(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
13 March 2025 (13.03.2025)

WIPO | PCT

(10) International Publication Number
WO 2025/054593 A2

(51) International Patent Classification:
G16B 15/30 (2019.01) *A61K 38/10* (2006.01)

(21) International Application Number:
PCT/US2024/045821

(22) International Filing Date:
09 September 2024 (09.09.2024)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
63/581,122 07 September 2023 (07.09.2023) US

(71) Applicant: **THE TRUSTEES OF THE UNIVERSITY OF PENNSYLVANIA** [US/US]; 3600 Civic Center Boulevard, 9th Floor, Philadelphia, Pennsylvania 19104 (US).

(72) Inventors: **DE LA FUENTE-NUNEZ, César**; 2119 Pine Street, Apt. 5, Philadelphia, Pennsylvania 19103 (US). **DER TOROSSIAN TORRES, Marcelo**; 1500 Locust Street, Apt. 3318, Philadelphia, Pennsylvania 19102 (US). **WAN, Fangping**; 1006 South 45th Street, Philadelphia, Pennsylvania 19104 (US).

(74) Agent: **HOFFMAN, David B** et al.; Baker & Hostetler LLP, 1735 Market Street, Suite 3300, Philadelphia, Pennsylvania 19103-7501 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:
— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*
— *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii))*

Published:
— *without international search report and to be republished upon receipt of that report (Rule 48.2(g))*

(54) Title: MOLECULAR DE-EXTINCTION OF ANTIBIOTICS ENABLED BY DEEP LEARNING

(57) **Abstract:** Provided herein are methods for identifying antimicrobial peptides deriving from extinct proteomes using a multitask deep learning approach. Also disclosed are antimicrobial peptides, and methods of treating an antimicrobial infection comprising contacting the infection with the present antimicrobial peptides. Also disclosed are methods of treating a microbial infection comprising administering to a subject in need thereof a pharmaceutically effective amount of an antimicrobial peptide according to the present disclosure.

WO 2025/054593 A2

MOLECULAR DE-EXTINCTION OF ANTIBIOTICS ENABLED BY DEEP LEARNING

GOVERNMENT RIGHTS

[0001] This invention was made with government support under GM138201 awarded by the National Institutes of Health and HDTRA1-18-1-0041, HDTRA1-21-1-0014, and HDTRA1-23-1-0001 awarded by the Defense Advanced Research Projects Agency. The government has certain rights in the invention.

CROSS-REFERENCE TO RELATED APPLICATIONS

[0002] The present application claims the benefit of priority to U.S. Provisional Application No. 63/581,122, filed September 7, 2023, the entire contents of which are incorporated herein by reference.

TECHNICAL FIELD

[0003] The present disclosure pertains to identification, synthesis, and use antibiotic peptides.

BACKGROUND

[0004] With antimicrobial resistant (AMR) infections causing approximately 1.27 million deaths annually worldwide, and projections indicating a potential 10 million annual fatalities by 2050[3] in the absence of effective new drugs, urgent measures are required to combat antibiotic resistance. Furthermore, according to the World Health Organization, by 2030, around 24 million individuals could face extreme poverty due to the high cost of treating these infections[3].

[0005] Computational approaches have been developed for the design and discovery of peptide antibiotics[4]. For example, machine learning (ML) models have been used to generate peptide sequences[5,6] and to predict antimicrobial activity[7,8], hemolysis[9], and AMR[10,11]. Recently, computational methods have been developed to discover new antibiotics through proteome mining[2,12].

SUMMARY

[0006] Provided herein are methods for identifying antimicrobial peptides deriving from extinct proteomes using a multitask deep learning approach.

[0007] Also disclosed are antimicrobial peptides, and methods of treating an antimicrobial infection comprising contacting the infection with the present antimicrobial peptides. Also disclosed are methods of treating a microbial infection comprising administering to a subject in need thereof a therapeutically effective amount of a peptide according to the present disclosure.

BRIEF DESCRIPTION OF THE DRAWINGS

[0008] FIG. 1 (SEQ ID NO: 42) depicts how molecular de-extinction of antibiotics from proteomes was accomplished using deep learning. All available proteomes of extinct organisms were mined by APEX, our deep learning algorithm. Amino acid sequences ranging from 8 to 50 amino acid residues within proteins from extinct organisms were inputted into a multitask deep learning model that trained on both public and in-house peptide data to evaluate the potential antimicrobial activity. The highest ranked peptides based on predicted antimicrobial activities were then selected and thoroughly characterized against clinically relevant pathogens both in vitro and in animal models. The mechanism of action, physicochemical features, and synergistic interactions of these peptides were also assayed. The dates reported the extinction date or period for the organisms studied.

[0009] FIGS. 2A and 2B illustrate how APEX accurately identified antimicrobials in extinct organisms. FIG. 2A provides a radar chart showing R-squared correlation in terms of species-specific antimicrobial activity prediction on an independent dataset (a held-out subset from our in-house peptide dataset) for various machine learning (ML) models. The radius reflects the R-squared value for each of the models. APEX variants outperformed the baseline ML methods with most of the pathogens analyzed. RF: random forest; GBDT: gradient boosting decision tree; ExtraTree: extra-tree regressor; ElasticNet: elastic net; LinearSVR: linear support vector regression. FIG. 2B depicts the mean of species-wise Pearson correlation of $\log_2$-transformed MICs between values obtained experimentally and predicted by various ML models. Evaluated dataset: 69 peptides were synthesized and tested.

[0010] FIGS. 3A-3B illustrate antimicrobials identified by APEX in extinct organisms and their composition and physicochemical properties. FIG. 3A provides a phylogenetic tree showing the extinct organisms scanned by APEX. Circular bars denote the $\log_{10}$-transformed average active (red) and inactive (blue) encrypted peptides discovered by APEX. A peptide was considered active when its predicted median MIC against the bacterial strains tested was ≤80 μmol L^{-1}. The values were normalized by the number of proteins per organism scanned. The organisms whose encrypted peptides were selected for validation are highlighted in bold type. Extinct organisms that presented active encrypted peptides (EPs) validated experimentally are indicated by a light red square and, within that group, those organisms encoding extinct sequences absent in extant organisms are highlighted with a dark red square. FIG. 3B shows the amino acid frequency in AEPs and MEPs compared with known AMPs from the DBAASP database. AEPs present a higher frequency of the basic residue K, the aliphatic residue V, and uncharged polar residues (M, Q, and T) than MEPs. FIGS. 3C and 3D show the distribution of two physicochemical properties for peptides with predicted antimicrobial activity (AEPs and MEPs) and AMPs from DBAASP: net charge in FIG. 3C; and normalized hydrophobicity in FIG. 3D. Net charge directly influences the initial electrostatic interactions between the peptide and negatively charged bacterial membranes, and hydrophobicity directly influences the interactions of the peptide with lipids in the membrane bilayers. Encrypted peptides from extinct organisms are slightly less hydrophobic, and similarly have a net positive charge, when compared with encrypted peptides from the modern human proteome[15] or peptides from DBAASP. Statistical significance in c and d was determined using two-tailed t-tests followed by Mann-Whitney test; p values are shown in the graph. The solid line inside each box represents the mean value obtained for each group.

[0011] FIGS. 4A and 4B pertain to mntimicrobial activity profiles of sequences from the proteomes of extinct organisms. FIG. 4A provides a heat map of the antimicrobial activities (μmol L^{-1}) of the active encrypted peptides from extinct organisms against 11 clinically relevant pathogens, including strains resistant to conventional antibiotics. Briefly, 10^6 bacterial cells and serially diluted encrypted peptides (0-128 μmol L^{-1}) were incubated at 37 °C. One day post-treatment, the optical density at 600 nm was measured in a microplate reader to evaluate bacterial growth in the presence of the encrypted peptides from extinct organisms. MIC values in the heat map are the arithmetic

mean of the replicates in each condition. FIG. 4B (SEQ ID NOS: 17, 15, 22, 25, 33 and 1) provides examples of active archaic encrypted peptides (AEPs) and modern encrypted peptides (MEPs) from various extinct organisms, their parent protein, and their activity profile against ESKAPE pathogens (*Enterococcus spp.*, *S. aureus*, *K. pneumoniae*, *A. baumannii*, *P. aeruginosa*, *E. coli*). Antimicrobial activity is expressed as the MIC (μmol L^{-1}) and activity bars are presented as $-\log_2$ MIC[13]. The data for the assays in a are the mean and the experiments were performed in three independent replicates. AEPs in a are indicated by an asterisk (*).

[0012] FIGS. 5A-5E illustrate the antimicrobial activity, mechanism of action, and synergy of antimicrobials from the proteomes of extinct organisms. FIG. 5A provides pan-bacterial Pearson and Spearman correlations of $\log_2$-transformed MICs between experimentally validated values and values predicted by APEX. FIG. 5B provides a comparison between the hit rates of APEX and the scoring function previously described by Torres *et al.*[15] to detect encrypted peptides in the modern human proteome. FIG. 5C shows cytoplasmic membrane depolarization by five encrypted peptides from extinct organisms. The *A. baumannii* membrane was more strongly depolarized by the encrypted peptides than by the antibiotic polymyxin B. FIG. 5D provides NPN permeabilization assays showing the effect of two encrypted peptides from extinct organisms on the outer membrane of *A. baumannii*. Higher permeability was observed with the encrypted peptides than with the antibiotic polymyxin B. FIG. 5E is a heat map showing interactions between encrypted peptides identified by APEX, expressed as the fractional inhibitory concentration index (FICI). Most of the tested encrypted peptide pairs from extinct organisms either synergized or had an additive effect against *A. baumannii* and *P. aeruginosa*; the latter was only tested against the peptide pair composed of equusin-1 and equusin-2 shown in the last row of the heatmap. The data for the assays in c, d, and e are the mean and the experiments were performed in three independent replicates. AEPs in e are indicated by an asterisk (*).

[0013] FIGS. 6A-6D pertain to studies of the anti-infective activity of encrypted peptides in animal models. FIG. 6A provides a schematic of the skin abscess mouse model used to assess the anti-infective activity of selected encrypted peptides from extinct organisms (n = 6) against *A. baumannii* ATCC 19606. FIG. 6B shows how the encrypted peptides mammuthusin-2 (*Mammuthus primigenius*), hydrodamin-1 (*Hydrodamalis*

gigas), megalocerin-1 (*Megalocerus sp.*), elephasin-2 (*Elephas antiquus*), and mylodonin-2 (*Mylodon darwinii*), administered at their MIC in a single dose inhibited the proliferation of the infection for up to four days after treatment compared to the untreated control group. Elephasin-2 and mylodonin-2 cleared the infection in some of the mice, with activity comparable to that of the antibiotic used as control, polymyxin B. FIG. 6C provides a schematic of the neutropenic thigh infection mouse model in which encrypted peptides from extinct organisms were injected intraperitoneally. Anti-infective activity against *A. baumannii* ATCC 19606 was assessed 2 and 4 days after intraperitoneal administration (n = 6). FIG. 6D shows how, two days after intraperitoneal injection, mylodonin-2 at its MIC reduced *A. baumannii* ATCC19606 infection as much as polymyxin B, compared to the untreated control group. Four days post-treatment, mammuthusin-2 and elephasin-2 showed the same level of activity as polymyxin B. Statistical significance in b and d was determined using one-way ANOVA followed by Dunnett's test; p values are shown in the graph. For the boxplots, the center line represents the mean, the box limits the first and third quartiles, and the whiskers (minima and maxima) 1.5× the interquartile range. The solid line inside each box represents the mean value obtained for each group.

[0014] FIG. 7 provides a schematic illustration of APEX.

[0015] FIG. 8. shows R-squared scores of various ML models on cross-validation (CV) set.

[0016] FIG. 9 shows Pearson correlation scores of various ML models on cross-validation (CV) set.

[0017] FIG. 10 shows Spearman correlation scores of various ML models on cross-validation (CV) set.

[0018] FIG. 11 depicts R-squared scores of various ML models on cross-validation (CV) set.

[0019] FIG. 12 shows Pearson correlation scores of various ML models on CV set.

[0020] FIG. 13 provides Spearman correlation scores of various ML models on CV set.

[0021] FIG. 14 shows the relationship between R-squared and the number of APEX models used in ensemble learning on the CV set.

[0022] FIG. 15 depicts the relationship between Pearson correlation and the number of APEX models used in ensemble learning on the CV set.

[0023] FIG. 16 shows the relationship between Spearman correlation and the number of APEX models used in ensemble learning on CV set.

[0024] FIG. 17 shows an ablation study of the multitask learning strategy of APEX in terms of R-squared on the CV set.

[0025] FIG. 18 shows an ablation study of the multitask learning strategy of APEX in terms of Pearson correlation on CV set.

[0026] FIG. 19 shows an ablation study of the multitask learning strategy of APEX in terms of Spearman correlation on CV set.

[0027] FIG. 20 provides Pearson correlation scores of various ML models on an independent set.

[0028] FIG. 21 provides Spearman correlation scores of various ML models on an independent set.

[0029] FIG. 22 provides R-squared scores of various ML models on an independent set.

[0030] FIG. 23 provides Pearson correlation scores of various ML models on an independent set.

[0031] FIG. 24 shows Spearman correlation scores of various ML models on an independent set.

[0032] FIG. 25 provides R-squared scores of ensemble APEX v2 and v1 on an independent set.

[0033] FIG. 26 shows Pearson correlation scores of ensemble APEX v2 and v1 on an independent set.

[0034] FIG. 27 provides Spearman correlation scores of ensemble APEX v2 and v1 on an independent set.

[0035] FIG. 28 shows Pearson correlation scores of ensemble APEX v2 and the scoring function used to identify modern human encrypted peptides.

[0036] FIG. 29 provides Spearman correlation scores of ensemble APEX v2 and the scoring function used to identify modern human encrypted peptides.

[0037] FIG. 30 shows Pearson correlation scores of various ML models used to identify modern human encrypted peptides.

[0038] FIG. 31 provides Spearman correlation scores of various ML models used to identify modern human encrypted peptides.

[0039] FIG. 32 shows sequence space exploration using a similarity matrix.

[0040] FIG. 33 depicts the relative abundance of the amino acid content of encrypted peptides (EPs) from the modern human proteome identified by APEX (top) and the scoring function (bottom).

[0041] FIG. 34 shows the relative abundance of the amino acid content of encrypted peptides (EPs) identified by APEX from the proteomes of extinct organisms (top) compared to known AMPs from DBAASP (bottom).

[0042] FIG. 35 shows the relative abundance of the amino acid content of archaic encrypted peptides (AEPs) identified by APEX from the proteomes of extinct organisms (top) compared to known AMPs from DBAASP (bottom).

[0043] FIG. 36 depicts the relative abundance of the amino acid content of MEPs identified by APEX from the proteomes of extinct organisms compared to known AMPs from DBAASP.

[0044] FIG. 37 shows the relative abundance of the amino acid content of AEPs and MEPs identified by APEX from the proteomes of extinct organisms.

[0045] FIGS. 38A-F depict physicochemical features of AEPs and MEPs identified by APEX in extinct organisms compared to AMPs from DBAASP.

[0046] FIGS. 39A-D concern the secondary structure of active EPs predicted by the scoring function and APEX in helical inducer medium.

[0047] FIG. 40 shows the antimicrobial activity of encrypted peptides from extinct organisms predicted by the scoring function.

[0048] FIG. 41 illustrates sequence space exploration using a similarity matrix containing the 69 encrypted peptides discovered by APEX selected for further experimental validation compared to peptide sequences from DBAASP.

[0049] FIG. 42 shows predicted vs. experimental MICs for A. baumannii ATCC 19606 of the encrypted peptides identified by APEX.

[0050] FIG. 43 shows predicted vs. experimental MICs for E. coli AIC221 of the encrypted peptides identified by APEX.

[0051] FIG. 44 shows predicted vs experimental MICs for E. coli AIC222 of the encrypted peptides identified by APEX.

[0052] FIG. 45 shows predicted vs. experimental MICs for E. coli ATCC 11775 of the encrypted peptides identified by APEX.

[0053] FIG. 46 shows predicted vs. experimental MICs for K. pneumoniae ATCC 13883 of the encrypted peptides identified by APEX.

[0054] FIG. 47 depicts predicted vs. experimental MICs for P. aeruginosa PA14 of the encrypted peptides identified by APEX.

[0055] FIG. 48 shows predicted vs. experimental MICs for P. aeruginosa PAO1 of the encrypted peptides identified by APEX.

[0056] FIG. 49 provides predicted vs. experimental MICs for methicillin-resistant S. aureus ATCC BAA-1556 of the encrypted peptides identified by APEX.

[0057] FIG. 50 shows predicted vs. experimental MICs for S. aureus ATCC 12600 of the encrypted peptides identified by APEX.

[0058] FIG. 51 provides predicted vs. experimental MICs for vancomycin-resistant E. faecalis ATCC 700802 of the encrypted peptides identified by APEX.

[0059] FIG. 52 provides predicted vs. experimental MICs for vancomycin-resistant E. faecium ATCC 700221 of the encrypted peptides identified by APEX.

[0060] FIGS. 53A-D pertain to cytoplasmic membrane depolarization of A. baumannii and P. aeruginosa triggered by AEPs and MEPs identified by APEX.

[0061] FIGS. 54A-D pertain to outer membrane permeabilization of A. baumannii and P. aeruginosa cell membranes caused by encrypted peptides from extinct organisms.

[0062] FIG. 55 depicts the synergy between peptide molecules from extinct organisms.

[0063] FIG. 56 shows the results of assays of the resistance to proteolytic degradation.

[0064] FIG. 57 depicts the results of weight change monitoring in the skin abscess mouse model infected with A. baumannii.

[0065] FIG. 58 depicts the results of weight change monitoring in the thigh mouse model infected with A. baumannii.

DETAILED DESCRIPTION OF ILLUSTRATIVE EMBODIMENTS

[0066] The presently disclosed inventive subject matter may be understood more readily by reference to the following detailed description taken in connection with the accompanying figures and examples, which form a part of this disclosure. It is to be understood that these inventions are not limited to the specific products, methods, conditions or parameters described and/or shown herein, and that the terminology used herein is for the purpose of describing particular embodiments by way of example only and is not intended to be limiting of the claimed inventions.

[0067] The entire disclosures of each patent, patent application, and publication cited or described in this document are hereby incorporated herein by reference.

[0068] As employed above and throughout the disclosure, the following terms and abbreviations, unless otherwise indicated, shall be understood to have the following meanings.

[0069] In the present disclosure the singular forms "a," "an," and "the" include the plural reference, and reference to a particular numerical value includes at least that particular value, unless the context clearly indicates otherwise. Thus, for example, a reference to "a treatment" is a reference to one or more of such treatments and equivalents thereof known to those skilled in the art, and so forth. Furthermore, when indicating that a certain element "may be" X, Y, or Z, it is not intended by such usage to exclude in all instances other choices for the element.

[0070] When values are expressed as approximations, by use of the antecedent "about," it will be understood that the particular value forms another embodiment. As usedherein, "about X" (where X is a numerical value) preferably refers to ±10% of the recited value, inclusive. For example, the phrase "about 8" preferably refers to a value of 7.2 to 8.8, inclusive; as another example, the phrase "about 8%" preferably refers to a value of 7.2% to 8.8%, inclusive. Where present, all ranges are inclusive and combinable. For example, when a range of "1 to 5" is recited, the recited range should be construed as optionally including ranges "1 to 4", "1 to 3", "1-2", "1-2 & 4-5", "1-3 & 5", and the like. In addition, when a list of alternatives is positively provided, such a listing can also include embodiments where any of the alternatives may be excluded. For example, when a range of "1 to 5" is described, such a description can support situations whereby any of 1, 2, 3, 4, or 5 are excluded; thus, a recitation of "1 to 5" may support "1 and 3-5, but not 2",

or simply "wherein 2 is not included." The phrase "at least about x" is intended to embrace both "about x" and "at least x". It is also understood that where a parameter range is provided, all integers within that range, and tenths thereof, are also provided by the invention. For example, "2-5 hours" includes 2 hours, 2.1 hours, 2.2 hours, 2.3 hours, etc., up to 5 hours.

[0071] Publications with potential relevance to the presently disclosed subject matter are cited in the present disclosure using superscripted numerals that correspond to the numbered references that are listed in the present disclosure under the heading "References", *infra*.

[0072] Biological molecules serve as invaluable records of evolutionary history[1] and may be able to provide blueprints for therapeutic design. The present inventors recently introduced the term molecular de-extinction[2], referring to the resurrection of extinct molecules of life that are no longer encoded by living organisms. By uncovering a new sequence space of previously unexplored molecules, molecular de-extinction offers a promising approach to tackle contemporary challenges. Specifically, this disclosure proposes molecular de-extinction as a framework for drug discovery, aiming to address the urgent global health issue of antimicrobial-resistant (AMR) infections.

[0073] This disclosure introduces **A**ntibiotic **P**eptide de-**Ex**tinction (APEX, **Fig. 1**), a new multitask deep learning (DL) approach. Using APEX, the inventors systematically mined all available proteomes of extinct organisms (the "extinctome") to discover potential antimicrobial peptides. This effort led to the identification of 37,176 EPs with predicted antibiotic activity (**Data S1**). Of these peptides, 11,035 were classified as archaic EPs (AEPs), meaning they were absent in extant organisms, while the rest, referred to as modern EPs (MEPs), were present in both extant and extinct organisms. These peptides (AEPs and MEPs) were computationally predicted to exhibit antimicrobial activity based on our broad-spectrum classification threshold (median MIC $\leq$80 μmol L^{-1}).

[0074] To validate their functionality, the inventors synthesized a diverse set of 69 EPs (**Data S2**), comprising 20 AEPs and 49 MEPs, covering a wide range of peptide sequences. These peptides were subjected to thorough *in vitro* characterization, including assessment of antimicrobial activity, mechanism of action, secondary structure, and synergy. The synthesis strategy ensured that no more than three peptides from the same extinct organism were included, both for AEPs and MEPs, to broaden organismal

coverage. The assays yielded 13 AEPs and 28 MEPs that displayed antimicrobial activity *in vitro*. These active peptides are described *infra* in Table 1.

[0075] To expand the analysis to *in vivo* settings, three AEPs and four MEPs were evaluated in two preclinical mouse models infected with *Acinetobacter baumannii*. Notably, two AEPs and one MEP exhibited anti-infective activity comparable to polymyxin B, a widely used antibiotic, under physiologically relevant conditions[13]. These findings demonstrate the potential of the identified peptides as effective antimicrobial agents.

[0076] This study illustrates the power of deep learning in propelling the field of molecular de-extinction and its potential as a drug discovery framework. The findings highlight the valuable insights that can be gained from molecules of the past, which hold promise for addressing present-day challenges and providing benefit in the present.

[0077] Accordingly, provided herein are antimicrobial peptides having an amino acid sequence of any of SEQ ID NOs:1-41. Also disclosed are compositions comprising an antimicrobial peptide having an amino acid sequence of any of SEQ ID NOs:1-41, or an antimicrobial peptide that has been identified using a presently disclosed method, and a pharmaceutically acceptable carrier, diluent, or excipient. In some embodiments, the compositions may include two or more peptides of SEQ ID NOs:1-41 or two more antimicrobial peptides that have been identified using a presently disclosed method.

[0078] The present disclosure also provides methods treating an antimicrobial infection comprising contacting the infection with a therapeutically effective amount of an antimicrobial peptide that has been identified according to a disclosed method, such as an antimicrobial peptide of any one of SEQ ID NOs:1-41, or a composition comprising a therapeutically effective amount of an antimicrobial peptide that has been identified according to a disclosed method, such as an antimicrobial peptide of any one of SEQ ID NOs:1-41.

[0079] As used herein, the phrase "therapeutically effective amount" refers to the amount of active agent (here, the antimicrobial peptide) that elicits the biological or medicinal response that is being sought in a tissue, system, animal, individual or human by a researcher, veterinarian, medical doctor or other clinician, which includes one or more of the following:

(1) at least partially preventing the disease or condition or a symptom thereof; for example, preventing a disease, condition or disorder in an individual who may be predisposed to the disease, condition or disorder but does not yet experience or display the pathology or symptomatology of the disease;

(2) inhibiting the disease or condition; for example, inhibiting a disease, condition or disorder in an individual who is experiencing or displaying the pathology or symptomatology of the disease, condition or disorder (i.e., including arresting further development of the pathology and/or symptomatology); and

(3) at least partially ameliorating the disease or condition; for example, ameliorating a disease, condition or disorder in an individual who is experiencing or displaying the pathology or symptomatology of the disease, condition or disorder (i.e., including reversing the pathology and/or symptomatology).

[0080] The antimicrobial peptides that are administered, contacted with a biofilm, or included in a composition to the present disclosure may be provided in a composition that is formulated for any type of administration. For example, the compositions may be formulated for administration orally, topically, parenterally, enterally, or by inhalation (*e.g.*, intranasally). The active agent may be formulated for neat administration, or in combination with conventional pharmaceutical carriers, diluents, or excipients, which may be liquid or solid. The applicable solid carrier, diluent, or excipient may function as, among other things, a binder, disintegrant, filler, lubricant, glidant, compression aid, processing aid, color, sweetener, preservative, suspensing/dispersing agent, tablet-disintegrating agent, encapsulating material, film former or coating, flavoring agent, or printing ink. Any material used in preparing any dosage unit form is preferably pharmaceutically pure and substantially non-toxic in the amounts employed. In addition, the active agent may be incorporated into sustained-release preparations and formulations. Administration in this respect includes administration by, *inter alia*, the following routes: intravenous, intramuscular, subcutaneous, intraocular, intrasynovial, transepithelial including transdermal, ophthalmic, sublingual and buccal; topically including ophthalmic, dermal, ocular, rectal and nasal inhalation via insufflation, aerosol, and rectal systemic.

[0081] In powders, the carrier, diluent, or excipient may be a finely divided solid that is in admixture with the finely divided active ingredient. In tablets, the active

ingredient is mixed with a carrier, diluent or excipient having the necessary compression properties in suitable proportions and compacted in the shape and size desired. For oral therapeutic administration, the active compound may be incorporated with the carrier, diluent, or excipient and used in the form of ingestible tablets, buccal tablets, troches, capsules, elixirs, suspensions, syrups, wafers, and the like. The amount of active agent(s) in such therapeutically useful compositions is preferably such that a suitable dosage will be obtained.

[0082] Liquid carriers, diluents, or excipients may be used in preparing solutions, suspensions, emulsions, syrups, elixirs, and the like. The active ingredient of this invention can be dissolved or suspended in a pharmaceutically acceptable liquid such as water, an organic solvent, a mixture of both, or pharmaceutically acceptable oils or fat. The liquid carrier, excipient, or diluent can contain other suitable pharmaceutical additives such as solubilizers, emulsifiers, buffers, preservatives, sweeteners, flavoring agents, suspending agents, thickening agents, colors, viscosity regulators, stabilizers, or osmo-regulators.

[0083] Suitable solid carriers, diluents, and excipients may include, for example, calcium phosphate, silicon dioxide, magnesium stearate, talc, sugars, lactose, dextrin, starch, gelatin, cellulose, methyl cellulose, ethylcellulose, sodium carboxymethyl cellulose, microcrystalline cellulose, polyvinylpyrrolidine, low melting waxes, ion exchange resins, croscarmellose carbon, acacia, pregelatinized starch, crospovidone, HPMC, povidone, titanium dioxide, polycrystalline cellulose, aluminum methahydroxide, agar-agar, tragacanth, or mixtures thereof.

[0084] Suitable examples of liquid carriers, diluents and excipients, for example, for oral, topical, or parenteral administration, include water (particularly containing additives as above, e.g. cellulose derivatives, preferably sodium carboxymethyl cellulose solution), alcohols (including monohydric alcohols and polyhydric alcohols, e.g. glycols) and their derivatives, and oils (e.g. fractionated coconut oil and arachis oil), or mixtures thereof.

[0085] For parenteral administration, the carrier, diluent, or excipient can also be an oily ester such as ethyl oleate and isopropyl myristate. Also contemplated are sterile liquid carriers, diluents, or excipients, which are used in sterile liquid form compositions for parenteral administration. Solutions of the active agents can be prepared in water

suitably mixed with a surfactant, such as hydroxypropylcellulose. A dispersion can also be prepared in glycerol, liquid polyethylene glycols, and mixtures thereof and in oils. Under ordinary conditions of storage and use, these preparations may contain a preservative to prevent the growth of microorganisms.

[0086] The pharmaceutical forms suitable for injectable use include, for example, sterile aqueous solutions or dispersions and sterile powders for the extemporaneous preparation of sterile injectable solutions or dispersions. In all cases, the form is preferably sterile and fluid to provide easy syringability. It is preferably stable under the conditions of manufacture and storage and is preferably preserved against the contaminating action of microorganisms such as bacteria and fungi. The carrier, diluent, or excipient may be a solvent or dispersion medium containing, for example, water, ethanol, polyol (for example, glycerol, propylene glycol, liquid polyethylene glycol and the like), suitable mixtures thereof, and vegetable oils. The proper fluidity can be maintained, for example, by the use of a coating, such as lecithin, by the maintenance of the required particle size in the case of a dispersion, and by the use of surfactants. The prevention of the action of microorganisms may be achieved by various antibacterial and antifungal agents, for example, parabens, chlorobutanol, phenol, sorbic acid, thimerosal and the like. In some instances, the antimicrobial peptides themselves may be sufficient to prevent contamination by microorganisms. In many cases, it will be preferable to include isotonic agents, for example, sugars or sodium chloride. Prolonged absorption of the injectable compositions may be achieved by the use of agents delaying absorption, for example, aluminum monostearate and gelatin.

[0087] Sterile injectable solutions may be prepared by incorporating the active agent in the pharmaceutically appropriate amounts, in the appropriate solvent, with various of the other ingredients enumerated above, as required, followed by filtered sterilization. Generally, dispersions may be prepared by incorporating the sterilized active ingredient into a sterile vehicle which contains the basic dispersion medium and the required other ingredients from those enumerated above. In the case of sterile powders for the preparation of sterile injectable solutions, the preferred methods of preparation may include vacuum drying and freeze drying techniques that yield a powder of the active ingredient or ingredients, plus any additional desired ingredient from the previously sterile-filtered solution thereof.

[0088] Thus, an antimicrobial peptide may be in the present compositions and methods in an effective amount by any of the conventional techniques well-established in the medical field. For example, the administration may be in the amount of about 0.1 mg/day to about 500 mg per day. In some embodiments, the administration may be in the amount of about 250 mg/kg/day. Thus, administration may be in the amount of about 0.1 mg/day, about 0.5 mg/day, about 1.0 mg/day, about 5 mg/day, about 10 mg/day, about 20 mg/day, about 50 mg/day, about 100 mg/day, about 200 mg/day, about 250 mg/day, about 300 mg/day, or about 500 mg/day.

[0089] Also disclosed are methods comprising contacting a biofilm with an effective amount of an antimicrobial peptide that has been identified according to a presently disclosed method. In some embodiments, the antimicrobial peptide comprises one or more of SEQ ID NOs: 1-41. Such methods may be effective to remove or reduce the presence of an unwanted biofilm, such as in hospitals or other medical settings, in sewer and filtration systems, in industrial settings, on equipment involved in food preparation or manufacture, in aquaculture or hydroponics, or in any other context that is prone to unwanted biofilm formation.

[0090] In accordance with the methods of treating a microbial infection in a subject or the methods comprising contacting a biofilm according to the present disclosure, microbes against which the present antimicrobial peptides are effective may be, for example, any unicellular organism, such as gram-negative bacteria, gram-positive bacteria, protozoa, viruses, bacteriophages, and archaea. The present peptides can have an antimicrobial effect with respect to any such microbe. Examples of bacteria against which the present compounds are effective to cause reduction in numbers include gram positive bacteria and gram negative bacteria, for example, *Salmonella enterica*, *Listeria monocytogenes*, *Escherichia coli*, *Clostridium botulinum*, *Clostridium difficile*, *Campylobacter*, *Bacillus cereus*, *Vibrio parahaemolyticus*, *Vibrio cholerae*, *Vibrio vulnificus*, *Staphylococcus aureus*, *Yersinia enterocolitica*, *Shigella*, *Moraxella* spp., *Helicobacter*, *Stenotrophomonas*, *Bdellovibrio*, *Legionella* spp. (*e.g.*, *pneumophila*), *Neisseria gonorrhoeae*, *Neisseria meningitidis*, *Haemophilus influenzae*, *Acinetobacter baumannii*, *Klebsiella pneumoniae*, *Pseudomonas aeruginosa*, *Proteus mirabilis*, *Enterobacter cloacae*, *Enterococcus faecium*, *Serratia marcescens*, *Helicobacter pylori*, *Salmonella enteritidis*, *Salmonella typhi*, and combinations thereof. Examples of

Salmonella enterica serovars that can be reduced using the compounds of the disclosure include, for example, *Salmonella enteriditis*, *Salmonella typhimurium*, *Salmonella poona*, *Salmonella heidelberg*, and *Salmonella anatum*. Exemplary viruses against which the present peptides are effective to cause reduction in numbers include coronaviruses, rhinoviruses, and influenza viruses.

[0091] Also disclosed herein are methods of identifying an antimicrobial peptide comprising: training a deep learning model using a peptide dataset; representing peptides as inputs into the deep learning model; utilizing a hybrid of recurrent and attention neural networks for the deep learning model in order to extract peptide sequence information from the peptide dataset; selecting hyperparameters for the deep learning model; mining candidate peptides from a proteome of an extinct organism; and, screening the candidate peptides in order to identify an antimicrobial peptide from the candidate peptides.

[0092] In some embodiments, the methods comprise further constraining the training of the deep learning model using bacterial distance curation.

[0093] In some embodiments, the mining includes setting a selectivity scoring method for selecting and filtering candidate antimicrobial peptides.

[0094] The present disclosure also pertains to methods of identifying an antimicrobial peptide comprising training and using a deep learning model according to steps disclosed in the present specification and figures.

[0095] Also provided are methods of identifying an antimicrobial peptide comprising training and using a deep learning model according to steps depicted in FIG. 1.

Examples

[0096] The present invention is further defined in the following Examples. It should be understood that the examples, while indicating preferred embodiments of the invention, are given by way of illustration only, and should not be construed as limiting the appended claims. From the above discussion and the examples, one skilled in the art can ascertain the essential characteristics of this invention, and without departing from the spirit and scope thereof, can make various changes and modifications of the invention to adapt it to various usages and conditions.

Example 1 – Antimicrobial activity prediction from sequence using deep learning

[0097] Recent computational advances have enabled the exploration of proteomes for antibiotic discovery[3,15]. To accelerate these efforts, the present inventors have developed APEX, a deep learning model that employs a multitask learning architecture to predict the antimicrobial activity of peptides (**FIG. 7**). APEX was trained on peptide data from both the inventors' in-house dataset and from the publicly available Database of Antimicrobial Activity and Structure of Peptides (DBAASP)[16]. APEX utilizes an encoder neural network, combining recurrent and attention neural networks (**FIG. 7**), to extract hidden features from peptide sequences. The encoder neural network was then coupled with multiple downstream neural networks to predict antimicrobial activity according to the peptide source (*i.e.*, from in-house or public datasets). Specifically, in the inventors' approach, the extracted hidden features were fed into two separate, fully connected neural networks (FCNNs): one neural network was trained on the inventors' in-house peptide dataset and used to predict antimicrobial activity against specific bacterial strains (*i.e.*, a regression problem); the other neural network was trained on publicly available antimicrobial peptides (AMPs) and inactive peptides (referred here as non-AMPs) derived from the DBAASP dataset[16] to perform a binary classification (*i.e.*, a classification problem). The present inventors defined as non-AMPs those peptides that were not active at the range of concentrations selected as threshold, *i.e.*, MIC >30 μmol L^{-1} (***Publicly available AMP sequences*** in **Methods**). Any publicly available sequences that overlapped with the inventors' in-house dataset were removed from the model training to prevent label information leakage. Since the encoder neural network was trained on both the in-house and public datasets, the incorporation of the latter FCNN served as a data augmentation strategy to improve prediction performance.

[0098] To train APEX, the present inventors utilized a combination of 988 in-house peptides and 5,093 and 5,500 publicly available AMPs and non-AMPs, respectively, obtained from DBAASP[16]. The inventors' in-house dataset included 14,738 antimicrobial activity data values obtained from 34 bacterial strains. To assess APEX's antimicrobial prediction performance, the present inventors randomly split the inventors' in-house dataset into a cross validation (CV) set and an independent set, consisting of 790 and 198 peptides, respectively. Five-fold CV was first used to tune the hyperparameters on

the CV set, while the independent set was used to evaluate the final prediction performance of ML models trained on the CV set with determined hyperparameters.

[0099] To compare the performance of the inventors' deep learning approach with simple ML predictors, the present inventors implemented several baseline ML models, including elastic net, linear support vector regression, extra-trees regressor, random forest, and gradient boosting decision tree, and trained and evaluated them on the same datasets. The hyperparameter ranges searched for each ML model are provided in **Supplementary Tables 1-4**. On the CV set, the inventors' APEX model with the best hyperparameter combination outperformed all baseline ML models in terms of predicted activity for most bacteria, focusing particularly on 11 bacterial pathogens known as the ESKAPEE pathogens. These pathogens are classified by the World Health Organization as the most dangerous threats to the inventors' society, and include *Enterococcus faecium, Staphylococcus aureus, Klebsiella pneumoniae, A. baumannii, P. aeruginosa, Enterobacter* spp., and *Escherichia coli*[17].

[00100] Specifically, APEX outperformed all baseline ML models on most pathogen-specific MIC predictions in terms of R-squared Pearson, and Spearman correlations (*i.e.*, single APEX in **FIGS. 8-13** and **Supplementary Tables 5-7)**. The average R-squared, Pearson, and Spearman correlations of the baseline ML methods were, at best, 0.378, 0.584, and 0.523, respectively. Compared to the baseline, the single best obtained similar R-squared = 0.369, and better Pearson correlation = 0.621, and Spearman correlation = 0.556 (**Supplementary Tables 5-7**).

[00101] To improve the prediction performance, the present inventors adopted an ensemble learning approach by selecting the top eight APEX models (with different neural network architectures and training strategies; for details, see subtopic "***Hyperparameter tuning, model evaluation and ensemble learning***" in the **Methods** section, **FIGS. 14-16**) ranked by the average R-squared on the CV set and obtained the final predictions by averaging the predictions from these models. The ensemble learning approach (*i.e.*, ensemble APEX v1) increased the prediction performance to 0.473, 0.669, and 0.594 in terms of R-squared, Pearson correlation, and Spearman correlation, respectively (**FIGS. 8-13** and **Supplementary Tables 5-7**).

[00102] APEX involved the following multitask training steps: (1) using a single FCNN to simultaneously predict peptide antimicrobial activity for the 34 strains tested; (2)

augmenting the training data by incorporating another FCNN to predict whether peptides from public databases (either AMPs or non-AMPs) were antimicrobial; and (3) the present inventors imposed a multitask training constraint on the learnable weights of the last layer in the species-specific antimicrobial prediction FCNN. Briefly, this last constraint encouraged the model to give similar prediction results for similar bacteria (defined by having shorter phylogenetic distance from each other).

[00103] To evaluate the effectiveness of adding publicly available AMPs/non-AMPs data into the inventors' training as well as that of using the multitask training constraint, the present inventors conducted an ablation study by dropping these two parts during training and evaluated the corresponding prediction performance on the CV set. The present inventors observed that dropping the publicly available AMPs/non-AMPs data from the training set substantially decreased prediction performance (**FIGS. 17-19 and Supplementary Tables 8-10**). Adding the multitask training constraint led to either increased or decreased prediction performance depending on the target bacterial strain (**FIGs. 17-19 and Supplementary Tables 8-10**). Thus, the present inventors treated the presence or absence of the multitask training constraint as an additional hyperparameter, and the present inventors allowed subsequent hyperparameter tuning to decide whether to use the constraint or not. Of note, among the top eight selected APEX models, six used the constraint during model training (**Supplementary Table 11**).

[00104] Because model selection was based on the CV set, evaluating the prediction performance on the CV set alone may overestimate the generalization ability of APEX and the baseline models. Therefore, the present inventors trained the ML models with the determined hyperparameters on the whole CV set and evaluated their prediction performance on the independent set. Similar to the results obtained on the CV set, ensemble APEX v1 achieved an R-squared value of 0.520, a Pearson correlation of 0.706, and a Spearman correlation of 0.582 on average (**FIG. 2a, FIGS. 16-20** and **Supplementary Tables 12-14**), outperforming all baseline ML methods. In practice, to make prediction results more robust, a single ML model may be trained with different random seeds. The present inventors averaged the prediction results from all model copies to counter the potential stochastic behavior caused by the choice of random seeds. For each APEX model selected, the present inventors trained five copies with different random seeds and created a second ensemble learning version (ensemble APEX v2) with 40 APEX

models (i.e., eight APEX models × five copies). This ensemble learning approach increased the prediction performance to 0.546, 0.728 and 0.607 in terms of R-squared, Pearson correlation, and Spearman correlation, respectively (**FIG. 2a, FIGS. 20-27** and **Supplementary Tables 12-14**).

[00105] The present inventors then tested APEX's predictive power compared to that of a scoring function used previously to discover encrypted peptide antibiotics in the modern human proteome[15]. Given a peptide sequence, the scoring function[18] uses hydrophobicity and net charge to compute a predictive score of antimicrobial potential. Since the 56 human peptide antibiotics validated experimentally for antimicrobial activity in the inventors' previous work[15] are part of the inventors' in-house dataset, the present inventors used them here as the test set and used the rest of the inventors' in-house dataset (932 peptides) for model training and selection. For the dataset consisting of the 56 validated human EPs, the ensemble APEX v2 achieved highest values for Pearson and Spearman correlations in most cases (**FIGS. 28-31** and **Supplementary Tables 15-16**). Lastly, during the subsequent prediction of the antimicrobial activities for the 69 synthesized EPs using ML models trained on the entire in-house dataset, APEX outperformed all baseline ML methods. Notably, APEX achieved the highest Pearson correlation for MIC prediction (**FIG. 2b**, more details on ***In vitro antimicrobial activity of antibiotic molecules from extinct organisms***). Collectively, these results substantiate the inventors' computational validation of APEX as the most accurate model for antimicrobial activity prediction in comparison to all the other models tested in this work (**FIG. 2a, b and FIGS. 8-31**). Based on these results, the present inventors decided to use the ensemble APEX v2 model (hereafter referred to as APEX) to mine extinct proteomes for EPs.

[00106] *Mining the extinctome for antibiotics and sequence space exploration.* To mine the extinctome using APEX, the inventors first collected 12,860 protein sequences from 208 extinct species obtained from NCBI (**FIG. 3a**). After removing redundant sequences, 5,190 proteins were left. EPs were defined as substrings ranging from 8-50 amino acid residues within the protein sequences, which align with the lengths of most active antimicrobial peptides reported previously[15]. This resulted in 10,311,899 EP sequences. The inventors then applied APEX to predict the antimicrobial activity of these compounds. This effort led to the identification of 37,176 peptides predicted to show

broad-spectrum antimicrobial activity with median minimal inhibitory concentrations (MIC) values ≤80 μmol L^{-1}. To test whether the compounds identified by APEX belonged to a new sequence space, the inventors compared the identified 37,176 sequences to previously described peptides in the literature and contained within the DBAASP[16] database. Briefly, for the EPs and DBAASP peptides, sequence alignment[19] was used to calculate the pairwise peptide sequence similarity (the sequence similarity calculation procedure can be found under ***Sequence similarity score*** in the **Methods**). For each sequence, the inventors used its sequence similarities to all peptides (i.e., DBAASP peptides and 37,176 EPs) as its feature representation. Next, the inventors used the uniform manifold approximation and projection (UMAP)[20] technique to reduce the dimension of the feature representations to a bidimensional (2D) space (**FIG. 32**). While DBAASP peptides mostly fell within the central area of the UMAP-derived 2D space, molecules identified by APEX had a much wider spread (**FIG. 32**), forming multiple distinct clusters that were not covered by DBAASP. These results revealed that sequences identified by APEX within the extinctome can belong to novel and previously unexplored parcels of sequence space.

[00107] *Differences between modern and archaic encrypted peptides.* To assess differences between modern encrypted peptides (MEPs; i.e., sequences present in both extinct and extant organisms), and archaic encrypted peptides (AEPs; i.e., sequences not found in the available proteomic data from extant organisms) and determine whether these sequences constitute new classes of antimicrobial peptides, the inventors compared sequences identified by APEX (i.e., 11,035 AEPs out of the 37,176 EPs identified above) with MEPs from the modern human proteome identified by an antimicrobial scoring function[15]. More details on the inventors' classification of MEPs and AEPs are provided in *Classification guidelines to identify archaic encrypted peptides (AEPs)* in **Methods**.

[00108] First, for a direct comparison with the scoring function, the inventors used APEX to find sequences within proteins from the modern human proteome, extracted these sequences, and determined their amino acid compositions (**FIG. 33**). In contrast to the scoring function, which considers and selects for net positive charge to identify sequences[15], APEX selected sequences with a higher frequency of negatively charged residues (i.e., aspartic acid and glutamic acid), as well as glycine and polar uncharged residues (i.e., asparagine, glutamine, and serine). These amino acids residues with net

charge and hydrophobicity features are not relevant to the scoring function, which mostly considers net positive charge and amphipathic peptide sequences. The EPs identified in modern human proteins[15] by the scoring function also showed a higher content of cysteine, methionine, phenylalanine, and arginine. Interestingly, lysine, which is preferentially scored by the scoring function, was slightly overrepresented in APEX-identified sequences (**FIG. 33**).

[00109] Furthermore, the inventors compared the amino acid composition of APEX's sequences to conventional AMPs from DBAASP (**FIG. 3b**). Generally, molecules identified by APEX presented lower cysteine, aspartic acid, and glycine content compared to AMPs from DBAASP (**FIG. 34**). Peptides derived from proteins of extinct organisms also had a lower asparagine and higher methionine and glutamine content compared to AMPs from DBAASP (**FIG. 35**). The MEPs identified by APEX had a much lower alanine, proline, and tryptophan content but a much higher isoleucine, leucine, asparagine, and serine content than peptides within DBAASP (**FIG. 36**). Comparative analysis between AEP and MEP sequences identified by APEX revealed an overrepresentation of methionine and glutamine in AEPs and of glycine in MEPs (**FIG. 37**).

[00110] The inventors then compared the physicochemical features contributing to antimicrobial properties[16] (**FIG. 38**). These analyses revealed that AEPs showed a lower amphiphilicity (amphiphilicity index <2; **FIG. 38a**) but a slightly higher propensity to be disordered (disordered conformation propensity score from -0.5 to 1) than MEPs or AMPs (**FIG. 38b**). These results indicate that interactions between AEPs and the bacterial membrane are likely to differ from those of standard AMPs, which are more amphiphilic and tend to assume a defined structure upon contact with the lipid from the membrane bilayer[16,21] (**FIGS. 38a-b and FIG. 39**). Additionally, to determine the potential toxicity and amphiphilicity of the peptides[21], the inventors assessed their theoretical tendency toward aggregation (**FIG. 38c)** and the angle of the hydrophobic residues upon adopting a secondary structure (**FIG. 38d)**. These physicochemical parameters are predictive of how the peptides interact with membrane lipids to exert antimicrobial activity[13]. Interestingly, when comparing AEPs with either MEPs from the human proteome or AMPs from DBAASP, the inventors found that AEPs were less prone to aggregate (*in vitro* aggregation propensity score <500) and presented a smaller predicted hydrophobic face

(<100°) (**FIG. 38c-d**). These results are a direct consequence of the higher frequency of uncharged polar residues in AEPs. To further investigate peptide structure, the inventors obtained the predicted normalized hydrophobic moment (**FIG. 38e)** and isoelectric point of AEPs (**FIG. 38f**), which presented a low range of normalized hydrophobic moment (0-0.6) and clustered within a short isoelectric point range (9.5 to 13). These values, found for sequences in extinct organisms (AEPs), overlapped with those obtained for sequences in extinct and extant organisms (MEPs) as well as in AMPs from DBAASP (**FIG. 38e-f**). The values aligned with the lower abundance of acidic residues compared to basic ones, particularly lysine, within AEPs (**FIG. 3c-d**).

[00111] Collectively, AEPs identified by APEX represent a distinct family of peptides with a higher abundance of uncharged polar residues and increased aliphatic content (particularly isoleucine and leucine) with respect to other classes of peptide antibiotics, including other EPs[15,22,23] and AMPs[24]. There are a few AMP families, such as he-1, brevenins, pleurains, and frog defensins[24], having a lower net charge than most standard AMPs, which have more uncharged polar residues or a balance of positively charged and acidic residues, and whose antimicrobial activity depends on how their electronic density is distributed[16]. Like the AMPs in these families but unlike previously described EPs[15] (**FIG. 33**) or conventional AMPs (**FIG. 34**), AEPs have a high abundance of uncharged polar residues[21]. Leucine and isoleucine, in particular, are structurally important: the stiffness of these bulky branched residues limits the internal flexibility of the peptide, whereas other aliphatic residues favor specific foldamers during the folding process[25]. The difference between the amino acid composition of known AMPs and that of AEPs and MEPs results in significantly different physicochemical features (**FIG. 3c-d** and **FIG. 38**), reaffirming that the EPs identified by APEX are different from known antimicrobial peptides.

Example 2 - In vitro antimicrobial activity of antibiotic molecules from extinct organisms

[00112] To further validate APEX's predictive power in identifying active encrypted peptide sequences from extinct organisms, the inventors synthesized and tested two non-overlapping sets of peptides: (i) 49 EPs predicted by a scoring function[15] (**FIG. 40**) and (ii) 69 EPs predicted by APEX (**FIG. 3a**) and found in 98 extinct species. While the 49 EPs predicted by the scoring function were found in both extinct and extant

organisms (*i.e.*, all were classified as MEPs), the APEX-predicted sequences included many that were unique to extinct organisms (20 AEPs and 49 MEPs). APEX was built to predict species-specific antimicrobial activities and 69 EPs were selected based on multiple selection criteria. Specifically, the inventors ranked the 10,311,899 sequences derived from the extinct proteomes by median antimicrobial activities (i.e., broad-spectrum activity), or selectivity against Gram-positive or Gram-negative pathogens. For each ranked list, the inventors used the following criteria to filter out compounds: (i) length not ranging from 8 to 30 amino acid residues, (ii) sequences that are present in the inventors' in-house dataset, (iii) with high sequence similarity to known AMPs from public databases, and (iv) EPs that are present in the modern human proteome. For each resulting list, the inventors grouped the EPs by their source organism and selected the top ranked EPs that were not too hydrophobic to be chemically synthesized by solid-phase peptide synthesis. To ensure the inventors explored a wider sequence space, EP sequences were selected from each list that were not sequentially too similar to each other (**FIG. 41**). Detailed selection and filtering criteria can be found in **Methods** subtopic *Encrypted peptide screening and selection*.

[00113] Among the 69 EPs identified by APEX, 21 (5 AEPs and 16 MEPs) were derived from secreted proteins, while 48 (15 AEPs and 33 MEPs) were from non-secreted proteins. The inventors included EPs from non-secreted proteins due to the limited annotations of extinct proteins. Filtering out unannotated secreted proteins would have restricted the sequence space explored, so EPs were considered from both secreted and non-secreted proteins. Out of the 21 peptides from secreted proteins identified by APEX, 4 were predicted to target Gram-positive bacteria selectively, 10 to target Gram-negative bacteria selectively, and 7 to exhibit broad-spectrum activity. Among the 48 peptides selected by APEX from non-secreted proteins, 19 were predicted to selectively target Gram-positive bacteria, 10 to selectively target Gram-negative bacteria, and 19 to display broad-spectrum activity.

[00114] Next, the inventors synthesized the 21 AEPs and 48 MEPs identified by APEX from extinct organisms and experimentally determined their MICs for 11 clinically relevant bacterial pathogens (seven Gram-negatives and four Gram-positives), ten of which are on the ESKAPEE pathogen list[17] (**FIG. 4a-b**). The name of each source organism was used as the basis for the inventors' molecular de-extinction nomenclature.

All experimentally determined MICs (log_2 transformed) were compared to predictions generated by APEX, yielding Pearson and Spearman correlation values of 0.448 and 0.404, respectively (**FIG. 5a**), underscoring APEX's substantial predictive power. In terms of species-specific antimicrobial activity prediction, APEX showed a high predicted versus experimentally validated activity correlation (Pearson correlation >0.3) for *A. baumannii* ATCC 19606, *Escherichia coli* strains AIC221, AIC222 (a colistin-resistant strain), and ATCC 11775, *P. aeruginosa* strains PAO1 and PA14, and *E. faecium* ATCC 700221 (a vancomycin-resistant strain). All correlation results obtained for the 11 experimentally validated strains are shown in **FIGS. 42-52**. Furthermore, when the average species-specific Pearson correlations of the log_2-transformed MIC predictions for the 11 pathogens of various baseline ML models were compared with that provided by APEX, the inventors' deep learning model yielded the most accurate predictions (**FIG. 2b**). Of the 69 synthesized peptides, 41 showed notable antimicrobial activity (i.e., MIC $\leq$128 μmol L^{-1}) against at least one bacterial strain, demonstrating a 59% hit rate for identifying peptides with antimicrobial activity (**FIG. 5b**). This hit rate is higher than that of the scoring function[15] (24%) when it was used for extracting antibiotics from the same extinct sources (**FIG. 5b**). In addition, 13 of the 41 active EPs identified by APEX were AEPs, meaning that they were present in extinct but not in extant organisms.

[00115] The names of the synthesized peptides and their sequences are provided below in Table 1.

Table 1

Name	Peptide	Sequence	SEQ ID NO:	AEP/ MEP
Equusin-1	AIO10918.1-FLK14	FLKLRWSRFARVLL	SEQ ID NO: 1	MEP
Hesperelin-1	CED79820.1-KLL26	KLLRKVLKETKKWVIKSVVFFKKIRK	SEQ ID NO: 2	AEP
Elephasin-1	AQU14158.1-LHL12	LHLKILKIIRLL	SEQ ID NO: 3	AEP
Arctodutin-1	CAQ68453.1-LLI15	LLINSIKRLLLGSIL	SEQ ID NO: 4	MEP
Arctoterin-1	ANA91291.1-GHL15	GHLLIHLIGKATLAL	SEQ ID NO: 5	MEP
Lophiosin-1	QYC36821.1-HWI16	HWITINTIKLSISLKI	SEQ ID NO: 6	AEP
Mammutin-1	ABQ86189.1-WMT15	WMTIHALKLSLSFKL	SEQ ID NO: 7	AEP
Ararin-1	AWH62781.1-ILL13	ILLTTAIAIKLGL	SEQ ID NO: 8	MEP

Mylodonin-1	AWK29290.1-WFH14	WFHFNSKILLLTGL	SEQ ID NO: 9	AEP
Mammuthusin-1	ABV45725.1-IFL14	IFLHLKTLKIIHLL	SEQ ID NO: 10	MEP
Paleopropin-1	AII98767.1-LTL14	LTLFIIFQLKISKL	SEQ ID NO: 11	MEP
Bisonin-1	AVE15294.1-LHT13	LHTINFIIKSLLL	SEQ ID NO: 12	MEP
Hesperelin-2	CED79820.1-KKW12	KKWVIKSVVFFK	SEQ ID NO: 13	MEP
Equusin-2	ADN88909.1-RAY26	RAYICRKKFLSLRKASIKLQSLVRMK	SEQ ID NO: 14	AEP
Mammuthusin-2	ABG37012.1-RAC26	RACLHARSIARLHKRWRPVHQGLGLK	SEQ ID NO: 15	MEP
Mammuthusin-3	BAF96956.1-KTL10	KTLKIIRLLF	SEQ ID NO: 16	MEP
Hydrodamin-1	AKN52354.1-LYC24	LYCRIYSLVRARGRRLTFRKNISK	SEQ ID NO: 17	AEP
Xenothrixin-1	AZB87500.1-TIK10	TIKLFLSFKL	SEQ ID NO: 18	MEP
Hesperelin-3	CAA38232.1-RQK17	RQKNHGIHFRVLAKALR	SEQ ID NO: 19	MEP
Ararin-2	AWH62785.1-RLA27	RLATLQLWTINKITKQLMIPLNKPGHK	SEQ ID NO: 20	AEP
Anomalopterin-1	AFS17509.1-RKI16	RKILGDLKFLESLKTY	SEQ ID NO: 21	MEP
Megalocerin-1	ARU77324.1-LIV13	LIVCFFRQLKFHF	SEQ ID NO: 22	MEP
Pinguinusin-1	ASB29243.1-KFI13	KFILNFKIPISFK	SEQ ID NO: 23	AEP
Ursusin-1	ANT45642.1-IFS13	IFSLHLAGISSIL	SEQ ID NO: 24	MEP
Elephasin-2	AQU14158.1-IFL14	IFLHLKILKIIRLL	SEQ ID NO: 25	AEP
Mammuthusin-4	ANH55260.1-LFI12	LFIGLTNLLGLL	SEQ ID NO: 26	MEP
Psephotellin-1	ANT45606.1-ISL14	ISLFIRPLALGVRL	SEQ ID NO: 27	MEP
Eudyptin-1	QBB10613.1-LHI15	LHIGLIKTYLGSFAL	SEQ ID NO: 28	MEP
	AII98767.1-ILT15	ILTLFIIFQLKISKL	SEQ ID NO: 29	MEP
Paleopropin-2	AKN79944.1-RMA19	RMARNLVRYVQGLKKKKVI	SEQ ID NO: 30	MEP
Hydrodamin-2	AKN79944.1-RNL29	RNLVRYVQGLKKKKVIVIPVGIGPHANIK	SEQ ID NO: 31	MEP
Hydrodamin-3	CAA38232.1-IVG16	IVGNVFGFKALRALRL	SEQ ID NO: 32	MEP
Hesperelin-4	SMQ11516.1-KRK18	KRKRGLKLATALSLNNKF	SEQ ID NO: 33	AEP
Mylodonin-2	AFS17541.1-KVG15	KVGAFLVDKVWKTLI	SEQ ID NO: 34	MEP
	ABN79624.1-KLY14	KLYQRILWRLISEL	SEQ ID NO: 35	MEP
Equusin-3	ANN03167.1-KWT13	KWTKIYLPLLLPL	SEQ ID NO: 36	MEP

Bisonin-2	ACX48939.1-KKP30	KKPPNPIKPKVPLSAPRKSPNTVKYRLKFR	SEQ ID NO: 37	MEP
	UEP15361.1-KKW12	KKWTKIYSPLSL	SEQ ID NO: 38	MEP
	SMQ11516.1-KIY25	KIYKKLSTPPFTLNIRTLPKVKFPK	SEQ ID NO: 39	AEP
Mylodonin-3	SMQ11516.1-RKR18	RKRGLKLATALSLNNKFV	SEQ ID NO: 40	MEP
Mylodonin-4	ABN79624.1-CVL25	CVLLFSQLPAVKARGTKHRIKWNRK	SEQ ID NO: 41	AEP

[00116] The inventors then used the selectivity score (see *Selectivity score* from the **Methods** section for details) to quantify peptide selectivity. Specifically, among the 69 peptides synthesized, 20 were computationally predicted to selectively target Gram-negative pathogens and 23 to selectively target Gram-positive pathogens. For peptides predicted to be selective for Gram-negative pathogens, the Pearson correlation of selectivity scores calculated by experimentally validated and predicted MICs was 0.295. In addition, the mean Gram-negative selectivity score derived from experimental MICs for peptides selective for Gram-negatives was 0.783, and 1.33 for the rest of peptides tested. A p-value of 0.013 for the one-sided Mann–Whitney U test on the selectivity scores from these two lists suggested a statistically significant difference. This result demonstrated APEX's ability to discover peptide sequences that selectively target Gram-negative pathogens. On the other hand, a weak Pearson correlation (0.11) was observed between selectivity scores calculated by predicted and validated MIC values of peptides that were selected to target Gram-positive bacteria. The mean selectivity score of these peptides for Gram-positive bacteria was 1.02, indicating that the peptides did not selectively kill Gram-positive pathogens. However, it was observed that the remaining peptides had a significantly higher mean selectivity score of 2.13 (p-value = 0.08, one-sided Mann–Whitney U test), suggesting that APEX can discover peptides that were relatively nonspecific against Gram-negative pathogens.

[00117] *Antibiotics identified in the extinctome.* Several molecules identified by APEX displayed excellent antimicrobial properties. These included anomalopterin-1, a peptide originating from the extinct moa species *Anomalopteryx didiformis*. This peptide is a fragment of the dynein axonemal heavy chain 3, which forms part of the microtubule-associated motor protein complex. Mylodonin-2, derived from the extinct South American giant sloth *Mylodon darwinii*, correspond to a fragment of apolipoprotein B, a lipoprotein that functions as a ligand for the low-density lipoprotein (LDL) receptor. Interestingly,

peptides derived from the modern human apolipoprotein B have also been described as depolarizers of the membranes of Gram-negative bacterial pathogens[26]. Mylodonin-1 and megalocerin-1, are both fragments from the cytochrome c oxidase subunit 3 from *Mylodon darwinii* and *Megaloceros sp.*, respectively. This protein is the last enzyme in the mitochondrial electron transport chain.

[00118] Equusins 1 (MEP) and 2 (AEP), originating from the extinct Grant's zebra *Equus quagga boehmi*, are derived from the natural resistance-associated macrophage protein 1 and the abnormal spindle-like microcephaly-associated protein, respectively. The natural resistance-associated macrophage protein 1 is essential for macrophage regulation and acts as a specific antiporter that fluxes metal ions in either direction against a proton gradient. The abnormal spindle-like microcephaly-associated protein is responsible for calmodulin-binding activity and plays a role in regulating the meiotic cell cycle, gamete generation, centrosome location maintenance, and nervous system development.

[00119] Mammuthusin-2, derived from *Mammuthus primigenius*, originated from the melanocyte-stimulating hormone receptor, which regulates all types of melanocyte-stimulating hormones (α, β, and γ). Elephasin-2, one of the most active antimicrobial molecules identified here, is found as a fragment of ATP synthase F0 subunit 8 from the extinct *Elephas antiquus*. This protein is one of the main subunits responsible for ATP synthesis.

[00120] Hesperelin-3 is produced by both an extinct magnolia species (*Magnolia latahensis*) and an extinct palm tree species (*Hesperelaea palmeri*). It is a component of the protein ribulose bisphosphate carboxylase large chain (RuBisCO). RuBisCO catalyzes two reactions: the carboxylation of D-ribulose 1,5-bisphosphate, which is the primary step in carbon dioxide fixation, and the oxidative fragmentation of the pentose substrate in photorespiration. The extinct manatee *Hydrodamalis gigas* yielded a fragment of the endothelial differential gene 1 (hydrodamin-1), which regulates endothelial cell differentiation, and a fragment from the von Willebrand factor (hydrodamin-2), a protein involved in hemostasis. EPs derived from the von Willebrand factor have also been previously found in modern humans[15].

Example 3 - Secondary structure of encrypted peptides from extinct organisms

[00121] Given that peptides identified by APEX were, on average, different from sequences predicted by the scoring function and previously reported AMPs in terms of physicochemical descriptors and amino acid residue composition, it was decided to determine their secondary structure. When AMPs come into contact with bacterial membranes, they typically adopt an α-helical conformation due to their amphipathicity, net charge, hydrophobicity, and length, all of which directly influence their secondary structure.

[00122] To computationally evaluate the alpha-helix structures and predict the secondary structures of peptides within the training dataset and the 69 EPs selected and validated, S4PRED[27] was utilized. The following ratio was then used to quantify the abundance of α-helix structures in a given peptide dataset:

$$\alpha - helix\ ratio = \frac{the\ number\ of\ amino\ acid\ residues\ predicted\ to\ be\ \alpha-helical\ in\ a\ dataset}{total\ number\ of\ amino\ acid\ residues\ in\ a\ dataset}.$$

[00123] For 988 in-house peptides and 5,093 publicly available AMPs used for APEX model training, the α-helix ratios were 25.99% and 40.88%, respectively. For the 37,176 EPs predicted to have broad-spectrum antimicrobial activity by APEX, the α-helix ratio was 40.0%. For the 69 EPs identified by APEX and selected for experimental validation, the ratio was 20.06%. The in-house peptides and publicly available AMPs used for APEX model training had a high α-helix ratio, confirming the dominance of α-helical structures in our training dataset. The high α-helix ratio for the 37,176 EPs identified by APEX further confirmed that our model captured this structure pattern from the training data. However, the ratio of α-helical structures for the 69 EPs that were selected and validated was lower than those present in the training data. It is believed this resulted from applying the second of five different filtering steps (***Encrypted peptide screening and selection***) to ensure the selected EPs were sequentially different (and consequently structurally diverse) from known AMPs and EPs[15].

[00124] To determine the secondary structure of the active sequences obtained from extinct organisms, they were exposed to a helix-inducing medium[28] (trifluoroethanol in water, 3:2, v:v). Interestingly, most molecules synthesized and tested that were identified through the scoring function were not α-helical, but instead had a relatively high content of anti-parallel β-structure and were largely unstructured (**FIG. 39a-b**). In

contrast, AEPs and MEPs identified by APEX demonstrated predominantly helical structures under the analyzed conditions, despite their unusual abundance of uncharged polar residues and low amphiphilicity (**FIG. 39c-d**). These results shed light into the considerably higher success rate achieved by APEX compared to the scoring function. The greater prevalence of α-helical peptides and amphipathic structures within the APEX sequences enhances their interactions with the membrane, resulting in more effective membrane damage [21,29](**FIG. 5b**).

Example 4 - Mechanism of action studies

[00125] The bacterial membrane is a common target for AMPs, where they engage in non-specific interactions with the lipid bilayer[21]. The antimicrobial activity of AMPs is influenced by their amino acid composition, distribution, and various physicochemical characteristics such as amphiphilicity and hydrophobicity. To investigate the underlying mechanisms by which the peptides identified by APEX kill bacteria, it was tested whether differences in the composition of AEPs and MEPs would affect their mechanism of action. The differences in composition assessed (**FIGS. 32 and 41**) include the content of uncharged polar and aliphatic residues (**FIG. 3b** and **FIGS. 34-37**) and different ranges of physicochemical features (**FIG. 3c-d** and **FIG. 38**).

[00126] First, it was tested whether AEPs and MEPs depolarized the cytoplasmic membrane of *A. baumannii*. The potentiometric fluorophore 3,3′-dipropylthiadicarbocyanine iodide ($DiSC_3$-5) was used, whose fluorescence is suppressed by its accumulation and aggregation within the cytoplasmic membrane. Upon disturbances in the transmembrane potential of the cytoplasmic membrane, this fluorophore migrates to the outer environment and emits fluorescence. Polymyxin B was used as a positive control in these experiments as it is a depolarizer that also permeabilizes and damages bacterial membranes. Notably, AEPs and MEPs depolarized the cytoplasmic membrane more effectively than polymyxin B (**FIG. 53**).

[00127] AEPs and MEPs depolarized the cytoplasmic membrane more effectively than peptides previously found in modern human proteins[15] (**FIG. 53**). The most potent depolarization effects were found for the following five peptides: anomalopterin-1, mylodonin-4, equusin-2, hesperelin-3, and hydrodamin-2 (**FIG. 5c**). It is hypothesized that this increased depolarization results from their different amino acid

composition, particularly the higher content of long aliphatic residues, such as leucine and isoleucine, in AEPs and MEPs compared to known AMPs.

[00128] To determine whether the compounds permeabilized the bacterial outer membrane, 1-(N-phenylamino)naphthalene (NPN) assays were performed. NPN, a lipophilic dye, fluoresces faintly in aqueous solutions but fluoresces substantially more when it encounters lipidic environments such as bacterial membranes. NPN can penetrate the bacterial outer membrane only if it is disrupted or compromised. In contrast to cells treated with the positive control polymyxin B or cells left untreated (untreated control group), bacteria exposed to the most active EPs at their MIC were, in general, not effectively permeabilized (**FIG. 5d and FIG. 54**). The EPs that permeabilized the outer membrane, causing a higher increase in the uptake of the fluorescent probe, were the MEP psephotellin-1, derived from the ATP synthase unit of the extinct parrot *Psephotellus pulcherrimus*, and arctodutin-1, a MEP that is part of the enzyme NADH-ubiquinone oxidoreductase chain 5 from the extinct bear *Arctodus simus*. Both enzymes would have played a crucial role in the metabolism of these extinct organisms. These results showed that the two MEPs, psephotellin-1 and arctodutin-1, were more effective permeabilizers than most known AMPs and previously reported EPs derived from human proteins[15]. Outer membrane permeabilization is the most common mechanism of action described for AMPs and a key mechanistic driver for EPs derived from modern human proteins, such as natriuretic peptide, SCUB1-SKE25, and SCUB3-MLP22[15]. However, the permeabilization effect exhibited by AEPs and MEPs from extinct organisms was not as potent as that shown by EPs from the modern human proteome.

Example 5 - Synergistic interaction studies

[00129] To investigate whether molecules from the same extinct organism (and close relatives) could synergize and thus potentiate each other's activity against pathogens, checkerboard assays[15] were performed at peptide concentrations ranging from twice the MIC to concentrations up to 64-times lower in the same conditions as used for the antimicrobial assays. First, peptides were selected according to their MIC values (**FIG. 4a and FIG. 55**) for two pathogenic strains, *A. baumannii* ATCC 19606 and *P. aeruginosa* PAO1. The former is an opportunistic nosocomial pathogen that is increasingly resistant to antibiotics, leading to considerable mortality worldwide[30]. The latter is an intrinsically resistant bacterium associated with infections of the urinary tract, gastrointestinal tissue,

skin, and soft tissues and a cause of bacterial pneumonia, as well as a common opportunistic pathogen in cystic fibrosis patients[31]. Most of the combinations tested resulted in synergistic or additive interactions, calculated by using the fractional inhibitory concentration index[32] (FICI, **FIG. 5e**). The MICs of combined EPs (**FIG. 55**) were mostly 2- to 3-fold lower than those of the individual peptides, but in some cases, for example, equusin-1 and equusin-3, the MICs decreased by 50 times (from 4 μmol L^{-1} to 78 nmol L^{-1}), reaching sub-micromolar concentrations that are comparable to the MICs of some of the most potent antibiotics[33].

[00130] Several pairs of EPs demonstrated particularly strong synergistic interactions, with FICI values as low as 0.38 for *A. baumannii*. These pairs included hesperelin-1 (AEP) and hesperelin-3 (MEP) from *Hesperelaea palmeri*, mammuthusin-1 (MEP) and mammuthusin-3 (MEP) from *Mammuthus primigenius*, equusin-2 (AEP) and equusin-4 (AEP) from *Equus quagga boehmi*, as well as hydrodamin-1 (AEP) and hydrodamin-3 (MEP) from *Hydrodamalis gigas* (**FIG. 5e**).

Example 7 - Cytotoxicity assays

[00131] All 41 active AEPs and MEPs identified in the antimicrobial assays (**FIG. 4a**) were tested for cytotoxic activity against human embryonic kidney (HEK293T) cells (**Supplementary Table 17**), an extensively characterized cell line. This assay is widely used to assess the toxicity of antimicrobials, yielding highly reproducible results[34–36]. Of the peptides tested, 39 displayed no notable cytotoxicity at the concentration range tested (8-128 μmol L^{-1}). Cytotoxicity was detected for the AEP lophisin-1 from the ancient crested rat (*Lophiomys imhausi maremortum*) and for the MEP xenothrixin-1 from the extinct Jamaican monkey (*Xenothrix mcgregori*). The peptide dose that led to 50% cytotoxicity (CC_{50}), estimated by non-linear regression, was 68.02 and 70.77 μmol L^{-1}, for lophisin-1 and xenothrixin-1, respectively. Despite their slight toxicity, the concentration of these two peptides needed to exert antimicrobial activity was, respectively, 8- and 8.84-times lower than their CC_{50} values, underscoring their potential as antimicrobials. The CC_{50} values for the other 39 peptides were higher than the maximum concentration analyzed, reinforcing the overall excellent safety profiles of this class of peptides.

Example 8 - Resistance to proteolytic degradation assays

[00132] The AEPs hydrodamin-1, elephasin-2, and mylodonin-2 and the MEPs mammuthusin-2 and megalocerin-1 were selected for further stability and animal studies

because of their potent antimicrobial activity (**FIG. 4a**) and good safety (**Supplementary Table 17**) profiles. To assess their stability in the presence of human proteases, these EPs were exposed to human serum and aliquots were collected and analyzed for 6 h at 37 °C. The AEP elephasin-2 and the MEP mammuthusin-2, both from organisms belonging to the same taxonomic order (i.e., *Proboscidea*) demonstrated the highest resistance to proteolytic degradation with ~40% peptide remaining after 6 h of exposure (**FIG. 56**). Mammuthusin-2 presented slower degradation kinetics, whereas elephasin-2 was present at ~40% within the first 30 min of the experiment. All other AEPs and MEP analyzed quickly degraded in the first 30 min to 1 h of the experiment (**FIG. 56**).

Example 9 - Anti-infective efficacy of encrypted peptides in animal models

[00133] To assess whether the active AEPs and MEPs had anti-infective efficacy *in vivo*, they were tested in preclinical mouse models of skin abscess[36] and thigh infection[15,37]. Five molecules were tested with a single dose at their MIC concentration after the infection was established. The following five compounds were tested, these having a wide range of MIC values (1-64 μmol L^{-1}) when tested *in vitro* against *A. baumannii*: the MEP elephasin-2 (MIC = 1 μmol L^{-1}, 1.3 μg Kg^{-1}) from *Elephas antiquus*, the AEP hydrodamin-1 (MIC = 4 μmol L^{-1}, 9.5 μg Kg^{-1}) from *Hydrodamalis gigas*, the MEP megalocerin-1 (MIC = 8 μmol L^{-1}, 10.9 μg Kg^{-1}) from *Megaloceros sp.*, the MEP mammuthusin-2 (MIC = 32 μmol L^{-1}, 78.1 μg Kg^{-1}) from *Mammuthus primigenius*, and the AEP mylodonin-2 (MIC = 64 μmol L^{-1}, 104.5 μg Kg^{-1}) from *Mylodon darwinii*.

[00134] In the skin abscess infection model, mice were infected with bacterial loads (10^6 cells of the pathogen *A. baumannii*) (**FIG. 6a**). Each molecule was administered as a single dose over the infected area. After two days, bacterial counts showed that all molecules tested, except hydrodamin-1, markedly reduced the bacterial load by 2-3 orders of magnitude. These results highlight the potential anti-infective activity of these antimicrobials (**FIG. 6b**). After four days, the EPs had either cleared the infection or reduced it by 3-5 orders of magnitude. The peptide hydrodamin-1, which was not active during the first two days post-infection, demonstrated activity by day 4 (**FIG. 6b**). The results obtained for the more active EPs tested (elephasin-2 and mylodonin-2) indicated antibacterial activity that was comparable to that of the widely used antibiotic polymyxin B, which was used as an antimicrobial control (**FIG. 6b**). Changes in weight, a surrogate measure of toxicity, were monitored from the time of the bacterial administration. No

variations in weight, damage to the skin tissue, or other harmful consequences induced by the molecules were observed in the mice throughout the experiments (**FIG. 57**).

[00135] Next, the anti-infective efficacy of the molecules was tested using an established preclinical model particularly suited to assess the translatability of potential antibiotics. Using a murine deep thigh infection model, the efficacy of elephasin-2, hydrodamin-1, megalocerin-1, mammuthusin-2, and mylodonin-2 were assessed, which were administered after the establishment of the intramuscular thigh infection (**FIG. 6c**). Briefly, mice were rendered neutropenic by cyclophosphamide treatment before intramuscular injection of 10^6 *A. baumannii* cells (**FIG. 6c**). Next, a single dose of each peptide at its MIC was injected intraperitoneally. Two- and four-days post-treatment, all peptides tested, except for hydrodamin-2, had reduced the bacterial load by 2-4 orders of magnitude compared to the untreated control group (**FIG. 6d**). Two days post-treatment, mylodonin-2 from *Mylodon darwinii* presented the most potent activity, which was comparable to that of the positive control antibiotic, polymyxin B (4-5 orders of magnitude reduction in bacterial counts). Four days post-treatment, elephasin-2 from *Elephas antiquus* and mammuthusin-2 from *Mammuthus primigenius* had decreased the bacterial loads by 3-4 orders of magnitude, resulting in similar levels as those in the mice treated with polymyxin B (**FIG. 6d**). None of the peptides tested were harmful to the mice based on weight monitoring during the experimental period (**FIG. 58**).

[00136] These substantial *in vivo* results with two different preclinical mouse models demonstrated that two AEPs (elephasin-2 and mylodonin-2) and one MEP (mammuthusin-2) displayed anti-infective efficacy comparable to that of a widely used antibiotic under physiologically relevant conditions, underscoring the potential of molecular de-extinction as an approach for antibiotic discovery.

[00137] This systematic analysis of all extinct organisms as a source of previously unrecognized antimicrobials demonstrates the concept of molecular de-extinction. Additionally, the inventive deep learning model APEX outperforms previous work in this emerging area[3] and constitutes an important method for antibiotic discovery through proteome mining. Molecular de-extinction enables the exploration of new sequence space, expanding the vision of molecular diversity and potentially unlocking new biology. It is hypothesized herein that encrypted peptides play a role in immunity throughout evolution and future work will be needed to further test this notion. Finally, the

present approach yielded preclinical candidates with activity comparable to the standard of care, such as polymyxin B, highlighting its broad potential applications in biotechnology and medicine. In sum, using deep learning, the present inventors have mined the proteomes of all available extinct organisms and have identified antibiotics effective against some of the bacterial pathogens most dangerous to human society.

[00138] The present inventors have leveraged deep learning (DL) to establish molecular de-extinction as a framework for antibiotic discovery. The present disclosure provides a proof-of-concept demonstration of the de-extinction of antimicrobial molecules from extinct organisms by combining deep learning with wet-lab validation both *in vitro* and in animals. The present approach of mining proteomes from extinct organisms unveils a previously untapped source of potential antibiotics. Furthermore, molecular de-extinction holds the potential to provide a source for other medicinal discoveries in the future.

Methods

[00139] Datasets

[00140] *Proteomes of extinct organisms.* Extinct organisms were collected from the NCBI taxonomy browser (https://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/index.cgi?chapter=extinct, access time: December 2021). For each species, we checked the corresponding Entrez records and downloaded the available protein sequences. In total, we retrieved 208 extinct species and a total of 12,860 protein sequences (5,190 non-redundant protein sequences) from them.

[00141] *Modern human proteome.* To construct the human proteome, we downloaded 20,388 reviewed *Homo sapiens* proteins (20,307 unique ones) from UniProt (https://www.uniprot.org/).

[00142] *In-house peptide dataset.* We utilized our high-quality in-house peptide dataset to train and evaluate APEX. In total, the dataset contained 14,738 antimicrobial activity measurements obtained by determining the minimum inhibitory concentration (MIC) of 988 peptides and 34 bacterial strains, including the following, which were used to train our model: *Escherichia coli* ATCC 11775, *Pseudomonas aeruginosa* PAO1, *Pseudomonas aeruginosa* PA14, *Staphylococcus aureus* ATCC 12600, *Escherichia coli* AIC221, *Escherichia coli* AIC222, *Klebsiella pneumoniae* ATCC 13883, *Acinetobacter*

baumannii ATCC 19606, *Akkermansia muciniphila* ATCC BAA-835, *Bacteroides fragilis* ATCC 25285, *Bacteroides vulgatus* (*Phocaeicola vulgatus*) ATCC 8482, *Collinsella aerofaciens* ATCC 25986, *Clostridium scindens* ATCC 35704, *Bacteroides thetaiotaomicron* ATCC 29148, *Bacteroides thetaiotaomicron* Δ*tdk* Δ*lpxF* (Background: VPI 5482)[33], *Bacteroides uniformis* ATCC 8492, *Bacteroides eggerthi* ATCC 27754, *Clostridium spiroforme* ATCC 29900, *Parabacteroides distasonis* ATCC 8503, *Prevotella copri* DSMZ 18205, *Bacteroides ovatus* ATCC 8483, *Eubacterium rectale* ATCC 33656, *Clostridium symbiosum* ATCC 14940, *Ruminococcus obeum* ATCC 29174, *Ruminococcus torques* ATCC 27756, methicillin-resistant *Staphylococcus aureus* ATCC BAA-1556, *vancomycin-resistant Enterococcus faecalis* ATCC 700802, *vancomycin-resistant Enterococcus faecium* ATCC 700221, *Escherichia coli* Nissle 1917, *Salmonella enterica* ATCC 9150 (BEIRES NR-515), *Salmonella enterica* (BEIRES NR-170), *Salmonella enterica* ATCC 9150 (BEIRES NR-174), *Listeria monocytogenes* ATCC 19111 (BEIRES NR-106).

[00143] Inactive data points, *i.e.*, MIC values higher than 128 μmol L^{-1}, were labeled as 140 μmol L^{-1}. All antimicrobial activities were transformed by $-log_{10}\frac{MIC\ value}{1{,}000{,}000}$ and were treated as labels to be predicted in the machine learning (ML) setting. To perform hyperparameter tuning and prediction performance evaluation, our in-house dataset was randomly split into a cross validation (CV) set and an independent set, which consisted of 790 and 198 peptides (*i.e.*, an 80%:20% split), respectively. Here, the CV set was used to determine the optimal hyperparameters for ML models. ML models trained with determined hyperparameters on the CV set were evaluated on the independent set to measure their generalizability.

Publicly available AMP sequences

[00144] The inventors augmented the peptide training data by incorporating publicly available AMPs and non-AMPs into our DL model training. Public AMP data was retrieved from the DBAASP[14]. Peptide sequences that consisted of only the 20 canonical or unknown amino acid residues were selected. The unknown amino acids were denoted by X, which corresponds to any possible canonical amino acid residues that usually occurred for the proteins having isoforms proteins whose composition was undetermined because there were issues with metagenomic studies. Any peptide sequences that overlapped with our in-house peptide database were removed. As a peptide may have

multiple MIC values for different bacterial species, we used the median MIC value to binarize the data. By using a stringent cutoff (i.e., AMPs with MIC ≤30 μmol L^{-1}), we created a balanced binary classification dataset (5,093 AMPs and 5,500 non-AMPs) for data augmentation and model training. To compare physicochemical properties, 14,114 peptides consisting of 20 canonical amino acids and having sequence length ≥4 amino acid residues were retrieved. We labeled this group of peptides as the DBAASP dataset.

Physicochemical properties of peptides

[00145] To analyze the physicochemical properties of the all peptide datasets (DBAASP, EPs generated by the scoring function[12], and EPs generated by APEX), we used the DBAASP server to calculate the following twelve physicochemical properties that are usually considered in the design and study of peptide antibiotics[14]: normalized hydrophobic moment, normalized hydrophobicity, net charge, isoelectric point, penetration depth, tilt angle, disordered conformation propensity, linear moment, propensity to aggregation *in vitro*, angle subtended by the hydrophobic residues, amphiphilicity index, and propensity to PPII coil. We used the Eisenberg and Weiss scale (the consensus scale) as the hydrophobicity scale[34].

Secreted protein labeling

[00146] As proteins from extinct organisms are not as well annotated as those from extant organisms, we resorted to Orthologous Matrix (OMA)[35] and DeepGOWeb[36] to predict gene ontology (GO) terms from protein sequences. Given a protein sequence, if any of its GO terms predicted by OMA or DeepGOWeb corresponds to an extracellular region (GO:0005576) or a child of extracellular region in a GO-directed acyclic graph, then we considered this sequence to be secreted. Among the proteins from the extinct organisms we collected, 157 sequences are labeled as secreted.

Peptide sequence encoding

[00147] A peptide was treated as a sequence of amino acids. The inventors further added two special characters as start (*i.e.*, '1') and terminal symbols (*i.e.*, '2') to the beginning and the end of this sequence, respectively. For each amino acid in the sequence, we used the AAindex[37], which is a 566-dimensional vector storing various physicochemical and biochemical properties of each amino acid to represent it. Non-amino acid symbols and unknown amino acids were represented by 566-dimensional zero vectors. In this work, we only considered peptides shorter than 50 residues to ensure that

they could be synthesized by solid phase peptide synthesis. We created a fixed size input by zero-padding each sequence to maximum length, so that each peptide sequence can be represented by a matrix $\boldsymbol{x} \in \mathbb{R}^{52\times 566}$ (52 = maximum peptide length + two special characters).

Bacterial distance

[00148] Taxonomy tree (bac120.tree) was downloaded from the Genome Taxonomy Database (GTDB)[38]. The phylogenetic distance matrix $\boldsymbol{D} \in \mathbb{R}^{g\times g}$, which stores distances between bacterial species, was calculated via the python package DendroPy[39]. Here, g denotes the number of bacterial species. We converted the distance matrix $\boldsymbol{D}$ to a bacterial similarity matrix $\boldsymbol{P} \in \mathbb{R}^{g\times g}$ using the following function:

$$\boldsymbol{P} = e^{\frac{-\boldsymbol{D}}{median(\boldsymbol{D})}},$$

where $median(\boldsymbol{D})$ stands for the median value of matrix $\boldsymbol{D}$. We further binarized the similarity matrix $\boldsymbol{P}$ using the K nearest neighbor algorithm, in which K was set to a heuristic number $Ceiling(\frac{\sqrt{g}}{2})$. Here, $Ceiling(\bullet)$ stands for the ceiling function.

Construction of APEX

[00149] *The encoder architecture.* The encoder started from a recurrent neural network (RNN) to process peptide sequence input $\boldsymbol{x}$ and extract its hidden features $\boldsymbol{h_{rnn}} \in \mathbb{R}^{52\times n}$:

$$\boldsymbol{h_{intermediate}} = \boldsymbol{GRU}(\boldsymbol{x}),$$

$$\boldsymbol{h_{rnn}} = \boldsymbol{Layer\_normalization}(\boldsymbol{h_{intermediate}}),$$

where n and $\boldsymbol{GRU}(\bullet)$ denotes the hidden feature dimension and gated recurrent unit[40], respectively. In addition, we added a layer normalization[41] to stabilize the model training. On top of the RNN, we designed a two-layer attention neural network to better model feature (i.e., amino acids) interactions globally and compress the hidden features to a lower-dimensional representation, respectively. Specifically, the first attention layer has the following form:

$$\boldsymbol{a_1} = \boldsymbol{softmax}(\boldsymbol{concat}(\boldsymbol{h_{rnn}}, \boldsymbol{x}) \times \boldsymbol{W_{att1}}),$$

$$\boldsymbol{h_{att1}} = \boldsymbol{a_1}^{\mathrm{T}} \times \boldsymbol{h_{rnn}},$$

where $\boldsymbol{concat}(\boldsymbol{h_{rnn}}, \boldsymbol{x}) \in \mathbb{R}^{52\times(n+566)}$ stands for concatenation operation along feature dimension and can be considered as a residual connection[42], $\boldsymbol{W_{att1}} \in \mathbb{R}^{(n+566)\times 52}$ is the learnable weights in this attention layer, $\boldsymbol{softmax}(\bullet)$ stands for the softmax

function, $\boldsymbol{a_1} \in \mathbb{R}^{52\times52}$ denotes the attention weights, and $\boldsymbol{h_{att1}} \in \mathbb{R}^{52\times n}$ is the output of this attention layer. For the second attention layer, it can be written as:

$$\boldsymbol{a_2} = \boldsymbol{softmax}(\boldsymbol{h_{att1}} \times \boldsymbol{w_{att2}}),$$

$$\boldsymbol{h_{att2}} = \boldsymbol{a_2^{\mathrm{T}}} \times \boldsymbol{h_{att1}},$$

where $\boldsymbol{w_{att2}} \in \mathbb{R}^{n\times1}$ is the learnable weights in this attention layer, $\boldsymbol{a_2} \in \mathbb{R}^{52\times1}$ denotes the attention weights, and $\boldsymbol{h_{att2}} \in \mathbb{R}^{1\times n}$ is the output of the second attention layer. In addition, we used a learnable linear transformation with weight matrix $\boldsymbol{W_{fc}} \in \mathbb{R}^{n\times m}$ and bias term $\boldsymbol{b_{fc}} \in \mathbb{R}^{1\times m}$ to create the final hidden representation $\boldsymbol{h} \in \mathbb{R}^{1\times m}$ for a peptide:

$$\boldsymbol{h} = \boldsymbol{h_{att2}} \times \boldsymbol{W_{fc}} + \boldsymbol{b_{fc}}.$$

[00150] *FCNNs on the prediction of sequences with antimicrobial activity.* The hidden representation $\boldsymbol{h}$ of a peptide generated by the encoder above could be fed into two separate FCNNs that predict species-specific antimicrobial activity or a binary AMP/non-AMP label, respectively. For convenience of hyperparameter tuning, both FCNNs were implemented as a 4-layer pyramid-like architecture:

$$\boldsymbol{h_{l,\ s}} = \boldsymbol{ReLu}(\boldsymbol{Layer\_normalization}(\boldsymbol{h_{l-1}} \times \boldsymbol{W_{l,\ s}} + \boldsymbol{b_{l,\ s}})),$$

where $s \in \{in-house,\ public\}$ denotes the training dataset for the FCNN, $l \in \{1,2,3,4\}$ denotes the layer index (note that $\boldsymbol{h_0} = \boldsymbol{h}$), $\boldsymbol{W_{l,\ s}}$ and $\boldsymbol{b_{l,\ s}}$ are weight matrix and bias term of lth layer, respectively. At lth layer, a linear transformation was first performed and followed by a layer normalization, a nonlinear transformation using rectified linear unit[43] (ReLU). In addition, if the lth layer is not an output layer, a dropout layer[44] that randomly set the input value to zero with probability p was added to its output side (we empirically set $p = 0.1$). The output dimensions of hidden layers (i.e., $\boldsymbol{h_{1,\ s}}$, $\boldsymbol{h_{2,\ s}}$, and $\boldsymbol{h_{3,s}}$) were set as $k, \frac{k}{2}, \frac{k}{4}$, respectively. The FCNN that was trained on our in-house data adopted a multitask learning strategy to predict species-specific antimicrobial activity. Suppose there are g bacterial species (*i.e.*, 34 in our context), the corresponding output $\boldsymbol{h_{4,\ in-house}}$ is a g-dimensional vector, in which each element is a predicted antimicrobial activity against a certain bacterial strain. The FCNN that was trained on public AMP data only outputted a scaler value $\in [0, 1]$ indicating the probability of the input peptide to be antimicrobial.

[00151] *Loss function.* The loss function for training the FCNN that performed binary classification was binary cross-entropy l_{BCE}. For the other FCNN, the loss function for predicting species-specific antimicrobial activity was the mean squared error:

$$l_{MSE} = \frac{1}{g}\sum_{i=1}^{g}\frac{1}{d_i}\sum_{j=1}^{d_i} Mask(i,j) * [\boldsymbol{h}_{4,\ \boldsymbol{in-house}}(i,j) - y(i,j)]^2,$$

where d_i denotes the number of training data points for ith bacterial strain, $\boldsymbol{h}_{4,\ \boldsymbol{in-house}}(i,j)$, $y(i,j)$ and $Mask(i,j)$ are the predicted antimicrobial activity, the experimentally validated antimicrobial activity, and the binary mask (1 for having antimicrobial activity, and 0 for not tested) between ith bacterial strain and jth peptide. In addition to these two loss functions on AMP prediction, we further imposed a constraint loss on the weights of output layer in the species-specific AMP prediction FCNN. Given a bacterial distance matrix $\boldsymbol{P} \in \mathbb{R}^{g\times g}$, and the weights $\boldsymbol{W}_{4,\ in-house}$ that we want to regularize, the constraint loss can be written as:

$$\boldsymbol{D}_{task} = 1 - cosine\_similarity(\boldsymbol{W}_{4,\ in-house})$$

$$l_{multitask\_constrain} = \frac{1}{2}\sum_{i=1}^{g}\sum_{j=1}^{g}\boldsymbol{P}(\boldsymbol{i},\boldsymbol{j}) * \boldsymbol{D}_{task}(i,j),$$

where $cosine\_similarity(\bullet)$ calculates the pairwise cosine similarity between two rows of the given input matrix, matrix $\boldsymbol{D}_{task} \in \mathbb{R}^{g\times g}$ stores the pairwise cosine distance between two learnable weights of two predictors. Intuitively, if two tasks are similar, their predictors should also be similar (*i.e.*, learnable weights have shorter distances). Adding $l_{multitask\_constrain}$ to the loss function encourages similar bacterial strains to have similar predictors and outputs. Taken together, the final loss function has the following form:

$$L = l_{MSE} + \lambda_{BCE}l_{BCE} + \lambda_{multitask\_constrain}l_{multitask\_constrain} + \lambda_{l2}l_2,$$

where l_2 is the L2 regularization for constraining DL model complexity, λ_{BCE}, $\lambda_{multitask\_constrain}$, and λ_{l2} are the weight parameters that balances different types of losses. To train the DL models, we used mini-batch training with an Adam optimizer[45]. Specifically, at each iteration, we selected B peptides from the in-house dataset and the same number of peptides from public AMP data we curated to perform feed-forwarding pass and back propagation. The training terminated when the procedure iterated the whole in-house dataset 5,000 times. The learning rate of the Adam optimizer was empirically set to 0.0001 and was scheduled to decay ten times every thousand training epochs. Batch size B was empirically set to 128.

Baseline methods

[00152] The prediction performance of APEX was compared to that of several baseline ML models, including elastic net, linear support vector regression, extra-trees regressor, random forest, and gradient boosting decision tree. For the baseline models, we represented each peptide sequence by the following features: (i) k-mer (*i.e.*, frequency of k-residue substrings, where k=1, 2 and 3) and (ii) ten peptide properties calculated by modlAMP[46], including sequence length, molecular weight, sequence charge, charge density, isoelectric point, instability index, aromaticity, aliphatic index, Boman index, and hydrophobic ratio. Note that for some of the bacterial strains, the trained Elastic Net outputted a constant prediction regardless of the peptide inputs. In this case, the Pearson and Spearman correlations could not be calculated, and we used 0 as pseudo correlation.

Hyperparameter tuning, model evaluation and ensemble learning

[00153] A five-fold cross validation on the CV set was conducted to select the hyperparameters of DL and baseline models. Specifically, the five-fold cross validation split the whole dataset evenly into five groups. At each time, one group was selected as the test dataset, while the rest was used for ML model training. We used averaged R-squared, Pearson correlation coefficient, and Spearman's rank correlation coefficient under five-fold cross validation to evaluate the prediction performance on the test set. We used grid search to find the best hyperparameters (see **Tables S1-S4** for hyperparameter range we searched). Hyperparameters were ranked by the averaged R-squared under cross validation. For baseline methods, we determined the best hyperparameters to be the ones resulting in the highest R-squared and trained the ML models with the selected hyperparameters. The trained models were then evaluated on the independent dataset. For APEX, we adopted an ensemble learning strategy. Specifically, the inventors averaged the prediction results from the top eight APEX models. After plotting the prediction performances versus the number of DL models involved in the ensemble learning, we observed that the elbow region (i.e., the area where the curve becomes smaller) was around 5-9 APEX models. After this step, improvement on prediction performance gradually became negligible. This observation led to the conclusion that we should average prediction results from no more than nine APEX models. We decided to select eight APEX models for ensemble learning, given our computational resources (*i.e.*, eight GPUs were available). To counter the potential stochastic behavior during mini-batch

training and in order to make prediction results more robust, we trained five copies of an APEX model with the same hyperparameters under different random seeds. In total, we trained 40 APEX models (*i.e.*, eight different hyperparameters × five different random seeds) and used the averaged predictions on the independent dataset for prediction performance evaluation. After the performance evaluation, we retrained the 40 APEX models on the entire in-house dataset and used the averaged antimicrobial activity prediction values from the trained models to discover encrypted peptide sequences from extinct organisms.

Selectivity score

[00154] Since APEX was designed to predict species-specific antimicrobial activity, we defined the following two selectivity scores that quantify peptides' ability to specifically target Gram-positive or Gram-negative bacteria:

$$Gram-positive\ selectivity\ score = \frac{Median\ MIC(gram\ positive\ pathogens)}{Median\ MIC(gram\ negative\ pathogens)},$$

$$Gram-negative\ selectivity\ score = \frac{Median\ MIC(gram\ negative\ pathogens)}{Median\ MIC(gram\ positive\ pathogens)},$$

where *Median MIC*(•) calculates the median value from a given input list. The input list consisted of the Gram-positive pathogens *S. aureus* ATCC 12600, methicillin-resistant *S. aureus* ATCC BAA-1556, vancomycin-resistant *E. faecalis* ATCC 700802, and vancomycin-resistant *E. faecium* ATCC 700221 and the Gram-negative pathogens *P. aeruginosa* PAO1, *P. aeruginosa* PA14, *E. coli* ATCC11775, *E. coli* AIC221, *E. coli* AIC222, *K. pneumoniae* ATCC 13883, and *A. baumannii* ATCC 19606. A selectivity score < 1.0 means that the median MIC of the target bacteria (numerator term) is smaller than that of the off-target bacteria (denominator term), yielding a selective peptide sequence. Thus, the closer to zero the better is the selective activity towards the specific bacterial target.

Sequence similarity score

[00155] Given two peptide sequences i and j, we used the Smith-Waterman algorithm[47] to calculate their sequence alignment score $SW(i,j)$. The sequence similarity score between these two peptides was defined as the normalized alignment score: $\frac{SW(i,j)}{\sqrt{SW(i,i)*SW(j,j)}} \in [0,1]$. A higher score reflects higher sequence similarity between two peptides than a lower score.

Encrypted peptide screening and selection from extinct proteomes

[00156] Given the proteome of extinct organisms, we considered as encrypted peptide (EP) sequences substrings ranging from 8 to 50 amino acid residues. In total, paleoproteome mining yielded 10,311,899 unique EPs from extinct proteomes of which 771,431 peptide sequences came from secreted proteins. EPs from secreted proteins would have had a higher likelihood of encountering bacterial cells, which are mostly found outside host cells, than EPs from non-secreted proteins Nevertheless, most EPs came from non-secreted proteins and consequently represented a more abundant peptide source. We hypothesized that these non-secreted proteins might also contain antibiotic-like substrings. Therefore, in this work, we selected, synthesized, and validated EPs originating from both secreted and non-secreted proteins.

[00157] Since we used 40 APEX models for the activity prediction (*i.e.*, ensemble APEX v2 or APEX in the main text), we averaged the prediction results from these 40 models to provide a final predictive output for each peptide. Based on these APEX predictions, we used multiple criteria to select EPs from our predictions for downstream validation. First, we mainly focused on EPs that could target a subset of the eleven pathogens that were listed in the ***Selectivity score*** section (i.e., *E. coli* ATCC 11775, *P. aeruginosa* PAO1, *P. aeruginosa* PA14, *S. aureus* ATCC 12600, *E. coli* AIC221, *E. coli* AIC222, *K. pneumoniae* ATCC 13883, *A. baumannii* ATCC 19606, methicillin-resistant *S. aureus* ATCC BAA-1556, vancomycin-resistant *E. faecalis* ATCC 700802, and vancomycin-resistant *E. faecium* ATCC 700221). Given the entire encrypted peptide list, we first generated two encrypted peptide lists; one with sequences predicted to selectively target Gram-positive and another with sequences predicted to target Gram-negative pathogens. To this end, the peptides in these lists were ranked increasingly by the $Gram-positive\ selectivity\ score$ and the $Gram-negative\ selectivity\ score$, respectively. We then generated another peptide list by ranking the peptides based on their median MIC predictions for the eleven pathogens of interest, indicating the level of broad-spectrum antimicrobial activity of the peptides.

[00158] As a result, three lists were generated for EPs derived from non-secreted proteins and another three lists were generated for EPs derived from secreted proteins, corresponding in each case to: (1) EPs predicted to selectively target Gram-positive bacteria; (2) EPs predicted to selectively target Gram-negative bacteria; and (3) EPs

predicted to have a broad-spectrum of activity. The six lists were filtered by: (1) including only peptides ranging from 8 to 30 residues, as they are easier to synthesize and as they comprise the range of sequence length in which most peptides with antimicrobial activity are active; (2) excluding sequences that appeared in our previous modern human proteome exploration[12], (3) excluding peptides that appeared in our in-house dataset, which is composed of natural and synthetic sequences, as well as computationally designed peptides; (4) excluding peptides that showed sequence similarity >0.75 to any peptide from the DBAASP dataset, which ensured that we would not explore peptides derived from modern natural versions of reported AMPs (i.e., we only kept EPs that were not similar to modern AMPs); (5) if two EPs in a list had a sequence similarity >0.75, keeping only the one that ranked higher (i.e., the more active or more selective peptide) in the list, which ensured that we explored the largest sequence space possible; and (6) for each of the lists, selecting a maximum of 1,000 peptides based on the ranking of predicted antimicrobial activity; (7) for broad-spectrum peptide lists, peptides whose median MIC predictions were ≥80 μmol L^{-1} were excluded as they were deemed inactive. For selectivity peptide lists, peptides whose selectivity scores were ≥0.75 were excluded as they were deemed as not selective enough. Finally, we grouped the selected peptides by their source organisms. The top (*i.e.*, more active or selective) EPs were selected while also taking into account species and sequence diversity (**Data S2**). Twenty-one EPs from secreted extinct proteins were selected for synthesis and subsequent experimental validation: 4 that were predicted to selectively target Gram-positive bacteria, 10 that were predicted to selectively target Gram-negative bacteria, and 7 that were predicted to have broad-spectrum activity. Forty-eight EPs from non-secreted extinct proteins were also selected for downstream experimental validation: 19 that were predicted to selectively target Gram-positive bacteria, 10 that were predicted to selectively target Gram-negative bacteria, and 19 that were predicted to display broad-spectrum activity.

Classification guidelines to identify archaic encrypted peptides (AEPs)

[00159] Since de-extinct sequences have not been previously identified, we developed our own classification guidelines to determine whether a particular peptide sequence qualified as truly de-extinct, *i.e.*, not present in the available proteomic data we had from living organisms. Since EP sequences from the proteomes of extinct organisms may also be found in modern organisms, we used the following procedure to classify an

EP sequence as an AEP: First, we accessed the NCBI taxonomy browser (https://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/index.cgi?chapter=extinct, access time: December, 2022) to get the most up-to-date taxonomy IDs of extinct organisms. Next, we created a protein sequence set with all possible organism sources by downloading protein sequences and their corresponding taxonomy IDs from Reviewed (Swiss-Prot), Unreviewed (TrEMBL), and Isoform sequences at UniProt[48] (https://www.uniprot.org/help/downloads). We excluded from the protein sequence set protein sequences whose taxonomy IDs belonged to extinct organisms. We labeled the resulting set as "extant protein set" as it contained protein sequences from extant organisms. If a given EP sequence was not present in any protein sequence from the modern protein set available, we defined it as an AEP.

Bacterial strains and growth conditions used in the experiments

[00160] The following Gram-negative bacteria were used in the present study: *Acinetobacter baumannii* ATCC 19606, *Escherichia coli* ATCC 11775, *Escherichia coli* AIC221 (*Escherichia coli* MG1655 phnE_2::FRT), *Escherichia coli* AIC222 [*Escherichia coli* MG1655 pmrA53 phnE_2::FRT (colistin-resistant)], *Klebsiella pneumoniae* ATCC 13883, *Pseudomonas aeruginosa* PAO1, and *Pseudomonas aeruginosa* PA14. The following Gram-positive bacteria were also used in our study: *Staphylococcus aureus* ATCC 12600, *Staphylococcus aureus* ATCC BAA-1556 (methicillin-resistant strain), *Enterococcus faecalis* ATCC 700802 (vancomycin-resistant strain), and *Enterococcus faecium* ATCC 700221 (vancomycin-resistant strain). Bacteria were grown from frozen stocks and plated on Luria-Bertani (LB) or *Pseudomonas* Isolation (*Pseudomonas aeruginosa* strains) agar plates and incubated overnight at 37 °C. After the incubation period, a single colony was transferred to 5 mL of LB medium, and cultures were incubated overnight (16 h) at 37 °C. The following day, an inoculum was prepared by diluting the overnight cultures 1:100 in 5 mL of the respective media and incubating them at 37 °C until bacteria reached logarithmic phase (OD_{600} = 0.3-0.5).

Antibacterial assays

[00161] The *in vitro* antimicrobial activity of the peptides was assessed by subjecting them to the broth microdilution assay[31]. Minimum inhibitory concentration (MIC) values of the peptides were determined with an initial inoculum of 2×10^6 cells mL^{-1} in LB in microtiter 96-well flat bottom transparent plates. Aqueous solutions of the

peptides were added to the plate at concentrations ranging from 1 to 64 μmol L^{-1}. The lowest concentration of peptide that inhibited 100% of the visible growth of bacteria was established as the MIC value in an experiment of 20 h of exposure at 37 °C. The optical density of the plates was measured at 600 nm using a spectrophotometer. All assays were done as three biological replicates.

Outer membrane permeabilization assays

[00162] The membrane permeability of the peptides was determined by using the 1-(N-phenylamino)naphthalene (NPN) uptake assay[12]. NPN is a hydrophobic fluorescent dye that does not readily permeate the bacterial outer membrane. However, when the membrane integrity is compromised, NPN can enter the cell and bind to the bacterial membrane lipids. This causes the dye to exhibit a strong fluorescence. *A. baumannii* ATCC19606 and *P. aeruginosa* PAO1 were grown (OD_{600} = 0.4), centrifuged (10,000 rpm at 4 ºC for 10 min), washed, and resuspended in buffer (5 mmol L^{-1} HEPES, 5 mmol L^{-1} glucose, pH 7.4). NPN solution (4 μL at the working concentration of 10 mmol L^{-1} after dilution) was added to 100 μL of the bacterial solution in a white 96-well plate. The fluorescence was recorded at λ_{ex} = 350 nm and λ_{em} = 420 nm. Aqueous solutions of the peptides (100 μL final volume at their MIC against the strain of interest) were added to a white 96-well plate, and fluorescence was recorded for 20 min after no further increase in fluorescence was observed. All assays were done as three biological replicates. The relative fluorescence values were calculated for the entire course of the experiment using non-linear fitting and the untreated control (buffer + bacteria + fluorescent dye) as baseline. The following equation was applied to show % difference between the fluorescence of the untreated control (baseline) and the sample:

$$Percentage\ difference = \frac{100 * (fluorescence_{sample} - baseline)}{baseline}$$

Cytoplasmic membrane depolarization assays

[00163] The depolarization of the bacterial cytoplasmic membrane was determined by fluorescence measurements of the membrane potential-sensitive dye, 3,3'-dipropylthiadicarbocyanine iodide $DiSC_3$-5[12]. Briefly, *A. baumannii* ATCC 19606 and *P. aeruginosa* PAO1 were grown at 37 ºC until mid-log phase (OD_{600} = 0.5). The cells were then centrifuged using the same conditions described for the NPN uptake assays, washed twice with washing buffer containing 20 mmol L^{-1} glucose and 5 mmol L^{-1} HEPES (pH

7.2). The cells were diluted 1:10 (OD_{600} = 0.05) in a buffer containing 0.1 mol L^{-1} KCl, 20 mmol L^{-1} glucose and 5 mmol L^{-1} HEPES (pH 7.2). One hundred µL of bacterial solution were then incubated for 15 min with 20 nmol L^{-1} of $DiSC_3$-5 until the fluorescence reached a plateau, i.e., the dye was fully internalized into the bacterial membrane. Transmembrane potential changes were monitored by observing the difference in the fluorescence emission intensity of $DiSC_3$-5 (λ_{ex} = 622 nm, λ_{em} = 670 nm), after the addition of 100 µL of peptide aqueous solution at its MIC. All assays were performed in three biological replicates. The relative fluorescence values were calculated for the course of the experiment using non-linear fitting and the untreated control (buffer + bacteria + fluorescent dye) served as baseline. The following equation was applied to show % difference between the fluorescence of the untreated control (baseline) and the sample:

$$Percentage\ difference = \frac{100 * (fluorescence_{sample} - baseline)}{baseline}$$

Synergy between encrypted peptides from extinct organisms

[00164] *P. aeruginosa* PAO1 and *A. baumannii* ATCC 19606 were used to assess the synergistic interactions of peptides derived from the same organisms because of their resistance to antimicrobials. The most active de-extinct EPs against *P. aeruginosa* PAO1 and *A. baumannii* ATCC 19606 were orthogonally diluted, using the microdilution technique, to concentrations ranging from 4-times MIC to 0.0625-times MIC in checkerboard assays. Plates were incubated for 20 h at 37 °C. All assays were done in three biological replicates.

Cytotoxicity assays

[00165] Human embryonic kidney (HEK293T) cells were obtained from the American Type Culture Collection (ATCC; CRL-3216™). The cells were cultured in high-glucose Dulbecco's modified Eagle's medium (DMEM) supplemented with 1% penicillin and streptomycin (antibiotics) and 10% fetal bovine serum (FBS) and grown at 37 °C in a humidified atmosphere containing 5% CO_2. One day before the experiment, an aliquot of 100 µL of the cells at 50,000 cells mL^{-1} was seeded into each well of the cell-treated 96-well plates used in the experiment (*i.e.*, 5,000 cells per well). The attached HEK293T cells were then exposed to increasing concentrations of the peptides (8-128 µmol L^{-1}) for one day. After the incubation period, we performed the (3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide) tetrazolium reduction assay (MTT

assay)[49]. The MTT reagent was dissolved at 0.5 mg mL^{-1} in medium without phenol red and was used to replace cell culture supernatants containing the peptides (100 µL per well), and the samples were incubated for 4 h at 37 °C in a humidified atmosphere containing 5% CO_2 yielding the insoluble formazan salt. The resulting salts were then resuspended in hydrochloric acid (0.04 mol L^{-1}) in anhydrous isopropanol and quantified by spectrophotometric measurements of absorbance at 570 nm. All assays were done as three biological replicates.

Resistance to proteolytic degradation assays

[00166] To assess the resistance of EPs to proteolysis, we incubated them in human serum[50]. The following five lead EPs: mammuthusin-2, hydrodamin-1, megalocerin-1, elephasin-2, and mylodonin-2 (at 3 mg mL^{-1}) were exposed to 25% human serum in water for 6 h at 37 °C. One hundred µL aliquots were collected after 0, 0.5, 1, 3, and 6 h, and 10 µL of 100% trifluoroacetic acid (TFA) was added to each sample to induce protein precipitation and incubated for 10 min on ice (at ~4 °C). Samples were then processed in a Waters Acquity UPLCMS equipped with a photodiode array detector (190-400 nm data collection) and a Waters TQD triple quadrupole MSMS, with 5 µL injections. The column used was a Waters Acquity UPLC HSS C_{18}, 1.8 µm (2.1 mm x 50 mm). The mobile phases used were A (100% water with 0.1%, v/v, formic acid) and B (100% acetonitrile with 0.1%, v/v, formic acid), Fisher optima grades. Measurements were made by ionization ESI +/- simultaneous over m/z 100-2,000 Da. The percentage of remaining peptide was calculated by integrating the area under the curve related to the peptide at time point zero. Experiments were performed in three independent replicates.

Time (min)	**A (%)**	**B (%)**	**Flow rate (mL min^{-1})**
0	95	5	0.5
0.5	95	5	0.5
2.5	5	95	0.5 (linear gradient)

3	5	95	0.5
3.25	5	95	0.5

Skin abscess infection mouse model

[00167] *A. baumannii* ATCC 19606 were grown in LB medium to an OD_{600} = 0.5. Cells were washed twice with sterile PBS (pH 7.4, 13,000 rpm for 2 min) and resuspended to a final concentration of 2×10^5 (*A. baumannii* cells) colony-forming units (CFU) mL^{-1}. Six-week-old female CD-1 mice from Charles River (stock number 18679700-022) were anesthetized with isoflurane and their backs were sterilized and shaved. A superficial linear skin abrasion was made with a needle to damage the stratum corneum and upper layer of the epidermis. An aliquot of 20 μL containing the bacterial load resuspended in PBS was inoculated over the scratched area. One hour after the infection, peptides diluted in water at their MIC value were administered to the infected area. Mice were euthanized and the area of scarified skin was excised two- and four-days post-infection, homogenized using a bead beater for 20 min (25 Hz), and 10-fold serially diluted for CFU quantification in MacConkey agar plates. The experiments were performed with 6 mice per group. All experiments were performed blindly, and no animal subjects were excluded from the analysis. The skin abscess infection mouse model was approved by the University Laboratory Animal Resources (ULAR) from the University of Pennsylvania (Protocol 806763). Statistical significance was determined using one-way ANOVA followed by Dunnett's test in a log_{10}-transformed data to mitigate the effect of outliers; p values are presented for each group, with all groups being compared to the untreated control group.

Thigh infection mouse model

[00168] Six-week-old female CD-1 mice from Charles River (stock number 18679700-022) were rendered neutropenic by two doses of cyclophosphamide (150 mg Kg^{-1}) applied intraperitoneally with an interval of 72 h. One day after the last dose of cyclophosphamide, the mice were injected intramuscularly in their right thigh with a bacterial load of 10^6 CFU mL^{-1} of *A. baumannii* ATCC 19606 cells. The bacteria had been grown in LB broth, washed twice with PBS (pH 7.4), and resuspended to the desired concentration. Two hours after bacterial injection, peptides resuspended in water were

administered intraperitoneally. Prior to each injection, mice were anesthetized with isoflurane and monitored for respiratory rate and pedal reflexes. Next, we monitored the establishment of the infection and euthanized the mice. The infected area was excised two days and four days post-infection, homogenized using a bead beater for 20 min (25 Hz), and 10-fold serially diluted for CFU quantification in MacConkey agar plates. The experiments were performed with 6 mice per group. All experiments were performed blindly, and no animal subjects were excluded from the analysis. The thigh infection mouse model was approved by the University Laboratory Animal Resources (ULAR) from the University of Pennsylvania (Protocol 807055). Statistical significance was determined using one-way ANOVA followed by Dunnett's test in a $\log_{10}$-transformed data to mitigate the effect of outliers; p values are presented for each group, with all groups being compared to the untreated control group.

Statistical analysis

[00169] Unless otherwise stated, all assays were performed in three independent biological replicates as indicated in each figure legend and Methods sections. The hemolytic and cytotoxic activities values were estimated using non-linear regression based on the range of concentrations screened andwere shown as the values that cause lysis of 50% of the cells in the experiment. Two technical replicates were performed in the cytotoxicity assays within each of the three biological replicates. In the mouse experiments, the statistical significance was determined using one-way ANOVA followed by Dunnett's test. All the p values are shown for each of the groups, all groups were compared to the untreated control group. The solid line inside each box represents the mean value obtained for each group. All calculation and statistical analyses of the experimental data were conducted using GraphPad Prism v.10.0.2 and computational data were performed in Python. Statistical significance between different groups was calculated using the tests indicated in each figure legend. No statistical methods were used to predetermine sample size.

[00170] Supplemental Tables

Table S1. Hyperparameter ranges explored for APEX.

Hyperparameters	number of RNN layers	n	m	λ_l2	λ_(multitask_constraint)	λ_BCE
Hyperparameter range	{1, 2, 3}	{128, 256}	{512, 1024, 2048}	{1e-5, 1e-6}	{0.1, 0.01, 0.001, 0.0}	{1.0, 0.1, 0.0}

Table S2. Hyperparameter ranges searched for elastic net.

Elastic Net hyperparameter	alpha
Hyperparameter range	{1.0, 0.5, 0.1, 0.05, 0.01, 0.005, 0.001, 0.0005, 0.0001}

Table S3. Hyperparameter ranges searched for linear support vector machine.

Linear support vector regression hyperparameter	C
Hyperparameter range	{1, 10, 20, 30, ..., 1000}

Table S4. Hyperparameter ranges searched for tree-based models.

Random forest, gradient boosting decision tree and, extra-trees regressor hyperparameter hyperparameters	Num estimators	Max depth
Hyperparameter range	{128, 256, 512, 1024}	{8, 16, 32, 64}

24-10510 (103241.007081)

Table S5. R-squared of various ML models on CV set. We divided our in-house peptide dataset into a CV set for hyperparameter tuning and ML model selection, and an independent dataset to evaluate prediction performance. The table shows the average species-specific prediction performance of 5-fold CV in terms of R-squared for various ML models on the CV set. RF: random forest; GBDT: gradient boosting decision tree; ExtraTree: extra-tree regressor; ElasticNet: elastic net; LinearSVR: linear support vector regression.

Strain	Ensemble APEX v1	Single APEX	RF	GBDT	ExtraTree	ElasticNet	LinearSVR
E. coli ATCC11775	0.588229	0.519899	0.494097	0.446439	0.434624	0.417793	0.128045
P. aeruginosa PAO1	0.520804	0.458351	0.490921	0.401458	0.430824	0.41167	-0.07728
P. aeruginosa PA14	0.461245	0.415884	0.402854	0.327131	0.38079	0.360004	0.02522
S. aureus ATCC12600	0.519007	0.443486	0.388353	0.348654	0.267847	0.271189	0.017144
E. coli AIC221	0.574544	0.527619	0.46737	0.471732	0.428688	0.351852	0.142403
E. coli AIC222	0.638583	0.578155	0.50825	0.497999	0.445026	0.408777	0.159808
K. pneumoniae ATCC13883	0.215734	0.078356	0.093001	0.025148	-0.17556	0.017496	-0.36587
A. baumannii ATCC19606	0.675531	0.597275	0.577164	0.581414	0.559864	0.420903	0.281722
Methicillin-resistant *S. aureus* ATCC BAA-1556	0.539041	0.442817	0.346027	0.355517	0.15017	0.237193	-0.2676
Vancomycin-resistant *E. faecalis* ATCC700802	-0.14859	-0.46429	-0.03412	-0.56905	-0.05681	-0.01311	-0.09197
Vancomycin-resistant *E. faecium* ATCC700221	0.615554	0.460972	0.422509	0.347594	0.345839	0.441332	0.385905
Average	**0.472699**	**0.368956**	**0.377857**	**0.294003**	**0.291936**	**0.302282**	**0.030685**

24-10510 (103241.007081)

Table S6. Pearson correlation of various ML models on CV set. We divided our in-house peptide dataset into a CV set for hyperparameter tuning and ML model selection, and an independent dataset to evaluate prediction performance. The table shows the average species-specific prediction performance of 5-fold CV in terms of Pearson correlation for various ML models on the CV set. RF: random forest; GBDT: gradient boosting decision tree; ExtraTree: extra-tree regressor; ElasticNet: elastic net; LinearSVR: linear support vector regression.

Strain	Ensemble APEX v1	Single APEX	RF	GBDT	ExtraTree	ElasticNet	LinearSVR
E. coli ATCC11775	0.767053	0.722816	0.703998	0.691083	0.661328	0.646422	0.556482
P. aeruginosa PAO1	0.72198	0.683082	0.701559	0.673932	0.656995	0.644568	0.450686
P. aeruginosa PA14	0.679904	0.654815	0.635966	0.598879	0.619795	0.606475	0.449355
S. aureus ATCC12600	0.720695	0.672914	0.623975	0.609784	0.548119	0.523365	0.469692
E. coli AIC221	0.758512	0.728434	0.688737	0.6902	0.65761	0.594226	0.490013
E. coli AIC222	0.799682	0.761297	0.715085	0.710545	0.667721	0.640531	0.52739
K. pneumoniae ATCC13883	0.477322	0.401615	0.319796	0.332036	0.155075	0.228322	0.250313
A. baumannii ATCC19606	0.822192	0.774607	0.761093	0.765975	0.749168	0.664209	0.58983
Methicillin-resistant *S. aureus* ATCC BAA-1556	0.734576	0.680613	0.589402	0.620643	0.500931	0.497521	0.424084
Vancomycin-resistant *E. faecalis* ATCC700802	0.096459	0.056318	0.029826	0.026884	0.029424	-0.15356	0.048211
Vancomycin-resistant *E. faecium* ATCC700221	0.784896	0.697111	0.655262	0.617407	0.600566	0.674514	0.622007
Average	**0.669388**	**0.621238**	**0.584063**	**0.576124**	**0.531521**	**0.506053**	**0.44346**

24-10510 (103241.007081)

Table S7. Spearman correlation of various ML models on CV set. We divided our in-house peptide dataset into a CV set for hyperparameter tuning and ML model selection, and an independent dataset to evaluate prediction performance. The table shows the average species-specific prediction performance of 5-fold CV in terms of Spearman correlation for various ML models on the CV set. RF: random forest; GBDT: gradient boosting decision tree; ExtraTree: extra-tree regressor; ElasticNet: elastic net; LinearSVR: linear support vector regression.

Strain	Ensemble APEX v1	Single APEX	RF	GBDT	ExtraTree	ElasticNet	LinearSVR
E. coli ATCC11775	0.686485	0.66331	0.622101	0.603803	0.571779	0.572859	0.446956
P. aeruginosa PAO1	0.636473	0.611234	0.606478	0.602546	0.579222	0.54594	0.368697
P. aeruginosa PA14	0.658294	0.625266	0.611538	0.592355	0.596603	0.569181	0.405066
S. aureus ATCC12600	0.546967	0.484758	0.433656	0.468351	0.297889	0.408671	0.260263
E. coli AIC221	0.733874	0.719559	0.673859	0.653163	0.681991	0.617147	0.469486
E. coli AIC222	0.737407	0.720051	0.672798	0.6632	0.626752	0.607942	0.452432
K. pneumoniae ATCC13883	0.527278	0.486858	0.260868	0.430102	0.058973	0.333547	0.159662
A. baumannii ATCC19606	0.726801	0.703749	0.687577	0.676496	0.696205	0.635361	0.510874
Methicillin-resistant *S. aureus* ATCC BAA-1556	0.525053	0.489835	0.462725	0.460328	0.309507	0.428855	0.348216
Vancomycin-resistant *E. faecalis* ATCC700802	0.107576	0.047122	0.141981	0.079331	0.169632	-0.08813	0.200453
Vancomycin-resistant *E. faecium* ATCC700221	0.646947	0.569296	0.484717	0.524636	0.444265	0.531174	0.468919
Average	**0.593923**	**0.556458**	**0.514391**	**0.523119**	**0.457529**	**0.469322**	**0.371911**

24-10510 (103241.007081)

Table S8. R-squared of single APEX variants on CV set. We divided our in-house peptide dataset into a CV set for hyperparameter tuning and ML model selection, and an independent dataset to evaluate prediction performance. The table shows the average species-specific prediction performance of 5-fold CV in terms of R-squared for various APEX variants, including the original APEX (*i.e.*, single APEX), APEX without multitask constraint, APEX without using public AMP data during training, and APEX without multitask constraint and public AMP data during training.

Strain	Single APEX	Single APEX without multitask constraint	Single APEX without public AMP data	Single APEX without multitask constraint and public AMP data
E. coli ATCC11775	0.519899	0.540226	0.515468	0.504094
P. aeruginosa PAO1	0.458351	0.411795	0.418864	0.328494
P. aeruginosa PA14	0.415884	0.33248	0.386427	0.253563
S. aureus ATCC12600	0.443486	0.418941	0.378119	0.391915
E. coli AIC221	0.527619	0.509886	0.468906	0.459699
E. coli AIC222	0.578155	0.586917	0.547914	0.559102
K. pneumoniae ATCC13883	0.078356	0.110211	0.12016	0.078548
A. baumannii ATCC19606	0.597275	0.597209	0.554144	0.57716
Methicillin-resistant *S. aureus* ATCC BAA-1556	0.442817	0.500468	0.432583	0.45067
Vancomycin-resistant *E. faecalis* ATCC700802	-0.46429	-0.18286	-0.48424	0.021012
Vancomycin-resistant *E. faecium* ATCC700221	0.460972	0.512912	0.479714	0.52491
Average	**0.368956**	**0.39438**	**0.347096**	**0.377197**

24-10510 (103241.007081)

Table S9. Pearson correlation of single APEX variants on CV set. We divided our in-house peptide dataset into a CV set for hyperparameter tuning and ML model selection, and an independent dataset to evaluate prediction performance. The table shows the average species-specific prediction performance of 5-fold CV in terms of Pearson correlation for various APEX variants, including the original APEX (i.e., single APEX), APEX without multitask constraint, APEX without using public AMP data during training, and APEX without multitask constraint and public AMP data during training.

Strain	Single APEX	Single APEX without multitask constraint	Single APEX without public AMP data	Single APEX without multitask constraint and public AMP data
E. coli ATCC11775	0.722816	0.745399	0.72278	0.724131
P. aeruginosa PAO1	0.683082	0.660926	0.657109	0.616645
P. aeruginosa PA14	0.654815	0.603326	0.633283	0.563075
S. aureus ATCC12600	0.672914	0.673138	0.62627	0.650051
E. coli AIC221	0.728434	0.716857	0.691317	0.687469
E. coli AIC222	0.761297	0.76824	0.744002	0.752025
K. pneumoniae ATCC13883	0.401615	0.419686	0.408857	0.390867
A. baumannii ATCC19606	0.774607	0.776853	0.749409	0.767272
Methicillin-resistant *S. aureus* ATCC BAA-1556	0.680613	0.717745	0.661616	0.67801
Vancomycin-resistant *E. faecalis* ATCC700802	0.056318	0.157377	0.010612	0.275715
Vancomycin-resistant *E. faecium* ATCC700221	0.697111	0.721689	0.699877	0.734134
Average	**0.621238**	**0.63284**	**0.600467**	**0.621763**

24-10510 (103241.007081)

Table S10. Spearman correlation of single APEX variants on CV set. We divided our in-house peptide dataset into a CV set for hyperparameter tuning and ML model selection, and an independent dataset to evaluate prediction performance. The table shows the average species-specific prediction performance of 5-fold CV in terms of Spearman correlation for various APEX variants, including the original APEX (i.e., single APEX), APEX without multitask constraint, APEX without using public AMP data during training, and APEX without multitask constraint and public AMP data during training.

Strain	Single APEX	Single APEX without multitask constraint	Single APEX without public AMP data	Single APEX without multitask constraint and public AMP data
E. coli ATCC11775	0.66331	0.624219	0.6493	0.640146
P. aeruginosa PAO1	0.611234	0.556474	0.576812	0.575338
P. aeruginosa PA14	0.625266	0.578997	0.609458	0.544803
S. aureus ATCC12600	0.484758	0.478392	0.456472	0.47019
E. coli AIC221	0.719559	0.659801	0.683423	0.661957
E. coli AIC222	0.720051	0.686905	0.691819	0.688436
K. pneumoniae ATCC13883	0.486858	0.370328	0.435828	0.396926
A. baumannii ATCC19606	0.703749	0.687933	0.703927	0.705705
Methicillin-resistant *S. aureus* ATCC BAA-1556	0.489835	0.469302	0.498565	0.508692
Vancomycin-resistant *E. faecalis* ATCC700802	0.047122	0.150755	0.055165	0.242589
Vancomycin-resistant *E. faecium* ATCC700221	0.569296	0.586361	0.577632	0.59948
Average	**0.556458**	**0.53177**	**0.539855**	**0.548569**

24-10510 (103241.007081)

Table S11. Hyperparameters of top eight APEX ranked by R-squared on CV set.

Top eight APEX models	number of RNN layers	n	m	λ_l2	λ_(multitask_constraint)	λ_BCE
1	3	128	2048	1.00E-05	0.1	1
2	3	256	2048	1.00E-06	0.1	1
3	2	128	512	1.00E-05	0.01	1
4	3	128	512	1.00E-05	0.001	1
5	2	128	2048	1.00E-06	0	1
6	3	256	512	1.00E-06	0	1
7	2	128	2048	1.00E-05	0.01	1
8	2	256	2048	1.00E-06	0.1	1

24-10510 (103241.007081)

Table S12. R-squared of various ML models on independent set. We divided our in-house peptide dataset into a CV set for hyperparameter tuning and ML model selection, and an independent dataset to evaluate prediction performance. The table shows the species-specific prediction performance in terms of R-squared for various ML models that were trained on the CV set and evaluated on the independent set. RF: random forest; GBDT: gradient boosting decision tree; ExtraTree: extra-tree regressor; ElasticNet: elastic net; LinearSVR: linear support vector regression.

Strain	Ensemble APEX v2	Ensemble APEX v1	Single APEX	RF	GBDT	ExtraTree	ElasticNet	LinearSVR
E. coli ATCC11775	0.718159	0.691395	0.668556	0.63738	0.648213	0.551901	0.520619	0.401562
P. aeruginosa PAO1	0.58922	0.550826	0.542209	0.585178	0.437675	0.583725	0.444805	-0.30991
P. aeruginosa PA14	0.577267	0.544865	0.523449	0.525311	0.469315	0.541576	0.464316	0.082555
S. aureus ATCC12600	0.654147	0.65415	0.613936	0.498575	0.496227	0.286083	0.309557	0.254665
E. coli AIC221	0.454362	0.423644	0.448303	0.380448	0.35911	0.338844	0.330079	-0.07492
E. coli AIC222	0.659494	0.642316	0.636876	0.540003	0.476026	0.456006	0.400548	0.208218
K. pneumoniae ATCC13883	0.388908	0.380317	0.439253	0.134236	0.174184	-0.16017	0.095197	-0.17792
A. baumannii ATCC19606	0.694184	0.697003	0.687141	0.664341	0.625813	0.627206	0.541836	0.250524
Methicillin-resistant *S. aureus* ATCC BAA-1556	0.725662	0.739452	0.60538	0.578394	0.587113	0.221953	0.325214	-0.12777
Vancomycin-resistant *E. faecalis* ATCC700802	0.002173	-0.07186	0.304024	-0.04891	-0.09351	-0.09296	-0.01733	-0.01604
Vancomycin-resistant *E. faecium* ATCC700221	0.538558	0.472891	0.514716	0.448163	0.375496	0.28212	0.34944	0.130954
Average	**0.545648**	**0.520455**	**0.543986**	**0.449374**	**0.414151**	**0.330572**	**0.342208**	**0.056538**

24-10510 (103241.007081)

Table S13. Pearson correlation of various ML models on independent set. We divided our in-house peptide dataset into a CV set for hyperparameter tuning and ML model selection, and an independent dataset to evaluate prediction performance. The table shows the species-specific prediction performance in terms of Pearson correlation for various ML models that were trained on the CV set and evaluated on the independent set. RF: random forest; GBDT: gradient boosting decision tree; ExtraTree: extra-tree regressor; ElasticNet: elastic net; LinearSVR: linear support vector regression.

Strain	Ensemble APEX v2	Ensemble APEX v1	Single APEX	RF	GBDT	ExtraTree	ElasticNet	LinearSVR
E. coli ATCC11775	0.847805	0.832404	0.824249	0.800211	0.808419	0.744944	0.722128	0.723382
P. aeruginosa PAO1	0.790273	0.769464	0.775476	0.771463	0.722799	0.771821	0.672709	0.58249
P. aeruginosa PA14	0.767182	0.746836	0.739504	0.725186	0.708727	0.736129	0.69213	0.593574
S. aureus ATCC12600	0.810676	0.81136	0.784346	0.70632	0.714723	0.551515	0.55703	0.598799
E. coli AIC221	0.688587	0.671029	0.693506	0.62584	0.654544	0.597245	0.580406	0.466094
E. coli AIC222	0.813293	0.803721	0.802091	0.736241	0.723161	0.676554	0.632948	0.607374
K. pneumoniae ATCC13883	0.633471	0.627873	0.67345	0.37494	0.451051	0.185642	0.325999	0.404414
A. baumannii ATCC19606	0.841729	0.842787	0.84471	0.818252	0.805014	0.796043	0.764364	0.674816
Methicillin-resistant *S. aureus* ATCC BAA-1556	0.859695	0.865534	0.786457	0.772415	0.78984	0.598027	0.578747	0.580319
Vancomycin-resistant *E. faecalis* ATCC700802	0.21569	0.101714	0.576916	0.119566	-0.51239	0	0	0.193324
Vancomycin-resistant *E. faecium* ATCC700221	0.736427	0.690436	0.741292	0.761104	0.642366	0.581336	0.684935	0.387822
Average	0.727712	0.705742	0.749272	0.655595	0.59166	0.567205	0.564672	0.528401

24-10510 (103241.007081)

Table S14. Spearman correlation of various ML models on independent set. We divided our in-house peptide dataset into a CV set for hyperparameter tuning and ML model selection, and an independent dataset to evaluate prediction performance. The table shows the species-specific prediction performance in terms of Spearman correlation for various ML models that were trained on the CV set and evaluated on the independent set. RF: random forest; GBDT: gradient boosting decision tree; ExtraTree: extra-tree regressor; ElasticNet: elastic net; LinearSVR: linear support vector regression.

Strain	Ensemble APEX v2	Ensemble APEX v1	Single APEX	RF	GBDT	ExtraTree	ElasticNet	LinearSVR
E. coli ATCC11775	0.742143	0.738599	0.695429	0.70242	0.686571	0.675388	0.618169	0.623012
P. aeruginosa PAO1	0.701432	0.671522	0.620927	0.631221	0.673711	0.592132	0.545197	0.481929
P. aeruginosa PA14	0.725533	0.695124	0.619324	0.665001	0.613896	0.63482	0.604422	0.484608
S. aureus ATCC12600	0.578662	0.528803	0.52903	0.435649	0.497982	0.518397	0.419481	0.407387
E. coli AIC221	0.707444	0.707646	0.651341	0.670901	0.670596	0.676166	0.614667	0.583788
E. coli AIC222	0.721146	0.711075	0.666143	0.723491	0.626716	0.638027	0.609947	0.570258
K. pneumoniae ATCC13883	0.603606	0.581167	0.53605	0.28463	0.483702	0.190172	0.379504	0.31371
A. baumannii ATCC19606	0.683889	0.664927	0.661991	0.673022	0.625251	0.695621	0.624478	0.600003
Methicillin-resistant *S. aureus* ATCC BAA-1556	0.474364	0.456122	0.482894	0.44037	0.418149	0.570131	0.418258	0.447617
Vancomycin-resistant *E. faecalis* ATCC700802	0.238607	0.206435	0.361932	0.146958	-0.4407	0	0	0.184652
Vancomycin-resistant *E. faecium* ATCC700221	0.495449	0.436833	0.3834	0.427797	0.359027	0.333684	0.392771	0.397954
Average	**0.60657**	**0.581659**	**0.564406**	**0.527405**	**0.474082**	**0.502231**	**0.475172**	**0.463174**

24-10510 (103241.007081)

Table S15. Pearson correlation of various ML models on human encrypted peptides. The human encrypted peptides from Torres et al.[1] constitute a subset of our in-house peptide dataset. We treated this subset as the test data and excluded it from ML model training. The table shows species-specific prediction performance in terms of Pearson correlation for various ML models.

Strain	Ensemble APEX v2	Ensemble APEX v1	Single APEX	RF	GBDT	ExtraTree	ElasticNet	LinearSVR	Torres *et al.*
E. coli ATCC11775	0.482757	0.496734	0.494723	0.435897	0.365328	0.407247	0.428736	0.320893	0.50776
P. aeruginosa PAO1	0.710869	0.686719	0.579761	0.731061	0.679557	0.699833	0.569554	0.143789	0.582418
P. aeruginosa PA14	0.734694	0.73291	0.671569	0.749065	0.68225	0.749057	0.725144	0.254412	0.712036
S. aureus ATCC12600	0.126253	0.186218	0.199264	0.076771	0.020007	0.07906	-0.01304	-0.26126	0.42482
E. coli AIC221	0.403527	0.404244	0.389027	0.521815	0.413622	0.483885	0.398567	0.200215	0.311953
E. coli AIC222	0.653576	0.655583	0.741654	0.680734	0.606786	0.762781	0.411687	0.316373	0.433937
K. pneumoniae ATCC13883	0.460277	0.466471	0.507818	0.456984	0.295749	-0.0989	0.22667	0.254919	0.282847
A. baumannii ATCC19606	0.773425	0.773293	0.73518	0.72433	0.760228	0.753661	0.510162	0.601455	0.430196
Average	**0.543172**	**0.550272**	**0.539874**	**0.547082**	**0.477941**	**0.479578**	**0.407185**	**0.228849**	**0.460746**

24-10510 (103241.007081)

Table S16. Spearman correlation of various ML models on human encrypted peptides. The human encrypted peptides from Torres et al.[1] constitute a subset of our in-house peptide dataset. We treated this subset as the test data and excluded it from ML model training. The table shows species-specific prediction performance in terms of Spearman correlation for various ML models.

Strain	Ensemble APEX v2	Ensemble APEX v1	Single APEX	RF	GBDT	ExtraTree	ElasticNet	LinearSVR	Torres et al.
E. coli ATCC11775	0.523727	0.535223	0.506701	0.492004	0.370203	0.473523	0.440344	0.220609	0.497825
P. aeruginosa PAO1	0.719366	0.718463	0.622332	0.673903	0.575407	0.648483	0.626982	0.03871	0.615923
P. aeruginosa PA14	0.76658	0.760304	0.657554	0.657068	0.683707	0.662736	0.688632	0.225784	0.721803
S. aureus ATCC12600	0.254791	0.233065	0.259042	0.244744	0.081066	0.177344	0.025934	-0.20224	0.4654
E. coli AIC221	0.403575	0.397753	0.346453	0.491014	0.378764	0.376929	0.409659	0.137742	0.31409
E. coli AIC222	0.584407	0.582971	0.583851	0.589457	0.487067	0.636713	0.55179	0.17471	0.477801
K. pneumoniae ATCC13883	0.420755	0.417876	0.315581	0.313113	0.332973	-0.10474	0.440967	0.066277	0.293077
A. baumannii ATCC19606	0.634678	0.64788	0.602808	0.6092	0.62634	0.611794	0.513264	0.458603	0.455221
Average	0.538485	0.536692	0.48679	0.508813	0.441941	0.435348	0.462197	0.140025	0.480142

24-10510 (103241.007081)

Table S17. Cytotoxic activity of AEPs and MEPs. The cytotoxic activity was expressed in terms of CC_{50} values (μmol L^{-1}), *i.e.*, cytotoxic concentration values needed to damage 50% of the HEK293T cells present in each condition. The values were estimated by non-linear regressions based on the screen of all active AEPs and MEPs at concentrations from 8 to 128 μmol L^{-1}, to ensure coverage of all tested antimicrobial activity concentrations. The experiments were done in three independent biological replicates with two technical replicates within each biological replicate. The therapeutic index (TI) was calculated to show the margin of safety obtained by comparing the lowest MIC values (μmol L^{-1}) obtained in the antimicrobial activity assays to the CC_{50} values of each active AEP or MEP.

Peptide	CC_{50} (μmol L^{-1})	MIC (μmol L^{-1})	TI	Peptide	CC_{50} (μmol L^{-1})	MIC (μmol L^{-1})	TI
Equusin-1	>128	1	>128	Megalocerin-1	>128	8	>16
Hesperelin-1	>128	2	>64	Pinguinusin-1	>128	4	>32
Elephasin-1	>128	4	>32	Ursusin-1	>128	64	>2
Arctodutin-1	>128	2	>64	Elephasin-2	>128	1	>128
Arctoterin-1	>128	64	>2	Mammuthusin-4	>128	2	>64
Lophiosin-1	68.02	16	>8	Psephotellin-1	>128	16	>8
Mammutin-1	>128	2	>64	Eudyptin-1	>128	64	>2
Ararin-1	>128	32	>4	Paleopropin-2	>128	16	>8
Mylodonin-1	>128	8	>16	Hydrodamin-2	>128	8	>16
Mammuthusin-1	>128	4	>32	Hydrodamin-3	>128	16	>8
Paleopropin-1	>128	32	>4	Hesperelin-4	>128	64	>2
Bisonin-1	>128	16	>8	Mylodonin-2	>128	32	>4
Hesperelin-2	>128	8	>16	Anomalopterin-2	>128	32	>4
Equusin-2	>128	8	>16	Equusin-3	>128	4	>32
Mammuthusin-2	>128	32	>4	Bisonin-2	>128	8	>16
Mammuthusin-3	>128	16	>8	Mammuthusin-5	>128	32	>4
Hydrodamin-1	>128	4	>32	Smilodin-1	>128	64	>2
Xenothrixin-1	70.77	8	8.84	Mylodonin-3	>128	16	>8
Hesperelin-3	>128	64	>2	Mylodonin-4	>128	32	>4
Ararin-2	>128	16	>8	Equusin-4	>128	16	>8

Anomalopterin-1	>128	64	>2

Supplementary Table 18. Method for the chromatography coupled to mass spectrometry experiments. The solvent gradient used is standard for small molecules and peptides and solvent used in the experiments are Fisher optima grades.

Time (min)	A (%)	B (%)	Flow rate (mL min^{-1})
0	95	5	0.5
0.5	95	5	0.5
2.5	5	95	0.5 (linear gradient)
3	5	95	0.5
3.25	5	95	0.5

References

[00171] The following publications, to which the superscripted numerals in the preceding disclosure refer, have potential relevance to the presently disclosed subject matter.

1. World Health Organization. *New Report Calls for Urgent Action to Avert Antimicrobial Resistance Crisis*. (2019).

2. Zuckerkandl, E. & Pauling, L. Molecules as documents of evolutionary history. *J Theor Biol* **8**, 357–366 (1965).

3. Maasch, J. R. M. A., Torres, M. D. T., Melo, M. C. R. & de la Fuente-Nunez, C. Molecular de-extinction of ancient antimicrobial peptides enabled by machine learning. *Cell Host Microbe* **31**, 1260–1274 (2023) doi:10.1016/j.chom.2023.07.001.

4. Wong, F., de la Fuente-Nunez, C. & Collins, J. J. Leveraging artificial intelligence in the fight against infectious diseases. *Science* **381**, 164–170 (2023).

5. Porto, W. F. *et al.* In silico optimization of a guava antimicrobial peptide enables combinatorial exploration for peptide design. *Nat Commun* **9**, 1490 (2018).

6. Torres, M. D. T. & de la Fuente-Nunez, C. Toward computer-made artificial antibiotics. *Curr Opin Microbiol* **51**, 30–38 (2019).

7. Wan, F., Kontogiorgos-Heintz, D. & de la Fuente-Nunez, C. Deep generative models for peptide design. *Digital Discovery* **1**, 195–208 (2022).

8. Das, P. *et al.* Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations. *Nat Biomed Eng* **5**, 613–623 (2021).

9. Ma, Y. *et al.* Identification of antimicrobial peptides from the human gut microbiome using deep learning. *Nat Biotechnol* **40**, 921–931 (2022).

10. Xu, J. *et al.* Comprehensive assessment of machine learning-based methods for predicting antimicrobial peptides. *Brief Bioinform* **22**, (2021).

11. Stokes, J. M. *et al.* A Deep Learning Approach to Antibiotic Discovery. *Cell* **180**, 688-702.e13 (2020).

12. Capecchi, A. *et al.* Machine learning designs non-hemolytic antimicrobial peptides. *Chem Sci* **12**, 9221–9232 (2021).

13. Green, A. G. *et al.* A convolutional neural network highlights mutations relevant to antimicrobial resistance in Mycobacterium tuberculosis. *Nat Commun* **13**, 3817 (2022).

14. Weis, C. *et al.* Direct antimicrobial resistance prediction from clinical MALDI-TOF mass spectra using machine learning. *Nat Med* **28**, 164–174 (2022).

proteome. *Nat Biomed Eng* **6**, 67–75 (2022).

16. 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).

17. Mulani, M. S., Kamble, E. E., Kumkar, S. N., Tawre, M. S. & Pardesi, K. R. Emerging Strategies to Combat ESKAPE Pathogens in the Era of Antimicrobial Resistance: A Review. *Front Microbiol* **10**, (2019).

18. Pane, K. *et al.* Antimicrobial potency of cationic antimicrobial peptides can be predicted from their amino acid composition: Application to the detection of "cryptic" antimicrobial peptides. *J Theor Biol* **419**, 254–265 (2017).

19. Zhao, M., Lee, W.-P., Garrison, E. P. & Marth, G. T. SSW Library: An SIMD Smith-Waterman C/C++ Library for Use in Genomic Applications. *PLoS One* **8**, e82138 (2013).

20. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J Open Source Softw* **3**, 861 (2018).

21. Torres, M. D. T., Sothiselvam, S., Lu, T. K. & de la Fuente-Nunez, C. Peptide Design Principles for Antimicrobial Applications. *J Mol Biol* **431**, 3547–3567 (2019).

22. Cesaro, A. *et al.* Synthetic Antibiotic Derived from Sequences Encrypted in a Protein from Human Plasma. *ACS Nano* **16**, 1880–1895 (2022).

23. Pizzo, E. *et al.* Novel bioactive peptides from PD-L1/2, a type 1 ribosome inactivating protein from Phytolacca dioica L. Evaluation of their antimicrobial properties and anti-biofilm activities. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **1860**, 1425–1435 (2018).

24. Dennison, S. R., Harris, F., Mura, M. & Phoenix, D. A. An Atlas of Anionic Antimicrobial Peptides from Amphibians. *Curr Protein Pept Sci* **19**, 823–838 (2018).

25. Deber, C. M. & Stone, T. A. Relative role(s) of leucine versus isoleucine in the folding of membrane proteins. *Peptide Science* **111**, e24075 (2019).

26. Cesaro, A. *et al.* Synthetic Antibiotic Derived from Sequences Encrypted in a Protein from Human Plasma. *ACS Nano* **16**, 1880–1895 (2022).

27. Moffat, L. & Jones, D. T. Increasing the accuracy of single sequence prediction methods using a deep semi-supervised learning framework. *Bioinformatics* **37**, 3744–3751 (2021).

28. Roccatano, D., Colombo, G., Fioroni, M. & Mark, A. E. Mechanism by which 2,2,2-trifluoroethanol/water mixtures stabilize secondary-structure formation in peptides: A

12184 (2002).

29. Tossi, A., Sandri, L. & Giangaspero, A. Amphipathic, alpha-helical antimicrobial peptides. *Biopolymers* **55**, 4–30 (2000).

30. Ayoub Moubareck, C. & Hammoudi Halat, D. Insights into Acinetobacter baumannii: A Review of Microbiological, Virulence, and Resistance Traits in a Threatening Nosocomial Pathogen. *Antibiotics* **9**, 119 (2020).

31. Pachori, P., Gothalwal, R. & Gandhi, P. Emergence of antibiotic resistance Pseudomonas aeruginosa in intensive care unit; a critical review. *Genes Dis* **6**, 109–119 (2019).

32. Tyers, M. & Wright, G. D. Drug combinations: a strategy to extend the life of antibiotics in the 21st century. *Nat Rev Microbiol* **17**, 141–155 (2019).

33. Lázár, V., Snitser, O., Barkan, D. & Kishony, R. Antibiotic combinations reduce Staphylococcus aureus clearance. *Nature* **610**, 540–546 (2022).

34. Nim, S. *et al.* Disrupting the α-synuclein-ESCRT interaction with a peptide inhibitor mitigates neurodegeneration in preclinical models of Parkinson's disease. *Nat Commun* **14**, 2150 (2023).

35. Silva, O. N. *et al.* Repurposing a peptide toxin from wasp venom into antiinfectives with dual antimicrobial and immunomodulatory properties. *Proceedings of the National Academy of Sciences* **118**, e2025351118 (2021).

36. Torres, M. D. T. *et al.* Structure-function-guided exploration of the antimicrobial peptide polybia-CP identifies activity determinants and generates synthetic therapeutic candidates. *Commun Biol* **1**, 221 (2018).

37. Ling, L. L. *et al.* A new antibiotic kills pathogens without detectable resistance. *Nature* **517**, 455–459 (2015).

38. Cullen, T. W. *et al.* Antimicrobial peptide resistance mediates resilience of prominent gut commensals during inflammation. *Science* **347**, 170–175 (2015).

39. Eisenberg, D., Weiss, R. M. & Terwilliger, T. C. The helical hydrophobic moment: a measure of the amphiphilicity of a helix. *Nature* **299**, 371–374 (1982).

40. Altenhoff, A. M. *et al.* OMA orthology in 2021: website overhaul, conserved isoforms, ancestral gene order and more. *Nucleic Acids Res* **49**, D373–D379 (2021).

41. Kulmanov, M., Zhapa-Camacho, F. & Hoehndorf, R. DeepGOWeb: fast and accurate protein function prediction on the (Semantic) Web. *Nucleic Acids Res* **49**, W140–W146 (2021).

(2000).

43. Parks, D. H. *et al.* GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res* **50**, D785–D794 (2022).

44. Sukumaran, J. & Holder, M. T. DendroPy: a Python library for phylogenetic computing. *Bioinformatics* **26**, 1569–1571 (2010).

45. Cho, K., van Merriënboer, B., Bahdanau, D. & Bengio, Y. On the properties of neural machine translation: Encoder-decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar. Association for Computational Linguistics.

46. Ba, J. L., Kiros, J. R. & Hinton, G. E. Layer normalization. *arXiv preprint arXiv:1607.06450* (2016).

47. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. in *Proceedings of the IEEE conference on computer vision and pattern recognition* 770–778 (2016).

48. Nair, V. & Hinton, G. E. Rectified linear units improve restricted boltzmann machines. in *Icml* (2010).

49. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* **15**, 1929–1958 (2014).

50. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. in *ICLR* (2015).

51. Müller, A. T., Gabernet, G., Hiss, J. A. & Schneider, G. modlAMP: Python for antimicrobial peptides. *Bioinformatics* **33**, 2753–2755 (2017).

52. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* **47**, D506–D515 (2019).

53. Cesaro, A., Torres, M. & de la Fuente-Nunez, C. Methods for the design and characterization of peptide antibiotics. in *Methods in Enzymology* **663**, 303–326 (Academic Press, 2022). doi:10.1016/bs.mie.2021.11.003.

54. Powell, M. F. *et al.* Peptide Stability in Drug Development. II. Effect of Single Amino Acid Substitution and Glycosylation on peptide Reactivity in Human Serum. *Pharm Res* **10**, 1268–1273 (1993).

55. Micsonai, A. *et al.* BeStSel: webserver for secondary structure and fold prediction for protein CD spectroscopy. *Nucleic Acids Res* **50**, W90–W98 (2022).

What is claimed:

1. An antimicrobial peptide having an amino acid sequence of any one of SEQ ID NOs:1-41.

2. A composition comprising a therapeutically effective amount of an antimicrobial peptide having an amino acid sequence of any one of SEQ ID NOs:1-41 and a pharmaceutically acceptable carrier, diluent, or excipient.

3. The composition comprising according to claim 2 comprising any two or more of the antimicrobial peptides.

4. A method of treating an antimicrobial infection comprising contacting the infection with a therapeutically effective amount of antimicrobial peptide according to claim 1 or a composition according to claim 2 or claim 3.

5. A method of treating an antimicrobial infection in a subject comprising administering to the subject in need thereof a therapeutically effective amount of an antimicrobial peptide according to claim 1 or a composition according to claim 2 or claim 3.

6. A method comprising contacting a biofilm with an effective amount of an antimicrobial peptide according to claim 1 or a composition according to claim 2 or claim 3

7. A method of identifying an antimicrobial peptide comprising:

training a deep learning model using a peptide dataset;

representing peptides as inputs into the deep learning model;

utilizing a hybrid of recurrent and attention neural networks for the deep learning model in order to extract peptide sequence information from the peptide dataset;

selecting hyperparameters for the deep learning model;

mining candidate peptides from a proteome of an extinct organism;

screening the candidate peptides in order to identify an antimicrobial peptide from the candidate peptides.

8. The method according to claim 7 comprising further constraining the training of the deep learning model using bacterial distance curation.

9. The method according to claim 7 wherein the mining includes setting a selectivity scoring method for selecting and filtering candidate antimicrobial peptides.

10. A method of identifying an antimicrobial peptide comprising training and using a deep learning model according to steps disclosed in the present specification and figures.

11. A method of identifying an antimicrobial peptide comprising training and using a deep learning model according to steps depicted in FIG. 1.

FIG. 1

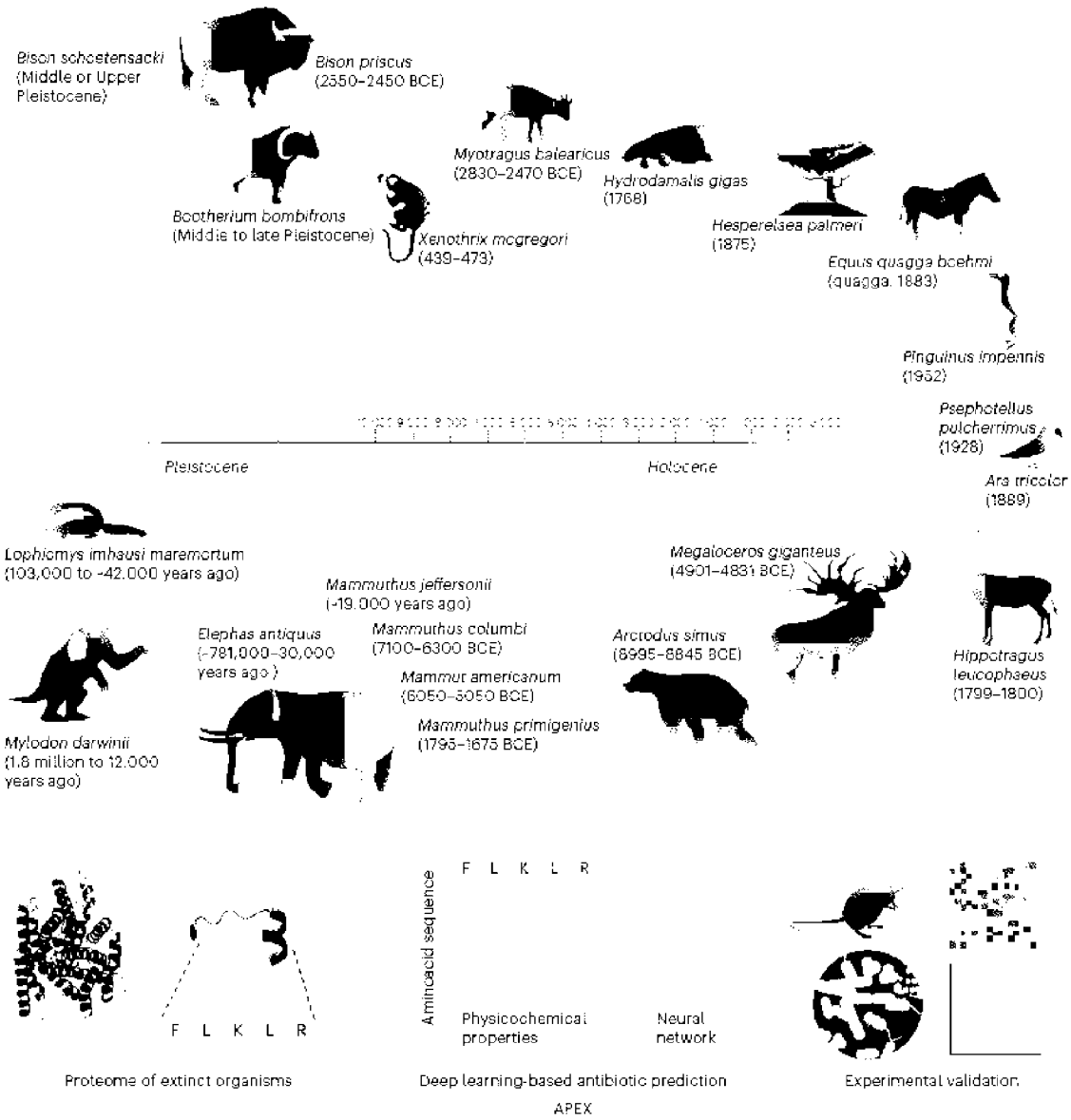

Bison schoetensacki
(Middle or Upper
Pleistocene)
Bison priscus
(2550–2450 BCE)
Myotragus balearicus
(2830–2470 BCE)
Hydrodamalis gigas
(1768)
Bootherium bombifrons
(Middle to late Pleistocene)
Xenothrix mcgregori
(439–473)
Hesperelaea palmeri
(1875)
Equus quagga boehmi
(quagga, 1883)
Pinguinus impennis
(1952)
Psephotellus
pulcherrimus
(1928)
Ara tricolor
(1889)
Pleistocene
Holocene
Lophiomys imhausi maremortum
(103,000 to ~42,000 years ago)
Mammuthus jeffersonii
(~19,000 years ago)
Megaloceros giganteus
(4901–4831 BCE)
Elephas antiquus
(~781,000–30,000
years ago)
Mammuthus columbi
(7100–6300 BCE)
Arctodus simus
(8995–8845 BCE)
Hippotragus
leucophaeus
(1799–1800)
Mammut americanum
(6050–5050 BCE)
Mammuthus primigenius
(1795–1675 BCE)
Mylodon darwinii
(1.6 million to 12,000
years ago)
F L K L R
Amino acid sequence
Physicochemical
properties
Neural
network
Proteome of extinct organisms
Deep learning-based antibiotic prediction
APEX
Experimental validation

FIG. 2

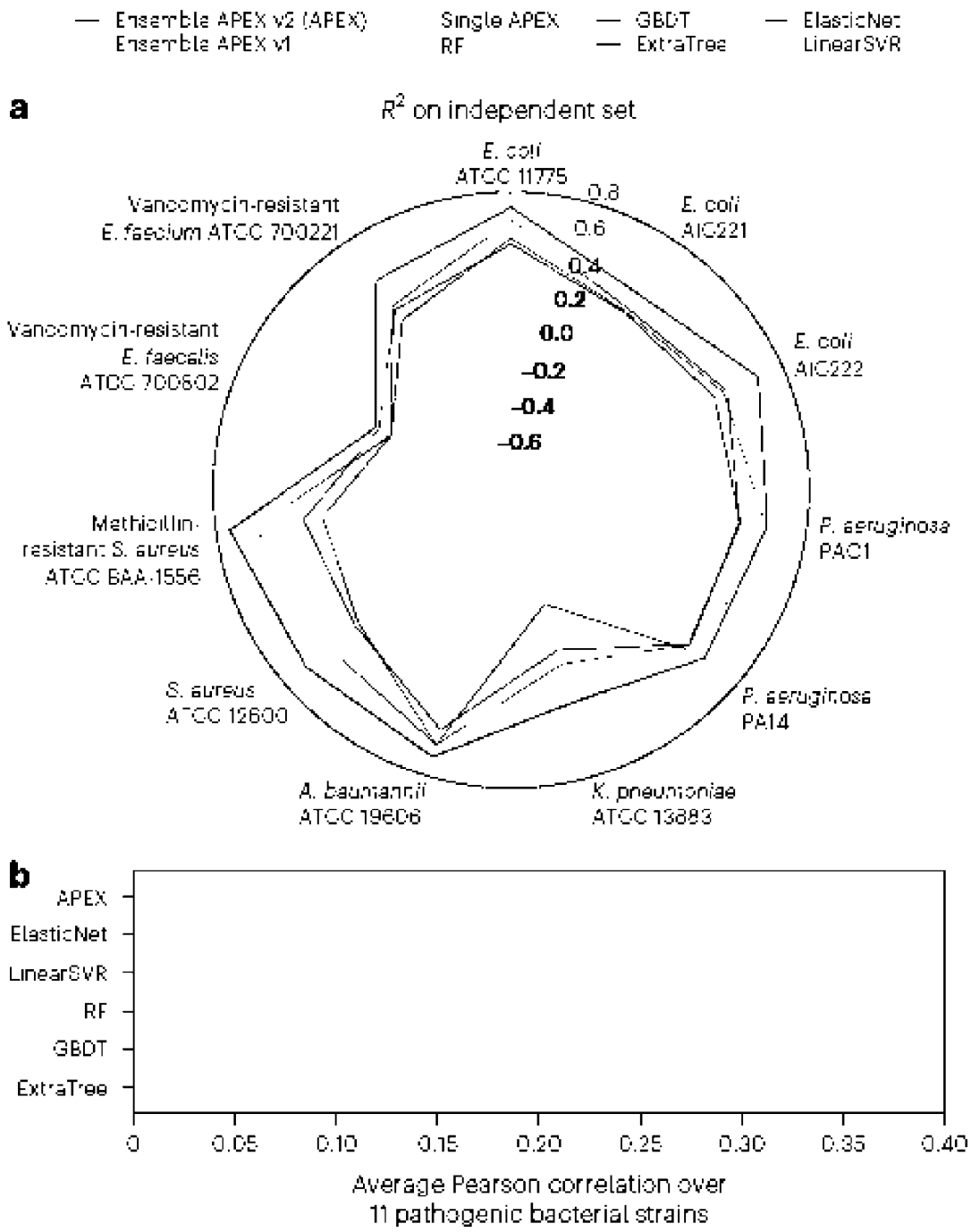

Ensemble APEX v2 (APEX)
Ensemble APEX v1
Single APEX
RF
GBDT
ExtraTree
ElasticNet
LinearSVR
a
R^2 on independent set
E. coli ATCC 11775
E. coli AIC221
E. coli AIC222
P. aeruginosa PAO1
P. aeruginosa PA14
K. pneumoniae ATCC 13883
A. baumannii ATCC 19606
S. aureus ATCC 12600
Methicillin-resistant S. aureus ATCC BAA-1556
Vancomycin-resistant E. faecalis ATCC 700802
Vancomycin-resistant E. faecium ATCC 700221
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
b
APEX
ElasticNet
LinearSVR
RF
GBDT
ExtraTree
0
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Average Pearson correlation over 11 pathogenic bacterial strains

FIG. 3

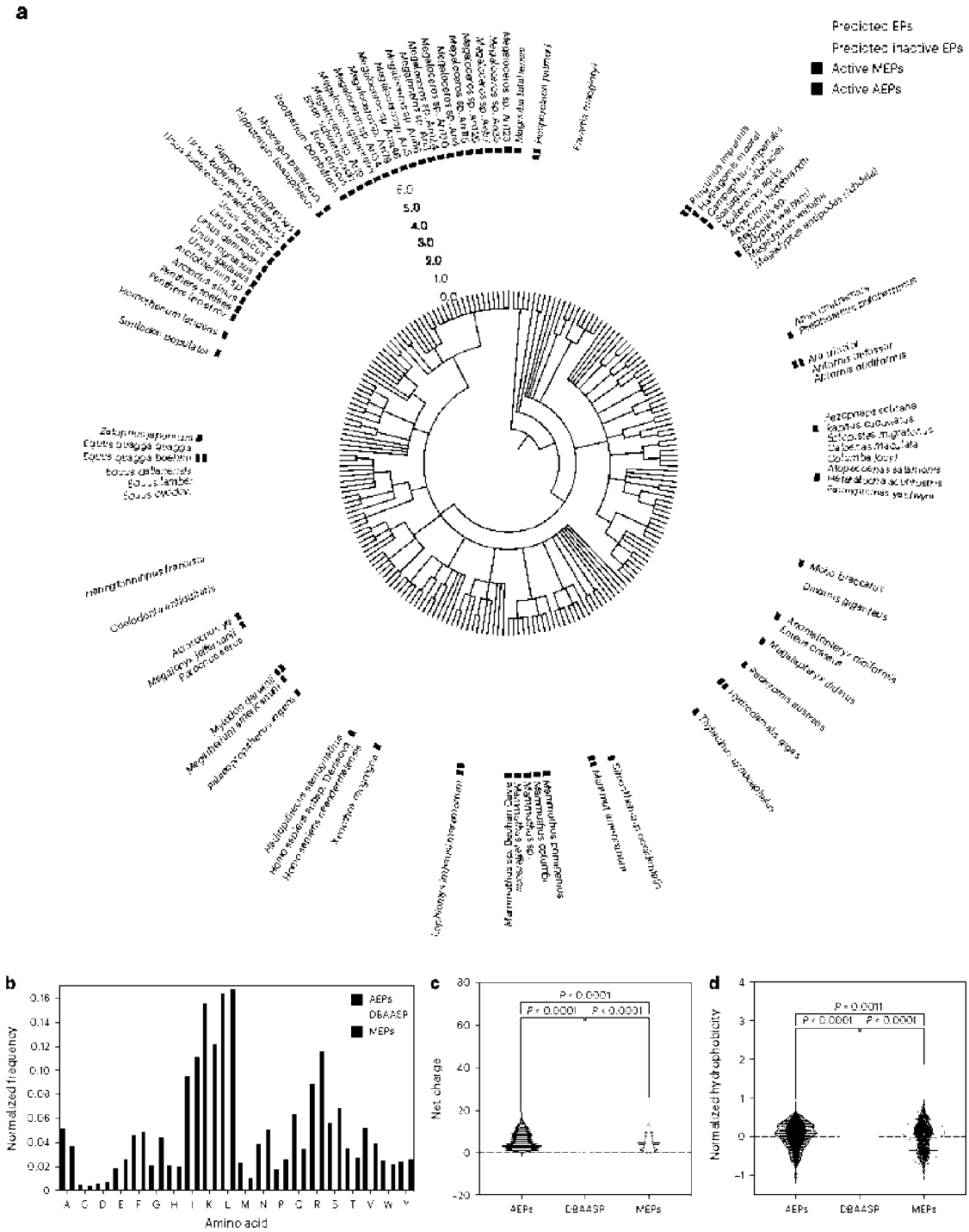

a
Predicted EPs
Predicted inactive EPs
Active MEPs
Active AEPs
b
AEPs
DBAASP
MEPs
Normalized frequency
Amino acid
c
P < 0.0001
Net charge
AEPs
DBAASP
MEPs
d
P = 0.0011
P < 0.0001
Normalized hydrophobicity
AEPs
DBAASP
MEPs

FIG. 4

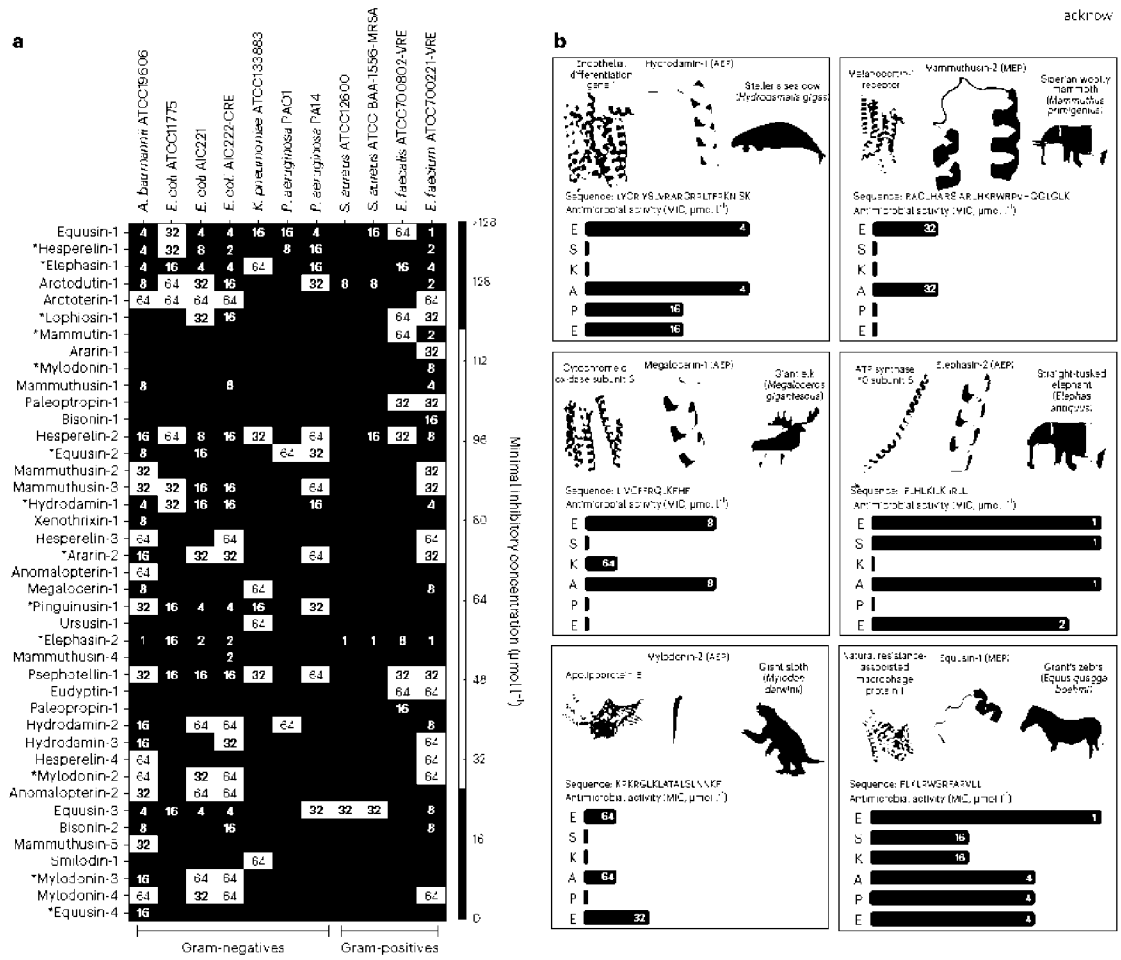

a
A. baumannii ATCC19606
E. coli ATCC11775
E. coli AIC221
E. coli AIC222/CRE
K. pneumoniae ATCC13883
P. aeruginosa PAO1
P. aeruginosa PA14
S. aureus ATCC12600
S. aureus ATCC BAA-1556-MRSA
E. faecalis ATCC700802/VRE
E. faecium ATCC700221/VRE
Equusin-1
*Hesperelin-1
*Elephasin-1
Arctodutin-1
Arctoterin-1
*Lophiosin-1
*Mammutin-1
Ararin-1
*Mylodonin-1
Mammuthusin-1
Paleoptropin-1
Bisonin-1
Hesperelin-2
*Equusin-2
Mammuthusin-2
Mammuthusin-3
*Hydrodamin-1
Xenothrixin-1
Hesperelin-3
*Ararin-2
Anomalopterin-1
Megalocerin-1
*Pinguinusin-1
Ursusin-1
*Elephasin-2
Mammuthusin-4
Psephotellin-1
Eudyptin-1
Paleopropin-1
Hydrodamin-2
Hydrodamin-3
Hesperelin-4
*Mylodonin-2
Anomalopterin-2
Equusin-3
Bisonin-2
Mammuthusin-5
Smilodin-1
*Mylodonin-3
Mylodonin-4
*Equusin-4
Gram-negatives
Gram-positives
Minimal inhibitory concentration (μmol L⁻¹)
>128
128
112
96
80
64
48
32
16
0
b
Hydrodamin-1 (AEP)
Steller's sea cow (Hydrodamalis gigas)
Sequence: LYQRYSLVRARQRSLTPRNLSK
Antimicrobial activity (MIC, μmol L⁻¹)
Mammuthusin-2 (MEP)
Siberian woolly mammoth (Mammuthus primigenius)
Sequence: RACLHARSARLHKRWRPV-QGLGLK
Megalocerin-1 (AEP)
Giant elk (Megaloceros giganteus)
Sequence: LVCFFRQLKFHF
Elephasin-2 (AEP)
Straight-tusked elephant (Elephas antiquus)
Sequence: FLHLKLKRLL
Mylodonin-2 (AEP)
Giant sloth (Mylodon darwinii)
Sequence: KFKRGLKLATALSLAAKF
Equusin-1 (MEP)
Grant's zebra (Equus quagga boehmi)
Sequence: FLYLFWGRFARVLL
E
S
K
A
P

FIG. 5

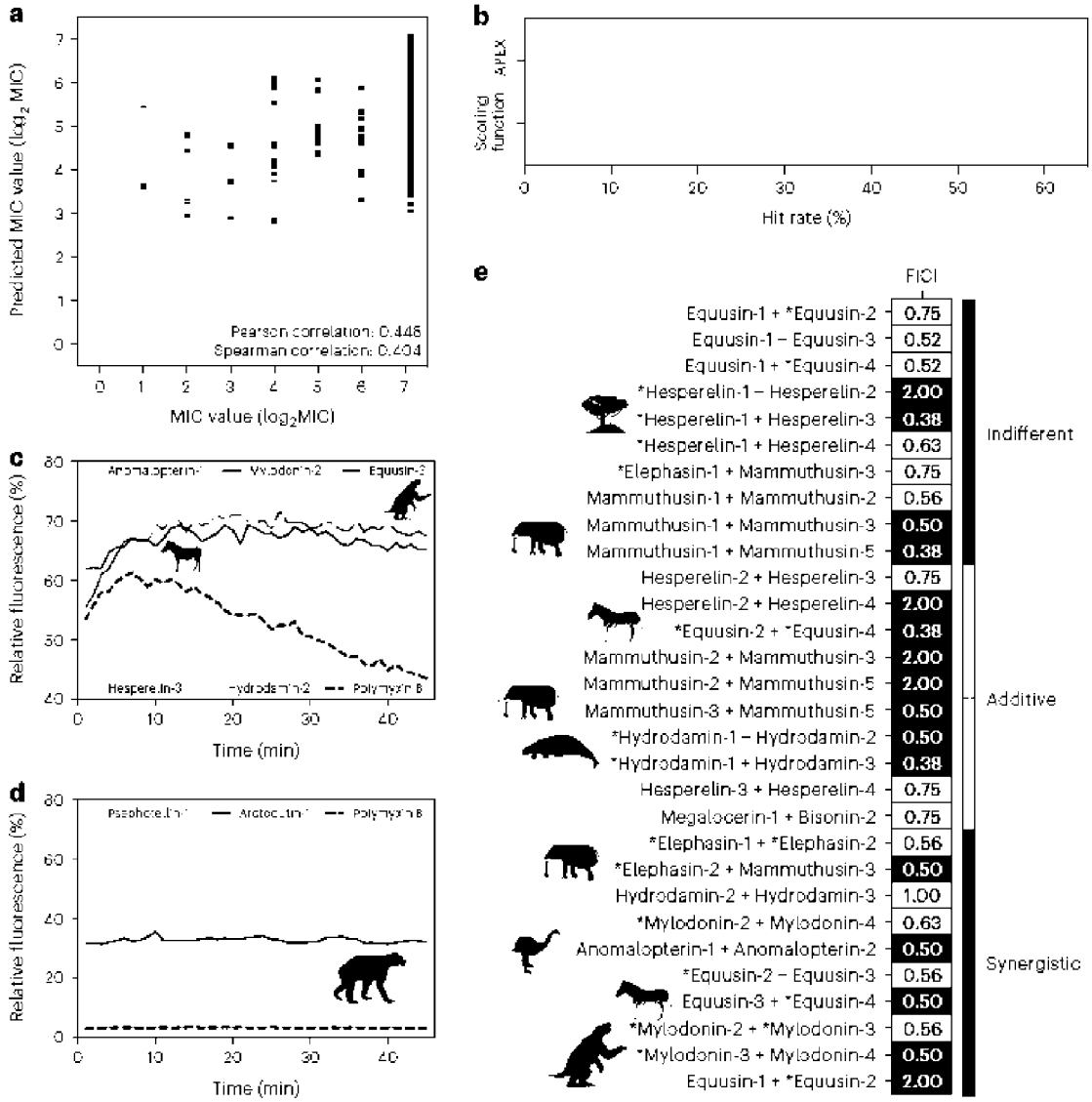
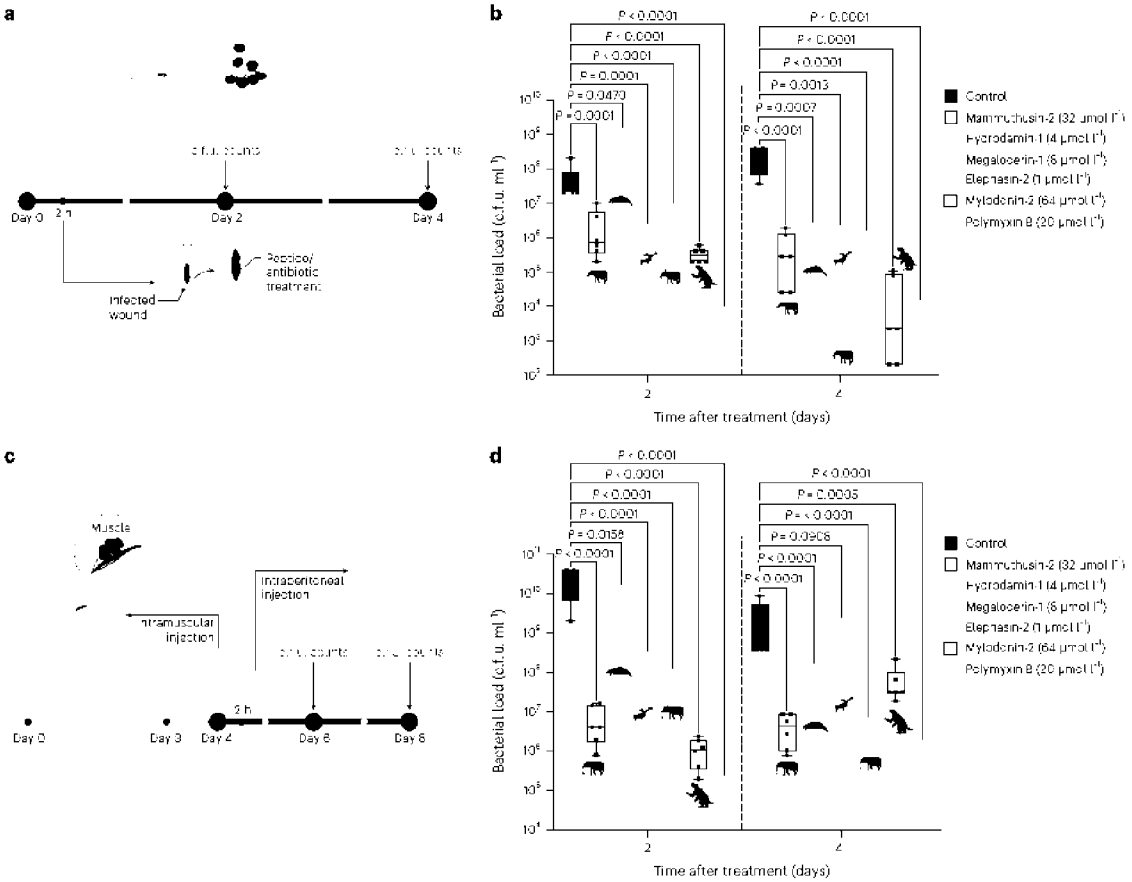

a
Predicted MIC value (log2 MIC)
MIC value (log2MIC)
Pearson correlation: 0.446
Spearman correlation: 0.404
b
Scoring function
APEX
Hit rate (%)
c
Relative fluorescence (%)
Time (min)
Hesperelin-3
Polymyxin B
d
Relative fluorescence (%)
Time (min)
Polymyxin B
e
FICI
Equusin-1 + *Equusin-2 0.75
Equusin-1 – Equusin-3 0.52
Equusin-1 + *Equusin-4 0.52
*Hesperelin-1 – Hesperelin-2 2.00
*Hesperelin-1 + Hesperelin-3 0.38
*Hesperelin-1 + Hesperelin-4 0.63
*Elephasin-1 + Mammuthusin-3 0.75
Mammuthusin-1 + Mammuthusin-2 0.56
Mammuthusin-1 + Mammuthusin-3 0.50
Mammuthusin-1 + Mammuthusin-5 0.38
Hesperelin-2 + Hesperelin-3 0.75
Hesperelin-2 + Hesperelin-4 2.00
*Equusin-2 + *Equusin-4 0.38
Mammuthusin-2 + Mammuthusin-3 2.00
Mammuthusin-2 + Mammuthusin-5 2.00
Mammuthusin-3 + Mammuthusin-5 0.50
*Hydrodamin-1 – Hydrodamin-2 0.50
*Hydrodamin-1 + Hydrodamin-3 0.38
Hesperelin-3 + Hesperelin-4 0.75
Megalocerin-1 + Bisonin-2 0.75
*Elephasin-1 + *Elephasin-2 0.56
*Elephasin-2 + Mammuthusin-3 0.50
Hydrodamin-2 + Hydrodamin-3 1.00
*Mylodonin-2 + Mylodonin-4 0.63
Anomalopterin-1 + Anomalopterin-2 0.50
*Equusin-2 – Equusin-3 0.56
Equusin-3 + *Equusin-4 0.50
*Mylodonin-2 + *Mylodonin-3 0.56
*Mylodonin-3 + Mylodonin-4 0.50
Equusin-1 + *Equusin-2 2.00
Indifferent
Additive
Synergistic

FIG. 6


a
Day 0
Day 2
Day 4
Infected wound
Peptide/ antibiotic treatment
b
Bacterial load (c.f.u. ml⁻¹)
Time after treatment (days)
Control
Mammuchusin-2 (32 μmol l⁻¹)
Hydradamin-1 (4 μmol l⁻¹)
Megalocerin-1 (8 μmol l⁻¹)
Elephasin-2 (1 μmol l⁻¹)
Mylodenin-2 (64 μmol l⁻¹)
Polymyxin B (20 μmol l⁻¹)
c
Muscle
Intraperitoneal injection
Intramuscular injection
2 h
Day 0
Day 3
Day 4
Day 6
Day 8
d
Bacterial load (c.f.u. ml⁻¹)
Time after treatment (days)
Control
Mammuchusin-2 (32 μmol l⁻¹)
Hydradamin-1 (4 μmol l⁻¹)
Megalocerin-1 (8 μmol l⁻¹)
Elephasin-2 (1 μmol l⁻¹)
Mylodenin-2 (64 μmol l⁻¹)
Polymyxin B (20 μmol l⁻¹)

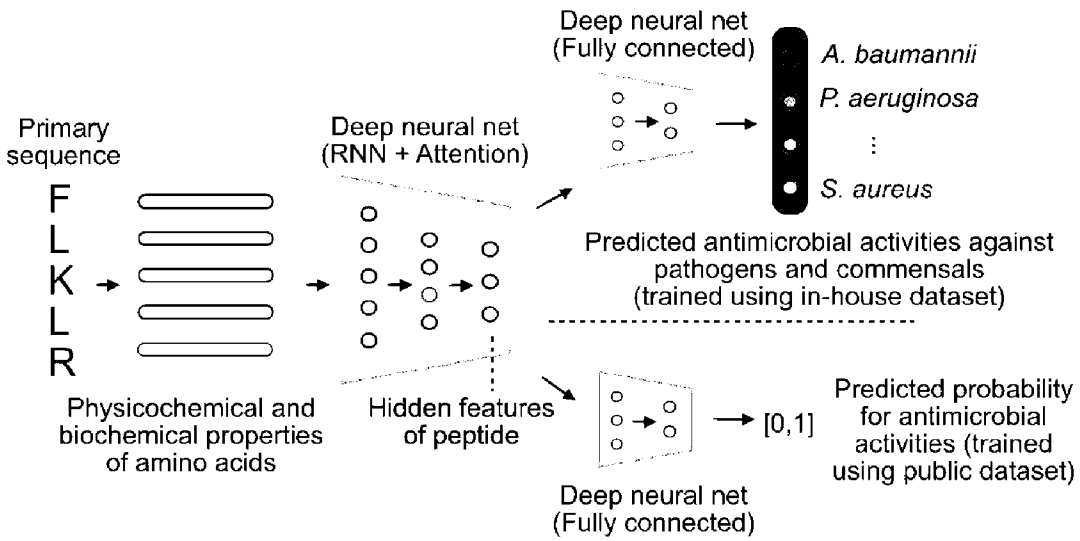


FIG. 7. Schematic illustration of APEX. APEX utilized a hybrid of recurrent and attention neural networks to extract peptide sequence information. Extracted hidden features for peptides from in-house or public datasets were processed by two fully connected neural networks to predict species-specific antimicrobial activities (i.e., a multi-output regression task) or general antimicrobial or not (i.e., a binary classification task), respectively. The inventors adopted this multitask framework to accurately predict whether a given peptide sequence was likely to have antimicrobial activity. This figure was created with BioRender.com.

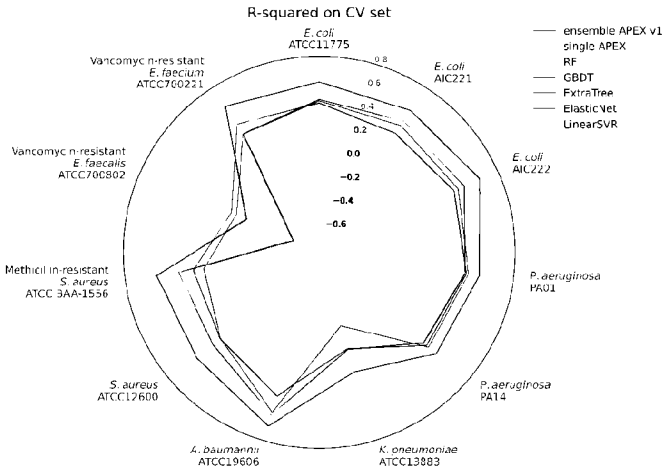


FIG. 8. R-squared scores of various ML models on cross-validation (CV) set. The inventors divided the inventive in-house peptide dataset into a CV set for hyperparameter tuning and ML model selection, and an independent dataset to evaluate the prediction performance. The figure shows the average species-specific prediction performance of 5-fold CV in terms of R-squared for various ML models on the CV set. The radius reflects the R-squared value. RF: random forest; GBDT: gradient boosting decision tree; ExtraTree: extra-tree regressor; ElasticNet: elastic net; LinearSVR: linear support vector regression.

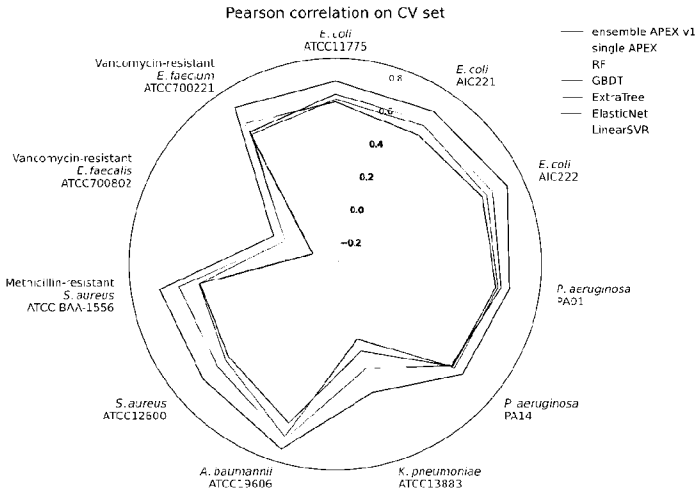


FIG. 9. Pearson correlation scores of various ML models on cross-validation (CV) set. The inventors divided the inventive in-house peptide dataset into a CV set for hyperparameter tuning and ML model selection, and an independent dataset to evaluate the prediction performance. The figure shows the average species-specific prediction performance of 5-fold CV in terms of Pearson correlation for various ML models on the CV set. The radius reflects the Pearson correlation value. RF: random forest; GBDT: gradient boosting decision tree; ExtraTree: extra-tree regressor; ElasticNet: elastic net; LinearSVR: linear support vector regression.

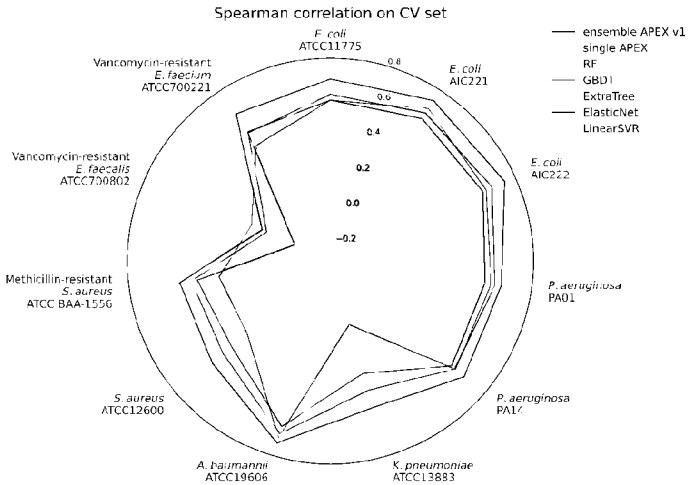


FIG. 10. Spearman correlation scores of various ML models on cross-validation (CV) set. The inventors divided the inventive in-house peptide dataset into a CV set for hyperparameter tuning and ML model selection, and an independent dataset to evaluate the prediction performance. The figure shows the average species-specific prediction performance of 5-fold CV in terms of Spearman correlation for various ML models on the CV set. The radius reflects the Spearman correlation value. RF: random forest; GBDT: gradient boosting decision tree; ExtraTree: extra-tree regressor; ElasticNet: elastic net; LinearSVR: linear support vector regression.

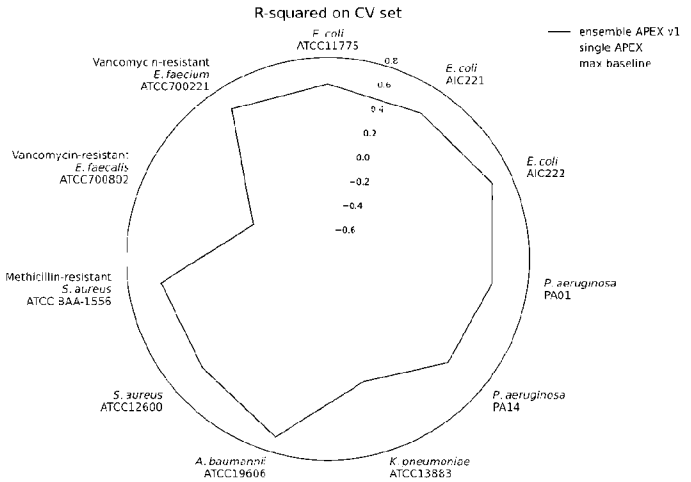


FIG. 11. R-squared scores of various ML models on cross-validation (CV) set. The inventors divided the inventive in-house peptide dataset into a CV set for hyperparameter tuning and ML model selection, and an independent dataset to evaluate the prediction performance. The figure shows the average species-specific prediction performance of 5-fold CV in terms of R-squared for APEX models and maximum performance from baseline ML models on the CV set. The radius reflects the R-squared value. Max baseline denotes the highest R-squared values from baseline ML models, including elastic net, linear support vector regression, extra-trees regressor, random forest and gradient boosting decision tree.

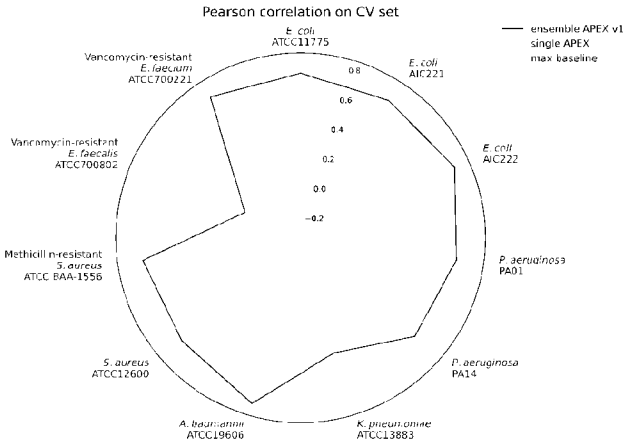


FIG. 12. Pearson correlation scores of various ML models on CV set. The inventors divided the inventive in-house peptide dataset into a CV set for hyperparameter tuning and ML model selection, and an independent dataset to evaluate the prediction performance. The figure shows the average species-specific prediction performance of 5-fold CV in terms of Pearson correlation for APEX models and maximum performance from baseline ML models on the CV set. The radius reflects the Pearson correlation value. Max baseline denotes the highest Pearson correlation values from baseline ML models, including elastic net, linear support vector regression, extra-trees regressor, random forest and gradient boosting decision tree.

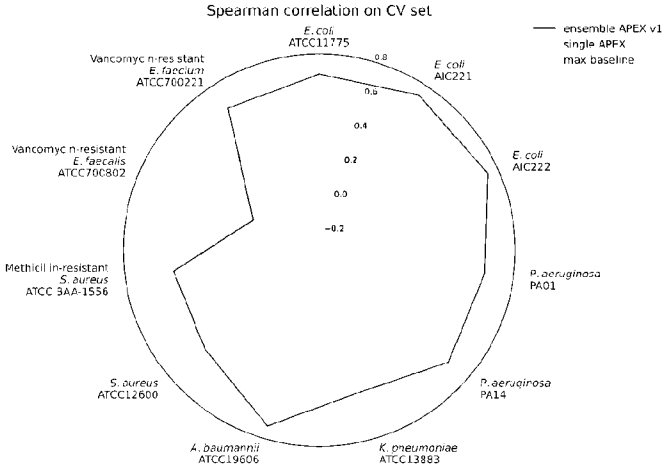


FIG. 13. Spearman correlation scores of various ML models on CV set. The inventors divided the inventive in-house peptide dataset into a CV set for hyperparameter tuning and ML model selection, and an independent dataset to evaluate the prediction performance. The figure shows the average species-specific prediction performance of 5-fold CV in terms of Spearman correlation for APEX models and maximum performance from baseline ML models on the CV set. The radius reflects the Spearman correlation value. Max baseline denotes the highest Spearman correlation values from baseline ML models, including elastic net, linear support vector regression, extra-trees regressor, random forest and gradient boosting decision tree.

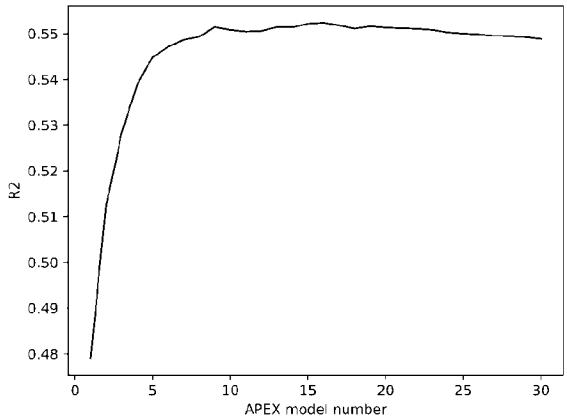


FIG. 14. Relationship between R-squared and the number of APEX models used in ensemble learning on the CV set. The inventors divided the inventive in-house peptide dataset into a CV set for hyperparameter tuning and ML model selection, and an independent dataset to evaluate prediction performance. On the CV set, the inventors trained several APEX models under different deep learning architectures and training strategies. The inventors ranked the APEX models by the averaged R-squared to predict species-specific antimicrobial activity. The inventors averaged the predictions from the top-ranked models to create the ensemble APEX to improve antibiotic prediction performance. The figure shows the relationship between the averaged R-squared for species-specific antimicrobial activity prediction and the number of APEX models being averaged.

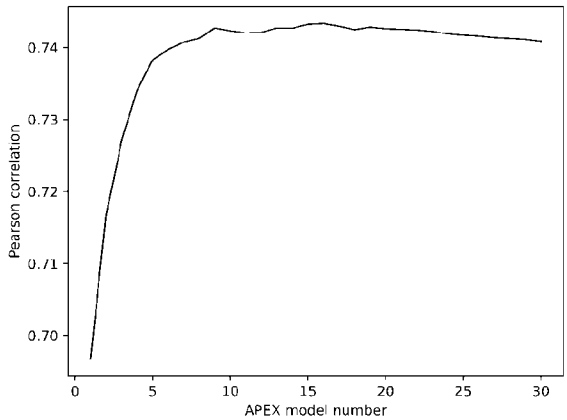


FIG. 15. Relationship between Pearson correlation and the number of APEX models used in ensemble learning on the CV set. The inventors divided the inventive in-house peptide dataset into a CV set for hyperparameter tuning and ML model selection, and an independent dataset to evaluate prediction performance. On CV set, the inventors trained the APEX models under different deep learning architectures and training strategies, then ranked them by the averaged R-squared for species-specific antimicrobial activity prediction. To improve antibiotic prediction performance, the inventors averaged the predictions from the top-ranked models creating ensemble APEX. The figure shows the relationship between the averaged Pearson correlation for species-specific antimicrobial activity prediction and the number of APEX models being averaged.

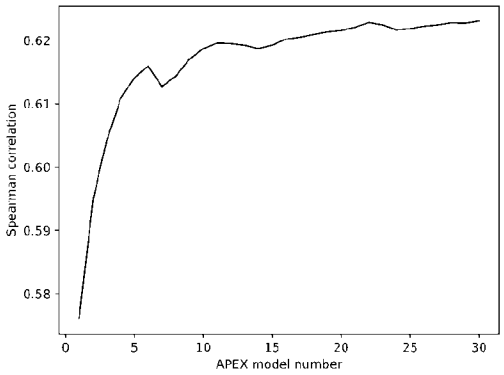


FIG. 16. Relationship between Spearman correlation and the number of APEX models used in ensemble learning on CV set. The inventors divided the inventive in-house peptide dataset into a CV set for hyperparameter tuning and ML model selection, and an independent dataset to evaluate prediction performance. On CV set, the inventors trained APEX models under different deep learning architectures and training strategies. The inventors ranked the APEX models by the averaged R-squared for species-specific antimicrobial activity prediction. To improve antibiotic prediction performance, the inventors averaged the predictions from the top-ranked models to create ensemble APEX to improve antibiotic prediction performance. The figure shows the relationship between averaged Spearman correlation for species-specific antimicrobial activity prediction and the number of APEX models being averaged.

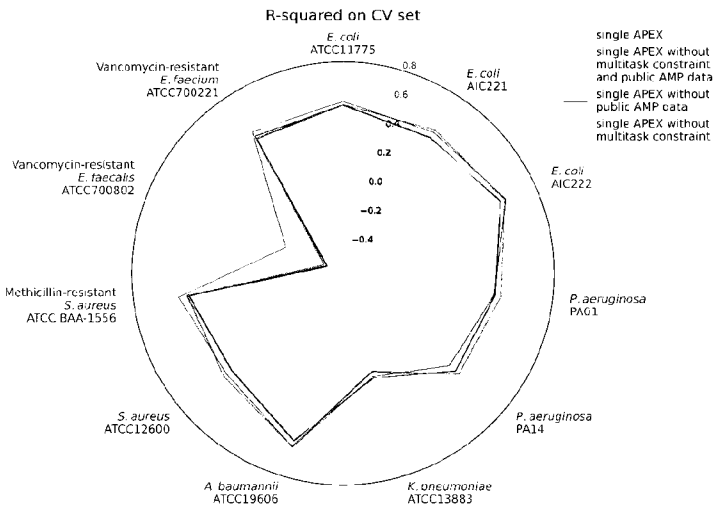


FIG. 17. Ablation study of the multitask learning strategy of APEX in terms of R-squared on the CV set. The inventors divided the inventive in-house peptide dataset into a CV set for hyperparameter tuning and ML model selection, and an independent dataset to evaluate prediction performance. The figure shows the average species-specific prediction performance of 5-fold cross validation in terms of R-squared for various APEX variants, including APEX without multitask constraint, APEX without using public AMP data during training, APEX without multitask constraint and public AMP data during training, and the original APEX (i.e., single APEX). The radius reflects the R-squared value.

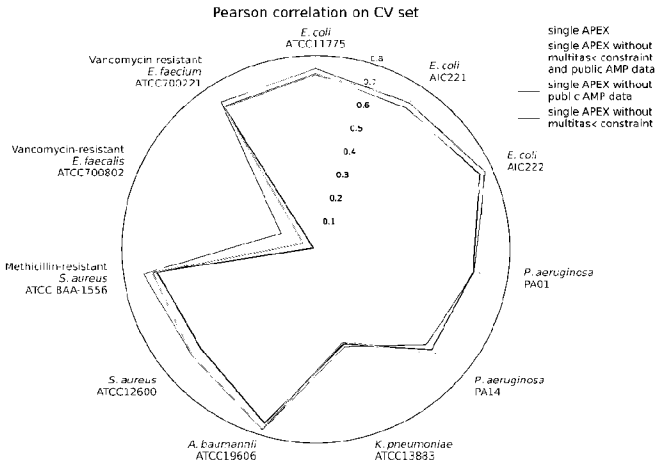


FIG. 18. Ablation study of the multitask learning strategy of APEX in terms of Pearson correlation on CV set. The inventors divided the inventive in-house peptide dataset into a CV set for hyperparameter tuning and ML model selection, and an independent dataset to evaluate prediction performance. The figure shows the average species-specific prediction performance of 5-fold cross validation in terms of Pearson correlation for various APEX variants, including APEX without multitask constraint, APEX without using public AMP data during training, APEX without multitask constraint and public AMP data during training, and the original APEX (i.e., single APEX). The radius reflects the Pearson correlation value.

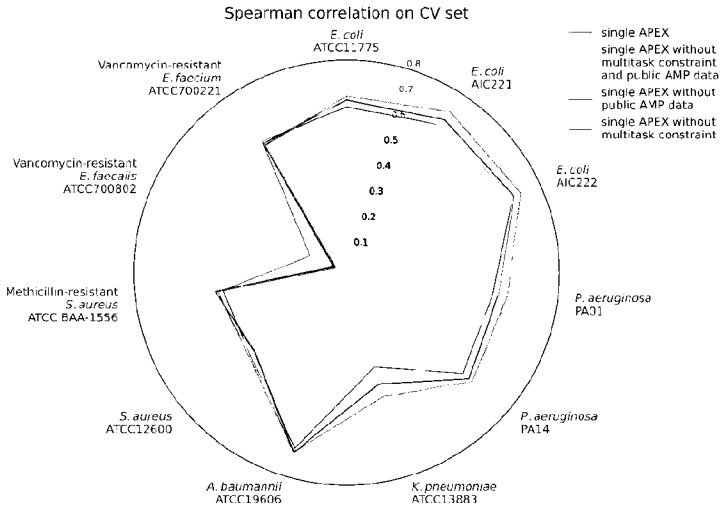


FIG. 19. Ablation study of the multitask learning strategy of APEX in terms of Spearman correlation on CV set. The inventors divided the inventive in-house peptide dataset into a CV set for hyperparameter tuning and ML model selection, and an independent dataset to evaluate prediction performance. The figure shows the average species-specific prediction performance of 5-fold cross validation in terms of Spearman correlation for various APEX variants, including APEX without multitask constraint, APEX without using public AMP data during training, APEX without multitask constraint and public AMP data during training, and the original APEX (i.e., single APEX). The radius reflects the Spearman correlation value.

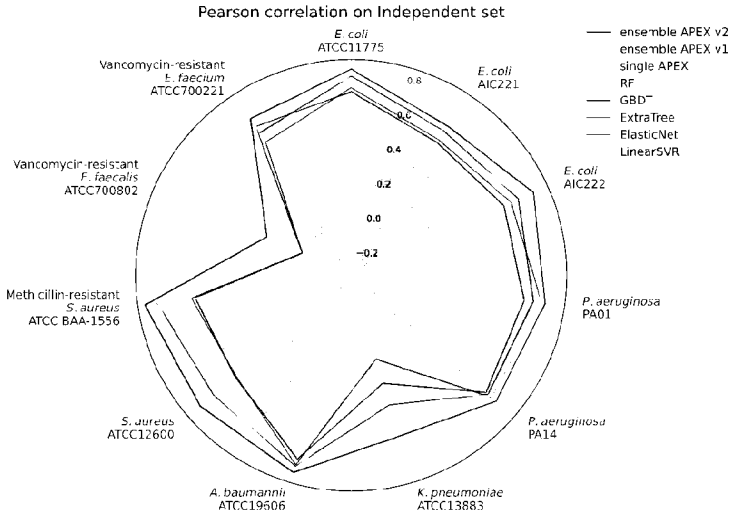


FIG. 20. Pearson correlation scores of various ML models on an independent set. The inventors divided the inventive in-house peptide dataset into a CV set for hyperparameter tuning and ML model selection, and an independent dataset to evaluate prediction performance. The figure shows the species-specific prediction performance in terms of Pearson correlation for various ML models that were trained on the CV set and evaluated on the independent set. The radius reflects the Pearson correlation value. RF: random forest; GBDT: gradient boosting decision tree; ExtraTree: extra-tree regressor; ElasticNet: elastic net; LinearSVR: linear support vector regression.

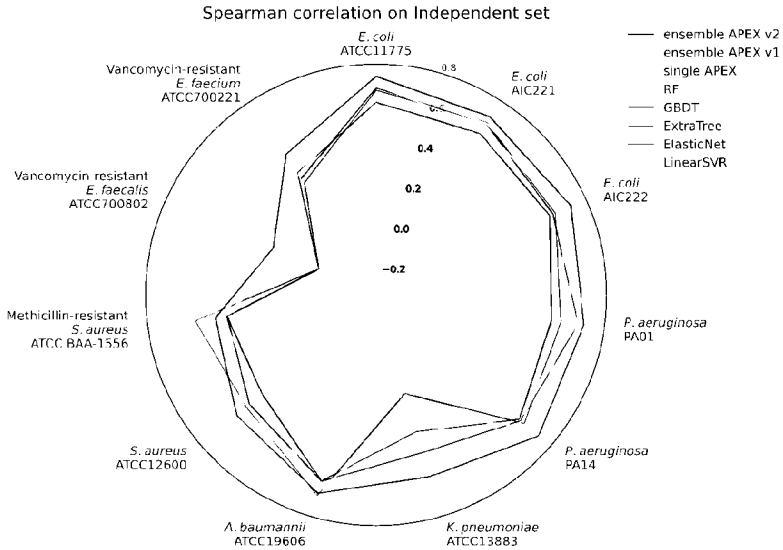


FIG. 21. Spearman correlation scores of various ML models on an independent set. The inventors divided the inventive in-house peptide dataset into a CV set for hyperparameter tuning and ML model selection, and an independent dataset to evaluate prediction performance. The figure shows the species-specific prediction performance in terms of Spearman correlation for various ML models that were trained on the CV set and evaluated on the independent set. The radius reflects the Spearman correlation value. RF: random forest; GBDT: gradient boosting decision tree; ExtraTree: extra-tree regressor; ElasticNet: elastic net; LinearSVR: linear support vector regression.

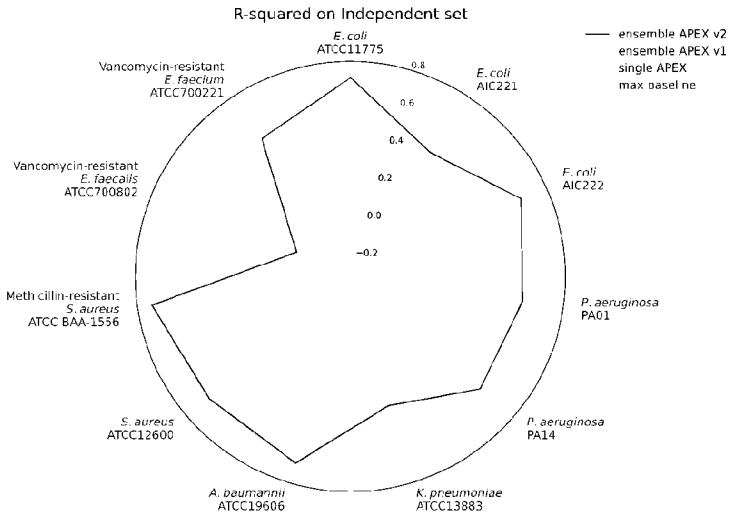


FIG. 22. R-squared scores of various ML models on an independent set. The inventors divided the inventive in-house peptide dataset into a CV set for hyperparameter tuning and ML model selection, and an independent dataset to evaluate prediction performance. The figure shows the species-specific prediction performance in terms of R-squared for APEX models and maximum performance from baseline ML models that were trained on the CV set and evaluated on the independent set. The radius reflects the R-squared value. Max baseline denotes the highest R-squared values from baseline ML models, including elastic net, linear support vector regression, extra-trees regressor, random forest and gradient boosting decision tree.

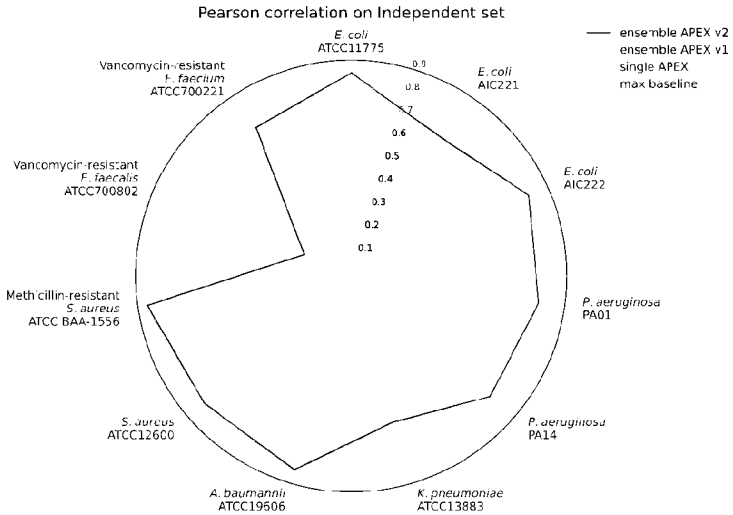


FIG. 23. Pearson correlation scores of various ML models on an independent set. The inventors divided the inventive in-house peptide dataset into a CV set for hyperparameter tuning and ML model selection, and an independent dataset to evaluate prediction performance. The figure shows the species-specific prediction performance in terms of Pearson correlation for APEX models and maximum performance from baseline ML models that were trained on the CV set and evaluated on the independent set. The radius reflects the Pearson correlation value. Max baseline denotes the highest Pearson correlation values from baseline ML models, including elastic net, linear support vector regression, extra-trees regressor, random forest and gradient boosting decision tree.

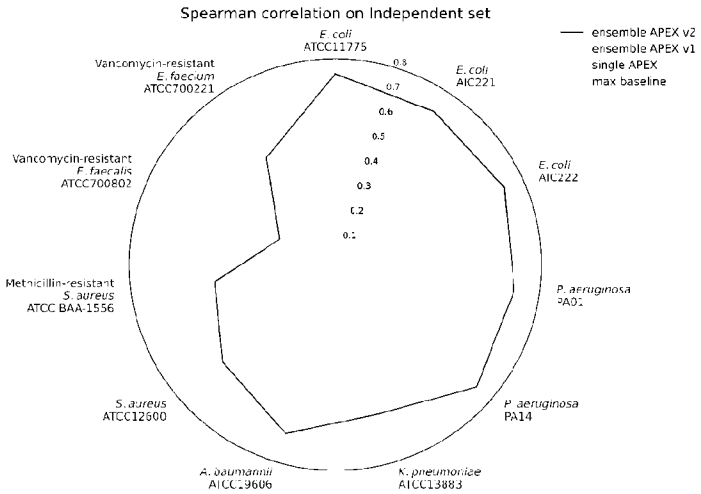


FIG. 24. Spearman correlation scores of various ML models on an independent set. The inventors divided the inventive in-house peptide dataset into a CV set for hyperparameter tuning and ML model selection, and an independent dataset to evaluate prediction performance. The figure shows the species-specific prediction performance in terms of Spearman correlation for APEX models and maximum performance from baseline ML models that were trained on the CV set and evaluated on the independent set. The radius reflects the Spearman correlation value. Max baseline denotes the highest Spearman correlation values from baseline ML models, including elastic net, linear support vector regression, extra-trees regressor, random forest and gradient boosting decision tree.

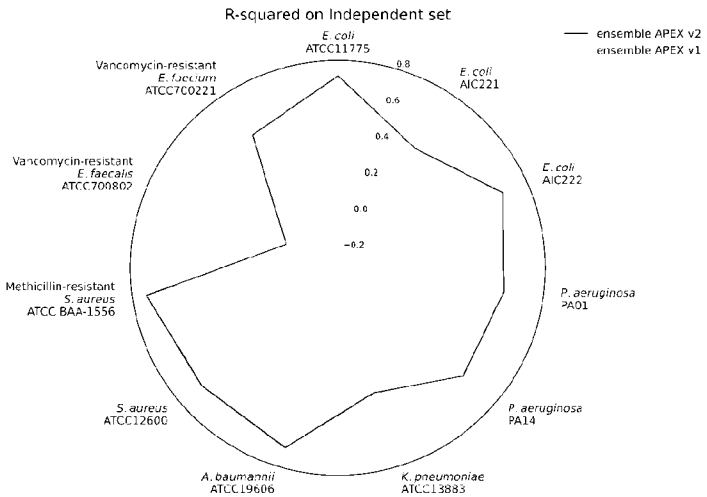


FIG. 25. R-squared scores of ensemble APEX v2 and v1 on an independent set. The inventors divided the inventive in-house peptide dataset into a CV set for hyperparameter tuning and ML model selection, and an independent dataset to evaluate prediction performance. The figure shows the species-specific prediction performance in terms of R-squared for two APEX variants that were trained on the CV set and evaluated on the independent set. The ensemble APEX v1 averaged the predictions from 8 different neural network architectures and training strategies. On top of ensemble APEX v1, ensemble APEX v2 further trained 5 copies under different random seeds for each base learner from ensemble APEX v1 to create $8 \times 5 = 40$ deep neural network predictors. The predictions from the 40 models were averaged to create the final prediction for ensemble APEX v2. The radius reflects the R-squared value.

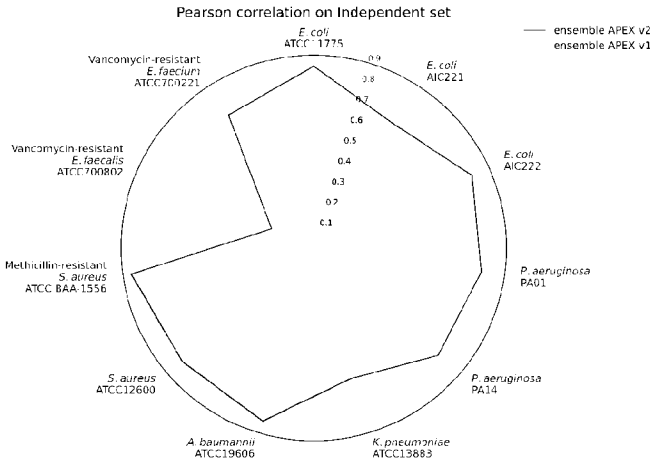


FIG. 26. Pearson correlation scores of ensemble APEX v2 and v1 on an independent set. The inventors divided the inventive in-house peptide dataset into a CV set for hyperparameter tuning and ML model selection, and an independent dataset to evaluate prediction performance. The figure shows the species-specific prediction performance in terms of Pearson correlation for two APEX variants that were trained on the CV set and evaluated on the independent set. The ensemble APEX v1 averaged the predictions from 8 different neural network architectures and training strategies. On top of ensemble APEX v1, ensemble APEX v2 further trained 5 copies under different random seeds for each base learner from ensemble APEX v1 to create 8×5 = 40 deep neural network predictors. The predictions from the 40 models were averaged to create the final prediction for ensemble APEX v2. The radius reflects the Pearson correlation value.

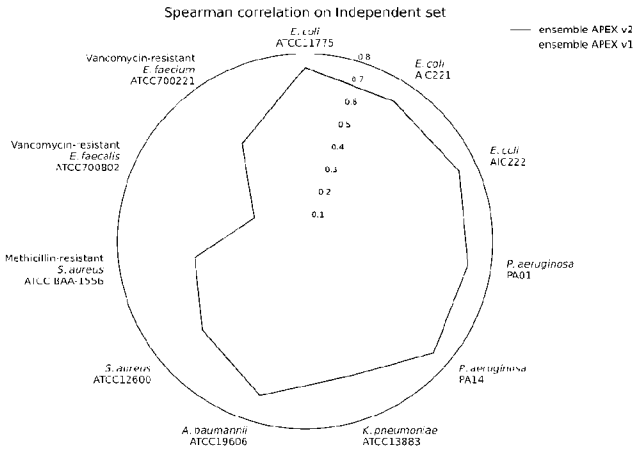


FIG. 27. Spearman correlation scores of ensemble APEX v2 and v1 on an independent set. The inventors divided the inventive in-house peptide dataset into a CV set for hyperparameter tuning and ML model selection, and an independent dataset to evaluate prediction performance. The figure shows the species-specific prediction performance in terms of Spearman correlation for two APEX variants that were trained on the CV set and evaluated on the independent set. The ensemble APEX v1 averaged the predictions from 8 different neural network architectures and training strategies. On top of ensemble APEX v1, ensemble APEX v2 further trained 5 copies under different random seeds for each base learner from ensemble APEX v1 to create 8×5 = 40 deep neural network predictors. The predictions from the 40 models were averaged to create the final prediction for ensemble APEX v2. The radius reflects the Spearman correlation value.

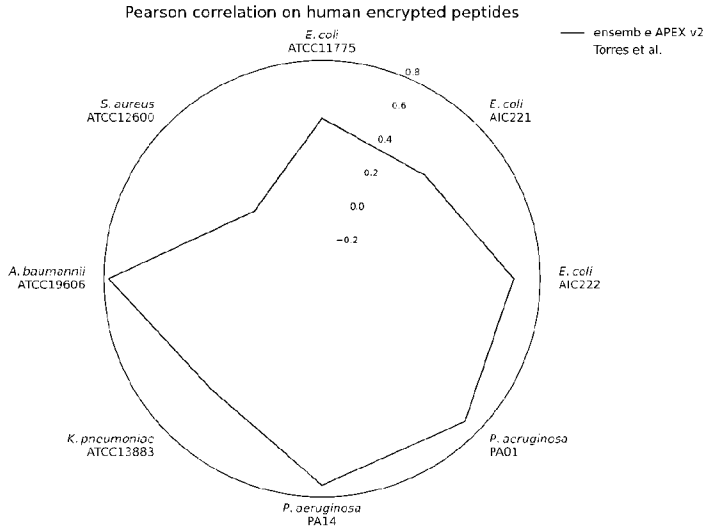


FIG. 28. Pearson correlation scores of ensemble APEX v2 and the scoring function used to identify modern human encrypted peptides. The human encrypted peptides from Torres et al.[1] constitute a subset of the inventive in-house peptide dataset. The inventors treated this subset as the test data and excluded it from ML model training. The figure shows the species-specific prediction performance in terms of Pearson correlation for ensemble APEX v2 and the AMP scoring function used by Torres et al.[1]. The radius reflects the Pearson correlation value.

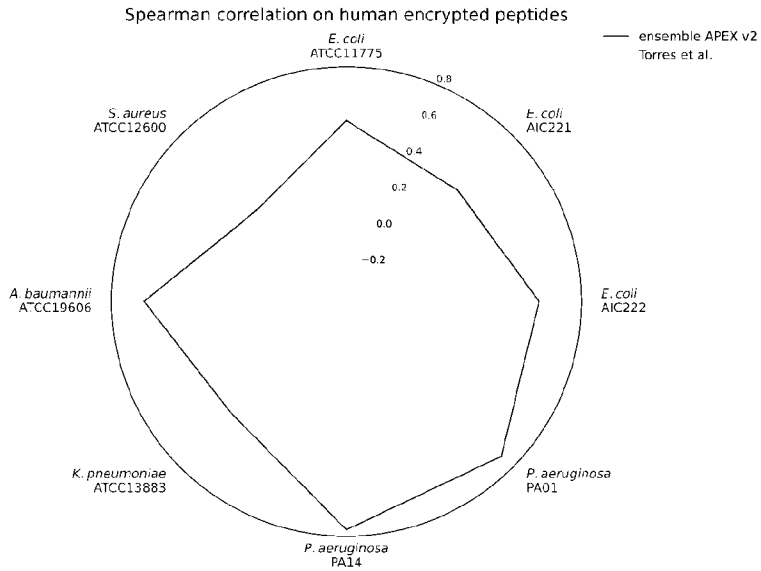


FIG. 29. Spearman correlation scores of ensemble APEX v2 and the scoring function used to identify modern human encrypted peptides. The human encrypted peptides from Torres et al.[1] constitute a subset of the inventive in-house peptide dataset. The inventors treated this subset as the test data and excluded it from ML model training. The figure shows the species-specific prediction performance in terms of Spearman correlation for ensemble APEX v2 and the AMP scoring function used in Torres et al.[1]. The radius reflects the Spearman correlation value.

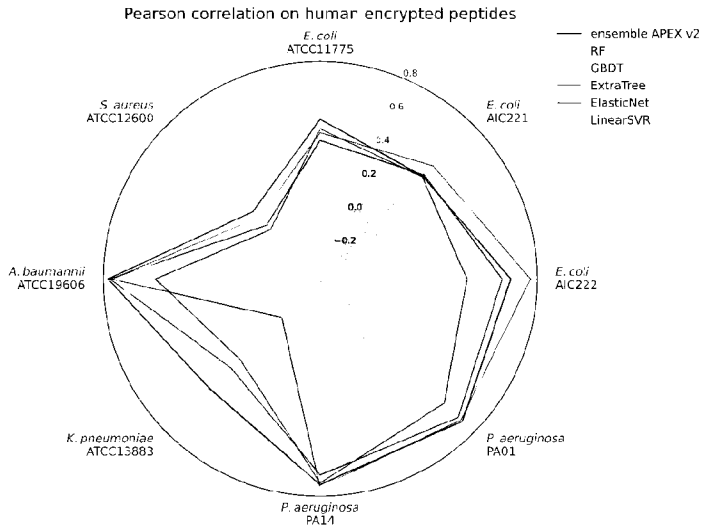


FIG. 30. Pearson correlation scores of various ML models used to identify modern human encrypted peptides. The human encrypted peptides from Torres et al.[1] constitute a subset of the inventive in-house peptide dataset. The inventors treated this subset as the test data and excluded it from ML model training. The figure shows the species-specific prediction performance in terms of Pearson correlation for various ML models. The radius reflects the Pearson correlation value. RF: random forest; GBDT: gradient boosting decision tree; ExtraTree: extra-tree regressor; ElasticNet: elastic net; LinearSVR: linear support vector regression.

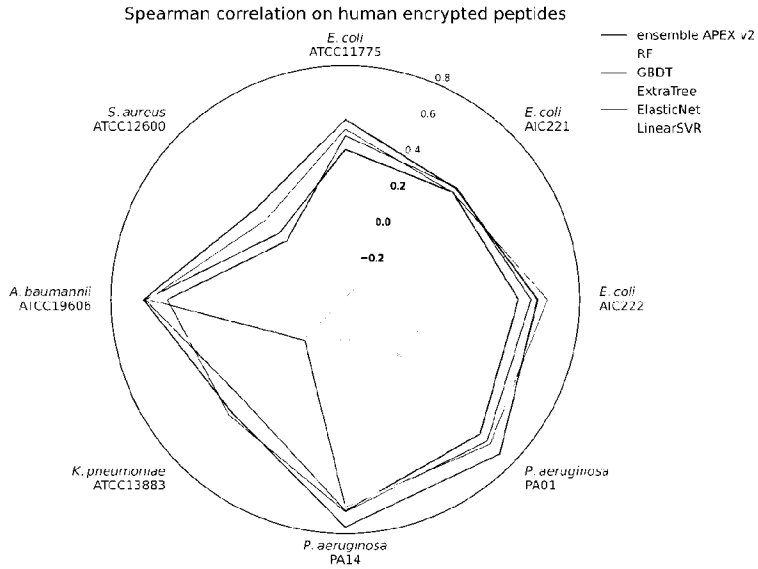


FIG. 31. Spearman correlation scores of various ML models used to identify modern human encrypted peptides. The human encrypted peptides from Torres et al.[1] constitute a subset of the inventive in-house peptide dataset. The inventors treated this subset as the test data and excluded it from ML model training. The figure shows the species-specific prediction performance in terms of Spearman correlation for various ML models. The radius reflects the Spearman correlation value. RF: random forest; GBDT: gradient boosting decision tree; ExtraTree: extra-tree regressor; ElasticNet: elastic net; LinearSVR: linear support vector regression.

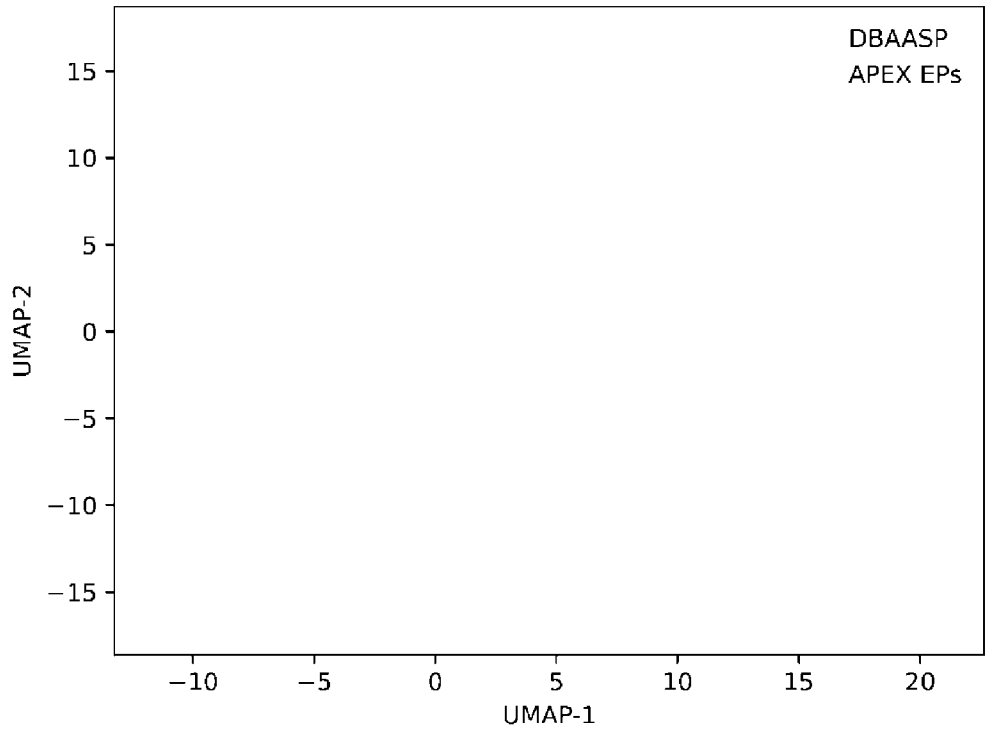


FIG. 32. Sequence space exploration using a similarity matrix. The graph represents a bidimensional sequence space visualization of peptide sequences found in DBAASP and antimicrobial EPs discovered by APEX. The inventors used sequence alignment to generate a sequence similarity matrix for all peptide sequences in DBAASP and the 37,176 antimicrobial EPs predicted by APEX. Each row in the similarity matrix corresponds to a feature representation of a peptide in terms of amino acid residues. Uniform Manifold Approximation and Projection (UMAP) was used to reduce the feature representation to two dimensions for visualization purposes.

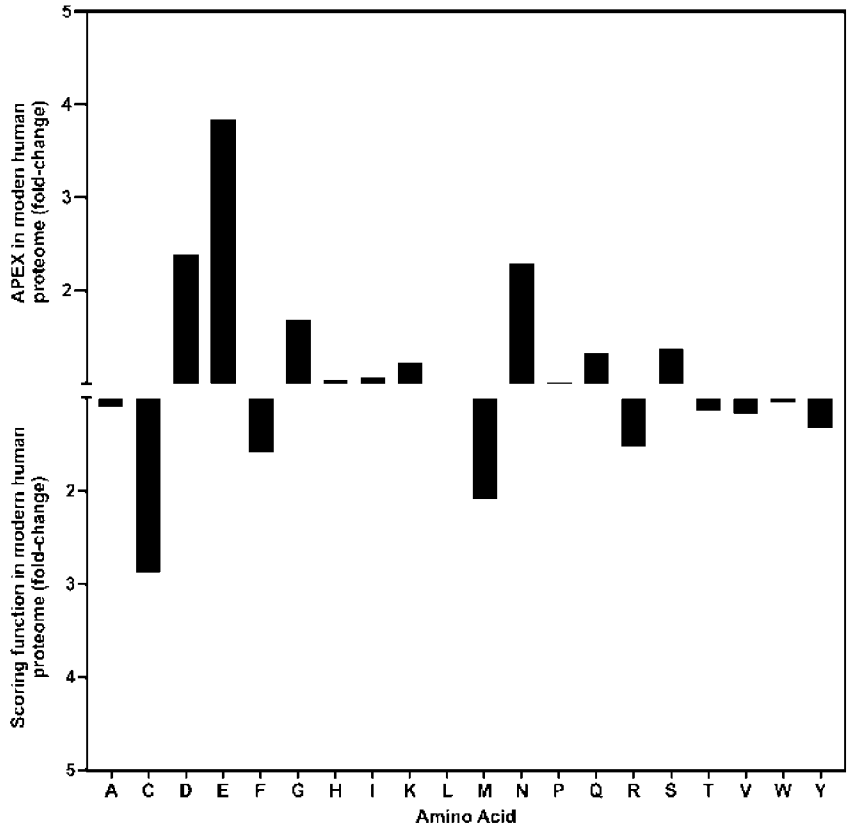


FIG. 33. Relative abundance of the amino acid content of encrypted peptides (EPs) from the modern human proteome identified by APEX (top) and the scoring function (bottom). The frequency of amino acid was normalized by the total number of amino acid residue counts. The ratio between normalized amino acid frequencies highlights the overrepresentation of negatively charged residues (D and E), glycine (G), and uncharged polar residues (N, Q, and S) in peptides identified by APEX. The scoring function preferably identified encrypted peptides with a high frequency of cysteine (C). methionine (M), arginine (R) and phenylalanine (F) residues.

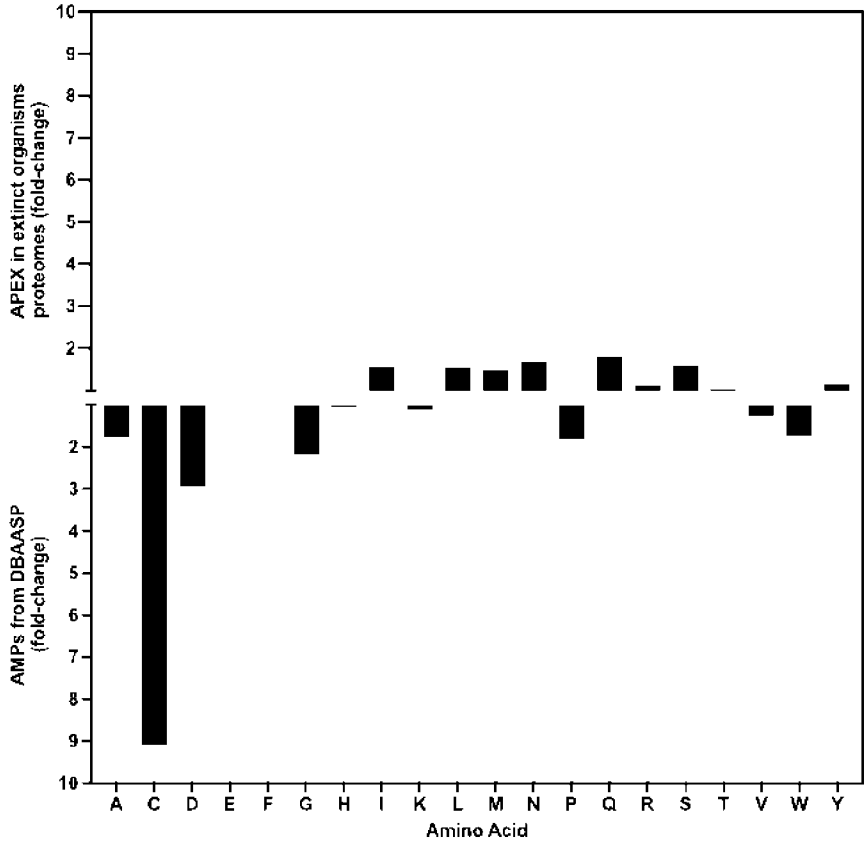


FIG. 34. Relative abundance of the amino acid content of encrypted peptides (EPs) identified by APEX from the proteomes of extinct organisms (top) compared to known AMPs from DBAASP (bottom). The frequency of each amino acid residue was normalized by the total number of amino acid residue counts. The ratio between AEPs identified by APEX and known peptide from DBAASP highlights the relative abundance of aliphatic (L and I) and uncharged polar (M, N, Q, and S) residues in sequences from extinct organisms, whereas peptides from DBAASP present a higher content of negatively charged (D) and aromatic (W) residues, as well as residues known for their role in secondary structure (C, G, P).

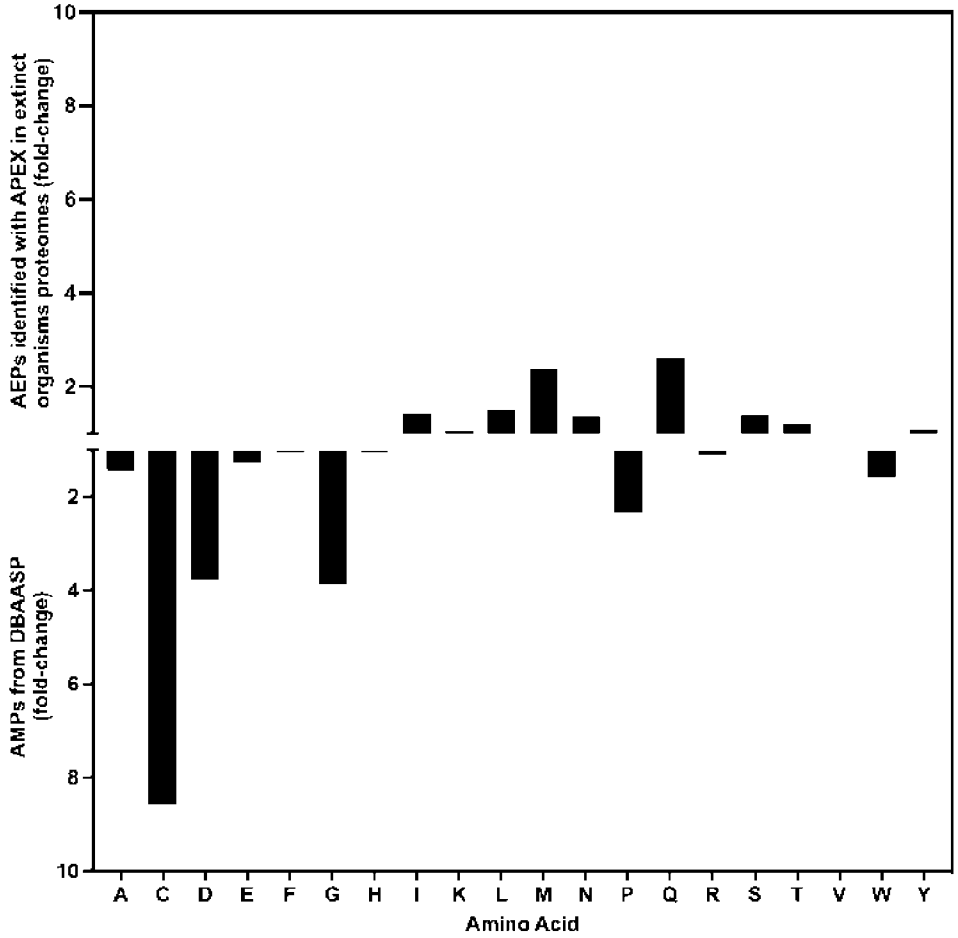


FIG. 35. Relative abundance of the amino acid content of archaic encrypted peptides (AEPs) identified by APEX from the proteomes of extinct organisms (top) compared to known AMPs from DBAASP (bottom). The frequency of each amino acid was normalized by the total number of amino acid residue counts. The ratio between encrypted peptides from extinct proteins identified by APEX and known peptides from the DBAASP highlights the relative abundance of aliphatic (L and I) and uncharged polar (M, N, Q, and S) residues in sequences from extinct proteins, whereas peptides from the DBAASP had a higher content of negatively charged (D and E) and aromatic (W) residues, as well as residues known for their role in secondary structure (A, C, G, P).

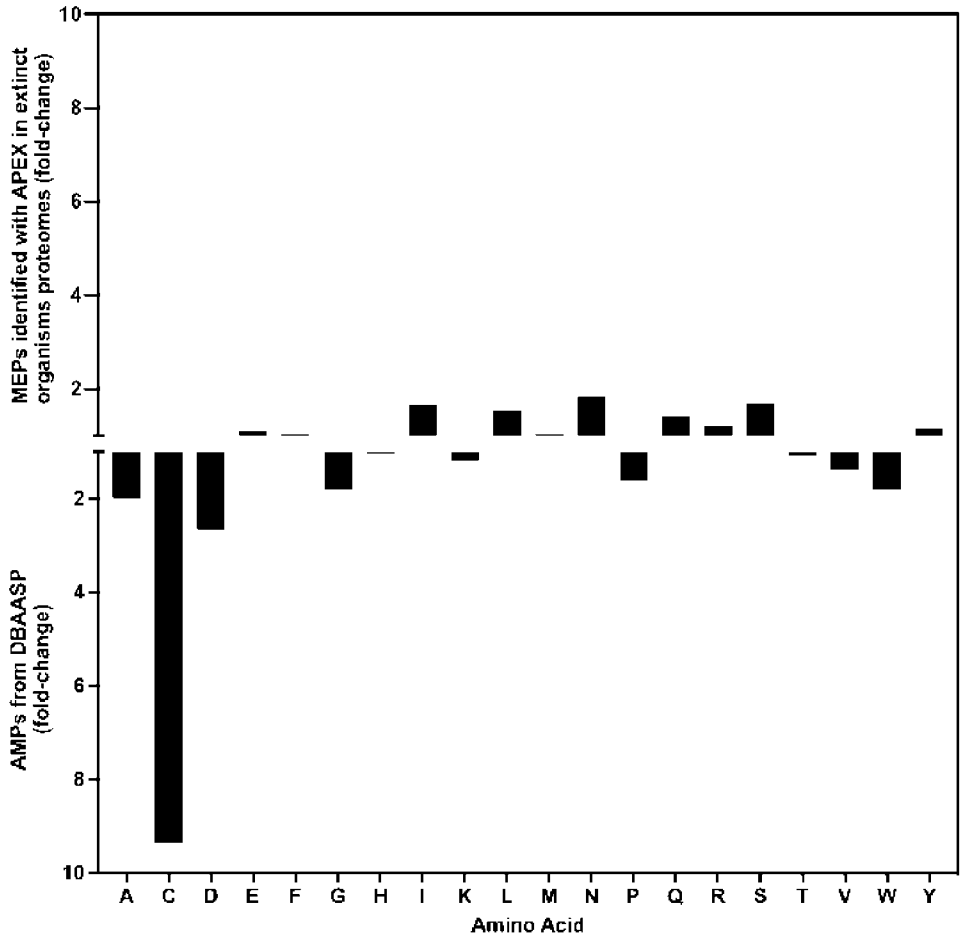


FIG. 36. Relative abundance of the amino acid content of MEPs identified by APEX from the proteomes of extinct organisms compared to known AMPs from DBAASP. The frequency of amino acid was normalized by the total number of amino acid residue counts. The ratio between encrypted peptides from extinct organisms that still exist in modern organisms proteomes identified by APEX and known peptide from DBAASP highlights the relative abundance of aliphatic (L and I) and polar non-charged (M, N, Q, and S) residues in sequences from extinct organisms proteins, whereas peptides from DBAASP present a higher content of negatively charged (D), aromatic (W), and residues known for their role on secondary structure (A, C, G, P).

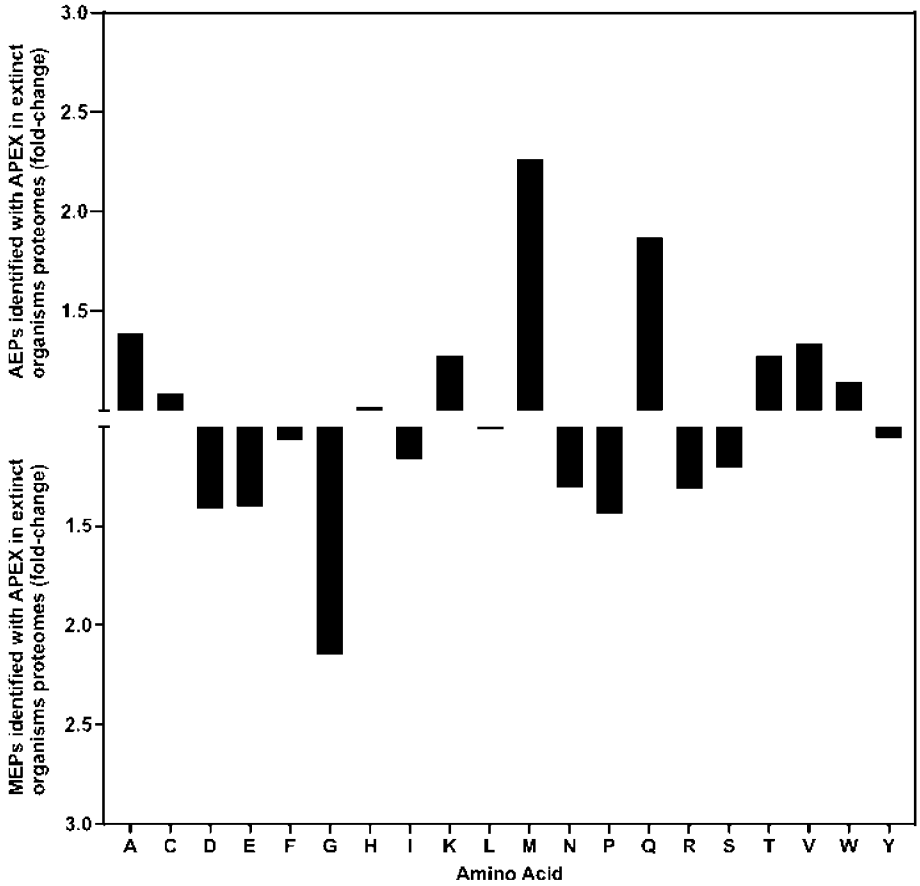


FIG. 37. Relative abundance of the amino acid content of AEPs and MEPs identified by APEX from the proteomes of extinct organisms. The frequency of amino acid was normalized by the total number of amino acid residue counts. The ratio between encrypted peptides from extinct proteins that are not present in modern organisms and from proteins that still exist in modern organisms identified by APEX highlights the relative abundance of polar non-charged (M and Q) residues in sequences from extinct proteins, whereas encrypted peptides from still existing proteins present a higher content of glycine.

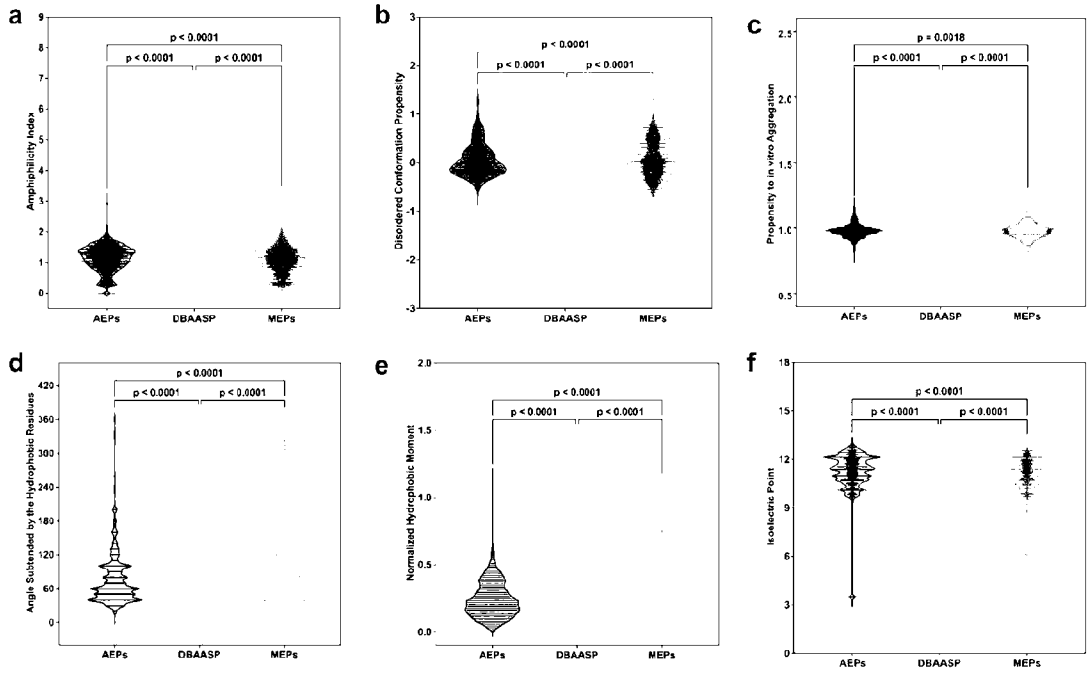
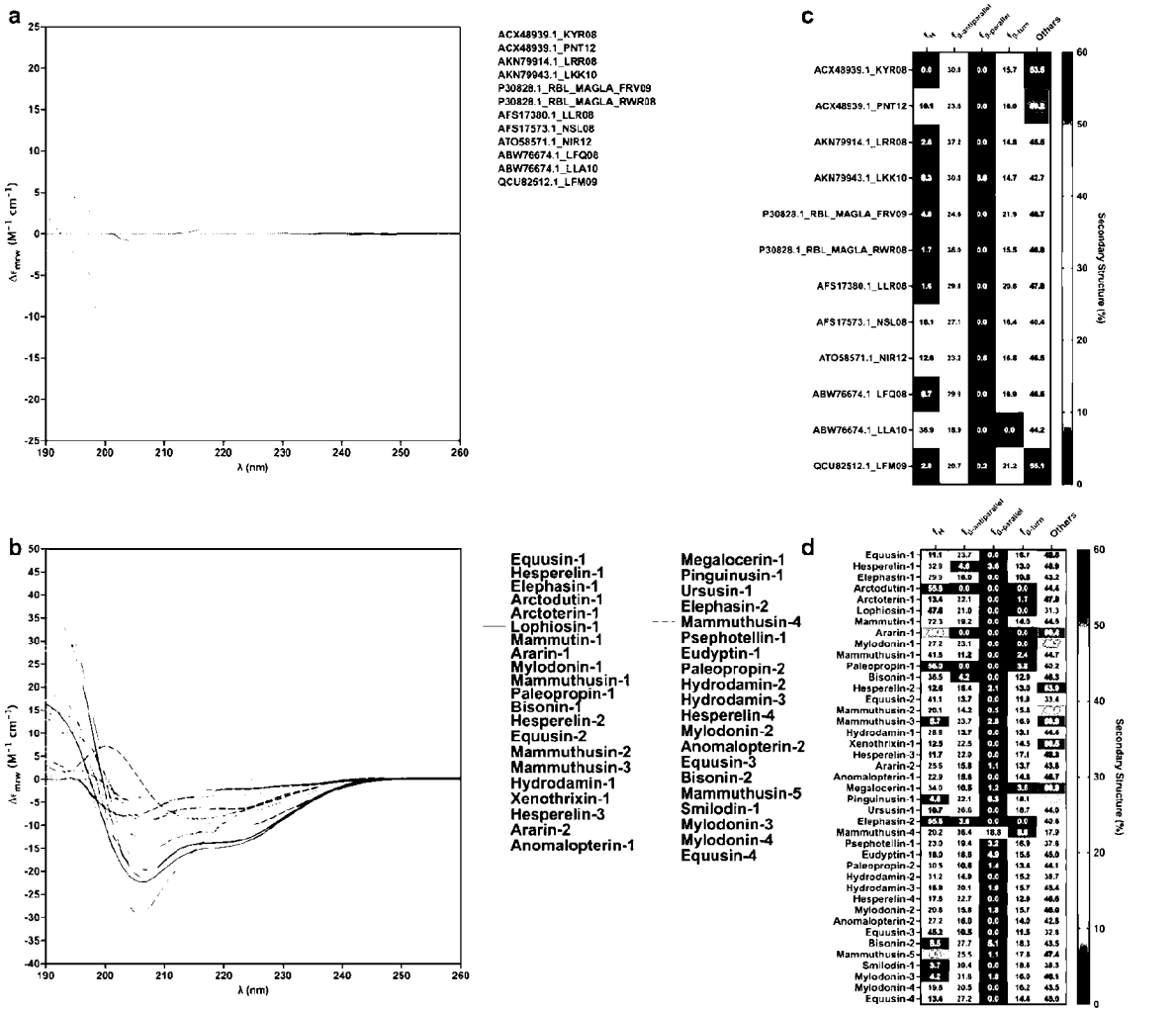


FIG. 38A-F. Physicochemical features of AEPs and MEPs identified by APEX in extinct organisms compared to AMPs from DBAASP. (a) Amphiphilicity index and **(b)** disordered conformation propensity; both properties closely correlated with mechanism of action, *i.e.*, how the peptides interact with membrane lipids to exert antimicrobial activity. **(c)** Propensity to aggregate *in vitro* and **(d)** angle subtended by the hydrophobic residues; both properties are correlated with supramolecular arrangement of the molecules and toxicity. **(e)** Hydrophobic moment normalized by peptide length and **(f)** isoelectric point; both properties are also related to the amphipathicity of the molecules that influence directly on their interactions with bacterial membranes. Statistical significance was obtained using two-tailed t-tests followed by Mann-Whitney test; p values are shown in the graph. The solid line inside each box represents the mean value obtained for each group.



FIGS. 39A-D. Secondary structure of active EPs predicted by the scoring function and APEX in helical inducer medium. Circular dichroism experiments with encrypted peptides (EPs) from extinct organisms generated by **(a)** the scoring function based on length, hydrophobicity, and net charge, which predominantly identified unstructured peptides, and **(b)** the deep learning model, which identified EPs having a higher helical content than the EPs predicted by the scoring function. **(c-d)** Heat maps of the secondary structure basis components values obtained using the algorithm BeStSel[2] for EPs identified by the scoring function **(c)** or APEX **(d)**. Assays were performed in a J-1500 (Jasco circular dichroism spectrophotometer), and the circular dichroism spectra were recorded after three accumulations at 25 °C, using a 1 mm path length quartz cell, between 260 and 190 nm at 50 nm min^{-1}, with a bandwidth of 0.5 nm.

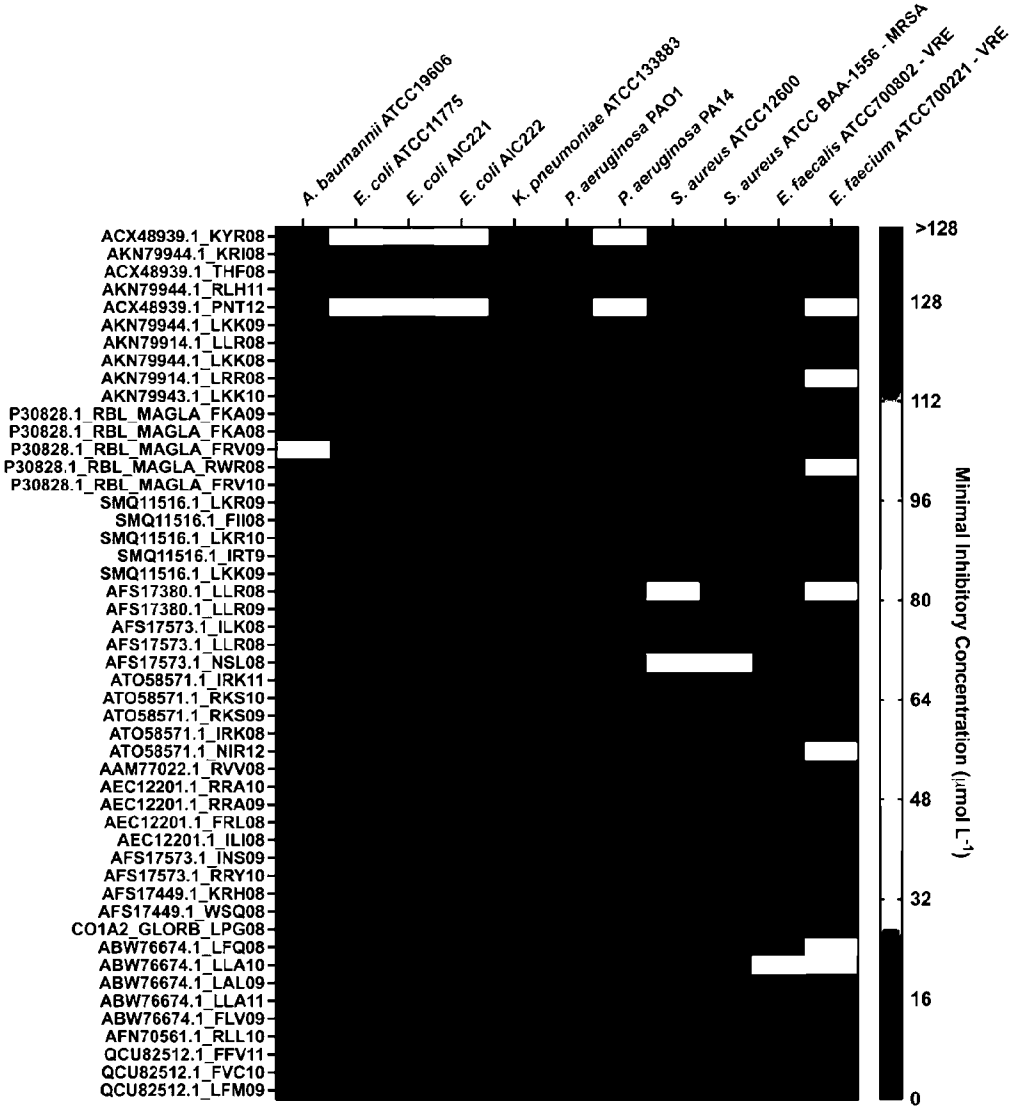


FIG. 40. Antimicrobial activity of encrypted peptides from extinct organisms predicted by the scoring function. Heat map of the antimicrobial activities (μmol L^{-1}) of the encrypted peptides (EPs) from extinct organisms predicted by the scoring function (Torres et al.[1]) based on length, hydrophobicity, and net charge, for 11 pathogens, including four strains resistant to conventional antibiotics. All the EPs identified by the scoring function were modern EPs (MEPs), *i.e.*, EPs identified in extinct organisms but also present in modern proteins. Briefly, 10^6 bacterial cells and serially diluted MEPs (0–128 μmol L^{-1}) were incubated at 37 °C. One day post-treatment, the optical density at 600 nm was measured in a microplate reader to determine whether the EPs from extinct organisms inhibited bacterial growth *in vitro*. Assays were performed in three independent replicates, and MIC values in the heat map are the arithmetic mean of the replicates in each condition.

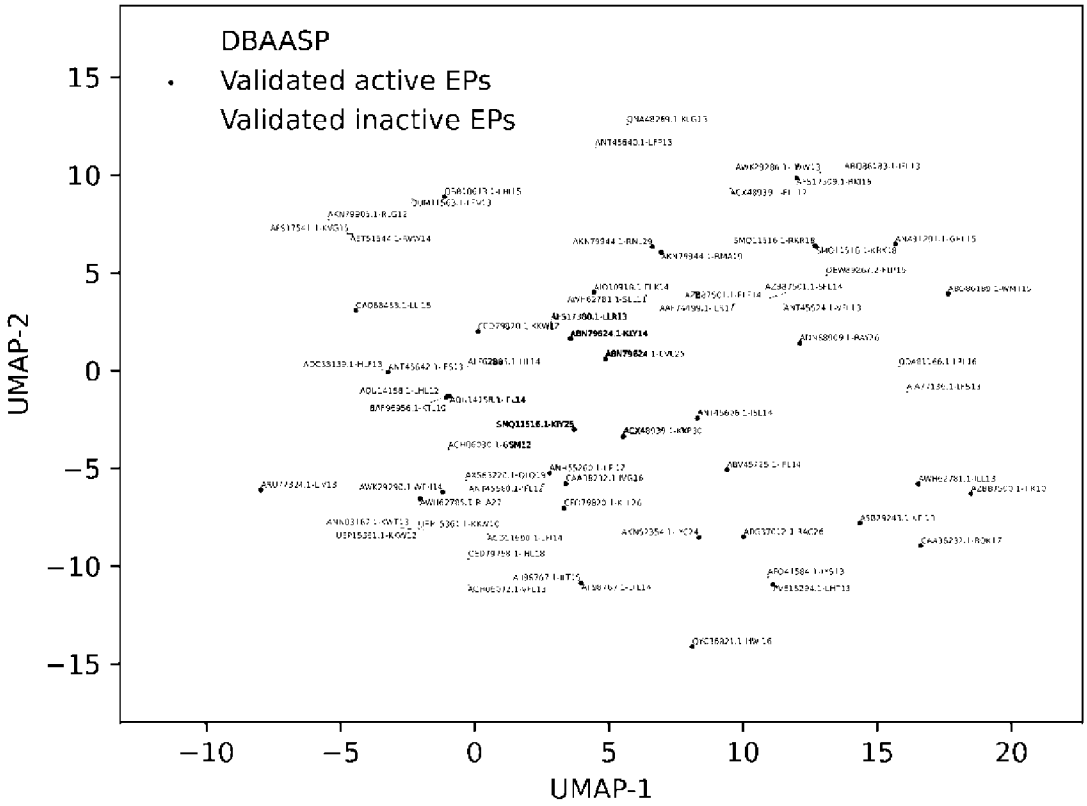


FIG. 41. Sequence space exploration using a similarity matrix containing the 69 encrypted peptides discovered by APEX selected for further experimental validation compared to peptide sequences from DBAASP. The inventors used the same methodology described in **Supplementary Fig. 26** to visualize DBAASP peptide sequences and the 69 EPs discovered by APEX and subsequently selected for synthesis and experimental validation. The 69 EP sequences are highlighted in the scatter plot. The inventors used different colors to show the EPs that were experimentally validated to be active (MIC ≤64 μmol L^{-1}) and inactive antimicrobials against the selected pathogens *in vitro*.

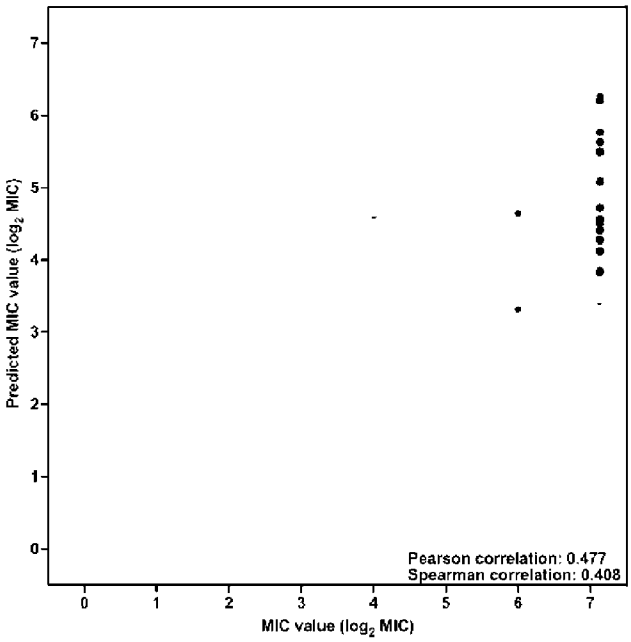


FIG. 42. Predicted vs. experimental MICs for *A. baumannii* ATCC 19606 of the encrypted peptides identified by APEX. Each peptide on the scatter plot is represented by a red circle. Pearson and Spearman correlations values are shown as an inset.

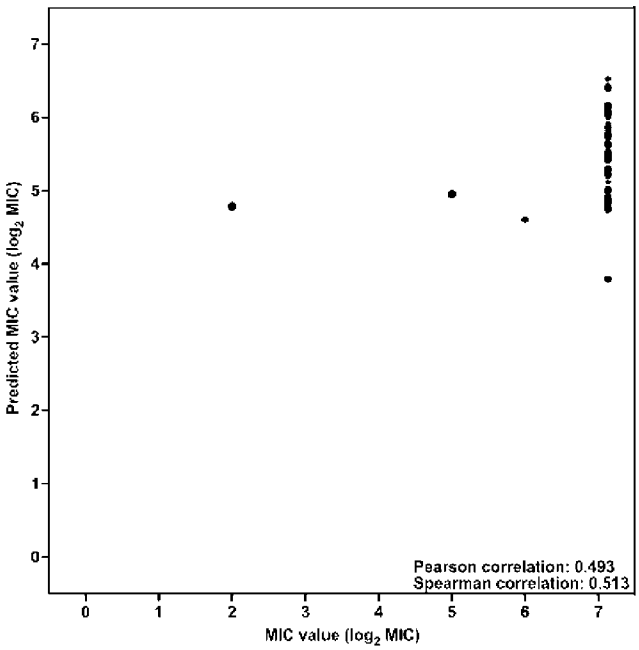


FIG. 43. Predicted vs. experimental MICs for *E. coli* AIC221 of the encrypted peptides identified by APEX. Each peptide on the scatter plot is represented by a red circle. Pearson and Spearman correlations values are shown as an inset.

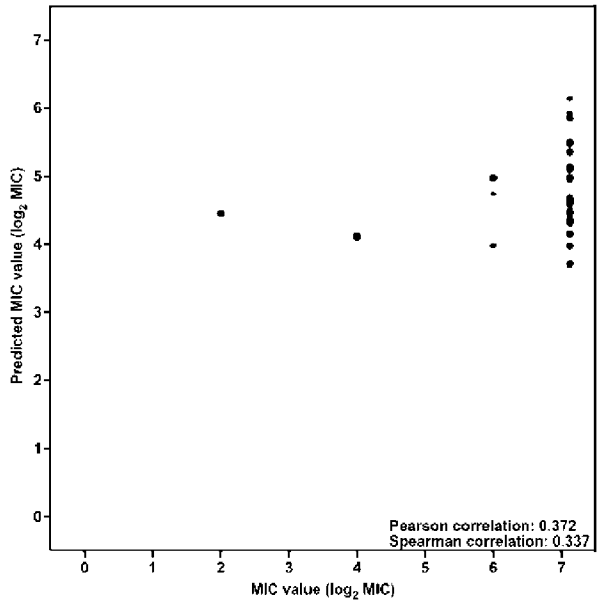


FIG. 44. Predicted *vs* experimental MICs for *E. coli* AIC222 of the encrypted peptides identified by APEX. Each peptide on the scatter plot is represented by a red circle. Pearson and Spearman correlations values are shown as an inset.

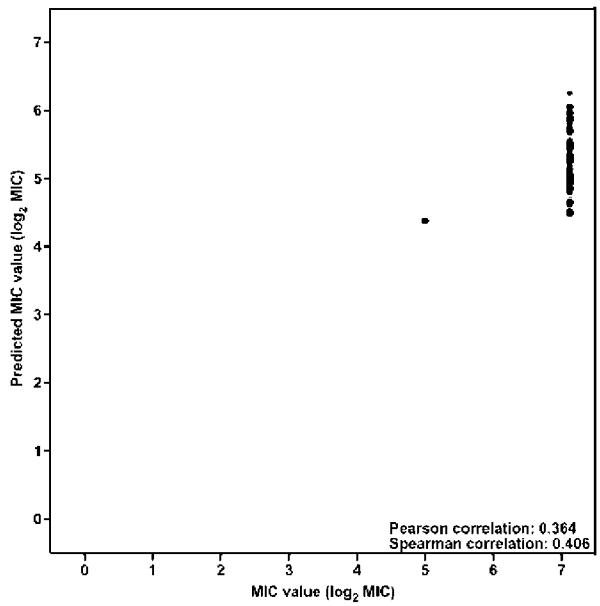


FIG. 45. Predicted vs. experimental MICs for *E. coli* ATCC 11775 of the encrypted peptides identified by APEX. Each peptide on the scatter plot is represented by a red circle. Pearson and Spearman correlations values are shown as an inset.

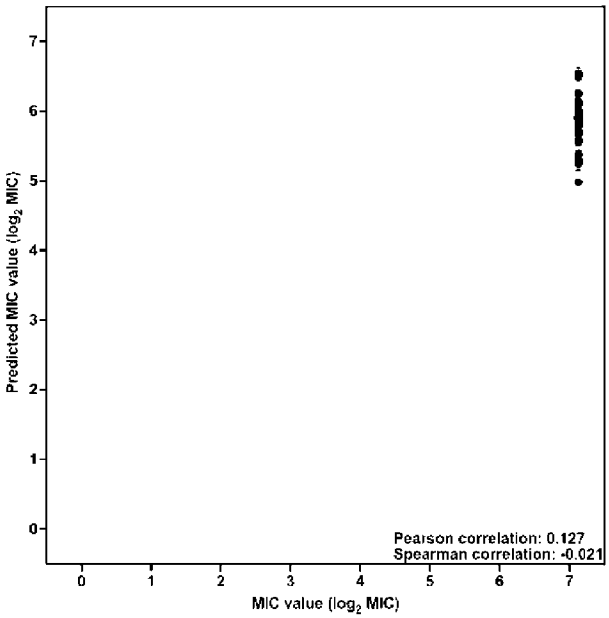


FIG. 46. Predicted vs. experimental MICs for *K. pneumoniae* ATCC 13883 of the encrypted peptides identified by APEX. Each peptide on the scatter plot is represented by a red circle. Pearson and spearman correlations values are shown as an inset.

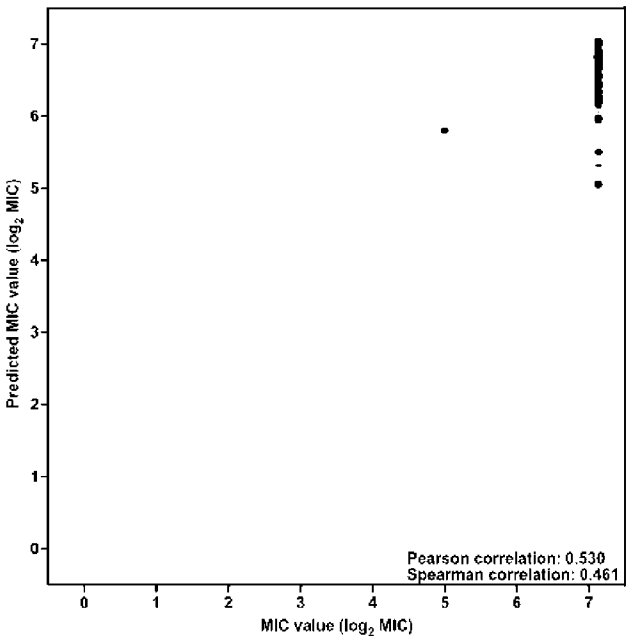


FIG. 47. Predicted vs. experimental MICs for *P. aeruginosa* PA14 of the encrypted peptides identified by APEX. Each peptide on the scatter plot is represented by a red circle. Pearson and spearman correlations values are shown as an inset.

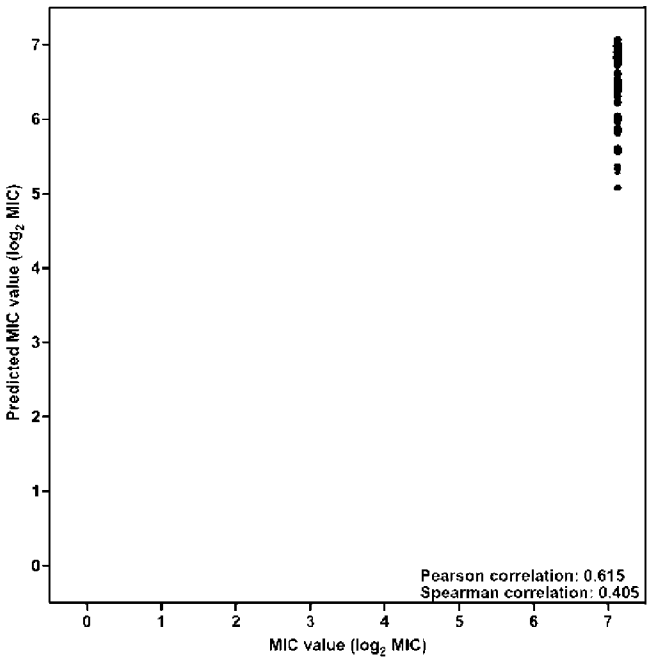


FIG. 48. Predicted vs. experimental MICs for *P. aeruginosa* PAO1 of the encrypted peptides identified by APEX. Each peptide on the scatter plot is represented by a red circle. Pearson and spearman correlations values are shown as an inset.

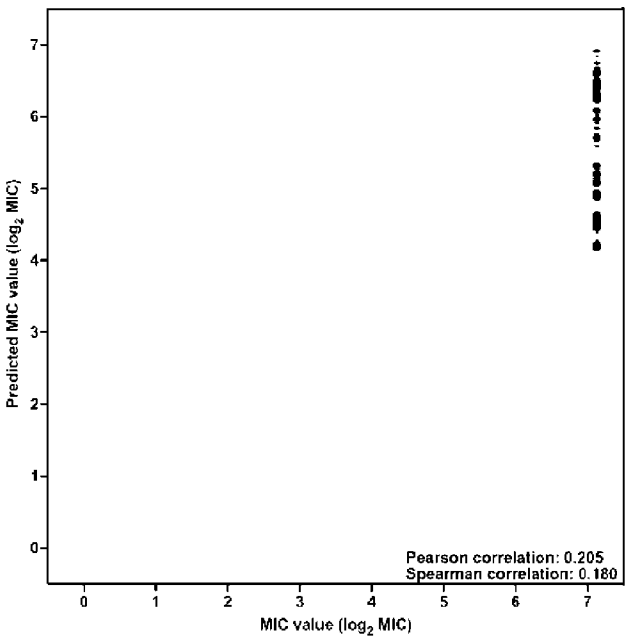


FIG. 49. Predicted vs. experimental MICs for methicillin-resistant *S. aureus* ATCC BAA-1556 of the encrypted peptides identified by APEX. Each peptide on the scatter plot is represented by a red circle. Pearson and spearman correlations values are shown as an inset.

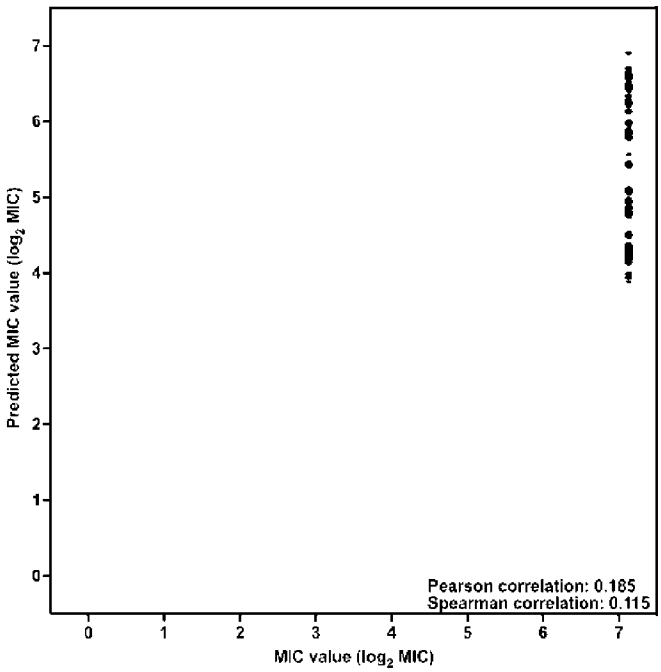


FIG. 50. Predicted vs. experimental MICs for *S. aureus* ATCC 12600 of the encrypted peptides identified by APEX. Each peptide on the scatter plot is represented by a red circle. Pearson and spearman correlations values are shown as an inset.

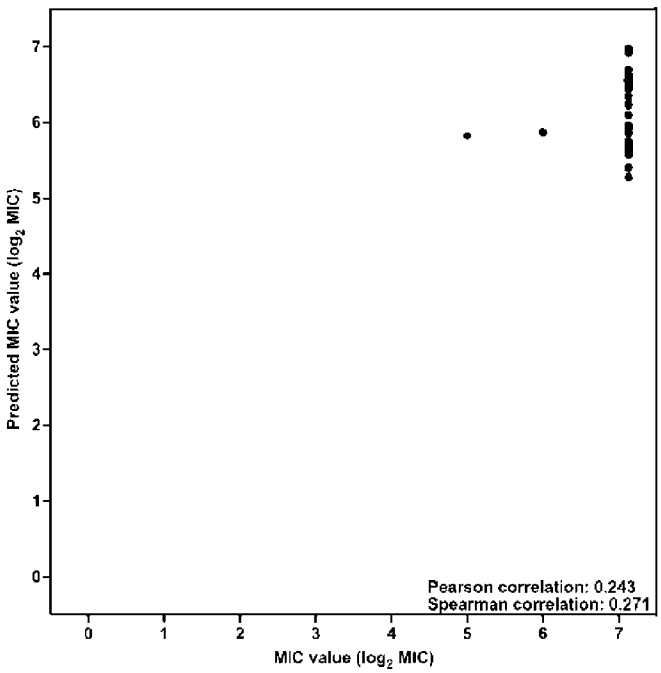


FIG. 51. Predicted vs. experimental MICs for vancomycin-resistant *E. faecalis* ATCC 700802 of the encrypted peptides identified by APEX. Each peptide on the scatter plot is represented by a red circle. Pearson and spearman correlations values are shown as an inset.

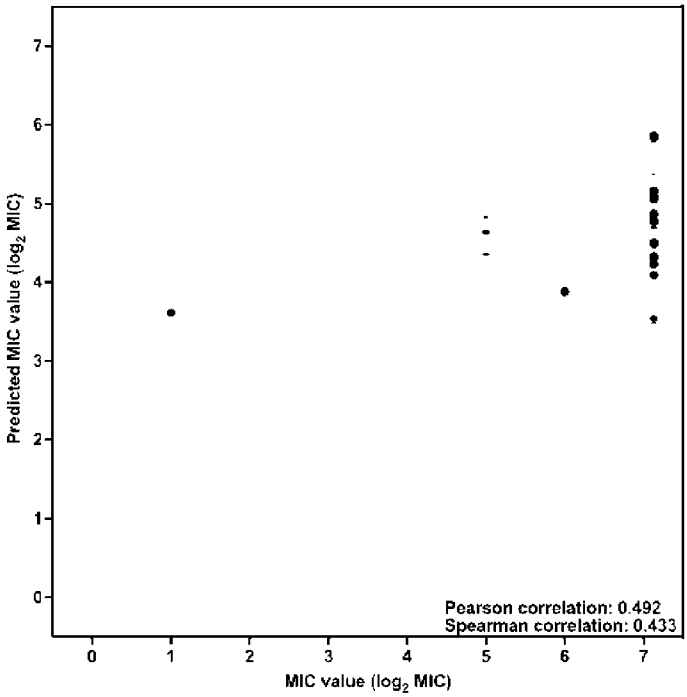
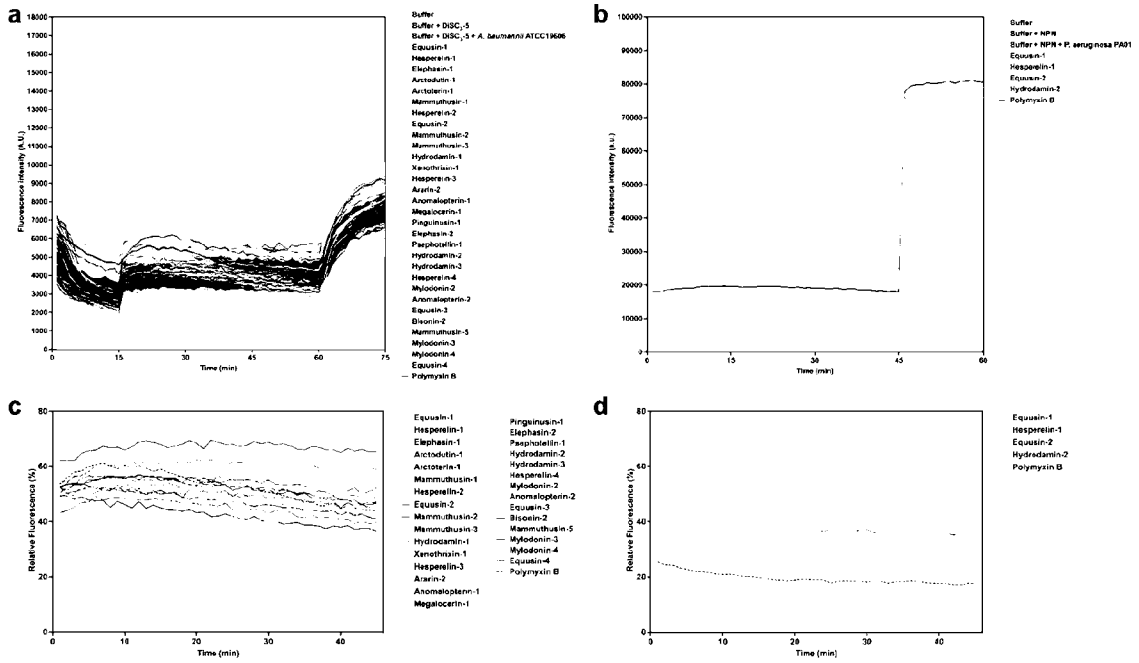
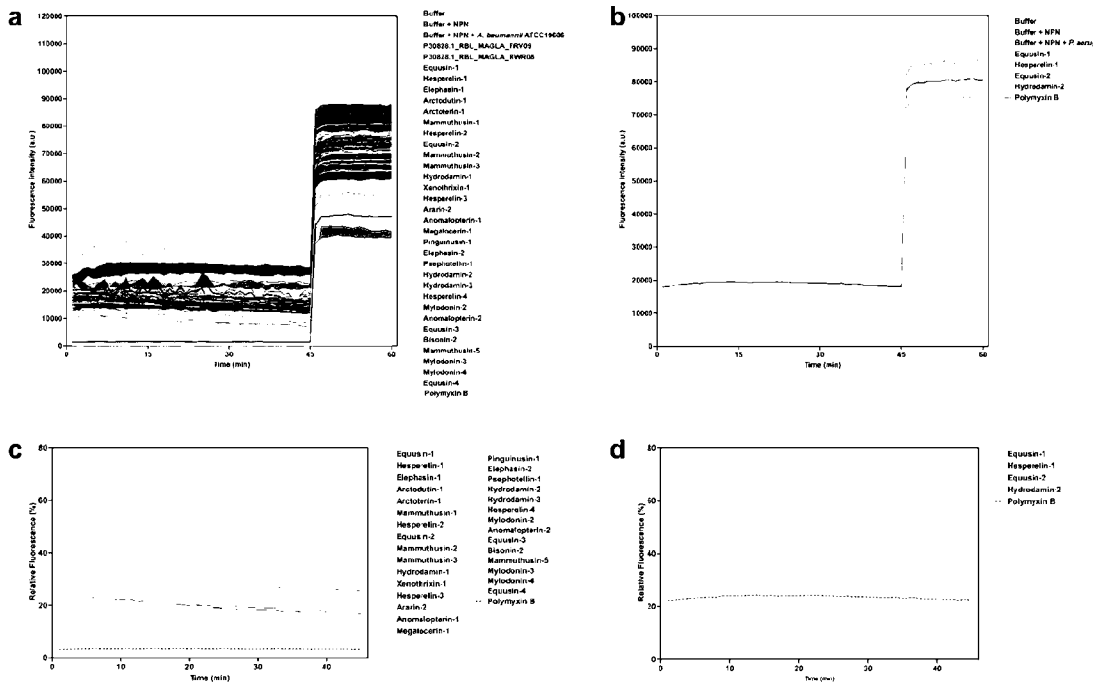


FIG. 52. Predicted vs. experimental MICs for vancomycin-resistant *E. faecium* ATCC 700221 of the encrypted peptides identified by APEX. Each peptide on the scatter plot is represented by a red circle. Pearson and spearman correlations values are shown as an inset.



FIGS. 53A-D. Cytoplasmic membrane depolarization of *A. baumannii* and *P. aeruginosa* triggered by AEPs and MEPs identified by APEX. (a-b) Depolarization assays with the hydrophobic probe 3,3′-dipropylthiadicarbocyanine iodide [$DiSC_3$-(5)] for all the encrypted peptides (EPs) from extinct organisms that were active against *A. baumannii* ATCC 19606 **(a)** and *P. aeruginosa* PAO1 **(b)**. Polymyxin B was used as positive control, and buffer, buffer with $DiSC_3$-(5), and buffer with $DiSC_3$-(5) and bacteria were used as baseline for fluorescence. The panels show the raw fluorescence intensity data obtained in the experiments. **(c-d)** Relative fluorescence values of the depolarization effect of EPs from extinct organisms compared to the untreated control on the cytoplasmic membranes of **(c)** *A. baumannii* ATCC 19606 and **(d)** *P. aeruginosa* PAO1.



FIGS. 54A-D. Outer membrane permeabilization of *A. baumannii* and *P. aeruginosa* cell membranes caused by encrypted peptides from extinct organisms. The probe 1-(N-phenylamino)naphthalene (NPN) was used to detect permeabilization of the outer membrane. **(a-b)** Permeabilization by encrypted peptides (EPs) from extinct organisms active against each one of the strains: on *A. baumannii* ATCC 19606 **(a)** and *P. aeruginosa* PAO1 **(b)**. Polymyxin B was used as positive control. Buffer was used as the baseline for fluorescence, and buffer with NPN, and buffer with NPN and bacteria were used as baseline for fluorescence. The panels show the raw fluorescence intensity data obtained in the experiments. **(c-d)** Relative fluorescence values of the permeabilization effect of encrypted peptides from extinct organisms compared to the untreated control on the outer membranes of **(c)** *A. baumannii* ATCC 19606 and **(d)** *P. aeruginosa* PAO1.

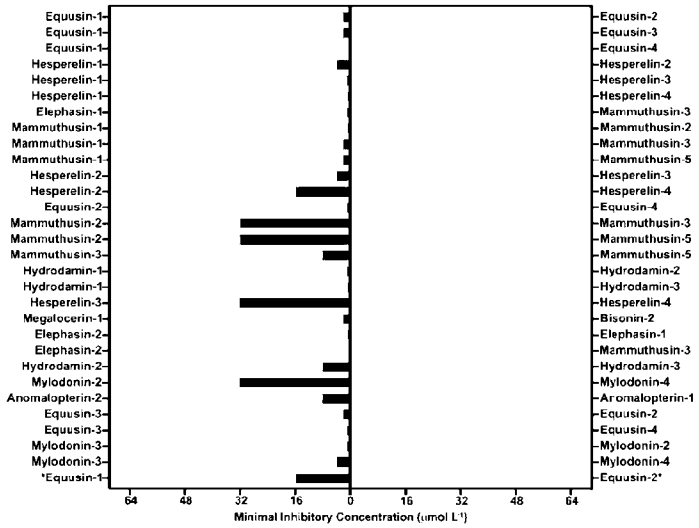


FIG. 55. Synergy between peptide molecules from extinct organisms. Bar plot showing the synergistic interactions between EPs from the same extinct organism against *A. baumannii* and *P. aeruginosa* (*P. aeruginosa* result is indicated with an asterisk). Stacked bars represent the MICs in µmol L^{-1} values in each condition. MICs of the individual peptides are shown in blue and light purple and when in combination, shown in brown and red. Each pair of peptides was placed in the same row for side-by-side comparison of MIC-fold change before and after they were tested in combination.

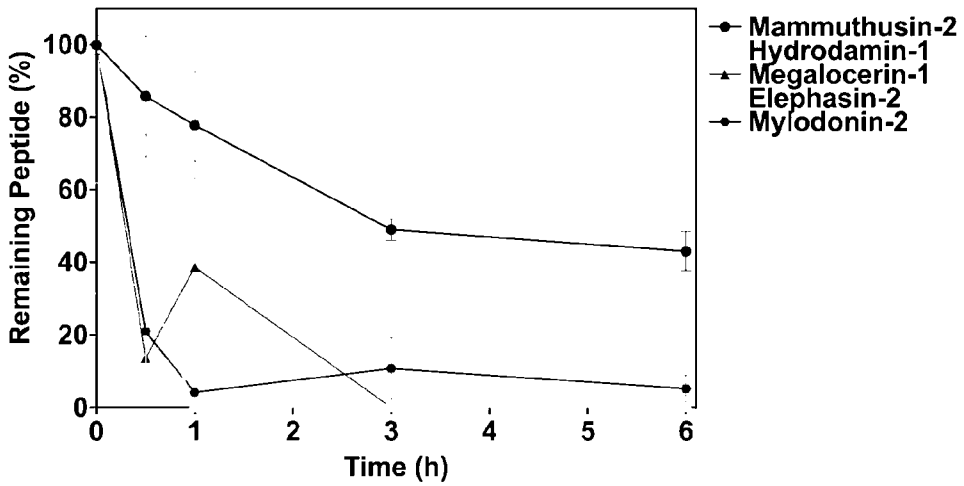


FIG. 56. Resistance to proteolytic degradation assays. The AEPs hydrodamin-1, elephasin-2, and mylodonin-2 and the MEPs mammuthusin-2 and megalocerin-1 were exposed for a total of 6 h to human serum, which contains proteases. Aliquots of the resulting solution were analyzed by liquid chromatography coupled to mass spectrometry. In summary, the AEP elephasin-2 and the MEP mammuthusin-2 exhibited the highest resistance to proteolytic degradation with ~40% peptide remaining after 6 h of exposure. All other AEPs and MEP tested degraded at varying degrees within the duration of the experiment. Experiments were performed in three independent replicates.

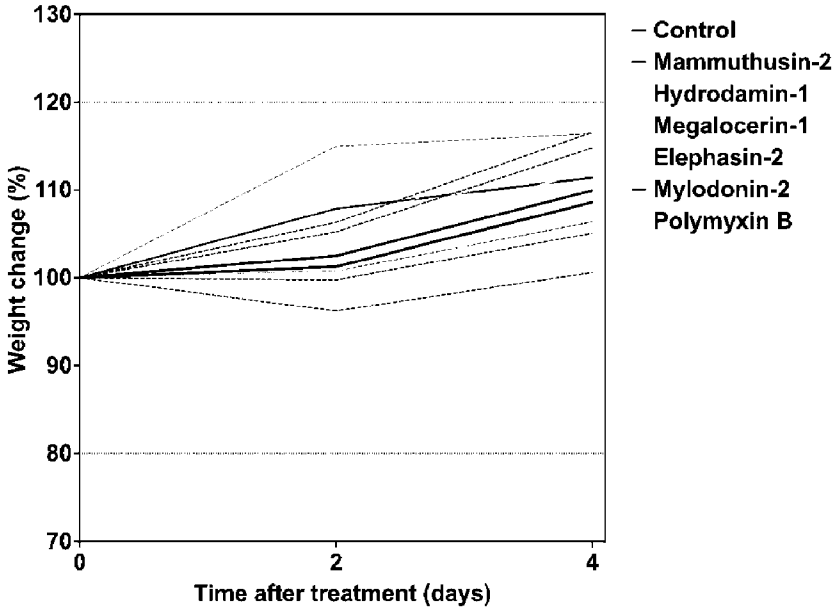


FIG. 57. Weight change monitoring in the skin abscess mouse model infected with *A. baumannii*. Mouse weight was monitored throughout the duration of the skin abscess assay (4 days total) to rule out potential toxic effects of the bacterial load and the encrypted peptides.

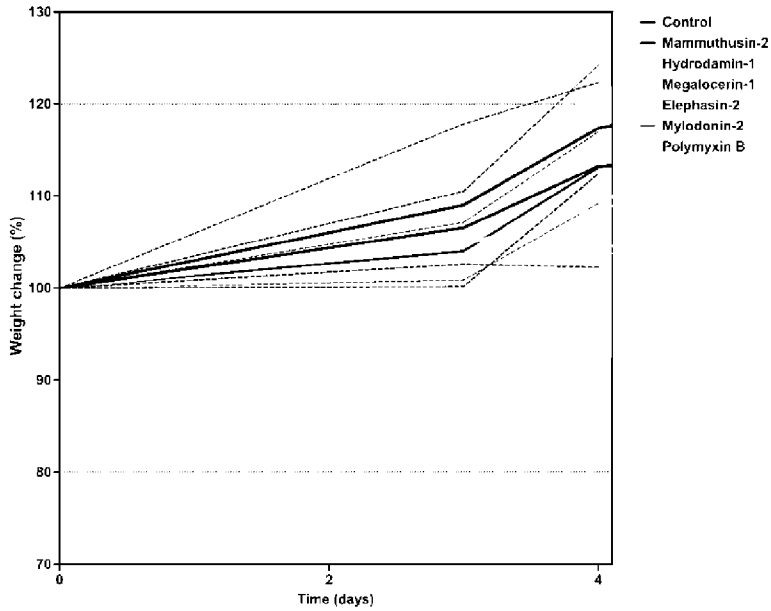


FIG. 58. Weight change monitoring in the thigh mouse model infected with *A. baumannii*. Mouse weight was monitored throughout the duration of the thigh infection (8 days total) to rule out the potentially toxic effects of bacterial load or the encrypted peptides.