Advanced Applications of Machine Learning in Bioinformatics

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

M. Sc. Wenhuan Zeng

aus Hongkong, China

Tübingen

2024

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	27.03.2025
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter/-in:	Prof. Dr. Daniel H. Huson
2. Berichterstatter/-in:	Prof. Dr. Nico Pfeifer

Abstract

Machine learning has evolved continuously since its first appearance, driving progress in science and technology while increasingly attracting interdisciplinary attention beyond the scope of traditional computer science. A number of studies have applied machine learning to bioinformatics, obtaining valuable findings and validating the feasibility of using machine learning algorithms to solve biological problems. The difficulties and challenges in using machine learning techniques for biological problem-solving arise from the differences in data and tasks between computer science and biology. Training a sophisticated learning-based model is a systematic task that involves a series of steps. These steps can be broadly categorized into data and task aspects, each influencing the overall performance of the model. Appropriately adjusting the model based on the specific task during training is essential for effectively adapting machine learning algorithms to domain-specific challenges, which motivates us to develop advanced methods tailored to the biological field, allowing machine learning to be more successfully applied to problems that were initially solved using traditional computational methods in bioinformatics. This thesis presents our studies addressing biological problems across three topics: epigenomics, metagenomics, and infectious diseases. Solving problems in each area, we developed advanced machine learning frameworks based on considering corresponding biological characteristics, thereby improving machine learning algorithms for better performance. First, we address DNA methylation status identification within epigenomics by developing two new frameworks inspired by natural language processing techniques. These frameworks implement the detection of DNA methylation sites and provide biological insights through model interpretation. Second, regarding problems in metagenomics, we present a study that predicts the source of microbiome samples among ten different origins by training a sophisticated ensemble model on taxonomic and functional profiles generated by whole-genome shotgun metagenomics sequencing. Finally, we introduce a study on an infectious disease problem in the context of the COVID-19 pandemic, focusing on predicting patients mortality and exploring factors associated

with disease severity. This study consists of multiple models trained on different types of datasets. These models jointly demonstrated the feasibility of predicting patients' status based on their diverse features and provided valuable insights during the early stages of a new infectious disease. In summary, this cumulative thesis assembles studies across multiple topics of biological problems, advancing current machine learning algorithms for more practical applications in bioinformatics.

Kurzfassung

Maschinelles Lernen hat sich seit seiner Entstehung kontinuierlich weiterentwickelt und treibt den Fortschritt in Wissenschaft und Technologie voran, während es zunehmend interdisziplinäre Aufmerksamkeit über den Bereich der traditionellen Informatik hinaus auf sich zieht. Zahlreiche Studien haben maschinelles Lernen in der Bioinformatik verwendet, um wertvolle Erkenntnisse zu gewinnen, was den Nutzen dieser Methoden zur Lösung biologischer Probleme unterstreicht. Die Schwierigkeiten und Herausforderungen bei der Anwendung von Techniken des maschinellen Lernens zur Lösung biologischer Probleme entstehen durch die unterschiedlichen Daten und Fragestellungen in der Informatik und Biologie. Das Training eines anspruchsvollen Modells ist ein systematischer Prozess, der eine Reihe von Schritten umfasst. Diese Schritte lassen sich grob in daten- und aufgabenbezogene Aspekte unterteilen, die die Gesamtleistung des Modells beeinflussen. Eine angemessene Anpassung der Modelle aus dem maschinellen Lernen an die jeweilige Aufgabe während des Trainings ermöglicht die Anwendung von ihnen zur Lösung von Problemen, für die bisher eher traditionelle Methoden der Bioinformatik genutzt wurden. Diese Dissertation stellt unsere Studien vor, die sich mit biologischen Problemen in drei Themenbereichen befassen: Epigenomik, Metagenomik und Infektionskrankheiten. Zur Lösung der Probleme in jedem Bereich haben wir fortschrittliche maschinelle Lernansätze entwickelt, die auf den entsprechenden biologischen Merkmalen basieren, um die Leistung der Algorithmen zu verbessern. Zunächst beschäftigen wir uns mit der Identifizierung des DNA-Methylierungsstatus in der Epigenomik, indem wir zwei Modelle vorstellen, die von Techniken der natürlichen Sprachverarbeitung inspiriert sind. Sie ermöglichen die Erkennung von DNA-Methylierungsstellen, deren Interpretation neue biologische Erkenntnisse ermöglicht. Zweitens präsentieren wir im Bereich der Metagenomik eine Methode, die die Quelle von Mikrobiomproben aus zehn verschiedenen Ursprüngen vorhersagt. Dafür haben wir ein Ensemblemodell auf Grund von taxonomischen und funktionellen Profilen trainiert, die mit Hilfe von vollständiger Sequenzierung der mikrobiellen Genome erzeugt wurden. Schließlich beschreiben wir

eine Studie im Zusammenhang mit der COVID-19-Pandemie, die die Sterblichkeitsrate von Patienten vorhersagt und die Faktoren untersucht, die mit der Schwere des Verlaufs der Krankheit zusammenhängen. Diese Studie umfasst mehrere Modelle, die auf unterschiedlichen Datensätzen trainiert wurden. Gemeinsam haben diese Modelle die Machbarkeit der Vorhersage des Patientenstatus basierend auf deren vielfältigen Merkmalen demonstriert und wertvolle Einblicke für die Erforschung einer neuen Infektionskrankheit geliefert. Zusammenfassend versammelt diese kumulative Dissertation Studien zu mehreren Themen biologischer Probleme und zeigt, wie aktuelle Algorithmen des maschinellen Lernens für die praktische Anwendung in der Bioinformatik weiterentwickelt werden können.

Acknowledgments

The completion of this dissertation has been a challenging yet rewarding journey, and I am deeply aware that it would not have been possible without the support, guidance, and encouragement of many individuals along the way.

I would like to express my deep gratitude to my supervisor, Prof. Dr. Daniel H. Huson, for the opportunity to join his lab and pursue my doctoral degree, and for his trust and support throughout my doctoral period. With his countless guidance and advice, as well as his patience, kindness, and expertise, he shaped my view of scientific work from the very beginning and the way and attitude for carrying out research. All I have learned from him has helped me to finish this dissertation and will undoubtedly be a lasting treasure in my future career and life.

I am deeply grateful to my TAC members, Prof. Dr. Nico Pfeifer and Prof. Dr. Ruth E. Ley, for the valuable time they spent supporting me in completing my research through their illuminated thought, practical guidance, and generous advice, and for taking the time to review my dissertation. My heartfelt thanks extend to all the professionals with whom I had valuable discussions, especially Prof. Dr. Peter J. Lockhart, Dr. James Marsh, Dr. Nicholas D. Youngblut, Dr. Rupashree Dass, and Dr. Cathy Westhues.

I am greatly thankful to Prof. Dr. Georg K. Gerber for hosting me as a visitor in his lab, for his expert advice and belief in my ability, and for the opportunity to be involved in multiple projects. The countless discussions we had during visiting significantly enhanced my understanding of relevant research. I would also like to thank the members of his lab, especially Dr. Christine Tataru and Dr. Jiening Zhu, for their generous help and the enriching conversations we shared.

I would like to extend my thanks to all of my collaborators, especially those who contributed to the research involved in this dissertation. Their expert insights and dedicated efforts were integral to the successful completion of this dissertation. I am fortunate to have worked with such kind colleagues of Huson lab: Bettina Kappler, Anupam Gautam, Dr. Xi Chen, Timo Lucas, Dr. Caner Bağci, Banu Çentikaya, Dr. Sascha Patz, and Dr.

Monika Zeller. I am especially grateful to those who have given me endless support and companionship during my time in Tübingen. The joyful moments we shared helped me through the darkness and created unforgettable memories.

Last but not least, I would like to express my deep gratitude to my family, Zhenbing Zeng, Xiaolan Dong, and Wenjiao Zeng, for their endless love, unconditional dedication, unwavering trust, and constant encouragement. All of these have been my greatest source of strength, helping me overcome challenges along the way, complete this journey, and become the person I am today.

Contents

1 Introduction

1.1 Bioinformatics and Its Applications in the Big Data Era

Advances in sequencing technology have made high-throughput analysis of biological systems possible, leading to the generation and processing of large amounts of genomic data in a cost-effective and time-saving manner [1, 2]. In addition, the accumulation of significant amounts of diverse biomedical data, including omics, image, and signal data [3], ushered in the biological "big data" era. While the accessibility of big data provides unprecedented opportunities, it also presents challenges, yielding the need for advanced methods and tools for mining, storing, and analyzing data. Adequately taking advantage of large amounts of biological data requires expertise across diverse areas.

Therefore, as an interdisciplinary science that integrates multiple disciplines to develop methods and software tools for understanding biological data, bioinformatics particularly plays a significant role in the big data era [4]. The term "Bioinformatics" was first proposed in the mid-1980s in order to describe the application of information science to biology [5], a practical subject used to enhance our understanding of biological systems and life. Bioinformatics derived expertise from diverse disciplines, including biology, computer science, and mathematics and is dedicated to the analysis and interpretation of biological data, particularly genomic and molecular data.

It encompasses a broad range of applications to life science, including but not limited to, genomic sequencing, protein structure prediction, evolutionary biology, drug discovery, and personalized medicine [5, 6]. Since the concept of bioinformatics was first proposed, the field has continually evolved alongside the growing availability of data and advancements in related disciplines. Unsurprisingly, machine learning techniques are increasingly being applied to bioinformatics research [3] and have shown comparable or even better performance to the traditional computational approaches. Additionally, they

are capable of illuminating the complex relationships hidden in large-scale biological and biomedical data [7].

1.2 Achievements and Challenges in Domain-Specific Machine Learning

The origins of the machine learning concept and term date back to the 1950s [8], with continuous innovation and optimization of machine learning algorithms, the popularization of computer science, and the enhancement of computing resources, machine learning approaches are gradually promoted to broad application scenarios, not only limited to the branch of pure computer science but also combined with different disciplines to solve domain-specific problems. Machine learning is applied across a broad range of disciplines, spanning natural science, including mathematics [9], chemistry [10], biology [11], social sciences, such as economics [12], linguistics [13], and psychology [14], and human science, including law [15] and philosophy [16].

Consensus holds that bioinformatics is the interdisciplinary field of science that develops methodologies to collect, store, and understand biological data and information using computational tools and techniques. Accordingly, it can be applied to various biological topics to aid and overcome the present hurdles in each topic and uncover hidden biological messages. Machine learning has been introduced into numerous application areas of bioinformatics to solve corresponding problems [11].

For example, the phenotype prediction [17], taxonomy classification [18], and binning tools development [19] in metagenomics, methylation status identification [20] and gene mutation classification [21] in epigenomics, protein structure prediction [22] and protein structure modeling [23] in structural biology. Machine learning methods are also used in the medical profession to better diagnose specific diseases by applying computational tools on medical imaging such as X-rays and magnetic resonance imaging (MRI) scans [24], to identify high-risk patients [25], and to aid in disease outbreak forecasting [26]. Machine learning techniques, with their predictive and generative capabilities, along with powerful computational efficiency, perform as a computational alternative to traditional wet-lab methods by optimizing time and cost efficiency, handling complicated data tasks, and providing new biological insights.

The term machine learning contains the traditional machine learning approach, which

performs inference and decision-making by learning from known data based on statistical principles, and the deep learning approach, which is inspired by the structure and mechanism of the human brain and uses neural networks to handle the data with more complex patterns and relationships. These two approaches are well integrated into the biological sciences, with their respective characteristics.

Deep learning models are adapted to diverse learning tasks, including supervised learning, non-supervised learning, and semi-supervised learning, especially showing incomparable performance on tasks containing large and high-quality datasets with their complicated but flexible network structure. Nevertheless, the outstanding performance of learning-based models depends heavily on the quantity and quality of data, and traditional machine learning models are more suitable in situations where the data quality remains suboptimal even after data engineering, such as data augmentation and feature selection.

Deciding the most appropriate model from existing types and structures based on the specific task is an unneglected step while applying machine learning to a specific domain. Still, current approaches usually overlook the differences between the data conditions and task characteristics across different fields, making it insufficient to maximize the efficacy of applying machine learning methods in biological scenarios. Many studies have demonstrated that tailoring algorithms specifically to the task or importing transfer learning is vital to adapting machine learning to a specific domain [27, 28, 29]. At the same time, researchers are continuously devoted to exploring and developing a better solution for applying machine learning to biological scenarios.

1.3 Methods for Incorporating Machine Learning into Bioinformatics

With its ability to leverage knowledge gained from a well-studied task to improve performance on another task [30], transfer learning can address problems while applying machine learning methods in bioinformatics caused by the unique characteristics of biological data [11, 3], such as insufficient data volume and severely skewed data distribution. Previous studies validated that transfer learning can contribute to biological science, including illuminating the functional role of microbial dark matter by employing domain knowledge of multi-dimensional features underlying microbial diversity, as well as being

applied to microbial community classification in diverse microbiome contexts [31, 32, 33]. Additionally, they also implement early diagnosis of different types of cancer using transfer learning based on gene-expression data or cancer histopathology and real-time images [34, 35].

The essence of transfer learning methods is to use an existing pre-trained model or custom a pre-trained model from scratch to address a new problem by continuing to train the pre-trained model with the necessary layers frozen or fine-tuning the pre-trained model on the dataset of the new problem and using the pre-trained models as feature extractors of newer models. It's worth mentioning that transfer learning is a technology applied to a selected model, whether it is built with the convolutional neural network (CNN) [36, 37], the graph neural network (GNN) [38] or the Transformer [39].

Transfer learning is widely used in the natural language processing (NLP) field, which is initially for comprehending and analyzing human language in text and audio formats, especially for those tasks implemented by transformer-based language models. With the ongoing development and accumulation of impressive achievements in pre-trained and fine-tuned paradigms across various NLP tasks [40, 41, 42, 43, 44], the application of NLP technology is gradually extending from human language to biological sequences, including DNA, RNA, and protein sequences.

This transition can effectively decipher biological sequences, enhancing our comprehensive and transferable understanding of genomic DNA sequences, and enabling the identification of methylation status, enhancers, and promoters within DNA sequences [45, 46, 47, 48]. The pre-trained and fine-tuned paradigms in RNA sequences are devoted to methylation status identification and cell type classification [49, 50]. Moreover, the model using the pre-trained and fine-tuned paradigms can decode the information hidden in protein sequences and apply it to multiple corresponding down-stream tasks by capturing local and global representations of proteins, showing promising performance in learning protein sequence structure, exploring protein-protein interaction, and aiding in protein engineering [51, 52, 53, 54, 55].

Beyond combining with the models that are applied for data in sequence format, transfer learning is equally important for the other models involved in bioinformatics scenarios. Choosing a model that takes the data belonging to matrix structure as input is a common approach in biological science since it suits biological images, count tables, and one-dimensional vectors. On the one hand, the pre-trained CNN-based neural networks are used as feature extractors to enhance biomedical image classification for human-host

disease [56]. On the other hand, these models can also be fine-tuned for specific domains using the corresponding domain-specific dataset. This leads to good performance in many subsequent tasks, such as identifying plant diseases and human-host diseases, by inheriting some feature weight from a model that has already been trained.[57, 58, 59, 60].

Some biological data and corresponding mechanisms can be encoded within the topological graph structure and subsequently utilized in graph-based neural networks. Adopting transfer learning rather than training graph-based neural networks from scratch, enables these networks to more accurately depict and capture the dynamics and structure of biological networks. This strategy also uncovers the correlations and interactions both within and between individuals, facilitating studies such as the discovery of protein-protein interactions, drug screening implementations, and enhancing the understanding of the biological process at the single cell level from cell types annotation in spatial transcriptomics [61, 62, 63].

Overall, transfer learning can effectively solve challenging tasks by leveraging the knowledge gained from pre-trained models and applying it to downstream tasks. This approach mitigates the difficulties caused by limited data or the rarity of a specific task when applying any neural network, regardless of network type or architecture. Besides the strategies mentioned above, many studies customize frameworks according to the particular task from aspects including feature engineering, model architecture, and the training process. These approaches address the challenges in applying machine learning technology to biological science.

Among them, customizing frameworks based on ensemble learning is a popular approach to improving model performance. This can be achieved by either increasing the number of models, in which they are trained on the same datasets [64] or different datasets of the same samples [65], or by increasing model complexity to capture as much relevant information as possible [66, 67]. Promising progress has been made in the topic of sequence analysis [68, 69], genome analysis [70, 71], and disease study [72, 73] in bioinformatics. The various ensemble strategies have evolved over time, leading to better incorporation of the learning models, where significant techniques encompass stacking, voting, and averaging ensemble learning [67].

The averaging strategy generates the final output by computing the weighted or unweighted average of the outputs of multiple models trained regarding the same task. In this strategy, seeing each model as a learner capturing data's unique characteristic, the

averaging of the base learners is carried out either on the outputs of the base learners directly or on the predicted probabilities of the classes computed by a softmax or sigmoid function. The framework integrated learners' ability has shown advantages while handling pair-end metagenomics reads in taxonomic classification [18] and pair-end nucleotide sequences in transcription factor binding identification of specific cell type [74].

In addition to the averaging strategy, stacking and majority voting methods are other ways to integrate multiple learners following respective rules to enhance overall performance in biological scenarios. Instead of taking the average of all learners' output, the final output of the majority voting is determined by basic learners' votes, either taking the label with the most learners' predictions [69, 75, 76] or summing learners' predicted probability and taking the one with a higher sum [77, 78, 79].

Unlike the strategies mentioned above, the stacking method ensembles multiple learners by training a meta-learner to output the final prediction, which takes the concatenation of each learner's output as input. The meta-learner attempts to learn how to combine the input predictions optimally to make a better output prediction. This method especially works well in the tasks involved in a high-dimensionality dataset, shows efficiency in disease prediction in human microbiome study [65, 77], biomarker discovery in microbiome [80], simultaneous or separate prediction of virulence factors and antibiotic resistance genes in microbial data [81, 82].

1.4 Machine Learning Interpretability Uncovers Biological Insights

In order to better adapt machine learning technology to biological sciences, optimizing model architecture is not the only task we need to complete. Learning-based models, especially those based on deep learning algorithms known as black-box models, mysteriously accomplished assigned tasks, whether classification and regression tasks in supervised learning or clustering and association analysis in unsupervised learning [83]. Interpreting the model helps us have a deeper understanding of the model's mechanism and the input data, no matter the models are employed for the original computer science field or biological science [84, 85]. Biological science involves significant sequence data like DNA, RNA, and protein sequences, and tabular datasets like operational taxonomic unit (OTU) tables. Unlike biological images, which provide more intuitive information

for models to capture, and they are easier to validate the interpretation obtained from models. Thereby, it's critical to import model interpretation approaches while applying machine learning technology to biological science.

Approaches to interpretable machine learning can be broadly grouped into post-hoc and intrinsic interpretation methods, each of which can be further classified into global and local interpretability [86, 87], where the global interpretability focuses on overall model behavior, and another aim to find out the reasons for sample's prediction made by model [88]. In general, the post-hoc interpretation method solves the black-box problem after a model is learned. That is, this method conducts an explanation on the trained model [89, 90].

SHapely Additive exPlanations (SHAP) [91] is a popular post-hoc interpretation method based on game theory. It converts the question of how to divide a prize so that every player gets a fair share based on their contribution to how to measure each feature's contribution as accurately as possible. SHAP offers model-specific explainers based on different algorithms to explain the output of any machine learning model, including tree-based models, linear models, and neural networks [92], providing reliable insights regarding both global and local interpretability. Researchers have found that the SHAP can help them figure out the hidden meaning and message in biological data. Resulting in a better understanding of the data or better model performance in a broad range of biological-related scenarios [93, 94, 95], including microbiome studies [96, 97, 98, 99], and joint studies like looking at both metagenomics and metabolomics [100, 101, 102].

On the other hand, Local Interpretable Model-Agnostic Explanations (LIME) [103] is a local explanation technique computing the features' contributions to each sample's prediction based on the trained model. For its applications in bioinformatics, LIME provides local explanations and analyses key factors that lead to model predictions for each sample for many studies, especially for disease studies based on metagenomics next-generation sequencing (mNGS) dataset [98], gene expression dataset [104], DNA methylation profiles [105]. SHAP and LIME obviously have advantages in unboxing machine learning models and analyzing features' importance regarding models and predictions [106]. However, they are both computationally expensive. In comparison, LIME has better time efficiency in explaining the prediction of a single sample.

Other than the post-hoc interpretation method, certain model architectures, often in traditional machine learning algorithms, are considered inherently interpretable [86]. The interpretability of these models benefits from their structures. Intrinsic interpretabil-

ity of traditional machine learning approaches represented by Bayesian models and tree-based models, including decision trees and random forests. Bayesian models offer a paradigm for interpretation based on probability theory [107], while tree models track and explain the decision through the tree by the contributions added at each decision node. These algorithms are gradually introduced to deep neural networks due to their intrinsic interpretability. A stochastic neural network combined with Bayesian inference in its training process also has interpretability, like the Bayesian neural network [108]. Same as the tree-based neural network, represented by deep neural decision trees [109].

Moreover, the emergence of the attention mechanism [110] also makes it possible for deep neural networks to be inherently interpretable by extracting the generated attention weights of the attention mechanism in a trained model and using them to explain the correlation between each token (element) in sequence (matrix), as well as perform reasoning upon the model's prediction. Although the interpretability of the attention mechanism and the plausibility of the approach to obtaining interpretation have been continuously controversial [111, 112, 87]. Meanwhile, some research demonstrated that it has the ability to provide an intuitive explanation to a certain extent [113]. Moreover, its interpretation ability also depends on how to define the interpretability [112].

The proposal for the attention mechanism was initially made to address the bottleneck that occurs in capturing the long-distance dependency of long sequences during encoding and decoding in NLP. Since the benefits of adding an attention mechanism to the model structure go far beyond its interpretability. Adding an attention mechanism to the model structure significantly enhances the ability to identify important input elements during the training process, resulting in a better model performance. Bioinformaticians have taken these advantages to expand their application object from tasks involving human language to tasks based on the genomes or metagenomes that are made of DNA [18, 114], RNA sequencing data [115, 116], and protein sequences [51]. This knowledge transfer helps domain-specific models make more accurate predictions and understand the contribution of features in reasoning the predictions [117, 7].

Naturally, employing an attention mechanism is not the only strategy for implementing intrinsic interpretation in deep learning. There are also model-specific approaches for interpreting deep neural networks. For instance, the gradient-based saliency map is widely employed on images to highlight features important for gaining insights into the decision-making of a neural network. This approach was originally developed for computer vision, and has gradually been transferred to biological science [118, 85], including

tasks based on images and tabular datasets that apply CNN as their neural network.

1.5 Model Optimization Using Personalized Loss Function

There are some necessary steps for thoroughly training a machine learning model, which can be broadly grouped into feature-wise and model-wise. Concerning model-related training tasks, training a machine learning model contains model construction, model training, and model evaluation. First of all, according to our task, we create a machine-learning model by either using current architectures or customizing model layers in the model construction step. Then, the model training part involves fitting the training data into the constructed model architecture to train a model, as well as optimizing the model to improve model performance by adjusting model parameters and tuning hyperparameters. After that, model evaluation steps evaluate model quality using evaluation metrics to quantify model performance using real labels and predicted values during each training epoch.

Determining a model architecture that fits specific tasks is one of the factors that maximize model performance for training a machine learning model. Additionally, model optimization also plays an important role in deciding model performance. Model optimization includes updating the attributes of the neural network, such as weights, biases, and learning rate, to maximize model efficiency. More specifically, the optimization function, represented by Adam and SGD optimizer [119], is used to update the model's parameters to minimize the calculated error estimated by the loss function during back-propagation [27].

The choice of loss function will influence the computed value used to measure the difference between real labels and predicted labels. This value is then used to guide the optimization process by providing feedback on how well the model fits the data. Consequently, using or designing an appropriate loss function according to the task is important since it helps the model improve its performance and training efficiency [120], which is particularly worth paying attention to while importing machine learning technology into bioinformatics.

There are correspondingly popular loss functions for various task types, and their effectiveness is suitable for the majority of scenarios belonging to that task, as demon-

strated by previous work. Binary and categorical cross-entropy loss functions are commonly used in classification tasks [121, 122, 123, 124], while mean square error (MSE) and mean absolute error (MAE) are essential for regression tasks [125]. During the process of gradually introducing machine learning technology to a particular field, such as bioinformatics, the algorithm for the loss function is developed with the increasing difficulty of the tasks and the varying features and distributions of the data.

Current improvement approaches are mainly based on the classic loss function algorithms [126]. One of the purposes of these improvements is to make the loss function better suited to the imbalanced dataset in universal fields. Examples include the development of the Dice loss [127] and Focal loss [128], both of which improve model performance by modifying loss function algorithms to increase attention to minority samples, thereby enabling models to have promising performance while transferring to bioinformatics tasks from original scenarios [129, 130, 46]. Moreover, sophisticated optimization strategies are designed when the general optimization algorithms of the loss function are unable to improve performance in unique scenarios. For example, extensions of Focal loss algorithms have been introduced to improve biological image segmentation in microscopy images [131, 132], and an extension of Laplacian graph regularization is defined for identifying disease-gene associations on graphs [133].

Previous work has demonstrated the effectiveness and success of personalized loss function in enhancing model performance, especially when the dataset with distinctive domain characteristics has an impact on model training. Nevertheless, in contrast to the major domains utilizing machine learning technology, bioinformatics has seen comparatively few studies that examined optimizing loss function techniques to overcome training bottlenecks, leaving considerable space for improvement.

Overall, even though the development of machine learning technology has received much attention and effort, adapting it to a specific domain is more than simply applying it to the target dataset. When attempting to combine them with bioinformatics, there remains room for improving machine learning algorithms throughout the training process for generating well-trained models. Especially when model performance plateaus during model training and no longer improves with additional training epochs, optimizing and customizing the steps that belong to model optimization can provide a breakthrough.

2 Objectives

The goal of this thesis is to address the challenges raised while applying machine learning to bioinformatics. This thesis encompasses a number of studies that develop and explore state-of-the-art machine learning approaches according to specific tasks of different bioinformatics application scenarios based on machine learning algorithms. It can be divided into three themes, each covering the application of machine learning methods in bioinformatics to a specific biological topic.

Epigenomics is the first application area of machine learning in bioinformatics introduced in this thesis. It presents two novel learning-based frameworks [134, 135] for identifying DNA methylation sites in genomics. DNA methylation and histone modification are two major epigenetic mechanisms, where DNA methylation happens when a methyl group gets added to the DNA molecule without altering the changes in DNA sequence, regulating gene expression. Although recent sequencing technologies can detect DNA methylation sites, computational approaches are more efficient in terms of time and cost. Previous works show that identification by applying machine learning technologies is possible. Our studies enhance the performance of computational approaches in different aspects by leveraging transfer learning. We obtain promising results by treating DNA sequences as human language and developing interpretable and flexible models on the basis of transformer-based language models.

The second topic of this thesis describes the adaptation of machine learning technology to metagenomics. The corresponding study focuses on predicting the source of microbiome samples using a novel ensemble learning framework based on diverse characteristics derived from the samples. It contributes a well-trained classifier and biological insights obtained by analysing the impact of features on the model outcome. Besides, it introduces a new approach to reducing the dimensionality of the biological tabular dataset, relieving the difficulty caused by the large number of features that prevent the model from performing better.

The third topic investigates the application of bioinformatics using machine learning

in infectious diseases. Advanced machine learning techniques contributed to fighting against the COVID-19 pandemic in different aspects. Against this background, we provide a novel computational approach to predict and interpret patient mortality based on multimodal patient data during the pandemic outbreak [136]. Our study demonstrates the viability of using genetic sequences to predict mortality and highlights the importance and efficiency of integrating other features with patients' clinic data in improving model performance.

3 Publications and Contributions

Publications mentioned this dissertation

1. **Zeng, Wenhuan**, Anupam Gautam, and Daniel H. Huson. “MuLan-Methyl—Multiple Transformer-Based Language Models for Accurate DNA Methylation Prediction.” GigaScience 12 (2023): giad054.

   Contributions: **W.Z.** and D.H.H. conceived the project. **W.Z.** collected and processed the dataset for the project, designed and implemented the architecture and algorithms of MuLan-Methyl, and conducted model analysis. A.G. and **W.Z.** designed and implemented the web server of MuLan-Methyl. **W.Z.**, D.H.H., and A.G. contributed to the manuscript.

2. **Zeng, Wenhuan**, and Daniel H. Huson. "Enhanced 5mC-Methylation-Site Recognition in DNA Sequences using Token Classification and a Domain-specific Loss Function" bioRxiv (2025): 2023.06.01.543218.

   Contributions: All the authors designed the concept of this project. **W.Z** collected and processed the database, **W.Z.** conceived and conducted the experiments, **W.Z** analyzed the results. **W.Z.** and D.H.H. discussed, wrote, and reviewed the original version of the manuscript.

3. **Zeng, Wenhuan**, Anupam Gautam, and Daniel H. Huson. "DeepToA: an ensemble deep-learning approach to predicting the theatre of activity of a microbiome." Bioinformatics 38, no. 20 (2022): 4670-4676.

   Contributions: D.H.H., A.G. and **W.Z.** designed the study. **W.Z.** developed the machinelearning approach and carried out model-based analysis. A.G. and **W.Z.** performed the microbiome analysis. D.H.H, **W.Z.** and A.G. and wrote and revised the manuscript. A.G. designed the web server. A.G. and **W.Z.** packaged the software. All authors discussed the results and edited the manuscript.

4. **Zeng, Wenhuan**, Anupam Gautam, and Daniel H. Huson. "On the application of advanced machine learning methods to analyze enhanced, multimodal data from persons infected with COVID-19." Computation 9, no. 1 (2021): 4.

   Contributions: D.H.H. proposed and guided the project. D.H.H., **W.Z.**, and A.G. wrote the manuscript. **W.Z.** and A.G. designed the work, wrote the code for data generation, conducted processing, and carried out the analysis. **W.Z.** carried out the machine learning and deep learning model development and analysis. All authors read and approved the final manuscript.

Publications not mentioned this dissertation

1. **Zeng, Wenhuan**, and Daniel H. Huson. "Leverage the Explainability of Transformer Models to Improve the DNA 5-Methylcytosine Identification (Student Abstract)." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 38. No. 21. 2024.

   Contributions: **W.Z.** and D.H.H. contributed to the conception of the project, **W.Z.** processed and analyzed the dataset, implemented the experiments, and analyzed the results. **W.Z.** and D.H.H. wrote and reviewed the manuscript.

2. Gautam, Anupam, Debaleena Bhowmik, Sayantani Basu, **Wenhuan Zeng**, Abhishake Lahiri, Daniel H. Huson, and Sandip Paul. “Microbiome Metabolome Integration Platform (MMIP): A Web-Based Platform for Microbiome and Metabolome Data Integration and Feature Identification.” Briefings in Bioinformatics 24, no. 6 (2023): bbad325.

   Contributions: A.G. performed the computations, implemented backend algorithms, and participated in data analysis. D.B. performed data selection, computations, and conducted thorough data analysis. S.B. and **W.Z.** implemented machine learning framework. A.L. participated in data analysis. D.H.H. participated in algorithm development. S.P. conceptualized, administered and coordinated the project, and participated in data analysis. A.G., D.B. and S.P. wrote the manuscript with inputs from other authors. All authors approved the manuscript.

3. Gautam, Anupam, **Wenhuan Zeng**, and Daniel H. Huson. “MeganServer: Facilitating Interactive Access to Metagenomic Data on a Server.” Bioinformatics 39,

no. 3 (2023): btad105.

Contributions: D.H.H. conceived and supervised the project. D.H.H. designed and implemented the programme that works on the server. A.G. and **W.Z.** conducted data processing and programme testing. A.G. performed data analysis. All authors edited and approved the manuscript.

4. Gautam, Anupam, **Wenhuan Zeng**, and Daniel H. Huson. "DIAMOND+ MEGAN Microbiome Analysis." In Metagenomic Data Analysis, pp. 107-131. New York, NY: Springer US, 2023.

   Contributions: D.H.H. conceived and supervised the project. D.H.H. contributed to the original draft of the manuscript. A.G. and **W.Z.** conducted data processing and analysis. All authors edited and approved the manuscript.

4 Results and Discussion

The objective of this chapter is to provide an overview of the problem addressed by the publications covered in this dissertation, as well as to summarize their contributions to advancing the better application of machine learning techniques to bioinformatics. As mentioned in Chapter 2, this dissertation contains the progress made in improving the adaptability of machine learning in bioinformatics across three application fields. In the following, we describe each application scenario.

4.1 Machine Learning in the Context of Epigenomics

Epigenomics is the study of epigenetic modifications on a genome-wide scale [137], which are chemical modifications to DNA that can affect gene expression and cellular function without changing the DNA sequence itself. DNA methylation and histone modification are two well-known epigenetic mechanisms [138, 139]. The identification of DNA methylation is critical due to its significance as an epigenetic phenomenon, involving the addition of methyl groups to particular cytosine residues in the DNA sequence, which regulates gene expression and controls the accessibility and organisation of chromatin [140]. Comprehensively and precisely identifying methylation sites has substantial implications, as DNA methylation regulates gene expression, influences cellular differentiation and development, and is associated with various diseases [141]. There are different patterns of methylation modifications, including 4-methylcytosine (4mC), 5-methylcytosine (5mC), and 6-methyladenine (6mA), which are distinguished by the specific bases where the methyl group adheres, that have been found in genomic DNA from diverse species. Of these, 4mC is restricted to prokaryotes and archaea, 5mC is the dominant pattern in eukaryotes, and 6mA is the most prevalent form in prokaryotes [142, 143].

Various sequencing-based methods have already been developed to detect and study

DNA methylation in the genome, based on both second and third-generation sequencing technology [144, 145]. The specific method used to analyze DNA methylation with short-read sequencing techniques depends on the initial treatments of DNA samples. The main processing methods include bisulfite conversion, endonuclease digestion, and affinity enrichment. These methods leverage distinct principles to distinguish methylated and unmethylated DNA. The bisulfite conversion-based sequencing methods, including bisulfite sequencing (BS-Seq), reduced representation bisulfite sequencing (RRBS), and whole genome bisulfite sequencing (WGBS), play a leading role in DNA methylation studies [146]. Unlike the steps for analysing DNA methylation with short-read sequencing techniques, the nanopore sequencer released by the Oxford Nanopore platform performs methylation analyses directly on DNA samples during sequencing, the modified bases are detected by analyzing the differences reflected in squiggles. The PacBio single molecular real-time (SMRT) sequencing methods can also be used to detect DNA modifications on the sequenced data [145].

Despite the many sequencing techniques available for determining DNA methylation, problems remain with cost, coverage differences, and the inability of some approaches to differentiate between different methylation forms. Therefore, computational approaches significantly compensate for the lack of sequencing methods. Bioinformatics methods that use machine learning techniques are equipped with the ability to provide efficient and accurate methylation analysis and solve issues with storing and retrieving large-scale epigenetics datasets. Synergies between sequencing technologies and computational methods are crucial for advancing our understanding of DNA methylation's functional implications in many areas.

4.1.1 Ensemble Transformer-Based Language Models for DNA Methylation Prediction

Although a significant amount of bioinformatics studies have been conducted on identifying methylation sites, more recent efforts have taken advantage of deep learning algorithms. Despite this, there still remains room for improvement in terms of both the accuracy and the scope of current methodologies. As indicated above, there are several different patterns of DNA modifications. However, most of the earlier studies concentrated on detecting a certain pattern of modifications [147, 148, 149]. Previous studies that tried to identify methylation sites in samples from different species only used DNA

sequences to represent the sample when constructing features, not taking into account the unique characteristics of each species [150, 151, 152].

Manuscript 1 introduced a deep learning-based computational framework called MuLan-Methyl[134] (Multiple Transformer-based Language Models for Accurate DNA Methylation Prediction) to enhance the performance of computational DNA methylation analysis (see Figure 4.1A). It implemented DNA methylation prediction for DNA sequences with 41 base pair (bp) length, which are from three different patterns of modification across multiple species. The underlying logic in building MuLan-Methyl was to adapt promising techniques for the NLP task to our scenario by treating DNA sequences like human language. Our method takes five transformer-based language models and uses their average prediction probabilities to generate an output based on the processed DNA sequences and the taxonomic identity of the corresponding genome for each sample.

It is a common practice to employ transformer-based language models to conduct supervised, unsupervised, or semi-supervised learning in various NLP tasks due to the transformer's capacity to capture long-distance dependency in sequences [39]. Training them according to the pre-train and fine-tune training paradigm, which splits the training process into two parts, enables them to demonstrate promising performance. Firstly, the language model is pre-trained through self-supervised learning tasks, such as the Next Sentence Prediction (NSP) task and the Masked Language Modelling (MLM) task, on a corpus containing various linguistic data related to its task, providing the model with a robust understanding of the semantics of textual inputs and contextual information. The subsequent fine-tuning step involves the supervised training of the pre-trained language model to adapt to specific downstream tasks. The pre-train and fine-tune paradigm provides a flexible framework that enables the creation of diverse transformer-based language models depending on variations in network structure and parameters.

Language models may perform differently on the same task since there are variations in pre-training methods and network architectures. MuLan-Methyl consists of five transformer-based language models: BERT, DistilBERT, ALBERT, XLNet, and ELECTRA. Each was trained using the "pre-train and fine-tune" paradigm on the custom corpus containing the processed training dataset that contains 41 bp DNA sequences and taxonomic lineage information downloaded from the National Center for Biotechnology Information (NCBI) database and the Genome Taxonomy Database (GTDB) (see Figure 4.1B). Instead of using the default tokenizer trained on the corpus that only consists of human language, we trained a custom tokenizer for each language model on the same

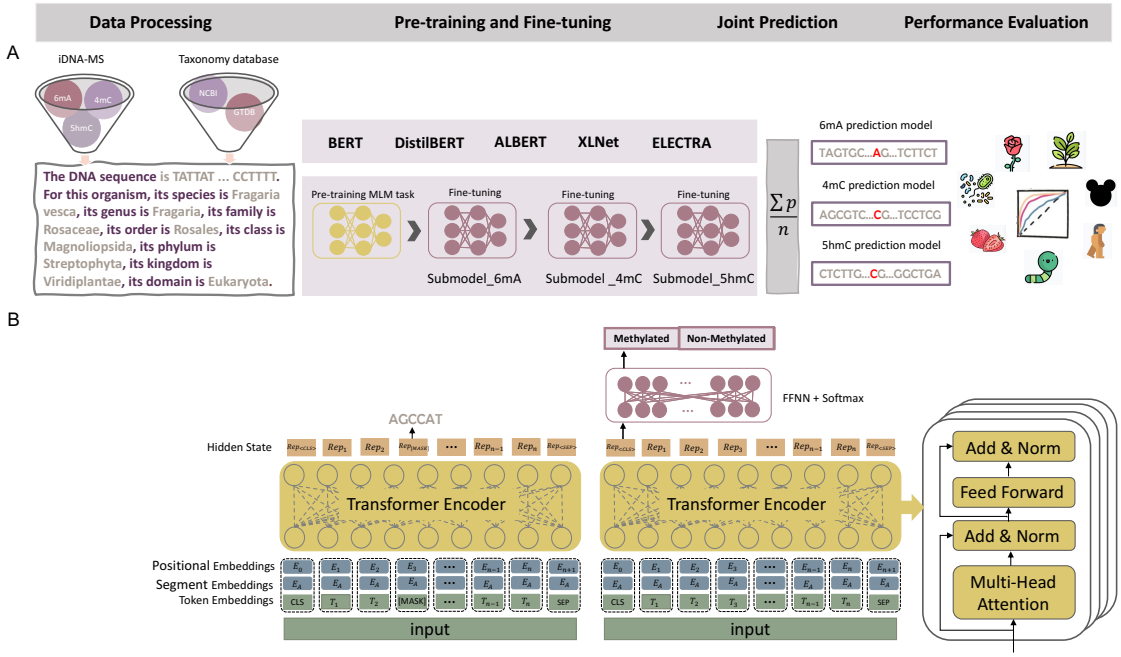


Figure 4.1: The MuLan-Methyl workflow. (A) The framework employs five fine-tuned language models for joint identification of DNA methylation sites. Methylation datasets (obtained from iDNA-MS) were processed as sentences that describe the DNA sequence as well as the taxonomy lineage, giving rise to the processed training dataset and the processed independent set. For each transformer-based language model, a custom tokenizer was trained based on a corpus of the processed training dataset and taxonomy lineage data. Pre-training and fine-tuning were both conducted on each methylation-site specific training subset separately. During model testing, the prediction of a sample in the processed independent test set is defined as the average prediction probability of the five fine-tuned models. We thus obtained three methylation type-wise prediction models. We evaluated the model performance on the genome type contained in the corresponding methylation type-wise dataset, respectively. In total, we evaluated 17 combinations of methylation types and taxonomic lineages. (B) The general transformer-based language model architecture for pre-training and fine-tuning. The model was pre-trained using the masked language modelling (MLM) task and then fine-tuned on the methylation type-wise processed training dataset. The figure is reproduced from Figure 1 of Manuscript 1.

custom corpus, with a parameter configuration identical to its default tokenizer, to accurately tokenize and capture the semantics of DNA sequences and associated taxonomic information in samples.

The pre-training phase involved training each language model on the processed training dataset specific for pre-training according to the corresponding pre-training task. The pre-trained language model was subsequently fine-tuned using the 6mA training dataset to produce the 6mA model. Similarly, the 4mC prediction model was obtained by fine-tuning the 6mA prediction model with the 4mC training dataset. Finally, the 5-hydroxymethylcytosine (5hmC) prediction model was generated by fine-tuning the 4mC prediction model using the 5hmC training dataset. Through this systematic approach, each language model in MuLan-Methyl is expertly customized to capture the hidden message carried by both genome sequences and their taxonomic lineages in epigenomics, improving its predictive ability in identifying DNA methylation sites.

The functionality of MuLan-Methyl is beyond DNA methylation status prediction. It overcomes the barriers in deep learning models caused by the lack of interpretability by capturing the characteristics of the transformer module and maximizing its advantages, providing intrinsic explanations for the features involved in model training, and helping us understand the decision-making process. The multi-head self-attention mechanism of transformer-based language models allows the model to show promising performance in capturing the relationship between elements and hidden messages in sequences, especially in sequences of significant length. Meanwhile, it is the primary factor for MuLan-Methyl equipped with explanability.

To illustrate, one of the five transformer-based language models in MuLan-Methyl, BERT, contains 12 encoder layers containing 12 attention heads each. Given a tokenized input sequence $X = (x_1, x_2, \cdots, x_m)$, the multi-head self-attention in one layer can be described as

$$\text{MultiHead}\,(Q,K,V) = \text{Concat}\,(head_1, \cdots, head_n)\,W^O. \tag{4.1}$$

Here, the i_{th} single attention head is computed as

$$\text{head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right), \tag{4.2}$$

$$\text{Attention}\,(Q,K,V) = \left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{4.3}$$

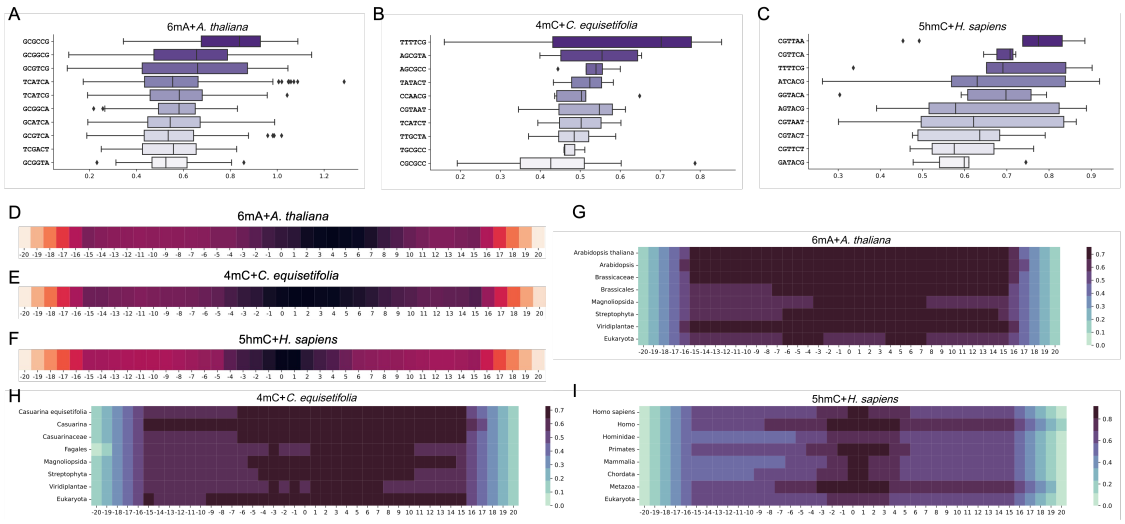


Figure 4.2: Interpretation of MuLan-Methyl by attention weights resulting from transformer self-attention mechanism. In **(A)**–**(C)**, box-plots show the distribution of attention weights for the ten 6-mer of highest average importance scores, for the combinations 6mA + *A. thaliana*, 5mC + *C. euisetifolia* and 5hmC + *H. sapiens*, respectively. In **(D)**–**(F)**, we indicate the importance score for each position in the DNA sequences of length 41, obtained by merging 6-mer fragments, for the same three combinations listed above, respectively. In **(G)**–**(I)**, for each taxonomic rank of a lineage, we indicate the attention weight assigned by MuLan-Methyl to each position of the sequence for generating the taxon of the given rank, for the same three combinations listed above, respectively. The figure is reproduced from Figure 3 of Manuscript 1.

where the projections represent parameter matrices $W_i^Q \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{model} \times d_v}$, and $W^O \in \mathbb{R}^{hd_v \times d_{model}}$. $Attention = \{a_{ij}\}$ is a scoring matrix, in which a_{ij} denotes the attention weight that the *Query* token x_i gets from then *Key* token x_j. This matrix is widely used to represent and explore the binding between tokens. We defined the explainability of MuLan-Methyl by averaging the attention matrix computed by MuLan-Methyl-BERT, DistilBERT, and ELECTRA, since their tokenizer employed the WordPiece algorithm, which allows viewing the attention weights of tokens as contribution scores of independent words.

The explainability of MuLan-Methyl improves the model's trustworthiness and provides us with new insights into the biological aspects of its task (see Figure 4.2). Analyzing and visualizing the contribution of each nucleotide in DNA sequences to outputs demonstrates the reasonability of model construction and the reliability of its outputs.

Scoring and ranking the average attention weights assigned from each token involving 6-mers DNA fragments to token [CLS] in the fine-tuned models discover motifs related to DNA methylation. Furthermore, MuLan-Methyl's explainability validates the value of constructing the input sequence consisting of the DNA sequences and its taxonomic lineages. The interaction between them was quantified by extracting the attention weights assigned from DNA tokens to the tokens that represent the taxonomic lineage. In some combinations of taxonomic lineages and methylation types, the taxonomic lineage token, especially the token for deeper hierarchy, had a stronger impact on the formation of the embedding of the center area of DNA sequences. This shows that taxonomic lineages do help with label decisions.

Compared to state-of-the-art techniques, MuLan-Methyl performed better at accurately detecting DNA methylation sites for 41 bp DNA sequences from known taxonomic lineages, particularly when closely related to the lineages involved in training. Compared to previous biological domain-adaption language models, MuLan-Methyl was the first study trained on multi-modal data containing DNA sequences and text-based taxonomy lineages.

4.1.2 Transferring Named Entity Recognition Techniques to DNA Methylation Sites Identification

DNA methylation at the fifth position of cytosine is the most dominant methylation type in mammals. Cytosine methylation occurs most frequently within CpG dinucleotides. However, non-CpG methylation is also detected in neurons and other cells [153]. In particular, 5mC DNA methylation is essential for controlling gene expression and many other biological processes, including genomic imprinting, X chromosome inactivation, and genome stability in humans [154]. DNA methylation patterns undergo coordinated changes during development, contributing to tissue-specific identities. Understanding 5mC DNA methylation is important because it sheds light on fundamental biological processes, disease mechanisms, and potential therapeutic approaches.

Previous research applying bioinformatics tools to detect DNA methylation mostly addressed whether the middle base of a DNA sequence of a certain length was methylated rather than focusing on the methylation status of a single base within a given DNA sequence. Manuscript 2 introduced a deep learning-based computational framework, MR-DNA (Methylation-site Recognition in DNA) that consists of MuLan-Methyl-

DistilBERT and a conditional random fields (CRF) algorithm. This study bridged the gap between computational approaches and traditional wet-lab approaches in DNA methylation analysis by adapting the algorithms designed for Named Entity Recognition (NER) tasks to epigenetic scenarios.

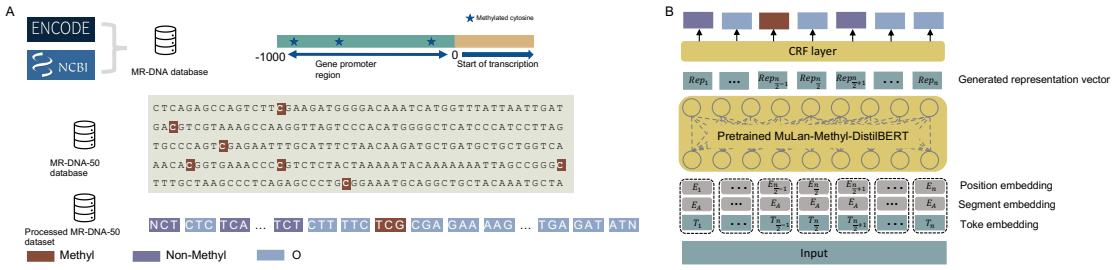


Figure 4.3: A. Database creation. For eight human cell lines in ENCODE, we extracted 1000 bp upstream gene promoter regions, and annotated cytosines by their reported methylation state. We call this the MR-DNA database. The MR-DNA-50 database was extracted from these sequences using a window size of 50 and stride of 25. The *processed* MR-DNA-50 database was obtained by extracting 3-mers from each sequence in the MR-DNA-50 database and labeling each such 3-mer as methylated, if its central nucleotide is so labeled. B. Model structure. MR-DNA is a stacked model that uses a CRF layer on top of pre-trained MuLan-Methyl-DistilBERT, assigning the most probable category to each token. The figure is reproduced from Figure 1 of Manuscript 2.

NER is a subfield of NLP that involves identifying and classifying entities within a given text, such as names of people, organisations, locations, dates, and other specific types of information. The main goal of NER is to extract and classify these named entities to improve the clarity of textual content for machines and humans. This ability is integrated into various NLP applications, including information retrieval, question answering, and text summarization. NER algorithms provide prediction to each token in given sequences. We utilize this advantage for our task by transferring the application object from human language to biological sequences. As a result, the NER technology-built model can predict methylation for specific bases in given sequences. The MuLan-Methyl framework [134] mentioned above consisted of five transformer-based language models, each trained on the iDNA-MS database [155], a comprehensive dataset containing DNA methylation sequences for three methylation types and twelve taxonomic lineages. MR-DNA was a stacked framework consisting of MuLan-Methyl-DistilBERT, one of the five finetuned MuLan models, as the model encoder, along with the CRF layer for assigning tags for each token according to computed probability (see Figure 4.3B). In addition,

in order to better adapt general NER approaches to biological scenarios, we developed a novel loss function based on the categorical cross-entropy loss function, methyl loss, which utilized known biological rules of 5mC methylation to help model training.

General NER tasks are trained on the datasets with text and its corresponding word-level annotations. As the first study to import the NER concept into DNA methylation analysis, we generated a dataset specifically for applying the NER algorithms to identify methylation status. The custom dataset consisted of DNA sequences of length 1000 bp, corresponding to gene promoter regions curated from eight human cell line projects in the ENCODE database, with differing numbers of experimentally verified methylation sites. We further divided each DNA sequence into smaller fragments based on a set stride during model training to improve model performance by increasing the number of samples (see Figure 4.3A). Statistics showed that methylated cytosine counts for a small portion of each record in the processed datasets compared to non-methylated cytosine and the other three nucleotides, leading to an unavoidable imbalance in the distribution of labels.

NLP models typically use the negative log-likelihood or cross-entropy loss function as a measurement of the training process performance based on their adaptability and empirical effectiveness [156] and guide the model to find the optimal parameters during the training process by attempting to minimize the values of the loss function. However, because of the skewed distribution of entity categories, the ability of the model to predict the minority entity is comparatively weak under the default loss function. We presented a novel loss function called "methyl loss" to address this problem, utilizing the biological rule of DNA methylation to improve the model's concentration on indistinguishable samples.

The methyl loss was built upon the categorical cross-entropy (CCE) loss function,

$$\mathcal{L}_{CCE} = -\sum_{i=1}^{N} y_i log(p_i) \tag{4.4}$$

where y_i represents the truth label, p_i is the predicted probability of the i_{th} class. Meanwhile, drawing inspiration from the focal loss, a widely-used variant of cross-entropy for imbalanced data,

$$\mathcal{L}_{Focal} = -\sum_{i=1}^{N} \alpha_i (1-p_i)^{\gamma} \log(p_i), \tag{4.5}$$

where $\gamma > 0$ is used to reduce the relative loss for entities well-classified during model training and to concentrate on challenging entity categories. Here, α_i is a hard hyper-parameter containing a list of weights corresponding to each category. The setting of its variables only depends on the ratio of classes, regardless of the scenario. Methyl loss closes this gap by utilizing the generated annotation vector, for each sample $X = \{x_t\}_{t \in T}$, where T is the number of tokens, and x_t is a token consisting of three elements. The annotation vector $\mathcal{A} = \{a_t\}_{t \in T}$ is defined based on whether a token is in the minority category in terms of the "Methyl" label.

$$a_t = \begin{cases} 0, & \text{if } C \neq x_{t1} \\ 1, & \text{if } C = x_{t1}, \text{ and } G = x_{t2} \\ 2, & \text{if } C = x_{t1}, \text{ and } G \neq x_{t2} \end{cases}$$

The proposed methyl loss function,

$$\mathcal{L}_{Methl} = -\sum_{i=1}^{N} \log(p_i) \lambda^{(1-p_i)\beta_i} \tag{4.6}$$

where β_i indicates whether y_i is identical with a_i, and $\lambda > 0$ controls the attention level on challenging objects, set to 2 in our experiments, aims to pay more attention to the token which belongs to the minority of its category during model training. For example, each token contains three nucleotides, whereas most tokens with the "Methyl" label contain CpG, where C is in the center. We assume that a token is difficult to classify if its nucleotides belong to the minority situation of its category.

The comparison experiments between the model trained using methyl loss and the model trained using focal loss indicated that the methyl loss function is more stable, improving the model performance by considering all categories and paying appropriate attention to the minority label without leading to severe false positive predictions. Our research is the first to attempt to forecast each base's methylation status for a specific DNA sequence. Because of this originality, it is challenging to compare this work to earlier research, which focused on forecasting the global methylation status of DNA sequences of a predetermined length where the target cytosine is located in the center. To fix this, we adjusted the MR-DNA prediction format output to enable comparisons with earlier research employing the iPromoter-5mC database [157] as the benchmark

database. MR-DNA models have shown comparable performance to those trained and tested on the benchmark database.

4.1.3 Discussion

DNA methylation is a significant and complicated epigenetic mechanism regulating gene expression and is tightly related to many biological processes. Accurately identifying DNA methylation patterns holds significant potential for both basic research and clinical applications. It sheds light on the complicated mechanisms regulating gene expression, advancing the fields of genomics, personalized medicine, and understanding disease mechanisms. Even though many bioinformatics tools using machine learning techniques were developed for detecting DNA methylation computationally in attempts to replace the role of wet-lab approaches to achieve better cost- and time-efficiency, there still remains some room for improvement in model performance.

This chapter introduced the studies carried out in Manuscript 1 and Manuscript 2. Both studies proposed novel bioinformatics tools, namely MuLan-Methyl and MR-DNA, respectively, enhancing the computational approaches for DNA methylation detection from different aspects.

MuLan-Methyl was an ensemble framework that consisted of five transformer-based language models trained using the "pre-train and fine-tune" paradigm, addressing the challenges of transferring mature technology of language models to scenarios involving DNA sequences instead of text. It was able to predict the methylation status of samples' 41 bp DNA sequences by taking their taxonomic lineages into consideration while building features and was able to improve the performance and training process compared to previous studies. To the best of our knowledge, MuLan-Methyl is the first language-model framework to take taxonomy information into consideration. Its interpretability successfully found DNA methylation motifs, which greatly benefited from the self-attention mechanism of the transformer structure. Additionally, the attention weights assigned to taxonomic lineages by DNA sequences uncovered the relationship between these factors.

MR-DNA was developed to determine the methylation status of each base of DNA sequences, thereby bridging the identification performance gap between computational and wet-lab approaches. Unlike previous methods that predicted the methylation state of DNA sequences of a fixed length, MR-DNA implemented the prediction on every

nucleotide of a sequence of a fixed length benefit from adapting the NER algorithms for DNA methylation analysis. Using a custom loss function, methyl loss, to update the model weight during training improves MR-DNA's classification ability. Optimizing the loss function to focus on challenging data by adding an exponential based on the categorical cross-entropy loss function improves minority category sensitivity. This work demonstrated the feasibility of machine learning techniques in determining a cytosine's methylation state and can be applied to identify methylation sites on larger DNA sequences.

4.2 Machine Learning in the Context of Metagenomics

The microbiome comprises diverse microorganisms residing in and on the human body or in various environments. Improving our knowledge of microbiomes provides valuable insights into the microbial community composition as well as sheds light on ecological dynamics and the significant interplay between microorganisms and their hosts, facilitating a comprehensive understanding of the microbial world and its impact on diverse ecosystems [158]. The rise of DNA sequencing and other genome-enabled technologies has introduced a new phase for studying microbial communities within environments and hosts [159]. In particular, high-throughput sequencing technologies enabled describing the genomes of microorganisms in a wide range of environments, such as the human gut, soil, aquatic ecosystems, and more, making huge strides in shedding light on the complicated relationships between microorganisms and their surroundings. Meanwhile, it avoids the need for microbial cultivation.

Metagenomics and marker-gene approaches are the two main approaches for studying the microbiome using high-throughput sequencing [160]. Marker-gene approaches involve sequencing a gene-specific region to show the diversity and composition of particular taxonomic groups present in an environmental sample. While high-throughput sequencing in metagenomics can characterize the entire diversity of habitat, encompassing viruses, bacteria, eukaryotes, and archaea, as well as uncover its potential and novel functions [161]. In practice, whole-genome shortgun (WGS) metagenomics sequencing can sequence the DNA from every organism present in a given sample, making it a common and culture-independent technique for studying the population structure and function of the microbiome.

The wide variety and heterogeneity inherent in microbial communities raise challenges in metagenomic studies. In addition, the high dimensionality and large scale of sequencing data, along with the presence of rare and unculturable microorganisms, are obstacles to accurate taxonomic classification and functional annotation. Many researchers have developed bioinformatics tools to overcome these hurdles and provide analysis services to study metagenomics from different aspects, including the bioinformatics tools built by machine learning techniques.

4.2.1 Ensemble Deep Learning Models for Classifying the Theatre of Activity of a Microbiome

There are always challenges in collecting high-quality samples, and the unstable quality of resulting metagenomics data has made metagenomics studies difficult. Benefiting from the ability of deep learning algorithms to handle complicated datasets and discover the hidden patterns and dependencies in the dataset, their use is becoming increasingly popular for gaining new insights into the genomes involved in microbial communities.

In microbiome research, it is often part of the research question to predict or track the source of a metagenomic sample [162], identifying the source habitat, whether it be the human gut, soil, oceans, or other ecosystems, is fundamental for understanding the ecological dynamics and functional significance of microbial communities within a variety of environments [163]. Tracking the source of microbial samples with precision allows the uncovering of patterns, biomarkers, and ecological principles that contribute to a deeper understanding of microbial ecology, biodiversity, and the implications for human health and environmental sustainability.

In 1988, the term “microbiome” was first proposed to stand for a combination of the words “micro” and “biome” [164]. When talking about the microbiome, the “theatre of activity” refers to the specific habitat or environment where the characteristic microbial community is actively functioning or carrying out its activities. Essentially, it’s the stage or setting where these microbes interact and thrive within their ecological niche [165]. Manuscript 3 introduced an ensemble deep-learning approach, namely DeepToA, to predict the theatre of activity of microbiome that are sequenced using the WGS sequencing technique. The WGS metagenomic samples used for deepToA were collected from MGnify [158].

The major advantages of WGS sequencing over 16s rRNA amplicon sequencing are

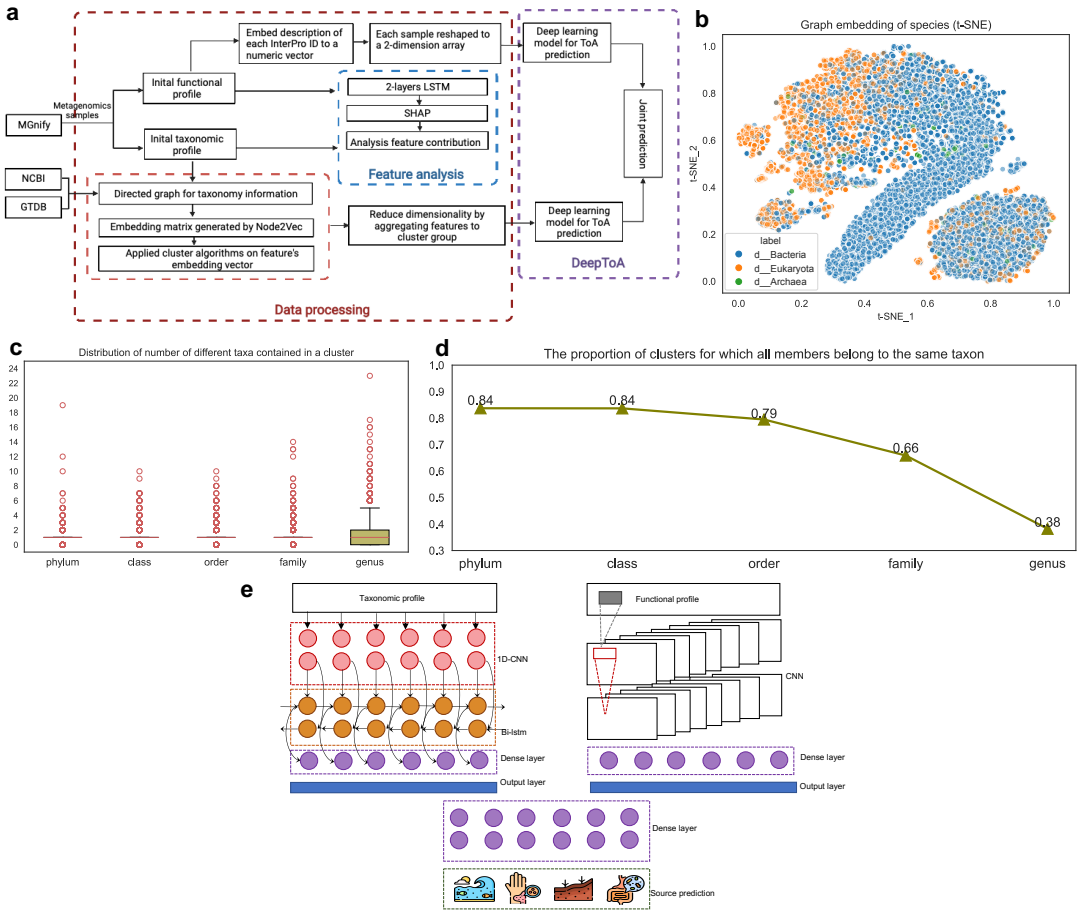


Figure 4.4: (a) Workflow for data collection, feature engineering and model building. Metagenomic samples are downloaded from MGnify, and their taxonomic and functional profiles are extracted. Taxonomic and functional profiles are processed in the lower and upper part of the workflow, respectively. In the middle of the workflow, feature importance is assessed using SHAP values. Joint prediction using DeepToA is performed on the right-hand side. (b) A t-SNE plot of species' embedding vectors, colored by taxonomic domain. (c) The distribution of the number of different taxa that are contained in a cluster, for each taxonomic rank. The mean value is 1 for all ranks. (d) The proportion of clusters that are pure for a given taxonomic rank, that is, for which all members of the cluster belong to the same taxon of that rank. (e) Structure of the DeepToA model, with the taxonomic model left, the functional model right, and the combining and prediction layers at the bottom. The figure is reproduced from Figure 1 of Manuscript 3.

that the taxa can be more accurately detected at the species level and the ability to extract whole genes and specify their function as part of the metagenome. To fully utilize the advantage of the WGS sequencing approach, DeepToA built an advanced framework consisting of two models, capturing the message in terms of taxonomy and function carried by metagenomic samples, respectively. Taxonomic profiles provide information about the composition of microbial communities, while functional profiles record the specific activities and roles these microorganisms play. By combining these two instances, deepToA demonstrated the feasibility and effectiveness of accurately predicting the theatre of activity by considering multiple characteristics of samples at the same time (see Fig 4.4a).

In addition to raw reads in sequence format directly obtained from the sequencing process, metagenomic information can be extracted and converted to diverse data types, including taxonomic profiling abundance tables, functional annotations, and associated metadata. Together, these elements provide a complete picture of the genomic landscape in various microbial ecosystems. Understanding and effectively leveraging these diverse data types is important for advancing our insights into the structure, function, and dynamics of microbial communities across various environments.

The tabular data format is widely used to store metagenomic data. Its processing and application are critical steps in metagenomic studies [166]. The count table in tabular data format can also be seen as a matrix that enumerates the abundance of specific DNA sequences or features across different samples. These count tables perform as a fundamental resource for quantifying the composition and diversity of microbial communities, enabling us to analyze the relative abundance of different taxa or functional elements. However, when metagenomic samples are stored in count tables, it consistently goes along with the trait that their dimensionality is significantly larger than the sample size [167] due to the limitations in collecting high-quality samples and massive biological features. This high-dimensional data phenomenon and its sparsity cause hurdles for training a sophisticated model. Typically, the initial dataset will undergo processing by decreasing its dimensionality, increasing its density, and filling in the missing values.

DeepToA developed a novel dimensionality reduction approach specifically for sequence count tables. Designing the algorithm for optimizing the taxonomic profile is inspired by the hierarchical structure of its features. We used the node2vec [168] on the reference tree that was built by mapping the taxonomic lineages of each feature to the tree structure to get an embedding vector for each taxon. We then calculated Euclidean dis-

tances between the embedding vectors and used the AGNES algorithm to cluster them. The resulting clusters were then used as input to a "taxonomy-based" deep-learning model for predicting the ToA of given samples. Meanwhile, the high-dimensionality problem of the functional profile was solved by expanding the two-dimensional table into a three-dimensional matrix, where the third dimension was obtained by encoding the features' textual description by doc2vec algorithms. The resulting data were further used as input to a "functional-based" deep-learning model for the ToA prediction.

Combining the output layer of the taxonomy-based and function-based models to create the ensemble model in DeepToA allowed for collective decision-making for microbial samples (see Figure 4.4e). Moreover, deepToA bridged the gap between the explainability and deep learning models by employing SHAP values to implement post-hoc interpretation on taxonomy-based and function-based models, respectively. Applying the post-hoc interpretability method enabled the exploration of the host-microbe interactions and the biological roles that particular microbial taxa play and identified key taxa and functions associated with specific environments.

4.2.2 Discussion

The capacity for extracting hidden messages from complicated and large datasets makes machine-learning approaches, whether traditional machine learning or deep learning, widely applicable to metagenomics data. While undergoing metagenomics problems, machine learning algorithms in bioinformatics can replace traditional computational approaches like the probabilistic statistics model to build tools with a better function, and can be used to explore unknown patterns and interactions in metagenomics samples.

Metagenomics data can be stored and displayed in many formats, including tables, sequences, and topological graphs. There are always difficulties in training a sophisticated machine learning model on the metagenomics tabular dataset, including taxonomic profile, functional profile, and gene expression data, due to their significant dimensionality. This problem likewise existed in our initial datasets collected from MGnify to implement ToA prediction.

What sets DeepToA apart is its new dimensionality-reduction algorithms for taxonomic and functional profiles based on the corresponding unique features that relieve the negative effects of their high dimensionality. Moreover, DeepToA was implemented based on WGS metagenomics data instead of the more commonly used 16S community

profile data. This choice provided a more informative dataset, contributed to the thorough training of machine learning models, and prevented the models from overfitting. It also enhanced the complexity and richness of this study, allowing for a more comprehensive understanding of the microbiome's dynamics.

On the other hand, one drawback of deep learning models is their lack of interpretability, which remains a puzzle for researchers. It is also why some researchers choose other approaches, such as Bayesian neural networks, statistical modeling, and tree-ensemble models, to obtain the explainability of model decisions and the importance of features. Here, we employed a post-hoc interpretation approach to explore separately the features' contribution level that is involved in taxonomic and functional profiles, revealing the taxonomic lineages and functions that are tightly associated with microbiomes from specific sources.

In conclusion, DeepToA pushes the boundaries in microbial source prediction by focusing on WGS metagenomic data, considering both the taxonomic and functional aspects of microbial samples for determining their source, and developing new dimensionality reduction methods for concentrating the information contained in each profile and improving model performance. Those innovations open up new avenues for understanding and predicting the intricate dynamics within microbial communities.

4.3 Machine Learning in the Context of Infectious Diseases

Even though medicine and basic research have come a long way in the last few decades, infectious diseases caused by bacteria, eukaryotes, viruses, and other pathogens that can be transmitted are still a problem for scientists and clinicians [169]. These diseases may manifest with distinct symptoms, asymptomatic, or asymptomatic during the early stages of infection. The complicated and multifactorial nature of infectious diseases leads to considerable challenges in accurately predicting patient mortality [170]. On the other hand, patient-specific variables, including age, overall health status, and immune state, significantly impact the development of the disease course, which makes it even harder to make predictive models that work for everyone.

Outbreaks of a novel coronavirus-associated pandemic of acute respiratory infectious disease originating from the SARS-CoV-2 virus, namely COVID-19, have been causing

worldwide health concerns since December 2019 [171, 172]. It once threatened public health and social safety and caused high pressure and crises in medical systems and healthcare services worldwide because of its high transmissibility among humans and pathogenicity [173]. Patients infected with COVID-19 have different degrees of severity, ranging from asymptomatic or mild symptoms to severe illness and death [174, 175], the differences that exist between individuals further add complexity to diagnosis and treatment, leading to a shortage of medical and intensive care resources.

As we stand in 2024, looking back on the COVID-19 pandemic, the factors that caused patients to have different symptoms and corresponding therapy gradually clearly, due to the accumulation of a vast number of cases [176, 177, 178]. However, in the early stages of the COVID-19 pandemic, there was a lack of available prognostic biomarkers to differentiate patient severity and estimate their associated mortality risk. Identifying cases with a high risk of death can help critically ill patients receive timely treatment, discharge low-risk patients, and reduce the burden on the medical system, which became a critical and necessary task at the time [172, 25].

4.3.1 Multimodal Machine Learning in Predicting Mortality of COVID-19

Owing to their capacity for data-driven insights and being independent of physicians' subjective evaluations, machine learning approaches have drawn much attention and become promising methods for participating in risk stratification and clinical decision-making [25]. Models based on machine learning techniques can discover patterns and dependencies between patients and diseases that traditional prognostic methods might miss by harnessing diverse information about patients, including but not limited to demographic data and medical data. Utilizing these advantages contributed to a deep comprehension of the interplay and individual role of features influencing COVID-19 outcomes, providing valuable information into the novel infectious disease at that time.

Manuscript 4 presented an ensemble framework containing two components, each implementing the prediction of COVID-19 patient mortality based on a distinct dataset type. It validated and compared the prediction performance between different combinations of models and datasets (see Figure 4.5). The first component of the ensemble framework contained a multimodal dataset, on which we trained models built by traditional machine learning algorithms and deep learning algorithms, respectively. Another

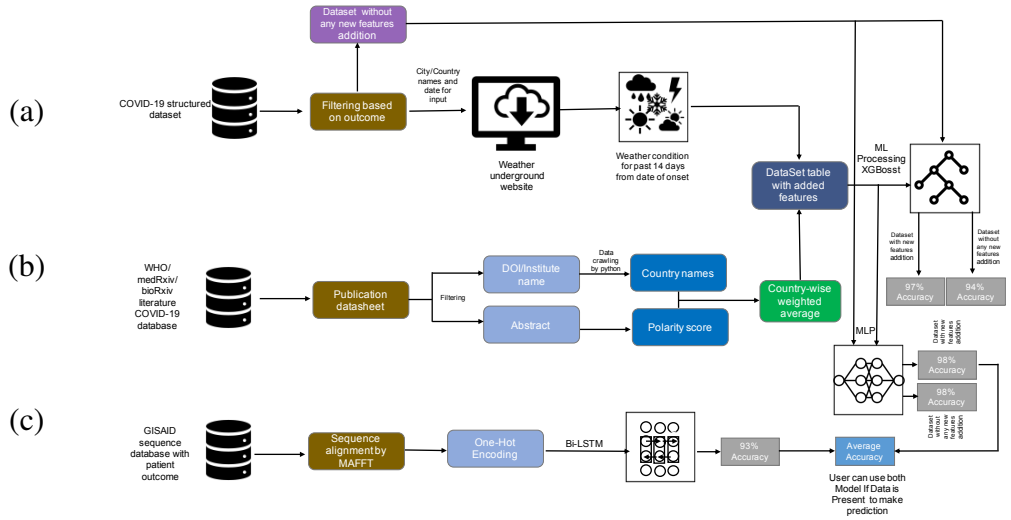


Figure 4.5: **Analysis summary.** (a) The initial COVID-19 structured dataset was filtered for patients for which the outcome has been recorded, and then, for these items, the weather was determined using the Weather Underground website [179]. (b) The WHO, medRxiv and bioRxiv COVID-19 literature database were filtered and pre-processed to extract author institute/address/country, and these were post-processed so as to obtain a country-wise research sentiment polarity score. XGBoost and MLP were trained on both the initial and the enhanced structured data and the accuracy of survival prediction was shown to be 94% and 97% (using XGBoost), and 98% and 98% (using MLP), respectively. (c) Bi-LSTM was used to train a classification model on the sequence dataset, the accuracy was 93%. Finally, the MLP and Bi-LSTM models were stacked to jointly predict outcomes. The figure is reproduced from Figure 1 of Manuscript 4.

contained the virus sequences dataset, which was trained using the Bidirectional Long Short Term Memory (Bi-LSTM) model. Moreover, this study also analyzed the factors that were critical to the patient's disease course, providing valuable insights into the factors that could cause the patients to have more severe symptoms in the initial phase of a pandemic.

Regarding the first component, the multimodal dataset was created based on the assumption that the variance of clinic data among the populations was not the only factor that causes the transmissibility and severity differences of infectious disease. Combining the insights from previous studies [180, 181] collectively, besides collecting the basic information of patients to form a dataset, we added the feature of local weather that was related to the patient in the dataset. Additionally, to explore if the country-wise attitude reflected by academic articles had an impact on the outcomes of patients, we computed the average polarity score of the academic articles for each country that we collected from a couple of professional databases and added it to the dataset as a feature to represent the country-wise researchers optimistic or pessimistic attitude towards COVID-19.

We trained both the Multi-Layer Perception (MLP) model and XGBoost on the multimodal dataset. They resulted in comparable performance on the test dataset. The application of different models provided a comparison of the performance between the traditional machine learning and deep learning models. Leveraging the explainability of the tree structure, XGBoost also quantified features' contributions. The most important feature was the weather textual description, followed by age, local temperature, humidity, and age. Analyzing the contributions of the factors provided insights into exploring the reasons that caused mortality in patients and contributed to the prevention and response strategy for COVID-19.

Another part of this study answered how much the specific genome sequence of the virus that affected patients would determine the outcome for the patient. The dataset for this part contained patients' virus sequences and their vital status, mainly collected from the GISAID sequence repository. We trained a Bi-LSTM model on this sequence dataset, and it resulted in a reliable classifier that was able to distinguish patients' outcomes based on their sequencing data, benefiting from the ability of the Bi-LSTM model to capture long-term dependency among elements in a sample. This part of our framework demonstrated that differences exist in the virus sequences of patients with different outcomes. Deep neural networks can identify these differences and can be used to help diagnose new patients' status.

This ensemble framework involved two components mentioned above, built based on distinct machine learning methodologies, providing novel insight into the question of how machine learning can contribute to infectious diseases from diverse aspects. On the one side, while emphasizing the reliability of the conclusion that traditional clinic data can benefit in predicting patients' status, our study also proposed and demonstrated that mining additional factors, extending beyond the basic patient information to increase sample informativeness, can enhance prediction performance. Moreover, our study demonstrated the virus sequences that were collected from patients also had the potential to distinguish the disease progression of patients. Comprehensive approaches showed enhanced value while fighting against novel, high-risk, and highly transmissible diseases.

4.3.2 Discussion

Manuscript 4 presented the study that applied machine learning to fight against COVID-19, providing insights into the vital factors that caused patients' mortality. It contributed an ensemble framework consisting of multiple models trained on two distinct datasets since the patients we collected data from were different. Each model demonstrated reliable performance in experiments, no matter the MLP, XGBoost that trained on the multimodal dataset, or Bi-LSTM trained on the virus sequence dataset.

The multimodal dataset was constructed innovatively. Besides patients' basic information, it contained both numeric features and text features that represented the local weather corresponding to the patient's location and polarity scores that were computed on region-wise academic articles to represent country-wise attitudes regarding COVID-19. It allowed for predicting the mortality of patients from different aspects, beyond demographic data, clinic data, and biomarkers.

Ideally, we would have liked to further enhance the above-mentioned multimodal dataset by adding virus genome sequences to each sample. However, limited to collection possibility, we chose the implementable plan that generated a sequence dataset from another source, which led to the success of collecting genome sequence data with annotation and let us explore the relationship and influence between genome sequences and patients' outcomes, though the patients were not identical to those we used in the multimodal dataset.

Overall, this study highlights the prediction of patient mortality based on machine

learning techniques by mining possible information from different aspects carried out by patients. It helps us to understand better the factors that relate to disease severity. The diverse features carry more potential for deciphering factors that are closely related to outcomes. This study also validates the feasibility of classifying patient severity according to the data from virus genome sequencing.

5 Conclusion

The application of computational methods in the biological sciences has led to rapid advances in biological research in many aspects, in terms of practical tool development and hidden biological message exploration. By bridging computer science and biology, traditional analysis methods become faster, more cost-effective, and scalable while also providing new biological insights and a deeper understanding of biological systems and their relationships [182, 183]. Bioinformatics is an interdisciplinary subject that encompasses the use of computational methods and tools to analyze biological data. With the emergence and continuous improvement of machine learning algorithms, they have shown impressive importance in different domains, benefiting from their learning and computing capabilities. Machine learning has also been used to adapt to biological scenarios and contributes to novel findings and high-efficiency tools for biological problems [27]. Proven achievements have made the application of machine learning algorithms to bioinformatics to solve wide-ranging biological problems a trend in the big data era.

Among several subtopics in biology research, our research concentrates on bioinformatics approaches in epigenomics, metagenomics, and infectious diseases. The application scenarios of the machine learning algorithms discussed in our paper within these topics are as follows. In epigenomics, machine learning methods are applied to detecting methylation status in biological sequences and discovering methylation motifs [134, 135]. In metagenomics, machine learning models classify the sources of metagenomic samples using both their taxonomic and functional profiles [184]. Lastly, concerning infectious diseases, machine learning models are utilized to predict disease severity in patients infected with COVID-19 and identify key features related to mortality [136]. Given that these three topics usually involve diverse and vast samples and complicated biological mechanisms. This complexity makes machine learning especially well-suited for modeling these problems due to its flexibility of model structure and the adaptability of data structure.

Regardless of the achievements that have been gained in applying machine learning al-

gorithms to these specific topics, applying mature machine learning algorithms properly to specific domains is critical to implementing promising models. There still remains room for optimizing algorithms during adaptation and opportunities for obtaining novel biological insights, because training a sophisticated model is a systematic task that involves both data-wise and model-wise processing.

Feature engineering is a term used to describe and summarize the steps involved in processing datasets, with considerable significance since the quality of the data determines model performance to some extent. After generating or collecting the dataset, methods for processing data, such as missing value filling, data dimensionality reduction, and data augmentation, are employed to improve its quality and enhance its informativeness [185]. There are differences between the mainstream data types involved in biological problems, represented by tabular data and sequence data, with the corresponding data types in the computer science field used to benchmark algorithms. These differences vary the performance or even lead to failure when applying machine learning algorithms to the biology domain, making data engineering for biological data more complicated and vital.

At the same time, model-wise engineering consists of model construction, model training, and model evaluation, each of which is necessary and possible to be adapted according to current problems while applying to bioinformatics. Classic model architectures proposed in the early stages are the fundamentals of advanced models and are still used for tasks in specific fields. They are developed for a corresponding data type and benchmarked using datasets collected from genre topics. For instance, LSTM and transformer-based language models are initially developed for text of long length, compensating for the defects of previous models that could not capture the long-time dependency of long sequences. When applied to genome sequences without any additional processing on either the model architecture or training procedure, their excellent performance is compromised compared to applying them to human language.

Our original intention in conducting a series of studies was to consider the necessity for machine learning algorithms to adapt to biological scenarios. There are still deficiencies in current research in fully considering the differences in datasets from different fields and in leveraging the unique information carried by specific fields to improve machine learning algorithms. Overall, we contributed advanced bioinformatics approaches by optimizing machine learning algorithms to solve problems across different biological topics, including epigenomics, metagenomics, and infectious diseases. The sophisti-

cated models and outcomes presented in this thesis demonstrate the value of advancing machine learning algorithms in alignment with specific domains and specific tasks. For each of the three concepts, we improved the adaptability of machine learning algorithms from multiple perspectives, like data-wise and model-wise approaches, as previously discussed. Altogether, this thesis bridges machine learning and bioinformatics by introducing novel algorithms for data processing and model construction, while also exploring the unique information carried by biological samples to improve model performance.

The unexpected ability and enormous potential of deep learning algorithms, especially NLP approaches, in handling challenges such as sparse data structure and biological problems with complicated rules are further demonstrated as our studies are conducted. We believe there remains room for improvement in obtaining new discoveries of biological significance by further developing machine learning approaches. As we discovered the DNA methylation motifs with the help of the interpretability of transformers, the interpretability of machine learning, especially deep learning algorithms, is widely used in various domains to explain models' decisions, find significant features, and explore the interaction between features and modeling results. We believe the continuous optimization and better application of interpretable machine learning algorithms will promote meaningful findings in biological science.

Bibliography

[1] Bertil Schmidt and Andreas Hildebrandt. “Next-generation sequencing: big data meets high performance computing”. In: *Drug discovery today* 22.4 (2017), pp. 712–717.

[2] Casey S Greene et al. “Big data bioinformatics”. In: *Journal of cellular physiology* 229.12 (2014), pp. 1896–1900.

[3] Seonwoo Min, Byunghan Lee, and Sungroh Yoon. “Deep learning in bioinformatics”. In: *Briefings in bioinformatics* 18.5 (2017), pp. 851–869.

[4] Ardeshir Bayat. “Science, medicine, and the future: Bioinformatics”. In: *BMJ: British Medical Journal* 324.7344 (2002), p. 1018.

[5] Jeremy Ramsden. *Bioinformatics: an introduction*. Springer Nature, 2023.

[6] Jeff Gauthier et al. “A brief history of bioinformatics”. In: *Briefings in bioinformatics* 20.6 (2019), pp. 1981–1996.

[7] Haoyang Li et al. “Modern deep learning in bioinformatics”. In: *Journal of molecular cell biology* 12.11 (2020), pp. 823–827.

[8] A. M. TURING. “I.—COMPUTING MACHINERY AND INTELLIGENCE”. In: *Mind* LIX.236 (Oct. 1950), pp. 433–460. ISSN: 0026-4423. DOI: 10.1093/mind/LIX.236.433. eprint: https://academic.oup.com/mind/article-pdf/LIX/236/433/30123314/lix-236-433.pdf. URL: https://doi.org/10.1093/mind/LIX.236.433.

[9] Kaushik Bhattacharya et al. “Model reduction and neural networks for parametric PDEs”. In: *The SMAI journal of computational mathematics* 7 (2021), pp. 121–157.

[10] Nongnuch Artrith et al. “Best practices in machine learning for chemistry”. In: *Nature chemistry* 13.6 (2021), pp. 505–508.

[11] Nicolae Sapoval et al. “Current progress and open challenges for applying deep learning across the biosciences”. In: *Nature Communications* 13.1 (2022), p. 1728.

[12] Mahinda Mailagaha Kumbure et al. “Machine learning techniques and data for stock market forecasting: A literature review”. In: *Expert Systems with Applications* 197 (2022), p. 116659.

[13] Gabe Dupre. “(What) Can deep learning contribute to theoretical linguistics?” In: *Minds and Machines* 31.4 (2021), pp. 617–635.

[14] Sijia Zhou, Jingping Zhao, and Lulu Zhang. “Application of artificial intelligence on psychological interventions and diagnosis: an overview”. In: *Frontiers in Psychiatry* 13 (2022), p. 811665.

[15] Haoxi Zhong et al. “How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020, pp. 5218–5230.

[16] Michael A Peters et al. “AI and the future of humanity: ChatGPT-4, philosophy and education–Critical responses”. In: *Educational Philosophy and Theory* (2023), pp. 1–35.

[17] Divya Sharma and Wei Xu. “phyLoSTM: a novel deep learning model on disease prediction from longitudinal microbiome data”. In: *Bioinformatics* 37.21 (2021), pp. 3707–3714.

[18] Qiaoxing Liang et al. “DeepMicrobes: taxonomic classification for metagenomics with deep learning”. In: *NAR Genomics and Bioinformatics* 2.1 (2020), lqaa009.

[19] Cédric G Arisdakessian et al. “CoCoNet: an efficient deep learning tool for viral metagenome binning”. In: *Bioinformatics* 37.18 (2021), pp. 2803–2810.

[20] Peng Ni et al. “Genome-wide detection of cytosine methylations in plant from Nanopore data using deep learning”. In: *Nature communications* 12.1 (2021), p. 5976.

[21] Hui Qu et al. “Genetic mutation and biological pathway prediction based on whole slide images in breast carcinoma using deep learning”. In: *NPJ precision oncology* 5.1 (2021), p. 87.

[22] John Jumper et al. “Highly accurate protein structure prediction with AlphaFold”. In: *Nature* 596.7873 (2021), pp. 583–589.

[23] Wenhao Gao et al. “Deep learning in protein structural modeling and design”. In: *Patterns* 1.9 (2020).

[24] Mounir Hamdi et al. “Evaluation of neuro images for the diagnosis of Alzheimer’s disease using deep learning neural network”. In: *Frontiers in Public Health* 10 (2022), p. 834032.

[25] Khadijeh Moulaei et al. “Comparing machine learning algorithms for predicting COVID-19 mortality”. In: *BMC medical informatics and decision making* 22.1 (2022), pp. 1–12.

[26] Nurul Absar et al. “The efficacy of deep learning based LSTM model in forecasting the outbreak of contagious diseases”. In: *Infectious Disease Modelling* 7.1 (2022), pp. 170–183.

[27] Joe G Greener et al. “A guide to machine learning for biologists”. In: *Nature Reviews Molecular Cell Biology* 23.1 (2022), pp. 40–55.

[28] Frank Noé et al. “Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning”. In: *Science* 365.6457 (2019), eaaw1147.

[29] Mark Alber et al. “Integrating machine learning and multiscale modeling—perspectives, challenges, and opportunities in the biological, biomedical, and behavioral sciences”. In: *NPJ digital medicine* 2.1 (2019), p. 115.

[30] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. “A survey of transfer learning”. In: *Journal of Big data* 3.1 (2016), pp. 1–40.

[31] A Hoarfrost et al. “Deep learning of a bacterial and archaeal universal language of life enables transfer learning and illuminates microbial dark matter”. In: *Nature communications* 13.1 (2022), p. 2606.

[32] Hui Chong et al. “EXPERT: transfer learning-enabled context-aware microbial community classification”. In: *Briefings in Bioinformatics* 23.6 (2022), bbac396.

[33] Maude M David et al. “Revealing general patterns of microbiomes that transcend systems: potential and challenges of deep transfer learning”. In: *Msystems* 7.1 (2022), e01058–21.

[34] Guillermo Lopez-Garcia et al. “Transfer learning with convolutional neural networks for cancer survival prediction using gene-expression data”. In: *PloS one* 15.3 (2020), e0230536.

[35] Khushboo Bansal, RK Bathla, and Yogesh Kumar. "Deep transfer learning techniques with hybrid optimization in early prediction and diagnosis of different types of oral cancer". In: *Soft Computing* 26.21 (2022), pp. 11153–11184.

[36] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *nature* 521.7553 (2015), pp. 436–444.

[37] Jürgen Schmidhuber. "Deep learning in neural networks: An overview". In: *Neural networks* 61 (2015), pp. 85–117.

[38] Franco Scarselli et al. "The graph neural network model". In: *IEEE transactions on neural networks* 20.1 (2008), pp. 61–80.

[39] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[40] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[41] Victor Sanh et al. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter". In: *arXiv preprint arXiv:1910.01108* (2019).

[42] Kevin Clark et al. "Electra: Pre-training text encoders as discriminators rather than generators". In: *arXiv preprint arXiv:2003.10555* (2020).

[43] Zhilin Yang et al. "Xlnet: Generalized autoregressive pretraining for language understanding". In: *Advances in neural information processing systems* 32 (2019).

[44] Zhenzhong Lan et al. "Albert: A lite bert for self-supervised learning of language representations". In: *arXiv preprint arXiv:1909.11942* (2019).

[45] Yanrong Ji et al. "DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome". In: *Bioinformatics* 37.15 (2021), pp. 2112–2120.

[46] Hüseyin Anil Gündüz et al. "A self-supervised deep learning method for data-efficient training in genomics". In: *Communications Biology* 6.1 (2023), p. 928.

[47] Hanyu Luo et al. "iEnhancer-BERT: A novel transfer learning architecture based on DNA-Language model for identifying enhancers and their strength". In: *International Conference on Intelligent Computing*. Springer. 2022, pp. 153–165.

[48] Nguyen Quoc Khanh Le et al. "BERT-Promoter: An improved sequence-based predictor of DNA promoter using BERT pre-trained model and SHAP feature selection". In: *Computational Biology and Chemistry* 99 (2022), p. 107732.

[49] Lu Zhang et al. "BERT-m7G: a transformer architecture based on BERT and stacking ensemble to identify RNA N7-Methylguanosine sites from sequence information". In: *Computational and Mathematical Methods in Medicine* 2021 (2021).

[50] Fan Yang et al. "scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data". In: *Nature Machine Intelligence* 4.10 (2022), pp. 852–866.

[51] Nadav Brandes et al. "ProteinBERT: a universal deep-learning model of protein sequence and function". In: *Bioinformatics* 38.8 (2022), pp. 2102–2110.

[52] Noelia Ferruz, Steffen Schmidt, and Birte Höcker. "ProtGPT2 is a deep unsupervised language model for protein design". In: *Nature communications* 13.1 (2022), p. 4348.

[53] Ali Madani et al. "Large language models generate functional protein sequences across diverse families". In: *Nature Biotechnology* (2023), pp. 1–8.

[54] Tuoyu Liu et al. "Protein–protein interaction and site prediction using transfer learning". In: *Briefings in Bioinformatics* 24.6 (2023), bbad376.

[55] Yuchi Qiu and Guo-Wei Wei. "Artificial intelligence-aided protein engineering: from topological data analysis to deep protein language models". In: *Briefings in Bioinformatics* (2023), bbad289.

[56] S Deepak and PM Ameer. "Brain tumor classification using deep CNN features via transfer learning". In: *Computers in biology and medicine* 111 (2019), p. 103345.

[57] P Bharat Siva Varma et al. "SLDCNet: Skin lesion detection and classification using full resolution convolutional network-based deep learning CNN with transfer learning". In: *Expert Systems* 39.9 (2022), e12944.

[58] Sk Mahmudul Hassan et al. "Identification of plant-leaf diseases using CNN and transfer-learning approach". In: *Electronics* 10.12 (2021), p. 1388.

[59] Nadiah A Baghdadi et al. “An automated diagnosis and classification of COVID-19 from chest CT images using a transfer learning-based convolutional neural network”. In: *Computers in biology and medicine* 144 (2022), p. 105383.

[60] Behrouz Rostami et al. “Multiclass wound image classification using an ensemble deep CNN-based classifier”. In: *Computers in Biology and Medicine* 134 (2021), p. 104536.

[61] Sazan Mahbub and Md Shamsuzzoha Bayzid. “EGRET: edge aggregated graph attention networks and transfer learning improve protein–protein interaction site prediction”. In: *Briefings in Bioinformatics* 23.2 (2022), bbab578.

[62] Milad Salem et al. “Transcreen: transfer learning on graph-based anti-cancer virtual screening model”. In: *Big Data and Cognitive Computing* 4.3 (2020), p. 16.

[63] Rongbo Shen et al. “Spatial-ID: a cell typing method for spatially resolved transcriptomics via transfer learning and spatial embedding”. In: *Nature Communications* 13.1 (2022), p. 7640.

[64] Jhabindra Khanal et al. “i6ma-stack: a stacking ensemble-based computational prediction of dna n6-methyladenine (6ma) sites in the rosaceae genome”. In: *Genomics* 113.1 (2021), pp. 582–592.

[65] Divya Sharma, Andrew D Paterson, and Wei Xu. “TaxoNN: ensemble of neural networks on stratified microbiome data for disease prediction”. In: *Bioinformatics* 36.17 (2020), pp. 4544–4550.

[66] Yue Cao et al. “Ensemble deep learning in bioinformatics”. In: *Nature Machine Intelligence* 2.9 (2020), pp. 500–508.

[67] Mudasir A Ganaie et al. “Ensemble deep learning: A review”. In: *Engineering Applications of Artificial Intelligence* 115 (2022), p. 105151.

[68] Somayah Albaradei et al. “Splice2Deep: An ensemble of deep convolutional neural networks for improved splice site prediction in genomic DNA”. In: *Gene* 763 (2020), p. 100035.

[69] Advait Balaji et al. “SeqScreen: accurate and sensitive functional screening of pathogenic sequences via ensemble learning”. In: *Genome Biology* 23.1 (2022), p. 133.

[70] Balachandran Manavalan et al. “4mCpred-EL: an ensemble learning framework for identification of DNA N4-methylcytosine sites in the mouse genome”. In: *Cells* 8.11 (2019), p. 1332.

[71] Fuyi Li et al. “Computational prediction and interpretation of both general and specific types of promoters in Escherichia coli by exploiting a stacked ensemble-learning framework”. In: *Briefings in bioinformatics* 22.2 (2021), pp. 2126–2140.

[72] Abhishek Das et al. “Breast cancer detection using an ensemble deep learning method”. In: *Biomedical Signal Processing and Control* 70 (2021), p. 103009.

[73] P Mohamed Shakeel, MA Burhanuddin, and Mohammad Ishak Desa. “Automatic lung cancer detection from CT image using improved deep neural network and ensemble classifier”. In: *Neural Computing and Applications* (2022), pp. 1–14.

[74] Daniel Quang and Xiaohui Xie. “FactorNet: a deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data”. In: *Methods* 166 (2019), pp. 40–47.

[75] Pablo Millán Arias et al. “DeLUCS: Deep learning for unsupervised clustering of DNA sequences”. In: *Plos one* 17.1 (2022), e0261531.

[76] Juntao Chen, Quan Zou, and Jing Li. “DeepM6ASeq-EL: prediction of human N6-methyladenosine (m 6 A) sites with LSTM and ensemble learning”. In: *Frontiers of Computer Science* 16 (2022), pp. 1–7.

[77] Sultan Imangaliyev et al. “Diagnosis of Inflammatory Bowel Disease and Colorectal Cancer through Multi-View Stacked Generalization Applied on Gut Microbiome Data”. In: *Diagnostics* 12.10 (2022), p. 2514.

[78] Xingjian Chen et al. “Human disease prediction from microbiome data by multiple feature fusion and deep learning”. In: *Iscience* 25.4 (2022).

[79] Mehdi Foroozandeh Shahraki et al. “MCIC: automated identification of cellulases from metagenomic data and characterization based on temperature and pH dependence”. In: *Frontiers in Microbiology* 11 (2020), p. 567863.

[80] Qiang Zhu et al. “Robust biomarker discovery for microbiome-wide association studies”. In: *Methods* 173 (2020), pp. 44–51.

[81] Boya Ji et al. “HyperVR: a hybrid deep ensemble learning approach for simultaneously predicting virulence factors and antibiotic resistance genes”. In: *NAR Genomics and Bioinformatics* 5.1 (2023), lqad012.

[82] Ruopeng Xie et al. “DeepVF: a deep learning-based hybrid framework for identifying virulence factors using the stacking strategy”. In: *Briefings in bioinformatics* 22.3 (2021), bbaa125.

[83] Arun Rai. “Explainable AI: From black box to glass box”. In: *Journal of the Academy of Marketing Science* 48 (2020), pp. 137–141.

[84] Christina B Azodi, Jiliang Tang, and Shin-Han Shiu. “Opening the black box: interpretable machine learning for geneticists”. In: *Trends in genetics* 36.6 (2020), pp. 442–455.

[85] Md Rezaul Karim et al. “Explainable AI for Bioinformatics: Methods, Tools and Applications”. In: *Briefings in bioinformatics* 24.5 (2023), bbad236.

[86] Mengnan Du, Ninghao Liu, and Xia Hu. “Techniques for interpretable machine learning”. In: *Communications of the ACM* 63.1 (2019), pp. 68–77.

[87] W James Murdoch et al. “Definitions, methods, and applications in interpretable machine learning”. In: *Proceedings of the National Academy of Sciences* 116.44 (2019), pp. 22071–22080.

[88] Cristian Munoz et al. “Local and Global Explainability Metrics for Machine Learning Predictions”. In: *arXiv preprint arXiv:2302.12094* (2023).

[89] Carla Piazzon Vieira and Luciano Antonio Digiampietri. “Machine Learning post-hoc interpretability: a systematic mapping study”. In: *XVIII Brazilian Symposium on Information Systems*. 2022, pp. 1–8.

[90] Andreas Madsen, Siva Reddy, and Sarath Chandar. “Post-hoc interpretability for neural nlp: A survey”. In: *ACM Computing Surveys* 55.8 (2022), pp. 1–42.

[91] Scott M Lundberg and Su-In Lee. “A unified approach to interpreting model predictions”. In: *Advances in neural information processing systems* 30 (2017).

[92] Scott M Lundberg et al. “From local explanations to global understanding with explainable AI for trees”. In: *Nature machine intelligence* 2.1 (2020), pp. 56–67.

[93] Olga Mineeva et al. “ResMiCo: Increasing the quality of metagenome-assembled genomes with deep learning”. In: *PLOS Computational Biology* 19.5 (2023), e1011001.

[94] Joshua E Lewis and Melissa L Kemp. “Integration of machine learning and genome-scale metabolic modeling identifies multi-omics biomarkers for radiation resistance”. In: *Nature Communications* 12.1 (2021), p. 2700.

[95] Joshua J Levy et al. “MethylNet: an automated and modular deep learning approach for DNA methylation analysis”. In: *BMC bioinformatics* 21.1 (2020), pp. 1–15.

[96] Ryza Rynazal et al. “Leveraging explainable AI for gut microbiome-based colorectal cancer classification”. In: *Genome Biology* 24.1 (2023), pp. 1–13.

[97] Youngro Lee, Marco Cappellato, and Barbara Di Camillo. “Machine learning–based feature selection to search stable microbial biomarkers: application to inflammatory bowel disease”. In: *GigaScience* 12 (2023), giad083.

[98] Fatma Hilal Yagin et al. “Explainable artificial intelligence model for identifying COVID-19 gene biomarkers”. In: *Computers in Biology and Medicine* 154 (2023), p. 106619.

[99] Kimmo Sirén et al. “Rapid discovery of novel prophages using biological feature engineering and machine learning”. In: *NAR genomics and bioinformatics* 3.1 (2021), lqaa109.

[100] Zhihai Shi et al. “Metagenomic and metabolomic analyses reveal the role of gut microbiome-associated metabolites in diarrhea calves”. In: *Msystems* (2023), e00582–23.

[101] Yeela Talmor-Barkan et al. “Metabolomic and microbiome profiling reveals personalized risk factors for coronary artery disease”. In: *Nature medicine* 28.2 (2022), pp. 295–302.

[102] Izhak Levi et al. “Potential role of indolelactate and butyrate in multiple sclerosis revealed by integrated microbiome-metabolome analysis”. In: *Cell Reports Medicine* 2.4 (2021).

[103] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “" Why should i trust you?" Explaining the predictions of any classifier”. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144.

[104] Md Sarwar Kamal et al. “Alzheimer’s patient analysis using image and gene expression data and explainable-AI to present associated genes”. In: *IEEE Transactions on Instrumentation and Measurement* 70 (2021), pp. 1–7.

[105] Vijayachitra Modhukur et al. “Machine learning approaches to classify primary and metastatic cancers using tissue of origin-based DNA methylation profiles”. In: *Cancers* 13.15 (2021), p. 3768.

[106] Daniel Fryer, Inga Strümke, and Hien Nguyen. “Shapley values for feature selection: The good, the bad, and the axioms”. In: *Ieee Access* 9 (2021), pp. 144352–144360.

[107] Bojan Mihaljević, Concha Bielza, and Pedro Larrañaga. “Bayesian networks for interpretable machine learning and optimization”. In: *Neurocomputing* 456 (2021), pp. 648–665.

[108] Igor Kononenko. “Bayesian neural networks”. In: *Biological Cybernetics* 61.5 (1989), pp. 361–370.

[109] Yongxin Yang, Irene Garcia Morillo, and Timothy M Hospedales. “Deep neural decision trees”. In: *arXiv preprint arXiv:1806.06988* (2018).

[110] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. “Neural machine translation by jointly learning to align and translate”. In: *arXiv preprint arXiv:1409.0473* (2014).

[111] Sarthak Jain and Byron C Wallace. “Attention is not explanation”. In: *arXiv preprint arXiv:1902.10186* (2019).

[112] Sarah Wiegreffe and Yuval Pinter. “Attention is not not explanation”. In: *arXiv preprint arXiv:1908.04626* (2019).

[113] Zhaoyang Niu, Guoqiang Zhong, and Hui Yu. “A review on the attention mechanism of deep learning”. In: *Neurocomputing* 452 (2021), pp. 48–62.

[114] Mohammad Arifur Rahman and Huzefa Rangwala. "IDMIL: an alignment-free Interpretable Deep Multiple Instance Learning (MIL) for predicting disease from whole-metagenomic data". In: *Bioinformatics* 36.Supplement_1 (2020), pp. i39–i47.

[115] Zitao Song et al. "Attention-based multi-label neural networks for integrated prediction and interpretation of twelve widely occurring RNA modifications". In: *Nature communications* 12.1 (2021), p. 4011.

[116] Zhengqiao Zhao et al. "Learning, visualizing and exploring 16S rRNA structure using an attention-based deep neural network". In: *PLoS computational biology* 17.9 (2021), e1009345.

[117] Amlan Talukder et al. "Interpretation of deep learning in genomics and epigenomics". In: *Briefings in Bioinformatics* 22.3 (2021), bbaa177.

[118] Muhammad Shoaib et al. "Deep learning for plant bioinformatics: an explainable gradient-based approach for disease detection". In: *Frontiers in Plant Science* 14 (2023), p. 1283235.

[119] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[120] Qi Wang et al. "A comprehensive survey of loss functions in machine learning". In: *Annals of Data Science* (2020), pp. 1–26.

[121] Anqi Mao, Mehryar Mohri, and Yutao Zhong. "Cross-entropy loss functions: Theoretical analysis and applications". In: *arXiv preprint arXiv:2304.07288* (2023).

[122] Rosario Vitale and Georgina Stegmayer. "Evaluating transfer learning for classification of proteins in bioinformatics". In: *Memorias de las JAIIO* 9.2 (2023), pp. 25–36.

[123] Luis Torada et al. "ImaGene: a convolutional neural network to quantify natural selection from genomic data". In: *BMC bioinformatics* 20 (2019), pp. 1–12.

[124] Mobeen Ur Rehman et al. "Bu-net: Brain tumor segmentation using modified u-net architecture". In: *Electronics* 9.12 (2020), p. 2203.

[125] Alex Chklovski et al. "CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning". In: *Nature Methods* 20.8 (2023), pp. 1203–1212.

[126] Vishal Rajput. "Robustness of different loss functions and their impact on networks learning capability". In: *arXiv preprint arXiv:2110.08322* (2021).

[127] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. "V-net: Fully convolutional neural networks for volumetric medical image segmentation". In: *2016 fourth international conference on 3D vision (3DV)*. Ieee. 2016, pp. 565–571.

[128] Tsung-Yi Lin et al. "Focal loss for dense object detection". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2980–2988.

[129] Henghui Fan et al. "Deep learning-based multi-functional therapeutic peptides prediction with a multi-label focal dice loss function". In: *Bioinformatics* 39.6 (2023), btad334.

[130] Antonio Pedro Camargo et al. "Identification of mobile genetic elements with geNomad". In: *Nature Biotechnology* (2023), pp. 1–10.

[131] Thomas Wollmann et al. "GRUU-Net: Integrated convolutional and gated recurrent neural network for cell segmentation". In: *Medical image analysis* 56 (2019), pp. 68–79.

[132] Michael Yeung et al. "Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation". In: *Computerized Medical Imaging and Graphics* 95 (2022), p. 102026.

[133] Peng Han et al. "GCN-MF: disease-gene association identification by graph convolutional networks and matrix factorization". In: *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2019, pp. 705–713.

[134] Wenhuan Zeng, Anupam Gautam, and Daniel H Huson. "MuLan-Methyl—multiple transformer-based language models for accurate DNA methylation prediction". In: *GigaScience* 12 (2023), giad054.

[135] Wenhuan Zeng and Daniel H. Huson. "Enhanced 5mC-Methylation-Site Recognition in DNA Sequences using Token Classification and a Domain-specific Loss Function". In: *bioRxiv* (2024). DOI: 10.1101/2023.06.01.543218.

[136] Wenhuan Zeng, Anupam Gautam, and Daniel H Huson. "On the application of advanced machine learning methods to analyze enhanced, multimodal data from persons infected with COVID-19". In: *Computation* 9.1 (2021), p. 4.

[137] Pauline A Callinan and Andrew P Feinberg. "The emerging science of epigenomics". In: *Human molecular genetics* 15.suppl_1 (2006), R95–R101.

[138] N Shenker and JM Flanagan. "Intragenic DNA methylation: implications of this epigenetic mechanism for cancer research". In: *British journal of cancer* 106.2 (2012), pp. 248–253.

[139] Lisa D Moore, Thuc Le, and Guoping Fan. "DNA methylation and its basic function". In: *Neuropsychopharmacology* 38.1 (2013), pp. 23–38.

[140] Zhenhui Zhong et al. "DNA methylation-linked chromatin accessibility affects genomic architecture in Arabidopsis". In: *Proceedings of the National Academy of Sciences* 118.5 (2021), e2023347118.

[141] Anna Portela and Manel Esteller. "Epigenetic modifications and human disease". In: *Nature biotechnology* 28.10 (2010), pp. 1057–1068.

[142] Robert A Gaultney et al. "4-Methylcytosine DNA modification is critical for global epigenetic regulation and virulence in the human pathogen Leptospira interrogans". In: *Nucleic Acids Research* 48.21 (2020), pp. 12102–12115.

[143] John Beaulaurier, Eric E Schadt, and Gang Fang. "Deciphering bacterial epigenomes using modern sequencing technologies". In: *Nature Reviews Genetics* 20.3 (2019), pp. 157–172.

[144] Ieva Rauluseviciute, Finn Drabløs, and Morten Beck Rye. "DNA methylation data by sequencing: experimental approaches and recommendations for tools and pipelines for data analysis". In: *Clinical epigenetics* 11.1 (2019), pp. 1–13.

[145] Liu Xu and Masahide Seki. "Recent advances in the detection of base modifications using the Nanopore sequencer". In: *Journal of human genetics* 65.1 (2020), pp. 25–33.

[146] Shizhao Li and Trygve O Tollefsbol. "DNA methylation methods: Global DNA methylation and methylomic analyses". In: *Methods* 187 (2021), pp. 28–43.

[147] Saleh Sereshki et al. "On the prediction of non-CG DNA methylation using machine learning". In: *NAR genomics and bioinformatics* 5.2 (2023), lqad045.

[148] Quanzhong Liu et al. "DeepTorrent: a deep learning-based approach for predicting DNA N4-methylcytosine sites". In: *Briefings in bioinformatics* 22.3 (2021), bbaa124.

[149] Sho Tsukiyama et al. "BERT6mA: prediction of DNA N6-methyladenine site using deep learning-based approaches". In: *Briefings in Bioinformatics* 23.2 (2022), bbac053.

[150] Yingying Yu et al. "iDNA-ABT: advanced deep learning model for detecting DNA methylation with adaptive features and transductive information maximization". In: *Bioinformatics* 37.24 (2021), pp. 4603–4610.

[151] Junru Jin et al. "iDNA-ABF: multi-scale deep biological language learning model for the interpretable prediction of DNA methylations". In: *Genome biology* 23.1 (2022), pp. 1–23.

[152] Xingyu Tang et al. "Deep6mAPred: A CNN and Bi-LSTM-based deep learning method for predicting DNA N6-methyladenosine sites across plant species". In: *Methods* 204 (2022), pp. 142–150.

[153] Nataliya Petryk et al. "Staying true to yourself: mechanisms of DNA methylation maintenance in mammals". In: *Nucleic acids research* 49.6 (2021), pp. 3020–3032.

[154] Michael Weber et al. "Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome". In: *Nature genetics* 39.4 (2007), pp. 457–466.

[155] Hao Lv et al. "iDNA-MS: an integrated computational tool for detecting DNA modification sites in multiple genomes". In: *Iscience* 23.4 (2020).

[156] Hanieh Poostchi and Massimo Piccardi. "A multi-constraint structured hinge loss for named-entity recognition". In: *Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association*. 2019, pp. 41–46.

[157] Lei Zhang, Xuan Xiao, and Zhao-Chun Xu. "iPromoter-5mC: a novel fusion decision predictor for the identification of 5-methylcytosine sites in genome-wide DNA promoters". In: *Frontiers in Cell and Developmental Biology* 8 (2020), p. 614.

[158] Alex L Mitchell et al. "MGnify: the microbiome analysis resource in 2020". In: *Nucleic acids research* 48.D1 (2020), pp. D570–D578.

[159] Elizabeth Stulberg et al. "An assessment of US microbiome research". In: *Nature microbiology* 1.1 (2016), pp. 1–7.

[160] Despoina D Roumpeka et al. "A review of bioinformatics tools for bio-prospecting from metagenomic sequence data". In: *Frontiers in genetics* (2017), p. 23.

[161] Ana Elena Pérez-Cobas, Laura Gomez-Valero, and Carmen Buchrieser. "Metagenomic approaches in microbial ecology: an update on whole-genome and marker gene sequencing analyses". In: *Microbial genomics* 6.8 (2020).

[162] Maxime Borry. "Sourcepredict: Prediction of metagenomic sample sources using dimension reduction followed by machine learning classification". In: *The Journal of Open Source Software* (2019).

[163] Ulzee An et al. "STENSL: Microbial Source Tracking with ENvironment SeLection". In: *Msystems* 7.5 (2022), e00995–21.

[164] Cooke RC. Whipps JM Lewis K. "Mycoparasitism and plant disease control". In: *Fungi in Biological Control Systems*. Ed. by NM Burge. P. 176. Manchester University Press, 1988, pp. 161–187.

[165] Gabriele Berg et al. "Microbiome definition re-visited: old concepts and new challenges". In: *Microbiome* 8.1 (2020), p. 103.

[166] Alex Coleman and Martin Callaghan. "Manipulating and Basic Analysis of Tabular Metagenomics Datasets Using R". In: *Metagenomic Data Analysis*. Ed. by Suparna Mitra. New York, NY: Springer US, 2023, pp. 339–357. ISBN: 978-1-0716-3072-3. DOI: 10.1007/978-1-0716-3072-3_18. URL: https://doi.org/10.1007/978-1-0716-3072-3_18.

[167] George Armstrong et al. "Applications and comparison of dimensionality reduction methods for microbiome data". In: *Frontiers in bioinformatics* 2 (2022), p. 821861.

[168] Aditya Grover and Jure Leskovec. *node2vec: Scalable Feature Learning for Networks*. 2016. arXiv: 1607.00653 [cs.SI].

[169] Felix Wong, Cesar de la Fuente-Nunez, and James J Collins. "Leveraging artificial intelligence in the fight against infectious diseases". In: *Science* 381.6654 (2023), pp. 164–170.

[170] Said Agrebi and Anis Larbi. "Use of artificial intelligence in infectious diseases". In: *Artificial intelligence in precision health*. Elsevier, 2020, pp. 415–438.

[171] Ben Hu et al. “Characteristics of SARS-CoV-2 and COVID-19”. In: *Nature Reviews Microbiology* 19.3 (2021), pp. 141–154.

[172] Li Yan et al. “An interpretable mortality prediction model for COVID-19 patients”. In: *Nature machine intelligence* 2.5 (2020), pp. 283–288.

[173] David S Hui et al. “The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health—The latest 2019 novel coronavirus outbreak in Wuhan, China”. In: *International journal of infectious diseases* 91 (2020), pp. 264–266.

[174] Andrea Lovato, Cosimo de Filippis, and Gino Marioni. “Upper airway symptoms in coronavirus disease 2019 (COVID-19)”. In: *American journal of Otolaryngology* (2020).

[175] Hanie Esakandari et al. “A comprehensive review of COVID-19 characteristics”. In: *Biological procedures online* 22.1 (2020), pp. 1–10.

[176] Petter Brodin. “Immune determinants of COVID-19 disease presentation and severity”. In: *Nature medicine* 27.1 (2021), pp. 28–33.

[177] Omid Dadras et al. “The relationship between COVID-19 viral load and disease severity: a systematic review”. In: *Immunity, inflammation and disease* 10.3 (2022), e580.

[178] Zahra Niknam et al. “Potential therapeutic options for COVID-19: an update on current evidence”. In: *European journal of medical research* 27 (2022), pp. 1–15.

[179] *WEATHER UNDERGROUND*. `https://www.wunderground.com/`.

[180] David N Prata, Waldecy Rodrigues, and Paulo H Bermejo. “Temperature significantly changes COVID-19 transmission in (sub) tropical cities of Brazil”. In: *Science of the Total Environment* 729 (2020), p. 138862.

[181] Jacques Demongeot, Yannis Flet-Berliac, and Hervé Seligmann. “Temperature decreases spread parameters of the new Covid-19 case dynamics”. In: *Biology* 9.5 (2020), p. 94.

[182] Florian Markowetz. “All biology is computational biology”. In: *PLoS biology* 15.3 (2017), e2002050.

[183] Anne E Carpenter and Shantanu Singh. “Bringing computation to biology by bridging the last mile”. In: *Nature cell biology* 26.1 (2024), pp. 5–7.

[184] Wenhuan Zeng, Anupam Gautam, and Daniel H Huson. "DeepToA: an ensemble deep-learning approach to predicting the theater of activity of a microbiome". In: *Bioinformatics* 38.20 (2022), pp. 4670–4676.

[185] Guozhu Dong and Huan Liu. *Feature engineering for machine learning and data analytics*. CRC press, 2018.

Abbreviations

4mC	4-Methylcytosine
5hmC	5-Hydroxymethylcytosine
5mC	5-Methylcytosine
6mA	6-Methyladenine
Bi-LSTM	Bidirectional Long Short-Term Memory
bp	Base Pair
BS-Seq	Bisulfite Sequencing
CCE	Categorical Cross-Entropy
CNN	Convolutional Neural Network
CRF	Conditional Random Fields
GNN	Graph Neural Network
GTDB	Genome Taxonomy Database
MAE	Mean Absolute Error
mNGS	Metagenomics Next-Generation Sequencing
MLM	Masked Language Modelling
MLP	Multilayer Perceptron
MRI	Magnetic Resonance Imaging
MSE	Mean Squared Error
NCBI	National Center for Biotechnology Information
NER	Named Entity Recognition
NLP	Natural Language Processing
NSP	Next Sentence Prediction
OTU	Operational Taxonomic Unit
RRBS	Reduced Representation Bisulfite Sequencing
SHAP	SHapely Additive exPlanations
SMRT	Single-Molecule Real-Time

LIME	Local Interpretable Model-Agnostic Explanations
WGBS	Whole Genome Bisulfite Sequencing
WGS	Whole-Genome Shotgun
XGBoost	eXtreme Gradient Boosting

Appendix

A Manuscript 1

Title: MuLan-Methyl—Multiple Transformer-Based Language Models for Accurate DNA Methylation Prediction

OXFORD (GIGA)n SCIENCE

GigaScience, 2023, **12**, 1–11
DOI: 10.1093/gigascience/giad054
Research

MuLan-Methyl—multiple transformer-based language models for accurate DNA methylation prediction

Wenhuan Zeng [1], Anupam Gautam [1,2,3] and Daniel H. Huson [1,2,3,*]

[1]Algorithms in Bioinformatics, Institute for Bioinformatics and Medical Informatics, University of Tübingen, 72076 Tübingen, Germany
[2]International Max Planck Research School "From Molecules to Organisms", Max Planck Institute for Biology Tübingen, 72076 Tübingen, Germany
[3]Cluster of Excellence: EXC 2124: Controlling Microbes to Fight Infection, University of Tübingen, 72076 Tübingen, Germany
*Correspondence address. Daniel H. Huson, Sand 14, University of Tübingen 72076 Germany. E-mail: daniel.huson@uni-tuebingen.de

Abstract

Transformer-based language models are successfully used to address massive text-related tasks. DNA methylation is an important epigenetic mechanism, and its analysis provides valuable insights into gene regulation and biomarker identification. Several deep learning–based methods have been proposed to identify DNA methylation, and each seeks to strike a balance between computational effort and accuracy. Here, we introduce MuLan-Methyl, a deep learning framework for predicting DNA methylation sites, which is based on 5 popular transformer-based language models. The framework identifies methylation sites for 3 different types of DNA methylation: N6-adenine, N4-cytosine, and 5-hydroxymethylcytosine. Each of the employed language models is adapted to the task using the "pretrain and fine-tune" paradigm. Pretraining is performed on a custom corpus of DNA fragments and taxonomy lineages using self-supervised learning. Fine-tuning aims at predicting the DNA methylation status of each type. The 5 models are used to collectively predict the DNA methylation status. We report excellent performance of MuLan-Methyl on a benchmark dataset. Moreover, we argue that the model captures characteristic differences between different species that are relevant for methylation. This work demonstrates that language models can be successfully adapted to applications in biological sequence analysis and that joint utilization of different language models improves model performance. Mulan-Methyl is open source, and we provide a web server that implements the approach.

Keywords: DNA methylation, natural language processing, model ensemble, model explainability, web server

Key Points:

- MuLan-Methyl aims at identifying 3 types of DNA methylation sites.
- It uses an ensemble of 5 transformer-based language models, which were pretrained and fine-tuned on a custom corpus.
- The self-attention mechanism of transformers give rise to importance scores, which can be used to extract motifs.
- The method performs favorably in comparison to existing methods.
- The implementation can be applied to chromosomal sequences to predict methylation sites.

Introduction

DNA methylation is an important biological process. It facilitates epigenetic regulation of gene expression, is associated with various medical disorders [1–3], and has other applications, such as a marker in metagenomic binning [4]. While DNA methylation is a dynamic process, existing machine learning techniques are able to predict DNA methylation states from genomic sequence with some degree of accuracy.

There are several types of DNA methylation that differ by which methyl group is attached to which type of nucleotide in the sequence. Here, we focus on 6-methyladenine (6mA), 5-hydroxymethylcytosine (5hmC), and 4-methylcytosine (4mC) methylation [5–7]. Different organisms exhibit different patterns of methylation, and this gives rise to the computational problem of predicting the location of methylation sites for a given genome sequence. While much algorithmic work has been done on the question, recent work has focused on the application of deep learning methods [8, 9]. However, there is room for improvement of accuracy and comprehensiveness.

A large number of studies address the problem of identifying methylation sites, but most of them focus on a specific form of modification [10–29], and only a few methods address all 3 types of methylation mentioned above [30–34], in particular iDNA-MS, iDNA-ABT, and iDNA-ABF. The database presented in [31, 35] is now widely used as a benchmark dataset for assessing model performance [21, 23, 32–34].

While different deep learning–based methods all address the same goal, they differ in the details of the features employed and the model structure. Input features include an encoding of the sequence, of course, but may also include biochemical properties [10, 12] or a DNA molecular graph representation [22]. Utilized model structures include convolutional neural networks, graph convolutional neural networks, bidirectional encoder representation from transformers (BERT) [36], and other types of machine learning algorithms. The specific choices made during feature engineering and model selection determine performance and are key to proposing a new framework.

Received: February 19, 2023. **Revised:** May 9, 2023. **Accepted:** July 18, 2023

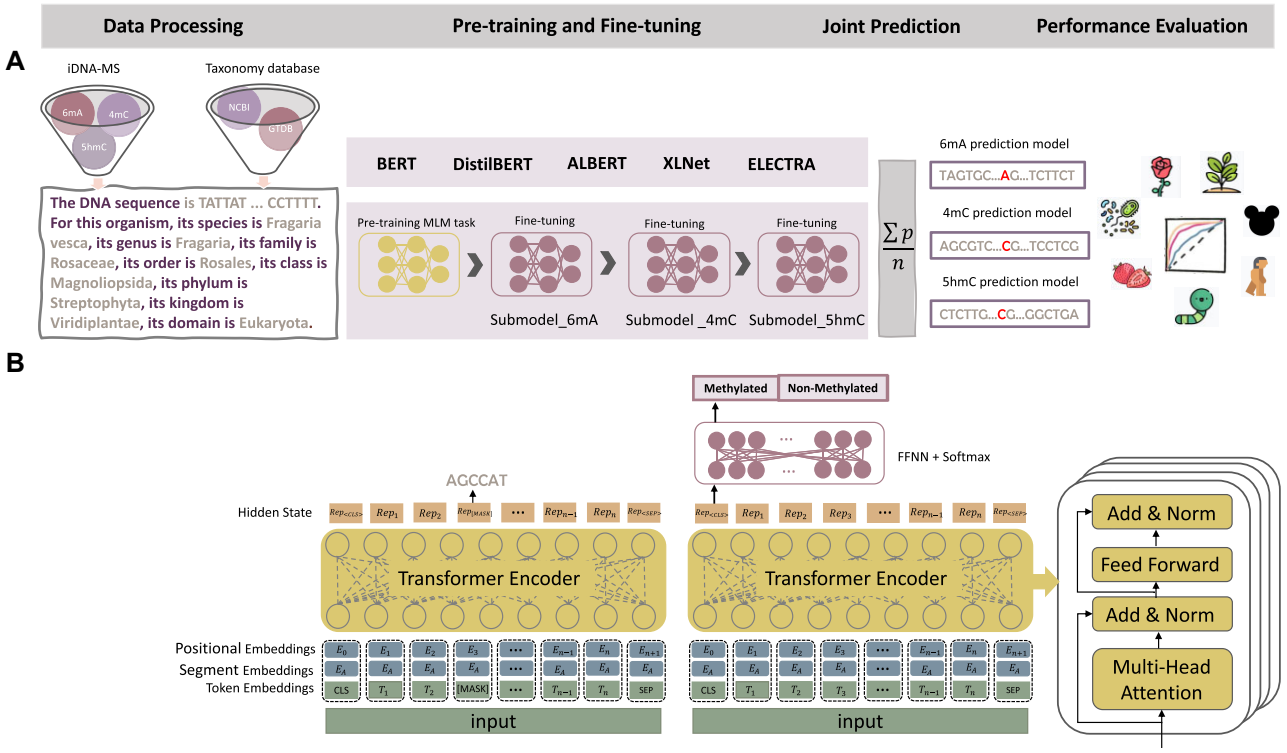


Figure 1: The MuLan-Methyl workflow. (A) The framework employs 5 fine-tuned language models for joint identification of DNA methylation sites. Methylation datasets (obtained from iDNA-MS) were processed as sentences that describe the DNA sequence as well as the taxonomy lineage, giving rise to the processed training dataset and the processed independent set. For each transformer-based language model, a custom tokenizer was trained based on a corpus of the processed training dataset and taxonomy lineage data. Pretraining and fine-tuning were both conducted on each methylation site-specific training subset separately. During model testing, the prediction of a sample in the processed independent test set is defined as the average prediction probability of the 5 fine-tuned models. We thus obtained 3 methylation type-wise prediction models. We evaluated the model performance on the genome type that contained the corresponding methylation type-wise dataset, respectively. In total, we evaluated 17 combinations of methylation types and taxonomic lineages. (B) The general transformer-based language model architecture for pretraining and fine-tuning. The model was pretrained using the masked language modeling (MLM) task and then fine-tuned on the methylation type-wise processed training dataset.

Here, we phrase DNA methylation site detection as a natural language processing (NLP) problem and propose a novel framework to address it. Previous studies for identifying methylation sites usually use BERT, a classic NLP approach, or, in the context of DNA sequences, the variant DNABERT [37], either as a model that accepts embeddings from Word2vec or as an encoder that generates embeddings for input to a deep neural network [23, 25, 32, 33, 38]. Only few published approaches aim at predicting multiple DNA modification sites. Moreover, many do not use taxonomic information as explicit features, although the taxonomic identity of an organism has an impact on DNA methylation [39]. Here we address both shortcomings by providing a new framework that uses a set of collective training language models, including but not limited to BERT, to predict 3 types of methylation sites from DNA sequences and taxonomic information.

Combining the transformer-based language model BERT with the "pretrain and fine-tune" paradigm has become the method of choice in NLP applications. In the pretraining step, self-supervised learning of the masked language modeling (MLM) task and the next sequence prediction task is usually performed on a corpus consisting of Wikipedia and books. This allows the transformer-based language model to capture the semantics of text input and contextual information exceptionally well. Transformer-based language models dynamically learn the input's representation through a multihead self-attention mechanism [40], and this leads to an improvement of prediction over classification models constructed using static embedding approaches [41]. The fine-tuning step involves supervised training of the pretrained language model to adapt to specific downstream tasks, here the prediction of 3 different types of methylation sites. Using BERT as a starting point, and then varying the network architecture and parameters, one can obtain 5 different language models, [42–46]. By pretraining on a domain-specific custom corpus, BERT can be adapted to a specific application scenario [47–50]. While the analysis of DNA sequences can be considered an application of NLP, using language models that are trained on human languages will not do well at capturing nucleotide rules. To address this, several approaches, such as BERTax, DNABERT, and LOGO [37, 51, 52], use large amounts of genomic sequence, instead of Wikipedia, as a corpus or similar structure.

The main aim of this article is to introduce MuLan-Methyl, a novel deep learning framework that combines 5 transformer-based language models to collectively predict sites for 3 different types of methylation (see Fig. 1A). In this approach, each methylation site sample is written as a sentence that represents the surrounding DNA sequence and the taxonomic identity of the corresponding genome. The output of our model is based on the average of the prediction probabilities obtained by 5 transformer-based language models: BERT [36], DistilBERT [42], ALBERT [46], XLNet [44], and ELECTRA [45].

Each of the 5 language models is trained according to the "pretrain and fine-tune" paradigm. For this, we used a custom corpus that contains the processed training dataset and taxonomic lineage information downloaded from NCBI [53] and GTDB [54]. For

each language model, we trained a custom tokenizer on the custom corpus, using the same configuration as the model's default tokenizer. We use a customized tokenizer to ensure that the represented DNA sequences and taxonomic information associated with each sample are captured effectively. Each language model was pretrained by training the MLM task on the processed training dataset. We then obtained the 6mA model by fine-tuning the pretrained language model using the 6mA training dataset. Next, the 4mC prediction model was obtained by fine-tuning the 6mA prediction model using the 4mC training dataset. Finally, the 5hmC prediction model was obtained by fine-tuning the 4mC prediction model using the 5hmC training dataset. In addition, we compared the performance of all models contained in MuLan-Methyl.

A main contribution of this work is that we use both DNA sequence and taxonomic identity as explicit features in the model. Using the iDNA-MS [31] independent test set as a benchmark, our approach shows improved performance over previous methods, especially for certain genomes. MuLan-Methyl is capable of making accurate predictions for genomes whose taxonomy lineage is not present in the training dataset. The interpretability of MuLan-Methyl facilitates the discovery of DNA motifs that are associated with DNA methylation and potential correlations between specific methylation sites and taxonomic lineages.

To the best of our knowledge, this is the first application in biology that achieves improved prediction performance by integrating multiple transformer-based language models. This shows that adding features to a model is not the only way to improve the accuracy of predictions.

Materials and Methods

Data processing

Data collection

We downloaded a DNA methylation dataset from the iDNA-MS web server [55]. This is an open resource that was published with the iDNA-MS method [31] and is widely used for benchmarking. The dataset contains 3 main types of DNA methylation sites (6mA, 4mC, and 5hmC) across 12 genomes (1 bacteria and 11 eukaryotes), in total 250,599 positive samples. In addition, the dataset provides the same number of nonmethylation sequences as negative samples.

The dataset is partitioned into a training set and an independent test set at a 1:1 ratio. The training dataset provides samples for methylation type 6mA present in 11 different species. In more detail, the numbers are 53,800 for *T. thermophile*, 15,937 for *Arabidopsis thaliana*, 9,168 for *Homo sapiens*, 8,608 for *Xoc. BLS256*, 5,596 for *Drosophila melanogaster*, 3,981 for *Caenorhabditis elegans*, 3,033 for *Casuarina equisetifolia*, 1,893 for *Saccharomyces cerevisiae*, 1,690 for *Tolypocladium*, 1,551 for *Fragaria vesca*, and 300 for *Rosa chinensis*. The 4mC methylation type is present in 4 species, where the numbers of samples are 7,899 for *F. vesca*, 7,664 for *Tolypocladium*, 990 for *S. cerevisiae*, and 183 for *C. equisetifolia*. Finally, the numbers of samples for the type 5hmC are 1,840 for *Mus musculus* sequences and 1,172 for *H. sapiens*. More detailed statistics are provided in Supplementary Table S1.

Each sample is a DNA segment of length 41, which is centered on an experimentally verified methylation site, in the case of a positive sample.

Dataset preparation

We processed each sample (DNA sequence of length 41) as follows. Using a sliding window of length 6, we extracted $36 = 41 - 6 + 1$ individual 6-mers from the DNA sequence and embedded these into a sentence, together with a description of the taxonomic lineage of the corresponding organism, which was phrased as follows: "For this organism, its species is *species*, its genus is *genus*, its family is *family*, its order is *order*, its class is *class*, its phylum is *phylum*, its kingdom is *kingdom*, its domain is *domain*." We refer to a set of sentences obtained from a set of samples as a "processed dataset." The full processed training dataset, containing all 3 types of methylation sites, was put into the custom corpus. For purposes of fine-tuning, both the processed training dataset and the processed independent test set were split into 3 sets by methylation type.

Corpus generation

We require a custom corpus for pretraining each language model to allow the model to learn and capture domain-specific words, which are not contained in a text corpus such as Wikipedia. The custom corpus contains the processed training dataset, as mentioned above. In addition, to enable the language to detect words about taxonomy, we incorporated all taxonomic lineages from the NCBI and GTDB taxonomies that are not already contained in the training datasets. In total, the corpus contains 2,440,894 sentences and uses a vocabulary of 25,000 words.

External dataset

We downloaded DNA methylation data published with the Hyb4mC method [16] and with the i6mA-pred method [56], respectively. As these "external" data are not contained in our training or independent datasets, nor do the associated taxonomic lineages coincide, it is ideal for evaluating the performance of MuLan-Methyl more broadly. In more detail, these data consist of samples (DNA sequences of length 41) representing 320 4mC site sequences in *Escherichia coli*, 1,926 4mC site sequences in *Geobacter pickeringii*, and 880 6mA site sequences in *Oryza sativa* L., each with the same number of negative samples, respectively.

Training transformer-based language models

We pretrained and fine-tuned 5 transformer-based language models. In the following, we first describe the architecture of each of the 5 employed language models. We then discuss the details of the training process for the first method, BERT (RRID:SCR_018008), including tokenization, pretraining, and fine-tuning (see Fig. 1B). The other 4 languages are trained in a similar way.

All code is written in Python 3.10, using the Pytorch and the Huggingface Transformers library [57]. The experiments were run on a Linux Virtual Machine (Ubuntu 20.04 LTS) equipped with 4 GPUs provided by de.NBI (flavor: de.NBI RTX6000 4 GPU medium).

Transformer-based language models

Our approach uses 5 transformer-based language models, which we introduce in the following.

(1) BERT is capable of modeling bidirectional contexts, using denoising and autoencoding-based pretraining. The transformer architecture of $\text{BERT}_{\text{base}}$ uses 12 layers in the encoder stack, 768 hidden units for feed-forward networks, and 12 attention heads; it has 110 million parameters in total.
(2) A distilled version of BERT, DistilBERT, is obtained by decreasing the number of layers. It has 40% the size of BERT and is 60% faster, while only being 3% less accurate.
(3) ALBERT adopts a cross-layer parameter sharing technique for 12 transformer encoder blocks and imports embedding

factorization between vocabulary and the hidden layer in order to reduce the parameter size of BERT.

(4) XLNet uses an innovative pretraining step; its generalized autoregressive pretraining method enables learning bidirectional contexts by maximizing the expected likelihood over all permutations of the factorization order, overcoming the issues caused by BERT's neglect of dependency between masked positions.

(5) Using a different architecture, ELECTRA trains 2 transformer models; a generator replaces tokens in a sequence and a discriminator tries to identify which tokens were replaced by the generator, instead of training on the MLM task.

Custom tokenizer

A tokenizer must be used to convert samples into the format that is expected by the transformer block of a language. In our study, such a tokenizer is obtained by training the language's default tokenizer on our custom corpus. Once trained, the tokenizer can capture any sample represented by a sentence consisting of 6-mer DNA words and a textual description of taxonomic lineage.

After tokenization, each input sample is represented by a list of tokens, starting and ending with special tokens [CLS] and [SEP], respectively, and padded to a length of 100 using padding tokens [PAD].

Model pretraining

The BERT language model is pretrained by performing unsupervised training of the MLM task on the custom corpus. Pretraining was conducted on the model using an architecture that is the same as *bert-base-uncase* but with setting the embedding size of input to 25,000 to match the vocabulary size of the corpus.

During training of the MLM task, 15% of all WordPiece tokens of a sample are selected at random as masking candidates. Of these, 80% are replaced by a special token [MASK], and 10% are replaced by a random token. Then the original tokens are predicted.

Pretraining was conducted using 8 epochs, a batch size of 64 per GPU, and a learning rate of 5e-4, which is achieved after 100 steps of warmup.

Model fine-tuning

Fine-tuning is performed for each of the 3 methylation site types separately, and so the processed training dataset is split into 3 training subsets, 6mA, 4mC, and 5hmC, listed in order of decreasing size. Each training subset is split into a training set and a validation set at a ratio of 8:2. The target model used to be fine-tuned depended on the subset's size. First, for the 6mA subset, we simply fine-tuned the pretrained language model that was trained on the custom corpus. Second, the 4mC fine-tuned model was then obtained by fine-tuning the 6mA fine-tuned model. Finally, the 5hmC fine-tuned model was obtained by fine-tuning the 4mC fine-tuned model. We fine-tune the fine-tuned models in this way to make the predictions more accurate on the smaller training subsets.

In all 3 cases, fine-tuning was performed using an early-stopping strategy, with a maximum of 32 epochs, a batch size of 64 per GPU, and a learning rate of 1e-5, which is achieved after 100 steps of warmup.

Multilanguage model

For each of the 3 types of methylation sites, 5 language models are trained and then the MuLan-Methyl framework integrates these, computing prediction probabilities that are obtained by averaging over the probabilities returned by the 5 models.

Interpretability of MuLan-Methyl

Transformer-based language models learn different and distant dependencies in the input, by virtue of the multihead self-attention mechanisms that are present in each encoding layer. For example, BERT contains 12 encoder layers containing 12 attention heads each. For 1 layer, the multihead self-attention can be described as

$$\text{MultiHead}\,(Q, K, V) = \text{Concat}\,(head_1, \cdots, head_n)\,W^O.$$

Here, the ith single attention head is computed as

$$\text{head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right),$$

$$\text{Attention}\,(Q, K, V) = \left(\frac{QK^T}{\sqrt{d_k}}\right)V,$$

where the projections represent parameter matrices $W_i^Q \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{model} \times d_v}$, and $W^O \in \mathbb{R}^{hd_v \times d_{model}}$. $Attention = \{a_{ij}\}$ is a scoring matrix, in which a_{ij} denotes the attention weight that the *Query* token t_i gets from then *Key* token t_j. This matrix is widely used for representing and exploring the binding between tokens [33, 50, 58].

While the language models are fine-tuned on the methylation sites prediction task, in the last layer of our model, a softmax function that acts as a classifier is placed on the special token [CLS] that is present at the beginning of each input sentence.

For each token, we sum the attention weights assigned to [CLS] over the 12 heads and regard this as the token's contribution to sample prediction.

To analyze the impact of the DNA sequence of a sample on the taxonomic lineage of the sample, we extract the attention weights assigned by the DNA tokens to the taxonomic hierarchy tokens.

Note that the WordPiece algorithm, which is used by the tokenizer employed in BERT, DistilBERT, and ELECTRA, provides word-wise tokens, so it makes sense to view the attention weights of tokens as contribution scores.

Here we conduct the above computation on the 3 fine-tuned models of each methylation type in MuLan-Methyl, respectively. The token importance score for MuLan-Methyl is obtained as the average score achieved on each of the 3 site-specific models.

Results

Comparison with encoders from language models

To illustrate the effectiveness of the approaches we proposed for training language models for DNA-based applications, we compared the encoder of our pretrained language model with that of both BERT and DNABERT (see Fig. 2A). Each pretrained language model was applied to 10% of the positive DNA sequences in the independent test set, obtaining their sentence representation by extracting the embedding of [CLS], with a dimension of (1, 768). The samples were then clustered and visualized using Uniform Manifold Approximation and Projection (UMAP) technique, colored by taxonomic lineage.

Since the original corpus that BERT is trained on does not explicitly includes DNA fragments, during tokenization, BERT will represent each DNA 6-mer with the special symbol [UNK], or cuts it into small pieces, unaware that it is a biological sequence. Consequently, the DNA sequences are embedded into a sparse space distribution by this encoder, with a poor ability to distinguish different species.

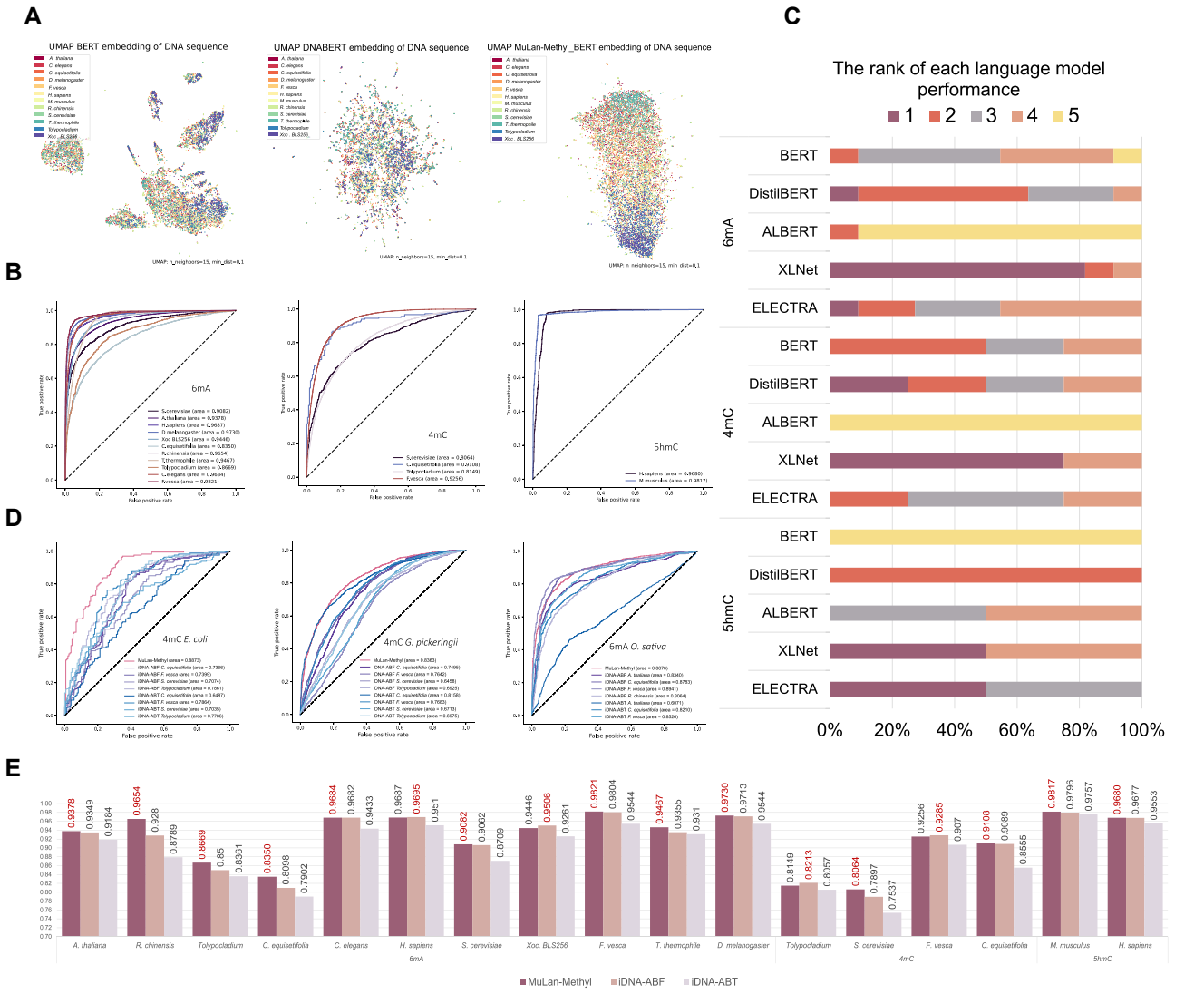


Figure 2: Model analysis and performance comparison of MuLan-Methyl. (A) UMAP clustering of sample representations encoded by different pretrained models: BERT, DNABERT, and MuLan-Methyl_BERT (from left to right). Samples are colored by taxonomic lineage. (B) For MuLan-Methyl predictions of the 3 methylation site types, 6mA, 4mC, and 5hmC, we present ROC curves for each of the 12 taxonomic types in the dataset. The AUC values are shown in brackets. (C) For each of the 3 methylation site types and each of the 5 language models, BERT, DistilBERT, ALBERT, XLNet, and ELECTRA, we show the ranking of models over all taxonomic lineages in terms of AUC scores. Moreover, the frequency with which each fine-tuned model appeared is indicated as the width of the corresponding block. (D) Comparison of MuLan-Methyl against 2 published methods, iDNA-ABF and iDNA-ABT, on an additional dataset that only contains taxonomic lineages that were not used to train the methods. From left to right, we show the ROCs obtained for the prediction of 4mC sites in *E. coli*, 4mC sites in *G. pickeringii* data, and 6mA sites in *O. sativa* L. data, respectively. (E) Comparison of MuLan-Methyl against iDNA-ABF and iDNA-ABT, on the iDMA-MS independent test set. We display the AUC scores for all 3 methods, for each of the 3 methylation site types and each of the 12 taxonomic lineages.

DNABERT is trained on genome sequences and has a better ability to capture DNA sequence features, as reflected in the absence of significant gaps between the distribution of DNA sequence representation obtained by its encoder. In the UMAP visualization, colors representing different taxonomic lineages appear to be randomly distributed.

In comparison, the MuLan-Methyl-BERT encoder is better at identifying DNA fragments and differentiating sequences by taxonomic lineage. Colors representing different taxonomic lineages exhibit a gradient from top to bottom. This suggests that pretraining the language model using a custom corpus that contains both DNA 6-mers and taxonomic lineages significantly improves the model's ability to capture potential information in this application scenario.

Comparison with single language submodels

The MuLan-Methyl framework uses 5 language models. In this section, we establish that the average prediction probability of this integrated approach is better than using any of the individual submodels, by comparing model performance using area under the curve (AUC) values. In summary, MuLan-Methyl outperforms the submodels, displaying the highest AUC across different taxonomic lineages and for each methylation site type.

In more detail, for 6mA site prediction, MuLan-Methyl is most beneficial when predicting on *Tolypocladium* and *S. cerevisiae*, with an AUC gain of 1.4% over the AUC calculated by ALBERT, which was the best-performing submodel. The average increase of AUC compared to the taxonomic lineage–specific best submodel is 0.7%. For 4mC site prediction, the average gain of AUC computed

from MuLan-Methyl is 0.8%, where the biggest improvement using MuLan-Methyl happened on *S. cerevisiae*, with an AUC increase of 1.1% over XLNet, the best submodel for this taxonomic lineage. MuLan-Methyl performs as well as taxonomic lineage–specific submodels at identifying 5hmC sites on both of the genomes. Moreover, we assessed the performance of MuLan-Methyl for each methylation site type and report on this for each taxonomic lineage using multiple metrics, including accuracy, F1-score, recall, area under the precision-recall curve (AUPR) and AUC (see Tables 1–3), as well as their receiver operating characteristic (ROC) curve (see Fig. 2B).

For each of the 3 methylation site types and for each of the 5 submodels included in MuLan-Methyl, we evaluated the performance of submodels on the corresponding independent test set. For each of the 12 taxonomic lineages, we ranked the given submodels based on their AUC values. Also, we determined the occurrence frequency of each submodel at each rank. This is shown in Fig. 2C and Supplementary Fig. S1.

We observed that XLNet usually shows better AUC than the other submodels for predicting 6mA sites, ranked first for 9 lineages. In contrast, ALBERT performs very poorly.

XLNet also performed best in 4mC site predictions, achieving the highest AUC on 3 of 4 taxonomic lineages. The lowest AUC from 4 taxonomic types concentratedly results from ALBERT. XLNet and ELECTRA performed best on the 5hmC site; BERT performs worst.

Comparison with existing methods

To demonstrate the advantage of MuLan-Methyl over existing methods, we compared the method against iDNA-ABF and iDNA-ABT, 2 state-of-the-art methods, that are both able to predict methylation sites for all 3 types, across different taxonomic lineages. (Note that all 3 frameworks were trained on the same training dataset, provided by iDNA-MS.) For this, we used the iDNA-MS independent test set, which is considered a benchmark dataset. We report the AUC scores in Fig. 2E, and more comprehensive evaluation metrics are displayed in Supplementary Table S2.

In this study, MuLan-Methyl outperforms the other 2 methods on 13 of 17 combinations of methylation types and taxonomic lineages. First, for 6mA site prediction, MuLan-Methyl improves over the other methods by between 0.02% and 3.74% AUC, whereas for *R. chinensis*, *C. equisetifolia*, *Tolypocladium*, and *T. thermophile*, the improvement is by more than 1%. Second, for 4mC site prediction, our method shows an increase of 1.67% and 0.02% AUC, on *S. cerevisiae* and *C. equisetifolia*, respectively. Finally, for 5hmC site prediction, our method shows an increase of 0.21% and 0.03% on *M. musculus* and *H. sapiens*, respectively.

The iDNA-ABF method has higher AUC scores in the remaining 4 cases—namely, for 6mA site prediction on *H. sapiens* and *Xoc.* BLS256, with an improvement of 0.08% and 0.6%, and for 4mC site prediction on *Tolypocladium* and *F. vesca*, with an improvment of 0.64%, and 0.3%, respectively, over MuLan-Methyl. A cursory comparison suggests that MuLan-Methyl and iDNA-ABF have similar reported runtimes (albeit using different GPUs), whereas iDNA-ABT runs about 10 times faster.

Explainability of MuLan-Methyl aids motifs discovery

To assess the contribution of each token toward correct methylation site detection, we use the average attention weight assigned by each token to [CLS] in the fine-tuned submodel, based on the positive sample from the independent test set.

The importance scores of each position in a DNA sequence has a Gaussian distribution across 17 different combinations of methylation site types and taxonomic lineages (see Fig. 3D–F and Supplementary Fig. S3). Positions of higher importance are concentrated around the center of the samples, and the central position always has high significance.

This observation supports the rationale used for constructing the iDNA-MS dataset—namely, to use, as positive samples, DNA segments of length 41 that are each centered on an experimentally verified methylation site. It also suggests the existence of DNA motifs that are closely associated with DNA methylation.

We also observed, for all 17 combinations, that the importance score starts low and then reaches a local maximum at position ±15. It then steadily increases from ±16 to the center of each sample (of length 41). This suggests that 41 is an ideal sample length for methylation detection, neither wasting resources to store unimportant positions nor missing important sequence.

The 6-mers with high importance may be considered to be DNA methylation "motifs" (see Fig. 3A–C and Supplementary Fig. S2). For a fixed taxonomic lineage, the 3 different methylation site types each have different motifs. However, for a fixed methylation site type, some motifs occur across different taxonomic lineages.

For example, the motif `CGAAGT` is important for 6mA methylation for several taxonomic lineages—namely, *S. cerevisiae*, *Tolypocladium*, and *Xoc.* BLS256. Note that the former 2 are eukaryotes, whereas the latter is a bacteria. Moreover, for 5hmC methylation, *H. sapiens* and *M. musculus* share many motifs. Similarly, for 4mC methylation, *C. equisetifolia* and *F. vesca* share many motifs.

Explainability of MuLan-Methyl reveals relationships between DNA sequence and taxonomic lineage

Integrating DNA sequences with taxonomic lineage as an explicit feature adds information and thus increases detection accuracy. Moreover, during fine-tuned model prediction, the association between DNA sequence and taxonomy can be measured by extracting the attention weights assigned from DNA tokens to the tokens that represent taxonomic lineage (see Fig. 3G–I and Supplementary Fig. S4).

The impact of DNA sequence on taxonomic lineage varies across the 17 combinations of methylation site types and taxonomic lineages. Generally, sequence locations that determine taxonomic lineage are concentrated around the center of samples, where the discussed DNA methylation-associated motifs are also clustered.

Of the 8 taxonomic ranks used to specify taxonomic lineage, the highest (kingdom) and lowest rank (species), in particular, are assigned larger attention weights by a wide range of positions in the sequence.

However, not all combinations follow this rule. For example, the impact of DNA sequence on species is weaker than on genus and family for the combinations 6mA + *D. melanogaster* and 5hmC + *M. musculus*. On combinations 6mA + *R. chinensis*, 6mA + *S. cerevisiae*, 6mA + *C. elegans*, 4mC + *S. cerevisiae*, and 5hmC + *H. sapiens*, we observed that the high scores assigned to the taxonomy lineages are quite sparsely distributed over the different ranks.

These observations demonstrate that the explainability of MuLan-Methyl can shed light on the relationships between DNA sequences and taxonomic lineage.

Table 1: MuLan-Methyl prediction performance on 6mA sites

Lineage	AUC	Accuracy	F1	Recall	AUPR
T. thermophile	0.9467	0.8840	0.8923	0.9611	0.9321
A. thaliana	0.9378	0.8649	0.8615	0.8401	0.9423
H. sapiens	0.9687	0.9077	0.9068	0.8975	0.9721
Xoc. BLS256	0.9446	0.8742	0.8712	0.8511	0.9421
D. melanogaster	0.9730	0.9276	0.9275	0.9258	0.9761
C. elegans	0.9684	0.9131	0.9138	0.9219	0.9674
C. equisetifolia	0.8350	0.7590	0.7481	0.7158	0.8492
S. cerevisiae	0.9082	0.8325	0.8233	0.7802	0.9198
Tolypocladium	0.8669	0.7895	0.7824	0.7567	0.8730
F. vesca	0.9821	0.9407	0.9403	0.9336	0.9831
R. chinensis	0.9654	0.9164	0.9167	0.9197	0.9691

Table 3: MuLan-Methyl prediction performance on 5hmC sites

Lineage	AUC	Accuracy	F1	Recall	AUPR
M. musculus	0.9817	0.9649	0.9651	0.9685	0.9782
H. sapiens	0.9680	0.9484	0.9500	0.9787	0.9485

Table 2: MuLan-Methyl prediction performance on 4mC sites

Lineage	AUC	Accuracy	F1	Recall	AUPR
C. equisetifolia	0.9108	0.8333	0.8272	0.7978	0.9221
F. vesca	0.9256	0.8522	0.8554	0.8739	0.9144
S. cerevisiae	0.8064	0.7376	0.7253	0.6926	0.8215
Tolypocladium	0.8149	0.7380	0.7285	0.7031	0.8089

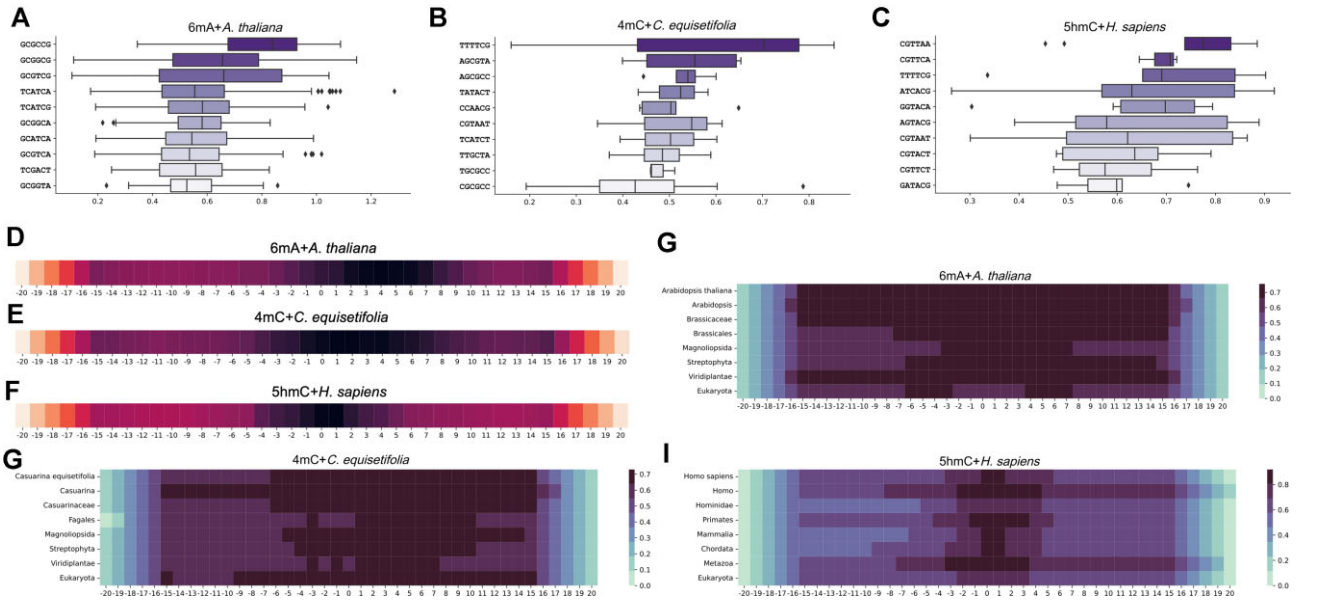


Figure 3: Interpretation of MuLan-Methyl by attention weights resulting from transformer self-attention mechanism. (A–C) Boxplots show the distribution of attention weights for the ten 6-mer of highest average importance scores, for the combinations 6mA + *A. thaliana*, 5mC + *C. equisetifolia* and 5hmC + *H. sapiens*, respectively. (D–F) We indicate the importance score for each position in the DNA sequences of length 41, obtained by merging 6-mer fragments, for the same 3 combinations listed above, respectively. (G–I) For each taxonomic rank of a lineage, we indicate the attention weight assigned by MuLan-Methyl to each position of the sequence for generating the taxon of the given rank, for the same 3 combinations listed above, respectively.

Performance on the external dataset

MuLan-Methyl was trained on 17 combinations of DNA methylation site types and taxonomic lineages. Fine-tuned models aim at performing well on input whose distribution is consistent with the training dataset but are not guaranteed to perform well on other data.

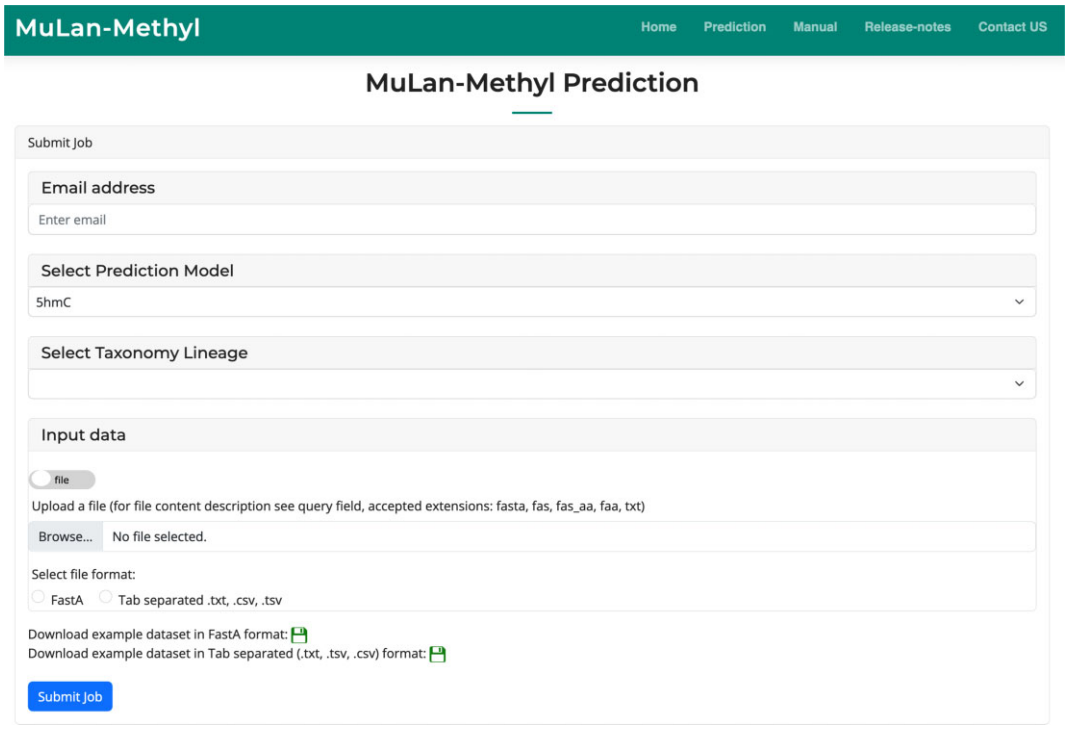


Figure 4: The Mulan-Methyl server hosted at [59] allows upload of DNA sequences and will perform methylation prediction along the whole sequence.

To explore the performance of MuLan-Methyl on other data, we applied the approach to an external dataset that contains 3 combinations of methylation types and taxonomic lineages—namely, 4mC + *E. coli*, 4mC + *G. pickeringii*, and 6mA + *O. sativa* L. Note that these 3 taxonomic lineages do not appear in the iDNA-MS datasets.

For the sake of comparison, we also calculated predictions using the servers provided by iDNA-ABF and iDNA-ABT. Since both approaches provide independent models for each combination, we ran all taxon-wise models for 4mC site detection and the appropriate ones for 6mA site detection.

MuLan-Methyl performed much better than the other 2 models on the 4mC + *E. coli* combination, achieving an AUC of 0.89, more than 10% better than the others. Our method also performed best on the 4mC + *G. pickeringii* combination, with an advantage of 2.05% over iDNA-ABT (using its *C. equisetifolia* model). On the third combination, 6mA + *O. sativa* L, MuLan-Methyl performed slightly worse (0.65%) than iDNA-ABF (using its *F. vesca* model). See Fig. 2D.

MuLan-Methyl server

We provide an implementation of MuLan-Methyl as a web server (see Fig. 4). Like other deep learning–based methylation services, this allows the user to upload DNA samples of length 41 and select the closet taxonomic lineage and the type of methylation site of interest. The uploaded samples will then be classified as methylation sites or not.

We also allow upload of longer DNA sequences, and in this case, the server will provide a list of all methylation sites that are predicted in the uploaded sequence.

To implement this extended functionality, we first extract all samples of length 41 that are centered on a nucleotide of the appropriate type (e.g., C when predicting 4mC or 5hmC sites) and then perform Mulan-Methyl prediction on these. The predicted positive samples are then filtered by feature importance analysis to resolve overlapping predictions. In more detail, we only retain samples for which the importance scores are highest at the center of the sample. Output is the list of all predicted methylation positions.

Discussion

Previous studies have focused on adapting BERT to specific biological tasks using the pretrain and fine-tune paradigm, with the aim of applying this popular NLP approach to tasks in genomics, phylogenetics, and other areas of computational biology.

However, BERT is not the only transformer-based language model, and it is important to choose the best model for a given task. Our proposed framework, MuLan-Methyl, consists of 5 transformer-based language models for identifying 3 types of DNA methylation sites across several taxonomic lineages, including both Eukaryota and Bacteria. With this work, we extend the list of transformed-based language models that have been successfully adapted to tasks involving biological sequences.

Each submodel in MuLan-Methyl is pretrained and fine-tuned on the training dataset, which then collectively predicts methylation sites on an independent test dataset. The performance of MuLan-Methyl was evaluated by multiple metrics and in comparison with 2 existing approaches, and the method showed very good performance.

Our study also indicates that models with enhanced algorithms in the pretraining step, such as XLNET, and models with fewer parameters and less memory consumption, such as DistilBERT, are more appropriate than BERT when storage or computational resources are limited.

In contrast to other biological domain adaption language models, the custom corpus that we trained MuLan-Methyl on contains multimodal data, consisting of both DNA sequences from iDNA-MS and taxonomy lineage in text format from the NCBI and GTDB taxonomies. To the best of our knowledge, MuLan-Methyl is the first language model framework to take taxonomy information into consideration.

This improves model accuracy and feature contribution analysis. The DNA methylation motifs found by MuLan-Methyl greatly benefited from the self-attention mechanism of transformer structure. In addition, the attention weights assigned to taxonomic lineages by DNA sequences help to analyze the relationship between nucleotide sequences and taxonomy lineage.

Previous approaches build a separate classifier for each taxonomic lineage and each methylation site type, giving rise to 17 different classifiers, for the data used here. In contrast, MuLan-Methyl considers taxonomic lineage as a feature and so only gives rise to 3 classifiers, one for each type of methylation site.

This study demonstrates that BERT is not the only choice when one wants to adapt a transformer-based language model to a specific domain; one should also consider its variants. It also shows that integrating multiple language models can offset the deficiencies of the individual models, to some extent, so as to obtain an improved ensemble prediction performance.

In conclusion, we have proposed a framework that integrates 5 popular NLP approaches to solve an important biological problem. MuLan-Methyl can be used to detect DNA methylation sites reliably for DNA sequences of 41 bp length from known taxonomic lineages, especially when closely related to the lineages involved in training, with slightly better performance than current state-of-the-art methods.

In practical applications, the input will usually be a chromosome or a set of assembled contigs, and the desired output will be a list of putative methylation sites. To address this, we designed a 2-step validation strategy for false-positive rate controlling and implemented it in the MuLan-Methyl server.

Availability of Source Code and Requirements

Project name: MuLan-Methyl
Project homepage: http://ab.cs.uni-tuebingen.de/software/mulan-methyl
Code GitHub: https://github.com/husonlab/mulan-methyl
Operating system(s): Platform independent
Programming language: Code: Python (3.10.6); WebServer: HTML5, Bootstrap, PHP (7.2.24), JavaScript
Other requirements: For WebServer MySQL (5.7.39)
License: Apache-2.0.
Any restrictions to use by nonacademics: None
Biotools ID: https://bio.tools/MuLan-Methyl
RRID: SCR_023591

Additional Files

Supplementary Fig. S1. Heatmap of Kendall tau distance matrix for exploring the ranking correlation, where the rank is obtained by comparing AUC with the other submodels on each methylation site type of each submodel.
Supplementary Fig. S2. Boxplot of top 10 tokens with the highest average attention scores for the remaining combinations of methylation types and taxonomic lineage.
Supplementary Fig. S3. Heatmap of the average importance score for each position of a 41-bp DNA sequence obtained by merging 6-mer fragments for each remaining combination of methylation types and taxonomic lineage.
Supplementary Fig. S4. Heatmap of the impact between DNA sequence and its taxonomy lineage for each remaining combination of methylation types and taxonomic lineage.
Supplementary Table S1. iDNA-MS dataset statistics.
Supplementary Table S2. The comparison of model performance on the iDNA-MS independent test set between MuLan-Methyl and its submodels, as well as with the previous studies.

Abbreviations

4mC: 4-methylcytosine; 5hmC: 5-hydroxymethylcytosine; 6mA: 6-methyladenine; AUC: area under the curve; BERT: bidirectional encoder representation from transformers; GTDB: Genome Taxonomy Database; AUPR: area under the precision-recall curve; NLP: natural language processing; MLM: masked language modeling; NCBI: The National Center for Biotechnology Information; ROC: receiver operating characteristic.

Data Availability

The benchmark dataset used in this study is available here [31, 55]. The processed dataset used for training MuLan-Methyl and the source code are available at [60]. A web server implementing the MuLan-Methyl approach is freely accessible at [59]; see also [61], RRID: SCR_023591. All supporting data and materials are available in the *GigaScience* GigaDB database [62].

Competing Interests

The authors declare that they have no competing interests.

Funding

We acknowledge support by the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A532B, 031A533A, 031A533B, 031A534A, 031A535A, 031A537A, 031A537B, 031A537C, 031A537D, 031A538A). We acknowledge support by the Open Access Publishing Fund of the University of Tübingen.

Authors' Contributions

W.Z. and D.H.H. conceived the project. W.Z. collected and processed the dataset for the project, designed and implemented the architecture and algorithms of MuLan-Methyl, and conducted model analysis. A.G. and W.Z. designed and implemented the web server of MuLan-Methyl. W.Z., D.H.H., and A.G. contributed to the manuscript.

References

1. Robertson KD, Wolffe AP. DNA methylation in health and disease. Nat Rev Genet 2000;1(1):11–9.
2. Moore LD, Le T, Fan G. DNA methylation and its basic function. Neuropsychopharmacology 2013;38(1):23–38.
3. Armstrong MJ, Jin Y, Allen EG, et al. Diverse and dynamic DNA modifications in brain and diseases. Hum Mol Genet 2019;28(R2):R241–53.
4. Tourancheau A, Mead EA, Zhang XS, et al. Discovering multiple types of DNA methylation from bacteria and microbiome using nanopore sequencing. Nat Methods 2021;18(5):491–8.
5. O'Brown ZK, Boulias K, Wang J, et al. Sources of artifact in measurements of 6mA and 4mC abundance in eukaryotic genomic DNA. BMC Genomics 2019;20(1):1–15.
6. Ito S, Shen L, Dai Q, et al. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. Science 2011;333(6047):1300–3.

7. Bilyard MK, Becker S, Balasubramanian S. Natural, modified DNA bases. Curr Opin Chem Biol 2020;57:1–7.
8. Rauluseviciute I, Drabløs F, Rye MB. DNA methylation data by sequencing: experimental approaches and recommendations for tools and pipelines for data analysis. Clin Epigenet 2019;11(1):1–13.
9. Ye P, Luan Y, Chen K, et al. MethSMRT: an integrative database for DNA N6-methyladenine and N4-methylcytosine generated by single-molecular real-time sequencing. Nucleic Acids Res 2016;45(D1):D85–89.
10. Xu H, Jia P, Zhao Z. Deep4mC: systematic assessment and computational prediction for DNA N4-methylcytosine sites by deep learning. Brief Bioinform 2021;22(3):bbaa099.
11. Zeng R, Cheng S, Liao M. 4mCPred-MTL: accurate identification of DNA 4mC sites in multiple species using multi-task deep learning based on multi-head attention mechanism. Front Cell Dev Biol 2021;9:664669.
12. Liu Q, Chen J, Wang Y, et al. DeepTorrent: a deep learning-based approach for predicting DNA N4-methylcytosine sites. Brief Bioinform 2021;22(3):bbaa124.
13. Hasan MM, Manavalan B, Shoombuatong W, et al. i4mC-Mouse: improved identification of DNA N4-methylcytosine sites in the mouse genome using multiple encoding schemes. Comput Struct Biotech J 2020;18:906–12.
14. Jin J, Yu Y, Wei L. Mouse4mC-BGRU: Deep learning for predicting DNA N4-methylcytosine sites in mouse genome. Methods 2022;204:258–62.
15. Zulfiqar H, Sun ZJ, Huang QL, et al. Deep-4mCW2V: a sequence-based predictor to identify N4-methylcytosine sites in Escherichia coli. Methods 2022;203:558–63.
16. Liang Y, Wu Y, Zhang Z, et al. Hyb4mC: a hybrid DNA2vec-based model for DNA N4-methylcytosine sites prediction. BMC Bioinformatics 2022;23(1):1–18.
17. Tran TA, Pham DM, Ou YY, et al. An extensive examination of discovering 5-methylcytosine sites in genome-wide DNA promoters using machine learning based approaches. IEEE/ACM T Comput Biol Bioinform 2021;19(1):87–94.
18. Cheng X, Wang J, Li Q, et al. BiLSTM-5mC: a bidirectional long short-term memory-based approach for predicting 5-methylcytosine sites in genome-wide DNA promoters. Molecules 2021;26(24):7414.
19. Li Z, Jiang H, Kong L, et al. Deep6mA: a deep learning framework for exploring similar patterns in DNA N6-methyladenine sites across different species. PLoS Comput Biol 2021;17(2):e1008767.
20. Rehman MU, Tayara H, Zou Q, et al. i6mA-Caps: a CapsuleNet-based framework for identifying DNA N6-methyladenine sites. Bioinformatics 2022;38(16):3885–91.
21. Zeng R, Liao M. 6mAPred-MSFF: a deep learning model for predicting DNA N6-methyladenine sites across species based on a multi-scale feature fusion mechanism. Appl Sci 2021;11(16):7731.
22. Liu M, Sun ZL, Zeng Z, et al. MGF6mARice: prediction of DNA N6-methyladenine sites in rice by exploiting molecular graph feature and residual block. Brief Bioinform 2022;23(3):bbac082.
23. Tsukiyama S, Hasan MM, Deng HW, et al. BERT6mA: prediction of DNA N6-methyladenine site using deep learning-based approaches. Brief Bioinform 2022;23(2):bbac053.
24. Tahir M, Hayat M, Ullah I, et al. A deep learning-based computational approach for discrimination of DNA N6-methyladenosine sites by fusing heterogeneous features. Chemometr Intell Lab Syst 2020;206:104151.
25. Le NQK, Ho QT. Deep transformers and convolutional neural network in identifying DNA N6-methyladenine sites in cross-species genomes. Methods 2022;204:199–206.
26. Tang X, Zheng P, Li X, et al. Deep6mAPred: a CNN and Bi-LSTM-based deep learning method for predicting DNA N6-methyladenosine sites across plant species. Methods 2022;204:142–50.
27. Hasan MM, Basith S, Khatun MS, et al. Meta-i6mA: an interspecies predictor for identifying DNA N6-methyladenine sites of plant genomes by exploiting informative features in an integrative machine-learning framework. Brief Bioinform 2021;22(3):bbaa202.
28. Chen J, Zou Q, Li J. DeepM6ASeq-EL: prediction of human N6-methyladenosine (m6A) sites with LSTM and ensemble learning. Front Comput Sci 2022;16(2):1–7.
29. Zhang Y, Liu Y, Xu J, et al. Leveraging the attention mechanism to improve the identification of DNA N6-methyladenine sites. Brief Bioinform 2021;22(6):bbab351.
30. Yang X, Ye X, Li X, et al. iDNA-MT: identification DNA modification sites in multiple species by using multi-task learning based a neural network tool. Front Genet 2021;12:663572.
31. Lv H, Dao FY, Zhang D, et al. iDNA-MS: an integrated computational tool for detecting DNA modification sites in multiple genomes. Iscience 2020;23(4):100991.
32. Yu Y, He W, Jin J, et al. iDNA-ABT: advanced deep learning model for detecting DNA methylation with adaptive features and transductive information maximization. Bioinformatics 2021;37(24):4603–10.
33. Jin J, Yu Y, Wang R, et al. iDNA-ABF: multi-scale deep biological language learning model for the interpretable prediction of DNA methylations. Genome Biol 2022;23(1):1–23.
34. Zheng Z, Le NQK, Chua MCH. MaskDNA-PGD: an innovative deep learning model for detecting DNA methylation by integrating mask sequences and adversarial PGD training as a data augmentation method. Chemometr Intell Lab Syst 2023;232:104715.
35. Lv H, Dao FY, Zhang D, et al. Supporting data for "iDNA-MS: An Integrated Computational Tool for Detecting DNA Modification Sites in Multiple Genomes." GigaScience Database. 2023. Http://dx.doi.org/10.5524/102395.
36. Devlin J, Chang MW, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT. Vol. 1. Association for Computational Linguistics.Minneapolis, Minnesota.2019;4171–86.
37. Ji Y, Zhou Z, Liu H, et al. DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. Bioinformatics 2021;37(15):2112–20.
38. Zhang Yz, Yamaguchi K, Hatakeyama S, et al. On the application of BERT models for NanoPore methylation detection. In: 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Houston, Texas: IEEE, 2021;320–7.
39. Seong HJ, Han SW, Sul WJ. Prokaryotic DNA methylation and its functional roles. J Microbiol 2021;59(3):242–8.
40. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Adv Neur Inf Process Syst 2017;30:5998–6008.
41. Zeng W, Gautam A, Huson DH. DeepToA: an ensemble deep-learning approach to predicting the theater of activity of a microbiome. Bioinformatics 2022;38(20):4670–6.
42. Sanh V, Debut L, Chaumond J, et al. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:191001108. 2019. https://arxiv.org/abs/1910.01108.
43. Liu Y, Ott M, Goyal N, et al. Roberta: a robustly optimized BERT pretraining approach. arXiv preprint arXiv:190711692. 2019. https://arxiv.org/abs/1907.11692.

44. Yang Z, Dai Z, Yang Y, et al. Xlnet: Generalized autoregressive pretraining for language understanding. Adv Neur Inf Process Syst 2019;32:5754–64.
45. Clark K, Luong MT, Le QV, et al. Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:200310555.2020. https://arxiv.org/abs/2003.10555.
46. Lan Z, Chen M, Goodman S, et al. Albert: a lite BERT for self-supervised learning of language representations. arXiv preprint arXiv:190911942. 2019. https://arxiv.org/abs/1909.11942.
47. Gururangan S, Marasović A, Swayamdipta S, et al. Don't stop pretraining: adapt language models to domains and tasks. arXiv preprint arXiv:200410964. 2020. https://arxiv.org/abs/2004.10964.
48. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;36(4):1234–40.
49. Conneau A, Lample G. Cross-lingual language model pretraining. Adv Neur Inf Process Syst 2019;32:7059–69.
50. Lupo U, Sgarbossa D, Bitbol AF. Protein language models trained on multiple sequence alignments learn phylogenetic relationships. Nat Commun 2022;13(1):6298.
51. Mock F, Kretschmer F, Kriese A, et al. Taxonomic classification of DNA sequences beyond sequence similarity using deep neural networks. Proc Natl Acad Sci 2022;119(35):e2122636119.
52. Yang M, Huang L, Huang H, et al. Integrating convolution and self-attention improves language model of human genome for interpreting non-coding regions at base-resolution. Nucleic Acids Res 2022;50(14):e81.
53. Schoch CL, Ciufo S, Domrachev M, et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database (Oxford) 2020;2020. http://doi.org/10.1093/database/baaa062.
54. Parks DH, Chuvochina M, Rinke C, et al. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. Nucleic Acids Res 2022;50.D785–94
55. iDNA-MS web server. 2020. http://lin-group.cn/server/iDNA-MS/download.html.
56. Chen W, Lv H, Nie F, et al. i6mA-Pred: identifying DNA N6-methyladenine sites in the rice genome. Bioinformatics 2019;35(16):2796–800.
57. Wolf T, Debut L, Sanh V, et al. Transformers: state-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Stroudsburg, Pennsylvania: Association for Computational Linguistics, 2020;38–45.
58. Yamada K, Hamada M. Prediction of RNA-protein interactions using a nucleotide language model. Bioinform Adv 2022;2(1):vbac023.
59. MuLan-Methyl web server. 2023. http://ab.cs.uni-tuebingen.de/software/mulan-methyl/.
60. GitHub repository of MuLan-Methyl. 2023. https://github.com/husonlab/mulan-methyl.
61. Biotools link of MuLan-Methyl. https://bio.tools/MuLan-Methyl.
62. Zeng W, Gautam A, Huson DH. Supporting data for "MuLan-Methyl—Multiple Transformer-Based Language Models for Accurate DNA Methylation Prediction." GigaScience Database. 2023. http://dx.doi.org/10.5524/102402.

Received: February 19, 2023. **Revised:** May 9, 2023. **Accepted:** July 18, 2023

B Manuscript 2

Title: Enhanced 5mC-Methylation-Site Recognition in DNA Sequences Using Token Classification and a Domain-specific Loss Function

Enhanced 5mC-Methylation-Site Recognition in DNA Sequences using Token Classification and a Domain-specific Loss Function

Wenhuan Zeng[1] **and Daniel H. Huson**[1,*]

[1]Algorithms in Bioinformatics, Institute for Bioinformatics and Medical Informatics, University of Tübingen, Sand 14, 72076, Tübingen, Germany
[*]daniel.huson@uni-tuebingen.de

ABSTRACT

DNA 5-methylcytosine modification has been widely studied in mammals and plays an important role in epigenetics. Several computational methods exist that attempt to determine the methylation state of a DNA sequence centered at a possible methylation site. Here, we introduce a novel deep-learning framework, MR-DNA, that predicts the methylation state of a single nucleotide located in a gene promoter region. The idea is to adapt the named-entity recognition approach to methylation-site prediction and to incorporate biological rules during model construction. MR-DNA has a stacked model architecture consisting of a pre-trained MuLan-Methyl-DistilBERT language model and a conditional random field algorithm, trained with a self-defined methyl loss function. The resulting fine-tuned model achieves an accuracy of 97.9% on an independent test dataset of samples. An advantage of this formulation of the methylation-site identification task is that it predicts on every nucleotide of a sequence of a given length, unlike previous methods that the predict methylation state of DNA sequences of a short fixed length. For training and testing purposes, we provide a database of DNA sequences containing verified 5mC-methylation sites, calculated from data for eight human cell lines downloaded from the ENCODE database.

Introduction

DNA methylation is an important epigenetic mechanism. Much attention has been given to 5-methylcytosine (5-mC) modifications in which cytosine is methylated at its fifth position. In mammals, this process is associated with a range of different biological processes, in particular gene regulation, gene silencing, X-chromosome inactivation, genomic imprinting and human cancer development[1–4].

Several experimental methods have been developed that aim at quantifying 5mC methylation, especially in mammalian genomes, such as bisulfite conversion second-generation sequencing (BS-seq)[5,6] or using third-generation sequencing technologies, namely Pacbio SMRT sequencing[7], Nanopore sequencing[8,9] and Pacbio circular consensus sequencing (CCS), particularly in repetitive genomic regions[10]. Several computational approaches based on machine-learning algorithms have been conceived to detect methylation state on data generated from the above-mentioned sequencing technologies (see Table 1). Most aim at predicting whether the central base of a short DNA sequence is methylated; all show reliable performance[11–14].

Previous studies have regarded DNA-methylation identification as a binary classification task. Most of them employ traditional machine learning approaches[11], classic deep learning techniques such as deep feed-forward neural networks (deepFNN)[11], bidirectional long short-term memory (BiLSTM) networks[12], or ensemble frameworks[13,14] are applied to a combination of features extracted from the DNA sequences. Also, transformer-based language models are trained on DNA sequences to provide a pre-trained model[15] that is then fine-tuned on a down-stream database for performing DNA-methylation identification[16]. These methods are all designed to predict whether a presented DNA sequence of fixed length 41 or others is methylated in the middle or not. Thus, these methods are not directly applicable to the main practical problem of interest, namely the prediction of individual methylation sites in a sequence.

Hence, there is a need for a machine-learning approach capable of making predictions for individual bases in DNA sequences of arbitrary length. To address this, here we model DNA sequences as "language", where each 3-mer is a token, and phrase methylation-site prediction as a token classification task, which is a fundamental task in natural language understanding (NLU). The named-entity recognition (NER) task is usually applied in information extraction (IE), for example, to locate characters, events, organizations, or locations[25,26], or to recognize other types of entities[27,28] in a text. This approach has been applied to biomedical problems that are focused on text[29–31], but has not yet been transferred to biological sequences.

Traditional statistical models[32–35] are gradually being superseded by transformer-based language models or other neural networks[31,36,37]. The stacked model architecture, which allows the stacking of additional layers, such as LSTMs and conditional

Table 1. Overview of work on the 5mC-methylation detection task in the human genome

Study	Database	Model architecture	Applied objective
Wang et al.[17]	Promoter sequences from the UCSC genome browser[18], iPromoter-5mC database[13].	Bidirectional Encoder Representations from Transformers (BERT)[19]	DNA fragment of 41 centered on a cytosine
Ni et al.[20]	PacBio CCS reads, Illumina and Nanopore data of human sample	Bidirectional Gated Recurrent Unit (GRU) with Bahdanau attention network	Provides both read-level (21-mer sequence that includes CpG in the center) and site-level prediction
Wang et al.[21]	iPromoter-5mC database	Transformer encoder and fully connected neural network	DNA fragment of 41 centered on a cytosine
Jia et al.[22]	iPromoter-5mC database	A stacked framework of densely connected convolutional networks (DenseNet) and bidirectional GRU with attention mechanism	DNA fragment of 41 centered on a cytosine
Jin et al.[16]	Human cell line	Pretrained DNABERT, fine-tuning using adversarial training method	DNA fragments of diverse length centered on a cytosine
Bonet et al.[23]	Nanopore data and BS data of Human cell line NA12878	Convolutional neural network (CNN)	Provides methylation calling on both read-level (length from 1 to 17) and site-level
Nyuyen et al.[11]	Human cell line	Machine learning techniques including XGBoost, random forest, deep forest, and deepFNN	DNA fragment of 41 centered on a cytosine
Cheng et al.[12]	iPromoter-5mC database	BiLSTM	DNA fragment of 41 centring the cytosine
Zhang et al.[13]	RRBS data in cell lines of lung cancer	Deep neural network	DNA fragment of 41 centered on a cytosine
Tian et al.[24]	WGBS data of H1 ESC, normal brain white matter, lung and colon tissue.	CNN	DNA fragment of 400 bps centered on a cytosine.

random fields (CRF), to a pre-trained transformer-based language model, has significantly improved the ability to capture dependencies between entities and corresponding labels[38–40].

As mentioned, previous approaches address the question of whether the middle base of a DNA sequence of length 41 is methylated or not. In our approach, we take advantage of the fact that NER assigns a label to each token and thus can predict methylation for individual bases in a given sequence. In addition, in order to better adapt general NER approaches to biological scenarios, we propose a novel loss function, methyl loss, that is based on the categorical cross-entropy loss function and uses known biological rules of 5-methylcytosine to help model training.

The MuLan-Methyl framework[41] consists of five transformer-based language models, each trained on the iDNA-MS database[42], a comprehensive dataset containing DNA methylation sequences for three methylation types and twelve taxonomic lineages. In this study, we use MuLan-Methyl-DistilBERT, one of five finetuned MuLan models, as the model encoder. We use this particular language model because it outperforms the other constituents of MuLan-Methyl for 5hmC-methylation site recognition in terms of both prediction performance and computational efficiency.

Here, we propose a new model architecture, MR-DNA (Methylation-site Recognition in DNA) that consists of MuLan-Methyl-DistilBERT and a conditional random fields (CRF) algorithm. The weights of MR-DNA are updated using methyl loss while training, which enhances the model prediction ability.

To evaluate the performance of MR-DNA, we created a database of DNA sequences, each of length 1000 bp, that correspond to gene promoter regions from eight human cell line projects in the ENCODE database[43,44], with differing numbers of experimentally-verified methylation sites. We further cut each such DNA sequence into small pieces according to a set stride during model training. The workflow of database construction is shown in Fig. 1A. The performance of the fine-tuned MR-DNA model was evaluated under the exact-match evaluation criterion[26].

Our main contributions are:

- This is the first study to phrase methylation-site recognition as a named entity recognition (NER) task.
- We propose a new loss function, methyl loss, that can effectively overcome the impact of skewed data distribution on the model.
- We present a novel framework, MR-DNA, that combines the statistical modelling method CRF with the fine-tuned MuLan-Methyl-DistilBERT model and is trained using methyl loss, showing promising performance on the test dataset.
- We provide a database of DNA gene-promoter sequences (each 1000 bp length) containing varying numbers of verified methylation sites.
- The MR-DNA method aims at recognizing the methylation state of each individual base in a given DNA sequence.

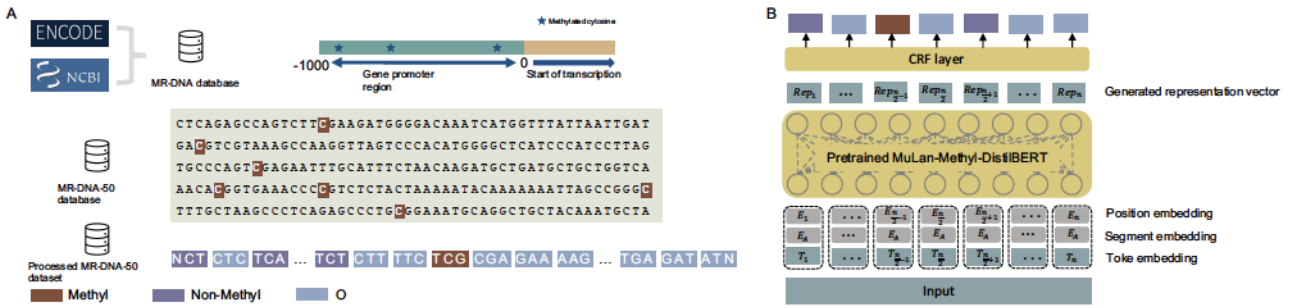


Figure 1. A. Database creation. For eight human cell lines in ENCODE, we extracted 1000 bp upstream gene promoter regions, and annotated cytosines by their reported methylation state. We call this the MR-DNA database. The MR-DNA-50 database was extracted from these sequences using a window size of 50 and stride of 25. The *processed* MR-DNA-50 database was obtained by extracting 3-mers from each sequence in the MR-DNA-50 database and labeling each such 3-mer as methylated, if its central nucleotide is so labeled.
B. Model structure. MR-DNA is a stacked model that uses a CRF layer on top of pretrained MuLan-Methyl-DistilBERT, assigning the most probable category to each token.

Results

Training MR-DNA

MR-DNA is implemented in Python 3.10, using the Pytorch, Huggingface[45], and pytorch-crf packages. It was run on a Linux Virtual Machine (Ubuntu 20.04 LTS) equipped with 4 GPUs, provided by de.NBI (flavor: de.NBI RTX6000 4 GPU medium). MR-DNA was trained on the above-mentioned processed MR-DNA-50 training and validation datasets using the methyl loss function, configured with a 2e-5 learning rate, early-stopping at 9 epochs, and a batch size of 256 for each GPU.

Experiment evaluation

Our study uses the exact-match evaluation criterion to assess MR-DNA performance instead of the seqeval[46] framework designed for sequence labelling evaluation. We use the following standard definitions.

$$\begin{aligned} Precision &= \frac{TP}{TP+FP}, \\ Recall &= \frac{TP}{TP+FN}, \\ F1-score &= \frac{TP}{TP+\frac{1}{2}(FP+FN)}, \\ Accuracy &= \frac{TP+TN}{TP+FN+TN+FP}. \end{aligned} \tag{1}$$

Here, TP, FP, TN, and FN are the true positive, false positive, true negative and false negative counts, respectively. Because the annotations in our database have a highly imbalanced distribution, We use the macro average method to compute the F1-score, precision, and recall, in order to gain a better understanding of the model's overall performance on imbalanced datasets. This is computed by taking the arithmetic mean of all the class-wise metrics. Each macro-average metric $metric \in \{Precision, Recall, F1-scores\}$ is computed as follows:

$$Macro\ metric = \frac{1}{|N|}\sum_{n\in N} metric_n \tag{2}$$

where $N = \{n_1, n_2, \cdots, n_j\}$ is the set of classes that appear in the datasets.

Performance of MR-DNA on multiple datasets with different lengths

First, we investigated the ability of MR-DNA, trained with the methyl loss function, to recognize methylation sites on the processed MR-DNA-50 test data. Here, each sample is a sentence of 50 whitespace-separated entities (3-mers), together with one of three possible annotations ("Methyl", "Non-Methly" or "O"). We used accuracy, F1-score, recall and precision to quantify the ability of MR-DNA to predict the label of each entity (3-mer). As this is a multiple classification task, we used macro averages to calculate the performance metrics.

Secondly, to explore the effect of sequence length on classification performance, we trained MR-DNA on additional processed MR-DNA-100 and MR-DNA-200 training datasets, as well, using sentences of length 100 and 200, respectively. We observed that MR-DNA with methyl loss trained on MR-DNA-50 outperforms the models trained on the two datasets of longer sequence lengths regarding F1-score, recall, and precision (See Table. 2).

Entities with annotation "O" are easier to identify correctly because they do not have a "C" at the center. To address this, in addition to evaluating model performance when distinguishing between three possible annotations, which are "Methyl", "Non-Methly" or "O", we also evaluated model performance distinguishing between only two annotations, "Methyl" and "Non-Methly" (see Table S1).

Methyl loss enhances MR-DNA performance

We evaluated the usefulness of the proposed loss function on each of the three datasets, MR-DNA-50, MR-DNA-100 and MR-DNA-200, comparing its performance with that of the categorical cross-entropy loss, which we view as a baseline. The datasets statistics of MR-DNA-100 and MR-DNA-200 are shown in Figure S1 and S2.

As shown in Table. 2 and Table S1, for each dataset, MR-DNA trained with the methyl loss outperforms the baseline, especially according to the macro F1-score. Confusion matrices (see Fig. 2) for MR-DNA-50, -100 and -200, show that MR-DNA trained with the methyl loss function on the former outperforms the model trained on the two latter, for the "Methyl" category. Additionally, the model trained with methyl loss has higher sensitivity than the model trained with cross-entropy on the same dataset.

Table 2. Performance evaluation. MR-DNA was trained on three datasets with different sequence lengths. The performance of the methyl loss is compared against that of models trained using a cross-entropy loss function and a focal loss function. For each test dataset, we report accuracy, macro F1-score, macro precision and macro recall. Best values are shown in bold.

Dataset	Loss function	Accuracy	Macro F1-score	Macro precision	Macro recall
MR-DNA-50	Cross-entropy	0.9780	0.8577	0.8265	0.9206
	Focal loss	0.9751	0.8557	0.8205	**0.9518**
	Methyl loss	**0.9787**	**0.8659**	**0.8319**	0.9390
MR-DNA-100	Cross-entropy	0.9845	0.8323	0.7985	0.9033
	Focal loss	0.9749	0.8046	0.7708	**0.9575**
	Methyl loss	**0.9851**	**0.8341**	**0.8012**	0.8996
MR-DNA-200	Cross-entropy	**0.9895**	0.7994	**0.7699**	0.8606
	Focal entropy	0.9783	0.7684	0.7394	**0.9569**
	Methyl loss	0.9893	**0.8003**	0.7695	0.8669

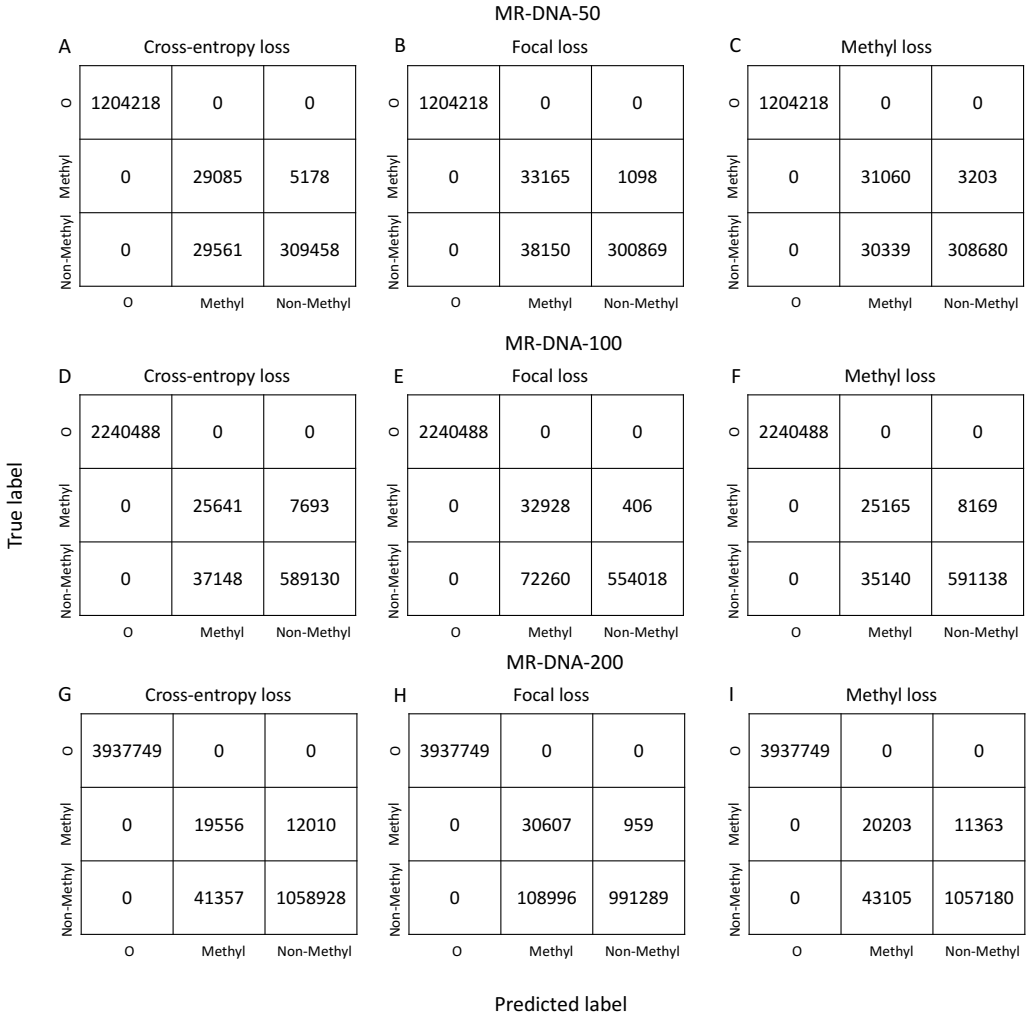


Figure 2. Confusion matrix visualization and comparison of the model trained on (A-C) the MR-DNA-50 training dataset, (D-F) MR-DNA-100 training dataset and (G-I) MR-DNA-200 training dataset using a categorical cross-entropy loss function, a focal loss function, and a methyl loss function, respectively. Performance evaluation was conducted on the corresponding test dataset.

Methyl loss is less complex and more stable

The focal loss function is widely used in scenarios that train models on an imbalanced dataset. Two hyperparameters, α and γ, are designed to assign different weights to categories according to their data distribution and control the weight according to the difficulty of predicting the sample, respectively. In comparison, the methyl loss function has one hyperparameter λ, used to assign weight according to the commonness of the 3-mer under its label.

We compared the performance of a model trained using the methyl loss function and one trained using the focal loss function on each of the three datasets. The α and γ values for focal loss are $[0.1, 0.7, 0.2]$ and 5, respectively.

For all three datasets, the model with focal loss function performs worse than the one with methyl loss with respect to accuracy, macro f1-score, and macro precision (see Table. 2). Using the hyperparameter α, the focal loss function puts more attention on the least category 'Methyl', however, it also leads to a poor false positive rate for this category as shown in Fig. 2. In comparison, the methyl loss function is more stable, improving the model performance in all categories.

Notably, as we mentioned above, we computed evaluation metrics in two ways, one is obtained by considering models' overall performance in identifying three annotations of entities (See Table. 2 and Table. S1), and another is only considering models' performance in identifying "Methyl" and "Non-Methly" entities (Table. S1). Comparing these two tables, the values in Table. S1 are comprehensively lower than the value in Table. 2, which is caused by excluding the promising performance in identifying "O" entities. however, the relative differences between the performance of the different models are consistent.

Stability and flexibility of MR-DNA

Our study is the first one that aims to predict the methylation state of every base in a given DNA sequence. Its novelty makes it difficult to compare to previous studies, which aim at predicting the global methylation state of DNA sequences with fixed length, where the target cytosine lies in the center of the sequences. To overcome this, we modified the prediction format output by MR-DNA to allows comparisons against previous studies using the iPromoter-5mC database[13] as the benchmark database, and applying our models to the independent test dataset of the benchmark database, namely the iPromoter-5mC test dataset.

The iPromoter-5mC test dataset consists of DNA sequences with a length of 41, labelled 'positive' if the centerd cytosine is methylated, and 'negative', otherwise. The DNA sequences in the iPromoter-5mC test dataset lack methylation state annotations for all bases except for the one located in the center.

To allow model comparison, the iPromoter-5mC test dataset was processed by converting the DNA sequence to a sequence that consists of 41 consecutive 3-mers, using the same processing approaches as the one used for our database. We applied the MR-DNA-50, -100, -200 that were trained on the corresponding MR-DNA database directly to the processed iPromoter-5mC test dataset for predicting each DNA sequence's global methylation state, respectively. MR-DNA is not trained on the training dataset of the iPromoter-5mC database since its data annotation is inadequate for training the MR-DNA model. Additional, we converted the output format of the NER task to the output format of the binary classification task, in which each DNA sequence is assigned '1' if the central 3-meris predicted as 'Methyl', and '0', otherwise.

We calculated accuracy, f1-score and specificity to evaluate our models' performance on the iPromoter-5mC test dataset, and for comparing against models from previous studies which are trained and tested on the iPromoter-5mC training dataset. Some evaluation metrics were computed using the weighted-average approach due to the imbalanced distribution of the iPromoter-5mC test dataset.

As shown in Fig. 3, among the MR-DNA models, the MR-DNA-100 performs best regarding recall, specificity, and accuracy. However, in comparison to the best performance model in previous studies, its accuracy, specificity, and recall are 10%, 5.4%, and 2.4% lower, respectively. We emphasise that the models we compared against were trained on the training set that corresponds to the iPromoter-5mC test dataset, while the MR-DNA models were trained on our own dataset, which was collected from different resources than the iPromoter-5mC test dataset. So, despite the difference between our training dataset and iPromoter-5mC test dataset, our models, especially MR-DNA-100, shown comparable performance on the iPromoter-5mC test dataset.

Discussion

The study of 5mC-modifications in mammalian genomes helps the understanding of several biological processes, such as gene expression control. Previous computational approaches based on machine-learning algorithms for identifying methylation sites are only able to categorize DNA sequences of a predetermined length rather than detect methylation sites of a single nucleotide, resulting in limited practical use.

Here, we presented MR-DNA, a novel framework based on natural language understanding technology, which transforms a naive classification task for DNA sequences into a NER problem with the ability to identify the methylation state of each entity. We enhanced MR-DNA's classification ability by using a custom loss function, methyl loss, to update model weights during training. The sensitivity of classifying the minority category is improved by optimizing the loss function, so as to put more attention on challenging data by adding an exponential based on the categorical cross-entropy loss function. This work

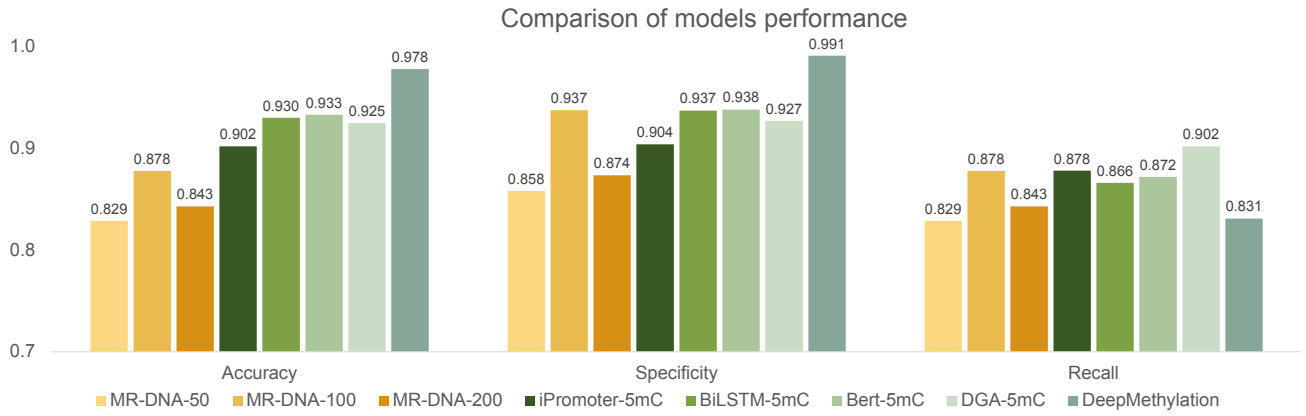


Figure 3. Comparison of model performance between MR-DNA and previous studies. The performance comparison is conducted by comparing the evaluation metrics computed by our models, namely MR-DNA-50, -100, -200, and those computed by previous studies, namely iPromoter-5mC and Bert-5mC, on the iPromoter-5mC test dataset.

shows how machine-learning approaches may be used to determine the methylation state of a specific cytosine, and allows the application for methylation site identification on longer DNA sequences.

The evaluation of MR-DNA's performance and the effectiveness of methyl loss was conducted using four different metrics compared with the model using the default loss function. Methyl loss shows superior performance in most comparisons among models trained on different datasets. Using methyl loss, all four evaluation metrics indicate that MR-DNA-50,in particular, shows high reliability. We illustrated how one might combine a pretrained language model and a classic statistical algorithm to improve predictive ability.

Moreover, by modifying the output of MR-DNA models, we were able to estimate how well our models work compared to other studies that focus on predicting the methylation state of a given DNA sequence, where the target cytosine is located in the center. The comparison demonstrated that our models, especially the MR-DNA-100 model have comparable prediction ability with the models that are specifically trained for predicting the global methylation state of DNA sequence. So, while the main aim of MR-DNA is to determine the methylation state of individual nucleotides, it also performs well when used to predict the global methylation state of a DNA sequence.

MR-DNA is the first study to transfer the NER approach from human language to biological sequences and to adapt the recognition target to the DNA methylation scenario, aiming to recognize the entities of "methylation", "non-methylation" and "other nucleotide" in DNA sequences, instead of "persons", "locations", and "organizations", in text.

Additionally, this study offers a custom MR-DNA database generated by obtaining methylation state from eight human cell lines in the ENCODE database, each record consists of 1000 bp in the original MR-DNA database. MR-DNA was trained on the processed dataset by extracting different sequence lengths following self-define standards. The MR-DNA database is provided as a benchmark dataset for researchers interested in optimizing methylation-site detection.

We would like to emphasize that, for a sequence of arbitrary length n, a prediction is made by partitioning the sequence into $n-50+1$ individual 50-mers and then applying MR-DNA to each such sequence.

Methods

Database construction

The MR-DNA database was constructed from DNA methylation-site information of eight human cell lines (K562, HepG2, GM12878, HeLa-S3, A549, SK-N-SH, H1, GM23248) downloaded from the ENCODE portal[43,44], with the following identifiers: ENCSR765JPC, ENCSR786DCL, ENCSR890UQO, ENCSR550RTN, ENCSR481JIW, ENCSR145HNT, ENCSR617FKV, ENCSR625HZA. Each experiment provided the methylation state at CpG, CHH, CHG locations, determined using whole-genome bisulfite sequencing (WGBS), in BED format. We filtered the BED files to keep only methylation sites that are on the same strand as the gene, have sequencing coverage of 10 or more, and for which 100% of the assembled reads are reported as methylated. For each of the human cell lines, the filtered BED files were split in a ratio of 9:1 to generate training and test sets, respectively.

The start and end positions of gene promoter regions were determined from the GRCH38 annotation file downloaded from NCBI as the 1000 bp upstream of a transcription start position, and the corresponding DNA sequences were extracted from

the GRCH38 genome reference. The filtered BED files were used to annotate the methylated sites in the promoter regions. The methylation site statistics on annotated gene promoter regions in terms of each human cell line project are reported in Fig. 4A-B and Table. S2.

We first populated the MR-DNA database with the described annotated promoter sequences of length 1000 bp. Then, for training purposes, we constructed a second database, MR-DNA-50, in which all sequences were cut into short sequences of length 50 bp (50-mers), using a stride of 25. Any such 50-mer that is annotated with at least one methylation site was used to generate the MR-DNA-50 training dataset, and the MR-DNA-50 test dataset, respectively (see Fig. 4C).

Entity category annotation

In preparation of the NER task, we processed and annotated all 50-mers as follows. First, each 50-mer was extended in either direction by a single "N" and each resulting 52-mer was converted into a sentence-like sequence consisting of 50 (=52-3+1) overlapping 3-mers obtained using a sliding window, separated by whitespace. Second, imitating the named-entity naming rule, for each 3-mer, we attached the label "Methyl", 'Non-Methyl", or "O", depending on whether the middle base is a methylated cytosine, a non-methylated cytosine, or neither, respectively. It is worth mentioning that every nucleotide will appear in the center of some 3-mer, so while a methylated cytosine might be regarded as Non-Methyl when encountered off-center, it will be annotated as Methyl when encountered in the center of a 3-mer.

In result, in both the processed MR-DNA-50 training dataset and test dataset, each sample is a sentence consisting of 50 individual 3-mers separated by whitespace, where each element is tagged as "Methyl", "Non-Methyl", or "O", depending on its methylation state reported in the MR-DNA database (see Fig. 1A).

The median value of the number of methylation sites for each sample in the initial MR-DNA-50 training dataset is 1; thus the number of "Methyl" entities is smaller than the number of "Non-Methyl" entities, which in turn is much smaller than the number of "O" entities (see Fig. 4D-E), and so the distribution of labels is skewed.

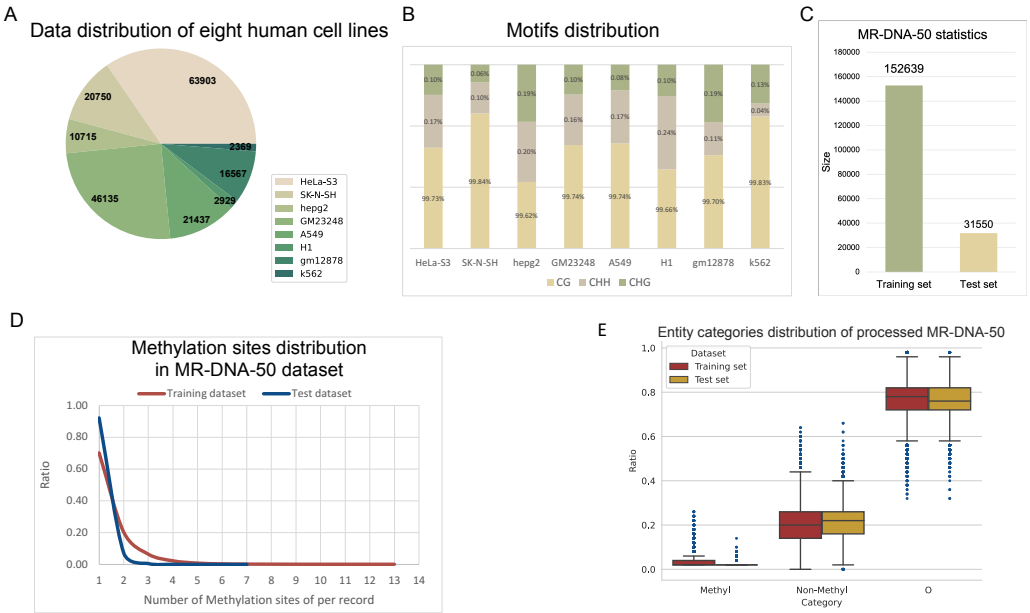


Figure 4. Database statistics. For annotated gene promoter regions, we show (A) a pie chart of the portion of each human cell line project in the annotated gene promoter regions, (B) the distribution of three types of methylation motifs of each human cell line project. For both the training and test dataset of MR-DNA-50, we show (C) a comparison of their sizes, (D) the distribution of ratios of methylation sites per sample, and (E) a box-plot of the entity category distributions.

Pre-trained MuLan-Methyl-DistilBERT

Our proposed framework, MR-DNA, consists of a token-level classifier based on the pre-trained MuLan-Methyl-DistilBERT[41], followed by a Linear-Chain CRF as its decoder layer (see Fig. 1B).

The MuLan-Methyl-DistilBERT classifier is an adaption of DistilBERT[47] to the problem of DNA-methylation identification, obtained by pre-training the masked language modelling (MLM) task on a custom corpus consisting of both DNA methylation and taxonomy data, and then fine-tuning on a DNA methylation-site dataset. The motivation here is that sequence representation

vectors encoded by MuLan-Methyl-DistilBERT that were trained taking domain adaption in to account can better capture the potential semantic relationships and information between tokens.

MuLan-Methyl-DistilBERT uses the same basic neural network architecture as DistilBERT, performing knowledge distillation of Bidirectional Encoder Representations from Transformers (BERT)[19], while reducing the number of transformer layers and adjusting pre-training tasks. The training objective of DistilBERT is a linear combination of distillation loss $\mathscr{L}_{ce}$, language modelling mask loss $\mathscr{L}_{mlm}$, and cosine-distance loss $\mathscr{L}_{cos}$[47] as (3):

$$\mathscr{L} = \alpha_{ce} \times \mathscr{L}_{ce} + \alpha_{mlm} \times \mathscr{L}_{mlm} + \alpha_{cos} \times \mathscr{L}_{cos}. \tag{3}$$

We now describe how to obtain representation vectors. Note that the pre-trained MuLan-Methyl-DistilBERT was trained on a custom corpus and that the vocabulary of the corresponding custom tokenizer contains all combinations of A, T, G, C of multiple lengths as well as words related to taxonomy lineages. However, the DNA sequences in the processed MR-DNA-50 dataset contain an additional character, N. So, the first step is to expand the vocabulary size of the original tokenizer from 25,000 to 25,032, where the additional 32 words are the combinations of A, T, G, C, N of length 3 that occur in the processed MR-DNA-50 training dataset. The embedding dimension of the pre-trained language model was expanded accordingly.

After tokenizing the MR-DNA-50 training dataset using the expanded tokenizer, we employed the updated pre-trained MuLan-Methyl-DistilBERT as an encoder to obtain an embedding representation vector in a three-dimensional array $W \in \mathbb{R}^{1\times t\times 768}$, for each input sequence $X = \{x_1, x_2, \cdots, x_t\}$, where t is the number of tokens.

Linear-chain CRF layer

For each input sequence, the embedding representation vectors W (generated by the encoder) serve as the input for the linear-chain CRF algorithm, for which the posterior probability y of the input sequence is defined as (4):

$$p(y|W) = \frac{1}{Z(W)} \exp\left\{\sum_{n=1}^{N}\sum_{k=1}^{K} \lambda_k f_k(y_n, y_{n-1}, W_n)\right\}, \tag{4}$$

where $\{f_k(y_n, y_{n-1}, W_n)\}_{k=1}^{K}$ is a set of real-valued feature functions, $\{\lambda_k\} \in \mathbb{R}^K$ is a parameter vector, and $Z(W)$ is a normalization factor of the form

$$Z(W) = \sum_{y} \exp\left\{\sum_{k=1}^{K} \lambda_k f_k(y_n, y_{n-1}, W_n)\right\}. \tag{5}$$

The objective for parameter learning is to maximize the conditional likelihood of the training data given by

$$l(\theta) = \log p(y, W) = \sum_{n=1}^{N}\sum_{k=1}^{K} \lambda_k f_k(y_n, y_{n-1}, W_n) - \log Z(W), \tag{6}$$

and the most probable assignment for each element in the input sequence

$$y^* = \arg\max_{y} p(y|W) \tag{7}$$

is further decoded by the Viterbi algorithm.

Methyl loss function

Because of the skewed distribution of entity categories, the model's ability to predict minority entities is comparatively weak. To mitigate this, we propose to use "methyl loss", a loss function that allows the model to focus more on challenging samples by embedding the biological rule of DNA methylation.

Methyl loss is based on the categorical cross-entropy loss function,

$$\mathscr{L}_{CCE} = -\sum_{i=1}^{N} y_i \log(p_i) \tag{8}$$

where y_i is the truth label, p_i is the predicted probability of the i_{th} class. This is inspired by focal loss, which is a variant of cross-entropy loss that has been widely used in the context of imbalanced data distribution:

$$\mathscr{L}_{Focal} = -\sum_{i=1}^{N} \alpha_i (1-p_i)^{\gamma} \log(p_i), \tag{9}$$

where $\gamma > 0$ is used to reduce the relative loss for entities well-classified during model training and to concentrate on challenging entity categories. Here, α_t is a hard hyper-parameter containing a list of weights corresponding to each category. The assignment of α and γ only depends on the ratio of classes, regardless of the scenario.

Methyl loss closes this gap by utilizing the generated annotation vector, for each sample $X = \{x_t\}_{t \in T}$, where T is the number of tokens, and x_t is a token consisting of three elements. We define the annotation vector $\mathscr{A} = \{a_t\}_{t \in T}$ as

$$a_t = \begin{cases} 0, & \text{if } x_{t1} \neq C \\ 1, & \text{if } x_{t1} = C, \text{ and } x_{t2} = G \\ 2, & \text{if } x_{t1} = C, \text{ and } x_{t2} \neq G \end{cases}$$

We assume a token is difficult to classify if its nucleotides belong to the minority situation of its category. For instance, each token contains three nucleotides, whereas most tokens with the “Methyl" label contain CpG, where C is in the center. The proposed methyl loss aims to pay more attention to the token which belongs to the minority of its category during model training. In consequence, the methyl loss function is given by

$$\mathscr{L}_{Methl} = -\sum_{i=1}^{N} \log(p_i) \lambda^{(1-p_i)\beta_i} \tag{10}$$

where

$$\beta_i = \begin{cases} 0, & \text{if } y_i = a_i \\ 1, & \text{else} \end{cases}$$

and $\lambda > 0$ is a variable for controlling the attention level on challenging objects. For choosing the optimal λ, we evaluate the impact of λ on model performance by setting λ from 2 to 5, where λ equals 2, resulting in the best accuracy and F1-score in our scenario.

Data availability

The data used in this study were downloaded from the ENCODE database at https://www.encodeproject.org/. The derived MR-DNA datasets and our code are available at https://github.com/husonlab/MR-DNA.

References

1. Luo, C., Hajkova, P. & Ecker, J. R. Dynamic DNA methylation: In the right place at the right time. *Science* **361**, 1336–1340 (2018).
2. Li, E. & Zhang, Y. DNA methylation in mammals. *Cold Spring Harb. perspectives biology* **6**, a019133 (2014).
3. Turpin, M. & Salbert, G. 5-methylcytosine turnover: Mechanisms and therapeutic implications in cancer. *Front. Mol. Biosci.* **9** (2022).
4. Schmutte, C. & Jones, P. A. Involvement of DNA methylation in human carcinogenesis. *Biol. Chem.* **379**, 377–388 (1998).
5. Li, Y. & Tollefsbol, T. O. DNA methylation detection: bisulfite genomic sequencing analysis. *Epigenetics Protoc.* 11–21 (2011).
6. Plongthongkum, N., Diep, D. H. & Zhang, K. Advances in the profiling of DNA modifications: cytosine methylation and beyond. *Nat. Rev. Genet.* **15**, 647–661 (2014).
7. Tse, O. O. *et al.* Genome-wide detection of cytosine methylation by single molecule real-time sequencing. *Proc. Natl. Acad. Sci.* **118**, e2019768118 (2021).
8. Simpson, J. T. *et al.* Detecting DNA cytosine methylation using nanopore sequencing. *Nat. methods* **14**, 407–410 (2017).
9. Beaulaurier, J., Schadt, E. E. & Fang, G. Deciphering bacterial epigenomes using modern sequencing technologies. *Nat. Rev. Genet.* **20**, 157–172 (2019).

10. Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. biotechnology* **37**, 1155–1162 (2019).

11. Tran, T.-A., Pham, D.-M., Ou, Y.-Y. *et al.* An extensive examination of discovering 5-Methylcytosine Sites in Genome-Wide DNA Promoters using machine learning based approaches. *IEEE/ACM Transactions on Comput. Biol. Bioinforma.* **19**, 87–94 (2021).

12. Cheng, X., Wang, J., Li, Q. & Liu, T. BiLSTM-5mC: a bidirectional long short-term memory-based approach for predicting 5-methylcytosine sites in genome-wide DNA promoters. *Molecules* **26**, 7414 (2021).

13. Zhang, L., Xiao, X. & Xu, Z.-C. iPromoter-5mC: a novel fusion decision predictor for the identification of 5-methylcytosine sites in genome-wide DNA promoters. *Front. Cell Dev. Biol.* **8**, 614 (2020).

14. Xiao, X., Shao, Y.-T., Luo, Z.-T. & Qiu, W.-R. m5C-HPromoter: An Ensemble Deep Learning Predictor for Identifying 5-methylcytosine Sites in Human Promoters. *Curr. Bioinforma.* **17**, 452–461 (2022).

15. Ji, Y., Zhou, Z., Liu, H. & Davuluri, R. V. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics* **37**, 2112–2120 (2021).

16. Jin, J. *et al.* idna-abf: multi-scale deep biological language learning model for the interpretable prediction of dna methylations. *Genome biology* **23**, 219 (2022).

17. Wang, S., Liu, Y., Liu, Y., Zhang, Y. & Zhu, X. Bert-5mc: an interpretable model for predicting 5-methylcytosine sites of dna based on bert. *PeerJ* **11**, e16600 (2023).

18. Kent, W. J. *et al.* The human genome browser at ucsc. *Genome research* **12**, 996–1006 (2002).

19. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

20. Ni, P. *et al.* DNA 5-methylcytosine detection and methylation phasing using PacBio circular consensus sequencing. *Nat. Commun.* **14**, 4054 (2023).

21. Wang, Z., Xiang, S., Zhou, C. & Xu, Q. DeepMethylation: a deep learning based framework with GloVe and Transformer encoder for DNA methylation prediction. *PeerJ* **11**, e16125 (2023).

22. Jia, J., Qin, L. & Lei, R. DGA-5mC: A 5-methylcytosine site prediction model based on an improved DenseNet and bidirectional GRU method. *Math. Biosci. Eng. MBE* **20**, 9759–9780 (2023).

23. Bonet, J. *et al.* DeepMP: a deep learning tool to detect DNA base modifications on Nanopore sequencing data. *Bioinformatics* **38**, 1235–1243 (2022).

24. Tian, Q. *et al.* MRCNN: a deep learning model for regression of genome-wide DNA methylation. *BMC genomics* **20**, 1–10 (2019).

25. Nadeau, D. & Sekine, S. A survey of named entity recognition and classification. *Lingvisticae Investig.* **30**, 3–26 (2007).

26. Li, J., Sun, A., Han, J. & Li, C. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowl. Data Eng.* **34**, 50–70 (2020).

27. Tabassum, J., Maddela, M., Xu, W. & Ritter, A. Code and named entity recognition in stackoverflow. *arXiv preprint arXiv:2005.01634* (2020).

28. Chang, Y., Kong, L., Jia, K. & Meng, Q. Chinese named entity recognition method based on BERT. In *2021 IEEE International Conference on Data Science and Computer Application (ICDSCA)*, 294–299 (IEEE, 2021).

29. Weber, L. *et al.* HunFlair: an easy-to-use tool for state-of-the-art biomedical named entity recognition. *Bioinformatics* **37**, 2792–2794 (2021).

30. Song, B., Li, F., Liu, Y. & Zeng, X. Deep learning methods for biomedical named entity recognition: a survey and qualitative comparison. *Briefings Bioinforma.* **22**, bbab282 (2021).

31. Naseem, U. *et al.* Bioalbert: A simple and effective pre-trained language model for biomedical named entity recognition. In *2021 International Joint Conference on Neural Networks (IJCNN)*, 1–7 (IEEE, 2021).

32. Morwal, S., Jahan, N. & Chopra, D. Named entity recognition using hidden Markov model (HMM). *Int. J. on Nat. Lang. Comput. (IJNLC) Vol* **1** (2012).

33. Zhang, L., Pan, Y. & Zhang, T. Focused named entity recognition using machine learning. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, 281–288 (2004).

34. Bender, O., Och, F. J. & Ney, H. Maximum entropy models for named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, 148–151 (2003).

35. Ju, Z., Wang, J. & Zhu, F. Named entity recognition from biomedical text using SVM. In *2011 5th international conference on bioinformatics and biomedical engineering*, 1–4 (IEEE, 2011).

36. Chiu, J. P. & Nichols, E. Named entity recognition with bidirectional LSTM-CNNs. *Transactions association for computational linguistics* **4**, 357–370 (2016).

37. Hakala, K. & Pyysalo, S. Biomedical named entity recognition with multilingual BERT. In *Proceedings of the 5th workshop on BioNLP open shared tasks*, 56–61 (2019).

38. Souza, F., Nogueira, R. & Lotufo, R. Portuguese named entity recognition using BERT-CRF. *arXiv preprint arXiv:1909.10649* (2019).

39. Sun, J., Liu, Y., Cui, J. & He, H. Deep learning-based methods for natural hazard named entity recognition. *Sci. reports* **12**, 4598 (2022).

40. Wu, X., Zhang, T., Yuan, S. & Yan, Y. One improved model of named entity recognition by combining BERT and BiLSTM-CNN for domain of Chinese railway construction. In *2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP)*, 728–732 (IEEE, 2022).

41. Zeng, W., Gautam, A. & Huson, D. H. MuLan-Methyl-Multiple Transformer-based Language Models for Accurate DNA Methylation Prediction. *bioRxiv* 2023–01 (2023).

42. Lv, H. *et al.* iDNA-MS: an integrated computational tool for detecting DNA modification sites in multiple genomes. *Iscience* **23**, 100991 (2020).

43. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

44. Luo, Y. *et al.* New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic acids research* **48**, D882–D889 (2020).

45. Wolf, T. *et al.* Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45 (Association for Computational Linguistics, Online, 2020).

46. Nakayama, H. seqeval: A python framework for sequence labeling evaluation (2018). Software available from https://github.com/chakki-works/seqeval, last accessedL: 16-Oct-2023.

47. Sanh, V., Debut, L., Chaumond, J. & Wolf, T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).

Acknowledgements

We acknowledge the support of the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A532B, 031A533A, 031A533B, 031A534A, 031A535A, 031A537A, 031A537B, 031A537C, 031A537D, 031A538A).

Author contributions statement

W.Z. conceived and conducted the experiments, W.Z analysed the results. W.Z. and D.H.H. wrote and reviewed the manuscript.

Funding

We acknowledge support by the Open Access Publishing Fund of the University of Tübingen.

Competing interests

No competing interests declared.

Additional information

The corresponding author is responsible for submitting a competing interests statement on behalf of all authors of the paper.

C Manuscript 3

Title: DeepToA: An Ensemble Deep-Learning Approach to Predicting the Theater of Activity of a Microbiome

Bioinformatics, 38(20), 2022, 4670–4676
https://doi.org/10.1093/bioinformatics/btac584
Advance Access Publication Date: 27 August 2022
Original Paper

OXFORD

Genome analysis

DeepToA: an ensemble deep-learning approach to predicting the theater of activity of a microbiome

Wenhuan Zeng[1], Anupam Gautam[1,2] and Daniel H. Huson [1,2,3,*]

[1]Department of Algorithms in Bioinformatics, Institute for Bioinformatics and Medical Informatics, University of Tübingen, Tübingen 72076, Germany, [2]International Max Planck Research School "From Molecules to Organisms", Max Planck Institute for Biology Tübingen, Tübingen 72076, Germany and [3]Cluster of Excellence: Controlling Microbes to Fight Infection, Tübingen, Germany

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on April 5, 2022; revised on July 19, 2022; editorial decision on August 19, 2022; accepted on August 26, 2022

Abstract

Motivation: Metagenomics is the study of microbiomes using DNA sequencing. A microbiome consists of an assemblage of microbes that is associated with a 'theater of activity' (ToA). An important question is, to what degree does the taxonomic and functional content of the former depend on the (details of the) latter? Here, we investigate a related technical question: Given a taxonomic and/or functional profile estimated from metagenomic sequencing data, how to predict the associated ToA? We present a deep-learning approach to this question. We use both taxonomic and functional profiles as input. We apply node2vec to embed hierarchical taxonomic profiles into numerical vectors. We then perform dimension reduction using clustering, to address the sparseness of the taxonomic data and thus make the problem more amenable to deep-learning algorithms. Functional features are combined with textual descriptions of protein families or domains. We present an ensemble deep-learning framework DeepToA for predicting the ToA of amicrobial community, based on taxonomic and functional profiles. We use SHAP (SHapley Additive exPlanations) values to determine which taxonomic and functional features are important for the prediction.

Results: Based on 7560 metagenomic profiles downloaded from MGnify, classified into 10 different theaters of activity, we demonstrate that DeepToA has an accuracy of 98.30%. We show that adding textual information to functional features increases the accuracy.

Availability and implementation: Our approach is available at http://ab.inf.uni-tuebingen.de/software/deeptoa.

Contact: daniel.huson@uni-tuebingen.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Deep-learning algorithms are known to perform well on a wide range of problems coming from different disciplines of science (Ardila *et al.*, 2019; Bukhari *et al.*, 2020; Jumper *et al.*, 2021; Reichstein *et al.*, 2019). To achieve good performance on new problems, the input data must satisfy certain requirements (Najafabadi *et al.*, 2015), and some engineering of the approach is required.

Several deep-learning approaches have been developed to address microbiome-related questions, such as disease prediction (Oh and Zhang, 2020; Sharma and Xu, 2021), the annotation of antibiotic resistance genes (ARGs) (Li *et al.*, 2021), microbial source tracking (Shenhav *et al.*, 2019; Wu *et al.*, 2021) and microbial community prediction (Thompson *et al.*, 2019; Zha *et al.*, 2022).

A microbiome consists of a collection of microbes that live in a specific 'theater of activity' (ToA), 'habitat' or 'ecological niche', and the total genomic information of the microbes is known as the metagenome of the microbiome (Handelsman, 2004). The term 'ToA' was introduced in Whipps and Lewis (1988), see also Berg *et al.* (2020); we prefer it because it avoids connotations that terms like *ecological niche* might have.

Computational analysis usually involves estimating the taxonomic and functional content of microbiomes, either from amplicon sequencing data (e.g. targeting 16S rRNA sequences, Caporaso *et al.*, 2010) or from metagenomic sequencing data (e.g. Huson *et al.*, 2016; Mitchell *et al.*, 2020).

The taxonomic and functional content of a microbiome is shaped, to a degree, by its ToA (Berg *et al.*, 2020; The Human Microbiome Project Consortium, 2012). An important question is, how strong is the influence?

Here, we address a related technical question: Can one accurately predict the ToA from the taxonomic and functional profile of a sample? We address this question using a deep-learning approach. We trained and tested our approach using 7560 metagenomic

datasets downloaded from MGnify (Mitchell *et al.*, 2020), classified into 10 different theaters of activity, namely *Animal Digestive System, Food Production, Freshwater, Human Respiratory System, Human Skin, Mammal Gastrointestinal Tract, Marine, Plants, Soil* and *Wastewater*. For a given query sample, represented by a taxonomic and/or functional profile, our classifier returns a probability of membership for each of the 10 classes.

In related work, the Earth Microbiome Project uses random forests to determine environmental factors (Smith *et al.*, 2010). SourcePredict (Borry, 2019) uses dimension reduction followed by a KNN algorithm to classify and predict the origin of metagenomics samples.

The MetaSUB consortium provides a 'metagenomic atlas important for understanding the ecology, virulence and antibiotic resistance of city-specific microbial communities' (Danko *et al.*, 2021). They describe the MetaGraph tool, which provides a *k*-mer graph that indexes all sequences found in different environments. This allows one to determine whether a given metagenomic sequence has been previously reported in some study. The article also explores the use of a Random Forest classifier to predict the city of origin from a *k*-mer-based taxonomic profile of a sample.

In a recent study (Wu *et al.*, 2021), several machine-learning algorithms are used to predict the dominant source of microbial contamination by using environmental and geographical data. The ONN4MST method uses an ontology-aware neural network (ONN) to embed biome ontology information into a hierarchical structure so as to improve the performance of community-based microbial source tracking (Zha *et al.*, 2022).

Some machine-learning approaches for predicting the ToA are based on taxonomic profiles of 16S rRNA sequencing data (Knights *et al.*, 2011), which have limited taxonomic resolution and cannot assess functional content. Here, we present an ensemble deep-learning framework DeepToA that is specifically designed for the analysis of taxonomic and functional profiles obtained from metagenomic data (see Fig. 1).

We provide an implementation of our ToA prediction for metagenomics samples on a server.

For each of the 7560 metagenomic samples, MGnify (Mitchell *et al.*, 2020) provides a taxonomic profile based on the NCBI taxonomy (Schoch *et al.*, 2020), a functional profile based on InterPro (Blum *et al.*, 2021), and a specific ToA. As described further below, we enhanced the samples by considering additional textual descriptions of the samples.

Taxonomic and functional profiles are usually represented by count tables, and for any given sample, the vast majority of entries will be zero. To address this, a dimensionality reduction is required (Oudah and Henschel, 2018; Sharma *et al.*, 2020; Zhou *et al.*, 2021).

The first step in our approach is to convert taxonomic lineages into numerical vectors. Popular pre-trained language models (Devlin *et al.*, 2018; Peters *et al.*, 2018) are not applicable here. Instead, we used the GTDB taxonomy (Parks *et al.*, 2021), supplemented by branches from the NCBI taxonomy (Schoch *et al.*, 2020), to obtain a single taxonomy for all archaea, bacteria, eukaryota and viruses. Additional taxa encountered in taxonomic profiles during training were also incorporated into the taxonomy.

We applied node2vec (Grover and Leskovec, 2016) to this reference tree to obtain an embedding vector for each taxon.

We then calculated Euclidean distances between the embedding vectors and used the AGNES algorithm to cluster them (which showed the best performance, as described below). The resulting clusters, which we will refer to as *processed taxonomic profiles*, were then used as input to a 'taxonomy-based' deep-learning model for ToA prediction. Below, we show that the clusters reflect taxonomic relationships.

The second step in our approach is to process functional profiles. These are initially given as InterPro count tables, which we expand into 3D tables by considering each features' textual description. The resulting *processed tables* are used as input to a 'function-based' deep-learning model for ToA prediction.

The taxonomy-based and function-based models are then combined into an ensemble deep-learning framework DeepToA, which achieves an accuracy of 98.30%, as discussed below.

To address explainable machine learning, we use a Bi-LSTM (bidirectional long short-term memory) model (Hochreiter and Schmidhuber, 1997) on the input taxonomic and functional profiles, respectively, and then compute SHAP (SHapley Additive exPlanations) values (Lundberg and Lee, 2017) to estimate feature importance, a widely used approach for 'explaining' deep-learning models (Arcadu *et al.*, 2019; Rajpurkar *et al.*, 2021; Yap *et al.*, 2021). This will help to identify key taxa and functions associated with specific environments.

2 Materials and methods

2.1 Dataset download and preparation

We downloaded 7560 metagenomic samples from MGnify (Mitchell *et al.*, 2020). We held back 20% of samples as a testing set for evaluating the models' performance. The remaining samples were split 9 : 1 into a training set and a validation set for model training.

Then, in the first step, for each sample selected for model training, we downloaded a taxonomic count table using the MGnify API. We used the QIIME script `merge_otu_tables.py` (Caporaso *et al.*, 2010) to merge all tables into a single, initial table, containing 6048 rows (samples) and 117 727 columns (taxa). Using the dimensionality reduction procedure described below, we clustered all taxa into 10 000 classes and this reduced the taxonomy table to a 'processed' table with 6048 rows (samples) and 10 000 columns (taxonomic clusters).

In a second step, for each sample selected for model training, we downloaded a functional count table using the MGnify API. These data were combined into a single initial table, containing 6048 rows (samples) and 13 041 columns (InterPro IDs). We also downloaded the textual descriptions associated with the InterPro IDs from MGnify and embedded each description into a numerical vector of length 10 using doc2vec, as described below. This gave rise to a 3D functional 'processed' table with 6048 rows (samples), 13 041 columns (InterPro IDs) and 10 additional values (the InterPro embedding vectors).

In this study, we used both the initial and processed datasets.

2.2 Computation of the taxonomy embedding matrix

2.2.1 Tree structure

We downloaded taxonomic details on 4316 archaea and 254 090 bacteria from the GTDB database release 202 (Parks *et al.*, 2021), which is based on 258 406 genomes organized into 47 894 species groups. We also downloaded taxonomy details from the NCBI taxonomy database (Schoch *et al.*, 2020), which cover 514 archaea, 5294 bacteria, 64 462 eukaryota and 1838 viruses.

Any taxon mentioned in a downloaded taxonomic profile that was not already included in the reference tree was incorporated into the tree, together with any corresponding higher-order taxa. In a result, we obtained an extended taxonomy that is based on both the GTDB taxonomy and the NCBI taxonomy and is organized in the eight usual taxonomic ranks, from 'domain' to 'species'. We implemented the taxonomy as a rooted, directed tree, using the Networkx Python package (see https://networkx.org).

2.2.2 Graph embedding

We ran the node2vec algorithm (Grover and Leskovec, 2016) on the taxonomic tree. In more detail, using 300 random walks per node and 20 nodes in each walk, we mapped each taxon t onto a 10D embedding vector $v(t)$. The embedding vectors for all species are shown in a t-SNE plot (van der Maaten and Hinton, 2008) in Figure 1b, indicating some clustering by domain.

We reshaped the data by assigning to each taxon t a lineage-based vector of length 80 that is obtained as the concatenation $V(t) = v(t_1) \oplus \ldots \oplus v(t_8)$ of all embedding vectors $v(t_1), \ldots (t_8)$ of

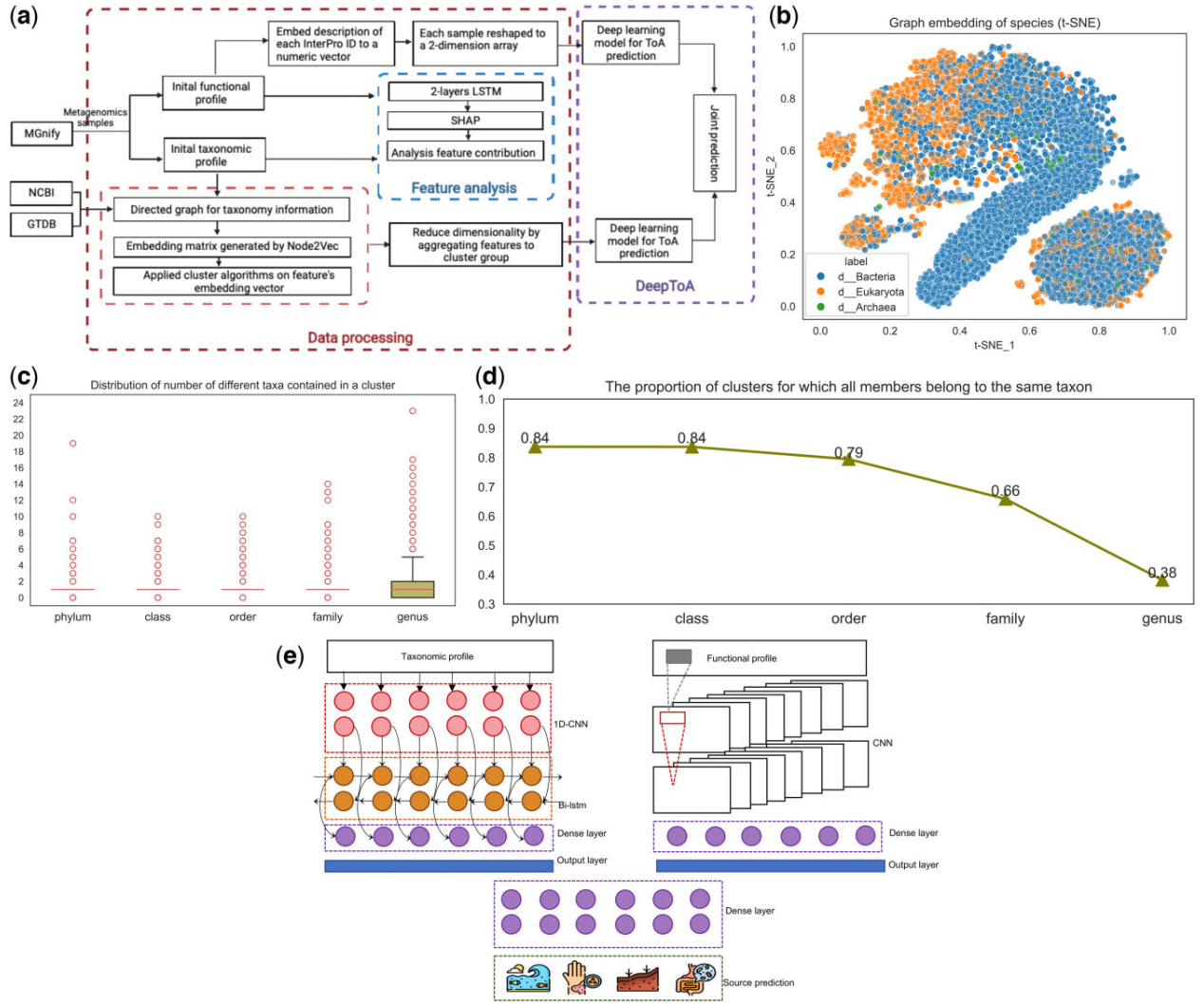


Fig. 1. (**a**) Workflow for data collection, feature engineering and model building. Metagenomic samples are downloaded from MGnify and their taxonomic and functional profiles are extracted. Taxonomic and functional profiles are processed in the lower and upper part of the workflow, respectively. In the middle of the workflow, feature importance is assessed using SHAP values. Joint prediction using DeepToA is performed on the right-hand side. (**b**) A t-SNE plot of species' embedding vectors, colored by taxonomic domain. (**c**) The distribution of the number of different taxa that are contained in a cluster, for each taxonomic rank. The mean value is 1 for all ranks. (**d**) The proportion of clusters that are pure for a given taxonomic rank, that is, for which all members of the cluster belong to the same taxon of that rank. (**e**) Structure of the DeepToA model, with the taxonomic model left, the functional model right, and the combining and prediction layers at the bottom

the taxa that lie on the path $t_1, t_2, \ldots t_8$ from the root of the taxonomy to the taxon t, one for each taxonomic rank, filling in missing data with zeros.

In total, we obtained 117 727 different 80D vectors representing taxa.

2.3 Dimension reduction

To reduce the number of different taxonomic features to a more manageable number that is similar to the number of available samples, we decided to cluster the taxonomic features into 10 000 clusters. To achieve this, we computed Euclidean distances between all vectors and then evaluated the performance of different distance-based clustering algorithms, see Table 1.

We used three metrics to evaluate the performance of the clustering techniques, namely the Calinski–Harabasz score (Calinski and Harabasz, 1974), the Silhouette score (Rousseeuw, 1987) and the Davies–Bouldin index (Davies and Bouldin, 1979). In addition, we also compared running time.

Based on the results reported in Table 1, we selected the AGNES clustering algorithm for use in this study, as it has a high Calinski–Harabasz score, the highest Silhouette score and low Davies–Bouldin score, suggesting more dense and well-separated clusters. The clustering gives rise to $r = 10\,000$ groups of taxa, which we denote by $C_1, \ldots, C_{10,000}$.

Table 1 For eight different methods considered for clustering taxonomic vectors, we report the Calinski–Harabasz score, Silhouette score, Davies–Bouldin index and total wall-clock time in minutes

	Calinski–Harabasz	Silhouette	Davies–Bouldin	Time (min)
K-means	4, 567.42	0.83	0.47	93.6
DBSCAN	61.16	0.75	1.01	0.83
GMM	4,548.56	0.83	0.46	464.62
Spectral	5.89	−0.43	0.78	1, 690.18
Birch	1, 878.28	0.77	0.41	0.88
AGNES	4, 978.53	0.85	0.48	26.39
Mini-batch K-means	1, 005.92	0.79	1.02	8.48
OPTICS	114.08	0.84	1.05	50.62

After this preprocessing, any given taxonomic profile t associated with a metagenomic dataset can be represented as a vector $D = (d_1, \ldots, d_r)$ of length $r = 10\ 000$, where $d_j = \sum_{k \in C_j} t(k)$ is the sum of counts over all taxa k that lie in cluster C_j.

Let $T_{m,n} = \{t_{ij}\}$ denote the matrix of all original input taxonomic profiles, with m the number of samples and n the number of taxa, in the dataset for model training, $m = 6048$ and $n = 117\ 727$. We will use $D_{m,r} = \{d_{ij}\}$ to denote the matrix of all 'processed' taxonomic profiles, where r is the number of clusters, and, for every sample i and cluster C_j, the entry $d_{ij} = \sum_{k \in C_j} t_{ik}$ is the sum of counts over all taxa k of sample i that lie in cluster C_j.

The GTDB and NCBI taxonomies are based on evolutionary relationships, and it is important that the clustering of taxa described here reflects these relationships. This appears to be the case. In Figure 1d, for each higher taxonomic rank, we report the proportion of clusters that are pure in the sense that all members of the cluster belong to the same taxon of the given rank. We see that a large proportion of clusters are pure at the rank of Phylum (84%) and this drops to 38% at the rank of Genus, as is to be expected. In Figure 1c, we show the distribution of number of different taxa that are contained in a cluster, for all higher taxonomic ranks. The mean count is 1 for all ranks.

2.4 Machine learning

We now describe the architecture of the neural network used for 'ToA' prediction. We then discuss how to determine feature importance.

2.4.1 Main neural network architecture in classification task

Taxonomic model. The processed taxonomic input data is represented as a table of counts, with rows representing samples and columns representing clusters of taxa. This is provided as input to a stacked combination of a 1D-CNN model (one-dimensional convolutional neural network) and an LSTM (long short-term memory) model. While a 1D-CNN architecture is usually used for text data and 1D signal data, an LSTM is specifically designed for processing long textual data. The combination of these two structures performs better than either model separately. We executed all training using the open-source framework Tensorflow-GPU Keras 2.6.0 (see https://www.tensorflow.org/); more implementation details are provided further below.

Functional model. The original functional input data are provided as a table of counts, with rows representing samples and columns representing (13 041) InterPro families. For each such family, we computed a 10D embedding vector of the textual description of the family, using doc2vec (Le and Mikolov, 2014), and thus obtained a 2D $13,041 \times 10$ matrix, which can be interpreted as a gray-scale image. This gives rise to the 'processed' functional data, to we which we apply a two layer CNN (to capture hidden rules), multiple max-pooling and dense layers, and an activation function (in the usual way).

Ensemble deep learning. In DeepToA, the taxonomic and functional deep-learning models are combined into an ensemble model to perform 'ToA' prediction together, as shown in Figure 1c. Each label in the model is assigned a weight, which is based on the sample-size distribution, to address the problem of data imbalance (Supplementary Fig. S1).

2.4.2 Explainable deep-learning prediction

For a given prediction of 'ToA', we would like to know which features play a role in the prediction. To address this, we designed a multi-categorical classification model that operates directly on the initial taxonomic and functional profiles.

For both types of profiles, taxonomic and functional, we use a two-layer Bi-LSTM network, built with Tensorflow-GPU Keras 2.6.0 (see https://www.tensorflow.org/). Output is the prediction of the 'ToA'.

The first Bi-LSTM layer, with 128 units, is fed by a tensor of shape (117 727, 1), in the case of taxonomy, or of shape (13 041, 1), in the case of function, respectively. Weights are initialized by setting `kernel_initializer` to `glorot_uniform`. The second Bi-LSTM has 64 units. Both layers employ L2 regularization to avoid overfitting.

Both of the two layers Bi-LSTM networks are each combined using a fully connected layer, in both cases using a softmax function to perform multi-category classification. After training the model on the initial taxonomic profile, or on the initial functional profile, respectively, in either case, the accuracy was 95.24%. Although lower than the accuracy achieved using our other models, this level of accuracy suffices for the purpose of determining feature importance.

We used SHAP values to determine which taxonomic and functional features are important for prediction (Lundberg and Lee, 2017). In more detail, we used the SHAP deep explainer module (see https://github.com/slundberg/shap) to analyze our model. We randomly sampled half of the input samples and used these for training (due to computational constraints). Then SHAP values were computed for all features using the full test set. The results are summarized in Figure 2 and Supplementary Figure S2.

3 Results

3.1 Performance of DeepToA

To develop an accurate classification model for determining the 'ToA' for a metagenomic dataset, we explored several ways of combining taxonomic and functional data with different neural network techniques. First, we considered using either the initial taxonomic profile, or the initial functional profile, separately, as input for neural network model. Second, we considered using either of the processed profiles as input. Third, we investigated two different approaches to combining both taxonomic and functional data. For each combination, we report Precision, Recall, F1-score, AUC (Area under the curve), MCC (Matthews correlation coefficient) and Acc (Accuracy), in Table 2. (We report on hyper-parameter optimization for the main models in Supplementary Table S1.)

Higher accuracy is achieved when building a model on the processed taxonomic profiles (96.76%) than when using a straightforward model built on the initial taxonomic profiles (95.24%). Similarly, higher accuracy is achieved when employing processed functional profiles (96.63%) than when using the initial functional profiles (95.24%). In comparison to these baseline values, the DeepToA model achieves an accuracy of 98.30%, with an AUC of 99.27%.

To evaluate the predictive performance of DeepToA for each class, we applied the DeLong test (DeLong *et al.*, 1988). (This is a non-parametric test designed for the evaluation of statistical differences of AUC between binary classification models, which we here adapted to the multi-label setting.) Comparing DeepToA to each of the other models, we obtained one P-value for each class-based comparison of AUCs. We observed that for each pair of models, there is always at least one class for which the difference is statistically significant ($P < 0.05$).

In Supplementary Figure S3, we report confusion matrices for each model. These suggest that DeepToA outperforms the other models, particularly on classes with small data volumes, and performs more consistently across all classes.

3.2 Additional information increases model performance

In the third row of Table 2, we report on the performance on processed taxonomic profiles. Here, we fed the input data into three 1D CNN layers, each equipped with a ReLU activation function and max-pooling layer. This is followed by two layers of LSTM, each with L2 regularizers with 0.001 and dropout rate 0.1. We use a dense layer with 10 units and softmax activation function as the first model's output layer.

This was then trained using the Adam optimizer with initial learning rate 0.003, decreasing conditionally on the accuracy on the validation set. An accuracy of 96.76% was achieved on the test

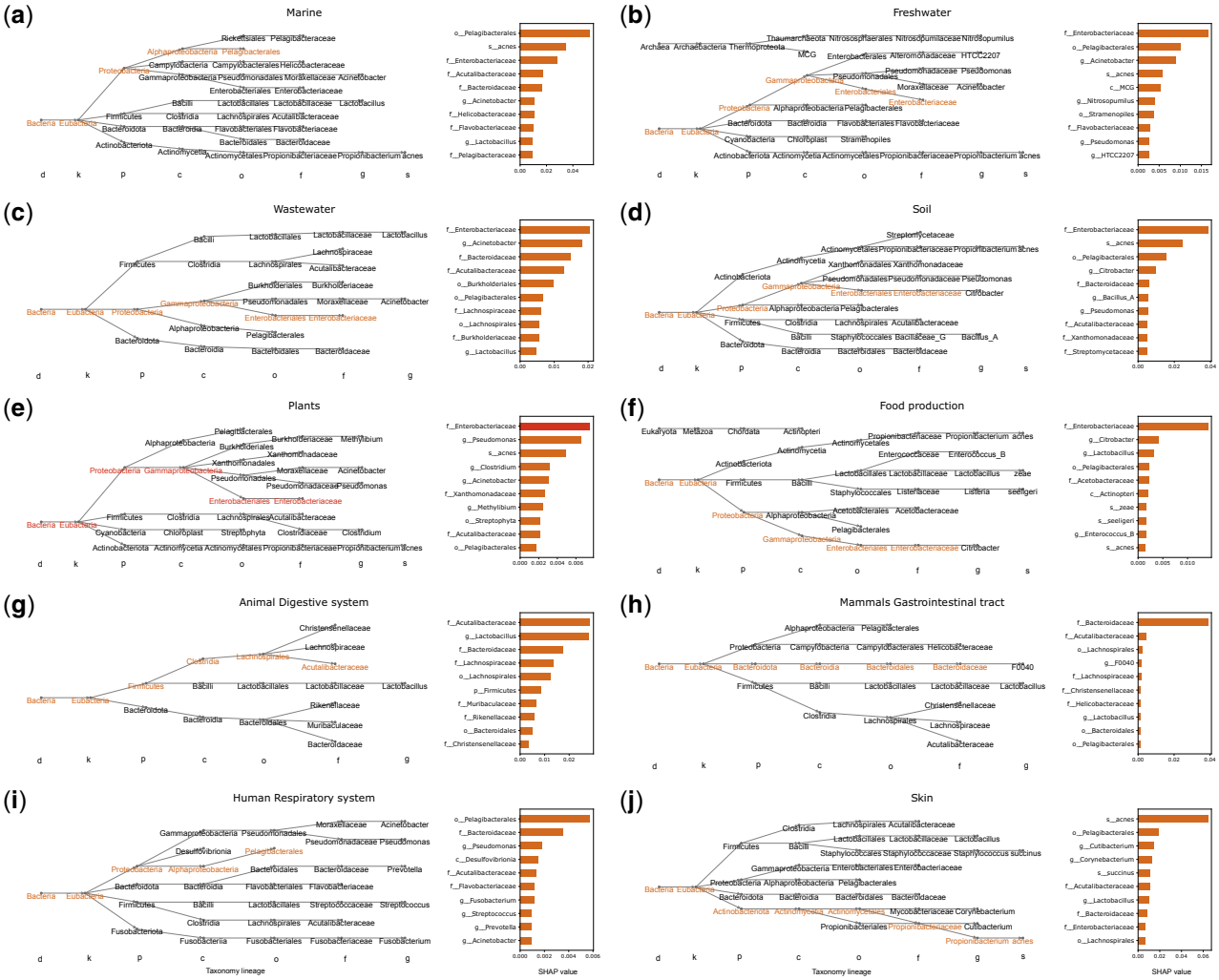


Fig. 2. For each of the named 10 'theaters of activity' under consideration, on the right, we use a bar chart to report the 10 largest SHAP importance values for taxonomic features, and on the left, we show the corresponding taxonomic lineages, indicating taxon ranks using letters, *d* for domain, *k* for kingdom, etc. The path to the taxon with highest SHAP value is highlighted

Table 2 Model evaluation.

Model	Dataset	Precision	Recall	F1-score	AUC	MCC	Acc
Bi-LSTM	Initial taxonomic profile	0.9161	0.9119	0.9131	0.9732	0.9368	0.9524
Bi-LSTM	Initial functional profile	0.9287	0.9305	0.9288	0.9735	0.9378	0.9524
Conv1D+LSTM	Processed taxonomic profile	0.9612	0.9345	0.9466	0.9798	0.9568	0.9676
Conv2D	Processed functional profile	0.9348	0.9368	0.9337	0.9813	0.9559	0.9663
Ensemble model	Processed taxonomic profile and initial functional profile	0.9622	0.9331	0.9464	0.9842	0.9628	0.9716
DeepToA	Processed taxonomic profile and processed functional profile	0.9671	0.9638	0.9622	0.9927	0.9742	0.9830

Note: For different deep-learning models and for different choices of dataset, we report the Precision, Recall, F1-score, AUC, MCC, Acc of ToA prediction. The best values are shown in bold.

dataset. This suggests that the processing of taxonomic features captures evolutionary information and this improves prediction ability.

In the fourth row of Table 2, we report the achieved performance on processed functional profiles, using a two layers CNN, as described above. Because the sample shape is rectangle, the filter size, strides and max-pooling size are set to a rectangle shape. An accuracy of 96.63% was achieved on the test dataset. This indicates that adding external textual information has a positive effect on prediction.

3.3 Model result interpretation

To determine which taxonomic and functional features play a major role when classifying the ToA of a sample, we built a two-layer Bi-

LSTM model on both the taxonomic and functional data, and then applied the Deep SHAP method to obtain SHAP feature importance values.

3.3.1 Feature importance for taxa

In Figure 2, for each of the 10 'theaters of activity' (ToA) under investigation, we list the 10 taxa that have the highest SHAP values for that particular ToA. In addition, we display the lineage of each such taxa using a part of the taxonomy.

The taxa listed for a particular ToA are often taxa that are known to be associated with the ToA. For example, the Enterobacteriaceae family shows high importance in freshwater, wastewater, soil, plants and food production. Likewise, the Pelagibacterales order shows high importance for the marine environment and is an order composed of free-living marine bacteria that make up roughly one in three cells at the ocean's surface (Wikipedia, https://en.wikipedia.org/w/index.php?title=Pelagibacterales, accessed 11 March 2022). However, it is unclear why it should appear as the most important taxonomic feature for human respiratory system, and also shows up in human skin and food production. Similarly, while *Propionibacterium acnes* has high importance for human skin, it is also listed for marine, freshwater, plants and soil.

3.3.2 Feature importance for function

Functional profiles considered here are based on InterPro families and domains, which are identified by IPR accession numbers. As shown in Supplementary Figure S2, IPR003514 (Microviridae F protein family) has the highest importance for the prediction of Marine, Freshwater, Wastewater, Soil, Animal Digestive System and Skin. IPR002513 (tract Tn3 transposase DDE domain) has the highest importance for the prediction of Plants, Food Production, Mammals Gastrointestinal and Human Respiratory System.

4 Discussion and conclusion

Here, we introduce DeepToA, an ensemble deep-learning framework that aims at predicting the 'ToA' of a microbiome from the taxonomic and functional profiles of its metagenome. To the best of our knowledge, DeepToA is one of the first deep-learning approaches to focus on metagenomic data, rather than 16S community profile data, and to utilize both taxonomic and functional profiles (Danko *et al.*, 2021; Shenhav *et al.*, 2019; Wu *et al.*, 2021; Zha *et al.*, 2022).

In addition to the ToA classifier, we also provide explanations in terms of the initial taxonomic and functional profiles. We see that, not surprisingly, taxa known to be associated with a particular ToA can have a high associated importance score. However, there are also puzzling appearances, such as *P.acnes* in Marine, and Pelagibacterales in Human Respiratory System and Skin.

We provide a pre-trained embedding matrix specifically for mapping textual taxonomy information to a numeric vector.

As a machine-learning approach, DeepToA will benefit from increases in the amount of data available for training. With a significant increase in the number of sequenced genomes and metagenomes, it will be possible to improve DeepToA so as to distinguish between a larger number of 'theaters of activity', including 'cryptic' ones that are not already obvious during sample collection. While we focus here on distinguishing between 10 diverse 'theaters of activity', we envision future classifiers addressing a much finer classification, between a 'healthy' and 'diseased' human respiratory system, say.

Data availability

Web server and data are available at http://ab.inf.uni-tuebingen.de/software/deeptoa. Details of the training and test data can be found in Supplementary Files S1 and S2. Code is available at https://github.com/husonlab/deeptoa.

Acknowledgements

We acknowledge hardware support by the High Performance and Cloud Computing Group at the Zentrum für Datenverarbeitung of the University of Tübingen, the state of Baden-Württemberg through bwHPC, the German Research Foundation (DFG) through grant no. INST 37/935-1 FUGG. We also acknowledge support of the BMBF-funded de. NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A532B, 031A533A, 031A533B, 031A534A, 031A535A, 031A537A, 031A537B, 031A537C, 031A537D and 031A538A).

Authors' contributions

D.H.H., A.G. and W.Z. designed the study. W.Z. developed the machine-learning approach and carried out model-based analysis. A.G. and W.Z. performed the microbiome analysis. D.H.H, W.Z. and A.G. and wrote and revised the manuscript. A.G. designed the web server. A.G. and W.Z. packaged the software. All authors discussed the results and edited the manuscript.

Conflict of Interest: none declared.

References

Arcadu,F. *et al.* (2019) Deep learning algorithm predicts diabetic retinopathy progression in individual patients. *NPJ Digit. Med.*, **2**, 1–9.

Ardila,D. *et al.* (2019) End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat. Med.*, **25**, 954–961.

Berg,G. *et al.* (2020) Microbiome definition re-visited: old concepts and new challenges. *Microbiome*, **8**, 103.

Blum,M. *et al.* (2021) The interpro protein families and domains database: 20 years on. *Nucleic Acids Res.*, **49**, D344–D354.

Borry,M. (2019) Sourcepredict: prediction of metagenomic sample sources using dimension reduction followed by machine learning classification. *J. Open Source Softw.*, **4**, 1540.

Bukhari,A.H. *et al.* (2020) Fractional neuro-sequential arfima-lstm for financial market forecasting. *IEEE Access*, **8**, 71326–71338.

Calinski,T. and Harabasz,J. (1974) A dendrite method for cluster analysis. *Commun. Stat. – Theory Methods*, **3**, 1–27.

Caporaso,J.G. *et al.* (2010) Qiime allows analysis of high-throughput community sequencing data. *Nat. Methods.*, **7**, 335–336.

Danko,D. *et al.*; International MetaSUB Consortium. (2021) A global metagenomic map of urban microbiomes and antimicrobial resistance. *Cell*, **184**, 3376–3393.e17.

Davies,D.L. and Bouldin,D.W. (1979) A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.*, **PAMI-1**, 224–227.

DeLong,E.R. *et al.* (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, **44**, 837–845.

Devlin,J. *et al.* (2018). BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, pp. 4171–4186.

Grover,A. and Leskovec,J. (2016). node2vec: scalable feature learning for networks.

Handelsman,J. (2004) Metagenomics: application of genomics to uncultured microorganisms. *Microbiol. Mol. Biol. Rev.*, **68**, 669–685.

Hochreiter,S. and Schmidhuber,J. (1997) Long short-term memory. *Neural Comput.*, **9**, 1735–1780.

Huson,D. *et al.* (2016) MEGAN community edition - interactive exploration and analysis of large-scale microbiome sequencing data. *PLoS Comput. Biol.*, **12**, e1004957.

Jumper,J. *et al.* (2021) Highly accurate protein structure prediction with alphafold. *Nature*, **596**, 583–589.

Knights,D. *et al.* (2011) Bayesian community-wide culture-independent microbial source tracking. *Nat. Methods.*, **8**, 761–763.

Le,Q. and Mikolov,T. (2014). Distributed representations of sentences and documents. In: *International Conference on Machine Learning*, *PMLR*. pp. 1188–1196.

Li,Y. *et al.* (2021) HMD-ARG: hierarchical multi-task deep learning for annotating antibiotic resistance genes. *Microbiome*, **9**, 40–12.

Lundberg,S.M. and Lee,S.-I. (2017). A unified approach to interpreting model predictions. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*, vol. 30, pp. 4768–4777.

Mitchell,A.L. *et al.* (2020) Mgnify: the microbiome analysis resource in 2020. *Nucleic Acids Res.*, **48**, D570–D578.

Najafabadi,M.M. *et al.* (2015) Deep learning applications and challenges in big data analytics. *J. Big Data*, **2**, 1–21.

Oh,M. and Zhang,L. (2020) Deepmicro: deep representation learning for disease prediction based on microbiome data. *Sci. Rep.*, **10**, 1–9.

Oudah,M. and Henschel,A. (2018) Taxonomy-aware feature engineering for microbiome classification. *BMC Bioinformatics*, **19**, 1–13.

Parks,D.H. *et al.* (2021) GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.*, **50**(D1), D785–D794.

Peters,M.E. *et al.* (2018). Deep contextualized word representations. In: *Proceedings of NAACL*.

Rajpurkar,A.R. *et al.* (2021) Deep learning connects DNA traces to transcription to reveal predictive features beyond enhancer–promoter contact. *Nat. Commun.*, **12**, 1–15.

Reichstein,M. *et al.* (2019) Deep learning and process understanding for data-driven earth system science. *Nature*, **566**, 195–204.

Rousseeuw,P.J. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, **20**, 53–65.

Schoch,C.L. *et al.* (2020) NCBI taxonomy: a comprehensive update on curation, resources and tools. *Database (Oxford)*, **2020**,

Sharma,D. and Xu,W. (2021) Phylostm: a novel deep learning model on disease prediction from longitudinal microbiome data. *Bioinformatics*, **37**, 3707–3714.

Sharma,D. *et al.* (2020) Taxonn: ensemble of neural networks on stratified microbiome data for disease prediction. *Bioinformatics*, **36**, 4544–4550.

Shenhav,L. *et al.* (2019) Feast: fast expectation-maximization for microbial source tracking. *Nat. Methods*, **16**, 627–632.

Smith,A. *et al.* (2010) Novel application of a statistical technique, random forests, in a bacterial source tracking study. *Water Res.*, **44**, 4067–4076.

The Human Microbiome Project Consortium (2012) Structure, function and diversity of the healthy human microbiome. *Nature*, **486**, 207–214.

Thompson,J. *et al.* (2019) Machine learning to predict microbial community functions: an analysis of dissolved organic carbon from litter decomposition. *PLoS One*, **14**, e0215502.

van der Maaten,L. and Hinton,G. (2008) Visualizing data using t-SNE. *J. Mach. Learn. Res.*, **9**, 2579–2605.

Whipps,J.M. and Lewis,K.C.R. (1988). Mycoparasitism and plant disease control. In: Burge,N. (ed.) *Fungi in Biological Control Systems*. Manchester University Press, Manchester, UK, pp. 161–187.

Wu,J. *et al.* (2021) Tracking major sources of water contamination using machine learning. *Front. Microbiol.*, **11**, 616692.

Yap,M. *et al.* (2021) Verifying explainability of a deep learning tissue classifier trained on rna-seq data. *Sci. Rep.*, **11**, 1–12.

Zha,Y. *et al.* (2022) Ontology-aware deep learning enables ultrafast and interpretable source tracking among Sub-million microbial community samples from hundreds of niches. *Genome Med.*, **14**, 1–17.

Zhou,J. *et al.* (2021) Kernel principal components based Cascade Forest towards disease identification with human microbiota. *BMC Med. Inform. Decis. Mak.*, **21**, 1–15.

D Manuscript 4

Title: On the Application of Advanced Machine Learning Methods to Analyze Enhanced, Multimodal Data from Persons Infected with COVID-19

computation

MDPI

Article

On the Application of Advanced Machine Learning Methods to Analyze Enhanced, Multimodal Data from Persons Infected with COVID-19

Wenhuan Zeng [1,*,†], Anupam Gautam [1,2,†] and Daniel H. Huson [1]

[1] Institute for Bioinformatics and Medical Informatics, University of Tübingen, Sand 14, 72076 Tübingen, Germany; anupam.gautam@uni-tuebingen.de (A.G.); daniel.huson@uni-tuebingen.de (D.H.H.)
[2] International Max Planck Research School 'From Molecules to Organisms', Max Planck Institute for Developmental Biology, Max-Planck-Ring 5, 72076 Tübingen, Germany
* Correspondence: wenhuan.zeng@uni-tuebingen.de
† These authors contributed equally to this work.

Abstract: The current COVID-19 pandemic, caused by the rapid worldwide spread of the SARS-CoV-2 virus, is having severe consequences for human health and the world economy. The virus affects different individuals differently, with many infected patients showing only mild symptoms, and others showing critical illness. To lessen the impact of the epidemic, one problem is to determine which factors play an important role in a patient's progression of the disease. Here, we construct an enhanced COVID-19 structured dataset from more than one source, using natural language processing to add local weather conditions and country-specific research sentiment. The enhanced structured dataset contains 301,363 samples and 43 features, and we applied both machine learning algorithms and deep learning algorithms on it so as to forecast patient's survival probability. In addition, we import alignment sequence data to improve the performance of the model. Application of Extreme Gradient Boosting (XGBoost) on the enhanced structured dataset achieves 97% accuracy in predicting patient's survival; with climatic factors, and then age, showing the most importance. Similarly, the application of a Multi-Layer Perceptron (MLP) achieves 98% accuracy. This work suggests that enhancing the available data, mostly basic information on patients, so as to include additional, potentially important features, such as weather conditions, is useful. The explored models suggest that textual weather descriptions can improve outcome forecast.

Keywords: COVID-19; machine learning; deep learning; NLP; weather; sentiment analysis

Citation: Zeng, W.; Gautam, A.; Huson, D.H. On the Application of Advanced Machine Learning Methods to Analyze Enhanced, Multimodal Data from Persons Infected with COVID-19. *Computation* **2021**, *9*, 4. https://doi.org/10.3390/computation9010004

Received: 8 December 2020
Accepted: 1 January 2021
Published: 7 January 2021

1. Introduction

The current COVID-19 pandemic, caused by the rapid worldwide spread of the SARS-CoV-2 virus, is affecting many aspects of society, in particular human health (at the time of writing, over 66 million diagnosed cases and 1.5 million deaths [1]), but also social issues [2,3], mental health, and the economy [4]. Researchers from different scientific fields, including immunology, genetics, and bioinformatics, are studying the pandemic to find ways to slow its progression.

Machine learning approaches are also part of this endeavor [5–9]. For example, Shahid et al. [10] use several models, including ARIMA, SVR, LSTM, and Bi-LSTM, for time series prediction of confirmed cases, deaths, and recoveries in ten major countries affected by COVID-19. Shreshth et al. [11] present a machine learning model to predict how the number of cases of COVID-19 will develop, and to forecast when a specific country can expect to see an end of the pandemic, using the FogBus framework. Other researchers have built machine learning models for the classification and diagnosis of COVID-19 that are based on medical images [12,13]. Further, Yan et al. [14] provide an interpretable mortality model that is based on a database of blood samples from 485 infected patients in the region of Wuhan,

China. To date, most machine learning and deep learning research [15,16] on COVID-19 build a classification model on various types of data to investigate which might be the important features to predict a specific outcome. One potential difficulty when running such approaches on publicly available dataset is that the features are originally collected so as to fulfill the needs of the data provider, which then can be a source of bias, when the data is used to address other questions. In particular, features that have high predictive value for the outcome for an infected patient might be missing. Generally speaking, the presence or absence of features will impact the accuracy of a model.

The COVID-19 data provided by Xu et al. [17] contain a large number of samples, but limited features that mainly provide basic information on patients. Here, we seek to improve the usefulness of this data by adding a number of features that might help to increase the accuracy of a predictive model.

Research indicates that local climate plays a roles in pandemic outbreaks [18]. Lowen et al. [19] demonstrated that aerosol spread of the influenza virus is dependent upon both ambient relative humidity and temperature, using guinea pig as a model host. Tan et al. [20] investigated the effect of weather in four cities in China and concluded that SARS outbreaks were significantly associated with the temperature and its variations. For the SARS-CoV-2 virus, there are some contradicting findings. Initial studies suggested a negative correlation between temperature and COVID-19 infection [21], or temperature-independence [22], while other research detected a positive relation between temperature and COVID-19 cases at temperatures below 3 °C [23], and also relates temperature to decrease in spread parameters of the case dynamics [24]. Therefore, local weather factors should be taken into consideration.

Infection and mortality rates differ between countries, as does the response to the pandemic. A study on news platforms and social media indicates that more than half (52%) of all news headlines evoked negative sentiments [25], on the one hand, whereas public positive tweets outweighed negative tweets on the other hand [26]. Application of machine learning algorithms on such data indicates a growth in fear and negative sentiment [27]. To explore this further, in this study we assume that a researchers attitude toward COVID-19, optimistic or pessimistic, will reflect the situation in their country, to some extent, and might be detectable in their publications on the pandemic.

While most previous work focuses on a single data type, in this study, we combine multiple data types. While a number of papers focus on country-wise pandemic prediction [28–30], here we develop a classification model that is based on worldwide data.

We first built an initial structured dataset on patients that tested positive for the virus, based on the work in [17]. We then constructed an enhanced structured dataset by adding new features based on (1) the local weather conditions when the patient was probably infected, and (2) the average weighted average polarity score for research abstracts on the pandemic, per country.

Another reasonable hypothesis is that the specific genome sequence of the virus that affected a given patient may help predict the outcome for the patient. There is research that associates genomic variations with mortality rate of COVID-19 [31], and further research [32] shows that the SARS-CoV-2 virus carries 7.23 mutations per sample compared to the reference, on average. There is work that attempts to predict outcome using machine learning and deep learning methods [33,34]. Both NCBI [35] and GISAID [36,37] provide genomic data for the virus.

Ideally, we would have liked to further enhance the initial dataset by adding virus genome sequences to each sample. Unfortunately, these sequences are not available. So, to explore the use of genomic sequences, we created an additional sequence dataset that consists of unknown patients and their virus sequence, obtained from GISAID.

In this paper, we investigated the application of two algorithms—XGBoost and MLP—to build models both on the initial structured dataset and also on the enhanced structured dataset. In addition, we built a Bi-LSTM model on the sequence dataset. The applied analysis pipelines are summarized in Figure 1.

Based on the initial dataset, we confirm that age is one of the most important factors for predicting survival. When considering the enhanced structured dataset, we find that the weather textual description, followed by local temperature, humidity, and age, arise as the most important features. On the enhanced data, we found that the Extreme Gradient Boosting (XGBoost) method achieved 97% accuracy in predicting a patient's survival. We describe how to predict patient's outcome using a combination of a Multi-Layer Perceptron (MLP) and Bidirectional Long Short-Term Memory (Bi-LSTM), using both the enhanced structured dataset, and the sequence dataset, respectively.

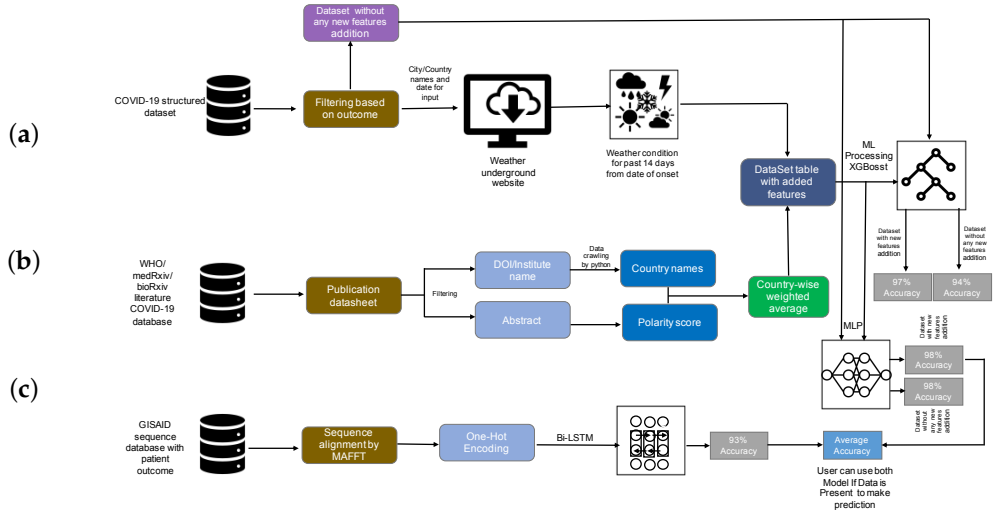


Figure 1. Analysis summary. (**a**) The initial COVID-19 structured dataset was filtered for patients for which the outcome has been recorded, and then, for these items, the weather was determined using the Weather Underground website [38]. (**b**) The WHO, medRxiv, and bioRxiv COVID-19 literature database were filtered and preprocessed to extract author institute/address/country, and these were postprocessed so as to obtain a country-wise research sentiment polarity score. XGBoost and Multi-Layer Perceptron (MLP) were trained on both the initial and the enhanced structured data, and the accuracy of survival prediction was shown to be 94% and 97% (using XGBoost), and 98% and 98% (using MLP), respectively. (**c**) Bidirectional Long Short-Term Memory (Bi-LSTM) was used to train a classification model on the sequence dataset, the accuracy was 93%. Finally, the MLP model and Bi-LSTM models were stacked to jointly predict outcome.

2. Materials

2.1. Data Collection

Data were collected from a number of sources.

2.1.1. COVID-19 Structured Dataset

We downloaded COVID-19 patient data provided by Xu et al. [17] from Github [39], on 21 August 2020 (file latestdata.csv). The dataset includes patient's basic information features, including ID, age, sex, city, province, country, etc. All rows that do not contain a value in the outcome column were dropped, resulting in 307,382 patient data rows out of 2,676,311. The final dataset contained 301,363 patients from 46 countries. All further processing was performed on this dataset.

2.1.2. WHO, medRxiv, and bioRxiv COVID-19 Literature Database

We downloaded a database of literature on COVID-19 from the World Health Organization (WHO) website [40] on 13 April 2020. Of the 5354 downloaded entries, we kept only those whose Journal Name and DOI fields were not blank, which resulted in 4683 publications in 590 journals. This list was extended with COVID-19 SARS-CoV-2 preprints published on medRxiv [41] and bioRxiv [42]. For this we used the bioRxiv API [43] to

download the paper information; a total of 8076 entries were downloaded on 27 August 2020. We then analyzed these publications to determine the authors' institute and country; when no country was explicitly given, we used Google Maps [44] and Wikipedia [45] to determine the country in which the author's institute is located. This gave rise to 9577 (1501 of 4683 WHO, 8076 of 8076 medRxiv and bioRxiv) entries. Finally, we merged the two datasets and removed all duplicates, obtaining 9542 (1484 of 1501 WHO, 8058 of 8076 medRxiv and bioRxiv, Additional File 1) entries in total.

2.1.3. GISAID CoV-19 Sequences Dataset

The GISAID sequence repository contains more than 244,000 genomic sequences for SARS-CoV-2. We downloaded all that were labeled as complete, with high coverage, and were found in a human host on 25 August 2020. This resulted in 4957 genome sequences (with metadata). Further, we included the reference SARS-CoV-2 Wuhan genome (NCBI Accession MN908947.3 [46]) to the dataset and collected the patient information from the publication [47]. Finally, we removed all those sequences that did not have a patient status in the metadata file. Our final dataset contained 4720 sequences (Additional File 2).

2.2. COVID-19-Enhanced Structured Dataset

In this paper, we present an enhanced COVID-19 structured dataset, which is based on the above described initial COVID-19 structured dataset. These data were enhanced by adding features that reflect the weather situation in the location of the infected person, and the research sentiment in units of country, as described in the following.

2.3. Addition Feature Construction

It has been demonstrated that there is a link between environmental factors and the development of COVID-19 [48]. It is reasonable to assume that weather plays a role in disease progression. Therefore, we collected temperature, humidity, and textual description of the weather for the city where the patient lives from the Weather Underground website [38]. Assuming that the incubation period of the virus is approximately 14 days, we collected weather data from 14 days before the patient exhibited relevant symptoms (as recorded in the initial structured dataset).

We also wanted to explore the assumption that researchers' attitudes toward COVID-19, either optimistic or pessimistic, reflect the situation in each country, to some extent, and might be detectable in their publications on the pandemic. Therefore, we collected journal publications from the WHO and from the medRxiv and bioRxiv COVID-19 literature database. For each abstract, we determined the author's institution with the help of the paper's DOI and address by institute name. We applied sentiment analysis to obtain a polarity score on each abstract, and then calculated an weighted average polarity score for each country. Figure 2 displays the weighted average polarity score inferred for different countries.

The weather and sentiment features were added to the initial structured dataset so as to produce the enhanced structured dataset, as outlined in Figure 1.

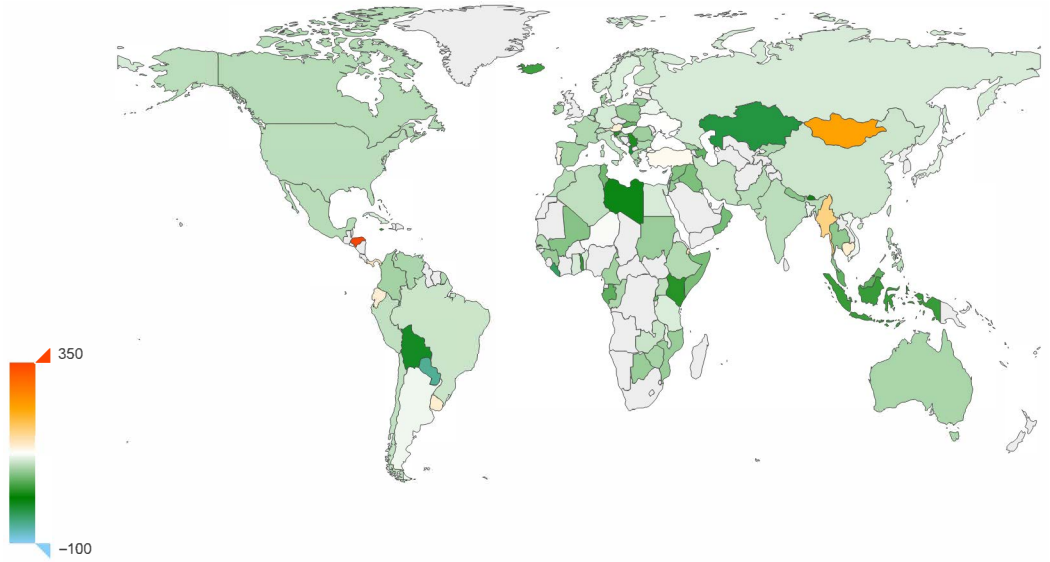


Figure 2. Sentiment polarity score. Average research sentiment polarity score of research, for different countries. Based on a sentiment analysis of abstracts of papers published on COVID-19. One-thousand times the real value.

2.4. Data Processing

2.4.1. Structured Data

The features present in the initial COVID-19 structured dataset include both categorical variables and discrete variables. Each sample in the dataset contains the variables sex, age, the time interval between the patient's onset date, confirmed infected date and admission date, symptoms description, presence of chronic disease, and outcome.

To this initial data, we then added local weather variables (temperature, humidity, and climate description) and the weighted polarity score of the country's scientific research sentiment. The result of this is called the enhanced structured dataset.

To prepare the datasets for building classification models using both XGBoost and MLP (as discussed below), we performed the following steps. We encoded all multi-value text features, such as symptom description (values such as fever, cough, and sputum) or climate description (values such as fair, light rain shower, and cloudy) into three-dimensional embedding vectors, using label encoding on categorical variables such as sex and history of chronic disease (Additional File 3).

We assigned the constant −999 to all missing values. After filtering for samples that have a valid outcome value and city record, we obtained 301,363 samples. Additionally, when we ran MLP, we treated sex and binary chronic disease as categorical features and all others as numerical features, and we normalized all numerical features.

2.4.2. Sequence Data

We performed multiple sequence alignment of the sequence dataset using MAFFT [49], run as follows.

```
mafft --retree 2 --maxiterate 1000 --thread 48 DeathAndAliveForMafft.fasta
  >DeathAndAliveForMafftAlignment1000Iterate.fasta
```

The program required 589 walk-clock minutes to align the 4720 virus genome sequences. The resulting alignment length was 32,015 (Additional File 4).

Furthermore, we applied character-level one-hot encoding on each sequence, mapping each position to a six-dimensional vector (one dimension for each of the four nucleotides, one for the gap character, and one for all ambiguity codes). Each sequence was padded to a fixed length of 33,100 (a multiple of 100), so as to allow us to use 100 time steps in the model described below.

2.5. Data Statistics

We built both a XGBoost model and an MLP model on both the initial structured dataset and on the enhanced structured dataset, respectively.

To evaluate the methods, we split each dataset into a training set and test set in proportion 8:2. Further, to prevent overfitting, we used cross-validation on our training datasets, instead of splitting additional validation sets from the original dataset. As shown in Table 1, the original dataset is typically imbalanced. To address this, we applied the Synthetic Minority Oversampling Technique (SMOTE) [50] to the minority group of each training set, attaining a ratio of positive to negative samples of 10:1. Note that here positive samples refer to patients that survive.

Table 1. Sampling statistics. For the enhanced structured dataset, we report the number of positive and negative samples both in the training set and test set, both before and after oversampling, respectively.

	Enhanced Data		**After Oversampling**	
	Training Set	**Test Set**	**Training Set**	**Test Set**
Positive samples	236,483	59,117	236,483	59,117
Negative samples	4607	1156	23,648	1156
Total	241,090	60,273	260,131	60,273

3. Methods and Experiment

3.1. Sentiment Analysis

A number of papers have studied the forecasting of pandemics using natural language processing on data obtained from various social media [51–53]. Along these lines, we performed sentiment analysis on the abstracts of research papers (associated with COVID-19) using the Python package Textblob [54], which operates by analyzing text content and assigning emotional values to words based on matches to a built-in dictionary.

3.2. Machine Learning Algorithm

Our focus was on the performance of prediction of survival of the infection, based on either the initial or the enhanced structured dataset.

Here, we use the Extreme Gradient Boosting (XGBoost) [55] method to build a prediction model. XGBoost is a powerful member of the gradient boosting family, which is designed to perform well on sparse features, and is known to perform well on Kaggle tasks. This approach avoids overfitting using its built-in L_1 and L_2 regularization on the target function:

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{i=1}^{t} \Omega(f_i). \tag{1}$$

As an additive model, XGBoost consists of k base models, and in most cases we choose the tree model as its base model. Suppose that, for the k-th of t iterations, we train the tree model $f_k(x)$, then

$$\hat{y}_i^t = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_{i-1}^t + f_t(x_i) \tag{2}$$

is the estimated result for the ith sample after t iterations. During construction of each tree, XGBoost minimizes the objective function, with the regularization term show in

Equation (1) in the split phase of each node. In each tree, we calculate the *Gain* of the feature and choose the tree that has the biggest value as the leaf node to be split:

$$Gain = \frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} - \lambda. \quad (3)$$

3.3. Deep Learning Algorithms

To broaden our research and to allow a comparison of methods, we also built deep learning models on both the initial and enhanced structured datasets, together with the sequence dataset, respectively (Figure 3).

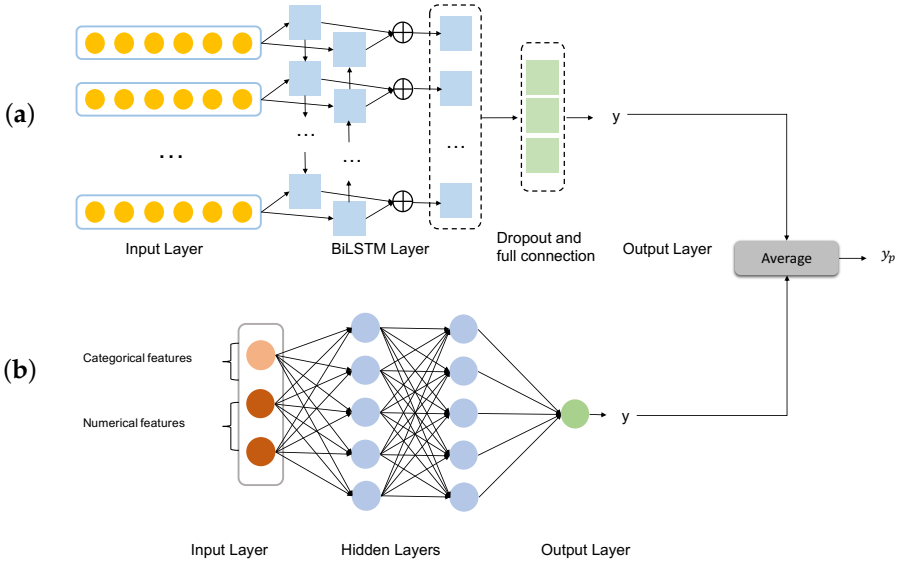


Figure 3. Ensemble deep learning model. (**a**) The MLP is trained on the structured dataset. (**b**) The Bi-LSTM model is trained on the sequence dataset. The two models are stacked in the prediction step.

3.3.1. Multi-Layer Perceptron

As indicated in Figure 3b, we use a simple Multi-Layer Perceptron (MLP) as neural network structure, which has an input layer, hidden layer, and output layer, to build a classification model on the structured dataset.

3.3.2. Bidirectional Long Short-Term Memory

Each sample in our sequence dataset has length 33,100 after alignment and data processing. We can interpret each sequence $X = (x_1, x_2, \cdots, x_n)$ as a time-series, where x_t is the data associated with the tth time point. Recurrent neural networks (RNN) proposed by Elman [56] are commonly used for time series; however, they are not suitable for our task due to the length of the alignments. Long short-term memory (LSTM) [57] is a special variant of RNN. It uses a gate structure in the hidden layer of each time step to protect and control the cell state.

An LSTM cell employs three gates, namely, a forget gate, an input gate, and an output gate, operating as shown in Figure 4. An LSTM learns to memorize and forget specific information during the training step. It provides the ability to capture long-term dependency relationships.

Each gate employs a sigmoid function that aims at producing output values of 0 or 1, defined as

$$\sigma(t) = \frac{1}{1 + e^{-t}} \quad (4)$$

An LSTM does not encode the information in inverse order, so it does not capture the impact of later words on previous words. A bidirectional long short-term memory (Bi-LSTM) overcomes this problem by combining a forward LSTM with a backward LSTM in each time step. This design addresses the issue of bidirectional semantic dependency during model building.

Therefore, we use a Bi-LSTM on our sequence data. Assume we are given a sequence $X = (x_1, x_2, \cdots, x_n)$, where x_t reflects the one-hot encoding. The hidden state of each time point is

$$h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}] \tag{5}$$

In summary, this allows us to consider the impact of the virus sequence information on the patient's condition.

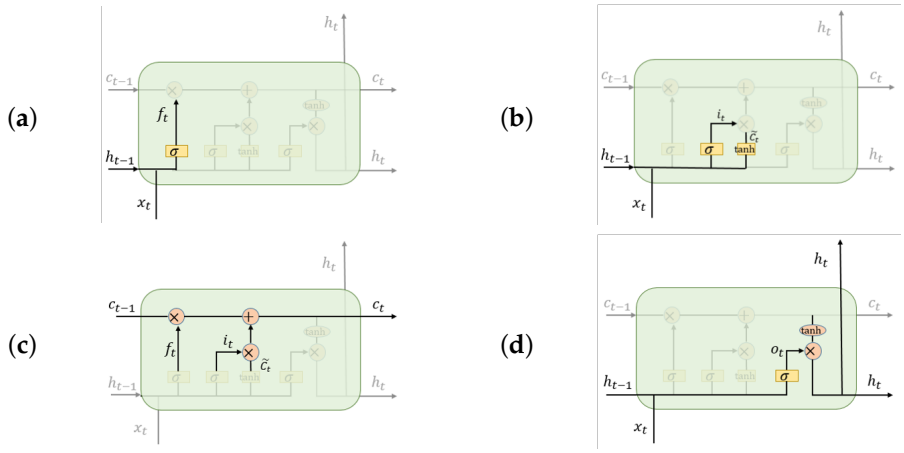


Figure 4. Operation of gates in an LSTM cell. The LSTM determines the hidden state and cell state at the present sequence location as follows. (**a**) A forget gate f_t controls the input of the $(t-1)$th hidden state, (**b**) an input gate i_t controls the input of x_t, (**c**) a transitional phase calculates the tth positions cell state, and then, finally, (**d**) an output gate O_t returns the tth position's hidden state h_t.

Finally, we stacked the MLP and Bi-LSTM deep learning classification models to jointly predict whether the infected patient will survive.

3.4. Implementation

3.4.1. Machine Learning Algorithms

In this study, we ran the XGBoost algorithm both on the initial structured dataset and also on the enhanced structured dataset, the latter additionally containing local weather and research sentiment. To determine the model parameters with the best capacity for prediction, we used GridSearchCV (a function of sklean) to systematically traverse multiple parameter combinations and determine the best parameters through cross-validation. Each subtree in our model is a complicated tree whose maximum depth is 10. Based on the result of model tuning, we set the learning rate to 0.05 and eta to 0.2. Further, we used 1500 estimators, and gamma, alpha, and lambda equal to 0.01, 0.5, and 0.8, respectively.

Each tree was trained on half of the features and half of the samples, chosen at random.

3.4.2. Deep Learning Algorithms

In Figure 3a we show the architecture of the model that accepts aligned sequences. It is a single Bi-LSTM with 128 hidden units and 100 time steps. After randomly dropping 1% of neurons, we use a fully connected layer and ReLU (rectified linear unit) activation function. Output is passed through a sigmoid function.

To model datasets that include both categorical features and normalized numerical features (Figure 3b), we used a 2-layer full connected neural network with 256 hidden units

for each layer. To prevent model overfitting, we dropped a neuron with 5% probability during the forward propagation. A sigmoid function was used to determine output.

During training of both models, we split validation set from training set as proportion 1:3, and to moderate bias created by imbalanced data distribution, we set the class weight ratio between positive samples and negative samples to 1:10. After training as described above, we stacked the two models together so as to obtained average probability, passed through a sigmoid function (Figure 3).

4. Results

We evaluated the algorithms' performance using multiple metrics (Table 2).

Table 2. Performance measures. We report accuracy (Acc.), area under the curve (AUC), F1 score, recall, and precision (Prec.) for the named models and datasets. To compare the performance of the models using the initial or enhanced structured datasets, superior values are shown in bold. (for confusion matrices see Additional file 5).

Model	Dataset	Acc.	AUC	F1 Score	Recall	Prec.
XGBoost	Initial structured dataset	0.94	0.61	0.97	0.96	0.98
	Enhanced structured dataset	**0.97**	**0.77**	**0.99**	**0.99**	0.98
MLP	Initial structured dataset	0.98	0.56	0.99	1.0	0.98
	Enhanced structured dataset	0.98	**0.59**	0.99	1.0	0.98
Bi-LSTM	Sequence dataset	0.93	0.73	0.96	1.0	0.93

4.1. Machine Learning Model

The accuracy of the model created by using the initial structured dataset (no added features) is 94%, whereas using the enhanced structured dataset (with added features), the model's accuracy is 97%. As accuracy on an imbalanced dataset is limited, we display the receiver operating characteristic (ROC) curve of both datasets in Figure 5 to provide a further comparison. The enhanced structured dataset has significantly higher area under the ROC curve (AUC) scores than the model built on the initial structured dataset. There also exist tiny differences between the F1 score, recall, and precision of the two models. The method we chose to evaluate the importance score of feature is based on counting the number of times that a feature occurred in a tree. The feature importance for both datasets is shown in Figure 6. For the initial structured dataset, age plays a more important role than other features. For the model based on the enhanced structured dataset, the weather description, temperature, and humidity are more important than age; moreover, the level of importance of weather is higher than that of age. We visualized the frequency of the textual weather description on survivors and non-survivors, respectively (Figure 7). The weighted average research sentiment polarity score does not have an exceptional f score.

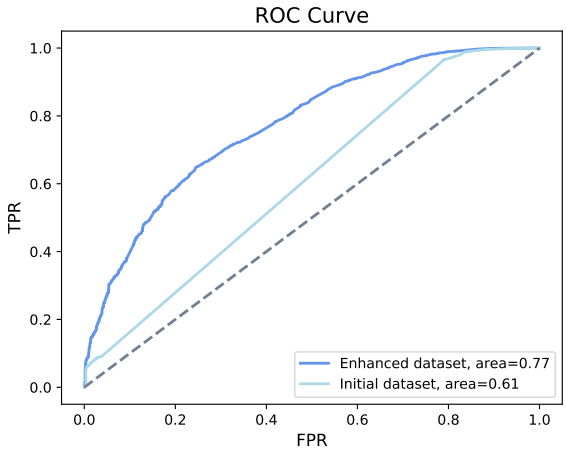
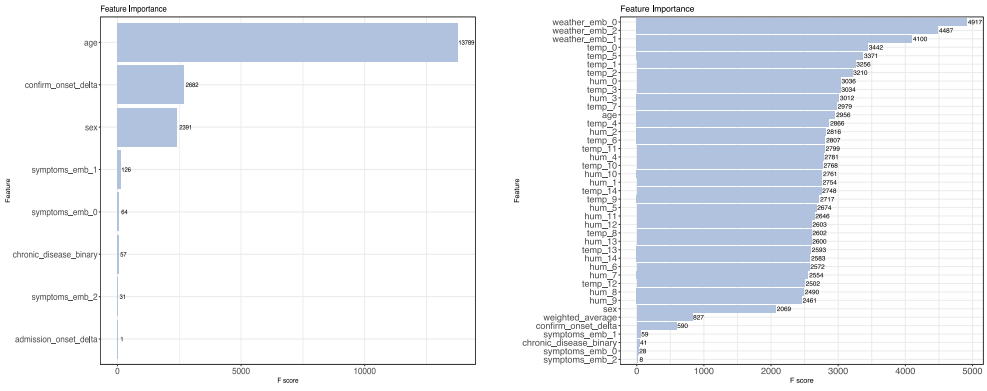


Figure 5. ROC of XGBoost. XGBoost shows an the accuracy of 94% on the initial structured dataset and an accuracy 97% on the enhanced structured dataset, with an increase of the area under the curve from 61% to 77%.



(**a**) (**b**)

Figure 6. Feature scores on the enhanced structured dataset. (**a**) XGBoost processing of the initial structured dataset identified age as an important feature. (**b**) XGBoost processing of the enhanced structured dataset identified in the weather as an important feature.

(**a**) (**b**)

Figure 7. Textual weather description. (**a**) Word cloud visualization of the frequency of textual weather description for survivors. (**b**) Word cloud visualization of the frequency of textual weather description for non-survivors.

4.2. Deep Learning Model

As shown in Table 2, on both the initial and enhanced structured datasets, the MLP method demonstrated higher accuracy than the XGBoost method. For both datasets,

the accuracy using MLP is 98%. However, the ROC curve (Figure 8) indicates that the model shows a better classification ability on the enhanced structured dataset.

Taking sequence data into account, we obtained 93% accuracy and the area under the ROC curve is 0.73, as shown in Figure 9. Among all the models we built, the AUC score was highest when using a Bi-LSTM on the sequence data.

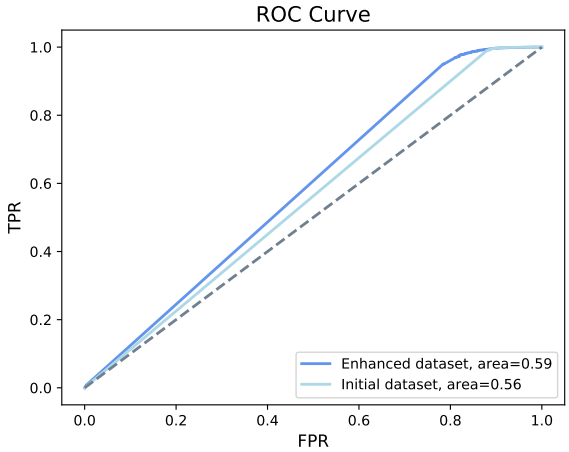


Figure 8. ROC of MLP. MLP shows an accuracy of 98% on both the initial and the enhanced structured dataset, with an increase in area under the curve from 56% to 59%.

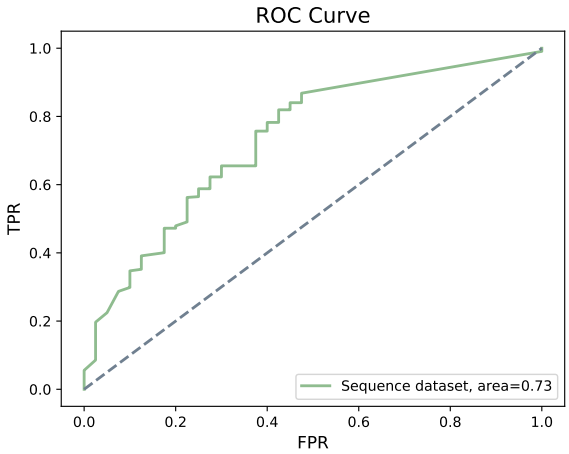


Figure 9. ROC of Bi-LSTM. Bi-LSTM shows an accuracy of 93% on sequence dataset, with an area under the curve of 0.73.

5. Discussion and Conclusions

The performance of machine learning and deep learning methods depends on the amount and quality of available features. Our analysis illustrates that current publicly available data can be enhanced, so as to increase the accuracy of survival prediction by 3% along with positive changes in other model validating metrics, such as AUC (16%), F1 score (2%), and Recall (3%) in case of XGBoost. For MLP the accuracy, F1 score, Recall, and Precision remained the same both for the initial and enhanced structured dataset, but the AUC increased by 3%.

To further evaluate the capability of the proposed models, we repeated the construction of all models on the same datasets, however, with the roles of positive and negative samples reversed, that is, this time considering patients who did *not* survive as positive samples. We observed that for XGBoost and MLP, the models based on the enhanced structured dataset perform better than those based on initial structured dataset in all aspects except recall (see Table 3). Further, it can be observed that even the best model has really poor performances in detecting patients who did not survive, as witnessed by the F1 score of 0.20.

Table 3. Performance measures (predicting death). Considering patients that die as positive samples, we report performance as in the previous Table (for confusion matrices see Additional file 5).

Model	Dataset	Acc.	AUC	F1 Score	Recall	Prec.
XGBoost	Initial structured dataset	0.96	0.60	0.15	**0.19**	0.12
	Enhanced structured dataset	**0.98**	**0.77**	**0.20**	0.13	**0.50**
MLP	Initial structured dataset	0.98	0.55	**0.15**	0.11	**0.21**
	Enhanced structured dataset	0.98	**0.59**	0.13	**0.21**	0.10
Bi-LSTM	Sequence dataset	0.93	0.64	0.21	0.35	0.14

Our study shows how one might enhance a dataset by adding informative features that are not available in the original dataset. Here we demonstrated this for local weather and country-wise research sentiment. Local weather conditions has been implicated as an important feature previous studies.

Our analysis also shows that age is an important factor for survival of COVID-19 as well. However, in the data considered here, the total number of deaths above age 60 were 793 and 2887 survived or were still alive, while in the age group between 40 and 60 there were 421 deaths and 10,346 alive or survived. Therefore, linking mortality to a particular age group is not appropriate based on the current data.

While this analysis suggests that elderly have a higher risk of death, which has already been observed [58,59], saying that mortality is associated with old age is probably generally true for any infectious disease. Age is one of the confounding factors that could be responsible for an increased COVID-19 mortality rate [60,61].

For the model based on the enhanced structured dataset, the weather textual description, followed by local temperature, humidity, and age, appear as the most important features and account for the increase in the accuracy of the model. The most apparent difference in the weather attributes for survivors and non-survivors (Figure 7) is "smoke". This suggests that environmental conditions, in particular air pollution, may play a role in determining the outcome of the disease.

In contrast, in our investigation, the research sentiment score did not show the importance that we had suspected. The values of this feature are never particular high or low, and the highest value of this feature is only 0.35, and thus the difference between the highest score and lowest score is also small. We assume that one of the reasons for this is that academic writing aims for a neutral tone.

The model that we developed on the virus genome dataset failed to provide added predictive power. We suspect that virus genome data would be much more useful, if it were available for the large, structured dataset. However, our study may provide a starting point for further work.

Further, this analysis confirms that enhancing a dataset, rather than just analyzing the originally given features, might lead to a better prediction of a particular outcome. Along with some of the features which should be paid more attention while collecting the data.

There are a number of possible directions for future work. As more viral genomes become available, more powerful Deep Learning methods can be applied to them to help predict patient survival. Additional features such as patient health status, weight, height, medical history should also be integrated. The effect of climate on patient survival warrants more investigation. Finally, methods such as a Recurrent Neural Network-based LSTM might help to study how mutations influence the transmissibility of the virus [62].

Supplementary Materials: Additional files, datasets and models analyzed during our study along with the supplementary materials (like scripts) can be accessed at https://github.com/husonlab/covid19paper. Additional file 1: Merge and processed publication data downloaded from WHO, medRxiv and bioRxiv COVID-19 literature database; Additional file 2: Sequences used for MAFFT alignment downloaded from GISAID and NCBI; Additional file 3: COVID-19 Enhanced structured dataset; Additional file 4: Aligned sequence used for Bi-LSTM; Additional file 5: Confusion matrices for all the build models.

Author Contributions: D.H.H. proposed and guided the project. D.H.H., W.Z., and A.G. wrote the manuscript. W.Z. and A.G. designed the work, wrote the code for data generation, conducted processing, and carried out the analysis. W.Z. carried out the machine learning and deep learning model development and analysis. All authors read and approved the final manuscript.

Funding: This work was supported by the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A537B, 031A533A, 031A538A,031A533B, 031A535A, 031A537C, 031A534A, 031A532B). Furthermore, we acknowledge support by the Open Access Publishing Fund of University of Tübingen.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. These datasets can be found here: Xu et al. dataset. [https://doi.org/10.1038/s41597-020-0448-0], preprints from medRxiv and bioRxiv were accessed by using API: [https://api.biorxiv.org/covid19/help], sequence dataset was download from GISAID: [https://www.gisaid.org/] and NCBI: [https://www.ncbi.nlm.nih.gov/search/all/?term=MN908947] and remaining processed and generated dataset can be downloaded from [https://github.com/husonlab/covid19paper].

Acknowledgments: We would like to thank Caner Bagči for helpful discussions on sequence analysis.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

COVID-19	coronavirus disease
SARS-CoV-2	severe acute respiratory syndrome coronavirus 2
ARIMA	Autoregressive Integrated Moving Average model
Bi-LSTM	Bidirectional Long Short-Term Memory
LSTM	Long Short-Term Memory
MLP	Multi-Layer Perceptron
NLP	Natural Language Processing
WHO	World Health Organization
NCBI	National Center for Biotechnology Information
GISAID	Global initiative on sharing all influenza data
RNN	Recurrent Neural Network
SVR	Support Vector Regression
XGBoost	Extreme Gradient Boosting

References

1. WHO Coronavirus Disease (COVID-19) Dashboard. Available online: https://covid19.who.int/ (accessed on 5 January 2021)
2. Torales, J.; O'Higgins, M.; Castaldelli-Maia, J.M.; Ventriglio, A. The outbreak of COVID-19 coronavirus and its impact on global mental health. *Int. J. Soc. Psychiatry* **2020**, *31*, 0020764020915212. [CrossRef]
3. Singh, J.; Singh, J. COVID-19 and its impact on society. *Electron. Res. J. Soc. Sci. Humanit.* **2020**, *2*, *102–105*.
4. Holmes, E.A.; O'Connor, R.C.; Perry, V.H.; Tracey, I.; Wessely, S.; Arseneault, L.; Ballard, C.; Christensen, H.; Silver, R.C.; Everall, I.; et al. Multidisciplinary research priorities for the COVID-19 pandemic: A call for action for mental health science. *Lancet Psychiatry* **2020**, *7*, 547–560. [CrossRef]
5. Lalmuanawma, S.; Hussain, J.; Chhakchhuak, L. Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. *Chaos Solitons Fractals* **2020**, *139*, 110059. [CrossRef] [PubMed]

6. Ramchandani, A.; Fan, C.; Mostafavi, A. Deepcovidnet: An interpretable deep learning model for predictive surveillance of covid-19 using heterogeneous features and their interactions. *IEEE Access* **2020**, *8*, 159915–159930. [CrossRef]
7. Wang, P.; Zheng, X.; Li, J.; Zhu, B. Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics. *Chaos Solitons Fractals* **2020**, *139*, 110058. [CrossRef] [PubMed]
8. Mirri, S.; Delnevo, G.; Roccetti, M. Is a COVID-19 Second Wave Possible in Emilia-Romagna (Italy)? Forecasting a Future Outbreak with Particulate Pollution and Machine Learning. *Computation* **2020**, *8*, 74. [CrossRef]
9. Alakus, T.B.; Turkoglu, I. Comparison of deep learning approaches to predict COVID-19 infection. *Chaos Solitons Fractals* **2020**, *140*, 110120. [CrossRef]
10. Shahid, F.; Zameer, A.; Muneeb, M. Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM. *Chaos Solitons Fractals* **2020**, *140*, 110212. [CrossRef]
11. Tuli, S.; Tuli, S.; Tuli, R.; Gill, S.S. Predicting the Growth and Trend of COVID-19 Pandemic using Machine Learning and Cloud Computing. *Internet Things* **2020**, *11*, 100222. [CrossRef]
12. Elaziz, M.A.; Hosny, K.M.; Salah, A.; Darwish, M.M.; Lu, S.; Sahlol, A.T. New machine learning method for image-based diagnosis of COVID-19. *PLoS ONE* **2020**, *15*, e0235187. [CrossRef] [PubMed]
13. Barstugan, M.; Ozkaya, U.; Ozturk, S. Coronavirus (Covid-19) classification using ct images by machine learning methods. *arXiv* **2020**, arXiv:2003.09424.
14. Yan, L.; Zhang, H.-T.; Goncalves, J.; Xiao, Y.; Wang, M.; Guo, Y.; Sun, C.; Tang, X.; Jing, L.; Zhang, M. An interpretable mortality prediction model for COVID-19 patients. *Nat. Mach. Intell.* **2020**, *2*, 283–288. [CrossRef]
15. Narin, A.; Kaya, C.; Pamuk, Z. Automatic detection of coronavirus disease (COVID-19) using X-ray images and deep convolutional neural networks. *arXiv* **2020**, arXiv:2003.10849.
16. Magar, R.; Yadav, P.; Farimani, A.B. Potential neutralizing antibodies discovered for novel corona virus using machine learning. *arXiv* **2020**, arXiv:22003.08447.
17. Xu, B.; Gutierrez, B.; Mekaru, S.; Sewalk, K.; Goodwin, L.; Loskill, A.; Cohn, E.L.; Hswen, Y.; Hill, S.C.; Cobo, M.M.; et al. Epidemiological data from the COVID-19 outbreak, real-time case information. *Sci. Data* **2020**, *7*. [CrossRef]
18. Lin, K.; Fong, D.Y.T.; Zhu, B.; Karlberg, J. Environmental factors on the SARS epidemic: Air temperature, passage of time and multiplicative effect of hospital infection. *Epidemiol. Infect.* **2006**, *134*, 223–230. [CrossRef] [PubMed]
19. Lowen, A.C.; Mubareka, S.; Steel, J.; Palese, P. Influenza virus transmission is dependent on relative humidity and temperature. *PLoS Pathog.* **2007**, *3*, 151. [CrossRef]
20. Tan, J.; Mu, L.; Huang, J.; Yu, S.; Chen, B.; Yin, J. An initial investigation of the association between the SARS outbreak and weather: With the view of the environmental temperature and its variation. *J. Epidemiol. Community Health* **2005**, *59*, 186–192. [CrossRef]
21. Prata, D.N.; Rodrigues, W.; Bermejo, P.H. Temperature significantly changes COVID-19 transmission in (sub)tropical cities of brazil. *Sci. Total. Environ.* **2020**, *729*, 138862. [CrossRef]
22. Jamil, T.; Alam, I.; Gojobori, T.; Duarte, C.M. No evidence for temperature-dependence of the COVID-19 epidemic. *Front. Public Health* **2020**, *8*, 436. [CrossRef] [PubMed]
23. Xie, J.; Zhu, Y. Association between ambient temperature and COVID-19 infection in 122 cities from China. *Sci. Total. Environ.* **2020**, *724*, 138201. [CrossRef] [PubMed]
24. Demongeot, J.; Flet-Berliac, Y.; Seligmann, H. Temperature decreases spread parameters of the new COVID-19 case dynamics. association between ambient temperature and COVID-19 infection in 122 cities from China. *Biology* **2020**, *9*, 94. [CrossRef]
25. Aslam, F.; Awan, T.M.; Syed, J.H.; Kashif, A.; Parveen, M. Sentiments and emotions evoked bynews headlines of coronavirus disease (covid-19) outbreak. *Humanit. Soc. Sci. Commun.* **2020**, *7*, 1–9. [CrossRef]
26. Hung, M.; Lauren, E.; Hon, E.S.; Birmingham, W.C.; Xu, J.; Su, S.; Hon, S.D.; Park, J.; Dang, P.; Lipsky, M.S. Social network analysis of covid-19 sentiments: Application of artificial intelligence. *J. Med. Internet Res.* **2020**, *22*, e22590. [CrossRef]
27. Samuel, J.; Ali, G.G.; Rahman, M.; Esawi, E.; Samuel, Y. Covid-19 public sentiment insights and machine learning for tweets classification. *Information* **2020**, *11*, 314. [CrossRef]
28. Souza, F.S.H.; Hojo-Souza, N.S.; Santos, E.B.; Silva, C.M.; Guidoni, D.L. Predicting the disease outcome in COVID-19 positive patients through Machine Learning: A retrospective cohort study with Brazilian data. *medRxiv* **2020**. [CrossRef]
29. Pinter, G.; Felde, I.; Mosavi, A.; Ghamisi, P.; Gloaguen, R. COVID-19 Pandemic Prediction for Hungary: A Hybrid Machine Learning Approach. *Mathematics* **2020**, *8*, 890. [CrossRef]
30. Arora, P.; Kumar, H.; Panigrahi, B.K. Prediction and analysis of COVID-19 positive cases using deep learning models: A descriptive case study of India. *Chaos Solitons Fractals* **2020**, *139*, 110017. [CrossRef]
31. Toyoshima, Y.; Nemoto, K.; Matsumoto, S.; Nakamura, Y.; Kiyotani, K. SARS-CoV-2 genomic variations associated with mortality rate of COVID-19. *J. Hum. Genet.* **2020**, *65*, 1075–1082. [CrossRef]
32. Mercatelli, D.; Giorgi, F.M. Geographic and genomic distribution of SARS-CoV-2 mutations. *Front. Microbiol.* **2020**. [CrossRef] [PubMed]
33. Bhonde, S.; Bhati, M.; Prasad, J. Predictive Analytics to Combat with COVID-19 Using Genome Sequencing. 2020. Available online: https://ssrn.com/abstract=3580692 (accessed on 5 January 2021).
34. Machine Learning for Biology: How Will COVID-19 Mutate Next? Available online: https://towardsdatascience.com/machine-learning-for-biology-how-will-covid-19-mutate-next-4df93cfaf544 (accessed on 5 January 2021).

35. National Center for Biotechnology Information. Available online: https://www.ncbi.nlm.nih.gov/ (accessed on 5 January 2021).
36. Elbe, S.; Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob. Challenges* **2017**, *1*, 33–46. [CrossRef] [PubMed]
37. Shu, Y.; McCauley, J. GISAID: Global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance* **2017**, *22*, 13. [CrossRef] [PubMed]
38. Weather Underground. Available online: https://www.wunderground.com/ (accessed on 5 January 2021).
39. nCoV2019. Available online: https://github.com/beoutbreakprepared/nCoV2019/tree/master/latest_data (accessed on 5 January 2021).
40. Global Research on Coronavirus Disease (COVID-19). Available online: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/global-research-on-novel-coronavirus-2019-ncov (accessed on 5 January 2021).
41. medRxiv. Available online: https://www.medrxiv.org/ (accessed on 5 January 2021).
42. bioRxiv. Available online: https://www.biorxiv.org/ (accessed on 5 January 2021).
43. API Summary for the Collection of COVID-19 SARS-CoV-2 Preprints from medRxiv and bioRxiv. Available online: https://api.biorxiv.org/covid19/help (accessed on 5 January 2021).
44. Google Map. Available online: https://www.google.com/maps/ (accessed on 5 January 2021).
45. WIKIPEDIA. Available online: https://www.wikipedia.org/ (accessed on 5 January 2021).
46. NCBI Accession MN908947.3. Available online: https://www.ncbi.nlm.nih.gov/search/all/?term=MN908947 (accessed on 5 January 2021).
47. Wu, F.; Zhao, S.; Yu, B.; Chen, Y.M.; Wang, W.; Song, Z.G.; Hu, Y.; Tao, Z.W.; Tian, J.H.; Pei, Y.Y.; et al. A new coronavirus associated with human respiratory disease in China. *Nature* **2020**, *579*, 265269. [CrossRef]
48. Triplett, M. Evidence that higher temperatures are associated with lower incidence of COVID-19 in pandemic state, cumulative cases reported up to March 27, 2020. *medRxiv* **2020**. [CrossRef]
49. Katoh, K.; Standley, D.M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **2013**, *30*, 772–780. [CrossRef]
50. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2020**, *16*, 321–357. [CrossRef]
51. Alessa, A.; Faezipour, M. A review of influenza detection and prediction through social networking sites. *Theor. Biol. Med Model.* **2018**, *15*, 2. [CrossRef]
52. Lee, K.; Agrawal, A.; Choudhary, A. Forecasting influenza levels using real-time social media streams. In Proceedings of the 2017 IEEE International Conference on Healthcare Informatics (ICHI), Park City, UT, USA, 23–26 August 2017; pp. 409–414.
53. Wang, Y.; Xu, K.; Kang, Y.; Wang, H.; Wang, F.; Avram, A. Regional influenza prediction with sampling Twitter data and PDE model. *Int. J. Environ. Res. Public Health* **2020**, *17*, 678. [CrossRef]
54. TextBlob. Available online: https://github.com/sloria/TextBlob (accessed on 5 January 2021).
55. 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*; ACM: New York, NY, USA, 2016; pp. 785–794.
56. Elman, J.L. Finding structure in time. *Cogn. Sci.* **1990**, *14*, 179–211. [CrossRef]
57. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780.
58. Verity, R.; Okell, L.C.; Dorigatti, I.; Winskill, P.; Whittaker, C.; Imai, N.; Cuomo-Dannenburg, G.; Thompson, H.; Walker, P.G.; Fu, H.; et al. Estimates of the severity of coronavirus disease 2019: A model-based analysis. *Lancet Infect. Dis.* **2020**. [CrossRef]
59. Glynn, J.R. Protecting workers aged 60–69 years from COVID-19. *Lancet Infect. Dis.* **2020**. [CrossRef]
60. Wang, H.; Li, T.; Barbarino, P.; Gauthier, S.; Brodaty, H.; Molinuevo, J.L.; Xie, H.; Sun, Y.; Yu, E. Dementia care during COVID-19. *Lancet* **2020**, *395*, 1190–1191.
61. Armitage, R.; Nellums, L.B. COVID-19 and the consequences of isolating the elderly. *Lancet Public Health* **2020**, *5*, e256.
62. Korber, B.; Fischer, W.M.; Gnanakaran, S.; Yoon, H.; Theiler, J.; Abfalterer, W.; Hengartner, N.; Giorgi, E.E.; Bhattacharya, T.; Foley, B.; et al. Tracking changes in SARS-CoV-2 Spike: Evidence that D614G increases infectivity of the COVID-19 virus. *Cell* **2020**, *182*, 812–827.