BIOTRAM: BIOCHEMISTRY-INFORMED TRANSFORMER FOR RELATIVE ABUNDANCE VIA METABARCODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Arthropod monitoring at scale begins with metabarcoding of bulk samples: trap collections of hundreds to thousands of specimens processed as one unit, extraction of DNA from the bulked specimens, then sequencing of the DNA from all specimens together. Metabarcoding reveals which taxa such a sample contains, but due to factors such as differing DNA amplification rates and body sizes, its read counts do not neatly correspond to how abundant the different taxa are relative to one another. This lack of relative abundance limits the data's utility for the multitude of ecological analyses that depend on it. We propose BIOTRAM: a neural network that estimates per-taxon specimen counts by dividing the observed read counts by a learned read yield per specimen, which it factors into the DNA each specimen contributes and how efficiently that DNA is amplified. On a paired dataset of 1,188 bottles (800k specimens) processed both by metabarcoding and by comprehensive individual barcoding for ground-truthing, BIOTRAM reduces the Jensen–Shannon divergence between the estimated and the true composition from 0.3586 for raw read proportions (the usual proxy for abundance) to 0.1200. Trained on the same objective and inputs as four generic deep architectures, all with tuned loss weights, it reaches a count error 3.2% below that of the strongest of them and the lowest composition divergence. It also generalizes to out-of-domain data better, beating all baselines on bottles from unseen regions, later collection dates and unseen sequencing batches, and transferring to data from an unseen laboratory. Structuring the mapping from reads to counts around the steps of the experimental pipeline thus yields more accurate abundance estimates than leaving that mapping unconstrained.

1 INTRODUCTION

Insects are by far the most diverse multicellular taxon on Earth (Colwell et al., 2026) and are fundamentally important for human existence given their foundational role in ecosystem processes (Yang & Gratton, 2014). Yet, they are undergoing extensive global declines (Wagner et al., 2021; Hallmann et al., 2017b); efforts to quantify this are limited by the paucity of reliable data (Didham et al., 2020). Biodiversity indices used to track populations rely on estimates of both species richness (number of species) and abundance (number of specimens per species) (Hallmann et al., 2017a; van Klink et al., 2020), while methods that provide only species lists but not abundances are of limited use. Malaise traps—tent-like structures that passively capture flying insects and direct them into a bottle of preservative—are widely used for capturing insects for study (Karlsson et al., 2020). Although identification of their contents by morphology was common in the past, many species can't be differentiated by morphology alone, and identifying species by manual or visual inspection is highly time consuming, expensive, and requires significant taxonomic expertise. Consequently DNA-based identification systems are increasingly used to identify captured species. Identifying specimens individually using DNA barcoding provides high-quality estimates of both richness and relative abundance, but is still relatively tedious and costly. Metabarcoding, a method wherein each bulk sample is processed as a unit by extracting and sequencing DNA from the entire collection, scales far better than DNA barcoding but does not produce individual specimen labels—only the number of amplicon reads—and so cannot directly estimate abundance (Yu et al., 2012; Ji et al., 2013) (see Figure 1).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

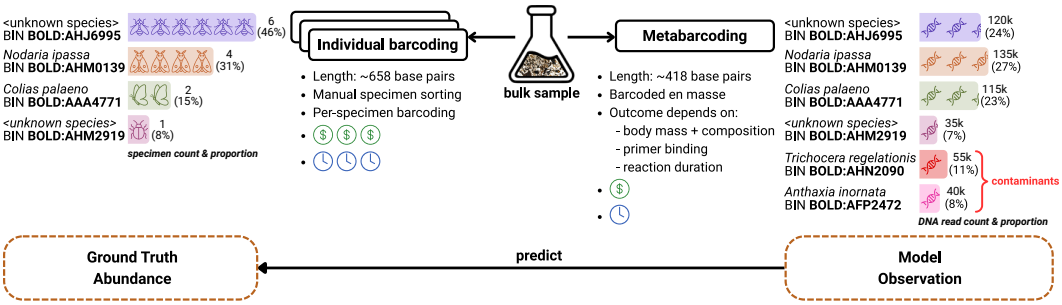


Figure 1: **Individual barcoding vs. metabarcoding for estimating species abundance.** Individual barcoding (manual sorting and per-specimen sequencing) gives accurate species counts but is slow and costly; metabarcoding (bulk amplification of the sample's DNA) is fast and cheaper, but its read counts are not linearly related to specimen abundances.

Rather than abundance, the output of metabarcoding is a vector of sequence read counts, one count per taxon. Besides the number of individuals, four key factors influence how many reads each taxon receives: (1) how much DNA each individual contributes to the bulk DNA mixture, (2) how efficiently its barcode sequence is amplified during PCR, (3) how long the PCR reaction runs before reagents are exhausted, and (4) what fraction of the amplified product is sequenced. Factors 3 & 4 vary by protocol, and factors 1 & 2 vary across taxa by orders of magnitude, so read counts track specimen counts only loosely. In bottles whose specimens were also barcoded one by one, raw read counts, taken as relative abundances, are on average further from the true composition than a guess that every taxon is equally abundant (Sec. 4).

Machine learning is changing how metabarcode reads are interpreted: foundation models trained on barcodes support identification and clustering even for unseen taxa (Millan Arias et al., 2026; Gao & Taylor, 2024; Safari et al., 2025). Abundance has resisted the same treatment: identification models learn from reference libraries of millions of individually barcoded specimens (deWaard et al., 2019; Gharaee et al., 2024), but abundance requires bulk samples whose composition is known specimen by specimen—the tedious, costly work that metabarcoding bypasses. The reads-to-specimens mapping also differs by taxon, and per-taxon correction factors (McLaren et al., 2019) cannot be calibrated for taxa that are rare or absent from calibration samples.

BIOTRAM, a **bio**chemistry-informed **t**ransformer for **r**elative **a**bundance via **m**etabarcoding, is instead structured around the steps of the experimental pipeline: it estimates specimen counts by dividing each taxon's reads by a learned read yield per specimen, factored into per-specimen DNA contribution and amplification efficiency, so it can produce estimates for taxa absent from training. Its taxonomic unit is the Barcode Index Number (BIN), a cluster of near-identical barcodes proxying for species that usually agrees with species assigned by morphology (Ratnasingham & Hebert, 2013). Our contributions are as follows:

(1) **A paired DNA barcoding-metabarcoding dataset** (Sec. 3). 1,188 bottles processed with both metabarcoding and (for ground-truthing) individual barcoding, yielding 414,640 matched (reads, count) pairs across 132,184 unique BINs and 578 sites—more than an order of magnitude larger than existing data.

(2) **A bespoke architecture built on biological and chemical knowledge** (Sec. 4.3). Rather than a naïve model, we propose a staged model based on the biochemical properties of the metabarcoding process.

(3) **Latent predictions of primer binding efficiency and body size** (Sec. 4.1). Our model produces physically interpretable latent estimates of both properties, which directly shape the relationship between specimen counts and read counts per BIN and bottle.

(4) **Controlled evaluation under distribution shift** (Sec. 5). We compare BIOTRAM against deep baselines trained with the same objective and inputs, generalizing to novel collection locations (spatial blocking), novel time-points (temporal cut-off), and across metabarcode sequencing methods (data from other labs).

2 RELATED WORK

Metabarcoding studies have long sought relative abundances from bulk samples, mostly with regression or other statistical models (Ji et al., 2013; Sickel et al., 2023; Lamb et al., 2019). For Malaise traps, Furneaux et al. (2026) analyzed a paired barcode-metabarcode dataset of 45 bulk samples from Sweden and Finland with mixed models, and found strong correspondence between read counts and abundance; they also modeled which BINs are present as individuals, even when undetected by metabarcoding, and how BIN-specific abundances affect detection. Iwaszkiewicz-Eggebrecht et al. (2026) applied a hierarchical Bayesian model that incorporates spike-in data to 15 Swedish samples with comprehensive barcode and metabarcode data. Both studies found close correspondence between read counts and relative abundances from barcoding, but were fit to a few dozen samples from northern Europe, vs. 1,188 bottles from 578 global sites here. We refit the hurdle mixed model of Furneaux et al. (2026) to our data as a baseline (Sec. 5 and Appx. I).

Metabarcoding studies of other taxa also estimate relative abundance, including that of plants (Matesanz et al., 2019), freshwater macroinvertebrates (Elbrecht & Leese, 2015), and fishes (Littlefair et al., 2026). Metagenomic workflows such as SPIKEPIPE (Ji et al., 2020a) avoid PCR bias but cost far more than metabarcoding. To our knowledge, none of these methods uses a neural network trained on paired data at this scale.

3 DATASET

Training requires metabarcoding data paired with ground-truth specimen counts—these come from separating and barcoding every specimen in a bottle, which is an expensive process. See Figure 1 for an overview of the respective individual barcoding and metabarcoding data pipelines. We assembled such data for Malaise-trap bottles from the LIFEPLAN project (deWaard et al., 2024; Hardwick et al., 2024) and the Global Malaise Program (Geiger et al., 2016), as described below. Permissions to use each bottle were arranged with the data owners (Appx. A).

Barcoding. DNA was extracted from every bottle by mild lysis, a chemical extraction that draws DNA from the specimens while leaving them intact, and every specimen was then barcoded individually. Metabarcoding amplified a $\sim$418 bp stretch of the cytochrome c oxidase subunit I gene (COI; BF3/BR2 primers; Elbrecht & Leese, 2017; Elbrecht et al., 2019a), in three PCR replicates, which were sequenced on an Illumina NextSeq 2000 and processed with the Metabarcoding Analysis Pipeline (MAP; Prosser et al., 2026a). MAP clustered the reads into operational taxonomic units (OTUs), each matched to a BIN in a reference library (Appx. B). OTUs with no BIN match within 2.3% sequence identity were assigned to their nearest BIN, because new BINs can only be created within BOLD. Individual barcoding yields a per-BIN specimen count for every bottle, which BIOTRAM learns to predict from the metabarcoding data alone. Individual barcoding fails for under 10% of specimens (Orsholm et al., 2026), so some BINs that metabarcoding detects but barcoding misses are present, not false detections.

At deployment, a monitoring program has only the metabarcoding results, without individual sorting or prior calibration. We therefore drop the BINs detected by individual barcoding but absent from metabarcoding, and keep the true zeros of BINs seen only in metabarcoding, which the model is trained and evaluated to predict.

The paired dataset spans 1,188 bottles from 578 sites collected between 2014 and 2025. It contains 414,640 rows, one per BIN per bottle, across 132,184 BINs and 800,875 counted specimens. A typical bottle holds 288 BINs and 375 specimens (medians; maxima 1,534 and 16,900), and the dataset covers 1,283 families.

As our goal is to predict the distribution of species within a bottle, the bottle is the model's unit of processing, and we partition the data by bottle. Specimens are almost entirely arthropods (Insecta 73.7%, Collembola 21.1%, Arachnida 5.1%). Most BINs lack genus and species labels, so we use BINs instead of named species.

Reads vs. counts. We call $n_{i,b}$ the specimen count of BIN i in bottle b and $p_{i,b} = n_{i,b} / \sum_j n_{j,b}$ its relative abundance; the vector p_b is the bottle's composition. Read shares are the usual proxy for composition, but they are biased (McLaren et al., 2019; Furneaux et al., 2026; Iwaszkiewicz-Eggebrecht et al., 2026), and in our data they track it poorly: on a $\log(1+x)$ scale, read counts explain

23% of the variance in specimen counts, rank BINs within a bottle at a mean Spearman correlation of 0.49, and separate present BINs from false detections at an AUC of 0.78. The most-sequenced BIN of a bottle is also its most abundant in only 13.8% of bottles.

Excess specimens. In 603 of the 1,188 bottles, technicians counted at least one highly abundant morphospecies by eye instead of barcoding each specimen (Figure 3), so these bottles lack complete per-BIN ground truth. We evaluate only on the 585 bottles without this *excess* and use the others for training only (Appx. E).

4 METHODS

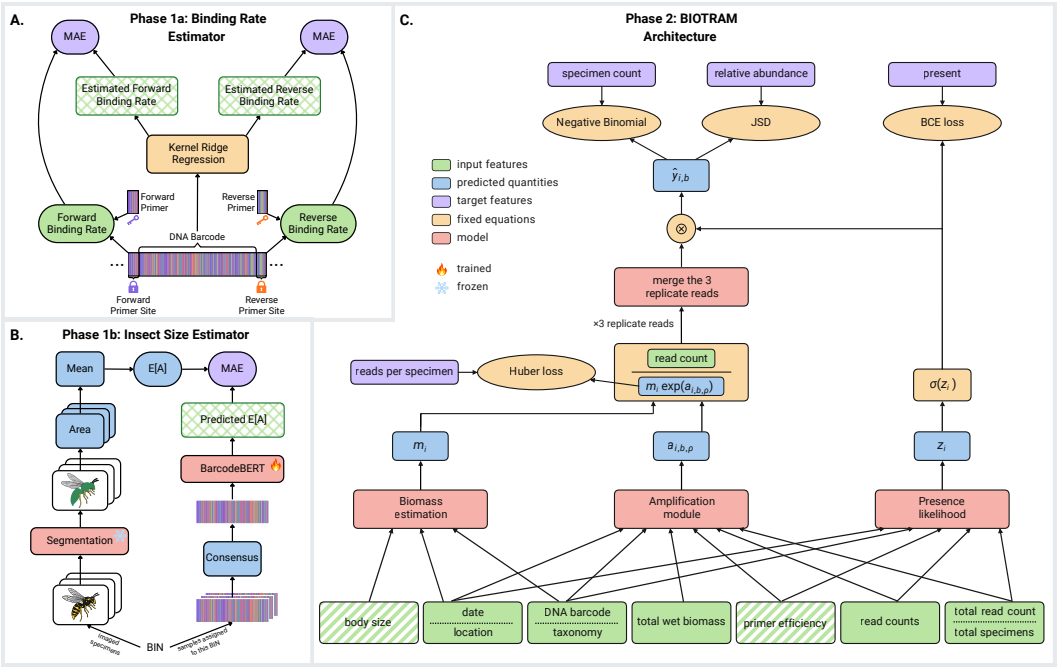


Figure 2: **BIOTRAM method overview.** (A) *Binding Rate Estimator* predicts each primer's binding rate from its barcode, trained against rates estimated for longer sequences with known DNA at primer sites. (B) *Insect Size Estimator* predicts a BIN's average specimen size from its consensus barcode sequence, trained on estimates found from paired image data. (C) *BIOTRAM's architecture* integrates barcode, taxonomic, and sample metadata to jointly predict per-BIN biomass and presence; an attention module corrects raw read counts for amplification bias across replicates before the corrected estimates are merged into a final abundance prediction, trained with negative binomial, Jensen–Shannon divergence (JSD), and binary cross-entropy (BCE) losses. For submodule component design, see Figure 6.

We assume that the read count of BIN i in bottle b for replicate ρ is, up to a bottle-level constant that absorbs the replicate's total read count, the product of three quantities: (1) the specimen count $n_{i,b}$, which we predict; (2) the amplifiable DNA per specimen, which on average scales with the BIN's expected body mass, the body-mass term m_i; and (3) the log amplification $a_{i,b,\rho}$ of its barcode in replicate ρ before the reagents ran out:

$$\text{reads}_{i,b,\rho} \;\propto\; n_{i,b} \cdot m_i \cdot \exp\big(a_{i,b,\rho}\big). \tag{1}$$

See Appx. C for our derivation of this forward relationship.

4.1 BODY SIZE AND PRIMER BINDING EFFICIENCY

Two per-BIN inputs inform the terms of Equation (1) and can be estimated from data outside metabarcoding: body size, which the body-mass term m_i should track, and primer binding efficiency,

which shapes the log amplification $a_{i,b,\rho}$. We estimate both from public specimen images and reference sequences, freeze the estimators, and give their outputs to BIOTRAM and every baseline. Sec. 5 evaluates these models and Appendices D and F detail them.

Body size. Larger individuals have more surface in contact with the DNA extraction fluid, so more of their DNA enters the amplification process. To estimate specimen sizes, we used the size of photographed insects as a proxy (Figure 2B). From the public specimen images in the Barcode of Life Data System (BOLD), we assembled 639,751 images associated with BIN labels and segmented each image using pretrained vision models (Appx. F). The pipeline segments the whole organism and then carves away spread wings, which add projected area but little tissue (Figure 5). For each BIN, we averaged the projected body area over specimens, giving an expected size for 78,259 BINs, representing 50.2% of our dataset's BINs. To generalize this model beyond BINs with imaged examples, we fine-tuned a pretrained barcode model–BarcodeBERT (Millan Arias et al., 2026)–to estimate the mean log area of a BIN and its within-BIN spread from a single barcode within the BIN; BIOTRAM reads these predicted values for every BIN.

Primer binding efficiency. The PCR process consists of copying the barcode region between two primers, short fixed sequences that bind to the DNA on either side of it (Figure 2A). The sequence at these binding sites varies across species, and a primer binds better the more closely the site matches it. These sites are not shown in reads, because amplification overwrites them with the primer sequences, so we take them from two external sources: full-length barcode records, which extend beyond the amplicon and give the forward site for 91.5% of BINs, and mitogenome records give both the forward and the reverse sites for another 1.2%. We score each retrieved 20-nucleotide site to estimate a binding efficiency between 0 and 1; for the remaining BINs, we predict the efficiency from the amplicon by ridge regression (Appx. D).

4.2 Sources of variation in reads per specimen

We quantify what distorts the relationship between reads and specimen counts: how much is attributable to the BIN itself (e.g., primer binding, body size), the reaction as a whole, and detection noise. Each source of variation motivates a specific component of the model.

Most of the variation is per-BIN, but it cannot be looked up. The number of reads one specimen produces spans 4.2 orders of magnitude across BINs, and BIN identity explains 80.4% of that variation. The classical remedy corrects each BIN by its own properties (McLaren et al., 2019), a structure our model keeps, but 60% of BINs appear in a single bottle and 19% of test rows carry a BIN absent from training, which a lookup table cannot handle. We, instead, compute the per-BIN quantities from the barcode sequence, since primer binding and body size are conserved among relatives.

False detections. 50.9% of the BINs detected in a bottle never occur: contamination, cross-talk between bottles sequenced together, traces of environmental DNA and clustering artifacts all leave reads where no specimen was present (Schnell et al., 2015; Zinger et al., 2019). A dedicated presence head, trained with its own loss, gates the abundance estimate, so a detected BIN can be reported absent.

Replicate disagreement. Only 46.8% of detections appear in all three PCR replicates and 29.4% in just one. When little of a BIN's DNA is present, whether it amplifies is partly chance, so replicate disagreement is itself evidence of how much DNA was there. The model takes the three replicate counts as inputs, not their sum.

Bottle-level saturation. A PCR amplifies exponentially until its fixed reagent budget is spent, then plateaus. How long that takes depends on the whole bottle. Because BINs amplify at different rates, the ratio between two BINs' reads depends on every BIN in the bottle (Suzuki & Giovannoni, 1996): the emptiest quarter of bottles yields a median of 5,896 reads per specimen, roughly $8\times$ the 720 of the fullest quarter. Sequencing then samples only a fraction of the amplified product. Both effects require bottle-level terms.

4.3 THE MODEL

BIOTRAM predicts m_i and $a_{i,\rho}^{(b)}$, together with the probability that the BIN is present at all ($\sigma(z_i)$), and inverts Equation (1) against the observed read count,

$$\begin{aligned} \log \tilde{n}_{i,\rho} \;&=\; \log\bigl(1+\mathrm{reads}_{i,b,\rho}\bigr) \;-\; \log T_{b,\rho} \;-\; a_{i,b,\rho} \;-\; \log m_i \;+\; \log \nu_b, \\ \log \hat{n}_i \;&=\; \mathcal{M}_i\bigl(\log \tilde{n}_{i,1},\, \log \tilde{n}_{i,2},\, \log \tilde{n}_{i,3};\; \log \nu_b\bigr) \;+\; \log \sigma\bigl(z_i\bigr), \end{aligned} \tag{2}$$

with $T_{b,\rho}$ the replicate's total reads, ν_b the total specimen count, z_i the logit that BIN i is truly present, σ the sigmoid function, and $\mathcal{M}_i$ a small learned network that merges the three replicate channels, starting from their mean. The channels are each BIN's smallest, median and largest read count, so they do not depend on the arbitrary order of the replicates.

Two trained networks predict the per-BIN quantities m_i and z_i. Each computes a frozen Barcode-Mamba embedding of the barcode (Gao & Taylor, 2024; Gao et al., 2025) through its own attention layer, adds a learned taxonomy embedding and the bottle's location and collection date (through a pretrained geographic encoder (Klemmer et al., 2025) and a seasonal encoding resp.) and passes the result through an MLP. The frozen embedding is what lets an unseen BIN inherit from its relatives. The body mass network also reads the barcode-predicted body size of Sec. 4.1, the presence network reads the primer efficiencies, the bottle's total specimen count and the BIN's replicate read counts, so $\sigma(z_i)$ is a read-informed presence probability.

The amplification term $a_{i,b,\rho}$ comes from a self-attention module over every BIN of the bottle, run once per replicate channel. Each BIN is characterized by its barcode, its two primer efficiencies (each with a binary input marking whether it was computed from a retrieved binding site or predicted; Sec. 4.1), its predicted body size, and its read count and read share within the bottle. Every BIN also receives a summary of the bottle: the mean of all its BINs' representations, weighted by a softmax over their amplification demand, estimated from the predicted body size and the two primer efficiencies with a learned temperature, along with the bottle's covariates and the replicate's total reads. The BINs expected to consume most of the shared reagents thus dominate the context from which every amplification term is estimated. A *read-blind* variant estimates $a_{i,b,\rho}$ without the BIN's own read (Appx. I), and Appx. P measures what each input contributes. The count is read out as the mean of a negative binomial whose dispersion is predicted per row. Apart from the replicate merge, the arithmetic that turns the predicted quantities into a count is fixed by Equation (2) (overview in Figure 2C, full architecture in Figure 6).

4.4 TRAINING OBJECTIVE

Let $\mathcal{B}$ be the bottles in a batch and $\mathcal{R}_b$ the per-BIN observation data for bottle b, and write $n_{i,b}$ for the observed count, $\hat{n}_{i,b}$ for Equation (2)'s prediction with the bottle index made explicit, and $\sigma(z_{i,b})$ for the predicted presence probability. Training minimizes a negative-binomial count likelihood (NB), which accommodates the over-dispersion of ecological counts, whose variance exceeds their mean, plus a per-bottle divergence on composition, a presence loss, and a term tying the predicted reads per specimen to its observed value:

$$\mathcal{L} \;=\; \mathcal{L}_{\mathrm{NB}} \;+\; \lambda_{\mathrm{comp}}\,\mathcal{L}_{\mathrm{comp}} \;+\; \lambda_{\mathrm{pres}}\,\mathcal{L}_{\mathrm{pres}} \;+\; \lambda_{\mathrm{rps}}\,\mathcal{L}_{\mathrm{rps}}, \tag{3}$$

$$\mathcal{L}_{\mathrm{NB}} \;=\; -\frac{1}{|\mathcal{R}|}\sum_{(i,b)\in\mathcal{R}} \log \mathrm{NB}\bigl(n_{i,b}\,;\,\hat{n}_{i,b},\,\alpha_{i,b}\bigr), \qquad \mathrm{Var}\left[n_{i,b}\right] \;=\; \hat{n}_{i,b} + \alpha_{i,b}\,\hat{n}_{i,b}^{2}, \tag{4}$$

$$\mathcal{L}_{\mathrm{comp}} \;=\; \frac{1}{|\mathcal{B}|}\sum_{b\in\mathcal{B}} \mathrm{JSD}\left(p_b \,\middle\|\, \hat{p}_b\right), \qquad p_{i,b} = \frac{n_{i,b}}{\sum_{j\in b} n_{j,b}}, \quad \hat{p}_{i,b} = \frac{\hat{n}_{i,b}}{\sum_{j\in b} \hat{n}_{j,b}}, \tag{5}$$

$$\mathcal{L}_{\mathrm{pres}} \;=\; -\frac{1}{|\mathcal{R}|}\sum_{(i,b)\in\mathcal{R}} \Bigl[\, y_{i,b}\log\sigma(z_{i,b}) + \bigl(1-y_{i,b}\bigr)\log\bigl(1-\sigma(z_{i,b})\bigr)\Bigr], \qquad y_{i,b} = \mathbf{1}\bigl[\, n_{i,b}\geq 1\,\bigr], \tag{6}$$

$$\mathcal{L}_{\mathrm{rps}} \;=\; \frac{1}{\sum_{(i,b)} w_i}\sum_{(i,b)\in\mathcal{R}} w_i\, H_\delta\!\left(\frac{\varepsilon_{i,b}-\bar{\varepsilon}}{s_i}\right), \qquad \varepsilon_{i,b} \;=\; \bigl(a_{i,b}+\log m_i\bigr) - \ell_i. \tag{7}$$

Table 1: **In-domain test performance.** Test split of the excess-free bottles; every model sees the same inputs. The hurdle mixed model of Furneaux et al. (2026) is fitted on its own likelihood with its published predictors (Appx. I); the deep models are trained on the objective of Equation (3), each with its auxiliary weights tuned for it (Appx. O). JSD scores the predicted composition, Spearman the ranking of the counts over all test rows, presence AUC how well the predicted counts separate present from absent BINs, and MAE/bottle the error on the counts within each bottle, averaged over bottles; $\text{MAE}_{=0}$ and $\text{MAE}_{>0}$ are the count error over all BINs truly absent or present. Values are means over seeds $\pm$ one standard deviation across seeds (the hurdle mixed model is a single fit). We highlight the **best** and <u>second best</u> model.

Model	JSD ↓	Spearman ↑	Pres. AUC ↑	MAE/bottle ↓	$\text{MAE}_{=0}$ ↓	$\text{MAE}_{>0}$ ↓
Hurdle GLMM (Furneaux et al., 2026)	0.1947	0.609	0.8419	1.498	0.730	2.286
Transformer @ 9.6M	0.1257 $\pm$ 0.0019	<u>0.741</u> $\pm$ 0.003	0.9156 $\pm$ 0.0017	0.723 $\pm$ 0.020	0.295 $\pm$ 0.023	<u>1.395</u> $\pm$ 0.048
MLP @ 11.1M	<u>0.1234</u> $\pm$ 0.0006	<u>0.741</u> $\pm$ 0.001	0.9154 $\pm$ 0.0007	0.737 $\pm$ 0.025	0.303 $\pm$ 0.013	1.409 $\pm$ 0.047
DeepSets @ 10.1M	0.1256 $\pm$ 0.0010	<u>0.742</u> $\pm$ 0.002	<u>0.9162</u> $\pm$ 0.0013	0.752 $\pm$ 0.036	0.325 $\pm$ 0.020	1.453 $\pm$ 0.110
1-D CNN @ 11.5M	0.1235 $\pm$ 0.0008	0.737 $\pm$ 0.001	0.9145 $\pm$ 0.0008	0.762 $\pm$ 0.016	0.321 $\pm$ 0.010	1.473 $\pm$ 0.036
BIOTRAM	**0.1200** $\pm$ 0.0009	<u>0.741</u> $\pm$ 0.001	0.9136 $\pm$ 0.0005	**0.699** $\pm$ 0.017	<u>0.279</u> $\pm$ 0.016	**1.341** $\pm$ 0.024
TabICLv2 (in-context, 28.5M)	0.1238 $\pm$ 0.0001	**0.751** $\pm$ 0.001	**0.9239** $\pm$ 0.0005	<u>0.704</u> $\pm$ 0.003	**0.249** $\pm$ 0.002	1.413 $\pm$ 0.006

The NB variance keeps a taxon whose count genuinely varies between bottles from having that variance absorbed as bias in the mean. $\mathcal{L}_{\text{comp}}$ scores the within-bottle composition, with $\text{JSD}(P\|Q) = \frac{1}{2}\,\text{KL}(P\|M) + \frac{1}{2}\,\text{KL}(Q\|M)$ at $M = \frac{1}{2}(P+Q)$; it is also the reported metric and the checkpoint-selection criterion. We use JSD rather than KL because it is symmetric — over-predicting a rare BIN and under-predicting an abundant one are penalized alike — and because it stays finite on the 50.9% of rows whose truth is zero, which $\text{KL}(p_b\|\hat{p}_b)$, weighted by the true abundance, does not directly penalize. $\mathcal{L}_{\text{pres}}$ trains the presence probability $\sigma(z_{i,b})$ of Equation (2) directly on the presence label, so that a BIN with reads but no individuals is treated as a false detection rather than a small count. Unlike the other terms, $\mathcal{L}_{\text{rps}}$ supervises an intermediate quantity rather than the output. By Equation (1), $a_{i,b} + \log m_i$ is the log number of reads per specimen of BIN i, where $a_{i,b}$ is the mean of $a_{i,b,\rho}$ over the three replicates. Its target ℓ_i is the mean of $\log(\text{reads}/\text{specimen})$ for BIN i over the training bottles, and s_i is the standard deviation of that quantity across those bottles. Only BINs observed in at least three training bottles contribute to the loss ($w_i = 1$, otherwise $w_i = 0$). Training batches are whole bottles (Appx. G).

5 Results

We evaluate BIOTRAM against a broad range of baselines, on random splits, under three kinds of distribution shift and on an external dataset with a different laboratory workflow.

Evaluation design and metrics. Data is partitioned at the bottle level. The 585 excess-free bottles are split 409 / 87 / 89 into train/val/test, and the bottles with excess specimens (Appx. E) join the training set only, for 864 training bottles in all. The baselines span linear models, nearest neighbours, tree-based models, zero-inflated and hurdle models, deep set and attention-based architectures. All receive the same inputs and are each tuned on their own hyperparameter search space against Jensen–Shannon divergence (JSD) between the predicted and the true within-bottle composition on the validation set, our primary metric. From the same predicted counts, we also score the mean absolute error (MAE; per bottle, on rows with no individuals, $\text{MAE}_{=0}$, on the rest, $\text{MAE}_{>0}$, and on log counts, MAE-log), presence AUC, Spearman correlation and two summary statistics used frequently in ecology (Appx. H). Two models are *level* when a two-sided Welch test over seeds cannot separate them.

Every deep model, BIOTRAM included, is trained on the objective of Equation (3), with its three auxiliary weights tuned for it under one protocol and budget (Appx. O). Its term $\mathcal{L}_{\text{rps}}$ supervises the log reads per specimen, which by Equation (1) is $a_{i,b} + \log m_i = \log(\text{reads}_{i,b}/n_{i,b})$. A baseline has no amplification term, but its predicted count $\hat{n}_{i,b}$ gives the same quantity as $\log(\text{reads}_{i,b}/\hat{n}_{i,b})$, so the term applies to every model.

Overall accuracy. Table 1 reports accuracy on the held-out test bottles. Among the deep models trained on our data, BIOTRAM is the most accurate on both the counts and the composition. Its mean

Table 2: **Out-of-domain test performance.** Each regime holds out one stratum of bottles from training; the collaterals rows are the external dataset of Appx. M. BIOTRAM is compared with the best baseline of each row, among the four matched deep baselines and the hurdle mixed model (Appx. L); Δ is each model's change in JSD from its own random-split score, MAE/bottle is the count error of Table 1, and the **best** value of each metric is in bold.

		BIOTRAM			Best baseline			
Regime	Simulated domain change	JSD ↓	Δ ↓	MAE/bottle ↓	Model	JSD ↓	Δ ↓	MAE/bottle ↓
Random	in-domain reference	**0.1194**	–	**0.700**	1-D CNN	0.1238	–	0.759
Spatial	a region never sampled in training	**0.1612**	+0.0419	**0.977**	Transformer	0.1702	+0.0445	1.076
Lane	a new sequencing batch	**0.1711**	+0.0517	**0.670**	Transformer	0.1758	+0.0501	0.848
Temporal	bottles collected later than any in training	**0.1856**	+0.0662	**0.896**	Transformer	0.1954	+0.0697	1.029
Collaterals dataset	different location, protocol (lysis)	0.2189	+0.0995	**2.781**	Hurdle GLMM	**0.2026**	+0.0078	4.146
Collaterals dataset	different location, protocol (homogenate)	0.2177	+0.0984	**2.544**	Hurdle GLMM	**0.1965**	+0.0018	3.981

absolute error per bottle is 0.699, which is 3.2% lower than that of the strongest baseline (0.723), and its JSD is 0.1200, against 0.1234 for the best baseline. Both differences are significant against each of the four baselines (two-sided Welch test over seeds). The first row of the table is the hurdle mixed model of Furneaux et al. (2026), the established statistical approach to this problem. It is fitted on its own likelihood rather than on Equation (3), so we show it for reference, outside the matched comparison. Simple predictors do much worse (Table 4). A uniform composition, with every taxon equally abundant, scores 0.3174 on JSD and the raw read shares score 0.3586. The classical per-BIN correction of Sec. 4 (McLaren et al., 2019) reaches only 0.3435, because 60% of BINs appear in a single bottle and most test rows therefore have no correction to apply.

The ranking metrics show a different pattern. On rank correlation, BIOTRAM is ahead of one of the baselines and level with three, and on presence AUC it trails four of them. This follows from the mechanism. Dividing the measured reads by an estimated amplification and body mass (Equation (2)) constrains the scale of each prediction, which is what MAE measures, but it offers no comparable help in deciding which of two taxa in a bottle is more abundant. That ranking is largely decided by presence, which the read counts already carry and which an unconstrained network can fit directly.

TabICLv2 (Qu et al., 2026) was kept apart since it works differently. It is pretrained on synthetic tabular datasets and is not trained on our data at all; instead, it predicts each test bottle by attending to examples from the training set. It has the best presence AUC, MAE-log and $\text{MAE}_{=0}$ of any model. BIOTRAM is level with it on count accuracy (MAE 0.699 against 0.704) and ahead of it on composition, with 35% of its parameters. Attending to the training set also makes TabICLv2 expensive to run: it needs 4×10^5 *times more compute per bottle* than BIOTRAM (Appx. I).

Error structure and distribution shift. By true abundance, BIOTRAM is most accurate on the rows with no individuals, ahead of four of the matched baselines, and on counts above 100, ahead of three of them and level with one: only 0.1% of rows, but they carry a large share of each bottle's individuals. It is also ahead of four of them on rows whose BIN was seen in training, but behind at least one of them on three of the four buckets between 1 and 100, and behind two of them on BINs never seen in training (Appx. J). A random split shares sites, seasons and sequencing batches between training and testing, which deployment does not, so we also retrain with a region never seen, bottles collected later than any in training, or a new sequencing batch held out (Table 2). The temporal shift is the hardest, adding +0.0662 in JSD against +0.0419 for the spatial and +0.0517 for the batch shift. Against each regime's best baseline (Table 2), BIOTRAM is ahead of it on the spatial split (-0.0090 JSD), ahead of it on the batch split (-0.0047) and ahead of it on the temporal one (-0.0099; Appx. L). On fifteen Malaise-trap bottles from Sweden (4,057 specimens, each barcoded individually), metabarcoded by another laboratory with its own protocol (unpublished data; Appx. M), BIOTRAM stays closer to the true composition than the read shares and is level with the Transformer and TabICLv2; the best baseline there is the hurdle mixed model of Furneaux et al. (2026), closer on composition but with the largest count error of any model (Table 2). Beyond point estimates, BIOTRAM's NB predictive distribution is conservative: its central 90% intervals cover 0.967 of held-out rows (Appx. K).

Robustness of the comparison. Table 1 rests on three choices for every deep model: loss weights tuned for it, a size close to BIOTRAM's and the full training set. We vary each in turn to test whether BIOTRAM's lead depends on them. With one set of loss weights shared by every model (Appx. O), only the architecture differs: BIOTRAM's MAE then separates from four of the four deep baselines

and its JSD from four, with an MAE 5.8% below the strongest of them, against 3.2% with tuned weights (Appx. I). Across a capacity ladder from 200k to 128M parameters, retuned at every rung, no model moves by more than 0.0062 JSD, and BIOTRAM at its worst size (0.121) is still ahead of the 1-D CNN at its best (0.123). Only training-set size reverses the lead: on nested subsets of the training bottles, BIOTRAM trails every other model up to 216 bottles and leads the four deep baselines from 648 on, and every model is still improving at the full set, BIOTRAM most steeply (Appx. Q). Its structure thus pays off once there are enough bottles to learn the per-BIN quantities it relies on.

External anchors. Mean log area and mean log dry mass correlate at $r = 0.88$ across the 587 weighed BINs, BarcodeBERT predicts the mean log area of held-out BINs from the barcode alone at $r = 0.86$ (against 0.82 for the nearest training barcode), and $\log(F \cdot R)$ correlates with a BIN's empirical $\log(\text{reads/specimen})$ at $r = 0.21$. The reverse primer site comes only from mitogenomes (1.2% of BINs) and is otherwise imputed, which can be a concern when deploying the model on new communities (Appx. D.1).

Ablations. A hierarchical ablation over 30 capability blocks, each removed from the reported configuration and compared seed-for-seed against a control, finds that the two largest effects are data and chemistry rather than architecture: the extra training bottles with excess (+0.0148 validation JSD) and the primer binding-efficiency block (+0.0147). The amplification module is next (+0.0072), and giving it the BIN's own read is worth +0.0049 on its own (full results in Appx. P).

Do the predicted quantities mean what they are named? The model's distinguishing property is its intermediate quantities, which should correspond to physical quantities and explain its predictions. Since the primer efficiency and body mass receive the corresponding external estimates as input features, correlating them with those estimates would be circular. Instead, we check how well the predicted log reads per specimen $a_{i,b} + \log m_i$ (Equation (1)) correlates with its empirical value: on the test bottles it reaches Pearson $+0.627 \pm 0.007$ over 12 seeds, and the read-blind variant, whose amplification term never sees the read and so cannot reproduce it mechanically, still reaches $+0.461 \pm 0.013$ (Appx. N).

6 Discussion

We have presented a new dataset that pairs metabarcoding read counts with specimen counts, and designed BIOTRAM, a machine learning model specifically imagined for the purpose of mapping those two quantities, trained and evaluated on that dataset.

Limitations. The number of training bottles, not model capacity, limits accuracy: larger variants of BIOTRAM do not improve it, no baseline gains more than 0.0062 JSD across the capacity ladder, and every model is still improving at the full training set (Appx. Q). The counts identify only the product of amplification and body mass, and the amplification module sees the BIN's own read, so the likelihood alone pins down neither term. What separates them is what each component reads: of the two, only the amplification module sees the primer features, the reads and the rest of the bottle. The latent quantities are therefore cleanest to interpret in the read-blind variant (Table 3), at a cost in accuracy. Body size varies within a BIN with life stage, sex and season, the image-based estimate depends on specimen pose, and the weighed-specimen dataset is small. The largest counts carry the least certain training labels: excess-affected taxa are 3.2% of BINs but 40% of the rows with more than a hundred specimens, and in the bottles with excess their counts are allocated to BINs by a model-based procedure (Appx. E). BIOTRAM does not yet use spike-ins, known quantities of specimens or DNA added to a sample, which improve quantification in metabarcoding and metagenomics (Chen et al., 2016; Harrison et al., 2021; Iwaszkiewicz-Eggebrecht et al., 2026; Ji et al., 2020b). Finally, the pretrained geographic encoder (Klemmer et al., 2025) contributes little, suggesting that general-purpose location embeddings do not transfer readily to biodiversity prediction.

Conclusions. BIOTRAM treats the biochemistry of metabarcoding as part of the model. Every quantity it predicts has a physical meaning, two are anchored to independent measurements, and the arithmetic that converts them into counts is fixed rather than learned. With every model's loss weights tuned, it reaches the lowest composition divergence and the lowest count error of the architectures we compared. Every model we tested degrades substantially under spatial, temporal and batch shift. Primer binding efficiency is among the largest gains in our ablation, which argues for single-specimen barcodes that retain the primer binding sites (Appx. D).

AI USE STATEMENT

We used generative AI tools, mainly Anthropic's Claude models through the Claude Code assistant. For research ideation and execution, they helped us refine hypotheses and experimental designs, for example, how to compare BIOTRAM with the baselines on an equal footing, and write and debug code for the experiments, including the loss-weight tuning and scripts to tabulate the experiment outputs. For writing, they drafted parts of some sections, proposed the structure of the appendix, and edited the text for clarity and length. We did not use them to develop the model's biochemical framework, to curate the dataset or to interpret the results. All AI-assisted text was reviewed and edited by the authors, all AI-written code was reviewed and run by them, and every result we report is generated by scripts from the experiment outputs. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The data used in this work was collected as part of the LIFEPLAN project and the Global Malaise Program (Appx. A). We obtained explicit permission from all data owners to use their data.

REPRODUCIBILITY STATEMENT

Upon acceptance, we will release all our source code, pretrained checkpoints, and our training and evaluation data.

REFERENCES

Patrick Aboyoun and Robert Gentleman. *pwalign: Perform pairwise sequence alignments*, 2024. URL https://bioconductor.org/packages/pwalign. R package version 1.0.0.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*, 2025. doi:10.48550/arXiv.2511.21631. URL http://arxiv.org/abs/2511.21631.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment Anything with Concepts. In *International Conference on Learning Representations*, pp. 138846–138923, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/e0982cbc81401df3430ee1ff780dc7a2-Abstract-Conference.html.

Scott A. Chamberlain and Eduard Szöcs. taxize: taxonomic search and retrieval in R. *F1000Research*, 2:191, 2013. doi:10.12688/f1000research.2-191.v2.

K. Chen, Z. Hu, Z. Xia, D. Zhao, W. Li, and J. K. Tyler. The overlooked fact: fundamental need for spike-in control for virtually all genome-wide analyses. *Molecular and Cellular Biology*, 36: 662–667, 2016. doi:10.1128/MCB.00970-14.

Robert K Colwell, Laura Melissa Guzman, Dirk Steinke, Anne Chao, Daniel H Janzen, Winnie Hallwachs, Austin Baker, José L Fernández-Triana, Paul D N Hebert, Frank Joyce, Robert

Puschendorf, Donald L J Quicke, Rodolphe Rougerie, M Alex Smith, Nelson Zamora, and Michael J Sharkey. Constructing a lower-bound estimate of the global number of insect species on a hyperdiverse empirical foundation. *Proc. Natl. Acad. Sci. U. S. A.*, 123(27):e2524283123, July 2026.

Jeremy R. deWaard, Sujeevan Ratnasingham, Evgeny V. Zakharov, Alex V. Borisenko, Dirk Steinke, Angela C. Telfer, Kate H. J. Perez, Jayme E. Sones, Monica R. Young, Valerie Levesque-Beaudin, Crystal N. Sobel, Arusyak Abrahamyan, Kyrylo Bessonov, Gergin Blagoev, Stephanie L. deWaard, Chris Ho, Natalia V. Ivanova, Kara K. S. Layton, Liuqiong Lu, Ramya Manjunath, Jaclyn T. A. McKeown, Megan A. Milton, Renee Miskie, Norm Monkhouse, Suresh Naik, Nadya Nikolova, Mikko Pentinsaari, Sean W. J. Prosser, Adriana E. Radulovici, Claudia Steinke, Connor P. Warne, and Paul D. N. Hebert. A reference library for Canadian invertebrates with 1.5 million barcodes, voucher specimens, and DNA samples. *Scientific Data*, 6(1):308, Dec 2019. ISSN 2052-4463. doi:10.1038/s41597-019-0320-2.

Jeremy R. deWaard, Stephanie deWaard, Maria Kuzmina, Jayme E Sones, Kate Perez, Amy Thompson, Bess Hardwick, Brendan Furneaux, Tomas Roslin, and Evgeny Zakharov. *LIFEPLAN Malaise sample metabarcoding v1*, March 2024.

Raphael K Didham, Yves Basset, C Matilda Collins, Simon R Leather, Nick A Littlewood, Myles H M Menz, Jörg Müller, Laurence Packer, Manu E Saunders, Karsten Schönrogge, Alan J A Stewart, Stephen P Yanoviak, and Christopher Hassall. Interpreting insect declines: seven challenges and a way forward. *Insect Conserv. Divers.*, 13(2):103–114, March 2020.

Vasco Elbrecht and Florian Leese. Can DNA-based ecosystem assessments quantify species abundance? Testing primer bias and biomass–sequence relationships with an innovative metabarcoding protocol. *PLoS One*, 10(7):e0130324, July 2015.

Vasco Elbrecht and Florian Leese. Validation and development of COI metabarcoding primers for freshwater macroinvertebrate bioassessment. *Front. Environ. Sci.*, 5:237020, 10 April 2017. doi:10.3389/fenvs.2017.00011.

Vasco Elbrecht, Thomas W A Braukmann, Natalia V Ivanova, Sean W J Prosser, Mehrdad Hajibabaei, Michael Wright, Evgeny V Zakharov, Paul D N Hebert, and Dirk Steinke. Validation of COI metabarcoding primers for terrestrial arthropods. *PeerJ*, 7(e7745):e7745, 7 October 2019a. doi:10.7717/peerj.7745.

Vasco Elbrecht, Thomas W A Braukmann, Natalia V Ivanova, Sean W J Prosser, Mehrdad Hajibabaei, Michael Wright, Evgeny V Zakharov, Paul D N Hebert, and Dirk Steinke. Validation of COI metabarcoding primers for terrestrial arthropods. *PeerJ*, pp. e7745, July 2019b.

B. Furneaux, T. Roslin, B. Hardwick, D. Kerdraon, H. Autto, G. Banelyte, J.R. deWaard, S.L. deWaard, A. Farrell, O. Kalttopää, and E. Kristensen. Is there a fly in my soup? To what extent do metabarcoding and individual barcoding tell the same story? *Molecular Ecology Resources*, 26(7): e70195, 2026. doi:10.1111/1755-0998.70195.

Tiancheng Gao and Graham Taylor. BarcodeMamba: State space models for biodiversity analysis. *arXiv preprint arXiv:2412.11084*, 2024. doi:10.48550/arXiv.2412.11084.

Tiancheng Gao, Scott C. Lowe, Brendan Furneaux, Angel X. Chang, and Graham W. Taylor. BarcodeMamba+: Advancing state-space models for fungal biodiversity research. *arXiv preprint arXiv:2512.15931*, 2025. doi:10.48550/arXiv.2512.15931.

Matthias F Geiger, Jerome Moriniere, Axel Hausmann, Gerhard Haszprunar, Wolfgang Wägele, Paul D N Hebert, and Björn Rulik. Testing the global malaise trap program – how well does the current barcode reference library identify flying insects in Germany? *Biodivers. Data J.*, 4(4):e10671, December 2016.

Zahra Gharaee, Scott C. Lowe, ZeMing Gong, Pablo Millan Arias, Nicholas Pellegrino, Austin T. Wang, Joakim Bruslund Haurum, Iuliia Zarubiieva, Lila Kari, Dirk Steinke, Graham W. Taylor, Paul Fieguth, and Angel X. Chang. BIOSCAN-5M: A multimodal dataset for insect biodiversity. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.),

Advances in Neural Information Processing Systems, volume 37, pp. 36285–36313. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/3fdbb472813041c9ecef04c20c2b1e5a-Paper-Datasets_and_Benchmarks_Track.pdf.

Caspar A Hallmann, Martin Sorg, Eelke Jongejans, Henk Siepel, Nick Hofland, Heinz Schwan, Werner Stenmans, Andreas Müller, Hubert Sumser, Thomas Hörren, Dave Goulson, and Hans de Kroon. More than 75 percent decline over 27 years in total flying insect biomass in protected areas. *PLoS One*, 12(10):e0185809, 18 October 2017a. doi:10.1371/journal.pone.0185809.

Caspar A Hallmann, Martin Sorg, Eelke Jongejans, Henk Siepel, Nick Hofland, Heinz Schwan, Werner Stenmans, Andreas Müller, Hubert Sumser, Thomas Hörren, Dave Goulson, and Hans de Kroon. More than 75 percent decline over 27 years in total flying insect biomass in protected areas. *PLoS One*, 12(10):e0185809, October 2017b.

Bess Hardwick, Deirdre Kerdraon, Hanna M K Rogers, Dimby Raharinjanahary, Eric Tsiriniaina Rajoelison, Tommi Mononen, Petteri Lehikoinen, Gaia Banelyte, Arielle Farrell, Brian L Fisher, Tomas Roslin, and Otso Ovaskainen. LIFEPLAN: A worldwide biodiversity sampling design. *PLoS One*, 19(12):e0313353, December 2024.

J. G. Harrison, W. John Calder, B. Shuman, and C. Alex Buerkle. The quest for absolute abundance: the use of internal standards for DNA-based community ecology. *Molecular Ecology Resources*, 21:30–43, 2021. doi:10.1111/1755-0998.13247.

Ela Iwaszkiewicz-Eggebrecht, Emma Granqvist, Karol H. Nowak, Catalina Valdivia, Mateusz Buczek, Amrita Srivathsan, Emily Hartop, Andreia Miraldo, Tomas Roslin, Ayco J. M. Tack, Piotr Łukasik, Rudolf Meier, and Fredrik Ronquist. Accuracy of occurrence and abundance estimates from insect metabarcoding. *Methods in Ecology and Evolution*, 2026. doi:10.1111/2041-210x.70351.

Yinqiu Ji, Louise Ashton, Scott M Pedley, David P Edwards, Yong Tang, Akihiro Nakamura, Roger Kitching, Paul M Dolman, Paul Woodcock, Felicity A Edwards, Trond H Larsen, Wayne W Hsu, Suzan Benedick, Keith C Hamer, David S Wilcove, Catharine Bruce, Xiaoyang Wang, Taal Levi, Martin Lott, Brent C Emerson, and Douglas W Yu. Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecol. Lett.*, 16(10):1245–1257, October 2013. doi:10.1111/ele.12162.

Yinqiu Ji, Tea Huotari, Tomas Roslin, Niels Martin Schmidt, Jiaxin Wang, Douglas W Yu, and Otso Ovaskainen. SPIKEPIPE: A metagenomic pipeline for the accurate quantification of eukaryotic species occurrences and intraspecific abundance change using DNA barcodes or mitogenomes. *Mol. Ecol. Resour.*, 20(1):256–267, January 2020a.

Yinqiu Ji, Tea Huotari, Tomas Roslin, Niels Martin Schmidt, Jiaxin Wang, Douglas W. Yu, and Otso Ovaskainen. SPIKEPIPE: A metagenomic pipeline for the accurate quantification of eukaryotic species occurrences and intraspecific abundance change using DNA barcodes or mitogenomes. *Molecular Ecology Resources*, 20(1):256–267, 2020b. doi:10.1111/1755-0998.13057.

Dave Karlsson, Emily Hartop, Mattias Forshage, Mathias Jaschhof, and Fredrik Ronquist. The Swedish malaise trap project: A 15 year retrospective on a countrywide insect inventory. *Biodivers. Data J.*, 8:e47255, January 2020.

Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan Liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment Anything in High Quality. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 29914–29934. Curran Associates, Inc., 2023. doi:10.52202/075280-1303.

Konstantin Klemmer, Esther Rolf, Caleb Robinson, Lester Mackey, and Marc Rußwurm. SatCLIP: Global, general-purpose location embeddings with satellite imagery. *Proc. Conf. AAAI Artif. Intell.*, 39(4):4347–4355, 11 April 2025. doi:10.1609/aaai.v39i4.32457.

Philip D Lamb, Ewan Hunter, John K Pinnegar, Simon Creer, Richard G Davies, and Martin I Taylor. How quantitative is metabarcoding: A meta-analytical approach. *Mol. Ecol.*, 28(2):420–430, January 2019.

Joanne E Littlefair, Lauren D Hayhurst, Matthew C Yates, Michael D Rennie, and Melania E Cristescu. Environmental DNA metabarcoding reflects rank-based abundance of fish. *J. Appl. Ecol.*, 63(3), March 2026.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with Grounded Pre-training for Open-Set Object Detection. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 38–55. Springer Nature Switzerland, 2025. doi:10.1007/978-3-031-72970-6_3.

Silvia Matesanz, David S Pescador, Beatriz Pías, Ana M Sánchez, Julia Chacón-Labella, Angela Illuminati, Marcelino de la Cruz, Jesús López-Angulo, Neus Marí-Mena, Antón Vizcaíno, and Adrián Escudero. Estimating belowground plant abundance with DNA metabarcoding. *Mol. Ecol. Resour.*, 19(5):1265–1277, September 2019.

Michael R McLaren, Amy D Willis, and Benjamin J Callahan. Consistent and correctable bias in metagenomic sequencing experiments. *eLife*, 8(e46923):e46923, 10 September 2019. doi:10.7554/eLife.46923.

Pablo Millan Arias, Niousha Sadjadi, Monireh Safari, Zeming Gong, Austin T Wang, Joakim Bruslund Haurum, Iuliia Zarubiieva, Dirk Steinke, Lila Kari, Angel X Chang, Scott C Lowe, and Graham W Taylor. BarcodeBERT: transformers for biodiversity analyses. *Bioinform. Adv.*, 6(1):vbag054, 19 February 2026. doi:10.1093/bioadv/vbag054.

Johanna Orsholm, John Quinto, Hannu Autto, Gaia Banelyte, Nicolas Chazot, Jeremy deWaard, Stephanie deWaard, Arielle Farrell, Brendan Furneaux, Bess Hardwick, Nao Ito, Amlan Kar, Oula Kalttopää, Deirdre Kerdraon, Erik Kristensen, Jaclyn McKeown, Tommi Mononen, Ellen Nein, Hanna Rogers, Tomas Roslin, Paula Schmitz, Jayme Sones, Maija Sujala, Amy Thompson, Evgeny V Zakharov, Iuliia Zarubiieva, Akshita Gupta, Scott C Lowe, and Graham W Taylor. A multi-modal dataset for insect biodiversity with imagery and DNA at the trap and individual level. *Sci. Data*, 13(1), April 2026.

Sameer Padhye, Liliana Ballesteros-Mejia, and Sujeevan Ratnasingham. *BOLDconnectR: Retrieve, Transform and Analyze the Barcode of Life Data Systems Data*, 2025. URL `https://CRAN.R-project.org/package=BOLDconnectR`. R package version 1.0.0.

Hervé Pagès, Patrick Aboyoun, Robert Gentleman, and Saikat DebRoy. *Biostrings: Efficient manipulation of biological strings*, 2024. URL `https://bioconductor.org/packages/Biostrings`. R package version 2.72.1.

M F Polz and C M Cavanaugh. Bias in template-to-product ratios in multitemplate PCR. *Appl. Environ. Microbiol.*, 64(10):3724–3730, 1 October 1998. doi:10.1128/AEM.64.10.3724-3730.1998.

S W J Prosser, R M Floyd, K A Thompson, S K Monckton, and P D N Hebert. BOLDistilled: Automated construction of comprehensive but compact DNA barcode reference libraries. *Mol. Ecol. Resour.*, 25(8):e70043, November 2025.

Sean W J Prosser, Nicholas W Bard, Ken A Thompson, Robin A Floyd, Sameer Padhye, Emine Ozsahin, Saeideh Jafarpour, and Paul D N Hebert. The metabarcoding analysis pipeline (MAP): Simple, accurate, and flexible metabarcoding. *bioRxiv*, pp. 2026.07.22.740107, 23 July 2026a. URL `https://www.biorxiv.org/content/10.64898/2026.07.22.740107v1.abstract`.

Sean WJ Prosser, Nicholas W Bard, Ken A Thompson, Robin M Floyd, Sameer Padhye, Emine Ozsahin, Saeideh Jafarpour, and Paul DN Hebert. The metabarcoding analysis pipeline (MAP): Simple, accurate, and flexible metabarcoding. *bioRxiv*, 2026b. doi:10.64898/2026.07.22.740107.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICLv2: A better, faster, scalable, and open tabular foundation model. In *Forty-third International Conference on Machine Learning*, 2026. URL `https://openreview.net/forum?id=SxsyLjIfWB`.

Sujeevan Ratnasingham and Paul D N Hebert. A DNA-based registry for all animal species: the barcode index number (BIN) system. *PLoS One*, 8(7):e66213, 8 July 2013. doi:10.1371/journal.pone.0066213.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment Anything in Images and Videos. In *International Conference on Learning Representations*, pp. 28085–28128, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/45c1f6a8cbf2da59ebf2c802b4f742cd-Abstract-Conference.html.

Monireh Safari, Pablo Millan Arias, Scott C. Lowe, Lila Kari, Angel X. Chang, and Graham W. Taylor. Enhancing DNA foundation models to address masking inefficiencies. In *ICLR 2025 Workshop on AI for Nucleic Acids*, 2025. doi:10.48550/arxiv.2502.18405. URL https://openreview.net/forum?id=DGs7lmqkkU.

B E Sample, R J Cooper, R D Greer, and R C Whitmore. Estimation of insect biomass by length and width. *Am. Midl. Nat.*, 129(2):234–240, 1 April 1993. doi:10.2307/2426503.

Eric W Sayers, Mark Cavanaugh, Linda Frisse, Kim D Pruitt, Valerie A Schneider, Beverly A Underwood, Linda Yankie, and Ilene Karsch-Mizrachi. GenBank 2025 update. *Nucleic Acids Res.*, 53(D1):D56–D61, January 2025.

Stefan Schneider, Graham W Taylor, Stefan C Kremer, Patrick Burgess, Jillian McGroarty, Kyomi Mitsui, Alex Zhuang, Jeremy R deWaard, and John M Fryxell. Bulk arthropod abundance, biomass and diversity estimation using deep learning for computer vision. *Methods Ecol. Evol.*, 13:346–357, 14 December 2021. doi:10.1111/2041-210X.13769.

Ida Baerholm Schnell, Kristine Bohmann, and M Thomas P Gilbert. Tag jumps illuminated–reducing sequence-to-sample misidentifications in metabarcoding studies. *Mol. Ecol. Resour.*, 15(6):1289–1303, 1 November 2015. doi:10.1111/1755-0998.12402.

Wiebke Sickel, Vera Zizka, Alice Scherges, Sarah J Bourlat, and Petra Dieker. Abundance estimation with DNA metabarcoding – recent advancements for terrestrial arthropods. *ARPHA Preprints*, July 2023.

M T Suzuki and S J Giovannoni. Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Appl. Environ. Microbiol.*, 62(2):625–630, February 1996. doi:10.1128/aem.62.2.625-630.1996.

Roel van Klink, Diana E Bowler, Konstantin B Gongalsky, Ann B Swengel, Alessandro Gentile, and Jonathan M Chase. Meta-analysis reveals declines in terrestrial but increases in freshwater insect abundances. *Science*, 368(6489):417–420, 24 April 2020. doi:10.1126/science.aax9931.

David L Wagner, Eliza M Grames, Matthew L Forister, May R Berenbaum, and David Stopak. Insect decline in the anthropocene: Death by a thousand cuts. *Proc. Natl. Acad. Sci. U. S. A.*, 118(2): e2023989118, January 2021.

Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. SegGPT: Towards Segmenting Everything in Context. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1130–1140, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Wang_SegGPT_Towards_Segmenting_Everything_in_Context_ICCV_2023_paper.html.

Hadley Wickham and Jennifer Bryan. *readxl: Read Excel Files*, 2025. URL https://CRAN.R-project.org/package=readxl. R package version 1.4.5.

Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the Tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019. doi:10.21105/joss.01686.

David Winter. rentrez: An R package for the NCBI eUtils API. *PeerJ Prepr.*, 2017.

Erik S. Wright. Fast and flexible search for homologous biological sequences with DECIPHER v3. *The R Journal*, 16(2):191–200, 2024.

Derrick Xin, Behrooz Ghorbani, Justin Gilmer, Ankush Garg, and Orhan Firat. Do Current Multi-Task Optimization Methods in Deep Learning Even Help? In *Advances in Neural Information Processing Systems*, volume 35, pp. 13597–13609. Curran Associates, Inc., 2022. doi:10.52202/068431-0988.

Louie H Yang and Claudio Gratton. Insects as drivers of ecosystem processes. *Curr. Opin. Insect Sci.*, 2:26–32, August 2014.

Douglas W Yu, Yinqiu Ji, Brent C Emerson, Xiaoyang Wang, Chengxi Ye, Chunyan Yang, and Zhaoli Ding. Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring: Biodiversity soup. *Methods Ecol. Evol.*, 3(4):613–623, August 2012. doi:10.1111/j.2041-210X.2012.00198.x.

Lucie Zinger, Aurélie Bonin, Inger G Alsos, Miklós Bálint, Holly Bik, Frédéric Boyer, Anthony A Chariton, Simon Creer, Eric Coissac, Bruce E Deagle, Marta De Barba, Ian A Dickie, Alex J Dumbrell, Gentile Francesco Ficetola, Noah Fierer, Luca Fumagalli, M Thomas P Gilbert, Simon Jarman, Ari Jumpponen, Håvard Kauserud, Ludovic Orlando, Johan Pansu, Jan Pawlowski, Leho Tedersoo, Philip Francis Thomsen, Eske Willerslev, and Pierre Taberlet. DNA metabarcoding-need for robust experimental designs to draw sound ecological conclusions. *Mol. Ecol.*, 28(8): 1857–1862, 1 April 2019. doi:10.1111/mec.15060.

APPENDICES

A Data provenance **17**

B Core dataset generation and assembly **17**

C Derivation of the forward DNA amplification relationship **18**

D Primer binding efficiency **19**

D.1 Evaluation 20

E 'Excess' specimens **20**

F Body size estimation: pipeline and evaluation **21**

F.1 Segmenting body area from specimen images 21

F.2 Estimating mass from segmented image surface area 22

F.3 Generalizing to unseen BINs 23

F.4 Evaluation 23

G Detailed model architecture **23**

H Evaluation methods **23**

H.1 Evaluation metrics 23

I The full baseline comparison **25**

J Error by abundance and by novelty **27**

K What a prediction looks like **29**

L Distribution shift **29**

M Validation on external "collaterals" dataset **30**

N The latent quantities **31**

O Loss-weight tuning **32**

P Ablation study **33**

Q Model scaling against baselines **33**

R How a read reaches the count **35**

A DATA PROVENANCE

We identified a set of Malaise trap samples (bottles) that met two conditions: (i) the bottle was metabarcoded upon receipt via mild lysis and (ii) the bottle was entirely sorted for single-specimen barcoding, with any 'excess' (see below) recorded. Samples derived from the LIFEPLAN project (deWaard et al., 2024; Hardwick et al., 2024) and the Global Malaise Program (Geiger et al., 2016). Permissions to use each bottle were arranged with the data owners. Figure 3 summarizes the DNA barcoding data pipeline.

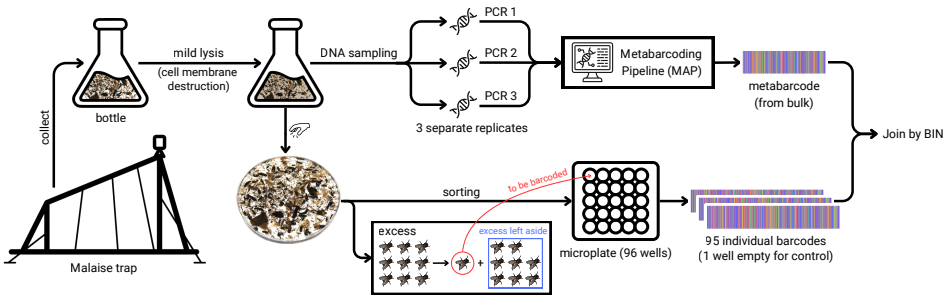


Figure 3: **DNA barcoding data pipeline.** Each bottle is collected from a Malaise trap and undergoes mild lysis to release DNA into solution while preserving the physical specimens. From this lysate, a DNA sample is drawn for metabarcoding: three independent PCR replicates are amplified from the bulk sample and passed through the Metabarcoding Pipeline to produce a single bulk metabarcode. In parallel, the physical sample is fully sorted; where several specimens belong to the same putative species, one representative is retained for individual barcoding and plated onto a 96-well microplate, yielding up to 95 individual barcodes per plate (one well is left empty as a negative control), while the remaining specimens are recorded and set aside as excess rather than individually barcoded. Individual barcode and metabarcode records are then joined by Barcode Index Number (BIN).

B CORE DATASET GENERATION AND ASSEMBLY

Raw metabarcode data were obtained as paired-end gzipped fastq files (i.e., .R1.fq.gz & .R2.fq.gz) accompanied by a plate map. Plate maps were converted into MAP parameter files and were analyzed with MAP v1.0 Prosser et al. 2026b with default settings for Illumina data. OTUs were queried against the June 2026 BOLDistilled COI library (Prosser et al., 2025) to provide BIN matches and taxonomy.

Once tabular metabarcode results were obtained, we integrated the data with data from BOLD to produce a single table containing all relevant information from both sources. Data import, manipulation, and file handling throughout relied on the `tidyverse` (Wickham et al., 2019) and `readxl` (Wickham & Bryan, 2025) packages, while retrieval of specimen-level Barcode Core Data Model (BCDM) records from the Barcode of Life Data System was performed with `BOLDconnectR` (Padhye et al., 2025).

Processing proceeded independently for each sequencing lane. For every lane, filtered OTU tables and per-replicate read counts were loaded into memory. Next, BCDM records for all process IDs associated with the lane were retrieved. Nucleotide sequences from the BOLD SSB records were deduplicated and used to generate a FASTA file, which was then queried against BOLDistilled to obtain BIN assignments. This was done because the 'true' BIN assignments on BOLD cannot be efficiently replicated for metabarcode data due to (i) the RESL algorithm being proprietary, (ii) the use of private data on BOLD, and (iii) BOLD's Barcode ID being only available via web browser and restricted to 100 sequences at a time during an exhaustive search. Since the reference library for metabarcode data becomes out of date every three months, metabarcoding OTUs assigned to Arthropoda were also matched to their corresponding BIN via the BOLDistilled lookup. Following BIN assignment, the metabarcode and single-specimen barcode data were joined by BIN in a single table that summarized total specimen counts and metabarcode read statistics within each bottle.

C Derivation of the forward DNA amplification relationship

Equation (1) follows from the kinetics of a PCR run on a mixture of templates (Suzuki & Giovannoni, 1996; Polz & Cavanaugh, 1998), which Figure 4 traces for two BINs. Consider one PCR replicate ρ of bottle b, and let $x_i(t)$ be the number of copies of BIN i's amplicon after t cycles, treating t as continuous. We drop the indices b and ρ until the end of the derivation.

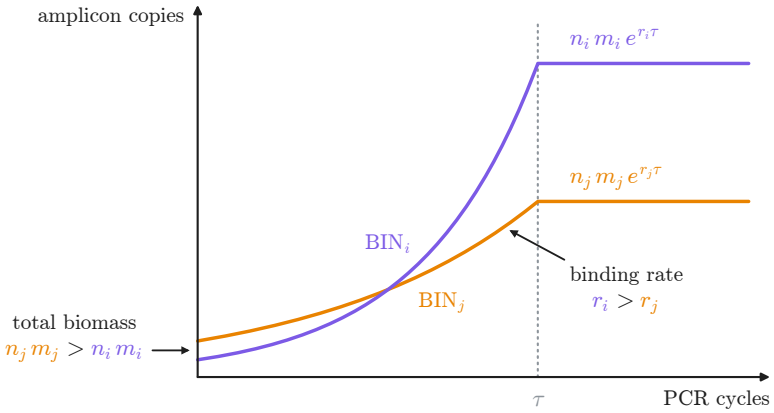


Figure 4: **Amplification from different starting points at different rates.** Two BINs amplified in the same PCR reaction (Equations (8) to (10)). BIN j starts from more template ($n_j m_j > n_i m_i$: heavier specimens, DNA more accessible, or both), but the primers bind BIN i better ($r_i > r_j$), so its copies grow faster and overtake those of BIN j. When the shared reagents run out at cycle τ, both plateau at $n\, m\, e^{r\tau}$, up to the factor κ common to the reaction, and sequencing returns reads in proportion to these final copy numbers (Equation (11)). Had the reaction stopped before the curves cross, BIN j would have yielded more reads. The axes are schematic: a real reaction multiplies its template by many orders of magnitude.

Starting point. Each specimen releases DNA into the lysate, on average in proportion to its biomass (Elbrecht & Leese, 2015), and the reaction draws a fixed fraction κ of that lysate. BIN i therefore enters the reaction with

$$x_i(0) \;=\; \kappa\, n_i\, m_i \tag{8}$$

copies of template. As κ is shared by every BIN of the reaction, the starting points of different BINs differ through the composition of the bottle (n_i) and through the biomass of their specimens (m_i).

Exponential growth. In each cycle, a fraction $E_i \in [0, 1]$ of BIN i's templates is copied. This efficiency depends on how well the primers bind the BIN's sequence (Polz & Cavanaugh, 1998; Elbrecht & Leese, 2015), so it differs between BINs (Appx. D), but not from one cycle to the next while reagents are plentiful. After t cycles,

$$x_i(t) \;=\; x_i(0)\,(1 + E_i)^t \;=\; \kappa\, n_i\, m_i\, e^{r_i t}, \qquad r_i \;=\; \ln(1 + E_i), \tag{9}$$

so each BIN grows exponentially, at a rate r_i between 0 (the primers do not bind) and $\ln 2$ (every template is copied in every cycle).

Plateau. The primers, nucleotides and polymerase are shared by every BIN in the reaction, and together they support at most some total number X of copies. Idealizing saturation as an abrupt stop shared by every BIN, amplification ends at the cycle τ at which the total product reaches this budget,

$$\sum_k x_k(\tau) \;=\; \kappa \sum_k n_k\, m_k\, e^{r_k \tau} \;=\; X, \tag{10}$$

and BIN i ends with $x_i(\tau) = \kappa\, n_i\, m_i\, e^{r_i \tau}$ copies. The duration τ is thus set by the whole bottle: a bottle with more template, or with more readily amplified BINs, exhausts the budget in fewer cycles, so each of its specimens is amplified less. This is the bottle-level saturation of Sec. 4.

Sequencing. The sequencer returns T reads for the replicate, each drawn from the pooled product, so in expectation BIN i receives reads in proportion to its share of that product,

$$\text{reads}_i \;=\; T\,\frac{x_i(\tau)}{\sum_k x_k(\tau)} \;=\; \frac{T\kappa}{X}\; n_i\, m_i\, e^{r_i \tau}. \tag{11}$$

The factor $T\kappa/X$ is common to every BIN of the replicate. Restoring the indices, Equation (11) is Equation (1), with the amplification term $a_{i,\rho}^{(b)} = r_i\, \tau_\rho^{(b)}$.

Consequences for the model. Two BINs of the same reaction compare as

$$\log \frac{\text{reads}_i}{\text{reads}_j} \;=\; \log \frac{n_i}{n_j} \;+\; \log \frac{m_i}{m_j} \;+\; (r_i - r_j)\,\tau, \tag{12}$$

so the read ratio confounds the ratio of specimen counts, which we want to recover, with the ratio of biomasses and with a difference in amplification that grows with the duration of the reaction. A BIN that starts from less template but binds the primers better overtakes the other if the reaction runs long enough (Figure 4), and whether it does depends, through Equation (10), on every other BIN in the bottle. The idealized amplification term factorizes into a per-BIN rate and a per-reaction duration. BIOTRAM keeps the product form of Equation (1) but does not impose this factorization: its attention module predicts $a_{i,\rho}^{(b)}$ from the content of the whole bottle (Sec. 4.3), so the term can also absorb departures from the idealization (Suzuki & Giovannoni, 1996).

D PRIMER BINDING EFFICIENCY

Retrieving the binding sites. How well a BIN amplifies directly depends on its primer binding efficiency that itself depends on the sequences of its two primer binding sites. During metabarcoding PCR, the BF3 and BR2 primers must bind to the 20 nucleotides that flank the amplicon DNA sequence. Since those binding sites are replaced by the primer sequence itself, we cannot use the sequenced amplicons to determine their binding affinity. To accomplish this, we took two approaches.

The first route to obtaining primer binding sites used full mitogenomes, which contain both sites. We downloaded 11,080 complete mitogenomes from GenBank (Sayers et al., 2025), restricting our search to Arthropoda records annotated as mitochondrial complete genomes. Records were queried and retrieved via the NCBI E-utilities using the `rentrez` R package (Winter, 2017), filtered to a minimum length of 14,503 bp (the smallest known mitogenome), and deduplicated to one record per named species. Sequences were parsed and written to per-species FASTA files using `Biostrings` (Pagès et al., 2024). For each downloaded mitogenome, we obtained taxonomic information using the `taxize` R package (Chamberlain & Szöcs, 2013), which queries the NCBI Taxonomy database. Species names were taken directly from each genome's GenBank organism field, so only the ranked classification required an additional network query. For each mitogenome, we searched for binding sites of the degenerate COI primers BF3 and BR2 (Elbrecht et al., 2019b), allowing up to four mismatches per primer, using `Biostrings`. Sites overlapping ambiguous (N) bases were discarded, and where a primer matched more than one site in a genome, the best-scoring site was retained via pairwise alignment implemented in the R package `pwalign` (Aboyoun & Gentleman, 2024). The amplicon between the retained BF3 and BR2 sites was extracted for each genome and written to a FASTA file, and candidate amplicons were filtered to a plausible COI length (388–424 bp) and in-frame codon length. 10,780 mitogenome records passed these filters and contain both sites, which gives forward and reverse sites for 9,396 BINs, of which 1,554 were in our dataset. When several genomes matched to one BIN, they were averaged.

The second route used full-length (658 bp) barcodes from single-specimen barcoding, which cover far more BINs, but only the forward site. The BF3/BR2 amplicon starts approximately 237 bases into the barcode region (median length 653 bases in BIOSCAN records), so the forward site lies inside a barcode record, whereas the reverse site lies beyond its 3' end. Forward sites extracted from these records cover 91.5% of the BINs in our dataset. Forward sites were located in the full-length record using the same method as was used on MitoGenomes.

From site to efficiency. Each site is scored into an efficiency between 0 and 1 by a function of its full 20-nucleotide sequence. For each retained amplicon, we estimated PCR efficiency based on the sequences of the binding sites and the primers, using the `CalculateEfficiencyPCR` function

in the R package `DECIPHER` (Wright, 2024), which calculates the Gibbs free energy, formamide melt point, and hybridization efficiency of the primer and its binding site.

Imputation. BINs without a retrieved site receive predicted efficiencies. The forward and reverse efficiency, predicted for 7.4% and 98.8% of BINs, respectively, is an equal blend of a ridge regression on the one-hot encoded amplicon and a phylogenetic kernel ridge regression whose kernel is defined by the pairwise alignment identity between amplicons. Each BIN carries its two efficiencies, the logarithm of their product and, for each primer, a flag marking efficiencies taken from a mitogenome or from a barcode record of documented provenance.

A note for sequencing practice. Primer binding sites flank the amplified fragment and are discarded by short-read metabarcoding, yet the primer binding-efficiency block is one of the two largest contributions in our ablation (Appx. P). Extending the region sequenced by single-specimen barcoding by about 50 base pairs in each direction would make this information directly available in future datasets, and as read lengths grow, the limitation may resolve on its own. In general, for effective estimation of abundance, the marker used for single-specimen barcoding should remain longer than the one used for metabarcoding, ideally capturing both the forward and the reverse binding site.

D.1 Evaluation

Reference values. Over the 9,396 BINs with a mitogenome, the forward site has a mean efficiency of 0.64 and the reverse site 0.58, and 24% and 30% of BINs, respectively, fall below 0.4. Most sites match the primer exactly (91.5% of forward and 97.3% of reverse sites), but an exact match does not guarantee efficient binding. The forward and reverse efficiencies of a BIN are nearly independent (Pearson $r = 0.15$), and efficiency differs between taxa. Among the BINs of our dataset with a retrieved forward site, the mean forward efficiency is 0.34 in spiders (Araneae) and 0.43 in Hymenoptera, with 72% and 52% of their BINs below 0.4, against 0.57 in Diptera and 0.68 in Lepidoptera. Across the 85,300 BINs with at least one specimen, $\log(F \cdot R)$ correlates with a BIN's mean empirical $\log(\text{reads/specimen})$ at $r = 0.21$ over 44,182 BINs, and at $r = 0.33$ over the 223 BINs whose two sites both come from mitogenomes.

Coverage. Mitogenomes cover only 1,554 of the BINs in our dataset (1.2%), and they are the only source of reverse sites. Full-length barcode records supply a forward site for 91.5% of BINs, which carry 97.9% of the individuals in the excess-free bottles. In total, 92.6% of BINs, carrying 99.4% of those individuals, have a forward site read from a reference sequence. The forward efficiency is imputed for the remaining 7.4% (0.6% of the individuals), and the reverse efficiency for 98.8% of BINs.

Imputation accuracy. We train the barcode-based predictor on the 7,663 mitogenome BINs absent from our dataset and on 60,000 forward sites extracted from barcode records. We then evaluated it on the 1,508 mitogenome BINs present in it. The forward predictor explains 23% of the variance of forward efficiency (Pearson $r = 0.49$, Spearman $\rho = 0.41$, MAE 0.20), slightly less than a shrunken taxonomic mean (26%). The reverse predictor explains 16% of the variance of reverse efficiency ($r = 0.41$, $\rho = 0.38$, MAE 0.21), against 9% for the taxonomic mean, and keeps its advantage on families absent from training ($R^2 = 0.17$ against 0.07). Its predictions are strongly shrunk toward the mean (standard deviation 0.09 for imputed reverse efficiencies, against 0.26 for those from mitogenomes), so an imputed value mainly carries the trend of the BIN's relatives.

E 'Excess' specimens

During routine single-specimen barcoding (SSB) workflows, technicians regularly encounter samples with many individuals of what appears to be the same species (Figure 3). In such cases, it is neither practical nor cost-effective to barcode them all. These specimens are counted and then archived, and technicians note the microplate(s) on which some representative specimens were barcoded and remark on their taxonomy. The existence of excess will always be a feature of DNA barcoding data gathered in a resource-limited environment. This *excess* is recorded as a count and a taxon label, not as BINs, so the 603 of our 1,188 bottles that contain excess lack complete per-BIN ground truth, even though their metabarcoding data include the DNA of every specimen. We evaluate only on the 585 bottles without excess; the others are used only for training, with the excess allocated to BINs as described below.

Since the excess were metabarcoded, it is essential to accurately update the BIN specimen counts to reflect it. To accomplish this, we obtained a copy of the excess database from the CBG's Specimens Unit, and filtered it to the bottles included in our dataset. To link identify the BINs associated with excess specimens, we obtained for all relevant microplates using the data previously downloaded from BOLDconnectR (Padhye et al., 2025). This BOLD-derived dataset serves as the first reference point to identify candidate BINs. Finally, the cleaned excess dataset of bottle, taxon, and count fields from the were written to a standardized output file that served as the authoritative excess-count reference for all subsequent analyses.

After assembling a clean database of all BINs on relevant microplates and a clean excess database, we identified potential candidate BINs. After refining the manually-entered taxonomic information and outdated nomenclature, we queried the unique sequences against the BOLDistilled library using VSEARCH to assign BINs. BIN assignments were then rejoined to the data by sequence identity.

To generate a candidate list of BINs consistent with each excess record, the full taxonomic hierarchy (phylum through species) associated with every BIN recovered from a given sample was reshaped into long format and matched, case-insensitively and within sample, against the corresponding excess taxon label, allowing a label recorded at any taxonomic rank to be resolved to the appropriate BIN(s). Data manipulation, reshaping, and joins throughout were performed using the `tidyverse` collection of packages (Wickham et al., 2019). Excess records for which no candidate BIN could be identified within the focal sample set were flagged and proceeded through a secondary path (detailed below), while all successfully matched sample–taxon–BIN combinations were retained as the candidate excess dataset used in downstream specimen identification.

We then used the model itself to educate our excess assignment. We used a model trained only on no-excess bottles to predict the relative abundance of taxa in bottles with excess. (In the final testing set, only bottles with no excess were included.) For each BIN in the dataset, we quantified its "model prediction factor" (MPF) as its predicted abundance divided by its true abundance, quantifying the degree to which the model over- or under-predicted abundance for that BIN in that sample. Samples with a true abundance of zero but nonzero read support were assigned the minimum nonzero true abundance observed within the sample to avoid division by zero. Within each sample–taxon group, MPFs were rescaled to relative weights that summed to 1. Excess counts were assigned to each BIN in proportion to their relative weights to produce corrected ground-truth abundances.

For taxa that did not have a candidate BIN, we assigned them by matching the excess taxon against family, order, class, and phylum—in that order of precedence until at least one candidate could be found. All BINs sharing the matched higher rank were then treated as above and the same excess allocation procedure was applied. Results were then merged with the main dataset.

F Body size estimation: pipeline and evaluation

Larger individuals contribute more DNA to the bulk DNA mixture, so the typical body size of a BIN is an important signal for our task. We estimate it in three steps, described below, and evaluate the estimates in Appx. F.4.

F.1 Segmenting body area from specimen images

During the individual barcoding pipeline, specimens are photographed individually. Following image-based estimates of arthropod size and biomass (Schneider et al., 2021), we segmented 639,751 of these photographs. Our segmentation pipeline combines pretrained vision models without fine-tuning, as follows.

1. We first locate the specimen with an open-vocabulary detector (Grounding DINO; Liu et al., 2025). **2.** The identified bounding-box is fed to two box-prompted segmentation models (SAM-HQ and SAM 3; Ke et al., 2023; Carion et al., 2026) which outline its overall shape ("envelope"; middle panels of Figure 5). We keep SAM-HQ's mask, which traces finer detail, unless the two masks disagree (IoU below 0.8), in which case we keep the one with the higher predicted IoU; if SAM-HQ misses much of the specimen, we add the large regions that only SAM 3 finds. This way we prioritize the more detailed mask while avoiding the risk of missing parts of the specimen. **3.** Next, we separate out the wings from the main body. To do this, we use two segmentation models focused on the body only: (a)

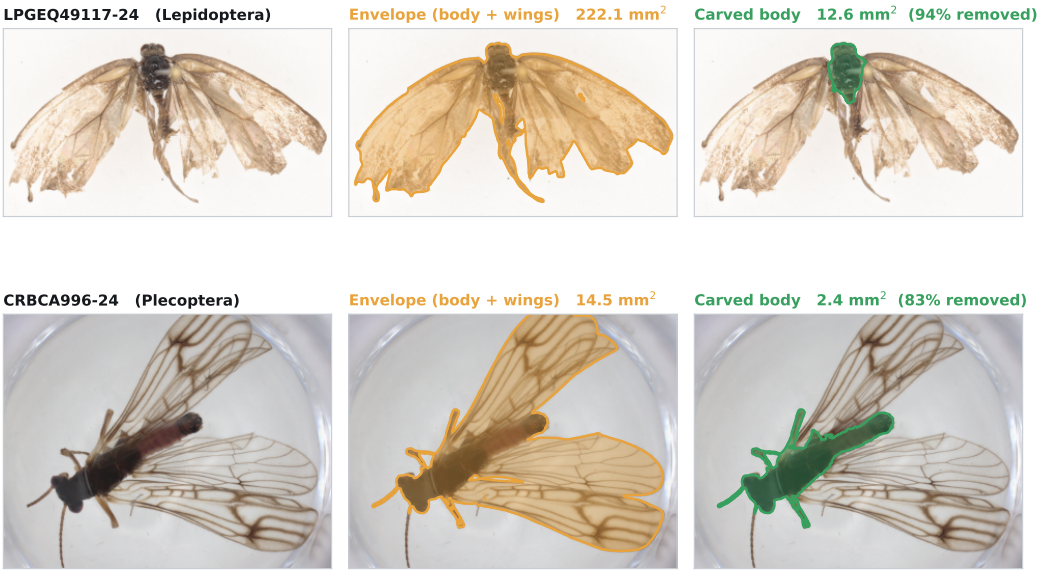


Figure 5: **Segmentation and the wing carve, on two winged specimens.** Each row is one specimen: the photograph, the envelope from the two full-organism models (SAM-HQ and SAM 3), and the carved body the pipeline keeps. Masks are drawn translucently over the photograph so the specimen stays visible underneath. LPGEQ49117-24 (Lepidoptera) goes from 222.1 to 12.6 mm^2, 94% of the silhouette removed; CRBCA996-24 (Plecoptera) from 14.5 to 2.4 mm^2, 83% removed. Areas are the calibrated values the size anchor is built from.

we ask a vision-language model (Qwen3-VL-8B-Instruct; Bai et al., 2025) for a bounding box around the compact body (head, thorax and abdomen, ignoring spread wings), which SAM 2 (Ravi et al., 2025) then segments. (b) SegGPT (Wang et al., 2023) prompted with five hand-annotated examples. Spread wings add projected area but little tissue, so we remove them only when a wing-presence score is high. The score combines the vision-language model's judgement of whether the wings are spread with the disagreement between the envelope and the body masks. We then cut from the envelope the regions that the body masks exclude: all of them if the two body masks agree (IoU at least 0.8); otherwise the larger of the two exclusions when the vision-language model or SAM 3 confirms the wings, and only the regions both exclude when neither does. The pixel count of the final mask, converted to mm^2 with the image's scale (known from the imaging system or read from a scale bar), is the specimen's projected body area, a two-dimensional projection of its body. Averaging the logarithm of this area over the images of each BIN gives a per-BIN mean log area and a within-BIN spread for the 78,259 BINs with images, which cover 50.2% of the BINs in our dataset.

Figure 5 shows two examples of specimens with spread wings. The envelope (middle panels) is what a general-purpose segmenter returns: the organism's outline, wings included. It is a faithful silhouette of the specimen as a whole, however it is a poor proxy to use for its mass. In particular, the wings provide a large cross-section of the image—for winged insects this is typically most of the surface area of the insect seen within the image—but contain very little tissue mass. We refine this segmentation mask The carve keeps the envelope's boundary, only where the two body-only models agree that the pixel is not wing. How much this changes is not a detail: across the 631,161 scored specimens, 22% carry spread wings, and on those the carve removes a median 48% of the silhouette, reaching 71% at the ninetieth percentile. Over the whole set, including specimens with folded or absent wings, the median removal is 13%.

F.2 Estimating mass from segmented image surface area

Insect mass scales as a power of 2D body dimensions (Sample et al., 1993), so log mass should be linear in log area. We calibrate this relation on a separate collection of 26,114 dried and weighed specimens from 13,287 BINs. Weighed and imaged specimens are different individuals, so the

relation cannot be fitted specimen by specimen; instead, we match the mean log mass of each genus to its mean log area, over the 332 genera with at least two specimens of each kind:

$$\overline{\log \text{mass}}_t \;=\; \alpha_t + \beta_t \, \overline{\log \text{area}}_t. \tag{13}$$

The offset α_t is estimated hierarchically from class down to genus, each level shrunk toward its parent. The slope is shared across taxa, except in the orders with enough genera to estimate their own. Applying Equation (13) to every imaged BIN gives a per-BIN dry mass.

F.3 Generalizing to unseen BINs

Half of the BINs in our data, and any BIN new to a monitoring program, have no image. To estimate their size as well, we fine-tune BarcodeBERT (Millan Arias et al., 2026) to predict, from the BIN's DNA barcode, both the BIN's mean log area and the within-BIN standard deviation of that log area, with a squared-error loss on each. Genera and BINs held out for evaluation are excluded from training (Appx. F.4). The same architecture trained on the allometric masses of step (2) predicts mass instead. Because the barcode carries nothing about an individual, its life stage or its sex, both networks predict the typical size of a taxon and the spread around it, not the size of any one specimen. The reported version of BIOTRAM reads the barcode-predicted log area and its spread for every BIN, including BINs with images; the measured areas are not among its inputs.

F.4 Evaluation

From area to mass. The fitted slope of Equation (13) is 1.37 overall and ranges from 1.27 (Coleoptera) to 1.47 (Diptera) across the orders with their own slope. All are below 1.5, the value expected if mass scaled with the cube of a linear dimension and area with its square, that is, if body shape and density did not change with size. To test the relation on taxa it was not fitted on, we predict each genus with its family's parameters withheld (five-fold cross-validation). Genus mean log mass is then predicted with Pearson $r = 0.815$ ($R^2 = 0.644$) and a mean absolute error of 0.90 natural-log units, against 1.61 and 1.77 when each genus is given the mean mass of its order or the global mean. At the level of single BINs, the mean log area and mean log weighed mass correlate at $r = 0.88$ across the 587 BINs that carry both measurements.

Body size from the barcode. On BINs held out from training whose genus is represented in training, the barcode network predicts the mean log area with Pearson $r = 0.86$ (RMSE 0.84 natural-log units), against $r = 0.82$ (RMSE 0.98) when each BIN is given the value of its nearest training barcode. On genera held out entirely it reaches $r = 0.74$ (RMSE 1.12), against 0.70 (1.28) for the nearest barcode. The predicted within-BIN spread is only weakly related to the observed one (Pearson $r \leq 0.10$ on held-out BINs), so it should be seen as a coarse taxon-level feature rather than a calibrated uncertainty.

G Detailed model architecture

Expanding on the model overview in the main text (Figure 2), we show the full model architecture in Figure 6.

Batches. A training example is a whole bottle rather than a single (bottle, BIN) row, since both the amplification module and $\mathcal{L}_{\text{comp}}$ act on a bottle's BINs jointly. A batch holds eight bottles, padded to the largest number of BINs among them and carrying a mask; padded entries take no part in the attention and are excluded from every term of Equation (3). Richness varies enough that the longest bottles are truncated: a bottle with more than 768 BINs, well above the 288 of a typical one but below the 1,534 of the largest, keeps the 768 BINs with the highest replicate read count.

H Evaluation methods

H.1 Evaluation metrics

BIOTRAM predicts a specimen count for each BIN in a bottle, and every metric derives from those same predicted counts. Some score the counts directly; others first normalize within the bottle to score the predicted composition. Since we care about the composition, about the counts themselves

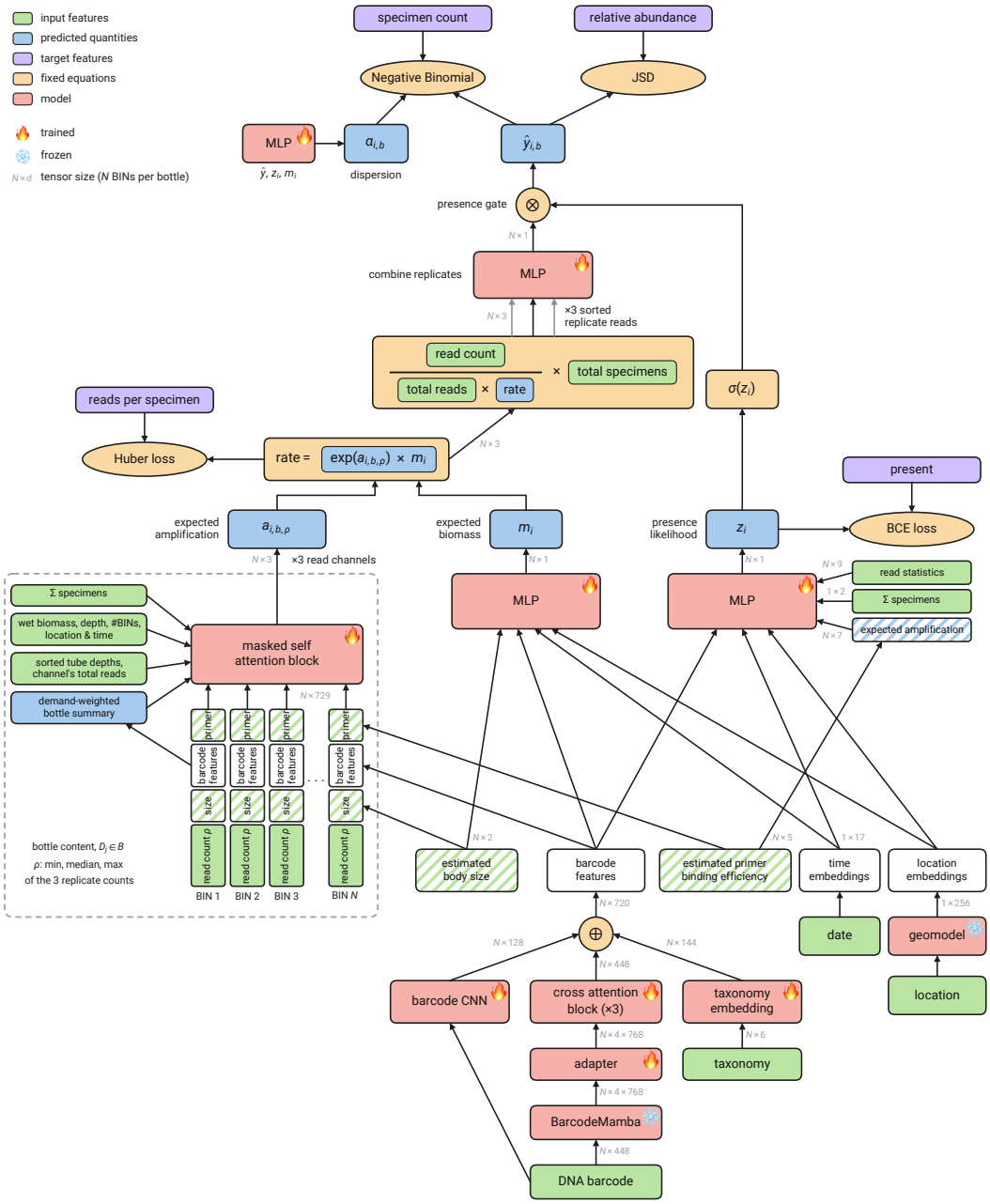


Figure 6: **Full BIOTRAM architecture.** A frozen barcode embedding, a learned taxonomy embedding and the bottle's location and date feed two per-BIN heads (the biomass $\log m_i$ and the presence logit z_i) and a masked-attention module that runs over every BIN of the bottle, once per replicate, and emits the amplification term $a_{i,b,\rho}$ from the BIN's identity, its primer and size features, its own read count and the replicate's total read count. Equation (2) then combines these quantities with the raw measured reads; the three per-replicate estimates are merged and gated by $\sigma(z_i)$.

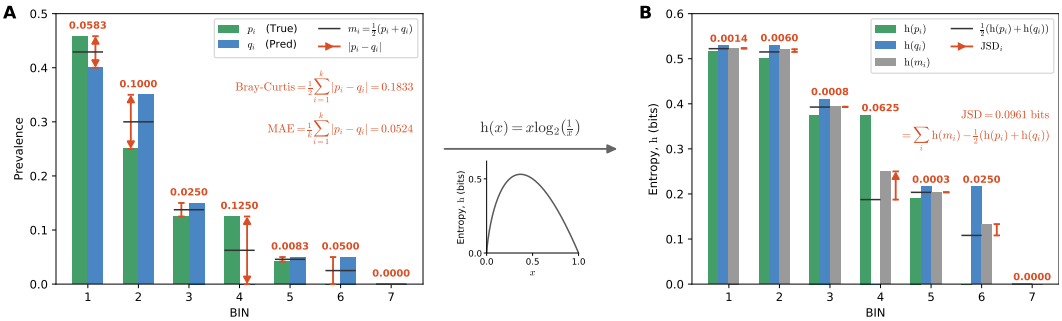


Figure 7: **Calculation of mean absolute error (MAE), Bray–Curtis, and Jensen–Shannon divergence (JSD) between a true and predicted BIN composition.** For a single bottle, the figure writes p_i and q_i for the true and predicted composition, $p_{i,b}$ and $\hat{p}_{i,b}$ in Equation (5), and m_i for their mean. **A**: Illustrative true (green) and predicted (blue) distributions over BINs. Their mean distribution is also shown (black line). Bray–Curtis is half the sum of BIN-wise absolute differences (red) between the two distributions, whereas MAE is the mean of the absolute differences. **B**: To compute JSD, we take the entropy of each BIN composition. We compare the entropy of the mean distribution (grey bars) against the mean of the entropies for the true and predicted compositions (black line). JSD is their difference (red), summed over BINs. Note that larger relative errors (BINs 4 and 6) contribute much more to the total JSD than smaller relative errors (BINs 2 and 1) despite having similar sized absolute difference between the predicted and true composition.

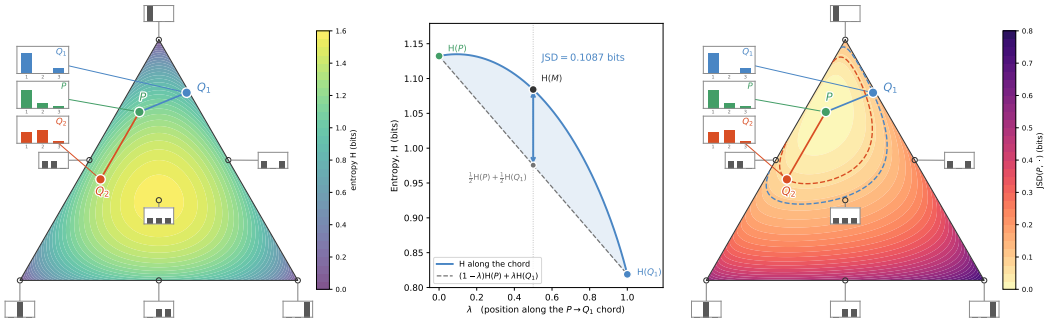
and about ecological usefulness, we report a set of metrics. Our primary metrics, JSD and MAE, are illustrated in Figure 7.

- **Jensen–Shannon divergence (JSD)**, the primary metric, defined with the composition term of the objective (Sec. 4.4). It turns predicted counts into proportions and measures how far the predicted composition sits from the true one. Figure 8 illustrates how this is calculated.
- **Mean absolute error (MAE) per bottle.** Taken within a bottle and averaged across bottles. This is the target quantity in its own units. Tables 1 and 3 also split it by truth, over all test rows: $\text{MAE}_{=0}$ on rows with no individuals, $\text{MAE}_{>0}$ on the rest, because a model can win the aggregate purely by predicting zeros well.
- **MAE on the log scale** (MAE-log), the same error on $\log$ counts. Counts span four orders of magnitude, so MAE is dominated by the abundant rows and MAE-log scores proportional rather than absolute error.
- **Presence AUC**, whether truly present BINs rank above absent ones; 0.5 is chance. The score ranked is the predicted count $\hat{n}_{i,b}$, for BIOTRAM exactly as for every baseline.
- **Spearman correlation per bottle**, whether a bottle's BINs are put in the right order; Tables 1 and 3 compute it over all test rows at once.
- **Shannon diversity** R^2, predicted against observed within-bottle Shannon diversity from single-specimen barcode data; a standard α-diversity quantity, and a direct test of whether the predicted composition is the right shape rather than merely close row by row.
- **Bray–Curtis Mantel correlation**, between-bottle dissimilarity on family-level abundances; a standard quantity in β-diversity analysis, assessing whether two bottles are correctly judged similar. It is computed at family level, because at the global scale of the dataset two bottles often share no BIN.

The last two are the quantities an ecological study would actually report, and they better represent the model's usefulness in practice.

I THE FULL BASELINE COMPARISON

Table 1 in the main text trains every deep model on the objective of Equation (3), each with the auxiliary weights tuned for it (Appx. O). Table 3 gives every deep model the same weights, those selected during development (Appx. O), so that only the architecture differs, and Table 4 lets every model use the objective it does best under, which answers the more practical question of what the



(a) Entropy of all distributions. (b) JSD between distr. P, Q_1. (c) JSD of all distr. w.r.t. P.

Figure 8: **Jensen–Shannon divergence calculation on the 3-simplex.** (a): The 3-simplex represents all possible distributions (summing to one) across three possible bins. The corners of the simplex correspond to distributions comprised of only a single bin, edges mixtures between two bins. The entropy of each possible distribution is show (heatmap), with the centre point [1/3, 1/3, 1/3] having the highest entropy and vertices [1, 0, 0], etc., the lowest. We indicate the distributions for three sample distributions, P, Q1, and Q2. (b): The Jensen–Shannon divergence between distributions P and Q1 is the difference between the entropy at their midpoint, M, and the mean of their individual entropies. (c): As with A, except the heatmap shows contours of equal JSD from distribution P. Note that distribution Q2 is closer to P than Q1 is under the JSD measurement, despite being further away in a Euclidean sense. This difference arises because Q1 underestimates bin 2 by too much; the true value of bin 2 in P is very surprising given the estimate in Q1.

Table 3: **Architecture only: every deep model at the same loss weights.** As Table 1, but every deep model is trained with the same loss weights, $(\lambda_{\text{comp}}, \lambda_{\text{pres}}, \lambda_{\text{rps}}) = (8, 0.15, 0.05)$, rather than its own tuned ones, so that only the architecture differs. The read-blind row is BIOTRAM without the BIN's own read in the amplification module (Sec. 4). The first two rows are the fixed reference predictors of Table 4, and Params is each model's number of parameters. We highlight the **best** and second best model.

Model	Params	JSD ↓	Spearman ↑	MAE/bottle ↓	$\text{MAE}_{=0}$ ↓	$\text{MAE}_{>0}$ ↓
Every taxon equally abundant	–	0.3174	–	1.219	1.000	1.723
Raw relative reads	–	0.3586	0.504	2.003	0.544	3.633
Hurdle GLMM (Furneaux et al., 2026)	–	0.1947	0.609	1.498	0.730	2.286
Transformer	9.6M	0.1257 ± 0.0025	0.737 ± 0.004	0.743 ± 0.029	0.306 ± 0.026	1.449 ± 0.084
1-D CNN	11.5M	0.1238 ± 0.0008	0.734 ± 0.002	0.759 ± 0.020	0.312 ± 0.010	1.459 ± 0.048
MLP	11.1M	0.1261 ± 0.0011	0.736 ± 0.001	0.792 ± 0.016	0.317 ± 0.013	1.533 ± 0.037
DeepSets	10.1M	0.1257 ± 0.0011	0.738 ± 0.002	0.845 ± 0.034	0.363 ± 0.020	1.642 ± 0.104
BIOTRAM	9.9M	**0.1194** ± 0.0007	0.740 ± 0.001	**0.700** ± 0.007	0.264 ± 0.008	**1.341** ± 0.016
BIOTRAM, read-blind ablation	9.9M	0.1229 ± 0.0012	0.737 ± 0.001	0.763 ± 0.017	0.286 ± 0.007	1.435 ± 0.031
TabICLv2 (in-context)	28.5M	0.1238 ± 0.0001	**0.751** ± 0.001	0.704 ± 0.003	**0.249** ± 0.002	1.413 ± 0.006

best available predictor is. The two are not comparable to one another and should not be read as a single ranking.

Architecture only. With the exact same loss, BIOTRAM's MAE per bottle is 5.8% below that of the strongest baseline, against 3.2% when each model has its own weights, and its JSD differs from the best baseline's by -0.0044, against -0.0034. Split by truth, over all rows, the MAE gain is 13.5% on rows with no individuals and 7.5% on the rest, against 5.4% and 3.9% with tuned weights. Of the four baselines, BIOTRAM is ahead of four on MAE and four on JSD; on rank correlation it is ahead of two and level with two, and on presence AUC it is level with three of them (Table 4).

The hurdle mixed model of Furneaux et al. (2026). They relate metabarcoding reads to individual counts in 45 Malaise-trap bottles with two generalized linear mixed models: a logistic regression of whether a BIN detected by metabarcoding is present as an individual, and a zero-truncated Poisson regression of how many individuals it has when it is. Both take the logarithms of the BIN's reads, pooled over PCR replicates, and of the bottle's total read count, with random intercepts nested by

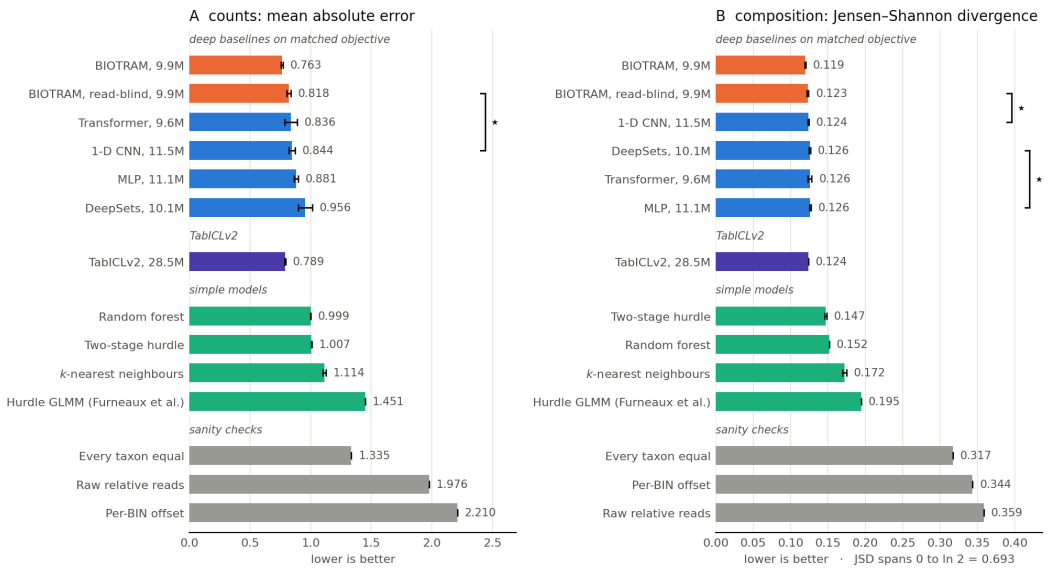


Figure 9: **The comparison of Tables 3 and 4 in bar form.** Bars start at zero, and error bars show one standard deviation across seeds. Rows are grouped because the groups are not comparable to one another: only the first holds the training objective fixed with no external model on either side, and so only it speaks to the architecture. **A starred bracket spans rows that are not distinguishable** by a two-sided Welch test over seeds at $\alpha = 0.05$; it is drawn for the matched-objective group only, since elsewhere the objective differs between rows and such a statement would not be like-for-like. On the counts BIOTRAM is clear of four of the four matched baselines and on composition of four; the read-blind variant shown beside it is clear of two on the counts and of three on composition. BIOTRAM is ahead of TabICLv2 on count accuracy, 3.2% apart, at 35% of its parameters.

order, family and genus and no effect at a rank the BIN is not identified to. Their unit of analysis, a BIN with reads in a bottle, is exactly one of our rows, so we fit both models to our training rows with the software they used (glmmTMB 1.1.13) and predict each row by the hurdle's mean, the presence probability times the truncated mean. The specification is fixed, so nothing is tuned, and the model sees only its published predictors. Under the three shifts of Table 2 it scores 0.2141 (spatial), 0.2366 (temporal) and 0.2267 (lane), behind the best matched baseline in each.

TabICLv2 and foundation-model pretraining. TabICLv2 (Qu et al., 2026) is kept apart from the other baselines in Tables 1, 3 and 4, partly because of its size: its 28.5M parameters are almost three times as many as BIOTRAM's, but also because TabICLv2 is a transformer pretrained on synthetic tabular datasets and used in context: without any gradient step on our data, it predicts each test row by attending to the training rows. It therefore brings prior knowledge from pretraining and direct access to the training set at inference, which no other model has. The pretraining matters: the same architecture trained from scratch on our data reached a JSD of 0.180, against 0.128 for the pretrained model in the same setting. TabICL has the best presence AUC (0.924), MAE-log (0.226) and $\text{MAE}_{=0}$ (0.249) of any model in our tables. BIOTRAM is level with it on count accuracy (MAE 0.699 against 0.704, 0.7% apart) and ahead of it on composition, with 35% of its parameters, while remaining behind it on $\text{MAE}_{=0}$ and rank correlation and on presence AUC (Table 1). Attending to the training rows also has a price at inference: a bottle costs TabICLv2 5.4×10^{15} floating-point operations against BIOTRAM's 1.4×10^{10}, a factor of 4×10^{5}.

J Error by abundance and by novelty

Tables 5 and 6 break the error of Table 3 down by test row: by true specimen count, and by whether the BIN occurs in any training bottle. Both compare BIOTRAM with the four matched baselines only. Table 5 scores MAE on $\log(1 + y)$, so that errors on very different counts can be compared; Table 6 scores it on the counts.

Table 4: Test set, occurrence-count target. Mean over seeds ± one standard deviation across seeds. Bold is the **best** value in each column and underline the second-best value. All models see the same rows, the same split and the same evaluation code, and are selected on validation JSD. Every deep model is at the 10.5M rung of the capacity ladder, BIOTRAM's own size.

Model	JSD ↓	MAE/bot. ↓	MAE-log ↓	Pres. AUC ↑	Spearman ↑	Shannon R^2 ↑	β Mantel ρ ↑
Reference							
Every taxon equally abundant	0.3174	1.22	0.502	0.500	–	0.627	0.724
Raw relative reads	0.3586 ± 0.0000	2.00	0.424	0.781	0.486	0.369	0.635
Per-BIN offset (McLaren et al., 2019)	0.3435 ± 0.0000	2.28	0.424 ± 0.000	0.803	0.519	0.443	0.588
Linear models							
Ridge regression	0.2159	1.25	0.424 ± 0.000	0.802	0.539	0.622	0.776
Mixed models							
Hurdle GLMM (Furneaux et al., 2026)	0.1947	1.50	0.430	0.842	0.583	0.731	0.830
Nearest neighbours							
k-nearest neighbours	0.1721 ± 0.0029	0.98 ± 0.01	0.319 ± 0.002	0.869 ± 0.001	0.626 ± 0.003	0.777 ± 0.006	0.839 ± 0.007
Tree-based models							
Random forest	0.1517 ± 0.0001	0.92 ± 0.00	0.288 ± 0.000	0.900 ± 0.000	0.669 ± 0.000	0.765 ± 0.001	0.854 ± 0.003
Gradient boosting	0.1653 ± 0.0040	0.92 ± 0.02	0.296 ± 0.002	0.884 ± 0.006	0.641 ± 0.011	0.751 ± 0.021	0.841 ± 0.005
Zero-inflated models							
Zero-inflated ridge	0.1742 ± 0.0000	0.95 ± 0.00	0.304 ± 0.000	0.889	0.646	0.680 ± 0.000	0.814 ± 0.000
Two-stage hurdle	0.1472 ± 0.0013	0.93 ± 0.00	0.290 ± 0.001	0.906 ± 0.000	0.678 ± 0.001	0.749 ± 0.004	0.852 ± 0.003
Deep models							
MLP	0.1247 ± 0.0016	0.74 ± 0.00	0.245 ± 0.003	0.916 ± 0.000	0.686 ± 0.002	0.850 ± 0.011	0.894 ± 0.005
DeepSets	0.1261 ± 0.0013	0.74 ± 0.01	0.244 ± 0.001	0.913 ± 0.000	0.686 ± 0.001	0.881 ± 0.007	0.876 ± 0.001
1-D CNN on the barcode	0.1250 ± 0.0010	0.73 ± 0.02	0.230 ± 0.004	0.916 ± 0.002	0.686 ± 0.003	0.865 ± 0.012	**0.895** ± 0.006
Transformer over the bottle	0.1256 ± 0.0008	0.71 ± 0.01	0.229 ± 0.003	0.913 ± 0.002	0.682 ± 0.002	0.886 ± 0.007	0.889 ± 0.008
BIOTRAM	**0.1194** ± 0.0007	**0.70** ± 0.01	0.232 ± 0.002	0.914 ± 0.001	0.688 ± 0.001	**0.894** ± 0.007	0.890 ± 0.006
BIOTRAM, read-blind ablation (no BIN's own read in the amplification module)							
BIOTRAM, read-blind (ablation)	0.1229 ± 0.0012	0.76 ± 0.02	0.242 ± 0.002	0.913 ± 0.000	0.686 ± 0.001	0.879 ± 0.007	0.886 ± 0.005
TabICLv2 (in-context, 28.5M)	0.1238 ± 0.0001	0.70 ± 0.00	**0.226** ± 0.001	**0.924** ± 0.000	**0.699** ± 0.001	0.820 ± 0.005	0.885 ± 0.002

Table 5: Error by true abundance, at the matched objective: MAE on $\log(1 + y)$ for BIOTRAM and for all four baselines of Table 3, each trained on BIOTRAM's own loss with the same inputs; columns ordered by overall MAE. Bold is the **best** value in each row and underline every value a two-sided Welch test over seeds cannot separate from it at $\alpha = 0.05$, or the second-best value where it separates them all.

True count	Rows	%	BIOTRAM ↓	Transformer ↓	CNN ↓	MLP ↓	DeepSets ↓
0 (absent)	14,449	53.6%	**0.173**	0.194	0.199	0.200	0.214
1	8,373	31.1%	0.214	**0.211**	0.214	0.216	0.263
2–5	3,275	12.2%	0.428	0.416	0.416	0.416	**0.388**
6–20	677	2.5%	0.681	0.663	0.694	0.710	**0.627**
21–100	146	0.5%	0.609	0.631	0.701	0.672	**0.568**
100+	14	0.1%	**0.415**	0.446	0.542	0.635	0.498

By abundance. BIOTRAM's structure acts most directly at the two ends of the range. Rows with no individuals, the most frequent bucket, are the ones the presence gate of Equation (2) can remove outright. On counts above 100, BIOTRAM's error is 7% below that of the Transformer, the matched baseline with the lowest overall MAE, a difference within seed noise. Few rows fall in this bucket, but they hold a large share of the specimens, so they weigh most on any estimate of total abundance or biomass. Between 1 and 100 specimens, BIOTRAM trails at least one matched baseline on three of the four buckets.

By novelty. Because the splits are by bottle, 19% of test rows carry a BIN that occurs in no training bottle. Their per-BIN quantities cannot be learned from training rows and must be inferred from the barcode and the taxonomy. Most of these BINs are rare, so every model makes a smaller error on them than on seen BINs. On seen BINs, BIOTRAM's MAE is 0.853 against 0.941 for the Transformer. On unseen BINs, where errors in the barcode-based estimates pass directly into the count, it is behind two of the matched baselines, level with one and ahead of one.

Table 6: Test rows split by whether the BIN appeared in training, at the matched objective. MAE in occurrences, for BIOTRAM and the baselines of Table 3; marked as in Table 5.

Test rows	Rows	BIOTRAM ↓	Transformer ↓	CNN ↓	MLP ↓	DeepSets ↓
Seen in training	21,903	**0.853**	0.941	0.955	1.004	1.068
Never seen	5,031	0.373	0.376	0.357	**0.343**	0.467

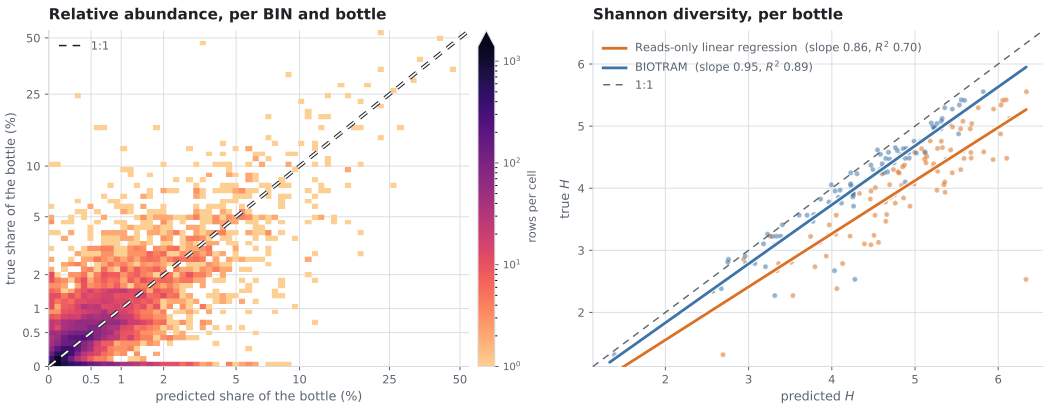


Figure 10: **Predictions, row by row and bottle by bottle.** One seed (200) of the reported arm on the test split: 26,934 rows across 89 bottles. **Left:** relative abundance per BIN and bottle, predicted against true, as a two-dimensional histogram on a $\log(1+x)$ axis in percent; colour is the number of rows in a cell, on a logarithmic scale, and the dashed line is 1:1. The cell at the origin holds every row that is zero in both, more than half the dataset; the colour scale is pinned to the largest cell beside it, so that one saturates (the arrow on the bar) rather than pushing the rest into the dark end. **Right:** Shannon diversity per bottle for BIOTRAM (blue, slope 0.95, R^2 0.89) and for an ordinary least-squares regression of the specimen count on log read count and log bottle total (orange, slope 0.86, R^2 0.70), each with its own fitted line against the 1:1 diagonal.

K What a prediction looks like

Figure 10 shows the predictions themselves for one seed of the reported arm. The left panel groups every test row by prediction and truth. The quantity is each BIN's share of the bottle it is in. The right panel aggregates each bottle to its Shannon diversity and compares BIOTRAM against a linear regression that sees only read counts, the comparator of Appx. I; each carries its own least-squares fit. The slope is the part worth reading: an estimator can have a respectable R^2 while systematically flattening diversity toward the mean, so we are looking for the model closest to the 1:1 line.

Calibration of the predictive distribution. The negative-binomial head gives a full predictive distribution for each count, not only a point estimate, so we check and score the distribution itself on the held-out rows over 12 seeds. For a discrete target, the plain probability integral transform $P(X \le y)$ is biased upward when about half of the rows are exactly zero (its mean here is 0.804), so we use its randomised mid-point version, whose mean is 0.500 under calibration. It averages 0.520: the predictive distribution is centred close to the observations. It is too wide, however, as nominal 90% intervals contain 0.967 of the observations. Part of this excess is unavoidable: on counts that are mostly zero or single-digit, a discrete interval cannot cover exactly 90%.

L Distribution shift

Regimes. Each shift regime of Table 2 replaces the random split by one that holds out a whole stratum of bottles for testing: the bottles of a region never sampled in training (spatial), those collected after every training bottle (temporal), or those of a sequencing batch absent from training (lane). Every deep model is retrained on the remaining bottles with the matched objective, so the rows of Table 2 read against Table 3.

Table 7: **Extended results on external Collaterals dataset.** 15 Malaise-trap bottles from a different country, processed by a different laboratory, with one PCR replicate per bottle and no bottle-level covariates (Appx. M). Lysis (L) and homogenate (H) are two DNA extractions of the same bottles, scored separately. JSD is the mean over seeds ± one standard deviation across seeds; MAE and presence AUC are the seed-averaged prediction. Every model sees identical rows. Extends bottom half of Table 2. Bold: **best** model; underlined: second-best.

	JSD ↓		MAE ↓		Presence AUC ↑	
Model	L	H	L	H	L	H
Read share (sanity check)	0.277	0.237	3.46	2.94	**0.772**	**0.839**
Uniform (sanity check)	0.255	0.255	3.23	2.98	0.540	0.504
BIOTRAM	0.219 ± 0.019	0.218 ± 0.015	2.94	2.60	0.722	0.740
BIOTRAM trained one replicate at a time	0.224 ± 0.024	0.217 ± 0.023	3.14	2.56	0.745	0.760
Transformer over the bottle	0.220 ± 0.018	0.229 ± 0.022	2.82	2.41	0.730	0.723
DeepSets	0.290 ± 0.020	0.296 ± 0.013	3.22	2.69	0.704	0.695
1-D CNN on the barcode	0.301 ± 0.007	0.321 ± 0.007	2.60	2.38	0.681	0.658
MLP	0.308 ± 0.005	0.335 ± 0.009	2.70	2.42	0.692	0.650
TabICLv2 (in-context) (Qu et al., 2026)	0.223 ± 0.016	0.232 ± 0.014	**2.46**	**2.24**	0.735	0.739
Hurdle GLMM (Furneaux et al., 2026)	**0.203**	**0.197**	4.00	3.75	0.770	0.823

Reading the table. Each row compares BIOTRAM with the best baseline of that row alone, chosen among the four matched deep baselines and the hurdle mixed model, so different rows can name different models: the best baseline in distribution need not be the best one under a given shift. The baseline of a row is chosen on JSD, and its MAE per bottle is shown beside BIOTRAM's. Δ is each model's change in JSD from its own random-split score (Tables 3 and 4), so the Δ of a row's baseline is not measured from the Random row whenever the two rows name different models.

Results. The shifts cost every model far more than the differences between models: BIOTRAM loses +0.0419, +0.0662 and +0.0517 JSD on the spatial, temporal and lane regimes, while its margin over the best baseline of each regime is -0.0090, -0.0099 and -0.0047, each tested with the same two-sided Welch test over seeds ($\alpha = 0.05$) as in Sec. 5. The hurdle mixed model of Furneaux et al. (2026) scores 0.2141, 0.2366 and 0.2267 JSD under these three regimes, behind the best matched baseline in each. On counts, BIOTRAM has the lower MAE per bottle in every row of Table 2, including the collaterals dataset, where the hurdle mixed model is closer on composition.

M Validation on external "collaterals" dataset

Our main results are evaluations using splits of data collected from one metabarcoding pipeline in one laboratory only. We also scored model performance on a "collaterals" dataset, comprised of fifteen Malaise-trap bottles from another country (Sweden), processed by a different laboratory. The dataset is unpublished and was shared by collaborators, whom we cannot name without breaking anonymity; we will identify its source and release it after the review period. Each bottle was metabarcoded in a single PCR replicate, without location, date or wet-biomass records, from two DNA extractions of the same material scored separately: a mild lysis (L), as in our data, and a homogenate (H). The specimens were also barcoded individually, so each bottle's true composition is known. Table 7 compares BIOTRAM, one of its variants, the four matched deep baselines, TabICLv2 and the hurdle mixed model of Furneaux et al. (2026) on these bottles; with fifteen of them, only large differences are interpretable.

Scoring. 534 of the BINs are absent from the 74,365-BIN vocabulary the models were trained on: they receive BarcodeMamba embeddings, taxonomy mapped through the training vocabulary, and body-size and primer-efficiency predictions from the same models as every other BIN (Sec. 4.1).

Results. Assigning each bottle's specimens in proportion to its reads gives a JSD of 0.277 on L and 0.237 on H, and spreading them uniformly over the detected BINs gives 0.255. BIOTRAM reaches 0.219 and 0.218 over twelve seeds, and its seed-averaged prediction is closer to the truth than the read shares in thirteen and twelve of the fifteen bottles. By a two-sided Welch test over seeds, it is level with the Transformer and TabICLv2 on both extractions and ahead of DeepSets, the 1-D CNN and

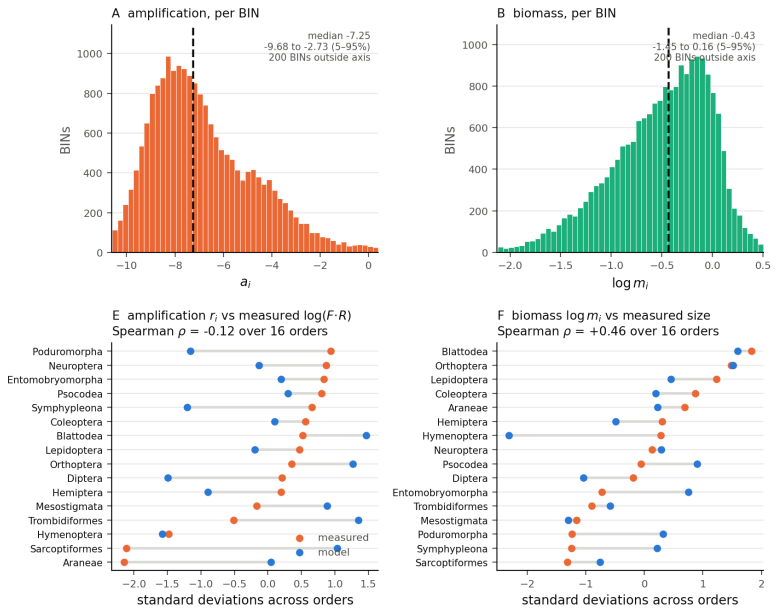


Figure 11: **The latent quantities of the reported model. Top:** their distributions across the test set, so the values the model actually emits are visible rather than only their rank agreement with an external table; axes are clipped to the central 99% and the count outside is stated. **Bottom:** the same quantities against the external measurements, aggregated to taxonomic order and expressed in standard deviations across orders so that both sit on one axis.

the MLP, which do worse than the read shares. It is behind the hurdle GLMM on composition (0.203 and 0.197) but clearly ahead on count error (MAE 4.00 and 3.75 against BIOTRAM's 2.94 and 2.60). With fifteen bottles that ordering is not firm: a bootstrap over bottles puts the mixed model's margin over BIOTRAM at $p = 0.069$ and $p = 0.045$. A variant trained with each PCR replicate as a separate sample, the format of that was used with this dataset, is level with BIOTRAM (0.224 and 0.217).

N THE LATENT QUANTITIES

Do the predicted quantities mean what they are named? The model's distinguishing property is its intermediate quantities that are supposed to correspond to physical quantities and give meaningful explanations on the model's prediction. Since the primer efficiency and body mass receive the corresponding external estimates as input features, we can't correlate them against those estimates without circularity. Instead, we can check that the empirical per-BIN $\log(\text{reads/specimen})$, measured from the metabarcoding pipeline correlates with the predicted $a_{i,b} + \log m_i$ (Equation (1)). The module also sees the BIN's own read, so part of that association is mechanical rather than discovered, and the read-blind variant is the cleaner test of the same question.

Over the reported model's 12 seeds it reaches Pearson $+0.627 \pm 0.007$, against $+0.593 \pm 0.013$ for $a_{i,b}$ alone and $+0.444 \pm 0.025$ for $\log m_i$ alone. The read-blind variant, whose amplification term never sees the read, still reaches $+0.461 \pm 0.013$ ($+0.397$ and $+0.411$ for the two terms alone, 12 seeds). Figure 11 shows the distribution of both quantities across the test set and their agreement with the external measurements at the level of taxonomic order. Note that the objective supervises $a_{i,b} + \log m_i$ toward this quantity at weight 0.05, so the number is a generalization measure over held-out bottles rather than an unsupervised discovery; and deleting that term outright moves validation JSD by $+0.0000$ (paired $t = +0.0$ across seeds; Table 9), which bounds how much of the association it can be creating.

Table 8: **Every model at its own loss weights.** The three auxiliary weights of Equation (3) are tuned separately for each model, with the same protocol and budget (Appx. O): every model is given the same 61 configurations, plus up to two additional steps for any weight whose best value lies at the edge of the explored range, each trained on the same five seeds. Test metrics at the selected weights, mean ± standard deviation over eight seeds disjoint from the search; the test set played no part in selection. We highlight the **best** and second best model.

Model	Budget configs	Selected weights			Test		
		λ_{comp}	λ_{pres}	λ_{rps}	JSD ↓	MAE/bottle ↓	AUC ↑
Transformer	63	1	0	0.1	0.1257 ± 0.0019	0.723 ± 0.020	0.9156 ± 0.0017
1-D CNN	61	2	0.009375	0.2	0.1235 ± 0.0008	0.762 ± 0.016	0.9145 ± 0.0008
MLP	63	1	0	0.00625	0.1234 ± 0.0006	0.737 ± 0.025	0.9154 ± 0.0007
DeepSets	65	1	0.0375	0.00625	0.1256 ± 0.0010	0.752 ± 0.036	0.9162 ± 0.0013
BIOTRAM	62	1	0.15	0	**0.1200** ± 0.0009	**0.699** ± 0.017	**0.9213** ± 0.0006

O LOSS-WEIGHT TUNING

Principle. A baseline comparison is only fair if every model's objective is tuned with the same care, since an under-tuned baseline makes the proposed model look better than it is (Xin et al., 2022). The three auxiliary weights $(\lambda_{\text{comp}}, \lambda_{\text{pres}}, \lambda_{\text{rps}})$ of Equation (3) are therefore tuned for each of the five models separately, under one protocol fixed before any run and with the same budget. The likelihood is the negative binomial of Equation (4) for every model.

Search. Every model runs the same four stages. (i) A full 3^3 factorial over the three weights, on the logarithmic grid $\lambda_{\text{comp}} \in \{2, 8, 32\}$, $\lambda_{\text{pres}} \in \{0.0375, 0.15, 0.6\}$ and $\lambda_{\text{rps}} \in \{0.0125, 0.05, 0.2\}$. (ii) At the best of these, each weight set to zero in turn, and KL in place of JSD for the composition term. (iii) A second 3^3 factorial at half the step, around the best configuration so far; a weight set to zero in (ii) is tested at zero against two values below the smallest one of (i), so that zero is found to be optimal rather than assumed. (iv) The learning rate at a third of and at three times the model's tuned value. Stage (iii) is a factorial rather than one weight at a time because the weights interact: four of the five models have a two-way interaction that is significant in a factorial ANOVA of stage (i), which a search moving one weight at a time can walk past. A best configuration on the edge of the range explored for a weight receives up to two further steps outward, under the same rule for every model.

Budget. Every model is given the same 61 configurations, plus between zero and four boundary steps granted by the rule described above: 61 to 65 configurations per model. BIOTRAM's 62 configurations lie inside the range of the baselines (Table 8). Each configuration is trained on the same five seeds, 1,650 runs in all. The budget is counted in configurations and seeds.

Selection. A configuration is scored on validation JSD, MAE and presence AUC together. Each metric is divided by its own seed-to-seed standard deviation, pooled over the stage-(i) factorial, so that the three enter on a common scale, and they are averaged per seed and then over seeds into a composite S; the configuration with the lowest S is selected. A configuration is admissible only if none of the three metrics is worse than at the centre of the stage-(i) grid by more than one noise band, so the composite cannot trade one metric for another.

Confirmation. The selected weights of each model are retrained on eight seeds disjoint from every seed used in the search, and from those on which the model's architecture was selected. The test set plays no part in any selection; Table 8 and Table 1 in the main text report it on these retrained models.

Result. With every model at its own weights, BIOTRAM is ahead of the best baseline on all three test metrics: JSD 0.1200 against 0.1234 for the MLP ($p = 8.4\times10^{-7}$), MAE per bottle 0.699 against 0.723 for the Transformer ($p = 0.025$), and presence AUC 0.9213 against 0.9162 for the DeepSets model ($p = 1.6\times10^{-6}$; two-sided Welch tests over seeds). The selected weights differ between architectures (Table 8): the MLP and the Transformer do best with no presence term, zero beating both smaller values, whereas the 1-D CNN keeps a small one, and BIOTRAM keeps its presence term and does best without $\mathcal{L}_{\text{rps}}$. No single weighting suits every model, which is why each is tuned separately. Everywhere else in the paper, except in Table 1, BIOTRAM and the matched baselines use the weights selected during development, $(\lambda_{\text{comp}}, \lambda_{\text{pres}}, \lambda_{\text{rps}}) = (8, 0.15, 0.05)$, rather

than those tuned here. BIOTRAM's tuned weights reach 0.1200 test JSD, slightly behind the 0.1194 its development weights reach in Table 3. That is expected rather than contradictory: a configuration is selected on a composite of validation JSD, MAE and presence AUC, never on test JSD, so the returned configuration need not be the one with the lowest test JSD — and the comparison in this appendix holds every model to that same rule.

P Ablation study

We ran a hierarchical ablation over 30 capability blocks (Table 9). Each block is removed from the reported configuration on its own and the model retrained, paired seed-for-seed against a control trained on the identical split; the screen runs at four seeds and survivors are confirmed at eight. Each t below is the paired t statistic of the change in validation JSD, the mean per-seed difference from the control divided by its standard error. The control scores 0.1239 validation JSD, and a block is load-bearing when removing it costs more than the noise band ($\pm$0.0021 JSD, $\pm$0.0200 relative error) on either metric. Reading the three groups of Table 9 together is the point of the study: most of the model is individually free to remove, a handful of blocks carry it, and two blocks that the criterion does not call load-bearing cost a large part of the presence AUC.

What carries the model. Of the 30 blocks, 9 are load-bearing. The two largest effects are data and chemistry rather than architecture: the extra training bottles with excess (+0.0148 validation JSD, $t = +25.9$) and the primer binding-efficiency block (+0.0147, $t = +60.5$). The amplification module itself is next (+0.0072, $t = +6.4$), and giving it the BIN's own reads is worth +0.0049 ($t = +6.3$) on its own.

What is free to remove. Every other block costs less than the noise band when removed. A few sit in a grey zone, where the seeds separate the change from zero ($|t| > 2$ for a paired t-test) but the band's floor decides the verdict; the rest are removable within noise, including the pretrained geographic encoder, the augmentation engine, per-BIN dispersion and the frozen-encoder choice (we tested BarcodeBERT and BarcodeMamba). The exceptions are the two forms of presence supervision: removing either stays inside the band on JSD and relative error but costs a large part of the presence AUC, which the criterion does not score.

Identity is redundant. Removing the frozen barcode while leaving the taxonomy embedding in place costs +0.0002 validation JSD ($t = +1.1$), inside the noise band; removing the taxonomy embedding while keeping the barcode costs +0.0017 ($t = +2.0$). Neither identity channel is individually load-bearing on this configuration, which is consistent with each being able to stand in for the other: genus is unknown for 76.9% of BINs and species for 87.5%, so for three quarters of the BINs a taxonomy lookup resolves only to family or coarser, yet what it does resolve is apparently enough to cover most of what the barcode carries.

Anchors as inputs. Supervising the body mass and primer binding efficiency heads directly toward the external estimates, instead of feeding those estimates as features the heads may correct, stays within noise on both validation JSD and MAE. We keep them as inputs, which lets the model discount them where the data disagree.

Q Model scaling against baselines

Capacity. Model capacity is not what separates the models. Every architecture is retrained at each rung of a capacity ladder, 200k to 128M parameters for the baselines and 3.2M upward for BIOTRAM, and tuned again at every rung (Figure 12). Capacity moves each of them by at most 0.0062 JSD across its own ladder, while BIOTRAM's margin over the best baseline of the same rung stays between 0.0018 and 0.0052: BIOTRAM at its worst size (0.121) is still ahead of the 1-D CNN at its best (0.123).

Compute. Figure 12 also places the same models against the cost of training one model and of predicting one bottle, both counted in floating-point operations rather than in wall clock, so that they do not depend on the hardware. At the 10.5M rung BIOTRAM trains for 4.4$\times$ the compute of the deep baselines. Predicting one bottle costs it 1.4×10^{10} operations, inside the range the deep baselines span at that rung (6.1×10^{9} to 9×10^{9}), so cost per prediction does not follow the parameter count.

Table 9: **Model component ablations.** All 30 component ablations, each removed from the full configuration on its own, ranked by what the removal costs in validation JSD (anchor: 0.1239). Δ is the change from the full configuration, so a positive Δ JSD or relative error, or a negative Δ AUC, means the component helps. Components are ranked by Δ JSD on the validation set, the split the study selects on; the relative error, $\sum |n - \hat{n}| / \sum n$, and the AUC of the presence probability $\sigma(z)$ are measured on the test set. t is the paired t statistic of Δ val JSD (the mean per-seed difference from the control divided by its standard error) across seeds and p its two-sided p-value, uncorrected ($^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$). The noise band, ± 0.0021 JSD and ± 0.0200 relative error, is two standard errors of the difference of seed means, floored at a fixed fraction of the anchor. Colour marks the three groups: load-bearing, removal costs more than the band on either metric; grey zone, inside the band but with $|t| > 2$, so only the band's floor keeps it from being load-bearing; and removable, inside the band with $|t| \leq 2$. N/A: removing this block swaps the head to predicting relative abundance, so its count-space error is against a different quantity and is not comparable. ‡: within the noise band on both gated metrics while costing a large part of the presence AUC, which the gate does not score.

	Component removed	Δ val JSD ↓	t	p	Δ relative error ↓	Δ AUC of $\sigma(z)$ ↑
Load-bearing: removal costs more than the noise band						
Data	Train on bottles with excess (GT partially approx.)	+0.0148	+25.9	$< 0.001^{***}$	+0.0203	−0.0055
Arch	Primer binding-efficiency block	+0.0147	+60.5	$< 0.001^{***}$	+0.0254	−0.0211
Input	Learned components on top of the frozen encoder	+0.0080	+6.5	0.007^{**}	+0.0391	−0.0068
Input	Cross-BIN amplification module	+0.0072	+6.4	0.008^{**}	+0.0421	−0.0047
Input	BIN's read count as module input	+0.0049	+6.3	0.008^{**}	+0.0334	−0.0038
Regul	Exponential moving average of the weights	+0.0035	+3.4	0.043^{*}	+0.0352	−0.0041
Input	Bottle occurrence total as a log-offset	+0.0030	+4.1	0.026^{*}	+0.0596	−0.0080
Loss	Count objective (against a compositional head)	+0.0029	+9.2	0.003^{**}	N/A	+0.0017
Loss	Composition auxiliary	+0.0027	+5.3	0.013^{*}	+0.0235	−0.0009
Grey zone: inside the noise band, but $\|t\| > 2$						
Input	Learned taxonomy embedding	+0.0017	+2.0	0.139	+0.0119	−0.0024
Input	Barcode-predicted body size	+0.0016	+12.8	0.001^{**}	+0.0037	−0.0005
Regul	Training budget of 100 epochs (50 when removed)	+0.0015	+3.7	0.034^{*}	+0.0013	−0.0005
Input	Per-replicate sequencing depth	+0.0013	+3.1	0.052	+0.0072	+0.0000
Removable: inside the band and $\|t\| \leq 2$						
Input	Bottle and primer features in the presence head	+0.0015	+1.6	0.208	+0.0032	−0.0047
Arch	Replicate merge as a residual	+0.0015	+0.9	0.438	+0.0178	−0.0030
Loss	Presence supervision‡	+0.0011	+1.2	0.319	+0.0038	−0.0523
Arch	Presence gate applied after the merge	+0.0009	+0.8	0.497	+0.0053	−0.0004
Loss	Learned Dirichlet-multinomial concentration	+0.0009	+1.0	0.393	−0.0001	−0.0001
Loss	Supervising the anchors instead of feeding them	+0.0008	+0.8	0.497	−0.0060	+0.0006
Arch	Per-tube and per-read channels in the merge	+0.0007	+1.1	0.343	+0.0056	+0.0006
Arch	Quantile read-out around the rate	+0.0006	+0.6	0.619	+0.0016	+0.0003
Input	Replicate order statistics	+0.0006	+0.5	0.652	+0.0095	+0.0014
Loss	Per-BIN dispersion	+0.0005	+0.6	0.568	+0.0016	+0.0006
Arch	De-leak correction on the bottle offset	+0.0004	+0.7	0.550	+0.0071	+0.0001
Input	Frozen barcode embedding‡	+0.0002	+1.1	0.342	+0.0062	+0.0006
Arch	Presence supervision‡	+0.0002	+0.2	0.829	+0.0108	−0.0835
Input	Pretrained geographic encoder	+0.0001	+0.1	0.939	−0.0043	+0.0007
Loss	RPO intermediate supervision	+0.0000	+0.0	0.999	+0.0084	+0.0011
Regul	Online augmentation engine	−0.0005	−0.8	0.505	+0.0134	+0.0015
Input	Frozen encoder choice (BarcodeMamba vs BarcodeBERT)	−0.0009	−1.1	0.369	+0.0098	−0.0014

TabICL has no training cost on our data and instead attends to every training row for each prediction, at 5.4×10^{15} operations per bottle.

Training-set size. Every model is also retrained on nested subsets of the 864 training bottles: the first k bottles of one fixed random permutation, for six values of k from 54 to 864, so that each subset contains the smaller ones and all models at a given k see the same bottles. Validation and test bottles are unchanged, and everything fitted on the training data, such as the feature standardisation or TabICLv2's context, is refitted on the subset. Each model keeps a single size and the configuration tuned on the full training set (Figure 13). Every curve is still falling at the full training set, and BIOTRAM's most steeply, from a JSD of 0.1841 to 0.1198: it is behind every other model up to 216 bottles, and from 648 bottles on it is ahead of the four deep baselines, level there with TabICLv2 and ahead of it on the full training set as well (0.1198 against 0.1227, $p = 0.001$). Its structure therefore does not make BIOTRAM more data-efficient at this size; it pays off once there are enough bottles to learn the per-BIN quantities it relies on.

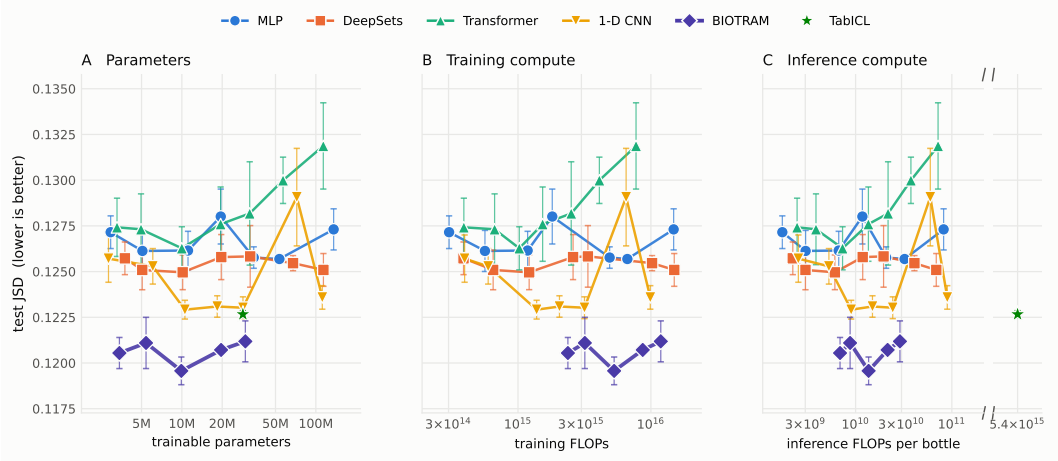


Figure 12: **Accuracy against cost.** Test JSD of every architecture at every rung of the capacity ladder, against the number of trainable parameters (**A**), the compute of one training run (**B**) and the compute of one prediction (**C**). FLOPs are measured on the real forward and backward passes. The baselines are carried to 128M parameters, past BIOTRAM's own inference and training compute, and stop improving well before that. TabICLv2 is fitted in context and has no training cost on our data; its inference cost is roughly 4×10^5 times BIOTRAM's.

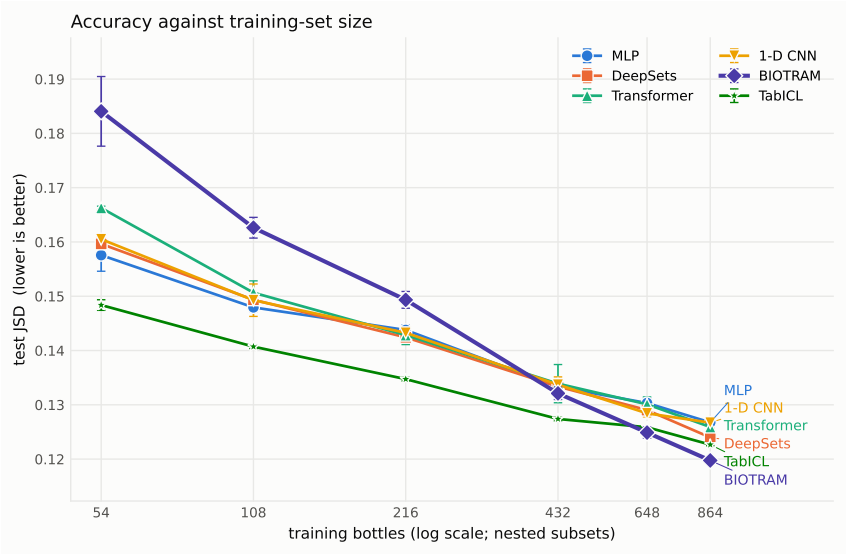


Figure 13: **Accuracy against the number of training bottles.** Test JSD of every model retrained on nested subsets of the training bottles (log scale), each with the configuration tuned on the full training set and at a single size: BIOTRAM's own for the four deep baselines, as in Table 4, and 28.5M parameters for TabICLv2. Error bars show one standard deviation across three seeds. With all 864 bottles, BIOTRAM reaches 0.1198 against TabICLv2's 0.1227 ($p = 0.001$).

R How a read reaches the count

Read elasticity. We measure how a change in a BIN's reads propagates to its predicted count, as the elasticity $\partial \log \hat{n}_{i,b} / \partial \log \mathrm{reads}_{i,b}$. In the read-blind variant, Equation (2) passes a change in reads to each replicate estimate one for one (measured: 1.00), and the learned replicate merge takes back 0.60, for an end-to-end elasticity of 0.40 (3 seeds). The merge sees only the three replicate estimates and the bottle's specimen total, so this discount cannot depend on the bottle's total read count or on the other BINs. In the reported model, the amplification term also responds to the BIN's own read: its

derivative with respect to the BIN's own read inputs (log read count and read share) has a per-seed mean absolute value between 0.37 and 0.43 over 12 seeds, at least 190 times its derivative with respect to another BIN's read (0.002). On average, when a BIN's reads double, its predicted amplification rises with an elasticity of 0.29, so part of the extra reads is attributed to stronger amplification rather than to more individuals. The replicate estimate then grows with an exponent of 0.71 instead of one, the merge takes back 0.40, and the end-to-end elasticity is 0.31, below the read-blind variant's 0.40. With the read in view, the discount is larger and can vary between BINs and with the bottle's total read count. Every baseline can fit such an interaction, since each row it receives carries the log read count, the BIN's share and the bottle's total read count; the read-blind variant gives it up by construction.

What the per-BIN read contributes. Estimating the amplification without the BIN's own read keeps the two sides of Equation (1) strictly separate, at a cost of +0.0036 JSD (Table 3). The read-blind variant is still ahead of two of the matched baselines and level with two on MAE per bottle, and ahead of three and level with one on JSD.

Attention across BINs. The amplification module is the only component in which BINs attend to one another; it was designed to express competition between BINs for shared reagents and sequencing budget. To test whether it uses this attention or mainly the BIN's own read, we make five comparisons, each paired over six seeds. Three compare the module with one of identical width, depth and feed-forward structure in which BINs do not attend to one another, varying which reads both see; every variant receives the total read count of the three replicates. The other two remove that total: one repeats the first comparison without it, the other measures what it adds to the module without attention. When the BIN's own read and those of the other BINs are visible, attention lowers JSD by 0.0033 (6/6 seeds, $p = 0.0067$; each p in this paragraph is a paired t-test, Holm-corrected across the five paired comparisons). Masking only the BIN's own read removes this gain (-0.0004 JSD, 3/6 seeds, $p = 1$), and with no reads at all, attention has no measurable effect (3/6 seeds, $p = 1$). Without the replicate total read count, attention over BINs with reads changes JSD by -0.0020 (6/6 seeds, $p = 0.045$), and the total read count without attention changes it by -0.0001 (3/6 seeds, $p = 1$). Attention therefore pays off only when the module sees the BIN's own read, and most when it also sees the replicate's total read count, which does nothing without attention. This suggests that it exploits an interaction between the per-BIN reads and the total reads rather than competition between taxa.