Digital phenotyping accelerates soil biodiversity discovery

Camila C. Filgueiras[1,5,*], **Yongwoon Kim**[2], **Daniel Gluesenkamp**[3], **and Denis S. Willett**[4,5]

[1]Department of Biology, University of North Carolina Asheville, Asheville, 28803, NC, USA
[2]Union Biometrica, Holliston, 01746, MA, USA
[3]California Institute for Biodiversity, Berkeley, 94709, CA, USA
[4]Department of Applied Ecology, North Carolina Institute for Climate Studies, NC State University, Raleigh, 27695, NC, USA
[5]SOIL, Asheville, 28803, NC, USA
[*]camila@unca.edu

ABSTRACT

Soils are among the most biodiverse ecosystems on the planet, home to staggering amounts of small organisms. Many of these organisms are not known to science. This biodiversity 'dark matter' remains largely unexplored because identifying small soil organisms requires either specialized taxonomic expertise or expensive molecular methods, creating a throughput bottleneck that restricts ecosystem-scale monitoring. To explore this dark matter, we developed a high-throughput digital phenotyping approach using multispectral flow cytometry to create digital 'fingerprints' of soil organisms. Analyzing 2,318 organisms spanning nematodes, collembola, mites, and tardigrades, we show that these digital fingerprints distinguish taxonomic groups with high accuracy and capture phylogenetic signal that explains 91% of variance in DNA barcode relationships. Machine learning alignment enables us to assess genetic similarity based solely on digital fingerprints, allowing prediction of relationships without sequencing. Smart sampling strategies guided by these projections achieve 6-fold improvements in species discovery efficiency compared to traditional approaches, with advantages that compound as sampling increases. Our smart sampling approach has applications across domains and provides a scalable pathway for rapid biodiversity assessment with immediate applications in agriculture, conservation, and ecosystem monitoring.

Introduction

Soil ecosystems harbor the majority of terrestrial biodiversity [1], with estimates suggesting that a single handful of soil contains thousands of species spanning bacteria, protists, fungi, nematodes, and microarthropods[2]. Despite this extraordinary diversity, most soil organisms remain uncharacterized, creating a fundamental obstacle to understanding ecosystem function, predicting responses to environmental change, and managing belowground communities for agriculture and conservation[3,4]. This biodiversity 'dark matter' persists not because soil organisms are difficult to collect, but because identifying them is prohibitively slow and expensive. Traditional morphological approaches require specialized taxonomic expertise and considerable time investment—often hours per specimen for microscopic fauna—while DNA barcoding and metabarcoding, though taxonomically precise, cost $5–9 per sample with processing times of 12–30 hours[5,6]. These constraints make comprehensive soil biodiversity assessment impractical at the landscape scales required for ecosystem monitoring, restricting most studies to small sample sizes that inadequately capture community diversity.

Nematodes exemplify both the importance and the challenge. As the most abundant metazoans on Earth, nematodes occupy all trophic levels—from bacterial feeders to plant parasites to predators—making them essential indicators of soil health and ecosystem function[7,8]. Yet morphological identification demands expert taxonomists capable of distinguishing cryptic species that differ in subtle cuticular patterns or reproductive structures, while molecular identification at scale remains economically prohibitive[9–11]. Plant-parasitic nematodes alone cause annual global crop losses in excess of $125 billion annually, underscoring the agricultural and economic stakes of rapid, accurate identification[12,13]. The fundamental problem is one of throughput: biodiversity discovery follows steep species accumulation curves where sampling effort increases exponentially with diversity, but identification bottlenecks prevent the high-throughput sampling required to climb these curves efficiently.

Flow cytometry offers a potential solution by capturing high-dimensional optical phenotypes—"digital fingerprints"—of individual organisms as they transit laser beams at rates exceeding 2,000 per minute[14]. Originally developed for medical diagnostics and cell sorting, multispectral flow cytometry has revolutionized microbial ecology by enabling cultivation-free quantification of bacterial community structure and diversity[15–17]. The technology generates rich phenotypic data: as cells pass through multiple laser interrogation points, detectors capture light scatter (encoding morphology) and multi-wavelength

fluorescence (encoding biochemical composition), creating unique spectral signatures for each individual. However, application to multicellular soil fauna has only begun to be explored[14], raising fundamental questions: Do optical measurements of invertebrate mesofauna contain sufficient information to distinguish taxonomic groups? If so, can digital phenotypes capture phylogenetic relationships—correlating with genetic distances in ways that enable prediction of molecular similarity without sequencing? If so, could these predictions guide intelligent sampling strategies that accelerate biodiversity discovery while reducing costs and enabling scaling?

Here we demonstrate that digital phenotyping with the Smart Soil Organism Detector (Smart SOD)—a custom large-bore flow cytometer—can accelerate soil biodiversity discovery by creating optical proxies for genetic relationships. Analyzing soil organisms across four major taxonomic groups (nematodes, collembola, mites, tardigrades), we show that multispectral digital fingerprints not only distinguish taxa with high accuracy but also capture phylogenetic signal, explaining variance in DNA barcode distances after machine learning alignment. This correspondence enables projection of unsequenced organisms into genetic similarity space based solely on optical features, allowing identification of phylogenetically novel specimens before molecular analysis. Applying this approach to California soils, we demonstrate that smart sampling strategies guided by digital embeddings achieve species discovery efficiencies 6-fold higher than random sampling, with advantages that compound at larger scales. Our results establish digital phenotyping as a scalable pathway for rapid biodiversity assessment, assisting in transitioning soil ecology from small-scale descriptive surveys to ecosystem-scale quantitative monitoring.

Results

Digital fingerprints distinguish major organism groups

Accelerating soil biodiversity discovery requires methods that can rapidly differentiate organisms without molecular sequencing. To test whether optical measurements contain sufficient taxonomic information, we analyzed 2,318 soil organisms spanning major taxonomic groups (collembola, mites, tardigrades, and nematodes) using the Smart SOD, a custom large-bore flow cytometer. Each organism was scanned at 2.5 MHz by four solid-state lasers (405, 488, 561, and 640 nm) capturing 12 optical channels: two extinction measurements (light attenuation at 561 and 488 nm), two forward scatter channels (morphology-dependent diffraction), and eight fluorescence emission channels spanning violet to far-red wavelengths (Figure 1A,D,G,J). Simultaneously, a high-speed 5 MP camera captured brightfield images at 5× magnification (Figure 1C,F,I,L), providing morphological confirmation of organism identity.

As organisms transit the flow cell, they generate unique temporal signatures—optical fingerprints that encode both external morphology and internal biochemistry. These multispectral signals can be assembled into characteristic heatmap images (Figure 1B,E,H,K) that visually distinguish taxonomic groups: collembola exhibit strong extinction signals from their compact chitinous bodies; mites show segmented patterns from appendages and cuticle autofluorescence; tardigrades display unique profiles from their cylindrical morphology; while nematodes generate elongated signatures reflecting their vermiform body plan. Combined with high-resolution photographs, digital fingerprints reveal both internal and external morphological features for multispectral phenotyping.

These digital fingerprints were reproducible within taxonomic groups despite substantial morphological variation (Figure 2). To visualize relationships among organisms, we extracted 37 aggregate optical features (peak height, width, and integrated area for each of 12 channels, plus time-of-flight) and projected this high-dimensional space into two dimensions using Uniform Manifold Approximation and Projection (UMAP). This projection was performed in an unsupervised manner, with species labels provided only after dimensionality reduction to avoid biasing the embedding. The resulting visualization demonstrates that taxonomic groups form distinct clusters: major groups (Nematoda, Collembola, Tardigrada, Arachnida) occupy distinct regions. Within nematodes, diverse genera separated into distinct subclusters, with multiple organisms of the same species (e.g., *Steinernema scarabei*, *Hoplolaimus galeatus*, *Belonolaimus longicaudatus*) clustering tightly together while remaining distinct from other species.

The unsupervised nature of this analysis suggest that digital fingerprints contain sufficient information to reconstruct taxonomic relationships without *a priori* labels—enabling label-free taxonomic grouping. This suggests digital fingerprints could be used to rapidly assess unknown organisms by projecting them into the same embedding space and evaluating their proximity to known groups. Grey triangles in Figure 2 show unclassified organisms from soil samples; their distances to characterized organisms provide a quantitative proxy for taxonomic relatedness. To validate this approach, we used k-nearest neighbor classification ($k = 3$) to assign species-level identities to held-out nematode specimens based solely on their digital fingerprints. Balanced classification accuracy exceeded 94% (out-of-sample balanced accuracy: 94.4%; cross-validated in-sample average balanced-accuracy: 95.6%), suggesting that optical phenotypes alone can achieve taxonomic resolution, but in seconds rather than hours per specimen.

These results suggest that digital fingerprinting can rapidly distinguish major soil organism groups and align with previously observed performance classifying nematode species[14].

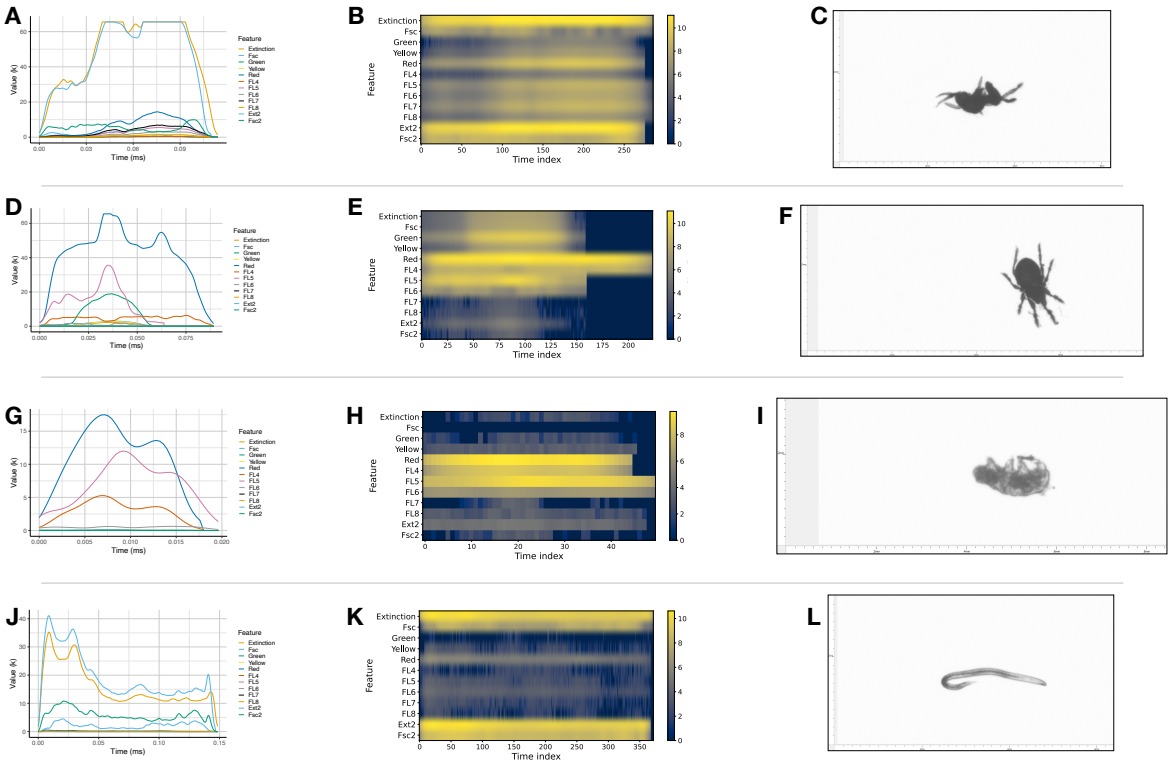

Figure 1. Digital fingerprints distinguish soil organism groups. Representative temporal signatures from Smart SOD multispectral laser scanning flow cytometry (2.5 MHz sampling, 12 selected optical features shown). (**A–C**) Collembola (Springtail); (**D–F**) Arachnida (Mite); (**G–I**) Tardigrade (Water Bear); (**J–L**) Nematode (Round Worm). Optical features recorded from these scans produce characteristic spectra along the body length of the organism (Left Panels: A, D, G, J). These spectra can be assembled into a layered representative image that forms a 'digital fingerprint' for that organism (Middle Panels: B, E, H, K). Each organism is also photographed with a 5MP camera at 5x magnification (Right Panels: C, F, I, L). Visually, the spectra and digital fingerprints of each organism are different and reflect external and internal morphology.

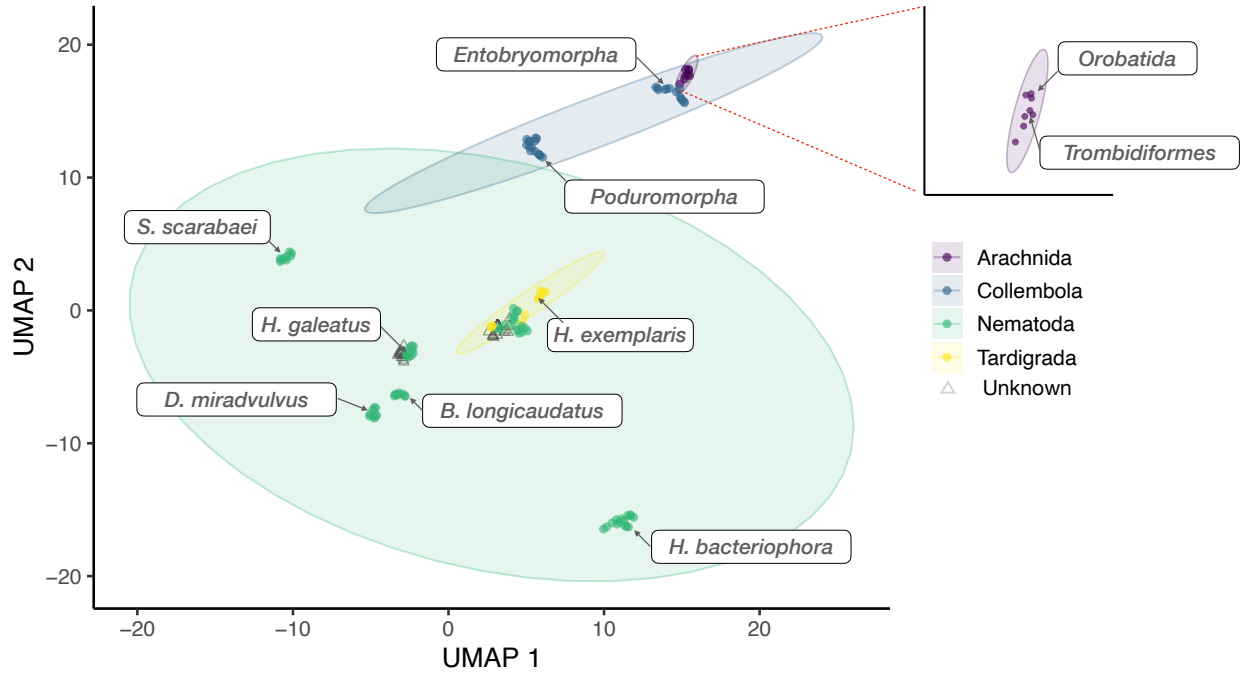

Figure 2. Digital fingerprints cluster taxonomic groups. Uniform Manifold Approximation and Projection of Smart SOD digital fingerprints. Major taxonomic groups (Nematoda, Collembola, Tardigrada, Arachnida) occupy distinct regions. Multiple organisms cluster tightly by species with separation between distinct genera. Unknown organisms (triangles) can be assessed taxonomically by their proximity to known groups. Colored points represent individual organisms, with colors indicating taxonomic group. Shaded ellipses indicate 99% confidence regions. Inset: Magnified view of Arachnida cluster showing subclustering by Orobatida and Trombidiformes. Individual example species labeled include the tardigrade *Hypsibius exemplaris* and nematodes *Steinernema scarabei*, *Hoplolaimus galeatus*, *Dolichodorus miradvulvus*, *Belonolaimus longicaudatus*, and *Heterorhabditis bacteriophora*.

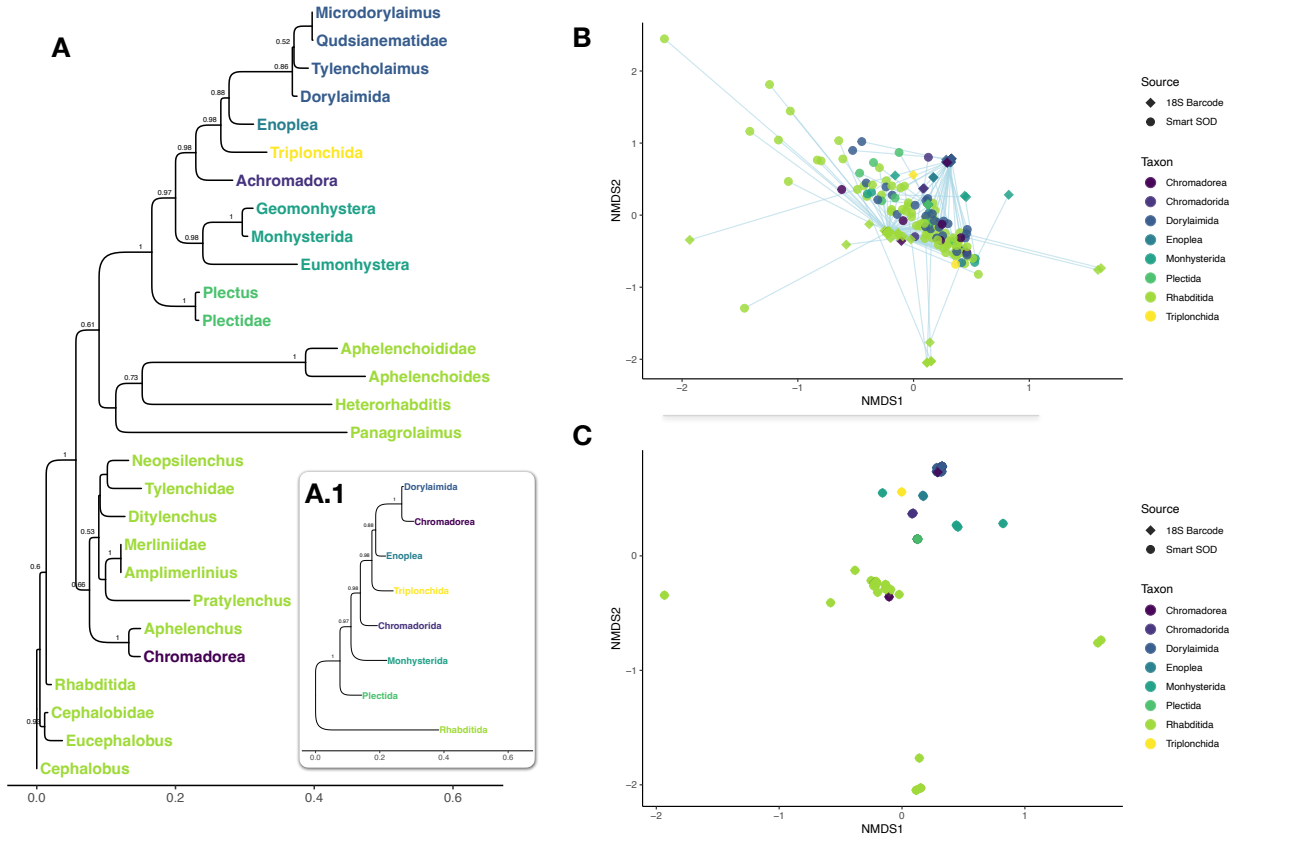

Figure 3. Machine learning aligns digital fingerprints and genetic phylogenies. (**A**) Maximum likelihood phylogenetic tree from 18S rRNA sequences showing lowest available taxonomic level relationships across soil based nematodes. Bootstrap (1,000) support values ≥50% shown at nodes. (**A.1**) Inset showing higher taxon resolution. (**B**) Original embedding: Digital fingerprints (filled circles) and genetic coordinates (diamonds) in 2D NMDS space before alignment. Lines show displacement between digital and genetic positions for each specimen. (**C**) Aligned embedding: Fused Gromov-Wasserstein optimal transport aligns digital fingerprints with genetic phylogenies. Digital fingerprints explain 91% of the observed variation in 18S barcode regions.

Digital fingerprints capture phylogenetic signal

While digital fingerprints enable rapid taxonomic clustering, accelerating biodiversity discovery requires predicting relationships without sequencing—enabling prioritization of which organisms to sequence for maximum phylogenetic coverage. To test whether optical phenotypes contain phylogenetic information, we performed a paired analysis: scanning soil nematodes with the Smart SOD, isolating individuals into 96-well plates, then sequencing the 18S rRNA and COI gene regions from those same specimens. This approach directly linked each organism's digital fingerprint to its molecular barcode, enabling quantitative comparison of optical and genetic similarity.

Sequencing 237 nematodes from California grassland and woodland soils recovered substantial phylogenetic diversity spanning eight orders (Chromadorida, Dorylaimida, Enoplea, Monhysterida, Plectida, Rhabditida, Triplonchida, plus Chromadorea at class level). Maximum likelihood phylogenetic reconstruction from 18S sequences revealed relationships across taxa (Figure 3A), with bootstrap support values ≥50% confirming robust topology. This phylogenetic diversity—ranging from bacterial-feeding *Plectus* to plant-parasitic *Pratylenchus* to entomopathogenic *Heterorhabditis*—provided a rigorous test dataset spanning spanning major nematode clades across the phylum's evolutionary history.

To assess whether digital fingerprints encode phylogenetic relationships, we computed pairwise genetic distances from the 18S phylogeny (maximum likelihood distances under Jukes-Cantor substitution model) and compared them to digital distances (Euclidean distances in 37-dimensional optical feature space). Both distance matrices were embedded into two-dimensional space using non-metric multidimensional scaling (NMDS) to visualize relationships. Initially, genetic and digital embeddings showed only weak correspondence (Mantel r: -0.06; $P = 0.9$; Figure 3B): although some closely related organisms had similar digital fingerprints, lines connecting each organism's genetic position (diamonds) to its digital position (circles) revealed

substantial misalignment.

This initial mismatch does not indicate absence of phylogenetic signal, but rather that the relationship between optical and genetic distances is non-linear and requires explicit alignment. We therefore applied optimal transport methods to find the transformation that best maps digital phenotypes onto genetic relationships. Starting with standard Gromov-Wasserstein (GW) optimal transport—which preserves geometric structure (pairwise distances) without considering taxonomic labels—we established a geometry-only baseline. We then applied Fused Gromov-Wasserstein (FGW) optimal transport, which augments geometric alignment with taxonomic information by penalizing mappings that pair organisms from different genera. FGW balances two objectives: preserving the internal structure of each distance space while respecting known taxonomic boundaries, formalized through a tunable parameter α that weights geometric versus taxonomic costs.

After optimizing α to maximize correlation between aligned spaces, FGW alignment dramatically improved correspondence between digital and genetic coordinates (Mantel r: 0.95; $P = 0.001$; Figure 3C), indicating that digital fingerprints explain 91% of the variance in 18S phylogenetic relationships. K-nearest neighbor overlap was 80% meaning that four out of five of each specimen's five nearest genetic neighbors were also among its five nearest digital neighbors after alignment, demonstrating that digital similarity reliably predicts genetic relatedness at fine taxonomic scales.

These results suggest that spectral signatures in digital fingerprints contain quantifiable phylogenetic information that can be computationally aligned with DNA barcode relationships. This alignment is not merely correlative: the optical features that distinguish taxonomic groups (autofluorescence spectra, morphometric ratios, scatter patterns) reflect heritable biological traits that covary with genetic divergence.

Smart sampling accelerates species discovery

Once the mapping between digital and genetic spaces is established using a reference set of sequenced organisms, it can be applied to predict genetic relationships for new, unsequenced specimens based solely on their digital fingerprints—the key requirement for accelerating biodiversity discovery through smart sampling. To test this, we developed a projection method that translates digital fingerprints from unsequenced organisms into predicted positions in genetic similarity space, then evaluated whether this projection enables more efficient biodiversity sampling strategies.

We used supervised k-nearest neighbor regression to project digital fingerprints into two-dimensional genetic coordinate space defined by 18S rRNA and COI gene distances (Figure 4A). Training this model on 80% of sequenced nematodes (18S) and 30% (COI) as anchors, we predicted genetic coordinates for held-out specimens from their optical features alone. Performance was high: mean absolute error reached only 1.9% of the 18S embedding range and 16.9% of the COI range, with error ellipses (Figure 4A, blue regions) showing tight correspondence between predicted and actual genetic positions. This low error reflects the model's ability to capture non-linear relationships between optical and genetic similarity while scaling to thousands of organisms. Importantly, projection uncertainty is quantified for each specimen, enabling statistically informed decisions about which organisms represent sufficiently distinct genetic lineages to warrant sequencing.

With this method for low error genetic position projection, we evaluated whether Smart SOD-guided sampling could accelerate biodiversity discovery relative to traditional approaches. We tested two complementary scenarios representing different field sampling contexts: sequential sampling (real-time prioritization as organisms are encountered) and batch optimization (selecting maximally diverse subsets from large collections with fixed sequencing budgets).

Sequential sampling simulates real-time biodiversity monitoring, where organisms are processed continuously and sequencing decisions must be made immediately. In traditional approaches without digital phenotyping, each extracted nematode would be sequenced to assess whether it represents a new taxonomic group, following steep species discovery curves where sampling effort increases exponentially with diversity. Analyzing California soil samples, random sequential sequencing required 108 nematodes to recover 15 unique genera (Figure 4B, gray line).

The Smart SOD improves this process through intelligent filtering. As each organism passes through the flow cytometer, its digital fingerprint is projected into genetic space and compared to all previously analyzed specimens using k-nearest neighbor searches. If the organism's projected position exceeds genetic distance thresholds in the 18S, COI space, it is flagged as potentially novel and prioritized for sequencing. Organisms similar to previously characterized specimens are recorded but not sequenced, as they are unlikely to represent new taxonomic groups. This high-throughput filtering dramatically accelerates discovery: Smart SOD-guided sampling recovered the same level of diversity with 93 less samples (Figure 4B, blue line), producing a **6-fold improvement** in sequencing efficiency.

This advantage is not static but **compounds with increasing diversity**: as more taxonomic groups are discovered, the probability that any random specimen represents a truly novel lineage decreases, while the Smart SOD continues to selectively enrich for distinctive organisms. The divergence between curves in Figure 4B demonstrates this compounding effect—efficiency gains grow as the number of taxonomic groups increase, and the gap continues widening. For comprehensive biodiversity inventories targeting hundreds of taxonomic groups, this compounding advantage would yield 10-fold or greater improvements, fundamentally changing the feasibility of ecosystem-scale monitoring.

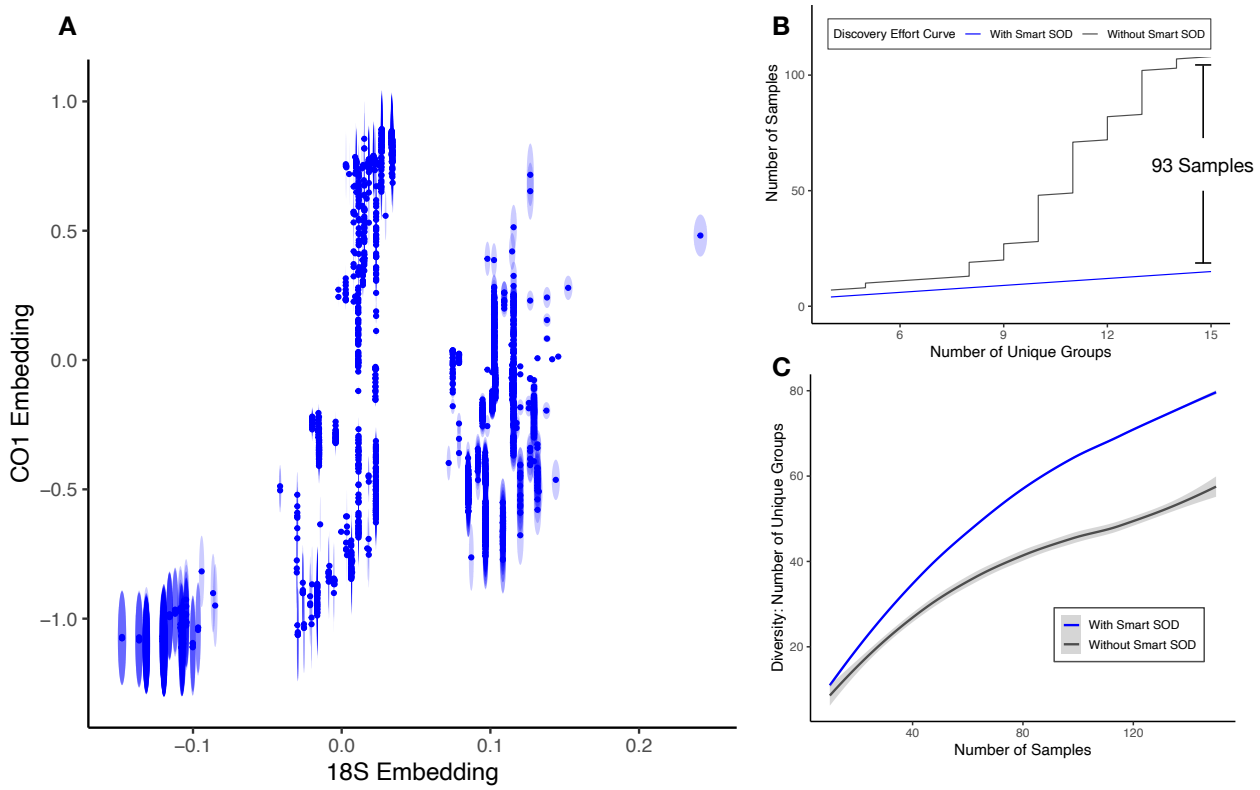

Figure 4. Smart sampling accelerates species discovery 6-fold. (**A**) Embedding dimensions showing projection of digital fingerprints into 18S and COI genetic spaces. Blue points represent individual organisms with error ellipses indicating projection uncertainty. (**B**) Effort curve: traditional sequential sampling (gray line, “Without Smart SOD") requires 108 specimens to discover 15 unique genera, Smart SOD guided sampling (blue line, “With Smart SOD") achieves the same diversity with just 15 specimens. (**C**) Diversity accumulation curves comparing random sampling (gray, “Without Smart SOD") from 10–150 specimens to Smart SOD maximin optimized sampling (blue, “With Smart SOD"). Smoothed curves are fitted with LOESS to actual sample values at each sampling density. Shaded regions denote 95% confidence intervals.

Batch optimization addresses scenarios with fixed sequencing capacity, where researchers must select which specimens to sequence from large collections to maximize phylogenetic coverage. Without digital phenotyping, random selection of 150 specimens from California soil samples recovered moderate diversity (Figure 4C, gray line). However, Smart SOD enables rational selection through maximin optimization: choosing the 150 organisms with maximum minimum pairwise distances in projected genetic space ensures the most phylogenetically dispersed sample possible. This strategy substantially outperforms random sampling across all sample sizes > 10 (Figure 4C, blue line), with advantages that grow consistently at larger sample sizes. At 150 specimens, Smart SOD maximin selection recovered 25% more unique taxonomic groups than random sampling. Importantly, the gap between optimized and random sampling widens continuously (Figure 4C), indicating that efficiency gains scale with sampling effort—the larger the survey, the greater the proportional advantage of digital phenotyping.

Both strategies demonstrate that digital phenotyping with the Smart SOD fundamentally changes the economics of biodiversity discovery in a scale-dependent manner. By linking digital fingerprints to genetic similarity, then using those predictions to identify, prioritize, and select organisms that maximize phylogenetic coverage, we can efficiently climb species discovery curves with minimal sequencing effort. The practical implications compound with survey scale, potentially reaching 10-fold or greater for comprehensive regional inventories. This scaling advantage contributes to moving soil biodiversity monitoring from small-scale studies to ecosystem-scale surveys—enabling spatial coverage, temporal resolution, and taxonomic breadth previously unattainable at any budget.

Discussion

Our results suggest that digital phenotyping can accelerate soil biodiversity discovery by creating optical proxies for genetic relationships, enabling intelligent sampling strategies that achieve 6-fold or greater efficiency gains over traditional approaches. This acceleration addresses a fundamental bottleneck: comprehensive soil biodiversity assessment has been impractical not because organisms are uncollectable, but because identification methods—whether morphological or molecular—cannot scale to match the extraordinary diversity present in even modest soil samples. By demonstrating that multispectral optical measurements capture sufficient phylogenetic information to guide sampling decisions, we provide a pathway toward ecosystem-scale biodiversity monitoring previously unattainable at any budget.

The economics of this approach favor scaling. The Smart SOD processes samples at rates up to 2,000 organisms per minute, with typical 50ml soil extracts analyzed in under 15 minutes while generating detailed digital fingerprints for each individual. In contrast to molecular workflows requiring DNA extraction, PCR amplification, and sequencing infrastructure, digital phenotyping requires only water as a reagent. The instrument simultaneously sorts prioritized organisms into individual wells of 96-well plates, directly linking high-resolution photographs and multispectral signatures to specific specimens for downstream molecular validation. Most importantly, the efficiency gains compound with survey scale: our results suggest 10-fold or greater advantages for comprehensive regional inventories targeting dozens of taxa. Applied to surveys of thousands or tens of thousands of specimens—the scale required for landscape-level monitoring—these compounding efficiencies translate to order-of-magnitude reductions in per-sample costs and processing time.

Several limitations warrant discussion. Our phylogenetic alignment and projection models were trained on nematodes from California grassland and woodland soils, representing substantial but geographically constrained diversity. The accuracy we report (1.9–16.9% projection error, 91% variance explained in genetic distances) applies specifically to organisms within or near the phylogenetic and morphological range of our training set. Organisms from unmapped regions of genetic similarity space will initially have higher prediction uncertainty, and organisms from dramatically different habitats or biogeographic regions may exhibit novel optical phenotypes that do not align well with our current models. This is not a fundamental limitation but rather a data requirement: like any learning approach, prediction accuracy depends on training set coverage.

This limitation diminishes with use through active learning. Each time the Smart SOD prioritizes an organism for sequencing based on its predicted novelty, that organism's digital fingerprint and molecular barcode become new training data, iteratively expanding the mapped region of phylogenetic-optical space and improving subsequent predictions. The maximin optimization strategy we demonstrate explicitly targets this expansion by selecting organisms at the boundaries of known diversity, ensuring that sequencing effort systematically fills gaps in coverage. Over time and across research groups, this creates a growing reference library linking digital phenotypes to genetic identities across broader taxonomic and geographic ranges—analogous to how DNA barcode databases expanded from initial focused efforts to comprehensive coverage through distributed contribution. The Smart SOD's ability to preserve sorted specimens for later sequencing means that even organisms initially deemed too uncertain to prioritize can be retrospectively analyzed as models improve, ensuring no diversity is permanently lost to initial model limitations.

Beyond nematodes, our demonstration with collembola, mites, and tardigrades (Figure 2) suggests that digital fingerprinting extends to other soil fauna, though taxon-specific training will be required. The optical features that distinguish taxonomic groups—cuticle autofluorescence, body morphology, scatter patterns, lipid content—are universal biological properties that should generalize across soil mesofauna. However, optimal laser configurations, fluorescence channels, and feature

extraction methods may differ for organisms with distinct size ranges or optical properties. Adapting the approach to soil microorganisms (bacteria, protists, fungi) represents a natural extension, as flow cytometry already enables high-throughput microbial community profiling[15,18], though sorting and imaging individual microbial cells presents technical challenges not present for larger organisms.

The practical implications of this technology span multiple domains. In agriculture, conservation, and ecology, the ability to process thousands of specimens per study scales questions about soil health, taxonomic diversity, functional diversity, trophic structure, and ecosystem responses to perturbation from correlative to mechanistic, as researchers can now afford sufficiently large sample sizes to detect subtle community shifts and test causal hypotheses.

Most fundamentally, digital phenotyping with the Smart SOD reframes biodiversity assessment from an endpoint to a continuous process. Traditional surveys produce static snapshots—species lists frozen at collection dates—because the time and cost of molecular identification precludes repeated sampling. In contrast, the Smart SOD's throughput and efficiency enable temporal monitoring with resolutions previously restricted to abiotic measurements: weekly, daily, or even continuous sampling becomes economically feasible. This temporal dimension is crucial for understanding ecosystem responses to climate change, agricultural intensification, and habitat fragmentation[3,19], where belowground communities may shift rapidly but remain invisible to conventional surveys. As soil ecosystems face unprecedented pressures, tools for real-time biodiversity assessment transition from convenient to essential.

Our results suggest that digital phenotyping can accelerate biodiversity assessment. Rather than replacing DNA barcoding, the Smart SOD makes sequencing resources more effective by concentrating them on phylogenetically distinct organisms, climbing species discovery curves with minimal effort. This synergy between optical and molecular approaches—where digital fingerprints guide intelligent sampling and sequencing validates and refines predictive models—creates a positive feedback loop that accelerates biodiversity discovery while building increasingly comprehensive reference libraries. As these libraries expand through distributed use, digital phenotyping could ultimately enable label-free biodiversity monitoring for many applications, reserving molecular validation for truly novel lineages. The Smart SOD allows us to move closer to ecosystem-scale biodiversity monitoring systems capable of tracking the hidden majority of terrestrial life in real time.

Methods

Instrument

Digital fingerprints of soil organisms were acquired using a COPAS VISION 500 large-bore flow cytometer (Union Biometrica) tuned to detect soil organisms[14]. This instrument was equipped with a 500 μm diameter flow cell and four solid-state lasers (405 nm, 488 nm, 561 nm, and 640 nm, each at 100 mW). As organisms transited the flow cell, temporal optical signals were recorded across 12 detection channels at 2.5 MHz: two extinction measurements (488 nm and 561 nm), two forward scatter measurements (488 nm at two gains), and eight fluorescence emission channels with the following bandpass filters: 445/50 nm (violet), 512/25 nm (green), 615/24 nm (red), 680/30 nm (far-red), and four additional fluorescence channels (FL4–FL7) spanning the emission spectrum. For each detection channel, three summary parameters were extracted from the temporal waveforms: peak height (maximum signal intensity), peak width (full width at half maximum), and peak content (integrated signal area), yielding 36 derived optical features per organism. Time-of-flight (TOF) through the detection volume provided an additional morphometric parameter. A high-speed 5 MP CMOS camera (Hamamatsu) with 1 μs exposure time captured brightfield images through a 5$\times$ objective as organisms passed through the flow cell.

Soil samples were suspended in distilled water within 50 ml conical tubes and delivered to the flow cell under 1.80 PSI pressure. Hydrodynamic focusing via calibrated sheath flow ensured single-file passage of organisms oriented lengthwise through the detection volume. Following optical interrogation and imaging, organisms were selectively sorted into individual wells of 96-well plates using a pneumatic air-diverter mechanism (3.00 PSI). The gentle fluidic handling preserved organism viability, with sorted specimens remaining intact and motile post-analysis.

Quality control procedures were implemented before and between sample runs. Instrument calibration employed fluorescent control beads of defined size (Union Biometrica GP Control Particle Concentrate, PN: 301-5071-001; 500 μm Control Particles, PN: 370-5071-000) to verify optical and imaging system performance. Between runs, the fluidics system was back-flushed and cleaned to prevent cross-contamination. Samples were pre-filtered through 400 μm mesh screens (pluriSelect pluriStrainer, Ref: 43-50400-01) to remove debris that could obstruct the flow cell.

PMT gain and voltage settings were optimized for each detection channel: extinction at 488 nm (gain 1.0), forward scatter at 488 nm (gain 1.9), 445/50 nm channel (gain 1.0, 180 V), 512/25 nm channel (gain 1.0, 350 V), 615/24 nm channel (gain 1.0, 450 V), and 680/30 nm channel (gain 1.0, 470 V). All temporal waveforms were sampled at 2.5 MHz.

Samples

Representative Samples Representative specimens from cultured populations of nematode species, tardigrades, collembola, and mites were obtained from laboratory colonies maintained at the University of North Carolina Asheville. All specimens

had been previously identified to species level through combined morphological assessment by taxonomic experts and DNA barcode confirmation using 18S rRNA and COI gene regions. Organisms were maintained in culture under standard laboratory conditions and analyzed as live specimens.

Soil Samples Soil samples were collected from natural grassland and woodland sites across the San Francisco Bay Area, California, in March 2024. Collection sites represented diverse habitat types to maximize diversity. Approximately 1 L of soil was collected from the upper 15 cm at each site, thoroughly homogenized in the field, and shipped overnight to UNC Asheville for analysis on the Smart Soil Organism Detector. A 50 ml subsample of the homogenized soil was processed using modified Baermann funnel extraction for 24 hours to concentrate motile organisms in 15 ml of distilled water. The resulting organism concentrate was pre-filtered through 400 μm mesh (pluriSelect pluriStrainer, Ref: 43-50400-01) to remove large debris before Smart SOD analysis.

Organisms from soil extracts were analyzed by Smart SOD flow cytometry and selectively sorted into individual wells of 96-well plates based on digital fingerprint diversity. Sorted organisms were preserved in 95% ethanol within 1 hour of sorting and shipped to the Canadian Centre for DNA Barcoding at the Biodiversity Institute of Ontario, University of Guelph, for DNA barcoding through the Barcode of Life Data System (BOLD). DNA extraction, PCR amplification, and bidirectional sequencing were performed following standard BOLD protocols[20]. For nematodes, the 18S small subunit rRNA gene was amplified using nematode-specific primers, producing amplicons of approximately 900–1000 bp. The mitochondrial COI gene was amplified using nematode-specific primer cocktails targeting the I3-M11 partition, yielding sequences of approximately 450–650 bp. A total of 237 sequences from individual organisms were successfully obtained and taxonomically assigned, with sequences deposited in BOLD under project code SSOD.

Analysis

Pre-processing Raw temporal spectral data from each organism were exported as tab-delimited text files containing waveforms for 12 optical channels sampled at 2.5 MHz. These files were processed using custom Python scripts (Python 3.12, Polars 1.34) to generate cleaned, standardized datasets.

Temporal waveforms were first converted to long-form data matrices with time indices calculated from the 2.5 MHz sampling frequency. For each of the 12 optical channels (two extinction, two forward scatter, and eight fluorescence channels), three summary parameters were extracted from the temporal waveforms: peak height (maximum signal intensity), peak width (temporal duration at half-maximum), and peak content (integrated signal area). Combined with time-of-flight measurements, this yielded 37 aggregate features per organism (12 channels $\times$ 3 parameters + TOF).

Data cleaning procedures included: (1) removal of trailing zero values resulting from organisms exiting the flow cell, (2) schema standardization to reconcile column naming variations across sample batches, and (3) conversion to Apache Parquet format for efficient downstream analysis. Hydrodynamic focusing ensured single-organism measurements.

Digital fingerprints for visualization were constructed directly from the 12 primary temporal optical channels by arranging signal intensities into matrices (features $\times$ time). Log-transformed intensities ($\log(value + 1)$) were visualized as heatmaps to facilitate pattern recognition while preserving the complete temporal dynamics of each organism's optical signature.

Quality Control Organisms were selected for analysis from background particles based on predefined gating regions in extinction versus time-of-flight parameter space[14]. These gates excluded events with insufficient time-of-flight and anomalous extinction profiles (indicating poor signal quality or multiple organisms).

Visualization To visualize high-dimensional relationships among organism digital fingerprints, we applied Uniform Manifold Approximation and Projection (UMAP)[21] to the optical features extracted from each organism. Prior to dimensionality reduction, feature matrices were standardized using z-score normalization to ensure comparable scales across optical channels.

UMAP was configured to project hyper-dimensional relationships between organism digital fingerprints into two dimensional space and applied using a fixed random seed, 5 nearest neighbors (local neighborhood size), and a minimum distance of 0.2 (minimum separation between embedded points). These parameters were selected to balance local structure preservation with global topology visualization.

Confidence ellipses for taxonomic groups in UMAP space were computed assuming bivariate Student's t-distributions fitted to the embedded coordinates of labeled specimens within each group, with ellipses representing 99% confidence regions. Unknown organisms were projected into the trained UMAP embedding allowing assessment of their similarity to known taxonomic groups based on Euclidean distances in the reduced dimensional space.

Phylogenetic Analysis DNA sequences from 18S rRNA barcodes were aligned using DECIPHER v3.4[22] with default parameters and parallel processing. The alignment was masked to remove ambiguous regions using a conservation threshold of 0.2, and columns with >50% gaps were excluded. Maximum likelihood phylogenetic trees were constructed using the phangorn package v2.12[23] in R. Model selection was performed via 'modelTest' evaluating JC, HKY, TN93, and GTR substitution

models with and without gamma-distributed rate heterogeneity (Γ) and invariant sites (I), selecting the best-fitting model by Akaike Information Criterion corrected for small sample sizes (AICc).

Initial tree topology was estimated using neighbor-joining on Jukes-Cantor distances, then evaluated under the best-fit substitution model selected by AICc. The tree was optimized via maximum likelihood uwith nearest-neighbor interchange (NNI) branch swapping, optimizing branch lengths, substitution rate parameters, gamma shape parameters, and proportion of invariant sites, with a maximum of 500 iterations. Bootstrap support was calculated from 1,000 replicates, each with topology optimization via NNI, with bootstrap values $\geq 50\%$ displayed on nodes. The final tree was midpoint-rooted for visualization.

Pairwise genetic distances between all sequenced specimens were computed as maximum likelihood distances under the Jukes-Cantor model using the 'dist.ml' function in phangorn, forming an $n \times n$ symmetric distance matrix. Digital distances between the same organisms were computed as pairwise Euclidean distances in the 37-dimensional aggregate optical feature space. Both distance matrices were normalized to [0,1] by min-max scaling of off-diagonal elements to enable direct comparison.

Alignment via Optimal Transport To quantify the correspondence between digital phenotypes and genetic relationships, we aligned the two distance spaces using optimal transport methods. Pairwise genetic and digital distance matrices ($\mathbf{C}_1$ and $\mathbf{C}_2$, respectively, both $n \times n$) were each embedded into 2D coordinate spaces using non-metric multidimensional scaling with normalized stress optimization to enable visualization and alignment assessment.

We first applied standard Gromov-Wasserstein (GW) optimal transport, which aligns distributions by preserving geometric structure (pairwise distances) without considering feature information. The GW optimization minimizes:

$$\min_{\mathbf{T}\in\Pi(\mathbf{p},\mathbf{q})} \sum_{i,j,k,\ell} |\mathbf{C}_1(i,k) - \mathbf{C}_2(j,\ell)|^2 \mathbf{T}_{ij}\mathbf{T}_{k\ell} \tag{1}$$

where $\mathbf{T}$ is a coupling matrix between the n genetic and n digital measurements, and $\mathbf{p}$ and $\mathbf{q}$ are uniform mass distributions. This serves as a geometry-only baseline.

To incorporate taxonomic information, we applied Fused Gromov-Wasserstein (FGW) optimal transport, which augments GW with a feature cost term:

$$\min_{\mathbf{T}\in\Pi(\mathbf{p},\mathbf{q})} (1-\alpha) \sum_{i,j,k,\ell} |\mathbf{C}_1(i,k) - \mathbf{C}_2(j,\ell)|^2 \mathbf{T}_{ij}\mathbf{T}_{k\ell} + \alpha \sum_{i,j} \mathbf{M}_{ij}\mathbf{T}_{ij} \tag{2}$$

where $\mathbf{M}$ is a feature cost matrix encoding taxonomic label disagreement (0 for same genus, 1 for different genera), and α balances geometric structure preservation versus feature alignment. Both GW and FGW objectives were optimized using the POT (Python Optimal Transport) library v0.9.6[24] with square loss, Armijo line search, convergence tolerance of 10^{-9}, and maximum iterations of 1,000 (GW) or 2,000 (FGW).

The resulting coupling matrix $\mathbf{T}$ was used to project digital coordinates into genetic coordinate space via barycentric mapping: $\mathbf{Y}_{digital} = \mathbf{T}^\top \mathbf{Y}_{genetic} / \sum_i \mathbf{T}_{ij}$, where $\mathbf{Y}_{genetic}$ are the MDS-embedded genetic coordinates. We systematically evaluated values of α, selecting the optimal α that maximized Mantel correlation between aligned distances.

Alignment quality was assessed using three complementary metrics: (1) Mantel correlation[25] between genetic distances and aligned digital distances (computed using Pearson correlation on z-score normalized upper-triangular distance vectors with 9,999 permutations), (2) Procrustes disparity (sum of squared distances between corresponding points after optimal rotation and scaling)[26], and (3) k-nearest neighbor ($k = 5$) overlap (proportion of shared neighbors between genetic and aligned digital spaces).

Accelerating Species Discovery To evaluate whether digital fingerprinting could accelerate biodiversity discovery, we developed two complementary sampling strategies guided by projections of digital phenotypes into genetic similarity space.

Projection into Genetic Space Digital fingerprints from all analyzed organisms were projected into both 18S and COI genetic embedding spaces using supervised k-nearest neighbor (KNN) regression trained on sequenced specimens. For each genetic marker, specimens were partitioned into anchor sets (80% for 18S, 30% for COI) and evaluation sets, with genetic distances computed as maximum likelihood distances under the Jukes-Cantor model following alignment and phylogenetic tree construction as described above. Digital distances between anchor and evaluation specimens were calculated as Euclidean distances in the 37-dimensional optical feature space.

Each evaluation specimen's genetic embedding coordinate (1D NMDS axis) was predicted from its digital distances to all anchor specimens. KNN regression models were implemented with hyperparameter optimization via 5-fold cross-validation. Hyperparameters (number of neighbors, distance weighting function, and Minkowski distance power) were tuned using space-filling designs across 20 parameter combinations, selecting optimal values by minimum mean absolute error (MAE). Projection uncertainty was visualized as error ellipses with semi-axes scaled by 1.9% (18S axis) and 16.9% (COI axis), representing the relative MAE as a fraction of the embedding range.

Sequential Sampling Strategy To simulate real-time biodiversity discovery, we evaluated organisms sequentially in the order they would be encountered during soil processing. For each organism i (starting at $i = 4$), its projected coordinates in 18S and COI space were compared to all previously analyzed organisms (1 to $i-1$) using k-nearest neighbor searches ($k = 3$). An organism was flagged as potentially representing a new taxonomic group if its nearest neighbors exceeded genetic distance thresholds: 0.04 for 18S (approximating family-level divergence) or 0.17 for COI (approximating genus-level divergence). These thresholds were empirically determined from median pairwise distances between known families (18S) and genera (COI) in sequenced specimens. Under this strategy, only flagged organisms would be prioritized for molecular sequencing, while organisms similar to previously characterized specimens would be recorded but not sequenced. Discovery efficiency was quantified as the cumulative number of unique taxonomic groups discovered as a function of specimens sequenced.

Maximin Optimization Strategy For scenarios with fixed sequencing capacity, we implemented maximin sampling to select maximally diverse subsets from all analyzed organisms. Given N total organisms with projected coordinates in 2D genetic space (18S, COI), we selected n specimens ($n \in [10, 150]$) to maximize the minimum pairwise distance among selected specimens. The algorithm initialized by selecting the organism furthest from the coordinate origin, then iteratively selected organisms with maximum minimum distance to all previously selected specimens, tracking distances in Euclidean space. For each sample size n, diversity was quantified by clustering organisms based on error ellipse overlap: two organisms were considered the same group if their normalized ellipse distance $\sqrt{(|\Delta x|/(a_1+a_2))^2 + (|\Delta y|/(b_1+b_2))^2} \leq 1$, where a and b denote semi-axes scaled by projection error. The number of unique groups was computed as the number of distinct clusters after merging overlapping ellipses. As a control, random sampling selected n organisms without replacement, with diversity computed identically. Both strategies were repeated to generate smoothed accumulation curves using LOESS regression with 95% confidence intervals.

Computational Tools and Statistical Analysis All data processing, statistical analyses, and visualizations were performed using R version 4.5.1 (2025-06-13, "Great Square Root")[27] and Python version 3.12. Analyses leveraged parallel processing across multiple CPU cores where applicable to optimize computational efficiency.

R Environment and Packages Data manipulation and visualization utilized the TIDYVERSE ecosystem[28], including DPLYR for data transformation, GGPLOT2 for graphics, and TIDYR for data restructuring. Machine learning workflows employed TIDYMODELS[29] for unified modeling interfaces, and THEMIS[30] for handling class imbalance. Parallel computing was implemented via FOREACH[31], DOPARALLEL[32], and FUTURE [33] packages for distributed processing across CPU cores.

Phylogenetic and sequence analyses used BIOSTRINGS [34] from Bioconductor for DNA sequence manipulation, DECIPHER[22] for sequence alignment, PHANGORN[23] for phylogenetic tree construction and distance calculations, and GGTREE/TREEIO[35] for phylogenetic tree visualization. Ecological and multivariate statistics employed VEGAN[36] for NMDS ordination, Mantel tests, and Procrustes analysis. Additional packages included NANOPARQUET[37] for efficient data serialization, VIRIDIS[38] for color scales, FNN[39] for fast k-nearest neighbor searches, and GLUE[40] for string interpolation.

Python Environment and Libraries Optimal transport calculations used the Python Optimal Transport (POT) library[24] implementing Gromov-Wasserstein and Fused Gromov-Wasserstein algorithms. Data manipulation employed PANDAS[41,42] for tabular data and POLARS [43] for high-performance DataFrame operations. Dimensionality reduction and manifold learning utilized UMAP-LEARN[44] for UMAP projections and SCIKIT-LEARN [45] for MDS embeddings, distance calculations (PAIRWISE_DISTANCES), data standardization (STANDARDSCALER), and nearest neighbor searches (NEARESTNEIGHBORS). Statistical analyses employed SCIPY [46] for Procrustes analysis (SCIPY.SPATIAL.PROCRUSTES), correlation calculations (SCIPY.STATS.PEARSONR), and distance matrix manipulations (SCIPY.SPATIAL.DISTANCE).

References

1. Anthony, M. A., Bender, S. F. & Van Der Heijden, M. G. Enumerating soil biodiversity. *Proc. Natl. Acad. Sci.* **120**, e2304663120 (2023).
2. Veresoglou, S. D., Halley, J. M. & Rillig, M. C. Extinction risk of soil biota. *Nat. communications* **6**, 8862 (2015).
3. Guerra, C. A. *et al.* Blind spots in global soil biodiversity and ecosystem function research. *Nat. communications* **11**, 3870 (2020).
4. Guerra, C. A. *et al.* Global hotspots for soil nature conservation. *Nature* **610**, 693–698 (2022).
5. Stein, E. D., Martinez, M. C., Stiles, S., Miller, P. E. & Zakharov, E. V. Is dna barcoding actually cheaper and faster than traditional morphological methods: results from a survey of freshwater bioassessment efforts in the united states? *PloS one* **9**, e95525 (2014).

6. Salis, L. *et al.* Performance of DNA metabarcoding, standard barcoding and morphological approaches in the identification of insect biodiversity. *Mol. Ecol. Resour.* **24**, e14018 (2024).
7. Van Den Hoogen, J. *et al.* Soil nematode abundance and functional group composition at a global scale. *Nature* **572**, 194–198 (2019).
8. Boag, B. & Yeates, G. W. Soil nematode biodiversity in terrestrial ecosystems. *Biodivers. & Conserv.* **7**, 617–630 (1998).
9. Bogale, M., Baniya, A. & DiGennaro, P. Nematode identification techniques and recent advances. *Plants* **9**, 1260 (2020).
10. Hunt, D. J. & Handoo, Z. A. Taxonomy, identification and principal species. In *Root-knot nematodes*, 55–97 (CABI Wallingford UK, 2009).
11. Palomares-Rius, J. E., Cantalapiedra-Navarrete, C. & Castillo, P. Cryptic species in plant-parasitic nematodes. *Nematology* **16**, 1105–1118 (2014).
12. Mesa-Valle, C. M., Garrido-Cardenas, J. A., Cebrian-Carmona, J., Talavera, M. & Manzano-Agugliaro, F. Global research on plant nematodes. *Agronomy* **10**, 1148 (2020).
13. Jones, J. T. *et al.* Top 10 plant-parasitic nematodes in molecular plant pathology. *Mol. plant pathology* **14**, 946–961 (2013).
14. Filgueiras, C. C. *et al.* The smart soil organism detector: An instrument and machine learning pipeline for soil species identification. *Biosens. Bioelectron.* **221**, 114417 (2023).
15. Props, R., Monsieurs, P., Mysara, M., Clement, L. & Boon, N. Measuring the biodiversity of microbial communities by flow cytometry. *Methods Ecol. Evol.* **7**, 1376–1385 (2016).
16. Besmer, M. D. *et al.* The feasibility of automated online flow cytometry for in-situ monitoring of microbial dynamics in aquatic ecosystems. *Front. microbiology* **5**, 265 (2014).
17. Gasol, J. M. & Del Giorgio, P. A. Using flow cytometry for counting natural planktonic bacteria and understanding the structure of planktonic bacterial communities. *Sci. Mar.* **64**, 197–224 (2000).
18. El Mujtar, V. A., Chirdo, F., Lagares, A., Wall, L. & Tittonell, P. Soil bacterial biodiversity characterization by flow cytometry: The bottleneck of cell extraction from soil. *Methods Ecol. Evol.* **13**, 1388–1401 (2022).
19. Archidona-Yuste, A., Ciobanu, M., Kardol, P. & Eisenhauer, N. Divergent alpha and beta diversity trends of soil nematode fauna along gradients of environmental change in the carpathian ecoregion. *Commun. biology* **8**, 587 (2025).
20. Ratnasingham, S. & Hebert, P. D. Bold: The barcode of life data system (http://www. barcodinglife. org). *Mol. ecology notes* **7**, 355–364 (2007).
21. Healy, J. & McInnes, L. Uniform manifold approximation and projection. *Nat. Rev. Methods Primers* **4**, 82 (2024).
22. Wright, E. S. Using decipher v2.0 to analyze big biological sequence data in r. *The R J.* **8**, 352–359 (2016).
23. Schliep, K. P. phangorn: phylogenetic analysis in r. *Bioinformatics* **27**, 592–593 (2011).
24. Flamary, R. *et al.* Pot: Python optimal transport. *J. Mach. Learn. Res.* **22**, 1–8 (2021).
25. Mantel, N. The detection of disease clustering and a generalized regression approach. *Cancer research* **27**, 209–220 (1967).
26. Gower, J. C. Generalized procrustes analysis. *Psychometrika* **40**, 33–51 (1975).
27. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2025).
28. Wickham, H. *et al.* Welcome to the tidyverse. *J. Open Source Softw.* **4**, 1686, DOI: 10.21105/joss.01686 (2019).
29. Kuhn, M. & Wickham, H. *Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles.* (2020).
30. Hvitfeldt, E. *themis: Extra Recipes Steps for Dealing with Unbalanced Data*, DOI: 10.32614/CRAN.package.themis (2025). R package version 1.0.3.
31. Microsoft & Weston, S. *foreach: Provides Foreach Looping Construct*, DOI: 10.32614/CRAN.package.foreach (2022). R package version 1.5.2.
32. Corporation, M. & Weston, S. *doParallel: Foreach Parallel Adaptor for the 'parallel' Package*, DOI: 10.32614/CRAN.package.doParallel (2022). R package version 1.0.17.
33. Bengtsson, H. A unifying framework for parallel and distributed processing in r using futures. *The R J.* **13**, 208–227, DOI: 10.32614/RJ-2021-048 (2021).

34. Pagès, H., Aboyoun, P., Gentleman, R. & DebRoy, S. *Biostrings: Efficient manipulation of biological strings*, DOI: 10.18129/B9.bioc.Biostrings (2025). R package version 2.76.0.

35. Yu, G. *Data integration, manipulation and visualization of phylogenetic trees* (Chapman and Hall/CRC, 2022).

36. Oksanen, J. *et al. vegan: Community Ecology Package*, DOI: 10.32614/CRAN.package.vegan (2025). R package version 2.7-1.

37. Csárdi, G. & Mühleisen, H. *nanoparquet: Read and Write 'Parquet' Files*, DOI: 10.32614/CRAN.package.nanoparquet (2025). R package version 0.4.2.

38. Garnier *et al. viridis(Lite) - Colorblind-Friendly Color Maps for R*, DOI: 10.5281/zenodo.4679423 (2024). Viridis package version 0.6.5.

39. Beygelzimer, A. *et al. FNN: Fast Nearest Neighbor Search Algorithms and Applications*, DOI: 10.32614/CRAN.package.FNN (2024). R package version 1.1.4.1.

40. Hester, J. & Bryan, J. *glue: Interpreted String Literals*, DOI: 10.32614/CRAN.package.glue (2024). R package version 1.8.0.

41. McKinney, W. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, vol. 445, 51–56 (Austin, TX, 2010).

42. pandas development team, T. pandas-dev/pandas: Pandas, DOI: 10.5281/zenodo.3509134 (2020).

43. Vink, R. & contributors. Polars: Blazingly fast dataframes in rust, python, node.js, r and sql (2024). Software available from https://www.pola.rs/.

44. McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

45. Pedregosa, F. *et al.* Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

46. Virtanen, P. *et al.* Scipy 1.0: fundamental algorithms for scientific computing in python. *Nat. Methods* **17**, 261–272 (2020).

Acknowledgements

Larry W. Duncan assisted with reference nematode specimens and provided helpful feedback on early iterations of experimental design.

Author contributions statement

C.C.F. and D.S.W. provided conceptualization and supervision. D.G. provided resources. Y.K. provided engineering and instrument expertise. C.C.F. and D.S.W. analyzed and visualized data. All authors contributed to project coordination and writing. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.