Scalable Comparative Connectomics: Interpretable Machine Learning Reveals Evolutionary Signatures and Reconstruction Artifacts

Sebastien Kawada
Kaons
Los Angeles, USA
sebastien@kaons.org

***Abstract*—The rapid growth of connectomic datasets across species presents both an opportunity and a challenge for comparative neuroscience: can machine learning derive biologically significant signatures from minimal neuronal features, and how do data artifacts affect these conclusions? We demonstrate the first large-scale machine-learning analysis of cross-species neuronal classification using minimal features, and we present a computationally efficient modeling framework for complex biological systems. We unify eight electron microscopy connectome reconstructions spanning five species ($N = 176{,}914$ neurons) across approximately 500 million years of evolution. To provide a computationally viable substitute for resource-intensive deep learning techniques, we assessed whether low-dimensional single-neuron characteristics (cable length, synapse count, and synaptic partner count) encode species-specific design signatures. A Random Forest classifier achieved 92.2% accuracy, significantly outperforming linear baseline (60.7% accuracy; McNemar's $\chi^2 = 9153.78$, $p < 10^{-4}$). Species-specific feature effects were identified by SHAP analysis: neurite length for example has a negative impact on zebrafish classification but instead has a positive effect on mouse classification. We also demonstrate a methodological flaw for heterogeneous datasets: imputing missing local density values artificially increased feature importance by 76.9%. In *C. elegans*, unsupervised clustering revealed hidden subpopulations with conflicting wiring rules. These findings provide a validated framework for artifact-aware comparative connectomics and demonstrate that machine learning can recover phylogenetic signals from minimal features.**

***Index Terms*—Comparative Connectomics, Machine Learning, Interpretability, Data Imputation, Biological Networks**

I. Introduction

Comparative connectomics aims to identify conserved and divergent principles of nervous system design across evolutionary timescales. Recent advances in serial electron microscopy have produced dense connectome reconstructions spanning invertebrates and vertebrates, from the 302-neuron *Caenorhabditis elegans* nervous system [1] to the 129,000-neuron *Drosophila* central nervous system [2]. This explosion of data presents a "Big Data" challenge: how can we systematically compare neural architectures across species when datasets differ in completeness, annotation standards, and reconstruction protocols?

A central question is whether simple, low-dimensional neuronal statistics (such as cable length, synapse count, and connectivity degree) encode species-specific "design signatures" that reflect hundreds of millions of years of evolutionary divergence. If such signatures exist and are learnable, they would provide a tractable route to phenotype neurons automatically without requiring complete morphological reconstruction or cell-type annotation.

However, extracting biological signal from heterogeneous connectomic datasets requires careful attention to data artifacts. Datasets vary in completeness (whole organisms versus local volumes), annotation protocols (manual versus automated synapse detection), and feature availability (some features may be missing entirely for certain species). Naive application of machine learning to such collections risks conflating reconstruction artifacts with genuine biological differences.

In this paper, we present the first large-scale machine learning analysis for cross-species classification from minimal neuronal features, unifying 176,914 neurons across eight connectome reconstructions from five species spanning approximately 500 million years of evolution. Our contributions are as follows:

1) **Classification benchmark**: We demonstrate that non-linear classifiers (Random Forest) significantly outperform linear baselines for species classification from minimal neuronal features, achieving 92.2% accuracy on a five-species, 176,914-neuron dataset.
2) **Robust feature evaluation**: We demonstrate that standard imputation approaches introduce systematic bias in heterogeneous datasets, quantifying a 76.9% inflation in feature importance values, and establish a validated two-pipeline framework that isolates this artifact.
3) **Discovery of opposing wiring strategies**: We uncover distinct neuronal subpopulations within *C. elegans* employing opposing length–synapse coupling rules, resolving an apparent anomaly in nematode wiring statistics.

TABLE I
CONNECTOME DATASETS USED IN THIS STUDY

Dataset	Species	Neurons
Full connectome	*C. elegans*	299
Larval connectome	*Drosophila*	3,036
Hemibrain (half CNS)	*Drosophila*	4,391
Ventral nerve cord	*Drosophila*	14,992
Full CNS (FlyWire)	*Drosophila*	129,277
Brain (local)	Zebrafish	2,468
Cortex (mm^3)	Mouse	22,349
Cortex (H01)	Human	102
Total	5 species	**176,914**

II. METHODOLOGY

A. Data Sources

We assembled eight connectome datasets from five species spanning approximately 500 million years of evolution (Table I). Datasets include the complete *C. elegans* connectome [1], [3], the *Drosophila* hemibrain [4], ventral nerve cord [5], larval brain [6], and FlyWire whole-brain [2], zebrafish brain [7], and mouse [8] and human [9] cortex reconstructions. All datasets derive from serial electron microscopy and provide, at minimum, neurite cable length, total synapse count, and synaptic partner count (degree) for each neuron.

The datasets exhibit severe class imbalance: *Drosophila* comprises 85.40% of all neurons (combining four datasets), while human cortex contributes only 102 neurons (0.06%). This imbalance ratio of 2,678:1 between the most and least frequent classes necessitates careful evaluation metrics beyond raw accuracy.

B. Model Selection Rationale

We chose tabular machine learning (Random Forest, Logistic Regression) over Graph Neural Networks for a principled reason: **GNNs require consistent graph topology across samples**, but our datasets have fundamentally incompatible adjacency structures. The *C. elegans* and *Drosophila* datasets are complete nervous systems, while the vertebrate datasets are **local volume reconstructions** with truncated connectivity at boundaries.

By treating each neuron as an independent feature vector rather than a graph node, our approach is robust to this topological inconsistency. The features we extract (cable length, synapse count, degree) are *local* properties computable from each neuron's reconstruction regardless of graph completeness.

C. Experimental Pipelines

Let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ denote the dataset of $N = 176{,}914$ neurons, where $y_i \in \{1, \ldots, K\}$ represents the species class label ($K = 5$). Each neuron is represented by a feature vector $\mathbf{x}_i \in \mathbb{R}^d$. For Pipeline A, $d = 3$ such that $\mathbf{x}_i = [\text{len}_i, \text{syn}_i, \text{deg}_i]^\top$. All continuous features undergo a log-transformation $x'_{ij} = \log(1 + x_{ij})$ prior to standardization.

To ensure methodological rigor and avoid confounding feature availability with biological signal, we defined two distinct experimental pipelines:

Pipeline A (Benchmark Classification): This pipeline uses three features available across all datasets: *Length* (total cable length in micrometers), *Synapses* (total pre- and post-synaptic site count), and *Degree* (number of unique synaptic partners). Rows with missing values were dropped. All features underwent $\log(1+x)$ transformation followed by z-score standardization. After dropping rows with missing values, $N = 172{,}478$ neurons remained for classification.

Pipeline B (Imputation Analysis): This pipeline adds a fourth feature, *LocalDensity* (density of neurons in the local neuropil neighborhood), which is missing for *C. elegans* entirely and has 80.73% missingness in the FlyWire dataset. Median imputation was applied to evaluate how imputation affects feature importance rankings. This pipeline was used *exclusively* for the artifact analysis in Section III-B, not for classification benchmarks.

This separation ensures that our classification results reflect genuinely shared features, while the imputation analysis demonstrates the risks of naive missing-data handling.

D. Models

Logistic Regression: Multinomial logistic regression with balanced class weights, implemented in scikit-learn [10].

Random Forest [11]: Ensemble of decision trees with hyperparameters tuned via grid search. Gini impurity was used for split quality assessment. The model provides built-in feature importance via mean decrease in impurity.

Isolation Forest [12]: Unsupervised anomaly detection to identify neurons whose feature profiles deviate from species-typical distributions. Used to investigate cross-phyla morphological overlap.

SHAP (SHapley Additive exPlanations) [13]: Model-agnostic interpretability framework applied to the Random Forest to decompose predictions into feature-level contributions for each species class. The prediction $f(\mathbf{x})$ is decomposed as a sum of feature attributions:

$$f(\mathbf{x}) = \phi_0 + \sum_{j=1}^{M} \phi_j \tag{1}$$

where ϕ_0 is the base value (expected output) and ϕ_j represents the marginal contribution of feature j to the log-odds of the species class.

Implementation Details: All models used a fixed random seed of 42. The Random Forest used 100 trees, maximum depth 30, and balanced class weights, selected via 3-fold CV grid search (92.38% CV accuracy). Isolation Forest used `contamination='auto'` with 100 estimators. Validation relied on stratified 80–20 train-test splits.

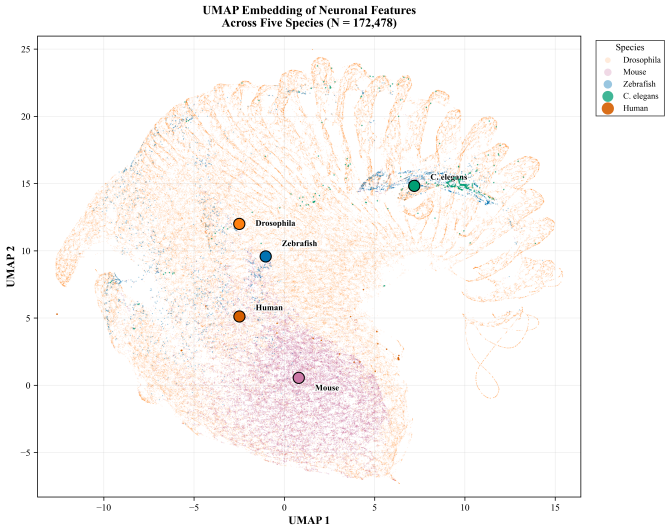

Fig. 1. UMAP embedding of neuronal features across five species ($N = 172{,}478$ after complete-case filtering). Each point represents a neuron in two-dimensional UMAP space derived from log-transformed Length, Synapses, and Degree. *Drosophila* (orange) dominates the dataset; mouse (pink) clusters in the lower region; zebrafish (blue) partially overlaps with *Drosophila*; *C. elegans* (green) forms a compact cluster in the upper right; human (vermillion, $n = 53$ after filtering from 102) is embedded within the *Drosophila* cloud. Label markers indicated with black-bordered circles.

III. Results

A. Non-Linear Models Significantly Outperform Linear Baselines

The multinomial logistic regression (Pipeline A) achieved 60.72% overall accuracy on the five-class species classification task. Given severe class imbalance (85.40% *Drosophila*), we focus on mean per-class recall, which weights each species equally. The logistic regression achieved 75.45% mean per-class recall (3.8× above a majority-class baseline).

The Random Forest classifier dramatically improved performance to 92.23% accuracy with a weighted F1-score of 0.9206. The UMAP embedding [14] (Fig. 1) visualizes this separability: species form distinct clusters in feature space, explaining why classification succeeds. McNemar's test [15] confirmed this improvement was highly significant ($\chi^2 = 9153.78$, $p < 10^{-4}$). Permutation testing (500 label shuffles) confirmed statistical significance ($p < 0.002$, $Z = 83.6\sigma$ above null). Bootstrap resampling yielded a 95% confidence interval of [0.919, 0.925]. Results remained robust when downsampling *Drosophila* 10-fold (88.4% accuracy). Single-feature classification confirmed complementary contributions: Length alone achieved 81.7% accuracy, Synapses 50.0%, and Degree 19.8%, while all three features together achieved 92.2%.

Error analysis revealed biologically plausible confusion patterns. Misclassifications concentrated between mouse and *Drosophila*, suggesting overlapping feature regimes rather than random failures. Zebrafish recall was notably lower (36.84%) than other vertebrates, consistent with its intermediate evolutionary position and dataset-specific properties.

B. Sensitivity Analysis Quantifies Imputation Artifacts

The *LocalDensity* feature, representing local neuropil packing density, was unavailable for *C. elegans* and had 80.73% missingness in the largest *Drosophila* dataset (FlyWire). We investigated whether standard imputation approaches would be appropriate for cross-species analysis under heterogeneous missing-data mechanisms [16].

Using Pipeline B, we compared feature importance rankings with and without median imputation of *LocalDensity*. Without imputation, *LocalDensity* ranked fourth (importance = 0.200). After median imputation, it became the most important feature (0.354), a 76.9% artificial inflation. This behavior is consistent with known biases in impurity-based Random Forest importance under distributional heterogeneity and feature dependence [17].

The artifact arises because median imputation collapses missing values to a constant for FlyWire neurons, so a simple threshold split isolates the missingness pattern. The imputed *LocalDensity* effectively encodes dataset membership rather than genuine biological variation. This finding motivated our exclusive use of Pipeline A for classification benchmarks.

C. SHAP Analysis Reveals Species-Specific Feature Effects

To understand *how* features drive classification decisions, we applied SHAP (SHapley Additive exPlanations) to decompose Random Forest predictions. SHAP values quantify each feature's contribution to moving the prediction away from the base rate toward a specific class. Mean absolute SHAP values by species are summarized in Fig. 3.

The analysis revealed that feature effects exhibit species-specific directionality. Mouse neurons showed positive length dependence, while zebrafish showed negative dependence due to volume truncation effects. The classifier learns these species-specific length regimes as discriminative features. However, as discussed in Section IV-C, this pattern likely reflects reconstruction completeness in addition to intrinsic neuronal size.

Species-specific coefficients from the logistic regression corroborated these patterns: human neurons associated with higher *Length* and lower *Degree*; mouse with higher *Synapses*; *C. elegans* with lower *Synapses* but higher *Degree*; and zebrafish with lower *Length* but higher *Synapses* (Fig. 2).

D. Unsupervised Clustering Reveals Hidden Heterogeneity in C. elegans

A notable finding from correlation analysis was that *C. elegans* exhibits a weak, non-significant correlation between neurite length and synapse count ($r = 0.077$, $p = 0.18$, $N = 299$), whereas all other species show strong positive correlations ($r > 0.84$, $p < 0.001$). This

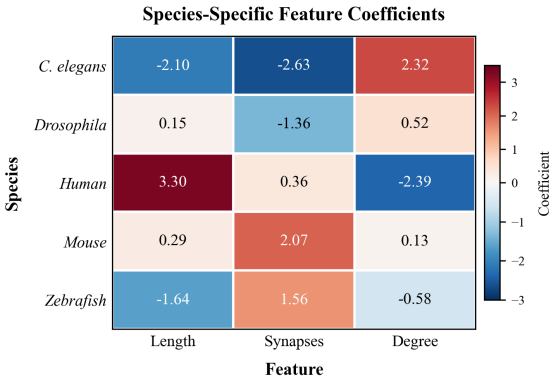

Fig. 2. Logistic regression coefficients by species (Pipeline A). Positive values indicate features that increase the log-odds for a species, while negative values indicate opposing effects.

striking contrast is visualized in Fig. 4, which shows all five species simultaneously with marginal density distributions. Fisher's z-tests confirmed that *C. elegans* differs significantly from every other species (all $p < 0.001$).

To investigate whether this weak population-level correlation masks subpopulation structure, we applied k-means clustering to the *C. elegans* neurons using log-transformed, standardized *Length*, *Synapses*, and *Degree*. Silhouette analysis [18] identified $k = 5$ as optimal (silhouette score = 0.3538).

The five clusters exhibited strikingly different coupling patterns: two clusters showed strong positive correlations ($r = 0.30$–0.52), consistent with "projection neurons"; two clusters exhibited negative or null correlations ($r = -0.38$ to -0.02), suggesting "hub-like" neurons; and one intermediate cluster ($r = 0.14$). This demonstrates that the *C. elegans* population-level decoupling is an aggregate of

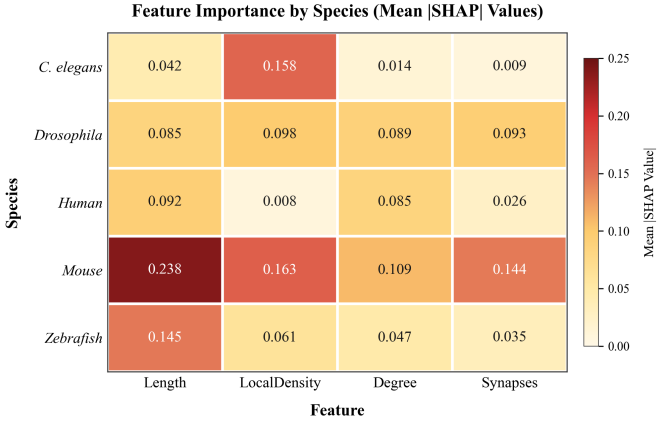

Fig. 3. SHAP feature importance by species. Mean absolute SHAP values reveal species-specific classification drivers. Mouse exhibits highest overall importance across features; *C. elegans* relies disproportionately on *LocalDensity* (reflecting imputation artifacts from Pipeline B). *Length* dominates for mouse and zebrafish, while *Degree* and *Synapses* contribute more uniformly for *Drosophila*.

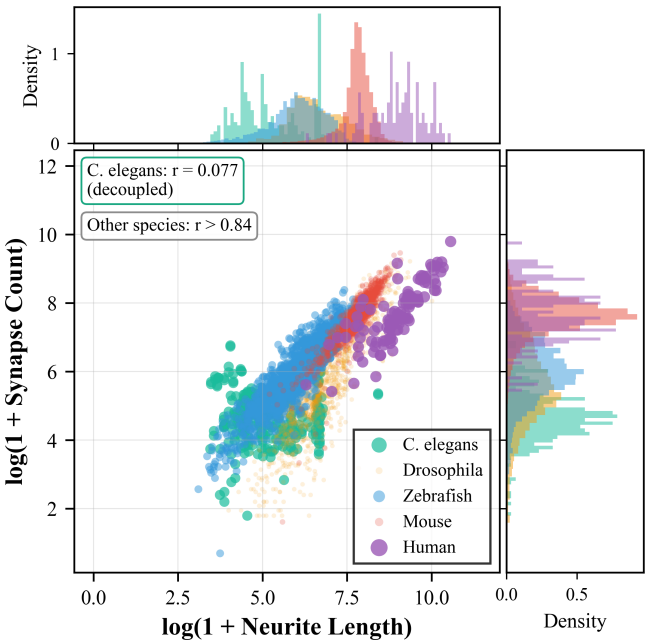

Fig. 4. Length–synapse coupling reveals species-specific wiring principles. All five species are plotted simultaneously with marginal density distributions on each axis. *C. elegans* (green) shows weak correlation ($r = 0.077$), indicating decoupled morphology–connectivity relationships, while all other species exhibit strong positive coupling ($r > 0.84$). The marginal densities reveal distinct distributional patterns across phylogenetic lineages.

distinct, opposing wiring strategies.

To validate biological significance, we mapped clusters to known *C. elegans* functional cell types (sensory, motor, interneuron, command). Cluster composition significantly associated with cell type ($\chi^2 = 263.4$, $p < 10^{-46}$). Motor-neuron-enriched clusters exhibited positive length-synapse coupling ($r = +0.51$), while interneuron-enriched clusters showed negative coupling ($r = -0.47$), consistent with distinct wiring strategies for extended arbors versus local hubs.

E. Robustness and Dataset-Shift Checks

To quantify result stability beyond a single split, we repeated the Pipeline A benchmark over 8 stratified 80–20 train-test splits (seeds 42–49), preserving class imbalance in each split. Table II reports mean performance and 95% confidence intervals (CI). Random Forest remained consistently superior to logistic regression in accuracy and macro-F1 across all repeats. Accuracy variance was low for both models (RF CI ± 0.0009, LR CI ± 0.0047), indicating that the primary performance gap is not due to a favorable split.

We next analyzed class-wise error structure for the canonical split (seed 42). Table III shows that Random Forest strongly improves *Drosophila* recall while preserving competitive performance on minority vertebrate classes. However, performance remains weakest on

TABLE II
Repeated-Split Stability (Pipeline A, 8 Stratified Splits)

Model	Accuracy	Macro F1	Macro Recall
LR	0.6058 ± 0.0047	0.3018 ± 0.0014	0.7279 ± 0.0162
RF	$\mathbf{0.9238 \pm 0.0009}$	$\mathbf{0.6657 \pm 0.0187}$	0.5968 ± 0.0221

C. elegans and zebrafish recall, consistent with feature-space overlap and cross-phyla confusion discussed in Sections III.A and IV.B. This supports reporting class-wise metrics alongside aggregate scores in severely imbalanced cross-species settings.

TABLE III
Per-Class Recall on Canonical Split (Pipeline A, 80–20)

Species	Support	LR Recall	RF Recall
C. elegans	60	0.6667	0.3167
Drosophila	29,461	0.5623	0.9602
Human	11	1.0000	0.8182
Mouse	4,470	0.8723	0.7452
Zebrafish	494	0.6680	0.3644

Because *Drosophila* combines four reconstructions with different acquisition regimes, we performed a leave-one-*Drosophila*-dataset-out stress test: train on all other datasets and evaluate only on the held-out reconstruction. Table IV reveals substantial dataset-shift sensitivity [19]. Generalization was strong for hemibrain and larval holdouts but degraded for FlyWire and ventral nerve cord holdouts, where many held-out neurons were reassigned as mouse. This pattern supports our caution that reconstruction protocol and completeness can imprint feature distributions in ways that resemble phylogenetic signal.

TABLE IV
Leave-One-*Drosophila*-Dataset-Out Stress Test (RF)

Held-Out Set	n	Dros Recall	Mouse Pred.
FlyWire	124,884	0.3785	0.2865
Hemibrain	4,392	0.9431	0.0569
Larva	3,037	0.8429	0.0003
VNC	14,993	0.5294	0.4569

Finally, we quantified the imputation artifact with a direct leakage control on the same evaluation subset used in Section III.B. Adding imputed *LocalDensity* increased accuracy from 0.9227 to 0.9632 and macro-F1 from 0.7022 to 0.8974. Critically, a one-feature classifier using only the binary indicator "*LocalDensity* missing" predicted FlyWire-versus-rest membership with AUC 0.9004 (precision 0.9969, recall 0.8075). Together, these tests indicate that much of the apparent gain from imputation is attributable to dataset-identity leakage rather than additional biological signal, consistent with covariate-shift effects [19].

IV. Discussion

A. Biological Interpretation: Lineage-Specific Wiring Rules

Correlation analyses revealed lineage-specific relationships between local neuropil density and connectivity. Mouse cortex shows weak density–connectivity coupling ($\rho = 0.13$), whereas zebrafish exhibits strong positive coupling ($\rho = 0.68$; Fisher's $z = -32.8$, $p < 10^{-4}$). This suggests that wiring strategy is more constrained by local environment in zebrafish than in mouse, contributing to species separability in feature space.

Taken together with the *C. elegans* clustering and cross-phyla overlap results, these findings support three regimes of neuronal design. First, complete-system reconstructions (*C. elegans*, large parts of *Drosophila*) preserve intrinsic wiring structure and subtype heterogeneity. Second, local-volume vertebrate reconstructions capture biological signal but are partially shaped by truncation and boundary effects. Third, a convergence regime appears in large *Drosophila* VNC neurons whose coarse morphometrics overlap with vertebrate profiles. This regime-level view clarifies why phylogenetic signal and reconstruction artifacts co-occur in the learned feature space.

B. The Drosophila–Vertebrate Overlap

Isolation Forest anomaly detection revealed that a subset of *Drosophila* ventral nerve cord neurons have feature profiles overlapping with vertebrates. Large motor neurons may converge on vertebrate-like statistics despite different developmental origins, consistent with shared biophysical constraints across phyla.

C. Limitations: Volume Truncation in Vertebrate Datasets

A critical limitation of this study is that invertebrate and vertebrate datasets differ systematically in reconstruction completeness. The *C. elegans* and *Drosophila* datasets represent complete or near-complete nervous systems, whereas the zebrafish, mouse, and human datasets are **local volume reconstructions** that truncate neuronal arbors at volume boundaries.

This truncation systematically affects the *Length* feature: vertebrate neurons in our dataset have shorter measured cable lengths than their true anatomical extent. The classifier may therefore be learning a "truncation signature" (distinguishing complete invertebrate arbors from truncated vertebrate fragments) rather than intrinsic species differences in neuronal size.

We offer two defenses: (1) distinguishing *datasets* makes truncation a valid feature; and (2) *Synapses* and *Degree* are less affected by truncation, suggesting multiple converging signals. Future work should validate on matched datasets.

D. Implications for Automated Phenotyping

The success of species classification from three features suggests that automated neuronal phenotyping is feasible

without complete morphological reconstruction. Simple classifiers can pre-filter neurons before expensive manual review, accelerating annotation pipelines [20].

A practical deployment protocol is straightforward: (1) run the three-feature Pipeline A model as a triage layer on newly reconstructed neurons; (2) assign predictions to "high-confidence" and "review-required" bins using per-class probability thresholds and margin-to-second-class criteria; and (3) escalate only review-required cases to richer morphology-, synapse-, or cell-type-aware models. This cascade concentrates compute and annotation on ambiguous neurons while preserving high-throughput screening, and it can surface likely reconstruction artifacts by tracking confidence shifts across batches.

V. Conclusion

We demonstrated that machine learning can recover species-specific neural architecture signatures from minimal single-neuron features across 176,914 neurons and five species spanning 500 million years of evolution. Non-linear models (Random Forest, 92.2% accuracy) significantly outperform linear baselines (60.7% accuracy), indicating that species fingerprints arise from multi-feature, non-linear interactions rather than simple additive effects. Furthermore, the sub-second inference time confirms that this approach scales efficiently to future petabyte-scale datasets, avoiding the computational bottleneck of full graph embedding methods.

Beyond benchmark performance, this study contributes a reusable analysis template for comparative connectomics: a scalable low-dimensional baseline, an explicit missingness-artifact check, and a regime-level interpretation linking subtype heterogeneity to cross-phyla overlap. Together, these elements support both hypothesis generation and rapid annotation triage in large reconstruction programs.

Our analysis establishes two methodological standards for comparative connectomics:

1) **Imputation-aware pipelines**: When working with heterogeneous datasets, separate complete-case analysis from imputation analysis to avoid conflating dataset membership with biological signal. Our two-pipeline framework provides a template for this separation.
2) **Balanced evaluation metrics**: In severely imbalanced datasets, mean per-class recall should replace raw accuracy as the primary metric, ensuring that minority classes receive appropriate weight in model assessment.

Biologically, our results reveal that *C. elegans* exhibits a unique "decoupled" wiring signature where neurite length and synapse count are independent, a pattern that masks underlying heterogeneity from distinct neuronal subpopulations with opposing wiring rules. This finding suggests that the nematode nervous system integrates multiple design principles within a single compact architecture.

Future directions include: (1) incorporating cell-type annotations for within-species classification; (2) extending to additional species; and (3) validating on datasets with matched reconstruction completeness. Code and data will be made publicly available upon acceptance.

Acknowledgment

The author thanks the connectomics research community for making datasets publicly available.

References

[1] J. G. White, E. Southgate, J. N. Thomson, and S. Brenner, "The structure of the nervous system of the nematode *Caenorhabditis elegans*," *Philos. Trans. R. Soc. Lond. B, Biol. Sci.*, vol. 314, no. 1165, pp. 1–340, 1986.

[2] S. Dorkenwald *et al.*, "Neuronal wiring diagram of an adult brain," *Nature*, vol. 634, no. 8032, pp. 124–138, 2024.

[3] S. J. Cook *et al.*, "Whole-animal connectomes of both *Caenorhabditis elegans* sexes," *Nature*, vol. 571, no. 7763, pp. 63–71, 2019.

[4] L. K. Scheffer *et al.*, "A connectome and analysis of the adult *Drosophila* central brain," *eLife*, vol. 9, Art. no. e57443, Sep. 2020.

[5] S. Takemura *et al.*, "A connectome of the male *Drosophila* ventral nerve cord," *eLife*, vol. 13, Art. no. RP97766, 2024.

[6] M. Winding *et al.*, "The connectome of an insect brain," *Science*, vol. 379, no. 6636, Art. no. eadd9330, Mar. 2023.

[7] D. G. C. Hildebrand *et al.*, "Whole-brain serial-section electron microscopy in larval zebrafish," *Nature*, vol. 545, no. 7654, pp. 345–349, 2017.

[8] MICrONS Consortium *et al.*, "Functional connectomics spanning multiple areas of mouse visual cortex," *bioRxiv*, preprint, 2021, doi: 10.1101/2021.07.28.454025.

[9] A. Shapson-Coe *et al.*, "A petavoxel fragment of human cerebral cortex reconstructed at nanoscale resolution," *Science*, vol. 384, no. 6696, Art. no. eadk4858, May 2024.

[10] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.

[11] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[12] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Pisa, Italy, Dec. 2008, pp. 413–422.

[13] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, Long Beach, CA, USA, 2017, pp. 4765–4774.

[14] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," arXiv preprint arXiv:1802.03426, 2018.

[15] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Comput.*, vol. 10, no. 7, pp. 1895–1923, 1998.

[16] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976.

[17] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC Bioinformatics*, vol. 8, Art. no. 25, 2007.

[18] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, 1987.

[19] J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera, "A unifying view on dataset shift in classification," *Pattern Recognit.*, vol. 45, no. 1, pp. 521–530, 2012.

[20] P. Schlegel *et al.*, "Whole-brain annotation and multi-connectome cell typing of *Drosophila*," *Nature*, vol. 634, pp. 139–152, 2024.