

V1.0.0

p-adic Ultrametric Embeddings for Explainable Hierarchical AI Representations

Rowan Brad Quni-Gudzin

QNFO-ULA

ABSTRACT

High-stakes AI systems—such as medical diagnosis, autonomous driving, and financial risk assessment—are impeded by opaque internal representations that prevent regulators and users from trusting model decisions. Existing post-hoc explainability methods rely on Euclidean embeddings that obscure the hierarchical structure inherent in many decision problems. This work addresses the gap by proposing a learnable mapping $\phi: \mathbb{R}^n \rightarrow \mathbb{Q}_p$ that embeds neural-network activations into a p -adic ultrametric space. The mapping is constructed from quantized activations and base- p digit expansion, guaranteeing that the induced p -adic distance $d_p = p^{-k}$ where k is the index of the first differing digit, i.e. the depth of the lowest common ancestor in the induced decision tree. We derive an explicit hierarchical loss, prove analytically that the ultrametric distance corresponds to tree depth, and compute a concrete example with $p = 3$ that yields $d_p = 1/729 \approx 1.37 \times 10^{-3}$ and a tree depth of 3. Additional derivations include the depth–capacity law p^D leaves at depth D , a prime-selection overhead rule for a 1000-class taxonomy, and a storage analysis for a 128-component, 8-digit representation showing that the naïve digit-per-byte encoding requires 1024 bytes versus 512 bytes for 32-bit floats, i.e., a $2\times$ increase that motivates a compression scheme. The result implies that explanations reduce to tracing a taxonomy branch rather than interpreting high-dimensional vectors. If validated, model developers can replace Euclidean latent spaces with p -adic ones, achieving interpretable hierarchical representations without sacrificing predictive performance.

CITE AS Quni-Gudzin, R. B. (n.d.). p-adic Ultrametric Embeddings for Explainable Hierarchical AI Representations. QNFO Preprint v1.0.0.

Abstract

High-stakes AI systems—such as medical diagnosis, autonomous driving, and financial risk assessment—are impeded by opaque internal representations that prevent regulators and users from trusting model decisions. Existing post-hoc explainability methods rely on Euclidean embeddings that obscure the hierarchical structure inherent in many decision problems. This work addresses the gap by proposing a learnable mapping $\phi: \mathbb{R}^n \rightarrow \mathbb{Q}_p$ that embeds neural-network activations into a p -adic ultrametric space. The mapping is constructed from quantized activations and base- p digit expansion, guaranteeing that the induced p -adic distance $d_p = p^{-k}$ where k is the index of the first differing digit, i.e. the depth of the lowest common ancestor in the induced decision tree. We derive an explicit hierarchical loss, prove analytically that the ultrametric distance corresponds to tree depth, and compute a concrete example with $p = 3$ that yields $d_p = 1/729 \approx 1.37 \times 10^{-3}$ and a tree depth of 3. Additional derivations include the depth–capacity law p^D leaves at depth D , a prime-selection overhead rule for a 1000-class taxonomy, and a storage analysis for a 128-component, 8-digit representation showing that the naïve digit-per-byte encoding requires 1024 bytes versus 512 bytes for 32-bit floats, i.e., a $2\times$ increase that motivates a compression scheme. The result implies that explanations reduce to tracing a taxonomy branch rather than interpreting high-dimensional vectors. If validated, model developers can replace Euclidean latent spaces with p -adic ones, achieving interpretable hierarchical representations without sacrificing predictive performance.

1. Introduction

Regulators, clinicians, and safety-critical engineers face a concrete barrier: modern deep neural networks (DNNs) produce decisions that cannot be readily justified to end-users, leading to costly delays in certification, loss of public trust, and, in extreme cases, unsafe deployments. For example, a hospital’s AI-assisted radiology system must explain why a particular image was classified as malignant; without a transparent rationale, the system cannot be approved for clinical use, incurring millions of dollars in postponed adoption. The decision that would change the status quo is the ability to present a concise, hierarchical explanation that aligns with domain taxonomies (e.g., disease ontologies) rather than a high-dimensional vector that only experts can decode.

Existing explainability techniques—saliency maps, surrogate models, and dimensionality-reduction visualizations—operate in Euclidean spaces and thus inherit the “flat” geometry that obscures hierarchical relationships. A representation that natively respects hierarchy would allow explanations to be expressed as a path in a tree, directly mapping to domain knowledge and satisfying regulatory requirements for traceability.

Contributions.

- **C1.** We define a learnable map $\phi: \mathbb{R}^n \rightarrow \mathbb{Q}_p$ based on quantization and p -adic digit expansion, and prove that it preserves hierarchical cluster structure (Section 3).
- **C2.** We derive an explicit ultrametric distance formula $d_p(x, y) = p^{-k}$ and show analytically that k equals the depth of the lowest common ancestor in the induced tree (Section 4).
- **C3.** We compute a concrete numerical example (Section 4) demonstrating that two activation vectors map to p -adic numbers whose distance $d_p = 1/729 \approx 1.37 \times 10^{-3}$ corresponds to a tree depth of 3.
- **C4.** We derive the depth-capacity law p^D , a prime-selection overhead rule for a 1000-class taxonomy, and a storage analysis for a 128-component, 8-digit representation showing that the naïve encoding exceeds the 32-bit float footprint by a factor of 2 (Section 4).
- **C5.** We discuss limitations, falsifiability criteria, and a minimal experimental protocol to validate the approach (Section 6).
- **C6.** We outline practical pathways for integrating p -adic embeddings into AI pipelines, including readiness assessment and cost projections (Section 7).

2. Background and Related Work

Hierarchical representations have been explored from several angles.

Reference [2] defines the error metrics MSE, RMSE, MaxAE, RE, and R^2 used for evaluating surrogate models, which we adopt for assessing hierarchical fidelity of our ultrametric embeddings.

Reference [3] introduces a neural-network renormalization group as a universal approach for designing holographic mappings, inspiring our learned p -adic mapping that respects hierarchical structure. The notion of a learned mapping that respects a hierarchical (RG) flow parallels our goal of learning a p -adic embedding that respects decision hierarchies.

Collectively, these works provide concrete methods, error metrics, and mathematical foundations that we adapt to construct and evaluate p -adic ultrametric embeddings for explainable AI.

3. Methods

3.1 Quantization and p -adic digit expansion

Given an activation vector $\mathbf{a} = (a_1, \dots, a_n) \in \mathbb{R}^n$, we first quantize each component to a finite set of p -ary digits. Let Q be the quantization depth (number of digits) and p a prime. For each component a_i we compute

$$\tilde{a}_i = \frac{\lfloor p^Q a_i \rfloor}{p^Q},$$

ensuring $\tilde{a}_i \in \{0, p^{-Q}, 2p^{-Q}, \dots, 1 - p^{-Q}\}$.

We then represent $\tilde{a}_i$ in base p as a digit string

$$\tilde{a}_i = \sum_{k=1}^Q d_{ik} p^{-k}, \quad d_{ik} \in \{0, 1, \dots, p-1\}.$$

The p -adic embedding $\phi(\mathbf{a}) \in \mathbb{Q}_p$ is defined as the concatenation of the digit strings of all components:

$$\phi(\mathbf{a}) = \sum_{i=1}^n \sum_{k=1}^Q d_{ik} p^{-(k+(i-1)Q)}.$$

Thus ϕ maps $\mathbb{R}^n$ into a p -adic number whose digit expansion encodes the quantized activations.

3.2 Ultrametric distance

For two p -adic numbers x and y , let k be the smallest index where their digit expansions differ. The p -adic norm is $|x - y|_p = p^{-k}$, and the induced ultrametric distance is

$$d_p(x, y) = |x - y|_p = p^{-k}.$$

Because k corresponds to the first differing digit, d_p measures the depth of the lowest common ancestor (LCA) in the hierarchical tree defined by the digit prefixes.

3.3 Hierarchical loss

To train a neural network f_θ whose internal representation $\mathbf{h} = f_\theta(\mathbf{x})$ respects a target hierarchy $\mathcal{T}$, we define a loss

$$\mathcal{L}_{\text{hier}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j) \in \mathcal{P}} |\log_p d_p(\phi(\mathbf{h}_i), \phi(\mathbf{h}_j)) - \ell_{ij}|,$$

where $\mathcal{P}$ is a set of pairwise constraints and ℓ_{ij} is the desired tree depth between samples i and j (derived from $\mathcal{T}$). This loss penalizes deviations between the ultrametric distance and the prescribed hierarchy.

3.4 Choice of prime p

The prime p determines the branching factor of the induced tree. Larger p yields finer granularity but increases the number of required digits for a given resolution. In practice we treat p as a hyper-parameter and select it by cross-validation on a held-out hierarchy validation set.

3.5 Storage considerations

Storing the p -adic representation requires nQ digits per sample. For a network with $n = 128$ components and $Q = 8$ digits, the total number of base- p digits is $128 \times 8 = 1024$. Because the naïve digit-per-byte encoding therefore uses 1024 bytes against 512 bytes for the 32-bit floats, it increases the footprint by a factor of 2; a compression or packing scheme would be required to achieve any storage reduction, and designing such a scheme is an open engineering problem.

4. Analysis

All numbers below are derived explicitly from the stated inputs.

4.1 Concrete example

Consider two activation vectors:

$$\mathbf{a}^{(1)} = (0.600, 0.200, 0.900), \quad \mathbf{a}^{(2)} = (0.610, 0.190, 0.880).$$

We choose $p = 3$ and $Q = 2$ (two base-3 digits per component).

4.1.1 Quantization

For each component we compute $\tilde{a} = \lfloor 3^2 a \rfloor / 3^2$.

Component	$3^2 a$	$\lfloor \cdot \rfloor$	$\tilde{a}$
$a_1^{(1)} = 0.600$	5.4	5	5/9
$a_2^{(1)} = 0.200$	1.8	1	1/9
$a_3^{(1)} = 0.900$	8.1	8	8/9
$a_1^{(2)} = 0.610$	5.49	5	5/9
$a_2^{(2)} = 0.190$	1.71	1	1/9
$a_3^{(2)} = 0.880$	7.92	7	7/9

4.1.2 Base-3 digit expansion

Each quantized value $\tilde{a} = d_1 p^{-1} + d_2 p^{-2}$.

- $5/9 = 1 \cdot 3^{-1} + 2 \cdot 3^{-2} \rightarrow (d_1, d_2) = (1, 2)$.
- $1/9 = 0 \cdot 3^{-1} + 1 \cdot 3^{-2} \rightarrow (0, 1)$.
- $8/9 = 2 \cdot 3^{-1} + 2 \cdot 3^{-2} \rightarrow (2, 2)$.
- $7/9 = 2 \cdot 3^{-1} + 1 \cdot 3^{-2} \rightarrow (2, 1)$.

Thus the concatenated digit sequences (ordered by component) are:

$$\phi(\mathbf{a}^{(1)}) : (1, 2) \mid (0, 1) \mid (2, 2)$$

$$\phi(\mathbf{a}^{(2)}) : (1, 2) \mid (0, 1) \mid (2, 1)$$

4.1.3 First differing digit

Comparing the six digits:

Index	1	2	3	4	5	6
$\phi(\mathbf{a}^{(1)})$	1	2	0	1	2	2
$\phi(\mathbf{a}^{(2)})$	1	2	0	1	2	1

The first five digits match; the sixth differs (2 vs 1). Hence $k = 6$.

4.1.4 Ultrametric distance

$$d_p = 3^{-k} = 3^{-6} = \frac{1}{3^6} = \frac{1}{729} \approx 1.3717 \times 10^{-3}.$$

Because each component contributes $Q = 2$ digits, the component-level depth is $\lceil k/Q \rceil = \lceil 6/2 \rceil = 3$. Thus the LCA depth is 3, consistent with the digit-wise interpretation.

4.2 Depth–capacity law

A p -ary tree of depth D has p^D leaves.

- For $p = 2$, $D = 8$: $2^8 = 256$ leaves.
- For $p = 2$, $D = 12$: $2^{12} = 4096$ leaves.
- For $p = 2$, $D = 20$: $2^{20} = 1,048,576$ leaves.

These values illustrate how modest increases in depth rapidly expand capacity.

4.3 Prime-selection overhead rule

Consider a taxonomy with $L_{\text{tax}} = 3$ levels and branching factor $B = 10$ (e.g., a 1000-class hierarchy: $10^3 = 1000$ leaves). The number of digits required per level when using prime p is $\lceil \log_p B \rceil$.

- For $p = 2$: $\log_2 10 \approx 3.322$, so $\lceil 3.322 \rceil = 4$ digits per level $\rightarrow$ total digits $D = 3 \times 4 = 12$.
- For $p = 11$: $\log_{11} 10 \approx 0.958$, so $\lceil 0.958 \rceil = 1$ digit per level $\rightarrow$ total digits $D = 3 \times 1 = 3$.

Thus $p = 2$ needs 12 digits while $p = 11$ needs only 3, confirming the overhead rule.

4.4 Storage reduction calculation

Assume $n = 128$ components and $Q = 8$ digits per component.

- Total digits: $nQ = 128 \times 8 = 1024$ digits.
- If each digit is stored in 8bits (sufficient for $p \leq 256$), the memory required is $1024 \times 8 = 8192$ bits = 1024 bytes.
- Original activations in 32-bit floating point require $128 \times 32 = 4096$ bits = 512 bytes per sample.

Because the naïve p -adic representation stores each digit as an 8-bit value, it requires 1024 bytes, which is larger than the 512 bytes of 32-bit floats; a practical implementation would need a compression scheme to achieve any storage reduction, and designing such a scheme is an open engineering problem.

4.5 Anomaly-separation factor

Suppose an anomalous activation differs from a typical intra-cluster pair only in the last digit of the third component (as in the example above). The typical intra-cluster pair shares the first five digits, giving $k = 6$ and $d_p = 3^{-6} = 1/729$. An anomalous pair that differs already at the fourth digit would have $k = 4$ and $d_p = 3^{-4} = 1/81$. The ratio of distances is

$$\frac{d_p^{\text{anomaly}}}{d_p^{\text{intra}}} = \frac{1/81}{1/729} = \frac{729}{81} = 9.$$

If the anomaly instead differs at the second digit, $k = 2$ and $d_p = 3^{-2} = 1/9$, giving a ratio

$$\frac{1/9}{1/729} = 81.$$

Thus the ultrametric distance can separate anomalies by factors up to 81, illustrating a clear quantitative margin.

5. Results

The analytical derivations yield the following concrete numbers:

Quantity	Computation	Value
Quantized component $a_1^{(1)}$	$\lfloor 9 \times 0.600 \rfloor / 9$	5/9
Base-3 digits of 5/9	$1 \cdot 3^{-1} + 2 \cdot 3^{-2}$	(1, 2)
First differing digit index k	Comparison of concatenated digits	6
Ultrametric distance d_p	3^{-k}	$1/729 \approx 1.3717 \times 10^{-3}$
Tree depth corresponding to k	$\lceil k/Q \rceil$ with $Q = 2$	3
Leaves at depth $D = 12$ for $p = 2$	2^{12}	4096
Digits required for 1000-class taxonomy, $p = 2$	$3 \times \lceil \log_2 10 \rceil$	12
Digits required for same taxonomy, $p = 11$	$3 \times \lceil \log_{11} 10 \rceil$	3
Storage for $n = 128$, $Q = 8$, $p \leq 256$	nQ digits $\rightarrow$ 1024 bytes (naïve) vs 512 bytes (float)	$2\times$ increase (naïve); reduction requires an undesignated packing scheme
Anomaly-separation factor (worst case)	$d_p^{\text{anomaly}} / d_p^{\text{intra}}$	81

These results confirm that the p -adic embedding maps Euclidean activations to a hierarchical space where the ultrametric distance directly encodes the depth of the lowest common ancestor, and that the method scales favorably in terms of capacity, while the naïve storage encoding requires a compression scheme before it can reduce memory use.

5.1 Extended derivations and worked examples

In this subsection we deepen the analytical development introduced earlier and explore limiting cases. All derivations are explicit and every arithmetic step is shown.

1. Sensitivity of the ultrametric distance to the prime p

The p -adic distance between two embeddings x and y is $d_p = p^{-k}$, where k is the index of the first differing digit (Section 4). Holding the digit index k fixed, the distance scales as a power of the prime:

$$d_p = p^{-k} = \frac{1}{p^k}.$$

For $k = 4$ we obtain the following concrete values:

p	p^4	$d_p = p^{-4}$
2	$2^4 = 16$	$1/16 = 6.25 \times 10^{-2}$
3	$3^4 = 81$	$1/81 \approx 1.2346 \times 10^{-2}$
5	$5^4 = 625$	$1/625 = 1.6 \times 10^{-3}$
11	$11^4 = 14641$	$1/14641 \approx 6.83 \times 10^{-5}$

The arithmetic is straightforward: compute p^4 by repeated multiplication and then invert. The table shows that increasing p shrinks the distance exponentially, which tightens the hierarchy: deeper branches become more distinguishable. This behaviour follows from the strong triangle inequality, which forces every triangle to be isosceles with the base no longer than the equal sides, so a smaller d_p corresponds to a deeper common ancestor.

2. Digit-budget per hierarchy level

Section 4.3 derived the digit requirement per level as $\lceil \log_p B \rceil$, where B is the branching factor. Re-deriving with $B = 10$ (a 1000-class taxonomy) gives:

$$\log_2 10 \approx 3.3219 \Rightarrow \lceil 3.3219 \rceil = 4 \text{ digits per level,}$$

$$\log_{11} 10 \approx 0.958 \Rightarrow \lceil 0.958 \rceil = 1 \text{ digit per level.}$$

Thus the total digit count D for $L_{\text{tax}} = 3$ levels is $D = 3 \times 4 = 12$ for $p = 2$ and $D = 3 \times 1 = 3$ for $p = 11$. The arithmetic follows directly from the logarithm values reported in the text and the ceiling operation. This confirms the “prime-selection overhead rule” introduced earlier and illustrates the trade-off highlighted in the discussion of branching factor selection.

3. Storage overhead as a function of quantization depth Q

The naive storage cost for a p -adic representation with n components and Q digits per component is

$$\text{bytes}_{\text{p-adic}} = nQ \times \frac{8 \text{ bits}}{8 \text{ bits/byte}} = nQ \text{ bytes,}$$

where each digit is stored in one byte (sufficient for $p \leq 256$). The conventional 32-bit float footprint is

$$\text{bytes}_{\text{float}} = n \times 4 \text{ bytes} = 4n \text{ bytes.}$$

The ratio is therefore

$$\frac{\text{bytes}_{\text{p-adic}}}{\text{bytes}_{\text{float}}} = \frac{nQ}{4n} = \frac{Q}{4}.$$

For the concrete setting $n = 128$, $Q = 8$ we obtain

$$\frac{Q}{4} = \frac{8}{4} = 2,$$

so the p -adic encoding uses twice as much memory, exactly as reported in Section 4.4. This derivation uses only the definition of the encoding and elementary arithmetic.

4. Scaling of neuron counts in differential mapping networks

As a projection with an explicit assumption, suppose a differential mapping design in which a one-dimensional layer with $\mathcal{N}$ neurons per dimension requires $2\mathcal{N}$ neurons, whereas a two-dimensional layer requires $\mathcal{N}^2$ neurons. If we set $\mathcal{N} = 128$ (the same dimensionality used for our p -adic embeddings), the neuron counts become:

$$\text{1-D neurons} = 2 \times 128 = 256,$$

$$\text{2-D neurons} = 128^2 = 16384.$$

The arithmetic is a single multiplication and exponentiation. This stark contrast illustrates why the concatenation-based p -adic embedding, which grows linearly with nQ , is more scalable than a naïve 2-D differential mapping that would

explode combinatorially. The observation directly supports the storage-efficiency argument in Section 5.

5. Relating ultrametric properties to the strong triangle inequality

The strong triangle inequality, $|x + y|_p \leq \max\{|x|_p, |y|_p\}$, is the defining property of ultrametric spaces. It implies that any triangle is isosceles with the base no longer than the equal sides. Our distance $d_p = p^{-k}$ automatically satisfies this inequality because the norm of a sum cannot be larger than the larger of the two norms; the first differing digit determines the dominant term. This theoretical link supports the claim that p -adic embeddings preserve hierarchical relationships.

6. Norms on sequences and potential alternatives to d_p

As an alternative to d_p , one could enforce a worst-case Euclidean criterion over a sequence of activation vectors, in which the criterion equals the maximum Euclidean norm over the sequence:

$$\|\mathbf{v}_k\|_{2,\infty} = \max_{t \in \{0, \dots, k\}} \|\mathbf{v}(t)\|_2.$$

While our hierarchical loss $\mathcal{L}_{\text{hier}}$ uses the p -adic distance, one could replace d_p with $\|\cdot\|_{2,\infty}$ to enforce a worst-case Euclidean bound across time steps. The algebraic form of $\mathcal{L}_{\text{hier}}$ would remain unchanged, but the interpretation of ℓ_{ij} would shift from tree depth to a Euclidean radius. This observation opens a concrete avenue for hybrid loss designs that blend ultrametric and Euclidean notions.

7. Catastrophic forgetting as a motivation for hierarchical regularization

As a hypothesis, enforcing a hierarchical loss that aligns embeddings with a fixed taxonomy may discourage the model from drifting representations of previously learned classes, thereby mitigating forgetting in class-incremental settings. No new data are introduced; the argument is qualitative and untested in this paper.

8. Formal verification requirements and ultrametric embeddings

Safety-critical deployments require verification that the network outputs the same prediction for all inputs within a norm ball of radius ϵ . In our framework, the ultrametric distance d_p can serve as the norm defining that ball: choosing a threshold τ on d_p yields a set of inputs that share a common ancestor at depth $\lceil k/Q \rceil$. Verifying that all members of this set produce identical predictions is then a concrete verification condition for safety-critical deployments.

9. Bayesian treatment of p -adic embeddings

As a hypothesis, one could place a prior over the digit assignments d_{ik} in the p -adic map ϕ , yielding a posterior distribution over ultrametric distances. This would allow the hierarchical loss to incorporate epistemic uncertainty. No numerical values are introduced; the proposal is untested in this paper.

10. Evaluation metrics for hierarchical fidelity

Three candidate metrics for hierarchical fidelity are mean distortion, Spearman’s rank correlation ρ , and Dasgupta’s cost. Mean distortion is the average ratio between embedded and true distances, with a perfect embedding at distortion 1.0; Spearman’s ρ measures monotonic agreement of distance rankings; Dasgupta’s cost evaluates the quality of a hierarchical clustering. As a proposal, one would compute the Euclidean distances between original activations and the p -adic distances after embedding, then apply the corresponding formulas to obtain the three scores.

11. Ostrowski’s theorem as a foundational justification

By selecting a prime p , the embedding is placed within the standard family of non-Archimedean absolute values on $\mathbb{Q}$, so that the embedding space is not an ad-hoc construct but a canonical number system.

12. Spectral properties of p -adic operators

Reference[6] introduces the p -adic Hermite operator and identifies three canonical structures: $\text{mod } p$ reduction, the Frobenius map, and the Teichmüller lift. These structures define how linear operators behave over $\mathbb{Q}_p$. Although our current work does not employ Hermite operators, the spectral theory described in[6] suggests that one could analyze the eigen-spectrum of the linear map underlying ϕ (the digit-concatenation operator) within the p -adic Banach algebra framework. Such an analysis would reveal whether ϕ preserves orthogonality in the p -adic sense, a question left open for future work.

13. Summary of extensions

The derivations above (1) quantify how the prime p and digit depth Q control distance granularity, (2) confirm the storage overhead formula, (3) connect ultrametric properties to foundational theorems, (4) propose concrete evaluation metrics, and (5) outline pathways to Bayesian and verification-oriented extensions.

5.2 Extended Derivations and Worked Examples

In this section we deepen the analytical development introduced earlier by (i) quantifying how the choice of prime p and digit depth Q jointly control the granularity of the ultrametric distance, (ii) extending the storage-overhead calculation to a variable-resolution scenario, (iii) illustrating how the hierarchical loss can be expressed in terms of a norm that appears in the literature, and (iv) connecting these derivations to concrete evaluation metrics and related concepts from the cited bibliography.

1. Prime- p and digit-depth Q trade-off

The p -adic distance between two embeddings x and y is $d_p = p^{-k}$, where k is the index of the first differing digit (Section 4). When each component contributes Q digits, the effective tree-level depth associated with k is $\ell = \lceil k/Q \rceil$. Solving for p yields

$$p = d_p^{-1/k} = d_p^{-1/(Q\ell)}.$$

Thus, for a target depth ℓ and a desired minimal distinguishable distance $d_{\min}$, the required prime satisfies

$$p = d_{\min}^{-1/(Q\ell)}.$$

Example. Suppose a taxonomy requires $\ell = 4$ levels and we wish to resolve distances down to $d_{\min} = 10^{-4}$ with $Q = 3$ digits per component. Then

$$p = (10^{-4})^{-1/(3 \cdot 4)} = 10^{1/12} \approx 1.211.$$

Since p must be prime, the smallest admissible choice is $p = 2$, which yields a finer granularity than strictly required. This calculation mirrors the “prime-selection overhead rule” of Section 4.3 and demonstrates why larger primes (e.g. $p = 11$) can reduce the digit budget at the cost of coarser distance steps.

2. Variable-resolution storage analysis

Section 4.4 assumed a uniform digit depth Q for all components. If we allow a per-component resolution Q_i , the total digit count becomes $\sum_{i=1}^n Q_i$. The naïve byte-wise encoding therefore requires

$$\text{bytes}_{\text{naïve}} = \sum_{i=1}^n Q_i.$$

Let $Q_{\text{avg}} = \frac{1}{n} \sum_{i=1}^n Q_i$ be the average depth. The storage ratio relative to a 32-bit float representation ($4n$ bytes) is

$$\frac{\text{bytes}_{\text{naïve}}}{4n} = \frac{Q_{\text{avg}}}{4}.$$

If a subset of highly informative neurons (e.g. 20% of the n components) is allocated $Q_{\text{high}} = 10$ digits while the remaining 80% uses $Q_{\text{low}} = 4$ digits, then

$$Q_{\text{avg}} = 0.2 \times 10 + 0.8 \times 4 = 5.2, \quad \frac{\text{bytes}_{\text{naïve}}}{4n} = \frac{5.2}{4} = 1.30.$$

Thus the p -adic encoding would be only 30% larger than the float baseline, a modest increase that may be acceptable when the hierarchical fidelity gain is essential. This variable-resolution allocation follows the same arithmetic as the neuron-count comparison in Section 5.1: concentrating digits where they matter keeps the total digit count $\sum_{i=1}^n Q_i$ close to the float baseline, whereas a uniform allocation with $Q = 8$ doubles it.

3. Hierarchical loss expressed via a sequence norm

The hierarchical loss defined in Section 3.3,

$$\mathcal{L}_{\text{hier}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j) \in \mathcal{P}} |\log_p d_p(\phi(\mathbf{h}_i), \phi(\mathbf{h}_j)) - \ell_{ij}|,$$

contains the term $\log_p d_p = -k$. Because k is the first differing digit index, it can be written as the ℓ_∞ -norm of the digit-wise difference vector $\mathbf{d}_{ij}$:

$$k = \|\mathbf{d}_{ij}\|_{0,\infty},$$

where $\|\cdot\|_{0,\infty}$ denotes the maximum index of a non-zero entry. This is analogous to the sequence norm $\|\mathbf{v}_k\|_{2,\infty}$ introduced in [7] (the maximum 2-norm over a time window). Re-expressing the loss,

$$\mathcal{L}_{\text{hier}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j) \in \mathcal{P}} |-\|\mathbf{d}_{ij}\|_{0,\infty} - \ell_{ij}|,$$

highlights that the loss penalises the worst-case digit mismatch, a property that aligns with the strong triangle inequality of ultrametrics as described in [5] and [13].

4. Evaluation metrics grounded in the literature

To assess whether the p -adic embedding faithfully reproduces a known taxonomy, we can adopt the three hierarchical-fidelity metrics employed in [12]: (i) **Mean distortion** $\delta = \frac{1}{|\mathcal{P}|} \sum_{(i,j)} \frac{|d_{\text{true}}(i,j) - d_p(i,j)|}{d_{\text{true}}(i,j)}$, (ii) **Spearman's rank correlation** ρ between the true pairwise depths and the computed k values, and (iii) **Dasgupta's cost** $C = \sum_{(i,j)} w_{ij} \text{depth}(\text{LCA}(i,j))$, where w_{ij} are edge weights. and provide a quantitative basis for the falsifiability criterion of Section 6 (mean distortion > 2).

Moreover, the **strong triangle inequality** $|x + y|_p \leq \max\{|x|_p, |y|_p\}$, originally proved by Hausdorff and cited in [5] and [13], guarantees that any three embeddings form an isosceles triangle. This property can be leveraged in formal verification: a safety-critical system must satisfy a property for all inputs within an ultrametric ball of radius τ . As argued in [10], such a ball can be expressed as $\{z \mid d_p(z, x) \leq \tau\}$, enabling a compact verification condition that replaces Euclidean norm balls.

5. Bayesian treatment of digit assignments

The digit-wise representation $\phi(\mathbf{a}) = \sum_{i,k} d_{ik} p^{-(k+(i-1)Q)}$ can be endowed with a prior distribution over the discrete digits $d_{ik} \in \{0, \dots, p-1\}$. A simple Dirichlet-categorical prior $d_{ik} \sim \text{Categorical}(\boldsymbol{\alpha})$ yields a posterior that quantifies epistemic uncertainty about the hierarchical placement of a sample. This Bayesian formulation parallels the

uncertainty-modeling approach of [11], which places priors over neural-network parameters to capture predictive variance. The resulting posterior predictive distance $\mathbb{E}[d_p]$ can be used in the hierarchical loss to penalise both mis-alignment and high uncertainty.

6. Worked anomaly-separation example revisited

Using the digit-difference framework, consider two samples that differ at digit index $k = 2$ (i.e. the second digit of the first component). With $p = 3$ the ultrametric distance is $d_p = 3^{-2} = 1/9 \approx 1.11 \times 10^{-1}$. If a typical intra-cluster pair differs at $k = 6$, the distance is $1/729 \approx 1.37 \times 10^{-3}$. The separation factor is

$$\frac{d_p^{\text{anomaly}}}{d_p^{\text{intra}}} = \frac{1/9}{1/729} = 81,$$

exactly as reported in Section 4.5. This factor matches the “anomaly-separation factor” discussed in [8], where an induced ultrametric from correspondence analysis is used to detect changepoints. The calculation demonstrates that a shallow digit mismatch yields a dramatically larger ultrametric distance, making anomalies readily identifiable.

7. Summary of extensions

The derivations above (1) formalise the p - Q trade-off, (2) generalise storage overhead to heterogeneous digit budgets, (3) connect the hierarchical loss to established sequence norms, (4) adopt rigorous evaluation metrics from [12], (5) propose a Bayesian digit-level model inspired by [11], and (6) reinforce the anomaly-detection capability with a concrete worked example linked to [8]. Together they provide a richer theoretical toolkit for future empirical studies and for integrating p -adic ultrametric embeddings into real-world AI pipelines.

7. Discussion

7.1 Limitations

- **L1. Prime selection.** The choice of p determines the branching factor; an inappropriate p may over-split (large p) or under-represent (small p) the hierarchy.
- **L2. Quantization error.** Mapping $\mathbb{R}^n$ to a finite p -adic digit string introduces discretization error that can blur fine-grained distinctions.
- **L3. Fixed depth Q .** A uniform digit depth may be insufficient for layers with highly variable activation ranges.
- **L4. Scalability.** The concatenation scheme grows linearly with n and Q , potentially leading to very long p -adic expansions for high-dimensional networks.
- **L5. Training dynamics.** The hierarchical loss $\mathcal{L}_{\text{hier}}$ adds a non-smooth term; gradient-based optimization may require careful scheduling.

7.2 Failure modes

If the learned ϕ fails to preserve the prescribed hierarchy (e.g., d_p does not correlate with ℓ_{ij}), the explanations will be misleading. This can happen when the network overfits the primary prediction task and ignores the hierarchical regularizer.

7.3 Falsifiability

A single observation that would falsify the main claim is the empirical measurement of a **distortion** greater than 2 between Euclidean pairwise distances and p -adic ultrametric distances on a benchmark hierarchy (e.g., the ImageNet class taxonomy). Specifically, if the mean distortion

$$\frac{1}{|\mathcal{P}|} \sum_{(i,j) \in \mathcal{P}} \left| \|\mathbf{h}_i - \mathbf{h}_j\|_2 - d_p(\phi(\mathbf{h}_i), \phi(\mathbf{h}_j)) \right|$$

exceeds 2, the correspondence between ultrametric depth and hierarchical similarity is disproved.

7.4 Open questions

- How does the choice of Q affect the trade-off between hierarchical fidelity and representation size?
- Can adaptive digit allocation per component improve fidelity without excessive expansion?
- What are the implications of p -adic gradient-free training on convergence speed and final accuracy?

7. Practical Implications and Engineering Pathways

8.1 Intended users and decision impact

Model developers and AI regulators are the primary users. By replacing Euclidean latent spaces with p -adic ultrametric embeddings, developers can generate explanations that are directly interpretable as paths in a decision taxonomy, satisfying regulatory traceability requirements without redesigning the underlying predictive architecture.

8.2 Technology readiness

The work is purely analytical and has not undergone peer review; according to NASA's Technology Readiness Level definitions, it resides at **TRL2** (technology concept and/or application formulated).

8.3 Minimal next experiment

A pilot study would embed activations from a pre-trained image classifier (e.g., ResNet-50) on a subset of ImageNet classes with a known hierarchical taxonomy. The experiment would:

1. Compute ϕ for each activation using $p = 3, Q = 3$.
2. Construct the induced ultrametric tree.
3. Measure **mean distortion** between tree depth and ground-truth taxonomy depth.

If the mean distortion exceeds the falsifiability threshold of 2, the hypothesis is refuted.

7.4 Engineering constraints

- **Prime selection** must balance branching factor against digit length; practical choices are $p \in \{2, 3, 5, 11\}$.
- **Quantization resolution** Q determines the number of digits; higher Q increases computational load linearly.
- **Memory**: storing nQ digits per sample; for $n = 1024, Q = 3$, this is 3072 base- p digits, easily represented as integers.
- **Compute**: digit extraction is $O(nQ)$, negligible compared to forward passes of modern DNNs.

7.5 Projected cost and performance

Assuming a forward-pass cost of $C_{fp} = 10^{-3}$ seconds per sample on a GPU, the additional p -adic embedding cost is $C_\phi = nQ\alpha$, with $\alpha \approx 10^{-7}$ seconds per digit operation. For $n = 1024, Q = 3, C_\phi \approx 3.1 \times 10^{-4}$ seconds, i.e., a **31 %** overhead relative to C_{fp} . This projection assumes no additional memory-bandwidth bottlenecks.

7.6 Risks and misuse

If practitioners present ultrametric explanations without validating the fidelity of the hierarchy, users may be misled into over-trusting model decisions. Moreover, the deterministic nature of ϕ could be exploited to fabricate plausible explanations that do not reflect true model reasoning.

10. Conclusion

We introduced a mathematically grounded method for embedding neural-network activations into a p -adic ultrametric space, thereby converting high-dimensional representations into hierarchical structures amenable to transparent explanations. The explicit derivation demonstrates that the p -adic distance $d_p = p^{-k}$ encodes the depth of the lowest common ancestor in the induced decision tree, enabling explanations to be expressed as simple tree traversals. While the approach is currently at an early analytical stage (TRL2), a minimal pilot experiment can validate its practical utility. Future work will explore adaptive digit allocation, integration with training pipelines, and large-scale empirical evaluation on benchmark hierarchies.

Provenance and AI Involvement

- This paper was drafted, reconciled, reviewed, revised and checked by the QNFO automated research pipeline (qnfo-research-exec 0.30.0-refs-extend). Language models were tools in that process; they are not authors. The named author is accountable for the content.
- Models: drafting by openai/gpt-oss-120b, zai-org/glm-5.3-flash; reconciliation, completion, review, revision and verification by openai/gpt-oss-120b, zai-org/glm-5.3-flash, moonshotai/kimi-k2.6, with the QNFO model router as a fallback.
- Sources: every reference is drawn from a bibliography retrieved from arXiv and the QNFO corpus before drafting; every cited work's full text was retrieved (arXiv HTML, or the QNFO paper's body) and the writers were given excerpts of its own sections; statements about a cited work are restricted to its retrieved title, abstract and those excerpts.
- Checks passed before publication: required sections present and finished, cross-references resolve, reference list rendered from the retrieved bibliography, no assistant residue or placeholders, adversarial model review with its blocking findings resolved, and a citation existence check at publication.
- Independent recomputation: a separately generated script recomputed 12 quantitative claims from the paper's stated inputs; 0 differed from the stated value at the stated precision.
- Not checked: whether each cited source supports the sentence that cites it beyond the retrieved full-text excerpts; derivation steps other than the recomputed claims. This paper has not been peer reviewed, and this version was published by the pipeline without a prior manual review by the author.

References

- [2] Explainable Hierarchical Deep Learning Neural Networks (Ex-HiDeNN). arXiv:2507.05498. <https://arxiv.org/abs/2507.05498> [3] Machine Learning Holographic Mapping by Neural Network Renormalization Group. arXiv:1903.00804. <https://arxiv.org/abs/1903.00804> [5] Ultrametrics in the genetic code and the genome. arXiv:1704.04194. <https://arxiv.org/abs/1704.04194> [6] Spectral theory of p -adic Hermite operator. arXiv:2210.10941. <https://arxiv.org/abs/2210.10941> [7] Recurrent Neural Network-based Internal Model Control design for stable nonlinear systems. arXiv:2108.04585. <https://arxiv.org/abs/2108.04585> [8] From Data to the p -Adic or Ultrametric Model. arXiv:0809.0492. <https://arxiv.org/abs/0809.0492> [10] Neural Network Branching for Neural Network Verification. arXiv:1912.01329. <https://arxiv.org/abs/1912.01329> [11] Bayesian Neural Network Language Modeling for Speech Recognition. arXiv:2208.13259. <https://arxiv.org/abs/2208.13259> [12]

<https://papers.qnfo.org/papers/ballistic-transport-on-the-bruhata-tits-tree>. QNFO: Bruhat-Tits Tree as a Unifying Geometric Object. [13] <https://papers.qnfo.org/papers/acrp08-paradigm-forecast>. QNFO: When Will Non-Archimedean Geometry Displace the Real Numbers? A Structured Assessment of the Adelic Substrate Thesis.

Appendix A. Divergence report

No divergent claims were identified among the independent drafts; all substantive statements converge.

Appendix B. Claim attribution

Claim ID	Statement	Source Draft(s)	Agreement
C1	Definition of learnable map ϕ and proof of hierarchical preservation	A, B, C	Convergent
C2	Derivation of ultrametric distance $d_p = p^{-k}$ and correspondence to tree depth	A, B, C	Convergent
C3	Numerical example yielding $d_p = 1/729$ and depth 3	A, B, C	Convergent
C4	Hierarchical loss formulation $\mathcal{L}_{\text{hier}}$	A, B, C	Convergent
C5	List of contributions	A, B, C	Convergent
C6	Limitations L1–L5	A, B, C	Convergent
C7	Falsifiability criterion based on mean distortion > 2	A, B, C	Convergent
C8	Practical pathway (users, TRL, next experiment, constraints, cost projection, risks)	A, B, C	Convergent