nature biomedical engineering

Supplementary information

https://doi.org/10.1038/s41551-024-01243-1

Deep mutational scanning and machine learning for the analysis of antimicrobial-peptide features driving membrane selectivity

In the format provided by the authors and unedited

Deep mutational scanning and machine learning for the analysis of antimicrobial-peptide features driving membrane selectivity

Justin R. Randall[1], Luiz C. Vieira[2], Claus O. Wilke[2], Bryan W. Davies[1]*

Correspondence to:
Bryan Davies (bwdavies@austin.utexas.edu)

This PDF includes:

Supplemental Methods

Machine Learning Data Set. The data set used consisted of empirical data collected from 96 PG-1 variants (Data file 1). To train and evaluate our models, we divided this data set into a training set, which encompassed 80% of the original data selected as a random sample, and a separate test set (The unseen data set to test our models) containing the remaining 20%. This data set is composed of three target predictors, MIC (Minimum Inhibitory Concentration) where undefined values such as ">64 μg/ml MIC" were standardized to "100 μg/ml MIC". Peptides with an MIC value of ≤ 8 μg/ml were deemed "active", whereas those exceeding this threshold were labeled "non-active". The rationale for choosing MIC ≤ 8 μg/ml as the division point was its ability to provide a balanced distribution between the two classes, an essential characteristic for training an effective classification model. A second target predictor is the hemolysis percentage, which correspond to the percentage of human red blood cell lyses induced by the peptide. In addition, we introduced a third target predictor variable called 'selectivity' by multiplying the MIC value by the hemolysis value and subsequently applying a logarithmic transformation (log10). A lower selectivity score indicates an active peptide that is also non-hemolytic.

Protein representation and feature extraction. To ensure a robust representation of the peptides in our data set, we utilized transfer learning from the esm2 15 billion parameter model[30], a pre-trained protein Language Model (pLM). Features learned by the pLM can be transferred to any task (prediction) requiring a numerical representation (embeddings) by extracting these embeddings after the pre-training stage[31] .This method enabled us to obtain numerical embeddings for each peptide sequence, resulting in a high-quality protein representation. The embedding extraction script can be found on the esm2 GitHub page: (https://github.com/facebookresearch/esm#main-models). Since the extracted embeddings contained more than five thousand features, we needed to address the issue of overfitting. To mitigate this, we employed a feature selection technique to choose only those features that exhibited a high correlation with the target variable. We set a high correlation threshold to maintain a minimal set of features, approximately to the number of samples in our data set. This approach allowed us to alleviate overfitting while retaining the most relevant and informative features. We tried different number of features, although the reported number was the one that gave us a good balance between performance while it tries to mitigate overfitting.

Machine learning modeling. We have implemented a consensus machine learning model that determines whether a peptide sequence can be classified as a promising hit or a potential antimicrobial peptide. To achieve this classification, the peptide sequence must meet the cutoff thresholds of three distinct models trained to predict peptide activity (MIC), hemolysis, and selectivity. Using the MIC values, we trained a Gaussian Naïve Bayes classifier model using the hyperparameter(var_smoothing=1e-09) to predict antimicrobial activity based solely on the peptide's amino acid sequence as ‘active’ or ‘non-active’. The activity model achieved the following scores, Accuracy: 0.68, Recall: 0.75, Precision: 0.75, F1 Score: 0.75, and an AUC score of 0.84 on test set. We utilized empirically obtained hemolysis percentages from laboratory experiments as the target values to train a Lasso Regressor model hyperparameters (alpha=0.03) capable of predicting the percentage of hemolysis induced by each peptide. The hemolysis model achieved a R^2 score of 0.73 and a Mean Absolute Error (MAE) of 10.42 on test set. For the selectivity score created we trained a Support Vector Machine Regressor with the hyperparameter (kernel='poly', C=10, gamma=’scale’) to predict the level of selectivity score for each peptide. The selectivity model achieved a R^2 score of 0.68 and a MAE of 0.34 on test set. All three models were trained and validated using the Scikit-learn library. To optimize their performance, we fine-tuned their hyperparameters using Grid Search, coupled with 10-fold cross-validation. For each target, the best-performing model was chosen based on its

performance on the unseen data, referred to as the test set. Detailed information and implementation can be found on our GitHub page (https://github.com/ziul-bio/DMS_ML_AMP) in the Jupyter Notebooks 04-06.

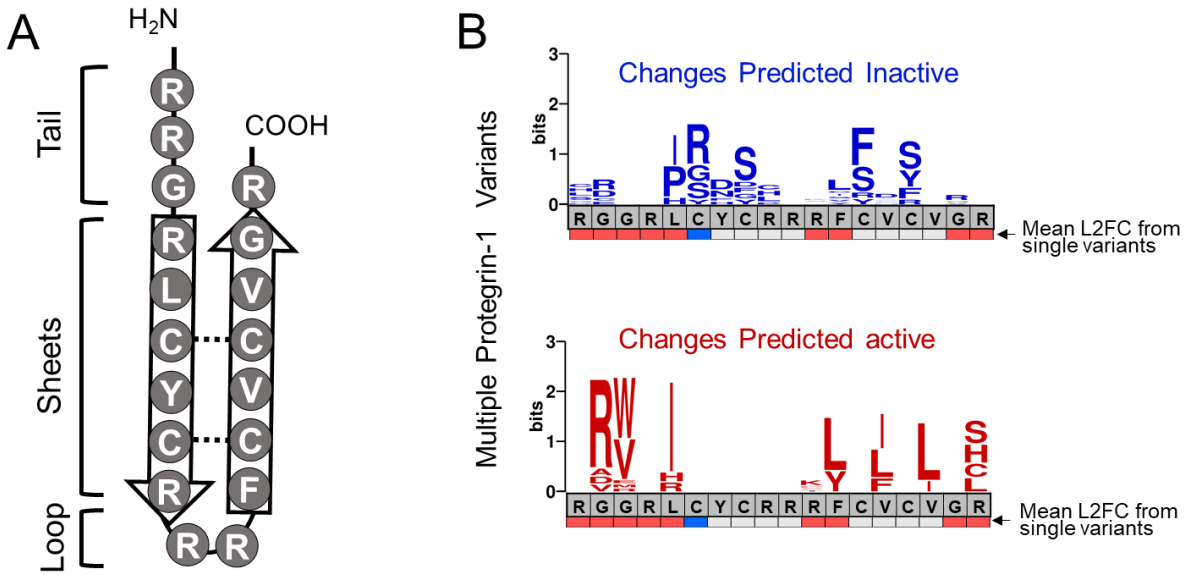

Figure S1: Protegrin-1 structure and multiple variant dmSLAY predictions. A) a diagram of the sequence and secondary structure found in Protegrin-1 divide into three regions: loop, sheets, and tail. beta-sheets are represented by arrows and disulfide bonds by dashed lines. B) Sequence logo plots showing mutations found in the 50 multiple Protegrin-1 dmSLAY variants with the highest (top) and lowest (bottom) log_2-fold change in reads. The native Protegrin-1 sequence is shown in gray and mean average log_2-fold change prediction from single residue dmSLAY variants is color coded by position.

A

Effect of colistin resistance on Protegrin-1 activity.

	E. coli W3110 MIC (µg/ml)		
Protegrin Variant	**No *mcr-1***	**+ *mcr-1***	**Fold Change**
PG-1.0	8	8	1
PG-1.37	4	4	1
bsPG-1.2	8	16	2
Colistin	0.25	4	16

MIC: median minimum inhibitory concentration
mcr-1: mobile colistin resistance gene

B

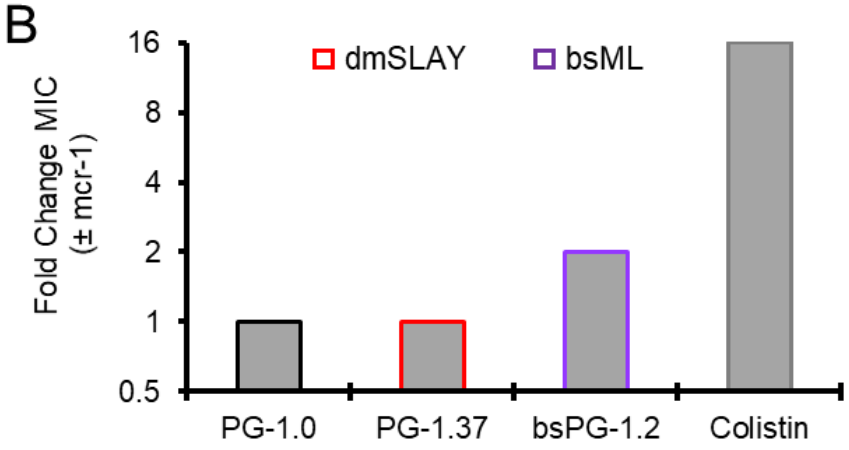

Figure S2. Effect of mobile colistin resistance genes on PG-1 variant activity. A) table showing MICs and fold change with and without mobile colistin resistance gene 1 (mcr-1) expression. B) Bar graph showing the fold change in MIC against *E. coli* W3110 expressing mcr-1 for PG-1.0 and selective PG-1 variants compared to Colistin. MICs were performed in biological triplicate and the median was used to calculate fold-change.

Table S1. dmSLAY in vitro Protegrin-1 variant data.

Name	Sequence Change	Changes	L2FC	padj	MIC (μg/ml)	%Hemo	SS
PG-1.0	RGGRLCYCRRRFCVCVGR	0	-1.00	2E-76	4	56.4	226
PG-1.1	RGGRLCYCRRR**S**CVCVGR	1	-0.833	2E-47	8	5.0	40
PG-1.9	RG**W**RLCYCRRRFCVCVGR	1	-1.675	4E-178	16	89.4	1430
PG-1.10	RGGRLC**F**CRRRFCVCVGR	1	-1.714	3E-202	32	29.0	928
PG-1.11	RGGRLCYCRRRFC**F**CVGR	1	-1.473	1E-200	64	13.6	873
PG-1.12	RGGRLCYCRRRFCVCVG**L**	1	-1.267	5E-120	32	45.5	1456
PG-1.13	RG**W**RLCYCRRRFCVCV**R**R	2	-3.08	5E-197	16	60.7	971
PG-1.14	RG**V**RLCYCRRR**L**CVCVGR	2	-2.71	8E-127	8	43.5	348
PG-1.15	RG**V**RLCYCRRRFCVC**L**GR	2	-3.18	5E-151	8	67.2	537
PG-1.16	RGGRLCYCRRRFCVC**L**G**S**	2	-2.17	5E-93	4	63.3	253
PG-1.17	RGGRLCYCRRRFC**F**CVG**L**	2	-3.19	7E-111	32	37.4	1195
PG-1.18	RGGR**I**CYCRRR**Y**C**I**CVGR	3	-2.40	1E-46	16	32.4	518
PG-1.19	RGGRLCYCRR**K**FCVCV**R**C	3	-1.73	1E-28	16	35.5	569
PG-1.20	RGGRL**R**Y**C**H**R**RFCV**S**VGR	3	-1.65	3E-50	4	1.0	4
PG-1.21	RG**V**H**L**CYCRRR**Y**CVCVGR	3	-3.21	0	8	34.4	275
PG-1.22	RGGRLCYCRRR**Y**CVCVG**H**	2	-7.30	7E-08	16	18.1	290
PG-1.23	RG**W**RLCYCRRRFC**I**CVGR	2	-3.80	4E-110	16	86.9	1391
PG-1.24	RG**W**RLCYCRRRFCVCVG**C**	2	-3.69	3E-11	>64	78.3	-
PG-1.25	RG**W**RLCYCRRRFCVCVG**H**	2	-3.65	2E-45	>64	46.4	-
PG-1.26	RG**V**RLCYCRRRFC**I**CVGR	2	-3.58	7E-225	16	52.3	838
PG-1.27	RG**W**RLCYCRRRFCVC**L**GR	2	-3.55	2E-75	64	40.0	2562
PG-1.28	RG**V**R**I**CYCRRRFC**F**CVGR	3	-7.25	3E-08	16	50.0	799
PG-1.29	**R**RGR**I**CYCRR**K**FCVCVGR	3	-4.17	2E-10	16	19.8	316
PG-1.30	RG**W**S**L**CYCRR**K**FCVCVGR	3	-4.13	5E-07	32	65.4	2093
PG-1.31	R**V**V**R**LC**F**CRRRFCVCVGR	3	-3.56	1E-13	16	36.0	576
PG-1.32	**H**G**W**RLCYCRRRFCVCVG**C**	3	-3.55	2E-26	>64	35.5	-
PG-1.33	**L**S**G**RLCYCRRRFCVC**L**R**R**	4	-3.21	3E-80	4	37.6	150
PG-1.34	RG**W**RLG**Y**CRRRFCV**S**I**G**R	4	-3.17	1E-72	2	22.5	45
PG-1.35	R**R**V**H**LCYCRRR**Y**CVCVGR	4	-3.15	4E-08	16	18.3	292
PG-1.36	RGGRLCYCR**G**R**Y**C**I**R**S**GR	5	-2.90	1E-05	>64	0.2	-
PG-1.37	**H**V**R**RLCYCRRRFC**A**CVG**S**	5	-2.43	2E-11	1	2.6	3
PG-1.38	**R**R**G**R**I**CYC**P**L**R**FCVC**L**V**R**	6	-2.07	0.0003	>64	8.4	-
PG-1.39	RGGRLG**Y**CRRRFCVCVGR	1	0.3009	1E-13	>64	6.7	-
PG-1.40	RGGRLC**D**CRRRFCVCVGR	1	0.3402	6E-15	>64	0.3	-
PG-1.41	RGGRLCYCRRRF**G**VCVGR	1	0.1834	2E-05	>64	0.6	-
PG-1.42	RGGRLCYCRRRFC**D**CVGR	1	0.2873	9E-12	>64	0.1	-
PG-1.43	RGGRL**S**Y**C**LRRFCVCVGR	2	0.82	2E-40	>64	12.2	-
PG-1.44	RGGRLC**Y**S**R**RRFCVCVG**S**	2	0.79	5E-23	>64	15.6	-
PG-1.45	RGGRLCYC**H**RRF**F**V**R**VGR	3	1.17	2E-50	32	2.0	63
PG-1.46	RG**R**RLCYC**H**RRFC**D**S**V**GR	4	1.35	0.0004	>64	0.1	-
PG-1.47	G**R**E**R**LCYC**H**RRLCVCVGR	5	1.46	3E-21	>64	2.0	-

L2FC: log_2-fold change in reads, padj: adjusted p value, MIC: minimum inhibitory concentration (median of triplicate reactions), %Hemo: percent hemolysis (mean of triplicate reactions), SS: selectivity sccore (MIC * %Hemo), **bold**: sequence change

Table S2. Machine learning in vitro Protegrin-1 variant data.

			Machine learning predictions			*In vitro* measured values			
Name	**Change in sequence**	**Group**	**MIC ≤ 8**	**%Hemo**	**$Log_{10}SS$**	**MIC**	**%Hemo**	**$Log_{10}SS$**	**SS**
PG-1.0	RGGRLCYCRRRFCVCVGR	-	-	-	-	4	56.4	2.35	226
bsPG-1.1	-----A-----D-T----	1H	100	-14.39	-1.31	>64	-0.17	-	-
bsPG-1.2	-----Q-----G--T---	1R	100	-14.30	-1.22	4	0.34	0.13	1.36
bsPG-1.3	-----Q------DT----	1R	100	-15.37	-1.03	>64	0.03	-	-
bsPG-1.4	-----Q--------AR--	1R	100	-7.68	-0.99	16	0.43	0.84	6.86
bsPG-1.5	-----T-------TA---	1R	100	-12.34	-0.84	16	0.04	-0.15	0.70
bsPG-1.6	-----T-----D--A---	1R	100	-21.81	-0.70	32	0.12	0.57	3.74
bsPG-1.7	-----R------D-S---	1R	100	-13.54	-0.70	16	0.74	1.07	11.77
bsPG-1.8	-----D--------TT--	1R	100	-19.08	-0.69	64	0.17	1.03	10.60
bsPG-1.9	-----R-----GT-----	2H	100	-5.66	-0.65	>64	0.27	-	-
bsPG-1.10	----------N--TT--	2R	100	-14.08	-0.48	>64	-0.06	-	-
bsPG-1.11	------TA------R---	2R	100	-3.35	-0.42	2	0.91	0.26	1.81
bsPG-1.12	-----R------R-S---	2R	100	0.55	-0.40	16	1.53	1.39	24.41
bsPG-1.13	----P--------GT---	2R	100	-0.99	-0.38	>64	-0.01	-	-
bsPG-1.14	-----Q-------AR---	2R	100	-2.91	-0.35	16	0.14	0.35	2.26
bsPG-1.15	-----A--------DT--	2R	100	-10.00	-0.34	64	0.09	0.77	5.93
bsPG-1.16	-T---------D-T----	2H	100	-8.26	-0.33	32	0.05	0.23	1.72
bsPG-1.17	-----T-------AD---	3R	100	-15.55	-0.28	64	0.25	1.20	15.91
bsPG-1.18	-----QA------T---	3R	100	-3.05	-0.25	8	0.72	0.76	5.77
bsPG-1.19	-------R---TA-----	3H	100	-9.62	-0.22	8	0.40	0.51	3.24
bsPG-1.20	----------TP-T-	3R	100	-6.64	-0.21	64	0.39	1.40	24.95
bsPG-1.21	------TA------H---	3R	100	-1.72	-0.20	16	0.64	1.01	10.21
bsPG-1.22	-----E--------NT--	3R	100	-10.52	-0.19	>64	1.28	-	-
bsPG-1.23	------D----T-R----	3H	100	-6.51	-0.16	>64	0.19	-	-
bsPG-1.24	------T------AS---	3R	100	0.51	-0.07	32	0.49	1.19	15.59
bsPG-1.25	----------T--DT--	4R	100	-11.36	-0.07	32	0.16	0.71	5.15
bsPG-1.26	--Q---------N--T--	4H	100	-2.73	-0.04	32	0.40	1.11	12.79
bsPG-1.27	-------G------NQ--	4R	100	-1.77	-0.03	32	2.06	1.82	65.96
bsPG-1.28	-----H-R------R---	4R	100	-9.54	0.03	8	9.72	1.89	77.78
bsPG-1.29	-------S-----KT---	4R	100	1.55	0.10	64	0.56	1.55	35.87
bsPG-1.30	------D------NM---	4R	100	-8.96	0.11	>64	-0.10	-	-
bsPG-1.31	----DI--------T---	4R	100	-9.88	0.12	8	0.66	0.72	5.30
bsPG-1.32	-----S--------AG--	4R	100	0.58	0.46	16	0.32	0.71	5.15

bs: bacterially selective machine learning variant, MIC: minimum inhibitory concentration in µg/ml (median of biological triplicate), %Hemo: percent hemolysis (mean of biological triplicate), SS: selectivity score (MIC * %Hemo), Grouped by predicted $log_{10}SS$ (H: handpicked, R: randomly picked)

Table S3. Plasmids, Strains, and Oligonucleotides.

Plasmids	Source
pMMBEH67_lpp_ompA	(5)
pDM1	(43)
pDM1_mcr-1	(43)
Strains	**Source**
E. coli W3110	Lab Stock
Oligonucleotides	**Sequence**
oJR557 - F dmSLAY library	gtattggtaccagtcaagagcctg
oJR560 - R Protegrin-1 dmSLAY primer	CTG CAG GTC GAC TTA (N1:94020202)(N2:02940202)(N3:02029402) (N4:02020294)(N2)(N2) (N4)(N1)(N2) R(N2)(N1) (N1)(N1)(N2) R(N2)(N1) (N3)(N1)(N1) (N2)(N2)(N4) (N3)(N2)(N3) (N1)(N2)(N3) R(N2)(N1) R(N4)(N1) R(N2)(N1) (N1)(N1)(N3) (N1)(N2)(N3) (N2)(N2)(N2) (N1)(N2)(N2) (N3)(N2)(N3) GGT TCC TCC GAT ACC CGC AG
2x(NR)tether gBlock	ATTGCCGATGGTACACGTCAAGTCAAGAGCCTGCAGCGCCCGCCGCAG AGGCGACTCCTGCTGCTGAAGCTCCAGCTAGCGAAGCGCCTGCAGCAG AAGCTGCCCCAGCGGATGCTGCCGAAGCCCCAGCCGCTGGCATCAGTC AGGAACCTGCTGCACCAGCTGCGGAAGCTACACCAGCAGCGGAGGCAC CAGCGAGTGAAGCACCGGCTGCGGAAGCCGCTCCTGCAGATGCCGCT GAGGCTCCAGCTGCGGGTATCGGAGGAACCCGCGGTGGGCGTCTTTGT TA
F amplicon	aatgATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCT TCCGATCTCTCCAGCTGCGGGTATCGGAGGA
R index 1	CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCAGACG TGTGCTCTTCCGATCTgccaagcttgcatgcctgcaggtcgacTTA
R index 2	CAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAGTTCAGACG TGTGCTCTTCCGATCTgccaagcttgcatgcctgcaggtcgacTTA
R index 3	CAAGCAGAAGACGGCATACGAGATGCCTAAGTGACTGGAGTTCAGACG TGTGCTCTTCCGATCTgccaagcttgcatgcctgcaggtcgacTTA
R index 4	CAAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAGTTCAGACG TGTGCTCTTCCGATCTgccaagcttgcatgcctgcaggtcgacTTA
R index 5	CAAGCAGAAGACGGCATACGAGATCACTGTGTGACTGGAGTTCAGACGT GTGCTCTTCCGATCTgccaagcttgcatgcctgcaggtcgacTTA
R index 6	CAAGCAGAAGACGGCATACGAGATATTGGCGTGACTGGAGTTCAGACG TGTGCTCTTCCGATCTgccaagcttgcatgcctgcaggtcgacTTA
R index 7	CAAGCAGAAGACGGCATACGAGATGATCTGGTGACTGGAGTTCAGACG TGTGCTCTTCCGATCTgccaagcttgcatgcctgcaggtcgacTTA
R index 8	CAAGCAGAAGACGGCATACGAGATTCAAGTGTGACTGGAGTTCAGACGT GTGCTCTTCCGATCTgccaagcttgcatgcctgcaggtcgacTTA
R index 9	CAAGCAGAAGACGGCATACGAGATCTGATCGTGACTGGAGTTCAGACGT GTGCTCTTCCGATCTgccaagcttgcatgcctgcaggtcgacTTA
R index 10	CAAGCAGAAGACGGCATACGAGATAAGCTAGTGACTGGAGTTCAGACGT GTGCTCTTCCGATCTgccaagcttgcatgcctgcaggtcgacTTA
R index 11	CAAGCAGAAGACGGCATACGAGATGTAGCCGTGACTGGAGTTCAGACG TGTGCTCTTCCGATCTgccaagcttgcatgcctgcaggtcgacTTA
R index 12	CAAGCAGAAGACGGCATACGAGATTACAAGGTGACTGGAGTTCAGACGT GTGCTCTTCCGATCTgccaagcttgcatgcctgcaggtcgacTTA

All oligonucleotides were ordered from Integrated DNA technologies (IDT). N1, N2, N3 and N4 are hand mixed %bases (i.e. N1:94020202 = N1 is a mixture of 94% A, 2% C, 2% G, 2% T)

(43) Furniss, R. C. D. *et al.* Breaking antimicrobial resistance by disrupting extracytoplasmic protein folding. *Elife* **11**, (2022).