AMAPEC: accurate antimicrobial activity prediction for fungal effector proteins

Fantin Mesny[1], Bart PHJ Thomma[1,2,*]

1. Institute for Plant Sciences, University of Cologne – Cologne, Germany
2. Cluster of Excellence on Plant Sciences (CEPLAS) – Cologne, Germany
*: Corresponding author. E-mail: bthomma@uni-koeln.de

Abstract

Fungi typically occur in environments where numerous and diverse other microbes occur as well, often resulting in fierce competition for nutrients and habitat. To support fungal fitness in these environments, they evolved various mechanisms that mediate direct antagonism towards niche competitors. Among these, the secretion of proteins with antimicrobial activities has been reported in fungi with diverse lifestyles. Recently, several plant-associated fungi were shown to rely on the secretion of antimicrobial effector proteins to antagonize certain members of plant hosts' microbiota and to successfully colonize plant tissues. Some of these effectors do not share homology with known antimicrobials and represent novel antibiotics. Accordingly, the occurrence and conservation of proteinaceous antimicrobials throughout the fungal tree of life remains enigmatic. Here we present a computational approach to annotate candidate antimicrobial effectors in fungal secretomes based on protein physicochemical properties. After curating a set of proteins that were experimentally verified to display antimicrobial activity and a set of proteins that lack such activity, we trained a machine learning classifier on properties of protein sequences and predicted structures. This predictor performs particularly well on fungal proteins (R^2=0.89) according to our validations and is delivered as a software package named AMAPEC, dedicated to **a**nti**m**icrobial **a**ctivity **p**rediction for **e**ffector **c**andidates. We subsequently used this novel software to predict antimicrobial effector catalogs in three phylogenetically distant fungi with distinct lifestyles, revealing relatively large catalogs of candidate antimicrobials for each of the three fungi, and suggesting a broad occurrence of such proteins throughout the fungal kingdom. Thus, AMAPEC is a unique method to uncover antimicrobials in fungal secretomes that are often sparsely functionally annotated, and may assist biological interpretations during omic analyses. It is freely available at https://github.com/fantin-mesny/amapec.

Introduction

In virtually any environment, organisms engage in fierce competition for limited resources and survival[1]. Mechanisms for direct antagonism support such competition and are essential for any organism's fitness[2]. Among these, the production of antimicrobial compounds, encompassing antibacterial as well as antifungal proteins, has been reported in each of the kingdoms of life[3–8]. Short peptides generally termed "antimicrobial peptides" (AMPs), have been given particular attention and were demonstrated to exert microbiocidal or microbiostatic activity by interacting

with microbial cell walls, causing their alteration or disruption[3,9–13]. However, also larger-sized secreted enzymes, such as lysozymes, chitinases, proteases and ribonucleases, may similarly display such antimicrobial activities and have been demonstrated to be essential to restrain the proliferation of microbial competitors[4,14–17].

As they spend most of their life cycles in microbe-rich environments, fungi secrete antimicrobial compounds, including antibiotic secondary metabolites but also AMPs and other antimicrobial proteins, to suppress niche competitors and promote their proliferation[18–20]. For instance, the secretion of antimicrobials is essential for the fitness of soil-dwelling fungi, as they need to compete for nutrients with a broad diversity of microbes and especially with bacteria[21,22]. Similarly, secreted antimicrobials sustain the competitiveness of wood-decaying fungi and are major determinants of microbial community compositions on deadwood substrates[23]. The secretomes of plant-associated fungi include diverse carbohydrate-active enzymes (CAZymes) that digest host cell walls, but also small proteins termed “effectors” that promote fungal colonization of plant tissues[24–26]. While modulation of host immunity and, in a broader sense, host physiology was long thought to be the main function of effector proteins[26,27], recent discoveries of antimicrobial activities displayed by effectors of diverse fungi suggest that microbial antagonism may be one of their key roles as well[5,28–33]. For instance, the plant-pathogenic fungus *Verticillium dahliae* secretes antimicrobial proteins during both the soil-dwelling stages of its life cycle and during host colonization, to antagonize niche competitors and foster the invasion of both soil and plant tissues[5,29,30]. Since plant microbiota may have been selected to protect the host from fungal intrusions, and therefore function as an additional defense layer, fungal effectors with antimicrobial activities can play essential roles in host colonization, by breaching this microbial barrier and foster plant tissue invasion[34]. Moreover, we hypothesized that effector-mediated microbial antagonism is an ancient trait that already evolved in fungal ancestors that encountered microbial competition long before the evolution of symbioses with land plants or other types of eukaryotic multicellular organisms[19]. However, thus far the occurrence and conservation of antimicrobial effectors throughout the fungal tree of life remains enigmatic.

To aid in the discovery of novel antimicrobial effectors, and to gain insights into their evolutionary dynamics, we aim to predict the potential antimicrobial activities of proteins and to annotate catalogs of candidate antimicrobial effector genes in fungal genomes. Accurate predictors of antimicrobial activity relying on amino acid properties and on sequence patterns have previously been developed[35–40]. However, these are inappropriate to annotate fungal effector proteins as they were trained on AMP databases[41–43] and are therefore essentially dedicated to short peptides of sizes up to 100 amino acids in length. With sequence lengths up to 850 amino acids[44], fungal effectors are considerably larger and thus require a dedicated predictor based on a training dataset that incorporates larger antimicrobial proteins that likely have different properties and mode-of-actions than AMPs. Here, we describe the development of such tool and introduce the software AMAPEC (for **a**nti**m**icrobial **a**ctivity **p**rediction for **e**ffector **c**andidates), a tool to annotate candidate antimicrobial effector proteins relying on their sequence and predicted structure properties.

Results

Composition of a literature-curated set of experimentally validated antimicrobial proteins

In order to develop an appropriate tool to predict antimicrobial activities among fungal effector proteins, it is important to determine the size range of typical effector proteins. By analyzing the predicted secretomes of three phylogenetically distant fungal strains with distinct lifestyles, namely the plant pathogenic ascomycete *Verticillium dahliae* (strain JR2)[45], the saprotrophic basidiomycete *Coprinopsis cinerea* (strain AmutBmut pab1-1)[46] and the arbuscular mycorrhizal glomeromycete *Rhizophagus irregularis* (strain DAOM197198)[47], we determined that fungi are predicted to secrete proteins from 24 up to 3445 amino acids in length (median=282 amino acids; Supplementary Fig 1). To assemble a set of experimentally validated antimicrobial proteins to assist our identification of novel fungal antimicrobial effectors, we performed a literature search and curated a set of similarly-sized proteins that have been experimentally verified to display antibacterial or antifungal activity *in vitro*. While paying attention not to enrich our dataset in short peptides that are overrepresented in literature but only occur in low amounts in fungal secretomes (Supplementary Fig 1), we identified 152 antimicrobial proteins, originating from a great diversity of organisms including mostly proteins of animal origin (n=81) but also a significant proportion of fungal proteins (n=29) (Fig 1a; Supplementary Table 1). These proteins display little genetic redundancy (Supplementary Fig 2ab), but structural similarity occurs more frequently (Supplementary Fig 2cd), based on Blastp[48]- and Foldseek[49] analyses of their sequences and structures predicted with AlphaFold2[50].

Next, we tested if our literature-curated set of proteins would allow us to identify novel candidate antimicrobials in fungal secretomes. Since certain fungal antimicrobial effectors have been discovered due to their sequence or structure similarity with known antimicrobials[5,29,30], we performed Blastp[48] and Foldseek[49] similarity searches using the secretomes of *V. dahliae*, *C. cinerea* and *R. irregularis* as queries and our literature-curated set of antimicrobials as subject. While sequence similarity searches revealed between 40 and 60 novel candidate antimicrobials per fungal secretome (E-value<0.05; Supplementary Table 2a), structural similarity analysis identified 120 to 270 candidates (E-value<0.05; Supplementary Table 2b). Our literature-curated set of 152 antimicrobial proteins is thus likely to support the discovery of novel fungal effectors with antimicrobial activity.

Accurate prediction of antimicrobial activity based on protein physicochemical properties

Since physicochemical properties of candidate proteins have previously supported the identification of AMPs[40,51], we hypothesized that physicochemical properties of fungal effector proteins may allow accurate prediction of their potential antimicrobial activities. To train such predictor, additionally to our set of experimentally verified antimicrobials, we curated a negative training set of proteins which according to their functional annotation, are unlikely to have antimicrobial activity (Supplementary Table 3). Since we anticipate that effector proteins lacking antimicrobial activity outnumber antimicrobial effector proteins in fungal secretomes, we took care that the

number of non-antimicrobial proteins exceeds the number of antimicrobials in the training set. Eventually, we doubled the size of the positive dataset (n=152), with a negative set of 304 members that include equivalent proteins in terms of sequence length and phylogenetic origin, and a significant proportion of secreted proteins (Fig 1ab). For each protein in the training set, we calculated properties from their amino acid sequences and from predicted high-confidence structures (AlphaFold2[50]; Supplementary Fig 3). In total, 70 numerical values reflecting diverse physicochemical properties were determined for each protein (Fig 1c; Supplementary Table 4a). Additionally, we queried for the presence/absence of six k-mers identified to be over- or underrepresented in the sequences of antimicrobial proteins in the training set (Fig 1c; Supplementary Table 4bc). We used these data to train a Support Vector Machines (SVM) classifier and subsequently estimated its quality through leave-one-out cross-validation. The quality assessment revealed that our classifier has high accuracy, recall and specificity, particularly for fungal proteins, although its precision value reveals a moderate bias towards false positive detections (Fig 1d). Whereas 42 of the 456 (9%) proteins in the training set were incorrectly classified as antimicrobials, only six out of 89 (6%) fungal proteins in this training set received such incorrect prediction. Thus, our predictor demonstrates that physicochemical properties of proteins can be correlated with their antimicrobial activity. Analysis of support vector coefficients, representing the importance of individual physiochemical properties for the prediction, revealed a particular role of properties linked to hydrophobicity, charge, secondary structures, disulphide bonds and structural cavities, and of the identity of exposed amino acids (Supplementary Fig 4).

The high accuracy of our predictor makes it a reliable tool to assist the identification of novel fungal antimicrobials. Thus, we further developed a software named AMAPEC, that relies on our SVM classifier. To help users to distinguish candidate antimicrobials with the highest confidence, we additionally trained a probability estimator by Platt scaling of our binary SVM classifier, allowing AMAPEC to return probability scores for protein antimicrobial activity (Fig 1c). Thus, using a (predicted) protein structure as input, AMAPEC returns (1) a mean confidence score for the predicted protein structure (the so-called “pLDDT”, introduced with AlphaFold2[50]), with the rationale that a low-confidence structure may obtain a predicted antimicrobial activity that is not biologically meaningful; (2) a classification as ‘Antimicrobial’ or ‘Non-antimicrobial’; and finally (3) a probability score for its antimicrobial activity, that ranges between 0 (no antimicrobial activity) and 1 (highly likely to be antimicrobial). AMAPEC is made available through GitHub with a GPL v3.0 license: https://github.com/fantin-mesny/amapec. Additionally, we provide a Google Colab notebook allowing to try AMAPEC and to perform online antimicrobial activity prediction: https://colab.research.google.com/github/fantin-mesny/amapec/blob/main/googleColab/AMAPEC.ipynb.

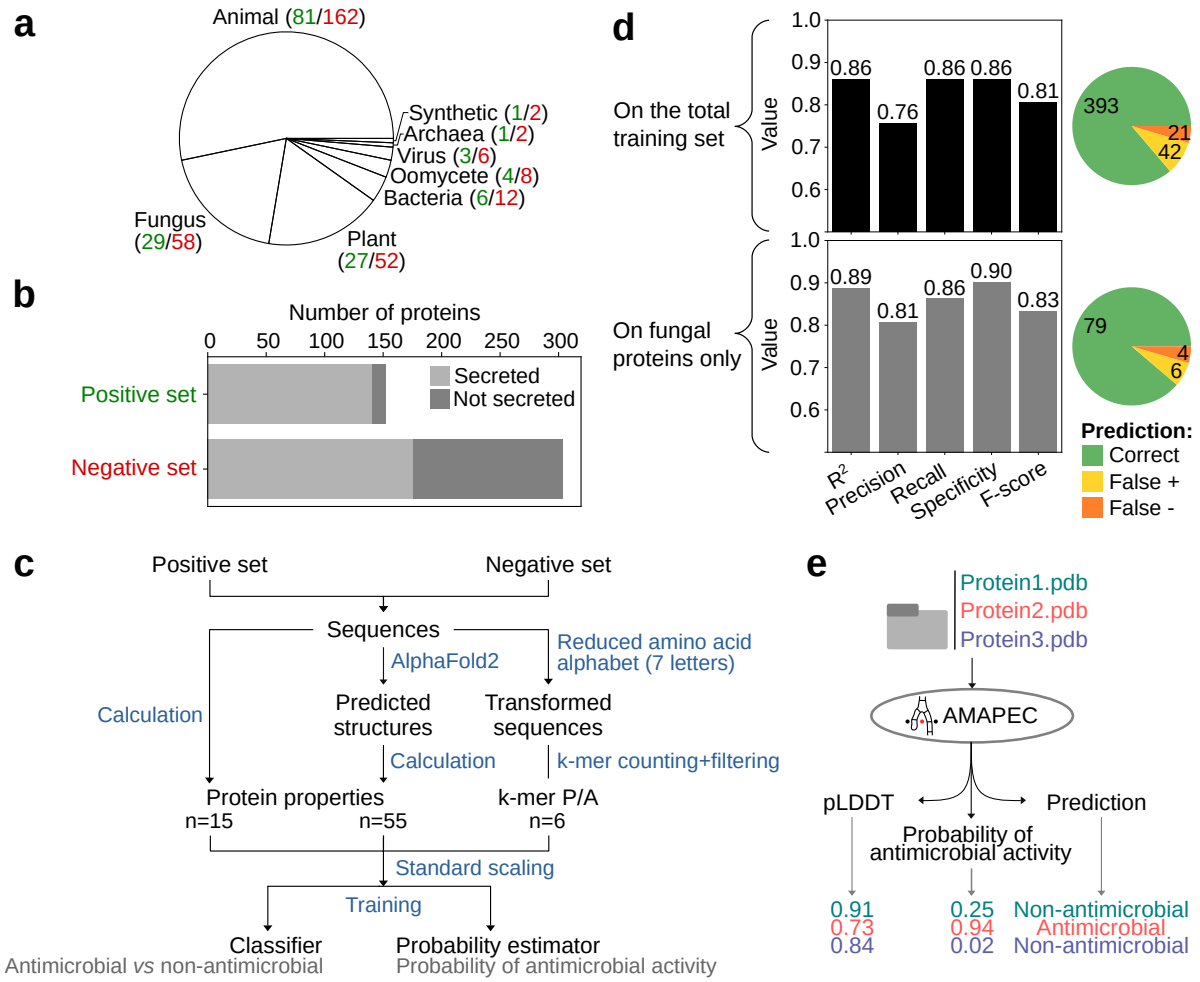


Figure 1: Description of the antimicrobial activity predictor AMAPEC v1.0. **a.** Phylogenetic origin of the proteins in the dataset (green: number of proteins in positive dataset; red: number in negative dataset) that was used to train the predictor. **b.** Number of proteins included in the training dataset and proportion of secreted proteins. **c.** Schematic overview of the training pipeline (P/A= presence/absence). **d.** Estimation of the classifier quality, based on "leave-one-out" cross-validation in the training dataset. The top bar plot and pie chart show quality estimates calculated on the total dataset (n=456), while the bottom charts analyze only the classifications of fungal proteins (n=87) during the "leave-one-out" cross validation. **e.** Schematic overview of AMAPEC v1.0 showing its inputs and outputs with an example of three proteins.

AMAPEC predicts numerous antimicrobials in fungal secretomes

Information on the size and composition of gene catalogs encoding secreted antimicrobial proteins in fungal genomes is scarce. To gain insights in the occurrence of secreted antimicrobial proteins in the fungal kingdom, we analyzed the secretomes of the three phylogenetically distant fungi *V. dahliae*, *C. cinerea* and *R. irregularis*. Similarity-based functional annotation with emapper[52] revealed that their secretomes include significant proportions of proteins that lack functional annotations (Fig 2), while several of the assigned annotations are poorly informative (Supplementary Table 5). Specific annotation of CAZymes with dbcan[53] revealed variable catalogs of these enzymes in fungal secretomes, ranging from less than 10% of the secretome of the mycorrhizal fungus *R. irregularis* up to 31% of the secretome of *V. dahliae*, which is consistent with

previous reports[24,25]. Since CAZyme families are well documented[54] and poorly represented in our training dataset (Supplementary Table 1), and also because it is difficult to determine whether individual CAZymes truly antagonize microbial growth, we excluded them from subsequent analyses. The structures of other proteins in the three secretomes were predicted with ESMFold[55], a tool that is considerably faster and computationally less demanding than AlphaFold2 and that is therefore better suited for structure prediction of large numbers of proteins such as complete effector catalogs, despite a minor drop in the quality of predicted structures (Supplementary Fig 5). On average, confident structures were predicted with ESMFold (Supplementary Fig 6), with secreted proteins of *R. irregularis* displaying the lowest confidence, possibly due to the low availability of structurally characterized homologs of proteins originating from the early-diverging Glomeromycetes clade. Using these structures as an input, AMAPEC was employed to predict antimicrobial effector catalogs in the three fungi. Interestingly, predictions revealed that one third to one half of total fungal secretomes is composed of proteins with antimicrobial properties (Fig 2), demonstrating the ability of AMAPEC to discover more candidates than similarity-based approaches do (Supplementary Table 2). While these large numbers corroborate the hypothetical importance of antimicrobial proteins for sustaining fungal fitness in diverse environments[19], they also reveal a broad occurrence of such secreted antimicrobials throughout the fungal tree of life. Functional enrichment analyses with GOATOOLS[56] did not identify significantly overrepresented functional terms in predicted antimicrobials (FDR<0.05), possibly due to the sparse annotation of secretomes. In line with this observation, one to two thirds of predicted antimicrobials could not be annotated based on sequence similarity (Fig 2). Thus, we conclude that AMAPEC offers unprecedented functional insights into fungal secretomes and may assist biological interpretations during genomic, transcriptomic and proteomic analyses.

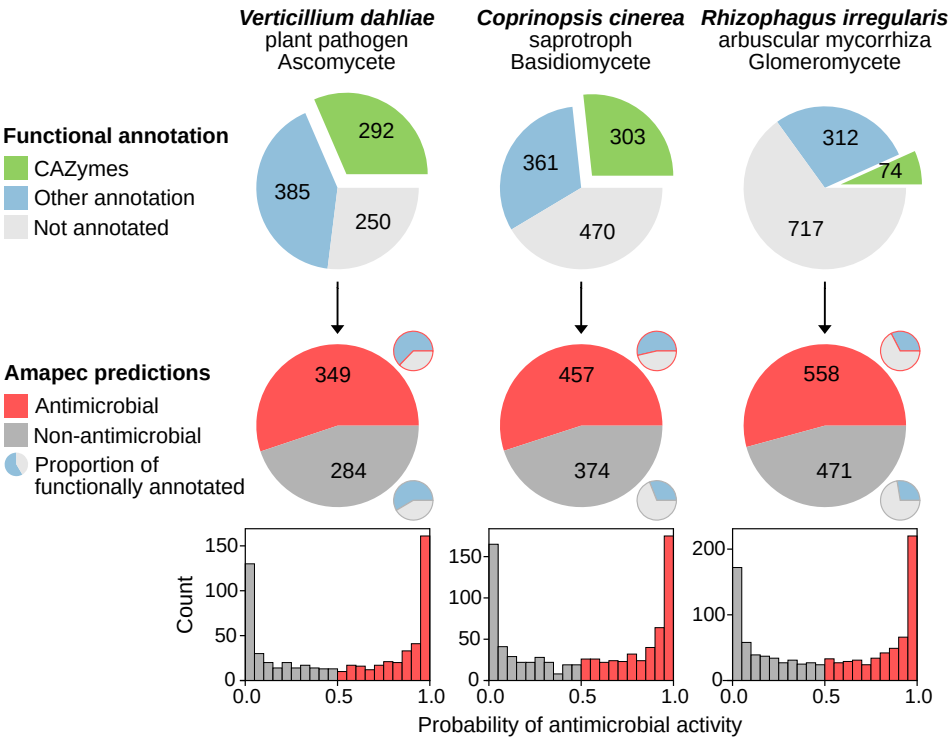


Figure 2: AMAPEC predicts numerous antimicrobials in the secretomes of diverse fungi.

The top row of pie charts shows proportions of functional annotations in the secretomes of three fungi. CAZymes were annotated with dbcan[53] and other functions were annotated with emapper[52]. The bottom row of pie charts shows results of antimicrobial activity prediction with AMAPEC, performed on all secreted proteins that are not CAZymes. On this same row, smaller pie charts depict proportions of predicted antimicrobials and non-antimicrobials that could be functionally annotated with emapper (in blue). Histograms showing the distributions of probabilities of antimicrobial activity estimated by AMAPEC are displayed at the bottom.

Discussion

Like other non-model organisms, fungal genomes comprise a large proportion of genes that lack functional annotation, which complicates biological interpretations of genomic and transcriptomic analyses. Fungal secretomes are particularly poorly annotated by similarity-based methods (Fig 2), since they include significant numbers of fast-evolving effectors that represent recent innovations and show little conservation across the fungal kingdom[44,57]. Recently, major efforts were made to develop new functional annotation methods, relying mostly on machine-learning prediction, but these were not aimed to discover antimicrobial activities[58–62]. Following the recent characterization of various fungal effectors with antimicrobial activities[5,28–33], we aimed to develop a method to predict antimicrobial activity in fungal secretomes and assist discoveries of such proteins across the fungal tree of life.

While the set of antimicrobial proteins curated to train our predictor is genetically diverse (Supplementary Fig 2), its size is limited by the number of effector-sized proteins with antimicrobial activity described in literature. Therefore, it does not cover the entire diversity of antimicrobial mechanisms existing in nature and AMAPEC is restricted to the prediction of previously described activities. Novel discoveries of proteinaceous antimicrobials will be implemented in the training set to expand the scope of the predictor in future versions of AMAPEC. Other biases in AMAPEC predictions may originate from our negative training set, that was manually curated to include non-antimicrobial proteins according to functional annotation. However, possible misjudgment during this curation, for instance due to bad annotations, may negatively affect predictions. Arguably, this limitation also concerns previously published AMP predictors for which negative training sets were assembled with a similar approach[36,40]. Nonetheless, the difficulty to experimentally demonstrate absence of antimicrobial activity makes it challenging, if not impossible, to overcome this limitation.

Antimicrobial activities of short AMPs were previously linked to physicochemical properties that can be inferred from amino acid sequence compositions[40,51]. We used such properties to predict the antimicrobial activity of effector-sized proteins, and relied on predicted structures to describe folded protein physicochemistry (Fig 1c). Our predictor corroborated studies of AMPs by revealing a particular importance of properties linked to hydrophobicity, charge, secondary structure and disulphide bonds for protein antimicrobial activity[40,51,63]. However, it also suggested a previously undescribed role of the identity of certain exposed amino acids and of structural cavities (Supplementary Fig 4), thereby raising new questions about the mode-of-action of proteinaceous antimicrobials.

These structural insights were obtained from AlphaFold2[50]-predicted protein structures in our training dataset. These structures have overall high confidence scores (Supplementary Fig 3). It is important to interpret AMAPEC output predictions while considering input structure quality. Depending on the tool used and the number of homologs with characterized structures in reference databases, structure predictions can result in low confidence geometries (Supplementary Fig 5; Supplementary Fig 6) which may impact prediction of antimicrobial activities. To assist in the interpretation of prediction results, AMAPEC returns the input structure confidence scores (pLDDT), which should be considered together with antimicrobial activity probabilities (Fig 1e). In other words, the predicted antimicrobial activity probability should be considered with caution if the input structure confidence score is low. In addition to structure quality-related limitations, the need

to compute structure predictions prior to the execution of AMAPEC can itself restrain the use of our software, since this process is computationally demanding in terms of time and resources. As input for AMAPEC, we strongly recommend to use structures predicted from mature amino acid sequences, after prediction and removal of signal peptides with tools like SignalP[64]. While the AlphaFold database provides numerous fungal effector protein structures[65], these have mostly been predicted on canonical amino acid sequences that include secretion signal peptides, resulting in biased geometries that do not represent state in which effectors are secreted from the organism that produces them and, therefore, not the "active state". The recent development of the relatively rapid structure prediction tool ESMFold[55] made it possible to compute antimicrobial activity predictions for complete fungal secretomes (Fig 2). Likely, future structure prediction algorithms will speed up this process even further. Moreover, initiatives like ColabFold[66] provide means for free online structure prediction and aim to make such computation accessible to all.

We analyzed the secretomes of three phylogenetically distant fungi with different lifestyles and identified large proportions of predicted antimicrobials, ranging from one third to half of the secreted proteins (Fig 2). CAZymes were excluded from this predictive analysis as they can be confidently recognized by dedicated annotation methods[53], and because antimicrobial activities have not been investigated in most CAZyme families (e.g. in plant cell wall-degrading enzymes), besides lysozymes and chitinases. Since AMAPEC especially aims to shed light on non-annotated portions of fungal secretomes, we suggest its users to annotate CAZymes separately and to exclude them from predictions. Our predictions (Fig 2) suggest a broad occurrence of proteinaceous antimicrobials across the fungal tree of life. Moreover, the large numbers of predicted antimicrobials corroborates the assumption that antimicrobial proteins are of major importance for fungal competitiveness, fitness, and survival in nature[19], although the low quality of certain structures (Supplementary Fig 6) and the intrinsic precision of AMAPEC (Fig 1d) may lead to a limited number of false positive predictions. Future experimental validation of predictions should confirm the estimated high accuracy of AMAPEC (Fig 1d) and is especially needed since quality estimation was performed by cross-validation using the training dataset exclusively. We hope the current version of AMAPEC will assist researchers in the characterization of new effectors and contribute to gaining more insights into fungal antagonistic mechanisms. Such novel characterizations will contribute to improve AMAPEC in return, since the software will be updated regularly, and its training dataset will be supplemented with recently discovered antimicrobials. Updates of the training dataset will also incorporate more high-confidence structures, following improvements of AlphaFold[50] and the addition of novel effector structures in its reference databases.

Methods

Reference fungal secretomes

Sets of proteins associated to the published genomes of three phylogenetically distant fungi with distinct lifestyles were downloaded: *Verticillium dahliae* JR2[45] (annotation *VDAG_JR2 v.4.0* downloaded from the database Ensembl Fungi[67]), *Coprinopsis cinerea* AmutBmut pab1-1[46] (annotation *Copci_AmutBmut1 v1.0* downloaded from the database JGI Mycocosm[68]) and *Rhizophagus irregularis* DAOM197198[47]. SignalP[64] v6.0 was then used to predict secretion signal peptides in protein sequences and thereby define the secretomes of these fungi. Sequences with removed signal peptides were used in all subsequent analyses. Functional annotation of proteins in

these secretomes was carried out using emapper[52] v2.0 and the database Eggnog[69] v5. CAZymes were specifically annotated in these secretomes using dbcan[53] v4.0. Structure predictions were computed with ESMFold[55] v1.0.3, using default parameters, of all secreted proteins besides CAZymes. The structures of two proteins from *V. dahliae* (VDAG_JR2_Chr4g10970 and VDAG_JR2_Chr1g22375) could not be predicted due to high computational requirements linked to their size (>2500 amino acids) and were excluded from our analyses. To compare the quality of protein structures predicted by AlphaFold2[50] and ESMFold (Supplementary Fig 5), we also computed structure prediction for 626/635 non-CAZyme secreted proteins of *V. dahliae*, using AlphaFold v2.0 with parameters --max_template_date=2021-05-14 --preset=casp14, with nine predictions failing due to high computational requirements.

Curation of a set of antimicrobial proteins

First, a positive training set of antimicrobial proteins (Supplementary Table 1) was curated from the literature. Only proteins which antimicrobial activity has been experimentally demonstrated *in vitro* (i.e. restricting the growth of bacteria and/or fungi in culture medium) were selected. While not restraining the dataset to proteins encoded by any phylogenetic group, we paid particular attention to include all the fungal antimicrobial proteins reported in scientific literature. Importantly, secretion signal peptides were removed from sequences (SignalP v6.0[64]), since the antimicrobial function of proteins generally occurs after secretion. Considering sequence lengths of fungal secreted proteins, peptide with mature sequence lengths below 40 amino acids were excluded (Supplementary Fig 1), not to enrich the protein set in AMPs, for which dedicated predictors exist[35–38,40]. By largely spanning the size range of typical effector proteins, this protein set should support the prediction of effector antimicrobial activity without bias towards the recognition of short AMPs, that are the most described antimicrobial proteins in the literature.

Similarity searches of secreted proteins from the three fungal secretomes (excluding CAZymes, since their function is well documented) were performed. For sequence similarity searches, blastp[48] v.2.5.0 was used with parameters "--max_target_seqs 1 --evalue 0.05". For structure similarity searches, ESMFold-predicted structures were used as inputs for Foldseek[49] v6.29e2557, run in function *foldseek search* then *foldseek convertalis* with default parameters. Results were then filtered to remove hits with E-values >0.05. Self-hits of known *V. dahliae* antimicrobials that were implemented in our literature-curated set of antimicrobials were manually removed.

Assembly of a negative training dataset

A negative training dataset was assembled by gathering presumable non-antimicrobial proteins. As previously suggested[36,40], this negative set was curated by retrieving proteins which functional annotation does not suggest any antimicrobial activity from the UniProt database[70]. To do so, Gene Ontology (GO) terms associated to antimicrobial activity were filtered out (i.e. GO:0090729, GO:0001878, GO:0045087, GO:0050830, GO:0050829, GO:0042742, GO:0071222, GO:0071224, GO:0001530, GO:0031640, GO:0050832). Additionally, only well-annotated proteins without any known function in microbial antagonism or immunity were selected. To prevent strong effects of potential misjudgment during the curation process, the negative training dataset includes twice as many non-antimicrobial proteins as there are antimicrobials in the positive set. For each antimicrobial in the positive set, two presumably non-antimicrobial proteins encoded by the same organism (or a close relative) and with similar sizes (+/- 4 amino acids) were included in the negative set. We paid attention to include at least as many secreted proteins (signal peptide detected

and removed with SignalP[64]) in the negative set as in the positive set, not to bias the prediction towards apoplastically released proteins. Finally, since 11 proteins in the positive training set were annotated or described as ribonucleases, 11 ribonucleases, unlikely to exert antimicrobial functions according to their annotation (for instance, involved in transfer RNA maturation) were included in the negative set.

Calculation of protein properties

The AMAPEC predictor was trained on a set of 70 numerical variables reflecting protein physicochemical properties (Supplementary table 4a). Some of these values (n=15) were calculated from amino acid sequences, using R v4.2.0 and the library Peptides v2.4.4[71]. However, to better describe the physicochemistry of proteins, their predicted structures were used to calculate 55 numerical values per protein that reflect structural properties. Protein structures were predicted using AlphaFold[50] v2.0 with parameters "--max_template_date=2021-05-14 –preset=casp14". Structure properties were calculated from AlphaFold best models (ranked_0.pdb output files) using Python v3.11.5 and the PDB parser implemented in Biopython v1.78. Some previously published code and formula from diverse sources[72–74] (details in Supplementary Table 4a) were implemented in AMAPEC's Python scripts. For properties linked to protein secondary structures, DSSP[75,76] v3.0.0 was used to assign individual amino acids to different types of secondary structures. Additionally, pocket structures in proteins were predicted using Fpocket[77] v4.0.2, and information related to their number, size and properties were implemented as variables.

Additional to sequence- and structure-derived physicochemical properties, the presence/absence in protein sequences of certain k-mers was implemented as variable. To reduce sequence complexity, a reduced 7-letter aminoacid alphabet was used, as previously implemented in various machine learning methods applied to protein sequences[36,78]. A novel alphabet based on amino acid properties was designed, to define k-mers that may represent key motifs in protein physicochemistry (Supplementary Table 4b). The k-mer compositions of transformed sequences in the training set was profiled using MerCat2[79], with k-mer sizes 3, 4, 5 and 6 amino acids. Then, chi-squared multiple testing was computed, as implemented in function feature_selection.SelectFdr(chi2, alpha=0.05) of Python library scikit-learn[80] v1.2.1, to identify k-mers that are over- or under-represented in antimicrobial protein sequences. This aimed to select for k-mers that are likely biologically meaningful, and to prevent later overfitting of our prediction model that can occur if training is performed on numerous variables which combination describes protein sequences in too much detail. With chi-squared testing, six k-mers (five 4-mers and one 3-mer) of interest were identified. Their presence/absence was implemented in the set of protein properties used to train the AMAPEC predictor (Supplementary table 4c).

Classifier training and quality estimation

Numerical variables reflecting properties of our 456 proteins were standardized using function preprocessing.StandardScaler() of Python library scikit-learn[80] v1.2.1. Then, a Support Vector Machines (SVM) classifier with a linear kernel was trained using function svm.SVC() from scikit-learn. The correct the imbalance of the training set (152 proteins in the positive set and 304 in the negative set), the weight of antimicrobials was set to 2 and the weight of non-antimicrobials to 1. A second model was trained to predict the probability of antimicrobial activity, by computing Platt

scaling over the SVM binary classifier. To do so, the function calibration.CalibratedClassifierCV(method='sigmoid', cv='prefit') from sci-kit learn was used. Both models were exported using function dump() from Python library joblib v1.2.0.

Due to the small size of the training dataset (n=456), classifier quality testing was performed through leave-one-out cross-validation. As implemented in function cross_val_score(cv=KFold(n_splits=456)) of scikit-learn, 456 SVM classifiers were trained with a train/test split of 455/1 to classify individual proteins using as a basis, protein properties in the rest of the dataset. Protein classifications into “antimicrobial” or “non-antimicrobial” were then analyzed by counting numbers of true positives, false positives, true negatives and false negatives. These counts allowed the estimation of the overall classifier accuracy (R^2) but also its precision, recall, specificity and F-score. Such quality estimates were also calculated by exclusively taking the classification correctness of fungal proteins into account, to identify if the predictor is suited for the annotation of fungal proteins.

A bash pipeline allowing both the calculation of protein properties and antimicrobial activity prediction using the trained predictors was written, resulting in the software AMAPEC v1.0, (developed and tested on operating system GNU/Linux Ubuntu v20.04.3 LTS).

Prediction of antimicrobial activity in fungal secretomes

AMAPEC v1.0 was used to predict the antimicrobial activity of proteins secreted by *V. dahliae*, *C. cinerea* and *R. irregularis*, while excluding dbcan[53]-annotated CAZymes. ESMFold[55]-predicted structures, which pLDDT confidence scores can be seen on Supplementary Figure 6 and in Supplementary Table 5, were used as an input. GO enrichment analyses were performed with software GOATOOLS[56] v1.3.1 and GO term annotation from emapper[52].

Author contributions

B.P.H.J.T. and F.M. initiated, designed and coordinated the project. All data analyses and software development were performed by F.M. This manuscript was written by F.M. with input from B.P.H.J.T.

Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), through the funding of F.M.’s Walter Benjamin position (Project ID: ME 6064/1-1). B.P.H.J.T. acknowledges funding by the Alexander von Humboldt Foundation in the framework of an Alexander von Humboldt Professorship endowed by the German Federal Ministry of Education and Research and is furthermore supported by DFG under Germany’s Excellence Strategy – EXC 2048/1 – Project ID: 390686111 and by the DFG – Project ID 458090666 / CRC1535/1. We thank Michael F. Seidl for his helpful suggestions and for proofreading this manuscript.

References

1. Lang, J. M. & Benbow, M. E. Species interactions and competition. *Nat. Educ. Knowl.* **4**, 8 (2013).
2. Gómez, J. M., Verdú, M. & Perfectti, F. Ecological interactions are evolutionarily conserved across the entire tree of life. *Nature* **465**, 918–921 (2010).
3. Pierre, J. F. *et al.* Peptide YY: A Paneth cell antimicrobial peptide that maintains *Candida* gut commensalism. *Science (80-.).* **381**, 502–508 (2023).
4. Huynh, Q. K. *et al.* Antifungal proteins from plants. Purification, molecular cloning, and antifungal properties of chitinases from maize seed. *J. Biol. Chem.* **267**, 6635–6640 (1992).
5. Snelders, N. C. *et al.* Microbiome manipulation by a soil-borne fungal plant pathogen using effector proteins. *Nat. Plants* **6**, 1365–1374 (2020).
6. Chiumento, S. *et al.* Ruminococcin C, a promising antibiotic produced by a human gut symbiont. *Sci. Adv.* **5**, eaaw9969 (2019).
7. Metcalf, J. A., Funkhouser-Jones, L. J., Brileya, K., Reysenbach, A.-L. & Bordenstein, S. R. Antibacterial gene transfer across the tree of life. *Elife* **3**, e04266 (2014).
8. Gómez-Pérez, D. *et al.* Proteins released into the plant apoplast by the obligate parasitic protist *Albugo* selectively repress phyllosphere-associated bacteria. *New Phytol.* **239**, 2320–2334 (2023).
9. Nguyen, L. T., Haney, E. F. & Vogel, H. J. The expanding scope of antimicrobial peptide structures and their modes of action. *Trends Biotechnol.* **29**, 464–472 (2011).
10. Lee, C. C., Sun, Y., Qian, S. & Huang, H. W. Transmembrane pores formed by human antimicrobial peptide LL-37. *Biophys. J.* **100**, 1688–1696 (2011).
11. Ling, L. L. *et al.* A new antibiotic kills pathogens without detectable resistance. *Nature* **517**, 455–459 (2015).
12. Müller, A. *et al.* Daptomycin inhibits cell envelope synthesis by interfering with fluid membrane microdomains. *Proc. Natl. Acad. Sci.* **113**, E7077--E7086 (2016).
13. Scheinpflug, K. *et al.* Antimicrobial peptide cWFW kills by combining lipid phase separation with autolysis. *Sci. Rep.* **7**, 44332 (2017).
14. Xue, Q. *et al.* A new lysozyme from the eastern oyster, *Crassostrea virginica*, and a possible evolutionary pathway for i-type lysozymes in bivalves from host defense to digestion. *BMC Evol. Biol.* **10**, 1–17 (2010).
15. Campanelli, D., Detmers, P. A., Nathan, C. F., Gabay, J. E. & others. Azurocidin and a homologous serine protease from neutrophils. Differential antimicrobial and proteolytic properties. *J. Clin. Invest.* **85**, 904–915 (1990).
16. Harder, J. & Schröder, J.-M. RNase 7, a novel innate immune defense antimicrobial protein of healthy human skin. *J. Biol. Chem.* **277**, 46779–46784 (2002).

17. Eitzen, K., Sengupta, P., Kroll, S., Kemen, E. & Doehlemann, G. A fungal member of the *Arabidopsis thaliana* phyllosphere antagonizes *Albugo laibachii* via a GH25 lysozyme. *Elife* **10**, 1 (2021).

18. Macheleidt, J. *et al.* Regulation and role of fungal secondary metabolites. *Annu. Rev. Genet.* **50**, 371–392 (2016).

19. Snelders, N. C., Rovenich, H. & Thomma, B. P. H. J. Microbiota manipulation through the secretion of effector proteins is fundamental to the wealth of lifestyles in the fungal kingdom. *FEMS Microbiol. Rev.* **46**, fuac022 (2022).

20. Sułkowska-Ziaja, K. *et al.* Natural compounds of fungal origin with antimicrobial activity—Potential cosmetics applications. *Pharmaceuticals* **16**, 1200 (2023).

21. De Boer, W., Folman, L. B., Summerbell, R. C. & Boddy, L. Living in a fungal world: impact of fungi on soil bacterial niche development. *FEMS Microbiol. Rev.* **29**, 795–811 (2005).

22. Kombrink, A. *et al.* Induction of antibacterial proteins and peptides in the coprophilous mushroom *Coprinopsis cinerea* in response to bacteria. *ISME J.* **13**, 588–602 (2018).

23. Hiscox, J. & Boddy, L. Armed and dangerous – Chemical warfare in wood decay communities. *Fungal Biol. Rev.* **31**, 169–184 (2017).

24. Mesny, F. *et al.* Genetic determinants of endophytism in the *Arabidopsis* root mycobiome. *Nat. Commun.* **12**, 1–15 (2021).

25. Miyauchi, S. *et al.* Large-scale genome sequencing of mycorrhizal fungi provides insights into the early evolution of symbiotic traits. *Nat. Commun.* **11**, 1–17 (2020).

26. Doehlemann, G., Ökmen, B., Zhu, W. & Sharon, A. Plant pathogenic fungi. *Microbiol. Spectr.* **5**, (2017).

27. Plett, J. M. & Plett, K. L. Leveraging genomics to understand the broader role of fungal small secreted proteins in niche colonization and nutrition. *ISME Commun.* **2**, 1–6 (2022).

28. Chavarro-Carrero, E. A. *et al.* The soil-borne white root rot pathogen *Rosellinia necatrix* expresses antimicrobial proteins during host colonization. *bioRxiv* 2023.04.10.536216 (2023) doi:10.1101/2023.04.10.536216.

29. Snelders, N. C., Petti, G. C., van den Berg, G. C. M., Seidl, M. F. & Thomma, B. P. H. J. An ancient antimicrobial protein co-opted by a fungal plant pathogen for in planta mycobiome manipulation. *Proc. Natl. Acad. Sci.* **118**, e2110968118 (2021).

30. Snelders, N. C. *et al.* A highly polymorphic effector protein promotes fungal virulence through suppression of plant-associated Actinobacteria. *New Phytol.* **237**, 944–958 (2023).

31. Kettles, G. J. *et al.* Characterization of an antimicrobial and phytotoxic ribonuclease secreted by the fungal wheat pathogen *Zymoseptoria tritici*. *New Phytol.* **217**, 320–331 (2018).

32. Ökmen, B., Katzy, P., Huang, L., Wemhöner, R. & Doehlemann, G. A conserved extracellular Ribo1 with broad-spectrum cytotoxic activity enables smut fungi to compete with host-associated bacteria. *New Phytol.* **240**, 1976–1989 (2023).

33. Fardella, P. A., Tian, Z., Clarke, B. B. & Belanger, F. C. The *Epichloë festucae* antifungal protein Efe-AfpA protects creeping bentgrass (*Agrostis stolonifera*) from the plant pathogen *Clarireedia jacksonii*, the causal agent of dollar spot disease. *J. Fungi* **8**, 1097 (2022).

34. Mesny, F., ephane Hacquard, S. & PHJ Thomma, B. Co-evolution within the plant holobiont drives host performance. *EMBO Rep.* **24**, e57455 (2023).

35. Meher, P. K., Sahu, T. K., Saini, V. & Rao, A. R. Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC. *Sci. Rep.* **7**, 1–12 (2017).

36. Veltri, D., Kamath, U. & Shehu, A. Deep learning improves antimicrobial peptide recognition. *Bioinformatics* **34**, 2740–2747 (2018).

37. Lin, T.-T. *et al.* AI4AMP: an antimicrobial peptide predictor using physicochemical property-based encoding method and deep learning. *mSystems* **6**, (2021).

38. Lee, H., Lee, S., Lee, I. & Nam, H. AMP-BERT: Prediction of antimicrobial peptide function based on a BERT model. *Protein Sci.* **32**, e4529 (2023).

39. Yan, J. *et al.* Deep-AmPEP30: Improve short antimicrobial peptides prediction with deep learning. *Mol. Ther. - Nucleic Acids* **20**, 882–894 (2020).

40. Torrent, M., Andreu, D., Nogués, V. M. & Boix, E. Connecting peptide physicochemical and antimicrobial properties by a rational prediction model. *PLoS One* **6**, e16968 (2011).

41. Pirtskhalava, M. *et al.* DBAASP v3: database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics. *Nucleic Acids Res.* **49**, D288–D297 (2021).

42. Kang, X. *et al.* DRAMP 2.0, an updated data repository of antimicrobial peptides. *Sci. Data 2019 61* **6**, 1–10 (2019).

43. Wang, G., Li, X. & Wang, Z. APD3: the antimicrobial peptide database as a tool for research and education. *Nucleic Acids Res.* **44**, D1087–D1093 (2016).

44. Todd, J. N. A., Carreón-Anguiano, K. G., Islas-Flores, I. & Canto-Canché, B. Fungal effectoromics: a world in constant evolution. *Int. J. Mol. Sci.* **23**, 13433 (2022).

45. de Jonge, R. *et al.* Tomato immune receptor Ve1 recognizes effector of multiple fungal pathogens uncovered by genome and RNA sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 5110–5115 (2012).

46. Muraguchi, H. *et al.* Strand-specific RNA-Seq analyses of fruiting body development in *Coprinopsis cinerea*. *PLoS One* **10**, e0141586 (2015).

47. Yildirir, G. *et al.* Long reads and Hi-C sequencing illuminate the two-compartment genome of the model arbuscular mycorrhizal symbiont *Rhizophagus irregularis*. *New Phytol.* **233**, 1097–1107 (2022).

48. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

49. van Kempen, M. *et al.* Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* 1–4 (2023).

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

51. Wang, G. The antimicrobial peptide database is 20 years old: Recent developments and future directions. *Protein Sci.* **32**, e4778 (2023).

52. Cantalapiedra, C. P., Hernandez-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).

53. Zheng, J. *et al.* dbCAN3: automated carbohydrate-active enzyme and substrate annotation. *Nucleic Acids Res.* gkad328 (2023).

54. Drula, E. *et al.* The carbohydrate-active enzyme database: functions and literature. *Nucleic Acids Res.* **50**, D571--D577 (2022).

55. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science (80-.).* **379**, 1123–1130 (2023).

56. Klopfenstein, D. V. *et al.* GOATOOLS: a Python library for Gene Ontology analyses. *Sci. Rep.* **8**, 1–17 (2018).

57. Möller, M. & Stukenbrock, E. H. Evolution and genome architecture in fungal plant pathogens. *Nat. Rev. Microbiol.* **15**, 756–771 (2017).

58. Sanderson, T., Bileschi, M. L., Belanger, D. & Colwell, L. J. ProteInfer, deep neural networks for protein functional inference. *Elife* **12**, (2023).

59. Kim, G. B. *et al.* Functional annotation of enzyme-encoding genes using deep learning with transformer layers. *Nat. Commun. 2023 141* **14**, 1–11 (2023).

60. Yu, T. *et al.* Enzyme function prediction using contrastive learning. *Science (80-.).* **379**, 1358–1363 (2023).

61. Yuan, Q., Tian, C. & Yang, Y. Genome-scale annotation of protein binding sites via language model and geometric deep learning. *bioRxiv* 2023.11.02.565344 (2023) doi:10.1101/2023.11.02.565344.

62. Sperschneider, J. & Dodds, P. N. EffectorP 3.0: prediction of apoplastic and cytoplasmic effectors in fungi and oomycetes. *Mol. Plant-Microbe Interact.* **35**, 146–156 (2022).

63. Pei, J., Xiong, L., Chu, M., Guo, X. & Yan, P. Effect of intramolecular disulfide bond of bovine lactoferricin on its molecular structure and antibacterial activity against *Trueperella pyogenes* separated from cow milk with mastitis. *BMC Vet. Res.* **16**, 1–10 (2020).

64. Teufel, F. *et al.* SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nat. Biotechnol.* **40**, 1023–1025 (2022).

65. Varadi, M. *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022).

66. Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat. Methods 2022 196* **19**, 679–682 (2022).

67. Kersey, P. J. *et al.* Ensembl Genomes 2016: more genomes, more complexity. *Nucleic Acids Res.* **44**, D574--D580 (2016).

68. Grigoriev, I. V. *et al.* MycoCosm portal: Gearing up for 1000 fungal genomes. *Nucleic Acids Res.* **42**, D699–D704 (2014).

69. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019).

70. Bateman, A. *et al.* UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).

71. Osorio, D., Rondón-Villarreal, P. & Torres, R. Peptides: A package for data mining of antimicrobial peptides. *R J.* **7**, 4–14 (2015).

72. Mih, N. *et al.* ssbio: a Python framework for structural systems biology. *Bioinformatics* **34**, 2155–2157 (2018).

73. Chen, H., Gu, F. & Huang, Z. Improved Chou-Fasman method for protein secondary structure prediction. *BMC Bioinformatics* **7**, 1–11 (2006).

74. Nagarajan, R. *et al.* PDBparam: online resource for computing structural parameters of proteins. *Bioinform. Biol. Insights* **10**, 73–80 (2016).

75. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolym. Orig. Res. Biomol.* **22**, 2577–2637 (1983).

76. Touw, W. G. *et al.* A series of PDB-related databanks for everyday needs. *Nucleic Acids Res.* **43**, D364--D368 (2015).

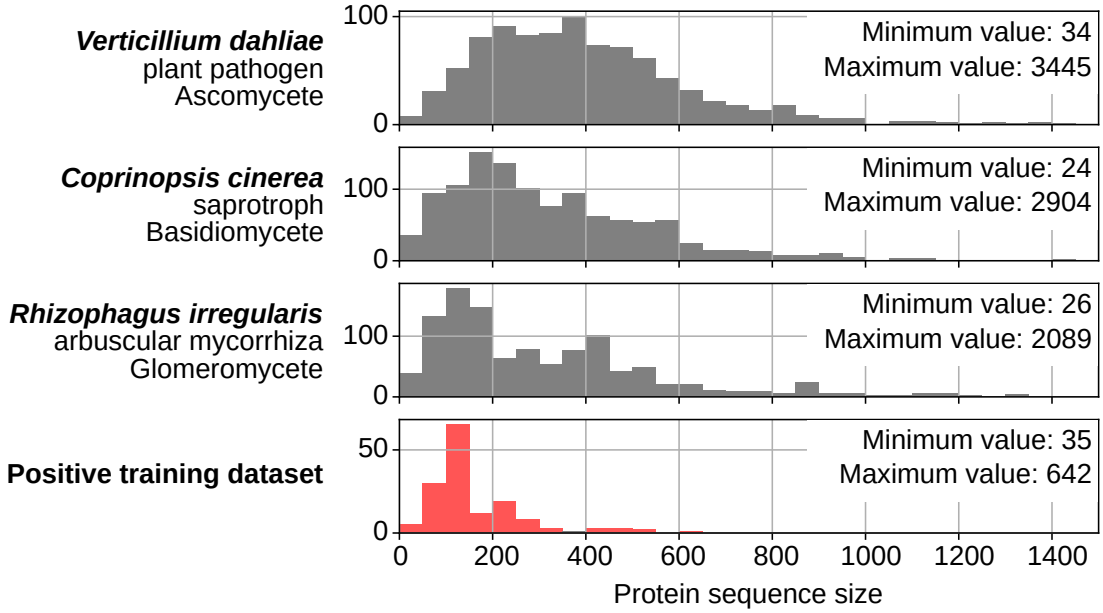
77. Le Guilloux, V., Schmidtke, P. & Tuffery, P. Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics* **10**, 1–11 (2009).

78. Liang, Y. *et al.* Research progress of reduced amino acid alphabets in protein analysis and prediction. *Comput. Struct. Biotechnol. J.* (2022).

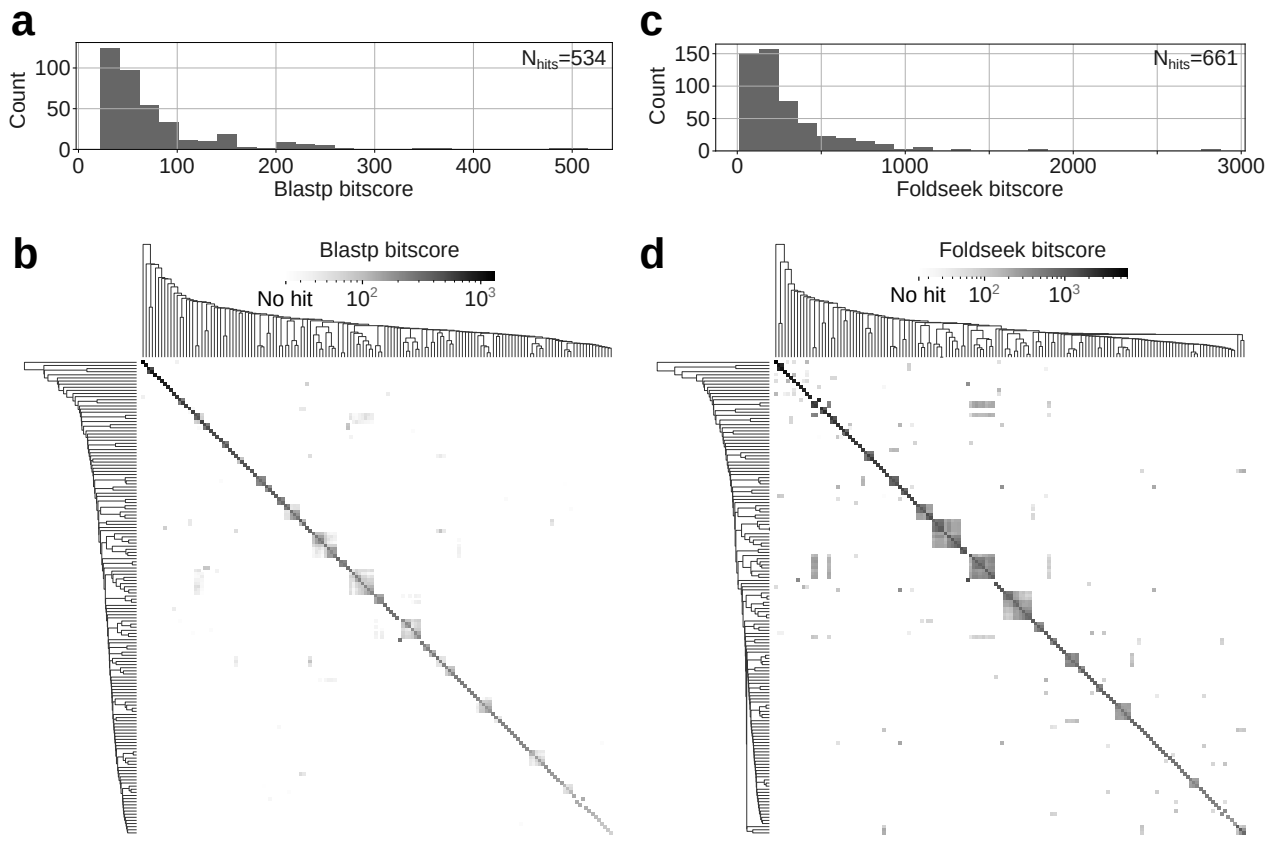
79. Figueroa, J. L. *et al.* MerCat2: a versatile k-mer counter and diversity estimator for database-independent property analysis obtained from omics data. *bioRxiv* 2022.11.22.517562 (2022). doi:10.1101/2022.11.22.517562

80. Pedregosa, F. *et al.* Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

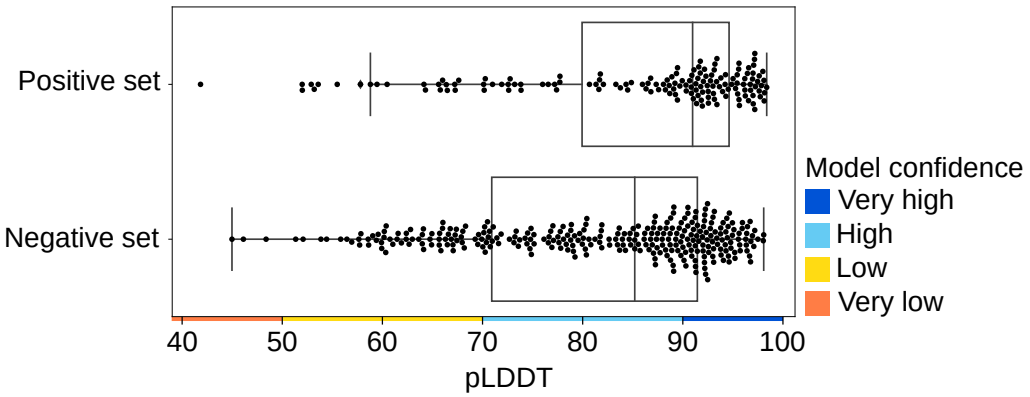
81. Hedges, L. V. Distribution theory for Glass's estimator of effect size and related estimators. *J. Educ. Stat.* **6**, 107–128 (1981).



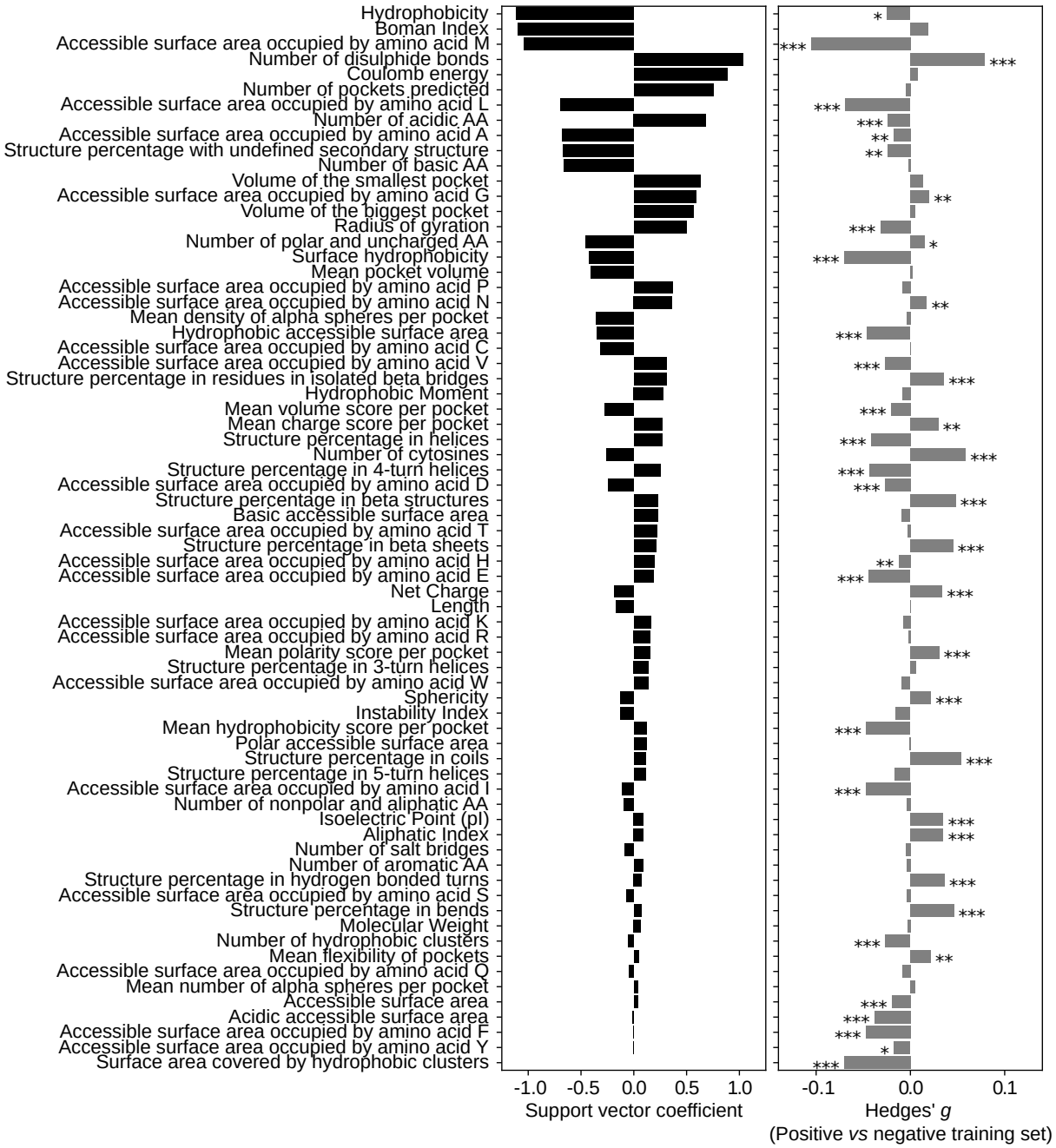
Supplementary Fig. 1: Sequence lengths in three fungal secretomes and in the literature-curated set of antimicrobial proteins. The top three histograms show mature sequence lengths in number of amino acids of secreted proteins (predicted with SignalP[64]) in three fungi selected based on their distance in the tree of life and their distinct lifestyles. Lifestyles and phyla are labeled together with species names on the right. The histogram at the bottom shows the length of sequences implemented in the literature-curated set of proteinaceous antimicrobials.



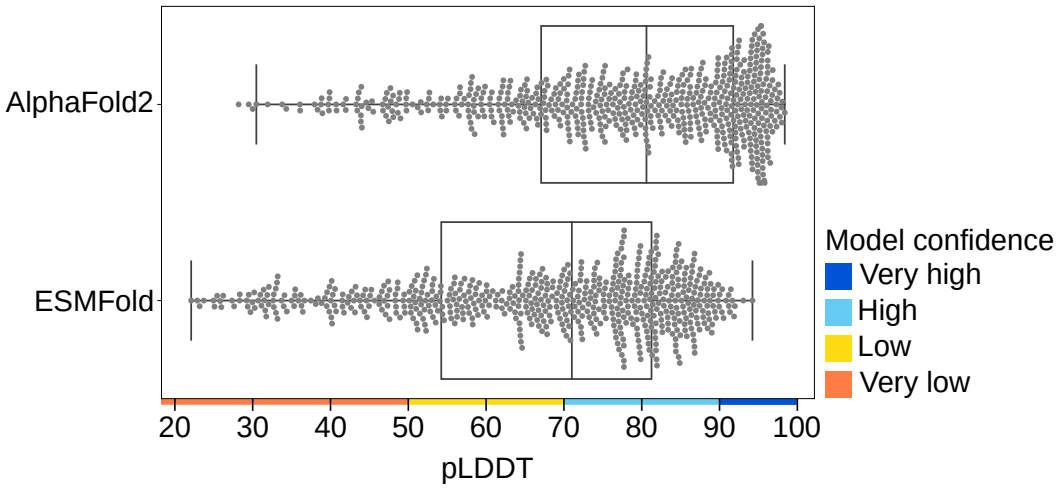
Supplementary Fig. 2: Sequence and structural similarities among proteins of the literature-curated set of antimicrobials. **a.** Distribution of blastp[48] bitscores revealing inter-protein sequence similarity in the dataset (534 non-self significant hits with evalue ≤0.05). **b.** Clustering of proteins in the dataset according to significant sequence similarity (blastp[48]; evalue ≤0.05; UPGMA hierarchical clustering by bitscore) revealing few small groups of similar proteins. **c.** Distribution of Foldseek[49] bitscores revealing inter-protein structural similarity in the dataset (661 non-self significant hits with E-value ≤0.05). d**.** Clustering of proteins in the dataset according to significant structural similarity (Foldseek[49]; E-value ≤0.05; UPGMA hierarchical clustering by bitscore) revealing more and larger groups of similar proteins than the analysis of sequence similarities (panel **b**).



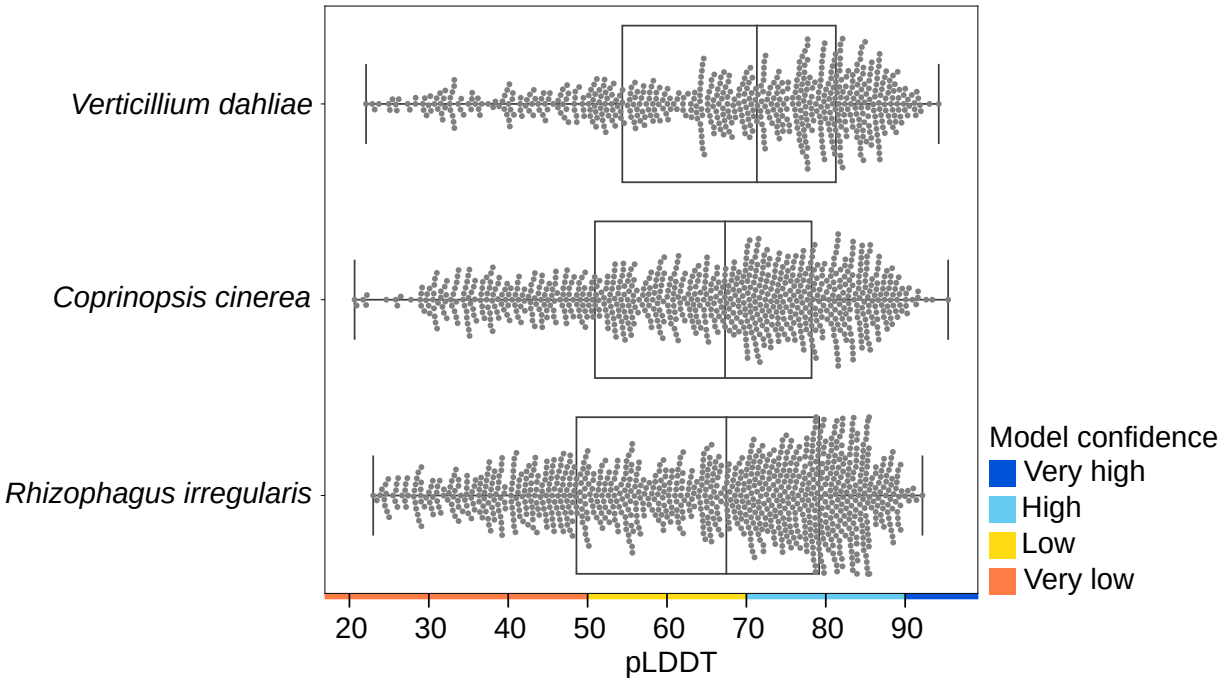
Supplementary Fig. 3: Confidence of predicted structures in our training datasets. Boxplots showing the distribution of mean pLDDT confidence scores of AlphaFold2[50]-predicted protein structures in our positive and negative training sets. The color code depicting model confidence originates from the documentation of AlphaFold2.



Supplementary Fig. 4: Physicochemical properties of proteins and their importance for antimicrobial activity prediction. Physicochemical properties implemented in our training pipeline (Fig 1c) are listed and ranked according to their importance for our SVM classifier (vector weights). The barplot in black (left) shows support vector coefficients, representing vector weights and orientation. The barplot in grey (right) shows the results of an enrichment analysis testing for significant differences between values in the positive and in the negative training set. This analysis was conducted by Mann-Whitney U test and Benjamini-Hochberg correction (FDR values depicted with asterisks: *: ≤0.05; **≤0.01; ***≤0.001) and we additionally calculated standard effect sizes (Hedges' g[81]).



Supplementary Fig. 5: Confidence of AlphaFold2- and ESMFold-predicted structures for secreted proteins of *Verticillium dahliae*. Boxplots showing the distribution of mean pLDDT confidence scores of AlphaFold2[50]- and ESMFold[55]-predicted structures for 626 non-CAZyme secreted proteins of *Verticillium dahliae*. While the secretome of *V. dahliae* includes 635 non-CAZyme proteins, AlphaFold2 failed at predicting the structures of nine of these proteins, which were therefore excluding from this analysis. The color code depicting model confidence originates from the documentation of AlphaFold2.



Supplementary Fig. 6: Confidence of predicted structures for three fungal secretomes analyzed with AMAPEC. Boxplots showing the distribution of mean pLDDT confidence scores of ESMFold[55]-predicted structures for the three fungal secretomes analyzed with AMAPEC. The color code depicting model confidence originates from the documentation of AlphaFold2[50].

Supplementary Table 1: Description of the literature-curated set of antimicrobial proteins. For each protein in the set, the table provides (1) a reference identifier; (2) the name of the protein in literature; (3) the reported antimicrobial activity in literature; (4) the group of organisms in which it is encoded; (5) the species producing the protein; (6) the publication that described the antimicrobial activity of this protein; (7) whether a secretion signal was identified and removed from the protein sequence; (8) the pLDDT confidence score for the AlphaFold2[50]-predicted structure.

Supplementary Table 2: Results of sequence and structure similarity searches between three fungal secretomes and our literature-curated set of antimicrobials. **a.** Output of a blastp[48] analysis using fungal secretomes (excluding CAZymes) as a query and the literature-curated set of antimicrobials as a subject. **b.** Output of a Foldseek[49] analysis using fungal secretomes (ESMFold[55]-predicted structures, excluding CAZymes) as a query and the literature-curated set of antimicrobials as a subject.

Supplementary Table 3: Description of the negative training set of presumably non-antimicrobial proteins. For each protein in the set, the table provides (1) a reference identifier; (2) the functional description of the protein in the UniProt database[70]; (3) the group of organisms in which it is encoded; (4) the species producing the protein; (5) the UniProt entry identifier; (6) whether a secretion signal was identified and removed from the protein sequence; (7) the pLDDT confidence score for the AlphaFold2[50]-predicted structure.

Supplementary Table 4: Properties and k-mers describing protein physicochemistry used to predict antimicrobial activity. **a.** List of 70 properties calculated from protein sequences and structures, with as a reference, the method implementing the calculation or publication introducing the formula. **b.** Reduced amino acid alphabet designed based on amino acid properties used for k-mer calling. **c.** Six k-mers found to be over- or under-represented in the sequences of the positive training set when compared to those of the negative training set, according to chi-squared testing.

Supplementary Table 5: Functional annotation of secretomes and antimicrobial activity prediction results. Secretome functional annotation outputs of emapper[52] and carbohydrate-active enzyme annotation from dbcan[53], together with the results of antimicrobial activity prediction with AMAPEC for *Verticillium dahliae* (**a**), *Coprinopsis cinerea* (**b**) and *Rhizophagus irregularis* (**c**).