Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A transparent soil-pH suitability index for arbuscular mycorrhizal fungal genera

Dmytro Tsyhliuk

dmy.cig@gmail.com

Vinnytsia National Agrarian University

Research Article

Keywords: Species distribution modeling, Ecological niche, Spatial cross-validation, Glomeromycota biogeography, Boyce index

Posted Date: June 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9972197/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Mycorrhiza on September 11th, 2026. See the published version at https://doi.org/10.1007/s00572-026-01312-0.

Abstract

Soil pH is among the most consistent abiotic axes structuring arbuscular mycorrhizal (AM) fungal communities, yet the pH dependence of individual taxa has not been turned into a transparent, parsimonious instrument that resolves preference at the genus level and can be projected globally. We introduce HESI (Hydrogenic Edaphic Suitability Index), a fuzzy-set trapezoidal pH-membership function for twelve *Glomeromycotina* genera whose four parameters are read directly from binned presence-rate profiles in GlobalFungi rather than fitted, and scored on an independent SoilGrids pH surface. Validated by spatial (continental) block cross-validation and the binned Boyce index, HESI returned a positive predictive signal for ten of twelve genera (mean Boyce index + 0.20; eight with confidence intervals above zero). The flexible pH spline matched or exceeded machine-learning models, most clearly at the species level, where the learners fell below zero while the parameter-free function remained positive. The two genera scoring negative — a cosmopolitan generalist and a genus pooling species of opposite preference — mark the diagnostic limits of a single-axis model. The genus niches were transferred to an independent database (GlobalAMFungi) without retraining and agreed in rank with virtual-taxon optima (cross-marker Spearman ρ up to + 0.90) on small-subunit (SSU) marker, demonstrating that they are not marker artifacts. Soil pH sorted which genera occur, while temperature governed richness, consistent with global richness maps. The per-genus suitability surfaces are offered as spatially explicit, empirically testable hypotheses.

Introduction

Arbuscular mycorrhizal (AM) fungi (*Glomeromycotina*) form symbioses with most terrestrial plants and are among the most consistently pH-structured groups of soil organisms (Smith and Read 2008; Tedersoo et al. 2018). Across intense local gradients and global surveys alike, soil pH repeatedly appears as the dominant edaphic axis sorting AM fungal community composition; on a sharply varying gradient, it can override host-plant identity in determining which fungi occur where (Dumbrell et al. 2010; Davison et al. 2021).

The niches of individual AM fungal taxa along this axis are now reasonably well characterized. Veresoglou et al. (2013) modeled the occurrence of two AM fungal families, linking *Acaulosporaceae* to acidic soils and *Gigasporaceae* to high precipitation; Davison et al. (2021) extended this to the whole phylum, estimating realized niche optima for approximately species-level virtual taxa (VT) defined on the SSU marker (Öpik et al. 2010) and identifying temperature and pH as the two leading abiotic drivers, with niche properties conserved at roughly the family level. In parallel, genus-level predictive instruments have been built for other gradients — for example, a nitrogen-association index for AM fungal genera (Treseder et al. 2018) — and global machine-learning models now map AM fungal richness at kilometer resolution (Van Nuland et al. 2025). These advances are complementary in target: they characterize niches descriptively, index a different axis, or predict the number of taxa per location rather than the identity of the taxa present.

A simple genus-level instrument that translates the established pH dependence of AM fungi into a per-genus suitability score — capable of validation for spatial transfer and global projection — remains absent. The incomplete resolution of ecological differentiation among individual AM fungal taxa along edaphic gradients is repeatedly highlighted in recent literature, which continues to identify soil pH and temperature as primary drivers (Šibanc et al. 2026). The most detailed niche characterization to date is correlative and resolved at the virtual taxon (VT, SSU) level (Davison et al. 2021), which does not provide a portable per-genus function. Family-level models address only a subset of the phylum (Veresoglou et al. 2013), and global maps focus on richness rather than the suitability of individual taxa (Van Nuland et al. 2025).

Here we develop such an instrument, HESI (Hydrogenic Edaphic Suitability Index). HESI does not re-establish AM fungal distributions with respect to soil pH (Dumbrell et al. 2010; Davison et al. 2021), but instead operationalizes this dependence in three novel ways. First, it represents the pH niche of each of twelve *Glomeromycotina* genera as a single fuzzy-set membership function, with four parameters read directly from the empirical presence-rate profile, resulting in a transparent, parameter-free per-genus suitability score that avoids overfitting. Second, it independently validates these genus-level, ITS-derived optima by transferring them to a second database (GlobalAMFungi; Větrovský et al. 2023) without additional training and by rank-testing them against the SSU-defined VT niches of Davison et al. (2021), providing a cross-marker validation not previously reported for AM fungal genus niches. Third, it projects the index as a global soil-pH suitability surface, which is appropriate given the near-global distributions and low endemism of AM fungi (Davison et al. 2015). The resulting contribution is an interpretable tool, an independent multi-source validation, and a set of global maps, building upon rather than competing with the niche characterizations of Davison et al. (2021) and the community-level evidence of Dumbrell et al. (2010). This approach also clarifies HESI's relationship to recent global machine-learning work: soil pH determines which genera occur, a transferable signal, while temperature primarily governs the accumulation of taxa, the richness signal emphasized by Van Nuland et al. (2025). These address distinct questions (see Discussion).

Specifically, this study (i) constructs HESI as a trapezoidal pH-membership function for twelve AM fungal genera calibrated on GlobalFungi (Větrovský et al. 2020); (ii) validates the index using spatial block cross-validation — at the continental (leave-one-region-out) and 5° grid scales — and the binned Boyce index against an independent SoilGrids pH surface; (iii) benchmarks HESI against flexible single-axis, multi-axis, and machine-learning alternatives at both the genus and species levels; (iv) tests transferability across an independent database — at the genus level by transfer without retraining, and at the species level by the concordance of independently estimated optima for shared taxa — and their agreement with SSU-defined VT niches; and (v) projects the index as global per-genus suitability maps.

Materials and methods

Occurrence data and the soil-pH predictor

HESI was calibrated on AM fungal occurrence records from the GlobalFungi database (Větrovský et al. 2020), a compilation of high-throughput metabarcoding studies indexing fungal internal transcribed spacer (ITS) sequences across more than 17,000 samples and 178 studies. Records assigned to twelve *Glomeromycotina* genera — *Acaulospora*, *Ambispora*, *Archaeospora*, *Claroideoglomus*, *Diversispora*, *Gigaspora*, *Paraglomus*, *Scutellospora*, *Entrophospora*, *Funneliformis*, *Septoglomus*, and *Glomus* — were retained as the calibration set. Each record carried a measured soil-pH value from the originating study metadata for calibration and was additionally matched to a soil-pH value from SoilGrids 2.0 (Poggio et al. 2021; see also Hengl et al. 2017) for validation. SoilGrids is a global, machine-learning-derived soil-property product at 250 m resolution that predicts pH in water at standard depth intervals together with quantified uncertainty; we used the median prediction (Q0.5) for the 0–20 cm layer throughout. Keeping calibration on measured pH and validating against the independent SoilGrids surface makes the validation a genuine test rather than a refit. Genus presence was scored against the pool of GlobalFungi records with a valid SoilGrids pH value (7,460 records). Because SoilGrids pH is a model output instead of a direct measurement, we first quantified its agreement with the measured pH at sampled locations (Table S5): the two were strongly rank-correlated for the GlobalFungi set (Spearman ρ = 0.89; root-mean-square deviation 0.42 pH), supporting the use of the SoilGrids surface as an independent scoring layer and, for the global projection.

The HESI membership function

HESI expresses the realized soil-pH niche of each genus as a fuzzy-set membership function (Zadeh 1965) of trapezoidal form, the shape used in fuzzy species-distribution modeling (Robertson et al. 2004). Each trapezoid is defined by four ordered node parameters $a \le b \le c \le d$ on the pH axis: suitability rises linearly from 0 at pH a to 1 at pH b, remains at its maximum of 1 across the plateau from b to c, and declines linearly back to 0 between c and d, so that HESI(pH) lies in [0,1]. Rather than fitting these nodes by optimization, we read them directly from the binned empirical presence-rate profile of each genus — the proportion of samples in which the genus is present within successive pH bins — taking the lower and upper limits (a, d) and the plateau shoulders (b, c) from the pH window over which presence rate is elevated. Because nothing is fitted, the function has no free parameters and no opportunity to overfit; triangular and single-shoulder shapes arise naturally as degenerate node settings (for example, b = c, or a = b) rather than as additional functional forms. The genera partition into four niche groups according to the position of their suitability window: acidophil (*Acaulospora*, *Ambispora*, *Archaeospora*), sub-acidophil (*Claroideoglomus*, *Diversispora*, *Gigaspora*, *Paraglomus*, *Scutellospora*), alkaliphil (*Entrophospora*, *Funneliformis*, *Septoglomus*), and a single cosmopolitan form (*Glomus*). The pooled records span an empirical calibration range of pH 3.9–9.2. Using the genus as the unit is supported by evidence that the functional and niche attributes of AM fungi are phylogenetically conserved at approximately the genus-to-family level (Treseder et al. 2018; Davison et al. 2021), so that a genus-level summary preserves much of the among-taxon signal while remaining portable across databases.

Percentile-node comparator

To test whether the read-from-profile (“canonical”) nodes were preferable to an automatic rule, we constructed an alternative percentile-based estimator in which the lower and upper limits (a, d) were placed at the 5th and 95th percentiles, and the plateau shoulders (b, c) at the 25th and 75th percentiles, of the presence-pH distribution of each genus. Both node sets were scored identically on the SoilGrids surface.

Validation: the binned Boyce index

The primary performance criterion was the binned Boyce index (Boyce et al. 2002; Hirzel et al. 2006), a presence-only measure of how well a continuous suitability score ranks observed occurrences. Suitability is divided into classes, and within each class, the ratio of observed to expected presence frequency (the predicted-to-expected, or P/E, ratio) is computed; the Boyce index is the Spearman rank correlation between class suitability and the P/E ratio. It ranges from − 1 to + 1: a value near + 1 indicates that presences accumulate monotonically with predicted suitability, 0 indicates a model no better than random, and negative values indicate a counter-predictive model. Bootstrap sampling of the records gave 95% confidence intervals for each genus. We treated the Boyce index as primary because it suits presence-only data and rewards monotone calibration — the property that matters when a single index is projected onto a pH surface. For comparison, we also computed the area under the receiver-operating-characteristic curve (AUC), but interpreted it only as an auxiliary diagnostic: in presence-only settings, AUC depends on the choice of pseudo-absences and rewards a kind of discrimination that is not realizable once suitability is reduced to a function of pH alone (Lobo et al. 2008; Phillips et al. 2009).

Spatial block cross-validation

To assess whether a calibrated membership function transfers to unsampled space, we used spatial block cross-validation (Roberts et al. 2017; Valavi et al. 2019), which partitions records into spatially coherent folds so that training and test data are geographically separated. Two blocking schemes were used — continental leave-one-region-out and 5° latitude–longitude grid blocks — re-reading the trapezoid nodes on the training fold of each split and scoring suitability on the held-out fold. Blocked rather than random partitioning is necessary because soil and occurrence data are spatially autocorrelated, so random cross-validation yields optimistic scores that do not reflect transferability to new regions (Ploton et al. 2020). In cross-validation, *Archaeospora* was omitted because its small sample size (n = 13 records) could not be partitioned across folds, and the large, widespread genus *Rhizophagus* was added, so that the benchmark comprised 12 genera with sufficient per-fold data.

Flexible single-axis and machine-learning benchmarks

Against the same spatial folds, we benchmarked HESI on the Boyce index and AUC at both the genus and species levels. The benchmarks were a flexible single-axis spline of presence on pH; a two-axis spline of presence on pH and mean annual temperature (MAT); a generalized additive model using all available soil and climatic axes; and two machine-learning learners, random forest and gradient boosting, trained on the full set of predictors. This design separates the contribution of a richer pH response (the single-axis spline) from that of the additional environmental axes and their interactions (the all-axis learners).

Cross-database transfer and cross-marker comparison

Transferability beyond GlobalFungi was tested in two ways. First, the GlobalFungi-calibrated genus nodes were applied without re-estimation to the independently assembled GlobalAMFungi database (Větrovský et al. 2023) using its ITS2 sequence subset and scored on the same SoilGrids surface; for comparison, nodes were also estimated internally within GlobalAMFungi. Species-level pH optima for the 22 taxa shared between the databases were compared using Spearman's rank correlation, root-mean-square deviation, and mean offset. Second, the genus-level optima recovered here from ITS data were compared, across the eight AM fungal families for which a match was possible, with the SSU virtual-taxon niche optima of Davison et al. (2021), whose taxonomy follows the MaarjAM virtual-taxon system (Öpik et al. 2010). Because the Davison optima were measured on KCl-extracted pH — systematically lower than the water pH used by SoilGrids — and are abundance-weighted, only rank order was interpreted in this cross-marker comparison.

Global projection

Each genus trapezoid was applied to the global SoilGrids pH surface (Q0.5, 0–20 cm) to produce a per-genus suitability layer. Pixels with no soil or non-vegetated cover were left unscored, as were pixels whose SoilGrids pH fell outside the calibration range (3.9–9.2); suitability beyond the calibration window is an extrapolation the index does not attempt.

Software, data and use of AI tools

Analyses were carried out in Python using NumPy, pandas, SciPy, and scikit-learn. The occurrence data and soil layers are openly available from the GlobalFungi (https://globalfungi.com), GlobalAMFungi (https://globalamfungi.com), and SoilGrids (https://soilgrids.org) repositories; the calibration and validation code and the per-genus node parameters are available at https://doi.org/10.5281/zenodo.20581579. During the preparation of this work, the author used a large language model (Claude Opus 4.8, Anthropic) to assist with drafting and language editing of the manuscript. The author reviewed and edited all content, made all scientific and analytical decisions, and takes full responsibility for the content of the publication.

Results

Genus pH-suitability functions

The twelve canonical trapezoids span the empirical calibration range of pH 3.9–9.2 and partition the genera into four niche groups (Table 1). The acidophil anchors *Acaulospora*, *Ambispora*, and *Archaeospora* switch on at the acid end (lower limit 3.9–4.0); *Acaulospora* is the narrowest and most acid, with its suitability plateau confined to pH 3.9–4.0, consistent with documented AM fungal adaptation to, and performance in, strongly acidic soils (Clark 1997). The sub-acidophil group (*Claroideoglomus*, *Diversispora*, *Gigaspora*, *Paraglomus*, *Scutellospora*) carries plateaus centered between pH 5.5 and 7.5, and the alkaliphil group (*Entrophospora*, *Funneliformis*, *Septoglomus*) between 7 and 8 — genera repeatedly recorded in and associated with alkaline and arid calcareous soils (Parihar et al. 2019; Guardiola-Márquez et al. 2022). *Glomus* is the single cosmopolitan form, rising over a broad ramp from pH 4.5 to a plateau reaching the upper calibration bound. All twelve functions are described by the same four parameters.

Table 1
Niche classification and full-dataset validation of the twelve *Glomeromycotina* genera. n (GF), number of GlobalFungi presence records used in calibration; the binned Boyce index is scored on the independent SoilGrids pH surface (Q0.5, 0–20 cm), values near + 1 indicating that presences accumulate with predicted pH-suitability. The membership function rises linearly from 0 at pH a to 1 at b, holds the plateau between b and c, and returns to 0 at d (a, b, c, d in pH units), with the nodes read from each genus's binned presence-rate profile (see Materials and methods)

Genus	Niche group	n (GF)	a	b	c	d	Boyce index
Acaulospora	acidophil	487	3.9	3.9	4.0	7.5	+ 0.98
Ambispora	acidophil	314	4.0	4.5	6.0	7.5	−0.82
Archaeospora	acidophil	13	3.9	3.9	6.5	7.5	+ 0.20
Claroideoglomus	sub-acidophil	253	5.0	6.0	7.5	8.5	+ 0.03
Diversispora	sub-acidophil	366	5.0	7.0	7.0	8.5	+ 0.78
Gigaspora	sub-acidophil	517	4.5	6.5	6.5	8.5	+ 0.25
Paraglomus	sub-acidophil	681	4.0	4.5	6.5	8.0	+ 0.53
Scutellospora	sub-acidophil	121	4.5	5.5	5.5	8.0	+ 0.58
Entrophospora	alkaliphil	488	5.5	6.5	8.0	9.0	+ 0.93
Funneliformis	alkaliphil	980	5.0	6.0	7.5	9.0	+ 0.89
Septoglomus	alkaliphil	291	5.5	6.5	7.0	8.5	+ 0.42
Glomus	cosmopolitan	1423	4.5	8.5	9.2	9.2	−0.73

Full-dataset validation

Scored on the independent SoilGrids pH surface across the 7,460-record GlobalFungi pool, the canonical trapezoids returned a positive binned Boyce index for ten of twelve genera, with eight having bootstrap confidence intervals entirely above zero (Fig. 1). The mean Boyce index across genera was + 0.34. The acidophil and alkaliphil anchors scored highest — *Acaulospora* + 0.98, *Entrophospora* + 0.93, *Funneliformis* + 0.89, *Diversispora* + 0.78 — and observed presence rate rose steeply from low to high pH-suitability: from 1% to 26% for *Acaulospora* (a 23-fold enrichment), and roughly three-fold for *Funneliformis* and *Paraglomus*.

Two genera returned negative Boyce indices: *Glomus* (− 0.73) and *Ambispora* (− 0.82). These mark the diagnostic limits of a single-axis model rather than calibration failures. *Glomus* is a genuine pH generalist — it occurs across the whole gradient at a near-constant rate (17% at both low and high suitability). Its canonical trapezoid is nonetheless shifted toward high pH, so it ranks alkaline pixels as most suitable, while Glomus occurs as readily in acidic soils; the index therefore anti-correlates with occurrence and is strongly negative.

This is the behavior expected of acid-tolerant AM fungi, which act as pH generalists and form nested assemblages along pH gradients (Kawahara et al. 2016). *Ambispora* combines species with opposite optima (an acidophil and an alkaliphil) that a single genus-level trapezoid cannot jointly capture; its negative genus score is a signature of within-genus niche heterogeneity, recovered explicitly at the species level (see below).

Robustness to node estimation

The read-from-profile canonical nodes outperformed the automatic percentile-based estimator. On the same SoilGrids surface, the canonical nodes gave a mean Boyce index of + 0.34 against + 0.28 for the percentile nodes (Fig. 2). The canonical windows are on average about 1.1 pH units wider, producing a smoother, monotone enrichment of presences along the gradient, whereas the percentile estimator cuts narrower niches with sharp edges that the data do not support. The clearest case is *Claroideoglomus*, whose Boyce index moved from − 0.48 under percentile nodes to + 0.03 under the canonical nodes once the truncated sub-acidophil tail was retained. The parameter-free nodes are therefore not only the most transparent choice but also the empirically stronger one.

Decomposing each genus into its component species (Table S3) confirmed the same logic from the opposite direction: the single-trapezoid summary is favorable for genera with narrow, concordant species niches — for which pooling raised the Boyce index, as in *Acaulospora*, *Paraglomus*, *Funneliformis*, and *Entrophospora* — but unfavorable for the generalist *Glomus* and for genera pooling species of opposite preference (*Claroideoglomus*, *Ambispora*), where the genus-level score fell below the mean of its species. The three *Ambispora* species each returned a positive species-level Boyce index (mean + 0.21), so the negative genus score reflects within-genus niche heterogeneity rather than an absence of signal.

Spatial transfer and benchmarking against machine learning

Under spatial block cross-validation (continental leave-one-region-out and 5° grid blocks, nodes re-read on training folds, suitability scored on SoilGrids pH), every model produced a positive mean Boyce index at the genus level (the cross-validated twelve-genus set differs in composition from the full-dataset set; see Materials and methods): the canonical trapezoid + 0.20, a flexible single-axis pH spline + 0.28, random forest + 0.25 and gradient boosting + 0.23 (Fig. 3). The flexible pH spline matched or exceeded the all-axis machine-learning models, indicating that the marginal contribution of the additional edaphic and climatic variables is absorbed by a richer pH response rather than by independent information.

At the species level, the contrast sharpened. The pH-only models retained positive transfer (trapezoid + 0.15, pH spline + 0.24) while the machine-learning benchmarks fell below zero (random forest − 0.03, gradient boosting − 0.09); the trapezoid returned positive transfer for 32 of 38 species, the random forest for only 17. With no parameters to tune, the trapezoid does not overfit the small per-species samples on which the flexible learners do. The area under the ROC curve reversed this ordering — the machine-learning models scored higher (for example, genus-level AUC 0.59 versus 0.54 for the trapezoid) — but this advantage is not realizable in mapping, which is driven by the pH raster rather than the full axis set, and AUC is, in any case, distorted by the choice of pseudo-absences in presence-only data. Both metrics are reported; the binned Boyce index is the primary criterion.

Concordance across databases and markers

Transferred to the independently assembled GlobalAMFungi database, the GlobalFungi-calibrated canonical nodes did not retain a positive cross-database Boyce index (mean − 0.04 across the nine genera with sufficient GlobalAMFungi records; Table S1), which is unsurprising given that GlobalAMFungi pH agrees far less well with the SoilGrids surface than GlobalFungi does (Spearman $\rho \approx 0.49$ versus 0.89; Table S5) and the available ITS2 subset is small. Data-driven nodes nonetheless transferred: percentile nodes fitted on GlobalFungi scored + 0.26 on GlobalAMFungi and outperformed nodes fitted internally on GlobalAMFungi (+ 0.18; Table S1), so the larger and pH-cleaner GlobalFungi remains the better calibration base. The transferability of the niches themselves is shown instead by the agreement of independently estimated optima. At the species level, pH optima of the 22 taxa shared between the two databases agreed in rank (Spearman $\rho = +0.45$, $p = 0.033$), with a root-mean-square deviation of 0.69 pH and a systematic offset of − 0.31 pH (GlobalAMFungi optima higher); 82% of species fell within 1 pH unit of the 1:1 line (Fig. 4). The offset is consistent with the marker (ITS versus ITS2) and pH-protocol differences between the databases.

A cross-marker comparison gave independent support. Across the eight AM fungal families for which ITS-genus and SSU-VT optima could be matched, the pH niche optima derived here agreed in rank order with those of Davison et al. (2021) (Spearman $\rho = +0.81$, $p = 0.015$; their Table S6), strengthening to $\rho = +0.90$ ($p = 0.04$) among the five families represented by a single HESI genus. HESI optima were

systematically higher than the Davison estimates by 1.36 pH units, consistent with the lower readings of KCl extraction relative to the water pH used here and with the abundance-weighted basis of the Davison optima; accordingly, only rank order is interpreted. The most conspicuous departure was *Gigasporaceae*, ranked low on pH by Davison et al. yet carrying the highest temperature and precipitation optima of any family (median VT optima 22.4°C and 2,124 mm) — the expected behavior of a taxon whose niche is structured by climate rather than soil pH (see Discussion). This concordance indicates that the genus-level ITS niches recovered here are not marker artifacts but track an independent SSU-based characterization, while resolving pH preference at the genus level — including within *Glomeraceae*, where the SSU analysis distinguishes only DNA-based clades.

Global suitability maps

Projecting each genus trapezoid onto the SoilGrids pH surface yielded twelve global per-genus suitability layers (Fig. 5), whose spatial structure follows the niche grouping of Table 1. The acidophil anchors (Acaulospora, Ambispora, Archaeospora) reach highest suitability across the humid, base-poor tropics and high-latitude soils — the Amazon and Congo basins, Southeast Asia, and the boreal zone — and fall to near zero across arid carbonate regions. The alkaliphil genera (Entrophospora, Funneliformis, Septoglomus) and the higher-centered sub-acidophils such as Diversispora show the inverse, peaking over the semi-arid calcareous belts of the Sahel margin, the Middle East, interior Australia, and the central-Asian and Andean steppe. The canonical *Glomus* trapezoid is shifted toward high pH, so its projected suitability peaks in arid, alkaline regions; its negative Boyce index reflects that this single shape does not match the broad distribution in which *Glomus* actually occurs — the expected outcome for a genus whose constituent species span the entire gradient.

White areas are unscored, and of two kinds: pixels with no soil or non-vegetated cover and — distinct from missing data — pixels whose SoilGrids pH falls outside the calibration window (3.9–9.2). The second kind removes the most strongly alkaline hyperarid surfaces, including the central Sahara, the Arabian Peninsula, and parts of interior Australia, which exceed the upper calibration bound.

Each layer is a projection of the per-genus pH membership function alone and is to be read as a spatially explicit hypothesis, rather than as predicted occurrence or abundance (see Discussion).

Discussion

A transparent, transferable pH-suitability index for AM fungal genera

HESI shows that the soil-pH niche of most AM fungal genera can be captured by a single membership function: it transfers to unsampled regions amid spatial cross-validation, transfers to an independently assembled database without further training, and agrees in rank with niche optima derived from a different marker by an independent study. The index is strongest for genera with well-defined acidophil or alkaliphil niches and weakest for the cosmopolitan generalist (*Glomus*) and for the genus that pools

species of opposite pH preference (*Ambispora*). That these two cases are recovered as the diagnostic limits of a one-axis model rather than as noise is itself evidence that the index behaves as a niche descriptor should.

Two findings carry the conceptual weight. First, a single read-not-fitted pH axis does not gain robust improvement by adding further environmental axes or by machine learning, on the presence-only metric by which suitability should be judged: the advantage of flexible learners on AUC does not survive on the binned Boyce index, and at the species level, the learners fall below zero while the parameter-free trapezoid continues positive. Second, the genus-level niches recovered from ITS data reproduce across databases and concord with the SSU-defined virtual-taxon niches of Davison et al. (2021).

Why one axis suffices, and how this reconciles with recent global maps

The benchmark clarifies what a richer model would and would not add. The small advantage that random forests hold on AUC traces to the shape of the pH response, rather than to the additional edaphic and climatic axes or to interactions among them, which contribute nothing under spatial transfer essentially; a flexible single-axis pH spline recovers the learners' AUC and exceeds them on the Boyce index. The transferable signal in these data is one-dimensional. This is not a claim that AM fungal distributions are governed by pH alone — much of where these fungi occur is captured by none of the abiotic layers available here (dispersal, host, biotic and abiotic interactions, history, etc.) — but that, among the measured environmental predictors, pH is the one that transfers for per-genus presence.

This places HESI in a coherent relationship with the two most relevant recent studies. Davison et al. (2021) identified temperature and pH as the leading niche axes of AM fungal VT in an explanatory, in-sample analysis; the global machine-learning maps of Van Nuland et al. (2025) report that soil pH is not among the top predictors of AM fungal richness, which temperature dominates. The three results describe one phenomenon at different levels of aggregation and under different evaluation paradigms. Richness sums across taxa whose per-genus pH optima are opposite, while temperature sets how many taxa accumulate. The evaluation paradigms compound the difference: richness is assessed by random cross-validation, which rewards spatially smooth climatic predictors, whereas HESI is assessed by spatial cross-validation, which penalizes predictors that do not transfer.

HESI does more than operationalize a known dependence: to our knowledge, it is the first to separate, under spatial transfer, which of Davison's two niche axes actually generalizes for per-taxon presence, and that axis is pH. Temperature's per-taxon association, although real in-sample, does not drastically improve blocked cross-validation results; this is consistent with a physiologically conservative pH niche and a regionally contingent temperature association. The cross-marker concordance reinforces the same boundary. It's one conspicuous rank departure, *Gigasporaceae*, is exactly the family whose niche Davison et al. (2021) place in high precipitation and high temperature rather than pH — the expected behavior of a taxon for which pH is not the operative axis, and a signal that a single-axis index will, correctly, misrank such taxa.

Validation as transferability, not map accuracy

We use spatial block cross-validation to test whether the membership function transfers to unsampled space, not to estimate the accuracy of the final maps. The distinction matters because the literature motivating blocked designs also cautions that spatial cross-validation should not be interpreted as a measure of map accuracy (Wadoux et al. 2021; Meyer and Pebesma 2022). That caution concerns the interpretation of a cross-validated score as a measure of map quality, which we do not claim; what blocked cross-validation does establish — that a function calibrated on some continents predicts presence on a withheld one — is exactly the property a transferable niche descriptor must possess. In the same spirit, we treat the binned Boyce index, a calibration and monotonicity criterion suited to presence-only data, as primary, and report AUC only as an auxiliary diagnostic, because AUC is distorted by the choice of pseudo-absences and, as shown here, rewards discrimination of a kind that is not realizable once the index is projected onto a single pH raster.

Limitations

HESI is a one-axis, descriptive index of the realized niche, and its scope should be read accordingly. It models soil pH alone; for taxa whose primary axis is moisture, texture, or climate — *Gigasporaceae* being the clearest case — a pH-only suitability is weak because the relevant axis is absent rather than mis-estimated. This is a deliberate scope choice, since pH is the axis that both transfers and can be mapped globally from a single product, but it means HESI characterizes rather than fully predicts AM fungal distributions and serves as a step toward, not a substitute for, a multi-axis model. The niche it fixes on is realized, not fundamental: the trapezoids reflect where genera are found in a sampled, presence-only dataset shaped also by dispersal, competition, and sampling bias, with a target-group background whose mode near pH 6.2 structurally reduces the contrast available to neutral-niche genera. The predictor surface, SoilGrids, is itself a machine-learning product, so comparison with it is at the level of the predictor rather than against ground truth. The genus is a heterogeneous unit — *Glomus*, *Claroideoglomus*, and *Ambispora* span species from acidic to alkaline — and a single trapezoid per genus is a summary that can mask opposing species-level signals, as the negative genus scores make explicit. Finally, the ITS-genus niches reconcile with SSU-VT niches only at the family level; within *Glomeraceae*, the resolution at which our genera differ is the resolution at which the cross-marker comparison cannot adjudicate, and AM fungal taxonomy at this level remains in flux (Błaszkowski et al. 2022; da Silva et al. 2025).

Significance and outlook

As a transparent per-genus instrument, HESI turns the established pH dependence of AM fungi into quantities that can be used and tested. The global suitability maps are best read as spatially explicit hypotheses, to be confronted with independent field sampling, not as maps of observed occurrence or abundance. The index also offers a direct way to anticipate how genus-level AM fungal composition might change under shifts in soil pH, whether due to liming, acidification, or land-use change.

Methodologically, HESI is one realised axis in a programme that both the wider literature and our own results point toward: characterising each niche axis as a transparent, separately validated transfer function — the modern descendant of indicator-value systems such as Ellenberg-type values (Tichý et al. 2023) — with temperature the next axis (Davison et al. 2021); treating the combination of axes as an open question rather than an assumed product; recalibrating at finer, curated VT resolution as databases such as GlobalAMFungi mature (Větrovský et al. 2023); and ultimately grounding the index in independent field validation and in the physiology that the read-from-profile nodes only approximate. The value of the present model is that it makes the pH axis transparent and validates it hard, leaving a clear interface for future work.

Conclusion

HESI condenses the soil-pH niche of twelve AM fungal genera into a single, transparent membership function whose parameters are read from the empirical presence-rate profile rather than fitted, and whose behavior is testable: it transfers to unsampled regions during spatial cross-validation, transfers to an independent database in the absence of retraining, and agrees in rank with the SSU-defined niches of an independent study. It does not rediscover that these fungi depend on soil pH but operationalizes that dependence into a usable per-genus instrument and, in doing so, separates the two axes of the established niche framework — soil pH, which determines which genera occur and their transfer between continents, from temperature, which governs how many taxa accumulate. The cosmopolitan and pH-heterogeneous genera that the index cannot resolve, and the climate-defined family it correctly misranks, delimit its scope rather than undermine it. The resulting global suitability maps are spatially explicit hypotheses for field testing, and the index provides a transparent template for extending niche modeling axis by axis as curated, finer-resolution data accumulate.

Declarations

Competing interests: The author declares no competing financial or non-financial interests.

Funding: No funding was received.

Author contributions: Tsyhliuk D. is the sole author and is responsible for all aspects of the work.

Data availability: The occurrence and soil data analyzed here are openly available from the GlobalFungi, GlobalAMFungi, and SoilGrids repositories; derived per-genus node parameters and analysis code are available at https://doi.org/10.5281/zenodo.20581579.

References

1. Błaszkowski J, Sánchez-García M, Niezgoda P, Zubek S, Fernández F, Vila A, Al-Yahya'ei MN, Symanczik S, Milczarski P, Malinowski R, Cabello M, Goto BT, Casieri L, Malicka M, Bierza W,

Magurno F (2022) A new order, *Entrophosporales*, and three new *Entrophospora* species in *Glomeromycota*. Front Microbiol 13:962856. https://doi.org/10.3389/fmicb.2022.962856

2. Boyce MS, Vernier PR, Nielsen SE, Schmiegelow FKA (2002) Evaluating resource selection functions. Ecol Model 157:281–300. https://doi.org/10.1016/S0304-3800(02)00200-4
3. Clark RB (1997) Arbuscular mycorrhizal adaptation, spore germination, root colonization, and host plant growth and mineral acquisition at low pH. Plant Soil 192:15–22. https://doi.org/10.1023/A:1004218915413
4. da Silva GA, Sieverding E, de Assis DMA, Goto BT, Corazon-Guivin MA, Oehl F (2025) Revision of *Entrophosporales*, with three genera and an identification key for all species currently attributed to this order. J Fungi 11:97. https://doi.org/10.3390/jof11020097
5. Davison J, Moora M, Öpik M, Adholeya A, Ainsaar L, Bâ A, Burla S, Diedhiou AG, Hiiesalu I, Jairus T et al (2015) Global assessment of arbuscular mycorrhizal fungus diversity reveals very low endemism. Science 349:970–973. https://doi.org/10.1126/science.aab1161
6. Davison J, Moora M, Semchenko M, Adenan SB, Ahmed T, Akhmetzhanova AA, Alatalo JM, Al-Quraishy S, Andriyanova E, Anslan S, Bahram M, Batbaatar A, Brown C, Bueno CG, Cahill J, Cantero JJ, Casper BB, Cherosov M, Chideh S, Coelho AP, Coghill M, Decocq G, Dudov S, Fabiano EC, Fedosov VE, Fraser L, Glassman SI, Helm A, Henry HAL, Hérault B, Hiiesalu I, Hiiesalu I, Hozzein WN, Kohout P, Kõljalg U, Koorem K, Laanisto L, Mander Ü, Mucina L, Munyampundu JP, Neuenkamp L, Niinemets Ü, Nyamukondiwa C, Oja J, Onipchenko V, Pärtel M, Phosri C, Põlme S, Püssa K, Ronk A, Saitta A, Semboli O, Sepp SK, Seregin A, Sudheer S, Peña-Venegas CP, Paz C, Vahter T, Vasar M, Veraart AJ, Tedersoo L, Zobel M, Öpik M (2021) Temperature and pH define the realised niche space of arbuscular mycorrhizal fungi. New Phytol 231:763–776. https://doi.org/10.1111/nph.17240
7. Dumbrell AJ, Nelson M, Helgason T, Dytham C, Fitter AH (2010) Relative roles of niche and neutral processes in structuring a soil microbial community. ISME J 4:337–345. https://doi.org/10.1038/ismej.2009.122
8. Guardiola-Márquez CE, Pacheco A, Mora-Godínez S et al (2022) *Septoglomus* species dominate the arbuscular mycorrhiza of five crop plants in an arid region of northern Mexico. Symbiosis 87:93–106. https://doi.org/10.1007/s13199-022-00851-2
9. Hengl T, Mendes de Jesus J, Heuvelink GBM, Ruiperez Gonzalez M, Kilibarda M, Blagotić A, Shangguan W, Wright MN, Geng X, Bauer-Marschallinger B, Guevara MA, Vargas R, MacMillan RA, Batjes NH, Leenaars JGB, Ribeiro E, Wheeler I, Mantel S, Kempen B (2017) SoilGrids250m: global gridded soil information based on machine learning. PLoS ONE 12:e0169748. https://doi.org/10.1371/journal.pone.0169748
10. Hirzel AH, Le Lay G, Helfer V, Randin C, Guisan A (2006) Evaluating the ability of habitat suitability models to predict species presences. Ecol Model 199:142–152. https://doi.org/10.1016/j.ecolmodel.2006.05.017
11. Kawahara A, An G-H, Miyakawa S, Sonoda J, Ezawa T (2016) Nestedness in arbuscular mycorrhizal fungal communities along soil pH gradients in early primary succession: acid-tolerant fungi are pH

generalists. PLoS ONE 11:e0165035. https://doi.org/10.1371/journal.pone.0165035

12. Lobo JM, Jiménez-Valverde A, Real R (2008) AUC: a misleading measure of the performance of predictive distribution models. Glob Ecol Biogeogr 17:145–151. https://doi.org/10.1111/j.1466-8238.2007.00358.x
13. Meyer H, Pebesma E (2022) Machine learning-based global maps of ecological variables and the challenge of assessing them. Nat Commun 13:2208. https://doi.org/10.1038/s41467-022-29838-9
14. Öpik M, Vanatoa A, Vanatoa E, Moora M, Davison J, Kalwij JM, Reier Ü, Zobel M (2010) The online database MaarjAM reveals global and ecosystemic distribution patterns in arbuscular mycorrhizal fungi (*Glomeromycota*). New Phytol 188:223–241. https://doi.org/10.1111/j.1469-8137.2010.03334.x
15. Parihar M, Rakshit A, Singh HB, Rana K (2019) Diversity of arbuscular mycorrhizal fungi in alkaline soils of hot sub humid eco-region of Middle Gangetic Plains of India. Acta Agric Scand Sect B Soil Plant Sci 69:386–397. https://doi.org/10.1080/09064710.2019.1582692
16. Phillips SJ, Dudík M, Elith J, Graham CH, Lehmann A, Leathwick J, Ferrier S (2009) Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecol Appl 19:181–197. https://doi.org/10.1890/07-2153.1
17. Ploton P, Mortier F, Réjou-Méchain M, Barbier N, Picard N, Rossi V, Dormann C, Cornu G, Viennois G, Bayol N, Lyapustin A, Gourlet-Fleury S, Pélissier R (2020) Spatial validation reveals poor predictive performance of large-scale ecological mapping models. Nat Commun 11:4540. https://doi.org/10.1038/s41467-020-18321-y
18. Poggio L, de Sousa LM, Batjes NH, Heuvelink GBM, Kempen B, Ribeiro E, Rossiter D (2021) SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. SOIL 7:217–240. https://doi.org/10.5194/soil-7-217-2021
19. Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillera-Arroita G, Hauenstein S, Lahoz-Monfort JJ, Schröder B, Thuiller W, Warton DI, Wintle BA, Hartig F, Dormann CF (2017) Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40:913–929. https://doi.org/10.1111/ecog.02881
20. Robertson MP, Villet MH, Palmer AR (2004) A fuzzy classification technique for predicting species' distributions: applications using invasive alien plants and indigenous insects. Divers Distrib 10:461–474. https://doi.org/10.1111/j.1366-9516.2004.00108.x
21. Šibanc N, Clark DR, Suhadolc M, Leštan D, Dumbrell AJ, Maček I (2026) Indigenous inoculant dampens the impact of remediation of heavy metal polluted soil on arbuscular mycorrhizal fungal communities. Mycorrhiza 36:23. https://doi.org/10.1007/s00572-026-01268-1
22. Smith SE, Read DJ (2008) Mycorrhizal symbiosis, 3rd edn. Academic, San Diego
23. Tedersoo L, Sánchez-Ramírez S, Kõljalg U, Bahram M, Döring M, Schigel D, May T, Ryberg M, Abarenkov K (2018) High-level classification of the Fungi and a tool for evolutionary ecological analyses. Fungal Divers 90:135–159. https://doi.org/10.1007/s13225-018-0401-0

24. Tichý L, Axmanová I, Dengler J et al (2023) Ellenberg-type indicator values for European vascular plant species. J Veg Sci 34:e13168. https://doi.org/10.1111/jvs.13168
25. Treseder KK, Allen EB, Egerton-Warburton LM, Hart MM, Klironomos JN, Maherali H, Tedersoo L, Wurzburger N (2018) Arbuscular mycorrhizal fungi as mediators of ecosystem responses to nitrogen deposition: a trait-based predictive framework. J Ecol 106:480–489. https://doi.org/10.1111/1365-2745.12919
26. Valavi R, Elith J, Lahoz-Monfort JJ, Guillera-Arroita G (2019) blockCV: an R package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. Methods Ecol Evol 10:225–232. https://doi.org/10.1111/2041-210X.13107
27. Van Nuland ME, Averill C, Stewart JD, Prylutskyi O, Corrales A, van Galen LG, Manley BF, Qin C, Lauber T, Mikryukov V, Dulia O, Furci G, Marín C, Sheldrake M, Weedon JT, Peay KG, Cornwallis CK, Větrovský T, Kohout P, Baldrian P, Tedersoo L, West SA, Crowther TW, Kiers ET, SPUN Mapping Consortium, van den Hoogen J (2025) Global hotspots of mycorrhizal fungal richness are poorly protected. Nature 645:414–422. https://doi.org/10.1038/s41586-025-09277-4
28. Veresoglou SD, Caruso T, Rillig MC (2013) Modelling the environmental and soil factors that shape the niches of two common arbuscular mycorrhizal fungal families. Plant Soil 368:507–518. https://doi.org/10.1007/s11104-012-1531-x
29. Větrovský T, Morais D, Kohout P, Lepinay C, Algora C, Awokunle Hollá S, Bahnmann BD, Bílohnědá K, Brabcová V, D'Alò F, Human ZR, Jomura M, Kolařík M, Kvasničková J, Lladó S, López-Mondéjar R, Martinović T, Mašínová T, Meszárošová L, Michalčíková L, Michalová T, Mundra S, Navrátilová D, Odriozola I, Piché-Choquette S, Štursová M, Švec K, Tláskal V, Urbanová M, Vlk L, Voříšková J, Žifčáková L, Baldrian P (2020) GlobalFungi, a global database of fungal occurrences from high-throughput-sequencing metabarcoding studies. Sci Data 7:228. https://doi.org/10.1038/s41597-020-0567-7
30. Větrovský T, Kohout P, Morais D, Větrovská J, Awokunle Hollá S, Choi J, Davison J, Tláskal V, Vályi K et al (2023) GlobalAMFungi: a global database of arbuscular mycorrhizal fungal occurrences from high-throughput sequencing metabarcoding studies. New Phytol 240:2151–2163. https://doi.org/10.1111/nph.19283
31. Wadoux AMJ-C, Heuvelink GBM, de Bruin S, Brus DJ (2021) Spatial cross-validation is not the right way to evaluate map accuracy. Ecol Model 457:109692. https://doi.org/10.1016/j.ecolmodel.2021.109692
32. Zadeh LA (1965) Fuzzy sets. Inf Control 8:338–353. https://doi.org/10.1016/S0019-9958(65)90241-X

Figures

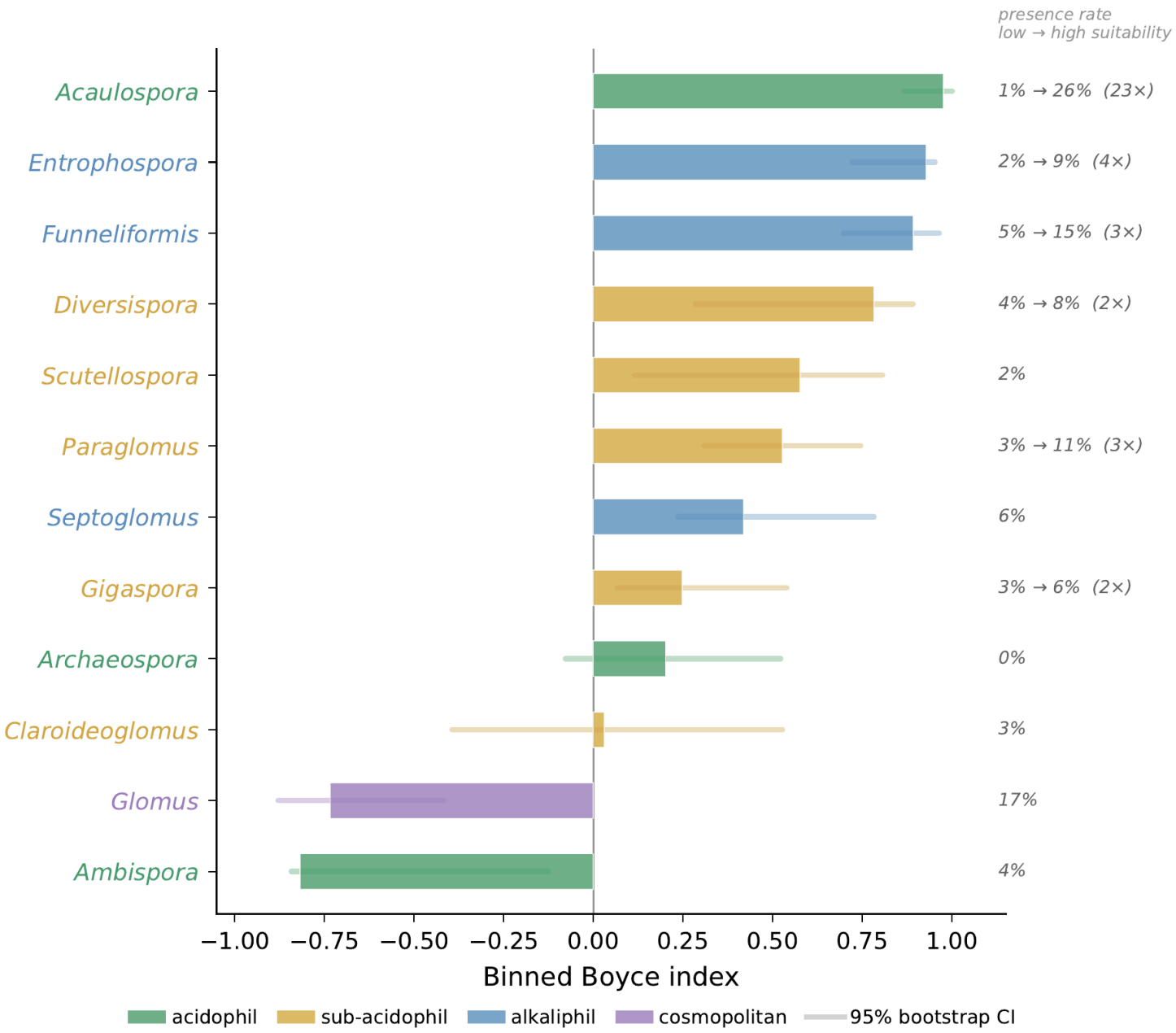


Figure 1

Full-dataset validation of the twelve genus pH-suitability functions. Binned Boyce index for each genus, scored on the independent SoilGrids pH surface (Q0.5, 0–20 cm) across the 7,460-record GlobalFungi pool; bars are coloured by niche group and horizontal lines show 95% bootstrap confidence intervals. Right-hand labels show the observed presence rate at low versus high predicted suitability, along with their fold change. Ten of twelve genera score positive, and eight have confidence intervals entirely above zero.

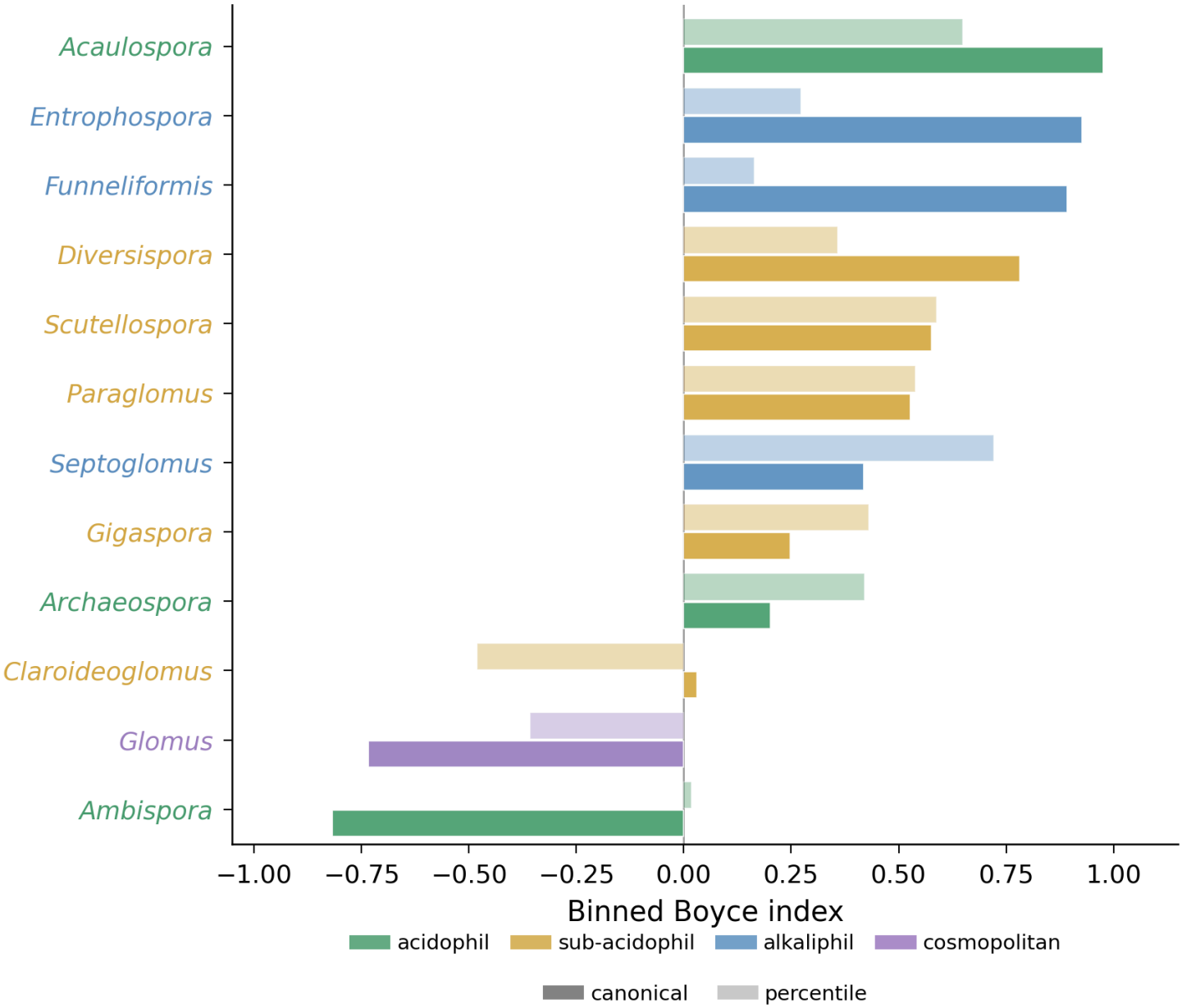


Figure 2

Sensitivity of the binned Boyce index to node estimation. Canonical nodes read directly from the binned presence-rate profile (darker bars) compared with an automatic percentile estimator (lighter bars), scored on the SoilGrids pH surface and coloured by niche group. The read-from-profile nodes give a higher mean index (+0.34 versus +0.28); the clearest gain is *Claroideoglomus*, which moves from −0.48 under percentile nodes to +0.03 under the canonical nodes.

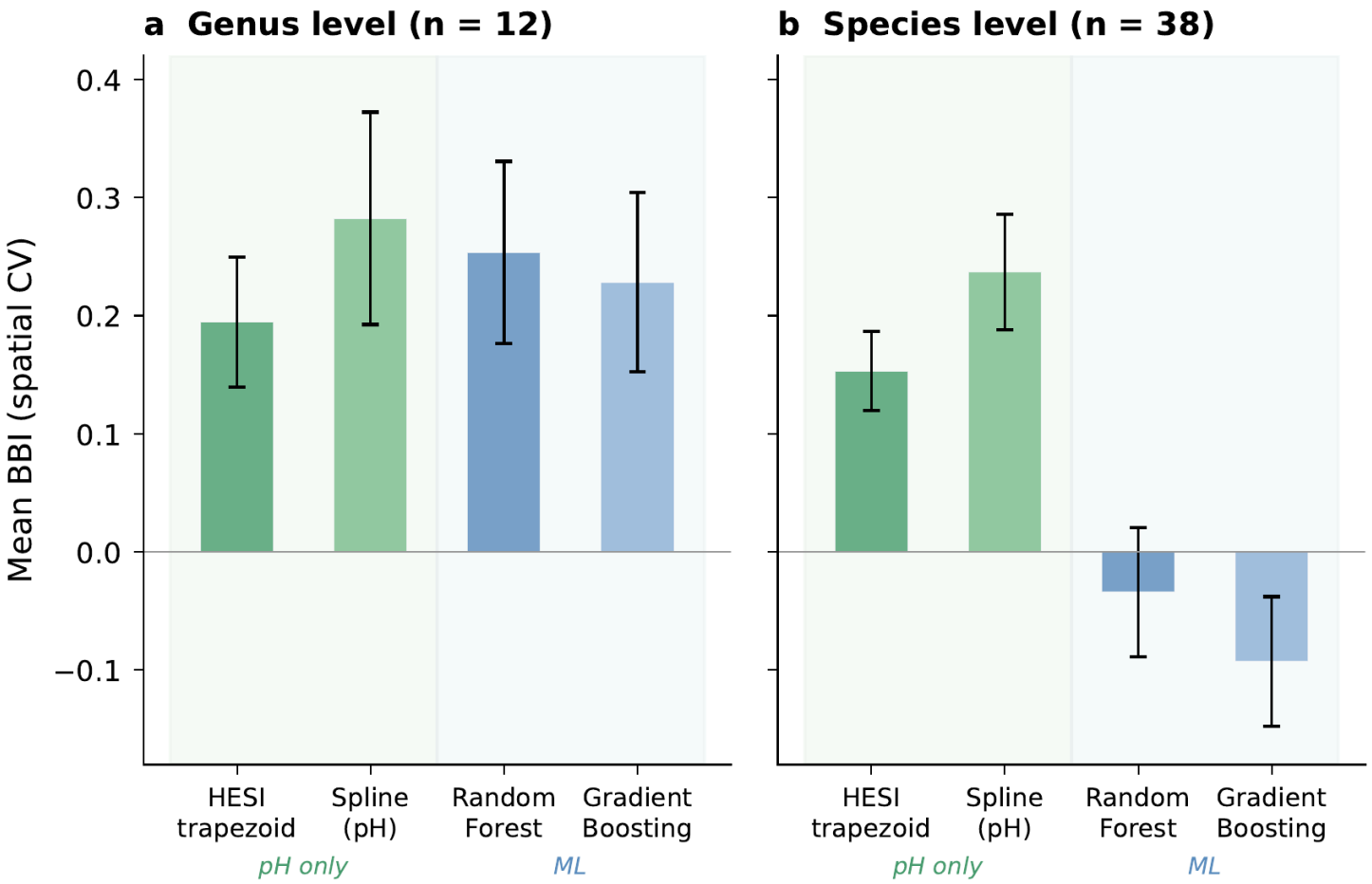


Figure 3

Spatial-transfer benchmark of HESI against flexible single-axis and machine-learning models. Mean binned Boyce index under spatial block cross-validation at the genus (a) and species (b) levels for the HESI trapezoid, a single-axis pH spline, random forest, along with gradient boosting; error bars show variation across folds. The pH-only models retain positive transfer at both levels, whereas the machine-learning learners fall below zero at the species level.

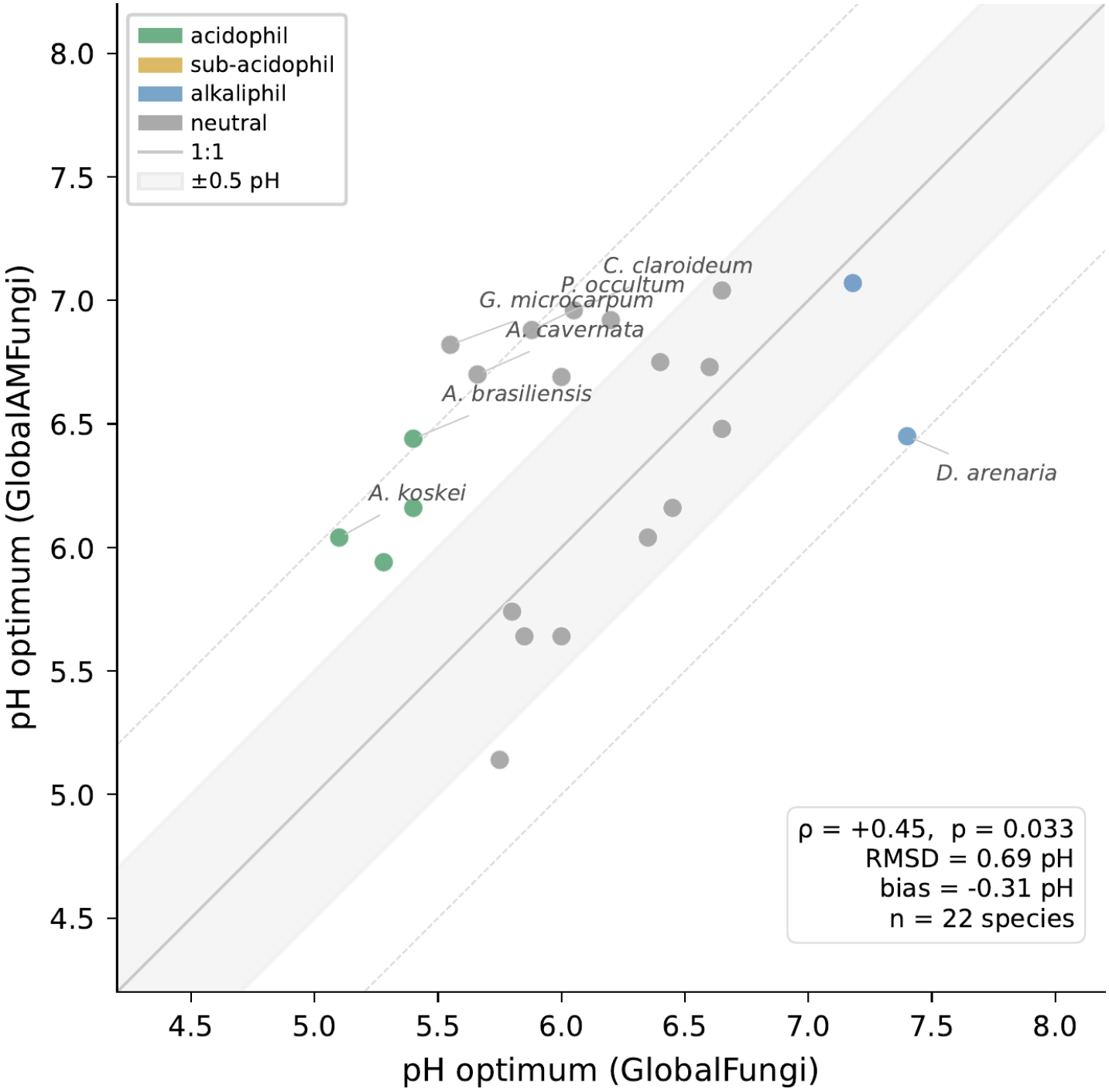


Figure 4

Cross-database concordance of species-level pH optima. pH optimum estimated independently in GlobalFungi (ITS) and GlobalAMFungi (ITS2) for the 22 species shared between the databases; the solid line is the 1:1 relationship, and the shaded band ±0.5 pH. Points are coloured by niche group. Optima agree in rank (Spearman ρ = +0.45, p = 0.033) with a root-mean-square deviation of 0.69 pH and a systematic offset of −0.31 pH, GlobalAMFungi optima being higher.

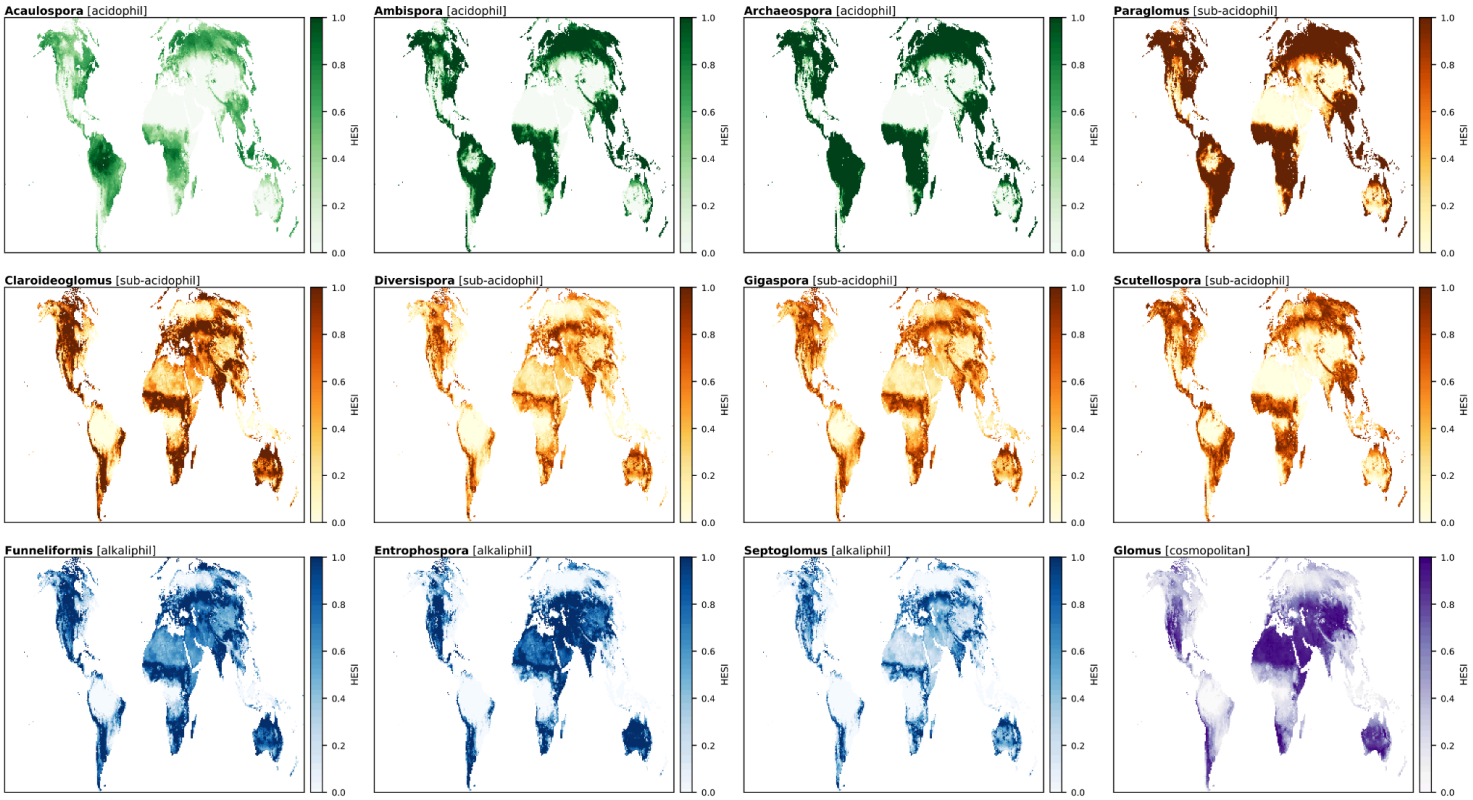


Figure 5

Global per-genus soil-pH suitability surfaces. Each panel projects one genus trapezoid onto the SoilGrids pH surface (Q0.5, 0–20 cm), with warmer shades indicating higher pH suitability; panels are grouped by niche group. White areas are unscored, comprising pixels with no soil or non-vegetated cover and pixels whose pH falls outside the calibration range (3.9–9.2). Acidophil and alkaliphil surfaces are close to spatial complements, while the cosmopolitan *Glomus*, on its high-pH-shifted trapezoid, reaches highest suitability over arid alkaline regions.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- HESIsupplementarytables.xlsx