TabRiboICL for read-specific P-site prediction and a unified framework for benchmarking P-site assignment in ribosome profiling

Eric Zhuang
Great Neck South High School
ezhuang2@student.gn.k12.ny.us

Manoj Kandpal
Rockefeller University
mkandpal@rockefeller.edu

Abstract

Ribosome profiling sequences messenger RNA (mRNA) fragments protected by ribosomes. Each sequenced fragment is represented as a read. Where a read aligns shows the region occupied by a ribosome, but it does not directly show which nucleotide lies at the ribosome's peptidyl site (P site). Identifying that nucleotide is necessary for analyses of individual codons, the three-nucleotide units read during translation. Existing P-site predictors require different inputs, define positions differently, report either one position per read or a signal summarized across many reads, and vary in whether they use information from the experimental sample being evaluated. These differences make fair comparison difficult.

We developed TabRiboICL, a novel predictor that assigns one P-site position to each read. It evaluates 13 possible positions located 8–20 nucleotides from the read's 5′ end. The published TabICL machine-learning model [10] scores each possible position. TabRiboICL combines each score with how often the corresponding distance occurred among similar training reads with expected P-site positions and then selects one P-site position. We also developed a benchmarking framework that converts the different forms of output to the same definition of a read-level P-site position and evaluates every predictor on the same reads. The benchmark includes 8 public datasets, 31 ribosome-profiling samples from *Drosophila melanogaster* and *Saccharomyces cerevisiae*, and 11 predictors. An expected P-site position was assigned to a read near the annotated beginning or end of a protein-coding region only when exactly one of the 13 possible positions matched the relevant boundary and fell on the first nucleotide of an annotated codon. Because these expected positions were inferred rather than directly observed, we call them proxy labels. Exact accuracy, the primary measure, required a prediction to match its proxy label. Accuracy within one nucleotide was the secondary measure.

During the held-out dataset experiment, TabRiboICL had the highest observed mean exact accuracy after every dataset was given equal weight: 37.6% (95% bootstrap confidence interval, 32.5%–43.8%). The interval was calculated by repeatedly redrawing datasets, samples, and reads. TabRiboICL had the highest observed exact accuracy in 5 of 8 datasets; Scikit-ribo had the highest in two datasets and LightGBM in one.

The framework also characterized predictor performance across datasets, organisms, and recorded protocol categories, providing a consistent, well-documented basis for comparing P-site predictors.

Keywords: ribosome profiling; Ribo-seq; P-site assignment; read-specific prediction; machine learning; computational benchmarking

Introduction

Ribosome profiling (Ribo-seq) sequences messenger RNA (mRNA) fragments protected by translating ribosomes to study translation at near-codon resolution [1,2]. Sequencing each protected fragment produces a read. A read identifies the region occupied by a ribosome, but it does not directly show which nucleotide lies at the ribosome's peptidyl site (P site), which holds the transfer RNA carrying the growing protein chain. That position must be inferred from the ends and length of the read, its reading frame (the grouping of nucleotides into successive three-nucleotide codons), its sequence, its genomic alignment (the location where the sequence matches the reference genome), and the experimental protocol. The adjacent aminoacyl site (A site), where the next amino acid is delivered during protein synthesis, lies one codon downstream of the P site in the direction of translation. An A-site prediction must therefore be shifted by one codon before it can be compared with a P-site position. P-site assignment influences measurements of the repeating three-nucleotide pattern of translating ribosomes, ribosome accumulation at particular codons, patterns near the beginning and end of translation, the movement of ribosomes along mRNA as proteins are made, and the identification of translated open reading frames, which are continuous nucleotide sequences with the potential to encode a protein or peptide.

Existing predictors infer P-site positions in several ways. One simple approach applies the same distance from the read end to every read, whereas another estimates a separate distance for each read length. This distance from a read end to the P site is commonly called an offset. Some predictors combine signals from many read ends near the boundaries of an annotated coding sequence, the part of a transcript marked as being translated into protein. Others mathematically separate overlapping signals to estimate positions or ribosome abundance at individual codons, require predictions to follow predefined structural rules, or predict each read

from measurements specific to that fragment [3–9,13,15]. Machine learning, in which a model learns patterns from examples with expected answers, has therefore already been used for read-specific and codon-level prediction. The question addressed here is how complete predictors perform when their required inputs, learned models, final rules for choosing a position, and permitted access to data from the sample being evaluated are considered together.

Fair comparison is difficult because predictors require different forms of data and return different forms of prediction. Some use reads aligned to a genome or to a collection of transcript sequences, the RNA sequences produced from genes. Some use position-by-position profiles that combine signals from many reads, and others require a description of every possible P-site position for each read. A predictor may report a P site or an A site, measure the distance from the beginning (5′ end) or the end (3′ end) of a read, return a combined profile instead of a position for each read, or omit reads that do not meet its requirements. Some predictors adjust to the sample being evaluated by using its aligned reads even when the expected P-site positions for those reads are hidden; others depend only on training datasets for which expected positions are supplied. If each predictor were scored only on the reads for which it produced a result, the evaluated reads would differ among predictors. Similarly, substituting results from different code when the specified software produced no prediction would change the predictor being evaluated.

The organization of the data creates another challenge. Each experimental sample contains many reads, and one dataset identified by a public database accession can contain several samples. Testing selected reads from samples that also supplied training examples asks how well a predictor handles additional reads from familiar samples; it does not show how the predictor will perform on a separately collected dataset. Withholding all expected P-site positions from one dataset gives a stronger test on a new dataset, although some predictors may still use alignments or summaries calculated from that evaluated dataset without using its expected positions. The expected P-site positions are also not directly observed. They are inferred for reads aligned near annotated coding-sequence boundaries and consistent with the annotated reading frame, so we call them proxy labels. A sound benchmark should estimate uncertainty without treating every read as an independent experiment. Reads from the same sample or dataset share experimental conditions, so a large

number of reads from one dataset does not provide the same evidence as data collected from many different datasets [14].

This paper makes two related but distinct contributions. First, we developed TabRiboICL, a novel predictor that assigns one P-site position to each read. For every read, it examines 13 possible distances from the 5′ end. It describes each possible position using the read sequence, alignment information, nearby aligned reads, and recorded information about the sample. The published TabICL model [10] scores each possible position, and TabRiboICL combines those scores with how often each distance occurred among labelled training reads before selecting one position.

Second, we developed a unified framework for comparing P-site predictors. We used it to evaluate TabRiboICL, XGBoost [11], LightGBM [12], two simple comparison predictors that learn either one offset for all reads or one offset for each read length, and six previously described predictors. Separate evaluations withheld proxy labels for selected reads within represented samples, for an entire dataset, for one organism, or for all samples in one protocol category, defined from recorded RNA-digestion and library-preparation information.

The expected P-site positions were hidden during evaluation. However, predictors that normally calculate patterns from all aligned reads in a sample could still use the evaluated sample's alignments, without access to its proxy labels. In the primary evaluation, all proxy labels from one dataset were withheld from model development at a time, and all eight datasets received equal weight in the overall average. Among the 11 evaluated predictors, TabRiboICL had the highest observed mean exact accuracy and had the highest observed exact accuracy in five of the eight datasets.

Materials and Methods

Benchmark datasets, samples, and reads

The benchmark includes eight public Ribo-seq datasets, each defined by a public database accession identifier, and 31 individual Ribo-seq samples (Table 1; Figure 2). Seventeen samples are from *D. melanogaster* and 14 are from *S. cerevisiae*. Two conventional RNA-sequencing (RNA-seq) records

associated with PRJNA842067 were excluded because they were not ribosome-profiling experiments; they are therefore not part of the 31-sample total.

Throughout this paper, a **dataset** contains all Ribo-seq samples grouped under the same public database accession, a **sample** is one retained Ribo-seq sequencing run recorded within that dataset, and a **read** is one sequenced ribosome-protected fragment.

Accuracy was first calculated from the reads in each sample. Sample accuracies were then averaged within each dataset, with every sample receiving equal weight. For results combining multiple datasets, every dataset also received equal weight. This prevented datasets with more samples or reads from dominating the overall result.

We summarized the available experimental information by assigning each sample to exactly one of five descriptive **protocol categories**. Three categories record the enzyme used to digest RNA not protected by ribosomes: ribonuclease I (RNase I), micrococcal nuclease (MNase), or nuclease P1 (P1). A fourth is based instead on the recorded OTTR library-preparation method and does not identify a nuclease. **Unreported** means that neither a digestion-enzyme category nor the OTTR library-preparation category could be assigned from the source records. The categories therefore combine two different kinds of information and should not be interpreted as five levels of one experimental variable. They were also unevenly represented and strongly linked to particular datasets and organisms. We used them only to describe predictor performance when proxy labels from one category were withheld during evaluation; differences between categories were not interpreted as evidence that a protocol caused a change in performance.

Table 1. Public benchmark datasets and recorded protocol categories.

Dataset	Organism	Ribo-seq samples	Ribo-seq data in the Sequence Read Archive (gigabytes)	Protocol category (number of samples)
SRP033366	*D. melanogaster*	5	6.67	RNase I (5)
PRJNA842067	*D. melanogaster*	4	3.18	RNase I (4)
SRP067542	*D. melanogaster*	8	13.29	MNase (8)
SRP419250	*S. cerevisiae*	8	8.97	P1 (2), OTTR library preparation (1), Unreported (5)
GSE53268	*S. cerevisiae*	1	2.63	Unreported (1)
GSE63789	*S. cerevisiae*	1	2.06	Unreported (1)

Dataset	Organism	Ribo-seq samples	Ribo-seq data in the Sequence Read Archive (gigabytes)	Protocol category (number of samples)
GSE58321	*S. cerevisiae*	3	3.26	Unreported (3)
SRP038992	*S. cerevisiae*	1	6.24	Unreported (1)

Note: Two conventional RNA-seq records associated with PRJNA842067 were excluded and are not included in the sample or data-size totals.

Read processing, reference genomes, and sample information

Public sequencing records were obtained from the Sequence Read Archive using the accessions in Table 1. The records were converted to FASTQ files, which store read sequences and a score indicating confidence in each recorded nucleotide. Cutadapt [16] removed the 3′ adapter recorded for each sample, which is an artificial sequence added during library preparation. Where specified in the sample information, Cutadapt also removed a fixed number of nucleotides from the read's 5′ end. Reads shorter than 20 nucleotides or longer than 120 nucleotides after trimming were removed.

Bowtie 2 [17] was used to identify and remove reads that aligned to reference sequences annotated as noncoding RNAs. The remaining reads were aligned to the appropriate genome. SAMtools [18] organized the alignments and created index files that allowed alignments to be retrieved efficiently. For each mapped read, we kept only the alignment marked as its main genomic placement and excluded alternative or supplementary placements. Each retained read received a unique identifier formed by combining its public sequencing-run accession with the read name in the FASTQ file. The adapter sequence and any required 5′ trimming were recorded separately for every sample.

For *D. melanogaster*, we used the BDGP6.32 genome and the Ensembl Metazoa release 52 annotation. For *S. cerevisiae*, we used the R64-1-1 genome and the Ensembl release 115 annotation [19]. For each organism, the genome sequence, complementary DNA (cDNA) transcript sequences, noncoding-RNA sequences, and gene annotations in Gene Transfer Format (GTF), a standard tabular format for describing genes and transcripts, were stored and processed as separate reference files.

P-site positions and annotation-derived labels used for training and evaluation

For every read, the prediction target was the number of nucleotides from the read's 5′ end to its predicted P-site position. We call this distance the P-site offset. Whether the read was aligned in the forward or reverse orientation was taken into account when the offset was converted to a genomic position. We considered 13 offsets at one-nucleotide intervals, from 8 through 20 nucleotides. Thus, each read was evaluated at 13 possible P-site positions and could receive no more than one final P-site assignment.

Some previously described predictors produced A-site positions, and others produced a profile, meaning a position-by-position signal combined across many reads, rather than a separate P-site position for every read. A conversion procedure specific to each predictor translated these outputs into the same read-level P-site definition before evaluation.

The true P-site positions were not directly observed. We therefore constructed expected positions from gene annotations for use in training and evaluation. Because these annotation-derived positions served as stand-ins for the unknown true positions, we call them **proxy labels**.

Coding-sequence (CDS) annotations were obtained from the Ensembl GTF files described above. A transcript is an RNA sequence produced from a gene, and its CDS is the part annotated as being translated into protein. When a CDS spanned multiple genomic segments, we ordered those segments in the transcript's 5′-to-3′ direction. This allowed every possible P-site position to be located within the CDS and classified as falling on the first, second, or third nucleotide of an annotated codon.

A possible position was used as the expected P site when it matched either the beginning of the annotated CDS or the first nucleotide of its final annotated codon and also fell on the first nucleotide of a codon according to the annotation. If exactly one of the 13 possible positions met these conditions, its distance from the read's 5′ end became the proxy label for that read. The other 12 positions were treated as incorrect alternatives during model training. Reads with more than one matching position were considered ambiguous, and reads with no matching position had no proxy label. Only reads with exactly one proxy-labelled position were used to train predictors that learn from labelled examples or to calculate accuracy against the proxy labels. We inferred the expected P-site positions by combining read alignments with annotated coding-

sequence boundaries and reading frames. Because the experiment did not directly measure the P-site nucleotide for each read, these expected positions are proxy labels rather than confirmed positions.

How the framework made predictions comparable

The framework performed five steps (Figure 1). First, it prepared the sample information, reference files, genome alignments, and information for individual reads in consistent formats. Second, because the predictors required different inputs and produced different kinds of results, it used a conversion procedure specific to each predictor. Third, it represented every result in the same form: a unique read identifier, either a selected P-site offset or a score for each of the 13 possible positions, and an indication of whether the predictor returned a result. Fourth, it matched every predictor's results to the same reads selected in advance, ensuring that reads without predictions remained in the evaluation. Finally, it saved the input files, settings, code version, command, output files, and random seed used for each evaluation. A random seed is the fixed starting value for randomized steps. These records allow the analysis to be checked and repeated.

For each evaluation, we separately recorded whether the predictor produced any P-site predictions and how accurate those predictions were. Evaluations with no predictions were retained, and the missing predictions were counted as incorrect. The benchmark contained 176 planned predictor evaluations: each of the 11 predictors was applied under 16 evaluation conditions. One condition withheld the proxy labels for selected reads while other labelled reads from the same samples remained available for training. Eight conditions each withheld all proxy labels from one dataset, five each withheld all proxy labels from one protocol category, and two each withheld all proxy labels from one organism. Of the 176 evaluations, 169 returned predictions. In the other seven, Asite-IP returned no prediction for any of the reads selected for evaluation, and those missing predictions were counted as incorrect in the accuracy calculations. Table 2 reports how many evaluations produced predictions for each predictor; it describes result availability, not prediction quality.

Predictors included in the comparison

The 11 predictors were divided into three classes (Table 2). Here, a candidate is one of the 13 possible P-site positions for a read. **Learned candidate predictors** used training reads with proxy labels to learn how to compare the 13 possible P-site positions for each read. They assigned a numerical prediction score to each position and used those scores to select one final position. This class included TabRiboICL, XGBoost, and LightGBM. The second class consisted of simple comparison predictors, called **fitted offset baselines** because they estimated P-site offsets from the training reads without fitting a more complex prediction model. Fixed offset estimated one offset and applied it to every read, whereas Read-length offset estimated a separate offset for each read length. **Previously described predictors** included PSite [8], riboWaltz [3], Ribodeblur [4,5], RiboBA [6], Scikit-ribo [7], and Asite-IP [9].

The previously described predictors did not all return one P-site position for every read. Some returned a position for each read, some returned offsets for groups of reads, and others returned a position-by-position profile that combined signals from many reads. The conversion procedure for each predictor translated its result into the common definition used in the comparison. For RiboBA, the comparison used the P-site profile function available in its software; the full RiboBA method is a broader framework for identifying open reading frames (ORFs). Outputs reported as A-site positions were shifted by one codon to obtain the corresponding P-site positions.

Table 2. Predictor classes, outputs used in the comparison, and evaluations producing predictions.

Predictor	Predictor class	Output used in the comparison	Evaluations producing predictions
TabRiboICL	Learned candidate predictor	Scores for 13 possible P-site positions; one selected position per read	16/16
XGBoost	Learned candidate predictor	Scores for 13 possible P-site positions; one selected position per read	16/16
LightGBM	Learned candidate predictor	Scores for 13 possible P-site positions; one selected position per read	16/16
Fixed offset	Fitted offset baseline	One P-site offset learned from training reads and applied to every read	16/16
Read-length offset	Fitted offset baseline	One P-site offset learned from training reads for each read	16/16

Predictor	Predictor class	Output used in the comparison	Evaluations producing predictions
		length	
PSite	Previously described predictor	One P-site position for each read	16/16
riboWaltz	Previously described predictor	P-site offsets estimated for each sample and read length, then applied to individual reads	16/16
Ribodeblur	Previously described predictor	Values from a P-site profile used to score each read's 13 possible positions	16/16
RiboBA	Previously described predictor	Values from a P-site profile used to score each read's 13 possible positions	16/16
Scikit-ribo	Previously described predictor	An A-site profile shifted to P-site positions and used to score each read's 13 possible positions	16/16
Asite-IP	Previously described predictor	A-site offsets shifted by one codon to obtain P-site offsets	9/16

How TabRiboICL predicts a P-site

TabRiboICL predicts one P-site position for each mapped ribosome-protected fragment by considering 13 possible positions and choosing the one with the strongest combined support (Figure 3). For each read, the possible positions lie 8–20 nucleotides from its 5′ end. Here, an offset is the number of nucleotides from the read's 5′ end to a proposed P site, so each read has 13 candidate offsets.

Each candidate is described using information supplied to the prediction model, referred to as features. Read features include the proposed offset and read length. They also include strand, which records the read's orientation, and mapping quality, which indicates confidence in the genomic alignment. Sequence features describe the seven nucleotides at each end of the read, oriented in the transcript's 5′-to-3′ direction. Features describing nearby alignments summarize the abundance of ribosome footprints within 15 nucleotides on either side of the proposed P site and the abundance of nearby 5′ and 3′ read ends. Sample information includes the organism, available experimental-protocol metadata, library-preparation method, and dataset. Some datasets did not clearly report the digestion enzyme. TabRiboICL used the available protocol category as a sample feature, but missing or ambiguous nuclease information was left unknown rather than inferred.

The expected P-site positions used for evaluation were derived from gene annotations. To avoid revealing information that directly contributed to these expected positions, we excluded annotation-based features that could disclose the answer. Predictors could still use unlabeled read alignments and general sample information when these were part of their usual inputs.

TabICL [10] produces four internal predictions for each candidate and combines them into one model score. A higher score indicates stronger support for that position. TabICL is one part of TabRiboICL, not an alternative name for the complete predictor. TabRiboICL also constructs the 13 possible positions and their features, calculates the training-based offset probabilities described below, combines the two scores, and selects the final P-site position.

TabRiboICL also estimates how often each offset occurs among labelled training reads with similar characteristics. These frequencies provide a second score: an offset that was common among similar training reads receives more support. To prevent an offset from receiving a probability of zero merely because it was absent from a small training group, every offset was given an initial count of one before the labelled training examples were counted. For each test read, the predictor used the most specific training group available and then moved to broader groups if necessary. In order, these groups were: sample and read length; dataset and read length; organism, protocol-related metadata, library-preparation method, and read length; organism, protocol-related metadata, and read length; organism and read length; read length alone; and all training reads. If none supplied an estimate, all 13 offsets received equal weight. Only training labels were used to estimate these probabilities.

Before combining the two values, TabRiboICL limited both to the range from 10^{-12} to 1 to prevent numerical problems caused by zero values. It then multiplied the training-based offset probability by the TabICL score raised to the power 1.5:

`combined score = TabICL score^1.5 × training-based offset probability`.

The exponent of 1.5 increased the influence of differences among the TabICL scores; the offset probability was used without an additional exponent. The 13 combined scores for a read were divided by their sum, and the candidate with the highest resulting score was selected as the P site. Randomized steps used a fixed pseudorandom seed to support repeatability.

Earlier development results informed the TabRiboICL feature set and score-combination settings, which were then fixed for all reported evaluations. XGBoost [11] and LightGBM [12] received the same candidate descriptions, and the benchmark compared all three as fixed, complete prediction methods.

Evaluation designs and information available for test data

We used four evaluation designs, each withholding a defined group of proxy labels from model training. In the held-out-read evaluation, a fixed, reproducible rule assigned read identifiers within each sample to training and test sets, and no identifier appeared in both. Because every test sample was also represented during training, this analysis measures performance on additional reads from samples already seen by the predictor. We did not separately check whether identical or nearly identical read sequences occurred under different identifiers in the two sets; we therefore call these held-out read identifiers rather than previously unseen sequences.

In the held-out-dataset evaluation, all proxy-labelled reads from one dataset were excluded from training, and that dataset was used for testing. The procedure was repeated for all eight datasets. Some predictors use summaries calculated from many reads in a sample without requiring its proxy labels. For those predictors, unlabeled alignments from the test sample remained available. This evaluation therefore measures performance on a dataset whose labels were withheld. It does not represent prediction with no information from the new dataset or validation on a completely independent collection.

The protocol-category evaluation withheld labelled reads from one of five metadata categories at a time: RNase I, MNase, P1, OTTR library preparation, or Unreported. Reads in the other categories remained available for training, including reads from the same dataset when a dataset contained more than one category. The organism evaluation similarly withheld all labelled reads from either *D. melanogaster* or *S. cerevisiae*. Because only two organisms were included, these tests show how the predictors performed when trained on one represented organism and evaluated on the other. They do not show average performance across other species.

Accuracy measures, weighting, and confidence intervals

The main accuracy measure was exact accuracy: a prediction was correct only when the predicted offset exactly matched the proxy-labelled offset. A secondary measure counted a prediction as correct when it was within one nucleotide of the proxy offset. All predictors were evaluated on the same preselected set of reads. A missing or invalid P-site prediction was counted as incorrect, preventing a predictor from improving its reported accuracy by returning results only for easier reads.

Accuracy was first calculated separately for each sample. Sample accuracies were then averaged with equal weight within each dataset, and dataset accuracies were averaged with equal weight across datasets. Consequently, a dataset with more reads did not have greater influence on the overall mean.

We estimated uncertainty with bootstrapping, a procedure that repeatedly redraws observations from the available data and recalculates the result. Each redraw permits the same observation to be selected more than once and preserves the organization of reads within samples and samples within datasets. The held-out-read confidence intervals used 10,000 redraws of samples and of each read's correct-or-incorrect result within each dataset. The overall held-out-dataset intervals also redrew datasets. The protocol-category and organism analyses used the same levels that were available in each condition. When a dataset contained only one sample, its interval reflects variation among the observed reads but cannot describe variation among samples.

Figure 1 also shows two optional measures supported by the software but not used as main outcomes here: frame accuracy, which asks whether a predicted P site falls in the annotated reading frame, and read-level negative log likelihood, which penalizes a predictor for assigning low probability to the proxy-labelled offset. Negative log likelihood can be calculated only for predictors that return a probability for every candidate offset.

Combining within-one-nucleotide results across evaluation designs

The single overall protocol-category result in Figure 8 was not obtained by averaging the five category percentages shown in Figure 7. Instead, we combined the read-level predictions from the five category-specific evaluations, confirmed that each read was counted once, grouped each sample with its original

dataset, averaged samples equally within each dataset, and then averaged the eight datasets equally. The held-out-dataset, protocol-category, and organism analyses include the same eight datasets and 31 samples but different sets of reads: 40,000, 23,044, and 10,000 reads, respectively. Differences among these three summaries therefore should not be interpreted as changes in performance on the same test reads.

Biological consistency check across the broader read set

To complement the exact accuracy benchmark, we used riboWaltz to examine broad biological patterns among reads that were not used to construct the proxy labels (Figure 9). We excluded all 496,515 reads used to construct proxy labels from coding-sequence starts and stops and retained the other 16,360,911 reads from the 31 samples; riboWaltz assigned a P site to 16,360,482 of them. For this check, riboWaltz estimated a separate offset for each read length in each sample, allowing assignments across the broader read set. This differs from the version used in the proxy-label accuracy evaluations, for which prediction was repeated within each training-and-test division of the data.

We constructed metagene profiles, which summarize P-site abundance at corresponding positions around start or stop codons across many transcripts. For each gene, we selected the transcript with the longest coding sequence; transcript identifier was used to break ties. Each profile covered 45 nucleotides on either side of the start or stop codon. A transcript was included only when at least five assigned reads fell within the 91-nucleotide window. Within each included transcript, the counts at each position were divided by the total count in that window and multiplied by 91, so a uniform profile had a value of 1. The transcript profiles were then averaged by giving samples equal weight within each dataset and datasets equal weight overall. The 95% bootstrap confidence intervals used 1,000 repeated redraws of datasets and samples, with random seed 6062026.

We also examined, for each read length, the fraction of assigned P sites that fell on the first nucleotide of an annotated codon. This is the expected position for a correctly framed P site. A sample contributed a value for a read length only when it contained at least 200 assigned reads within coding sequences of that length. Values were averaged equally among eligible samples within each dataset and then equally among datasets. An average was displayed for a read length only when at least two datasets contributed. Finally, assigned P

sites were classified as lying within the selected coding sequence, immediately upstream of its start, immediately downstream of its stop, or elsewhere in the transcriptome; these fractions were again averaged equally among samples and then datasets.

These checks ask whether the assignments show familiar overall properties of ribosome-profiling data, such as concentration in coding sequences, three-nucleotide reading-frame structure, and peaks near start and stop codons. Figure 9 presents riboWaltz as a case study of this framework capability; it is not a comparison of predictors.

Records kept for reproducibility

For each analysis, we retained the identifiers of the evaluation reads, their assignments to training and test groups, predictor settings, random seeds, predictions, result tables, code-version information, and digital fingerprints (checksums) used to detect changes in source files. Operational status was recorded directly for every planned analysis, including whether it produced an output file.

Results

A common workflow standardized predictor outputs and recorded missing results

The framework converted the outputs of all 11 predictors to one definition of P-site position: the number of nucleotides from the read's 5′ end, interpreted consistently on both strands (Figure 1). Starting from sample information, reference sequences, genome annotations, and Ribo-seq reads, the workflow generated the possible P-site positions for each read, ran each predictor, converted all predictions to the same table format, and evaluated every predictor on the same preselected reads. Predictions reported for the A site were shifted by one codon before comparison with the P-site labels. We report exact accuracy and accuracy allowing a one-nucleotide difference. The framework can also calculate reading-frame accuracy and a probability-based error measure, but these were not primary outcomes in this analysis.

Across 176 planned predictor runs, each representing one predictor applied to one evaluation condition, 169 produced complete, nonempty output files and seven produced empty files (Table 2). Every expected read

remained in the accuracy calculation, so an empty file counted as missing predictions rather than disappearing from the analysis.

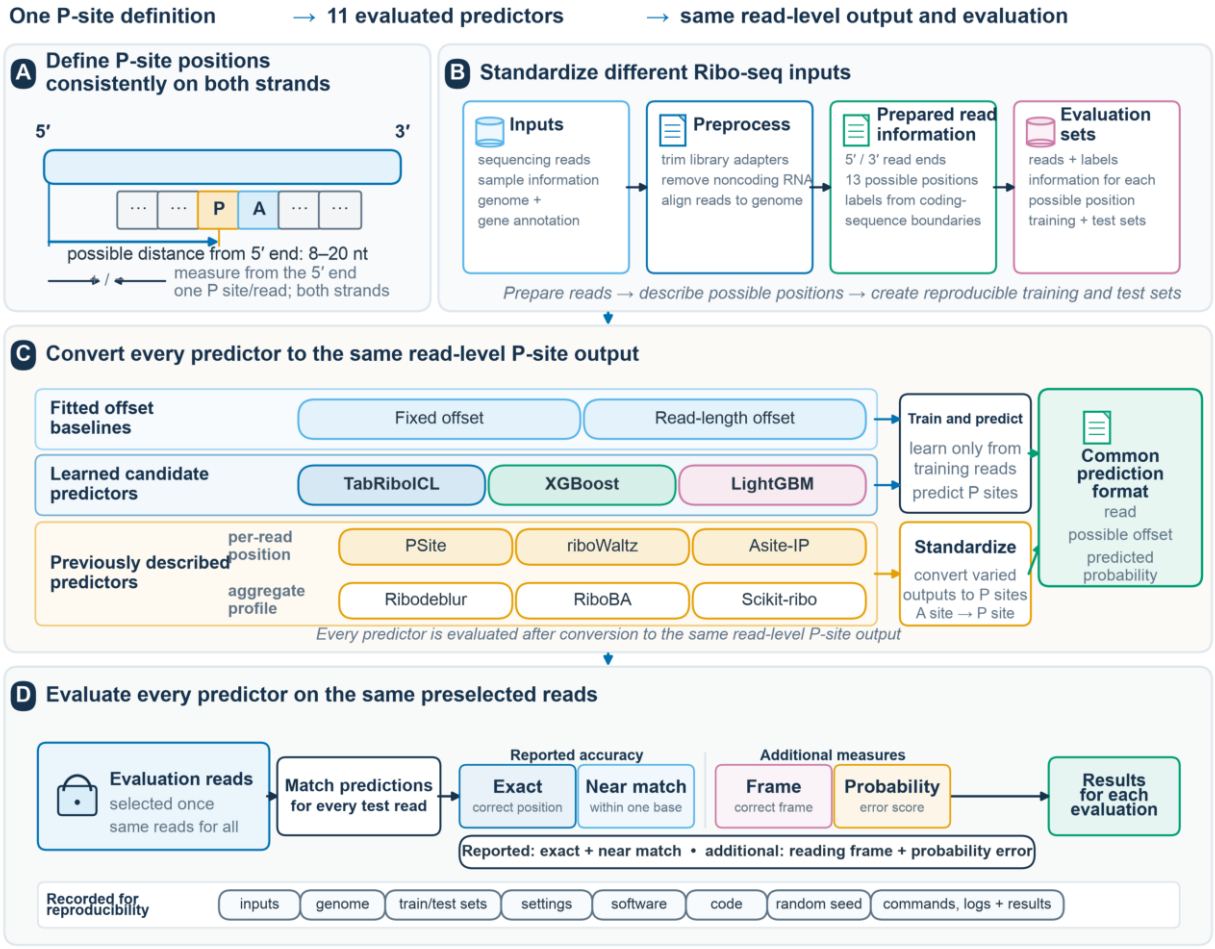


Figure 1. Common workflow for P-site prediction and benchmarking. (A) Every P-site position is represented as an offset of 8–20 nucleotides from the read's 5′ end, interpreted consistently on both strands, and each read receives at most one final prediction. (B) Ribo-seq inputs are converted into standard read tables, tables containing the possible P-site positions, and training-and-test groups. (C) Method-specific conversion steps place all outputs in the same format and shift A-site positions by one codon when needed. (D) Every predictor is evaluated on the same preselected reads; missing predictions and cases in which a method returns no prediction remain in the accuracy calculation, and each result is linked to records of its inputs, settings, code, and outputs. Exact and within-one-nucleotide accuracy are the reported outcomes; reading-frame accuracy and the probability-based error measure are optional.

The benchmark spanned eight datasets, 31 samples, two organisms, and five recorded protocol categories

The benchmark contained 31 Ribo-seq samples from eight datasets after exclusion of two RNA-seq companions (Figure 2). The datasets comprised 17 *D. melanogaster* samples in three datasets and 14 *S.*

cerevisiae samples in five datasets. The included Sequence Read Archive files totaled 46.30 GB. Dataset sizes ranged from one to eight samples.

Protocol representation was highly uneven: RNase I accounted for 9 samples from 2 datasets, MNase for 8 samples from 1 dataset, P1 for 2 samples from 1 dataset, OTTR library preparation for 1 sample from 1 dataset, and Unreported metadata for 11 samples from 5 datasets. All MNase samples belonged to one *D. melanogaster* dataset, while P1 and OTTR occurred only in SRP419250.

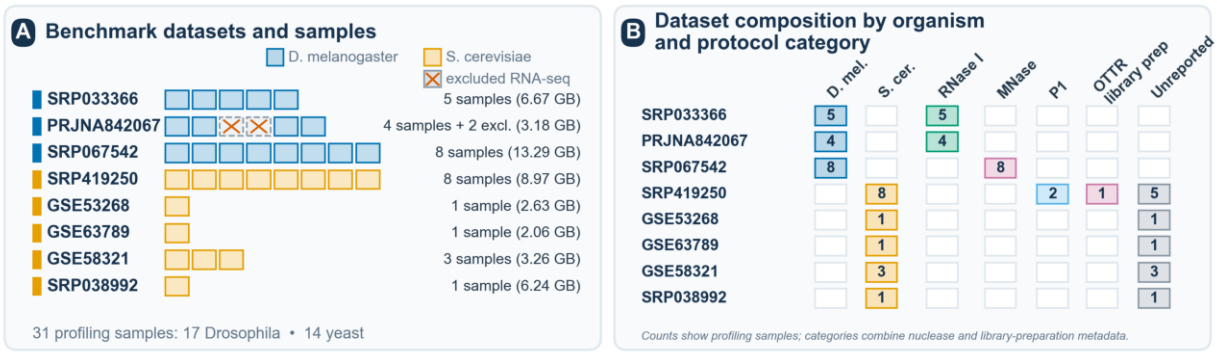


Figure 2. Benchmark composition. (A) Eight datasets contribute 31 Ribo-seq samples; two PRJNA842067 RNA-seq companion records are excluded. (B) Composition by organism and protocol category. RNase I, MNase, and P1 denote nucleases; OTTR denotes library preparation; Unreported denotes missing protocol metadata. Together, these counts show the distribution of two organisms and five recorded protocol categories across the benchmark's eight datasets.

TabRiboICL scored 13 possible positions and selected one P-site position per read

For each read, TabRiboICL evaluated 13 possible P-site offsets, ranging from 8 to 20 nucleotides from the 5′ end (Figure 3). Each possible position was described using sequence, alignment, nearby Ribo-seq signal, and sample information. Features based on annotated coding-sequence boundaries or reading frame were excluded because that information was also used to construct the proxy labels. TabICL assigned a score to each possible position. TabRiboICL then adjusted these scores using offset frequencies observed in the training data, using patterns specific to the sample and read length when matching training examples were available. The position with the highest adjusted score was selected as the predicted P site.

XGBoost and LightGBM evaluated the same 13 possible positions, providing comparisons among three complete learned predictors.

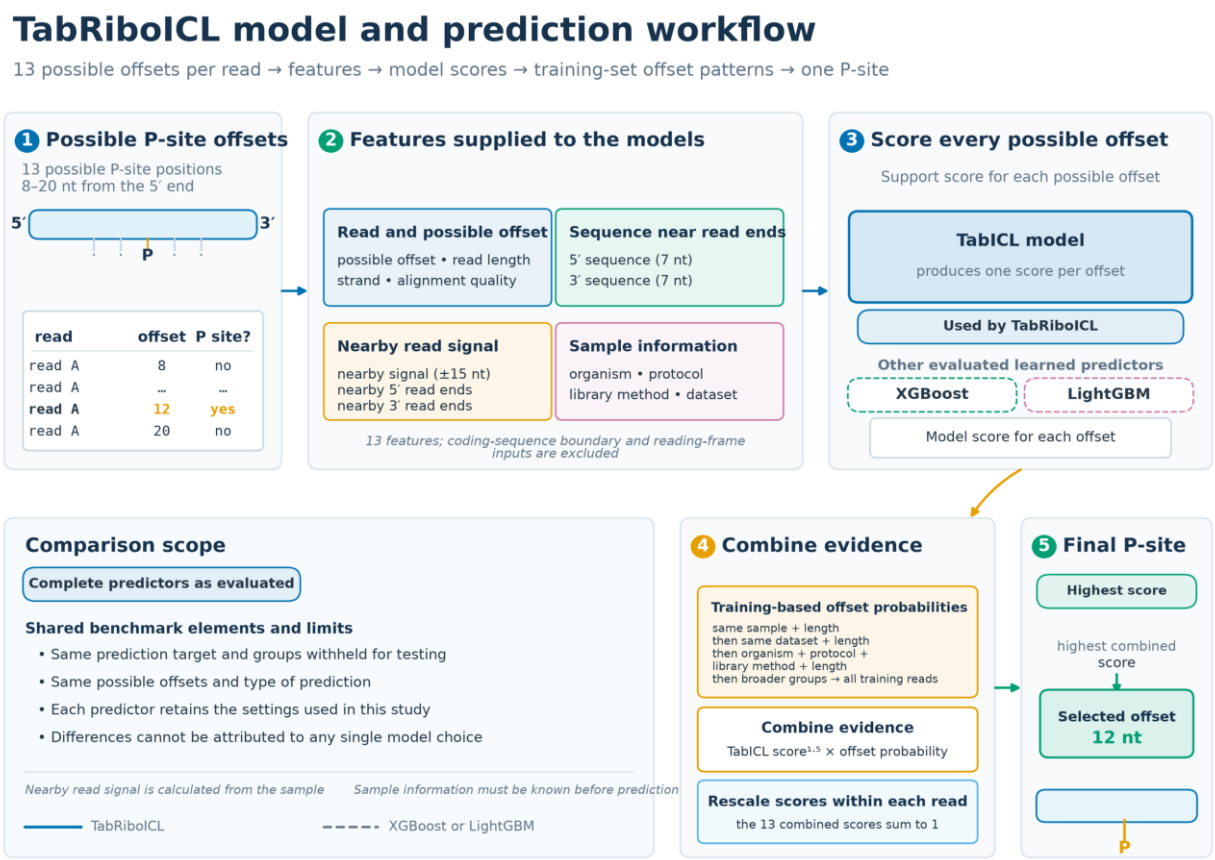


Figure 3. How TabRiboICL predicts a P-site position. For each read, 13 possible offsets are described using read, sequence, nearby-alignment, and sample information. Features based on annotated coding-sequence boundaries and reading frames are excluded. TabICL assigns a score to every possible position, and TabRiboICL combines that score with offset frequencies learned from the training labels. The 13 combined scores are converted to relative values whose total is one, and the highest-scoring position is selected. XGBoost and LightGBM use the same possible positions but are evaluated as separate complete methods.

TabRiboICL led exact accuracy on held-out reads in seven of eight datasets

The held-out-read analysis used one preselected set of 5,000 reads distributed unevenly across the eight datasets: 137 in SRP033366, 1,683 in PRJNA842067, 1,070 in SRP067542, 677 in SRP419250, 391 in GSE53268, 275 in GSE63789, 548 in GSE58321, and 219 in SRP038992. Accuracy was first calculated for each sample, and samples were then given equal weight within each dataset.

For exact accuracy, TabRiboICL had the highest estimated accuracy in seven of eight datasets (Figure 4). Its estimates ranged from 25.1% in SRP067542 to 98.9% in GSE63789. RiboBA led SRP067542 at 37.6%, compared with 25.1% for TabRiboICL. TabRiboICL's exact estimates were 45.8% in SRP033366, 59.8% in

PRJNA842067, 33.1% in SRP419250, 96.9% in GSE53268, 98.9% in GSE63789, 52.4% in GSE58321, and 96.3% in SRP038992. Accuracy varied substantially among datasets.

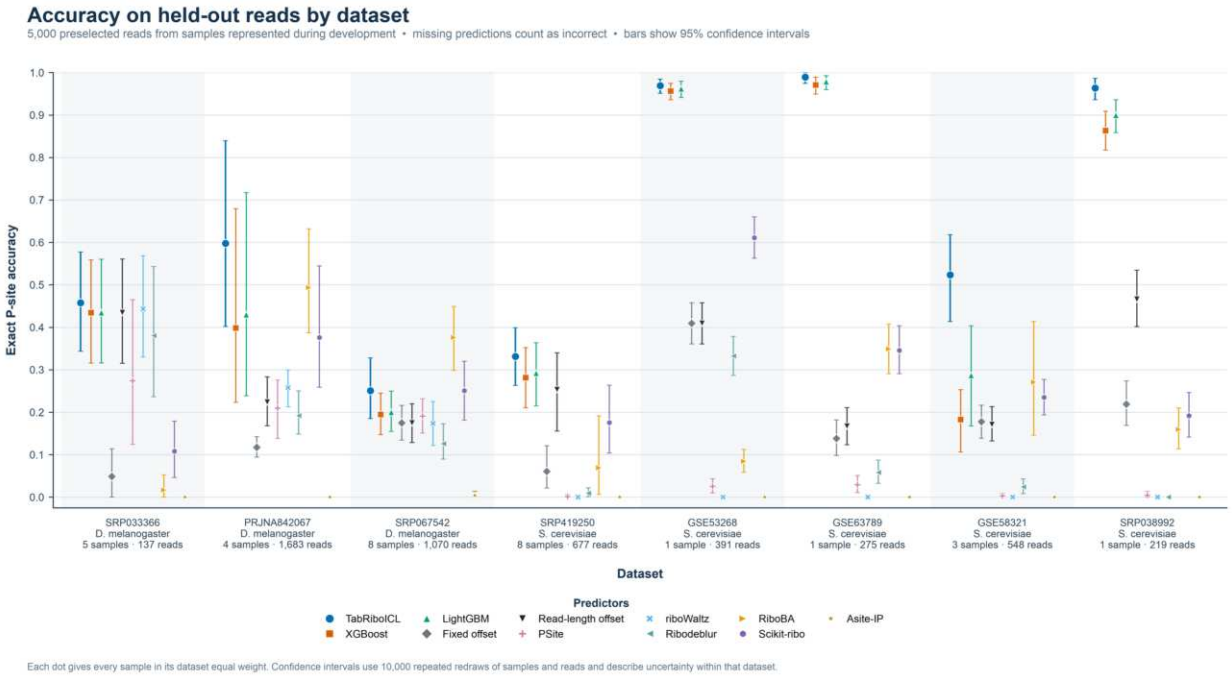


Figure 4. Exact accuracy on held-out reads by dataset. The analysis contains 5,000 preselected reads from samples represented during both development and evaluation; the number contributed by each dataset is shown. Each point is an accuracy estimate calculated after giving every sample in that dataset equal weight. Bars are 95% bootstrap confidence intervals from 10,000 repeated redraws. For datasets containing one sample, the intervals reflect variation among reads only. Missing predictions and cases in which a method returned no prediction are counted as incorrect. This analysis evaluates reads that were not used during model development, although other reads from the same samples were used.

Accuracy within one nucleotide was higher than exact accuracy for most predictor–dataset combinations (Figure 5). TabRiboICL had the highest or tied-highest estimated accuracy in all eight datasets, tying riboWaltz at 67.5% in SRP033366. TabRiboICL estimates ranged from 50.6% in SRP419250 to 99.0% in GSE53268; the corresponding SRP067542 estimate was 51.7%, despite the lower exact result. The difference between exact accuracy and accuracy within one nucleotide shows that many errors were close to the proxy offset, but the biological consequences of a one-nucleotide displacement depend on the downstream analysis.

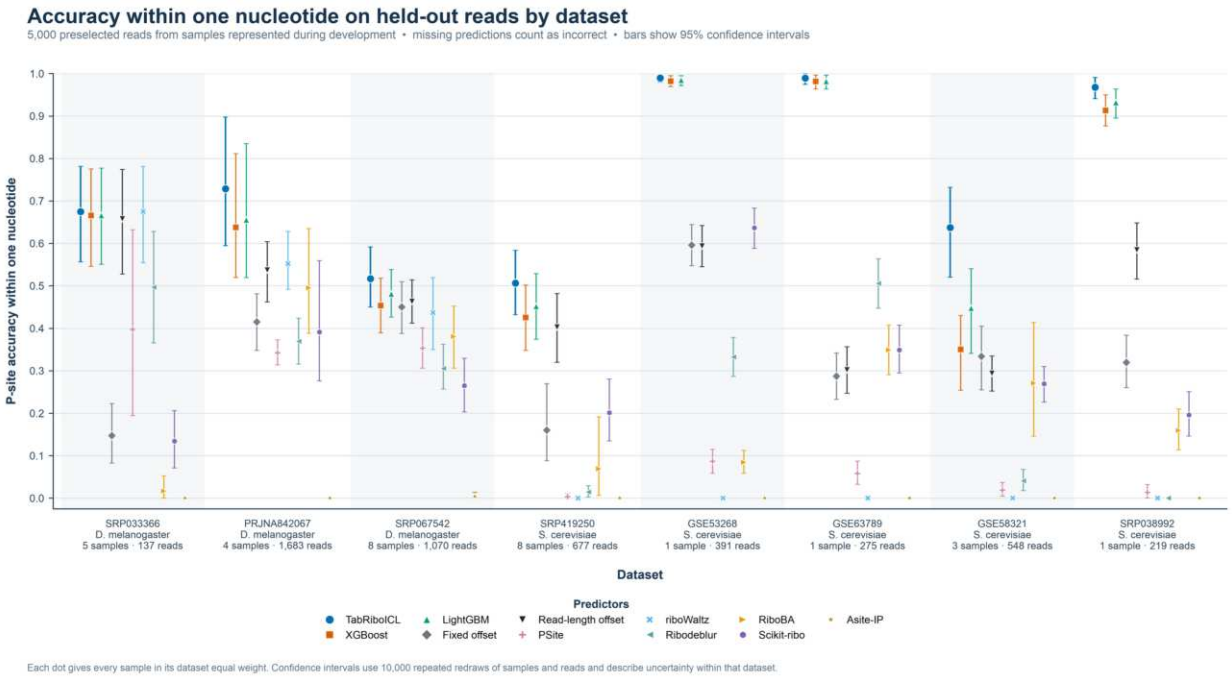


Figure 5. Accuracy within one nucleotide on held-out reads by dataset. The evaluated reads, weighting, and confidence-interval calculation match Figure 4. A prediction is counted as correct when it differs from the proxy offset by no more than one nucleotide. TabRiboICL has the highest or tied-highest estimated accuracy in all eight datasets. Each confidence interval describes one predictor and does not test the difference between two predictors.

TabRiboICL had the highest observed mean exact accuracy across held-out datasets

Each dataset was held out in turn. For every held-out dataset, all proxy-labelled reads from that dataset were excluded during model training, and each predictor was evaluated on the same preselected 5,000 reads. This produced 40,000 evaluated reads across the eight datasets. A missing prediction was counted as incorrect. Each sample contributed equally to its dataset result, and each dataset contributed equally to the overall average (Figure 6).

TabRiboICL had a mean exact accuracy of 37.6% after giving each dataset equal weight (95% bootstrap confidence interval, 32.5%–43.8%), the highest observed mean among the evaluated predictors (Figure 6B; Table 3). It led five datasets (Figure 6A): SRP033366 (52.3%), PRJNA842067 (37.8%), SRP067542 (31.8%), SRP419250 (34.3%), and GSE58321 (29.3%). Scikit-ribo led GSE53268 (59.1%) and GSE63789 (34.3%), and LightGBM led SRP038992 (57.3%). The best-performing predictor and the accuracy of each predictor varied among datasets.

Table 3. Mean exact accuracy when every held-out dataset was given equal weight.

Predictor	Exact accuracy (%)	95% bootstrap confidence interval (%)
TabRiboICL	37.6	32.5–43.8
LightGBM	28.5	16.1–42.7
Scikit-ribo	26.4	17.0–37.9
riboWaltz	24.8	17.9–32.9
PSite	20.9	12.8–30.2
Ribodeblur	17.0	7.9–27.1
XGBoost	13.6	7.7–20.6
Read-length offset	13.4	9.7–17.0
Fixed offset	9.7	7.0–12.2
RiboBA	8.8	4.1–14.5
Asite-IP	1.1	0.0–3.0

The analysis reports predictor-specific accuracy and 95% bootstrap confidence intervals, with comparisons stated as observed differences. Under the inputs permitted for each method, TabRiboICL achieved the highest observed equal-dataset mean when each dataset's proxy labels were withheld.



Figure 6. Exact accuracy when proxy labels from each dataset were held out. (A) Accuracy for each of eight datasets, calculated from 5,000 preselected reads after giving every sample equal weight. (B) Gray dots show individual-dataset accuracies; colored markers and error bars show overall means, with every dataset given equal weight, and their 95% bootstrap confidence intervals. Missing predictions and cases in which a method returns no prediction are counted as incorrect. Bars are 95% bootstrap confidence intervals from 10,000 repeated redraws; the overall intervals redraw datasets, samples, and read-level correct-or-incorrect results. All labelled reads from the displayed dataset were excluded from model development, although some predictors could still use unlabeled alignments from that dataset.

The framework compared predictors after withholding entire organisms and protocol categories

The predictor with the highest exact accuracy differed across the seven organism and protocol-category conditions (Figure 7). TabRiboICL led held-out RNase I at 38.1% (30.5%–44.2%) and held-out *D. melanogaster* at 34.2% (27.5%–38.7%). Scikit-ribo led Unreported at 31.6% (18.6%–49.2%) and held-out *S. cerevisiae* at 31.3% (17.9%–48.8%). RiboBA led MNase at 23.0% (17.2%–29.3%) and OTTR library preparation at 46.1% (44.3%–47.9%), while Ribodeblur led P1 at 25.0% (23.3%–26.6%).

The amount of available data differed sharply. MNase contained 1 dataset, 8 samples, and 5,000 evaluated reads; OTTR library preparation contained 1 dataset, 1 sample, and 3,044 evaluated reads; P1 contained 1 dataset, 2 samples, and 5,000 evaluated reads; RNase I contained 2 datasets, 9 samples, and 5,000 evaluated

reads; and Unreported contained 5 datasets, 11 samples, and 5,000 evaluated reads. The organism evaluations used 3 datasets and 17 samples for *D. melanogaster* and 5 datasets and 14 samples for *S. cerevisiae*, each with 5,000 evaluated reads.

The bootstrap intervals summarize the uncertainty supported by the datasets, samples, and reads available for each condition. Together, these seven evaluations demonstrate that the framework can compare predictor performance when labels from an entire organism or recorded protocol category are withheld, including conditions with substantially different amounts of available data.

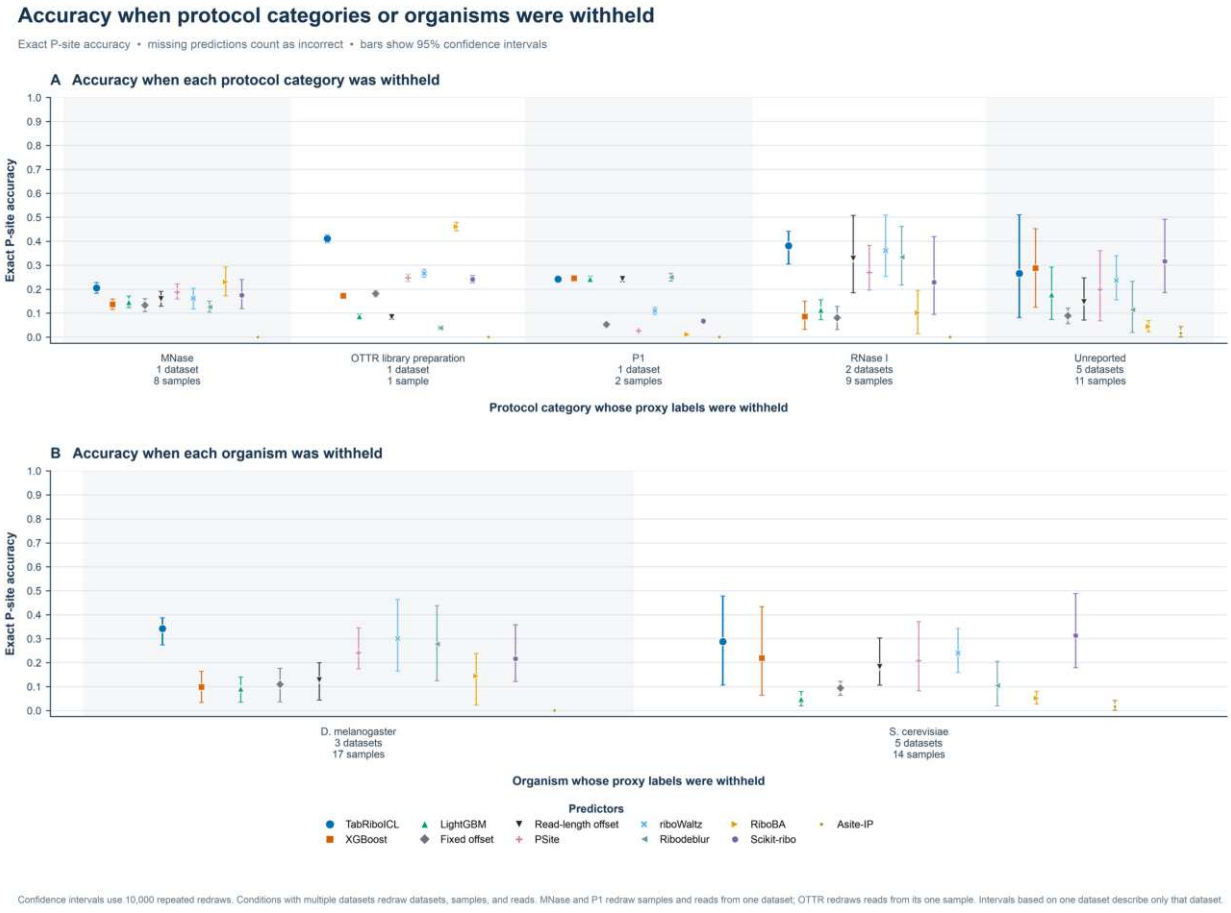


Figure 7. Exact accuracy when an organism or protocol category was held out. (A) Results after withholding proxy labels for MNase, OTTR library preparation, P1, RNase I, or samples with unreported protocol metadata. (B) Results after withholding proxy labels for either represented organism. Missing predictions and cases in which a method returned no prediction are counted as incorrect. Bars show 95% bootstrap confidence intervals obtained by repeatedly redrawing the available datasets, samples, and reads. OTTR is a library-preparation protocol, and Unreported denotes missing metadata.

To obtain one overall protocol-category result for accuracy within one nucleotide, predictions from the five protocol-category evaluations were returned to their original datasets before accuracy was averaged across datasets (Figure 8). TabRiboICL had the highest observed mean for the held-out-dataset evaluation at 55.5%

(47.4%–64.6%) and the held-out-organism evaluation at 51.2% (38.0%–63.0%). XGBoost had the highest observed protocol-category mean at 48.9% (40.1%–59.9%), compared with 47.2% (37.3%–58.4%) for TabRiboICL.

Together, these three evaluations provide complementary views of performance when datasets, protocol categories, or organisms were withheld. TabRiboICL had the highest observed mean in two evaluations, and XGBoost had the highest observed mean in the protocol-category evaluation.

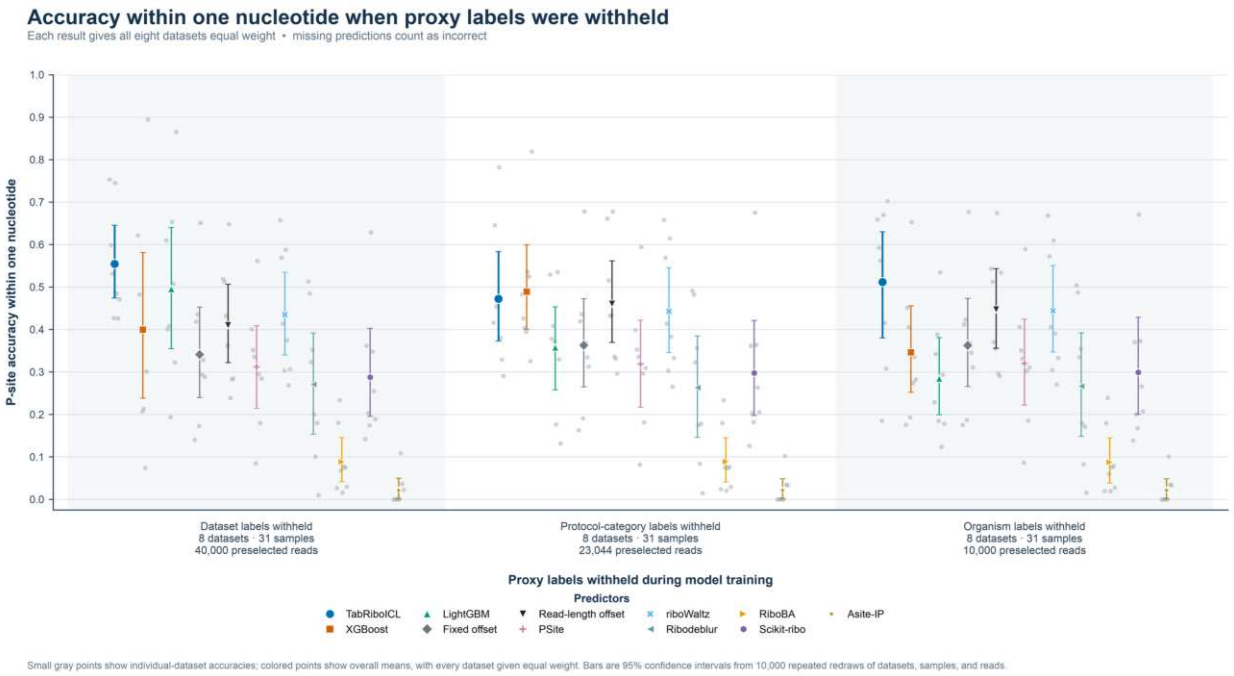


Figure 8. Accuracy within one nucleotide when datasets, protocol categories, or organisms were held out. Small gray points show accuracy for individual datasets; colored points show overall means calculated after giving every dataset equal weight, with 95% bootstrap confidence intervals. The three evaluations include 40,000, 23,044, and 10,000 reads, respectively. For the protocol-category result, reads from the five separate evaluations were returned to their original datasets before averaging; the overall value is therefore not the simple mean of the five percentages in Figure 7.

riboWaltz assignments showed characteristic Ribo-seq patterns across the broader read set

Using the broader set of 16,360,911 reads that were not used to create the proxy labels, we ran riboWaltz separately on each sample and allowed it to estimate a different offset for each read length (Figure 9). To examine broad biological patterns, predicted P sites from many transcripts were aligned relative to annotated coding-sequence starts and stops and then averaged; these averaged profiles are commonly called metagene profiles. The profiles showed peaks at the expected positions and reading frames near coding-sequence boundaries. Each profile was scaled so that a uniform distribution across the displayed window had a value

of 1. For P sites at the first base of an annotated codon, the highest value was 3.92 at 12 nucleotides downstream of the coding-sequence start (95% bootstrap confidence interval, 2.50–5.92) and 4.09 at 33 nucleotides upstream of the stop (3.06–5.25).

The proportion of P sites aligned with the annotated coding-sequence reading frame depended strongly on read length. Among the read lengths included in the summary, this proportion reached 68.5% at 28 nucleotides, with data from seven datasets; values for lengths from 27 to 32 nucleotides ranged from 41.8% to 68.5%. Most predicted P sites fell within the coding sequence of the representative transcript selected for each gene. This proportion ranged from 81.0% to 99.9% across datasets and averaged 94.5% when each dataset was given equal weight. The remaining positions were located upstream of the coding sequence (0.5%), downstream of it (0.2%), or elsewhere (4.8%).

Across more than 16 million reads that were not used to create the benchmark proxy labels, riboWaltz assignments showed characteristic Ribo-seq patterns, including three-nucleotide reading-frame structure, enrichment within coding sequences, and structured profiles around start and stop codons. This complementary analysis demonstrates how the framework can assess broad biological consistency when exact expected P-site positions are unavailable.

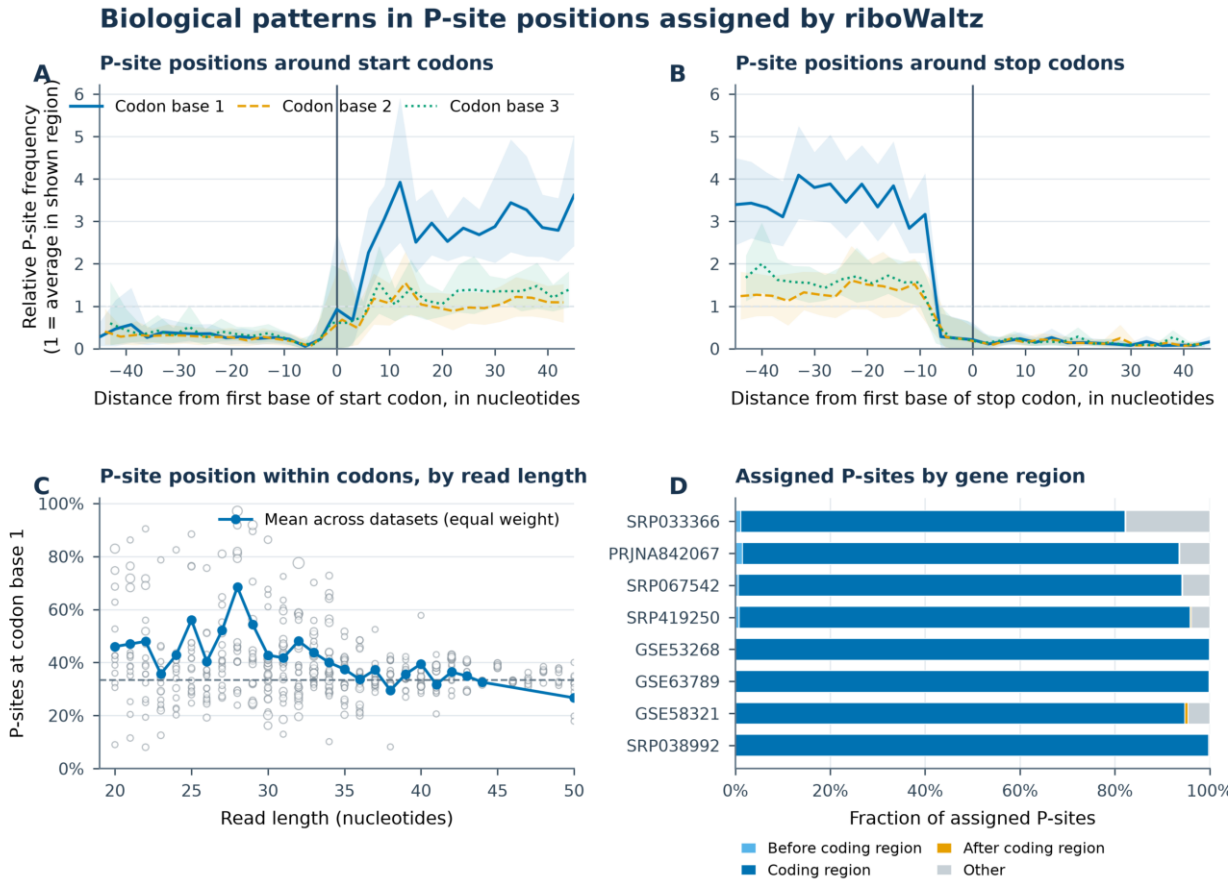


Figure 9. Expected biological patterns in riboWaltz P-site assignments. Reads used to create proxy labels from coding-sequence starts and stops were excluded before analysis. (A,B) Average P-site profiles around coding-sequence start and stop codons. Each transcript profile was scaled so that a uniform distribution across the window had a value of 1; samples and then datasets were given equal weight. Shaded bands are 95% bootstrap confidence intervals from 1,000 repeated redraws of datasets and samples. (C) Percentage of P sites at the first nucleotide of an annotated codon, shown by read length. A sample contributes only when it has at least 200 coding-sequence reads of that length, and the mean is shown only when at least two datasets contribute. (D) Percentage of predicted P sites within the selected coding sequence, immediately upstream, immediately downstream, or elsewhere.

Discussion

This study contributes both a read-specific P-site predictor and a framework for comparing P-site prediction methods. In the primary held-out-dataset analysis, TabRiboICL had the highest observed mean exact accuracy when every dataset was given equal weight, and it had the highest accuracy in five of eight datasets. Scikit-ribo and LightGBM had the highest accuracy in the remaining datasets, and the leading predictor also changed across organism and protocol-category evaluations. Taken together, these results support TabRiboICL as a strong predictor in this benchmark and demonstrate the value of comparing methods across multiple datasets and evaluation settings.

The benchmark evaluates TabRiboICL, XGBoost, and LightGBM as complete prediction methods and therefore provides an end-to-end comparison of their performance. TabRiboICL defines 13 possible positions, describes them with features, scores them with TabICL, adjusts the scores using offset patterns from the training data, and selects the position with the highest score. XGBoost and LightGBM were evaluated in their own complete configurations. Future controlled component studies can build on this benchmark to determine how individual design choices contribute to performance.

In the held-out-read analysis, training and evaluation reads came from the same samples. Predictors could therefore use patterns calculated from those samples, including typical offsets for each read length and nearby Ribo-seq signals. This analysis estimates performance on additional reads from samples already represented during development. In the held-out-dataset analysis, proxy labels from the evaluated dataset were hidden, but some methods could still use its unlabeled read alignments. Both settings differ from applying a fixed predictor to an entirely new dataset without adjustment. Researchers should therefore select the result that best matches how they intend to use the predictor.

The framework defines one P-site position, gives each read a unique identifier that remains the same throughout the analysis, converts every predictor output to the same format, evaluates every predictor on the same reads, and averages results at the sample and dataset levels. Each output file is linked to records of the input data, settings, code, and resulting predictions. These steps ensure that missing predictions remain in the accuracy calculation instead of being silently omitted. These records document result availability for every planned analysis under the software configurations and computing environment evaluated here.

The study's defined scope also identifies priorities for broader validation. First, annotation-derived proxy labels enabled read-level evaluation across all predictors; independently measured P-site positions would provide complementary validation beyond annotation patterns. Second, the eight public datasets from two organisms establish a foundation that can be expanded to additional organisms and protocol categories represented across multiple datasets, allowing protocol and organism effects to be examined separately. Third, the benchmark preserved each predictor's usual access to unlabeled alignments and sample signals, allowing sample-adaptive methods to operate as intended; a complementary fixed-model evaluation could examine performance without sample-specific adaptation. Fourth, the reported TabRiboICL configuration

was selected after examining preliminary development results and then fixed for all reported evaluations. A future sequence-level audit could provide an even stricter held-out-read split by identifying identical or highly similar sequences across training and test identifiers.

Future benchmarks should use P-site positions established by independent experiments, synthetic datasets with known positions, or experimental measurements not used by the predictors. This would reduce the risk that a method appears accurate because it relies on the same annotation patterns used to create the proxy labels. Broader evaluations should include more organisms, protocols represented across several independent datasets, and test datasets selected before model development. Evaluations should separately examine methods allowed to use unlabeled data from a new sample and methods applied without any adjustment to that sample. Future studies could also test whether allowing a predictor to return no result for a low-confidence read improves practical usefulness, using evaluation criteria defined in advance, and could change one TabRiboICL part at a time to determine the contributions of its input information, TabICL model, training-based offset patterns, and final selection step. Finally, the standardized predictions produced by this framework also provide a basis for testing how P-site assignment errors influence downstream analyses of codon-specific ribosome pausing, translation initiation, translated open reading frames, and differences in translation between conditions.

Conclusion

TabRiboICL achieved the highest observed average exact accuracy when each held-out dataset was given equal weight. The unified framework enabled 11 P-site predictors to be compared using consistent P-site definitions, reads, and scoring rules across multiple evaluation settings. Together, these results support TabRiboICL as a strong read-specific P-site predictor and provide a practical framework for future P-site benchmarking.

Data and code availability

The code is available at https://github.com/ericzhuang2010/tabriboicl.

The study used previously published ribosome-profiling datasets available through the NCBI Sequence Read Archive (SRA) and Gene Expression Omnibus (GEO) under accessions SRP033366, PRJNA842067, SRP067542, SRP419250, GSE53268, GSE63789, GSE58321, and SRP038992. Raw sequencing data can be retrieved using the SRA Toolkit through scripts/download_sra_fastq.sh; reference files are downloaded using scripts/download_references.sh.

Author contributions

Eric Zhuang conceived the study, developed the methodology, implemented the software, conducted the computational analyses and evaluation, and wrote the original manuscript. Manoj Kandpal contributed to reviewing and revising the manuscript.

Funding

No funding was received for this study.

Competing interests

The authors declare no competing interests.

Ethics statement

The analysis used publicly available sequencing data and did not involve new human participants or animal experiments.

References

1. Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*. 2009;324(5924):218–223. doi:10.1126/science.1168978.

2. Ingolia NT, Brar GA, Rouskin S, McGeachy AM, Weissman JS. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nature Protocols*. 2012;7:1534–1550. doi:10.1038/nprot.2012.086.
3. Lauria F, Tebaldi T, Bernabò P, Groen EJN, Gillingwater TH, Viero G. riboWaltz: optimization of ribosome P-site positioning in ribosome profiling data. *PLoS Computational Biology*. 2018;14(8):e1006169. doi:10.1371/journal.pcbi.1006169.
4. Wang H, McManus J, Kingsford C. Accurate recovery of ribosome positions reveals slow translation of wobble-pairing codons in yeast. *Journal of Computational Biology*. 2017;24(6):486–500. doi:10.1089/cmb.2016.0147.
5. Wang H, Kingsford C, McManus CJ. Using the Ribodeblur pipeline to recover A-sites from yeast ribosome profiling data. *Methods*. 2018;137:67–70. doi:10.1016/j.ymeth.2018.01.002.
6. Bai J, Yang R. RiboBA: a bias-aware probabilistic framework for robust ORF identification across diverse ribosome profiling protocols. *bioRxiv* [Preprint]. 2026. doi:10.64898/2026.03.17.712439.
7. Fang H, Huang Y-F, Radhakrishnan A, Siepel A, Lyon GJ, Schatz MC. Scikit-ribo enables accurate estimation and robust modeling of translation dynamics at codon resolution. *Cell Systems*. 2018;6(2):180–191.e4. doi:10.1016/j.cels.2017.12.007.
8. Chang Y, Lei T, Zhang H. PSite: inference of read-specific P-site offsets for ribosomal footprints. *bioRxiv* [Preprint]. 2023:2023.06.27.546788. doi:10.1101/2023.06.27.546788.
9. Ahmed N, Sormanni P, Ciryam P, Vendruscolo M, Dobson CM, O'Brien EP. Identifying A- and P-site locations on ribosome-protected mRNA fragments using integer programming. *Scientific Reports*. 2019;9:6256. doi:10.1038/s41598-019-42348-x.
10. Qu J, Holzmüller D, Varoquaux G, Le Morvan M. TabICL: a tabular foundation model for in-context learning on large data. In: *Proceedings of the 42nd International Conference on Machine Learning*. Proceedings of Machine Learning Research. 2025;267:50817–50847.
11. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016:785–794. doi:10.1145/2939672.2939785.

12. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T-Y. LightGBM: a highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*. 2017;30.
13. Dunn JG, Weissman JS. Plastid: nucleotide-resolution analysis of next-generation sequencing and genomics data. *BMC Genomics*. 2016;17:958. doi:10.1186/s12864-016-3278-x.
14. Weber LM, Saelens W, Cannoodt R, Soneson C, Hapfelmeier A, Gardner PP, Boulesteix A-L, Saeys Y, Robinson MD. Essential guidelines for computational method benchmarking. *Genome Biology*. 2019;20:125. doi:10.1186/s13059-019-1738-8.
15. Popa A, Lebrigand K, Paquet A, Nottet N, Robbe-Sermesant K, Waldmann R, Barbry P. RiboProfiling: a Bioconductor package for standard Ribo-seq pipeline processing. *F1000Research*. 2016;5:1309. doi:10.12688/f1000research.8964.1.
16. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*. 2011;17(1):10–12. doi:10.14806/ej.17.1.200.
17. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nature Methods*. 2012;9(4):357–359. doi:10.1038/nmeth.1923.
18. Danecek P, Bonfield JK, Liddle J, et al. Twelve years of SAMtools and BCFtools. *GigaScience*. 2021;10(2):giab008. doi:10.1093/gigascience/giab008.
19. Cunningham F, Allen JE, Allen J, et al. Ensembl 2022. *Nucleic Acids Research*. 2022;50(D1):D988–D995. doi:10.1093/nar/gkab1049.