Naïve Bayes Classifier as an Out-of-Distribution Detector of Novel Taxa

Keyush Ramjattun[1] and Gail L. Rosen[1]

[1]**Ecological and Evolutionary Signal Processing and Informatics (EESI) Laboratory, Drexel University, 3141 Chestnut St., 19104, Philadelphia, PA, USA**

Corresponding author:
Gail L. Rosen[1]

Email address: glr26@drexel.edu

ABSTRACT

Detecting sequences from novel taxa remains a key challenge in metagenomic classification, as reference databases rarely capture the full extent of microbial diversity. We investigate the Naïve Bayes Classifier (NBC++) as an out-of-distribution (OOD) detector by analyzing its log-likelihood scores across simulated and real metagenomic datasets. By partitioning reference databases and introducing taxonomic novelty, we derive thresholds that distinguish known from unknown reads at multiple taxonomic levels. These thresholds remain consistent across database sizes, indicating that once a lineage is represented, novelty detection performance stabilizes. Applied to a human gut metagenome, the thresholds reflect differences in database density and classification confidence. This work characterizes how NBC++ responds to novelty and illustrates its use in evaluating unclassified metagenomic reads.

INTRODUCTION

Metagenomics has transformed our ability to profile microbial communities, yet one of the biggest challenges is detecting sequences originating from novel taxa—organisms absent from existing reference databases. It remains a difficult problem due the challenge of identifying something unknown from a reference point of your known. Compounding that, is the taxonomic hierarchy of kingdoms, where each clade may be evolving at a different evolutionary rate. In machine learning, this problem is known as "out-of-distribution" (OOD) detection. Large-scale surveys consistently reveal that a meaningful fraction of metagenomic reads cannot be classified systematically analyzed human metagenomes and found that reads, which appeared across multiple studies, resulted in "unknown contigs" that assembled into novel viral genomes that bear no similarity to sequences in the public databases. Similar evidence is reported in global-scale metagenomic studies, which highlight extensive "functional dark matter" in the form of uncharacterized protein families. These findings show that unclassified sequences represent real biological diversity rather than random sequencing artefacts. Metagenomic taxonomic tools have tried to mitigate novelty detection. In Kraken2, for example, any query sequence that does not contain "enough k-mers" (a threshold that can be defined) belonging to a taxa in the reference database cannot be classified and is therefore marked as "unclassified." Choosing the appropriate confidence threshold is a trade-off between increasing accuracy vs. the proportion of classified reads that are lost. Also, the percent of reads that could be classified increases as a function of the database size. When NBC incorporates minimizers like Kraken2, it is able to obtain a reasonable confidence threshold to filter unknown reads. However, "unknown reads" were only simulated from the *Arabidopsis thaliana* plant, which may limit the feasibility on unknown microbial sequences. Therefore, a natural problem to benchmark metagenomic classifiers is generating feasible datasets that take into consideration various types of novelty. Nickols et al. tested over different types of novelty: a) taxa missing from a database, newly added NCBI taxa since the last database update, and completely novel species genome bins (SGBs) without NCBI IDs. Taxonomic classifier methods performed better on higher levels of taxonomy than lower levels, especially for real samples. One of Nickols et al.'s suggestions is to use MAGs in reference databases, which is also supported by another study. As more genomes are added to a database, they do not necessarily represent more diversity. Therefore, it is important to have a way to measure diversity growth within

clades and pangenomes. A delta measure, based on compression, has been introduced and extended to suggest optimal k-mers, which can be critical to minimizer hash-based methods. In this study, we examine how well the metagenomic NBC++ can identify novel taxa. NBC is competitive and especially, for 16S, newer deep learning architectures cannot outperform it. While there are other classifiers, we want to experiment with NBC++'s score, a log-likelihood probability, that can be used to intelligently compare the probability of sequence reads against a database, and see how well it extends to novelty detection of various taxa. This will be vital to NBC++, since it is an incremental-class learner and to incrementally learn to distinguish between a growing number of objects or classes will be vital, especially with novel classes being identified before being integrated.

METHODOLOGY

Our goal was to derive a log-likelihood threshold that can be used to identify novel taxa not present in a reference database. The general workflow involved partitioning a database, training NBC, classifying a test set, and calculating a threshold using ROC analysis. This process was applied independently to a "basic" database containing one genome per species and a much larger "extended" database with multiple genomes per species, each requiring specific variations during the data preparation stage.

1. **Initial Database Filtering:** Both databases were first prepared at the phylum, class, order, and family level. To ensure adequate representation, classes with a low number of genomes were removed. The minimum threshold was set to 30 representatives for the "basic" database and 400 for the "extended" database.

2. **Training and Holdout Set Partitioning:** To create "known" (training) and "unknown" (holdout) sets for the ROC analysis, the remaining genomes were partitioned. This step differed significantly between the two databases to account for their different structures:

   - For the "basic" database: 50% of the species were randomly selected to form the training set, with the remaining 50% forming the holdout set.

   - For the balanced "extended" database: A two-step balancing approach was used. First, 50% of the classes were randomly selected. Then, from within each of those selected classes, 400 genomes were randomly sampled to create a balanced training set. All genomes from the unselected classes formed the holdout set.

   This entire partitioning process was repeated five times for each database to generate independent trials for statistical robustness.

3. **Model Training and Classification:** For each trial, NBC was trained on the corresponding training set for various k-mer lengths (3, 6, 9, 12, 15). These models were then used to classify a standardized test set containing reads from both the training ("known") and holdout ("unknown") sets, generating a log-probability score for each read.

4. **Threshold Derivation:** An ROC analysis was performed using the log-probability scores to distinguish between the "known" and "unknown" reads. The optimal novelty threshold was identified as the point on the ROC curve that maximizes the Youden's J statistic.

5. **Real-world validation:** The derived 9-mer threshold were applied to classify reads from a real human gut metagenomic sample to evaluate the method's practical performance.

Data Collection and Database Construction

This study uses the Refseq retrieval pipeline developed as a part of the Woltka software to sample the NCBI RefSeq database at high microbiome diversity at varying degrees of depth, and then test the performance of NBC++ on the resulting databases. The following databases with distinct profiles were created:

- "Basic" database: comprising one genome per genus (with exceptions of Baxterfoxvirus and Betanucleorhabdovirus) resulting in a compilation of 4,634 genomes as of July 24, 2023.

- "Extended" database: featuring one genome per species with a Latinate name and higher ranks, accumulating 319,554 genomes by July 26, 2023. The extended database also includes reference, representative, and type material genomes. The genomes were then grouped by species resulting in: **4,634 species classes** for "Basic," and **58,978 species classes** for "Extended."

Data Sources and Experimental design

Simulated Data: For all simulated experiments, we prepare the data at four taxonomic levels (phylum, class, order, family). To ensure sufficient representation for training, classes with a low number of genomes are excluded. This filtering threshold is set to 30 genomes for the basic database and 400 genomes for the extended database.

After the initial filtering, the remaining genomes were partitioned to create "known" (training) and "unknown" (holdout) sets for the subsequent ROC analysis. The partitioning method differed between the two databases to account for their distinct structures. For the basic database, which contains only one genome per species, a simple random split was performed where 50% of the species were assigned to the training set and the other 50% formed the holdout set. In contrast, the extended database, with multiple genomes per species, required a different approach to ensure the training set was balanced. For these experiments, 50% of the taxonomic classes were first randomly selected, and then 400 genomes were randomly sampled from within each of those selected classes. All genomes from the unselected classes were subsequently designated as the holdout set. This random partitioning is performed five times, creating five independent trials.

Next, for each of the five trials and for each k-mer length tested ($k = 3, 6, 9, 12, 15$), NBC++ was trained using the genomes from the respective training set.

To derive a novelty threshold, we scored simulated reads generated from all reference species (1 genome for Basic and many genomes per species for Extended) as in **?**. Therefore, a proportion of the test reads were from species in the training database, and many were from species out-of-distribution. The InSilicoSeq simulator generated 100 reads per species. Then, each simulated read was classified using the trained models to obtain a log-probability score – the log-probability of that read given each training species. Then the read was assigned the maximum likelihood probability as the final score. If the read was computed on a training database that contained the species belonging to the same taxa as the read, its max log-probability score was put into a "known bin". Conversely, if the read's taxa was not in the training database, the maximum likelihood score was assigned to an "unknown bin".

ROC analysis is then performed to find the optimal threshold to separate the "known" reads, that are related to the training data, from the "unknown" reads, from the holdout set. The optimal log-probablity threshold is then determined as the point on the ROC curves that maximized Youden's J statistic.

Real Human Gut Metagenomic Sample: To validate our method, we analyze a human gut metagenomic dataset described in [. In contrast to the fixed-length sequences used for model training (126 bp), the real-world sample consists of reads with varying lengths. To address this, we first apply a length-normalization to the novelty threshold for each read. We then classify each normalized read using the 9-mer models derived from both the basic and extended databases. This process allowed us to assess the practical performance of our novelty detection thresholds and directly compare the models' outputs on complex, real-world data.

RESULTS AND DISCUSSION

We evaluate the performance of our novelty detection method across a range of key parameters, specifically investigating the impact of k-mer length and database scale. The results of this evaluation are then used to assess our framework's practical application on the human gut metagenome.

K-mer length and classification accuracy

We assessed NBC's ability to distinguish "known" sequences from holdout "unknown" sequences. A noticeable trend emerged for both the basic and extended databases: as the k-mer length increased from 3 to 15, the classifier's performance improved significantly. There is a consistent increase in the Area Under Curve (AUC) values from the ROC analysis, confirming that longer, more specific k-mers provide greater distinguishing power for classification (Fig. 1).

When comparing the two databases, the model trained on the basic database often yielded slightly higher AUCs than the model trained on the extended database. This suggests that while both models

improve with longer k-mers, the classification task itself becomes more challenging with a denser, more complex reference set. With many more closely related species in the extended database, the potential for ambiguity increases, leading to a slightly lower, but likely more realistic, measure of classification performance.

Novelty thresholds across Database Scales

After determining the optimal log-probability for each model using Youden's J statistic, we compared the values derived from the basic database against those from the extended database.

Despite the vast difference in size and taxonomic complexity between the two databases, the calculated novelty thresholds were remarkably stable (Fig. 2). For any given k-mer length and taxonomic level, the threshold value remained nearly identical regardless of which database was used for training. This finding suggests that the novelty thresholds (at various taxonomic levels) do not significantly change with database depth – as long as a taxonomic lineage has representation, the threshold value for determining that it is "in-distribution" remains fairly fixed.

Application to a Human Gut Metagenome

To assess the practical performance of our framework, we applied the 9-mer models from both the basic and extended databases to classify reads from a real human gut metagenome.

Interestingly, the model trained on the smaller basic database classified a higher proportion of reads as "known" (98.09%) compared to the extended model (92.77%). To confirm this finding was driven by the models' intrinsic properties and not minor variations in their thresholds, a sensitivity analysis was performed. Applying a single, common threshold to both models' output consistently resulted in the basic model classifying more reads as "known". (Table 1.)

Threshold Used	Model Used	% Known
-1049.62 (extended phylum)	Extended	92.77
-1049.62 (extended phylum)	Basic	96.41
-1053.81 (basic phylum)	Extended	95.35
-1053.81 (basic phylum)	Basic	98.09

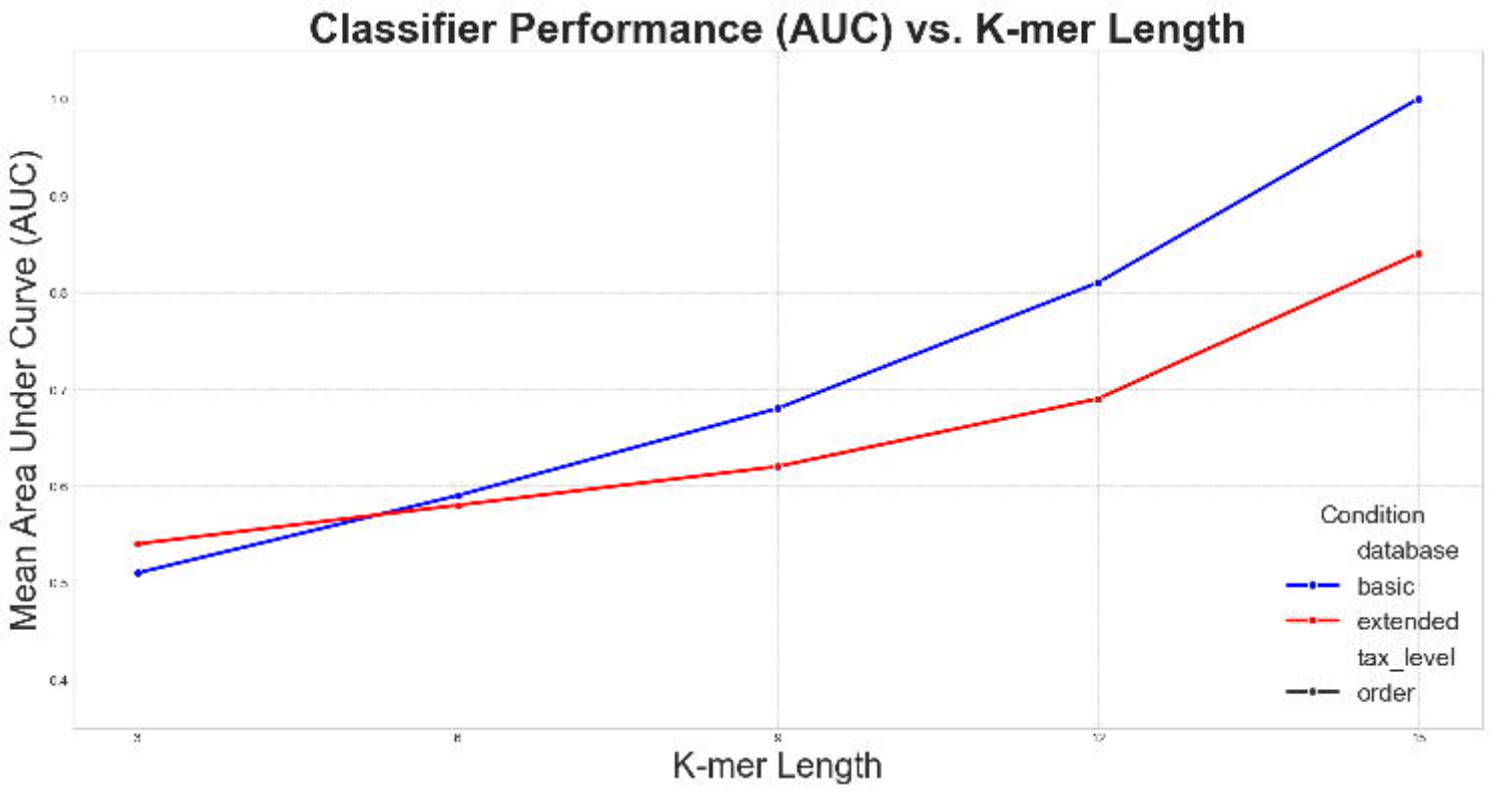
Table 1. Proportion of reads classified as "known" under different thresholds and models.

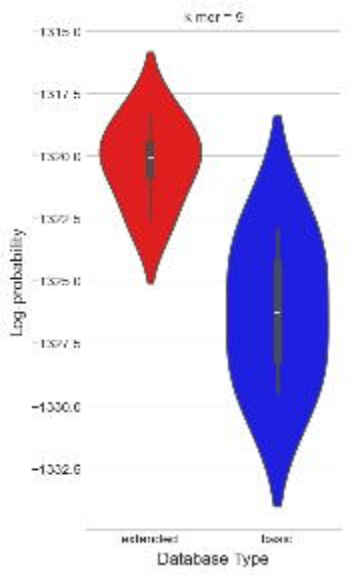
To better understand the mechanism behind this disparity, we found that 75% of the reads classified by the extended model produced a more negative log-probability score than the basic model, indicating a lower classification confidence in its predictions (Fig. 3).

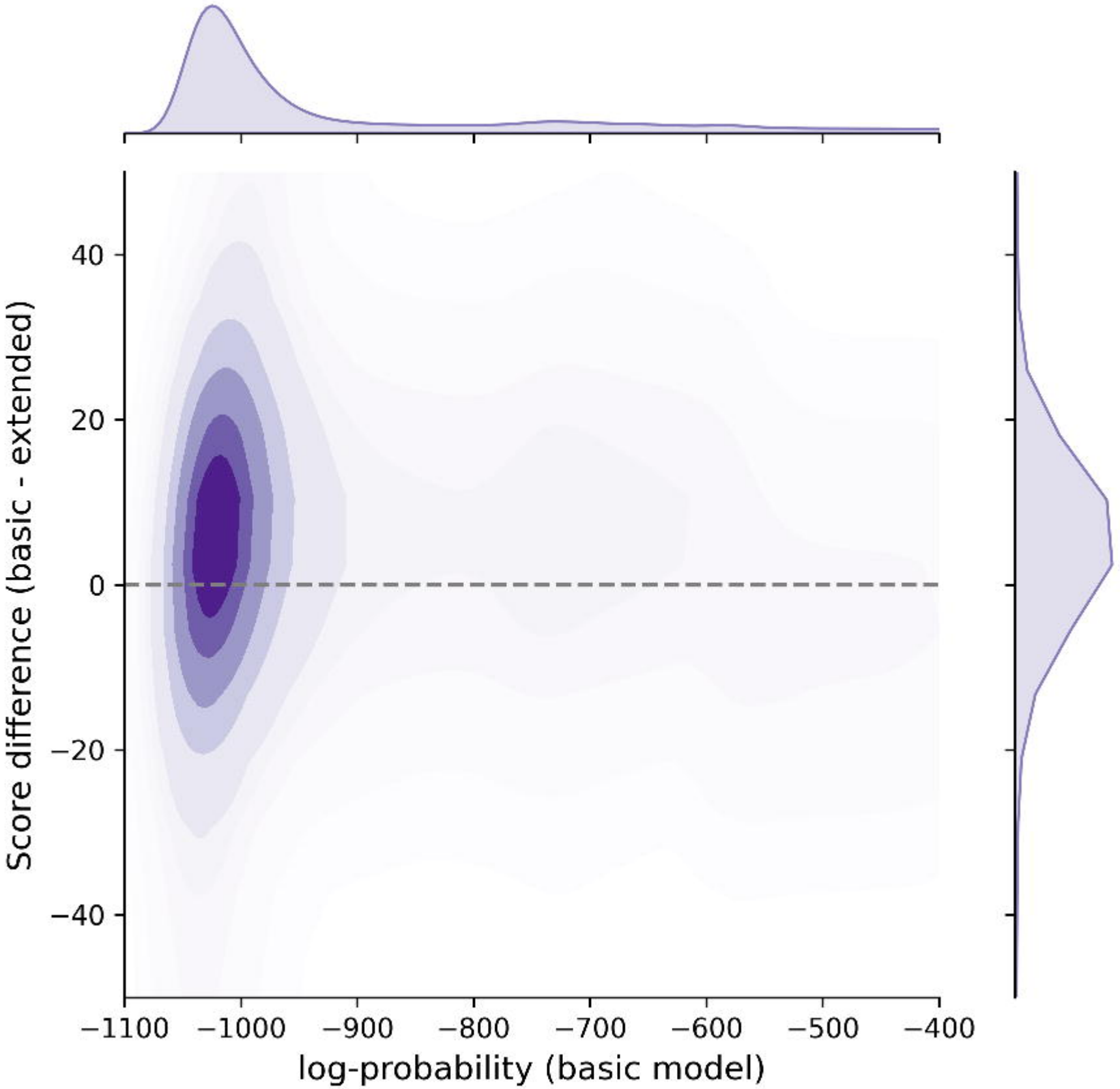
Taken together, these results provide strong evidence that the extended model operates with a higher standard of certainty. This higher standard is a direct result of its training on a denser taxonomic space, leading to a more conservative and likely more accurate estimate of the sample's "known" fraction.

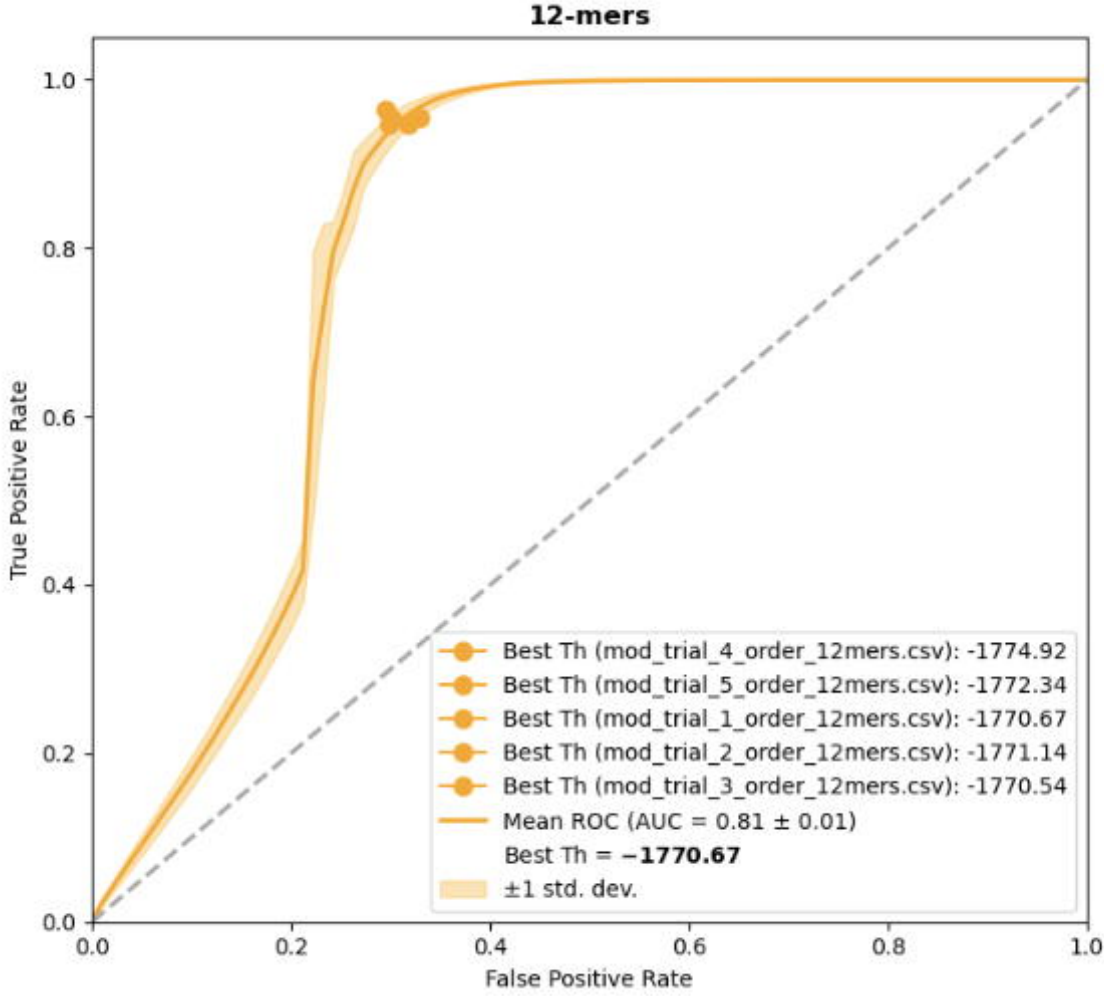
Determining Optimal Thresholds

To pinpoint the most effective log-probability score for identifying novel sequences, we performed a Receiver Operating Characteristic (ROC) analysis on each model's output. We generated an ROC curve using the scores from our test set, treating reads from the training set as the "known" positive class and reads from the holdout set as the "unknown" negative class. The optimal threshold was then identified as the point on this curve that maximized the **Youden's J statistic**, calculated as $J = \text{sensitivity} + \text{specificity} - 1$. This process was repeated for each of the five independent trials for a given experimental condition (e.g., order level with 12-mers). The final, definitive threshold reported for that condition was then taken from the single trial that yielded the highest Youden's J statistic, representing the best classification performance observed across the replicates.


Classifier Performance (AUC) vs. K-mer Length
Mean Area Under Curve (AUC)
K-mer Length
Condition
database
basic
extended
tax_level
order


k mer = 9
-1315.0
-1317.5
-1320.0
-1322.5
-1325.0
-1327.5
-1330.0
-1332.0
Log probability
extended
base
Database Type


40
20
0
−20
−40
Score difference (basic - extended)
−1100
−1000
−900
−800
−700
−600
−500
−400
log-probability (basic model)


12-mers
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Best Th (mod_trial_4_order_12mers.csv): -1774.92
Best Th (mod_trial_5_order_12mers.csv): -1772.34
Best Th (mod_trial_1_order_12mers.csv): -1770.67
Best Th (mod_trial_2_order_12mers.csv): -1771.14
Best Th (mod_trial_3_order_12mers.csv): -1770.54
Mean ROC (AUC = 0.81 ± 0.01)
Best Th = −1770.67
±1 std. dev.