Bioinformatics, 38, 2022, ii168–ii174
https://doi.org/10.1093/bioinformatics/btac495
Proceedings of ECCB2022

Detecting DNA of novel fungal pathogens using ResNets and a curated fungi-hosts data collection

Jakub M. Bartoszewicz [1,2,*,†], Ferdous Nasri [1,2,†], Melania Nowicka [1,2] and Bernhard Y. Renard [1,*]

[1]Hasso Plattner Institute for Digital Engineering, Digital Engineering Faculty, University of Potsdam, Potsdam 14482, Germany and [2]Department of Mathematics and Computer Science, Free University of Berlin, Berlin 14195, Germany

*To whom correspondence should be addressed.
†The authors wish it to be known that these authors contributed equally.

accepted on July 8, 2022

Abstract

Background: Emerging pathogens are a growing threat, but large data collections and approaches for predicting the risk associated with novel agents are limited to bacteria and viruses. Pathogenic fungi, which also pose a constant threat to public health, remain understudied. Relevant data remain comparatively scarce and scattered among many different sources, hindering the development of sequencing-based detection workflows for novel fungal pathogens. No prediction method working for agents across all three groups is available, even though the cause of an infection is often difficult to identify from symptoms alone.
Results: We present a curated collection of fungal host range data, comprising records on human, animal and plant pathogens, as well as other plant-associated fungi, linked to publicly available genomes. We show that it can be used to predict the pathogenic potential of novel fungal species directly from DNA sequences with either sequence homology or deep learning. We develop learned, numerical representations of the collected genomes and visualize the landscape of fungal pathogenicity. Finally, we train multi-class models predicting if next-generation sequencing reads originate from novel fungal, bacterial or viral threats.
Conclusions: The neural networks trained using our data collection enable accurate detection of novel fungal pathogens. A curated set of over 1400 genomes with host and pathogenicity metadata supports training of machine-learning models and sequence comparison, not limited to the pathogen detection task.
Availability and implementation: The data, models and code are hosted at https://zenodo.org/record/5846345, https://zenodo.org/record/5711877 and https://gitlab.com/dacs-hpi/deepac.
Contact: jakub.bartoszewicz@hpi.de or bernhard.renard@hpi.de
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Many species of fungi are dangerous plant, animal or human pathogens. Importantly, even usually harmless opportunists can be deadly in susceptible populations. For example, *Candida albicans* causes common, relatively benign infections like thrush and vulvovaginal candidosis, affecting up to 75% of women at least once in their lifetime and often re-occurring multiple times (Sobel, 2007). It is also frequently found in healthy humans without leading to any disease and has been reported to be capable of stable colonization (Raimondi *et al.*, 2019). However, invasive *Candida* infections, especially bloodstream infections, can reach mortality rates of up to 75%, rivaling those of bacterial and viral sepsis (Brown *et al.*, 2012). A related species, *Candida auris*, has been first recognized in a human patient in 2009 (Satoh *et al.*, 2009) and quickly became one of the most urgent threats among the drug-resistant pathogens (CDC, 2019), reaching mortality rates of up to 60% (Spivak and Hanson, 2018). It might have originally been a plant saprophyte which has adapted to avian, and then also mammalian hosts, possibly prompted by climate change (Casadevall *et al.*, 2019). Strikingly, it seems to have emerged in three different clonal populations on three continents at the same time, for reasons that currently remain unexplained (Lockhart *et al.*, 2017).

Even though fungal infections are estimated to kill 1.6 million people a year, they remain understudied and underreported (Chowdhary *et al.*, 2016; Huseyin *et al.*, 2017; No author, 2017). Estimates suggest that between 1.5 million (Hawksworth, 2001) and 5.1 million (Blackwell, 2011), or even 6 million (Taylor *et al.*, 2014) different species of fungi exist, but only a small fraction of them has been sequenced. This poses a major challenge especially for pathogen detection workflows based on next-generation sequencing (NGS). Standard methods are based on recognition of known taxonomic units by homology detection, using either sequence alignment (Ahn *et al.*, 2015; Altschul *et al.*, 1990; Andrusch *et al.*, 2018; Camacho *et al.*, 2009; Hong *et al.*, 2014; Langmead and Salzberg, 2012; Li and Durbin, 2010; Li, 2018; Naccache *et al.*,

2014), k-mer-based approaches (Breitwieser *et al.*, 2018; Piro *et al.*, 2020; Wood *et al.*, 2019) or combinations thereof (Piro *et al.*, 2017). This in turn requires curated databases of fungal, as well as bacterial, viral and other species labelled with information regarding the corresponding pathogenic phenotype or host information. Limited host information is available in the NCBI Genome browser (Sayers *et al.*, 2021a), Database of Virulence Factors in Fungal Pathogens (Lu *et al.*, 2012) and the US National Fungus Collections Fungus-Host Database (Farr and Rossman, 2021). Those resources are partially complementary and none of them encompasses all the available data. What is more, multiple literature sources describe fungal pathogens and their hosts without referring to the corresponding genomes, even if they are indeed available in databases such as GenBank (Sayers *et al.*, 2021b) or FungiDB (Basenko *et al.*, 2018), which store genomic data without clear-cut host annotation. The ENHanCEd Infectious Diseases Database (EID2) (Wardeh *et al.*, 2015) aims to detect all 'carrier'–'cargo' relationships, not limited to fungi or pathogens specifically, although it does contain fungal pathogens as well. It relies on automatically mining the 'host' field in NCBI Taxonomy (Schoch *et al.*, 2020) and finding co-occurrences of species names in articles indexed by PubMed (Sayers *et al.*, 2021a), providing links to the associated nucleotide sequences. This method is efficient and scalable, but automated processing based on a concise set of simplifying assumptions may sometimes lead to spurious results. Many 'cargo' and 'carrier' species can be mentioned in the same paper even though one is not really a host of the other. This is often the case in literature reviews, taxonomy updates and holds also for this work. The 'host' field in a database as large as NCBI Taxonomy may also contain outdated, inaccurate or incomplete information. For example, *Pneumocystis jirovecii*, the causative agent of deadly pneumocystis pneumonia, was previously called *Pneumocystis carinii*. While the latter name is now reserved for a species infecting exclusively rats and not humans (Stringer *et al.*, 2002), records in Taxonomy (and, possibly by consequence, EID2) still list humans as the hosts of *P.carinii* at the time of writing. What is more, many sequences included in EID2 are not genome assemblies, but single genes, which are not enough for open-view fungal pathogen detection based on shotgun sequencing. For this and similar applications, a new resource is needed.

We compiled a collection of metadata on a comprehensive selection of fungal species, annotated according to their reported host groups and pathogenicity. We store the metadata in a flat-file database and link them to the corresponding representative (as defined in GenBank) or reference genomes, if available. To showcase the possible applications of the database, we model a scenario of novel fungal pathogen detection. While to our knowledge, this is a first systematic evaluation of feasibility of this task, we note that it mirrors similar problems in bacterial and viral genomics (Barash *et al.*, 2019; Bartoszewicz *et al.*, 2020, 2021b; Bergner *et al.*, 2021; Brierley and Fowler, 2021; Deneke *et al.*, 2017; Gałan *et al.*, 2019; Guo *et al.*, 2021; Mock *et al.*, 2020; Tang *et al.*, 2015; Wardeh *et al.*, 2021; Zhang *et al.*, 2019). We expect new agents to emerge due to environmental changes, host-switching events and growing human exposition to the unexplored diversity of potentially harmful fungi, as shown by the example of *C.auris*. Further, advances in engineering of fungal genomes (Amores *et al.*, 2016; Burgess, 2017; Dai *et al.*, 2020; Luo *et al.*, 2018; Martins-Santana *et al.*, 2018; Richardson *et al.*, 2017; Szymanski and Calvert, 2018) could lead to new risks and screening of synthetic sequences relies on of methods developed originally for pathogen detection (Balaji *et al.*, 2021; Diggans and Leproust, 2019). Therefore, we evaluate if detecting homology between previously unseen species and their known relatives accurately predicts if a DNA sequence originates from a novel fungus capable of colonizing and infecting humans. BLAST (Altschul *et al.*, 1990; Camacho *et al.*, 2009) represents the gold standard in pathogen detection via taxonomic assignment to the closest relative. Although read mappers or k-mer-based taxonomic classifiers are more computationally efficient on large NGS datasets (Alser *et al.*, 2021; Breitwieser *et al.*, 2019; Ye *et al.*, 2019), BLAST has been shown to be more accurate in similar tasks of detecting novel bacterial and viral pathogens (Bartoszewicz *et al.*, 2021b; Deneke *et al.*, 2017). However, convolutional neural networks of the DeePaC package have been proven to outperform BLAST in both those scenarios (Bartoszewicz *et al.*, 2020, 2021b) for both isolated NGS reads and full genomes, and a recently presented variant of residual neural networks (ResNets) outperforms all alternatives on short NGS reads and their fragments (Bartoszewicz *et al.*, 2021a). We trained similar ResNets to predict if a novel DNA sequence originates from a human-infecting fungus. To visualize the dataset, we developed trained numerical representations of all genomes in the database.

2 Materials and methods

2.1 Data description

We collected metadata on species infecting humans, animals or plants, supplemented with information on other plant-associated species. To this end, we integrated multiple literature and database sources (see Supplementary Information for citations), relying on manual curation, but also including the automatically extracted data for future reference. Supplementary Table S1 summarizes the data we collected for each species. We describe the curation procedure in detail in Supplementary Note S1. The database contains 14 555 records in total. Here, we will focus on what we will call the core database, comprising metadata on genomes of 954 manually confirmed pathogens (including 332 species reported to cause disease in humans), available on October 9, 2021. This forms a collection of species most relevant to the pathogen detection task, belonging to 6 phyla, 37 classes, 82 orders and 182 families. A 'temporal benchmark' subset contains 15 further pathogens (including one infecting humans), collected in a database update on January 2, 2022. We also include records on 486 plant-associated fungi. The supplementary part of the database contains information on 481 putatively labelled genomes, 1147 unlabelled species with available genomes, 8 synonyms (with 6 alternative genomes) derived from the Atlas of Clinical Fungi (de Hoog *et al.*, 2020), 885 labelled species without genomes (including 284 species without TaxIDs) and 10 579 putatively labelled species without genomes (including 9 without TaxIDs). This subset will enable easy updating of the database in the future, as more genomes of already labelled species are sequenced. It also serves as a record of all screened genomes and species to ensure reproducibility and facilitate future extensions (e.g. adding new data or sources of evidence). Supplementary Figure S1 presents the numbers of genomes with manually confirmed labels and genomes for which putative labels could be found in EID2 (Wardeh *et al.*, 2015).

2.2 Training, validation and test sets

While we envision a wide range of possible applications of the database, we present example usecases allowing one to take advantage of the wealth of collected data—detection of novel fungal pathogens from NGS data. The core of the database contains 332 genomes of human pathogens (including opportunists), forming the positive class. The negative class comprises 622 species not reported to infect humans; this includes 565 plant pathogens and 58 non-human animal pathogens. To evaluate the performance of the selected pathogenic potential prediction methods, we divided the corresponding genomes into non-overlapping training, validation and test sets. In this setup, the training set is used as a reference database for the methods based on sequence homology and to train the neural networks, while the performance metrics are calculated on the held-out test set. The validation set is used for hyperparameter tuning and to select the best training epoch. While evaluating performance on genuinely unknown species is by definition impossible, we effectively model the 'novel species' scenario by testing on a wide range of sequences removed from the database, including selected important species (Brown *et al.*, 2012; Dean *et al.*, 2012; Satoh *et al.*, 2009; Scheele *et al.*, 2019; Skamnioti and Gurr, 2009). We simulated low- and high-coverage read sets using a protocol outlined in Supplementary Note S2 (Holtgrewe, 2010). We considered two methods of balancing the number of samples between taxa: setting

the read number per species to be proportional to the respective genome's length ('linear-size') or its logarithm ('logarithmic-size').

2.3 Phenotype prediction and genome representations

Next, we evaluated the feasibility of pathogenic potential prediction for novel fungal species. We used a ResNet architecture implemented in the DeePaC package, previously shown to outperform alternatives based on deep learning, traditional machine learning and sequence homology in the context of novel bacteria and viruses (Bartoszewicz *et al.*, 2021a). We also adapted a BLAST-based pipeline used by Bartoszewicz *et al.* (2020) for benchmarking. Details of the pipeline, the ResNet architecture, hyperparameter tuning and the evaluation procedures for read pairs and genomes are described in Supplementary Note S3.

To visualize the structure of the dataset as learned by the trained classifier, we developed numerical representations for the collected genomes. This poses a challenge, as the networks are trained on reads rather than full genomes. However, we observe that final outputs of the network can be averaged over all reads originating from a single genome to generate a prediction for the genome in question (Bartoszewicz *et al.*, 2020). Analogously, we can average the activations of the intermediate layers to construct vector representations for whole genomes based on the corresponding reads. Note that averaging the activations of the penultimate layer is approximately equivalent to using a full genome as input (assuming full coverage), as our architecture uses global average pooling just before the output layer. Classifier outputs based on average activations indeed approximate classifier outputs averaged over all reads originating from a given species (see Supplementary Note S4 and Supplementary Figs S2 and S3). Hence, we extracted penultimate activations for all simulated reads in the low-coverage, 'linear-size' training, validation and test sets. We then used the averaged activation vectors for each species to map the distances between them as learned by our networks. We used UMAP (McInnes *et al.*, 2020) to visualize the dataset (Supplementary Note S6).

2.4 Multi-class evaluation

Finally, we investigated an application requiring merging the 'positive' subset of our database with previously available resources for pathogenic potential prediction in bacteria and viruses. We aimed to integrate the separate classifiers for fungi, bacteria and viruses into a single, multi-class model capable of predicting whether unassembled NGS reads originate from (possibly novel) pathogens present in a human-derived sample. To this end, we extended the DeePaC package adding the multi-class classification functionality. The resulting architectures differ from the previously described ResNets (Bartoszewicz *et al.*, 2021a) only by the output layer, which has as many units as the number of considered classes and uses a softmax activation (see Supplementary Note S3). In practice, only human-hosted fungi are expected to be found in clinical samples. In this context, a slightly constrained view is admissible: we assume that only human-pathogenic fungi, human-hosted bacteria (pathogenic or commensal), human viruses and non-human viruses (mainly bacteriophages) will be present in the sample. Further, bacteriophage sequences tend to be very similar to the sequences of their bacterial hosts (Zielezinski *et al.*, 2021; Zielezinski *et al.*, 2022) and difficult to differentiate, but both commensal bacteria and non-human viruses can be viewed here as a joint 'negative' (i.e. harmless) class. Hence, learning a precise decision boundary between them can be omitted. Human reads can be ignored, as they can be relatively easily filtered out with traditional methods based on read mapping or k-mers (Ahmed *et al.*, 2021; Loka *et al.*, 2018; Wood *et al.*, 2019). Therefore, we fused the previously published datasets used in DeePaC (Bartoszewicz *et al.*, 2021a) for bacteria (pathogens versus commensals) and viruses (human versus non-human) with the 'positive' (human-pathogenic) class of our database (Supplementary Notes S2 and S3). The final result is a dataset divided into four classes: nonpathogenic bacteria and non-human viruses, bacterial pathogens, human-infecting viruses and human-pathogenic fungi, in either the 'linear-size' or the 'logarithmic-size' variant. Note that even in this case, the 'negative' part of the presented database is useful, allowing us to constrain our view to a curated set of clinically relevant fungi only. Using this dataset, we trained two models including all four classes (using the 'linear-size' or the 'logarithmic-size' variant of the fungal training set). We further evaluated the one resulting in higher validation accuracy and a simple ensemble averaging the predictions of both models. Then, we trained a three-class model including only the bacterial and viral classes. This allows us to measure the 'difficulty' of integrating the fungal dataset with the others within a single network in terms of resulting differences in prediction accuracy on the original DeePaC datasets. By comparing the performance of our models to the performance of the original binary classifiers (Bartoszewicz *et al.*, 2021a), we can disentangle the 'difficulty' of adding the fungal class from the 'difficulty' of integrating the bacterial and viral classes, and assess how much performance is 'lost' by using a more open-view classifier. Note that in the case of the purely viral dataset, spurious assignments to the bacterial pathogens class may be treated as detection of bacteriophages infecting the bacterial species of this class and hence reassigned into predictions for the non-pathogen class by adding the predicted probabilities for both classes. This effectively merges the non-pathogen and bacterial pathogen classes at test time when appropriate, but still keeps the possibility to use the trained networks in a fully open-view setting (with all classes) without the need for retraining. We performed an additional comparison to BLAST with a pre-selected training database [bacterial for bacteria (Bartoszewicz *et al.*, 2020) and viral for viruses (Bartoszewicz *et al.*, 2021b)]. This resulted in an estimated upper bound on the performance of non-machine-learning approaches on those datasets (as extending the training database with irrelevant reference genomes can only lower BLAST's performance). Finally, we evaluated the neural networks on the full dataset of all four classes and a real *C.auris* sequencing run. We also analysed the latter with STAT (Katz *et al.*, 2021), used by the SRA database (Supplementary Note S7).

3 Results

3.1 Fungal pathogenic potential prediction

The best network, trained on the high-coverage, 'logarithmic-size' dataset without dropout, required 8 days of training on four Tesla V100 GPUs and was selected for further evaluation. Proper retuning of the classification threshold for species-level predictions appears to be a necessary step for an independent, viral dataset (Supplementary Table S2), so we also retuned the threshold (0.46 instead of the default 0.5) for the respective fungal ResNet setup. Overall, prediction accuracy for reads and read pairs is suboptimal for both BLAST and the ResNet, probably reflecting the extreme difficulty of the task (Supplementary Note S5 and Supplementary Table S3). The error estimates based on the held-out test dataset are consistent with the results of the temporal benchmark (Supplementary Table S4). Balanced accuracy is much higher on full genomes (88.4–90.3), suggesting that the main performance bottleneck is the total amount of information (total sequence length) available as input. This is consistent with previous observations (Bartoszewicz *et al.*, 2020, 2021b; Deneke *et al.*, 2017), although more drastic than in the case of bacteria or viruses. All approaches correctly classify the single human pathogen in the temporal benchmark dataset, but the ResNet achieves higher specificity (92.9) than read-based (78.6) and contig-based BLAST (85.7). Despite the low coverage, the test set seems to be indeed representative. The genome-wide predictions are equally accurate if all test reads are used and when only a half of the dataset (corresponding to either the first or the second mate) is analysed. Therefore, computations for single-species samples can be sped up by considering first mates only, as they are enough to deliver an accurate prediction. Strikingly, this is the case even though they correspond to a mean coverage below 0.08. As a result, the read-based ResNet yields only slightly less accurate predictions than contig-based BLAST, but requires 700-fold less time if a GPU is used (Supplementary Note S5 and Supplementary Table S3).

3.2 The landscape of fungal pathogenicity

Good overall accuracy of the ResNet is reflected in the visualization of learned genome representations for the entirety of the core

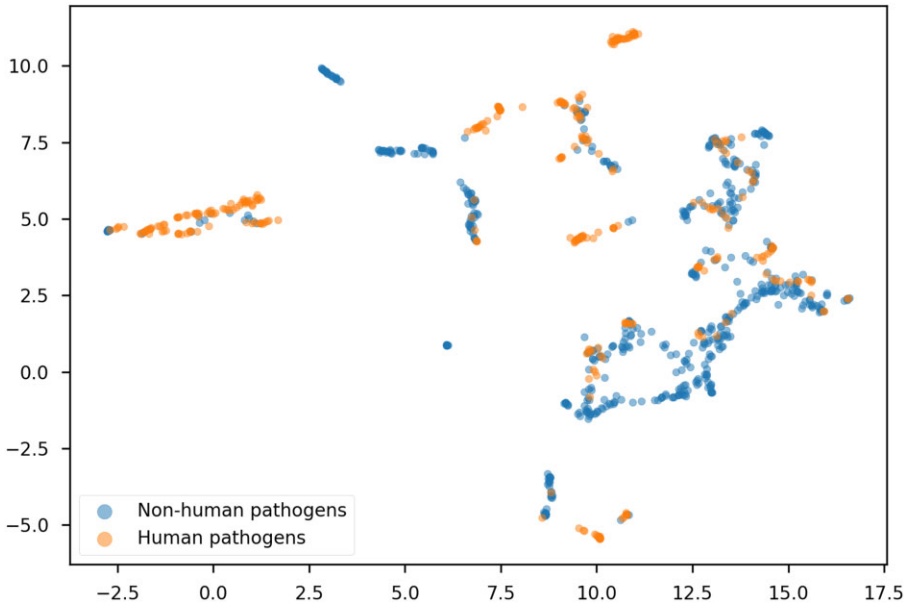

Fig. 1. UMAP embeddings of the learned genome representations for the core database, enlarged in Supplementary Figure S4. Each point represents a genome of a single species, coloured by its ground truth label. The learned representations offer a way of visualizing the core database along relevant labels for each genome. The ResNet correctly classifies most of the genomes (Supplementary Fig. S5). The clusters are related, but not fully reducible to the taxonomic classification of the analysed species (Supplementary Note S6 and Supplementary Figs S7–S10)

database. Figure 1 and Supplementary Figures S4–S10 present UMAP embeddings of the extracted representations for all labelled genomes, that is, a sum of the training, validation and test datasets. Although some noise is present, the positive and the negative class are mostly separated. Several clusters of human pathogens and non-human pathogens are present. The ResNet correctly recovers most of the labels, including many of the 'positive' members of the otherwise 'negative' clusters (Supplementary Fig. S5). To measure this, we visually identified 14 clusters, which could be easily retrieved automatically using single-linkage agglomerative clustering (Supplementary Fig. S10). Cluster purity for the whole dataset was high (0.90). We also measured it for the members of the large, mixed clusters 4 (pink in Supplementary Fig. S10) and 6 (red in Supplementary Fig. S10), which achieved purity of 0.85 and 0.88, respectively. Classification errors seem to originate from an interpolation based on neighbouring data points—within clusters, the predicted labels are more homogeneous than the ground truth annotations. This is expected, as the clusters represent similarity in the space of learned representations. The network should in general assign similar labels to inputs similar in this space. In contrast, BLAST works analogously to a *k*-nearest neighbours classifier in the input sequence space (finding the single closest match for each query). The ResNet, interpolating between multiple data points, may be less efficient in modelling situations where a small set of 'negative' data points is embedded within a larger 'positive' cluster of similar species or vice versa. This hypothesis is supported by the visualization of BLAST-predicted labels in the learned representation space (Supplementary Fig. S6). BLAST recovers mixed, contrasting labels within cluster more accurately and its errors seem to be more evenly distributed across the space. At the same time, its slightly lower sensitivity is especially visible within the diverse *Sordariomycetes* class placed in the rightmost cluster. The clusters themselves are noticeably related to the taxonomic units represented in the database, although this is importantly not a simple one-to-one mapping (Supplementary Note S6 and Supplementary Figs S7–S10).

3.3 Multi-class models

For the final evaluation of our database, we aimed to develop a model capable of classifying NGS reads originating from novel viruses, bacterial and fungal species into appropriate pathogen and non-pathogen classes. We trained the multi-class ResNets on data including four classes (human-pathogenic fungi, bacterial pathogens, human viruses and non-pathogens). The network trained on the dataset containing the 'logarithmic-size' version of the fungal positive class achieved slightly better validation accuracy and was selected for further evaluation, but the difference was small (<0.5%). We then evaluated a simple ensemble of both four-class ResNets. First, we used the DeePaC datasets consisting of bacteria and viruses to compare the four-class models to a classifier including the three non-fungal classes only, as well as the binary ResNets (Bartoszewicz *et al.*, 2021a) and BLAST. This procedure allows us to (i) measure the effect of integrating the fungal dataset with the bacterial and viral data in one task and (ii) disentangle the effects of adding the fungal data from the effects of merging the bacterial and viral datasets. We expected the fungal sequences to be relatively easy to differentiate from the others, but whether the ResNet architecture would be expressive enough to accurately represent all those diverse sequences was unclear. As shown in Supplementary Table S5, integrating the fungal dataset with three bacterial and viral classes indeed does not negatively influence the prediction accuracy. BLAST, using an appropriate reference database and representing the estimated upper bound on performance of homology-based approaches, is still outperformed by a significant margin. The fungal dataset can be integrated with the other classes without causing any significant performance hits on the full, multi-class dataset as well (Supplementary Table S6). Consistently with the results presented in Supplementary Table S5, performance is lower on the non-pathogen class, since many bacteriophage reads can be confused with pathogenic bacteria. While this issue requires further research, we expect future solutions to remain compatible with our database. The four-class ensemble achieves the most balanced performance on non-pathogen data, the best recall on fungal reads and is also the most accurate overall, cutting the average error rate by over 40% compared with BLAST (Table 1). As expected, distinguishing human-pathogenic fungi from the other classes is easier than predicting fungal hosts, so the performance of both BLAST and the ResNet is higher than in Supplementary Table S3. This holds also for real data (Supplementary Note S7 and Table 2). If the correct reference genome is missing, STAT is unable to classify most of the reads (Table 2 and Supplementary Table S7). This is true even though genomes of related species are present in the database. The ResNets perform markedly better than BLAST. Although the simulated test sets are more representative and effectively model mock metagenomic

Table 1. Performance on the multi-class dataset, read pairs

		Acc.	F1	Prec.	Rec.	AUPR
All classes	Four-class ens. (ours)	**87.6**	**87.7**	87.7	**87.6**	**93.4**
	BLAST	78.3	84.0	**90.6**	78.3	–
Non-pathogens	Four-class ens. (ours)	**77.4**	**78.7**	**80.1**	**77.4**	**86.7**
	BLAST	66.5	71.6	77.5	66.5	–
Path. bacteria	Four-class ens. (ours)	**87.2**	85.1	83.2	**87.2**	**90.4**
	BLAST	83.8	**87.5**	**91.6**	83.8	–
Human viruses	Four-class ens. (ours)	**90.9**	**93.7**	96.7	**90.9**	**98.4**
	BLAST	78.9	87.9	**99.2**	78.9	–
Fungi	Four-class ens. (ours)	**95.0**	**92.9**	90.9	**95.0**	**97.9**
	BLAST	84.1	88.9	**94.2**	84.1	–

Notes: The four-class classifier includes the fungi class along the three viral and bacterial classes included in the three-class classifier. The best performance for each class is marked in bold. In this setting, the true positive rate corresponds to the rate of correct assignments within a given class. Hence, recall is equal to accuracy for each class. We use the F1 score as an additional measure. As expected, BLAST's predictions are very precise, since when it finds a match, it is usually a relevant one. This does not hold for the non-pathogen class, which could indicate confusion between bacteriophage and bacterial pathogen reads. Our classifier significantly outperforms BLAST in terms of recall and prediction accuracy for all classes. BLAST, representing the estimated upper bound on performance of homology-based approaches, yields no predictions for 12.5% of all read pairs. Acc., accuracy; F1, F1 score; Prec., precision; Rec., recall and AUPR, area under the PR curve (best in bold).

Table 2. *C.auris* sequencing run, SRA accession SRR17577041

		Acc.	Rec.	Pred.
First read	Four-class ens. (ours)	**86.7**	**86.7**	**100.0**
	Four-class (ours)	**86.7**	**86.7**	**100.0**
	BLAST	49.8	49.8	50.2
	STAT	2.5	2.5	7.7
Both reads	Four-class ens. (ours)	**93.2**	**93.2**	**100.0**
	Four-class (ours)	92.6	92.6	**100.0**
	BLAST	56.4	56.4	57.1
	STAT	1.7	1.7	5.4

Notes: As this is a pure pathogen sample, accuracy and recall are equivalent. We report performance metrics for the first mate and both mates. The mean sequencing quality for the second mate was low (below 28). This is a problem especially for STAT, which performs worse if both reads are counted than if only the first, higher-quality mate is considered. BLAST and ResNets are more robust. ResNets are the best methods overall. Acc., accuracy; Rec., recall and Pred., prediction rate (a fraction of reads with any hits).

samples, this case study shows that our methods accurately classify real data as well. The database enabled us to accurately predict whether NGS reads originate from novel pathogens.

4 Discussion

Fungal pathogens have been under-studied compared with human-infecting bacteria and viruses, leading to repeated calls for more research in this area (Huseyin *et al.*, 2017; No author, 2017). What is more, a large part of the research effort has been focused on plant pathogens due to their agricultural significance. A subset of them could in principle also have an unreported or undetected ability to infect a human host. An analogous problem applies also to incomplete data regarding pathogenicity towards non-human animals or plants. For this reason, we do not claim that species not listed as potential pathogens are indeed non-pathogens. In our database, we include confirmed labels alongside appropriate sources; in the case of lack of evidence, we treat the respective label as missing. It is therefore possible that some of the fungi currently labelled as 'non-human pathogens' would have to be reclassified as the state of science evolves. This may be especially important as it has been suggested that the ongoing climate change will lead to more frequent host-switching events, including expansion of host range to mammals, which are usually relatively resistant to fungal infections (Garcia-Solache and Casadevall, 2010). Even though the very goal of the presented classifiers is to generalize to newly emerging species, large, comprehensive datasets are crucial—often more important than the actual analysis method used. This has been shown before for metagenomic data (Piro *et al.*, 2020) and likely applies to the tasks analysed here as well. Therefore, extending the database to include more species, as more genomes are sequenced in the future, could facilitate the downstream tasks. To support future extensions, we include all considered species in the database—even those without assigned TaxIDs, genomes or labels (in the case of screened GenBank genomes). This broadens the scope of the data from 1455 labelled genomes to over 14 500 records, enabling easy labelling of newly published genomes and minimizing the workload needed for addition of new, non-redundant records. It is also possible to link the species TaxIDs to taxa below the species level. However, it should be kept in mind that the fungal taxonomy is in constant flux—taxa previously considered variants of a single species may be reclassified into separate species in the future. While automatically curated databases like EID2 (Wardeh *et al.*, 2015) are relatively easy to update and scale, we note that they may be prone to errors introduced by the automated protocol used. Manual curation is not fully error-free either, but we see it as a necessary step to maximize the quality of the collected labels. Both approaches are complementary and may be best suited for different use-cases.

We show that both BLAST and ResNet can accurately predict if a fungus is a human pathogen based on its genome. Notably, ResNets offer a major reduction in inference time (up to 700-fold), both when used on GPUs and on CPUs. Although speed-optimized tools such as read mappers and *k*-mer-based taxonomic classifiers could be even faster, they have been shown to underperform in the context of novel pathogen detection (Bartoszewicz *et al.*, 2021b; Deneke *et al.*, 2017). The read-level performance is admittedly low for predicting a fungal host, but the trained representations allowed us to visualize the taxonomic diversity of the database along its phenotypic landscape. As expected, the apparent fungal host-range signal seems to be related to, but not fully reducible to the fungal taxonomy. Most importantly, multi-class networks detecting fungal, bacterial or viral pathogens noticeably outperform the homology-based approach. Different extraction protocols can affect the relative yield of bacterial and fungal DNA (Fiedorová *et al.*, 2019), but the methods investigated here process one sequence at a time, so are not affected by this kind of bias. Further work could extend the presented multi-class setup to Nanopore reads, as shown for bacterial and viral models (Bartoszewicz *et al.*, 2021a), enabling selective sequencing of mixed-pathogen samples.

Full genomes can be represented by aggregating representations of reads originating from each genome. In addition to that, we

observe that coverage as low as 0.08 is enough to correctly classify a species. Taken together, those two facts warrant a view of a species genome as a distribution generating subsequences (i.e. reads) originating from it; such a distribution can also be considered in an abstract representation space (Supplementary Note S4). This concept is very similar to that of a k-mer spectrum, where an *empirical* distribution of k-mers is used as a signature of a longer sequence to enable alignment-free comparisons (Zielezinski *et al.*, 2017), including being used as input features for machine-learning approaches as in Deneke *et al.* (2017). However, k-mer spectra operate in the sequence space only. Classifiers based on aggregated *representations* are approximately equivalent to classifiers based on aggregated *predictions*, although this relation is modulated by the standard deviation of the respective, genome-specific distribution. A somewhat related effect was reported in the context of competing design choices for neural networks equivariant to DNA reverse-complementarity—models averaging the predictions for both DNA strands were found to be approximately equivalent to models applying a sigmoid transformation to an average of logits (Zhou *et al.*, 2021). The probabilistic view of genome representations presented here deserves deeper investigation; this could potentially lead to a development of useful embeddings also for whole, multi-species samples.

Although we focus on using the collected data in a pathogenic potential prediction task, the database itself can find future applications beyond this particular problem. Genomes collected here can be a valuable resource for functional and comparative genomics of fungi. For example, fungal genomes could be scanned for regions associated with their ability to colonize and infect humans, as shown previously for bacteria and viruses (Bartoszewicz *et al.*, 2021b). On the other hand, the multitude of genomic features present in fungal genomes, often including intron features and regions without obvious functional annotation, renders the validation of such an approach a challenging project on its own. This could be perhaps facilitated by focusing exclusively on coding regions, which should in principle carry a stronger phenotype-related signal, at the risk of omitting potentially relevant, non-coding (e.g. regulatory) elements. As a source of curated labels, the dataset could also support application of proteomics to fungal pathogen research. Computational metaproteomics and proteogenomics approaches enable analysis of microbial communities based on mass spectrometry data and can be co-opted for pathogen detection workflows independent of DNA sequencing (Renard *et al.*, 2012; Schiebenhoefer *et al.*, 2019, 2020).

In conclusion, we compiled a comprehensive database of fungal species linked to their host group (human, non-human animal or plant), evidence for their pathogenicity and publicly available genomes. To highlight the potential uses of the dataset, we benchmark two most promising approaches to novel fungal pathogen detection: a deep neural network capable of fast inference directly from DNA sequences and the gold standard in homology-based pathogen identification—BLAST. The database, hosted at https://zenodo.org/record/5846345, can be reused for future research on fungal pathogenicity. The models, read sets and code are available at https://zenodo.org/record/5711877, https://zenodo.org/record/5846397 and https://github.com/dacs-hpi/deepac.

Acknowledgements

We thank the NCBI help desk for assistance and helpful suggestions, as well as Katharina Baum (HPI) for valuable discussions and comments.

Funding

This paper was published as part of a special issue financially supported by ECCB2022. This work was supported by the Computational Life Science initiative funded by Bundesministerium für Bildung und Forschung [project DeepPath, 031L0208, to B.Y.R.] and the de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) funded by Bundesministerium für Bildung und Forschung [031A537B, 031A533A, 031A538A, 031A533B, 031A535A, 031A537C, 031A534A, 031A532B].

Conflict of Interest: none declared.

References

No author. (2017) Stop neglecting fungi. *Nature Microbiology*, **2**, 17120.
Ahmed,O. *et al.* (2021) Pan-genomic matching statistics for targeted nanopore sequencing. *iScience*, **24**, 102696.
Ahn,T.-H. *et al.* (2015) Sigma: strain-level inference of genomes from metagenomic analysis for biosurveillance. *Bioinformatics*, **31**, 170–177.
Alser,M. *et al.* (2021) Technology dictates algorithms: recent developments in read alignment. *Genome Biol.*, **22**, 249.
Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
Amores,G.R. *et al.* (2016) Recent progress on systems and synthetic biology approaches to engineer fungi as microbial cell factories. *Curr. Genomics*, **17**, 85–98.
Andrusch,A. *et al.* (2018) PAIPline: pathogen identification in metagenomic and clinical next generation sequencing samples. *Bioinformatics*, **34**, i715–i721.
Balaji,A. *et al.* (2021) SeqScreen: accurate and sensitive functional screening of pathogenic sequences via ensemble learning. bioRxiv, page 2021.05.02.442344 (Cold Spring Harbor Laboratory Section: New Results).
Barash,E. *et al.* (2019) BacPaCS—bacterial pathogenicity classification via Sparse-SVM. *Bioinformatics*, **35**, 2001–2008.
Bartoszewicz,J.M. *et al.* (2020) DeePaC: predicting pathogenic potential of novel DNA with reverse-complement neural networks. *Bioinformatics*, **36**, 81–89.
Bartoszewicz,J.M. *et al.* (2021a) Deep learning-based real-time detection of novel pathogens during sequencing. *Brief. Bioinform.*, **22**, (bbab269).
Bartoszewicz,J.M. *et al.* (2021b) Interpretable detection of novel human viruses from genome sequencing data. *NAR Genom. Bioinform.*, **3**, lqab004.
Basenko,E.Y. *et al.* (2018) FungiDB: an integrated bioinformatic resource for fungi and oomycetes. *J. Fungi*, **4**, 39.
Bergner,L.M. *et al.* (2021) Characterizing and evaluating the zoonotic potential of novel viruses discovered in vampire bats. *Viruses*, **13**, 252.
Blackwell,M. (2011) The fungi: 1, 2, 3... 5.1 million species? *Am. J. Bot.*, **98**, 426–438.
Breitwieser,F.P. *et al.* (2018) KrakenUniq: confident and fast metagenomics classification using unique k-mer counts. *Genome Biol.*, **19**, 198.
Breitwieser,F.P. *et al.* (2019) A review of methods and databases for metagenomic classification and assembly. *Brief. Bioinform.*, **20**, 1125–1136.
Brierley,L. and Fowler,A. (2021) Predicting the animal hosts of coronaviruses from compositional biases of spike protein and whole genome sequences through machine learning. *PLoS Pathog.*, **17**, e1009149.
Brown,G.D. *et al.* (2012) Hidden killers: human fungal infections. *Sci. Transl. Med.*, **4**, 165rv13.
Burgess,D.J. (2017) Synthetic biology: building a custom eukaryotic genome de novo. *Nat. Rev. Genet.*, **18**, 274–274.
Camacho,C. *et al.* (2009) BLAST+: architecture and applications. *BMC Bioinform.*, **10**, 421.
Casadevall,A. *et al.* (2019) On the emergence of *Candida auris*: climate change, azoles, swamps, and birds. *MBio*, **10**, e01397–19.
CDC. (2019) *Antibiotic Resistance Threats in the United States, 2019*. U.S. Department of Health and Human Services, CDC.
Chowdhary,A. *et al.* (2016) Filamentous fungi in respiratory infections. *PLoS Pathog.*, **12**, e1005491.
Dai,J. *et al.* (2020) Sc3.0: revamping and minimizing the yeast genome. *Genome Biol.*, **21**, 205.
de Hoog,G. *et al.* (2020). *Atlas of Clinical Fungi*. 4th edn. Hilversum.
Dean,R. *et al.* (2012) The top 10 fungal pathogens in molecular plant pathology. *Mol. Plant Pathol.*, **13**, 414–430.
Deneke,C. *et al.* (2017) PaPrBaG: a machine learning approach for the detection of novel pathogens from NGS data. *Sci. Rep.*, **7**, 39194.
Diggans,J. and Leproust,E. (2019) Next steps for access to safe, secure DNA synthesis. *Front. Bioeng. Biotechnol.*, **7**, 86.
Farr,D.F. and Rossman,A.Y. (2021) *Fungal Databases* (https://nt.ars-grin.gov/fungaldatabases/, retrieved October 9, 2021, and January 2, 2022).
Fiedorová,K. *et al.* (2019) The impact of DNA extraction methods on stool bacterial and fungal microbiota community recovery. *Front. Microbiol.*, **10**, 821.
Gałan,W. *et al.* (2019) Host taxon predictor—a tool for predicting taxon of the host of a newly discovered virus. *Sci. Rep.*, **9**, 3436.
Garcia-Solache,M.A. and Casadevall,A. (2010) Global warming will bring new fungal diseases for mammals. *mBio*, **1**, e00061–10.
Guo,Q. *et al.* (2021) Predicting hosts based on early SARS-CoV-2 samples and analyzing later world-wide pandemic in 2020. bioRxiv, p. 2021.03.21.436312.

Hawksworth,D.L. (2001) The magnitude of fungal diversity: the 1.5 million species estimate revisited. *Mycol. Res*., **105**, 1422–1432.
Holtgrewe,M. (2010) Mason—a read simulator for second generation sequencing data. *Technical Report*. FU Berlin.
Hong,C. *et al.* (2014) PathoScope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples. *Microbiome*, **2**, 33.
Huseyin,C.E. *et al.* (2017) Forgotten fungi—the gut mycobiome in human health and disease. *FEMS Microbiol. Rev*., **41**, 479–511.
Katz,K.S. *et al.* (2021) STAT: a fast, scalable, MinHash-based k-mer tool to assess sequence read archive next-generation sequence submissions. *Genome Biol*., **22**, 270.
Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with bowtie 2. *Nat. Methods*, **9**, 357–359.
Li,H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**, 3094–3100.
Li,H. and Durbin,R. (2010) Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, **26**, 589–595.
Lockhart,S.R. *et al.* (2017) Simultaneous emergence of multidrug-resistant *Candida auris* on 3 continents confirmed by whole-genome sequencing and epidemiological analyses. *Clin. Infect. Dis*., **64**, 134–140.
Loka,T.P. *et al.* (2018) PriLive: privacy-preserving real-time filtering for next-generation sequencing. *Bioinformatics (Oxford, England)*, **34**, 2376–2383.
Lu,T. *et al.* (2012) DFVF: database of fungal virulence factors. *Database (Oxford)*, **2012**, bas032.
Luo,Z. *et al.* (2018) Identifying and characterizing SCRaMbLEd synthetic yeast using ReSCuES. *Nat. Commun*., **9**, 1930.
Martins-Santana,L. *et al.* (2018) Systems and synthetic biology approaches to engineer fungi for fine chemical production. *Front. Bioeng. Biotechnol*., **6**, 117.
McInnes,L. *et al.* (2020) UMAP: uniform manifold approximation and projection for dimension reduction. *arXiv*, 1802.03426 [cs, stat]. arXiv: 1802.03426.
Mock,F. *et al.* (2020) VIDHOP, viral host prediction with deep learning. *Bioinformatics*, **37**, 318–325.
Naccache,S.N. *et al.* (2014) A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. *Genome Res*., **24**, 1180–1192.
Piro,V.C. *et al.* (2017) Metameta: integrating metagenome analysis tools to improve taxonomic profiling. *Microbiome*, **5**, 1–11.
Piro,V.C. *et al.* (2020) Ganon: precise metagenomics classification against large and up-to-date sets of reference sequences. *Bioinformatics*, **36**, i12–i20.
Raimondi,S. *et al.* (2019) Longitudinal survey of fungi in the human gut: its profiling, phenotyping, and colonization. *Front. Microbiol*., **10**, 1575.
Renard,B.Y. *et al.* (2012) Overcoming species boundaries in peptide identification with Bayesian information criterion-driven error-tolerant peptide search (biceps). *Mol. Cell. Proteomics*, **11**, M111.014167-1–014167.
Richardson,S.M. *et al.* (2017) Design of a synthetic yeast genome. *Science*, **355**, 1040–1044.
Satoh,K. *et al.* (2009) *Candida auris* sp. nov., a novel ascomycetous yeast isolated from the external ear canal of an inpatient in a Japanese hospital. *Microbiol. Immunol*., **53**, 41–44.
Sayers,E.W. *et al.* (2021a) Database resources of the national center for biotechnology information. *Nucleic Acids Res*., **49**, D10–D17.
Sayers,E.W. *et al.* (2021b) GenBank. *Nucleic Acids Res*., **49**, D92–D96.
Scheele,B.C. *et al.* (2019) Amphibian fungal panzootic causes catastrophic and ongoing loss of biodiversity. *Science*, **363**, 1459–1463.
Schiebenhoefer,H. *et al.* (2019) Challenges and promise at the interface of metaproteomics and genomics: an overview of recent progress in metaproteogenomic data analysis. *Expert Rev. Proteomics*., **16**, 375–390.
Schiebenhoefer,H. *et al.* (2020) A complete and flexible workflow for metaproteomics data analysis based on MetaProteomeAnalyzer and prophane. *Nat. Protoc*., **15**, 3212–3239.
Schoch,C.L. *et al.* (2020) NCBI taxonomy: a comprehensive update on curation, resources and tools. *Database*, **2020**.
Skamnioti,P. and Gurr,S.J. (2009) Against the grain: safeguarding rice from rice blast disease. *Trends Biotechnol*., **27**, 141–150.
Sobel,J.D. (2007) Vulvovaginal candidosis. *Lancet*, **369**, 1961–1971.
Spivak,E.S. and Hanson,K.E. (2018) *Candida auris*: an emerging fungal pathogen. *J. Clin. Microbiol*., **56**, 1–10.
Stringer,J.R. *et al.* (2002) A new name for *Pneumocystis* from humans and new perspectives on the host–pathogen relationship. *Emerg. Infect. Dis*., **8**, 891–896.
Szymanski,E. and Calvert,J. (2018) Designing with living systems in the synthetic yeast project. *Nat. Commun*., **9**, 2950.
Tang,Q. *et al.* (2015) Inferring the hosts of coronavirus using dual statistical models based on nucleotide composition. *Sci. Rep*., **5**, 17155.
Taylor,D.L. *et al.* (2014) A first comprehensive census of fungi in soil reveals both hyperdiversity and fine-scale niche partitioning. *Ecol. Monogr*., **84**, 3–20.
Wardeh,M. *et al.* (2015) Database of host–pathogen and related species interactions, and their global distribution. *Sci. Data*, **2**, 150049.
Wardeh,M. *et al.* (2021) Predicting mammalian hosts in which novel coronaviruses can be generated. *Nat. Commun*., **12**, 780.
Wood,D.E. *et al.* (2019) Improved metagenomic analysis with kraken 2. *Genome Biol*., **20**, 257.
Ye,S.H. *et al.* (2019) Benchmarking metagenomics tools for taxonomic classification. *Cell*, **178**, 779–794.
Zhang,Z. *et al.* (2019) Rapid identification of human-infecting viruses. *Transbound. Emerg. Dis*., **66**, 2517–2522.
Zhou,H. *et al.* (2021) Towards a better understanding of reverse-complement equivariance for deep learning models in regulatory genomics. bioRxiv, p. 2020.11.04.368803.
Zielezinski,A. *et al.* (2017) Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biol*., **18**, 186.
Zielezinski,A. *et al.* (2021) Taxonomy-aware, sequence similarity ranking reliably predicts phage–host relationships. *BMC Biol*., **19**, 223.
Zielezinski,A. *et al.* (2022) PHIST: fast and accurate prediction of prokaryotic hosts from metagenomic viral sequences. *Bioinformatics*, **38**, 1447–1449.