Multi-label pathway prediction based on active dataset subsampling

Abdur Rahman M. A. Basher and Steven J. Hallam

September 14, 2020

[1] Graduate Program in Bioinformatics, University of British Columbia, Genome Sciences Centre, 100-570 West 7th Avenue, Vancouver, British Columbia V5Z 4S6, Canada.
[2] Department of Microbiology & Immunology, University of British Columbia, 2552-2350 Health Sciences Mall, Vancouver, British Columbia V6T 1Z3, Canada.
[3] Genome Science and Technology Program, University of British Columbia, 2329 West Mall, Vancouver, BC V6T 1Z4, Canada
[4] Life Sciences Institute, University of British Columbia, Vancouver, British Columbia, Canada V6T 1Z3
[5] ECOSCOPE Training Program, University of British Columbia, Vancouver, British Columbia, Canada V6T 1Z3
∗ To whom correspondence should be addressed

Abstract

Machine learning methods show great promise in predicting metabolic pathways at different levels of biological organization. However, several complications remain that can degrade prediction performance including inadequately labeled training data, missing feature information, and inherent imbalances in the distribution of enzymes and pathways within a dataset. This class imbalance problem is commonly encountered by the machine learning community when the proportion of instances over class labels within a dataset are uneven, resulting in poor predictive performance for underrepresented classes. Here, we present leADS, multi-label learning based on active dataset subsampling, that leverages the idea of subsampling points from a pool of data to reduce the negative impact of training loss due to class imbalance. Specifically, leADS performs an iterative process to: (i)- construct an acquisition model in an ensemble framework; (ii) select informative points using an appropriate acquisition function; and (iii) train on selected samples. Multiple base learners are implemented in parallel where each is assigned a portion of labeled training data to learn pathways. We benchmark leADS using a corpora of 10 experimental datasets manifesting diverse multi-label properties used in previous pathway prediction studies, including manually curated organismal genomes, synthetic microbial communities and low complexity microbial communities. Resulting performance metrics equaled or exceeded previously reported machine learning methods for both organismal and multi-organismal genomes while establishing an extensible framework for navigating class imbalances across diverse real world datasets.

Availability and implementation: The software package, and installation instructions are published on github.com/leADS

Contact: shallam@mail.ubc.ca

1 Introduction

Metabolic pathways are composed of interconnected reactions catalyzed by enzymes. The set of reactions within and between cells comprises a reactome. Pathways and reactomes can be predicted from annotated genes encoded within organismal or multi-organismal genomes. This pathway prediction problem presents a fundamental challenge in biology that connects hereditary information contained within the DNA of living things e.g. genotype, to its expression and activity at the individual, population and community levels of organization e.g. phenotype ([29, 17, 23]). The rise of increasingly powerful sequencing technologies has motivated corresponding innovations in the methods used to predict metabolic pathways at different levels of genome complexity and completion ([1]). These encompass rule-based or heuristic methods including PathoLogic ([22]) and MinPath ([41]), and more recently, machine learning (ML) methods including PtwML ([10]), mlLGPR ([5]) and triUMPF ([4]). While ML methods overcome issues of probability and scale associated with rule-based methods, several complications remain that can degrade prediction performance including inadequately labeled training data, missing feature information, and inherent imbalances in the distribution of pathways within a dataset.

The *class imbalance* problem arises when the proportion of instances over class labels within a dataset are uneven, resulting in poor predictive performance for underrepresented classes e.g. training loss. Such skewed distributions are encountered across a wide range of real world datasets, from environmental monitoring and fraud detection to medical diagnosis and facial recognition ([19]). In the case of metabolic pathways, a similar problem exists where certain pathways are more common than others because they conduct core metabolic functions conserved across the tree of life. These functions are overrepresented in labeled training data relative to more niche-defining or accessory metabolic functions.

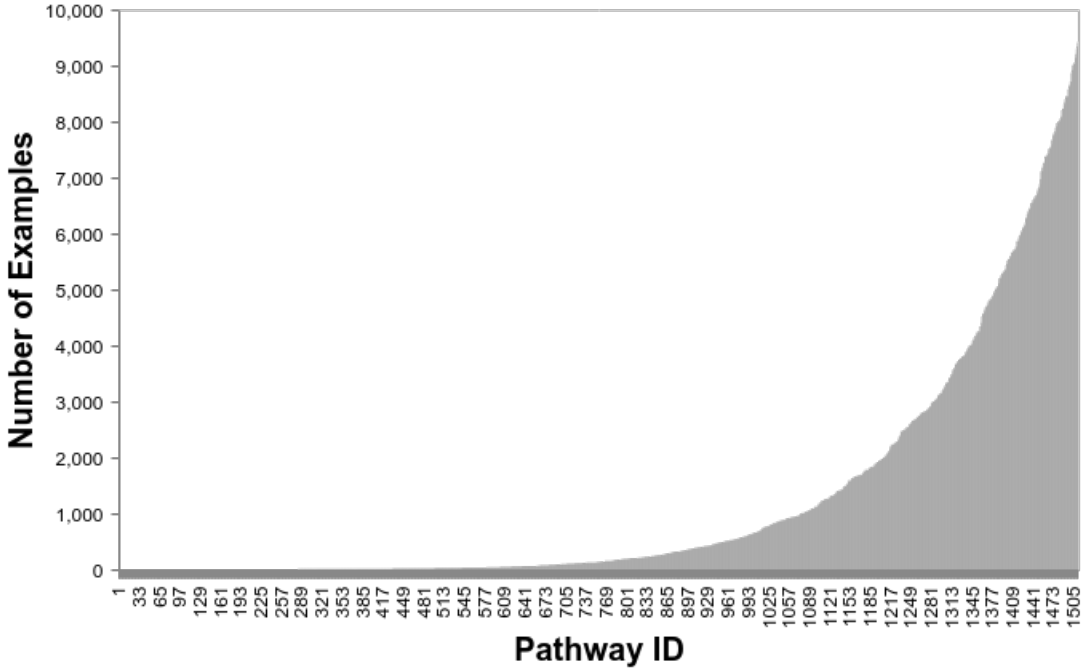


Figure 1: Number of PGDBs (or samples) for each pathway in BioCyc v21 T2-3 training data. The horizontal axis indicates the indices of pathways while the vertical axis represents the number of associated PGDBs.

Basher and colleagues described an information hierarchy based on the BioCyc curation-tiered structure of Pathway/Genome Databases (PGDBs) ([8]) that traverses four tiers of genome completion and complexity (T1-4) in descending order of curation and functional validation ([5]). Labeled pathways associated with T1-3 genomes were incorporated into synthetic datasets and used to train supervised ML pathway prediction methods ([5, 4]). During the benchmarking process class imbalances were recognized that limited recovery of underrepresented pathways in the training data. For example, labeled T2-3 pathways follow a power law distribution (Fig. 1) where $30-35\%$ of pathways were observed to occur in less than 25 PGDBs within the BioCyc collection. This class imbalance extended to closely related genotypes e.g. *E. coli* with potential implications for resolving metabolic differences between symbiotic, commensal or pathogenic strains.

Different class imbalance learning methods have been developed that take into account skewed distributions including sampling, algorithm modification and ensemble learning ([24]). Sampling methods attempt to balance input data prior to training through random under-sampling, one sided-selection or a combination of over-sampling less common classes while under-sampling more common ones ([11]). In relation to PGDBs with numerous shared pathways, noisy class labels or missing pathway information (e.g. T2-4), subsampling presents a more tractable solution than oversampling. Two distinct modes of subsampling have been developed that are effective under different training scenarios: i) incremental learning from easier to harder examples, and ii) hard example mining. While the incremental mode may be effective when learning from noisy data by gradually removing hard examples ([6, 30]), sampling hard examples directly can accelerate the learning process ([36, 26]). Given that BioCyc (T2 &3) contains more than 9000 instances (corresponding to over 1500 organismal genomes) hard example mining is expected to reduce training loss resulting from pathway class imbalance.

Here we describe leADS, multi-label learning based on active dataset subsampling, that builds on prior work in active dataset subsampling (ADS) ([9]) by incorporating an ensemble of multi-label learners ([42]) to perform hard example mining. Specifically, leADS executes, in parallel, a group of multi-label base learners (constituting an ensemble) where each is allocated to learn from a portion of randomly selected samples ([40]). Then, each member in the ensemble selects data according to predefined choices of: i)- sample size and ii)- an acquisition function. Samples from all base learners are aggregated for subsequent rounds of learning.

To verify the effectiveness of leADS, we conducted three experimental studies: parameter sensitivity, scalability, and metabolic pathway prediction. Overall, leaADS significantly improved pathway prediction results in relation to other inference methods including MinPath ([41]), PathoLogic ([22]), mlLGPR ([5]) and triUMPF ([4]) on a corpora of 10 organismal and multi-organismal datasets including T1 PGDBs from the BioCyc collection, symbiont genomes encoding distributed metabolic pathways for amino acid biosynthesis [27], genomes used in the Critical Assessment of Metagenome Interpretation (CAMI) initiative [33], and whole genome shotgun sequences from the Hawaii Ocean Time Series (HOTS) [38].

2 Definitions and Problem Formulation

Here the default vector is considered to be a column vector and is represented by a boldface lowercase letter (e.g., $\mathbf{x}$) while the matrices are represented by boldface uppercase letters (e.g., $\mathbf{X}$). If a subscript letter i is attached to a matrix, such as $\mathbf{X}_i$, it indicates the i-th row of $\mathbf{X}$, which is a row vector. A subscript character to a vector, $\mathbf{x}_i$, denotes an i-th cell of $\mathbf{x}$. Occasional superscript, $\mathbf{X}^{(i)}$, suggests an index to a sample or current epoch during a learning period. With these notations in mind, we introduce information integral to the problem formulation, starting by defining the multi-label data.

Definition 2.1. Multi-label Pathway Dataset ([5]). A pathway dataset is represented by $\mathcal{S}_A = \{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}) : 1 < i \leqslant n\}$ consisting of n examples, where $\mathbf{x}^{(i)}$ is a vector indicating the abundance information corresponding to enzymatic reactions. An enzymatic reaction is denoted by c, which is an element of a set $\mathcal{E} = \{c_1, c_2, ..., c_r\}$, having r possible enzymatic reactions, hence, the vector size $\mathbf{x}^{(i)}$ is r. The abundance of an enzymatic reaction for an example i, say $c_l^{(i)}$, is defined as $a_l^{(i)} (\in \mathbb{R}_{\geq 0})$. The class labels $\mathbf{y}^{(i)} = [y_1^{(i)}, ..., y_t^{(i)}] \in \{-1, +1\}^t$ is a pathway label vector of size t representing the total number of pathways derived from a set of universal metabolic pathway $\mathcal{Y}$. The matrix form of $\mathbf{x}^{(i)}$ and $\mathbf{y}^{(i)}$ are $\mathbf{X}$ and $\mathbf{Y}$, respectively. ■

Both $\mathcal{E}$ and $\mathcal{Y}$ can be retrieved from trusted sources, such as KEGG ([21]) or MetaCyc ([7]). Although the input space is assumed to be encoded as r-dimensional vector, symbolized as $\mathcal{X} = \mathbb{R}^r$, through features engineering it can be represented as $\mathcal{X} = \mathbb{R}^d$.

Problem Statement. Given a multi-label dataset, $\mathcal{S}_A$, the goal is to select a subset of $\mathcal{S}_A$, denoted by $\mathcal{S}_{\mathbf{per}\%}$, where $\mathbf{per}\%$ is a prespecified hyperparameter, indicating the proportion of samples to be chosen from $\mathcal{S}_A$, such that learning on $\mathcal{S}_{\mathbf{per}\%}$ incurs similar predictive score (or better) as if it was trained on full multi-label dataset, $\mathcal{S}_A$.

3 The leADS Method

In this section, we provide a description of leADS components including: i)- building an acquisition model, ii)- active dataset sub-sampling, and iii)- learning using the reduced sub-sampled data. These three steps interact with each other in an iterative process as illustrated in Fig. 2. At the very first iteration, a set $\mathcal{S}^0_{\mathbf{per}\%}$ is initialized with randomly selected data. In the next iteration q, instead of re-initializing $\mathcal{S}^q_{\mathbf{per}\%}$ with randomly selected samples, $\mathcal{S}^{q-1}_{\mathbf{per}\%}$ data collected from the previous iteration $q-1$ is used, constituting a *build-up scheme* implemented in many active learning methods ([9]). This process is repeated until the maximum number of rounds τ is reached.

3.1 Building an Acquisition Model

Given $\mathcal{S}_A$, the objective of this step is to estimate posterior predictive uncertainty given a new test point $\mathbf{x}^*$ for a pathway $\mathbf{y}_j$ as:

$$p(\mathbf{y}_j = +1|\mathbf{x}^*, \mathcal{S}_A) = \int p(\mathbf{y}_j = +1|\mathbf{x}^*, \Theta_j) p(\Theta_j|\mathcal{S}_A) \partial\Theta_j \tag{3.1}$$

where $\Theta \in \mathbb{R}^{t \times r}$ denotes pathway's parameters. Notice that Eq 3.1 involves marginalization over Θ_j parameters, which is hard to compute ([28]). One way to mitigate this issue is to approximate the above equation using Monte Carlo (MC) techniques ([24]) by constructing an ensemble, denoted by E, which consists of $g (\in \mathbb{Z}_{\geq 1})$ models (Fig. 2c) where each generates multiple samples according to the following formula:

$$p(\mathbf{y}_j = +1|\mathbf{x}^*, \mathcal{S}_A) \approx \frac{1}{g} \sum_{s \in g} p^s(\mathbf{y}_j = +1|\mathbf{x}^*, \Theta_j^s)$$

where, (3.2)

$$p(\mathbf{y}_j = +1|\mathbf{x}^*, \Theta_j^s) = \frac{1}{1 + e^{-\Theta_j^{s,\mathrm{T}} \mathbf{x}^*}}$$

where $\Theta^s_{[.]}$ is sampled from $q(\Theta^s)$ which is considered to be in the same family distribution as the true hidden variables $p(\Theta^s_j|\mathcal{S}_A)$. The parameters Θ^s for the s-th model can be estimated according to the multi-label 1-vs-All approach ([42]).

Although the computed MC error is expected to decrease by incorporating more samples and members in E, label correlation increases computational complexity during training and pathway prediction (see Section 6.2). Moreover, a single multi-label learner (Fig. 3a) suffers from generalization error due to overfitting despite being able to exploit label correlations. In contrast the ensemble learning method (Fig. 3b) is robust given a group of multi-label base learners that are both accurate and *diverse* (with regard to the allocated samples), potentially reducing overfitting.

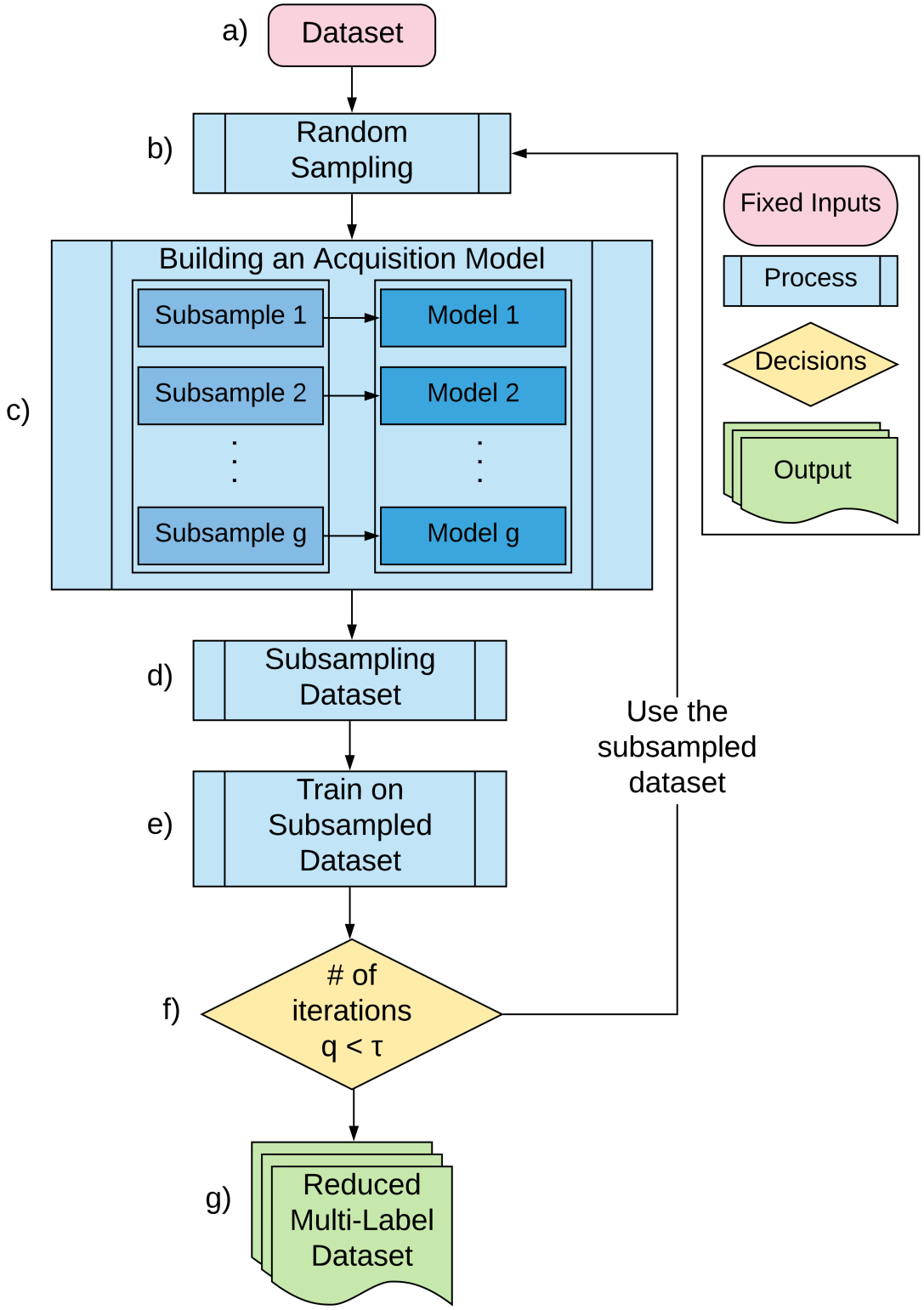


Figure 2: A schematic diagram indicating the leADS workflow. Using a multi-label pathway dataset (a), leADS randomly selects samples at the very first iteration (b) then builds g members of an ensemble (c), where each is trained on a randomly selected portion of the training set. Next, leADS applies an acquisition function (d), based on either: entropy, mutual information, variation ratios, or normalized propensity scored precision at k, to select **per**% sub-samples. Following subsample selection, leADS performs parallel training steps (e). The process (b-e) is repeated τ times (f), where during each iteration **per**% samples are used in addition to another set of samples for training. If the current iteration q reaches a desired number of rounds τ, training is terminated and final **per**% results are presented (g).

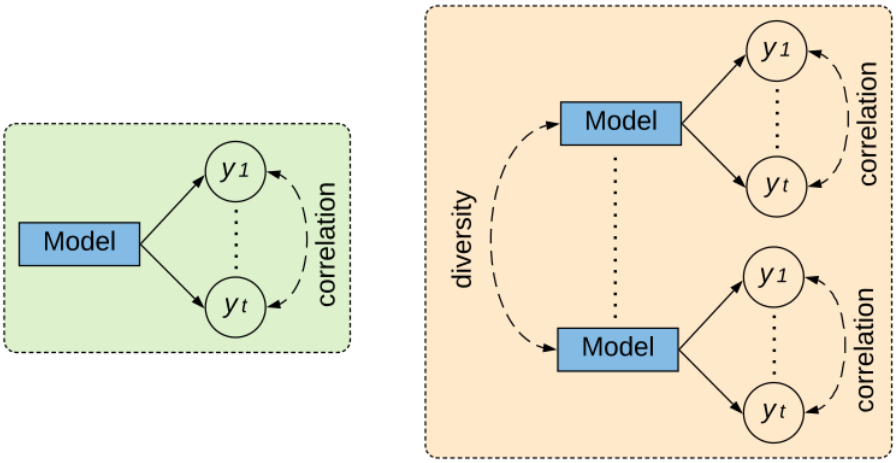


Figure 3: The two approaches for constructing multi-label learning algorithm. The individual multi-label learner (on the left) and the ensemble based multi-label learning (on the right).

3.2 Subsampling Dataset

During this step, a subset of $\mathcal{S}_A$, denoted as $\mathcal{S}^{q-1}_{\mathbf{per}\%} \subseteq \mathcal{S}_A$, is picked for each member in E using an acquisition function $f : \mathbf{x} \rightarrow \mathbb{R}$ where $\mathbf{per}\%$ is a pre-specified threshold, indicating the proportion of samples to be chosen from $\mathcal{S}_A$, at iteration $q-1$.

Four acquisition functions used in subsampling are described that incorporate predictive uncertainty distribution from the previous step: *entropy*, *mutual information*, *variation ratios*, and *normalized PSP@k*. For each function, we retrieve top $\mathbf{per}\%$ samples that contain high acquisition (or uncertainty) values.

1. **Entropy ($\mathcal{H}$)** ([34]). This function measures the uncertainty of a sample given the predictive distribution of that sample:
$$\mathcal{H} = -\mathbf{p}^{\intercal} \log(\mathbf{p}) \tag{3.3}$$
where $\mathbf{p}$ is a vector of predictive probabilities over t pathways.

2. **Mutual information ($\mathcal{M}$)** ([37]). This function looks for low mutual information between g models, encouraging samples with high disagreement to be selected during the data acquisition process:
$$\mathcal{M} = \mathcal{H} - \frac{1}{g} \sum_{s \in g} \mathcal{H}^s \tag{3.4}$$
where $\mathcal{H}^s$ denotes the entropy obtained from an individual member of E for a sample before marginalization. Since entropy is always positive, the maximum possible value for $\mathcal{M}$ is $\mathcal{H}$. However, when the models make similar predictions, then $\frac{1}{e}\sum_{s=1}^{s=e} \mathcal{H}^s \rightarrow \mathcal{H}$, resulting in $\mathcal{M} \rightarrow 0$, which is its minimum value ([9]). Note that this formula is similar to multi-label negative correlation learning ([35]), which estimates pairwise negative correlation of each learner's error with respect to errors of other members in E.

3. **Variation ratios ($\mathcal{V}$)** ([14]). This function measures the number of members in E that disagree with the majority vote for a sample according to k desired pathway size, where larger values indicate higher uncertainty:
$$\begin{aligned} \mathcal{V} =& 1 - \frac{1}{|V|g} \sum_{s \in g} \left| \Big(\{ \arg\, p^s_j : 1 \leq j \leq k \} \Big) \cap V \right| \\ V =& \underset{s \in g}{\text{Mode}} \Big(\{ \arg\, p^s_j : 1 \leq j \leq k \} \Big) \end{aligned} \tag{3.5}$$
where V corresponds the disagreement of k pathways across g models, where $k \in \mathbb{Z}_{>0}$ is a pre-specified number of pathways to be considered in computing the mode operation.

4. **Normalized propensity scored precision at** k (nPSP@k). This is a modified version of PSP@k ([20]), which measures the average precision of top k relevant pathways given an instance i where larger values indicate less uncertainty:
$$\begin{aligned} \text{nPSP@}k =& 1 - \text{Norm}\left(\frac{1}{k} \sum_{j \in \text{rank}_k(\mathbf{p})} \frac{\mathbf{y}_j}{\mathbf{ps}_j} \right) \\ \mathbf{ps}_j =& \frac{1}{1 + (n_j + 1)^{-1}} \end{aligned} \tag{3.6}$$
where Norm(.) scales the score within $[0, 1]$. The term $\mathbf{p}$ is a vector of predictive probabilities over t pathways, $\text{rank}_k(\mathbf{p})$ returns the indices of k largest value in $\mathbf{p}$, ranked in a descending order, where

$k \in \mathbb{Z}_{>0}$ is a hyperparameter. $\mathbf{ps}_j$ is the propensity score for the j-th pathway, where n_j is the number of the positive training instances for the pathway j. In the context of extreme multi-label problems, PSP@k was used to derive an upper bound for missing/miss-classified labels ([39]), and is reported to be a good performance metric for long-tail distribution in which a significant portion of labels are tail labels ([31, 2]).

3.3 Training on the Reduced Dataset

As described above, each member in E is assigned to train on randomly selected samples from $\mathcal{S}^{q-1}_{\mathbf{per}\%}$, which is expected to contain hard examples that are difficult to learn and classify. The process is repeated τ times, where during each iteration the top **per**% are selected based on their acquisition values for the next round of training.

4 Optimization and Prediction

The objective function in Eq. 3.2 can be solved by decomposing into t independent binary classification problems according to the multi-label 1-vs-All approach enabling parallel training. Consider optimization for a member s:

$$\min_{\Theta^s} \sum_{i \in n^s} \sum_{j \in t} \log\left(1 + e^{-\mathbf{y}_j^{(i)} \Theta_j^{s\intercal} \mathbf{x}^{(i)}}\right) + \sum_{j \in t} \lambda ||\Theta_j^s||_{2,1} \tag{4.1}$$

where $||.||^2_{2,1}$ is the $L_{2,1}$ regularization term, which is the sum of the Euclidean norms of columns of Θ. The $L_{2,1}$ norm imposes sparsity on the model's parameters to minimize the negative effect of label correlations, where $\lambda (\in \mathbb{R}_{>0})$ is employed to govern relative contributions of $L_{2,1}$ and the log-loss term. Although the joint formula in Eq 4.1 is convex, the logistic log-loss function still posses a problem where there exists no analytical solution for it. To address this problem, we apply mini-batch gradient descent ([25]), which begins with some initial random guess for leADS parameters, and performs iterative updates to each individual parameter to minimize Eq. 4.1 where the derivative for each $\Theta_j^s \in \Theta^s$ has the following formula:

$$\nabla \Theta_j^s = \frac{1}{n^s} \sum_{i=1}^{i=n^s} \left(\frac{-\mathbf{y}_j^{(i)} \mathbf{x}^{(i)}}{1 + e^{\mathbf{y}_j^{(i)} \Theta_j^{s\intercal} \mathbf{x}^{(i)}}} \right) + \lambda \frac{\Theta_j^s}{2||\Theta_j^s||_2} \tag{4.2}$$

For prediction, we apply a cut-off threshold $\xi \in \mathbb{R}_{\geq 0}$ to retain only pathways having higher probability values than ξ, i.e., $\mathcal{L}(\mathbf{x}) = \{j : p(y_j = +1 | \mathbf{x}, \Theta_j^s) \geq \xi, \forall j \in t, \forall s \in g\}$, where $p(y_j = +1 | \mathbf{x}, \Theta_j^s) = \frac{1}{1+e^{-\Theta_j^{s\intercal} \mathbf{x}^{(i)}}}$.

5 Experimental Setup

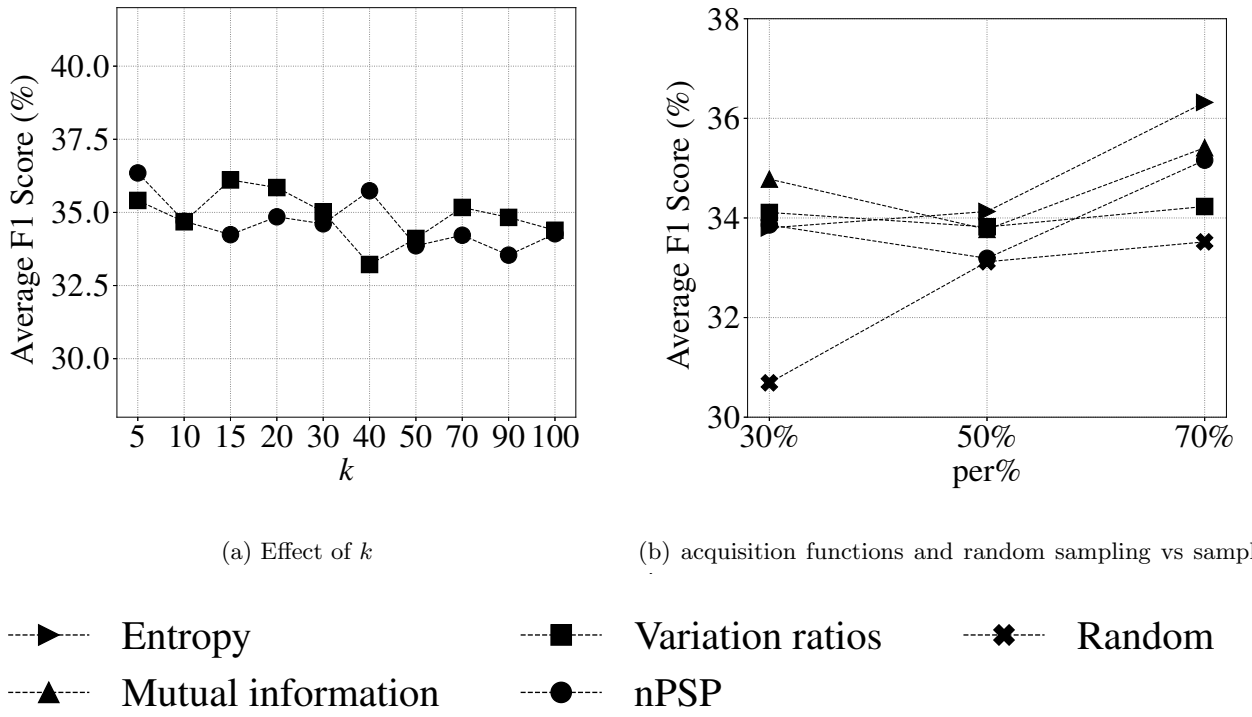
In this section, we describe an experimental framework used to demonstrate leADS pathway prediction performance across multiple datasets spanning the genomic information hierarchy ([5]). leADS was written in the Python programming language (v3). Unless otherwise specified all tests were conducted on a Linux server using 10 cores of Intel Xeon CPU E5-2650.

5.1 Description of Datasets

We used a corpora of 10 organismal and multi-organismal datasets including T1 PGDBs from the BioCyc collection (*AraCyc*, *EcoCyc*, *HumanCyc*, *LeishCyc*, *TrypanoCyc*, and *YeastCyc*), symbiont genomes describing distributed metabolic pathways for amino acid biosynthesis between the two symbiotic bacteria: *Moranella* (GenBank NC-015735) living inside *Tremblaya* (GenBank NC-015736) ([27]), genomes used in the CAMI initiative [33], and whole genome shotgun sequences from HOTS at 25m, 75m, 110m (sunlit) and 500m (dark) ocean depth intervals [38] to benchmark leADS. The general characteristics of the datasets are summarized Supp. Table 1. We used BioCyc (v21 T2 &3) ([8]) to train leADS and triUMPF (using default settings). This version of BioCyc consists of 9429 PGDBs with 1512 distinct pathway labels assigned using Pathway Tools with or without manual curation ([22]).

5.2 Parameter Settings

We used pathway2vec ([3]) to obtain pathway and EC features using "crt" as the embedding method with the following settings: the number of memorized domain was 3, the explore and the in-out hyperparameters were 0.55 and 0.84, respectively, the number of sampled path instances was 100, the walk length was 100, the embedding dimension size was $m = 128$, the neighborhood size was 5, the size of negative samples was 5, and the used configuration of MetaCyc was "uec", indicating links among ECs were trimmed. The obtained features were used to leverage correlations among ECs and pathways for training leADS (see Supp. Section 4). We then trained leADS using the following default settings (unless otherwise mentioned): the learning



(a) Effect of k

(b) acquisition functions and random sampling vs sample

Figure 4: The impact of k on leADS performance on the CAMI dataset by varying $k \in \{5, 10, 15, 20, 30, 40, 50, 70, 90, 100\}$ using variation ratios and nPSP as acquisition functions is demonstrated in Fig. 4a while the effect of four acquisition functions and random sampling by varying sample size according to **per**% $\in \{30\%, 50\%, 70\%\}$ is shown in Fig. 4b.

rate was 0.0001, the batch size was 50, the number of epochs was 3, the number of models was $g = 3$, the proportion of samples (**per**%) to be selected was 30%, the number of subsampled pathways for each member was 500, and the cutoff threshold ξ for predictions was 0.5. For the regularized hyperparameter λ, we performed 10-fold cross-validation on BioCyc T2 &3 data and found the settings $\lambda = 10$ to be optimum according to results obtained on golden T1 and CAMI datasets.

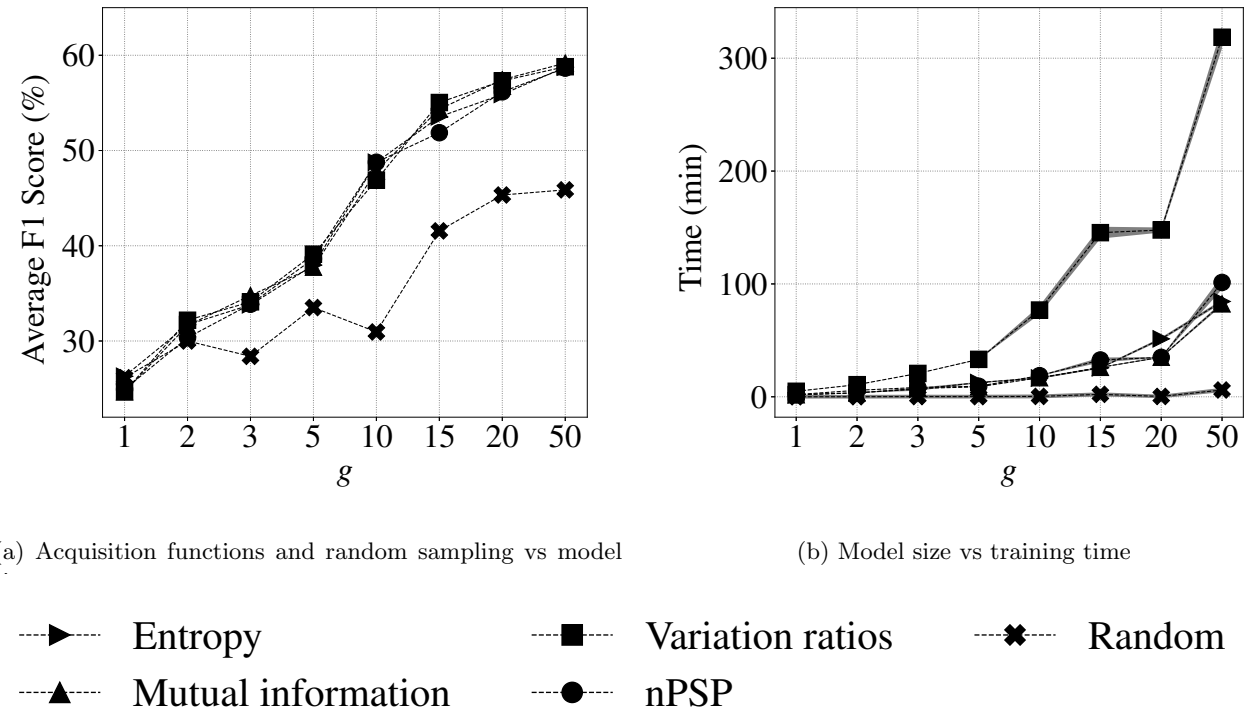
6 Experimental Results and Discussion

To verify the effectiveness of leADS, we conducted three experimental studies: parameter sensitivity, scalability, and metabolic pathway prediction.

6.1 Parameter Sensitivity

Experimental setup. In this section, the impact of two user defined hyperparameters (k and **per**%) were evaluated on the CAMI dataset using acquisition functions described in Section 3.2. In the case of k, a range of values between $\{5, 10, 15, 20, 30, 40, 50, 70, 90, 100\}$ was tested in relation to pathway size for variation ratios in Eq. 3.5 or top k relevant pathways for nPSP in Eq. 3.6. In the case of **per**% different subsampling proportions between $\{30\%, 50\%, 70\%\}$ were tested by selecting BioCyc T2 &3 data at random. For variation ratios and nPSP, the values of k were fixed based on the optimum results obtained from the previous experiment. All other hyperparameters, were set according to the configurations described in Section 5.2 and results were reported using average F1 scores.

Experimental results. Fig. 4a shows the impact of k for both variation ratios and nPSP acquisition functions. Although both functions have similar disagreement metrics, the optimum performance for variation ratios is at $k = 15$ while the optimum for nPSP is at $k = 40$. This discrepancy in k values likely results from the effects of subsampling pathways and examples that are allocated randomly to each member in E. After several rounds of experiments, we found $k = 50$ to be optimum for both variation ratios and nPSP. Next, we examined the effect of **per**% on leADS's performances using four acquisition functions and random sampling, where we fixed $k = 50$ for variation ratios and nPSP. From Fig. 4b, it is evident that leADS performance generally improves by including more samples for each acquisition function, although the entropy function resulted in a marginal improvement. In contrast, random sampling had no performance benefit across the sample size range tested.



(a) Acquisition functions and random sampling vs model

(b) Model size vs training time

Figure 5: Fig. 5a shows the average F1 score reported on CAMI data as the ensemble size g varies across $\{1, 2, 3, 5, 10, 15, 20, 50\}$ while the elapsed computational time (in minutes) per epoch (averaged over 3 epochs) is demonstrated in Fig. 5b based on the same ensemble size variation.

6.2 Scalability to the Ensemble Size

Experimental setup. In this section, time complexity of training was determined when the model size varied as $g \in \{1, 2, 3, 5, 10, 15, 20, 50\}$, simultaneously. Performance was evaluated on the CAMI dataset as described above using the average F1 score metric for each configuration of g. **per**%*wassetto*30% of BioCyc T2 &3 data for training under the four acquisition functions. In the case of random sampling, leADS was trained on 30% of randomly selected BioCyc T2 &3 data. Performance was expected was expected to improve proportionally to the member size in E (due to the dual effects of pathways and examples that are being allocated randomly to each base learner) with concomitant increase in computational time. See section 5.2 for configuration settings.

Experimental results. Results in Fig. 5a are consistent with expectations, with gradual inclusion of more members in E improving leADs performance. Although random sampling reduced time complexity when compared to the four acquisition functions under all model size configurations, it resulted in the lowest performance (Fig. 5b). Among the four acquisition functions, variation ratios required an additional mode operation, contributing to increased training time. Based on these results, setting the model size between $[3, 10] \in \mathbb{Z}_{>0}$ while increasing pathway subsampling size accordingly (e.g. 2000 for 10 members) is recommended to improve prediction outcomes and reduce both computational complexity (training and inference) and parameter storage needs.

6.3 Metabolic Pathway Prediction

Experimental setup. In this section, pathway prediction performance was evaluated using parameter settings described in Section 5.2. Three training configurations were tested: i)- **per**% = 70% under the four acquisition functions, ii)- random sampling corresponding to 70% of BioCyc T2 &3 selected at random, and iii)- full configuration where all BioCyc T2 &3 data were utilized without subsampling. After training, pathway prediction results were reported on golden T1 data using four evaluation metrics: *Hamming loss*, *average precision*, *average recall*, and *average F1 score*. leADS performance was compared to four extant pathway prediction algorithms: i)- **MinPath** v1.2 ([41]), ii)- **PathoLogic** v21 ([22]); iii)- **mlLGPR** ([5]) and iv)- **triUMPF** ([4]) on the T1 data. In addition, we compared leADS performance to other methods on multi-organismal datasets including symbiont, CAMI low complexity and HOTS datasets. For all experiments,the number of epochs was 10, the member size was $g = 3$, the subsampled pathway size was 2000, and k was 50 (for variation ratios and nPSP). See section 5.2 for additional configuration settings.

Experimental results. As shown in Table 1, leADS resulted in competitive performance compared to other pathway inference algorithms based on average F1 scores. For each column in Table 1, a boldface number represents the best evaluation metric score while an underlined number indicates the best score between

leADS variants. Among the four acquisition functions, leADS+nPSP resulted in the highest average F1 scores for EcoCyc (0.8874) and HumanCyc (0.8333) which are also the highest scores among all models tested. Consistent with previous sections, random sampling resulted in the poorest overall performance scores. Interestingly, leADS+Full in Table 1 was on par with random sampling, reinforcing the idea that BioCyc T2 &3 contain noisy data that hampered proper estimation of leADS coefficients. Through subsampling informative data in an ensemble based framework, leADS was able to reduce noise and improve the prediction performance on golden T1 data.

To evaluate leADS performance on metabolic pathways distributed between organisms we used the reduced genomes of the mealybug symbionts *Moranella* (GenBank NC-015735) and *Tremblaya* (GenBank NC-015736) ([27]). The two symbiont genomes in combination encode intact biosynthetic pathways for 9 essential amino acids. Pathologic, mlLGPR, triUMPF, and leADS were used to predict pathways on individual symbiont genomes and a concatenated dataset consisting of both symbiont genomes, and resulting amino acid biosynthetic pathway distributions were determined (Supp. Fig. 1). Pathologic, triUMPF, and leADS predicted 6 of the expected amino acid biosynthetic pathways on the composite genome while mlLGPR predicted 8 pathways. The phenylalanine biosynthesis (*L-phenylalanine biosynthesis I*) pathway was excluded from analysis because the associated genes were reported to be missing during the ORF prediction process. All models inferred false positive pathways for individual symbiont genomes (*Moranella* and *Tremblaya*) despite reduced pathway coverage information (mapping enzymes onto associated 9 amino acid biosynthetic pathways) relative to the composite genome. Although it is possible for leADS to reduce type I error by incorporating taxonomy-based predictions using rules, such pruning can also increase false-negative (type II error) pathway predictions in multi-organismal datasets ([18]).

To evaluate performance on more complex multi-organismal genomes we compared leADS to mlLGPR and triUMPF using the CAMI low complexity dataset ([33]) and to PathoLogic, mlLGPR, triUMPF using the HOTS dataset ([38]). In the case of CAMI, leADS+nPSP outperformed other methods resulting in an average F1 score of 0.6214 (Supp. Table 2). In the case of HOTS, leADS+Random, leADS+Full, leADS+$\mathcal{H}$, leADS+$\mathcal{M}$, leADS+$\mathcal{V}$, and leADS+nPSP predicted a total of 60, 67, 63, 68, 67, and 68 pathways among a subset of 180 selected water column pathways ([18]), while PathoLogic, mlLGPR, and triUMPF (using BioCyc v21) inferred 54, 62 and 67 pathways, respectively. These observations indicate that leADS with subsampling improves pathway prediction outcomes by reducing training loss due to pathway class imbalance (Supp. Fig. 10).

7 Conclusion

In this paper we present leADS, a novel ensemble-based ML approach for hard example mining that constructs a set of diverse multi-label base learners to jointly improve the subselection of samples and overcome class imbalance during metabolic pathway prediction from genomic sequence information at different levels of complexity and completion. leADS performs an iterative process to: (i)- construct an acquisition model in an ensemble framework; (ii) select informative points using an appropriate acquisition function including entropy, mutual information, variation ratios, and normalized PSP@k; and (iii) train on selected samples.

We evaluated leADS performance using a corpora of experimental datasets manifesting diverse multi-label properties comparing pathway prediction outcomes to other prediction methods including MinPath ([41]), PathoLogic ([22]), mlLGPR ([5]) and triUMPF ([4]). Resulting performance metrics indicated that leADS equaled or exceeded pathway prediction outcomes on organismal and multi-organismal datasets with increased sensitivity on T1 golden data. This indicates that active subsampling can overcome pathway class imbalance. At the same time, it is important to emphasize that the acquisition functions used in subsampling tend to reduce the number of pathways used in training $\mathcal{Y}$ in Def. 2.1. For example, leADS+$\mathcal{H}$, leADS+$\mathcal{M}$, leADS+$\mathcal{V}$, and leADS+nPSP returned 1380, 1378, 1431 and 1404 distinct pathways, respectively, from a total of 1512 pathways in BioCyc T2 &3. This reduction reveals a fundamental limitation of subsampling based approaches ([11]).

Members of an ensemble in leADS have the following two important properties: *representativeness* (each member has a different set of candidate examples) and *diversity* (each member has different overlapping pathways across examples) ([16]). Having these properties implies that a member trained on a subset of examples containing a more diverse subset of pathways should be given more weights when predicting those subsets of pathways. Unfortunately, leADS does not utilize such weighting which can be resolved in part by adopting a better voting scheme ([32, 15]), or by incorporating an additional learner that integrates weights obtained from all the base learners into global weights ([12, 13]). Looking forward, an integrated ensemble or meta-learning framework is needed that can estimate the confidence of multiple training methods to provide an optimal balance between sensitivity and precision when predicting pathways across different levels of genome complexity and completion.

Acknowledgments

We would like to thank Connor Morgan-Lang, Julia Glinos, Kishori Konwar and Aria Hahn for lucid discussions on the function of the leADS model and Ryan MacLaughlin for his participation in preliminary performance evaluations and graphics design.

Methods	Hamming Loss ↓					
	EcoCyc	HumanCyc	AraCyc	YeastCyc	LeishCyc	TrypanoCyc
PathoLogic	0.0610	0.0633	0.1188	**0.0424**	0.0368	**0.0424**
MinPath	0.2257	0.2530	0.3266	0.2482	0.1615	0.2561
mlLGPR	0.0804	0.0633	**0.1069**	0.0550	0.0380	0.0590
triUMPF	0.0317	0.0523	0.1560	0.0740	0.0530	0.0515
leADS+Random	0.0574	0.0796	0.1528	0.0796	0.0515	0.0685
leADS+Full	0.0471	0.0732	0.1576	0.0736	0.0396	0.0566
leADS+$\mathcal{H}$	0.0265	0.0610	0.1453	0.0756	0.0471	0.0606
leADS+$\mathcal{M}$	0.0289	0.0499	0.1425	0.0657	0.0408	0.0542
leADS+$\mathcal{V}$	0.0301	0.0424	0.1394	0.0649	0.0368	0.0507
leADS+nPSP	**0.0261**	**0.0364**	0.1457	0.0653	**0.0333**	0.0499

Methods	Average Precision Score ↑					
	EcoCyc	HumanCyc	AraCyc	YeastCyc	LeishCyc	TrypanoCyc
PathoLogic	0.7230	0.6695	0.7011	0.7194	0.4803	0.5480
MinPath	0.3490	0.3004	0.3806	0.2675	0.1758	0.2129
mlLGPR	0.6187	0.6686	0.7372	0.6480	0.4731	0.5455
triUMPF	0.9158	0.7094	0.6801	0.6040	0.3819	0.5789
leADS+Random	0.7516	0.6383	0.7199	0.5714	0.3799	0.5039
leADS+Full	0.7994	0.6767	0.7171	0.6352	0.4606	0.5611
leADS+$\mathcal{H}$	**0.9380**	0.6997	0.7299	0.5872	0.4192	0.5423
leADS+$\mathcal{M}$	0.9239	0.7508	0.7757	0.6684	0.4529	0.5779
leADS+$\mathcal{V}$	0.9231	0.7654	0.8110	0.6720	0.4828	0.6009
leADS+nPSP	0.9319	**0.8425**	**0.8198**	0.7078	**0.5102**	**0.6061**

Methods	Average Recall Score ↑					
	EcoCyc	HumanCyc	AraCyc	YeastCyc	LeishCyc	TrypanoCyc
PathoLogic	0.8078	0.8423	0.7176	0.8734	0.8391	0.7829
MinPath	**0.9902**	**0.9713**	**0.9843**	**1.0000**	**1.0000**	**1.0000**
mlLGPR	0.8827	0.8459	0.7314	0.8603	0.9080	0.8914
triUMPF	0.8143	0.8925	0.4294	0.5328	0.8736	0.9429
leADS+Random	0.7883	0.6452	0.3980	0.4891	0.7816	0.7429
leADS+Full	0.8176	0.6452	0.3627	0.4410	0.8736	0.8400
leADS+$\mathcal{H}$	0.8371	0.7849	0.4451	0.5590	0.9540	0.8057
leADS+$\mathcal{M}$	0.8306	0.8208	0.4137	0.5459	0.8851	0.8057
leADS+$\mathcal{V}$	0.8208	0.8889	0.4039	0.5546	0.9655	0.8000
leADS+nPSP	0.8469	0.8244	0.3569	0.4760	0.8621	0.8000

Methods	Average F1 Score ↑					
	EcoCyc	HumanCyc	AraCyc	YeastCyc	LeishCyc	TrypanoCyc
PathoLogic	0.7631	0.7460	0.7093	**0.7890**	0.6109	0.6447
MinPath	0.5161	0.4589	0.5489	0.4221	0.2990	0.3511
mlLGPR	0.7275	0.7468	**0.7343**	0.7392	0.6220	0.6768
triUMPF	0.8621	0.7905	0.5264	0.5661	0.5315	**0.7174**
leADS+Random	0.7695	0.6417	0.5126	0.5271	0.5113	0.6005
leADS+Full	0.8084	0.6606	0.4818	0.5206	0.6032	0.6728
leADS+$\mathcal{H}$	0.8847	0.7399	0.5530	0.5727	0.5825	0.6483
leADS+$\mathcal{M}$	0.8748	0.7842	0.5396	0.6010	0.5992	0.6730
leADS+$\mathcal{V}$	0.8690	0.8226	0.5393	0.6077	**0.6437**	0.6863
leADS+nPSP	**0.8874**	**0.8333**	0.4973	0.5692	0.6410	0.6897

Table 1: Predictive performance of each comparing algorithm on 6 benchmark datasets. leADS+Full: leADS with full data, leADS+Random: leADS with random sampling, leADS+$\mathcal{H}$: leADS with entropy, leADS+$\mathcal{M}$: leADS with mutual information, leADS+$\mathcal{V}$: leADS with variation ratios, and leADS+nPSP: leADS with normalized propensity scored precision. For each performance metric, '↓' indicates the smaller score is better while '↑' indicates the higher score is better. Values in boldface represent the best performance score while the underlined score indicates the best performance among leADS variances.

Funding: This work was performed under the auspices Genome Canada, Genome British Columbia, the Natural Science and Engineering Research Council (NSERC) of Canada, and Compute/Calcul Canada). ARMAB was supported by a UBC four-year doctoral fellowship (4YF) administered through the UBC Graduate Program in Bioinformatics.
Conflict of Interest: SJH is a co-founder of Koonkie Inc., a bioinformatics consulting company that designs and provides scalable algorithmic and data analytics solutions in the cloud.

References

[1] Wilhelm J Ansorge. Next-generation dna sequencing techniques. *New biotechnology*, 25(4):195–203, 2009.

[2] Rohit Babbar and Bernhard Schölkopf. Adversarial extreme multi-label classification. *arXiv preprint arXiv:1803.01570*, 2018.

[3] Abdur Rahman MA Basher and Steven J Hallam. Leveraging heterogeneous network embedding for metabolic pathway prediction. *bioRxiv*, feb 2020.

[4] Abdur Rahman MA Basher, Ryan J McLaughlin, and Steven J Hallam. Incorporating triple nmf with community detection to metabolic pathway inference. *bioRxiv*, 2020.

[5] Abdur Rahman MA Basher, Ryan J McLaughlin, and Steven J Hallam. Metabolic pathway inference using multi-label classification with rich pathway features. *bioRxiv*, February 2020.

[6] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pp. 41–48. ACM, 2009.

[7] Ron Caspi, Richard Billington, Luciana Ferrer, Hartmut Foerster, Carol A. Fulcher, Ingrid M. Keseler, Anamika Kothari, Markus Krummenacker, Mario Latendresse, Lukas A. Mueller, Quang Ong, Suzanne Paley, Pallavi Subhraveti, Daniel S. Weaver, and Peter D. Karp. The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic Acids Research*, 44(D1):D471–D480, 2016.

[8] Ron Caspi, Richard Billington, Hartmut Foerster, Carol A Fulcher, Ingrid Keseler, Anamika Kothari, Markus Krummenacker, Mario Latendresse, Lukas A Mueller, Quang Ong, et al. Biocyc: Online resource for genome and metabolic pathway analysis. *The FASEB Journal*, 30(1 Supplement):lb192–lb192, 2016.

[9] Kashyap Chitta, Jose M Alvarez, Elmar Haussmann, and Clement Farabet. Less is more: An exploration of data redundancy with active dataset subsampling. *arXiv preprint arXiv:1905.12737*, 2019.

[10] Joseph M Dale, Liviu Popescu, and Peter D Karp. Machine learning methods for metabolic pathway prediction. *BMC bioinformatics*, 11(1):1, 2010.

[11] Alberto Fernández, Salvador Garcia, Francisco Herrera, and Nitesh V Chawla. Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research*, 61:863–905, 2018.

[12] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1126–1135. JMLR. org, 2017.

[13] Chelsea Finn, Aravind Rajeswaran, Sham Kakade, and Sergey Levine. Online meta-learning. In *ICML*, 2019.

[14] Yarin Gal. *Uncertainty in deep learning*. PhD thesis, PhD thesis, University of Cambridge, 2016.

[15] Heitor Murilo Gomes, Jean Paul Barddal, Fabrício Enembreck, and Albert Bifet. A survey on ensemble learning for data stream classification. *ACM Computing Surveys (CSUR)*, 50(2):23, 2017.

[16] Yuhong Guo and Dale Schuurmans. Discriminative batch mode active learning. In *Advances in neural information processing systems*, pp. 593–600, 2008.

[17] Aria S Hahn, Kishori M Konwar, Stilianos Louca, Niels W Hanson, and Steven J Hallam. The information science of microbial ecology. *Current opinion in microbiology*, 31:209–216, 2016.

[18] Niels W Hanson, Kishori M Konwar, Alyse K Hawley, Tomer Altman, Peter D Karp, and Steven J Hallam. Metabolic pathways for the whole community. *BMC genomics*, 15(1):1, 2014.

[19] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

[20] Himanshu Jain, Yashoteja Prabhu, and Manik Varma. Extreme multi-label loss functions for recommendation, tagging, ranking & other missing label applications. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 935–944. ACM, 2016.

[21] Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. Kegg: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45(D1):D353–D361, 2017.

[22] Peter D Karp, Mario Latendresse, Suzanne M Paley, Markus Krummenacker, Quang D Ong, Richard Billington, Anamika Kothari, Daniel Weaver, Thomas Lee, Pallavi Subhraveti, et al. Pathway tools version 19.0 update: software for pathway/genome informatics and systems biology. *Briefings in bioinformatics*, 17(5):877–890, 2016.

[23] Hiroaki Kitano. Computational systems biology. *Nature*, 420(6912):206, 2002.

[24] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pp. 6402–6413, 2017.

[25] Jason D Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I Jordan, and Benjamin Recht. First-order methods almost always avoid saddle points. *arXiv preprint arXiv:1710.07406*, 2017.

[26] Ilya Loshchilov and Frank Hutter. Online batch selection for faster training of neural networks. *arXiv preprint arXiv:1511.06343*, 2015.

[27] John P McCutcheon and Carol D Von Dohlen. An interdependent metabolic patchwork in the nested symbiosis of mealybugs. *Current Biology*, 21(16):1366–1372, 2011.

[28] Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

[29] Zoltán N Oltvai and Albert-László Barabási. Life's complexity pyramid. *Science*, 298(5594):763–764, 2002.

[30] Te Pi, Xi Li, Zhongfei Zhang, Deyu Meng, Fei Wu, Jun Xiao, and Yueting Zhuang. Self-paced boost learning for classification. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pp. 1932–1938. AAAI Press, 2016.

[31] Yashoteja Prabhu, Anil Kag, Shilpa Gopinath, Kunal Dahiya, Shrutendra Harsola, Rahul Agrawal, and Manik Varma. Extreme multi-label learning with label features for warm-start tagging, ranking & recommendation. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pp. 441–449. ACM, 2018.

[32] Omer Sagi and Lior Rokach. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1249, 2018.

[33] Alexander Sczyrba, Peter Hofmann, Peter Belmann, David Koslicki, Stefan Janssen, Johannes Dröge, Ivan Gregor, Stephan Majda, Jessika Fiedler, Eik Dahms, et al. Critical assessment of metagenome interpretation—a benchmark of metagenomics software. *Nature methods*, 14(11):1063, 2017.

[34] Claude Elwood Shannon. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423, 1948.

[35] Chuan Shi, Xiangnan Kong, S Yu Philip, and Bai Wang. Multi-label ensemble learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 223–239. Springer, 2011.

[36] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 761–769, 2016.

[37] Lewis Smith and Yarin Gal. Understanding measures of uncertainty for adversarial example detection. *arXiv preprint arXiv:1803.08533*, 2018.

[38] Frank J Stewart, Adrian K Sharma, Jessica A Bryant, John M Eppley, and Edward F DeLong. Community transcriptomics reveals universal patterns of protein sequence conservation in natural microbial communities. *Genome biology*, 12(3):R26, 2011.

[39] Tong Wei and Yu-Feng Li. Learning compact model for large-scale multi-label data. 2019.

[40] Liang Yang, Xi-Zhu Wu, Yuan Jiang, and Zhi-Hua Zhou. Multi-label learning with deep forest. *arXiv preprint arXiv:1911.06557*, 2019.

[41] Yuzhen Ye and Thomas G Doak. A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. *PLoS Comput Biol*, 5(8):e1000465, 2009.

[42] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837, 2014.

Supplementary - Multi-label pathway prediction based on active dataset subsampling

Abdur Rahman M. A. Basher[1] and Steven J. Hallam [1,2,3,4,5*]

[1] Graduate Program in Bioinformatics, University of British Columbia, Genome Sciences Centre, 100-570 West 7th Avenue, Vancouver, British Columbia V5Z 4S6, Canada.
[2] Department of Microbiology & Immunology, University of British Columbia, 2552-2350 Health Sciences Mall, Vancouver, British Columbia V6T 1Z3, Canada.
[3] Genome Science and Technology Program, University of British Columbia, 2329 West Mall, Vancouver, BC V6T 1Z4, Canada
[4] Life Sciences Institute, University of British Columbia, Vancouver, British Columbia, Canada V6T 1Z3
[5] ECOSCOPE Training Program, University of British Columbia, Vancouver, British Columbia, Canada V6T 1Z3
* To whom correspondence should be addressed

1 Dataset Characteristics

Experiments were conducted on a corpora of 10 high-dimensional pathway datasets with diverse multi-label properties, ranging from organismal to multi-organismal genomes. These datasets are: i)- golden T1 composed of six databases, retrieved from biocyc website: *EcoCyc (v21)*, *HumanCyc (v19.5)*, *AraCyc (v18.5)*, *YeastCyc (v19.5)*, *LeishCyc (v19.5)*, and *TrypanoCyc (v18.5)*, and are refined to include only those pathways that cross-intersect with the *MetaCyc* database (v21) [2]; ii)- reduced complexity data from *Moranella* (GenBank NC-015735) and *Tremblaya* (GenBank NC-015736) mealybug symbiont genomes encoding distributed metabolic pathways for amino acid biosynthesis [5]; iii)- the Critical Assessment of Metagenome Interpretation (CAMI) initiative low complexity dataset (edwards.sdsu.edu/research/cami-challenge-datasets/), consisting of 40 genomes [6]; iv)- whole genome shotgun sequences from the Hawaii Ocean Time Series (HOTS) at 25m, 75m, 110m (sunlit) and 500m (dark) ocean depth intervals downloaded from the NCBI Sequence Read Archive under accession numbers SRX007372, SRX007369, SRX007370, SRX007371 [7]; and v)- BioCyc (v21 T2 &3) [3], which consists of 9429 PGDBs (Pathway/Genome Databases) with 1512 distinct pathways. The detailed characteristics of the datasets are summarized in Table 1. For each dataset $\mathcal{S}$, we use $|\mathcal{S}|$ and L($\mathcal{S}$) to represent the number of instances and pathway labels, respectively. In addition, we also present some characteristics of the multi-label datasets, which are denoted as:

1. Label cardinality (LCard($\mathcal{S}$) $= \frac{1}{n}\sum_{i=1}^{i=n}\sum_{j=1}^{j=t}\mathbb{I}[\mathbf{Y}_{i,j} \neq -1]$), where $\mathbb{I}$ is an indicator function. It denotes the average number of pathways in $\mathcal{S}$.
2. Label density (LDen($\mathcal{S}$) $= \frac{LCard(\mathcal{S})}{L(\mathcal{S})}$). This is simply obtained through normalizing LCard($\mathcal{S}$) by the number of total pathways in $\mathcal{S}$.
3. Distinct label sets (DL($\mathcal{S}$)). This notation indicates the number of distinct pathways in $\mathcal{S}$.
4. Proportion of distinct label sets (PDL($\mathcal{S}$) $= \frac{DL(\mathcal{S})}{|\mathcal{S}|}$). It represents the normalized version of DL($\mathcal{S}$), and is obtained by dividing DL(.) with the number of instances in $\mathcal{S}$.

The notations R($\mathcal{S}$), RCard($\mathcal{S}$), RDen($\mathcal{S}$), DR($\mathcal{S}$), and PDR($\mathcal{S}$) have similar meanings for the enzymatic reactions $\mathcal{E}$ in $\mathcal{S}$. Finally, PLR($\mathcal{S}$) represents a ratio of L($\mathcal{S}$) to R($\mathcal{S}$). The preprocessed experimental datasets can be obtained from zenodo.org/record/3993874#.X2BLO4ZlDeQ

2 Incorporating EC Features

For pathway prediction, we first obtain node features using "crt" embedding method from pathway2vec [1] with settings provided in Section 5.2 of the main manuscript. Then, we exclusively use EC features to concatenate each example i according to:

$$\tilde{\mathbf{x}}^{(i)} = \mathbf{x}^{(i)} \oplus \frac{1}{r}\mathbf{x}^{(i)}\mathbf{E} \tag{2.1}$$

Dataset	$\lvert\mathcal{S}\rvert$	L($\mathcal{S}$)	LCard($\mathcal{S}$)	LDen($\mathcal{S}$)	DL($\mathcal{S}$)	PDL($\mathcal{S}$)	R($\mathcal{S}$)	RCard($\mathcal{S}$)	RDen($\mathcal{S}$)	DR($\mathcal{S}$)	PDR($\mathcal{S}$)	PLR($\mathcal{S}$)	Domain
AraCyc	1	510	510	1	510	510	2182	2182	1	1034	1034	0.2337	Arabidopsis thaliana
EcoCyc	1	307	307	1	307	307	1134	1134	1	719	719	0.2707	Escherichia coli K-12 substr.MG1655
HumanCyc	1	279	279	1	279	279	1177	1177	1	693	693	0.2370	Homo sapiens
LeishCyc	1	87	87	1	87	87	363	363	1	292	292	0.2397	Leishmania major Friedlin
TrypanoCyc	1	175	175	1	175	175	743	743	1	512	512	0.2355	Trypanosoma brucei
YeastCyc	1	229	229	1	229	229	966	966	1	544	544	0.2371	Saccharomyces cerevisiae
Symbiont	3	119	39.6667	0.3333	59	19.6667	304	101.3333	0.3333	130	43.3333	0.3914	Composed of Moranella and Tremblaya
CAMI	40	6261	156.5250	0.0250	674	16.8500	14269	356.7250	0.0250	1083	27.0750	0.4388	Simulated microbiomes of low complexity
HOT	4	2178	311.1429	0.1429	781	111.5714	182675	26096.4286	0.1429	1442	206.0000	0.0119	Metagenomic Hawaii Ocean Time-series (10m, 75m, 110m, and 500m)
BioCyc	9429	1833617	194.4657	0.0001	1512	0.1604	9000227	954.5261	0.0001	2766	0.2934	0.2037	BioCyc version 21 (tier 2 & 3)

Table 1: Experimental data set properties. The notations $\lvert\mathcal{S}\rvert$, L($\mathcal{S}$), LCard($\mathcal{S}$), LDen($\mathcal{S}$), DL($\mathcal{S}$), and PDL($\mathcal{S}$) represent: number of instances, number of pathway labels, pathway labels cardinality, pathway labels density, distinct pathway labels set, and proportion of distinct pathway labels set for $\mathcal{S}$, respectively. The notations R($\mathcal{S}$), RCard($\mathcal{S}$), RDen($\mathcal{S}$), DR($\mathcal{S}$), and PDR($\mathcal{S}$) have similar meanings for the enzymatic reactions $\mathcal{E}$ in $\mathcal{S}$. PLR($\mathcal{S}$) represents a ratio of L($\mathcal{S}$) to R($\mathcal{S}$). The last column denotes the domain of $\mathcal{S}$.

Metric	mlLGPR	triUMPF	leADS+Random	leADS+Full	leADS+$\mathcal{H}$	leADS+$\mathcal{M}$	leADS+$\mathcal{V}$	leADS+nPSP
Hamming Loss (↓)	0.0975	0.0423	0.0577	0.0553	0.0402	0.0398	0.0399	**0.0397**
Average Precision Score (↑)	0.3570	0.7308	0.5245	0.5468	0.7515	0.7558	0.7550	**0.7569**
Average Recall Score (↑)	**0.7827**	0.5030	0.5212	0.5284	0.5260	0.5306	0.5268	0.5334
Average F1 Score (↑)	0.4866	0.5915	0.5174	0.5320	0.6151	0.6199	0.6167	**0.6214**

Table 2: Predictive performance of mlLGPR with elastic net penalty, triUMPF, and leADS on CAMI low complexity data. leADS+Full: leADS with full data, leADS+Random: leADS with random sampling, leADS+$\mathcal{H}$: leADS with entropy, leADS+$\mathcal{M}$: leADS with mutual information, leADS+$\mathcal{V}$: leADS with variation ratios, and leADS+nPSP: leADS with normalized propensity scored precision. Values in boldface represent the best performance score while the underlined score indicates the best performance among leADS variances.

where $\oplus$ indicates the vector concatenation operation, $\mathbf{E} \in \mathbb{R}^{r \times m}$ corresponds the feature matrix of ECs and $m = 128$. The addition of features results in a dimension of size $r + m$, where $r = 3650$. We expect by incorporating enzymatic reaction features into the original r dimensional example $\mathbf{x}^{(i)}$, the modified $\tilde{\mathbf{x}}^{(i)}$ summarizes informative characteristics, which are expected to be useful in pathway prediction.

3 Metabolic Pathway Prediction

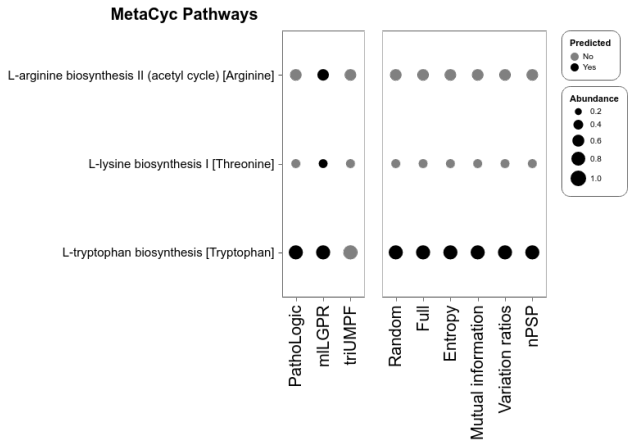
Here, we investigate the effectiveness of leADS for the pathway prediction task on mealybug symbiont genomes, CAMI low complexity, and HOTS datasets.

3.1 Predicted Pathways on Symbiont data

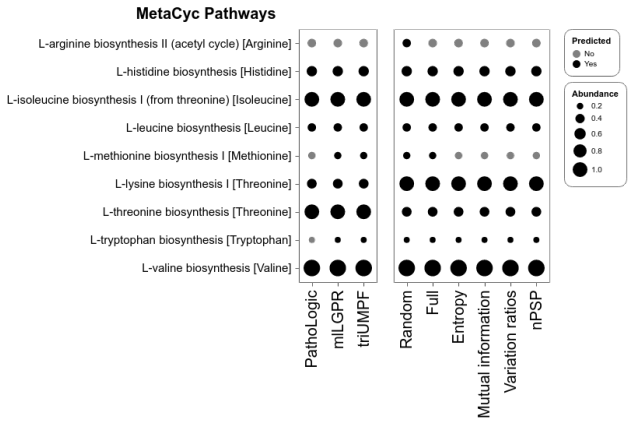
We analyzed pathways from each individual genome of symbiotic data and their combinations. Fig. 1 shows that leADS (with all strategies), triUMPF, and PathoLogic predicted 6 of the expected amino acid biosynthetic pathways on the composite genome while mlLGPR predicted 8 pathways.

3.2 Pathway Prediction from CAMI data

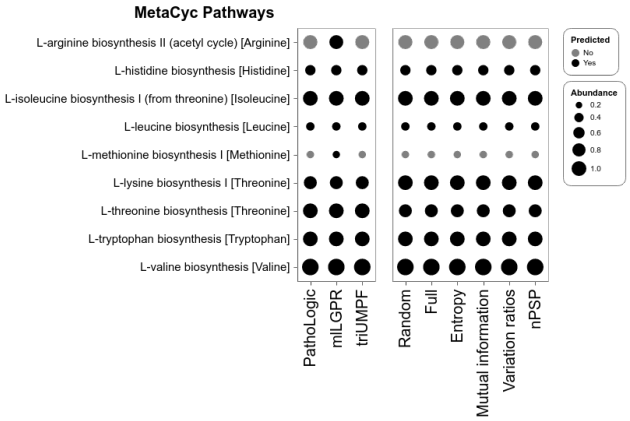
In this section, we contrast leADS (using four acquisition functions and random sampling) with triUMPF and mlLGPR (using elastic net penalty with reaction and pathway evidence features) on CAMI low complexity dataset. From Table 2, we observe that leADS+nPSP outperformed other algorithms with regard to the average F1 score, achieving 0.6214.



(a) Moranella



(b) Tremblaya



(c) Composite

Figure 1: Comparative study of predicted pathways for symbiont data between PathoLogic, mlLGPR, triUMPF, and leADS (with random sampling, full data, and four acquisition functions). Black circles indicate predicted pathways by associated models while grey circles indicate pathways that were not recovered by models. The size of circles corresponds the pathway abundance information.

3.3 Predicted Pathways from HOTS data

We used leADS to infer a set of pathways from HOTS dataset, where leADS+Random, leADS+Full, leADS+$\mathcal{H}$, leADS+$\mathcal{M}$, leADS+$\mathcal{V}$, and leADS+nPSP were able to recover a total of 60, 67, 63, 68, 67, and 68 pathways while triUMPF, mlLGPR, and PathoLogic detected 67, 62, and 54 pathways, respectively, from 180 previously reported pathways ([4]). The results of leADS are presented in Figs. 2, 3, 4 & 5.

4 Visualization

We applied leADS for **per**% = 70% subsampling, which reduced the number of selected samples (more than half for some species related instances) as illustrated in Fig. 10. Figs 6, 7, 8, and 9 show the selected examples by leADS+$\mathcal{H}$, leADS+$\mathcal{M}$, leADS+$\mathcal{V}$, and leADS+nPSP, respectively, in relation to top 100 occurring species in BioCyc T2 & 3.

References

[1] Abdur Rahman MA Basher and Steven J Hallam. Leveraging heterogeneous network embedding for metabolic pathway prediction. *bioRxiv*, feb 2020.

[2] Ron Caspi, Richard Billington, Luciana Ferrer, Hartmut Foerster, Carol A. Fulcher, Ingrid M. Keseler, Anamika Kothari, Markus Krummenacker, Mario Latendresse, Lukas A. Mueller, Quang Ong, Suzanne Paley, Pallavi Subhraveti, Daniel S. Weaver, and Peter D. Karp. The metacyc database of metabolic

pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic Acids Research*, 44(D1):D471–D480, 2016.

[3] Ron Caspi, Richard Billington, Hartmut Foerster, Carol A Fulcher, Ingrid Keseler, Anamika Kothari, Markus Krummenacker, Mario Latendresse, Lukas A Mueller, Quang Ong, et al. Biocyc: Online resource for genome and metabolic pathway analysis. *The FASEB Journal*, 30(1 Supplement):lb192–lb192, 2016.

[4] Niels W Hanson, Kishori M Konwar, Alyse K Hawley, Tomer Altman, Peter D Karp, and Steven J Hallam. Metabolic pathways for the whole community. *BMC genomics*, 15(1):1, 2014.

[5] John P McCutcheon and Carol D Von Dohlen. An interdependent metabolic patchwork in the nested symbiosis of mealybugs. *Current Biology*, 21(16):1366–1372, 2011.

[6] Alexander Sczyrba, Peter Hofmann, Peter Belmann, David Koslicki, Stefan Janssen, Johannes Dröge, Ivan Gregor, Stephan Majda, Jessika Fiedler, Eik Dahms, et al. Critical assessment of metagenome interpretation—a benchmark of metagenomics software. *Nature methods*, 14(11):1063, 2017.

[7] Frank J Stewart, Adrian K Sharma, Jessica A Bryant, John M Eppley, and Edward F DeLong. Community transcriptomics reveals universal patterns of protein sequence conservation in natural microbial communities. *Genome biology*, 12(3):R26, 2011.

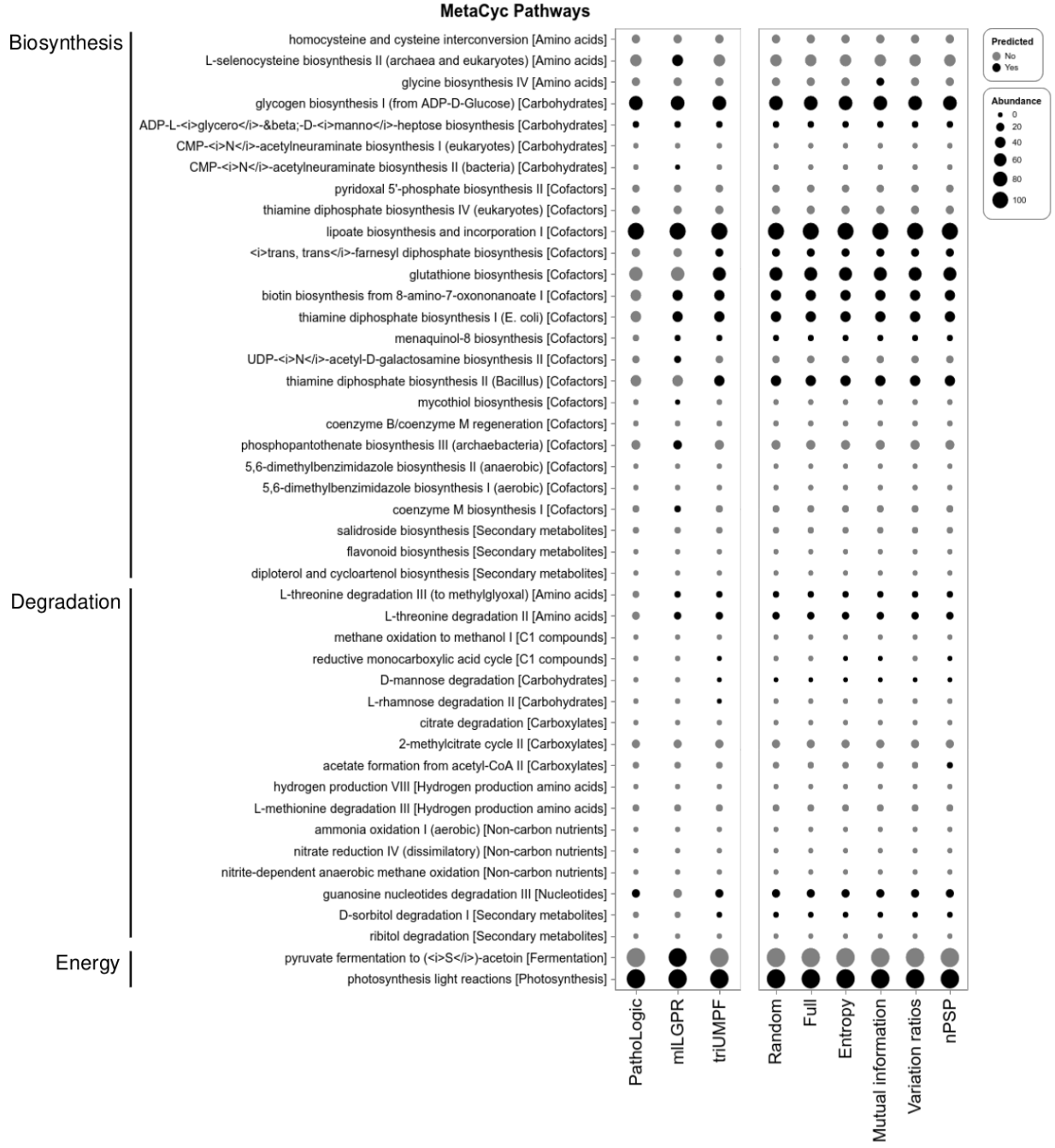


Figure 2: Comparative study of predicted pathways for HOTS 25m dataset between PathoLogic, mlLGPR, triUMPF, and leADS (with random sampling, full data, and four acquisition functions). Black circles indicate predicted pathways by the associated models while grey circles indicate pathways that were not recovered by models. The size of circles corresponds the pathway abundance information.

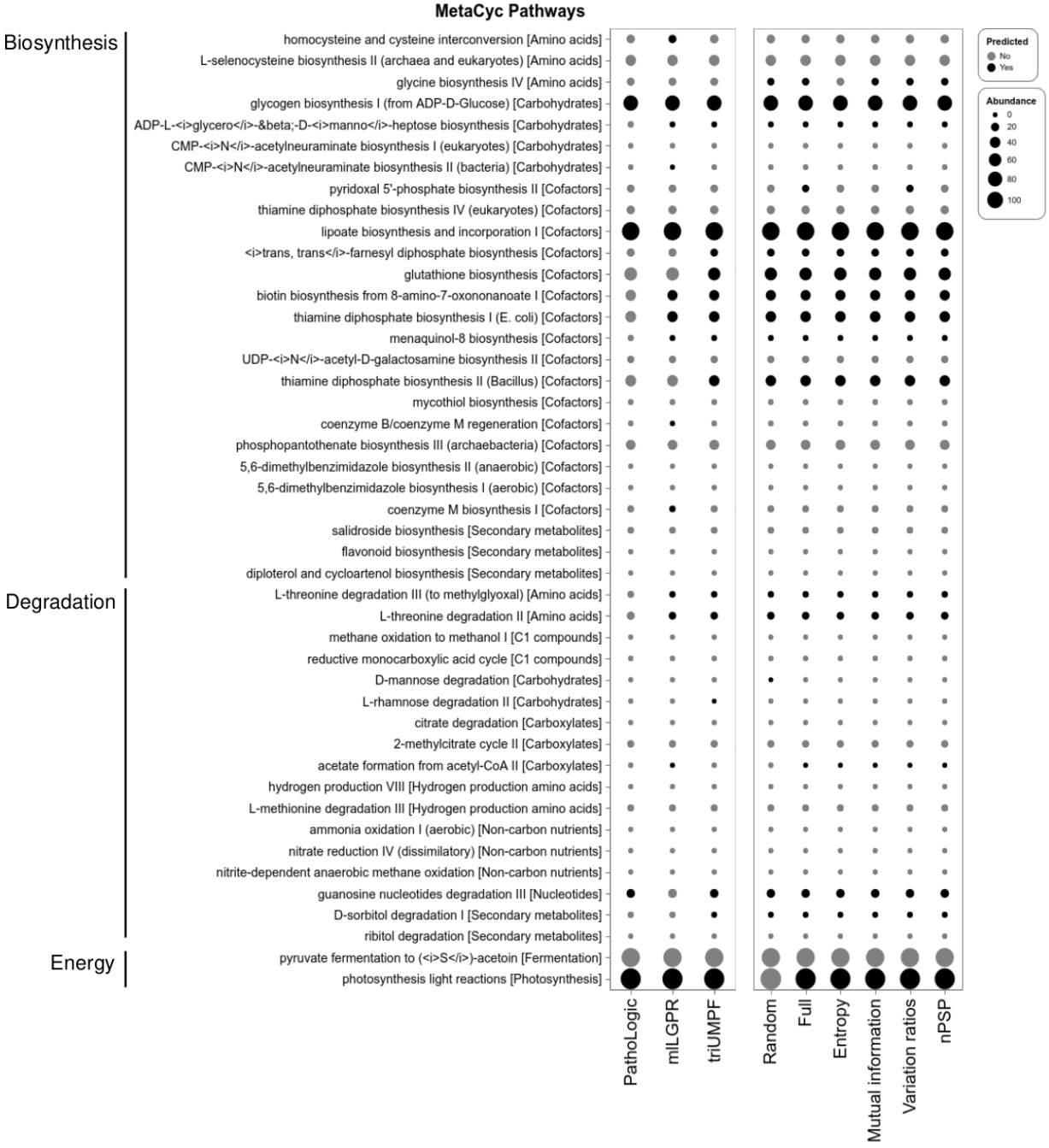


Figure 3: Comparative study of predicted pathways for HOTS 75m dataset between PathoLogic, mlLGPR, triUMPF, and leADS (with random sampling, full data, and four acquisition functions). Black circles indicate predicted pathways by the associated models while grey circles indicate pathways that were not recovered by models. The size of circles corresponds the pathway abundance information.

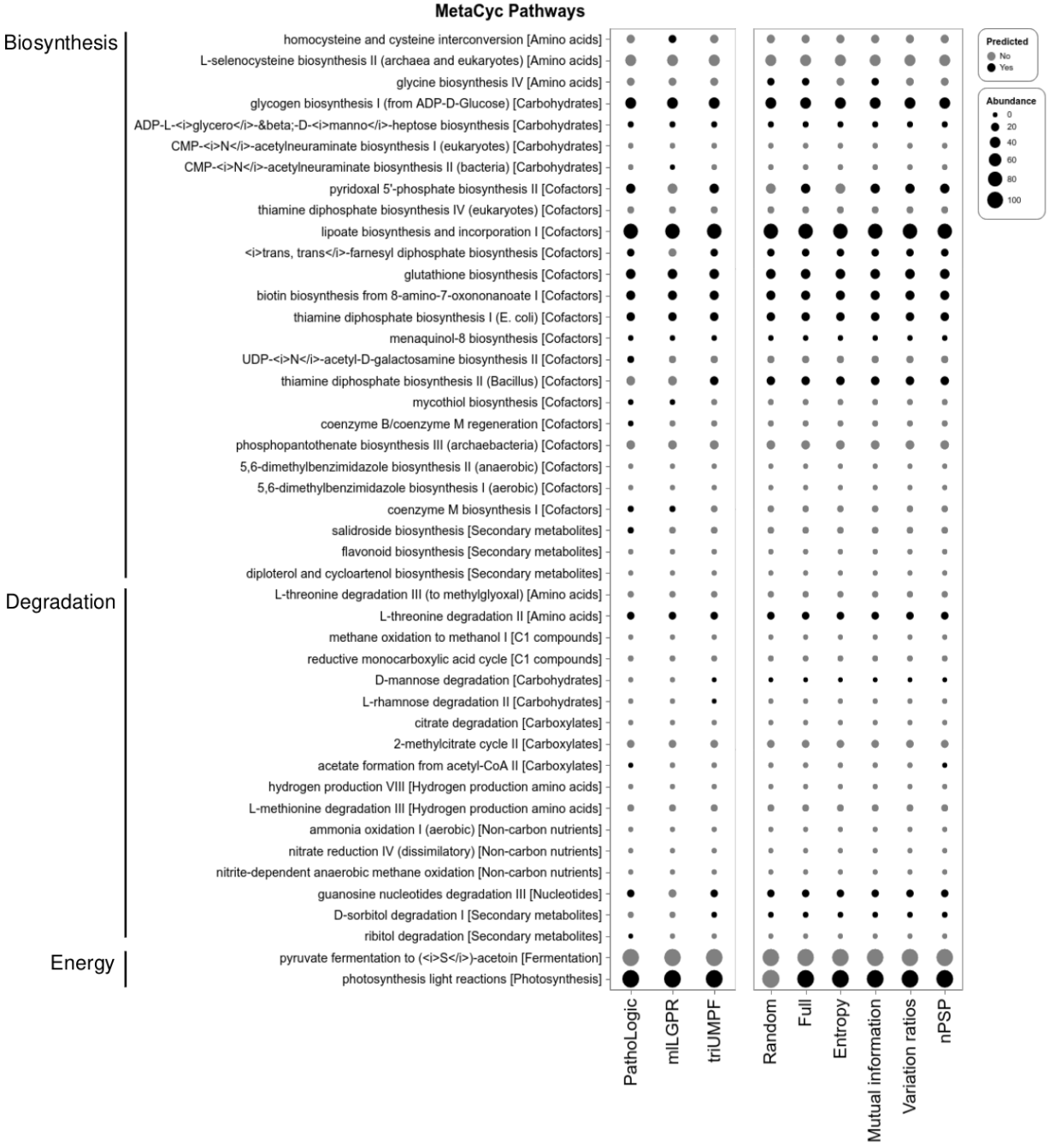


Figure 4: Comparative study of predicted pathways for HOTS 110m dataset between PathoLogic, mlLGPR, triUMPF, and leADS (with random sampling, full data, and four acquisition functions). Black circles indicate predicted pathways by the associated models while grey circles indicate pathways that were not recovered by models. The size of circles corresponds the pathway abundance information.

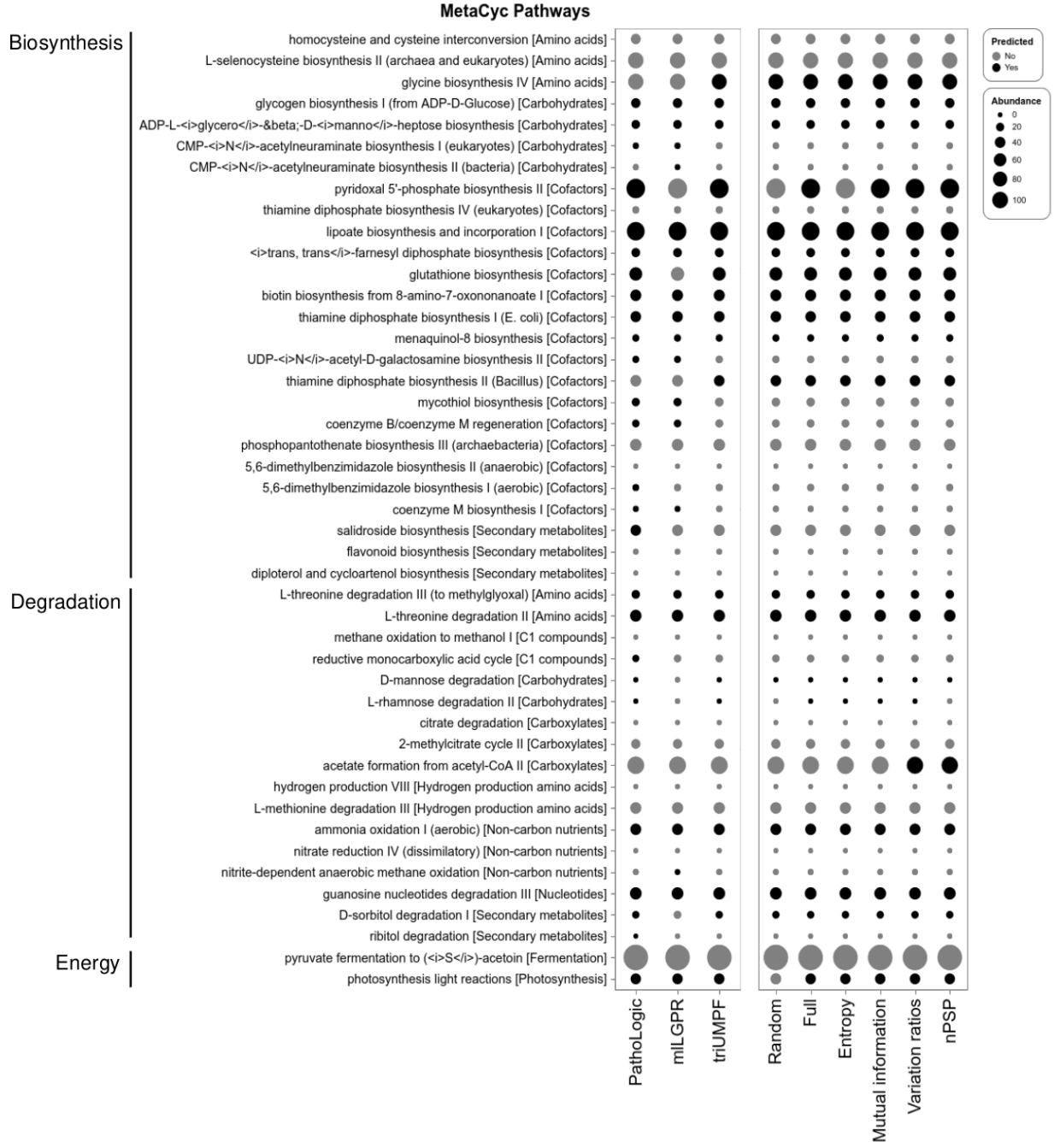


Figure 5: Comparative study of predicted pathways for HOTS 500m dataset between PathoLogic, mlLGPR, triUMPF, and leADS (with random sampling, full data, and four acquisition functions). Black circles indicate predicted pathways by the associated models while grey circles indicate pathways that were not recovered by models. The size of circles corresponds the pathway abundance information.

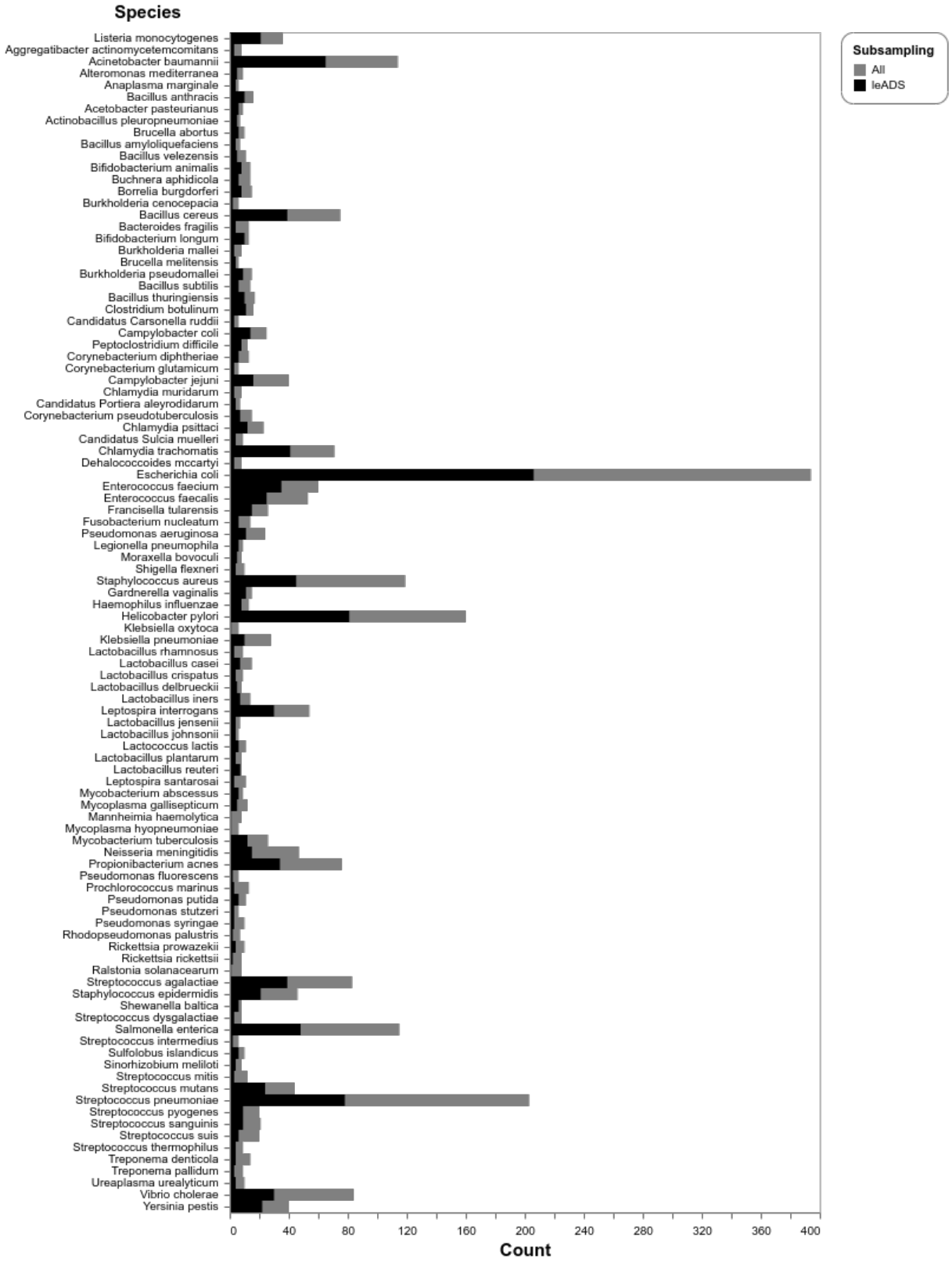


Figure 6: Samples corresponding top 100 species in BioCyc T2 &3. The black colored bars represent leADS+$\mathcal{H}$ (**per**% = 70%) selected samples while the grey colored bars indicates an overall number of samples associated with species in BioCyc T2 &3. For example, leADS+$\mathcal{H}$ selected 48 *Salmonella* related instances (out of 115) comprising a total of 496 distinct pathways (out of 548 distinct pathways).

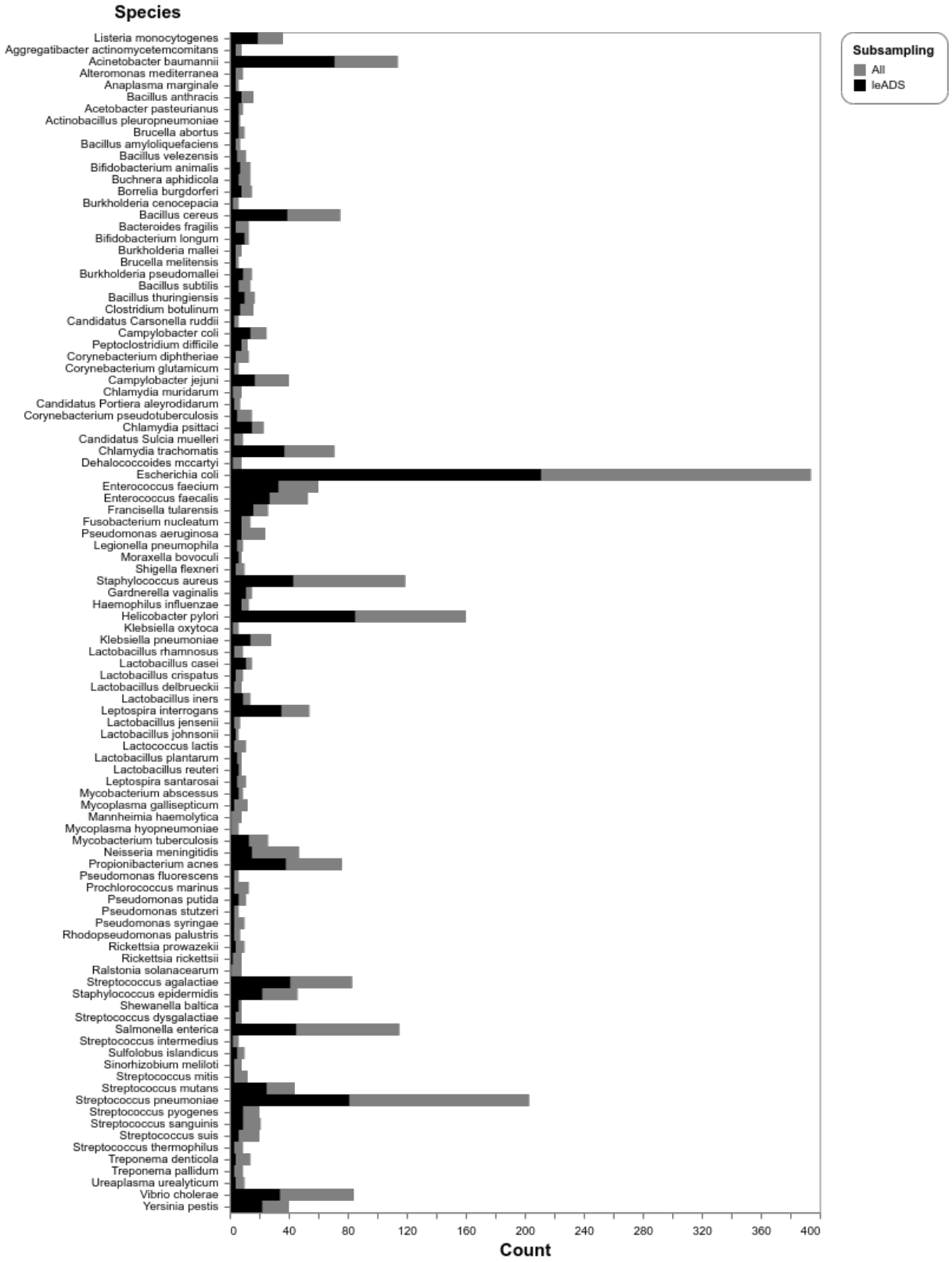


Figure 7: Samples corresponding top 100 species in BioCyc T2 &3. The black colored bars represent leADS+$\mathcal{M}$ (**per**% = 70%) selected samples while the grey colored bars indicates an overall number of samples associated with species in BioCyc T2 &3. For example, leADS+$\mathcal{M}$ selected 45 *Salmonella* related instances (out of 115) comprising a total of 494 distinct pathways (out of 548 distinct pathways).

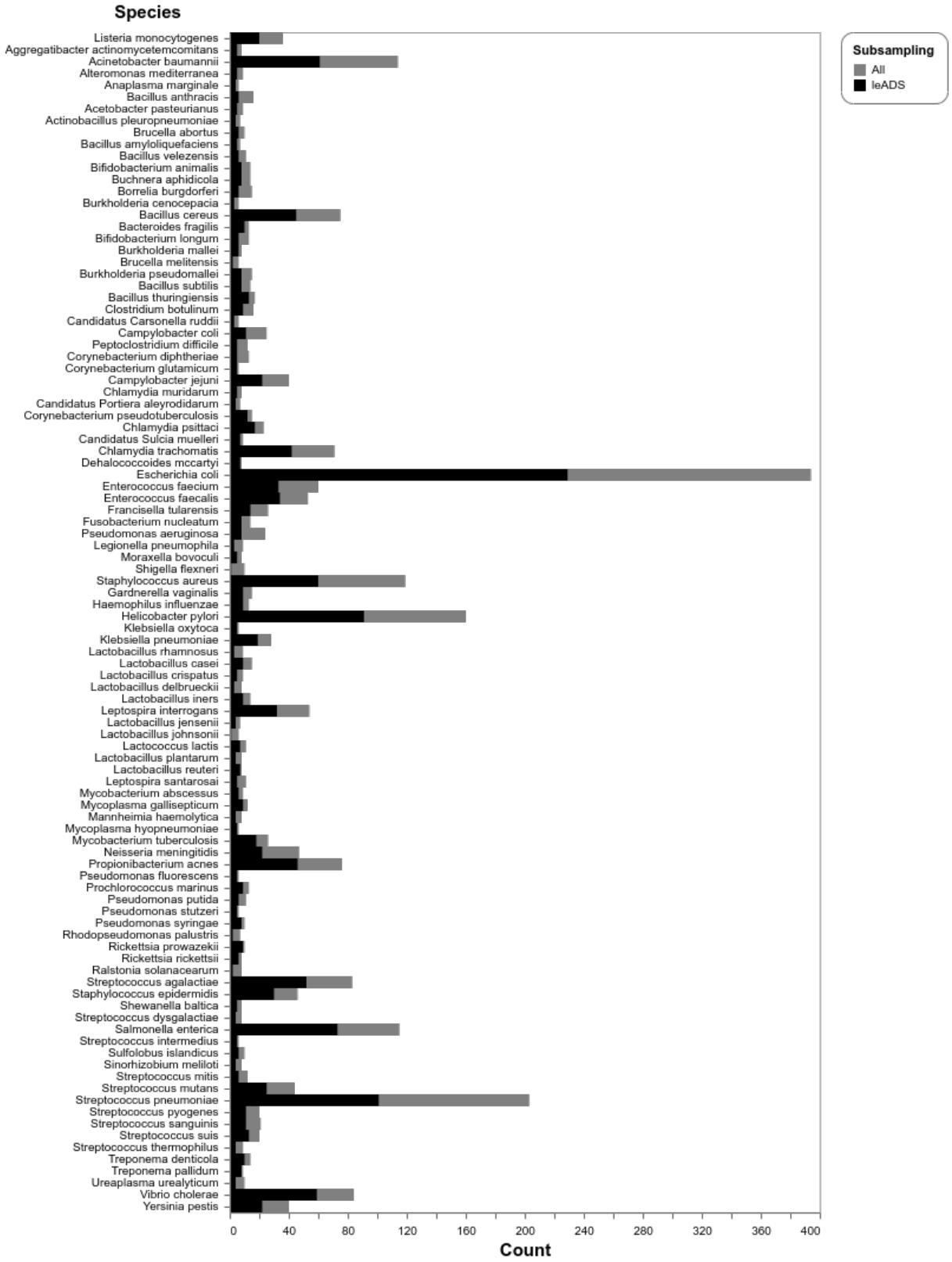


Figure 8: Samples corresponding top 100 species in BioCyc T2 &3. The black colored bars represent leADS+$\mathcal{V}$ (**per**% = 70%) selected samples while the grey colored bars indicates an overall number of samples associated with species in BioCyc T2 &3. For example, leADS+$\mathcal{V}$ selected 73 *Salmonella* related instances (out of 115) comprising a total of 537 distinct pathways (out of 548 distinct pathways).

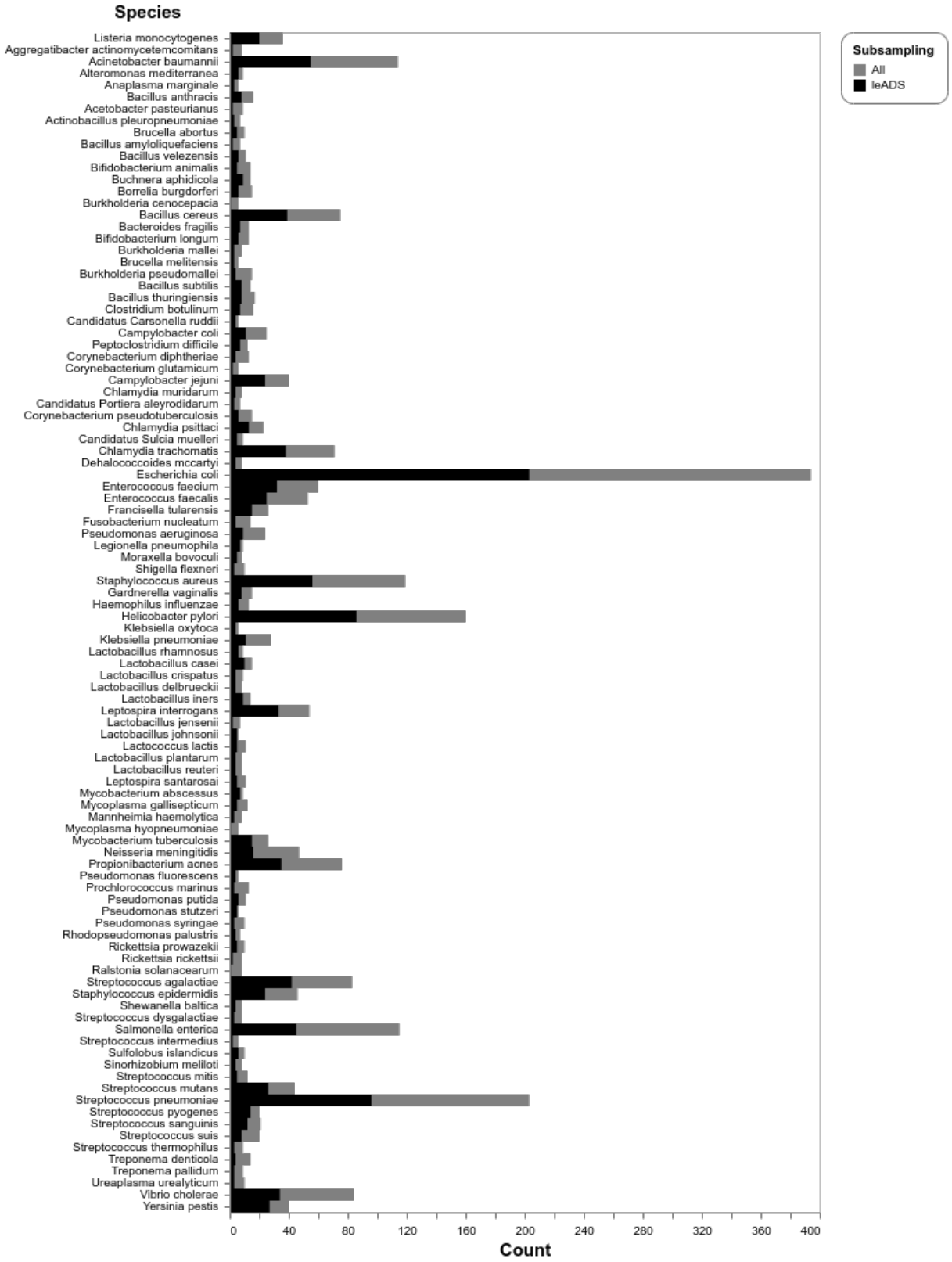
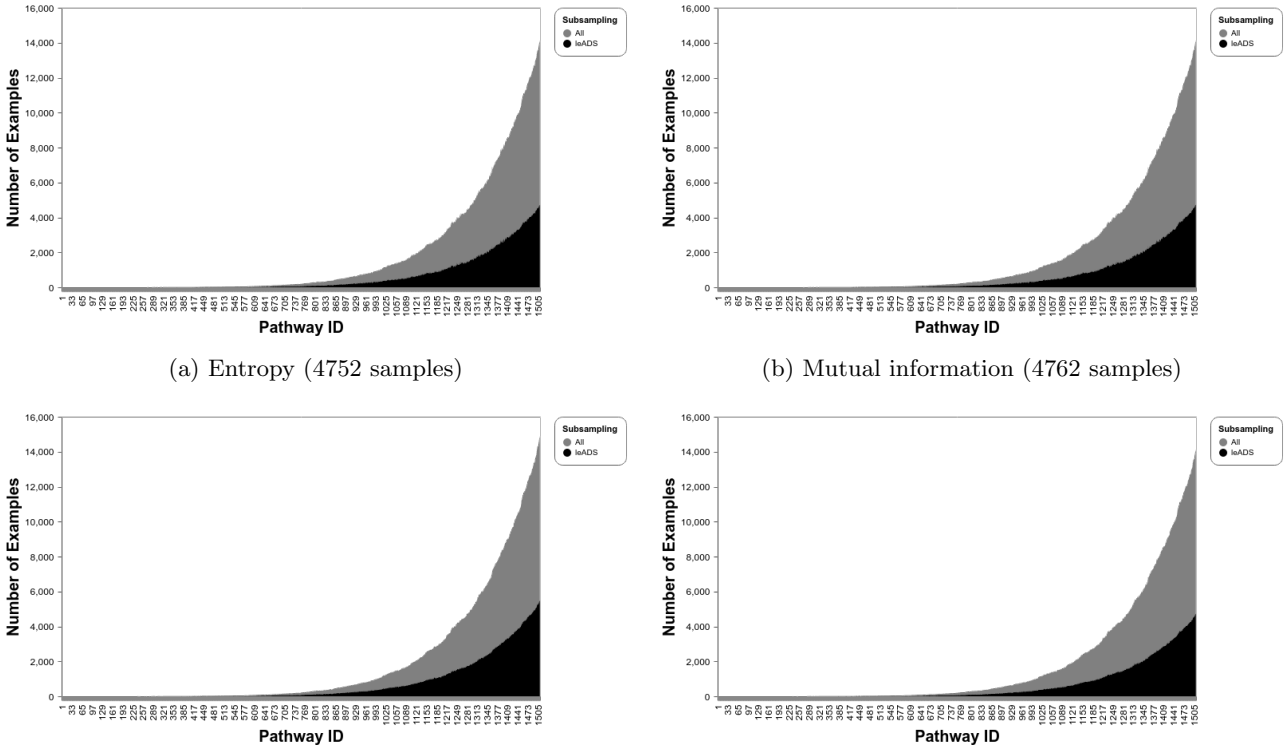


Figure 9: Samples corresponding top 100 species in BioCyc T2 &3. The black colored bars represent leADS+nPSP (**per**% = 70%) selected samples while the grey colored bars indicates an overall number of samples associated with species in BioCyc T2 &3. For example, leADS+nPSP selected 45 *Salmonella* related instances (out of 115) comprising a total of 505 distinct pathways (out of 548 distinct pathways).



(a) Entropy (4752 samples)

(b) Mutual information (4762 samples)

(c) Variation ratios (5506 samples)

(d) nPSP (4752 samples)

Figure 10: The number of reduced samples for each pathway in BioCyc T2 &3 data. For each figure, the horizontal axis indicates the indices of pathways while the vertical axis represents the number of associated examples in BioCyc T2 &3 collection. The black colored area represents leADS ($\mathbf{per}\% = 70\%$) selected instances while the grey colored area indicates an overall number of samples for pathways in BioCyc T2 &3.