
PathoLM: Identifying pathogenicity from the DNA sequence through the Genome Foundation Model

Sajib Acharjee Dip¹ Uddip Acharjee Shuvo² Tran Chau³ Haoqiu Song¹ Petra Choi⁴ Xuan Wang¹
Liqing Zhang^{1,3}

Abstract

Pathogen identification is pivotal in diagnosing, treating, and preventing diseases, crucial for controlling infections and safeguarding public health. Traditional alignment-based methods, though widely used, are computationally intense and reliant on extensive reference databases, often failing to detect novel pathogens due to their low sensitivity and specificity. Similarly, conventional machine learning techniques, while promising, require large annotated datasets and extensive feature engineering and are prone to overfitting. Addressing these challenges, we introduce PathoLM, a cutting-edge pathogen language model optimized for the identification of pathogenicity in bacterial and viral sequences. Leveraging the strengths of pre-trained DNA models such as the Nucleotide Transformer, PathoLM requires minimal data for fine-tuning, thereby enhancing pathogen detection capabilities. It effectively captures a broader genomic context, significantly improving the identification of novel and divergent pathogens. We developed a comprehensive data set comprising approximately 30 species of viruses and bacteria, including ESKAPEE pathogens, seven notably virulent bacterial strains resistant to antibiotics. Additionally, we curated a species classification dataset centered specifically on the ESKAPEE group. In comparative assessments, PathoLM dramatically outperforms existing models like DciPatho, demonstrating robust zero-shot and few-shot capabilities. Furthermore, we expanded PathoLM-Sp for ESKAPEE species classification, where it showed superior performance compared to other advanced deep learning methods, despite the complexities of the task.

1. Introduction

Pathogens, ranging from flu viruses to severe disease agents like tuberculosis, can cause significant health challenges, leading to high morbidity and mortality, particularly in areas with limited healthcare. The emergence of antibiotic resistance further complicates treatment, elevating the risk of minor infections becoming fatal. Highlighted by the COVID-19 pandemic, pathogen surveillance using next-generation sequencing (NGS) is critical for monitoring public health across various settings. Effective pathogen identification is essential for timely disease management, informing treatment strategies, and fostering advancements in medical research.

Pathogens, including viruses, bacteria, fungi, and parasites, are major causes of infectious diseases worldwide. Their rapid mutation and reproduction make timely pathogen identification crucial for effective interventions. However, labeled data for these pathogens is scarce. To address this, we curated a dataset focused on ESKAPEE (Ruekit et al., 2022) and viral pathogens, using pathogenic strains from the PATRIC database and non-pathogenic strains from NCBI. This strategic compilation provided a robust dataset allows the PathoLM model to effectively distinguish between pathogenic and non-pathogenic strains. To enhance PathoLM's pathogen detection, we expanded our dataset with additional viral sequences from NCBI, targeting human pathogens for the positive sample pool and using animal pathogens as negatives. This strategy broadened our training scope to include about 30 diverse species, creating a well-balanced dataset for our pathogen language model.

¹Department of Computer Science, Virginia Tech, USA ²Institute of Information Technology, Dhaka University, Bangladesh ³Department of Genetics, Bioinformatics and Computational Biology program (GBCB), Virginia Tech, USA ⁴Department of Civil and Environmental Engineering, Virginia Tech, USA. Correspondence to: Sajib Acharjee Dip <sajibacharjeedip@vt.edu>, Liqing Zhang <lqzhang@vt.edu>.

Previous pathogen detection methods used alignment-based techniques that struggle with novel pathogens and require intensive computation. To address these challenges, recent advancements have integrated machine learning and deep learning strategies, aiming to improve data classification and analysis. However, these advanced methods often require complex feature engineering and large well-annotated datasets for effective training. Many utilize k-mer frequency-based features, which despite their efficacy struggle with scalability and sensitivity, particularly in the detection of new pathogens. Specifically, k-mer-based techniques are adept at analyzing short reads but face obstacles with large-scale data integration and are prone to overfitting (Lorenzi et al., 2020; Watts et al., 2019; Liebhoff et al., 2023). As for our knowledge, the latest method, DciPatho, enhances pathogen identification by combining k-mer frequency features ranging from 3 to 7 k-mers and integrates three computational models—ResNet, DeepNet, and CrossNet—to create a robust intermediate representation. This approach significantly outperforms previous methods like PaPrBaG (Deneke et al., 2017), BacPaCS (Barash et al., 2019), and DeePac (Bartoszewicz et al., 2020). Despite its improvements, DciPatho still faces limitations related to computational time for training from scratch, and its performance heavily depends on the quality and size of the training dataset.

Recent advancements in large foundation models have shown significant promise in various fields, including biomedicine and genomics (Li et al., 2021; Luu & Buehler, 2024; Bi et al., 2024). These models, trained on extensive datasets, capture complex patterns and relationships in data, which can be leveraged to improve pathogen detection. Utilizing pre-trained language models strategically mitigates the need for extensive domain-specific datasets and computational power. In addressing these challenges, we introduced PathoLM, a genome modeling tool that uses the pre-trained Nucleotide Transformer v2 50M (Dalla-Torre et al., 2023) for enhanced pathogen detection in bacterial and viral genomes, both improving accuracy and addressing data limitations. To the best of our knowledge, we are the first to leverage the pre-trained knowledge of DNA language model for pathogen prediction tasks. Moreover, we have demonstrated PathoLM’s capabilities using different foundation models as backbone, in zero-shot and few-shot scenarios, showcasing its robust performance compared to traditional machine learning and deep learning methods in species classification. This model not only addresses the challenges of limited data availability but also sets a new standard in pathogen detection technology.

2. Materials and Methods

2.1. Data collection

ESKAPEE genome For the pathogen dataset, we downloaded 49,642 whole genome assemblies from the PATRIC website (Gillespie et al., 2011), encompassing approximately 27 species of bacteria. From this collection, we retained only the genomes of the seven ESKAPEE pathogens: Escherichia coli, Enterococcus faecium, Staphylococcus aureus, Klebsiella pneumoniae, Acinetobacter baumannii, Pseudomonas aeruginosa, and all species from the Enterobacter genus. These genomes varied significantly in size, ranging from 200 to 1,147,640 base pairs. For the Eskapee non-pathogen data, we sourced our dataset from the NCBI database, ensuring to exclude any strains identified as pathogens in the PATRIC database. We obtained 10,633 whole genome sequences, which originally included around 22 different species.

Table 1. Viral data set for pathogenic and non-pathogenic species

VIRUS	PATHOGENIC	NONPATHOGENIC	REFERENCES
CORONAVIRUS	SARS, MERS, OC43, NL63, 229E, HKU1	BOVINE CORONAVIRUS, CANINE CORONAVIRUS, FELINE CORONAVIRUS, EQUINE CORONAVIRUS, BAT CORONAVIRUS, MINK CORONAVIRUS, MURINE CORONAVIRUS, AVIAN CORONAVIRUS	(SAIB ET AL., 2021)
INFLUENZA	TYPE A, B, C	TYPE D	(CENTERS FOR DISEASE CONTROL AND PREVENTION, 2024)
NOROVIRUS	GI, GII, GIV	GIII, GV, GVI, GVII, GVIII	(HASSAN & BALDRIDGE, 2019)
IMMUNODEFICIENCY VIRUS	HIV	FIV, BIV, SIV	(EGBERINK & HORZINEK, 1992)

Viral genome All viral genome data were downloaded from NCBI. For specific binary prediction, we limited our train dataset to include species that has both human pathogenic and non pathogenic virus. For coronavirus, SARS, MERS, OC43, NL63, 229E, HKU1 strains were considered as pathogenic (Saib et al., 2021). Bovine Coronavirus, Canine coronavirus, Feline coronavirus, Equine Coronavirus, Bat coronavirus, Mink coronavirus, Murine coronavirus, Avian coronavirus were considered as nonpathogenic. For influenza, type A, B, and C were considered human pathogenic while type D was considered as nonpathogenic (Centers for Disease Control and Prevention, 2024). For norovirus, strain GI, GII, GIV were considered as human pathogenic strains while GIII, GV, GVI, GVII, GVIII were considered as non pathogenic virus (Hassan & Baldrige, 2019) (Egberink & Horzinek, 1992). To address data disparities in health surveillance, where non-pathogenic data are less abundant than human pathogenic data due to a focus on clinical patients, we expanded our dataset. We included non-pathogenic viruses often found in wastewater metagenomics, such as Escherichia phage T1 & T7, Bacteriophage sp., Klebsiella pneumoniae phage, and Crassphage, sourced from NCBI. Additionally, we gathered genomes of plant pathogens like Tomato yellow leaf curl virus, pepper mild mottle virus, and tomato mosaic virus to diversify our dataset further.

2.2. Data Processing

2.2.1. DATA FILTERING AND INTEGRATION

Our data preparation process began with the cleaning and organization of ESKAPEE pathogen and non-pathogen datasets separately. We simplified this data set to include only seven species, mirroring the selection criteria used for the pathogen data. The sequence lengths in this dataset were more varied, ranging from 294 to 7,411,863 base pairs. We expanded our binary prediction model by incorporating viral genomes alongside bacterial ones. The viral dataset included 25,229 complete genomes across 765 strains from species with at least 10 genomes each, such as Influenza A and B, Norovirus, HIV, and SARS-CoV-2, among others. Species names were standardized, and for nonpathogenic viruses, we analyzed 1,782 genomes, focusing on the top 14 species. We then integrated these with the ESKAPEE bacterial data, creating a comprehensive dataset for our binary prediction model. Training and testing sets were organized using MMseqs2 clustering to capture the genetic diversity, with detailed clustering methods outlined later.

2.2.2. DATA PARTITIONING USING SEQUENCE CLUSTERING

To ensure the creation of a robust and generalizable evaluation dataset, we employed MMseqs2 (Steinegger & Söding, 2017) for clustering based on sequence similarity. This approach allowed us to systematically explore the impact of similarity thresholds on model performance. Specifically, we establish clusters using coverage and similarity thresholds set at 80%, 60%, and 40%, resulting in three distinct datasets for subsequent experimentation. By selecting the highest sensitivity settings for the clustering process, we aimed to capture the most detailed relationships within our data. We subsequently allocated the generated clusters into training and test sets using an 8:2 split, with the assurance that both sets comprised different clusters. This methodical division was designed to foster a well-generalized dataset, leveraging the advanced capabilities of the MMseqs2 clustering suite to ensure that our train-test splits reflected true biological variations and complexities.

2.2.3. CREATING VARIED SEQUENCE LENGTHS FOR PERFORMANCE SENSITIVITY ANALYSIS

To evaluate the impact of sequence length on model performance, we conducted experiments using a range of sequence lengths derived from whole genome assemblies. Specifically, we fragmented these assemblies into segments of varying lengths, including 150, 500, 2k, 5k, 10k, 50k, as well as using the entire genome sequence as input. For each specified length, we generated a separate dataset. To ensure the integrity of our experimental data, we applied MMseqs2 (Steinegger & Söding, 2017) clustering to each dataset to eliminate highly similar sequences, utilizing both coverage and similarity thresholds. This step was crucial for maintaining diversity within our data and ensuring that our findings were not biased by over represented sequences.

2.3. Pretrained model

2.3.1. PRETRAINING DATASET

We utilized the nucleotide transformer multispecies v2 (50 million parameter) model (Dalla-Torre et al., 2023). This model, pre-trained on a dataset including 3,202 human genomes and 850 genomes from diverse taxa but excluding plants and viruses, shown in Figure 1A, was chosen for its balance of performance and computational efficiency compared to

its 250 million parameter counterpart. The curated dataset from NCBI featured one species per genus to ensure wide phylogenetic representation, totaling 174 billion nucleotides. It spans several taxa, including key organisms like *Escherichia coli*, *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, various *Plasmodium* species, and vertebrates like *Homo sapiens* and *Mus musculus*.

2.3.2. K-MER BASED TOKENIZATION

In natural language processing, we split sentences into words using grammar and semantics, but genomic sequences lack clear delimiters, making tokenization challenging. A typical method is k-mer tokenization, where sequences are divided into overlapping substrings of a fixed length. In this study, we used a specialized tokenizer for nucleic acids, capable of recognizing 4,096 unique 6-mer combinations plus the nucleotides A, T, C, G, and N. This tokenizer also includes special tokens like '[PAD]', '[MASK]', and '[CLS]' for padding, masking, and classification, resulting in a total vocabulary of 4,104 tokens. The process begins with a '[CLS]' token and progresses by converting sequences into tokens from left to right, giving priority to 6-mer tokens and defaulting to individual nucleotide tokens where necessary. This method preserves biological information for deep learning applications, enhancing the pretrained Nucleotide Transformer model's accuracy in pathogen prediction by fine-tuning it on these tokenized sequences.

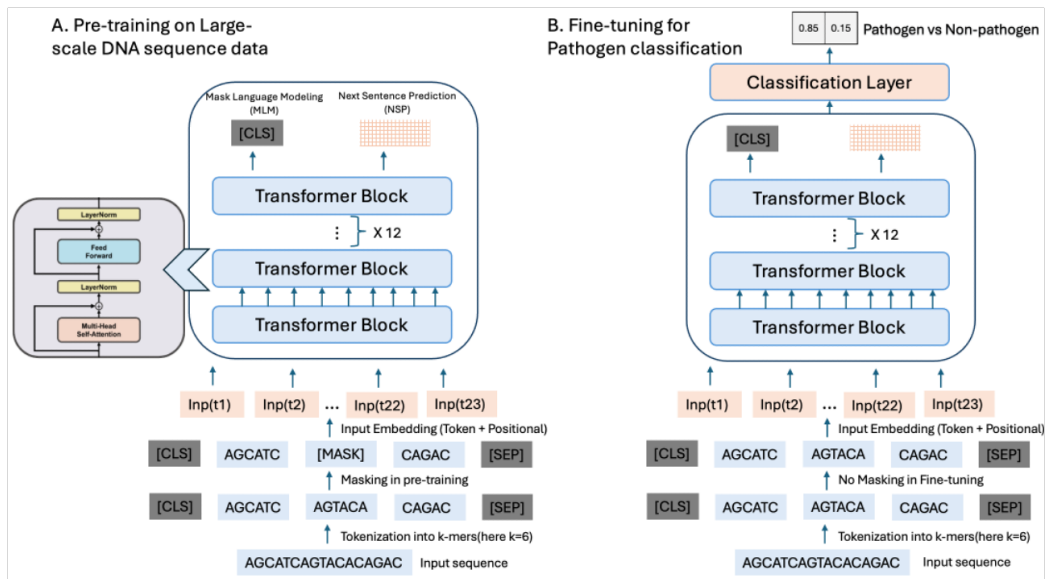


Figure 1. Model Architecture of PathoLM. (A) Pretraining steps of the genome foundation model on a large-scale dataset, where tokenization and masking are applied before feeding the input sequence into the transformer block. (B) Finetuning steps for the pathogen identification task using specific datasets, where no masking is applied.

2.3.3. BASIC ARCHITECTURE OF NUCLEOTIDE TRANSFORMER

The Nucleotide Transformer v2 50M, designed for genomic data, utilizes an encoder-only transformer architecture. It starts by embedding 6-mer sequences into dense vectors, enhanced with Rotary Positional Embeddings for positional context up to 12kb. The model employs multi-head self-attention within each transformer layer to analyze sequence interdependencies, incorporating residual connections to retain information. After normalization, the sequences proceed through a feed-forward network using Gated Linear Units and Swish activation to boost processing efficiency. In training, it predicts nucleotide occurrences to improve genomic data interpretation.

2.4. Fine-tuning

We fine-tuned our pre-trained Nucleotide Transformer model using the Hugging Face Transformers library, selected for its robust support for genomic sequences. We adapted the model to manage different sequence lengths, padding shorter sequences and truncating longer ones beyond the 12,000 nucleotide context limit shown in Figure 1B. The model was optimized for both binary classification of pathogens and detailed classification among ESKAPEE species, using dual and

septenary output labels, respectively. Training involved the Adam optimizer with a specific learning rate and a warm-up phase to prevent premature convergence, coupled with an early stopping mechanism to avoid overfitting. This strategy effectively leveraged the Nucleotide Transformer’s strengths, yielding precise and robust models for pathogen genomics.

2.5. Evaluation

We evaluated our model using key metrics: Accuracy, F1-Score, MCC, and AUC-ROC. Accuracy measures the rate of correct predictions, F1-Score assesses precision and recall balance, MCC provides a robust correlation coefficient regardless of class size, and AUC-ROC evaluates the model’s ability to distinguish between classes at various thresholds. These metrics collectively ensure a thorough assessment of precision, recall, classification quality, and discriminative capacity.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Where TP = True Positives, TN = True Negatives, FP = False Positives, FN = False Negatives.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Where Precision = $\frac{TP}{TP+FP}$ and Recall = $\frac{TP}{TP+FN}$.

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

3. Results

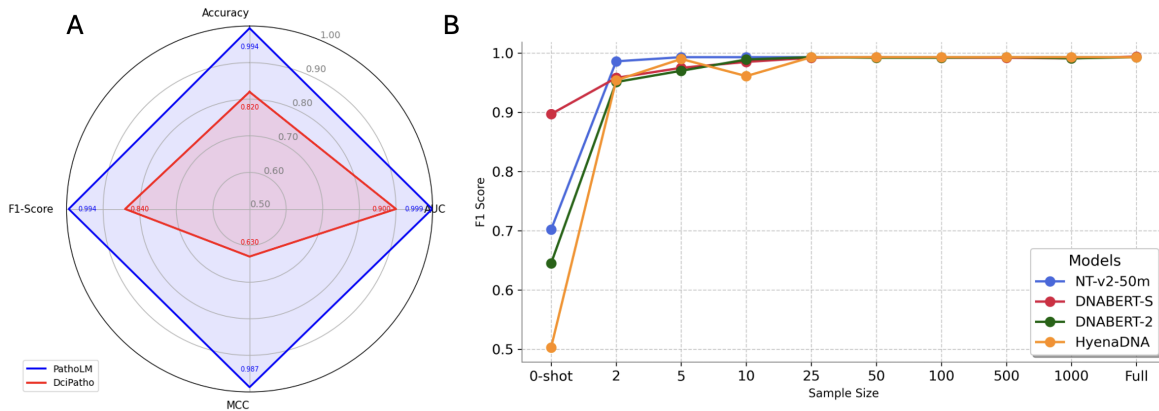


Figure 2. A: Performance comparison between PathoLM and state-of-the-art method DciPatho on the binary pathogen prediction task in terms of Accuracy, F1-score, AUC-ROC, MCC. B: Comparison of F1-scores for various genome language models, highlighting their performance in zero-shot and few-shot scenarios.

3.1. Identifying pathogenicity from PATRIC ESKAPEE species and human pathogen virus

The PathoLM model is built upon the pre-trained foundation model, Nucleotide Transformer. To adapt the pre-trained model parameters for our specific pathogen identification task, we employed transfer learning techniques. Specifically, we used the frozen weights from the Nucleotide Transformer and fine-tuned them on our pathogen identification dataset. To assess the effectiveness of PathoLM, we conducted a comparison with state-of-the-art method, including DciPatho (Jiang et al., 2023) on a binary prediction task. Both PathoLM and DciPatho were trained on the same training dataset and evaluated on the same held-out test dataset to ensure a fair and consistent evaluation. As shown in Figure 2A, PathoLM significantly outperforms DciPatho across all evaluation metrics, thanks to the pre-trained Nucleotide Transformer. This foundation

enhances PathoLM’s ability to detect subtle genomic patterns essential for accurate pathogen identification. Leveraging pre-trained representations allows PathoLM to deliver high accuracy and reliability in pathogen detection, highlighting the effectiveness of transfer learning in genomic applications.

3.2. Zero-Shot and Few-Shot Capabilities in Large Language Models for Pathogenicity Identification

While traditional machine learning and deep learning models require supervised training with labeled data, large language foundation models facilitate zero-shot classification, classing sequences without labeled data. We compared the zero-shot pathogen classification performance of four different language models pre-trained in a self-supervised manner on large-scale DNA sequence data, each differing in dataset and training methodologies. For instance, DNABERT and Nucleotide Transformer primarily utilize k-mer tokenization, while DNABERT-2 (Zhou et al., 2023) employs byte-pair encoding. DNABERT-S (Zhou et al., 2024) boosts performance through Manifold Instance Mixup and Curriculum Contrastive Learning. Our findings indicate that DNABERT-S excels in zero-shot tasks with a 90% F1-score, while NT-v2-50m achieves a commendable 70%, likely due to its extensive multi-species training. DNABERT-2 and HyenaDNA lag with performances less than 70% shown in Figure 2B. Additionally, we assessed the few-shot capabilities of these models by training with incrementally larger samples, observing optimal performance at 25 samples, showcasing the strong few-shot learning potential of large language models. Despite NT-v2-50m’s lower initial zero-shot performance, it rapidly improves with minimal data input, surpassing others with only two samples from pathogen and non-pathogen categories, making it our model of choice due to its broad training on multiple species and lower computational demands compared to DNABERT-S.

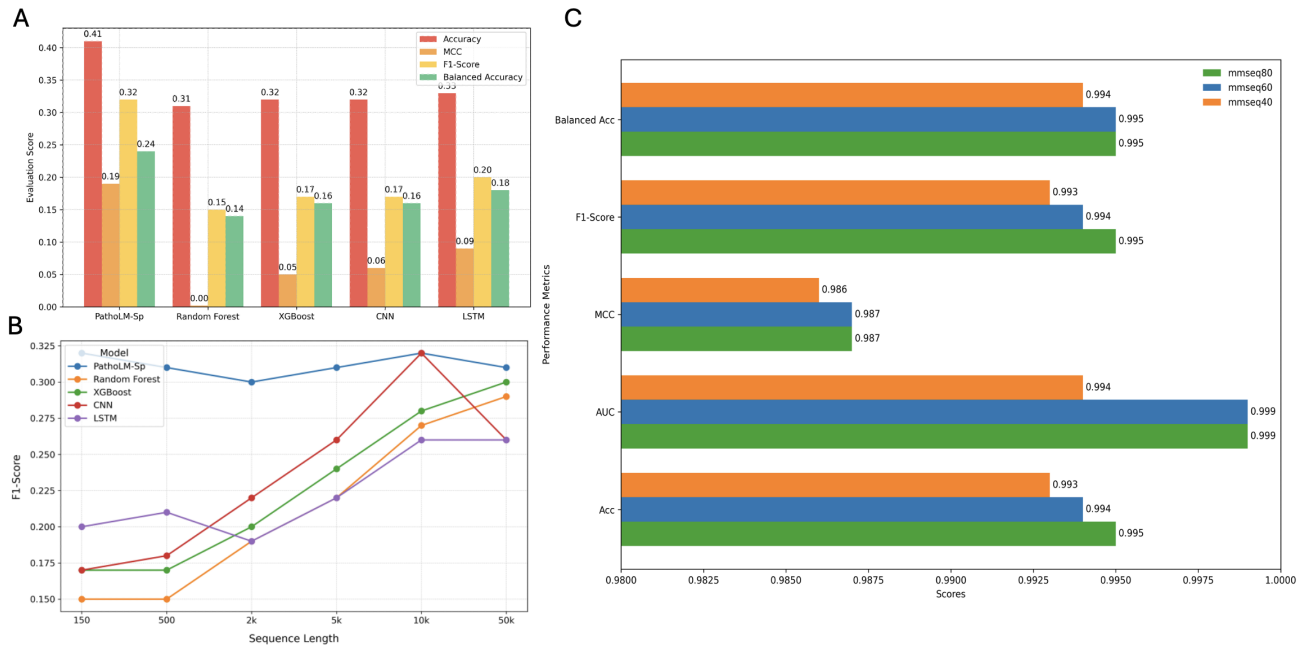


Figure 3. A: Performance comparison between PathoLM-Sp and machine learning Random Forest, XGBoost and deep learning methods CNN, LSTM on the ESKAPEE species classification task. PathoLM-Sp outperforms all other methods across multiple evaluation metrics. B: Performance comparison of PathoLM-Sp and other methods for varied sequence length from 150 bp to 50k bp. PathoLM-Sp consistently performs well for all sequence lengths. C: Performance comparison of PathoLM for different train and test set data created using mmseq2 sequence clustering. The similarity-threshold and coverage 80, 60, 40 are shown as mmseq80, mmseq60, mmseq40.

3.3. Species classification with PathoLM-Sp

We expanded our PathoLM model to classify seven ESKAPEE species, naming this variant PathoLM-Sp. Since there were no pre-trained models for ESKAPEE species, we compared PathoLM-Sp with machine learning methods like Random Forest and XGBoost, and deep learning methods including a 1-layer CNN and a 1-layer LSTM. Sequences were transformed into one-hot encodings and utilized as input features for the model. The CNN had a one-dimensional convolutional layer

with 64 filters, ReLU activation, and max pooling, followed by two dense layers. The LSTM model used a single layer with 50 units. Both were optimized with Adam and trained on the same dataset for consistent evaluation. PathoLM-Sp demonstrated superior accuracy, F1-score, MCC, and balanced accuracy, as shown in Figure 3A, confirming its effectiveness despite the inherent challenge of species classification with these data.

3.4. Performance comparison for varied sequence length

We evaluated PathoLM-Sp against other models using datasets of various sequence lengths (150 bp to 50k bp) to determine how sequence length impacts performance. PathoLM-Sp consistently outperformed all models at each sequence length, maintaining superior accuracy and robustness. While models like Random Forest, XGBoost, CNN, and LSTM improved with longer sequences, PathoLM-Sp excelled even with shorter sequences, leveraging its pre-trained capabilities effectively shown in Figure 3B. This performance consistency across various sequence lengths highlights PathoLM-Sp's reliability for genomic analysis, particularly when dealing with diverse data conditions.

3.5. Performance Comparison on Different Train-Test Sets Created Using MMseq2

To ensure a robust and generalized evaluation, we used MMseq2 to create different train-test splits based on sequence clustering. MMseq2 was employed to generate clusters using similarity threshold and coverage criteria. The clusters were divided into training and testing sets at an 8:2 ratio, ensuring samples from different clusters for each set. We created three distinct datasets with similarity thresholds and coverage values of 80, 60, and 40. The performance of our PathoLM-Sp model was compared against other models using these datasets shown in Figure 3C. The results demonstrate that model performance improves as the similarity threshold increases. Specifically, with an 80% similarity threshold, the model achieves higher evaluation metrics compared to the 60% and 40% thresholds. This comparison highlights the impact of sequence similarity on model performance. A reduced similarity threshold enhances learning robustness and generalization, as evidenced by the variation in performance metrics.

4. Discussion

In this study, we introduce PathoLM, a novel language model tailored for pathogen identification, drawing significantly from a pre-trained foundational model and a meticulously curated dataset of high-risk pathogens. This foundation imbues PathoLM with a deep genomic understanding, enabling it to discern subtle differences between pathogenic and nonpathogenic sequences across viruses and ESKAPEE bacteria, thereby enhancing its performance. Our database, concentrated on about 30 bacterial and viral species with a focus on ESKAPEE pathogens, supports extensive pathogen prediction tasks and offers a valuable resource for further research. Additionally, our PathoLM-Sp variant, designed for ESKAPEE species classification, surpasses conventional machine and deep learning methods in accuracy, highlighting the effectiveness and flexibility of our model.

Despite these advances, our model has limitations, notably the Nucleotide Transformer v2 50M's maximum context length of 12k bp. Sequences exceeding this length must be fragmented for analysis, potentially missing broader sequence interactions. Moreover, training high-parameter language models demands extensive computational resources. While the 50 million parameter model is manageable, the more accurate 250 million parameter version is even more resource-intensive. Additionally, PathoLM-Sp's species classification accuracy around 41% suggests room for improvement, particularly for species beyond *E. coli*. Future work will explore integrating diverse features to enhance model performance. Despite these challenges, our work significantly advances pathogen identification using pre-trained language models and provides a valuable dataset for ESKAPEE species, underscoring the potential for further enhancements and the current effectiveness of our approach.

5. Code and Data Availability

We downloaded the ESKAPEE dataset from PATRIC website (<https://www.bv-brc.org/>) and downloaded viral and non-pathogenic ESKAPEE data from the NCBI website (<https://www.ncbi.nlm.nih.gov/>). The PathoLM source code and data can be found in this github repository (<https://github.com/Sajib-006/Patho-LM>).

References

- Barash, E., Sal-Man, N., Sabato, S., and Ziv-Ukelson, M. Bacpacs—bacterial pathogenicity classification via sparse-svm. *Bioinformatics*, 35(12):2001–2008, 2019.
- Bartoszewicz, J. M., Seidel, A., Rentzsch, R., and Renard, B. Y. Deepac: predicting pathogenic potential of novel dna with reverse-complement neural networks. *Bioinformatics*, 36(1):81–89, 2020.
- Bi, Z., Dip, S. A., Hajjaligol, D., Kommu, S., Liu, H., Lu, M., and Wang, X. Ai for biomedicine in the era of large language models. *arXiv preprint arXiv:2403.15673*, 2024.
- Centers for Disease Control and Prevention. Influenza virus. <https://search.cdc.gov/search/?query=influenza%20virus&dpag=1>, 2024. Accessed: 2024-05-07.
- Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J., Carranza, N. L., Grzywaczewski, A. H., Oteri, F., Dallago, C., Trop, E., de Almeida, B. P., Sirelkhatim, H., et al. The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *bioRxiv*, pp. 2023–01, 2023.
- Deneke, C., Rentzsch, R., and Renard, B. Y. Paprbag: A machine learning approach for the detection of novel pathogens from ngs data. *Scientific reports*, 7(1):39194, 2017.
- Egberink, H. and Horzinek, M. Animal immunodeficiency viruses. *Veterinary microbiology*, 33(1-4):311–331, 1992.
- Gillespie, J. J., Wattam, A. R., Cammer, S. A., Gabbard, J. L., Shukla, M. P., Dalay, O., Driscoll, T., Hix, D., Mane, S. P., Mao, C., et al. Patric: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. *Infection and immunity*, 79(11):4286–4298, 2011.
- Hassan, E. and Baldrige, M. T. Norovirus encounters in the gut: multifaceted interactions and disease outcomes. *Mucosal immunology*, 12(6):1259–1267, 2019.
- Jiang, G., Zhang, J., Zhang, Y., Yang, X., Li, T., Wang, N., Chen, X., Zhao, F.-J., Wei, Z., Xu, Y., et al. Dcipatho: deep cross-fusion networks for genome scale identification of pathogens. *Briefings in Bioinformatics*, 24(4):bbad194, 2023.
- Li, H.-L., Pang, Y.-H., and Liu, B. Bioseq-blm: a platform for analyzing dna, rna and protein sequences based on biological language models. *Nucleic acids research*, 49(22):e129–e129, 2021.
- Liebhoff, A.-M., Menden, K., Laschtowitz, A., Franke, A., Schramm, C., and Bonn, S. Pathogen detection in rna-seq data with pathonoia. *BMC bioinformatics*, 24(1):53, 2023.
- Lorenzi, C., Barriere, S., Villemin, J.-P., Dejardin Bretones, L., Mancheron, A., and Ritchie, W. imoka: k-mer based software to analyze large collections of sequencing data. *Genome biology*, 21:1–19, 2020.
- Luu, R. K. and Buehler, M. J. Bioinspiredllm: Conversational large language model for the mechanics of biological and bio-inspired materials. *Advanced Science*, 11(10):2306724, 2024.
- Ruekit, S., Srijan, A., Serichantalergs, O., Margulieux, K. R., Mc Gann, P., Mills, E. G., Stribling, W. C., Pimsawat, T., Kormanee, R., Nakornchai, S., et al. Molecular characterization of multidrug-resistant escapee pathogens from clinical samples in chonburi, thailand (2017–2018). *BMC infectious diseases*, 22(1):695, 2022.
- Saib, I., Aleisa, S., Ardah, H., Mahmoud, E., Alharbi, A. O., Alsaedy, A., Aljohani, S., Alshehri, A., Alharbi, N. K., and Bosaeed, M. Non-sars non-mers human coronaviruses: clinical characteristics and outcome. *Pathogens*, 10(12):1549, 2021.
- Steinegger, M. and Söding, J. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, 35(11):1026–1028, 2017.
- Watts, G. S., Thornton Jr, J. E., Youens-Clark, K., Ponsero, A. J., Slepian, M. J., Menashi, E., Hu, C., Deng, W., Armstrong, D. G., Reed, S., et al. Identification and quantitation of clinically relevant microbes in patient samples: Comparison of three k-mer based classifiers for speed, accuracy, and sensitivity. *PLoS Computational Biology*, 15(11):e1006863, 2019.

Zhou, Z., Ji, Y., Li, W., Dutta, P., Davuluri, R., and Liu, H. Dnabert-2: Efficient foundation model and benchmark for multi-species genome. *arXiv preprint arXiv:2306.15006*, 2023.

Zhou, Z., Wu, W., Ho, H., Wang, J., Shi, L., Davuluri, R. V., Wang, Z., and Liu, H. Dnabert-s: Learning species-aware dna embedding with genome foundation models. *ArXiv*, 2024.