nature communications

Article

https://doi.org/10.1038/s41467-025-59253-9

The powerbend distribution provides a unified model for the species abundance distribution across animals, plants and microbes

Received: 24 July 2024

Accepted: 16 April 2025

Published online: 29 April 2025

Check for updates

Yingnan Gao[1,2], Ahmed Abdullah[1] & Martin Wu[1]

Remarkably, almost every ecological community investigated to date is composed of many rare species and a few abundant species. While the precise nature of this species abundance distribution is believed to reflect fundamental ecological principles underlying community assembly, ecologists have yet to identify a single model that comprehensively explains all species abundance distributions. Recent studies using large datasets have suggested that the logseries distribution best describes animal and plant communities, while the Poisson lognormal distribution is the best model for microbes, thereby challenging the notion of a unifying species abundance distribution model across the tree of life. Here, using a large dataset of ~30,000 globally distributed communities spanning animals, plants and microbes from diverse environments, we show that the powerbend distribution, predicted by a maximum information entropy-based theory of ecology, emerges as a unifying model that accurately captures species abundance distributions of all life forms, habitats and abundance scales. Our findings challenge the notion of pure neutrality, suggesting instead that community assembly is driven by a combination of random fluctuations and deterministic mechanisms shaped by interspecific trait variation.

The species abundance distribution (SAD) follows one of ecology's oldest and most universal laws. It describes the commonness and rarity in ecosystems, namely, the abundance (number of individuals) of each species in a community. Remarkably, almost every animal or plant community investigated to date is dominated by a few species, and most species in the communities are rare[1]. This universal hollow curve of SAD holds true in communities of different spatial scales, habitat types, and taxonomic groups. In recent years, such a hollow-curve pattern, known as 'rare biosphere' to microbiologists[2], has also been found to be universal in microbial communities.

The universality of the hollow curve SADs is both surprising and informative. It suggests that there might be universal principles operating across habitats, taxonomic groups, and spatial scales. In particular, the shape of SAD is thought to reflect key ecological processes involved in community assembly. If we can explain this high degree of unevenness, then we can gain insight into the mechanisms that structure communities, whether they involve stochastic processes, deterministic processes (e.g., species traits and niche partitioning), or a combination of both. As a result, SAD has been extensively studied in animals and plants, and dozens of models have been proposed to explain the shape of the SAD (see ref. 1 for a review). Among them, the best-known models are logseries[3], lognormal[4], broken-stick[5], geometric series[6], and Zipf power law[7].

[1]Department of Biology, University of Virginia, Charlottesville, VA, USA. [2]Present address: Department of Genome Sciences, University of Washington, Seattle, WA, USA. e-mail: mw4yv@virginia.edu

These models range from purely statistical ones selected for optimal data fitting[3] to those explicitly incorporating ecological processes[8–10]. For example, while logseries was initially developed by Fisher to fit empirical data[3], it has subsequently been predicted by Hubbell's neutral theory[11] and Harte's maximum entropy theory of ecology (METE)[12,13]. Hubbell's Neutral Theory proposes that species abundances and distributions are shaped by random processes such as birth, death, dispersal, and speciation rather than ecological differences. It assumes all species are functionally equivalent, ignoring trait differences. Similarly, METE, based on the maximum information entropy principle (MaxEnt), also assumes no prior trait differences among species. It posits that the most likely form of an ecological pattern is the one that represents the most unbiased or least-informative distribution given a set of ecological constraints (e.g., the average species abundance). These constraints represent the deterministic factors that are believed to be imposed on an ecosystem. This MaxEnt framework therefore offers a platform for investigating and assessing factors that underlie the universal hollow-curve nature of SADs. For example, using MaxEnt and modeling intrinsic species trait differences, a new SAD model has been proposed[14,15]. This model, termed 'powerbend' in the R package 'sads', is a modified version of the power law, distinguished by its establishment of an upper limit on the abundances of the most dominant species within a community[16]. This model, exceptionally versatile in nature, encompasses all aforementioned SAD models, with the exception of the Poisson lognormal model (Supplementary Note 1). Nonetheless, powerbend remains relatively obscure compared to other models and has not been widely tested.

Due to the fundamental importance of SAD in biodiversity studies, there has been a long-standing quest for a unifying SAD model for all life forms[1]. The Poisson lognormal and logseries distributions are the two most successful SAD models, and they are often used as benchmarks to test other models. For example, White et al. and Baldridge et al. tested several commonly used SAD models in about 16,000 animal and plant communities from terrestrial, aquatic, and marine environments[17,18]. They found that logseries is the overall best SAD model based on the Akaike Information Criterion (AIC). However, previous studies have also indicated that the Poisson lognormal model[19–21], as well as the less commonly used gambin[22,23] and Weibull models[24], are preferred in some animal and plant communities. In another large-scale study of over 20,000 bacterial and archaeal communities, Shoemaker et al. found that Poisson lognormal, not logseries, best describes microbial SADs[25]. In contrast to animals and plants, bacteria and archaea reproduce asexually and have relatively large population sizes, high dispersal rates, and short generation times. Given these key differences, the finding that microorganisms and macroorganisms may have distinct SADs raises a key question: are there unifying macroecological rules and ecological theories that can explain SADs across the tree of life?

To address this question, we test whether a universal SAD model unites all types of organisms, large and small. Using a large dataset encompassing animals, plants, and microbes, we establish the emergence of powerbend as a unifying SAD model across communities of broad scales, habitats and taxonomic groups. Our findings support the existence of universal ecological principles governing the assembly of animal, plant, and microbial communities, driven by both deterministic and neutral processes.

Results

Powerbend accurately captures SADs of animal and plant communities

In this study, we focused on four SAD models—Poisson lognormal, logseries, power law, and powerbend (Supplementary Table S1). The selection of the Poisson lognormal, logseries, and power law models was based on their widespread use and extensive testing in large-scale studies of animal, plant, and microbial communities[17,18,25]. In contrast, the more flexible powerbend model has received little attention previously[26]. To enable direct comparisons with previous findings, we utilized the datasets compiled by Baldridge et al.[18] and Shoemaker et al.[25] in our study. In terms of the goodness of fit, as measured by the modified coefficient of determination r_m^2 [17,25,27], powerbend explains an average of 93.2% of the variation (weighted by the size of datasets representing different taxonomic groups, Fig. 1a) in 13,819 animal and plant SADs (Supplementary Table S2). In comparison, Poisson lognormal explains 94.7% of the variation, while logseries explains 73.2% (Fig. 1a). Using Monte-Carlo simulations, we found that powerbend, Poisson lognormal and logseries have r_m^2 values not significantly different from 1.0 ($r_m^2 = 1.0$ indicates a perfect fit) in 99.5%, 100% and 88.7% of SADs, respectively. Furthermore, compared to the other models, powerbend produces unbiased predictions regardless of the scale of species abundance (Fig. 1a). Poisson lognormal, while equal to powerbend in terms of the overall predictive power, tends to overestimate the abundance of the most abundant taxa (Fig. 1a, b), and performs poorly in predicting the evenness and rareness of the SAD (see below). In contrast to the other models, power law fits the data poorly ($r_m^2 = -0.079$). r_m^2 calculated using unweighted samples showed similar results (Supplementary Table S3).

In addition to evaluating goodness of fit (r_m^2), we also compared models using AIC, an approach strongly recommended for SAD model testing[28] and employed in the Shoemaker et al. study[25]. Our simulations show that when the number of observed species in a SAD is less than 40, AIC-based model selection does not have enough power to distinguish between SAD models, often favoring the simpler models even when the simpler model is incorrect (Supplementary Fig. S1). The small number of species in the Baldridge et al. dataset[18] (weighted mean: 36.8 species per SAD) limited the power of AIC-based model selection in animal and plant communities. According to AIC, powerbend is significantly better than logseries in 20.88% of animal and plant SADs (ΔAIC ≥ 2), while logseries is significantly better only in 0.04% of animal and plant communities. Similarly, powerbend is significantly better than Poisson lognormal in 16.44% of SADs, while Poisson lognormal is significantly better in 11.17%. Moreover, powerbend significantly outperforms the Weibull model in 84.27% of SADs, while the Weibull model is significantly better in only 1.33% of cases.

Our results are consistent with previous findings that distinguishing between some SAD models is challenging in animal and plant communities[1,18,28]. Next, we test the models in microbial communities, which have substantially greater species richness.

Powerbend provides the best fit to microbial communities

Although 16S rRNA sequencing is widely employed for determining the microbial species abundance, there are certain challenges when using 16S rRNA sequence data to test SAD models. One key challenge arises from the fact that we only count the number of reads from a species (commonly defined as an operational taxonomic unit, or OTU, at 97% sequence identity threshold), rather than the actual number of individual cells present. To establish a connection between 16S rRNA read numbers and absolute species abundances, it is essential to account for the sampling effort within the 16S rRNA sequencing pipeline by incorporating a sampling error such as the Poisson distribution.

Shoemaker et al. have shown that the Poisson lognormal model appears to be the best SAD model for microbial communities[25]. However, it is important to note that the Poisson lognormal model was the only model in that study to incorporate a Poisson sampling error. This could confer an inherent advantage to the Poisson lognormal model over the other SAD models because 16S rRNA sequencing inherently involves multiple sampling processes. To select the best SAD model in microbial communities, we fitted Poisson lognormal and three models (logseries, power law, and powerbend) with and without a Poisson sampling error to 15,329 microbial SADs (Supplementary Table S2).

https://doi.org/10.1038/s41467-025-59253-9

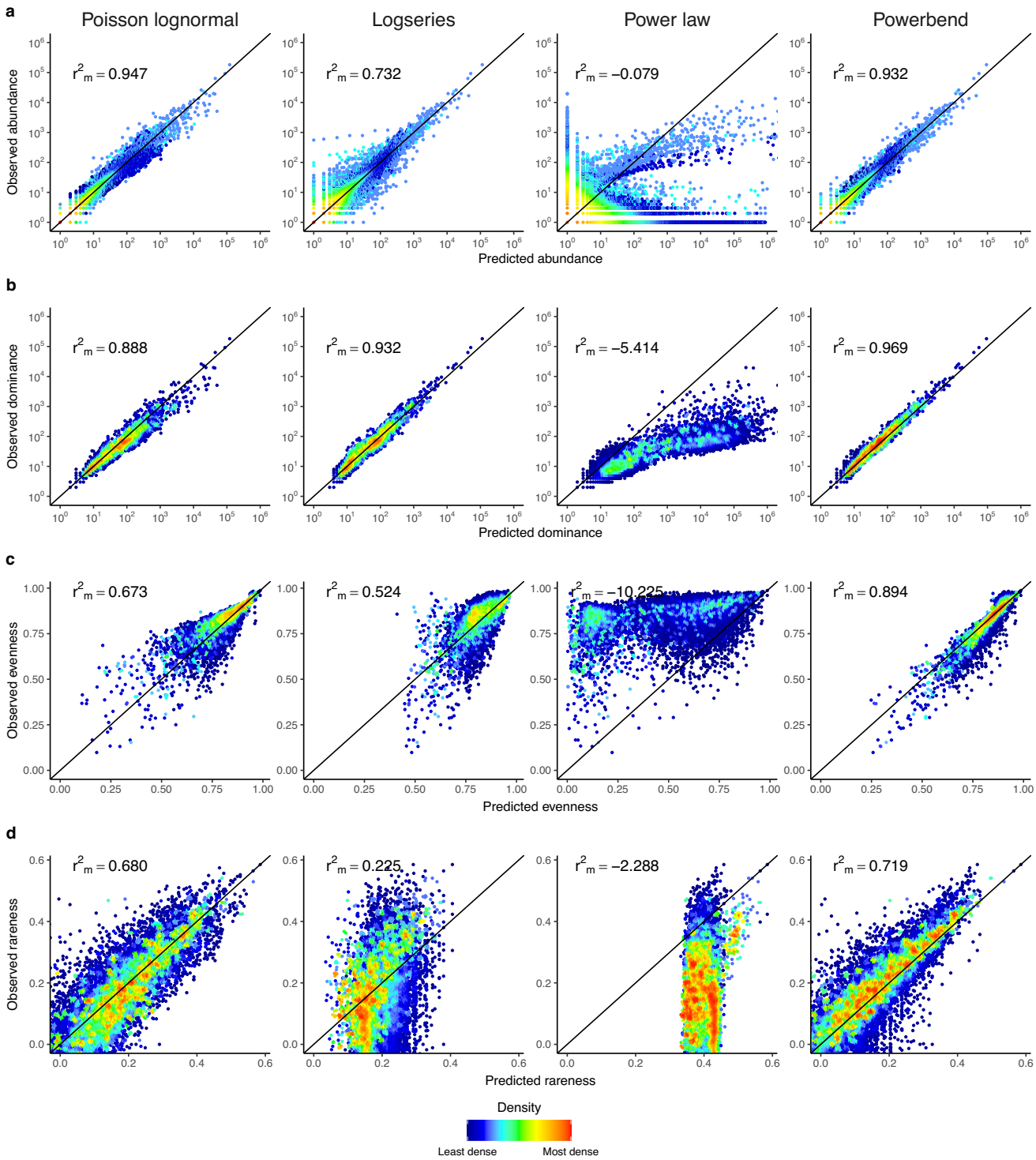

Fig. 1 | Goodness of fit of SAD models in 13,819 animal and plant communities. The predicted values are plotted against observed values for (**a**) species abundance. **b** SAD dominance. **c** SAD evenness, and **d** SAD rareness. Each dot in (**a**) represents a species, and each dot in (**b**–**d**) represents a community. The color represents the density of the dots: red represents the densest dots, and dark blue represents the least dense ones. The black diagonal line is the 1:1 line that represents a perfect fit. Goodness of fit for each SAD model is determined by the modified coefficient of determination against the 1:1 line (r^2_m) and its mean value, weighted by the sample sizes of the datasets is shown. Because r^2_m is calculated against the 1:1 line with a fixed intercept of 0, it is possible that its value drops below zero, which indicates a poor fit. Source data is provided as a Source Data file.

Incorporating a Poisson sampling error substantially improves the fit of power law and powerbend (Table 1). The powerbend model outperforms all other models in 67.0% of communities tested, while the next best model, Poisson lognormal, is superior in 18.3% of the communities (Table 1). In stark contrast, logseries, though a decent model for animal and plant SADs, is the best model in only 0.2% of microbial communities. In a direct comparison of the two best models, the Poisson powerbend model outperforms the Poisson lognormal model in 76.1% of communities, 87.0% of which are statistically significant (ΔAIC ≥ 2). This finding remained consistent regardless of the

Table 1 | Frequencies of each SAD model and model family being selected as the best model by AIC in 15,329 microbial communities

SAD model	Sampling error structure	Best model (statistically significant)	Best model family (statistically significant)
Lognormal	Poisson	18.3% (12.7%)	18.3% (12.7%)
Logseries	None	0.2% (0.0%)	0.2% (0.0%)
	Poisson	0.0% (0.0%)	
Power Law	None	1.7% (0.0%)	14.5% (0.5%)
	Poisson	12.8% (0.5%)	
Powerbend	None	6.7% (2.0%)	67.0% (56.1%)
	Poisson	60.3% (49.4%)	

A best model is considered statistically significant when its AIC difference to the second-best model is greater than 2.

sequence similarity thresholds used to define microbial species (Supplementary Table S4), whether 16S rRNA copy number variation was taken into account in determining the species abundance (Supplementary Table S5), and when using Bayesian Information Criterion (BIC) that applies greater penalty to more complex models (Supplementary Table S6).

Poisson powerbend exhibits an excellent overall fit to the observed SAD data, on average explaining 99.3% of the variation (Fig. 2a). The excellent fit is consistent throughout the entire range of species abundance spanning 7 orders of magnitude. In contrast, although Poisson lognormal and Poisson power law also provide great overall fits, they substantially overestimate the abundance of the abundant species in the communities (Fig. 2a). Using Monte-Carlo simulations, we found that Poisson powerbend exhibits r^2_m values not significantly different from 1.0 in 90.1% of SADs. This suggests that powerbend accurately describes the vast majority of microbial SADs. In contrast, Poisson lognormal and Poisson power law models have r^2_m values not significantly different from 1.0 only in 67.3% and 43.0% of SADs, respectively. Poisson logseries fits the data poorly, only explaining 12.9% of variation (Fig. 2a).

We noted that Fig. 2 of the study by Shoemaker et al. did not seem to show the overprediction of species abundance by the Poisson lognormal model[25], as we have demonstrated here. This discrepancy arises because the overprediction is most apparent in SADs with 16S rRNA reads exceeding 10^5, and these SADs were excluded from Shoemaker et al.'s Fig. 2[25] (Supplementary Note 1, Supplementary Fig. S2).

We were interested in the performance of the gambin model—a simple, one-parameter model—relative to the other more complex SAD models. Because the Gambin model requires species abundance data to be binned into $\log_2$ octaves, resulting in a substantial loss of information, we did not test it in animal and plant communities, which have limited data points. In microbial communities, however, the Gambin model was significantly inferior to the Poisson powerbend and lognormal models (Supplementary Table S7).

Powerbend accurately captures SAD skewness, evenness, and dominance while other models fail

SAD is often considered a weak test due to its limited ability to distinguish between ecological models using benchmarks such as AIC and R^2 that measure the overall fit to the species abundance data[1]. One issue with the overall fit measurements is that they are heavily weighted by rare species that make up most of the data points. To overcome this problem, additional SAD metrics can be used to evaluate SAD models[25,29]. They encompass various features of the SAD including rareness (the asymmetry in the distribution of species abundance as a measure of rarity), evenness (the level of uniformity in species abundance), and dominance (N_{max}, the abundance of the most abundant species). A good SAD model should not only capture the overall species abundance distribution in the raw data but also accurately reflect key scalar metrics such as evenness, rareness, and richness. Figures 1b–d and 2b-d show that the powerbend model performs very well in capturing these additional SAD metrics of animal, plant, and microbial communities, while the other models all perform poorly. For example, although Poisson lognormal provides a good overall fit to the species abundance data in microbial communities (Fig. 2a), it fits poorly to the SAD dominance ($r^2_m = -1.122$, Fig. 2b), evenness ($r^2_m = -1.519$, Fig. 2c) and rareness ($r^2_m = 0.600$, Fig. 2d). This result demonstrates again that the powerbend model is overwhelmingly superior to the other SAD models.

Discussion

The distribution of species abundance stands as one of the oldest, most universal, and fundamental laws in ecology. Despite decades of research and the development of numerous models, a universally accepted SAD model applicable across the tree of life remains elusive. Consistent with previous findings[1,18,28], we found it challenging to identify a single best model for animal and plant communities based on the AIC and r^2_m criteria. The Poisson lognormal and powerbend models are essentially tied, although the powerbend model is superior when evaluated with additional SAD metrics such as dominance, evenness, and rareness. This difficulty in distinguishing among SAD models is partly due to the relatively low number of species within animal and plant communities, which limits the statistical power for model comparison. This challenge can be mitigated by testing SAD models within microbial communities, which typically exhibit significantly greater species richness. In this study, for instance, microbial communities have a median value of 3,246 species per SAD, facilitating more robust model testing. We boosted the power of SAD testing by concurrently evaluating models across animal, plant, and microbial communities, covering a range of abundance scales spanning 7 orders of magnitude. Incorporating additional SAD metrics, including evenness, dominance, and rareness, enabled us to identify a superior model that would otherwise be indistinguishable from others using only AIC and r^2_m. Moreover, we explicitly modeled the sampling effort of surveys, an important but often overlooked factor in SAD model testing[1]. As a result, we demonstrated that, among the models tested in this study, powerbend is the only one that provides a good fit for both microorganism and macroorganism communities, establishing it as a promising unifying SAD model. Our simulations show that when sampling or species number is sufficient, SAD data have enough power to distinguish SAD models and recover the true model (Supplementary Fig. S1). Conversely, poor sampling or the small number of species tends to favor simpler models such as the logseries and Poisson lognormal over the powerbend (Supplementary Fig. S1). Therefore, we concluded that the superiority of the powerbend model is unlikely to be a result of poor sampling. While the success of the powerbend model does not by itself prove the existence of universal ecological principles, it provides key evidence supporting this hypothesis.

It is important to point out that powerbend is in essence a "bent" power law[16]. While the abundance distribution of most species within a community follows a power law, the bending in the powerbend model imposes an upper limit on the abundance of dominant species. This bending reflects the inherent constraint on the population size of a community[12,13] and is a key feature that distinguishes powerbend as a superior model to the power law. In contrast, the power law fits SADs poorly because it consistently overestimates the abundance of dominant species within a community (Figs. 1a, b, 2a and b).

The existence of a unifying SAD model suggests some common mechanisms at work. Although a-mechanistic models such as the powerbend model do not explicitly incorporate the underlying mechanisms or processes, they can still be useful for discerning and

https://doi.org/10.1038/s41467-025-59253-9

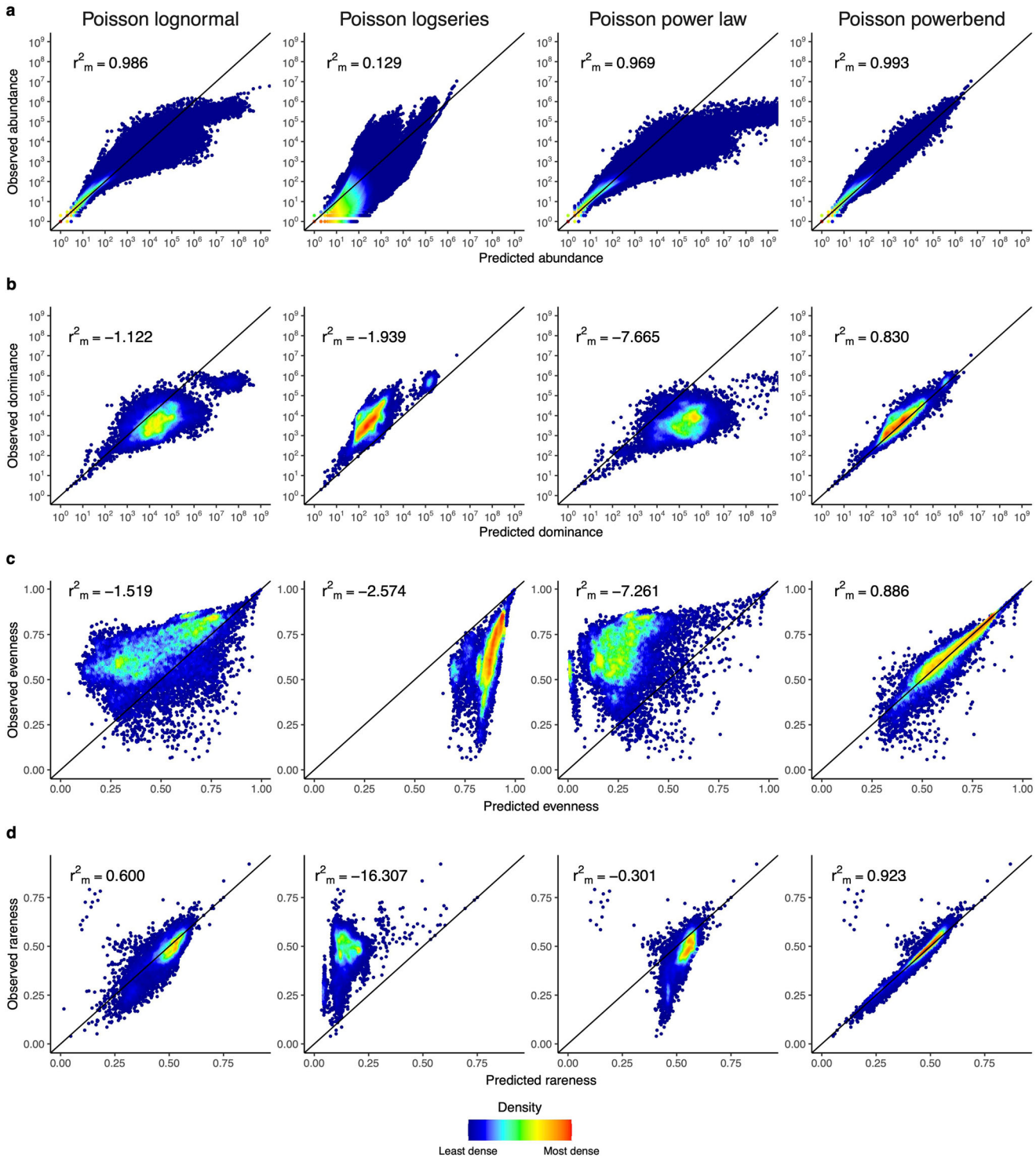

Fig. 2 | Goodness of fit of SAD models in 15,329 microbial communities. The predicted values are plotted against observed values for (**a**) species abundance. **b** SAD dominance. **c** SAD evenness, and **d** SAD rareness. Each dot in **a** represents a species, and each dot in (**b**–**d**) represents a community. The color represents the density of the dots: red represents the densest dots and dark blue represents the least dense ones. The black diagonal line is the 1:1 line that represents a perfect fit. Goodness of fit for each SAD model is determined by the modified coefficient of determination against the 1:1 line (r^2_m) and its mean value, weighted by the sample sizes of the datasets is shown. Because r^2_m is calculated against the 1:1 line with a fixed intercept of 0, it is possible that its value drops below zero, which indicates a poor fit. For illustrative purposes, the x-axis is truncated to 10^9 in (**a**) and (**b**) because the power law overpredicts the abundance of the most abundant species. Source data are provided as a Source Data file.

revealing dominant mechanistic drivers of the ecosystem[12,13,30]. For example, because the powerbend model was derived by modeling interspecific trait variation[14,15], it supports the idea that deterministic mechanisms are important in driving the community assembly. Therefore, our result challenges purely neutral models such as Hubbell's neutral theory, which predicts logseries SAD. On the other hand, the fact that the powerbend model is derived using

a MaxEnt framework implies that stochasticity also plays an important role in shaping macroecological patterns.

The *s* parameter in the powerbend model can be conceptualized as representing the number of limiting resources driving the community assembly[30]. Alternatively, it could be seen as analogous to the α parameter in the Gambin model, representing the 'dimensionality' of the community's realized niche space. Our study indicates that in animal and plant communities, the weighted mean of the *s* parameter is 0.9, close to 1.0 when the powerbend model degenerates into the logseries model. In this scenario, total available energy can be considered the limiting resource[30] or as one dimension of niche space driving the macroecological patterns. In contrast, the mean *s* parameter of microbial communities is 1.6. This suggests that microbe communities are constrained by a greater number of limiting resources or occupy a higher-dimensional 'realized' niche space, likely because microbial surveys typically span a much broader taxonomic range than animal and plant surveys. A greater number of limiting resources or an expanded niche dimensionality provides more specialized opportunities for rare species to survive, resulting in more uneven SADs or a more pronounced 'rare biosphere' in microbial communities. The fractional nature of the *s* parameter also implies that the limiting resource or niche space is hierarchically partitioned among community members. This aligns with findings showing that cross-feeding between species in a microbial community is a key driver of community assembly[31].

The prediction of the powerbend SAD model is rooted in a maximum entropy-based framework. The success of the powerbend SAD model provides evidence that the assembly of both microorganism and macroorganism communities adheres to the same governing principles. This study should stimulate additional testing of maximum entropy-based theories as potential unifying frameworks for understanding other macroecological patterns, including the species-area relationship and the metabolic rate-abundance relationship[12,30,32].

Methods

SAD data

To test the universality of SAD models in animals and plants, the species abundance data were downloaded from Baldridge et al. study[18], which includes the Breeding Bird Survey (BBS)[33], Alwyn H. Gentry's Forest Transect Data Set (GENTRY)[34], the Mammal Community Database (MCDB)[35], the Forest Inventory and Analysis (FIA)[36] and SAD data compiled from literature for an assortment of taxa[37]. To test microbial SADs, the bacterial and archaeal species abundance data were downloaded from Shoemaker et al. study[25], which includes data from the Earth Microbiome Project (EMP)[38], the first phase of Human Microbiome Project (HMP1)[39], and the MG-RAST repository (MGRAST)[40]. Because the rare tail of the SAD contains critical information for model fitting, we included species with a single read, known as singletons, in our analysis to avoid potential bias in model selection, consistent with the practice of Shoemaker et al.[25] Our simulations showed that when the number of observed species in a SAD is small or when the sampling effort is low, model selection using AIC does not have sufficient power to distinguish SAD models, often favoring the simpler models even when the simpler model is incorrect (Supplementary Fig. S1). The more complex the model is, the more data points (i.e., species) are required to recover the true model. To strike a balance between the number of SADs we can analyze and the power of model selection, and following the practice of the previous studies[17,18], we filtered out SADs with less than 10 species to test base SAD models for animal and plant SAD data, and SADs with less than 100 Operational Taxonomic Units (OTUs) to test compound SAD models for microbial SAD data. In addition, in the microbial SAD data, we identified 687 outlier SADs where the number of doubleton species exceeded the number of singleton species by more than 10-fold (Supplementary Fig. S3). These outliers all originated from the EMP dataset compiled by Shoemaker et al.[25] but are no longer present in the current EMP dataset (year 2017 version). Consequently, we excluded these outliers from our analyses. This resulted in a total of 13,819 animal and plant SADs and 15,329 microbial SADs (see Supplementary Table S2).

The animal and plant dataset encompasses diverse taxonomic groups (see Supplementary Table S2). To ensure equitable representation across these groups, we applied weighting to each SAD based on the size of the dataset it originated from when assessing the frequency of a model being the best by AIC or its goodness of fit. For instance, a SAD within the Mammal Community Database (MCDB, 103 samples) carried 26.9 times the weight of a SAD from the Breeding Bird Survey (BBS, 2,769 samples).

SAD models

The lognormal SAD model has the probability density function:

$$\Phi_{lognorm}(n;\mu,\sigma)=\frac{1}{n\sigma\sqrt{2\pi}}e^{-\frac{(\log(n)-\mu)^2}{2\sigma^2}}\ (n\in\mathbb{R}^+) \tag{1}$$

where n represents the species abundance, and μ and σ are the mean and standard deviation of log-transformed n, respectively. The logarithm is calculated with the natural base e. Given that the lognormal distribution is continuous while the SAD is inherently discrete, it is necessary to convolute it with a sampling error when fitting it to SAD data.

The logseries distribution has the probability mass function:

$$\Phi_{ls}(n;\lambda)=-\frac{1}{\log(1-e^{-\lambda})}\cdot\frac{e^{-\lambda n}}{n}\ (n\in\mathbb{Z}^+) \tag{2}$$

where λ is the exponential rate at which the numerator decays.

The power law distribution has the probability mass function:

$$\Phi_{power}(n;s)=\frac{n^{-s}}{\zeta(s)}\ (n\in\mathbb{Z}^+) \tag{3}$$

where $s>1$ is the scaling parameter that controls the distribution's decay rate, and $\zeta(s)$ is the Riemann zeta function, serving as a normalization factor to ensure the sum of the probabilities over all possible values of n equals 1.

The powerbend distribution takes a more general form compared to the logseries and power law distributions. It is a hybrid of a power law and an exponential function, in which the exponential function bends the power law by setting an upper bound to the power law[16]. It has the probability mass function:

$$\Phi_{powbend}(n;s,\lambda)=\frac{1}{Z}\cdot\frac{e^{-\lambda n}}{n^s}\ (n\in\mathbb{Z}^+) \tag{4}$$

where s is the order of the denominator, λ is the exponential rate at which the numerator decays, and Z is the probability normalizer. It should be noted that the powerbend distribution is a generalized distribution that can degenerate into the logseries, geometric, broken-stick, and power law distributions by setting its parameters to certain fixed values (Supplementary Note 1).

Modeling sampling effort in 16S rRNA survey

To relate the number of observed reads in 16S rRNA profiling data to the number of individual cells in the community, we explicitly modeled the sampling effort by convoluting a sampling error to the base SAD models. The 16S rRNA profile of a community typically comprises thousands of reads. Assuming no bias, the sampling error can be approximated by a Poisson distribution whose mean represents the

expected number of reads given the number of individual cells. We used parameter η to denote the expected number of reads per individual cell, which represents the sampling effort. By convoluting the sampling error to the base SAD models, we derived the sampling-explicit SAD models defined on non-negative integers $k \in 0 \cup \mathbb{Z}^+$:

$$\Phi_{poils}(k;\lambda,\eta) = \sum_{n=1}^{\infty} \frac{(\eta n)^k e^{-\eta n}}{k!} \cdot \Phi_{ls}(n;\lambda) \quad (5)$$

$$\Phi_{poipower}(k;s,\eta) = \sum_{n=1}^{\infty} \frac{(\eta n)^k e^{-\eta n}}{k!} \cdot \Phi_{power}(n;s) \quad (6)$$

$$\Phi_{poipowbend}(k;s,\lambda,\eta) = \sum_{n=1}^{\infty} \frac{(\eta n)^k e^{-\eta n}}{k!} \cdot \Phi_{powbend}(n;s,\lambda) \quad (7)$$

It should be noted that for Poisson lognormal distribution, because the number of individuals n and the sampling effort η always appear together in the formula as their product, there is no way to disentangle the two parameters. Therefore, the sampling effort η is fixed to 1, and the probability mass function of the Poisson lognormal distribution has the same number of parameters as the base distribution:

$$\Phi_{poilog}(k;\mu,\sigma) = \int_0^{\infty} \frac{n^k e^{-n}}{k!} \cdot \Phi_{lognorm}(n;\mu,\sigma) \cdot dn \quad (8)$$

In practice, there may be bias in 16S rRNA sequencing resulting from biases in DNA extraction and PCR amplification. Thus, a Poisson sampling error may not be sufficient to accurately relate the sequence reads to the number of individual cells. Because the direction and magnitude of such bias are often unknown, its effect on the distribution of sequence reads can be seen as an inflation of the variance or over-dispersion without changing the mean. As a result, we used the negative binomial distribution to model such bias[41], and derived the corresponding probability mass function for SAD models:

$$\Phi_{nblog}(k;\mu,\sigma,r) = \int_0^{\infty} \frac{\Gamma(r+k) \cdot r^k \cdot n^r}{\Gamma(r) \cdot k! \cdot (r+n)^{r+k}} \cdot \Phi_{lognorm}(n;\mu,\sigma) \cdot dn \quad (9)$$

$$\Phi_{nbls}(k;\lambda,\eta,r) = \sum_{n=1}^{\infty} \frac{\Gamma(r+k) \cdot r^k \cdot (\eta n)^r}{\Gamma(r) \cdot k! \cdot (r+\eta n)^{r+k}} \cdot \Phi_{ls}(n;\lambda) \quad (10)$$

$$\Phi_{nbpowber}(k;s,\eta,r) = \sum_{n=1}^{\infty} \frac{\Gamma(r+k) \cdot r^k \cdot (\eta n)^r}{\Gamma(r) \cdot k! \cdot (r+\eta n)^{r+k}} \cdot \Phi_{powber}(n;s) \quad (11)$$

$$\Phi_{nbpowbend}(k;s,\lambda,\eta,r) = \sum_{n=1}^{\infty} \frac{\Gamma(r+k) \cdot r^k \cdot (\eta n)^r}{\Gamma(r) \cdot k! \cdot (r+\eta n)^{r+k}} \cdot \Phi_{powbend}(n;s,\lambda) \quad (12)$$

In the above equations, r measures the degree of over-dispersion or bias in the negative binomial distribution and Γ is the gamma function.

Using 200 randomly selected microbial SADs, we compared the use of Poisson and negative binomial distributions to model the sampling effort in the 16S rRNA sequencing pipeline. Our result indicated that SAD models with the Poisson error structure are superior to those with the negative binomial error structure (Supplementary Table S8). Therefore, we used the Poisson error structure for testing the full microbial SAD dataset.

Supplementary Table S1 lists the models tested in this study and their free parameters.

Fitting SAD models to empirical data

Because all SAD models (both base models and their sampling-explicit counterparts) are formulated as probability distributions, we fitted them to the observed abundances in empirical data using the maximum likelihood (ML) framework. Because species with zero observations are not recorded in empirical data, the likelihood of a single species with k observations (denoted as $L(k)$) for sampling-explicit SAD models is normalized by the cumulative probability of any positive observation:

$$L(k) = \frac{P(k)}{\sum_{m=1}^{\infty} P(m)} \quad (13)$$

In the above formula, $P(k)$ is the probability mass function for one of the sampling-explicit SAD models described in the previous section. We compared the performance of SAD models using AIC.

We fitted all models with a Poisson or negative binomial error structure using the R package 'microSAD'. Additionally, we fitted the logseries, power law, powerbend, and Weibull models using the R package 'sads'. For the gambin model, we used the R package 'gambin'[23].

Evaluating the goodness of fit of SAD models

The goodness of fit of SAD models was evaluated by comparing predicted and observed species abundances in the rank-abundance distribution (RAD). Briefly, to generate the expected RAD for a SAD with S observed species, we placed S quantiles on the cumulative distribution function (CDF) of the fitted SAD model, evenly dividing the cumulative probability (the y-axis of the CDF curve). For comparison between the expected and the observed RADs, we used the modified coefficient of determination around the 1:1 line r_m^2 as described in previous studies[17,25,27] to quantify the goodness of fit:

$$r_m^2 = 1 - \frac{\sum_{i=1}^{S} (log(x_i) - log(y_i))^2}{\sum_{i=1}^{S} \left(log(x_i) - \frac{1}{S}\sum_{j=1}^{S} log\left(x_j\right)\right)^2} \quad (14)$$

In the above formula, x and y are the observed and the predicted abundances (e.g., number of reads in 16S rRNA profiling data), respectively. The subscripts denote the rank of species, and S is the total number of observed species in the SAD. Because the coefficient of determination here is calculated against the 1:1 line with a fixed intercept of 0, it is possible that its value drops below zero, which indicates a poor fit.

The value of r_m^2 reaches 1 if and only if the observed and the predicted abundances match perfectly, indicating a perfect fit of SAD. To determine if the r_m^2 of a model fitted to a SAD (observed r_m^2) is significantly lower than 1, we established a baseline distribution of r_m^2 values from 1000 iterations of Monte-Carlo simulation. In each iteration, we simulated a RAD based on the fitted SAD model and calculated r_m^2 between the simulated RAD and the expected RAD. We then employed a one-sided test to assess the lack-of-fit by calculating the empirical frequency at which the baseline r_m^2 values were smaller than the observed r_m^2.

In addition, the goodness of fit of SAD models was evaluated by comparing observed and predicted SAD evenness, rareness, and dominance. Evenness of a SAD was measured using Shannon's evenness E_H, also known as Pielou's evenness J[42]. Rareness was measured by the log-modulo of skewness of a SAD, as described in refs. 25,29, with log-transformed species abundances utilized for the skewness calculation. Dominance is measured simply as the abundance of the most abundant species (N_{max}) in the SAD, also as described in[25,29].

Evaluating the effects of observed species number and sampling effort on SAD model selection

We investigated the influence of the number of observed species on the efficacy of model selection by AIC. We generated simulated communities employing either a logseries ($\lambda = 10^{-2}$) or powerbend ($s = 1.5$, $\lambda = 10^{-2}$) SAD model. The simulated communities varied in the observed species number, ranging from 10 to 100 species per community. For each level of species count, we simulated 100 communities under each SAD model. Subsequently, we fitted the logseries and powerbend models to the simulated data and recorded the frequency at which each model emerged as the best-fit model by AIC.

Likewise, we examined the effect of sampling effort on SAD model selection. We conducted simulations using communities generated according to either a powerbend ($s = 1.5$ and $\lambda = 10^{-5}$) or lognormal ($\mu = 0$ and $\sigma = 3.25$) SAD model. For each SAD model, we simulated 100 communities, each consisting of 10^4 species, by randomly drawing species abundances from the respective SAD distribution. Next, we simulated the sampling process by randomly drawing the number of observations for each species from a Poisson distribution, whose mean was set to be the product of the species abundance and the sampling effort. We simulated 9 levels of sampling effort, evenly spaced from 10^{-1} to 10^{-5} on a log-scale. Subsequently, we fitted the Poisson powerbend and Poisson lognormal models to the simulated data and recorded the frequency at which each model appeared as the best model at each level of sampling effort.

Evaluating the effect of OTU sequence identity threshold on SAD model selection

To investigate the effect of OTU sequence identity threshold on SAD model selection, we clustered OTU at different identity thresholds for the 565 SADs from the HMP1 dataset using Mothur (version 1.48)[43]. Specifically, we extracted all aligned 16S rRNA gene sequences in the HMP1 dataset from the summary table of the Mothur pipeline. We calculated the pairwise distance between unique sequences using the function "dist.seqs" (with arguments cutoff=0.20 and output=lt) and clustered the unique sequences with the average neighbor algorithm using the function "cluster" (with arguments: method=average and cutoff=0.20). OTU tables were generated at the sequence identity threshold of 95% and 99% using the function "make.shared" (with argument: label=0.01-0.05). We fitted SAD models to these two datasets, conducted model selection through AIC, and compared the outcome with that of the original HMP1 dataset.

Evaluating the effect of 16S rRNA gene copy number (GCN) correction on SAD model selection

Because 16S rRNA GCN variation can bias the species composition estimated using 16S rRNA read counts[44], we assessed its impact on the outcome of SAD model selection by performing model selection on the 565 SADs from the HMP1 dataset. We predicted GCN for each OTU and then fitted models on the species abundance data corrected for GCN variation. Specifically, we selected the most abundant sequence in each OTU as its representative sequence. We then predicted the 16S GCN of each OTU using RasperGade16S[45], an R package that predicts 16S GCN based on phylogenetic relatedness. In SAD model fitting, we modeled 16S GCN as an OTU-specific multiplier to the sampling effort η. Consequently, only SAD models with a Poisson sampling error structure were included in this analysis. Model selection was performed through AIC, and the outcomes were compared with those obtained without 16S GCN correction.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The SAD data that support the findings of this study are available at figshare https://doi.org/10.6084/m9.figshare.25711257. Source data are provided with this paper.

Code availability

The code and instructions (a Readme file) for replicating the analyses in this study are available at figshare https://doi.org/10.6084/m9.figshare.25711257. The R package '*microSAD*' can be downloaded from Github: https://github.com/wu-lab-uva/microSAD or https://doi.org/10.5281/zenodo.14845910.

References

1. McGill, B. J. et al. Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. *Ecol. Lett.* **10**, 995–1015 (2007).
2. Sogin, M. L. et al. Microbial diversity in the deep sea and the underexplored "rare biosphere. *Proc. Natl. Acad. Sci. USA* **103**, 12115–12120 (2006).
3. Fisher, R. A., Corbet, A. S. & Williams, C. B. The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Anim. Ecol.* **12**, 42 (1943).
4. Preston, F. W. The commonness, and rarity, of species. *Ecology* **29**, 254–283 (1948).
5. MacArthur, R. H. On the relative abundance of bird species. *Proc. Natl. Acad. Sci. USA* **43**, 293–295 (1957).
6. Motomura, I. A statistical treatment of ecological communities. *Zool. Mag.* **44**, 379–383 (1932).
7. Frontier, S. Diversity and structure in aquatic ecosystems. *Oceanogr. Mar. Biol.* **23**, 253–312 (1985).
8. Hubbell, S. P. *The Unified Neutral Theory of Biodiversity and Biogeography*. (Princeton University Press, Princeton, 2001).
9. Volkov, I., Banavar, J. R., Hubbell, S. P. & Maritan, A. Neutral theory and relative species abundance in ecology. *Nature* **424**, 1035–1037 (2003).
10. Alroy, J. The shape of terrestrial abundance distributions. *Sci. Adv.* **1**, e1500082 (2015).
11. Leigh, E. G. *Tropical Forestecology A View from Barro Colorado Island*. (Oxford University Press, Oxford, 1999).
12. Harte, J., Zillio, T., Conlisk, E. & Smith, A. B. Maximum entropy and the state-variable approach to macroecology. *Ecology* **89**, 2700–2711 (2008).
13. Harte, J. *Maximum Entropy and Ecology A Theory of Abundance, Distribution, and Energetics*. (Oxford University Press, New York, 2011).
14. Pueyo, S., He, F. & Zillio, T. The maximum entropy formalism and the idiosyncratic theory of biodiversity. *Ecol. Lett.* **10**, 1017–1028 (2007).
15. Bertram, J., Newman, E. A. & Dewar, R. C. Comparison of two maximum entropy models highlights the metabolic structure of metacommunities as a key determinant of local community assembly. *Ecol. Model* **407**, 108720 (2019).
16. Pueyo, S. Diversity: between neutrality and structure. *Oikos* **112**, 392–405 (2006).
17. White, E. P., Thibault, K. M. & Xiao, X. Characterizing species abundance distributions across taxa and ecosystems using a simple maximum entropy model. *Ecology* **93**, 1772–1778 (2012).
18. Baldridge, E., Harris, D. J., Xiao, X. & White, E. P. An extensive comparison of species-abundance distribution models. *Peerj* **4**, e2823 (2016).
19. Magurran, A. E. & Henderson, P. A. Explaining the excess of rare species in natural species abundance distributions. *Nature* **422**, 714–716 (2003).
20. Ulrich, W., Ollik, M. & Ugland, K. I. A meta-analysis of species–abundance distributions. *Oikos* **119**, 1149–1155 (2010).

21. Ulrich, W., Kusumoto, B., Shiono, T. & Kubota, Y. Climatic and geographic correlates of global forest tree species–abundance distributions and community evenness. *J. Veg. Sci.* **27**, 295–305 (2016).
22. Ugland, K. I. et al. Modelling dimensionality in species abundance distributions: description and evaluation of the Gambin model. *Evolutionary Ecology Research* 313–324 (2007).
23. Matthews, T. J. et al. The Gambin model provides a superior fit to species abundance distributions with a single free parameter: evidence, implementation and interpretation. *Ecography* **37**, 1002–1011 (2014).
24. Ulrich, W., Nakadai, R., Matthews, T. J. & Kubota, Y. The two-parameter Weibull distribution as a universal tool to model the variation in species relative abundances. *Ecol. Complex.* **36**, 110–116 (2018).
25. Shoemaker, W. R., Locey, K. J. & Lennon, J. T. A macroecological theory of microbial biodiversity. *Nat. Ecol. Evol.* **1**, 0107 (2017).
26. Enquist, B. J. et al. The commonness of rarity: Global and future distribution of rarity across land plants. *Sci. Adv.* **5**, eaaz0414 (2019).
27. Xiao, X., McGlinn, D. J. & White, E. P. A strong test of the maximum entropy theory of ecology. *Am. Nat.* **185**, E70–E80 (2015).
28. Matthews, T. J. & Whittaker, R. J. Fitting and comparing competing models of the species abundance distribution: assessment and prospect. *Front. Biogeogr.* **6**, (2014).
29. Locey, K. J. & Lennon, J. T. Scaling laws predict global microbial diversity. *Proc. Natl Acad. Sci.* **113**, 5970–5975 (2016).
30. Harte, J. & Newman, E. A. Maximum information entropy: a foundation for ecological theory. *Trends Ecol. Evol.* **29**, 384–389 (2014).
31. Goldford, J. E. et al. Emergent simplicity in microbial community assembly. *Science* **361**, 469–474 (2018).
32. Shade, A. et al. Macroecology to unite all Life, large and small. *Trends Ecol. Evol.* **33**, 731–744 (2018).
33. Ziolkowski, D., Lutmerding, M., Aponte, V. I. & Hudson, M.-A. R. *North American Breeding Bird Survey Dataset (1966-2021)*. https://doi.org/10.5066/P97WAZE5.
34. Phillips, O. & Miller, J. S. *Global Patterns of Plant Diversity: Alwyn H. Gentry Forest Transect Data Set*. (Missouri Botanical Garden Press, 2002).
35. Thibault, K. M., Supp, S. R., Giffin, M., White, E. P. & Ernest, S. K. M. Species composition and abundance of mammalian communities. *Ecology* **92**, 2316–2316 (2011).
36. *Forest inventory and analysis national core field guide (Phase 2 and 3)*. https://www.fs.usda.gov/research/programs/fia.
37. Baldridge, E. *Community abundance data*. https://figshare.com/articles/dataset/Community_abundance_data/769251/1.
38. Gilbert, J. A., Jansson, J. K. & Knight, R. The Earth Microbiome project: successes and aspirations. *Bmc Biol.* **12**, 69 (2014).
39. Turnbaugh, P. J. et al. The human microbiome project. *Nature* **449**, 804–810 (2007).
40. Meyer, F. et al. The metagenomics RAST server–a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinforma.* **9**, 386 (2008).
41. Bliss, C. I. & Fisher, R. A. Fitting the negative binomial distribution to biological data. *Biometrics* **9**, 176 (1953).
42. Pielou, E. C. *Ecological Diversity*. (John Wiley & Sons, New York, 1975).
43. Schloss, P. D. et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* **75**, 7537–7541 (2009).
44. Kembel, S. W., Wu, M., Eisen, J. A. & Green, J. L. Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. *PLoS Comput. Biol.* **8**, e1002743 (2012).
45. Gao, Y. & Wu, M. Accounting for 16S rRNA copy number prediction uncertainty and its implications in bacterial diversity analyses. *ISME Commun.* **3**, 59 (2023).

Acknowledgements

We thank Research Computing at The University of Virginia for providing computational resources and technical support that have contributed to the results reported within this publication.

Author contributions

Y.G. and A.A. developed the R package 'microSAD'. Y.G., A.A. and M.W. conducted data analyses. Y.G. and M.W. wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-59253-9.

Correspondence and requests for materials should be addressed to Martin Wu.

Peer review information *Nature Communications* thanks Micah Brush, Joaquin Hortal, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.