Supplementary Materials for

Molecular de-extinction of ancient antimicrobial peptides enabled by machine learning

Jacqueline R. M. A. Maasch, Marcelo D. T. Torres, Marcelo C. R. Melo, Cesar de la Fuente-Nunez

Correspondence to: cfuente@upenn.edu

This PDF file includes:

Figs. S1 to S3
Tables S1 to S5

Supplementary Tables and Figures:

Supplementary Tables

Table S1 related to Fig. 2. panCleave accuracy on proteases with at least 100 test set observations and on protease families and clans with at least 50 test set observations. Protease family and clan IDs are those reported in MEROPS[S1]. Total test set observations (*Total*) and total predictions with predicted labels that are concordant with true labels (*Concordant*) are noted. Mean probability refers to the predicted probability of class assignment averaged over all observations, as returned by the panCleave random forest (RF) model. Table is ordered by descending test set accuracy in each case.

Protease MEROPS ID	Protease name	Concordant	Total	Mean probability	Accuracy
C14.003	Caspase-3	117	118	0.8266	0.9915
C14.005	Caspase-6	217	220	0.8635	0.9864
S01.010	Granzyme B	300	322	0.7647	0.9317
C13.004	Legumain	387	427	0.6714	0.9063
C01.034	Cathepsin S	471	575	0.6403	0.8191
C01.009	Cathepsin V	235	300	0.6362	0.7833
C01.036	Cathepsin K	313	410	0.6353	0.7634
M10.003	Matrix metallopeptidase-2	353	471	0.6363	0.7495
C01.032	Cathepsin L	406	543	0.6336	0.7477
C01.060	Cathepsin B	75	101	0.6480	0.7426
M10.005	Matrix metallopeptidase-3	337	454	0.6264	0.7423
A01.010	Cathepsin E	188	275	0.6101	0.6836
A01.009	Cathepsin D	86	134	0.6180	0.6418
S01.139	Granzyme M	86	136	0.6031	0.6324
M12.004	Meprin beta subunit	108	183	0.6254	0.5902
M12.002	Meprin alpha subunit	91	168	0.6232	0.5417
	Protease family				
-	C14	411	422	0.82512289	0.97393365
-	C13	388	429	0.6712052	0.9044289
-	S8	62	70	0.66046515	0.88571429
-	C1	1026	1360	0.63477728	0.75441176
-	M10	728	1008	0.63091031	0.72222222
-	S1	616	872	0.67541674	0.70642202
-	A1	270	433	0.61190775	0.62355658
-	C2	56	98	0.63135532	0.57142857
-	M12	179	328	0.62257116	0.54573171
-	S26	34	69	0.59704888	0.49275362
-	T1	17	51	0.64760432	0.33333333
	Protease clan				
-	CD	785	836	0.7465014	0.93899522
-	SB	62	70	0.66046515	0.88571429
-	CA	1081	1459	0.63442325	0.74091844
-	PA	616	872	0.67541674	0.70642202
-	MA	910	1346	0.62844619	0.67607727
-	AA	270	433	0.61190775	0.62355658
-	SF	34	69	0.59704888	0.49275362
-	PB	18	52	0.65295276	0.34615385

Table S2 related to Fig. 2. Protease-specific accuracy of panCleave as compared to published cleavage site models. The following table is adapted from Table S6 in[S2]. Reported panCleave values are test set accuracy. Best accuracy per protease is represented in bold type. Note that all comparisons are relative, with accuracies as reported in[S2]; direct comparisons were not possible, as training and testing data were not released for prior models. Models are Cascleave[S3], CAT3[S4], CleavPredict[S5], ScreenCap3[S6], SitePrediction[S7], PROSPERous[S8], and DeepCleave[S2]. Results for DeepCleave are reported with and without transfer learning (TL), as reported in the original publication.

Protease name	MEROPS ID	panCleave	Cascleave	CAT3	CleavPredict	ScreenCap3	SitePrediction	PROSPERous	DeepCleave (without TL)	DeepCleave (with TL)
Caspase-1	C14.001	**0.9268**	0.5119	0.6786	-	0.7368	0.8571	0.8750	0.8315	0.9205
Caspase-3	C14.003	**0.9915**	0.5920	0.7731	-	0.8741	0.9275	0.9656	0.9759	0.9856
Caspase-7	C14.004	0.9524	0.6143	0.8378	-	0.8684	0.9605	0.9359	0.9625	**0.9744**
Caspase-6	C14.005	0.9864	0.5286	0.6583	-	0.7802	0.8767	0.9850	**0.9935**	0.9901
Caspase-2	C14.006	**1.0000**	0.5519	0.7632	-	0.8333	-	-	0.9878	0.9880
Matrix metallopeptidase-8	M10.002	0.6667	-	-	0.7273	-	**0.8750**	0.8250	0.6250	0.7879
Matrix metallopeptidase-2	M10.003	0.7495	-	-	0.6032	-	0.7133	0.8736	-	**0.8899**
Matrix metallopeptidase-9	M10.004	0.6613	-	-	0.6081	-	0.8378	0.8077	0.5734	**0.8613**
Matrix metallopeptidase-3	M10.005	0.7423	-	-	0.7273	-	0.8571	**0.8810**	0.7027	0.8780
Matrix metallopeptidase-7	M10.008	0.7805	-	-	-	-	0.7375	0.8523	0.6264	**0.9318**
Matrix metallopeptidase-12	M10.009	0.7143	-	-	-	-	0.6786	0.8378	0.6316	**0.9079**
Matrix metallopeptidase-1	M10.014	0.4815	-	-	-	-	-	0.8125	0.6275	**0.8519**

Table S3 related to Fig. 3. Antimicrobial and cytotoxic activities of modern and archaic secreted protein fragments. Minimum inhibitory concentration (MIC) values (μmol L^{-1}) for peptides screened against pathogenic strains (dash indicates no activity) in BM2 with glucose (B) and LB (L) media. Curation methods are machine learning model consensus vote (ML), random selection (RS), and human expert (HE). Predicted label (Pred. label) indicates predicted antimicrobial activity (1), no predicted activity (0), or no prediction (NA). Peptides were classified as archaic encrypted peptides (AEPs) and modern encrypted peptides (MEPs). The hemolytic and cytotoxic activities are expressed in terms of HC_{50} and CC_{50} values (μmol L^{-1}), respectively. The values were estimated by non-linear regressions based on the screen of peptides in a gradient of concentrations and represent the hemolytic and the cytotoxic concentration values needed to lyse and kill 50% of the cells present in the experiment. The experiments were done in three independent biological replicates, and for the cytotoxic activity assays, two technical replicates were performed within each biological replicate.

ID	Classification	Fragment sequence	Length	Curation method	Pred. label	Antimicrobial activity, MIC (μmol L^{-1})														Hemolytic Activity vs RBCs (HC_{50}, μmol L^{-1})	Cytotoxic Activity vs HEK293T cells (CC_{50}, μmol L^{-1})
						PA01		PA14		*Ec* AIC221		*Ec* AIC222		*Ab*		*Sa*		MRSA			
						B	L	B	L	B	L	B	L	B	L	B	L	B	L		
CBPZ-GSK24	MEP	GSKPWWWSYFTSLSTHRPRWLLKY	24	ML	1	8	-	4	-	4	-	2	-	-	16	-	-	-	-	19.42	44.19
XDH-AVA32	MEP	AVAKLPAQKTEVFRGVLEQLRWFAGKQVKSVA	32	ML	1	-	-	-	-	32	-	32	-	-	-	-	-	-	-	-	-
LYSC-AVA39	MEP	AVACAKRVVRDPQGIRAWVAWRNRCQNRDVRQYVQGCGV	39	ML	1	-	-	128	-	128	-	128	-	-	-	-	-	-	-	>128	>128
ISK5-GKI32	MEP	GKIHGNTCSMCEAFFQQEAKEKERAEPRAKVK	32	RS	NA	-	-	-	-	128	-	128	-	-	-	-	-	-	-	>128	>128
CALR-GWT20	MEP	GWTSRWIESKHKSDFGKFVL	20	HE	1	-	-	-	-	-	-	128	-	-	64	-	-	-	-	>128	19.28
CO7A1-AIG15	MEP	AIGPKGDRGFPGPLG	15	CD	1	-	-	-	32	-	-	-	-	-	-	-	-	-	-	>128	>128
TKN1-SSI27	MEP	SSIEKQVALLKALYGHGQISHKRHKTD	27	CD	1	-	-	-	-	-	-	-	-	-	64	-	-	-	-	>128	40.27
A7E2T1-SPR29	MEP	SPRYHTVGRAAGLLM	29	CD	1	-	-	-	-	-	64	-	64	-	8	-	-	-	-	112	12.78

		GLRRSPYL WRRALR																			
PDB6I34D-ALQ29	AEP	ALQLCYRH NKRRKFFV DPRCHPQTI AVVQ	29	-	-	64	-	32	-	128	128	-	-	-	-	-	-	-	-	>128	>128
A0A384E0N4-DLI09	AEP	DLIERIQAD	9	-	-	-	-	-	-	-	-	-	-	-	128	-	128	-	128	>128	>128
A0A343EQH4-LAM11	AEP	LAMVIPLW AGA	11	-	-	-	-	-	-	-	-	-	-	-	128	-	-	-	-	>128	>128
A0A343AZS4-FMA25	AEP	FMAEYTNII MMNTLTTT IFLGTTYN	25	-	-	-	-	-	-	-	-	-	-	-	128	-	-	-	-	-	-
A0A343EQH0-NVK38	AEP	NVKMKWQ FEHTKPTPF LPTLITLTT LLLPISPFM LMIL	38	-	-	-	128	-	-	-	-	-	-	-	-	-	-	-	-	54.72	>128
A0A0S2IB02-AYT38	AEP	AYTTWNIL SSAGSFISL TAVMLMIF MIWEAFAS KRKVL	38	-	-	-	128	-	-	-	-	-	-	-	-	-	-	-	-	88.11	>128

PA01: *P. aeruginosa* PA01; PA14: *P aeruginosa* PA14; *Ec* AIC221: *E. coli* AIC221; *Ec* AIC222: *E. coli* AIC222; *Ab*: *A. baumannii* ATCC19606; *Sa*: *S. aureus* ATCC12600; MRSA: methicillin-resistant *S. aureus* ATCC BAA-1556.

Table S4 related to Fig. 2. Physicochemical properties of archaic and modern encrypted peptides, and previously described encrypted and antimicrobial peptides. Medians are reported with standard deviations in parentheses. Physicochemical properties were calculated using the DBAASP[S9] (https://dbaasp.org/tools?page=property-calculation) and the Eisenberg and Weiss hydrophobicity scale[S10]. Reported properties are: Normalized Hydrophobic Moment (NHM), Normalized Hydrophobicity (NH), Net Charge (NC), Isoelectric Point (IP), Penetration Depth (PD), Tilt Angle (TA), Disordered Conformation Propensity (DCP), Linear Moment (LM), Propensity to *in vitro* Aggregation (PA), Angle Subtended by the Hydrophobic Residues (AS), Amphiphilicity Index (AI), and Propensity to PPII coil (PC).

Fragment ID	NHM	NH	NC	IP	PD	TA	DCP	LM	PA	AS	AI	PC
CBPZ-GSK24	0.11	0.02	4	10.87	15	90	-0.19	0.31	356.51	40	2.15	1.13
XDH-AVA32	0.32	-0.02	4	11.07	18	114	0.12	0.14	0.00	90	1.03	0.95
LYSC-AVA39	0.28	0.25	6	10.98	30	141	-0.06	0.24	12.06	60	1.15	1.01
ISK5-GKI32	0.18	0.30	2	8.93	30	42	-0.16	0.31	0.00	30	1.05	0.97
CALR-GWT20	0.10	0.06	2	10.43	24	159	-0.08	0.33	0.00	40	1.50	0.96
CO7A1-AIG15	0.33	-0.14	1	10.17	22	140	0.06	0.50	0.00	170	0.41	1.01
TKN1-SSI27	0.14	0.15	3	10.39	19	70	-0.11	0.33	4.11	40	1.12	0.95
A7E2T1-SPR29	0.27	0.23	7	12.23	16	97	-0.15	0.31	1.58	50	1.23	1.02
PDB6I34D-ALQ29	0.07	0.23	5	10.75	30	16	-0.1	0.34	208.02	50	0.99	1.09
A0A384E0N4-DLI09	0.54	0.16	-2	3.57	17	72	0.33	0.31	0.00	130	0.55	0.93
A0A343EQH4-LAM11	0.18	-0.77	0	3.5	5	82	0.82	0.00	0.00	360	0.63	0.99
A0A343AZS4-FMA25	0.06	-0.35	-1	3.22	13	89	0.43	0.27	1043.42	60	0.46	0.98
A0A343EQH0-NVK38	0.10	-0.34	2	10.38	13	148	0.38	0.42	892.42	50	0.58	1.11
A0A0S2IB02-AYT38	0.09	-0.39	2	10.28	3	40	0.48	0.39	1447.14	90	0.79	0.99
Archaic encrypted peptides *(n = 6)*	0.10 (0.18)	-0.35 (0.38)	1 (2.53)	6.93 (3.86)	13 (9.67)	77 (45.36)	0.41 (0.3)	0.33 (0.15)	550.22 (612.11)	75 (119.94)	0.61 (0.19)	0.99 (0.07)
Modern encrypted peptides *(n = 8)*	0.23 (0.09)	0.11 (0.15)	3.50 (2.07)	10.65 (0.94)	20.50 (5.87)	105.50 (39.6)	-0.10 (0.11)	0.31 (0.1)	0.79 (125.22)	45 (46.29)	1.14 (0.49)	0.99 (0.06)
Previously reported encrypted peptides [S11] *(n = 35)*	0.23 (0.15)	0.20 (0.19)	7 (3.34)	11.58 (0.78)	21 (7.19)	88 (30.89)	-0.17 (0.21)	0.23 (0.05)	0 (127.27)	60 (31.48)	1.23 (0.31)	1.03 (0.12)
AMPs in the DBAASP [S9] *(n = 14,995)*	0.37 (0.24)	0.07 (0.42)	4 (3.18)	11.15 (2.21)	15 (6.98)	88 (33.21)	-0.05 (0.46)	0.29 (0.11)	0 (149.89)	110 (69.95)	1.23 (0.91)	0.97 (0.13)

Table S5 related to Fig. 2. Precursor proteins of modern and archaic encrypted peptides with antimicrobial activity. The precursor protein data for the archaic encrypted peptides was obtained from UniProt[S12] and NCBI Protein Databases (https://www.ncbi.nlm.nih.gov/protein/). The precursor protein data for the modern encrypted peptides was obtained from the PANTHER Classification System (http://www.pantherdb.org/)[S13]. PANTHER protein class is indicated by an asterisk (*). Percent identity shared of archaic precursor proteins with modern human proteins (% ID) and query coverage (QC) were computed by BLAST analysis[S14], with values reported for the top modern human hits. Dashes indicate that data were unavailable.

Classification	Precursor name	Hominin	UniProt ID	NCBI ID	% ID (QC)	Length	Keywords	Features	Fragment ID	Fragment sequence
AEP	Chain D, Neanderthal Glycine decarboxylase	*Homo sapiens neanderthalensis* (Neanderthal)		gi|1777435468|pdb|6I34|D	99.90% (100%)	984	-	-	PDB6I34D-ALQ29	ALQLCYRHNKRRKFFV DPRCHPQTIAVVQ
AEP	Adenylosuccinate lyase (ASL) (EC 4.3.2.2) (Adenylosuccinase)	*Homo sapiens neanderthalensis* (Neanderthal)	A0A384E0N4	gi|1393955578|pdb|5NXA|H	99.59% (100%)	487	3D-structure; Coiled coil; Lyase; Purine biosynthesis	Coiled coil (1); Domain (1)	A0A384E0N4-DLI09	DLIERIQAD
AEP	ATP synthase subunit a	*Homo sapiens neanderthalensis* (Neanderthal)	A0A343EQH4	gi|1214786277|gb|ASK06270.1|	99.56% (100%)	226	ATP synthesis; CF(0); Hydrogen ion transport; Ion transport; Membrane; Mitochondrion; Mitochondrion inner membrane; Transmembrane; Transmembrane helix; Transport	Transmembrane (6)	A0A343EQH4-LAM11	LAMVIPLWAGA
AEP	NADH-ubiquinone oxidoreductase chain 1 (EC 7.1.1.2)	*Homo sapiens* subsp. 'Denisova' (Denisova hominin)	A0A343AZS4	gi|1141958635|gb|AQD17584.1|	99.06% (100%)	318	Membrane; Mitochondrion; NAD; Transmembrane; Transmembrane helix; Transport; Ubiquinone	Transmembrane (8)	A0A343AZS4-FMA25	FMAEYTNIIMMNTLTTT IFLGTTYN
AEP	NADH-ubiquinone oxidoreductase chain 2 (EC 7.1.1.2)	*Homo sapiens neanderthalensis* (Neanderthal)	A0A343EQH0	gi|1578894740|gb|ASK06266.2|	99.38% (92%)	347	Electron transport; Membrane; Mitochondrion; Mitochondrion inner membrane; NAD; Respiratory chain; Translocase; Transmembrane; Transmembrane	Domain (2); Transmembrane (8)	A0A343EQH0-NVK38	NVKMKWQFEHTKPTP FLPTLITLTTLLLP ISPF MLMIL

							helix; Transport; Ubiquinone			
AEP	Cytochrome c oxidase subunit 1 (EC 7.1.1.9)	*Homo sapiens* subsp. 'Denisova' (Denisova hominin)	A0A0S2IB02	gi|1141958637|gb|AQD17586.1|	99.42% (100%)	513	Calcium; Copper; Electron transport; Heme; Iron; Magnesium; Membrane; Metal-binding; Mitochondrion; Mitochondrion inner membrane; Respiratory chain; Sodium; Translocase; Transmembrane; Transmembrane helix; Transport	Domain (1); Transmembrane (12)	A0A0S2IB02-AYT38	AYTTWNILSSAGSFIS LTAVMLMIFMIW EAF ASKRKVL
MEP	Calreticulin	*Homo sapiens sapiens*	CALR_HUMAN	HUMAN|HGNC=1455|UniProtKB=P27797	-	-	Chaperone*	-	CALR-GWT20	GWTSRWIESKHK SDFGKFVL
MEP	Xanthine dehydrogenase/oxidase	*Homo sapiens sapiens*	XDH_HUMAN	HUMAN|HGNC=12805|UniProtKB=P47989	-	-	Oxidoreductase*	-	XDH-AVA32	AVAKLPAQKTEV FRGVLE QLRWFAGKQVKS VA
MEP	Serine protease inhibitor Kazal-type 5	*Homo sapiens sapiens*	ISK5_HUMAN	HUMAN|HGNC=15464|UniProtKB=Q9NQ38	-	-	Protease inhibitor*	-	ISK5-GKI32	GKIHGNTCSMCE AFFQQE AKEKERAEPRAK VK
MEP	Carboxypeptidase Z	*Homo sapiens sapiens*	CBPZ_HUMAN	HUMAN|HGNC=2333|UniProtKB=Q66K79	-	-	Protease*	-	CBPZ-GSK24	GSKPWWWSYFTS LST HRPRWLLKY
MEP	Lysozyme C	*Homo sapiens sapiens*	LYSC_HUMAN	HUMAN|HGNC=6740|UniProtKB=P61626	-	-	-	-	LYSC-AVA39	AVACAKRVVRDP QGIRA WVAWRNRCQNR DVRQY VQGCGV
MEP	Collagen alpha-1(VII) chain (Long-chain collagen) (LC collagen)	*Homo sapiens sapiens*	CO7A1_HUMAN	HUMAN|HGNC=2214|UniProtKB=Q02388	-	-	Extracellular matrix -structural -protein*	-	CO7A1-AIG15	AIGPKGDRGFPGP LG
MEP	Protachykinin-1 (PPT) [Cleaved into: Substance P; Neurokinin A (NKA) (Neuromedin L) (Substance K); Neuropeptide K (NPK); Neuropeptide	*Homo sapiens sapiens*	TKN1_HUMAN	HUMAN|HGNC=11517|UniProtKB=P20366	-	-	-	-	TKN1-SSI27	SSIEKQVALLKAL YGHGQIS HKRHKTD

	gamma; C-terminal-flanking peptide]									
MEP	Uncharacterized protein (Fragment)	*Homo sapiens sapiens*	A7E2T1_HUMAN		-	-	-	-	A7E2T1-SPR29	SPRYHTVGRAAGLLMGLRRSPYLWRRALR

96

Supplementary Figures

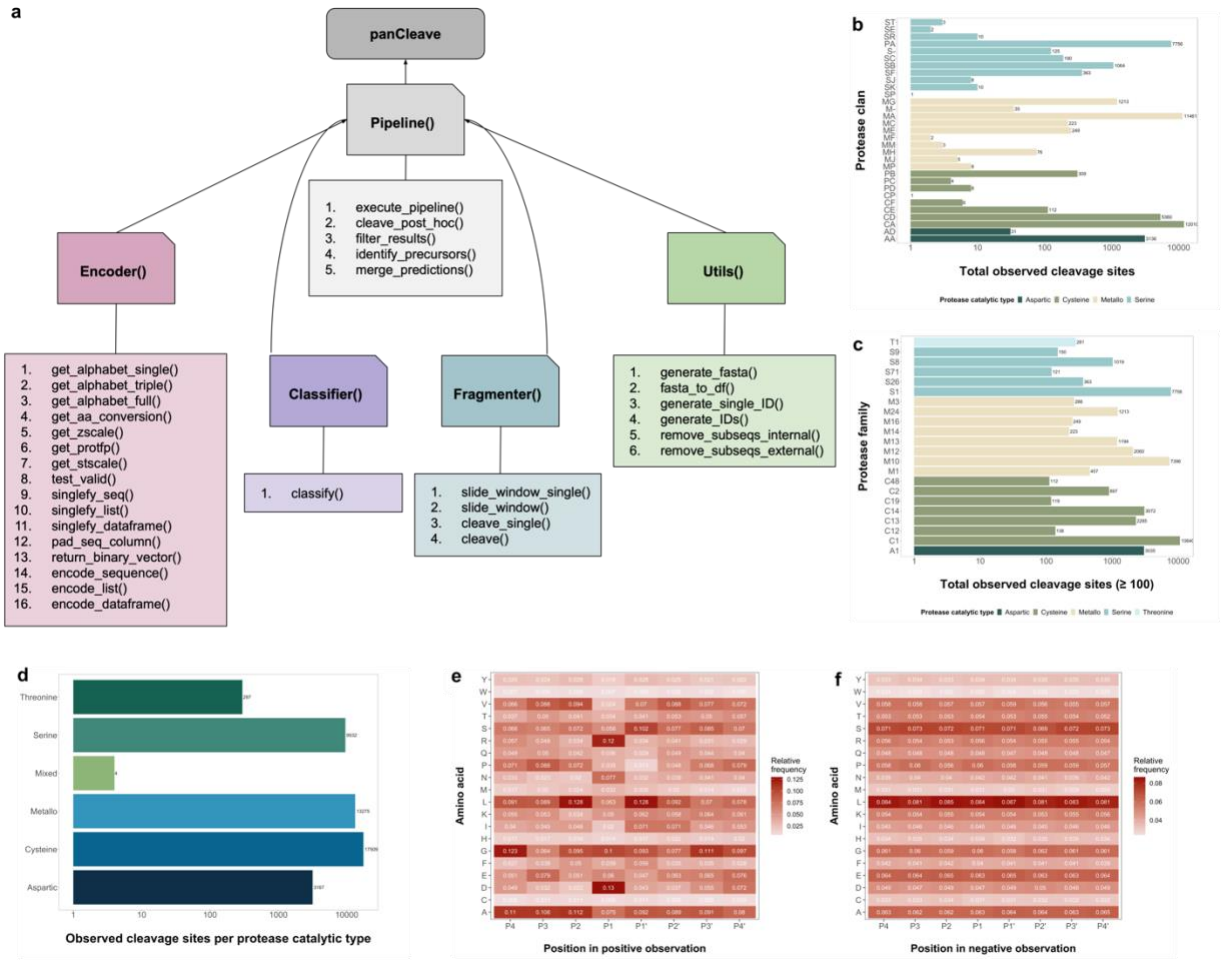
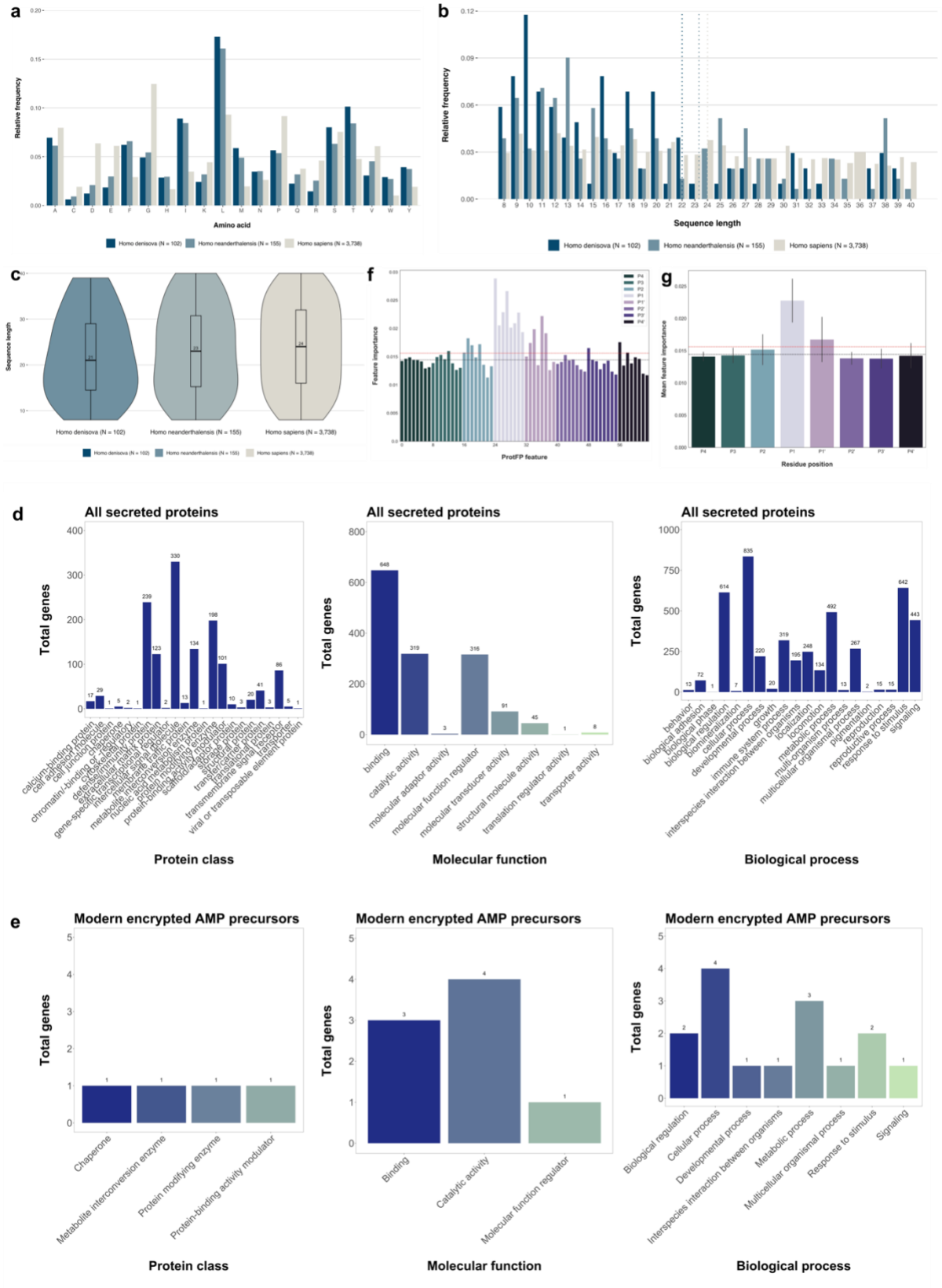


Fig. S1 related to Fig. 2. Domain model for the panCleave pipeline in Python, representation of proteases among substrate cleavage sites in panCleave training and testing data (*n* = 24,817), and amino acid frequencies by residue position for all training and testing data. (a) The class *Pipeline* is dependent on classes *Encoder*, *Classifier*, *Fragmenter*, and *Utils*. Each class features the methods enumerated here. Plots represent total cleavage sites arranged by **(b)** protease clan, **(c)** family, and **(d)** catalytic type, as defined by the MEROPS Peptidase Database[S1]. Amino acid relative frequencies are reported by residue position in **(e)** 8-residue positive observations (*n* = 24,817) and **(f)** 8-residue negative observations (*n* = 24,817). Cleavage takes place between positions P1 and P1' in positive observations.


a
Relative frequency
Amino acid
Homo denisova (N = 102)
Homo neanderthalensis (N = 155)
Homo sapiens (N = 3,738)
b
Relative frequency
Sequence length
c
Sequence length
f
Feature importance
ProtFP feature
g
Mean feature importance
Residue position
d
All secreted proteins
Total genes
Protein class
Molecular function
Biological process
e
Modern encrypted AMP precursors
Chaperone
Metabolite interconversion enzyme
Protein modifying enzyme
Protein-binding activity modulator
Binding
Catalytic activity
Molecular function regulator
Biological regulation
Cellular process
Developmental process
Interspecies interaction between organisms
Metabolic process
Multicellular organismal process
Response to stimulus
Signaling

118

Fig. S2 related to Fig. 2. Relative frequencies of amino acids and fragment length for all unique panCleave fragments per taxon; origin, molecular functions, and biological processes represented by all queried human secreted proteins and by precursors of modern encrypted peptides discovered in the present work, and panCleave feature importance based on mean decrease in impurity. (a) Relative amino acid frequency for all generated fragments, not just those filtered for synthesis or with demonstrated activity. **(b)** Relative frequency of sequence length of all secreted proteins available in UniProt[S12]. **(c)** Sequence length distribution of EPs across the different hominids. Plots shown in **(a)**, **(b),** and **(c)** include all generated fragments, not just those filtered for synthesis or with demonstrated activity. Protein classes, molecular functions, and biological processes represented by **(d)** all queried human secreted proteins available in UniProt[S12] and **(e)** by precursors of modern encrypted peptides discovered in the present work. Data were obtained from PANTHER (http://www.pantherdb.org/)[S27[S13]. **(f and g)** Features importance were calculated by in-built functions provided by scikit-learn for random forests. Panel **(f)** plots the importance of each individual ProtFP feature. Each of the eight residues in a given P4:P4' cleavage flanking site is encoded by eight floating point features, as computed under the ProtFP encoding scheme. Thus, each input sequence is represented by 64 features. Panel **(g)** plots the average importance of each residue position, with error bars signifying standard deviation. Mean and median importance across all ProtFP features are visualized as red and black dashed lines, respectively.

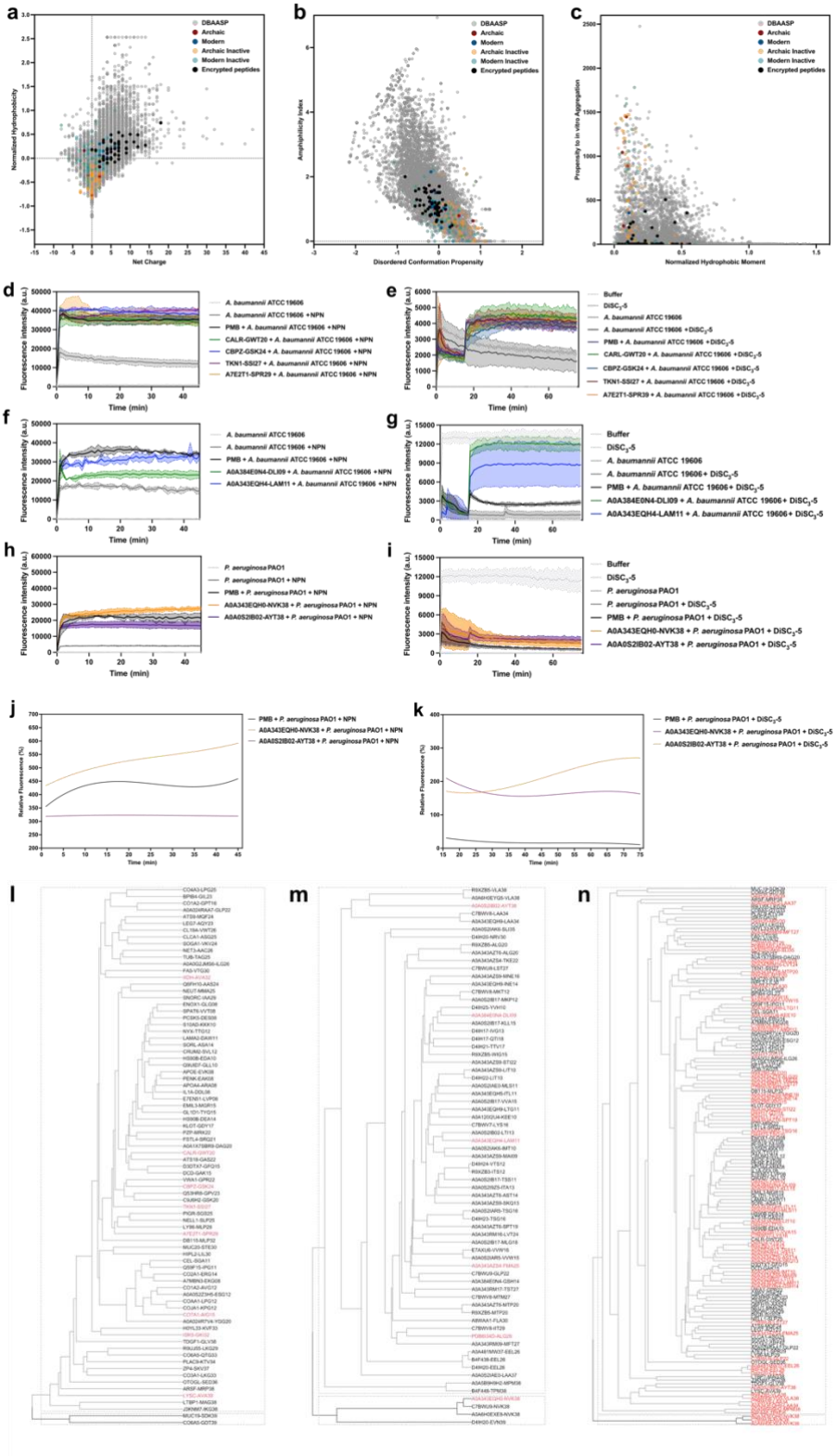


Fig. S3 related to STAR Methods and Fig. 3. Physicochemical features, mechanism of action, and clustering according to antimicrobial activity of encrypted peptides identified by panCleave. (a) Net charge vs. hydrophobicity normalized according to the length of the peptide.

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. **(b)** Amphiphilicity index vs. disordered conformation propensity; both properties closely correlated with AMP mechanism of action. **(c)** Propensity to aggregate *in vitro* vs. hydrophobic moment normalized by peptide length; propensity to aggregate correlates with AMP toxicity. All properties were calculated using the DBAASP property calculator tool[S9]. AEPs and MEPs were compared to known AMPs and other previously described encrypted peptides from the human proteome. Panels **(d)** to **(i)**: Permeabilization assays with the fluorescent probe 1-(N-phenylamino)naphthalene (NPN); effect of **(d)** modern encrypted peptides and **(e)** archaic encrypted peptides on against *A. baumannii* cells, and **(f)** archaic encrypted peptides on *P. aeruginosa* PA01 cells. Depolarization assays with the hydrophobic probe 3,3′-dipropylthiadicarbocyanine iodide [$DiSC_3$-(5)]; effects of **(g)** modern encrypted peptides and **(h)** archaic encrypted peptides on *A. baumannii* cells, and **(i)** archaic encrypted peptides on *P. aeruginosa* PA01 cells. All panels show the raw fluorescence intensity data obtained in the experiments. Panels **(j)** and **(k)**: Relative fluorescence values of archaic encrypted peptides compared to the untreated control**. (j)** Permeabilization of the outer membrane using the probe 1-(N-phenylamino)naphthalene (NPN) and **(k)** depolarization of the cytoplasmic membrane indicated by the probe 3,3′-dipropylthiadicarbocyanine iodide [$DiSC_3$-(5)] of *P. aeruginosa* PA01 cells. Both peptides depolarized membranes more strongly than polymyxin B (control). A0A343EQH0-NVK38 permeabilized outer membranes more strongly than PMB or A0A0S2IB02-AYT38. Archaic **(l)** and modern **(m)** encrypted peptides did not cluster neatly according to the presence (red labels) or absence (black labels) of antimicrobial activity. Likewise, hierarchical *k*-means clustering of archaic (red labels) and modern (black labels) in panel **(n)** did not reveal clean separation of archaic and modern fragments. Values for *k* (*i.e.*, total clusters; $k = 2$ for all subfigures) were selected based on joint evaluation of gap statistic, average silhouette score, and within-cluster sum of squares methods. Data were represented using the ProtFP encoding method[S15] and scaled prior to clustering method and scaled prior to clustering.

References

S1. Rawlings, N. D., Barrett, A. J., Thomas, P. D., Huang, X., Bateman, A., & Finn, R. D. (2018). The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. *Nucleic Acids Research*, *46*(D1), D624–D632. https://doi.org/10.1093/nar/gkx1134

S2. Li, F., Chen, J., Leier, A., Marquez-Lago, T., Liu, Q., Wang, Y., … Song, J. (2020). DeepCleave: a deep learning predictor for caspase and matrix metalloprotease substrates and cleavage sites. *Bioinformatics*, *36*(4), 1057–1065. https://doi.org/10.1093/bioinformatics/btz721

S3. Wang, M., Zhao, X.-M., Tan, H., Akutsu, T., Whisstock, J. C., & Song, J. (2014). Cascleave 2.0, a new approach for predicting caspase and granzyme cleavage targets. *Bioinformatics*, *30*(1), 71–80. https://doi.org/10.1093/bioinformatics/btt603

S4. Ayyash, M., Tamimi, H., & Ashhab, Y. (2012). Developing a powerful In Silico tool for the discovery of novel caspase-3 substrates: a preliminary screening of the human proteome. *BMC Bioinformatics*, *13*(1), 14. https://doi.org/10.1186/1471-2105-13-14

S5. Kumar, S., Ratnikov, B. I., Kazanov, M. D., Smith, J. W., & Cieplak, P. (2015). CleavPredict: A Platform for Reasoning about Matrix Metalloproteinases Proteolytic Events. *PLOS ONE*, *10*(5), e0127877. https://doi.org/10.1371/journal.pone.0127877

S6. Fu, S., Imai, K., Sawasaki, T., & Tomii, K. (2014). ScreenCap3: Improving prediction of caspase-3 cleavage sites using experimentally verified noncleavage sites. *PROTEOMICS*, *14*(17–18), 2042–2046. https://doi.org/10.1002/pmic.201400002

S7. Verspurten, J., Gevaert, K., Declercq, W., & Vandenabeele, P. (2009). SitePredicting the cleavage of proteinase substrates. *Trends in Biochemical Sciences*, *34*(7), 319–323. https://doi.org/10.1016/j.tibs.2009.04.001

S8. Song, J., Li, F., Leier, A., Marquez-Lago, T. T., Akutsu, T., Haffari, G., … Pike, R. N. (2018). PROSPERous: high-throughput prediction of substrate cleavage sites for 90 proteases with improved accuracy. *Bioinformatics*, *34*(4), 684–687. https://doi.org/10.1093/bioinformatics/btx670

S9. Pirtskhalava, M., Amstrong, A. A., Grigolava, M., Chubinidze, M., Alimbarashvili, E., Vishnepolsky, B., … Tartakovsky, M. (2021). DBAASP v3: database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics. *Nucleic Acids Research*, *49*(D1), D288–D297. https://doi.org/10.1093/nar/gkaa991

S10. Eisenberg, D., Schwarz, E., Komaromy, M., & Wall, R. (1984). Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *Journal of Molecular Biology*, *179*(1), 125–142. https://doi.org/https://doi.org/10.1016/0022-2836(84)90309-7

S11. Torres, M. D. T., Melo, M. C. R., Flowers, L., Crescenzi, O., Notomista, E., & de la Fuente-Nunez, C. (2022). Mining for encrypted peptide antibiotics in the human proteome. *Nature Biomedical Engineering*, *6*(1), 67–75. https://doi.org/10.1038/s41551-021-00801-1

S12. The UniProt Consortium. (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, *47*(D1), D506–D515. https://doi.org/10.1093/nar/gky1049

S13. Mi, H., Ebert, D., Muruganujan, A., Mills, C., Albou, L.-P., Mushayamaha, T., & Thomas, P. D. (2021). PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Research*, *49*(D1), D394–D403. https://doi.org/10.1093/nar/gkaa1106

S14. Altschul, S. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402. https://doi.org/10.1093/nar/25.17.3389

S15. van Westen, G. J., Swier, R. F., Cortes-Ciriano, I., Wegner, J. K., Overington, J. P., IJzerman, A. P., … Bender, A. (2013). Benchmarking of protein descriptor sets in proteochemometric modeling (part 2): modeling performance of 13 amino acid descriptor sets. *Journal of Cheminformatics*, *5*(1), 42. https://doi.org/10.1186/1758-2946-5-42