applied sciences

MDPI

Article

Integrating Biological Domain Knowledge with Machine Learning for Identifying Colorectal-Cancer-Associated Microbial Enzymes in Metagenomic Data

Burcu Bakir-Gungor [1,†], Nur Sebnem Ersoz [2,†] and Malik Yousef [3,4,*]

[1] Department of Computer Engineering, Faculty of Engineering, Abdullah Gul University, Kayseri 38080, Türkiye; burcu.gungor@agu.edu.tr
[2] Department of Bioengineering, Graduate School of Engineering and Science, Abdullah Gul University, Kayseri 38080, Türkiye; ersoz.nursebnem@gmail.com
[3] Department of Information Systems, Zefat Academic College, Zefat 1320611, Israel
[4] Galilee Digital Health Research Center (GDH), Zefat Academic College, Zefat 1320611, Israel
* Correspondence: malik.yousef@gmail.com
† These authors contributed equally to this work.

Abstract: Advances in metagenomics have revolutionized our ability to elucidate links between the microbiome and human diseases. Colorectal cancer (CRC), a leading cause of cancer-related mortality worldwide, has been associated with dysbiosis of the gut microbiome. This study aims to develop a method for identifying CRC-associated microbial enzymes by incorporating biological domain knowledge into the feature selection process. Conventional feature selection techniques often evaluate features individually and fail to leverage biological knowledge during metagenomic data analysis. To address this gap, we propose the enzyme commission (EC)-nomenclature-based Grouping-Scoring-Modeling (G-S-M) method, which integrates biological domain knowledge into feature grouping and selection. The proposed method was tested on a CRC-associated metagenomic dataset collected from eight different countries. Community-level relative abundance values of enzymes were considered as features and grouped based on their EC categories to provide biologically informed groupings. Our findings in randomized 10-fold cross-validation experiments imply that glycosidases, CoA-transferases, hydro-lyases, oligo-1,6-glucosidase, crotonobetainyl-CoA hydratase, and citrate CoA-transferase enzymes can be associated with CRC development as part of different molecular pathways. These enzymes are mostly synthesized by *Eschericia coli*, *Salmonella enterica*, *Klebsiella pneumoniae*, *Staphylococcus aureus*, *Streptococcus pneumoniae*, and *Clostridioides dificile*. Comparative evaluation experiments showed that the proposed model consistently outperforms traditional feature selection methods paired with various classifiers.

Keywords: metagenomic analysis of colorectal cancer; machine learning; feature grouping; functional profiling of metagenomes; community-level enzyme commission (EC) abundances

Academic Editor: Nikolaos Kourkoumelis

Received: 31 December 2024
Revised: 28 February 2025
Accepted: 3 March 2025
Published: 8 March 2025

Citation: Bakir-Gungor, B.; Ersoz, N.S.; Yousef, M. Integrating Biological Domain Knowledge with Machine Learning for Identifying Colorectal-Cancer-Associated Microbial Enzymes in Metagenomic Data. *Appl. Sci.* **2025**, *15*, 2940. https://doi.org/10.3390/app15062940

1. Introduction

Colorectal cancer (CRC) stands out as one of the most prevalent cancer types, and it is a significant contributor to cancer-related mortality worldwide. CRC is notable as the second most prevalent cancer among women and the third most prevalent among men [1]. The onset of CRC is closely linked to several attributes, including age, history of chronic illnesses, lifestyle choices, and genetic factors. CRC development is characterized by various mutations affecting oncogenes and tumor suppressor genes involved in DNA

repair mechanisms. Colorectal carcinomas are classified based on the origin of mutations as sporadic, inherited, or familial. At the genetic level, the progression of CRC is influenced by chromosomal alterations, genetic mutations, and translocations. Additionally, non-coding RNAs (ncRNAs), such as microRNAs (miRNAs) and long non-coding RNAs (lncRNAs), play significant roles at different stages of carcinogenesis [2,3]. Recent studies have demonstrated that gut microbiome composition plays a crucial role in CRC development and progression by influencing intestinal inflammation, tumorigenesis, and the anti-cancer immune response [4]. Gut microbiota have the potential to find biomarkers for predicting immunotherapy outcomes and enhancing treatment efficacy through modulation [5].

The human body, particularly the gut, harbors over 100 trillion microbes, collectively forming the human microbiome [6,7]. This diverse microbial community includes various microorganisms, such as archaea, viruses, fungi, and bacteria [8,9]. These microbial communities contribute to maintaining bodily homeostasis and are predominantly acquired at birth [10]. The interaction between the host and the microbiome is dynamic and influenced by genetic and environmental factors, such as age, geographical location, alcohol or drug consumption, and diet [11–13]. Metagenomics is a valuable tool used to understand the interactions between the host and intestinal microbiome. These studies analyze genomic data inferred from clinical and environmental samples. However, the mechanisms underlying host–pathogen interactions have remained poorly understood. Metagenomics can reveal alterations in microbial functions and detect dysbiosis in the human intestinal microbiome associated with particular diseases. Additionally, functional metagenomics is able to find functional dysbiosis, new microbial pathways, functional genes, and antibiotic resistance genes in the intestinal microbiome by considering microbiota vs. host interactions and their co-evolution. Furthermore, metatranscriptomics, metaproteomics, and metabolomics are used to understand the complexity of the human gut microbiome [14]. The rapid advancement of computational methods has accelerated metagenomic data analysis studies. The increasing number of studies performing functional metagenomic analysis has highlighted several disease and microbiome relationships in the human gut [12,15,16].

Metagenomics is categorized into amplicon (targeted gene) and shotgun (nontargeted gene) metagenomics, based on the types of data obtained [17]. Amplicon metagenomics involves sequences from specific marker genes like 16S, 18S, 26S rRNA, and intergenic transcribed spacer (ITS), obtained through amplification [18,19]. On the other hand, shotgun metagenomics encompasses all DNA sequences present in the samples, providing a comprehensive view. The classification of metagenomics also distinguishes between functional and sequencing metagenomics [20]. Functional metagenomics focuses on identifying novel functional genes and bioactive compounds, while sequencing metagenomics delves into microbial community diversity. The development of tools for functional profiling has significantly advanced the field of metagenomics. One such example is MetaPhlAn, which harnesses a vast database comprising over one million prokaryotic reference genomes to enable detailed metagenomic taxonomic profiling, as highlighted by Blanco-Míguez et al. [21]. Additionally, tools like HUMAnN, StrainPhlAn, PanPhlAn, and PhyloPhlAn have been proven to be highly effective in various aspects of microbial community analysis, including strain-level resolution, phylogenetic insights, taxonomic classification, and functional profiling, as demonstrated in [22]. These innovative tools, exemplified by MetaPhlAn for comprehensive taxonomic profiling and by HUMAnN for precise functional analysis, significantly contribute to unraveling the intricate structure and functionality of microbial communities through sophisticated sequence-based and functional metagenomic approaches. The utilization of these cutting-edge tools not only enhances our understanding of the complex interactions within microbial ecosystems, but also paves the way for

more in-depth research and insights into the role of microbiomes in various biological processes, including the development of diseases like colorectal cancer.

Biomarker detection is crucial for early disease diagnosis, prognosis, and predicting treatment responses, facilitating medical interventions. Identifying actionable biomarkers can guide the selection of effective targeted therapies and avoid less effective treatments [23]. For biomarker detection, relying solely on statistical analysis can be insufficient [24]. To this end, the Grouping-Scoring-Modeling (G-S-M) approach offers to evaluate groups of features rather than assessing individual features. The G-S-M technique was recently proposed to integrate biological knowledge into machine learning models [25]. This method generates feature sets using either of the following: (i) pre-existing biological knowledge stored in databases such as mirTarBase [26], DisGeNET [27], or KEGG pathways [28]; or (ii) a data-driven approach relying on statistical measures, such as Pearson's correlation. The current study seeks to integrate external biological knowledge into the colorectal-cancer-associated enzyme selection process in metagenomic data analysis, aiming to deliver results that are both statistically robust and biologically meaningful. Along this line, enzyme commission (EC) nomenclature was developed by the International Union of Biochemistry and Molecular Biology (IUPAC-IUBMB) in 1999. This system assigns each enzyme an EC number, consisting of the prefix "EC" followed by four digits separated by periods. For example, an EC number like EC 1.1.1.1 identifies a specific enzyme. The first three digits of the EC number represent the enzyme's class, subclass, and sub-subclass, respectivley, providing insight into its general function. The seven main EC classes, including oxidoreductases (EC: 1), transferases (EC: 2), hydrolases (EC: 3), lyases (EC: 4), isomerases (EC: 5), ligases (EC: 6), and translocases (EC:7), categorize enzymes based on the type of chemical reactions that they facilitate (as depicted in Figure 1). To delve deeper into enzyme specificity, subclass and sub-subclass numbers further refine the classification, specifying the exact nature of the reaction and the substrates involved. For example, within the oxidoreductases class (EC: 1), there are 23 subclasses (e.g., EC 1.1–1.22 and EC 1.97), each representing a distinct group of enzymes with specific functions. By utilizing these detailed classification systems, researchers can gain a comprehensive understanding of the enzyme activities and their contributions to biochemical pathways, paving the way for advancements in fields such as metagenomics and disease–microbiome relationships.

The research is centered on the identification of sets of enzymes linked by enzyme commission (EC) terms that are relevant to a specific disease. The main objective in this study is to help CRC diagnosis via classifying enzymes into their functional groups and then utilizing the Grouping-Scoring-Modeling (G-S-M) methodology to pinpoint highly correlated EC terms associated with the disease. In this regard, the G-S-M algorithm is introduced as an innovative approach that enhances accuracy in classification by incorporating EC terms as external biological data during the analysis of disease-associated metagenomic datasets, including enzyme abundance values. The analysis involves the application of the 10-fold cross-validation technique, where datasets are divided repeatedly into training and testing sets. During the training phase, the most informative EC term is identified, and enzymes linked to it are grouped for model training. Comparative assessments with established methodologies highlight the benefits of this novel approach. What distinguishes this method is its ability to leverage EC terms for classification while identifying the most relevant sets of EC terms associated with the disease, in contrast to traditional selection approaches that emphasize individual enzymes. Several experiments have been conducted in this study to assess the efficacy of different feature selection techniques in conjunction with various machine learning algorithms. The objective is to compare and assess the influence of feature selection methods and machine learning algorithms on model performance through a consistent experiment conducted over 10 iterations.

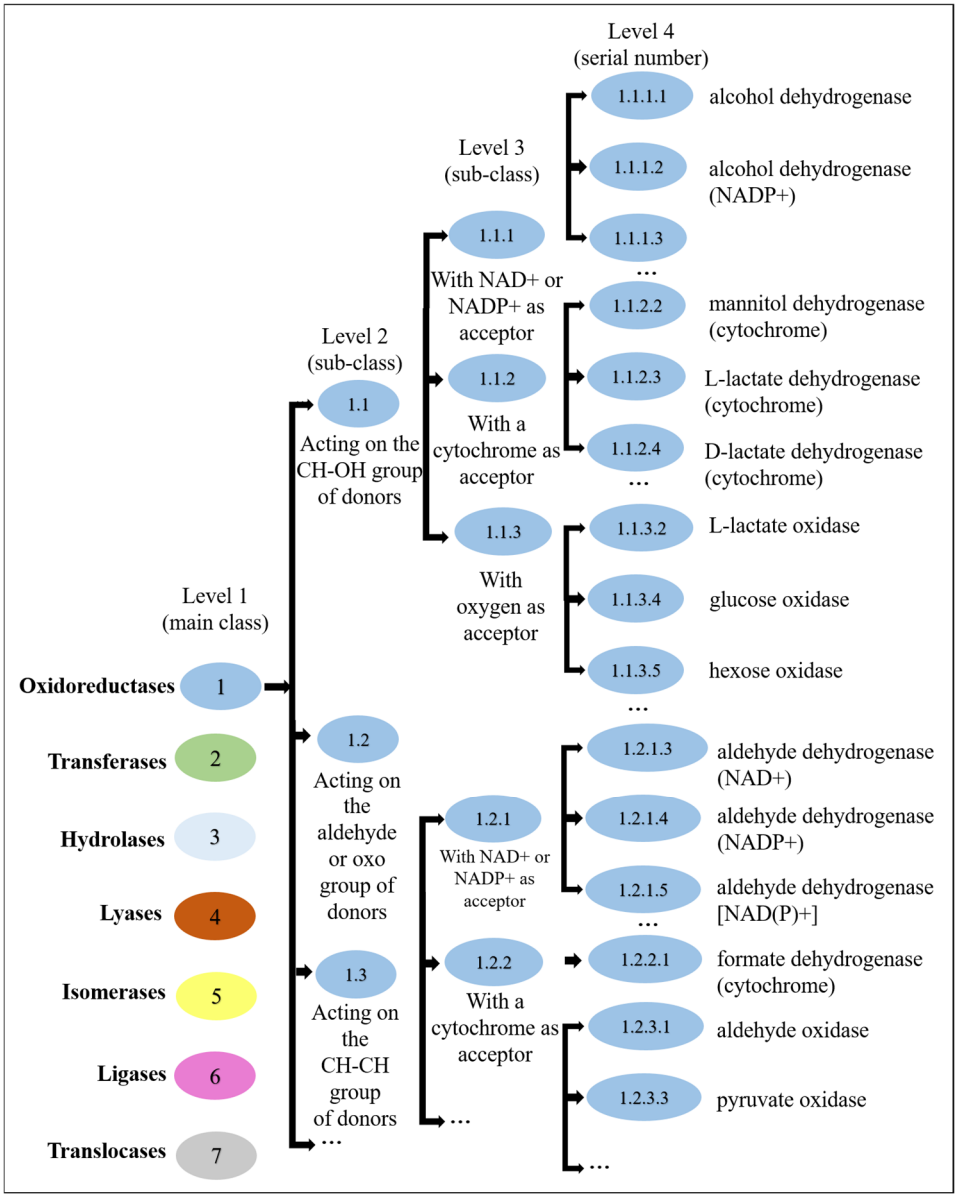

Figure 1. Enzyme commission (EC) nomenclature involves seven main enzyme groups with many subclasses, each related to a specific enzyme activity.

2. Materials and Methods

2.1. Datasets

2.1.1. CRC-Associated Metagenomic Dataset

The relative enzyme abundance dataset provided by Beghini et al. includes 1262 samples (600 CRC patients and 662 controls) collected from 8 different countries (as shown in Table 1) [22]. The functional profiles of these samples are provided in the same study (Beghini et al., 2021 [22]), where they employ HUMAnN3 to estimate community-level enzyme commission (EC) abundance values. The functional profile of the CRC-associated metagenomic relative enzyme abundance dataset is represented as a matrix. In this matrix, the EC numbers of the enzymes are shown in the columns, and the rows represent the samples. This matrix contains a special column called "label", which indicates the class annotation for each row. Here, the class labels are either positive, indicating the CRC, or negative, indicating the control.

Table 1. Description of the CRC-associated metagenomic datasets collected from eight different countries.

Country	# of Controls	# of CRC Patient	Total
Austria (AUT)	61	46	107
China (CHN)	53	75	128
Germany (DEU)	65	60	125
France (FRA)	61	53	114
Indian (IND)	30	30	60
Italy (ITA)	49	57	106
Japan (JP)/(JPN)	291	227	518
United State of America (USA)	52	52	104
Total	662	600	1262

2.1.2. Enzyme Commission Dataset

The classification of enzymes plays a crucial role in understanding their functions and how they contribute to biological processes. Enzymes are pivotal in catalyzing specific chemical reactions, and various classification systems have been established to categorize them based on their distinct roles. Structural Classification of Proteins (SCOP) [29] and the Class, Architecture, Topology, Homologous superfamily (CATH) [30] classify enzymes structurally based on shared protein domains. On the other hand, the enzyme nomenclature system, referred to as the enzyme commission (EC), is a vital resource utilized for the classification and prediction of enzymatic reactions. This structured annotation system plays a crucial role in facilitating the development and validation of biological hypotheses [31]. The enzyme commission (EC) system classifies enzymes based on their catalytic activities, organizing them into seven main categories, each of which is further divided into subclasses, as illustrated in Figure 1. In this study, the enzyme groups were defined at the third-level subclass of the EC hierarchy. A total of 243 enzyme groups were identified and utilized as biological domain knowledge, which was integrated into the proposed G-S-M model.

2.2. Our Proposed Method: EC-Nomenclature-Based G-S-M Approach

The main idea of the G-S-M technique is to evaluate a group or groups of features rather than assessing individual features. Biological knowledge is used as a function applied to the feature space to create sets of groups, where each of the groups is a set of features (i.e., enzymes, genes). Figure 2 illustrates the main workflow of the proposed EC-nomenclature-based G-S-M approach. It involves three main components, i.e., the G Component, S Component, and M component. While the G component generates sub-datasets for each EC term group and the S Component scores the EC terms, the M Component trains the classifier using different classifiers, such as Random Forest (RF) and XGBoost, to build the model.

The general G-S-M technique was developed by Yousef et al. [32] and was embedded in different computational tools, including the following: maTE [32], which uses microRNA target gene information for grouping genes; miRModuleNet [33], which analyzes mRNA and miRNA expression datasets concurrently to detect feature sets; CogNet [34] and PriPath [35], which use KEGG pathway information for gene grouping; miRdisNET [36], which uses miRNA target gene information to assign genes into groups; and GeNetOntology [37], which uses gene ontology (GO) terms to find disease-related gene ontology groups. AMP-G-S-M is another G-S-M-based approach used for the prediction of antimicrobial peptides [38]. The main idea and most of the relevant tools are reviewed in [39].

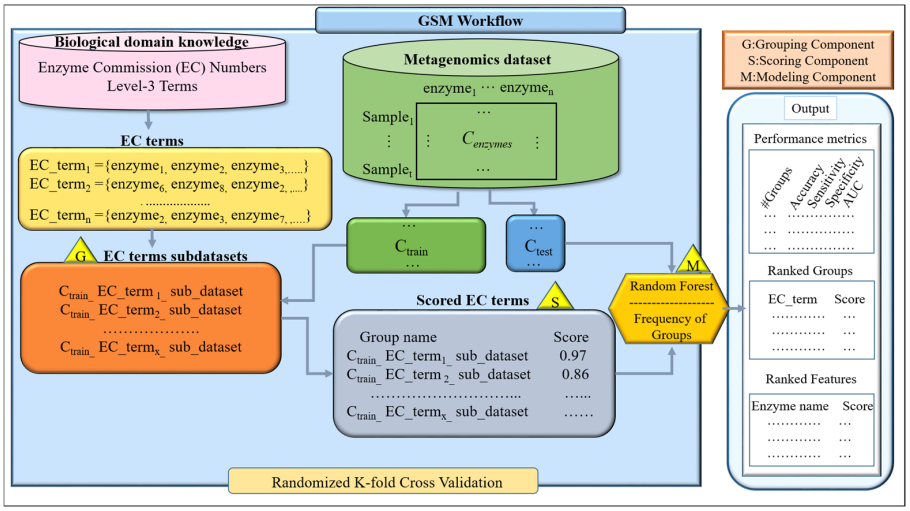

Figure 2. The proposed EC-nomenclature-based G-S-M workflow for analyzing enzyme abundance values obtained from disease-associated metagenomics relative enzyme abundance datasets.

The main purpose of the proposed EC-nomenclature-based G-S-M method is to identify the most significant enzyme groups and enzyme terms (as scored in the S Component) to be used for training the classifier (realized in the M component) (Figure 2). In order to evaluate a set of features, for each EC term, a sub-dataset is created by only including the abundance values of the enzymes that are associated with that particular EC term (represented with the orange box). The enzyme abundance values of samples are represented by the C matrix (represented with the green box), which is divided into Ctrain and Ctest matrices. While Ctrain has been utilized for scoring the EC terms and training the classifier, Ctest has been used to test and report the final performance. The M component (represented with the yellow box) trains a classifier using the enzyme abundance values that are annotated with the top-scoring EC terms. Finally, the G-S-M model provides performance metrics, ranked groups, and ranked features as outputs. The G-S-M KNIME workflow is available publicly at: https://github.com/malikyousef/The-G-S-M-Grouping-Scoring-Modeling-Approach (Accessed on 7 March 2023)

2.3. Implementation of the EC-Nomenclature-Based G-S-M Model

The first component (G component) of the proposed G-S-M plays a crucial role in creating sub-datasets for each enzyme commission (EC) term. This component focuses on the annotation of several enzymes with specific EC terms and then extracting a sub-dataset for each EC term from the relative enzyme abundance dataset. In essence, these sub-datasets are tailored to include abundance values solely for the enzymes associated with a particular EC term, along with the class labels of the samples, whether positive or negative. It is important to highlight that, while each sub-dataset contains an equal number of samples, the number of features varies based on the number of enzymes annotated with the specific EC term. This process is further elucidated in Figure 2, which visually represents a flowchart for generating sub-datasets based on the enzymes linked to a specific EC term. Each sub-dataset is uniquely named as "ECterm(i)_sub_dataset", where 'i' starts from 1 and increments up to k, the total number of EC terms considered.

The G component operates with two key tables, as follows: the EC terms table, where each EC term is connected to a set of enzymes, and the relative enzyme abundance dataset. Figures 2 and 3 depict the creation of four distinct sub-datasets, each corresponding to a particular EC term based on this process. Following the G component, component S steps in, denoted as "S" in the latter part of Figure 2, to conduct the scoring phase using the generated sub-datasets. Component S leverages machine learning algorithms, specifically

RF in this study. It implements an internal Monte Carlo Cross-Validation (MCCV) process, which is repeated r times, as indicated in Figure 3 (represented with a yellow triangle on the top right). This scoring mechanism assesses the classification performance of each EC term, indicating the accuracy of classification solely based on the enzyme abundance values associated with that specific EC term. The classification accuracy is averaged across r iterations of MCCV to derive a mean accuracy value, which serves as the final importance score for the respective EC term. Subsequently, S ranks all EC terms based on their scores, with the top-scoring EC terms moving to the next phase for model training.

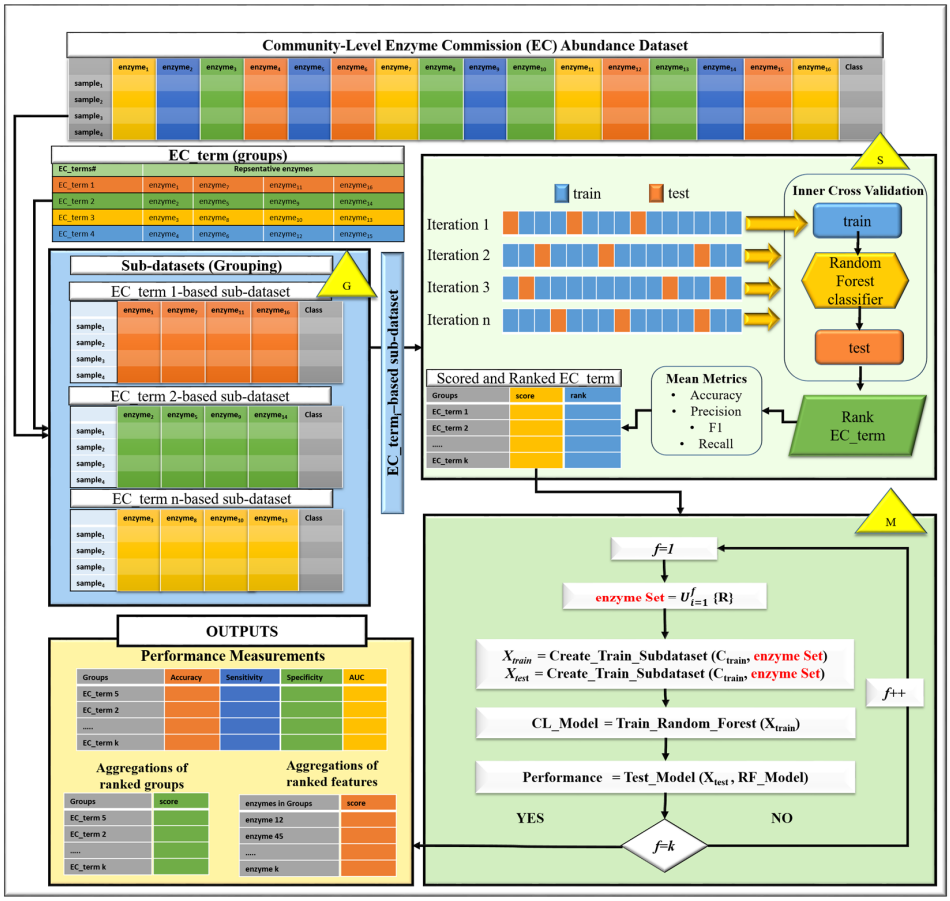

Figure 3. Detailed description of the grouping, scoring, and modeling components in the proposed EC-nomenclature-based G-S-M approach.

Furthermore, the M component trains the classifier and constructs the model, following a cumulative approach, as illustrated in Figure 3 (represented with a yellow triangle on the bottom right). This component utilizes the enzyme abundance values linked to the top-scoring EC terms for model training. The iterative process starts with building an RF model using enzymes annotated with the highest scoring EC term and continues by progressively incorporating enzymes associated with subsequent high-scoring EC terms. This cumulative strategy enables the calculation of the model's overall performance, facilitating the comparison of performance results across different feature sets, ranging from the highest scoring EC term to the top 10 highest scoring EC terms. By evaluating the model's performance over various feature sets, one can identify the optimal combination of enzymes annotated with the top-scoring EC terms. The M component provides insights into the average performance metrics of classification using enzymes tied to the top 10 highest scoring EC terms, averaged over a 10-fold MCCV, offering a comprehensive evaluation of the model's efficacy. The output of the M component includes the performance

measurements (accuracy, sensitivity, specificity, and AUC metrics), ranked EC groups, and ranked features (enzymes).

2.4. Comparative Evaluation with Traditional Feature Selection Methods and Classifiers

Feature selection is a fundamental step in machine learning and data analysis, designed to enhance model performance, improve interpretability, and reduce computational complexity by identifying the most informative features while eliminating redundant or irrelevant ones. Feature selection methods are generally categorized into the following three main types: filter methods, wrapper methods, and embedded methods, each with distinct advantages and applications [40]. Filter methods assess feature relevance using statistical measures, selecting features independently of any specific model. These methods evaluate the relevance of features based on statistical or information-theoretic measures without involving any machine learning model. Statistics-based methods include Pearson correlation, Spearman correlation, Chi-square test, and Anova (analysis of variance) [41]. Information-theory-based methods are Information Gain (IG) [42], Minimum Redundancy Maximum Relevance (MRMR) [43], Conditional Mutual Information Maximization (CMIM), and Fast Correlation-Based Filter (FCBF) [44,45]. Also, ranking-based methods like Select K Best (SKB) and variance threshold are among the widely used filter methods. The advantages of filter methods are their being computationally efficient and scalable to high-dimensional datasets. Also, independent of the learning algorithm, the selected features are generalizable across different models. These methods reduce overfitting by removing irrelevant or redundant features early in the pipeline. However, these methods ignore feature dependencies and interactions and may not always lead to the best feature subset for a specific model. Their selection is based on individual feature relevance rather than on assessing their collective impact.

Wrapper methods evaluate subsets of features by iteratively testing their impact on model performance by using a specific learning algorithm [46]. Recursive Feature Elimination (RFE) is one of the popular approaches. While this method often yields high accuracy, it can be computationally intensive, especially for large datasets. Recursive Cluster Elimination (RCE) is another wrapper method [47]. RCE integrates clustering with recursive elimination to enhance classification performance and iteratively removes less informative feature clusters while training an SVM model. To improve predictive accuracy, RCE selects only the most relevant features, and discriminative features are retained [48]. Wrapper methods can capture feature dependencies and interactions, leading to potentially higher accuracy, and can be optimized for a specific learning algorithm, improving predictive performance. However, they are computationally expensive, especially for large datasets, and prone to overfitting, as they tailor feature selection to a specific model [46]. It is also worth noting that they require extensive cross-validation to avoid biased results.

Embedded methods integrate feature selection directly into the model training process, allowing the algorithm to determine the most important features as it learns [49]. Examples include decision-tree-based methods, such as RF feature importance and XGBoost feature selection [50]. These methods strike a balance between efficiency and predictive performance. Embedded methods are more efficient than wrapper methods, as feature selection is performed during model training; moreover, they balance feature selection and predictive performance and are less prone to overfitting compared to wrapper methods. However, they are model-dependent. In other words, the selected features may not generalize well across different models. They can also be computationally demanding for complex models. Limited flexibility in applying selection criteria compared to filter methods creates a disadvantage for the model [40].

Several metagenomic data analysis studies have been performed using machine learning algorithms [51,52]. Marcos-Zambrano et al. conducted a review on the use of machine learning in human microbiome studies, summarizing its applications in feature selection, biomarker discovery, disease prediction, and treatment strategies [53]. The review identified Random Forest, Support Vector Machines (SVM), and logistic regression as the most commonly employed classification algorithms for microbiome analysis [54]. The use of different feature selection methods on metagenomic data analysis is reviewed in [55]. IG, CMIM, XGBoost, MRMR, FCBF, and SKB feature selection methods have been previously utilized for type 2 diabetes [56], inflammatory bowel disease [45], and CRC-associated metagenomic datasets to identify disease-associated taxonomic biomarkers [57].

In this study, we have performed several experiments to define the effectiveness of different feature selection techniques in combination with various classification algorithms. To conduct a comparative performance analysis, traditional feature selection (TFS) methods, such as XGBoost (eXtreme Gradient Boosting), IG (Information Gain), SKB (Select K Best), and FCBF (Fast Correlation-Based Filter), have been applied to CRC-associated metagenomic relative enzyme abundance. RF is an ensemble learning algorithm that aggregates multiple decision trees (DT) to enhance predictive accuracy. By constructing decision trees on random subsets of training data and features, RF can handle high-dimensional data effectively. Adaptive Boosting (Adaboost), another ensemble method, iteratively combines weak classifiers to create a robust classifier, focusing on misclassified samples to boost overall prediction accuracy. LogitBoost, tailored for binary classification tasks, utilizes logistic regression as the base learner to iteratively adjust instance weights for model refinement. In this study, while evaluating TFS, we have employed different classifiers, such as RF, LogitBoost, DT, SVM_opt (Support Vector Machines with Optimization), Support Vector Machines (SVM), Stack_Logitboost_Kmeans, Stack_SVM_Kmeans, and XGBoost. We have aimed to evaluate the impact of feature selection approaches and classifiers on model performance by performing a consistent experiment repeated in 10 iterations of Monte Carlo Cross-Validation.

2.5. Performance Evaluation Metrics

We have evaluated a set of statistical measures, such as specificity, sensitivity, and accuracy, for each model. The following formulations were used to calculate the statistics:

Accuracy = (True Positive + True Negative)/# of all examples;

Specificity = True Negative/(True Negative + False Positive);

Sensitivity (Recall) = True Positive/(True Positive + False Negative).

Moreover, the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC) has been employed to estimate the likelihood of a classifier scoring a randomly chosen positive instance higher than a randomly selected negative instance. All performance metrics reported in this study represent the average values derived from 10-fold Monte Carlo Cross-Validation.

2.6. Molecular/Metabolic Pathways

Information on molecular/metabolic pathways and enzymatic activities are obtained from KEGG (https://www.genome.jp/kegg (Accessed on 29 October 2023)) and BRENDA (https://www.brenda-enzymes.org (Accessed on 29 October 2023)). GO annotations of enzymes are obtained from QuickGO (https://www.ebi.ac.uk/QuickGO/annotations (Accessed on 29 October 2023)).

3. Results

Metagenomics is a vital field in understanding the microbiome's role in disease development. The functional profiling of metagenomic samples has been significantly advanced to uncover the links between diseases and the microbiome through the burgeoning techniques of biological profiling via shotgun metagenomic sequencing. Each gene within metagenomic data undergoes annotation and quantification based on its functional role, which is intricately linked to established biological pathways and functions. The increasing utilization of shotgun metagenomics technology in biological profiling has propelled studies in metagenomics research, establishing a vital connection between the microbiome and disease development and progression. The primary aim of this research was to detect the enzymes produced by microbial communities that are linked to diseases. To contribute to CRC diagnosis, we constructed a proficient classification model utilizing the functional profile of CRC-associated metagenomic data, primarily grounded in community-level EC abundance.

3.1. Performance Evaluation of the Proposed EC-Nomenclature-Based G-S-M Approach

EC-nomenclature-based G-S-M was tested on a CRC-associated metagenomic relative enzyme abundance dataset including 1262 samples, where 600 belong to CRC patients and 662 belong to controls. For each dataset, for different numbers of groups, accuracy, sensitivity, specificity, and AUC values have been calculated as the mean of the values obtained in 10 iterations of the cross-validation procedure. For each group, G-S-M reports the number of enzymes included in the group. Also, the average enzyme numbers over 10 iterations are reported. Table 2 shows the performance metrics of the EC-nomenclature-based G-S-M method applied to the CRC-associated metagenomic dataset, including relative enzyme abundance values. For example, there are 59.5 enzymes on average, as shown in the # (number) of enzmes column of the top row (highest scoring group) in Table 2. There are 90.7 enzymes on average, as shown in the # of features in the column of the 2nd row (the top two groups are combined). In other words, the model that is generated via using the enzyme values of 90.7 can successfully predict CRC with a 0.769 AUC score. Upon scrutinizing different performance metrics of EC-nomenclature-based G-S-M, it becomes evident that our proposed method consistently excels, as delineated in Table 2.

Table 2. Averages over 10-fold MCCV and standard deviations are presented for different performance metrics. These were obtained cumulatively for the top 10 ranked enzyme groups, when applied to the CRC-associated metagenomic dataset, including relative abundance values of the enzymes.

# of Groups	# of Enzymes	AUC	Accuracy	Specificity	Sensitivity
1	59.5 ± 26.517	0.728 ± 0.031	0.673 ± 0.028	0.749 ± 0.08	0.588 ± 0.058
2	90.7 ± 38.257	0.769 ± 0.041	0.695 ± 0.037	0.767 ± 0.050	0.615 ± 0.049
3	147.9 ± 45.261	0.763 ± 0.031	0.704 ± 0.032	0.769 ± 0.054	0.632 ± 0.033
4	200.9 ± 52.821	0.765 ± 0.032	0.704 ± 0.035	0.773 ± 0.039	0.627 ± 0.052
5	244.9 ± 64.824	0.770 ± 0.029	0.706 ± 0.029	0.773 ± 0.034	0.630 ± 0.057
6	281.9 ± 68.709	0.763 ± 0.033	0.694 ± 0.026	0.766 ± 0.024	0.613 ± 0.046
7	330.4 ± 79.269	0.761 ± 0.029	0.683 ± 0.036	0.760 ± 0.034	0.598 ± 0.060
8	372.1 ± 82.131	0.766 ± 0.034	0.692 ± 0.037	0.755 ± 0.038	0.622 ± 0.050
9	391.1 ± 84.069	0.767 ± 0.026	0.706 ± 0.032	0.772 ± 0.030	0.633 ± 0.048
10	457.0 ± 79.538	0.763 ± 0.035	0.688 ± 0.045	0.739 ± 0.047	0.632 ± 0.062

The robust rank aggregation approach was applied (i.e., ranked enzyme groups and enzymes) to aggregate them into a final list, as shown in Table 3. The ranked enzyme groups and the enzymes that are annotated with these enzyme groups are shown in the final list.

All of those final lists are visualized in the output panel of the G-S-M approach, as shown in Figure 4. The EC-nomenclature-based G-S-M approach identified glycosidases (3.2.1), CoA-transferases (2.8.3), and hydro-lyases (4.2.1) as the top three most important enzyme groups when the CRC-associated metagenomic dataset was analyzed (as shown in Figure 4A). Oligo 1,6 glucosidase, crotonobetainyl-CoA hydratase, and citrate CoA-transferase are found to be the top three important enzymes when the EC-nomenclature-based G-S-M approach was applied to the CRC-associated metagenomic dataset (as shown in Figure 4B).

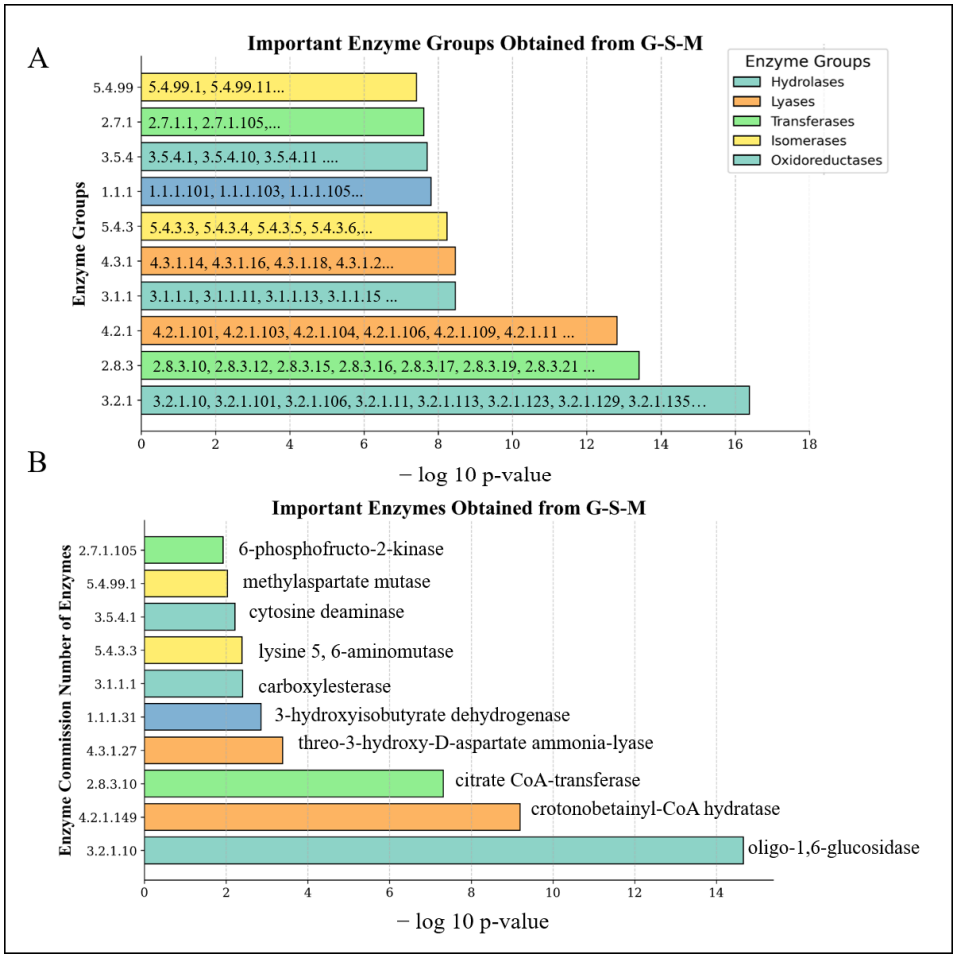

Figure 4. (**A**) Top 10 important enzyme groups and (**B**) top scored enzyme in each enzyme group identified by the EC-nomenclature-based G-S-M method applied to the CRC-associated metagenomic datasets, including relative abundance values of the enzymes. The −log 10 *p*-values indicate the significance values assigned by the robust rank aggregation method. Each color represents the enzyme commission (EC) activity.

We have also applied the EC-nomenclature-based G-S-M method to population-specific CRC-associated metagenomic datasets, including the relative abundance values of the enzymes. The AUC values and # of enzymes of the top two scored groups of each country are represented in Figure 5. The model generated for the Indian population yielded high accuracy values using low # of enzymes. As plotted in Figure 5, the EC-nomenclature-based G-S-M approach showed similar AUC metrics when applied to the following population-

specific datasets: France, Australia, Japan, China, and Italy. The EC-nomenclature-based G-S-M model resulted in the lowest AUC value for the USA population dataset.

Table 3. Identified enzyme groups and their associated enzymes when the robust rank aggregation method in G-S-M is applied to the CRC-associated metagenomic dataset, including relative abundance values of the enzymes.

Enzyme Group	Enzyme Group Name	*p*-Value	# of Enzymes	Enzymes (EC)
3.2.1	Glycosidases	4.13×10^{-17}	74	3.2.1.1, 3.2.1.10, 3.2.1.101...
2.8.3	CoA-transferase	3.86×10^{-14}	12	2.8.3.10, 2.8.3.12, 2.8.3.15...
4.2.1	Hydro-lyases	1.52×10^{-13}	73	4.2.1.101, 4.2.1.103, 4.2.1.104...
3.1.1	Carboxylic-ester Hydrolases	3.43×10^{-9}	28	3.1.1.1, 3.1.1.11, 3.1.1.13...
4.3.1	Ammonia-lyases	3.43×10^{-9}	16	4.3.1.14, 4.3.1.16, 4.3.1.18...

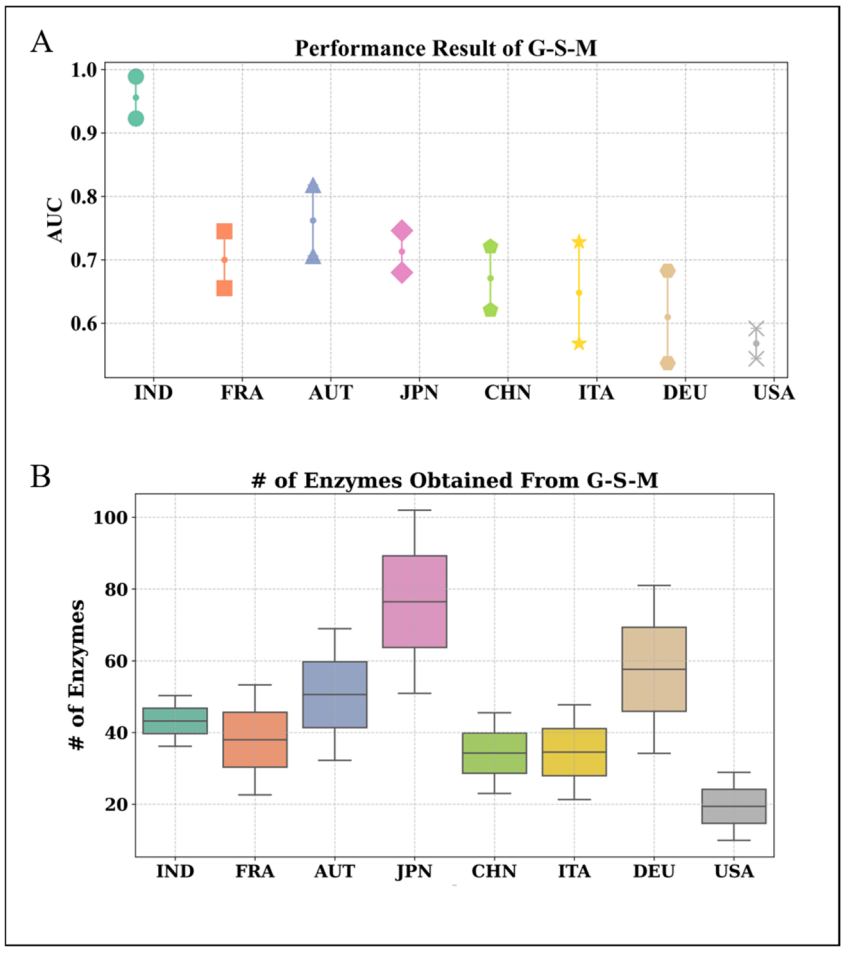

Figure 5. Performance metrics of the EC-nomenclature-based G-S-M method when applied to population-specific CRC-associated metagenomic dataset, including relative abundance values of the enzymes. (**A**) AUC values and (**B**) # of enzymes (features) selected for population-specific datasets.

3.2. Comparative Performance Evaluation of G-S-M with Traditional Feature Selection Methods

We have comparatively evaluated the performance of EC-nomenclature-based G-S-M models with the performances of traditional feature selection methods, including XGB,

SKB, IG, CMIM, MRMR, and FCBF, coupled with different classifiers, such as Adaboost, DT, LogitBoost, RF, SVM_opt, Stack_Logitboost_Kmeans, Stack_SVM_Kmeans, and XGBoost (as shown in Figure 6). As shown in Table 2, there are 90 enzymes on average within the top two scoring groups selected by the EC-nomenclature-based G-S-M model when applied to the CRC-associated metagenomic dataset. To conduct a comparative analysis, we have focused on the performance metrics that were obtained using the 90 features selected by different methods. In our previous studies, we have shown that, for different GSM architectures, RF classifiers outperform other classifiers [25]. Also, in the literature, it is shown that RF classifiers are widely used for human microbiome studies, including selecting relevant features, categorizing patients, identifying disease-associated biomarkers, and revealing host–microbial patterns [53]. Therefore, we have incorporated RF as the main classifier in the EC-nomenclature-based G-S-M approach. Still, to investigate the performance of the EC-nomenclature-based G-S-M using different classifiers, we have run G-S-M using RF, XGBoost, and DT classifiers for 10 iterations. Figure 6 summarizes the AUC metrics obtained using the EC-nomenclature-based G-S-M when coupled with RF, XGBoost, and DT classifiers and the AUC metrics of TFS methods coupled with different classifiers and tested on the CRC-associated metagenomic dataset. Different classifiers generated different performance metrics for different TFS methods, however, Figure 6 consistently shows the dominance of the RF and XGBoost classifiers. The EC-nomenclature-based G-S-M with RF generated a 0.769 AUC value, and the TFS methods with the RF classifier have AUC values as of 0.763 (XGB), 0.762 (SKB), 0.725 (IG), 0.495 (FCBF), 0.519 (MRMR), and 0.584 (CMIM), using 90 features. As can be seen from Figure 6, when compared with the RF classifier, the other classifiers reported either similar or lower AUC values. Our results demonstrate that the proposed method outperforms the conventional approaches, delivering more reliable insights into the molecular and biological associations between the human gut microbiome and CRC.

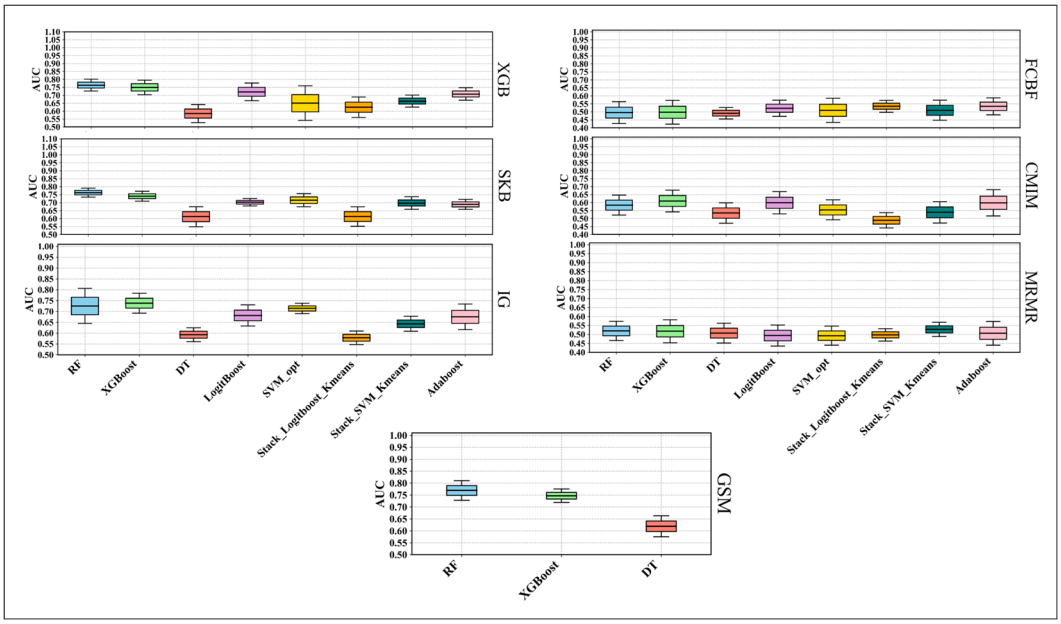

Figure 6. AUC values of traditional feature selection methods, including XGB, SKB, IG, MRMR, CMIM, and FCBF, when coupled with different classifiers, including Adaboost, DT, LogitBoost, RF, SVM_opt, Stack_Logitboost_Kmenas, Stack_SVM_Kmeans, and XGBoost, compared with the AUC metrics of the EC-nomenclature-based G-S-M approach when coupled with RF, XGBoost, and DT classifiers and tested on the CRC-associated metagenomic dataset.

Additionally, using the same CRC-associated enzyme abundance value dataset, we have tested the performance of the Recursive Cluster Elimination (RCE) algorithm with an SVM classifier as one of the wrapper methods. The SVM-RCE method resulted in 0.65 (±0.04) accuracy, 0.59 (±0.06) sensitivity, 0.70 (±0.05) specificity, and a 0.70 (±0.05) AUC value.

As shown in Figure 6, the XGB and SKB feature selection methods have higher AUC values than the other tested methods. The top 10 most important enzymes selected by the XGB and SKB feature selection methods when coupled with the RF classifier are shown in Figure 7. One can infer from Figure 7 that glycine dehydrogenase (EC: 1.4.4.2) and glycine C-acetyltransferase activity (EC: 2.3.1.29) have been commonly identified by both XGB and SKB. In addition to that, the methylaspartate mutase activity (EC: 5.4.99.1) enzyme has been commonly selected by the EC-nomenclature-based G-S-M approach, XGB, and SKB (as shown in Figures 4 and 7). The citrate CoA-transferase (EC: 2.8.3.10) enzyme has been commonly detected by the EC-nomenclature-based G-S-M approach and SKB.

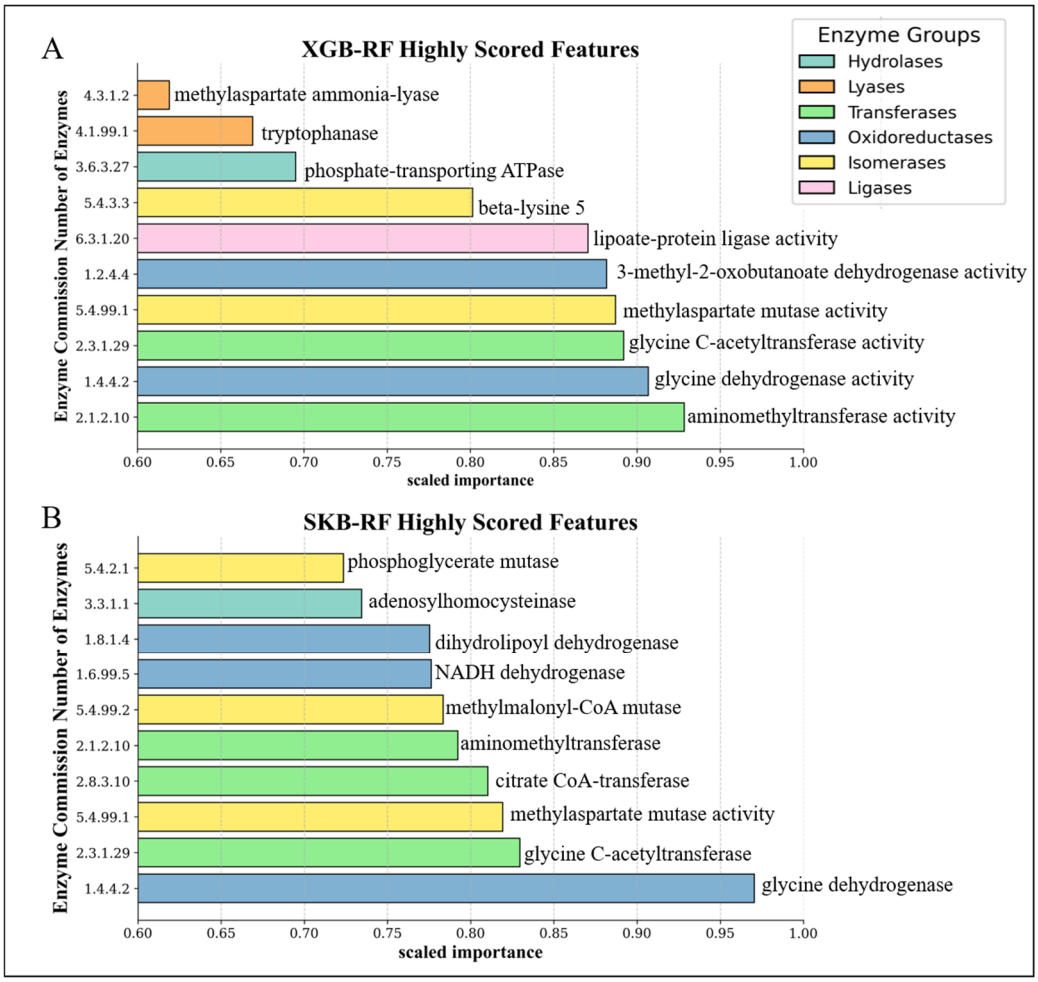

Figure 7. Top 10 most important enzymes detected by (**A**) XGB and (**B**) SKB feature selection methods. The colors represent different enzyme functions, as defined by the enzyme commission.

3.3. Metabolic Pathways That Are Associated with the Top Scoring Enzyme Groups

Metagenomic studies play an important role in identifying biomarkers for many diseases associated with gut health, such as CRC, inflammatory bowel disease (IBD), and irritable bowel disease [58–60]. Therefore, it is important to highlight the mechanisms behind the disease formation and progression. Enzyme commission (EC) terms provide information about enzymes and their chemical activities. The top 10 significant enzymes and the top 10 significant enzyme groups detected by the EC-nomenclature-based

G-S-M approach when applied to CRC-associated metagenomic dataset are shown in Figure 4. Among these, the top three enzyme groups are as follows: The glycosidases group (EC: 3.2.1), which is represented with blue color, belongs to hydrolases (EC: 3) as its main class and glycosylases (EC: 3.2) as its subclass. The hydro-lyases group (EC: 4.2.1), which is represented with orange color, belongs to lyases (EC: 4) as its main class and carbon–oxygen lyases (EC: 4.2) as its subclass. The CoA-transferases group (EC: 2.8.3), which is represented with green color, belongs to transferases (EC: 2) as its main class and transferring sulfur-containing groups (EC: 2.8) as its subclass (as illustrated in Figure 8). Within these top three identified enzyme groups, Figure 8 also lists the top 10 most important enzymes identified by the EC-nomenclature-based G-S-M approach when applied to the CRC-associated metagenomic dataset.

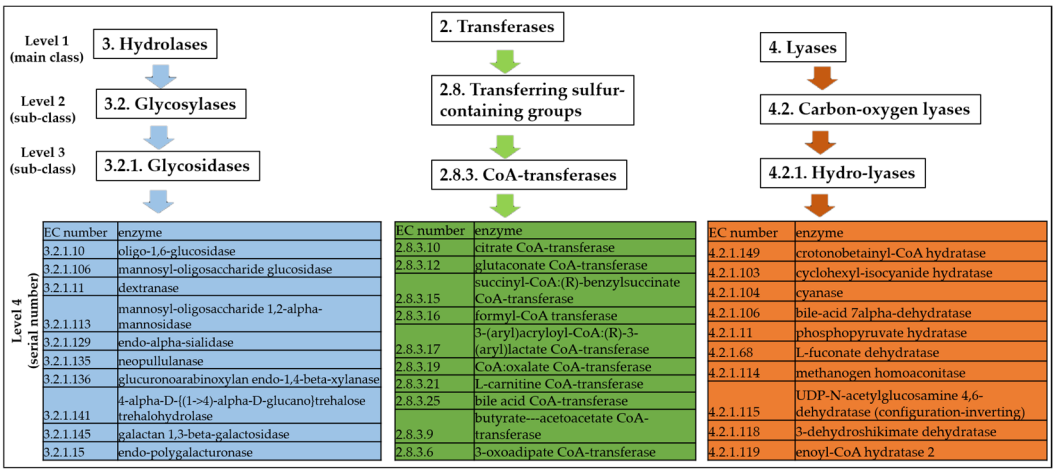

Figure 8. Top 3 scoring enzyme groups and top 10 scoring enzymes included in these groups, which are identified by the EC-nomenclature-based G-S-M model when applied to the CRC-associated metagenomic dataset. Each color represents the related enzyme commission (EC) activity.

Gene ontology (GO) provides a meticulously curated vocabulary for delineating gene products [61]. In Figure 9, we show the related GO MF (molecular function) terms for the top three scoring enzymes that belong to the top three scoring enzyme groups, which are identified by the EC-nomenclature-based G-S-M model applied to the CRC-associated metagenomic dataset. Our proposed method identified glycosidases (3.2.1), CoA-transferases (2.8.3), and hydro-lyases (4.2.1) as the top three most important enzyme groups when the CRC-associated metagenomic dataset was analyzed (as shown in Figure 4A). These enzymes are associated with hydrolase activity, transferase activity, and lyase activity, respectively, as shown in Figure 9. Glycosidases (EC: 3.2.1), hydro-lyases (4.2.1), and CoA-transferases (2.8.3) also commonly belong to the catalytic activity term (GO:0003824), which is located right under the root molecular function term (GO:0003674) (Figure 9). Glycosidases enzymes (EC: 3.2.1) branch from hydrolase activity term (GO:0016787) and hydrolase activity, acting on glucosyl terms (GO:0016798)(as illustrated in Figure 9). Hydro-lyases (4.2.1) branched through lyase activity term (GO:0016829), carbon–oxygen lyase activity term (GO:0016835), and hydro-lyase activity term (GO:0016836). CoA-transferases enzymes (2.8.3) take place under the transferase activity term(GO:0016740), transferring term (GO:0016782), and CoA transferase activity term (GO:0008410) (as shown in Figure 9).

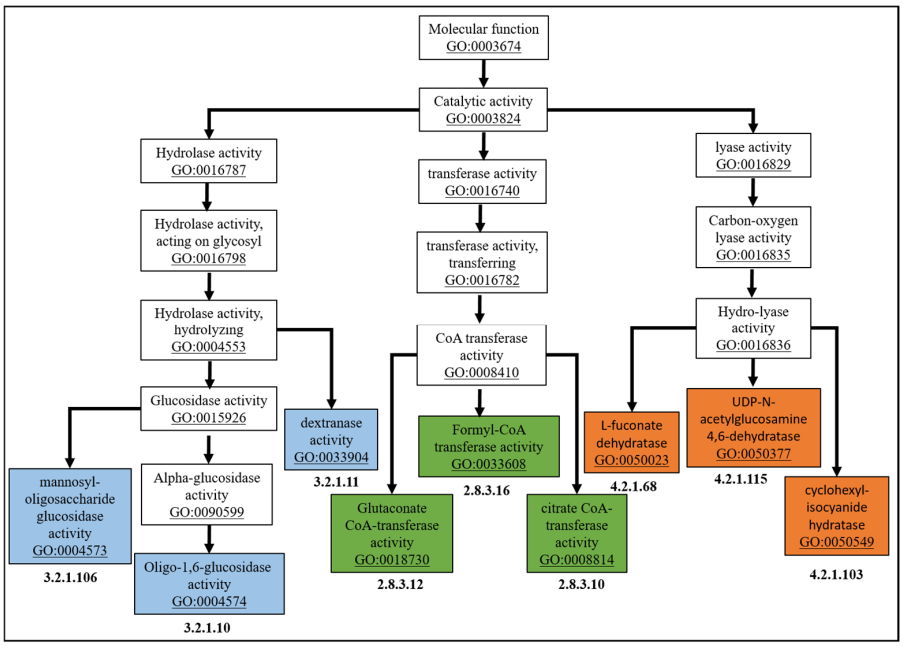

Figure 9. GO MF terms for the top 3 scoring enzymes that belong to the top 3 scoring enzyme groups, which were identified by the EC-nomenclature-based G-S-M model applied to the CRC-associated metagenomic dataset. GO hierarchy was obtained from the Quick-GO annotations provided by NCBI.

Furthermore, the Kyoto Encyclopedia of Genes and Genomes (KEGG) furnishes comprehensive annotations of biochemical pathways, thereby enhancing our comprehension of gene functions and gene interactions within biological pathways [28]. In Figure 10, we present the related KEGG pathways, where the top 100 scoring enzymes that were identified by the proposed model have chemical activities. It noteworthy that these top 100 scoring enzymes belong to the top three scoring enzyme groups (glycosidases (3.2.1), CoA-transferases (2.8.3), and hydro-lyases (4.2.1)) identified by the EC-nomenclature-based G-S-M approach when applied to the CRC-associated metagenomic dataset. The top 100 scoring enzymes play roles in different pathways, such as starch and sucrose metabolism, biosynthesis of secondary metabolites, amino sugar and nucleotide sugar metabolism, galactose metabolism, N-Glycan biosynthesis, fructose and mannose metabolism, sphingolipid metabolism, and butanoate metabolism.

The enzymes in the glycosidases (EC: 3.2.1) group take part in different metabolisms. For example, oligo-1,6-glucosidase (EC: 3.2.1.10) plays a role in galactose metabolism and starch and sucrose metabolism (as illustrated in Figure 11A,C). On the other hand, mannosyl-oligosaccharide glucosidase (EC: 3.2.1.106) and mannosyl-oligosaccharide 1,2-alpha-mannosidase (EC: 3.2.1.113) play a role in N-Glycan biosynthesis (as depicted in Figure 11D). Dextranase (EC: 3.2.1.11) plays a role in starch and sucrose metabolism (as shown in Figure 11A). In addition to that, many of the enzymes in the glycosidases (EC: 3.2.1) group have activity in sphingolipid metabolism (showcased in Figure 11B). In starch and sucrose metabolism (Figure 11A), the energy source of starch is converted to maltodextrin by isoamylase (EC: 3.2.1.68), alpha-amylase (EC: 3.2.1.1), beta-amylase (EC: 3.2.1.2), and glucan 1,4-alpha-maltohydrolase (EC: 3.2.1.133) enzymes. Later, maltose is converted to D-Glucose with the help of the maltase (EC: 3.2.1.20) enzyme. D-Glucose is synthesized to be used as an energy source. In sphingolipid metabolism (Figure 11B), beta-galactosidase (EC: 3.2.1.23), hexosaminidase (EC: 3.2.1.52), sialidase (EC: 3.2.1.18), and alpha-galactosidase (EC: 3.2.1.22) enzymes contribute to increase the amount of lactosylceramide in the cell. The beta-galactosidase (EC: 3.2.1.23) enzyme converts Lactosylceramide to Glucosyl-ceramide, which is converted to ceramide by the glucosylceramidase

enzyme (EC: 3.2.1.45). The amount of sphingosine is increased through the conversion of ceramide, glucosylsphingosine, and psychosine by ceramidase (EC: 3.5.1.23), glucosylceramidase (EC: 3.2.1.45), and galactosylceramidase (EC: 3.2.1.46), respectively (Figure 11B). Alpha-galactosidase (EC: 3.2.1.22), alpha-glucosidase (EC: 3.2.1.20), oligo-1,6-glucosidase (EC: 3.2.1.10), and saccharase (EC: 3.2.1.26) contribute to the production of D-Galactose, D-Glucose, and D-Fructose in galactose metabolism (Figure 11C). *N*-glycan biosynthesis begins in the endoplasmic reticulum (ER), where a precursor glycan is assembled and attached to nascent proteins. This glycan is then processed by glycosidases, including mannosyl-oligosaccharide glucosidase (EC: 3.2.1.106), mannosyl-oligosaccharide alpha-1, 3-glucosidase (EC: 3.2.1.207), endoplasmic reticulum Man9GlcNAc2 1,2-alpha-mannosidase (EC: 3.2.1.209), and mannosyl-oligosaccharide 1,2-alpha-mannosidase (EC: 3.2.1.113). Further maturation and modification of N-glycans occur in the Golgi apparatus.(Figure 11D).

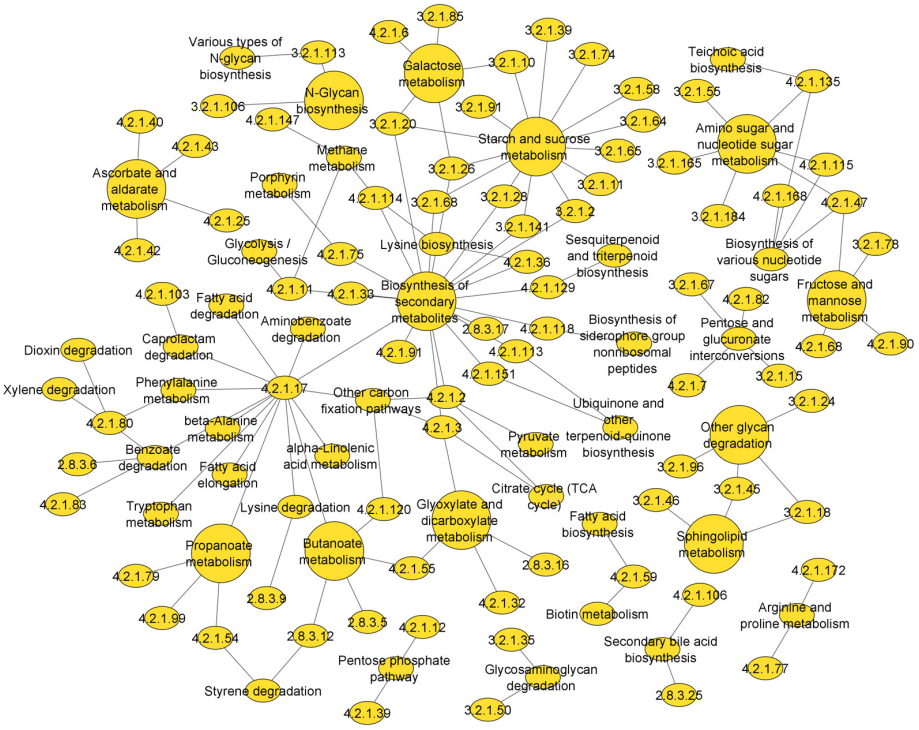

Figure 10. Top 100 scoring enzymes obtained by the EC-nomenclature-based G-S-M method applied to the CRC-associated metagenomic dataset and their related KEGG pathways.

The enzymes in hydro-lyases (EC: 4.2.1) and CoA-transferases (EC: 2.8.3) cordially play roles in many different metabolic pathways, such as styrene degradation, butanoate metabolism, carnitine metabolism, and the citric acid cycle (as shown in Figure 12). Nitrile hydratase (EC: 4.2.1.84) and lactoyl-CoA dehydratase (EC: 4.2.1.54) enzymes, which are part of the hydro-lyases group (EC: 4.2.1), have a role in the styrene degradation pathway. The metabolites from styrene degradation transform into pyruvate metabolism (as shown in Figure 12A). Some metabolites from pyruvate metabolism are transferred into butanoate metabolism to produce different metabolites. Acetoacetyl-CoA processing is also carried out by the conversion of Acetoacetate by acetate CoA-transferase (EC: 2.8.3.8) and 3-oxoacid CoA-transferase (EC: 2.8.3.5) enzymes. Crotonoyl-CoA formation is also performed by enoyl-CoA hydratase (EC: 4.2.1.17) and 4-hydroxybutanoyl-CoA dehydratase (EC: 4.2.1.120) (as illustrated in Figure 12B). In carnitine metabolism, L-carnitinyl-CoA is synthesized by L-carnitine CoA-transferase (EC: 2.8.3.21) and crotonobetainyl-CoA hydratase (EC: 4.2.1.149) from gamma-butyrobetainyl-CoA and Crotonobetanily-CoA, respectively. In the citric acid

cycle, succinyl-CoA:acetate CoA-transferase (EC: 2.8.3.18) and 3-oxoacid CoA-transferase (EC: 2.8.3.5) catalize the reaction involving succinyl-CoA.

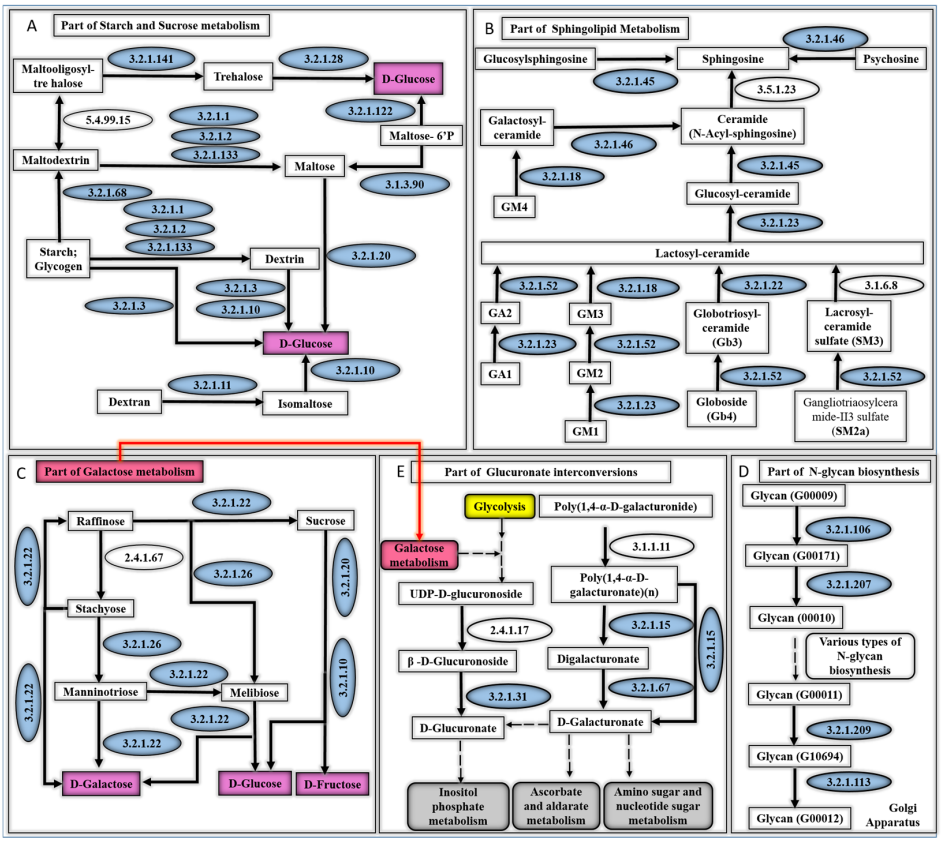

Figure 11. The enzymes in the glycosidases (EC: 3.2.1) group and their metabolic pathways, i.e., (**A**) starch and sucrose metabolism, (**B**) sphingolipid metabolism, (**C**) galactose metabolism, (**D**) *N*-glycan biosynthesis, and (**E**) glucuronate interconversions. All pathway information is excerpted from the KEGG database.

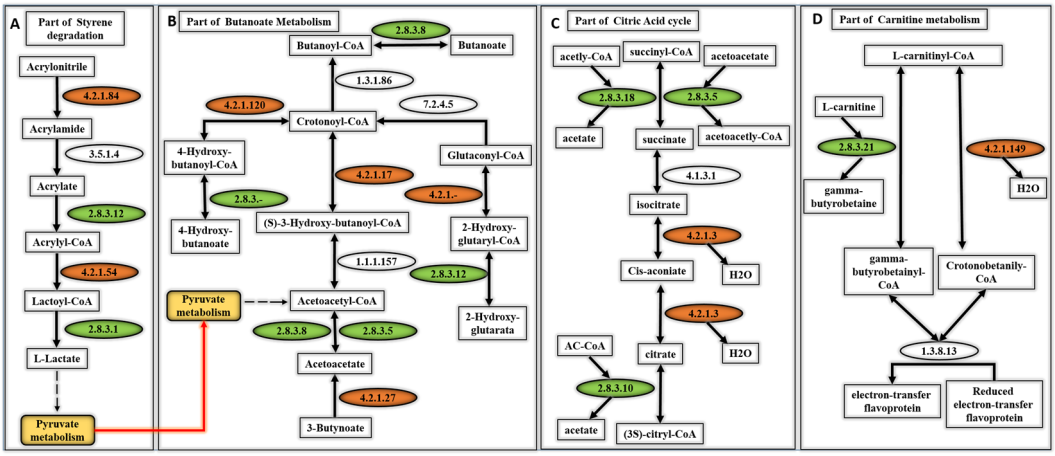

Figure 12. Metabolic pathways of top 3 scoring enzyme groups found using the EC-nomenclature-based G-S-M approach applied to the CRC-associated metagenomic dataset, including relative abundance values of the enzymes. EC: 4.2.1 and EC: 2.8.3 groups of enzymes perform activities in (**A**) styrene degradation (KEGG database), in (**B**) butanoate metabolism (KEGG database), in (**C**) the citric acid cycle (BRENDA database), and in (**D**) carnitine metabolism (BRENDA database). Each color represents the related enzyme commission (EC) activity.

3.4. Top Scored Enzyme-Associated Species Obtained from CRC Dataset

The human gut microbiome is considered to have a role in disease formation and prognosis [62]. Therefore, it is important to understand the microbiome–disease associations, including microorganisms, enzymes, and pathways. It is worth noting that the enzymes that have been identified in metagenomic studies may be synthesized by different microorganisms. To enlighten these relations, for the top 100 scoring enzymes that have been identified by the EC-nomenclature-based G-S-M in the CRC-associated metagenomic data, the organisms that synthesize these enzymes have been obtained from Uniprot. We eliminated the species if they were not included in the CRC-associated metagenomic data at the taxonomic level. In total, 268 species have been identified for 100 enzymes, and their relations are visualized in Figure 13. As shown in Figure 13, most of the top 100 identified enzymes are synthesized by *Eschericia coli*, *Salmonella enterica*, *Klebsiella pneumoniae*, *Staphylococcus aureus*, *Streptococcus pneumoniae*, and *Clostridioides dificile*.

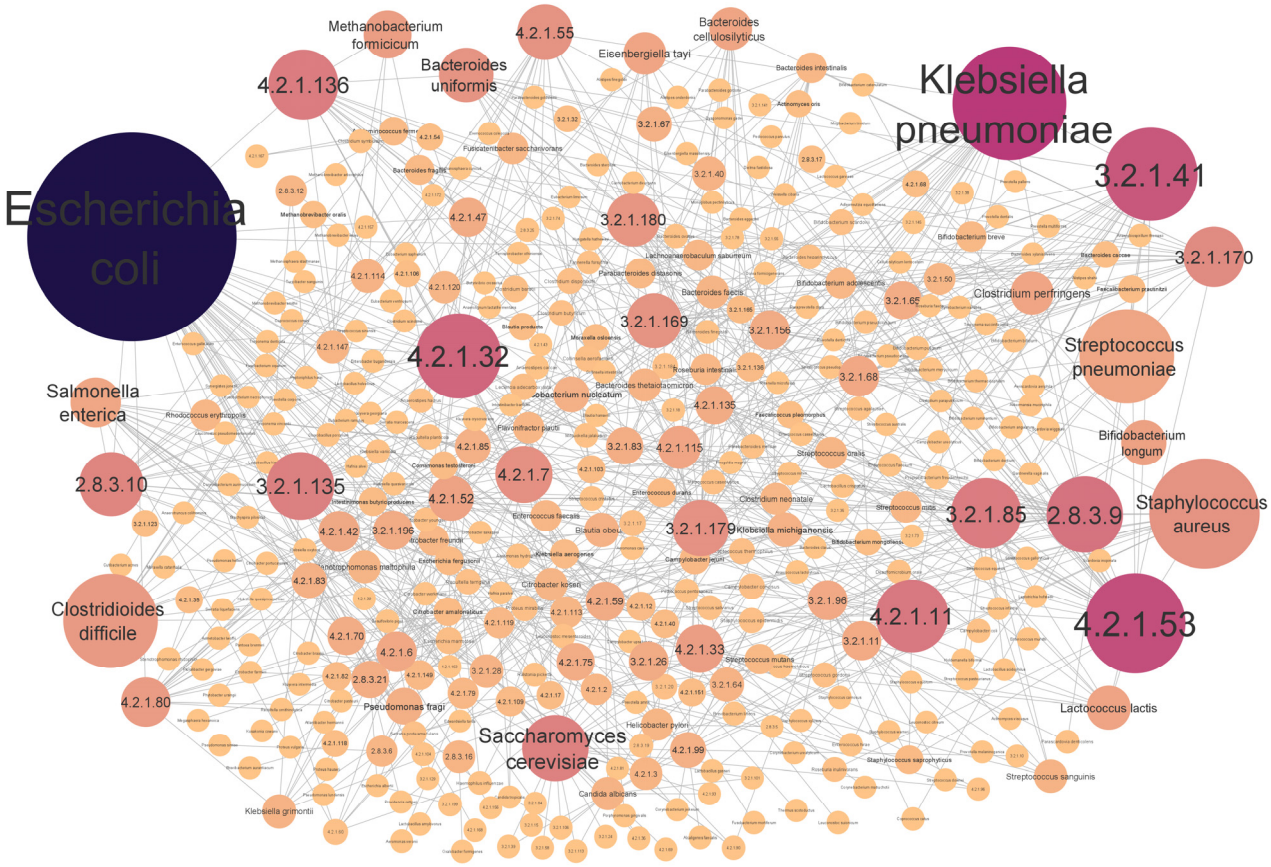

Figure 13. Network of top 100 enzymes that have been identified by the EC-nomenclature-based G-S-M on the CRC-associated metagenomic data and associated organisms. A total of 268 species that synthesize the top scoring 100 enzymes are presented. Node size is associated with betweenness centrality.

Glycine dehydrogenase (EC: 1.4.4.2) and glycine C-acetyltransferase activity (EC: 2.3.1.29) have been commonly identified by both XGB and SKB. In addition to that, the methylaspartate mutase activity (EC: 5.4.99.1) enzyme has been commonly selected by the EC-nomenclature-based G-S-M approach, XGB, and SKB (as shown in Figures 4 and 7).

Similarly, for the top 10 scoring enzymes that were identified by each of XGB and SKB feature selection methods (glycine dehydrogenase (EC: 1.4.4.2), glycine C-acetyltransferase activity (EC: 2.3.1.29), citrate CoA-transferase (EC: 2.8.3.10), methylaspartate mutase activity (EC: 5.4.99.1), aminomethytransferase (EC: 2.1.2.10), glycine C-acetyltransferase activity (EC:2.3.1.29), and the related species were obtained from Uniprot. We eliminated species if they were not included in CRC-associated metagenomic data at the taxonomic

level. In total, 85 species have been detected as being associated with the top 16 enzymes (4 enzymes have been found common in XGB and SKB), as reported by the XGB and SKB FS methods. As shown in Figure 14, *Eschericia coli*, *Salmonella enterica*, *Klebsiella pneumoniae*, and *Citrobacter koseri* have been found to be important. As one can notice in Figure 14, *Eschericia coli* is the most significant species by having the highest betweenness centrality.

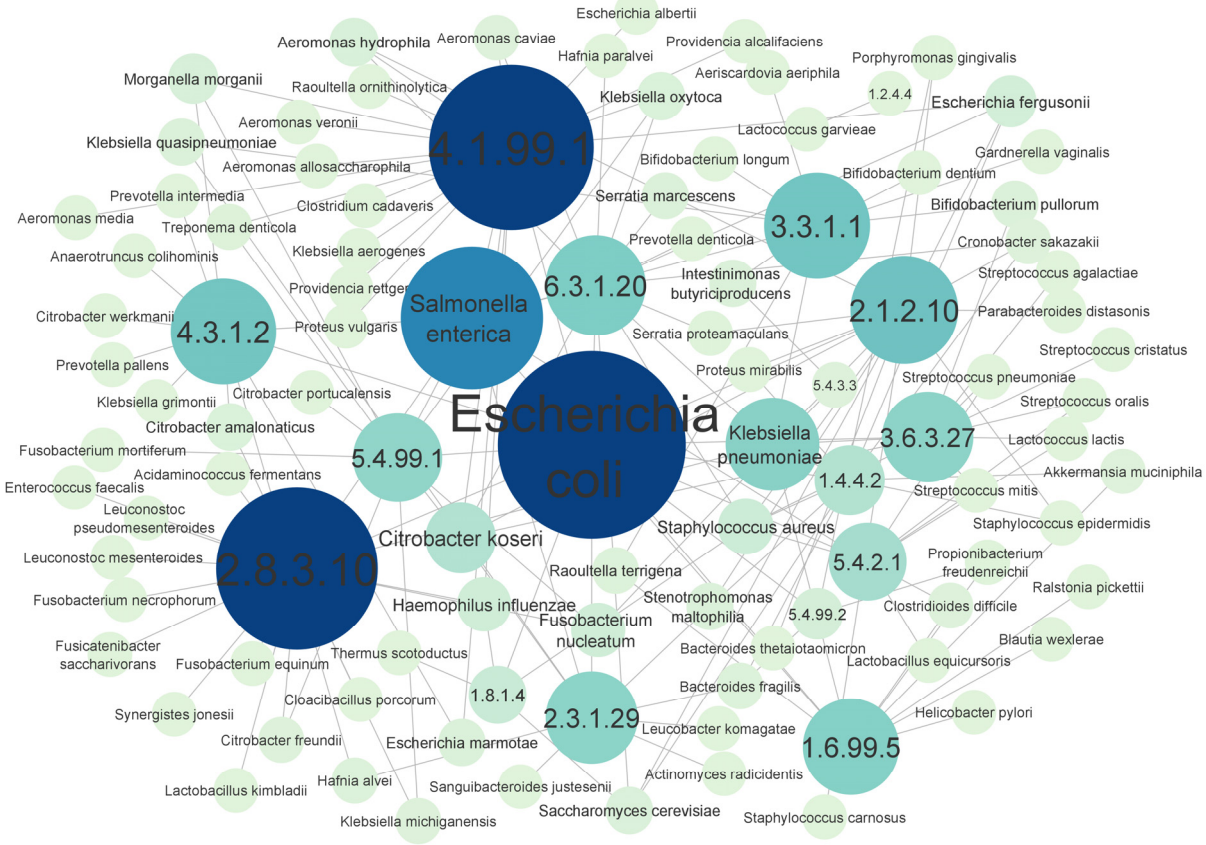

Figure 14. Network of top 16 scoring enzymes that were identified either by the XGB or SKB feature selection methods as part of their top 10 scoring lists and their associated organisms. A total of 85 species that synthesize the top scoring enzymes are presented. Node size is associated with betweenness centrality.

4. Discussion

Metagenomics has extensive potential in unveiling the mechanisms and correlations between the human intestinal microbiome and diseases. Recent advances in functional profiling of metagenomic samples, encompassing genes and biochemical pathways, have revolutionized our ability to elucidate the links between the functional capabilities of microbial communities and human diseases [63]. The functional characterization of metagenomic samples enables the estimation of relative abundance values for genes encoding enzymes, which is critical for uncovering the molecular underpinnings of diseases such as CRC, a leading cause of cancer-related mortality worldwide. The gut microbiota plays a pivotal role in this context, with dysbiosis potentially triggering colonic carcinogenesis through chronic inflammation, involving specific bacterial strains in this multifaceted process. Various factors, such as antibiotic usage and specific dietary habits, have been recognized as key drivers in the development of dysbiosis [64]. Despite the precise mechanisms underlying how dysbiosis leads to colonic carcinogenesis not being fully elucidated, chronic inflammation is widely acknowledged as a primary instigator [65,66]. Different risk factors associated with life styles, such as obesity, smoking, and alcohol consumption, have been found to contribute to CRC formation and accelerate the worsening of the disease [2].

Although dysbiosis of the gut microbiome has been strongly associated with CRC, leveraging biological insights in metagenomic data analysis is essential for understanding the disease mechanisms.

To address the limitations of the conventional feature selection techniques that fail to utilize the biological insights embedded in multi-omics data, in this study, a novel method called "EC-nomenclature-based Grouping-Scoring-Modeling (G-S-M)" was developed. This approach integrates biological domain knowledge into feature grouping and selection. Specifically, in this study, community-level EC abundances were used as features and were grouped based on their third-level EC categories to provide biologically informed groupings. The EC-nomenclature-based G-S-M method was evaluated through randomized 10-fold cross-validation and compared against traditional feature selection methods such as IG, SKB, XGBoost, CMIM, MRMR, and FCBF. These were paired with classifiers like Adaboost, decision tree, LogitBoost, RF, XGBoost, an optimized Support Vector Machine, and ensemble models like Stack_LogitBoost_KMeans and Stack_SVM_KMeans. Our experiments demonstrated that the EC-nomenclature-based G-S-M approach outperforms the conventional methods, offering more reliable insights into the molecular and biological associations between the microbiome and CRC.

4.1. Computational Performance Evaluation of the EC-Nomenclature-Based G-S-M Model

The EC-nomenclature-based G-S-M approach demonstrates its efficacy in predicting disease-causing enzymes by modeling and analyzing enzyme abundance data utilizing EC numbers. Within the G-S-M framework, a machine learning algorithm was employed to identify the most significant EC groups. The EC-nomenclature-based G-S-M method was evaluated using 1262 CRC-associated metagenomic samples, including the relative abundance values of enzymes. In Table 2, we present the performance metrics of the EC-nomenclature-based G-S-M model averaged over 10-fold MCCV iterations for the aggregated top 10 scoring EC groups identified for the CRC-associated metagenomic dataset, including a functional profile of the samples. Table 2 also demonstrates the effect of cumulatively adding the top scoring EC groups in terms of model performance. For example, the EC-nomenclature-based G-S-M model has an AUC value of 0.729 when on average 59.5 enzymes are used, as shown in the first row of Table 2 (using only the enzymes found in the top scoring EC group). On the other hand, the AUC value of the EC-nomenclature-based G-S-M model is reported as 0.769 when 90.7 enzymes from the top two scoring EC groups are used (as shown in the second row of Table 2). Additionally, the AUC value of the EC-nomenclature-based G-S-M model becomes 0.763 when 457 enzymes from the top-10 scoring EC groups are cumulatively used (as shown in the last row of Table 2). Instead of checking the functional profile values of 457 enzymes, one could check the functional profile values for only 90 enzymes to predict with high AUC values. The model that is generated via only using the enzyme abundance value of 90 can successfully predict CRC-associated enzymes with an AUC score of 0.769, which is quite satisfying. As shown in Table 2, the inclusion of additional EC groups did not show a statistically significant improvement in performance metrics. Consequently, constructing a model with a reduced number of enzymes becomes feasible, thereby facilitating the interpretation of the resultant model.

In order to evaluate population-based differences, we have applied the EC-nomenclature-based G-S-M method to population-specific CRC-associated metagenomic datasets, including the elative abundance values of the enzymes. Figure 5 presents the number of selected enzymes and AUC values recorded in this experiment. Moreover, when the EC-nomenclature-based G-S-M method is applied, we assess the correlation among the identified EC groups between different population datasets using the Jaccard index (as

shown in Figure 15). The Jaccard index refers to the number of enzyme groups that are commonly identified in two different populations divided by the total number of enzyme groups detected by either one of these populations. The maximum Jaccard index value between the identified enzyme groups from different populations is calculated as 0.25 between the French and Indian datasets.

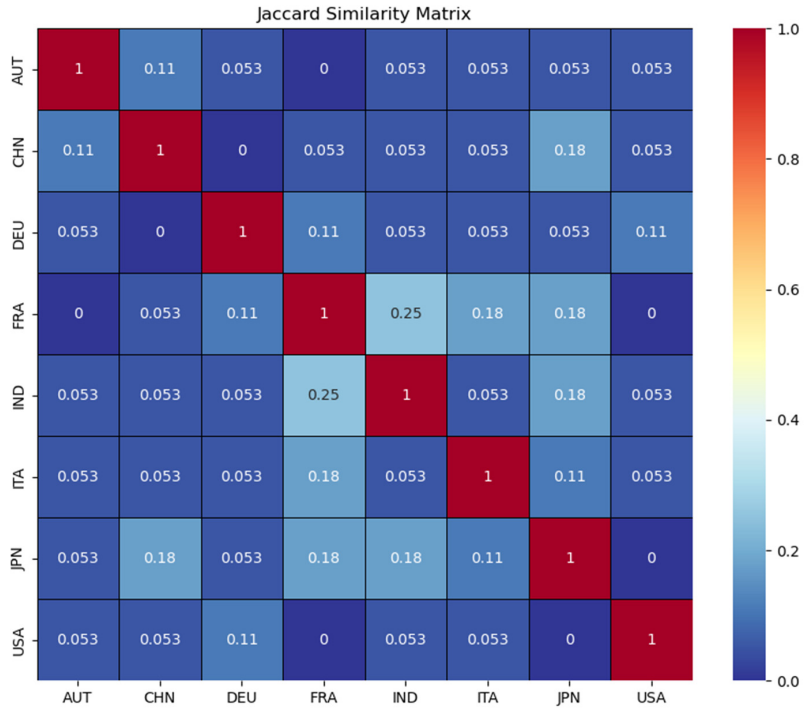

Figure 15. Correlations among the top 10 enzyme groups that were selected by the EC-nomenclature-based G-S-M for different CRC-associated metagenomic datasets, including the relative abundance values of the enzymes obtained from the samples belonging to different populations. The Jaccard index is used to calculate the correlation between the identified EC groups among two populations.

Figure 16 presents the commonalities among the top 10 enzyme groups that are selected for different datasets obtained from different populations. The methyltransferases (EC: 2.1.1) group has been identified among the top 10 groups for the five different population datasets obtained from Germany (DEU), France (FRA), India (IND), Italy (ITA), and Japan (JPN).

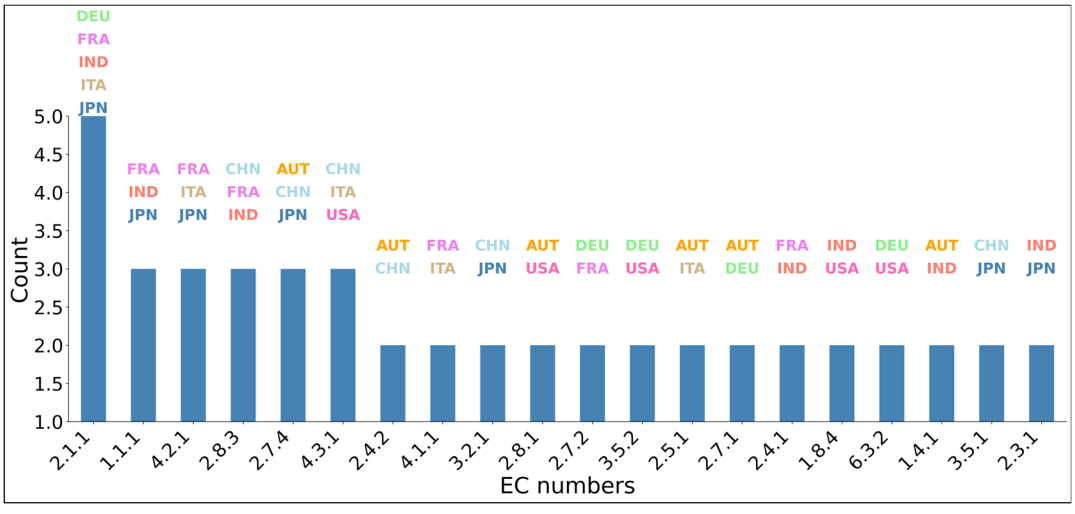

Figure 16. Commonalities among the top 10 enzyme groups that were selected by the EC-nomenclature-based G-S-M for different CRC-associated metagenomic datasets, including the relative abundance values of the enzymes obtained from the samples belonging to different populations.

It is demonstrated in Figure 5 that the CRC-associated metagenomic dataset specific to the Indian population yielded the highest AUC value when the EC-nomenclature-based G-S-M is run, as compared with AUC values of the other datasets belonging to the other populations. There may be several reasons behind the differences in performance metrics between the different populations. The geographic distribution of colorectal cancer (CRC) is strongly influenced by variations in dietary habits, emphasizing the role of diet in shaping the gut microbiome and its functions [67]. Gut microbes rely on residues from high-fiber foods to maintain their normal structure and metabolism, which underscores the importance of diet in microbial homeostasis. Among the metabolites produced by gut microbiota, butyrate plays a critical role in preserving intestinal homeostasis and epithelial health. This short-chain fatty acid acts as the primary energy source for intestinal epithelial cells and possesses anti-inflammatory properties. Furthermore, as a histone deacetylase inhibitor (HDACi) and a byproduct of fiber fermentation, butyrate mediates the protective effects of dietary fiber against CRC through mechanisms such as the modulation of gene expression, the promotion of apoptosis in cancer cells, and the regulation of immune responses.

4.2. Comparative Performance Evaluation of the EC-Nomenclature-Based G-S-M Model

We have comparatively evaluated the EC-nomenclature-based G-S-M with traditional feature selection methods, including XGB, SKB, IG, CMIM, MRMR, and FCBF, coupled with different classifiers, such as Adaboost, DT, LogitBoost, RF, SVM_opt, Stack_Logitboost_Kmeans, Stack_SVM_Kmeans, and XGBoost (as shown in Figure 6). In our earlier studies using the Grouping-Scoring-Modeling approach, we have shown that RF classifiers outperform other classifiers [25]. Therefore, in this study, we utilize the RF classifier within the EC-nomenclature-based G-S-M. Additionally, to show the performance of different classifiers on the G-S-M model, we experimented with XGBoost and DT. The TFS methods with the RF classifier yielded the following AUC values: 0.763 (XGB), 0.762 (SKB), 0.725 (IG), 0.495 (FCBF), 0.519 (MRMR), and 0.584 (CMIM), using 90 features. As one can notice from Figure 6, the other classifiers have either similar performance results or lower performance results compared with the RF classifier.

On the other hand, our experimental findings imply that the EC-nomenclature-based G-S-M model using RF has the capability to compete with other tested FS methods (as shown in Figure 6). Furthermore, as plotted in Figure 17, the superiority of the EC-nomenclature-based G-S-M approach becomes apparent (in terms of accuracy, specificity, and AUC performance metrics) when it is compared with the performance metrics of different feature selection methods on the CRC-associated metagenomic dataset. The experimental results underscore that, while the performance metrics of both the traditional feature selection methods and the EC-nomenclature-based G-S-M model are not very high, due to the complexity of the problem, the EC-nomenclature-based G-S-M model consistently outperforms the traditional feature selection methods, as illustrated in Figures 6 and 17. In addition to its performance, the EC-nomenclature-based G-S-M model also has the capability to select informative groups, which can be associated with the disease under investigation, which was CRC in this study.

The correlation among the top 100 features (enzymes) that are selected by different feature selection methods and the EC-nomenclature-based G-S-M approach is illustrated in Figure 18. It is shown that XGB and SKB have the highest Jaccard similarity (0.14).

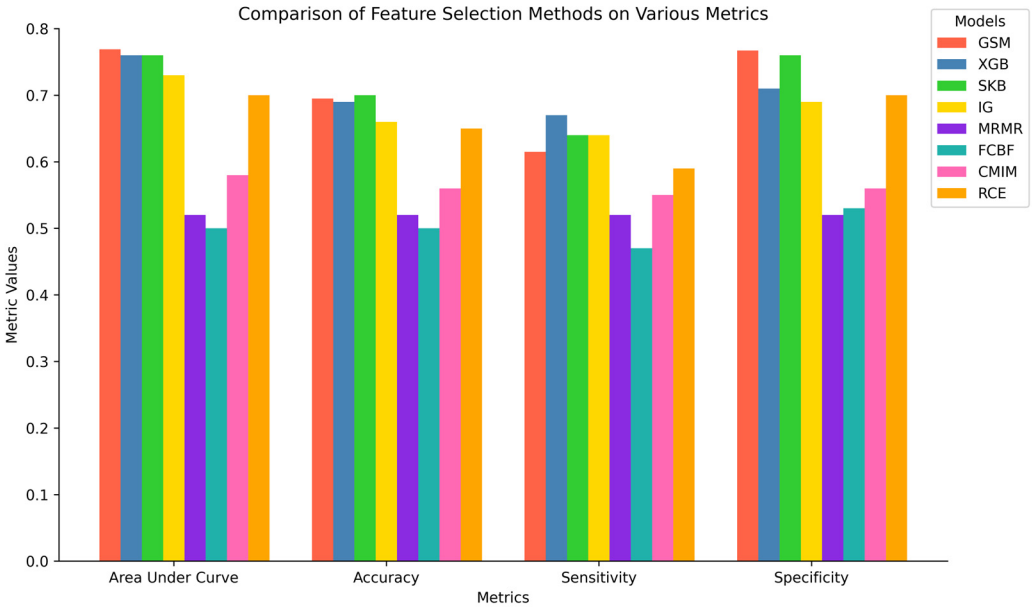

Figure 17. Performance metrics of different feature selection methods (XGB, SKB, IG, MRMR, FCBF, and CMIM) coupled with the RF classifier, RCE with the SVM classifier, and the EC-nomenclature-based G-S-M approach with RF when tested on the CRC-associated metagenomic dataset, including relative abundance values of the enzymes.

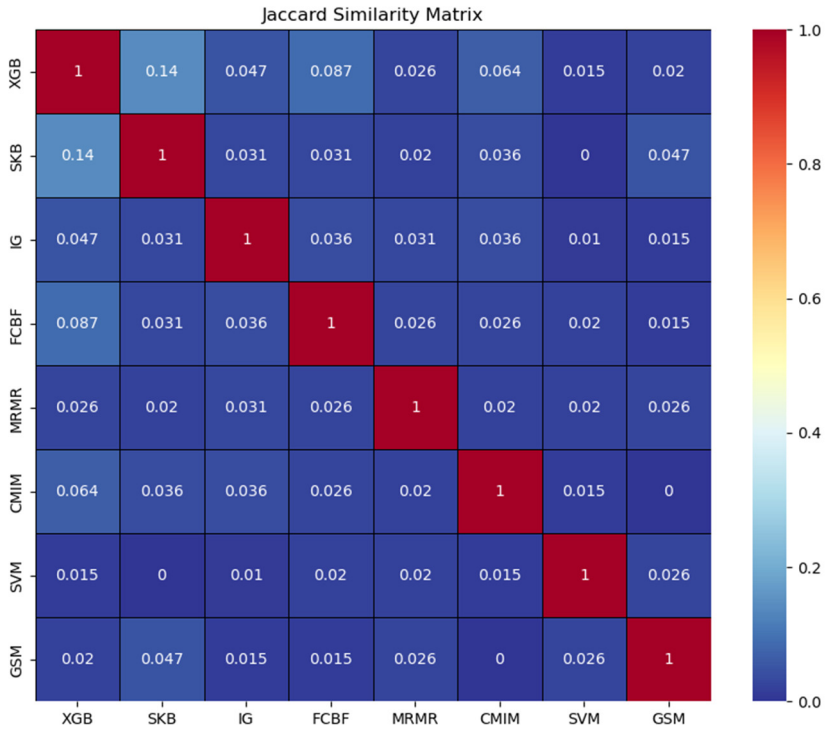

Figure 18. Correlations among the top 100 features that are selected by different feature selection algorithms and the EC-nomenclature-based G-S-M approach when tested on the CRC-associated metagenomic dataset, including relative abundance values of the enzymes.

Figure 19 presents the commonalities among the top 100 features (enzymes) that are selected by different feature selection algorithms and the EC-nomenclature-based G-S-M approach when tested on the CRC-associated metagenomic dataset, including the relative abundance values of the enzymes. Malate dehydrogenase (EC: 1.1.1.37) has been detected as an important enzyme by the XGB, SKB, FCBF, and CMIM methods. Similarly, citrate

CoA-transferase (EC: 2.8.3.10) has been identified as another important enzyme by XGB, SKB, FCBF, and the EC-nomenclature-based G-S-M approach (as shown in Figure 19).

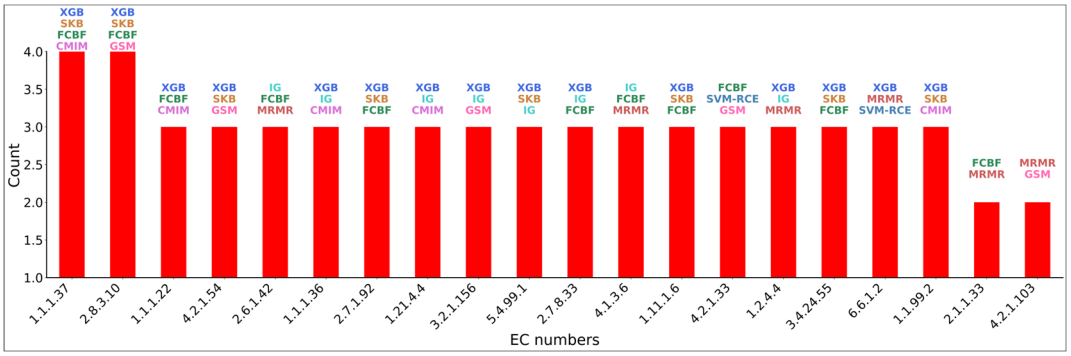

Figure 19. Commonalities among the top 100 features that were selected by different feature selection algorithms and the EC-nomenclature-based G-S-M approach when tested on CRC-associated metagenomic dataset, including relative abundance values of the enzymes.

4.3. Metabolic Pathways of Top Scoring Enzymes

The etiology of CRC is influenced by the bacterial communities that colonize the gastrointestinal tract. These microorganisms, through their intricate interactions with the gut environment, play a crucial role in shaping the physiology and pathology of the host. Within the intricate ecosystem of the gut microbiome, various bacterial species derive essential nutrients from otherwise indigestible dietary components or host-derived compounds. This process not only aids in the maintenance of microbial populations, but also contributes to the overall health of the host. Moreover, these bacteria have been found to activate the molecular signaling pathways that are vital for maintaining normal tissue function and supporting immune responses within the gut. By modulating these molecular pathways, the gut bacteria can influence the host's susceptibility to various diseases, including colorectal cancer. The dynamic interplay between the gut microbiome and host physiology underscores the importance of understanding the intricate relationships between bacterial communities and disease development. The tumor microenvironment and the gut microbiota (microenvironment) are play important role in the formation and progression of colorectal cancer. The balance of bacterial metabolism plays a crucial role in the health of the intestine. Bacterial side-products may affect the colon surface area and result in cancer formation.

The increased incidence of CRC can be attributed to various factors, such as an aging population, the prevalence of unhealthy dietary practices, smoking, a sedentary lifestyle, and obesity. There is a documented positive correlation between alcohol consumption and elevated rates of colorectal cancer globally, especially in developed nations [68]. In terms of dietary factors, certain anti-inflammatory components have been associated with potential benefits in reducing the risk of colorectal cancer. These components include the following: monounsaturated fatty acids, polyunsaturated fatty acids, omega-3 fatty acids, vitamins (B12, A, C, D, E, folic acid), minerals (zinc, magnesium, selenium), and bioactive compounds such as flavonoids, anthocyanidins, and certain herbs and spices (e.g., garlic, ginger, turmeric, and tea) [69–71]. Conversely, pro-inflammatory food components have been linked to a potential increase in the risk of developing colorectal cancer. These components include energy dense foods, total fat, trans fat, cholesterol, saturated fatty acids [72]. As shown in Figure 8, glycosidases (EC: 3.2.1), hydro-lyases (EC:4.2.1), and CoA-transferases (EC: 2.8.3) have been found to be the top three scoring CRC-associated enzyme groups. Glycosidases (EC: 3.2.1) play a role in several metabolic activities, especially for starch and sucrose metabolism (Figure 11A), sphingolipid metabolism (Figure 11B),

galactose metabolism (Figure 11C), N-glycan biosynthesis (Figure 11D), and glucuronate interconversions (Figure 11E).

In starch and sucrose metabolism and galactose metabolism, glycosidases (EC: 3.2.1), such as oligo-1,6-glucosidase (EC: 3.2.1.10) and alpha-glucosidase (EC: 3.2.1.20), result in an increasing amount of D-Glucose from different sources (Figure 11A,C). Especially in galactose metabolism, glycosidases (EC: 3.2.1) and alpha-galactosidase (EC: 3.2.1.22) activation result in the formation of D-Galactose and D-Glucose. Additionally, oligo-1,6-glucosidase (EC: 3.2.1.10) and alpha-glucosidase (EC: 3.2.1.20) activation results in increasing amounts of D-Fructose (Figure 11C). Also, other glycosidases (EC: 3.2.1) have different roles in these pathways in terms of contributing to D-Glucose, D-Galactose, and D-Fructose formation.

Blood glucose is controlled by the production of insulin, which is a hormon that allows the regulation activity of blood glucose in the body. Bacteria in the gut break down food substances like fiber to synthesize short-chain fatty acids like butyrate and propionate, which can influence pancreatic function and insulin regulation. However, when this balancing act mechanism is blocked, it causes a massive dysregulation in blood glucose levels. When pre-diabetic people consume high fiber diet to stimulate bacteria to produce SCFA, it may enhance their insulin sensitivity and glucose regulation.

The surfaces of gastrointestinal (GI) organs have diverse tasks, such as nutritional uptake and defense. The immune cells within the mucosa maintain functional balance to prevent inflammation and uphold barrier function. The rise in GI diseases is linked to dietary changes causing gut microbiome imbalance and chronic inflammation. Gut microbes exhibit a profound connection to host metabolism, although the precise mechanisms underpinning this relationship remain elusive. Understanding immuno-nutrition, nutritional compounds, the role of diet, gut microbiota, and the host immune response is crucial for disease prevention and treatment. Lipids, especially bioactive sphingolipids (SLs) like sphingomyelin (SM), sphingosine (Sph), ceramide (Cer), sphingosine-1-phosphate (S1P), and ceramide-1-phosphate (C1P), from the diet play a key role in GI barrier function and immune homeostasis [73].

As shown in Figure 11B, the galactosylceramidase (EC: 3.2.1.46) enzyme is responsible for the synthesis of Sph from psychosine and Cer synthesis from galacto-sylceramide, while glucosylceramidase (EC: 3.2.1.45) enzymes synthesize Sph from glucosylsphingosine and Cer from glucosyl-ceramide. Lactosyl-ceramide is hydrolyzed by beta-galactosidase (EC: 3.2.1.23) to form glucosyl-ceramide. Lactosyl-ceramide formation is performed by sialidase (EC: 3.2.1.18), beta-hexosaminidase (EC: 3.2.1.52), and alpha-galactosidase (EC: 3.2.1.22) enzymes (Figure 11B).

SLs, a diverse bioactive lipid class, are present in cellular membranes, lipoproteins, and lipid-rich structures like the skin, influencing apoptosis, cell growth, and migration. They impact pro- and anti-inflammatory immune responses and are regulated through endogenous metabolic pathways. SLs are also in food, affecting immune activation in the GI tract and related diseases. The actions of SLs include inhibiting intestinal lipid uptake, activating inflammatory receptors, and influencing lymphocyte chemotaxis. They neutralize bacterial endotoxins and alter intestinal microbiota, and SL metabolites affect cell functions and can influence the gut microbiota composition [73]. SLs determine cell fate by regulating cell proliferation and growth [74], and they are associated with many cancers, including acute myeloid leukemia [75]. Ceramide, the pivotal apoptotic molecule within the sphingolipid metabolism pathway, is synthesized via the de novo pathway facilitated by serine palmitoyl transferase (SPT) [76]. Studies have shown that targeting ceramide and S1P in SL metabolism for AML treatment regulates the apoptotic mechanisms and cell proliferation [77].

SLs are crucial in mammals, also playing roles in metabolic disorders, ranging from insulin resistance (IR) to hepatic steatosis. While SLs are acquired from the diet and synthesized de novo in mammalian tissues, a yet unexplored potential source of mammalian SLs may be their production by Bacteroidetes, which are a predominant phylum within the gut microbiome. The genomes of *Bacteroides* spp. and their counterparts encode SPT, enabling them to synthesize SLs [78]. The impact of SL production by gut *Bacteroides* on host SL homeostasis and related findings underscore the influence of gut-derived bacterial SLs on the host's lipid metabolism. In human cell cultures, bacterial SLs undergo processing by host SL metabolic pathways and, in murine models, lipids derived from Bacteroides are transferred to host epithelial tissue and the hepatic portal vein. These findings show that *Bacteroides* SLs may impact host lipid homeostasis [78].

In animal models, reducing ceramide in the liver improves insulin sensitivity, while inhibition of intestinal ceramide pathways enhances glucose metabolism. Human studies link hepatic or plasma ceramide levels to IR. SL regulation in tissues relies on multiple signaling pathways, controlled by synthesis, recycling, and intestinal uptake. Dietary SLs impact cholesterol absorption, hepatic lipid accumulation, obesity, and insulin sensitivity, partly by altering hepatic ceramide levels. Diet influences gut microbiota-derived SLs, in that way modulate some inflammation related pathways in the colon. However, their impact on the host lipid metabolism pathways remains unknown [78].

Studies indicate that bacterial sphingolipids from Bacteroides are reduced in IBD stool samples and correlate with lower gut inflammation, whereas increased host-derived SLs are associated with reduced Bacteroides levels in IBD patients. Mouse studies have supported these findings, showing disruptions in both bacterial and host sphingolipid pathways during inflammation and IBD. This discovery is crucial, due to the key role of sphingolipid signaling in IBD-related pathways and the notable decrease in Bacteroidetes in various IBD groups [79–82].

Cer content is linked to inflammation and metabolic diseases. Increasing Cer generation or inhibiting degradation can lead to Cer accumulation, inducing excessive apoptosis. Recent studies using HPLC, GC-MS, and MALDI-MS methods have revealed insights into dietary Cer structures, indicating a more complex role of the sphingolipid metabolite. Higher cellular Cer levels have been found to prevent inflammatory responses, while genetically modified yeast-produced Cers inhibit TNF-α signaling, maintaining cell viability. Plant-derived Cer-precursor sphingolipids have been used as dietary supplements to restore skin barrier function in humans [73]. C1P, a Cer metabolite, is a bioactive SL metabolite linked to cell proliferation, macrophage migration, and inflammatory response. C1P acts as an extracellular ligand through a G(i) protein-coupled plasma membrane receptor, potentially mediating chemotaxis, and regulates the inflammation in response to Cer stimuli [73].

Recent findings suggest an imbalance in SLs favoring specific types like S1P, C1P, Cers, Sph, and SM over more complex cerebrosides and gangliosides, promoting intestinal inflammation [83]. SLs interact with epithelial cells and immune cells, showing promise for treating immune diseases and as biomarkers. These lipids affect intestinal mucus function, compete with commensals for binding sites, and boost defense against pathogens. Intestinal SL levels are influenced by diet, endogenous production, and SL-producing bacteria. Bacterial and host SLs are structurally alike, affecting the immune responses and signaling pathways. Gut bacteria lacking sphingolipids are linked to inflammatory bowel disease, while human-produced sphingolipids in the gut increase. Gut microbiota metabolites signal to the host, impacting the health outcomes in inflammatory bowel disease [84].

In the literature, it has been also shown that an alteration in the SL metabolism is associated with several inherited human diseases, where defect in the lysosomal SL degradation results in an accumulation of non-degradable storage material in the organs as seen in as seen in GM1-gangliosidosis (GM1 beta-galactosidase deficiency), Tay–Sachs disease (mutations in β-hexosaminidase), Fabry disease (alpha-galactosidase deficiency), Gaucher disease (β-glucoceramidase I deficiency), and Krabbe disease (β-galactoceramidase deficiency) [85–88].

Also, hydro-lyases (EC:4.2.1) and CoA-transferases (EC: 2.8.3) perform metabolic activity together in styrene degradation (Figure 12A), butanoate metabolism (Figure 12B), the citric acid cycle (Figure 12C), and carnitine metabolism (Figure 12D). These enzymes contribute to key biochemical processes regulating energy metabolism and detoxification.

Butyrate is an essential mediator between the gut microbiota and host metabolic health. Therefore, it has regulatory effects on metabolic functions such as body weight, body composition, and glucose homeostasis, absorbing and metabolizing though tissues and cells beyond the colon [89].

4.4. The Microorganisms That Synthesize the Enzymes Identified by the EC-Nomenclature-Based G-S-M and Their Association with CRC Development

The microbial community of the gut has a crucial role in safeguarding the host from harmful microorganisms. Also, the microbial community influences the immune system and regulates metabolic processes, therefore, it is recognized as an endocrine organ. This dynamic interplay between the host and the microbiome is pivotal in human physiology and metabolism, encompassing functions such as synthesizing essential vitamins like vitamin K, extracting energy from indigestible carbohydrates like pectin, modulating the human immune system, and preventing colonization by enteropathogenic bacteria [7,90–92]. The system's homeostasis and the abundance of microbiota in the gut are regulated by factors such as the optimal pH range, suitable oxygen concentration, and abundant nutrients. Disruptions within the typical composition of the intestinal microbiome pose significant challenges [93–95]. Dysbiosis, characterized by an imbalance in the symbiotic relationship between the host and the intestinal microbiota, has been identified as a potential factor contributing to several diseases, such as inflammatory bowel disease (IBD) [96], hepatocellular carcinoma (HCC) [97], and colorectal cancer (CRC) [98,99]. The onset of these diseases is often noticed by disruptions in the normal microbiota, prompting immune system activation and subsequent inflammation [100]. Given the increased susceptibility to CRC in individuals with IBD, and the observed dysbiosis in some cases, it is reasonable to infer that CRC associated with IBD is preceded by a phase of dysbiosis. Consequently, the microbiome is increasingly implicated as a significant factor in the onset and/or progression of colonic carcinogenesis [99,101,102].

The gut microbiome can be related to disease formation and prognosis, and hence it is important to understand the underlying mechanisms involved. It is important to note that, while some bacteria are known to cause infections in humans, others are part of the normal microbiota. The pathogenicity of these bacteria can vary, and their impact on human health depends on various factors, including the specific strain and the host's immune system.

The enzymes within the top three significant enzyme groups (glycosidases (EC: 3.2.10), hydro-lyases (EC: 4.2.1), and CoA-transferases (2.8.3)) identified by the EC-nomenclature-based G-S-M on the CRC-associated metagenomic dataset are synthesized by different organisms. As shown in Figure 13, *Eschericia coli* (*E. coli*), Salmonella enterica, Klebsiella pneumoniae, Staphylococcus aureus, Streptococcus pneumoniae, and Clostridioides dificile can be relevant for CRC development. Additionally, as shown in Figure 14, *E. coli*, *Salmonella enterica*, *Klebsiella pneumoniae*, and *Citrobacter koseri* were highlighted by the XGB and SKB feature selection methods coupled with the RF classifier.

E. coli is reported in the literature to have a higher abundance in CRC patients compared with controls [103]. Recent studies have strengthened the association between certain strains of *E. coli* and CRC. Notably, specific *E. coli* strains and *Klebsiella pneumoniae* produce colibactin, a genotoxin that can damage DNA, potentially leading to cancer development [104]. The research indicates that these colibactin-producing *E. coli* strains are more prevalent in industrialized countries, correlating with higher CRC incidence rates. This suggests that targeting these bacterial strains through vaccines or other treatments could be a viable strategy to reduce cancer risk. Additionally, studies have found that mucosa-associated *E. coli* are more prevalent in CRC biopsies compared to those of healthy controls. These strains can survive within macrophages and induce a pro-inflammatory response, contributing to increased cell proliferation and cancer progression. The production of colibactin by these *E. coli* strains leads to DNA damage, cell cycle arrest, and chromosomal rearrangements in mammalian cells, further implicating them in CRC development [105]. A study found that the colonic mucosa of patients with colorectal carcinoma was colonized by intracellular *E. coli*, whereas normal colonic mucosa was not. This suggests a potential role of *E. coli* in colorectal carcinogenesis. Researchers observed an increased level of mucosa-associated and internalized *E. coli* in tumor tissues compared to normal tissues [106]. Pathogenic cyclomodulin-positive *E. coli* strains were more prevalent in patients with advanced-stage colon cancer. Infection with a colon-cancer-associated *E. coli* strain led to a significant increase in colonic polyps in mice, supporting the hypothesis that pathogenic *E. coli* could be a cofactor in CRC pathogenesis [106]. The evidence and mechanistic explanations for the role of colibactin-producing *E. coli* strains in CRC has been reported, and the authors also present observations that challenge this hypothesis, suggesting that *E. coli* may preferentially colonize cancerous lesions as an effect rather than a cause [107].

A study conducted in the Netherlands found that patients diagnosed with severe *Salmonella* infections had a significantly increased risk of developing cancer in the ascending and transverse parts of the colon. This risk was particularly associated with *Salmonella Enteritidis* infections, suggesting a potential role of this pathogen in colon cancer development. Research has indicated that repeated exposure to *non-typhoidal Salmonella* is significantly associated with an increased risk of colon cancer. In mouse models, repetitive low-dose NTS infections accelerated tumor growth, with the bacteria preferentially infecting the pre-transformed cells and exponentially increasing the rate of cellular transformation with each infection round [108]. It is also shown that infection with genotoxin-producing *Salmonella enterica* collaborates with the loss of the tumor suppressor APC gene to enhance genomic instability in colonic epithelial cells through the PI3K pathway [109].

Clostridioides difficile is a toxin-producing bacteria. The relationship between *Clostridioides difficile* and CRC has been shown in the literature. In addition to that, a statistically significant correlation between obese patients with *Clostridioides difficile* and an increased incidence of CRC has been shown [110].

In the literature, a study found that a variant of *Citrobacter freundii* reduced the latent period of colon carcinogenesis in mice treated with 1,2-dimethylhydrazine, suggesting a potential role in promoting colorectal cancer development. These findings suggest that certain strains of *Citrobacter*, such as *Citrobacter freundii*, may influence colorectal carcinogenesis, potentially by reducing the latency period for tumor development [111]. Further research is needed to elucidate the mechanisms by which *Citrobacter* species may contribute to CRC and to determine their potential as targets for therapeutic intervention. The review discusses various bacterial species implicated in CRC development, including *Salmonella*, *E.coli*, and *Citrobacter rodentium.* This study highlights that *Salmonella* and *C. rodentium* can induce epithelial–mesenchymal transition (EMT) in the intestinal cells, a

process associated with cancer progression. Additionally, *C. rodentium* has been shown to activate signaling pathways such as PI3K/AKT, WNT-β-catenin, and TLR-4-based NF-κB, which are involved in colorectal carcinogenesis. This study shows that E. coli produces a genotoxin that induces interstrand crosslinks in host cells, resulting in a mutational signature detected in CRC genomes [62]. The same study demonstrates the role of various bacteria in CRC development and indicates that *Citrobacter rodentium* may contribute to tumor development by inducing epithelial–mesenchymal transition in the intestinal cells. The research highlights the importance of microbial metabolites and signaling pathways in CRC progression [112]. These findings suggest that *Citrobacter* species may play a role in colorectal cancer development through mechanisms such as inducing epithelial–mesenchymal transition and activating oncogenic signaling pathways.

In addition to other *Salmonella* strains, bacteria associated with colorectal cancer such as *Klebsiella pneumoniae* have also been found. *Klebsiella pneumoniae* is an opportunistic pathogen known to cause infections, particularly in immunocompromised individuals [113]. Chronic infections and inflammation are major risk factors for cancer, including CRC [114]. *Klebsiella pneumoniae* has been implicated in chronic intestinal inflammation, which is thought to promote an environment conducive to cancer development [115]. Long-term inflammation can lead to DNA damage, immune system dysfunction, and the activation of oncogenic pathways [100]. The bacterium may promote tumorigenesis by altering gut microbiota composition and fostering a pro-inflammatory environment [97]. Certain strains of *Klebsiella pneumoniae* produce carcinogenic metabolites, which can directly damage the host DNA, potentially leading to mutations and cancer development. Some studies have shown that *Klebsiella* can induce DNA damage in host cells, possibly through the production of toxic metabolites or the secretion of virulence factors that interfere with host cell signaling [116]. *Klebsiella pneumoniae* produces several virulence factors, such as lipopolysaccharides (LPS), siderophores, and capsular polysaccharides, which may play a role in CRC development. These virulence factors can trigger immune responses, cause inflammation, and interact with the host's cells in different ways that could support cancer development. The ability of *Klebsiella* to form biofilms is another important factor. Biofilms protect the bacteria from immune responses and antibiotics, potentially allowing the bacteria to persist in the gut and may contribute to chronic inflammation. *Klebsiella pneumoniae* has been linked to increased tumorigenesis in animal models of CRC. It has been suggested that the bacterium may enhance the development of colonic tumors in genetically predisposed mice by promoting chronic inflammation and altering immune responses. There is evidence that *Klebsiella* can interact with the host cell pathways that regulate inflammation, immune responses, and tumor growth, thus potentially playing a role in promoting CRC. Another study revealed that *Esherichia* and *Klebsiella* were found to be enriched in the intestines of CRC patients [67]. The authors also showed that *Klebsiella*-induced inflammation may contribute to tumor formation by influencing immune cell activity and cytokine release. Another study showed that *Klebsiella pneumoniae* interacts with the host's immune system in several ways that could affect both local inflammation and systemic immune responses, contributing to cancer progression [117].

Immune responses play a critical role in both the prevention and progression of CRC, and changes in the microbial populations in the gut could contribute to an environment conducive to CRC development. Gut microbiomes can modulate the immune response in a way that might influence cancer progression. It is important to note that the human gut microbiome is complex, and research is ongoing in order to fully understand the relationships between various microorganisms, their enzymatic activities in different pathways, and cancer development.

4.5. Study Limitations and Future Directions

This study may have some limitations, such as a reliance on prior biological knowledge, generalizability, validation, experimental limitations, and dietary and environmental factors. The CRC-associated metagenomic dataset is inherently complex and high-dimensional, making it challenging to interpret the results comprehensively. The functional profiles derived from the metagenomic data may contain noise or biases, which could affect the accuracy of the findings. While incorporating enzyme categories as domain knowledge provides insights, it may introduce biases by prioritizing specific pathways or enzymes, potentially overlooking novel or less-studied biomarkers. The EC-nomenclature-based G-S-M method depends heavily on accurate and comprehensive biological databases, which may not always reflect the latest discoveries. The findings, such as the associations with specific enzymes (e.g., glycosidases and CoA-transferases) and microbial taxa, may not generalize across diverse populations, due to differences in microbiome composition influenced by geography, diet, and genetics. This study relies primarily on computational modeling and cross-validation, lacking experimental validation to confirm the causal roles of the identified enzymes or microbes in CRC progression. The precise molecular mechanisms linking the identified microbial enzymes, metabolites, and CRC progression remain unresolved, highlighting a gap in mechanistic understanding. The interactions between microbiota and host genetics, immune responses, and environmental factors (e.g., diet) are not fully explored in this study.

The traditional feature selection methods might fail to capture the full biological complexity of the data, but even the EC-nomenclature-based G-S-M approach has limitations in terms of performance, as noted in the results. The dataset appears to be cross-sectional, which limits insights into the temporal changes in the microbiome and their role in CRC development or progression. This study does not account for the influence of external factors like diet, medication, or lifestyle, which are known to significantly impact the gut microbiome. While some key taxa (e.g., *Escherichia coli* and *Salmonella enterica*) are identified, this study does not deeply explore the microbial interactions or the broader ecological dynamics within the gut microbiome. Given the complexity of the data and the number of classifiers tested, there is a risk of overfitting, which could reduce the reproducibility of the findings when applied to new datasets. By addressing these limitations in future studies, researchers can enhance the robustness and translational potential of metagenomic analyses in understanding CRC and other microbiome-associated diseases. Current metagenomic datasets may lack adequate representation of diverse populations, leading to limited generalizability. Incomplete functional annotation in microbial databases may hinder the identification of novel biomarkers.

To enlighten the utility of metagenomics in CRC, our research highlights specific enzymes and biological pathways for further exploration. Future research should focus on addressing unresolved questions, including the precise molecular mechanisms linking microbial metabolites to CRC development, the role of host genetics in shaping microbiome interactions, and the long-term impact of dietary interventions on CRC risk and gut health. In such studies, it is also important to include data from diverse populations in order to capture the geographical, genetic, and dietary variations in microbiome composition. Combining metagenomic data with transcriptomic, metabolomic, and proteomic analyses can also provide new insights into CRC-associated pathways. To this end, it is important to develop multi-omics models to study the interplay between microbial metabolites and host pathways. By combining metagenomics, functional genomics, and computational modeling, these efforts will contribute to the development of personalized interventions, microbiome-based therapeutics, and targeted modulation of microbial metabolites, ultimately advancing the prevention, diagnosis, and treatment of colorectal cancer.

5. Conclusions

Colorectal cancer (CRC) stands as one of the most prevalent cancer types globally, and the significant role of the gut microbiome in CRC development is well established. To comprehend the involvement of the gut microbiome in CRC formation and progression, computational methodologies and analyses are imperative. In this study, we attempt to build a classification model based on the community-level enzyme commission (EC) abundance values that are obtained from the functional profile of the CRC-associated metagenomic data. Conventional feature selection methods fail to leverage the prior biological knowledge available in different databases. Alternatively, the proposed EC-nomenclature-based Grouping-Scoring-Modeling (G-S-M) method incorporates enzyme categories as the domain knowledge into the group selection process, yielding biological insights.

Our experimental findings imply that glycosidases (EC: 3.2.1), CoA-transferases (EC: 2.8.3), hydro-lyases (EC: 4.2.1), oligo-1,6-glucosidase (EC:3.2.1.10), crotonobetainyl-CoA hydratase (EC:4.2.1.149), and citrate CoA-transferase (EC: 2.8.3.10) enzymes can be associated with CRC development as part of different molecular pathways. Along this line, our findings indicate that *Escherichia coli*, *Salmonella enterica*, *Clostridioides difficile*, *Klebsiella pneumoniae*, *Citrobacter koseri*, *Staphylococcus aureus*, and *Streptococcus pneumoniae* can contribute directly or indirectly to CRC development. This research endeavor highlights the transformative potential of integrating biological domain knowledge into machine learning algorithms during metagenomic analyses. By leveraging novel methods, such as the EC-nomenclature-based G-S-M, researchers can gain deeper insights into CRC-associated microbial enzymes, metabolic pathways, and disease mechanisms.

Author Contributions: All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication. Conceptualization, N.S.E., B.B.-G. and M.Y.; Data curation, N.S.E., B.B.-G. and M.Y.; Formal analysis, N.S.E., B.B.-G. and M.Y.; Investigation, N.S.E., B.B.-G. and M.Y.; Software, N.S.E., B.B.-G. and M.Y.; Validation, N.S.E., B.B.-G. and M.Y.; Writing—original draft, N.S.E., B.B.-G. and M.Y. All authors have read and agreed to the published version of the manuscript.

Funding: We would like to thank The Scientific and Technological Research Council of Türkiye (TÜBİTAK) 2211A BIDEP program for supporting the work of N.S.E. The work of B.B.-G. has also been supported by the Abdullah Gul University Support Foundation (AGUV). B.B.-G. would like to express her gratitude for the L'Oréal-UNESCO Young Women Scientist Award. This research was made possible by the support of the L'Oréal-UNESCO Young Women Scientist Program. The work of M.Y. has been supported by Zefat Academic College.

Data Availability Statement: Metagenomic data is downloaded from https://pmc.ncbi.nlm.nih.gov/articles/PMC8096432/#supp5, accessed on 7 March 2023.

Acknowledgments: We extend our gratitude to COST Action ML4Microbiome, which has played a pivotal role in advancing microbiome research and facilitating the expansion of these research endeavors.

Conflicts of Interest: The authors declare that this research was conducted in the absence of any commercial or financial relationships that could be construed as potential conflicts of interest.

References

1. Li, J.; Ma, X.; Chakravarti, D.; Shalapour, S.; DePinho, R.A. Genetic and biological hallmarks of colorectal cancer. *Genes Dev.* **2021**, *35*, 787–820. [CrossRef] [PubMed]
2. Mármol, I.; Sánchez-De-Diego, C.; Pradilla Dieste, A.; Cerrada, E.; Rodriguez Yoldi, M. Colorectal Carcinoma: A General Overview and Future Perspectives in Colorectal Cancer. *Int. J. Mol. Sci.* **2017**, *18*, 197. [CrossRef]
3. Ryan, B.M.; Wolff, R.K.; Valeri, N.; Khan, M.; Robinson, D.; Paone, A.; Bowman, E.D.; Lundgreen, A.; Caan, B.; Potter, J.; et al. An analysis of genetic factors related to risk of inflammatory bowel disease and colon cancer. *Cancer Epidemiol.* **2014**, *38*, 583–590. [CrossRef] [PubMed]

4. Wong, C.C.; Yu, J. Gut microbiota in colorectal cancer development and therapy. *Nat. Rev. Clin. Oncol.* **2023**, *20*, 429–452. [CrossRef] [PubMed]
5. Kim, J.; Lee, H.K. Potential Role of the Gut Microbiome In Colorectal Cancer Progression. *Front. Immunol.* **2022**, *12*, 807648. [CrossRef]
6. McNally, L.; Brown, S.P. Building the microbiome in health and disease: Niche construction and social conflict in bacteria. *Philos. Trans. R. Soc. B Biol. Sci.* **2015**, *370*, 20140298. [CrossRef]
7. Ursell, L.K.; Metcalf, J.L.; Parfrey, L.W.; Knight, R. Defining the human microbiome. *Nutr. Rev.* **2012**, *70* (Suppl. 1), S38–S44. [CrossRef]
8. Scarpellini, E.; Ianiro, G.; Attili, F.; Bassanelli, C.; De Santis, A.; Gasbarrini, A. The human gut microbiota and virome: Potential therapeutic implications. *Dig. Liver Dis.* **2015**, *47*, 1007–1012. [CrossRef]
9. Stearns, J.C.; Lynch, M.D.J.; Senadheera, D.B.; Tenenbaum, H.C.; Goldberg, M.B.; Cvitkovitch, D.G.; Croitoru, K.; Moreno-Hagelsieb, G.; Neufeld, J.D. Bacterial biogeography of the human digestive tract. *Sci. Rep.* **2011**, *1*, 170. [CrossRef]
10. Matamoros, S.; Gras-Leguen, C.; Le Vacon, F.; Potel, G.; de La Cochetiere, M.-F. Development of intestinal microbiota in infants and its impact on health. *Trends Microbiol.* **2013**, *21*, 167–173. [CrossRef]
11. Yadav, D.; Ghosh, T.S.; Mande, S.S. Global investigation of composition and interaction networks in gut microbiomes of individuals belonging to diverse geographies and age-groups. *Gut Pathog.* **2016**, *8*, 17. [CrossRef]
12. Yatsunenko, T.; Rey, F.E.; Manary, M.J.; Trehan, I.; Dominguez-Bello, M.G.; Contreras, M.; Magris, M.; Hidalgo, G.; Baldassano, R.N.; Anokhin, A.P.; et al. Human gut microbiome viewed across age and geography. *Nature* **2012**, *486*, 222–227. [CrossRef] [PubMed]
13. Xu, Z.; Knight, R. Dietary effects on human gut microbiome diversity. *Br. J. Nutr.* **2015**, *113*, S1–S5. [CrossRef] [PubMed]
14. Gao, B.; Chi, L.; Zhu, Y.; Shi, X.; Tu, P.; Li, B.; Yin, J.; Gao, N.; Shen, W.; Schnabl, B. An Introduction to Next Generation Sequencing Bioinformatic Analysis in Gut Microbiome Studies. *Biomolecules* **2021**, *11*, 530. [CrossRef] [PubMed]
15. Qin, J.; Li, R.; Raes, J.; Arumugam, M.; Burgdorf, K.S.; Manichanh, C.; Nielsen, T.; Pons, N.; Levenez, F.; Yamada, T.; et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **2010**, *464*, 59–65. [CrossRef]
16. Turnbaugh, P.J.; Hamady, M.; Yatsunenko, T.; Cantarel, B.L.; Duncan, A.; Ley, R.E.; Sogin, M.L.; Jones, W.J.; Roe, B.A.; Affourtit, J.P.; et al. A core gut microbiome in obese and lean twins. *Nature* **2009**, *457*, 480–484. [CrossRef]
17. Nam, N.N.; Do, H.D.K.; Trinh, K.T.L.; Lee, N.Y. Metagenomics: An Effective Approach for Exploring Microbial Diversity and Functions. *Foods* **2023**, *12*, 2140. [CrossRef]
18. Liu, Y.-X.; Qin, Y.; Chen, T.; Lu, M.; Qian, X.; Guo, X.; Bai, Y. A practical guide to amplicon and metagenomic analysis of microbiome data. *Protein Cell* **2021**, *12*, 315–330. [CrossRef]
19. Kinoshita, Y.; Niwa, H.; Uchida-Fujii, E.; Nukada, T. Establishment and assessment of an amplicon sequencing method targeting the 16S-ITS-23S rRNA operon for analysis of the equine gut microbiome. *Sci. Rep.* **2021**, *11*, 11884. [CrossRef]
20. Zhang, L.; Chen, F.; Zeng, Z.; Xu, M.; Sun, F.; Yang, L.; Bi, X.; Lin, Y.; Gao, Y.; Hao, H.; et al. Advances in Metagenomics and Its Application in Environmental Microorganisms. *Front. Microbiol.* **2021**, *12*, 766364. [CrossRef]
21. Blanco-Míguez, A.; Beghini, F.; Cumbo, F.; McIver, L.J.; Thompson, K.N.; Zolfo, M.; Manghi, P.; Dubois, L.; Huang, K.D.; Thomas, A.M.; et al. Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat. Biotechnol.* **2023**, *41*, 1633–1644. [CrossRef] [PubMed]
22. Beghini, F.; McIver, L.J.; Blanco-Míguez, A.; Dubois, L.; Asnicar, F.; Maharjan, S.; Mailyan, A.; Manghi, P.; Scholz, M.; Thomas, A.M.; et al. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *eLife* **2021**, *10*, e65088. [CrossRef] [PubMed]
23. Hirsch, F.R.; Kim, C. The Importance of Biomarker Testing in the Treatment of Advanced Non-Small Cell Lung Cancer: A Podcast. *Oncol. Ther.* **2024**, *12*, 223–231. [CrossRef]
24. Perscheid, C. Integrative biomarker detection on high-dimensional gene expression data sets: A survey on prior knowledge approaches. *Briefings Bioinform.* **2021**, *22*, bbaa151. [CrossRef]
25. Yousef, M.; Inal, Y.; Gungor, B.B.; Allmer, J. G-S-M: A Comprehensive Framework for Integrative Feature Selection in Omics Data Analysis and Beyond. *bioRxiv* **2024**. [CrossRef]
26. Chou, C.-H.; Shrestha, S.; Yang, C.-D.; Chang, N.-W.; Lin, Y.-L.; Liao, K.-W.; Huang, W.-C.; Sun, T.-H.; Tu, S.-J.; Lee, W.-H.; et al. miRTarBase update 2018: A resource for experimentally validated microRNA-target interactions. *Nucleic Acids Res.* **2018**, *46*, D296–D302. [CrossRef]
27. Piñero, J.; Queralt-Rosinach, N.; Bravo, À.; Deu-Pons, J.; Bauer-Mehren, A.; Baron, M.; Sanz, F.; Furlong, L.I. DisGeNET: A discovery platform for the dynamical exploration of human diseases and their genes. *Database* **2015**, *2015*, bav028. [CrossRef] [PubMed]
28. Kanehisa, M.; Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **2000**, *28*, 27–30. [CrossRef]

29. Hubbard, T.J.P.; Ailey, B.; Brenner, S.E.; Murzin, A.G.; Chothia, C. SCOP, Structural classification of proteins database: Applications to evaluation of the effectiveness of sequence alignment methods and statistics of protein structural data. *Acta Crystallogr. Sect. D Struct. Biol.* **1998**, *54*, 1147–1154. [CrossRef]
30. Orengo, C.; Michie, A.; Jones, S.; Jones, D.; Swindells, M.; Thornton, J. CATH—A hierarchic classification of protein domain structures. *Structure* **1997**, *5*, 1093–1109. [CrossRef]
31. Matsuta, Y.; Ito, M.; Tohsato, Y. ECOH: An Enzyme Commission number predictor using mutual information and a support vector machine. *Bioinformatics* **2013**, *29*, 365–372. [CrossRef] [PubMed]
32. Yousef, M.; Abdallah, L.; Allmer, J. maTE: Discovering expressed interactions between microRNAs and their targets. *Bioinformatics* **2019**, *35*, 4020–4028. [CrossRef]
33. Yousef, M.; Goy, G.; Bakir-Gungor, B. miRModuleNet: Detecting miRNA-mRNA Regulatory Modules. *Front. Genet.* **2022**, *13*, 767455. [CrossRef] [PubMed]
34. Yousef, M.; Ülgen, E.; Sezerman, O.U. CogNet: Classification of gene expression data based on ranked active-subnetwork-oriented KEGG pathway enrichment analysis. *PeerJ Comput. Sci.* **2021**, *7*, e336. [CrossRef]
35. Yousef, M.; Ozdemir, F.; Jaaber, A.; Allmer, J.; Bakir-Gungor, B. PriPath: Identifying Dysregulated Pathways from Differential Gene Expression via Grouping, Scoring and Modeling with an Embedded Machine Learning Approach. *Preprint* 2022. [CrossRef]
36. Jabeer, A.; Temiz, M.; Bakir-Gungor, B.; Yousef, M. miRdisNET: Discovering microRNA biomarkers that are associated with diseases utilizing biological knowledge-based machine learning. *Front. Genet.* **2023**, *13*, 1076554. [CrossRef]
37. Ersoz, N.S.; Bakir-Gungor, B.; Yousef, M. GeNetOntology: Identifying affected gene ontology terms via grouping, scoring, and modeling of gene expression data utilizing biological knowledge-based machine learning. *Front. Genet.* **2023**, *14*, 1139082. [CrossRef]
38. Söylemez, Ü.G.; Yousef, M.; Bakir-Gungor, B. AMP-GSM: Prediction of Antimicrobial Peptides via a Grouping–Scoring–Modeling Approach. *Appl. Sci.* **2023**, *13*, 5106. [CrossRef]
39. Yousef, M.; Kumar, A.; Bakir-Gungor, B. Application of Biological Domain Knowledge Based Feature Selection on Gene Expression Data. *Entropy* **2021**, *23*, 2. [CrossRef]
40. Pudjihartono, N.; Fadason, T.; Kempa-Liehr, A.W.; O'Sullivan, J.M. A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. *Front. Bioinform.* **2022**, *2*, 927312. [CrossRef]
41. Kuzudisli, C.; Bakir-Gungor, B.; Bulut, N.; Qaqish, B.; Yousef, M. Review of feature selection approaches based on grouping of features. *PeerJ* **2023**, *11*, e15666. [CrossRef]
42. Prasetiyowati, M.I.; Maulidevi, N.U.; Surendro, K. Determining threshold value on information gain feature selection to increase speed and prediction accuracy of random forest. *J. Big Data* **2021**, *8*, 84. [CrossRef]
43. Radovic, M.; Ghalwash, M.; Filipovic, N.; Obradovic, Z. Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. *BMC Bioinform.* **2017**, *18*, 9. [CrossRef] [PubMed]
44. Gopika, N.; A. Meena Kowshalaya, M.E. Correlation Based Feature Selection Algorithm for Machine Learning. In Proceedings of the 2018 3rd International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 15–16 October 2018; pp. 692–695. [CrossRef]
45. Bakir-Gungor, B.; Hacılar, H.; Jabeer, A.; Nalbantoglu, O.U.; Aran, O.; Yousef, M. Inflammatory bowel disease biomarkers of human gut microbiota selected via different feature selection methods. *PeerJ* **2022**, *10*, e13205. [CrossRef] [PubMed]
46. Ghosh, M.; Guha, R.; Sarkar, R.; Abraham, A. A wrapper-filter feature selection technique based on ant colony optimization. *Neural Comput. Appl.* **2020**, *32*, 7839–7857. [CrossRef]
47. Yousef, M.; Jung, S.; Showe, L.C.; Showe, M.K. Recursive Cluster Elimination (RCE) for classification and feature selection from gene expression data. *BMC Bioinform.* **2007**, *8*, 144. [CrossRef]
48. Kuzudisli, C.; Bakir-Gungor, B.; Qaqish, B.; Yousef, M. RCE-IFE: Recursive cluster elimination with intra-cluster feature elimination. *bioRxiv* **2024**. [CrossRef]
49. Wang, L.; Wang, Y.; Chang, Q. Feature selection methods for big data bioinformatics: A survey from the search perspective. *Methods* **2016**, *111*, 21–31. [CrossRef]
50. 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, San Francisco, CA, USA, 13–17 August 2016; ACM: New York, NY, USA, 2016; pp. 785–794. [CrossRef]
51. Mathieu, A.; Leclercq, M.; Sanabria, M.; Perin, O.; Droit, A. Machine Learning and Deep Learning Applications in Metagenomic Taxonomy and Functional Annotation. *Front. Microbiol.* **2022**, *13*, 811495. [CrossRef]
52. Cammarota, G.; Ianiro, G.; Ahern, A.; Carbone, C.; Temko, A.; Claesson, M.J.; Gasbarrini, A.; Tortora, G. Gut microbiome, big data and machine learning to promote precision medicine for cancer. *Nat. Rev. Gastroenterol. Hepatol.* **2020**, *17*, 635–648. [CrossRef]
53. Marcos-Zambrano, L.J.; Karaduzovic-Hadziabdic, K.; Turukalo, T.L.; Przymus, P.; Trajkovik, V.; Aasmets, O.; Berland, M.; Gruca, A.; Hasic, J.; Hron, K.; et al. Applications of Machine Learning in Human Microbiome Studies: A Review on Feature Selection, Biomarker Identification, Disease Prediction and Treatment. *Front. Microbiol.* **2021**, *12*. [CrossRef] [PubMed]

54. Chandrashekar, G.; Sahin, F. A survey on feature selection methods. *Comput. Electr. Eng.* **2014**, *40*, 16–28. [CrossRef]
55. Pödör, Z.; Hekfusz, M. Comparing Feature Selection Methods on Metagenomic Data using Random Forest Classifier. *Trans. Mach. Learn. Artif. Intell.* **2024**, *12*, 175–187. [CrossRef]
56. Bakir-Gungor, B.; Bulut, O.; Jabeer, A.; Nalbantoglu, O.U.; Yousef, M. Discovering Potential Taxonomic Biomarkers of Type 2 Diabetes From Human Gut Microbiota via Different Feature Selection Methods. *Front. Microbiol.* **2021**, *12*, 628426. [CrossRef]
57. Bakir-Gungor, B.; Temiz, M.; Inal, Y.; Cicekyurt, E.; Yousef, M. CCPred: Global and population-specific colorectal cancer prediction and metagenomic biomarker identification at different molecular levels using machine learning techniques. *Comput. Biol. Med.* **2024**, *182*, 109098. [CrossRef] [PubMed]
58. Dai, Z.; Coker, O.O.; Nakatsu, G.; Wu, W.K.K.; Zhao, L.; Chen, Z.; Chan, F.K.L.; Kristiansen, K.; Sung, J.J.Y.; Wong, S.H.; et al. Multi-cohort analysis of colorectal cancer metagenome identified altered bacteria across populations and universal bacterial markers. *Microbiome* **2018**, *6*, 70. [CrossRef]
59. Xu, X.; Ocansey, D.K.W.; Hang, S.; Wang, B.; Amoah, S.; Yi, C.; Zhang, X.; Liu, L.; Mao, F. The gut metagenomics and metabolomics signature in patients with inflammatory bowel disease. *Gut Pathog.* **2022**, *14*, 26. [CrossRef]
60. Jacobs, J.P.; Lagishetty, V.; Hauer, M.C.; Labus, J.S.; Dong, T.S.; Toma, R.; Vuyisich, M.; Naliboff, B.D.; Lackner, J.M.; Gupta, A.; et al. Multi-omics profiles of the intestinal microbiome in irritable bowel syndrome and its bowel habit subtypes. *Microbiome* **2023**, *11*, 5. [CrossRef]
61. Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; et al. Gene ontology: Tool for the unification of biology. *Nat. Genet.* **2000**, *25*, 25–29. [CrossRef]
62. Dougherty, M.W.; Jobin, C. Intestinal bacteria and colorectal cancer: Etiology and treatment. *Gut Microbes* **2023**, *15*, 2185028. [CrossRef]
63. Hera, M.R.; Liu, S.; Wei, W.; Rodriguez, J.S.; Ma, C.; Koslicki, D. Metagenomic functional profiling: To sketch or not to sketch? *Bioinformatics* **2024**, *40*, ii165–ii173. [CrossRef] [PubMed]
64. David, L.A.; Maurice, C.F.; Carmody, R.N.; Gootenberg, D.B.; Button, J.E.; Wolfe, B.E.; Ling, A.V.; Devlin, A.S.; Varma, Y.; Fischbach, M.A.; et al. Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **2014**, *505*, 559–563. [CrossRef]
65. Chai, E.Z.P.; Siveen, K.S.; Shanmugam, M.K.; Arfuso, F.; Sethi, G. Analysis of the intricate relationship between chronic inflammation and cancer. *Biochem. J.* **2015**, *468*, 1–15. [CrossRef] [PubMed]
66. Hung, R.J.; Ulrich, C.M.; Goode, E.L.; Brhane, Y.; Muir, K.; Chan, A.T.; Le Marchand, L.; Schildkraut, J.; Witte, J.S.; Eeles, R.; et al. Cross Cancer Genomic Investigation of Inflammation Pathway for Five Common Cancers: Lung, Ovary, Prostate, Breast, and Colorectal Cancer. *JNCI J. Natl. Cancer Inst.* **2015**, *107*, djv246. [CrossRef]
67. Pandey, H.; Tang, D.W.T.; Wong, S.H.; Lal, D. Gut Microbiota in Colorectal Cancer: Biological Role and Therapeutic Opportunities. *Cancers* **2023**, *15*, 866. [CrossRef]
68. Fedirko, V.; Tramacere, I.; Bagnardi, V.; Rota, M.; Scotti, L.; Islami, F.; Negri, E.; Straif, K.; Romieu, I.; La Vecchia, C.; et al. Alcohol drinking and colorectal cancer risk: An overall and dose–response meta-analysis of published studies. *Ann. Oncol.* **2011**, *22*, 1958–1972. [CrossRef] [PubMed]
69. Little, C.H.; Combet, E.; McMillan, D.C.; Horgan, P.G.; Roxburgh, C.S.D. The role of dietary polyphenols in the moderation of the inflammatory response in early stage colorectal cancer. *Crit. Rev. Food Sci. Nutr.* **2017**, *57*, 2310–2320. [CrossRef]
70. Shivappa, N.; Zucchetto, A.; Montella, M.; Serraino, D.; Steck, S.E.; La Vecchia, C.; Hébert, J.R. Inflammatory potential of diet and risk of colorectal cancer: A case–control study from Italy. *Br. J. Nutr.* **2015**, *114*, 152–158. [CrossRef]
71. Tojjari, A.; Choucair, K.; Sadeghipour, A.; Saeed, A.; Saeed, A. Anti-Inflammatory and Immune Properties of Polyunsaturated Fatty Acids (PUFAs) and Their Impact on Colorectal Cancer (CRC) Prevention and Treatment. *Cancers* **2023**, *15*, 4294. [CrossRef]
72. Thanikachalam, K.; Khan, G. Colorectal Cancer and Nutrition. *Nutrients* **2019**, *11*, 164. [CrossRef]
73. Rohrhofer, J.; Zwirzitz, B.; Selberherr, E.; Untersmayr, E. The Impact of Dietary Sphingolipids on Intestinal Microbiota and Gastrointestinal Immune Homeostasis. *Front. Immunol.* **2021**, *12*, 635704. [CrossRef] [PubMed]
74. Ersöz, N.Ş.; Adan, A. Cytotoxic Effects of Resveratrol and Its Combinations with Ceramide Metabolism Inhibitors on FLT3 Positive Acute Myeloid Leukemia. *Erzincan Üniversitesi Fen Bilim. Enstitüsü Derg.* **2020**, *13*, 1205–1216. [CrossRef]
75. Ersöz, N.Ş.; Adan, A. Resveratrol triggers anti-proliferative and apoptotic effects in FLT3-ITD-positive acute myeloid leukemia cells via inhibiting ceramide catabolism enzymes. *Med Oncol.* **2022**, *39*, 35. [CrossRef]
76. Ersöz, N.Ş.; Adan, A. Resveratrol Targets Sphingolipid Metabolism to Induce Growth Inhibition in FLT3 ITD Acute Myeloid Leukemia. *Proceedings* **2019**, *40*, 4. [CrossRef]
77. Ersöz, N.Ş.; Adan, A. Differential in vitro anti-leukemic activity of resveratrol combined with serine palmitoyltransferase inhibitor myriocin in FMS-like tyrosine kinase 3-internal tandem duplication (FLT3-ITD) carrying AML cells. *Cytotechnology* **2022**, *74*, 271–281. [CrossRef]
78. Johnson, E.L.; Heaver, S.L.; Waters, J.L.; Kim, B.I.; Bretin, A.; Goodman, A.L.; Gewirtz, A.T.; Worgall, T.S.; Ley, R.E. Sphingolipids produced by gut bacteria enter host metabolic pathways impacting ceramide levels. *Nat. Commun.* **2020**, *11*, 2471. [CrossRef]

79. Gevers, D.; Kugathasan, S.; Denson, L.A.; Vázquez-Baeza, Y.; Van Treuren, W.; Ren, B.; Schwager, E.; Knights, D.; Song, S.J.; Yassour, M.; et al. The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe* **2014**, *15*, 382–392. [CrossRef] [PubMed]
80. Bryan, P.-F.; Karla, C.; Edgar Alejandro, M.-T.; Sara Elva, E.-P.; Gemma, F.; Luz, C. Sphingolipids as Mediators in the Crosstalk between Microbiota and Intestinal Cells: Implications for Inflammatory Bowel Disease. *Mediat. Inflamm.* **2016**, *2016*, 9890141. [CrossRef]
81. Zhou, Y.; Zhi, F. Lower Level of *Bacteroides* in the Gut Microbiota Is Associated with Inflammatory Bowel Disease: A Meta-Analysis. *BioMed Res. Int.* **2016**, *2016*, 5828959. [CrossRef]
82. Brown, E.M.; Ke, X.; Hitchcock, D.; Jeanfavre, S.; Avila-Pacheco, J.; Nakata, T.; Arthur, T.D.; Fornelos, N.; Heim, C.; Franzosa, E.A.; et al. Bacteroides-Derived Sphingolipids Are Critical for Maintaining Intestinal Homeostasis and Symbiosis. *Cell Host Microbe* **2019**, *25*, 668–680.e7. [CrossRef]
83. Lee-Sarwar, K.; Kelly, R.S.; Lasky-Su, J.; Moody, D.B.; Mola, A.R.; Cheng, T.-Y.; Comstock, L.E.; Zeiger, R.S.; O'Connor, G.T.; Sandel, M.T.; et al. Intestinal microbial-derived sphingolipids are inversely associated with childhood food allergy. *J. Allergy Clin. Immunol.* **2018**, *142*, 335–338.e9. [CrossRef] [PubMed]
84. Wlodarska, M.; Kostic, A.D.; Xavier, R.J. An integrative view of microbiome-host interactions in inflammatory bowel diseases. *Cell Host Microbe* **2015**, *17*, 577–591. [CrossRef] [PubMed]
85. Sano, R.; Trindade, V.M.; Tessitore, A.; D'Azzo, A.; Vieira, M.B.; Giugliani, R.; Coelho, J.C. GM1-ganglioside degradation and biosynthesis in human and murine GM1-gangliosidosis. *Clin. Chim. Acta* **2005**, *354*, 131–139. [CrossRef]
86. Kytzia, H.; Hinrichs, U.; Maire, I.; Suzuki, K.; Sandhoff, K. Variant of GM2-gangliosidosis with hexosaminidase A having a severely changed substrate specificity. *EMBO J.* **1983**, *2*, 1201–1205. [CrossRef] [PubMed]
87. Kolter, T.; Sandhoff, K. Sphingolipid metabolism diseases. *Biochim. Biophys. Acta (BBA)-Biomembr.* **2006**, *1758*, 2057–2079. [CrossRef]
88. Jmoudiak, M.; Futerman, A.H. Gaucher disease: Pathological mechanisms and modern management. *Br. J. Haematol.* **2005**, *129*, 178–188. [CrossRef]
89. Zhang, L.; Liu, C.; Jiang, Q.; Yin, Y. Butyrate in Energy Metabolism: There Is Still More to Learn. *Trends Endocrinol. Metab.* **2021**, *32*, 159–169. [CrossRef]
90. Geuking, M.B.; Köller, Y.; Rupp, S.; McCoy, K.D. The interplay between the gut microbiota and the immune system. *Gut Microbes* **2014**, *5*, 411–418. [CrossRef]
91. Chung, H.; Kasper, D.L. Microbiota-stimulated immune mechanisms to maintain gut homeostasis. *Curr. Opin. Immunol.* **2010**, *22*, 455–460. [CrossRef]
92. Krishnan, S.; Alden, N.; Lee, K. Pathways and functions of gut microbiota metabolism impacting host physiology. *Curr. Opin. Biotechnol.* **2015**, *36*, 137–145. [CrossRef]
93. Zhang, Y.-J.; Li, S.; Gan, R.-Y.; Zhou, T.; Xu, D.-P.; Li, H.-B. Impacts of gut bacteria on human health and diseases. *Int. J. Mol. Sci.* **2015**, *16*, 7493–7519. [CrossRef] [PubMed]
94. Serino, M.; Blasco-Baque, V.; Nicolas, S.; Burcelin, R. Far from the eyes, close to the heart: Dysbiosis of gut microbiota and cardiovascular consequences. *Curr. Cardiol. Rep.* **2014**, *16*, 540. [CrossRef] [PubMed]
95. Kim, Y.-G.; Udayanga, K.G.S.; Totsuka, N.; Weinberg, J.B.; Núñez, G.; Shibuya, A. Gut dysbiosis promotes M2 macrophage polarization and allergic airway inflammation via fungi-induced PGE_2. *Cell Host Microbe* **2014**, *15*, 95–102. [CrossRef]
96. Yang, W.; Cong, Y. Gut microbiota-derived metabolites in the regulation of host immune responses and immune-related inflammatory diseases. *Cell. Mol. Immunol.* **2021**, *18*, 866–877. [CrossRef]
97. Wang, X.; Fang, Y.; Liang, W.; Cai, Y.; Wong, C.C.; Wang, J.; Wang, N.; Lau, H.C.-H.; Jiao, Y.; Zhou, X.; et al. Gut–liver translocation of pathogen Klebsiella pneumoniae promotes hepatocellular carcinoma in mice. *Nat. Microbiol.* **2025**, *10*, 169–184. [CrossRef] [PubMed]
98. Fantini, M.C.; Guadagni, I. From inflammation to colitis-associated colorectal cancer in inflammatory bowel disease: Pathogenesis and impact of current therapies. *Dig. Liver Dis.* **2021**, *53*, 558–565. [CrossRef]
99. Nagao-Kitamoto, H.; Kitamoto, S.; Kuffa, P.; Kamada, N. Pathogenic role of the gut microbiota in gastrointestinal diseases. *Intest. Res.* **2016**, *14*, 127–138. [CrossRef]
100. Zhao, H.; Wu, L.; Yan, G.; Chen, Y.; Zhou, M.; Wu, Y.; Li, Y. Inflammation and tumor progression: Signaling pathways and targeted intervention. *Signal Transduct. Target. Ther.* **2021**, *6*, 263. [CrossRef]
101. Peloquin, J.M.; Nguyen, D.D. The microbiota and inflammatory bowel disease: Insights from animal models. *Anaerobe* **2013**, *24*, 102–106. [CrossRef]
102. Tomasello, G.; Tralongo, P.; Damiani, P.; Sinagra, E.; Di Trapani, B.; Zeenny, M.N.; Hussein, I.H.; Jurjus, A.; Leone, A. Dismicrobism in inflammatory bowel disease and colorectal cancer: Changes in response of colocytes. *World J. Gastroenterol.* **2014**, *20*, 18121–18130. [CrossRef]

103. Chattopadhyay, I.; Dhar, R.; Pethusamy, K.; Seethy, A.; Srivastava, T.; Sah, R.; Sharma, J.; Karmakar, S. Exploring the Role of Gut Microbiome in Colon Cancer. *Appl. Biochem. Biotechnol.* **2021**, *193*, 1780–1799. [CrossRef] [PubMed]
104. Yu, I.; Wu, R.; Tokumaru, Y.; Terracina, K.P.; Takabe, K. The Role of the Microbiome on the Pathogenesis and Treatment of Colorectal Cancer. *Cancers* **2022**, *14*, 5685. [CrossRef] [PubMed]
105. Rezaee, M.A.; Nouri, R.; Hasani, A.; Shirazi, K.M.; Alivand, M.R.; Sepehri, B.; Sotoodeh, S.; Hemmati, F. Escherichia coli and Colorectal Cancer: Unfolding the Enigmatic Relationship. *Curr. Pharm. Biotechnol.* **2022**, *23*, 1257–1268. [CrossRef]
106. Bonnet, M.; Buc, E.; Sauvanet, P.; Darcha, C.; Dubois, D.; Pereira, B.; Déchelotte, P.; Bonnet, R.; Pezet, D.; Darfeuille-Michaud, A. Colonization of the human gut by *E. coli* and colorectal cancer risk. *Clin. Cancer Res.* **2014**, *20*, 859–867. [CrossRef]
107. Wassenaar, T.M. *E. coli* and colorectal cancer: A complex relationship that deserves a critical mindset. *Crit. Rev. Microbiol.* **2018**, *44*, 619–632. [CrossRef]
108. Mughini-Gras, L.; Schaapveld, M.; Kramers, J.; Mooij, S.; Neefjes-Borst, E.A.; van Pelt, W.; Neefjes, J. Increased colon cancer risk after severe Salmonella infection. *PLoS ONE* **2018**, *13*, e0189721. [CrossRef]
109. Martin, O.C.; Bergonzini, A.; D'Amico, F.; Chen, P.; Shay, J.W.; Dupuy, J.; Svensson, M.; Masucci, M.G.; Frisan, T. Infection with genotoxin-producing *Salmonella enterica* synergises with loss of the tumour suppressor *APC* in promoting genomic instability via the PI3K pathway in colonic epithelial cells. *Cell. Microbiol.* **2019**, *21*, e13099. [CrossRef]
110. Patel, R.K.; Cardeiro, M.; Frankel, L.; Kim, E.; Takabe, K.; Rashid, O.M. Incidence of Colorectal Cancer After Intestinal Infection Due to *Clostridioides difficile*. *World J. Oncol.* **2024**, *15*, 279–286. [CrossRef] [PubMed]
111. Coleman, O.I.; Nunes, T. Role of the Microbiota in Colorectal Cancer: Updates on Microbial Associations and Therapeutic Implications. *BioResearch Open Access* **2016**, *5*, 279–288. [CrossRef]
112. Narayanan, V.; Peppelenbosch, M.P.; Konstantinov, S.R. Human Fecal Microbiome–Based Biomarkers for Colorectal Cancer. *Cancer Prev. Res.* **2014**, *7*, 1108–1111. [CrossRef]
113. Karampatakis, T.; Tsergouli, K.; Behzadi, P. Carbapenem-Resistant *Klebsiella pneumoniae*: Virulence Factors, Molecular Epidemiology and Latest Updates in Treatment Options. *Antibiotics* **2023**, *12*, 234. [CrossRef] [PubMed]
114. Dubois, R.N. Role of inflammation and inflammatory mediators in colorectal cancer. *Trans. Am. Clin. Climatol. Assoc.* **2014**, *125*, 358–372, discussion 372–373. [PubMed]
115. Zhang, Q.; Su, X.; Zhang, C.; Chen, W.; Wang, Y.; Yang, X.; Liu, D.; Zhang, Y.; Yang, R. Klebsiella pneumoniae Induces Inflammatory Bowel Disease Through Caspase-11–Mediated IL18 in the Gut Epithelial Cells. *Cell. Mol. Gastroenterol. Hepatol.* **2022**, *15*, 613–632. [CrossRef] [PubMed]
116. Strakova, N.; Korena, K.; Karpiskova, R. Klebsiella pneumoniae producing bacterial toxin colibactin as a risk of colorectal cancer development—A systematic review. *Toxicon Off. J. Int. Soc. Toxinol.* **2021**, *197*, 126–135. [CrossRef]
117. Chiang, M.-K.; Hsiao, P.-Y.; Liu, Y.-Y.; Tang, H.-L.; Chiou, C.-S.; Lu, M.-C.; Lai, Y.-C. Two ST11 *Klebsiella pneumoniae* strains exacerbate colorectal tumorigenesis in a colitis-associated mouse model. *Gut Microbes* **2021**, *13*, 1980348. [CrossRef]