kendall's τ on the AGNews-Transformers dataset. Error bars indicate the standard error over 5 seeds.

D.5 Number of parameters comparison

Here we show the comparison of the number of parameters for different popular models:

Method	INR classification			CNN regression	Transformer zoo
	MNIST	FMNIST	CIFAR-10	CIFAR-10-GS	MNIST-Trans.
MLP (Monomial-NFN setup)	2.00	2.00	2.00	–	–
MLP (Transformer setup)	–	–	–	–	0.933
StatNN	–	–	–	1.06	0.203
DWS	–	0.553	–	–	–
NG-GNN (no probes)	–	0.331	–	–	–
NFN_{NP}	15.0	15.0	16.0	2.03	–
NFN_{HNP}	22.0	22.0	42.0	2.81	–
ScaleGMN	–	1.128	–	–	–
ScaleGMN-B	–	1.689	–	–	–
Monomial-NFN	22.0	20.0	16.0	0.25	–
Monomial-NFN Quasi	22.2	20.5	16.3	0.26	–
Transformer-NFN	–	–	–	–	1.812
Transformer-NFN Large	–	–	–	–	2.857
Transformer-NFN Quasi	–	–	–	–	1.894
ProbeGen	0.4	0.4	0.45	1.0	1.8
HiddenProbe	0.4	0.4	0.45	1.0	1.8

Table 8: Number of trainable parameters (millions, M) for published baselines on three INR classification tasks, CIFAR-10-GS regression, and MNIST-Transformers. INR and CNN counts are taken from Tran et al. (2024; 2026) and the ScaleGMN supplement (Kalogeropoulos et al., 2024); MNIST-Transformers counts are from Tran et al. (2025; 2026). Published model configurations may vary across sources and need not exactly match those used in our accuracy comparisons. ProbeGen and HiddenProbe counts are provided by the authors. A dash indicates that a count has not been verified for the specified task.

D.6 Time measurement

Here we show the time comparison of popular deep weight space models and ProbeGen and HiddenProbe on the MNIST INR benchmark.

Table 9: Runtime on the full MNIST-INR training split (55,000 target INRs, batch size 32).

Method	Time (s)	Relative to MLP
MLP	103.461	1.00×
NFN	90.755	0.88×
ProbeGen	109.122	1.05×
HiddenProbe	155.508	1.50×
NFT	164.774	1.59×
DWS	174.766	1.69×

D.7 Tuning budget

For benchmarks on which ProbeGen had not previously been evaluated, we ran ProbeGen ourselves. We used the same hyper-parameter selection procedure for ProbeGen and for HiddenProbe, ensuring that both methods were tuned under the same validation-based protocol. In all cases, we used the original benchmark train/validation/test splits and selected hyper-parameters using only the validation split.

For both methods, we fixed the batch size to 32, the number of learned probes to 128, and the weight decay to zero. All models were optimized with Adam.

For Transformer experiments, we used 256 probes.

For each dataset, we performed a 27-configuration sweep over the learning rate and learning-rate scheduler. We considered learning rates

$$\{3 \times 10^{-4},\, 5 \times 10^{-4},\, 7 \times 10^{-4}\}.$$

For the ReduceLROnPlateau scheduler, we evaluated all combinations of

$$\text{patience} \in \{3, 5\},\ \text{factor} \in \{0.2, 0.3, 0.5, 0.7\},$$

for each learning rate, resulting in 24 configurations. We additionally evaluated a cosine-annealing schedule for each of the three learning rates, yielding 3 further configurations and 27 configurations in total.

Each hyper-parameter configuration was first trained for 20 epochs using seed 0. For test-accuracy prediction, configurations were ranked exclusively according to validation Kendall's τ; test performance was not used for model selection. The best configuration for each method was then retrained using five seeds, $0, \ldots, 4$, for 60 epochs, and we report the mean and standard error of the resulting test Kendall's τ.

For CIFAR-10 INR classification, we use the same 27-configuration optimization grid for both ProbeGen and HIDDENPROBE, but select the best configuration according to validation classification accuracy. The selected configuration is subsequently evaluated over five seeds. The final CIFAR-10 INR runs are trained for 30 epochs.