Briefings in Bioinformatics, 2026, 27, bbag241
https://doi.org/10.1093/bib/bbag241
Published: 19 May 2026
Problem Solving Protocol

OXFORD

DCVBin: a novel binning method for single-sample metagenomes based on DNA language model and variational autoencoder

Jingyuan Wang[1,2,‡], Yifan Liu[3,‡], Fu Liu[3], Tao Hou[3], Siming Chen[3], Shengxi Liu[3], Yun Liu[3,*]

[1]School of Artificial Intelligence, Jilin University, Qianjin Street No. 3003, 130000, Changchun, Jilin, China
[2]School of Electrical and Information Engineering, Jilin Engineering Normal University, Kaixuan Road No. 3050, 130000, Changchun, Jilin, China
[3]College of Communication Engineering, Jilin University, Remin Street No. 5988, 130000, Changchun, Jilin, China

*Corresponding author. College of Communication Engineering, Jilin University, Remin Street No. 5988, 130000, Changchun, Jilin, China. E-mail: liuyun313@jlu.edu.cn
‡Jingyuan Wang and Yifan Liu contributed equally to this work.

Abstract

DNA contigs binning is necessary to reconstruct metagenome-assembled genomes. Current metagenomic DNA contigs binning methods often leverage coverage profiles across multiple related metagenomes and have demonstrated strong performance on co-assembled contigs. However, in single-sample scenarios where coverage information is rare, their performance drops significantly, limiting the in-depth development of metagenomics at the individual sample level. To address this issue, we propose DCVBin, a novel single-sample metagenomic contigs binning method that incorporates semantic features extracted from a DNA language model. Specifically, our approach continues pretraining on a DNA language model to capture more domain-specific semantic representations, which are then integrated with 4-mer frequencies using a variational autoencoder. Clustering is subsequently performed using the k-means algorithm, in which the number of clusters is determined by single copy genes. Experimental results on six publicly available datasets demonstrate that DCVBin achieves high-accuracy single-sample metagenomic binning and outperforms other state-of-the-art methods. Furthermore, DCVBin is included into a disease diagnostic framework that is evaluated on a cohort of gut metagenomes from people with colorectal cancer and healthy people. The framework is shown to be accurate in predicting colorectal cancer using gut metagenomes and has identified a list of potential microbial biomarkers.

Keywords **metagenomic binning, single-sample binning, DNA language model, semantic features, variational autoencoder, clustering**

Introduction

Microorganisms are the most diverse, abundant, and widely distributed organisms on Earth, playing a crucial role in many ecosystems [1]. For example, human gut microbes have been discovered to be closely associated with several human diseases, such as colorectal cancer (CRC) [2], showing potential value for disease diagnosis and treatment [3]. Metagenomic techniques can acquire the genetic material of all microorganisms from environmental samples without laboratory cultivation, making them widely used in the study of microbial communities. However, due to factors preventing the assembly of complete genomes from metagenomic raw reads, binning contigs into draft genomes becomes a necessary step for downstream metagenomic analysis.

Current mainstream metagenomic binning methods mainly depend on two key features: the compositional characteristics of DNA contigs and coverage profiles across multiple related metagenomes [4]. While these methods have demonstrated effectiveness on both simulated and real datasets, their performance is heavily contingent upon the availability of multiple related samples—a requirement that presents significant practical challenges. The dependence on multi-sample coverage manifests three critical limitations: first, the dimensionality of coverage features increases with the number of metagenomes, creating computational burdens for large datasets [5]. Second, empirical studies indicate that at least five related metagenomes are required for meaningful binning results [6]. Most importantly, matched multi-sample datasets are frequently unavailable in key research scenarios, including clinical investigations, retrospective studies, and longitudinal monitoring. In these constrained contexts, single-sample analysis has emerged as both a necessary alternative and a valuable methodological approach. Its efficacy is particularly evident in CRC research, where single-sample strategies have enabled (i) identification of *Escherichia coli* as the dominant antibiotic resistance reservoir through individual patient profiling [7]; (ii) detection of tumor progression-associated microbiome shifts, including oral bacterial translocation [8]; and (iii) derivation of microbial signatures predictive of adenoma–carcinoma

Received: November 5, 2025. **Revised:** February 4, 2026. **Accepted:** April 21, 2026

transition from solitary biopsies [9]. These accomplishments collectively demonstrate that single-sample metagenomics represents not merely a fallback option when multi-sample analysis is impossible, but rather an essential paradigm for microbiome research when sample matching is clinically impractical or biologically unfeasible.

To overcome the limitations of multi-sample dependency, we propose DCVBin, a method that incorporates semantic features of contigs into the binning process. First, DCVBin extracts compositional and semantic features of DNA contigs. The semantic features are obtained by a DNA language model, DNABERT_S, which is fine-tuned on the DNA contigs to be binned in an unsupervised manner. Then, the two sets of features are pooled into a feature learning module based on a variational autoencoder (VAE). Finally, a k-means clustering algorithm is performed on the latent embeddings of the VAE to produce draft genomes. Experiments on several single-sample metagenomes demonstrate that DCVBin reconstructs more high-quality draft genomes than other tools. DCVBin is also integrated into a diagnostic framework for CRC and is shown to outperform existing prediction models in terms of accuracy.

The main contributions of this paper are summarized as follows. First, it leverages a DNA language model to extract semantic features from DNA contigs, overcoming the limitations of traditional methods that rely on coverage features. By further pretraining the language model, deeper and more specific semantic features are extracted, allowing for a more precise representation of DNA contigs in single-sample metagenomes. Secondly, a joint feature fusion strategy using VAE [10] and k-means clustering combines semantic features with traditional k-mer frequencies. By combining VAE with a target-adaptive continued pretraining (CPT) strategy, the model is able to capture domain-specific semantic representations, thereby effectively compensating for the lack of abundance information. Simultaneously, it eliminates the reliance on multi-sample coverage profiles, which grants it unique advantages in single-sample scenarios where such data are unavailable.

The remainder of this paper is organized as follows: Section 2 discusses related work, Section 3 presents the details of the proposed model, Section 4 covers the experimental results, and Section 5 concludes the paper.

Related work

Existing metagenomic binning methods can be divided into three categories based on the features used: composition feature-based methods, fusion of composition and coverage feature-based methods, and graph feature-based methods.

(i) **Composition feature-based methods**
The early binning methods primarily rely on composition features of DNA sequences. TETRA [11] calculates tetra-nucleotide frequency differences and uses Pearson correlation [12] for clustering. CompostBin [13] utilizes k-mer frequencies and phylogenetic information for clustering, while MetaCluster [14] applies unsupervised learning based on k-mer distributions. However, these methods struggle to distinguish sequences from closely related genomes.

(ii) **Fusion of composition and coverage feature-based methods**
These methods assume that sequences from the same genome have similar abundance distributions across multiple metagenomes. MetaBAT2 [15] combines composition and coverage features to compute distance matrices for clustering using graph algorithms. VAMB [16] employs a VAE to map concatenated inputs of TNF and multi-sample coverage profiles into low-dimensional embeddings for k-medoids clustering. MetaDecoder [17] uses a two-layer model: an improved Dirichlet Process Gaussian Mixture Model (DPGMM) [18] and a probabilistic model based on k-mer frequencies and single-copy marker genes. These methods are sensitive to sequencing depth and sample size, often suffering from reduced performance when metagenomic data are sparse.

(iii) **Graph feature-based methods**
Graph-based methods like GraphBin [19] and HiCBin [20] have improved metagenomic binning. GraphBin uses graph theory to partition contigs based on overlap, while HiCBin utilizes Hi-C data to build contact maps and applies HiCZin normalization along with the Leiden algorithm [21]. Though effective, these methods are sensitive to assembly errors and heavily rely on high-quality Hi-C data for accurate binning.

In summary, metagenomic contig binning methods have made significant progress but still face some limitations. For example, in the case of single-sample metagenomes, most binning methods perform poorly due to the lack of effective coverage features.In contrast, our proposed DCVBin is expressly designed for single-sample scenarios by substituting coverage information with semantic features extracted from a DNA language model. DCVBin tackles the computational challenge of fusing compositional features with high-dimensional semantic vectors, effectively managing the sparsity and complexity of these embeddings within a unified VAE framework.

Materials and methods

The structure of the DCVBin model is shown in Fig. 1, consisting of three core stages. First, DCVBin extracts 4-mer frequencies and semantic features from DNA contigs. The semantic features are obtained from a DNA language model, DNABERT_S, through a two-stage process. First, DNABERT_S is continuously pretrained using DNA contigs to be binned, followed by a feature extraction stage where the trained model is used to generate the final semantic features. Next, the VAE network is constructed to fuse and jointly learn the semantic and k-mer features. Finally, the latent embeddings of the VAE are clustered using k-means algorithm whose initial clusters number is determined by single-copy genes (SCGs).

Preprocessing and features extraction

Compositional feature extraction

We utilize 4-mer frequency distributions to characterize sequence composition, limiting k to 4 to balance phylogenetic signal retention against computational complexity. Our approach adheres to the *de facto* standards set by widely adopted binning tools such as MetaBAT2 [15].

Following the analytical framework established by the VAMB method [16], we employ a constraint-based dimensionality reduction strategy to process k-mer frequencies. Instead of empirical reduction, we explicitly construct a linear constraint system based on biological invariants, where each constraint constitutes a specific row in the coefficient matrix $\boldsymbol{A}$. The theoretical space of tetranucleotides is

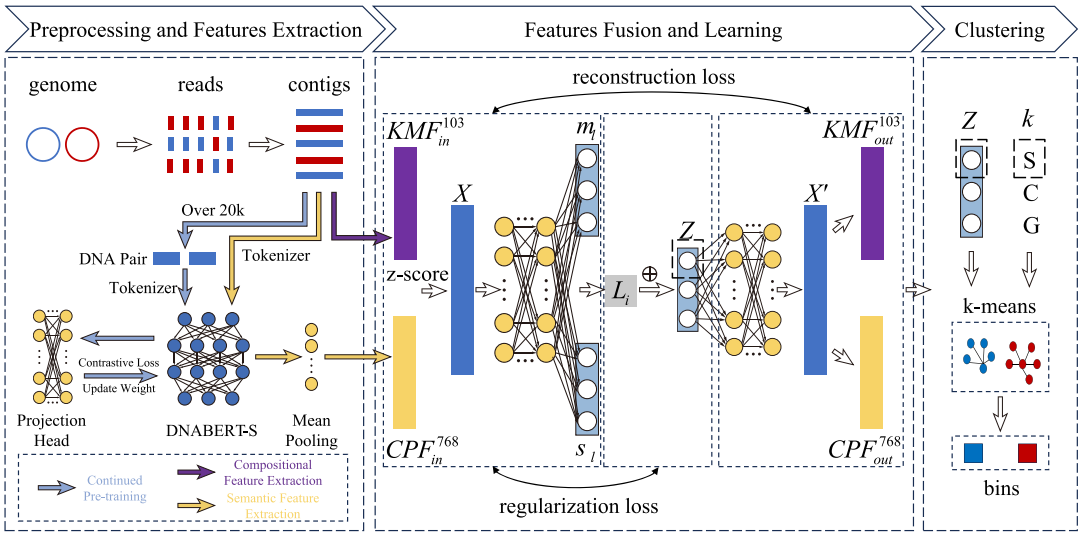

Figure 1 The structure of the DCVBin model.

governed by three fundamental dependency constraints that define the structure of $\boldsymbol{A}$.

The first is the (i) **Sum constraint**: the sum of all k-mer frequencies must equal 1, i.e.

$$P_{AAAA} + P_{AATT} + \cdots + P_{CCGC} = 1 \tag{1}$$

This global summation invariant forms the first row of our linear system. The second is the (ii) **Complementarity constraint**: the frequency of a k-mer should be equal to the frequency of its reverse complement sequence, i.e.

$$P_{AAAA} = P_{TTTT}, P_{CCCC} = P_{GGGG} \tag{2}$$

Incorporating these equality constraints allows us to group reverse complements, effectively reducing the initial parameter space to 136 canonical components. The third is the (iii) **Edge constraint**: The marginal probabilities at different k-mer boundaries should remain consistent, i.e.

$$P_{AAAA} + P_{AAAT} + P_{AAAC} + P_{AAAG} = P_{AAA} \tag{3}$$

This constraint represents the flux conservation law on the de Bruijn graph, ensuring that the frequency influx into any trimer node equals its efflux.

Collectively, these constraints define a linear system $\boldsymbol{Ap} = \boldsymbol{b}$, where $\boldsymbol{p}$ represents the vector of the 136 canonical components. A rank analysis of the coefficient matrix $\boldsymbol{A}$ reveals a rank of 33, indicating 33 linearly independent constraints (comprising the 1 sum constraint and 32 independent edge constraints derived from canonical trimers). To resolve the intrinsic feature space, we compute the **null space** of $\boldsymbol{A}$ using Singular Value Decomposition, yielding an orthonormal basis matrix $\boldsymbol{B} \in \mathbb{R}^{136\times 103}$. We then project the canonical frequencies onto $\boldsymbol{B}$, resulting in a compact 103D representation ($136 - 33 = 103$) that preserves all independent compositional information while eliminating redundancy.

Semantic feature extraction

To capture high-level semantic information tailored to the specific microbial community structure, we utilize the DNABERT_S model [22] with a target-adaptive continued pretraining strategy. Unlike standard approaches that rely solely on static pretrained weights, our method dynamically adapts the model to the target metagenomic dataset prior to feature extraction.

We generate a domain-specific pretraining dataset directly from the input contigs. Contigs longer than 20 kbp are retained and segmented into nonoverlapping 10 kbp fragments. To construct positive pairs for contrastive learning, we exhaustively combine segments derived from the same contig, enforcing a "same-genome" constraint. The resulting pair dataset is randomly partitioned into training and validation sets with a 90:10 ratio to monitor convergence.

The model undergoes continued pretraining for three epochs using a curriculum learning framework designed to progressively enhance feature robustness. Stage I (Epoch 0): the model is optimized using a standard Hard Contrastive Loss, focusing on fundamental instance discrimination without feature mixing. Stage II (Epochs 1–2): we introduce Manifold Instance Mixup (MI-Mix) with a mixing coefficient $\alpha = 1.0$. This stage interpolates hidden states between samples, regularizing the model to learn smoother decision boundaries in the semantic space.

Hyperparameters are configured to balance computational efficiency with adaptation performance. The model is trained with a batch size of 8 and a maximum sequence length of 2000 tokens. We employ the AdamW optimizer with a base learning rate of 1×10^{-6} and a scaling factor of 100. The contrastive temperature (τ) is set to 0.05. A projection head maps the transformer embeddings to a 128D space for loss calculation. Following training, the checkpoint with the lowest validation loss is selected as the optimal encoder for the subsequent extraction phase.

For the final representation, input contigs are tokenized into n k-mers, denoted as $I = (I_1, I_2, \ldots, I_n)$. We extract the hidden states from the last transformer layer, yielding a 768D vector for each token:

$$\mathbf{O}_{I_i} \in \mathbb{R}^{768}, \quad i = 1, 2, \ldots, n \tag{4}$$

A global semantic vector is obtained via mean pooling across the sequence dimension:

$$\mathbf{CPF} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{O}_{I_i} \tag{5}$$

This 768D vector *CPF* serves as the semantic input for the downstream VAE fusion module.

Features fusion and learning

Encoder: hierarchical feature encoding and data preprocessing

To effectively integrate the heterogeneous modalities of sequence composition and semantic information, we employ an early fusion strategy via feature concatenation. Specifically, the dimensionality-reduced compositional feature and the semantic feature are concatenated to form a unified input vector. Here, we denote the 103D compositional feature vector as KMF_{in}^{103}) and the 768D semantic feature vector as CPF_{in}^{768}. We opt for a concatenation-based multilayer perceptron architecture. Our input consists of two fixed-length global feature vectors. The subsequent fully connected layers enable the model to automatically learn complex nonlinear interactions between the compositional and semantic domains through backpropagation, providing a computationally efficient and robust fusion capability.

The encoder is implemented as a three-layer fully connected neural network. The input layer processes the 871D concatenated vector, followed by two successive hidden layers with 256 neurons each. The architecture culminates in two parallel output branches that generate 32D mean ($\boldsymbol{\mu}$) and standard deviation ($\boldsymbol{\sigma}$) vectors, respectively, parameterizing the latent space distribution.

Prior to feature encoding, input data undergo z-score normalization to ensure feature scale uniformity:

$$x_z = \frac{x - \mu_x}{\sigma_x}, \tag{6}$$

where μ_x and σ_x denote the feature-wise mean and standard deviation.

Each layer of the encoder applies Leaky ReLU activation, which is defined as

$$f_L(c) = \begin{cases} c, & c > 0 \\ \alpha c, & c \le 0 \end{cases}, \tag{7}$$

where c is the input value and α is a small constant (typically 0.01) used to control the slope of the negative part. When c is less than or equal to 0, the output of the Leaky ReLU is αc, where the negative part is not completely discarded, ensuring that the neuron remains updated during training.

Latent space representation: reparameterization and sampling

The encoder outputs mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_l)$ and standard vector $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_l)$ for the l-dimensional latent space. Each latent dimension z_i is independently sampled from its corresponding Gaussian distribution $\mathcal{N}(\mu_i, \sigma_i^2)$ using the reparameterization trick:

$$z_i = \mu_i + \sigma_i \cdot \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0,1) \tag{8}$$

This per-dimension formulation ensures that the complete latent vector $\mathbf{z} = (z_1, z_2, \ldots, z_l)$ follows a multivariate Gaussian distribution, with each dimension sampled independently from $\mathcal{N}(\mu_i, \sigma_i^2)$. The reparameterization separates the stochastic component (ϵ_i) from the deterministic network outputs (μ_i, σ_i), allowing gradient backpropagation through the sampling operation. After model convergence, these latent representations $\mathbf{z}$ are clustered using K-means for downstream analysis.

Decoder: feature reconstruction and output generation

The decoder is responsible for reconstructing the original input features from the latent representation $\mathbf{Z}$. This is achieved by using fully connected layers symmetric to the encoder but in reverse order. The output is a vector $\mathbf{X}'$ with dimensions 103+768, which is then split into two vectors: KMF_{out}^{103} (103 dimensions) and CPF_{out}^{768} (768 dimensions). The decoder's goal is to generate outputs as close as possible to the input features KMF_{in}^{103} and CPF_{in}^{768}, with the reconstruction error being minimized during the training process.

Loss function: optimization and regularization

To train the VAE model, a loss function is designed that optimizes both the reconstruction of the input features and the validity of the latent representation. The objective is to maximize the log-likelihood of the data, ensuring that the encoded data's probability distribution matches the original data. The loss function is composed of three main components:

$$\begin{aligned} Loss =& \alpha \sum (KMF_{out}^{103} - KMF_{in}^{103})^2 \\ &+ \beta \sum (CPF_{out}^{768} - CPF_{in}^{768})^2 \\ &+ \gamma \sum \frac{1}{2}\left(\sigma_i^2 + \mu_i^2 - \log\sigma_i^2 - 1\right) \end{aligned}, \tag{9}$$

where the first two terms are reconstruction errors of the two vectors, using ℓ_2 loss as the reconstruction error metric. α and β are weighting coefficients controlling the influence of the reconstruction errors of the two feature types in the overall loss function. The last term is the KL divergence, which is a measure of the difference between two probability distributions. Here, γ is the weight coefficient for the KL divergence term, controlling the influence of the regularization term, and $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are the mean and standard deviation of the latent distribution, respectively.

To introduce nonlinearity and improve the stability of the training process, Batch Normalization (BatchNorm) is applied. Additionally, Dropout is employed to mitigate overfitting during training.

Clustering

We apply k-means clustering to obtain final draft genomes, randomly selecting k points from the VAE latent embeddings as initial centroids. Since k-means requires predefining the number of clusters, single-copy marker genes are typically used to determine this number. These genes appear once per genome, so the ideal cluster count matches the number of bins containing the same gene. However, due to assembly errors and sequencing depth, some genomes may lack SCGs, meaning using their number to define clusters could discard genomes, impacting binning performance.

The method proposed in this paper determines the optimal cluster number through a two-stage process. First, we identify SCGs by

Table 1 MAGs quality criteria.

Quality grade	Completeness	Contamination
High quality	≥ 90%	< 5%
Medium quality	≥ 50%	< 10%
Low quality	< 50%	< 10%

predicting ORFs with FragGeneScan [23] and verifying them against reference databases. Note that FragGeneScan identifies gene regions but does not directly determine SCGs; SCGs are estimated by comparing gene copy frequencies and validating against genomic databases. Then, we use the third quartile (Q_3) of marker gene counts as the lower bound (k_1) and the maximum count as the upper bound (k_2). Subsequently, we perform k-means clustering for each integer $k \in [k_1, k_2]$ and select the value maximizing the silhouette score:

$$S = \frac{1}{N}\sum_{i=1}^{N}\frac{b(i)-a(i)}{\max(a(i),b(i))}, \tag{10}$$

where $a(i)$ is the average intra-cluster distance for bin i, $b(i)$ is its average distance to the nearest neighboring cluster, and N is the total bin count. This data-driven approach ensures robust cluster number determination while minimizing outlier effects through the use of Q_3 for initial estimation.

Results

Evaluation indicators

In metagenomics, reliable species analysis based on metagenome bins requires accurate assessment of genome completeness and contamination. To achieve this, we employ CheckM2 [24], a metagenome quality assessment tool, to evaluate the quality of genomes assembled through clustering. According to the quality standards set by CheckM2, we quantify the number of draft genomes with contamination levels below 5% and categorize them based on completeness thresholds (≥ 50%, ≥ 60%, ≥ 70%, ≥ 80%, ≥ 90%), which serve as key metrics for evaluating clustering performance on real datasets. Furthermore, using the metagenome-assembled genomes (MAGs) criteria [25], we categorize draft genomes into high-quality and medium-quality groups. The specific MAGs criteria are outlined in Table 1.

For simulated datasets with known labels, we evaluate clustering performance using three metrics: FMI (Fowlkes–Mallows Index), ARI (Adjusted Rand Index), and NMI (Normalized Mutual Information), offering a comprehensive assessment for further optimization.

Dataset description

To validate the effectiveness of the proposed method, this study utilizes four real datasets (SRR17635647, SRR17635658, SRR17635503, SRR17858159) and two simulated datasets (Sim5g, Sim10g). Table 2 summarizes the detailed information of the datasets used in this study.

The real datasets used in this study are sourced from the Human Vaginal Metagenome (HVM) project. Four single-sample metagenome datasets are randomly selected from this project for the experiments; all datasets are publicly available on the NCBI website (https://www.ncbi.nlm.nih.gov/). Additionally, to further assess the performance of the method across different datasets, two simulated datasets, Sim-5G and Sim-10G, are employed. These datasets are preconstructed within the GraphBin [19] framework and contain 5 and 10 species, respectively. Both datasets are generated using the InSilicoSeq tool.

Table 2 Detailed information of datasets used in experiments.

Data set	Size (MB)	Number of contigs	N50 (bp)
SRR17858159	34.95	5822	1758
SRR17635658	31.91	3542	2619
SRR17635503	59.14	7830	2063
SRR17635647	78.54	11 110	1532
Sim5g	27.10	544	258 393
Sim10g	43.80	981	315 263

Table 3 Clustering metrics of four methods on two simulated datasets.

Data set	Indicator	VAMB	MetaBAT2	MetaDecoder	DCVBin
Sim5g	ARI	0.529	0.571	0.862	**0.887**
	NMI	0.670	0.730	0.852	**0.876**
	FMI	0.651	0.665	0.894	**0.915**
Sim10g	ARI	0.688	0.687	0.772	**0.876**
	NMI	0.824	0.819	0.825	**0.901**
	FMI	0.728	0.727	0.802	**0.892**

Note: Bold values indicate the best performance for each metric.

DCVBin outperforms other binning methods

In this study, the proposed method is compared with three mainstream binning methods: VAMB, MetaBAT2, and MetaDecoder. To evaluate the clustering performance of each method across various datasets, two simulated datasets and four real datasets are used. All datasets are assembled using the MetaSpades assembler and undergo identical preprocessing and normalization procedures. The three compared clustering methods are executed with their default parameters, as specified in their respective publications, to ensure consistency in the evaluation.

(i) **Simulated datasets**

Table 3 presents the clustering results of the four methods on the Sim5g and Sim10g datasets, showing that DCVBin consistently outperforms the other three methods on both. On the simpler Sim5g dataset, DCVBin achieves the highest scores across all three metrics (ARI, NMI, and FMI), indicating superior clustering accuracy. Furthermore, on the more complex Sim10g dataset, DCVBin maintains the best performance across all evaluation metrics, confirming its robustness. Notably, DCVBin's advantage in NMI and FMI is significant; its FMI score of 0.892 substantially surpasses MetaDecoder's 0.802, representing a nearly 9% improvement in clustering consistency and accuracy. Overall, DCVBin demonstrates strong superiority and robustness across datasets of varying complexity.

(ii) **Real datasets**

The number of genomes with different completeness thresholds (≥ 50%, ≥ 60%, ≥ 70%, ≥ 80%, ≥ 90%) is separately counted when the contamination is below 5%. The CheckM2 evaluation results on the real datasets are shown in Fig. 2. For the SRR17635658 and SRR17858159 datasets, our method obtains the largest total number

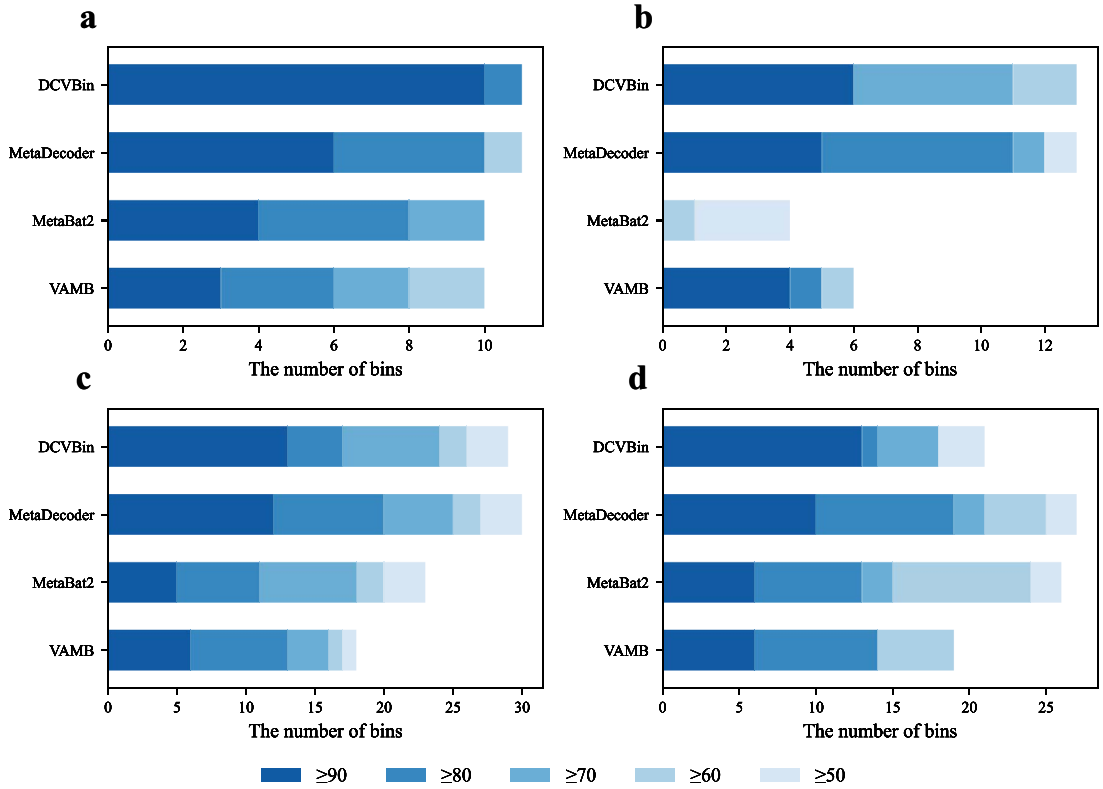

Figure 2 The number of bins with contamination below 5% and different completeness levels. (a) SRR17635658; (b) SRR17858159; (c) SRR17635647; (d) SRR17635503.

of genomes, tied with MetaDecoder as the highest among the four methods. Moreover, our method recovers 10 and 6 genomes with completeness ≥ 90% in these datasets, respectively, while MetaDecoder recoveres six and five genomes, respectively. This indicates that our method not only recovers more genomes but also produces the greatest number of high-quality genomes on these two datasets.

For the SRR17635647 and SRR17635503 datasets, although MetaDecoder and MetaBAT2 recover a larger total number of genomes, most of these genomes have completeness below 90%, indicating limited ability to recover high-quality genomes. In contrast, our method recovers more genomes with high completeness, demonstrating its superior recovery capability for high-quality genomes.

To further validate the clustering performance of the model, we use the MAGs evaluation criteria to count and analyze the number of high-quality and medium-quality genomes. The experimental results are shown in Fig. 3. The results indicate that, on the SRR17858159 dataset, our method achieves the highest number of both high-quality and medium-quality genomes, demonstrating its superior ability to recover high-quality genomes from single metagenome samples in this dataset. Additionally, our method continues to perform outstandingly on the other three real datasets, achieving the greatest number of high-quality genomes. For instance, on the SRR17635658 dataset, our method successfully recovers 10 high-quality genomes, while the second-best method, MetaDecoder, recovered only six. In terms of medium-quality genomes, MetaBAT2 and MetaDecoder show relatively better performance, suggesting that these methods may be more suitable for recovering genomes with lower completeness in certain scenarios. Nevertheless, the exceptional performance of our method in recovering high-quality genomes underscores its effectiveness in species identification and genome reconstruction tasks.

Statistical significance analysis

To rigorously evaluate the clustering performance across multiple datasets, we apply a Friedman test followed by a Nemenyi *post hoc* analysis on six datasets, which include two simulated datasets and four real-world metagenomic datasets. The four methods are ranked based on their primary evaluation metrics: ARI for the simulated datasets and the number of high-quality bins for the real-world datasets.

As shown in Fig. 4a, DCVBin consistently performs the best across all six datasets, achieving an average rank of 1.00. MetaDecoder ranked second with an average rank of 2.00, followed by VAMB and MetaBAT2, which displayed relatively weaker performance.

The Friedman test reveals statistically significant differences among the methods ($\chi^2 = 16.53, P < 0.001$), confirming that the observed performance variations are unlikely to be attributed to random chance. The subsequent Nemenyi *post hoc* analysis, using a critical difference (CD) of 1.91, confirmed that DCVBin significantly outperformed VAMB ($P = 0.007$) and MetaBAT2 ($P = 0.003$). As illustrated in Fig. 4b, the pairwise comparison *P*-values from the Nemenyi test further validate these significant differences between DCVBin and VAMB, as well as between DCVBin and MetaBAT2.

Although the difference between DCVBin and MetaDecoder does not reach statistical significance under the Nemenyi test ($P = 0.536$), DCVBin consistently ranked higher than MetaDecoder across

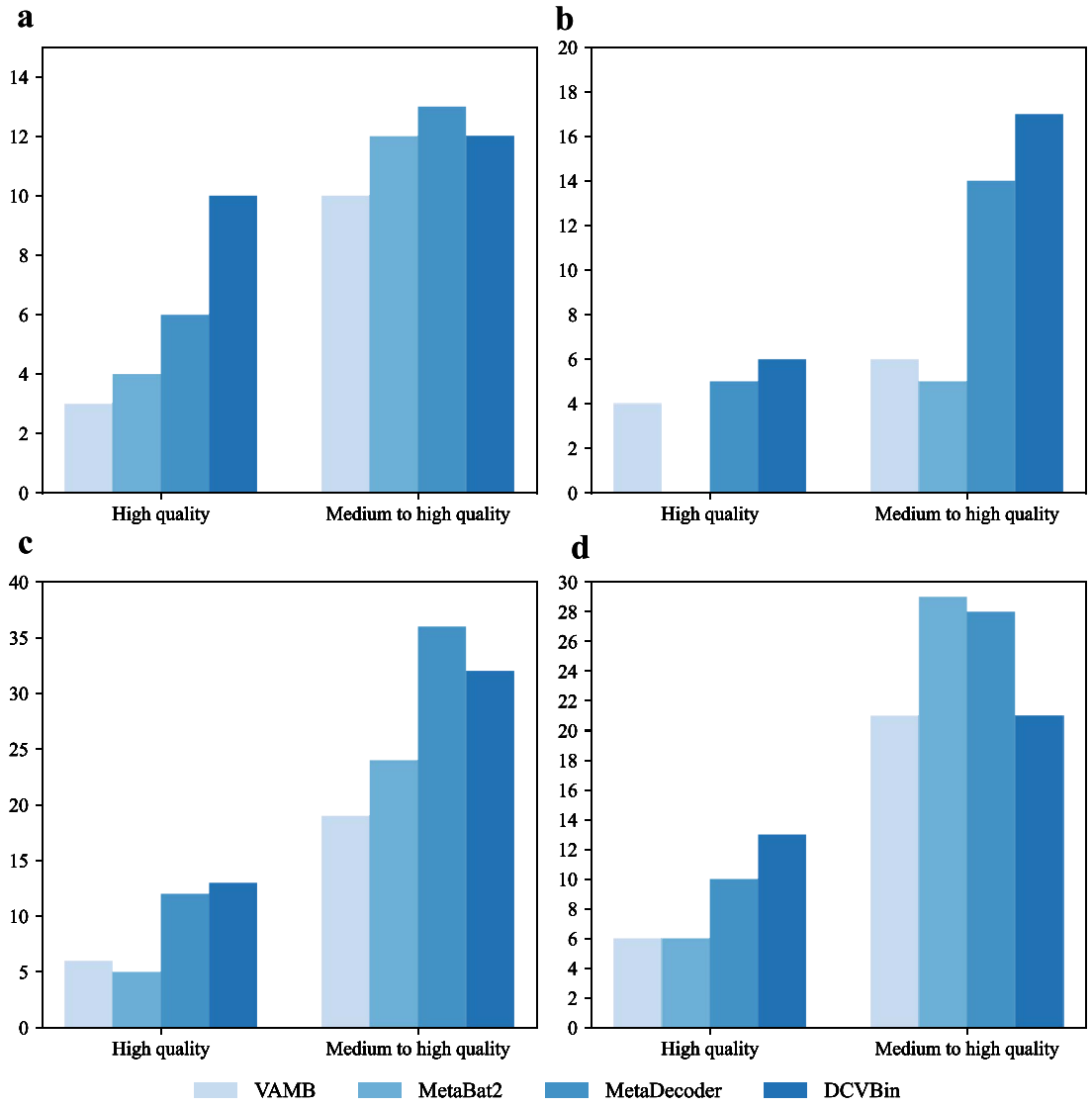

Figure 3 The number of high- and medium-quality bins. (a) SRR17635658; (b) SRR17858159; (c) SRR17635647; (d) SRR17635503.

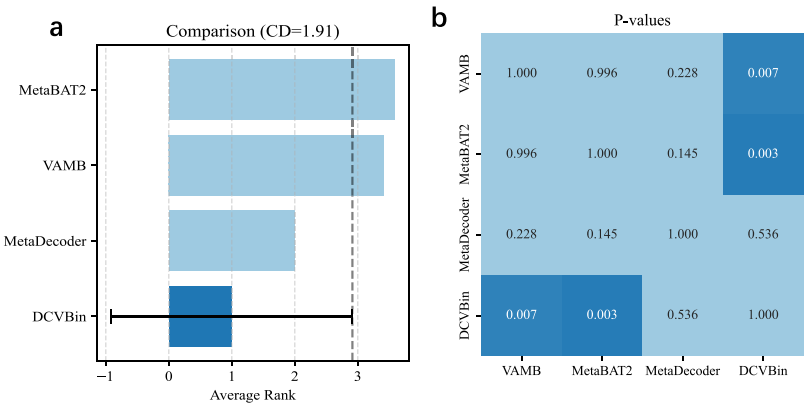

Figure 4 Statistical significance analysis. (a) Average rankings from the Friedman test across six datasets. (b) *P*-values from the Nemenyi *post hoc* test.

all datasets. This consistent ranking advantage demonstrates that DCVBin maintains a stable and robust performance across diverse experimental conditions.

Ablation study

We conduct ablation experiments on the *Sim5g* and *Sim10g* datasets to evaluate the contributions of CPT and VAE-based feature fusion.

Table 4 Clustering metrics of four methods on two simulated datasets.

Dataset	Method	ARI	NMI	FMI
Sim5g	K-mer	0.748	0.813	0.808
	DNABERT_2	0.445	0.588	0.563
	DNABERT_S	0.773	0.832	0.825
	DNABERT_S + CPT	0.854	0.869	0.889
	DNABERT_S + CPT + VAE	**0.887**	**0.877**	**0.915**
Sim10g	K-mer	0.749	0.831	0.785
	DNABERT_2	0.513	0.691	0.576
	DNABERT_S	0.811	0.872	0.835
	DNABERT_S + CPT	0.850	0.893	0.870
	DNABERT_S + CPT + VAE	**0.876**	**0.902**	**0.892**

As summarized in Table 4, we compare five configurations: (1) **k-mer**: a baseline using solely 4-mer compositional features; (2) **DNABERT_2**: a general-purpose comparison model; (3) **DNABERT_S**: the unadapted backbone model; (4) **DNABERT_S + CPT**: the backbone enhanced with domain-specific pretraining on target contigs; and (5) **DNABERT_S + CPT + VAE**: the full model that jointly encodes and reconstructs compositional and semantic features via a VAE to optimize latent embeddings.

Our ablation analysis **highlights** the contribution of each component. As shown in Table 4, the **k-mer** baseline **demonstrates** relatively robust performance across both datasets (ARI $\approx$ 0.75), surprisingly outperforming the general-purpose **DNABERT_2**. Notably, **DNABERT_2 struggles** to capture the specific genomic patterns effectively without adaptation, yielding the lowest scores on both Sim5g (ARI 0.445) and Sim10g (ARI 0.513). In contrast, **DNABERT_S proves** to be a superior foundation model, consistently surpassing both K-mer and DNABERT_2. The introduction of the continued pretraining strategy (**DNABERT_S + CPT**) **yields** substantial gains, improving ARI from 0.773 to 0.854 on Sim5g, which **confirms** that domain-adaptive pretraining effectively **refines** feature alignment to target distributions. Crucially, the full model (**DNABERT_S + CPT + VAE**) **achieves** optimal clustering results across all metrics and datasets. By integrating adaptive semantic encoding with compositional features through probabilistic fusion, the full model significantly **enhances** the discriminative power of the latent embeddings.

We further assessed the generalization capability of CPT using cross-dataset transfer experiments on both simulated and real datasets. The detailed results are provided in Supplementary Note 1 and Supplementary Figs S1–S3.

DCVBin assists analysis of potential disease-associated microbial biomarkers in colorectal cancer

CRC is one of the most common malignant tumors in humans, primarily driven by genomic alterations that lead to abnormal cell proliferation [26–28]. Annually, ~1 million new cases are diagnosed globally, with >600 000 deaths directly or indirectly attributed to the disease [29, 30]. To address the challenge of low prediction accuracy for CRC based on human gut microbiome data, we apply our single-sample metagenomic contig binning model, DCVBin, to this field, demonstrating its research significance and practical value.

The prediction model framework is illustrated in Fig. 5a. First, the CRC metagenomic dataset CRC-AT (PRJEB7774) is obtained from the ENA database. This dataset includes 109 human samples, comprising 46 healthy individuals and 63 CRC patients. After quality control, raw reads from these metagenomes are assembled into contigs. The dataset is then split into training and testing sets in an 8:2 ratio, resulting in 87 samples for training and 22 samples for testing. Next, the DCVBin model is employed to perform binning on the contigs from the 87 training samples, generating 1891 draft genomes with completeness $\geq$ 50% and contamination $<$ 10%. These 1891 bins are further clustered into 596 species-level genome bins, which were subsequently annotated using the GTDB-TK tool [31]. Finally, a binary classifier based on the random forest (RF) algorithm [32] is developed to predict disease status.

Using the constructed species database, species feature extraction is performed on both the training and testing sets, focusing on two main aspects: (1) the extraction of relative abundance features and (2) batch effect correction (batch correction) for the extracted features.

The calculation formula for the relative abundance of species is as follows:

$$Ab(i) = \frac{rc_i}{\sum_{j=1}^{N} rc_j}, \tag{11}$$

where Ab represents the relative abundance of species, which is calculated as the number of reads aligned to the species divided by the species genome length, and N is the number of species.

Since the exact genome length of species is unknown, an alternative calculation method is employed [33], where the normalized read count for each contig is calculated by dividing the number of reads aligned to each contig by the length of that contig. The average value of all contigs assigned to a specific species is then used as the relative abundance estimate for that species. Using this method, relative abundance features are extracted for both the training and testing sets, serving as the input features. The final feature dimensions are 87×596 for the training set and 22×596 for the testing set, where 87 and 22 represent the number of samples, respectively, and 596 denotes the number of species in the constructed database.

Batch effects can arise in multiple datasets due to factors such as different experimental run batches and instrument variations [34, 35]. These effects may introduce noise into the data, making it necessary to remove the bias caused by batch differences to ensure the accuracy and reliability of the analysis results. The batch-corrected feature vector matrix is then used as the input data for the subsequent classification model.

We compare our method with two widely used disease prediction approaches based on metagenomic species composition: DeepMicro [36] and GDmicro [37]. The model performance is evaluated using three metrics: accuracy (ACC), recall, and area under the curve (AUC). In the comparative experiments, both baseline methods are run using their default parameter settings. Figure 5b summarizes the comparison results of the three methods across these metrics. Our proposed method outperforms the other two methods in terms of AUC, ACC, and recall values. The superior predictive performance of our method can be attributed to the use of the DCVBin model for constructing a species-level genome database. This database supplements species information lost in traditional feature extraction methods based on reference sequences or binning methods. By providing more comprehensive species information, it enhances the feature space and, consequently, improves the accuracy of disease prediction.

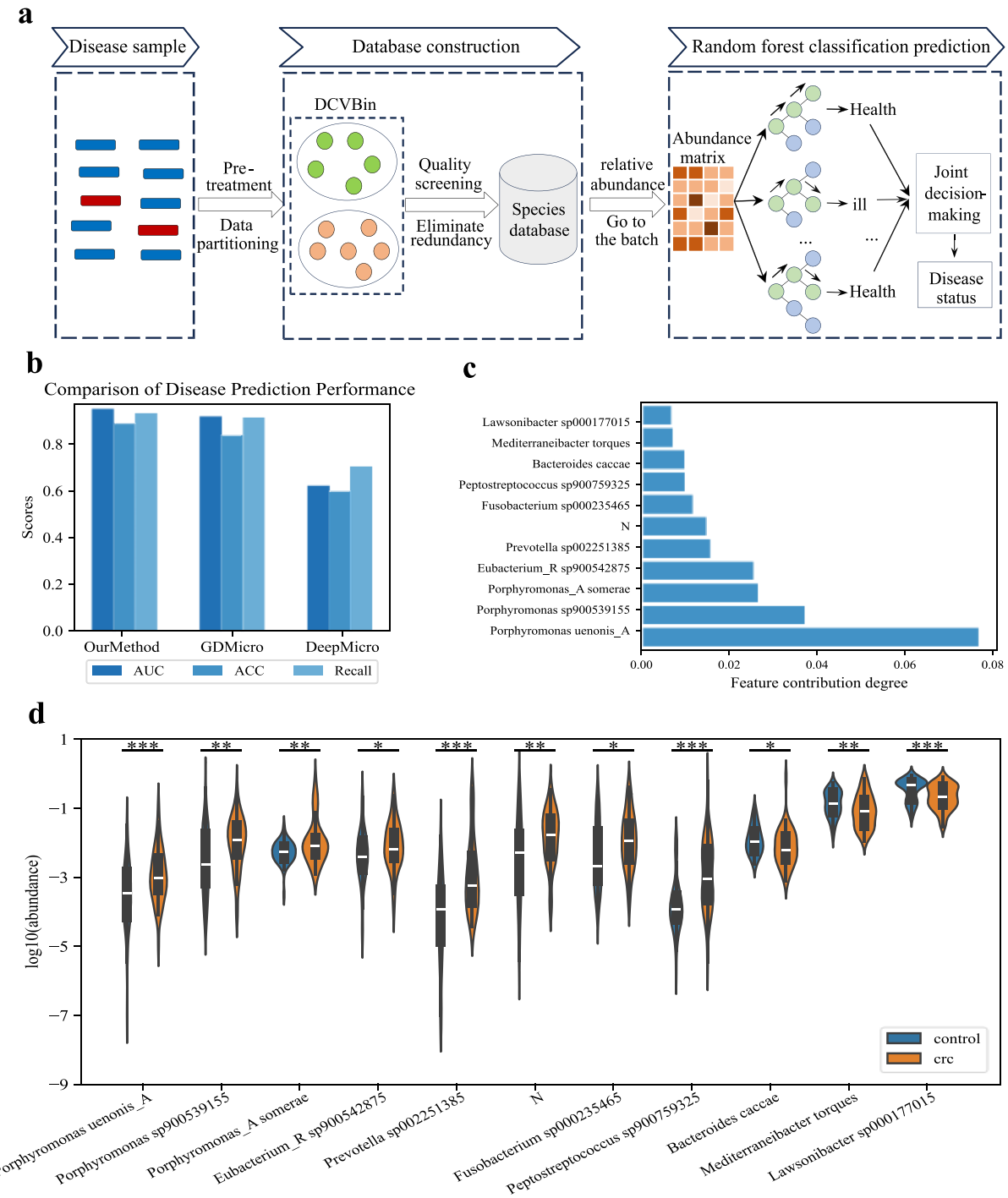

Figure 5 Disease prediction flowchart.

The contribution of each species feature is extracted from the trained RF model. Figure 5c displays the top 11 most important species in the disease prediction model. The species names are generated by the GTDB annotation tool, based on the current reference tree and classification standards. Some species names end with genus names and letter suffixes, indicating that these taxonomic units are not fully resolved and require further classification based on clustering rules and majority voting. Additionally, GTDB uses NCBI accession numbers (e.g. sp9005339155) as placeholders for unnamed species. These species may not yet be registered in the NCBI official database but require confirmation through further research.

Statistical analysis is performed between the CRC group and the control group. The Mann–Whitney U test is applied, along with False Discovery Rate and Bonferroni corrections, to identify species with significantly different abundances between the two groups. As shown in Fig. 5d, the plot illustrates the statistical analysis of the 11 species from Fig. 5c between the CRC and control groups. Our analysis reveals that certain species, confirmed in previous studies, play

significant roles in CRC diagnosis. Species such as *Porphyromonas uenonis_A*, Fusobacterium sp000235465, Lawsonibacter sp000177015, and Mediterraneibacter torques exhibit notable differences in abundance across the tests, highlighting their potential key roles in CRC diagnosis. *Porphyromonas uenonis_A* is enriched in cancer samples and belongs to the genus *Porphyromonas*, which is known for its association with multiple diseases, including CRC [38]. In contrast, *Lawsonibacter sp000177015* and *Mediterraneibacter torques* are more abundant in normal samples. *Lawsonibacter sp000177015*, belonging to the genus *Lawsonibacter*, is known for butyrate production [39], a key metabolite that fuels intestinal epithelial cells and is involved in carbohydrate metabolism and sodium absorption in the gut [40, 41], making it essential for gut health. *Mediterraneibacter torques* is a dominant gut symbiont that helps maintain gut barrier function and exhibits anti-inflammatory effects [42]. Notably, *Peptostreptococcus sp900759325* and *Prevotella sp002251385* are found to be more abundant in CRC samples than in control samples. These species, belonging to the genera *Porphyromonas*, *Peptostreptococcus*, and *Prevotella*, are widely distributed in the oral cavity and gut and are closely linked to respiratory infections, CRC, and other diseases [38].

In addition to these previously validated species, we also identify some strains not yet reported in the literature, which may hold significant diagnostic potential for CRC. The marked differences in the abundance of these species suggest they could play crucial roles in the development or progression of CRC and may have potential as microbial biomarkers for CRC diagnosis. For instance, an unknown species *N*, within the class *Alphaproteobacteria*, shows increased abundance in cancer samples, suggesting a possible association with disease progression. The changes in this species' abundance may affect the gut microbial balance and it could serve as a potential biomarker for early CRC diagnosis and mechanistic studies.

Conclusion

Single-sample metagenomic binning is a pivotal technique in microbial research. DCVBin distinguishes itself from standard binning methods by eliminating the reliance on multi-sample coverage profiles, making it uniquely effective for scenarios where such data are absent. By integrating high-dimensional semantic features from a DNA language model (DNABERT-S) with k-mer frequencies, DCVBin demonstrates that deep learning embeddings can effectively compensate for the lack of abundance information inherent in single-sample data. This study further underscores the power of continued pretraining, which allows the VAE-based architecture to capture domain-specific representations tailored to complex metagenomic landscapes.

Our generalization analysis reveals that while the model captures transferable semantic features—particularly when pretrained on large-scale datasets—this capability is scale-dependent. Ultimately, the consistent superiority of the Self-CPT strategy validates that target-specific refinement is indispensable for optimal precision in single-sample binning. Crucially, the methodology underlying DCVBin is versatile and holds significant potential for application across diverse ecological contexts, from human microbiomes to varied environmental samples.

While the proposed model establishes a robust baseline, certain challenges remain. The computational overhead and GPU dependency of Transformer-based architectures are higher than pure k-mer methods like MetaBAT2. To fully capture the inherent complexity of microbial communities, future work will focus on optimizing inference efficiency and integrating broader feature types, such as structural elements derived from assembly graphs. This multidimensional feature fusion strategy is expected to enhance model generalizability and binning continuity, thereby making a significant contribution to community structure analysis, pathogen discovery, and environmental biotechnology.

Key Points

- DCVBin is a novel method for metagenomic genome reconstruction from a single sample, overcoming the need for multiple samples.
- The approach integrates DNA language model-derived semantic features with traditional sequence composition (4-mer frequencies) using a variational autoencoder to improve contig representation.
- Extensive benchmarks across six datasets show that DCVBin outperforms existing tools in single-sample binning accuracy.
- DCVBin is successfully integrated into a diagnostic framework for colorectal cancer prediction from gut microbiomes and for the identification of potential microbial biomarkers.

Acknowledgments

The authors thank the anonymous reviewers for their valuable suggestions.

Supplementary material

Supplementary material is available at *Briefings in Bioinformatics* online.

Conflicts of interest

None declared.

Funding

This work was supported by funds from the National Natural Science Foundation of China (62303193), Science and Technology Development Plan Project of Jilin Province, China (20230101064JC), and the Fundamental Research Funds for the Central Universities.

Data availability

All data used are from public repositories: Real datasets are available on NCBI (Human Vaginal Metagenome project), and simulated datasets (Sim5g, Sim10g) were sourced from the GraphBin framework (generated by InSilicoSeq). The DCVBin code is available at https://github.com/dengdengf/dcvbin.

References

1. Dobrovol'skaya TG, Zvyagintsev DG, Chernov IY *et al.* The role of microorganisms in the ecological functions of soils. *Eurasian Soil Sci* 2015;**48**:959–67. https://doi.org/10.1134/S1064229315090033

2. Gao R, Gao Z, Huang L *et al.* Gut microbiota and colorectal cancer. *Eur J Clin Microbiol Infect Dis* 2017;**36**:757–69. https://doi.org/10.1007/s10096-016-2881-8
3. Pandey H, Tang DWT, Wong SH *et al.* Gut microbiota in colorectal cancer: biological role and therapeutic opportunities. *Cancers (Basel)* 2023;**15**:866. https://doi.org/10.3390/cancers15030866
4. Herazo-Álvarez J, Mora M, Cuadros-Orellana S *et al.* A review of neural networks for metagenomic binning. *Brief Bioinform* 2025;**26**:bbaf065. https://doi.org/10.1093/bib/bbaf065
5. Mattock J, Watson M. A comparison of single-coverage and multi-coverage metagenomic binning reveals extensive hidden contamination. *Nat Methods* 2023;**20**:1170–3.
6. Pasolli E, Asnicar F, Manara S *et al.* Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell* 2019;**176**:649–662.e20. https://doi.org/10.1016/j.cell.2019.01.001
7. Liu C, Li Z, Ding J *et al.* Species-level analysis of the human gut microbiome shows antibiotic resistance genes associated with colorectal cancer. *Front Microbiol* 2021;**12**:765291. https://doi.org/10.3389/fmicb.2021.765291
8. Loftus M, Hassouneh SA, Yooseph S. Bacterial community structure alterations within the colorectal cancer gut microbiome. *BMC Microbiol* 2021;**21**:98.
9. Piccinno G, Thompson KN, Manghi P *et al.* Pooled analysis of 3,741 stool metagenomes from 18 cohorts for cross-stage and strain-level reproducible microbial biomarkers of colorectal cancer. *Nat Med* 2025;**31**:2416–29. https://doi.org/10.1038/s41591-025-03693-9
10. Rybkin O, Daniilidis K, Levine S. Simple and effective VAE training with calibrated decoders. *Proceedings of the 38 th International Conference on MachineLearning*, PMLR 139, Virtual metting, pp. 9179–89, 2021.
11. Teeling H, Waldmann J, Lombardot T. *et al.* TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in dna sequences. *BMC Bioinformatics* 2004;**5**:163. https://doi.org/10.1186/1471-2105-5-163
12. Edelmann D, Móri TF, Székely GJ. On relationships between the Pearson and the distance correlation coefficients. *Stat Probab Lett* 2021;**169**:108960.
13. Chatterji S, Yamazaki I, Bai Z, Eisen JA. CompostBin: a DNA composition-based algorithm for binning environmental shotgun reads. *Research in Computational Molecular Biology*, RECOMB 2008, vol **4955**, pages 17–28, Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-78839-3_3
14. Yang B, Peng Y, Leung HCM *et al.* MetaCluster: unsupervised binning of environmental genomic fragments and taxonomic annotation. *Proceedings of the first ACM international conference on bioinformatics and computational biology*, Association for Computing Machinery, New York, NY, USA, pp. 170–9, 2010.
15. Kang DD, Li F, Kirton E *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* 2019;**7**:e7359. https://doi.org/10.7717/peerj.7359
16. Nissen JN, Johansen J, Allesœ RL *et al.* Improved metagenome binning and assembly using deep variational autoencoders. *Nat Biotechnol* 2021;**39**:555–60. https://doi.org/10.1038/s41587-020-00777-4
17. Liu C-C, Dong SS, Chen JB *et al.* MetaDecoder: a novel method for clustering metagenomic contigs. *Microbiome* 2022;**10**:46. https://doi.org/10.1186/s40168-022-01237-8
18. Heck M, Sakti S, Nakamura S. Supervised learning of acoustic models in a zero resource setting to improve DPGMM clustering. *INTERSPEECH*, San Francisco, USA, pp. 1310–4, 2016.
19. Mallawaarachchi V, Wickramarachchi A, Lin Y. GraphBin: refined binning of metagenomic contigs using assembly graphs. *Bioinformatics* 2020;**36**:3307–13. https://doi.org/10.1093/bioinformatics/btaa180
20. Du Y, Sun F. HiCBin: binning metagenomic contigs and recovering metagenome-assembled genomes using Hi-C contact maps. *Genome Biol* 2022;**23**:63.
21. Traag VA, Waltman L, Van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep* 2019;**9**:1–12.
22. Zhou Z, Wu Weimin, Ho Harrison *et al.* Dnabert-s: Learning species-aware dna embedding with genome foundation models. *arXiv preprint arXiv:2402.08777*, 2, 2024.
23. Rho M, Tang H, Ye Y. FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res* 2010;**38**:e191.
24. Chklovski A, Parks DH, Woodcroft BJ *et al.* CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat Methods* 2023;**20**:1203–12. https://doi.org/10.1038/s41592-023-01940-w
25. Bowers RM, Kyrpides NC, Stepanauskas R *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol* 2017;**35**:725–31. https://doi.org/10.1038/nbt.3893
26. Hundt S, Haug U, Brenner H. Comparative evaluation of immunochemical fecal occult blood tests for colorectal adenoma detection. *Ann Intern Med* 2009;**150**:162–9.
27. Leddin D, Enns R, Hilsden R *et al.* Colorectal cancer surveillance after index colonoscopy: guidance from the Canadian association of gastroenterology. *Can J Gastroenterol Hepatol* 2013;**27**:224–8.
28. Wolf AM, Fontham ETH, Church TR *et al.* Colorectal cancer screening for average-risk adults: 2018 guideline update from the american cancer society. *CA Cancer J Clin* 2018;**68**:250–81. https://doi.org/10.3322/caac.21457
29. Haug U, Hundt S, Brenner H. Quantitative immunochemical fecal occult blood testing for colorectal adenoma detection: evaluation in the target population of screening and comparison with qualitative tests. *Am J Gastroenterol.* 2010;**105**:682–90.
30. Wang H, Tian T, Zhang J. Tumor-associated macrophages (TAMs) in colorectal cancer (CRC): from mechanism to therapy and prognosis. *Int J Mol Sci* 2021;**22**:8470.
31. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* 2020;**36**;6:1925–1927. https://doi.org/10.1093/bioinformatics/btz848
32. Belgiu M, Drăguţ L. Random forest in remote sensing: a review of applications and future directions. *ISPRS J Photogramm Remote Sens* 2016;**114**:24–31.
33. Zhu Z *et al.* MicroPro: using metagenomic unmapped reads to provide insights into human microbiota and disease associations. *Genome Biol* 2019;**20**:1–13.
34. Wang Y, LêCao KA. Managing batch effects in microbiome data. *Brief Bioinform* 2020;**21**:1954–70.
35. Han W, Li L. Evaluating and minimizing batch effects in metabolomics. *Mass Spectrom Rev* 2022;**41**:421–42.
36. Oh M, Zhang L. DeepMicro: deep representation learning for disease prediction based on microbiome data. *Sci Rep* 2020;**10**:6026.

37. Liao H, Shang J, Sun Y. GDmicro: classifying host disease status with GCN and deep adaptation network based on the human gut microbiome data. *Bioinformatics* 2023;**39**:btad747.
38. Wu Y, Jiao N, Zhu R *et al.* Identification of microbial markers across populations in early detection of colorectal cancer. *Nat Commun* 2021;**12**:3063. https://doi.org/10.1038/s41467-021-23265-y
39. Sakamoto M, Iino T, Yuki M *et al.* Lawsonibacter asaccharolyticus gen. nov., sp. nov., a butyrate-producing bacterium isolated from human faeces. *Int J Syst Evol Microbiol* 2018;**68**:2074–81. https://doi.org/10.1099/ijsem.0.002800
40. Valdes AM, Walter J, Segal E *et al.* Role of the gut microbiota in nutrition and health. *BMJ* 2018;**361**:k2179. https://doi.org/10.1136/bmj.k2179
41. Hu J, Lin S, Zheng B *et al.* Short-chain fatty acids in control of energy metabolism. *Crit Rev Food Sci Nutr* 2018;**58**:1243–9. https://doi.org/10.1080/10408398.2016.1245650
42. Shuwen H, Yinhang W, Xingming Z *et al.* Using whole-genome sequencing (WGS) to plot colorectal cancer-related gut microbiota in a population with varied geography. *Gut Pathog* 2022;**14**:50. https://doi.org/10.1186/s13099-022-00524-x

Briefings in Bioinformatics, 2026, **27**, bbag241
https://doi.org/10.1093/bib/bbag241
Problem Solving Protocol