Briefings in Bioinformatics, 2025, **26(3)**, bbaf271
https://doi.org/10.1093/bib/bbaf271
Problem Solving Protocol

VirulentHunter: deep learning-based virulence factor predictor illuminates pathogenicity in diverse microbial contexts

Chen Chen[1,2], Yong Xu[1], Jian Ouyang[1,3], Xiangyi Xiong[1], Paweł P. Łabaj[4], Agnieszka Chmielarczyk[5], Anna Różańska[5], Hao Zhang[1], Keyang Liu[1], Tieliu Shi[1,6,7,*], Jun Wu[1,7,*]

[1]Center for Bioinformatics and Computational Biology, and The Institute of Biomedical Sciences, School of Life Sciences, East China Normal University, Dongchuan Road 500, Shanghai 200241, China
[2]School of Mathematics and Computer Science, Ningxia Normal University, College Road, Guyuan City, Ningxia 756099, China
[3]Henan International Joint Laboratory of Infection and Immunity, Henan Key Laboratory of Critical Care Medicine, Department of Emergency Medicine, The First Affiliated Hospital, Zhengzhou University, East Jianshe Road No. 1, Zhengzhou 450052, China
[4]Małopolska Centre of Biotechnology, Jagiellonian University, Gronostajowa 7A, 30-387 Kraków, Poland
[5]Faculty of Medicine, Department of Microbiology, Jagiellonian University, ul. Czysta 18, 31-121, Poland
[6]Key Laboratory of Advanced Theory and Application in Statistics and Data Science—Ministry of Education, School of Statistics, East China Normal University, Zhongshan North Road 3663, Shanghai 200062, China
[7]Shanghai Institute of Wildlife Epidemics, East China Normal University, Dongchuan Road 500, Shanghai 200062, China

*Corresponding authors. Tieliu Shi, Center for Bioinformatics and Computational Biology, and the Institute of Biomedical Sciences, School of Life Sciences, East China Normal University, Shanghai 200241, China. E-mail: tieliushi@yahoo.com; Jun Wu, Center for Bioinformatics and Computational Biology, and the Institute of Biomedical Sciences, School of Life Sciences, East China Normal University, Shanghai 200241, China. E-mail: jwu@bio.ecnu.edu.cn

Abstract

Virulence factors (VFs) are critical determinants of bacterial pathogenicity, but current homology-based identification methods often miss novel or divergent VFs, and many machine learning approaches neglect functional classification. Here, we present VirulentHunter, a novel deep learning framework that enable simultaneous VF identification and classification directly from protein sequences by leveraging the crucial step of fine-tuning pretrained protein language model. We curate a comprehensive VF database by integrating diverse public resources and expanding VF category annotations. Our benchmarking results demonstrate that VirulentHunter outperforms existing methods, particularly in identifying VFs lacking detectable homologs. Additionally, strain-level analysis using VirulentHunter highlights distinct pathogenicity profiles between *Mycobacterium tuberculosis* and *Mycobacterium avium*, revealing enrichment in VFs related to adherence, effector delivery systems, and immune modulation in *M. tuberculosis*, compared to biofilm formation and motility in *M. avium*. Furthermore, metagenomic profiling of gut microbiota from inflammatory bowel disease patient reveals a depletion of VFs associated with immune homeostasis. These results underscore the versatility of VirulentHunter as a powerful tool for VF analysis across diverse applications. To facilitate broader accessibility, we provide a freely accessible web service for VF prediction (http://www.unimd.org/VirulentHunter), accommodating protein sequences, genomes, and metagenomic data.

Keywords: virulence factors; deep learning; microbial pathogenicity; homology independent; multi-purpose tool

Introduction

Virulence factors (VFs) are molecules expressed or secreted by microorganisms that enable host colonization, immune evasion, and nutrient acquisition, facilitating infections [1]. Understanding bacterial VFs is crucial for elucidating pathogenesis and developing therapeutic strategies [2–5]. High-throughput sequencing has revolutionized microbiology, enabling insights into microbial communities and their functional potential [6], including VFs. However, a significant proportion of the proteins identified, including potential VFs, remain unannotated, posing substantial challenges. Thus, developing robust methods to analyze and interpret this complex data is essential for advancing microbial research and therapeutic innovation.

To address these challenges, various computational methods have been developed for VF identification. These range from sequence homology searches against known VF databases (e.g. VFDB [7], Victors [8], BV-BRC [9]), to advanced machine learning algorithms leveraging protein signatures. Homology-based methods [10, 11] rely on sequence similarity but are limited in identifying novel VFs without known homologs. To overcome this, features such as amino acid composition, physicochemical properties, PSSM, and one-hot encoding have been integrated with machine-learning methods to improve VF identification. For instance, Gupta *et al.* introduced MP4, incorporating dipeptide frequency and pepstats features to classify pathogenic proteins into three groups: Non-pathogenic Proteins, Antibiotic Resistance and Toxins, and Secretory System

Received: January 4, 2025. **Revised:** May 2, 2025. **Accepted:** May 21, 2025

and Capsular Proteins [12]. Additionally, Singh *et al.* developed VF-Pred, an ensemble learning model predicting VFs using sequence alignment and 982 engineered features [13]. Similarly, Ji *et al.* introduced HyperVR, a hybrid deep ensemble approach combining machine and deep learning with sequence-based and evolutionary features [14]. Moreover, Xie *et al.* proposed a stacking-based hybrid framework integrating four machine leaning and three deep learning algorithms for VF prediction from protein sequences [15]. The 3D structural information is crucial for understanding protein functions and has been applied in tools like GTAE-VF [16], its utility is limited by the scarcity of experimentally resolved structures. However, the availability of 3D structures for VFs is limited, and computational prediction of these structures with high accuracy (e.g. AlphaFold2 [17], Rosetta [18], and ESMFold [19]) is resource-intensive, posing challenges for metagenomic studies that generate vast amounts of novel proteins.

Although these methods reliably identify VFs, they struggle to classify VFs into specific functional categories. Categorizing VFs is crucial for understanding how pathogens interact with hosts. Different VF types, such as toxins, adhesins, and secretion systems, have unique roles in infection. For example, toxins harm host tissues, adhesins help bacteria attach to hosts, and secretion systems deliver effector molecules. This gap highlights the need for advanced methods to classify VFs into specific functional groups, enhancing our understanding of microbial pathogenicity and resistance mechanisms.

Recent advances in protein language models, such as ESM2 [19], ProteinBERT [20] and ProtT5 [21] have been pre-trained and our previous studies has demonstrated their ability to extract complex sequence–structure function relationships [22, 23]. Here, we developed VirulentHunter, a novel framework for VF identification and classification. By fine-tuning the pre-trained ESM2 model, VirulentHunter effectively extracts meaningful features from protein sequences, enabling precise prediction of VF categories. We validated its utility by analyzing pathogenic differences between two *Mycobacterium* species with distinct virulence profiles and contrasting VF variations in the gut microbiomes of healthy individuals and inflammatory bowel disease (IBD) patients. VirulentHunter outperforms state-of-the-art methods, demonstrating the critical role of pre-trained models in advancing VF analysis and providing a powerful tool for studying bacterial pathogenesis and disease stratification.

Materials and methods

Data collection and curation

In this study, we focused exclusively on VFs found in bacteria. We gathered all bacterial VF protein sequences from three public databases: VFDB 2022 [7], Victors [8] and BV-BRC [9]. Subsequently, these sequences were clustered with CD-HIT [24] v4.8.1, and the duplicates were removed with 100% sequence identity and 80% coverage, yielding 30,483 non-redundant VFs.

Since many collected VFs lacked category information, we implemented a rigorous label propagation strategy to annotate them using the 14 primary categories defined by VFDB. Initially, we performed sequence-based clustering using DIAMOND [25] with an 80% sequence identity and an 80% coverage threshold. Within each cluster, all member VFs were assigned combined labels from the union of their existing annotations. To further enhance category assignment, we employed TM-Vec [26], a deep learning tool for structural similarity detection. VFs were clustered using TM-Vec with a threshold of 0.9. The same label propagation strategy was then applied, allowing each VF to be assigned combined labels derived from the union of their existing annotations.

Fine-tuning of ESM2 for virulence factor prediction and classification

VirulentHunter was developed by fine-tuning the state-of-the-art ESM2 model (esm2_t30_150M_UR50D) to predict VFs and their categories directly from protein sequences (Fig. 1). This fine-tuning process utilized Low-Rank Adaptation [27] to adjust the query, key, and value matrices of its self-attention modules. This approach preserved pre-trained representations while enhancing VF-specific feature extraction. Subsequently, a classification head comprising two fully connected layers and a dropout layer was added to perform binary VF prediction and multi-class categorization. Considering class imbalance in VF categories, which could bias predictions and confound performance evaluation, we employed Focal Loss [28] for the VF classification task. Focal Loss is designed to upweight hard-to-classify examples and those belonging to minority classes, promoting more accurate and robust categorization, especially for infrequent VF categories. To determine the optimal hyperparameters, we utilized the W&B Sweeps [29], a reliable and widely recognized method for exploring the hyperparameter space. The detailed search spaces are provided in Supplementary Tables S1 and S2. Besides that, we also fixed the random seeds for all experiments to ensure reproducibility.

Performance metrics and evaluation

To comprehensively assess VirulentHunter's performance, we used five standard metrics: accuracy, precision, recall, F1-score, and Matthews correlation coefficient (MCC). While accuracy, precision, recall, and F1-score provide valuable insights into prediction performance, MCC offers a more robust evaluation, particularly when dealing with imbalanced datasets. The equations for these metrics are as follows:

- $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$
- $Precision = \frac{TP}{FT+FN}$
- $Recall = \frac{TN}{FP+TN}$
- $F1-Score = 2 \times \frac{TP}{2TP+FP+FN}$
- $MCC = \frac{(TP \times TN)-(FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$

Where:

- $TP = True\ Positives$
- $TN = True\ Negatives$
- $FP = False\ Positives$
- $FN = False\ Negatives$

Given that VF category classification is a multi-label task, we employed micro-averaged metrics: precision, recall, F1-score, and MCC. In micro-averaging, each instance-label pair is considered a separate prediction.

Webserver construction

We utilized the Apache HTTP server as a web server, developed by PHP (Version: 7.0.12, https://www.php.net/) programming. Data interaction was implemented by HTML5, JavaScript, jQuery. All data in VirulentHunter are stored and managed in MySQL database (Version: 5.7.17, https://www.mysql.com/). Data analyses were mainly carried out by the R (Version 4.4.0, https://www.r-project.org/) or python (Version 3.9.10, https://www.python.org/) script.

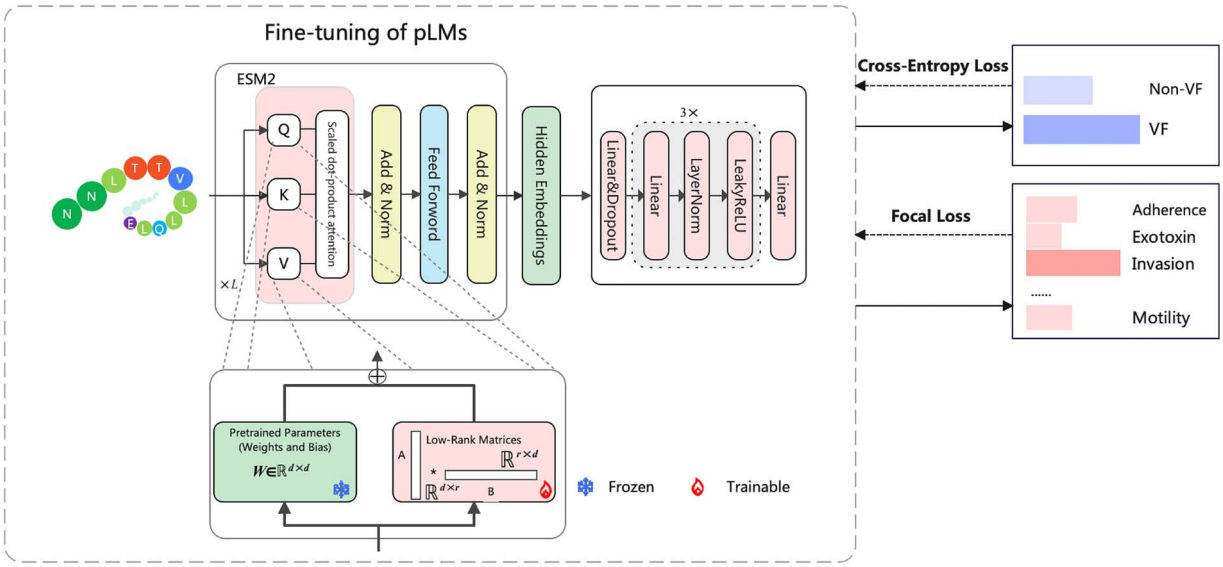

Figure 1. Illustration of the proposed VirulentHunter framework for virulence factor identification and classification.

Results

Unifying public virulence factors databases for enhancing VirulentHunter training

We developed a reproducible data curation schema for benchmark VFs and its category datasets, encompassing manual review, redundancy elimination, and category label propagate (Fig. 2a and b, see Materials and methods). The finally curated VF dataset comprises 30 483 VFs across 14 distinct categories. The majority of the VFs are associated with only one category (90.1%), while 7.62% are related to two categories. Additionally, five out of the 14 VFs categories constitute approximately 55% of all VFs, highlighting a significant imbalance among the known VF categories. The three most prevalent categories—effector delivery systems (25.67% of VFs), immune modulation (15.12% of VFs), and adherence (14.1% of VFs)—collectively account for nearly half of the VFs in the dataset (Fig. 2c).

To construct a negative dataset for VF identification model training, we extracted non-virulent protein sequences from the Swiss-Prot [29] using the 23 keywords used in previous study [14] (e.g. 'NOT Virulence', 'NOT Toxin' and 'Not Capsule', Supplementary Table S3). To match the length distribution of the collected VFs (Fig. 2d), we retained only non-virulent protein sequences with lengths between 50 and 2000 amino acids. After removing duplicates, 30 215 non-virulent proteins were randomly selected to match the number of VFs, creating a negative dataset for VF identification. Although standard approaches for negative dataset construction are lacking, we recognized its significant impact on model performance. Therefore, we additionally explored a Gene Ontology (GO) annotation-based method [29] (NExIGO) for negative sample selection. However, compared to the keyword filtering method, the NexIGO-based method resulted in lower performance (Supplementary Fig. S1).

Comparative analysis and sequence similarity evaluation of the VirulentHunter method for virulence factors identification

We initially assessed the performance of our proposed method, VirulentHunter, by comparing it with current state-of-the-art models: MP4 [12], VirulentPred 2.0 [30], and DeepVF [15]. Prior to training, we randomly selected 3048 VFs and 3021 non-VFs to form an independent validation dataset. The remaining proteins were allocated to training and validation set. It should be noted that DeepVF and MP4 are available exclusively as online services and do not provide open-source code, thus preventing us from retraining these models with our constructed dataset, potentially influencing their performance in the comparison. The 10-fold cross-validation results demonstrated that VirulentHunter achieved significantly improved robustness and consistently higher mean performance metrics across all evaluation criteria compared to baseline methods, highlighting its stability and generalizability (Supplementary Fig. S2). Validation on an independent validation set demonstrated that VirulentHunter outperformed competing methods across most metrics. However, its recall score of 0.809 was slightly lower than those achieved by DeepVF and MP4. (Fig. 3a). The ROC analysis demonstrated that VirulentHunter achieved improvements of 47.76% and 68.37% over MP4 and DeepVF, respectively, in terms of the AUC value. Since VirulentPred 2.0 only provides binary outcomes, it was excluded from the ROC comparison. Additionally, the comparison between the VirulentHunter with and without the fine-tuning step indicated that incorporating a fine-tuning strategy can significantly enhanced performance, as evidenced by the generation of more distinctive embedding outcomes (Fig. 3b, Supplementary Fig. S3).

We subsequently assessed the capacity of VirulentHunter to identify VFs with a range of sequence identities to the training set. Hence, we initially categorized the independent validation set into seven groups based on their sequence identity to the training set. The results revealed that most of the methods provide accurate predictions for queried proteins with high sequence identities (>70%) to the training set (Fig. 3c, Supplementary Fig. S4). Notably, VirulentHunter exhibited more consistent performance than other methods, even when sequence identity was below 40%. This underscores its superiority in characterizing VFs within metagenomic studies, which are known for containing numerous unknown proteins.

Classification performance of VirulentHunter in diverse virulence factors categories

Having established the superior performance of VirulentHunter in identifying VFs, we further evaluated its performance to

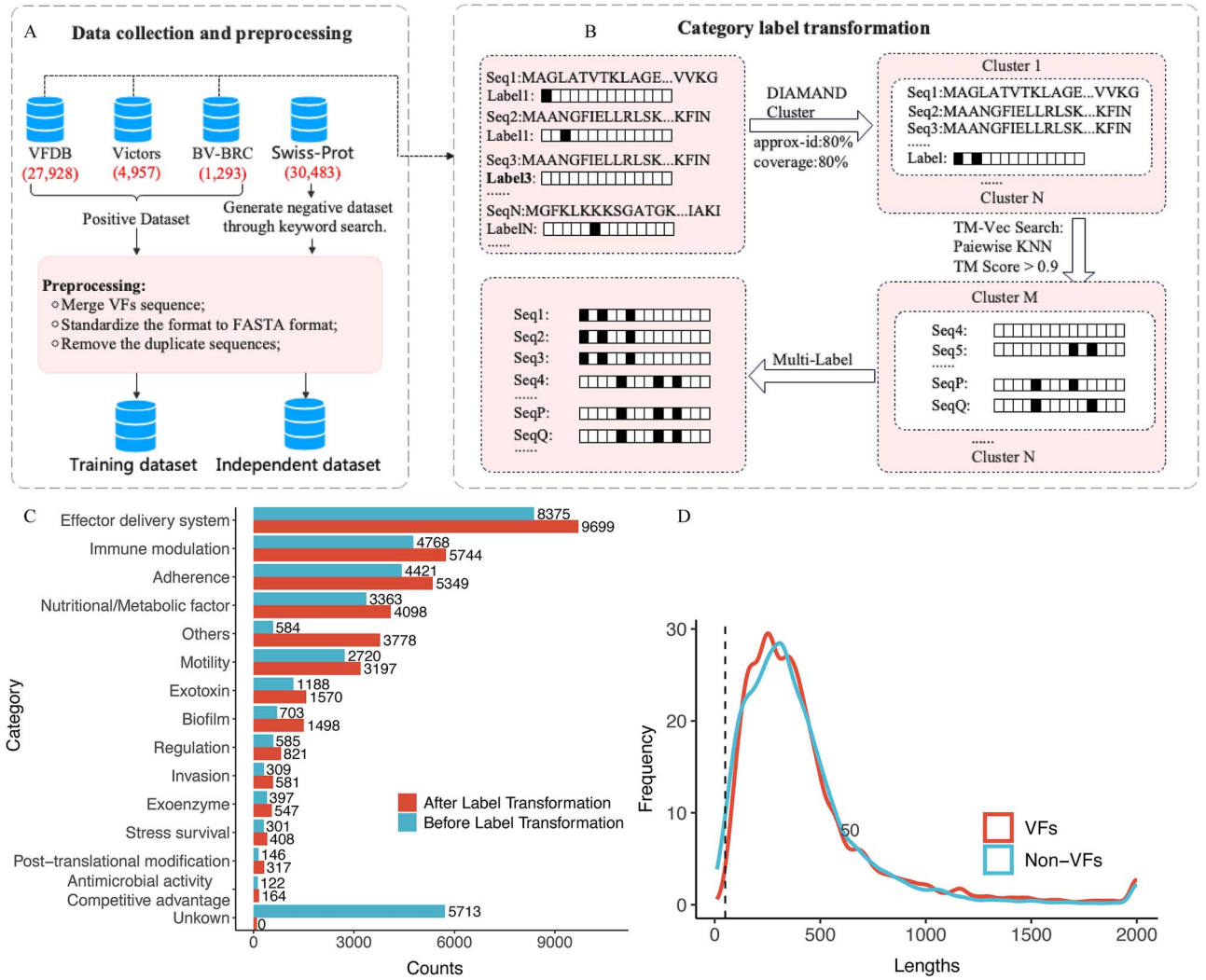

Figure 2. Workflow of VF data curation and the composition of VirulentHunter database: (a) preprocessing of VFs collected from three publicly available VF databases; (b) illustration of VF category propagation based on both the sequence and structure similarity; (c) expansion of the VF categories after label propagation; (d) Distribution of the length of VFs and non-VFs selected as negative set.

classifying VFs into different functional categories. Following the categorization schemes adopted in the VFDB database, we aimed to further classify the identified VFs into 14 basal categories.

Given the absence of direct competitors for this specific task, we evaluated the proposed VirulentHunter method against a baseline approach using BLAST and a naïve ESM2 + XGBoost method. The ESM2 + XGBoost method involved using ESM2 to embed the proteins and then employing XGBoost for the classification task, a strategy that has been shown to effectively handle the complexity of protein classification [23]. Consistent with our previous methodology, employed 10-fold cross-validation and stratified the validation set into five groups based on sequence similarity. The comparison results revealed that both VirulentHunter and the ESM2 + XGBoost methods offered higher and more stable performance compared to the baseline BLAST approach across all sequence identity groups (Fig. 4a and Supplementary Fig. S5a). Notably, VirulentHunter demonstrated superior performance to the ESM2 + XGBoost method, particularly in groups with low sequence identity, underscoring its robustness in classifying VFs even when sequence similarity to known VFs is limited. This capability also suggests that VirulentHunter is exceptionally well-suited for diverse metagenomic datasets, where sequence diversity is prevalent.

We further conducted a comprehensive evaluation of the performance of VirulentHunter across various VF categories. The results demonstrated that our method consistently achieved high and robust performance across all categories, with mean Accuracy values ranging from 0.973 to 0.996, mean MCC values from 0.753 to 0.934, and mean F1-scores from 0.704 to 0.946 (Fig. 4b). However, when examining query proteins with less than 40% sequence identity, we observed a decline in performance for categories with a limited number of training proteins, particularly when the number was below 2000 (Fig. 4c, Supplementary Fig. S5b and c).

VirulentHunter reveals the distinct patterns in virulence factors between *Mycobacterium avium* and *mycobacterium tuberculosis*

To further validate the reliability of VirulentHunter in real-world microbiological studies, we conducted a comparative analysis focusing on *Mycobacterium tuberculosis* (*M. tuberculosis*) and *Mycobacterium avium* (*M. avium*). These two species, while

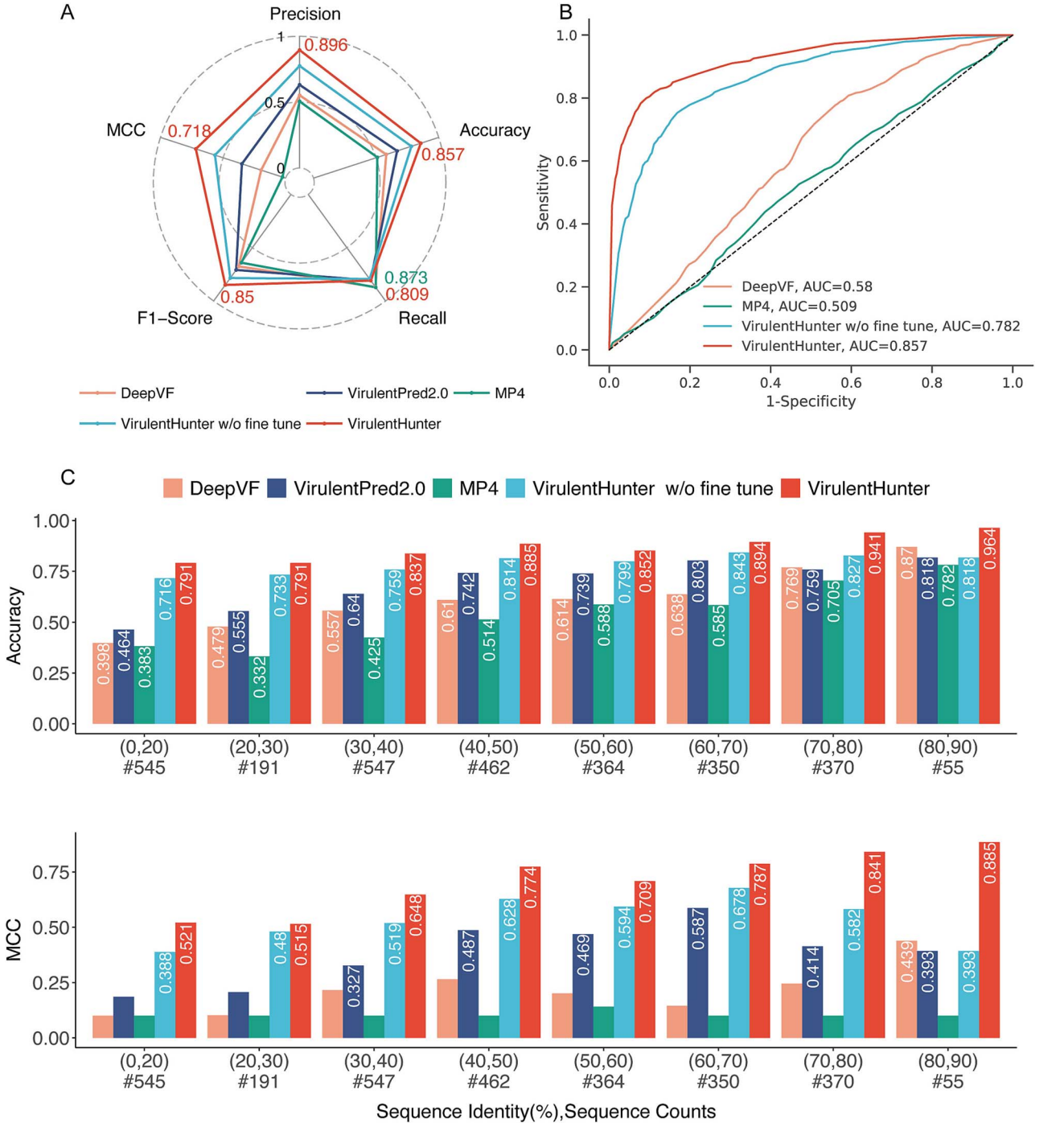

Figure 3. Performance comparison of VirulentHunter against state-of-the-art (SOTA) methods for identifying VFs on the independent test set. (a) Radar plot comparing VirulentHunter with SOTA methods and results from ablation experiments. (b) ROC curves of VirulentHunter, VirulentHunter w/o fine tune, DeepVF, and MP4 on the independent test set. (c) Comparative performance of VirulentHunter and SOTA methods across varying sequence identities relative to reference proteins. The bar chart illustrates performance differences between VirulentHunter and other SOTA methods, measured by accuracy and MCC. The x-axis of the bar chart represents different similarity intervals and the number of protein sequences within each interval.

belonging to the same genus, *Mycobacterium*, display distinct pathogenic profiles in humans, with *M. tuberculosis* mainly infects individuals with normal immune function and *M. avium* typically associated with opportunistic infections, particularly in immunocompromised individuals.

We retrieved the complete genomic sequences of 93 *M. avium* and 88 *M. tuberculosis* strains, along with their corresponding protein sequences, from the NCBI database (release date: 20 July 2024, Supplementary Table S4). Using VirulentHunter method, we identified and categorized 212,432 VFs across these strains. Each *M. avium* strain contains 972–1425 VFs, while each *M. tuberculosis* harbored 1122–1333 VFs. The predominant VF categories in both species were Effector Delivery Systems, Nutritional/Metabolic Factors, and Immune Modulation (Fig. 5a). These categories underscore the multifaceted strategies utilized by these pathogens to establish infection, acquire essential

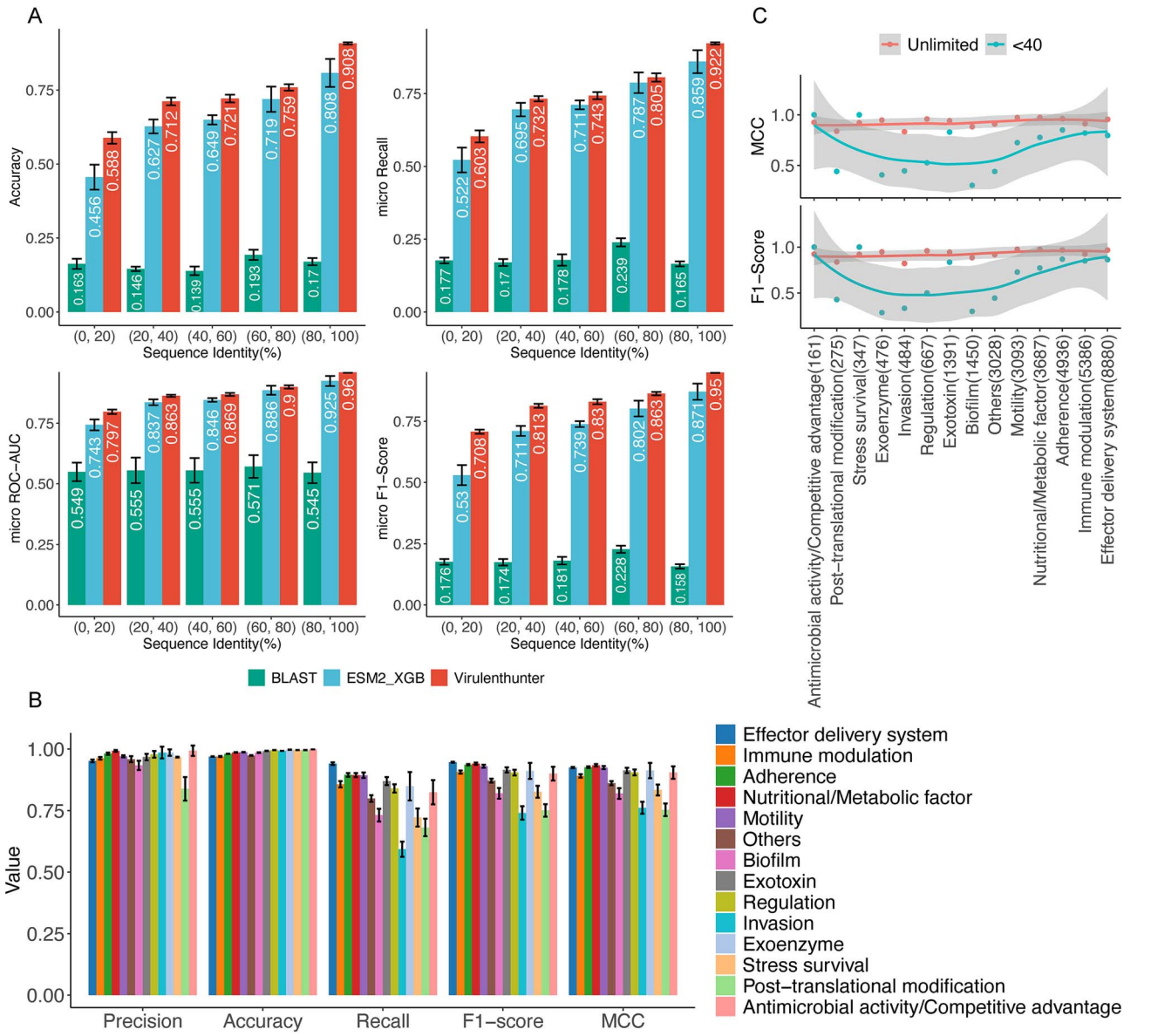

Figure 4. Performance of VirulentHunter in the VFs categorization task. (a) Comparison of VFs category identification performance via 10-fold cross-validation among VirulentHunter, BLAST and ESM2_XGBoost under varying sequence identity thresholds relative to the reference VFs dataset. (b) Performance of VirulentHunter on each VF category evaluated with five criteria. (c) Performance decline of VirulentHunter with decreasing training numbers for VFs lacking highly similar homologous sequences in the reference dataset. 'Unlimited': No sequence identity cutoff (independent test versus reference); '<40': Identity <40% (independent test validation versus reference).

nutrients, and manipulate host immune responses. The prevalence of VFs associated with Effector Delivery Systems, such as type VII secretion systems, highlights the importance of these systems in delivering VFs into host cells. Additionally, VFs involved in Nutritional/Metabolic Factors suggests these pathogens have evolved mechanisms to acquire essential nutrients from the host environment. Finally, the presence of VFs associated with Immune Modulation indicates that these pathogens can actively subvert host immune responses to promote their survival and replication. Intriguingly, we observed a distinct separation of *M. avium* and *M. tuberculosis* strains when visualizing their VF profiles (Fig. 5b). This suggests that the proportion of different VF categories within each strain constitutes a unique 'virulence fingerprint.' This 'fingerprint,' based on the relative representation of different VF classes, effectively differentiates these closely related *mycobacterial* species. This highlights the potential of VF proportion profiles as a valuable tool for strain-level classification and for gaining insights into the evolutionary trajectories of pathogenicity within the *Mycobacterium* genus.

We further compared VF prevalence between *M. tuberculosis* and *M. avium*, revealing a significantly higher prevalence of VF in *M. tuberculosis* compared to *M. avium* (Mann–Whitney U test, $P < .01$, Fig. 5c). However, the direction of this difference varied across VF categories (Fig. 5d). Specifically, VFs associated with Adherence, Effector Delivery System and Immune Modulation were more prevalent in *M. tuberculosis*, suggesting greater emphasis on host cell invasion and immune evasion. Conversely, a higher proportion of VFs related to Biofilm Formation and Motility were observed in *M. avium*, indicating potential adaptation to diverse ecological niche and reliance on extracellular survival strategies.

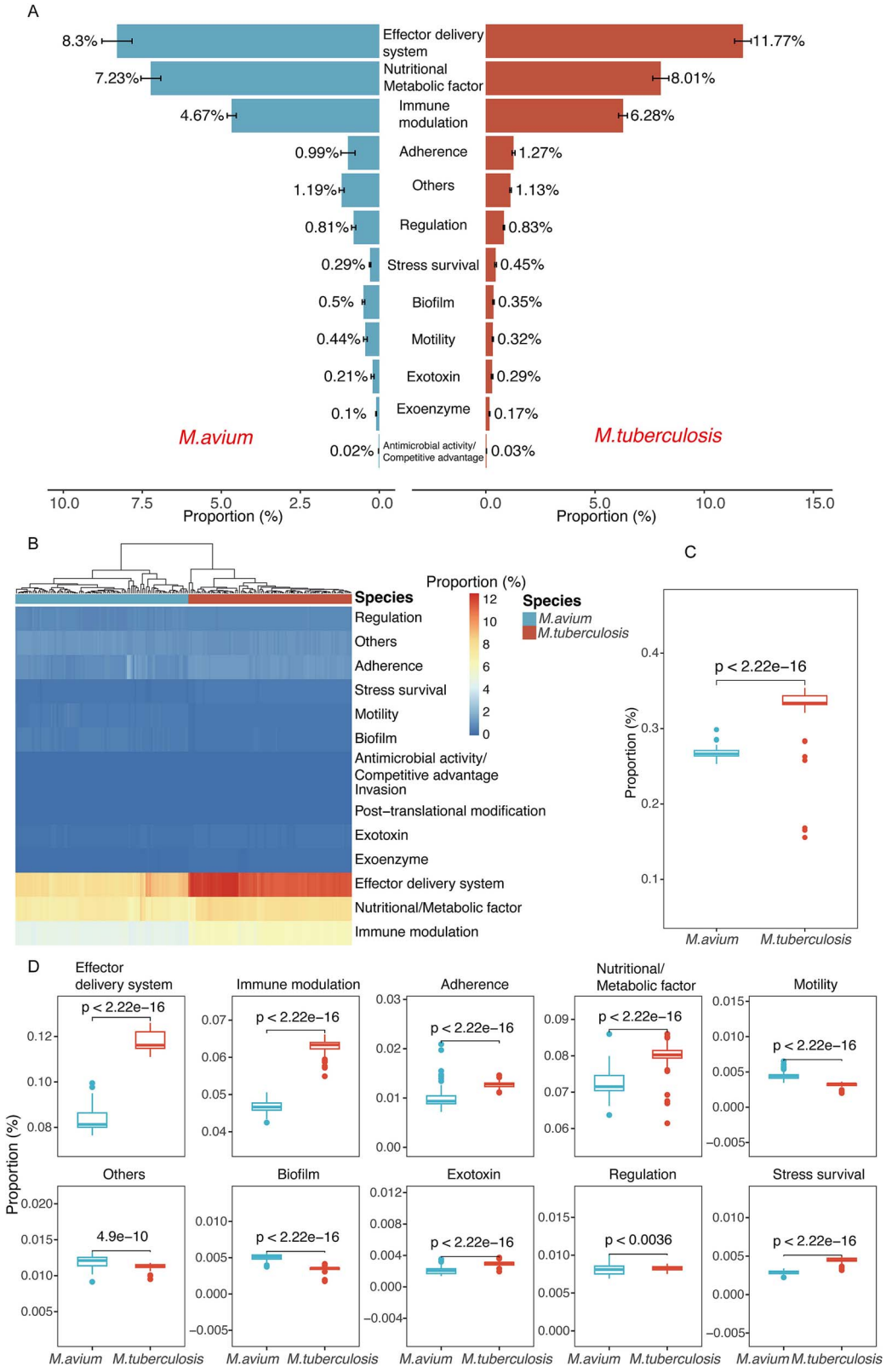

Figure 5. The application of VirulentHunter to large-scale assembled bacterial genomes reveals significant differences in VFs between *M. avium* and *M. tuberculosis*. (a) Histogram showing the genome size distribution of *M. avium* and *M. tuberculosis* groups. (b) Heatmap illustrating the distinct separation of VF proportion profiles between *M. avium* and *M. tuberculosis* groups. (c) Comparison of the overall proportion of VFs between *M. avium* and *M. tuberculosis*. (d) Comparison of the proportions of VFs in each category between *M. avium* and *M. tuberculosis*. Statistical significance was determined using the Mann–Whitney U test.

These findings align with the distinct pathogenic strategies employed by these two *Mycobacterium* species. *M. avium*, often infecting immunocompromised hosts, harbors a higher proportion of VFs related to Biofilm Formation and Motility. These VFs enhance its ability to form biofilms, adapt to various environments, and survive in hosts with weakened immune systems. In contrast, *M. tuberculosis*, a primary pathogen infecting immunocompetent individuals, possesses more prevalent VFs involved in Immune Modulation and Effector Delivery Systems. This adaptation enables it to evade and manipulate the host's immune response, establishing a chronic infection. Furthermore, the higher proportion of Adherence-related VFs in *M. tuberculosis* may facilitate its transmission and colonization within the host.

VirulentHunter identified differences in the relative abundance of virulence factor gene expression among patients with different inflammatory bowel diseases

Building on the established effectiveness of VirulentHunter in classifying VFs across diverse microbial strains, we further applied this method to the complex landscape of metagenomic research, with a particular focus on IBD—a spectrum of chronic intestinal inflammatory disorders encompassing Crohn's disease (CD) and ulcerative colitis (UC). Prior research have revealed that specific VFs within the gut microbiota substantially affect the pathogenesis and progression of IBD [3, 31–33]. Consequently, we utilized VirulentHunter to analyze metagenomic data from IBD patients, thereby identifying VFs associated with this disease.

To achieve the objectives outlined above, we utilized the dataset from Lloyd-Price *et al.* [34], randomly selecting 220 gut metagenomic samples from 106 participants, including 60 samples from non-IBD individuals, 80 from patients with CD, and 80 from patients with UC. We initiated our analysis with a standard metagenomic data processing pipeline, including quality control, genome assembly, and gene prediction, on these samples (see Materials and methods). To reduce the bias leading by sequence length, 209 samples with more than 10,000 predicted genes were retained for downstream analysis (Supplementary Table S5). The relative abundances of these genes were estimated using Salmon [35]. To optimize computational efficiency, we performed clustering using CD-HIT with a sequence identity threshold of at least 90% before applying VirulentHunter model to identify and classify VFs among the selected representative proteins. This approach identified 64 642 representative VFs, of which only 42.21% matched known reference VFs with sequence identity greater than 40%. This significant discovery underscores the presence of numerous previously unrecognized or divergent VFs, which are likely overlooked by commonly used sequence-alignment-based methodologies. The assigned VF categories were then propagated to their corresponding protein members.

Our results revealed significant differences in VF prevalence across different cohort groups. Specifically, gut microbiome samples from patients with CD and UC exhibited higher proportions of VFs compared to samples from non-IBD individuals (Fig. 6a). To evaluate this finding using a traditional approach, we also employed BLAST-based method (DIAMOND) to quantify VF proportions across the cohorts (positive hits were defined using established reference criteria of sequence identity >50% and query coverage >80%, as employed in previous studies [39, 40]). While this traditional method also identified higher proportions of VFs in the IBD cohorts, the estimated VF proportion was significantly lower than that determined by the proposed VirulentHunter method, demonstrating its high sensitivity (Supplementary Fig. S6a). These findings are consistent with previous studies, including those reviewed by [41], which reported an increased presence of VFs in IBD patients. Our results further reinforce the hypothesis that VFs may play a role in the pathogenesis of IBD and that patients with IBD tend to harbor a higher abundance of VFs in their gut microbiome compared to healthy individuals. To further investigate, we examined the VF dispersion across the three cohorts, focusing on their estimated relative abundances. The non-metric multidimensional scaling (NMDS) analysis did not reveal significant differences (Fig. 6b). However, specific VF analysis identified numerous VF with significantly altered abundances in patients with UC and CD compared to healthy individuals. Between the UC and CD cohorts, only 4 VFs showed differential expression. Notably, 79 VFs were significantly different in the UC cohort compared to the non-IBD cohort, with 47 having higher abundance and 32 having lower abundance Similarly, 97 VFs were differentially expressed in the CD cohort compared to the non-IBD cohort, with 29 having higher abundance and 68 having lower abundance (Fig. 6c, Supplementary Table S6). Of the VFs with significant differences in abundance, 29 were commonly dysregulated in abundance in both the UC and CD cohorts compared to the non-IBD cohort (Fig. 6d). Notably, 19 have higher abundance VFs, primarily linked to biofilm formation and immune modulation (Fig. 6e), suggesting a potential mechanism for enhanced bacterial persistence and chronic inflammation. Conversely, 10 VFs with lower abundance may be associated with immune modulation (Fig. 6f), indicating a potential disruption in the delicate balance between the gut microbiota and the host immune response in IBD. This categorization underscores the multifaceted roles these VFs may play in the pathogenesis and progression of IBD, providing insight into the potential mechanisms through which these factors contribute to the disease. In contrast, the traditional BLAST-based method failed to identify any significantly differentially abundant VFs between the non-IBD and CD cohorts, or between the CD and UC cohorts, highlighting a key advantage of VirulentHunter in revealing disease-relevant functional differences (Supplementary Fig. S6b).

Conclusion and discussion

VFs are central to infectious disease pathogenesis. Current homology-based VF identification methods suffer from high false-negative rates, particularly for novel or divergent VFs. While machine learning approaches improve detection, they often neglect functional classification, limiting mechanistic insights. Here, we present VirulentHunter, a novel deep-learning framework that simultaneously identifies and classifies VFs directly from protein sequences. We constructed a comprehensive and curated VF database by integrating public resources and expanding annotations via label propagation strategy. Benchmarking demonstrated VirulentHunter significantly outperforms state-of-the-art methods, particularly for VFs lacking detectable homology—a major hurdle in microbial pathogenesis research. We further validated VirulentHunter's practical utility through two distinct applications: strain-level analysis, revealing key differences in pathogenicity between two *Mycobacterium* species, and metagenomic profiling, providing novel insights into VF distribution within the gut microbiota of IBD patients.

The superior performance of VirulentHunter, particularly in identifying non-homologous VFs, highlights the power of ESM2 to capture complex sequence patterns beyond homology. This capability is crucial for uncovering novel virulence mechanisms in emerging pathogens and characterizing the 'virulome' of

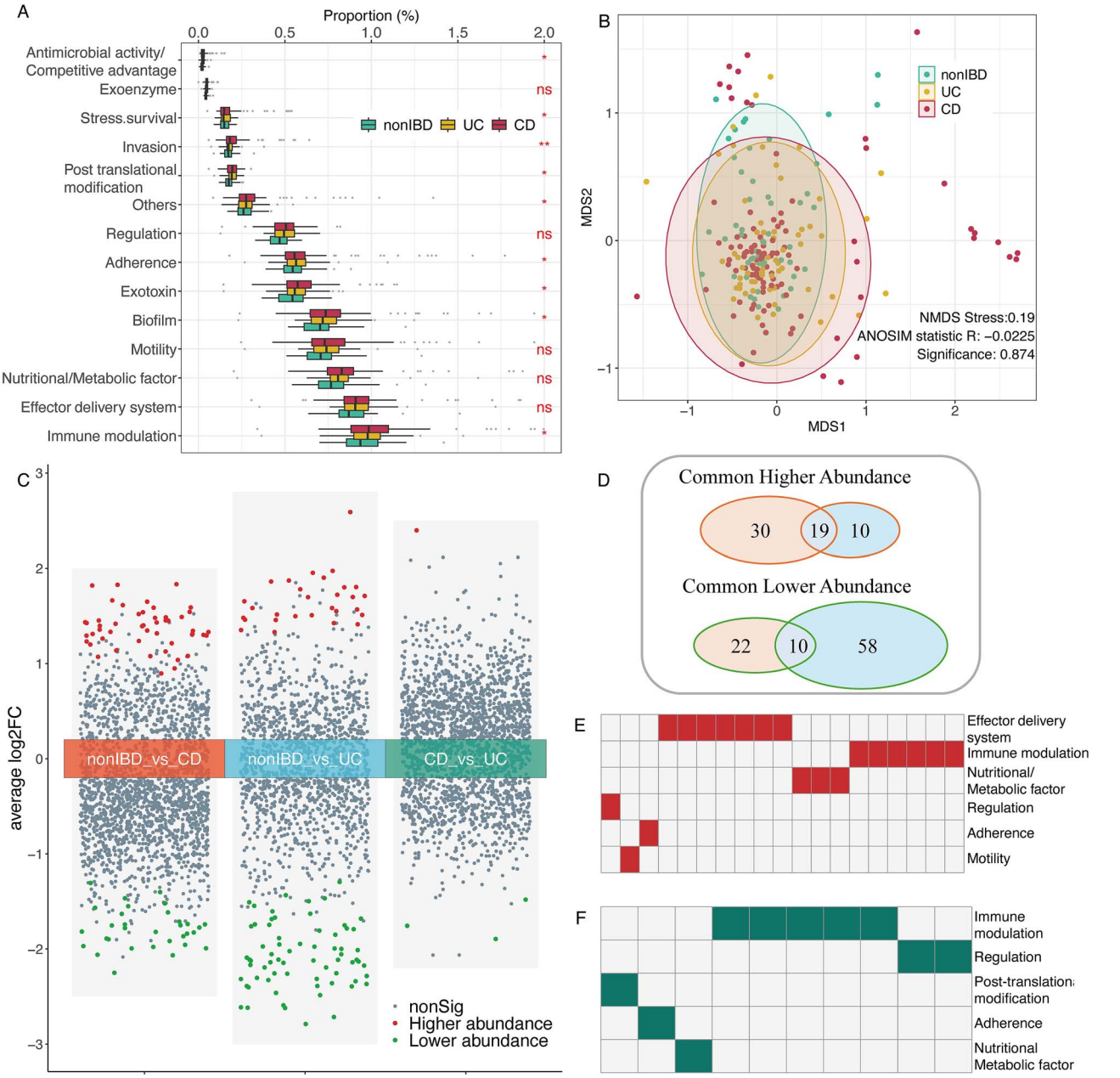

Figure 6. Dispersion of VFs across the non-IBD, UC, and CD cohorts. (a) Comparison of the proportion of VFs identified in the gut microbial samples from the three cohorts. (ANOVA test, FDR-adjusted. 'Ns' indicates no significant difference; ∗ and ∗∗ represent *P*-values of ≤.05 and ≤.01, respectively.). (b) NMDS plots of VFs grouped by individual status (non-IBD, UC and CD). The stress value of the NMDS ordination is 0.19. The 95% confidence ellipses are displayed around the sample groups. (c) Comparison of relative VF abundance between each pair of cohorts. (Wilcoxon test was used for comparisons. Significance was defined as FDR-adjusted *P*-value <.1, consistent with thresholds applied in previous studies [36–38]). (d) Overlap of VFs with higher and lower abundance in the UC and CD cohorts compared to the healthy individuals. (e and f) Categories of VFs with higher abundance and lower abundance in the UC and CD cohort compared to the healthy individuals, respectively.

understudied microbes. Our strain-level analysis of *Mycobacterium* species demonstrates the VirulentHunter's ability to dissect subtle differences. By accurately identifying and classifying VFs, we gain insights into disease mechanisms, such as variations in immune modulation, invasion and biofilm formation, which may explain differences in disease severity or clinical manifestations. This approach is particularly valuable for understanding drug resistance evolution and the emergence of more virulent strains. Furthermore, applying VirulentHunter to metagenomic data from IBD patients provides novel insights into gut microbiome pathogenesis. Profiling VF composition revealed dysbiosis in IBD patients, marked by a reduction in VFs linked to immune modulation. This suggests that the altered microbial community in IBD may impair immune regulation, contributing to chronic inflammation. This approach could also be used to investigate the impact of dietary interventions or therapeutic treatments on the gut virulome.

While VirulentHunter advances in VF identification and classification, several key challenges remain: (i) limited training data for certain VF categories. Expanding the VF database with experimentally validated entries, leveraging augmented functional annotation [42, 43], and integrating structural information [44] could enhance accuracy; (ii) potential data leakage from pLMs like ESM2, which we mitigated through sequence similarity subgroup

analysis (Figs 3c and 4a), particularly focusing on low-similarity proteins (<40%) to assess generalization capability, it serves as an approximation of the model's behavior on out-of-distribution proteins; and (iii) the selection of the negative class is a critical factor that can significantly influence the performance of machine learning classifiers. We argue that the choice of a negative sampling strategy should be carefully tailored to the specific application, as different approaches may be optimal depending on the research context and the nature of the available data. Additionally, exploring the interplay between different VFs and their regulatory mechanisms will contribute to more comprehensive understanding of microbial pathogenesis.

Key Points

- VirulentHunter is a novel deep learning framework that overcomes limitations of current methods by directly predicting VFs and classifying them into functional categories from protein sequences, without relying on sequence homology.
- VirulentHunter leverages a comprehensive, curated VF database integrating diverse public resources and employs a rigorous label propagation strategy to expand VF annotations, enabling more accurate and inclusive VF identification.
- Benchmarking shows VirulentHunter outperforms existing methods, particularly in identifying VFs with limited sequence homology, which is critical for studying emerging pathogens and less-studied microbial species.
- Two key applications demonstrate its utility: strain-level analysis reveals distinct VF patterns between *M. tuberculosis* and *M. avium*, while metagenomic profiling links VF depletion in IBD gut microbiomes to disease pathogenesis.
- VirulentHunter advances microbiome research by enabling comprehensive VF detection, functional categorization, high-resolution strain comparisons, and metagenome-scale insights into disease mechanisms.

Acknowledgements

JW. and T.S. conceived and designed the study. C.C., J.W. developed the model. C.C. J.W., Y.X., and J.O. performed data analysis and visualization. X.Y.X., J.O., H.Z. and K.L. built the web server. C. C and J.W. wrote the manuscript. P.P.Ł., A.C., A. R, and T.S. revised the manuscripts. All authors read and approved the final manuscripts.

Supplementary data

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

Conflict of interest

The authors declare no conflict of interest.

Funding

This work was supported by the National Key Research and Development Program of China (2023YFC2706503), the National Natural Science Foundation of China (32370720), the Open Research Fund of Key Laboratory of Advanced Theory and Application in Statistics and Data Science-MOE, ECNU, and the Open Research Fund of Key Laboratory of MEA, Ministry of Education, ECNU.

Data availability

The VirulentHunter model development source code is available at https://github.com/mini-ops996/VirulentHunter. The online platform http://www.unimd.org/VirulentHunter provides a free, user-friendly service for the efficient identification and classification of VFs from protein sequences, genomes, and even metagenomic data.

Ethics statement

No animals or humans were involved in this study.

References

1. Josenhans C, Suerbaum S. The role of motility as a virulence factor in bacteria. *Int J Med Microbiol* 2002;**291**:605–14. https://doi.org/10.1078/1438-4221-00173.
2. Kumar A, Prasoon P, Kumari C. *et al.* SARS-CoV-2 specific virulence factors in COVID-19. *J Med Virol* 2021;**93**:1343–1350. https://doi.org/10.2139/ssrn.3697116.
3. Fan L, Xia Y, Wang Y. *et al.* Gut microbiota bridges dietary nutrients and host immunity. *Sci China Life Sci* 2023;**66**:2466–514. https://doi.org/10.1007/s11427-023-2346-1.
4. Dickey SW, Cheung GYC, Otto M. Different drugs for bad bugs: antivirulence strategies in the age of antibiotic resistance. *Nat Rev Drug Discov* 2017;**16**:457–71. https://doi.org/10.1038/nrd.2017.23.
5. Que H, Chen L, Wei X. SARS-CoV-2 variants, immune escape, COVID-19 vaccine, and therapeutic strategies. *Sci China Life Sci* 2023;**66**:406–10. https://doi.org/10.1007/s11427-021-2164-6.
6. Pinto Y, Bhatt AS. Sequencing-based analysis of microbiomes. *Nat Rev Genet* 2024;**25**:829–45. https://doi.org/10.1038/s41576-024-00746-6.
7. Liu B, Zheng D, Zhou S. *et al.* VFDB 2022: a general classification scheme for bacterial virulence factors. *Nucleic Acids Res* 2022;**50**:D912–7. https://doi.org/10.1093/nar/gkab1107.
8. Sayers S, Li L, Ong E. *et al.* Victors: a web-based knowledge base of virulence factors in human and animal pathogens. *Nucleic Acids Res* 2019;**47**:D693–700. https://doi.org/10.1093/nar/gky999.
9. Olson RD, Assaf R, Brettin T. *et al.* Introducing the bacterial and viral bioinformatics resource Center (BV-BRC): a resource combining PATRIC IRD and ViPR. *Nucleic Acids Research* 2023;**51**:D678–89. https://doi.org/10.1093/nar/gkac1003.
10. Allen JP, Snitkin E, Pincus NB. *et al.* Forest and trees: exploring bacterial virulence with genome-wide association studies and machine learning. *Trends Microbiol* 2021;**29**:621–33. https://doi.org/10.1016/j.tim.2020.12.002.
11. Martínez-García PM, Ramos C, Rodríguez-Palenzuela P. T346Hunter: a novel web-based tool for the prediction of type III, type IV and type VI secretion systems in bacterial genomes. *PloS One* 2015;**10**:e0119317. https://doi.org/10.1371/journal.pone.0119317.
12. Gupta A, Malwe AS, Srivastava GN. *et al.* MP4: a machine learning based classification tool for prediction and

functional annotation of pathogenic proteins from metagenomic and genomic datasets. *BMC Bioinformatics* 2022;**23**:507. https://doi.org/10.1186/s12859-022-05061-7.
13. Singh S, Le NQK, Wang C. VF-Pred: predicting virulence factor using sequence alignment percentage and ensemble learning models. *Comput Biol Med* 2024;**168**:107662. https://doi.org/10.1016/j.compbiomed.2023.107662.
14. Ji B, Pi W, Liu W. *et al.* HyperVR: a hybrid deep ensemble learning approach for simultaneously predicting virulence factors and antibiotic resistance genes. *NAR Genom Bioinform* 2023;**5**:lqad012. https://doi.org/10.1093/nargab/lqad012.
15. Xie R, Li J, Wang J. *et al.* DeepVF: a deep learning-based hybrid framework for identifying virulence factors using the stacking strategy. *Brief Bioinform* 2021;**22**:bbaa125. https://doi.org/10.1093/bib/bbaa125.
16. Li G, Bai P, Chen J. *et al.* Identifying virulence factors using graph transformer autoencoder with ESMFold-predicted structures. *Comput Biol Med* 2024;**170**:108062. https://doi.org/10.1016/j.compbiomed.2024.108062.
17. Yang Z, Zeng X, Zhao Y. *et al.* AlphaFold2 and its applications in the fields of biology and medicine. *Signal Transduct Tar* 2023;**8**:1–14. https://doi.org/10.1038/s41392-023-01381-z.
18. Leman JK, Weitzner BD, Lewis SM. *et al.* Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nat Methods* 2020;**17**:665–80. https://doi.org/10.1038/s41592-020-0848-2.
19. Lin Z, Akin H, Rao R. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 2023;**379**:1123–30. https://doi.org/10.1126/science.ade2574.
20. Brandes N, Ofer D, Peleg Y. *et al.* ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics* 2022;**38**:2102–10. https://doi.org/10.1093/bioinformatics/btac020.
21. Elnaggar A, Heinzinger M, Dallago C. *et al.* ProtTrans: toward understanding the language of life through self-supervised learning. *IEEE Trans Pattern Anal Mach Intell* 2022;**44**:7112–27. https://doi.org/10.1109/TPAMI.2021.3095381.
22. Wu J, Qing H, Ouyang J. *et al.* HiFun: homology independent protein function prediction by a novel protein-language self-attention model. *Brief Bioinform* 2023;**24**:bbad311. https://doi.org/10.1093/bib/bbad311.
23. Wu J, Ouyang J, Qin H. *et al.* PLM-ARG: antibiotic resistance gene identification using a pretrained protein language model. *Bioinformatics* 2023;**39**:btad690. https://doi.org/10.1093/bioinformatics/btad690.
24. Fu L, Niu B, Zhu Z. *et al.* CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012;**28**:3150–2. https://doi.org/10.1093/bioinformatics/bts565.
25. Buchfink B, Reuter K, Drost HG. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods* 2021;**18**:366–8. https://doi.org/10.1038/s41592-021-01101-x.
26. Hamamsy T, Morton JT, Blackwell R. *et al.* Protein remote homology detection and structural alignment using deep learning. *Nat Biotechnol* 2024;975–85.
27. Hu EJ, Shen Y, Wallis P. *et al.* LoRA: low-rank adaptation of large language models. *The International Conference on Learning Representations* 2022;1–13.
28. Lin T-Y, Goyal P, Girshick R. *et al.* Focal loss for dense object detection. *IEEE Trans Pattern Anal Mach Intell* 2020;**42**:318–27.
29. Weights & Biases: The AI Developer Platform.
30. Sharma A, Garg A, Ramana J. *et al.* VirulentPred 2.0: an improved method for prediction of virulent proteins in bacterial pathogens. *Protein Sci* 2023;**32**:e4808. https://doi.org/10.1002/pro.4808.
31. Lavelle A, Sokol H. Gut microbiota-derived metabolites as key actors in inflammatory bowel disease. *Nat Rev Gastroenterol Hepatol* 2020;**17**:223–37. https://doi.org/10.1038/s41575-019-0258-z.
32. Vich Vila A, Imhann F, Collij V. *et al.* Gut microbiota composition and functional changes in inflammatory bowel disease and irritable bowel syndrome. *Sci Transl Med* 2018;**10**:eaap8914. https://doi.org/10.1126/scitranslmed.aap8914.
33. Jiang C. Progress in gut microbiota-host interaction. *Sci China Life Sci* 2024;**67**:851–3. https://doi.org/10.1007/s11427-024-2577-0.
34. Lloyd-Price J, Arze C, Ananthakrishnan AN. *et al.* Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* 2019;**569**:655–62. https://doi.org/10.1038/s41586-019-1237-9.
35. Patro R, Duggal G, Love MI. *et al.* Salmon: fast and bias-aware quantification of transcript expression using dual-phase inference. *Nat Methods* 2017;**14**:417–9. https://doi.org/10.1038/nmeth.4197.
36. Lomakin A, Svedlund J, Strell C. *et al.* Spatial genomics maps the structure, nature and evolution of cancer clones. *Nature* 2022;**611**:594–602. https://doi.org/10.1038/s41586-022-05425-2.
37. Li Z, Wang T, Liu P. *et al.* SpatialDM for rapid identification of spatially co-expressed ligand–receptor and revealing cell–cell communication patterns. *Nat Commun* 2023;**4**:3995. https://doi.org/10.1038/s41467-023-39608-w.
38. Huang Y, Shan Y, Zhang W. *et al.* Deciphering genetic causes for sex differences in human health through drug metabolism and transporter genes. *Nat Commun* 2023;**14**:175. https://doi.org/10.1038/s41467-023-35808-6.
39. MetaHIT Consortium, Li J, Jia H. *et al.* An integrated catalog of reference genes in the human gut microbiome. *Nat Biotechnol* 2014;**32**:834–41. https://doi.org/10.1038/nbt.2942.
40. Mi J, Jing X, Ma C. *et al.* Massive expansion of the pig gut virome based on global metagenomic mining. *NPJ Biofilms Microbi* 2024;**10**:76. https://doi.org/10.1038/s41522-024-00554-0.
41. Abdelhalim KA, Uzel A, Gülşen ÜN. The role of major virulence factors and pathogenicity of adherent-invasive Escherichia coli in patients with Crohn's disease. *pg* 2020;**15**:279–88. https://doi.org/10.5114/pg.2020.93235.
42. Maranga M, Szczerbiak P, Bezshapkin V. *et al.* Comprehensive functional annotation of metagenomes and microbial genomes using a deep learning-based method. *Msystems* 2023;**8**:e01178–22. https://doi.org/10.1128/msystems.01178-22.
43. Szydlowski LM, Bulbul AA, Simpson AC. *et al.* Adaptation to space conditions of novel bacterial species isolated from the international Space Station revealed by functional gene annotations and comparative genome analysis. *Microbiome* 2024;**12**:190. https://doi.org/10.1186/s40168-024-01916-8.
44. Koehler Leman J, Szczerbiak P, Renfrew PD. *et al.* Sequence-structure-function relationships in the microbial protein universe. *Nat Commun* 2023;**14**:2351. https://doi.org/10.1038/s41467-023-37896-w.

Briefings in Bioinformatics, 2025, **26(3)**, bbaf271
https://doi.org/10.1093/bib/bbaf271
Problem Solving Protocol