Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

Subject Section

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, 14482 Potsdam, Germany and [2]Department of Mathematics and Computer Science, Free University of Berlin, 14195 Berlin, Germany.

*To whom correspondence should be addressed. †Contributed equally.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

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 remains 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 1,400 genomes with host and pathogenicity metadata supports training of machine learning models and sequence comparison, not limited to the pathogen detection task.
Availability: 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`.

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; noa, 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 (Li and Durbin, 2010; Langmead and Salzberg, 2012; Li, 2018; Altschul *et al.*, 1990; Camacho *et al.*, 2009; Hong *et al.*, 2014; Naccache *et al.*, 2014; Ahn *et al.*, 2015; Andrusch *et al.*, 2018), k-mer based approaches (Wood *et al.*, 2019; Piro *et al.*, 2020; Breitwieser *et al.*, 2018) 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

© The Author 2021.

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 U.S. 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 *P. 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 (Deneke *et al.*, 2017; Mock *et al.*, 2020; Zhang *et al.*, 2019; Barash *et al.*, 2018; Guo *et al.*, 2021; Gałan *et al.*, 2019; Wardeh *et al.*, 2021; Brierley and Fowler, 2021; Bergner *et al.*, 2021; Tang *et al.*, 2015; Bartoszewicz *et al.*, 2020, 2021b). 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 (Richardson *et al.*, 2017; Burgess, 2017; Dai *et al.*, 2020; Szymanski and Calvert, 2018; Luo *et al.*, 2018; Amores *et al.*, 2016; Martins-Santana *et al.*, 2018) could lead to new risks, and screening of synthetic sequences relies on of methods developed originally for pathogen detection (Diggans and Leproust, 2019; Balaji *et al.*, 2021). 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 (Breitwieser *et al.*, 2017; Ye *et al.*, 2019; Alser *et al.*, 2021), BLAST has been shown to be more accurate in similar tasks of detecting novel bacterial and viral pathogens (Deneke *et al.*, 2017; Bartoszewicz *et al.*, 2021b). 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 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 Supp. Information for citations), relying on manual curation, but also including the automatically extracted data for future reference. Table S1 summarizes the data we collected for each species. We describe the curation procedure in detail in Supp. Note 1. 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, 1,147 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). 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 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 (Satoh *et al.*, 2009; Brown *et al.*, 2012; Dean *et al.*, 2012; Skamnioti and Gurr, 2009; Scheele *et al.*, 2019). We simulated low- and high-coverage read sets using a

protocol outlined in Supp. Note 2 (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 and the hyperparameter tuning procedure are described in Supp. Note 3.

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 Supp. Note 4 and Figure S2-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 (Supp. Note 6).

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 Supp. Note 3). 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.*, 2021b,a) 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 (Loka *et al.*, 2018; Ahmed *et al.*, 2021; Wood *et al.*, 2019). Therefore, we fused the previously published datasets used in DeePaC (Bartoszewicz *et al.*, 2021a) for bacteria (pathogens vs. commensals) and viruses (human vs. non-human) with the 'positive' (human-pathogenic) class of our database (Supp. Notes 2-3). The final result is a dataset divided into 4 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 3-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), 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 analyzed the latter with STAT (Katz *et al.*, 2021), used by the SRA database (Supp. Note 7).

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 (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 (Supp. Note 5, Table S3). The error estimates based on the held-out test dataset are consistent with the results of the temporal benchmark (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 (Deneke *et al.*, 2017; Bartoszewicz *et al.*, 2020, 2021b), 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 analyzed. Therefore, computations for single-species samples can be sped up by considering first mates only, as they are enough to deliver an accurate

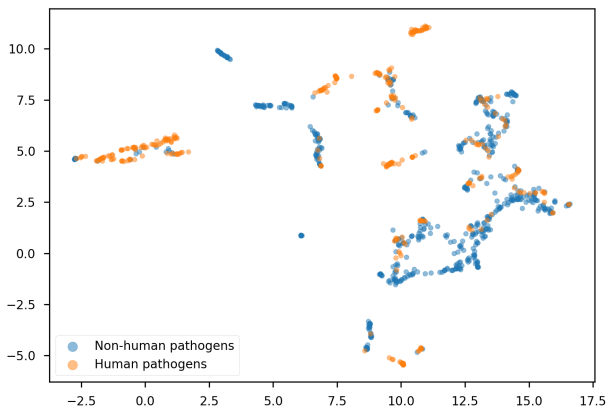


Fig. 1: UMAP embeddings of the learned genome representations for the core database, enlarged in Figure S4. Each point represents a genome of a single species, coloured by its ground truth label. The learned representations offer a way of visualizing the core database along relevant labels for each genome. The ResNet correctly classifies most of the genomes (Figure S5). The clusters are related, but not fully reducible to the taxonomic classification of the analysed species (Supp. Note 6, Figures S7–S10).

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

		Acc.	F1	Prec.	Rec.	AUPR
All classes	4-class ens. (ours)	**87.6**	**87.7**	87.7	**87.6**	**93.4**
	BLAST	78.3	84.0	**90.6**	78.3	-
Non-pathogens	4-class ens. (ours)	**77.4**	**78.7**	**80.1**	**77.4**	**86.7**
	BLAST	66.5	71.6	77.5	66.5	-
Path. bacteria	4-class ens. (ours)	**87.2**	85.1	83.2	**87.2**	**90.4**
	BLAST	83.8	**87.5**	**91.6**	83.8	-
Human viruses	4-class ens. (ours)	**90.9**	**93.7**	96.7	**90.9**	**98.4**
	BLAST	78.9	87.9	**99.2**	78.9	-
Fungi	4-class ens. (ours)	**95.0**	**92.9**	90.9	**95.0**	**97.9**
	BLAST	84.1	88.9	**94.2**	84.1	-

Table 1. The 4-class classifier includes the fungi class along the three viral and bacterial classes included in the 3-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, AUPR – area under the PR curve (best in bold).

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 (Supp. Note 5, Table S3).

3.2 The landscape of fungal pathogenicity

Good overall accuracy of the ResNet is reflected in the visualisation of learned genome representations for the entirety of the core database. Figure 1 and Figures S4–S10 presents UMAP embeddings of the extracted representations for all labelled genomes, i.e. 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 (Figure S5). To measure this, we visually identified 14 clusters, which could be easily retrieved automatically using single-linkage agglomerative clustering (Figure 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 Figure S10) and 6 (red in Figure 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 (Figure 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 (Supp. Note 6, Figures 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 4-class ResNets. First, we used the DeePaC datasets consisting of bacteria and viruses to compare the 4-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 a) measure the effect of integrating the fungal dataset with the bacterial and viral data in one task, and b) 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 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 (Table S6). Consistently with the results presented in 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 4-class ensemble achieves the most

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

		Acc.	Rec.	Pred.
First read	4-class ens. (ours)	**86.7**	**86.7**	**100.0**
	4-class (ours)	**86.7**	**86.7**	**100.0**
	BLAST	49.8	49.8	50.2
	STAT	2.5	2.5	7.7
Both reads	4-class ens. (ours)	**93.2**	**93.2**	**100.0**
	4-class (ours)	92.6	92.6	**100.0**
	BLAST	56.4	56.4	57.1
	STAT	1.7	1.7	5.4

Table 2. 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, Pred. – prediction rate (a fraction of reads with any hits).

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 to 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 Table S3. This holds also for real data (Supp. Note 7, Table 2). If the correct reference genome is missing, STAT is unable to classify most of the reads (Table 2, 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 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 to human-infecting bacteria and viruses, leading to repeated calls for more research in this area (Huseyin *et al.*, 2017; noa, 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 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 analyzed 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 1,455 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. 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 (Supp. Note 4). 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`.

Funding

This work was supported by the BMBF-funded Computational Life Science initiative (project DeepPath, 031L0208, to B.Y.R.) and the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A537B, 031A533A, 031A538A, 031A533B, 031A535A, 031A537C, 031A534A, 031A532B).
Conflict of interest: None declared.

Acknowledgements

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

References

(2017). Stop neglecting fungi. *Nature Microbiology*, **2**(8), 1–2.

Ahmed, O. *et al.* (2021). Pan-genomic matching statistics for targeted nanopore sequencing. *iScience*, **24**(6), 102696.

Ahn, T.-H. *et al.* (2015). Sigma: Strain-level inference of genomes from metagenomic analysis for biosurveillance. *Bioinformatics*, **31**(2), 170–177.

Alser, M. *et al.* (2021). Technology dictates algorithms: recent developments in read alignment. *Genome Biology*, **22**(1), 249.

Altschul, S. F. *et al.* (1990). Basic local alignment search tool. *Journal of Molecular Biology*, **215**(3), 403–410.

Amores, G. R. *et al.* (2016). Recent Progress on Systems and Synthetic Biology Approaches to Engineer Fungi As Microbial Cell Factories. *Current Genomics*, **17**(2), 85–98.

Andrusch, A. *et al.* (2018). PAIPline: pathogen identification in metagenomic and clinical next generation sequencing samples. *Bioinformatics*, **34**(17), 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. Publisher: Cold Spring Harbor Laboratory Section: New Results.

Barash, E. *et al.* (2018). BacPaCS—Bacterial Pathogenicity Classification via Sparse-SVM. *Bioinformatics*.

Bartoszewicz, J. M. *et al.* (2020). DeePaC: predicting pathogenic potential of novel DNA with reverse-complement neural networks. *Bioinformatics*, **36**(1), 81–89.

Bartoszewicz, J. M. *et al.* (2021a). Deep learning-based real-time detection of novel pathogens during sequencing. *Briefings in Bioinformatics*, (bbab269).

Bartoszewicz, J. M. *et al.* (2021b). Interpretable detection of novel human viruses from genome sequencing data. *NAR Genomics and Bioinformatics*, **3**(1).

Basenko, E. Y. *et al.* (2018). FungiDB: An Integrated Bioinformatic Resource for Fungi and Oomycetes. *Journal of Fungi*, **4**(1), 39.

Bergner, L. M. *et al.* (2021). Characterizing and Evaluating the Zoonotic Potential of Novel Viruses Discovered in Vampire Bats. *Viruses*, **13**(2), 252.

Blackwell, M. (2011). The fungi: 1, 2, 3 … 5.1 million species? *American Journal of Botany*, **98**(3), 426–438.

Breitwieser, F. P. *et al.* (2017). A review of methods and databases for metagenomic classification and assembly. *Briefings in Bioinformatics*, **20**(4), 1125–1136.

Breitwieser, F. P. *et al.* (2018). KrakenUniq: confident and fast metagenomics classification using unique k-mer counts. *Genome Biology*, **19**(1), 198.

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 Pathogens*, **17**(4), e1009149.

Brown, G. D. *et al.* (2012). Hidden killers: Human fungal infections. *Science Translational Medicine*, **4**(165).

Burgess, D. J. (2017). Synthetic Biology: Building a custom eukaryotic genome de novo. *Nature Reviews Genetics*, **18**(5), 274–274.

Camacho, C. *et al.* (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, **10**(1), 421.

Casadevall, A. *et al.* (2019). On the emergence of candida auris: climate change, azoles, swamps, and birds. *MBio*, **10**(4), 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. What Lies Beyond Aspergillosis and Mucormycosis? *PLOS Pathogens*, **12**(4), e1005491.

Dai, J. *et al.* (2020). Sc3.0: revamping and minimizing the yeast genome. *Genome Biology*, **21**(1), 205.

de Hoog, G. *et al.* (2020). *Atlas of Clinical Fungi, 4th edition*. Hilversum.

Dean, R. *et al.* (2012). The Top 10 fungal pathogens in molecular plant pathology. *Molecular Plant Pathology*, **13**(4), 414–430.

Deneke, C. *et al.* (2017). PaPrBaG: A machine learning approach for the detection of novel pathogens from NGS data. *Scientific Reports*, **7**, 39194.

Diggans, J. and Leproust, E. (2019). Next Steps for Access to Safe, Secure DNA Synthesis. *Frontiers in Bioengineering and Biotechnology*, **7**.

Farr, D. F. and Rossman, A. Y. (2021). Fungal databases. 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. *Frontiers in Microbiology*, **10**.

Garcia-Solache, M. A. and Casadevall, A. (2010). Global warming will bring new fungal diseases for mammals. *mBio*, **1**(1), e00061–10.

Gałan, W. *et al.* (2019). Host Taxon Predictor - A Tool for Predicting Taxon of the Host of a Newly Discovered Virus. *Scientific Reports*, **9**(1), 3436.

Guo, Q. *et al.* (2021). Predicting Hosts Based on Early SARS-CoV-2 Samples and Analyzing Later World-wide Pandemic in 2020. *bioRxiv*, page 2021.03.21.436312.

Hawksworth, D. L. (2001). The magnitude of fungal diversity: the 1.5 million species estimate revisited. *Mycological Research*, **105**(12), 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**(1), 33.

Huseyin, C. E. *et al.* (2017). Forgotten fungi—the gut mycobiome in human health and disease. *FEMS Microbiology Reviews*, **41**(4), 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 Biology*, **22**(1), 270.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature methods*, **9**(4), 357–359.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**(18), 3094–3100.

Li, H. and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, **26**(5), 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. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, **64**(2), 134–140.

Loka, T. P. *et al.* (2018). PriLive: privacy-preserving real-time filtering for next-generation sequencing. *Bioinformatics (Oxford, England)*, **34**(14), 2376–2383.

Lu, T. *et al.* (2012). DFVF: database of fungal virulence factors. *Database*, **2012**(bas032).

Luo, Z. *et al.* (2018). Identifying and characterizing SCRaMbLEd synthetic yeast using ReSCuES. *Nature Communications*, **9**(1), 1930.

Martins-Santana, L. *et al.* (2018). Systems and Synthetic Biology Approaches to Engineer Fungi for Fine Chemical Production. *Frontiers in Bioengineering and Biotechnology*, **6**.

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*.

Naccache, S. N. *et al.* (2014). A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. *Genome Research*, **24**(7), 1180–1192.

Piro, V. C. *et al.* (2017). Metameta: integrating metagenome analysis tools to improve taxonomic profiling. *Microbiome*, **5**(1), 1–11.

Piro, V. C. *et al.* (2020). ganon: precise metagenomics classification against large and up-to-date sets of reference sequences. *Bioinformatics*, **36**(Supplement_1), i12–i20.

Raimondi, S. *et al.* (2019). Longitudinal survey of fungi in the human gut: Its profiling, phenotyping, and colonization. *Frontiers in Microbiology*, **10**.

Renard, B. Y. *et al.* (2012). Overcoming species boundaries in peptide identification with bayesian information criterion-driven error-tolerant peptide search (biceps). *Molecular & cellular proteomics*, **11**(7), M111–014167.

Richardson, S. M. *et al.* (2017). Design of a synthetic yeast genome. *Science*, **355**(6329), 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. *Microbiology and Immunology*, **53**(1), 41–44.

Sayers, E. W. *et al.* (2021a). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, **49**(D1), D10–D17.

Sayers, E. W. *et al.* (2021b). GenBank. *Nucleic Acids Research*, **49**(D1), D92–D96.

Scheele, B. C. *et al.* (2019). Amphibian fungal panzootic causes catastrophic and ongoing loss of biodiversity. *Science*, **363**(6434), 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 Review of Proteomics*, **16**(5), 375–390.

Schiebenhoefer, H. *et al.* (2020). A complete and flexible workflow for metaproteomics data analysis based on MetaProteomeAnalyzer and Prophane. *Nature Protocols*, **15**(10), 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 in Biotechnology*, **27**(3), 141–150.

Sobel, J. D. (2007). Vulvovaginal candidosis. *The Lancet*, **369**(9577), 1961–1971.

Spivak, E. S. and Hanson, K. E. (2018). Candida auris: an emerging fungal pathogen. *Journal of clinical microbiology*, **56**(2).

Stringer, J. R. *et al.* (2002). A New Name for Pneumocystis from Humans and New Perspectives on the Host-Pathogen Relationship. *Emerging Infectious Diseases*, **8**(9), 891–896.

Szymanski, E. and Calvert, J. (2018). Designing with living systems in the synthetic yeast project. *Nature Communications*, **9**(1), 2950.

Tang, Q. *et al.* (2015). Inferring the hosts of coronavirus using dual statistical models based on nucleotide composition. *Scientific Reports*, **5**(1), 17155.

Taylor, D. L. *et al.* (2014). A first comprehensive census of fungi in soil reveals both hyperdiversity and fine-scale niche partitioning. *Ecological Monographs*, **84**(1), 3–20.

Wardeh, M. *et al.* (2015). Database of host-pathogen and related species interactions, and their global distribution. *Scientific Data*, **2**(1), 150049.

Wardeh, M. *et al.* (2021). Predicting mammalian hosts in which novel coronaviruses can be generated. *Nature Communications*, **12**(1), 780.

Wood, D. E. *et al.* (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology*, **20**(1), 257.

Ye, S. H. *et al.* (2019). Benchmarking Metagenomics Tools for Taxonomic Classification. *Cell*, **178**(4), 779–794.

Zhang, Z. *et al.* (2019). Rapid identification of human-infecting viruses. *Transboundary and Emerging Diseases*, **66**(6), 2517–2522.

Zhou, H. *et al.* (2021). Towards a Better Understanding of Reverse-Complement Equivariance for Deep Learning Models in Regulatory Genomics. *bioRxiv*, page 2020.11.04.368803.

Zielezinski, A. *et al.* (2017). Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biology*, **18**(1), 186.

Zielezinski, A. *et al.* (2021a). PHIST: fast and accurate prediction of prokaryotic hosts from metagenomic viral sequences. *bioRxiv*.

Zielezinski, A. *et al.* (2021b). Taxonomy-aware, sequence similarity ranking reliably predicts phage–host relationships. *BMC Biology*, **19**(1), 223.

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

Supplementary Information

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

Supplementary Note 1: Database construction

First, we accessed the Database of Virulence Factors in Fungal Pathogens (DFVF) [Lu et al., 2012] on October 9, 2021. The database contains records on virulence factors of a wide selection of fungal species together with their NCBI TaxIDs and information on their hosts and the disease caused. It also includes separate parts for animal and plant pathogens. We searched the records for all proteins in the database; if the "Disease" or "Disease host" fields contained the word "human", we added the corresponding species to the list of human pathogens. We added remaining species from the animal part of the database to the list of animal pathogens. Finally, we put all species mentioned in the plant part of the database on the list of plant pathogens, irrespective of whether humans were also mentioned as a possible host. For all those species, we also extracted the TaxID. As some of those TaxIDs corresponded to taxa below the species level, we linked them to their species-level ancestors in the NCBI Taxonomy tree. This resulted in a list of unique, labelled species names and species-level TaxIDs. Note that as DFVF was built using manual curation of text-mining results, we treat the obtained records as manually curated as well.

Further, we searched the Taxonomy database [Schoch et al., 2020] using the query '"host human"[Properties] AND Fungi [orgn]', obtaining a list of species with a putative human host and their TaxIDs. We then used a comprehensive selection of literature reviews and comparative studies [Taylor et al., 2001, Chowdhary et al., 2014, Bombassaro et al., 2020, Brunner-Mendoza et al., 2019, Franzen and Müller, 2001, Colombo et al., 2011, Han and Weiss, 2017, Hu et al., 2015, Jančič et al., 2015, Kjærbølling et al., 2018, Köhler et al., 2015, Kwon-Chung et al., 2017, Paulussen et al., 2017, Prakash et al., 2017, Dellière et al., 2020, Teixeira et al., 2017, Seyedmousavi et al., 2018] to collect additional evidence and extend the collected list of human pathogens. We manually extracted the species names and searched the NCBI Taxonomy (including synonyms) to link them to their names as reported in Taxonomy and the respective species-level TaxIDs. In both cases (searching with extracted TaxIDs or species names), we handled the possibility of multiple, ambiguous matches by preferring exact name matches to any of the synonymous names recorded in Taxonomy. If this was not possible (e.g. because of additional annotations in the "name" field in Taxonomy), we considered matches where the first two words exactly match the query name, and varieties of a query species (i.e. matches containing the abbreviation 'var.'). If no exact match was found, we selected the first record containing the query species name. We ignored species hybrids and *formae speciales*, excluded the organism names explicitly expressing taxonomic ambiguity (marked with 'sp.', 'cf.' or 'aff.'), and filtered out fungal viruses. If the species name was not mentioned in Taxonomy at all, we still included it in the database along its associated host group label. This will enable easy extension of the database in the future, as more fungal sequences become available.

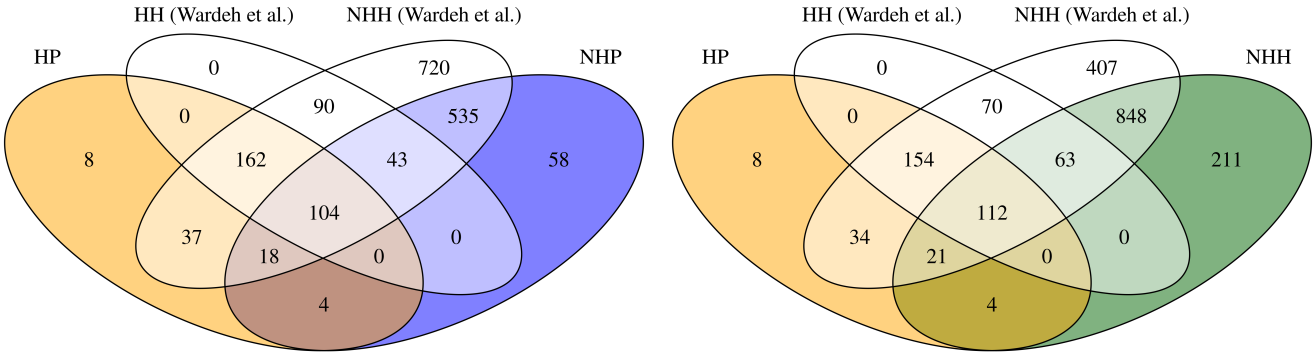
Similarly, we consulted the literature on animal [Seyedmousavi et al., 2018, Smith, 2006, Bałazy et al., 2008, Becnel and Andreadis, 2014, Franzen and Müller, 2001, Cissé et al., 2021, Chandler et al., 2000, Evans et al., 2011, Gerson et al., 2008, Han and Weiss, 2017, Schulenburg and Félix, 2017, Leonhardt et al., 2018, Lovett and St. Leger,

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

2017, Rehner and Buckley, 2005, Seyedmousavi et al., 2015, Shang et al., 2015, Wu et al., 2021, St. Leger and Wang, 2020, Sung et al., 2007, Teixeira et al., 2017, van der Geest et al., 2000] and plant pathogens [Vacher et al., 2008, Doehlemann et al., 2017, Barbara and Clewes, 2003, Costa et al., 2021, Coutinho et al., 2017, Gurung et al., 2015, Jones and Baker, 2007, Kim et al., 2012, 2019, Leonhardt et al., 2018, Schardl et al., 2013, Stukenbrock et al., 2012] and linked the results to the respective species TaxIDs wherever possible. Further, we accessed the Fungal Databases of the Germplasm Resources Information Network (GRIN). On October 9, 2021, we searched for all records with 'Homo sapiens' as a host [Farr and Rossman, 2021]. We also added the species listed in the Nomenclature Fact Sheets for plant-associated fungi with quarantine importance and the Fungal Diagnostic Fact Sheets of invasive and emerging fungal pathogens [Farr and Rossman, 2021] to the list of plant-pathogenic species, and searched NCBI Taxonomy as described earlier.

We then linked the resulting lists of species TaxIDs with their representative (as defined in GenBank) or reference genomes in GenBank, accessed on October 9, 2021. Further, we selected all species with available genomes, but without a label assigned by any of the sources used. On the same day, we manually searched the GRIN Fungal Database for each of those species, allowing for synonyms. We checked the "Host, "Disease" and "Notes" fields; if the disease of a plant or animal host was clearly described, we added the query species to the appropriate list. If no disease was mentioned for a confirmed plant host, we added the species to a list of plant-associated fungi. Next, we used the EID2 database [Wardeh et al., 2015] to extract species with putative human, animal, or plant hosts. Note that those labels were collected automatically and are therefore prone to errors. Moreover, they do not consist of pathogens only, and may include commensals or symbionts. We nevertheless added them to our database for completeness, albeit noting that they represent only a putative, automatically extracted carrier-cargo relationship, rather than a manually confirmed, true pathogenic potential of a species. The same, 'putative' category includes human-hosted species found in NCBI Taxonomy and GRIN, unless confirmed in other literature sources as well. We then selected all species with a putative human host and manually searched the Atlas of Clinical Fungi [de Hoog et al., 2020] with their names and synonyms noted in NCBI Taxonomy, considering also additional species referred to in the search results. If the Atlas confirmed a species to be a pathogen, we added it to the appropriate category. Note that only three of the human-hosted species (all belonging to the genus *Malassezia*) were clearly described as human commensals without any reports of causing disease (and hence, we excluded them from our list of pathogens). In cases when the Atlas mentioned two names to be synonyms, even though NCBI Taxonomy listed them as two separate species, we followed the nomenclature suggested by the Atlas. We retained both records to explicitly reflect this in the database, keeping one name unlabelled and linking it to the main record with an appropriate annotation regarding the synonym TaxID and name. Finally, we linked all labelled species to their GenBank genomes, if available.

After 12 weeks, we updated the database by extracting the records on reference and representative fungal genomes present in GenBank on January 2, 2022, selecting those with accession numbers and TaxIDs not already present in our database, and filtering the organism names as described above. We then linked the 67 new genomes to the previously collected metadata (29 of those species were absent from the first version of the database), and manually searched GRIN following the same procedure as for the main part of the database. If a human host was mentioned, we added the corresponding putative label to the appropriate list. Finally, we checked all putative human host labels in the Atlas of Clinical Fungi [de Hoog et al., 2020]. The update added 27 new labelled genomes, including 12 plant-associated fungi without proven pathogenic labels. One plant pathogen and four plant-associated species were absent from the first version of the database. We used the added pathogen genomes for our temporal benchmark.



(a) Core database

(b) Core database + genomes of plant-associated species

Figure S1: A Venn diagram of the manually confirmed labels and putative labels from EID2 [Wardeh et al., 2015] for all labelled genomes (except synonyms). HP – human pathogens (manually curated); NHP – non-human pathogens (manually curated); NHH – species with a non-human host (manually curated); HH (Wardeh et al.) – species with a putative human host (EID2); NHH (Wardeh et al.) – species with a putative non-human host (EID2).

Table S1: Key information on each species included in the database.

Name	Species name, GenBank organism name and infraspecific name for the representative genome.
NCBI TaxID	Taxonomic identifiers of the species and the representative genome in the NCBI Taxonomy database.
Assembly accession & name	GenBank assembly accession number and assembly name for the representative genome.
Pathogenicity & host group	Manually curated information regarding pathogenicity and the relevant host groups.
Putative host group	Automatically extracted information suggesting a host group. Includes non-pathogenic associations.
Label sources	References to resources used to label the species for all host groups and evidence levels.
Source name and TaxID	Species name and TaxID as mentioned in the first resource used to label it.
Assembly level	Assembly level of the representative genome.
Sequence release date	Sequence release date as reported in GenBank.
FTP path	GenBank FTP path to the representative genome.
Label date	Date of adding the species to the database or updating its labels.

We use genomes labelled as representative or reference in GenBank.

Supplementary Note 2: Training, validation and test read set simulation

Fungal datasets

Pathogen 'novelty' is an inherently ambiguous concept. We operationalize it in taxonomic terms, measuring generalization to 'novel' (i.e. unseen in training) species. This is consistent with previous works (e.g. Deneke et al. [2017], Zhang et al. [2019], Bartoszewicz et al. [2020, 2021b], Barash et al. [2018], Katz et al. [2021], Wood et al. [2019]), some of which also investigated 'novelty' measured at other levels of taxonomy (e.g. strain). Alternatively, novelty could be expressed in terms of sequence similarity. However, designation of novel species is done by experts considering multiple criteria, including, but not limited to genome similarity. Hence, we follow the established operational definition of a novel pathogen as a novel *species*. Note that BLAST detects sequence similarity, so its performance (or lack thereof) can be interpreted as an alternative measure of 'novelty' of the test genomes relative to the training database.

We manually placed two clinically important pathogens – *Candida auris* (a recently emerged threat [Satoh et al., 2009, CDC, 2019]) and *Aspergillus fumigatus* (a causative agent of aspergillosis, reaching up to 95% mortality rates [Brown et al., 2012]) – in the test set. We also added two non-human pathogens: *Pyricularia oryzae* (syn. *Magnaporthe oryzae*; voted the most important fungal pathogen by a panel of almost 500 fungal pathologists [Dean et al., 2012] as it causes up to 30% of global rice production losses [Skamnioti and Gurr, 2009]) and *Batrachochytrium dendrobatidis* (blamed for recent decline of 500 amphibian species and complete extinction of 90 of them [Scheele et al., 2019]). The rest of the species were assigned randomly. Overall, we placed 10% of all species in test and validation sets each, while the remaining 80% formed the training set.

To model the task of predicting pathogenic potentials from NGS samples, we simulated training, validation and test reads using Mason [Holtgrewe, 2010]. First, we generated 10 million single reads, 1.25 million single reads and 1.25 million paired reads per class for the training, validation and test sets respectively, keeping the number of reads per genome proportional to each genome's length. This mirrors the protocol used for bacteria [Bartoszewicz et al., 2020] and viruses [Bartoszewicz et al., 2021b], resulting in equal mean coverage for all genomes. A side effect of such an approach is that reads originating from longer genomes could be over-represented compared to those from shorter or incomplete genomes, possibly causing generalization problems for trained machine learning models. To tackle this issue, we also simulated a second version of the training set, where the number of reads per genome is proportional to a logarithm of a given genome's length. We call this version of the training set the 'logarithmic-size' set to differentiate it from the previous ('linear-size') approach. Note that while the 'logarithmic-size' training set may help balance the resulting classifiers' performance, in a real sample, we expect the coverage to be approximately equal for equally abundant species. Therefore, we only used the 'linear-size' versions of validation and test sets. In both cases, the datasets maintain the 8-1-1 proportions on both read and genome level, with a total count of 25 million reads per class.

A crucial difference compared to previous pathogenic potential prediction studies is that fungal genomes are orders of magnitude longer than bacterial or viral ones. While the procedure mentioned above produced a mean coverage of 1.82 for the bacterial dataset presented in [Bartoszewicz et al., 2020], it results in a mean coverage of only 0.15 on our data. We expected this could cause training problems, as the machine learning models would only have access to a small fraction of the overall sequence diversity of the training set. Therefore, we also simulated 'high-coverage' versions of both the 'linear-size' and 'logarithmic-size' training sets. In this setup, we increased the total number of training reads to 240 million, compared to 20 million for the 'low-coverage' versions described before. The 'high-coverage' versions keep the mean coverage of 1.82. As the reads were simulated randomly, we expected the 'low-coverage' validation and test sets to be representative, and correctly model a common case of low abundance of the target pathogen compared to other DNA sources in the sample (e.g. the host). We therefore used only 'low-coverage' validation and test sets. In summary, we generated four versions of the training read set and one version of the validation and test set each. The datasets contain the same number of reads for each of the classes and can be easily reused for future machine learning and benchmarking applications. Finally, we simulated a temporal test

set based on the genomes added to GenBank up to 12 weeks after the main part of the database was compiled (see Supplementary Note 1). We generated 625,000 reads using the 'linear-size' setup and balancing the number of reads per class; the average coverage matched our main test set.

Joint bacterial, viral and fungal datasets

As the bacterial and viral datasets [Bartoszewicz et al., 2021a] contain only 12.5 million reads per class (with the same 8-1-1 proportion of the training, validation and test sets, stratified by bacterial species or virus), we used the low-coverage versions of our 'positive' read sets, matching this number. We also merged the original negative classes (commensal bacteria and non-human viruses) into a joint non-pathogen class, downsampling accordingly. Further, the original datasets contain reads of between 25 and 250bp instead of 250bp only, which makes the problem more challenging, but increases robustness on very short sequences [Bartoszewicz et al., 2021a]. To make our fungal reads compatible with this setup, we randomly shortened the reads in our validation and low-coverage training sets to 25–250bp. We only used the test sets of 250bp in order to highlight the upper performance limit of the resulting classifier.

Each of the classes contains 12.5 million reads divided in the training, validation and test subsets while keeping separate species or viruses in each of the subsets. This is intended to model a scenario of analysing a clinical sample containing a mixture of previously unseen pathogens and non-pathogens, while keeping the task feasible by resigning from differentiating between bacteriophages and commensal bacteria, as well as fungi with human and non-human hosts.

Supplementary Note 3: Prediction algorithms and hyperparameter tuning

Following the convention introduced previously for bacteria [Deneke et al., 2017] and viruses [Bartoszewicz et al., 2021b], we focus on predicting the pathogenic potential of the analyzed species, rather than the pathogenicity itself. This distinction captures the fact that the pathogenic phenotype may or may not be manifested depending on specific host-pathogen interactions; whether this happens cannot be fully predicted outside the context of a particular host. We can only predict if a given sequence originates from a species that could cause disease under some circumstances (i.e. its pathogenic potential), not whether an invasive infection actually occurred. However, the task remains inherently difficult – a DNA sequence by itself, outside of the context of the host, can be probabilistically associated with a certain trait, but the trait itself is only realized in the biological system comprising both the host and the colonizing microbe as a whole. Therefore, we intentionally include opportunistic pathogens in the 'positive' class of human pathogens. These considerations mirror challenges previously described for other pathogen classes [Deneke et al., 2017, Bartoszewicz et al., 2020, 2021b,a]. Given limited information content of a short NGS read, predicting if a fungus can infect humans directly (and only) from reads is therefore an extremely challenging task.

The ResNet architecture used in this work consists of 17 convolutional layers of between 64 and 512 filters (with filter size of 7 for the first layer and 5 for the following layers) using skip connections, followed by a global average pooling layer and a single-neuron output layer with sigmoid activation. It returns an output score between 0 and 1, where a threshold of 0.5 is used by default as a boundary between 'positive' and 'negative' predictions. The network uses batch normalization and input dropout, which can be understood as randomly switching a predefined fraction of input nucleotides in training samples to Ns. It also guarantees identical predictions for any given sequence and its reverse-complement via parameter sharing. For more details, we refer the reader to the corresponding publication [Bartoszewicz et al., 2021a].

As previous research reported that the input dropout rate is an important hyperparameter [Bartoszewicz et al., 2020, 2021b,a], we investigated two different values – the default 0.25 and 0 (no dropout). To this end, we trained the networks using the 'linear-size' and 'logarithmic-size' versions of the low-coverage training set for the default maximum of 30 epochs with an early stopping patience of 10 epochs, selecting the epoch and input dropout rate

resulting in the highest accuracy on the validation set. We then retrained the networks using the selected input dropout rates and both versions of the high-coverage training set. The multi-class networks use the default input dropout rate (0.25), as required by the addition of viral and bacterial data. We trained those models for 50 epochs due to the increased difficulty of the task.

As shown in [Bartoszewicz et al., 2020], predictions for two reads in a read pair can be averaged, improving prediction accuracy; training the networks on single reads allows to use them for both single reads and read pairs. Further, averaging over predictions for all reads originating from the same organism (e.g. for classification of whole genomes or single-species samples) has been demonstrated to yield accurate species-level predictions [Deneke et al., 2017, Bartoszewicz et al., 2020]. We adapted those steps to our models as well. However, since the networks are explicitly trained for read-based classification only, the default classification threshold may be suboptimal for the genome-wise classification scenario. Therefore, we hypothesized that it should be retuned for optimal performance according to a preselected metric, e.g. balanced accuracy. To test this assumption on a fully independent dataset first, we used the 'novel viral species' dataset from [Bartoszewicz et al., 2021b] and their convolutional neural network (CNN) without any retraining. We selected the threshold optimizing the balanced accuracy on the appropriate validation set, rounded to two decimal places. We compared the original and retuned models and applied the procedure resulting in the better test set performance to our fungal models.

To classify novel fungal sequences based on their closest taxonomic matches, we used BLAST [Altschul et al., 1990, Camacho et al., 2009] as described for bacterial data in [Bartoszewicz et al., 2020]. The reference database was created from the training set genomes. We used discontiguous megablast with an E-value cutoff of 10 and default parameters, selecting the label of the top hit ('pathogenic to humans' or not) as a predicted label for each query sequence. Note that if no hits are found, no prediction can be returned for a given read. This lowers the true positive and true negative rates, defined as the ratios of correct positive or negative predictions to all reads in the positive or negative class, respectively. For read pairs, a single match is enough to assign a label to a pair, but conflicting matches result in no prediction. Similarly to [Bartoszewicz et al., 2021b], we considered two approaches for generating species-level predictions. First, we found BLAST matches of contigs from the test set, using a majority vote over all contigs belonging to the same species to assign a corresponding label. Second, we used the majority vote over predictions for all reads originating from a given species. The latter case is especially relevant for a use-case of a sequencing sample containing a single-species, and is more directly comparable to the genome-level evaluation of the ResNet.

Table S2: Classification threshold tuning for full available genomes on the 'novel viral species' dataset of DeePaC-vir [Bartoszewicz et al., 2021b].

	BAcc.	Prec.	Rec.	Spec.
CNN (retuned)	**75.9**	31.1	**70.8**	81.1
CNN [Bartoszewicz et al., 2021b]	64.9	31.0	40.6	89.1
BLAST (reads)	61.8	**46.8**	30.2	**93.5**
BLAST (genome)	64.0	44.9	36.5	91.5

Average prediction over all reads from a genome are used as a prediction for a given species. BLAST (reads) corresponds to a majority vote over all reads from a genome, and BLAST (genome) to the majority vote over all contigs. The negative class is more numerous in the dataset, which affects the precision values. The retuned classifier uses the threshold of 0.45 instead of the default 0.5 and achieves the best balanced accuracy. BAcc. – balanced accuracy, Prec. – precision; Rec. – recall, Spec. – specificity.

Supplementary Note 4: Relationship between $s(E[z])$ and $E[s(z)]$

Methods

Averaging the activations of the penultimate layer for a set of reads belonging to the same genome and passing the output to the final layer, we obtain the value $s(E[z])$, where $s(x) = 1/(1 + e^{-x})$ is the sigmoid activation and $z = \mathbf{w}^\top \mathbf{h} + b$, for the penultimate activation vector $\mathbf{h}$, weights of the last layer $\mathbf{w}$ and its bias b. Note that this is *not* mathematically equivalent to $E[s(z)]$, or the expected output of the network for reads originating from a given genome, which was used for the genome-level predictions in previously published literature [Bartoszewicz et al., 2020, 2021b], as well as in this study. However, we suspected that although usually $s(E[z])$ and $E[s(z)]$ are not identical, those terms could have similar values in practice, as they both express a notion of aggregating information contained in individual reads to yield a prediction for a full genome. If this is the case, mean activations for a set of reads originating from the same genome can be used as if they were the internal genome representations of our species-level classifiers, even though the models rely on aggregating read-level predictions. To investigate the relationship between $s(E[z])$ and $E[s(z)]$, we performed additional experiments.

We speculated that average activations of the penultimate layer $E[\mathbf{h}]$ could be used as representations of full genomes, even though our genome-level classifiers do not explicitly use $E[\mathbf{h}]$, but rely on average per-read predictions $E[s(z)]$. Note that $E[s(z)]$ represents averaging in in the space of predicted probabilities ('proba'-average), while $s(E[z])$ is averaging in the logit space ('logit'-average). To check if $s(E[z]) \approx E[s(z)]$, we plotted the 'logit'-average predictions for each species against their 'proba'-average equivalents. As any effects found could be dataset-dependent, we perform this not only on the fungal validation and test sets, but the 'novel viruses' dataset used for the DeePaC-vir networks (and consequently, also for our multi-class classifiers). Further, we use simulated logit values (z) to gain deeper insight into the relationship between both versions of read averaging. To accurately model the problem of aggregating predictions for individual species in the context of binary classifications, we first defined two classes, '0' and '1'. We then simulated 100 'species' belonging to each class. Each species S is described by a distribution of logit values $z \sim \mathcal{N}(\mu_S, \sigma_S^2)$, where μ_S and σ_S are the mean and standard deviation of z for that particular species and $\mathcal{N}$ is the normal distribution. To generate μ_S and σ_S for each species, we sample from additional distributions: $\mu_S \sim \mathcal{N}(\mu_c, \sigma_c^2)$ and $\sigma_S \sim HalfNormal(\mu_\sigma, \sigma_\sigma^2)$, where $HalfNormal$ is the half-normal distribution and c is the class index for classes '0' and '1'. Note that for simplicity, we sample σ_S from the same distribution irrespective of the species's class. Finally we generate 1000 values of z per species (corresponding to 1000 reads) and plot the 'logit'-averages $s(E[z])]$ against the 'proba'-averages $E[s(z)]$. Hence, to each species belonging to a given class, we assign its own mean and standard deviation used to generate read representations according to the species's class. First, we perform two experiments using the following parameters:

- $\mu_0 = -3, \mu_1 = 3, \sigma_0 = \sigma_1 = 2, \mu_\sigma = 0.25, \sigma_\sigma = 0.25$
- $\mu_0 = -3, \mu_1 = 3, \sigma_0 = \sigma_1 = 2, \mu_\sigma = 3, \sigma_\sigma = 1$

This models two well-separated classes, and the only difference between the two settings is the distribution of within-species variances of z.

We then test if the relationship between 'logit'- and 'proba'-average changes if no separable classes are present in the dataset. To this end, we perform two additional experiments with only one 'class', '0':

- $\mu_0 = 0, \sigma_0 = 3, \mu_\sigma = 0.25, \sigma_\sigma = 0.25$
- $\mu_0 = 0, \sigma_0 = 3, \mu_\sigma = 3, \sigma_\sigma = 1$

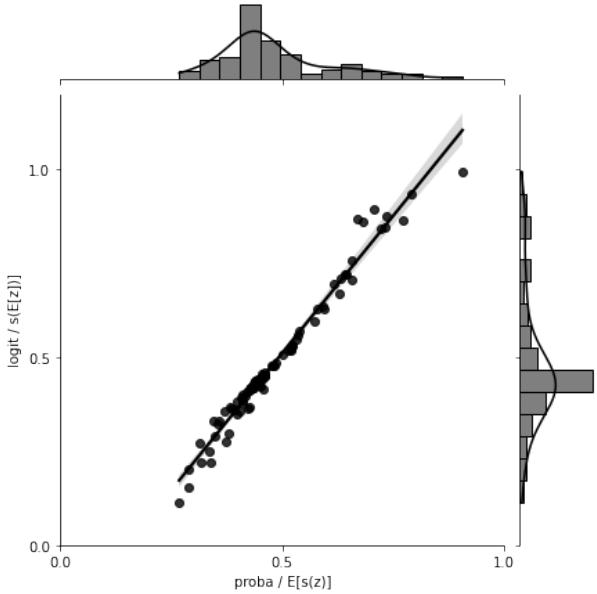
Finally, we investigate if any effects found could be conditional on the presence of a species-level signal, i.e. on differences in the distribution of z for each species S. We generate 100,000 values of $z \sim \mathcal{N}(\mu_r, \sigma_r^2)$, where $\mu_r = 0$ and $\sigma_r = 20$, and assign the 'reads' to 100 simulated 'species' arbitrarily (first 1000 reads to the first species, second 1000 reads to the second species, etc.). We also repeat the experiment simulating 5000 'reads' per 'species' (corresponding to a 5-fold increase in coverage).

Results

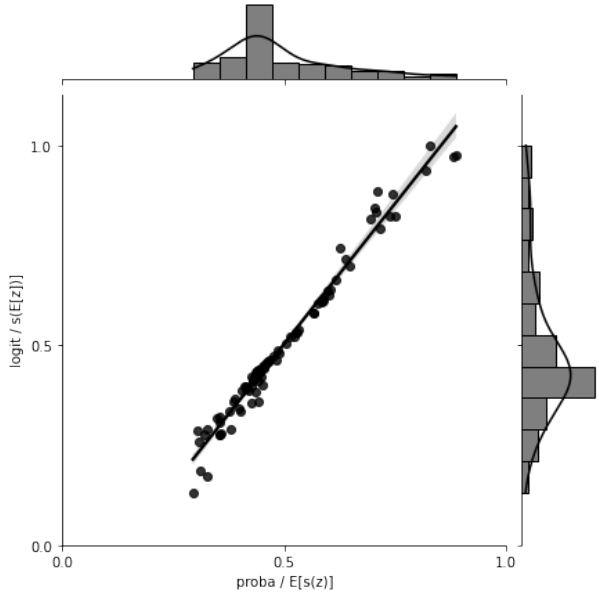
As shown in Figure S2a and Figure S2b, the values of $s(E[z])$ and $E[s(z)]$ are indeed similar for both the validation and test set. They are highly correlated (Spearman's $\rho > 0.99$, $p < 10^{-15}$) and linearily dependent. This justifies using $E[\mathbf{h}]$ as our genome representations. However, for the 'novel viruses' dataset, the relationship actually resembles a sigmoid function (Figure S2c). To investigate the relationship between the two versions of read averaging, we performed further experiments using simulated logit values z. In the binary classification case, if σ_S is low, the relationship between 'logit'- and 'proba'-averages is linear (Figure S3a), and if σ_S is high, the relationship is sigmoid (Figure S3b). We suggest an intuitive explanation of this effect. High within-species variance leads to many large absolute values of z within a given species. Note that extreme outliers disproportionately influence the mean, or $E[z]$. The sigmoid function applied *before* averaging 'squashes' those extreme values, diminishing their influence. This explains why $s(E[z])$ values tend to be more extreme than $E[s(z)]$). The effect is stable even if no separable classes are present in the dataset. Strikingly, we can easily reproduce the linear (Figure S3c) and sigmoid (Figure S3d) behaviour just by manipulating the within-species variance.

What is more, the effect requires species-level signal to be present, i.e. each species must have its own mean and standard deviation assigned. If the distributions of z are the same for each species, the differences in $E[s(z)]$ are expected to be minimal. However, if the variance of z is extremely high, sampling effects can cause large dispersion of $s(E[z])$. As shown in Figure S3e, 'logit'-averaging indeed causes spurious, high-confidence predictions in this setting (since $\mu_r = 0$, correct predictions should be close to 0.5). Those are most probably artifacts caused by the extreme values of some z outliers. 'Proba'-averaging is more robust in this case and correctly yields low confidence predictions close to 0.5. Note that this can be also understood as the 'logit'-averaging amplifying noise generated by sampling effects. Therefore, the problem can be at least partially mitigated with higher read numbers per genome – if we increase the number of 'reads' per 'species', we see the expected decrease in effect strength (Figure S3f). In general, higher sensitivity of 'logit'-averaging could also be useful in some applications, depending on the levels of different sources of variance in the dataset.

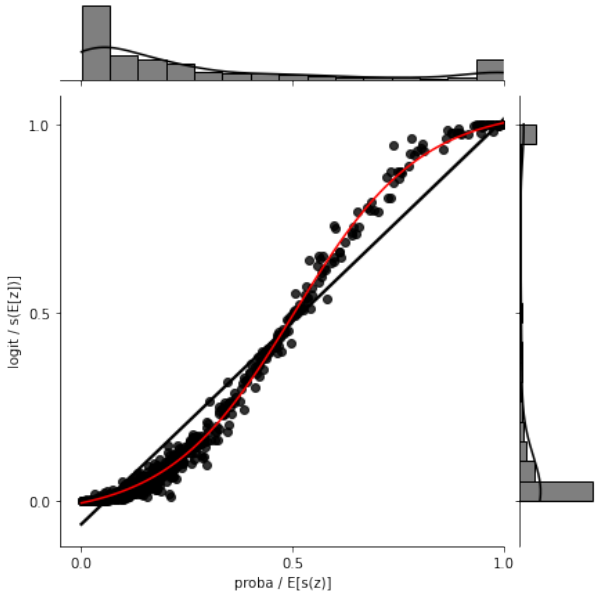
Note that even in the settings with high within-species variance, where the relationship between $s(E[z])$ and $E[s(z)]$ becomes sigmoid, its approximate monotonicity is maintained (Spearman's $\rho > 0.98$, $p < 10^{-15}$ for both Figure S3b and Figure S3d). This suggests that even in this case, $E[\mathbf{h}]$ could potentially be used as genome representations. However, one must exercise caution – the more steep the sigmoid relationship, the more distorted the distances between such representations would become. In our fungal dataset, the within-species variance is seemingly low enough to guarantee a linear relationship, so this problem does not affect it.



(a) Fungal validation dataset

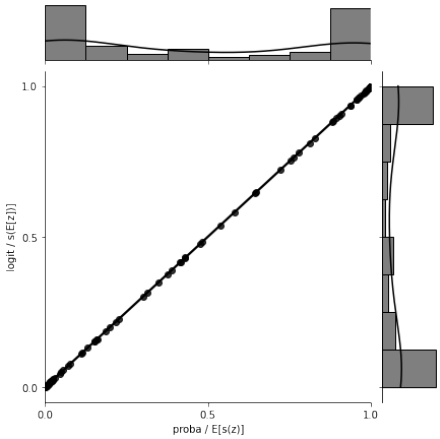


(b) Fungal test dataset

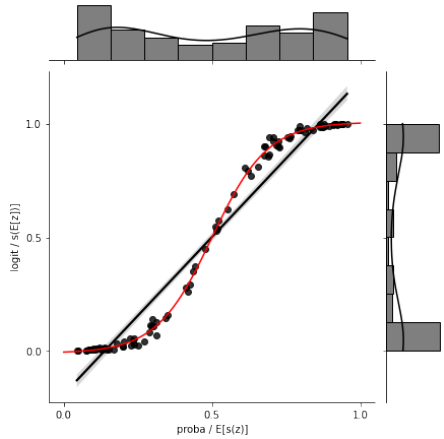


(c) 'Novel viruses' test dataset

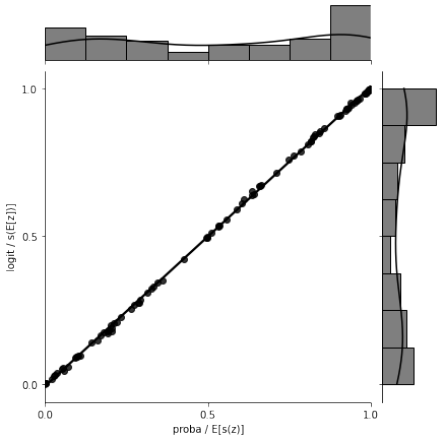
Figure S2: Comparison of relationships between 'logit'-average ($s(E[z])$ and 'proba'-average ($E[s(z)]$) predictions. Real data.



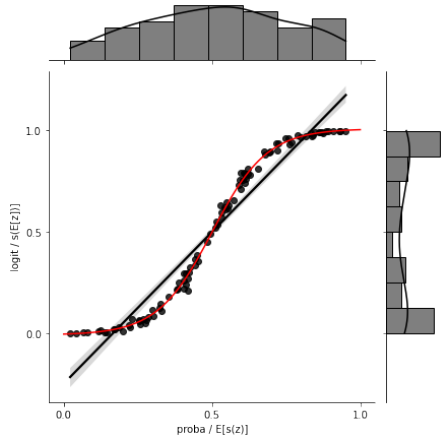
(a) Two classes, well separated, low within-species variances



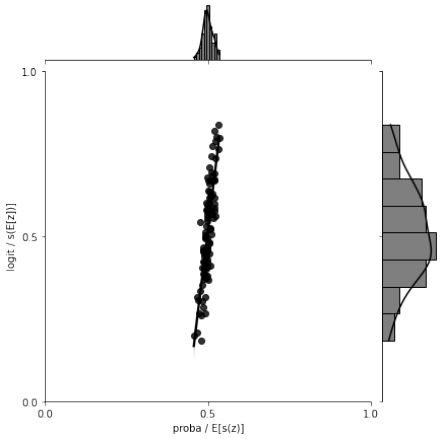
(b) Two classes, well separated, high within-species variances



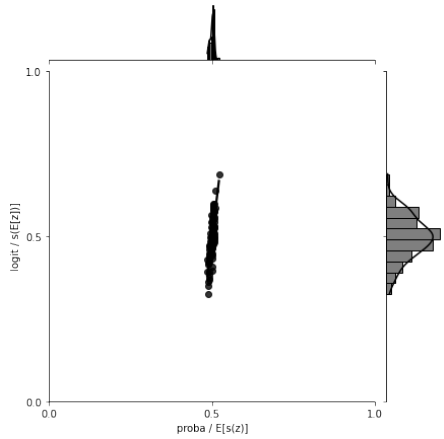
(c) No class distinction, low within-species variances



(d) No class distinction, high within-species variances



(e) No species signal, extremely high per-read variance



(f) No species signal, extremely high per-read variance, increased coverage

Figure S3: Comparison of relationships between 'logit'-average ($s(E[z])$ and 'proba'-average ($E[s(z)]$) predictions. Simulated z values.

Supplementary Note 5: Fungal pathogenic potential prediction

The networks trained without input dropout achieved higher validation accuracy compared to those trained with dropout, and the 'logarithmic-size' training set resulted in higher accuracy than the 'linear-size' set in all cases. As expected, the 'high-coverage' training set also improved performance, although the training time of around 8h45' per epoch on four Tesla V100 GPUs was much higher than around 45 minutes per epoch for the 'low-coverage' variant on the same hardware.

The difference in performance for the first and second mate of the read pairs is negligible; we present the mean values. Interestingly, while BLAST's accuracy on the test set is higher, its accuracy on the validation set is lower by a similar margin (61.6%, compared to 63.3% for the ResNet). While this comparison should not be overinterpreted, as the ResNet epoch maximizing validation accuracy was chosen, those results suggest that the performance differences between the two approaches are small and could be dependent on a particular composition of species in the test set. Although this could be explicitly tested using a nested cross-validation scheme, such a setup would be computationally prohibitive.

We do not observe much overfitting (the single read training accuracy for the selected epoch is 68.9%), which suggests that the unsatisfactory performance is a form of underfitting. More expressive neural architectures could possibly help alleviate this problem, but it is also possible that prediction accuracy known for other pathogen classes is simply impossible for fungi, for example due to the higher complexity of eukaryotic genomes with abundant regions of non-coding DNA. However, we speculate that even those predictions could be useful in some circumstances. High precision of BLAST proves that its positive predictions are trustworthy. On the other hand, the ResNet offers only slightly worse accuracy, but for all reads in the sample (BLAST finds no matches for over 20% of reads even if read pairs are used). What is more, the ResNet can process 1.25 million reads in just 3 minutes on a single GPU, while BLAST needs over three hours and 256 threads on a high-performance computing node for the same task. The ResNet is slower when used on CPUs, but still outperforms BLAST more than sevenfold. This is very important in practice, as screening of large NGS datasets must be performed quickly to remain feasible. For this reason, BLAST is usually too slow for NGS analysis. Note that in this work, it represents the upper bound on accuracy of homology-based approaches; faster alternatives like NGS mappers or k-mer apporaches have been shown to underperform in the pathogenic potential tasks for other pathogen groups [Deneke et al., 2017, Bartoszewicz et al., 2021b].

Interestingly, BLAST's precision is actually a bit lower for full genomes than for reads, while the ResNet becomes more precise when more information (more reads) is considered. Contrary to what was reported for viruses, using raw, unassembled reads improves BLAST's performance compared to relying on assembled contigs. A probable reason is that when a majority vote over all contigs is taken, all contigs have the same voting weight, which means that the ratio of voting weight to sequence length can be much higher for short contigs of potentially lower quality than for the very long ones. In contrast, when reads are used, all regions of the genome weight the same. Using contigs also markedly increases the processing time. This slowdown is probably best explained by the computational difficulty of the local alignment of long contigs compared to a low-coverage set of representative reads.

Both BLAST and the ResNet correctly predict *C. auris* and *A. fumigatus* to be pathogens, even though they were kept in the held-out test dataset unused during training. *P. oryzae* is also recognized as a a member of the negative class. The ResNet (but not BLAST) incorrectly classifies *B. dendrobatidis*, a deadly amphibian pathogen, as able to infect humans. This prediction could be in principle indicative of a hidden zoonotic potential that could be realized under some yet unreported circumstances (e.g. in susceptible individuals or tissues). However, without further evidence, we assume this is simply an error of the model – perfect generalization is rarely possible, even though as noted in Table S3, the prediction performance of the ResNet on full genomes is still high. It is possible that the erroneous prediction for *B. dendrobatidis* is caused by underrepresentation of related species in the training set, as the core database contains only two other species of the same order, *Chytridiomycota*, hindering the classifier from correctly modelling their pathogenic potential.

Table S3: Classification performance comparison between BLAST and the ResNet model on single reads, read pairs and genomes from the test dataset.

		BAcc.	Prec.	Rec.	Spec.	AUC	AUPR	Pred.
Single reads	ResNet	64.7	65.8	**61.3**	68.1	70.9	72.0	**100.0**
	BLAST	**66.2**	**92.6**	61.1	**71.3**	-	-	77.6
Read pairs	ResNet	68.5	69.8	**65.1**	71.9	75.7	76.5	**100.0**
	BLAST	**69.7**	**94.5**	62.3	**77.2**	-	-	79.8
Genomes	ResNet	88.4	77.5	**91.2**	85.7	95.1	90.1	**100.0**
	BLAST (reads)	**90.3**	**90.6**	85.3	**95.2**	-	-	**100.0**
	BLAST (contigs)	89.5	87.8	85.3	93.7	-	-	**100.0**

		Time (CPU)	Time (GPU)
Single reads	ResNet	**25 min.**	**3 min.**
	BLAST	181 min.	-
Read pairs	ResNet	**50 min.**	**6 min.**
	BLAST	362 min.	-
Genomes	ResNet	**25–50 min.**	**3–6 min.**
	BLAST (reads)	181–362 min.	-
	BLAST (contigs)	2190 min.	-

The best performance value for each performance metric (row) and setting is displayed in bold. Read datasets are balanced, and the differences in performance for the first and second mate were negligible for all metrics. In genome datasets, negative class is the majority class. We use BLAST as shown in [Bartoszewicz et al., 2020]; since this only returns binary labels, AUC and AUPR are not defined. Overall, BLAST is much more precise, but the ResNet yields only slightly less accurate predictions for all reads, even those impossible to match with BLAST, in a fraction of the time. We comment more on this fact in Supplementary Note 5. AUC and AUPR are marginally higher (95.2 and 90.2) for the genome-level ResNet if only first mates are used compared to using both mates (95.1 and 90.1). Other metrics are equal, and the computation time is 50% lower. BAcc. – balanced accuracy (equivalent to accuracy on read sets), Prec. – precision; Rec. – recall, Spec. – specificity, AUC – area under the ROC curve, AUPR – area under the PR curve, Pred. – prediction rate; Time (CPU) – prediction time on 2x AMD EPYC 7742 (256 threads); Time (GPU) – prediction time on 1x Tesla V100 GPU (not possible for BLAST).

Table S4: Classification performance comparison between BLAST and the ResNet model on single reads and read pairs from the temporal dataset.

		BAcc.	Prec.	Rec.	Spec.	AUC	AUPR	Pred.
Single reads	ResNet	68.9	69.6	66.9	**70.8**	74.9	74.6	**100.0**
	BLAST	**71.2**	**87.4**	**81.3**	61.2	-	-	82.6
Read pairs	ResNet	**74.2**	74.6	73.4	**75.1**	81.8	82.1	**100.0**
	BLAST	71.9	**87.4**	**77.5**	66.3	-	-	80.1

The best performance value for each performance metric (row) and setting is displayed in bold. While the read number per class is balanced, the dataset is small and extremely imbalanced on the genome level – only a single human pathogen genome and 14 non-human pathogen genomes were added to GenBank in the time window of our temporal benchmark. Therefore, the presented results are not fully representative, and the higher performance of the ResNet on read pairs cannot be guaranteed in general. However, the overall results are similar to those obtained for our main held-out test dataset. BAcc. – balanced accuracy (equivalent to accuracy on read sets), Prec. – precision; Rec. – recall, Spec. – specificity, AUC – area under the ROC curve, AUPR – area under the PR curve, Pred. – prediction rate.

Table S5: Effects of integrating multiple classes on the classification performance on non-fungal datasets, read pairs.

		Acc.	Prec.	Rec.	Spec.	AUPR
Bacteria	4-class ensemble	85.1	**84.1**	87.2	**83.0**	88.3
	4-class	83.6	80.9	89.0	78.1	88.6
	3-class	84.0	80.6	89.7	78.2	**90.0**
	2-class [Bartoszewicz et al., 2021a]	**87.3**	83.6	**92.7**	81.8	89.0
	BLAST	70.0	**84.1**	86.6	53.5	-
Viruses	4-class ensemble	81.4	94.3	90.9	71.8	**98.1**
	4-class	79.0	93.4	**91.2**	66.7	97.9
	3-class	81.9	93.3	91.1	72.7	97.8
	4-class ensemble, reassigned	89.6	95.3	89.3	90.0	**98.1**
	4-class, reassigned	89.6	94.7	89.6	89.7	97.9
	3-class, reassigned	**91.8**	95.1	88.1	**95.4**	97.8
	2-class [Bartoszewicz et al., 2021a]	90.3	94.7	85.5	95.2	97.2
	BLAST	80.6	**98.4**	79.1	82.2	-

The 4-class classifier includes the human-pathogenic fungi class along the three viral and bacterial classes included in the 3-class classifier. The 2-class classifiers are the original DeePaC ResNets [Bartoszewicz et al., 2021a] for bacteria and viruses accordingly. We use BLAST as shown in [Bartoszewicz et al., 2020, 2021b]; since this only returns binary labels, AUPR is not defined. The multi-class models do not assume a purely bacterial or viral sample, so they are more flexible than the 2-class networks – a single classifier can be used for both datasets without retraining, although a minor cost in performance should be expected of a more general approach. For the viral dataset, assignments to the bacterial pathogens class may be assumed to actually reflect bacteriophages infecting those bacteria due to high bacteriophage-host sequence similarity. Classifiers reassigning the bacterial assignments into non-pathogen class based on this assumption are listed as 'reassigned'.

The multi-class models and the specialized binary classifiers perform comparably well, even though the latter can only be used in very specific situations. Integrating fungi into the classifiers does not significantly impact the overall accuracy. The main challenge in multi-class pathogenic potential prediction seems to be actually differentiating between bacteriophages and the pathogenic bacteria that they infect. The multi-class networks confuse 18–21% of viral non-pathogen reads for pathogenic bacteria. However, their accuracy is still similar to BLAST's, even though BLAST was used with a purely viral reference database, and hence would always classify ambiguous bacterial or viral sequences as viral. This problem is unrelated to the fungal database presented here, and can be circumvented by assuming that spurious assignments to bacteria should be reassigned to bacteriophages in the case of a purely viral dataset. The resulting networks achieve accuracy similar to that of the simple binary classifiers, while still being capable of multi-class predictions for more complex datasets. They also outperform BLAST, which represents the estimated upper limit on the accuracy of pathogen detection by finding the closest taxonomic matches, by a wide margin.

Acc. – accuracy, Prec. – precision; Rec. – recall, Spec. – specificity, AUPR – area under the PR curve (calculated for the respective positive class).

Table S6: Effects of integrating multiple classes on the classification performance on the multi-class dataset, read pairs.

		Acc.	F1	Prec.	Rec.	AUPR
All classes	4-class ensemble	87.6	87.7	87.7	87.6	93.4
	4-class	86.6	86.7	86.8	86.6	92.8
	3-class	85.5	85.6	85.8	85.5	92.4
Non-pathogens	4-class ensemble	**77.4**	**78.7**	80.1	**77.4**	86.7
	4-class	72.5	76.7	81.5	72.5	85.3
	3-class	75.5	78.6	**81.8**	75.5	**87.6**
Path. bacteria	4-class ensemble	87.2	**85.1**	**83.2**	87.2	90.4
	4-class	89.0	83.8	79.1	89.0	90.4
	3-class	**89.7**	84.2	79.3	**89.7**	**91.1**
Human viruses	4-class ensemble	90.9	**93.7**	**96.7**	90.9	98.4
	4-class	**91.2**	93.4	95.7	**91.2**	98.1
	3-class	91.1	93.6	96.3	91.1	**98.5**
Fungi	4-class ensemble	**95.0**	**92.9**	**90.9**	**95.0**	**97.9**
	4-class	93.7	92.2	90.7	93.7	97.4

The 4-class classifier includes the fungi class along the three viral and bacterial classes included in the 3-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 individual class. We use the F1 score as an additional measure. Integrating fungi into the classifiers does significantly impact the overall accuracy for any of the non-fungal classes, and the 4-class ensemble offers a more balanced performance on putative non-pathogen reads than the single-network alternative. The metrics labelled as 'All classes' were calculated for all four classes for the 4-class models, and for the three relevant classes for the 3-class model. Acc. – accuracy, F1 – F1 score, Prec. – precision, Rec. – recall, AUPR – area under the PR curve.

Supplementary Note 6: Genome embeddings

The last convolutional layer of our fungal ResNet has 512 filters, but we removed the filters with activations equal to zero for all samples in the dataset. We then used UMAP [McInnes et al., 2020] to embed the resulting representations in a 2-dimensional space for visualization. We used the Euclidean distance metric with a minimum possible distance of 0.1 and the neighbourhood size of 15. The inputs had been randomly shuffled beforehand to avoid artifacts that can appear if an embedding is learned based on representations ordered by class.

The number of units at most levels of taxonomy poses a challenge for intuitive visualization of the dataset structure. Nevertheless, some correspondence between cluster membership, label and the taxonomic rank of a class is clearly visible (Figure S7; note that the 'class' as a taxonomic term is distinct from the concept of a positive or negative class in machine learning). This extends also to the lower and higher taxonomic ranks of order and phylum, respectively (Figure S8-Figure S9).

Some orders of one class may belong to different clusters. For example, the orders *Chaetothyriales*, *Eurotiales* (including the genus *Aspergillus*) and *Onygenales* all belong to the class *Eurotiomycetes* (Figure S7, in red), but are members of four different clusters, with *Eurotiales* split into two (Figure S8). Even more strikingly, the leftmost cluster contains at least one member of all six phyla, with two most noticeable groups corresponding to the orders *Saccharomycetales* and *Mucorales* in phyla *Ascomycota* and *Mucormycota*, respectively (Figure S7-Figure S9). 89% of the members of both those orders are at least opportunistic human pathogens; this includes *Candida* species in *Saccharomycetales* and causative agents of mucormycosis, known also as the 'black fungus', in *Mucorales*. Therefore, we hypothesize that the grouping represents a cluster of 'conserved' potential pathogens, relatively easy to classify with an acceptable accuracy. A common signal at the functional level or technical artifacts could be alternative explanations; they are however less likely due to large phylogenetic distance on one side, and a consistent grouping of almost all species from those well-represented orders on the other.

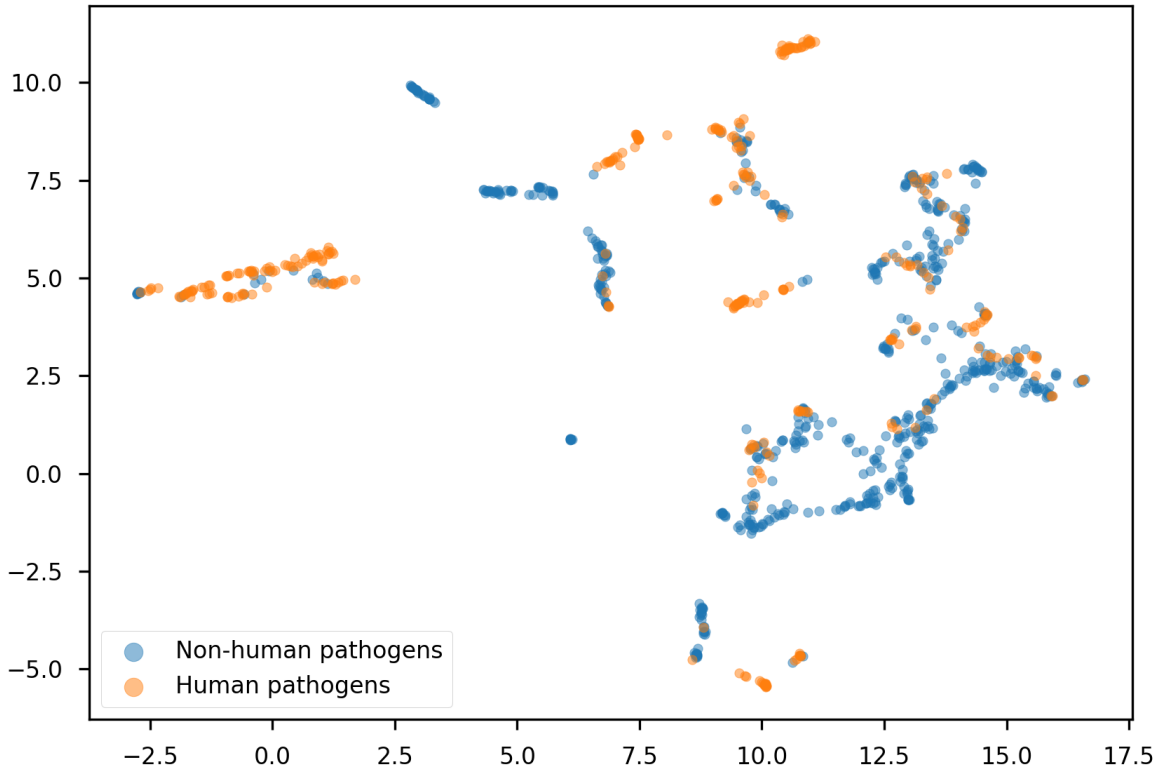


Figure S4: UMAP embeddings of the learned genome representations for the core database; true labels.

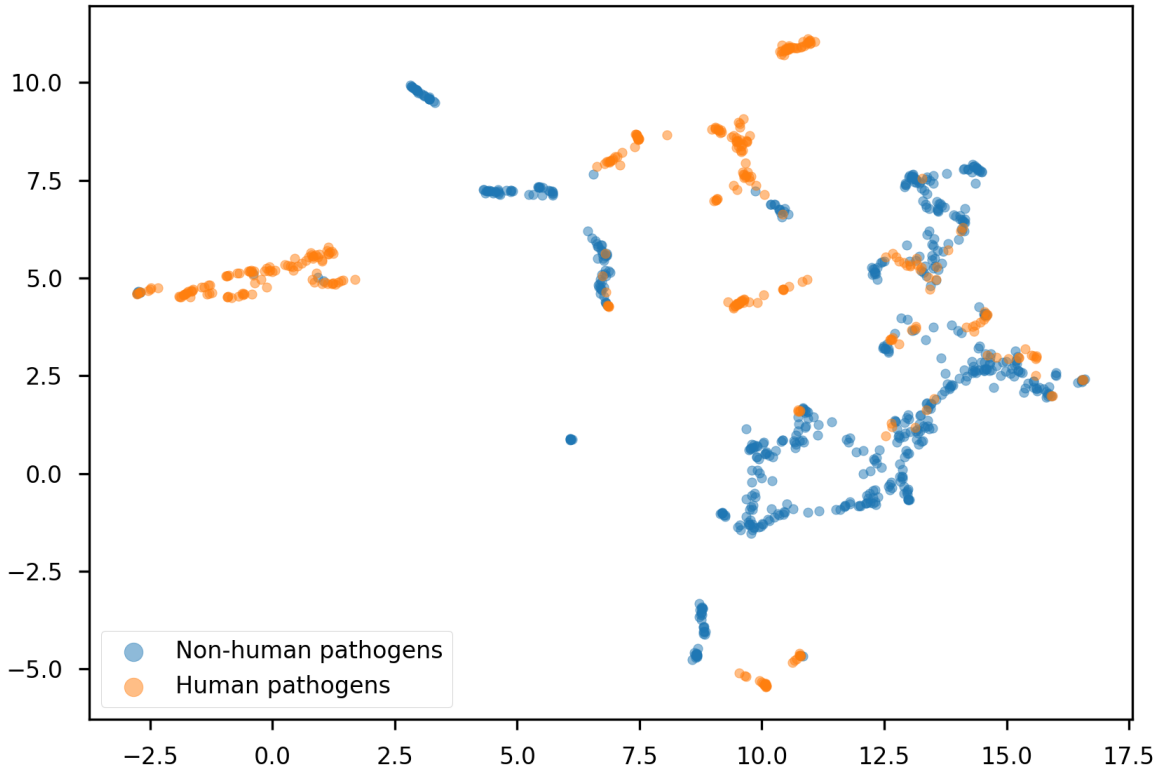


Figure S5: UMAP embeddings of the learned genome representations for the core database; labels predicted by the ResNet (retuned threshold). Most 'positive' members of otherwise 'negative' clusters are correctly detected.

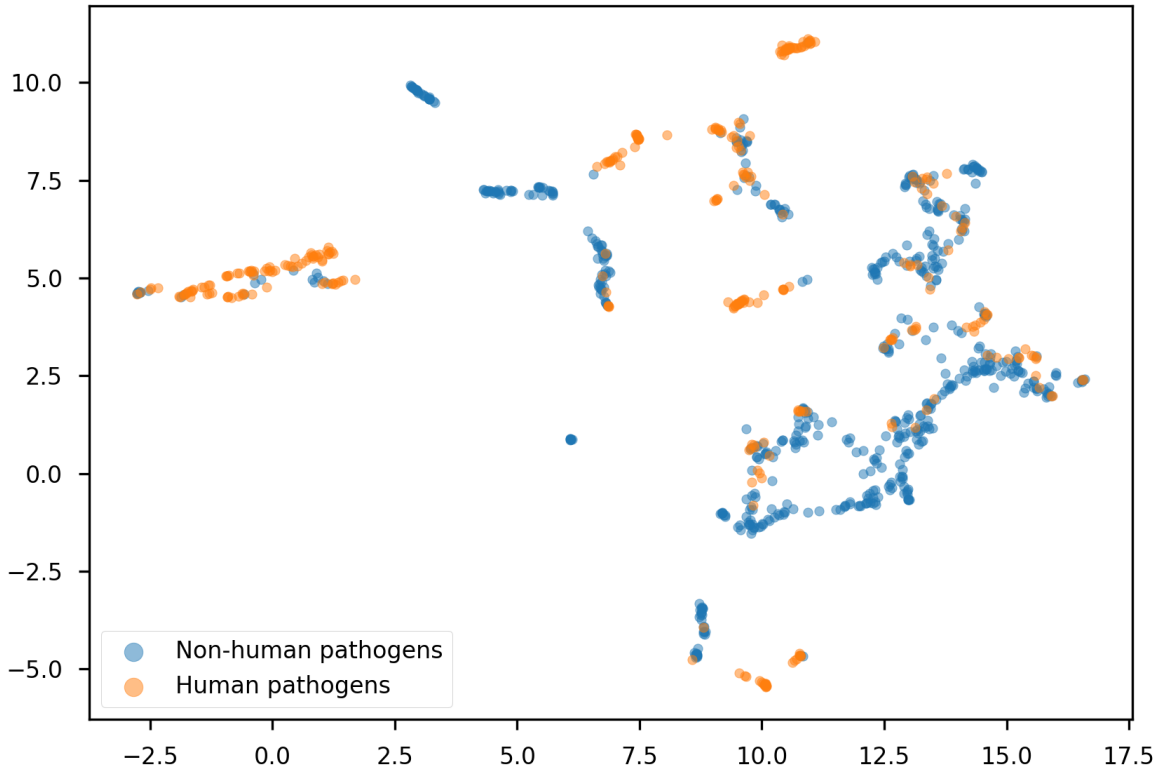


Figure S6: UMAP embeddings of the learned genome representations for the core database; labels predicted by BLAST. Training species are present in the reference database, so predicting labels for them is easy (99.9% accuracy).

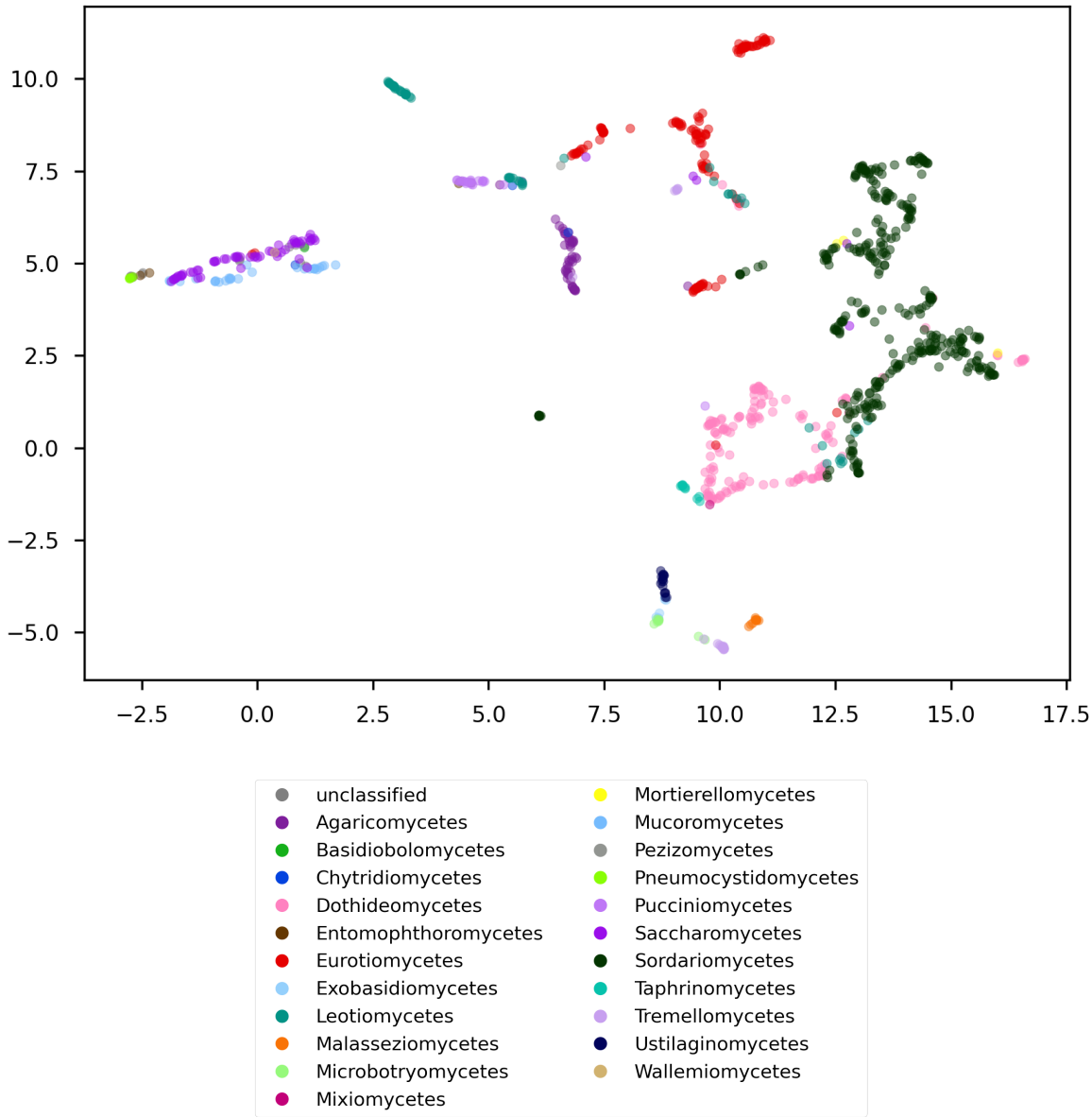


Figure S7: UMAP embeddings of the learned genome representations for the core database; taxonomic rank: class. In general, related species are close to each other in this space, but some taxonomic units are distributed among more than one cluster (e.g. *Eurotiomycetes*, in red), and some clusters contain members of distant taxa (e.g. the leftmost cluster).

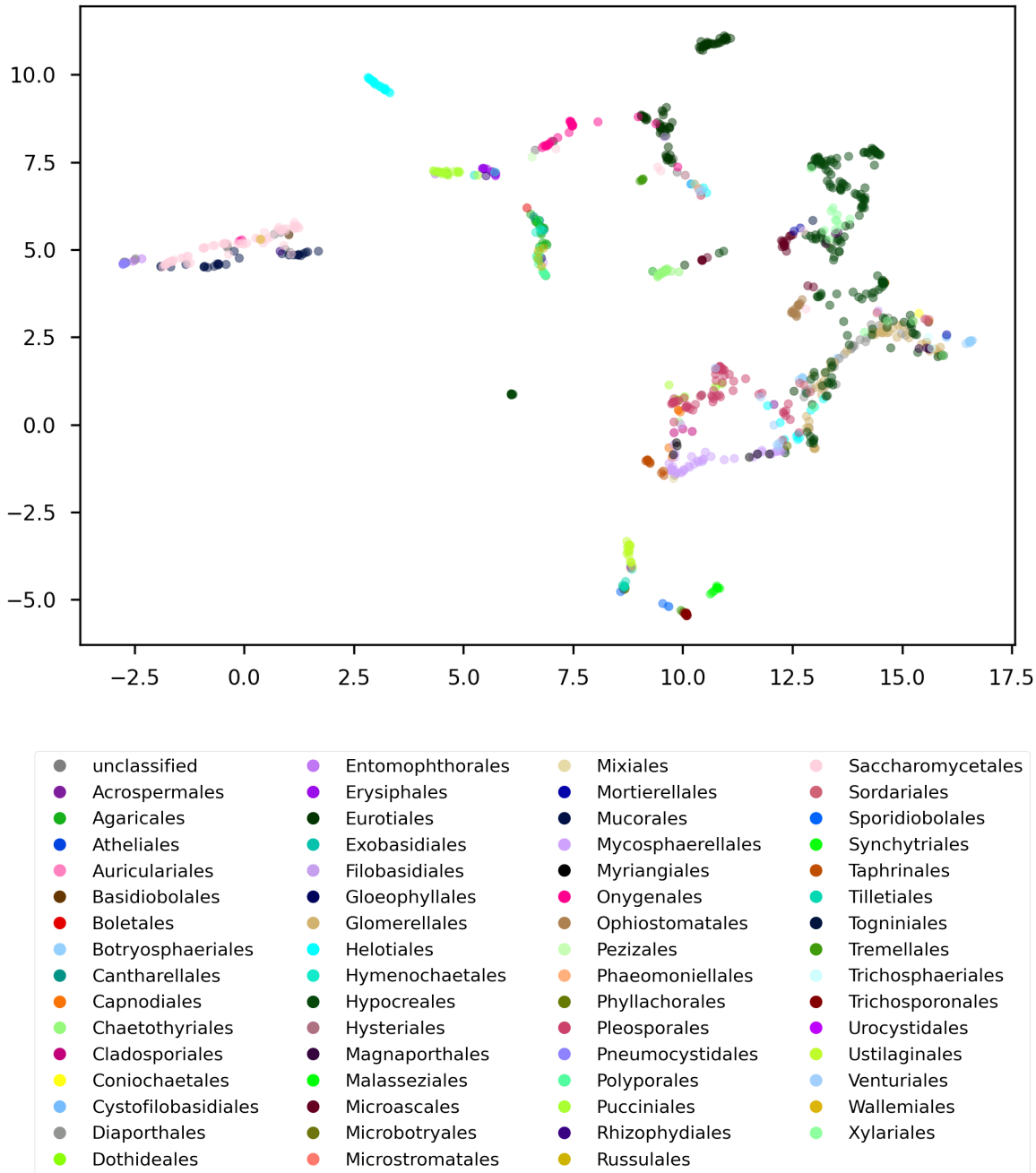


Figure S8: UMAP embeddings of the learned genome representations for the core database; taxonomic rank: order.

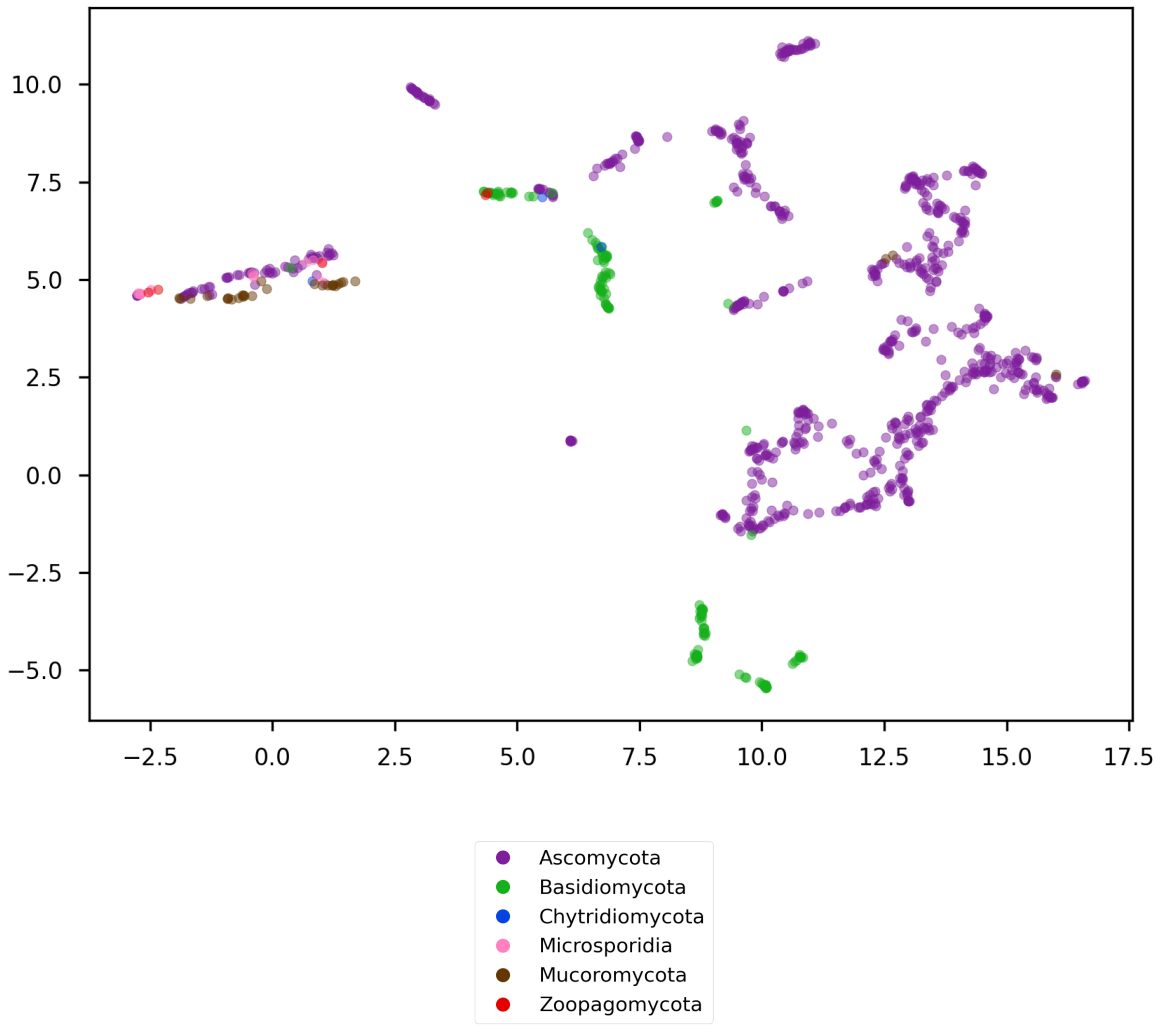


Figure S9: UMAP embeddings of the learned genome representations for the core database; taxonomic rank: phylum.

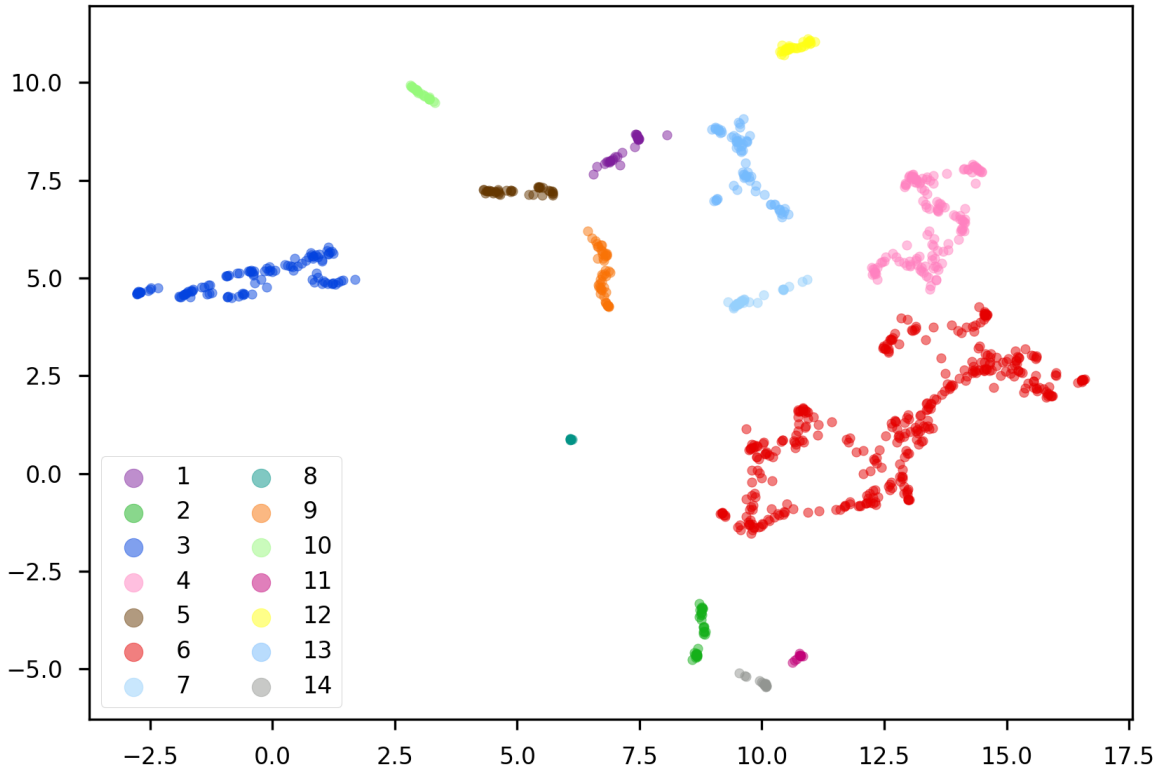


Figure S10: UMAP embeddings of the learned genome representations for the core database. Assignment to clusters detected with single-linkage agglomerative clustering. Clusters numbered arbitrarily.

Table S7: *C. auris* sequencing run, SRA accession SRR17577041. Binary classifiers (host group prediction).

		Acc.	Rec.	Pred.
First read	2-class (ours)	**77.4**	**77.4**	**100.0**
	BLAST	46.9	46.9	52.0
	STAT	0.1	0.1	7.7
Both reads	2-class (ours)	**78.4**	**78.4**	**100.0**
	BLAST	50.1	50.1	56.9
	STAT	0.1	0.1	5.4

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. Interestingly, 2-class ResNet performance for this species is above average (see Table S3). Perhaps this is because *Candida* species are placed in a cluster of conserved pathogens (see Supp. Note 6). Acc. – accuracy, Rec. – recall, Pred. – prediction rate (a fraction of reads with any hits).

Supplementary Note 7: Performance on real data

Obtaining ground truth for genuine metagenomic samples is very challenging without spiking them with known organisms or corresponding reads. As our test sets consist of reads simulated from a mixture of different species (and viruses), they effectively function as mock metagenomic samples. Simulated reads are commonly used to evaluate computational approaches for read classification (e.g. [Wood et al., 2019, Katz et al., 2021]. To show that the methods discussed here work also on real data, we focus on a single-species sequencing run, where ground truth is available. We selected a *C. auris* sequencing run (SRA accession SRR17577041) and analyzed it in two scenarios: a) assuming it is a novel fungal species, predict whether it could infect humans (binary classification, as in Table S3-Table S4) b) assuming it is an unknown agent originating from a patient, predict its correct pathogen group (multi-class setting, as in Table 1 and Table S6). In addition to the corresponding ResNets and BLAST databases, we also used STAT [Katz et al., 2021], a k-mer tool used by the NCBI to generate taxonomy reports for all SRA archives.

To evaluate STAT, we downloaded its reference database published along the paper and run a *C. auris* strain exclusion study as described by the authors. In the host prediction setting Table S7, we counted all STAT species-level hits to any pathogenic fungal species as positives, but in the multi-class pathogen detection task (Table 2), we assumed that all reads classified as fungi can be counted as successful detection. As STAT has no way of jointly classifying read pairs, we treated both mates as independent reads also when comparing to BLAST and ResNet predictions for read pairs. Note that in contrast to our BLAST evalutation, STAT reference database contained 248,426 TaxIds (see Katz et al. [2021]), and only *C. auris* was effectively removed from it for the strain exclusion study.

References

S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, Oct. 1990. ISSN 0022-2836. doi: 10.1016/S0022-2836(05)80360-2.

E. Barash, N. Sal-Man, S. Sabato, and M. Ziv-Ukelson. BacPaCS—Bacterial Pathogenicity Classification via Sparse-SVM. *Bioinformatics*, 2018. doi: 10.1093/bioinformatics/bty928. URL `https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/bty928/5165373`.

D. J. Barbara and E. Clewes. Plant pathogenic Verticillium species: how many of them are there? *Molecular Plant Pathology*, 4(4):297–305, 2003. ISSN 1364-3703. doi: 10.1046/j.1364-3703.2003.00172.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1364-3703.2003.00172.x`.

J. M. Bartoszewicz, A. Seidel, R. Rentzsch, and B. Y. Renard. DeePaC: predicting pathogenic potential of novel DNA with reverse-complement neural networks. *Bioinformatics*, 36(1):81–89, Jan. 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz541. URL `https://doi.org/10.1093/bioinformatics/btz541`.

J. M. Bartoszewicz, U. Genske, and B. Y. Renard. Deep learning-based real-time detection of novel pathogens during sequencing. *Briefings in Bioinformatics*, (bbab269), July 2021a. ISSN 1477-4054. doi: 10.1093/bib/bbab269. URL `https://doi.org/10.1093/bib/bbab269`.

J. M. Bartoszewicz, A. Seidel, and B. Y. Renard. Interpretable detection of novel human viruses from genome sequencing data. *NAR Genomics and Bioinformatics*, 3(1), Mar. 2021b. ISSN 2631-9268. doi: 10.1093/nargab/lqab004. URL `https://doi.org/10.1093/nargab/lqab004`.

S. Bałazy, R. Miętkiewski, C. Tkaczuk, R. Wegensteiner, and M. Wrzosek. Diversity of acaropathogenic fungi in Poland and other European countries. *Experimental and Applied Acarology*, 46(1):53–70, Dec. 2008. ISSN 1572-9702. doi: 10.1007/s10493-008-9207-1. URL `https://doi.org/10.1007/s10493-008-9207-1`.

J. J. Becnel and T. G. Andreadis. Microsporidia in Insects. In *Microsporidia*, pages 521–570. John Wiley & Sons, Ltd, 2014. ISBN 978-1-118-39526-4. doi: 10.1002/9781118395264.ch21. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118395264.ch21`.

A. Bombassaro, G. X. Schneider, F. F. Costa, A. C. R. Leão, B. S. Soley, F. Medeiros, N. M. da Silva, B. J. F. S. Lima, R. J. A. Castro, A. L. Bocca, V. A. Baura, E. Balsanelli, V. C. S. Pankievicz, N. M. C. Hrysay, R. H. Scola, L. F. Moreno, C. M. P. S. Azevedo, E. M. Souza, R. R. Gomes, S. de Hoog, and V. A. Vicente. Genomics and Virulence of Fonsecaea pugnacius, Agent of Disseminated Chromoblastomycosis. *Frontiers in Genetics*, 11:822, 2020. ISSN 1664-8021. doi: 10.3389/fgene.2020.00822. URL `https://www.frontiersin.org/article/10.3389/fgene.2020.00822`.

G. D. Brown, D. W. Denning, N. A. R. Gow, S. M. Levitz, M. G. Netea, and T. C. White. Hidden killers: Human fungal infections. *Science Translational Medicine*, 4(165), 2012. doi: 10.1126/scitranslmed.3004404.

C. Brunner-Mendoza, M. d. R. Reyes-Montes, S. Moonjely, M. J. Bidochka, and C. Toriello. A review on the genus Metarhizium as an entomopathogenic microbial biocontrol agent with emphasis on its use and utility in Mexico. *Biocontrol Science and Technology*, 29(1):83–102, Jan. 2019. ISSN 0958-3157. doi: 10.1080/09583157.2018.1531111. URL `https://doi.org/10.1080/09583157.2018.1531111`.

C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden. BLAST+: architecture and applications. *BMC Bioinformatics*, 10(1):421, Dec. 2009. ISSN 1471-2105. doi: 10.1186/1471-2105-10-421. URL `https://doi.org/10.1186/1471-2105-10-421`.

CDC. *Antibiotic resistance threats in the United States, 2019*. U.S. Department of Health and Human Services, CDC, 2019. URL https://stacks.cdc.gov/view/cdc/82532.

D. Chandler, G. Davidson, J. K. Pell, B. V. Ball, K. Shaw, and K. D. Sunderland. Fungal Biocontrol of Acari. *Biocontrol Science and Technology*, 10(4):357–384, Aug. 2000. ISSN 0958-3157. doi: 10.1080/09583150050114972. URL https://doi.org/10.1080/09583150050114972.

A. Chowdhary, S. Kathuria, K. Agarwal, and J. F. Meis. Recognizing filamentous basidiomycetes as agents of human disease: A review. *Medical Mycology*, 52(8):782–797, Nov. 2014. ISSN 1369-3786. doi: 10.1093/mmy/myu047. URL https://doi.org/10.1093/mmy/myu047.

O. H. Cissé, L. Ma, J. P. Dekker, P. P. Khil, J.-H. Youn, J. M. Brenchley, R. Blair, B. Pahar, M. Chabé, K. K. A. Van Rompay, R. Keesler, A. Sukura, V. Hirsch, G. Kutty, Y. Liu, L. Peng, J. Chen, J. Song, C. Weissenbacher-Lang, J. Xu, N. S. Upham, J. E. Stajich, C. A. Cuomo, M. T. Cushion, and J. A. Kovacs. Genomic insights into the host specific adaptation of the Pneumocystis genus. *Communications Biology*, 4(1):1–14, Mar. 2021. ISSN 2399-3642. doi: 10.1038/s42003-021-01799-7. URL https://www.nature.com/articles/s42003-021-01799-7.

A. L. Colombo, A. C. B. Padovan, and G. M. Chaves. Current knowledge of Trichosporon spp. and Trichosporonosis. *Clinical Microbiology Reviews*, 24(4):682–700, Oct. 2011. ISSN 1098-6618. doi: 10.1128/CMR.00003-11.

M. M. Costa, B. A. A. S. Silva, G. M. Moreira, and L. H. Pfenning. Colletotrichum falcatum and Fusarium species induce symptoms of red rot in sugarcane in Brazil. *Plant Pathology*, 70(8):1807–1818, 2021. ISSN 1365-3059. doi: 10.1111/ppa.13423. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ppa.13423.

I. B. L. Coutinho, F. C. O. Freire, C. S. Lima, J. S. Lima, F. J. T. Gonçalves, A. R. Machado, A. M. S. Silva, and J. E. Cardoso. Diversity of genus Lasiodiplodia associated with perennial tropical fruit plants in northeastern Brazil. *Plant Pathology*, 66(1):90–104, 2017. ISSN 1365-3059. doi: 10.1111/ppa.12565. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ppa.12565.

G. de Hoog, J. Guarro, J. Gené, S. Ahmed, A. Al-Hatmi, M. Figueras, and R. Vitale. *Atlas of Clinical Fungi, 4th edition*. Hilversum, 2020.

R. Dean, J. A. L. Van Kan, Z. A. Pretorius, K. E. Hammond-Kosack, A. Di PIEietro, P. D. Spanu, J. J. Rudd, M. Dickman, R. Kahmann, J. Ellis, and G. D. Foster. The Top 10 fungal pathogens in molecular plant pathology. *Molecular Plant Pathology*, 13(4):414–430, Apr. 2012. ISSN 1464-6722. doi: 10.1111/j.1364-3703.2011.00783.x. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6638784/.

S. Dellière, O. Rivero-Menendez, C. Gautier, D. Garcia-Hermoso, A. Alastruey-Izquierdo, and A. Alanio. Emerging mould infections: Get prepared to meet unexpected fungi in your patient. *Medical Mycology*, 58(2):156–162, Feb. 2020. ISSN 1369-3786. doi: 10.1093/mmy/myz039. URL https://doi.org/10.1093/mmy/myz039.

C. Deneke, R. Rentzsch, and B. Y. Renard. PaPrBaG: A machine learning approach for the detection of novel pathogens from NGS data. *Scientific Reports*, 7:39194, 2017. ISSN 2045-2322. doi: 10.1038/srep39194. URL https://www.nature.com/articles/srep39194.

G. Doehlemann, B. Ökmen, W. Zhu, and A. Sharon. Plant Pathogenic Fungi. *Microbiology Spectrum*, 5(1), Jan. 2017. ISSN 2165-0497. doi: 10.1128/microbiolspec.FUNK-0023-2016.

H. C. Evans, S. L. Elliot, and D. P. Hughes. Hidden Diversity Behind the Zombie-Ant Fungus Ophiocordyceps unilateralis: Four New Species Described from Carpenter Ants in Minas Gerais, Brazil. *PLOS ONE*, 6(3):e17024, Mar. 2011. ISSN 1932-6203. doi: 10.1371/journal.pone.0017024. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0017024.

D. F. Farr and A. Y. Rossman. Fungal databases, 2021. URL https://nt.ars-grin.gov/fungaldatabases/. Retrieved October 9, 2021 and January 2, 2022.

C. Franzen and A. Müller. Microsporidiosis: human diseases and diagnosis. *Microbes and Infection*, 3(5):389–400, Apr. 2001. ISSN 1286-4579. doi: 10.1016/s1286-4579(01)01395-8.

U. Gerson, A. Gafni, Z. Paz, and A. Sztejnberg. A tale of three acaropathogenic fungi in Israel: Hirsutella, Meira and Acaromyces. *Experimental and Applied Acarology*, 46(1):183–194, Dec. 2008. ISSN 1572-9702. doi: 10.1007/s10493-008-9202-6. URL https://doi.org/10.1007/s10493-008-9202-6.

S. Gurung, D. P. G. Short, X. Hu, G. V. Sandoya, R. J. Hayes, S. T. Koike, and K. V. Subbarao. Host Range of Verticillium isaacii and Verticillium klebahnii from Artichoke, Spinach, and Lettuce. *Plant Disease*, 99(7):933–938, July 2015. ISSN 0191-2917. doi: 10.1094/PDIS-12-14-1307-RE.

B. Han and L. M. Weiss. Microsporidia: Obligate Intracellular Pathogens Within the Fungal Kingdom. *Microbiology Spectrum*, 5(2), Apr. 2017. ISSN 2165-0497. doi: 10.1128/microbiolspec.FUNK-0018-2016.

M. Holtgrewe. Mason – a read simulator for second generation sequencing data. *Technical Report FU Berlin*, 2010. URL http://publications.imp.fu-berlin.de/962/.

W. Hu, Y. Ran, K. Zhuang, J. Lama, and C. Zhang. Alternaria arborescens Infection in a Healthy Individual and Literature Review of Cutaneous Alternariosis. *Mycopathologia*, 179(1):147–152, Feb. 2015. ISSN 1573-0832. doi: 10.1007/s11046-014-9822-9. URL https://doi.org/10.1007/s11046-014-9822-9.

S. Jančič, H. D. T. Nguyen, J. C. Frisvad, P. Zalar, H.-J. Schroers, K. A. Seifert, and N. Gunde-Cimerman. A Taxonomic Revision of the Wallemia sebi Species Complex. *PLOS ONE*, 10(5):e0125933, May 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0125933. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0125933.

D. R. Jones and R. H. A. Baker. Introductions of non-native plant pathogens into Great Britain, 1970–2004. *Plant Pathology*, 56(5):891–910, 2007. ISSN 1365-3059. doi: 10.1111/j.1365-3059.2007.01619.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-3059.2007.01619.x.

K. S. Katz, O. Shutov, R. Lapoint, M. Kimelman, J. R. Brister, and C. O'Sullivan. STAT: a fast, scalable, MinHash-based k-mer tool to assess Sequence Read Archive next-generation sequence submissions. *Genome Biology*, 22 (1):270, Sept. 2021. ISSN 1474-760X. doi: 10.1186/s13059-021-02490-0. URL https://doi.org/10.1186/s13059-021-02490-0.

S. W. Kim, J. H. Jung, K. Lamsal, Y. S. Kim, J. S. Min, and Y. S. Lee. Antifungal Effects of Silver Nanoparticles (AgNPs) against Various Plant Pathogenic Fungi. *Mycobiology*, 40(1):53–58, Mar. 2012. ISSN 1229-8093. doi: 10.5941/MYCO.2012.40.1.053. URL https://doi.org/10.5941/MYCO.2012.40.1.053.

W. Kim, B. Cavinder, R. H. Proctor, K. O'Donnell, J. P. Townsend, and F. Trail. Comparative Genomics and Transcriptomics During Sexual Development Gives Insight Into the Life History of the Cosmopolitan Fungus Fusarium neocosmosporiellum. *Frontiers in Microbiology*, 10:1247, 2019. ISSN 1664-302X. doi: 10.3389/fmicb.2019.01247. URL https://www.frontiersin.org/article/10.3389/fmicb.2019.01247.

I. Kjærbølling, T. C. Vesth, J. C. Frisvad, J. L. Nybo, S. Theobald, A. Kuo, P. Bowyer, Y. Matsuda, S. Mondo, E. K. Lyhne, M. E. Kogle, A. Clum, A. Lipzen, A. Salamov, C. Y. Ngan, C. Daum, J. Chiniquy, K. Barry, K. LaButti, S. Haridas, B. A. Simmons, J. K. Magnuson, U. H. Mortensen, T. O. Larsen, I. V. Grigoriev, S. E. Baker, and M. R. Andersen. Linking secondary metabolites to gene clusters through genome sequencing of six diverse Aspergillus species. *Proceedings of the National Academy of Sciences*, 115(4):E753–E761, Jan. 2018. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1715954115. URL https://www.pnas.org/content/115/4/E753.

K. J. Kwon-Chung, J. E. Bennett, B. L. Wickes, W. Meyer, C. A. Cuomo, K. R. Wollenburg, T. A. Bicanic, E. Castañeda, Y. C. Chang, J. Chen, M. Cogliati, F. Dromer, D. Ellis, S. G. Filler, M. C. Fisher, T. S. Harrison, S. M. Holland, S. Kohno, J. W. Kronstad, M. Lazera, S. M. Levitz, M. S. Lionakis, R. C. May, P. Ngamskulrongroj, P. G. Pappas, J. R. Perfect, V. Rickerts, T. C. Sorrell, T. J. Walsh, P. R. Williamson, J. Xu, A. M. Zelazny, and A. Casadevall. The Case for Adopting the "Species Complex" Nomenclature for the Etiologic Agents of Cryptococcosis. *mSphere*, 2(1):e00357–16, Feb. 2017. ISSN 2379-5042. doi: 10.1128/mSphere.00357-16.

J. R. Köhler, A. Casadevall, and J. Perfect. The Spectrum of Fungi That Infects Humans. *Cold Spring Harbor Perspectives in Medicine*, 5(1):a019273, Jan. 2015. ISSN , 2157-1422. doi: 10.1101/cshperspect.a019273. URL `http://perspectivesinmedicine.cshlp.org/content/5/1/a019273`.

S. Leonhardt, E. Büttner, A. M. Gebauer, M. Hofrichter, and H. Kellner. Draft Genome Sequence of the Sordariomycete Lecythophora (Coniochaeta) hoffmannii CBS 245.38. *Genome Announcements*, 6(7):e01510–17, 2018. doi: 10.1128/genomeA.01510-17. URL `https://journals.asm.org/doi/10.1128/genomeA.01510-17`.

B. Lovett and R. J. St. Leger. The Insect Pathogens. *Microbiology Spectrum*, 5(2):5.2.01, Mar. 2017. doi: 10.1128/microbiolspec.FUNK-0001-2016. URL `https://journals.asm.org/doi/full/10.1128/microbiolspec.FUNK-0001-2016`.

T. Lu, B. Yao, and C. Zhang. DFVF: database of fungal virulence factors. *Database*, 2012(bas032), Jan. 2012. ISSN 1758-0463. doi: 10.1093/database/bas032. URL `https://doi.org/10.1093/database/bas032`.

L. McInnes, J. Healy, and J. Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv:1802.03426 [cs, stat]*, Sept. 2020. URL `http://arxiv.org/abs/1802.03426`. arXiv: 1802.03426.

C. Paulussen, J. E. Hallsworth, S. Álvarez Pérez, W. C. Nierman, P. G. Hamill, D. Blain, H. Rediers, and B. Lievens. Ecology of aspergillosis: insights into the pathogenic potency of Aspergillus fumigatus and some other Aspergillus species. *Microbial Biotechnology*, 10(2):296–322, 2017. ISSN 1751-7915. doi: 10.1111/1751-7915.12367. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/1751-7915.12367`.

P. Y. Prakash, L. Irinyi, C. Halliday, S. Chen, V. Robert, and W. Meyer. Online Databases for Taxonomy and Identification of Pathogenic Fungi and Proposal for a Cloud-Based Dynamic Data Network Platform. *Journal of Clinical Microbiology*, 55(4):1011–1024, Apr. 2017. ISSN 1098-660X. doi: 10.1128/JCM.02084-16.

S. A. Rehner and E. Buckley. A Beauveria phylogeny inferred from nuclear ITS and EF1-alpha sequences: evidence for cryptic diversification and links to Cordyceps teleomorphs. *Mycologia*, 97(1):84–98, Feb. 2005. ISSN 0027-5514. doi: 10.3852/mycologia.97.1.84.

K. Satoh, K. Makimura, Y. Hasumi, Y. Nishiyama, K. Uchida, and H. Yamaguchi. Candida auris sp. nov., a novel ascomycetous yeast isolated from the external ear canal of an inpatient in a Japanese hospital. *Microbiology and Immunology*, 53(1):41–44, 2009. ISSN 1348-0421. doi: 10.1111/j.1348-0421.2008.00083.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1348-0421.2008.00083.x`.

C. L. Schardl, C. A. Young, U. Hesse, S. G. Amyotte, K. Andreeva, P. J. Calie, D. J. Fleetwood, D. C. Haws, N. Moore, B. Oeser, D. G. Panaccione, K. K. Schweri, C. R. Voisey, M. L. Farman, J. W. Jaromczyk, B. A. Roe, D. M. O'Sullivan, B. Scott, P. Tudzynski, Z. An, E. G. Arnaoudova, C. T. Bullock, N. D. Charlton, L. Chen, M. Cox, R. D. Dinkins, S. Florea, A. E. Glenn, A. Gordon, U. Güldener, D. R. Harris, W. Hollin, J. Jaromczyk, R. D. Johnson, A. K. Khan, E. Leistner, A. Leuchtmann, C. Li, J. Liu, J. Liu, M. Liu, W. Mace, C. Machado, P. Nagabhyru, J. Pan, J. Schmid, K. Sugawara, U. Steiner, J. E. Takach, E. Tanaka, J. S. Webb, E. V. Wilson, J. L. Wiseman, R. Yoshida, and Z. Zeng. Plant-Symbiotic Fungi as Chemical Engineers: Multi-Genome Analysis of the Clavicipitaceae Reveals Dynamics of Alkaloid Loci. *PLOS Genetics*, 9(2):e1003323, Feb. 2013. ISSN 1553-7404. doi: 10.1371/journal.pgen.1003323. URL `https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1003323`.

B. C. Scheele, F. Pasmans, L. F. Skerratt, L. Berger, A. Martel, W. Beukema, A. A. Acevedo, P. A. Burrowes, T. Carvalho, A. Catenazzi, I. De la Riva, M. C. Fisher, S. V. Flechas, C. N. Foster, P. Frías-Álvarez, T. W. J. Garner, B. Gratwicke, J. M. Guayasamin, M. Hirschfeld, J. E. Kolby, T. A. Kosch, E. La Marca, D. B. Lindenmayer, K. R. Lips, A. V. Longo, R. Maneyro, C. A. McDonald, J. Mendelson, P. Palacios-Rodriguez, G. Parra-Olea, C. L. Richards-Zawacki, M.-O. Rödel, S. M. Rovito, C. Soto-Azat, L. F. Toledo, J. Voyles, C. Weldon, S. M. Whitfield, M. Wilkinson, K. R. Zamudio, and S. Canessa. Amphibian fungal panzootic causes catastrophic and ongoing loss of biodiversity. *Science*, 363(6434):1459–1463, Mar. 2019. doi: 10.1126/science.aav0379. URL `https://www.science.org/doi/10.1126/science.aav0379`.

C. L. Schoch, S. Ciufo, M. Domrachev, C. L. Hotton, S. Kannan, R. Khovanskaya, D. Leipe, R. Mcveigh, K. O'Neill, B. Robbertse, et al. NCBI taxonomy: a comprehensive update on curation, resources and tools. *Database*, 2020, 2020.

H. Schulenburg and M.-A. Félix. The Natural Biotic Environment of Caenorhabditis elegans. *Genetics*, 206(1):55–86, May 2017. ISSN 1943-2631. doi: 10.1534/genetics.116.195511. URL `https://doi.org/10.1534/genetics.116.195511`.

S. Seyedmousavi, J. Guillot, P. Arné, G. S. de Hoog, J. W. Mouton, W. J. G. Melchers, and P. E. Verweij. Aspergillus and aspergilloses in wild and domestic animals: a global health concern with parallels to human disease. *Medical Mycology*, 53(8):765–797, Nov. 2015. ISSN 1369-3786. doi: 10.1093/mmy/myv067. URL `https://doi.org/10.1093/mmy/myv067`.

S. Seyedmousavi, S. d. M. G. Bosco, S. de Hoog, F. Ebel, D. Elad, R. R. Gomes, I. D. Jacobsen, H. E. Jensen, A. Martel, B. Mignon, F. Pasmans, E. Piecková, A. M. Rodrigues, K. Singh, V. A. Vicente, G. Wibbelt, N. P. Wiederhold, and J. Guillot. Fungal infections in animals: a patchwork of different situations. *Medical Mycology*, 56(suppl_1):165–187, Apr. 2018. ISSN 1460-2709. doi: 10.1093/mmy/myx104.

Y. Shang, P. Feng, and C. Wang. Fungi That Infect Insects: Altering Host Behavior and Beyond. *PLOS Pathogens*, 11(8):e1005037, Aug. 2015. ISSN 1553-7374. doi: 10.1371/journal.ppat.1005037. URL `https://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1005037`.

P. Skamnioti and S. J. Gurr. Against the grain: safeguarding rice from rice blast disease. *Trends in Biotechnology*, 27(3):141–150, Mar. 2009. ISSN 0167-7799, 1879-3096. doi: 10.1016/j.tibtech.2008.12.002. URL `https://www.cell.com/trends/biotechnology/abstract/S0167-7799(09)00014-6`.

J. M. Smith. Fungal Pathogens of Nonhuman Animals. In *eLS*. American Cancer Society, 2006. ISBN 978-0-470-01590-2. doi: 10.1038/npg.els.0004235. URL `https://onlinelibrary.wiley.com/doi/abs/10.1038/npg.els.0004235`.

R. J. St. Leger and J. B. Wang. Metarhizium: jack of all trades, master of many. *Open Biology*, 10(12):200307, 2020. doi: 10.1098/rsob.200307. URL `https://royalsocietypublishing.org/doi/10.1098/rsob.200307`.

E. H. Stukenbrock, W. Quaedvlieg, M. Javan-Nikhah, M. Zala, P. W. Crous, and B. A. McDonald. Zymoseptoria ardabiliae and Z. pseudotritici, two progenitor species of the septoria tritici leaf blotch fungus Z. tritici (synonym: Mycosphaerella graminicola). *Mycologia*, 104(6):1397–1407, Dec. 2012. ISSN 0027-5514. doi: 10.3852/11-374.

G.-H. Sung, N. L. Hywel-Jones, J.-M. Sung, J. J. Luangsa-ard, B. Shrestha, and J. W. Spatafora. Phylogenetic classification of Cordyceps and the clavicipitaceous fungi. *Studies in Mycology*, 57:5–59, Jan. 2007. ISSN 0166-0616. doi: 10.3114/sim.2007.57.01. URL `https://www.sciencedirect.com/science/article/pii/S0166061614601305`.

L. H. Taylor, S. M. Latham, and M. E. Woolhouse. Risk factors for human disease emergence. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 356(1411):983–989, July 2001. ISSN 0962-8436. doi: 10.1098/rstb.2001.0888.

M. M. Teixeira, L. F. Moreno, B. J. Stielow, A. Muszewska, M. Hainaut, L. Gonzaga, A. Abouelleil, J. S. L. Patané, M. Priest, R. Souza, S. Young, K. S. Ferreira, Q. Zeng, M. M. L. da Cunha, A. Gladki, B. Barker, V. A. Vicente, E. M. de Souza, S. Almeida, B. Henrissat, A. T. R. Vasconcelos, S. Deng, H. Voglmayr, T. A. A. Moussa, A. Gorbushina, M. S. S. Felipe, C. A. Cuomo, and G. S. de Hoog. Exploring the genomic diversity of black yeasts and relatives (Chaetothyriales, Ascomycota). *Studies in Mycology*, 86:1–28, Mar. 2017. ISSN 0166-0616. doi: 10.1016/j.simyco.2017.01.001. URL `https://www.sciencedirect.com/science/article/pii/S0166061617300167`.

C. Vacher, D. Piou, and M.-L. Desprez-Loustau. Architecture of an Antagonistic Tree/Fungus Network: The Asymmetric Influence of Past Evolutionary History. *PLOS ONE*, 3(3):e1740, Mar. 2008. ISSN 1932-6203. doi: 10.1371/journal.pone.0001740. URL `https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0001740`.

L. van der Geest, S. Elliot, J. Breeuwer, and E. Beerling. Diseases of Mites. *Experimental & Applied Acarology*, 24 (7):497–560, July 2000. ISSN 1572-9702. doi: 10.1023/A:1026518418163. URL `https://doi.org/10.1023/A:1026518418163`.

M. Wardeh, C. Risley, M. K. McIntyre, C. Setzkorn, and M. Baylis. Database of host-pathogen and related species interactions, and their global distribution. *Scientific Data*, 2(1):150049, Sept. 2015. ISSN 2052-4463. doi: 10.1038/sdata.2015.49. URL `https://www.nature.com/articles/sdata201549`.

D. E. Wood, J. Lu, and B. Langmead. Improved metagenomic analysis with Kraken 2. *Genome Biology*, 20 (1):257, Nov. 2019. ISSN 1474-760X. doi: 10.1186/s13059-019-1891-0. URL `https://doi.org/10.1186/s13059-019-1891-0`.

S. Wu, M. D. Toews, L. A. Castrillo, A. K. Barman, T. E. Cottrell, and D. I. Shapiro-Ilan. Identification and Virulence of Cordyceps javanica Strain wf GA17 Isolated From a Natural Fungal Population in Sweetpotato Whiteflies (Hemiptera: Aleyrodidae). *Environmental Entomology*, 50(5):1127–1136, Oct. 2021. ISSN 0046-225X. doi: 10.1093/ee/nvab061. URL `https://doi.org/10.1093/ee/nvab061`.

Z. Zhang, Z. Cai, Z. Tan, C. Lu, T. Jiang, G. Zhang, and Y. Peng. Rapid identification of human-infecting viruses. *Transboundary and Emerging Diseases*, 66(6):2517–2522, 2019. doi: 10.1111/tbed.13314. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/tbed.13314`.