

A generative artificial intelligence approach for the discovery of antimicrobial peptides against multidrug-resistant bacteria

In the format provided by the
authors and unedited

This file contains:

Supplementary Text

The details of model structure.

Supplementary Tables

Supplementary Table 1 Model metrics of AMPSorter on the test set and external validation set.

Supplementary Table 2 The first amino acid distribution of known AMPs collected.

Supplementary Table 3 F/B values of intestinal flora in different treatment groups.

Supplementary Table 4 Blast results of m_AMP76 in NCBI.

Supplementary Table 5 The species distribution of *non-AMPs*.

Supplementary Table 6 Catalogue of reagents used.

Supplementary Table 7 Time gradient for mobile phase composition.

Supplementary Text

The details of model structure.

The most fundamental model is ProteoGPT pre-trained model based on GPT-2 architecture. And the model structure of AMPGenix is identical that of ProteoGPT. AMPSorter and BioToxiPept are created by adding an additional classification module at the end of ProteoGPT to reduce the dimension to 2. For ProteoGPT, initial parameters from “gpt2” were then fine-tuned with Cross entropy as loss function, epochs=8, batch size = 8 and learning rate = 1.5×10^{-3} . Then AMPGenix was generated by fine-tuning ProteoGPT with epochs=30, batch size = 8, learning rate = 1.5×10^{-3} ; AMPSorter with epochs=20, batch size = 512, learning rate = 1.5×10^{-3} ; BioToxiPept with epochs=20, batch size = 256, learning rate = 1.5×10^{-3} . AdamW optimization algorithm with default parameters were employed to regulate the scale of model parameters to mitigate the risk of overfitting. For individual models, they were fixed and unchanged after the training, and a peptide with prediction probability greater than 0.5 (positive) is considered a candidate AMP in AMPSorter or a poisonous peptide in BioToxiPept. The four models above were built with the Transformer library (version 4.27.1).

● ProteoGPT and AMPGenix

```
GPT2LMHeadModel(  
  (transformer): GPT2Model(  
    (wte): Embedding(50257, 768)  
    (wpe): Embedding(1024, 768)  
    (drop): Dropout(p=0.1, inplace=False)  
    (h): ModuleList(  
      (0): GPT2Block(  
        (ln_1): LayerNorm((768,), eps=1e-05, elementwise_affine=True)  
        (attn): GPT2Attention(  
          (c_attn): Conv1D()  
          (c_proj): Conv1D()  
          (attn_dropout): Dropout(p=0.1, inplace=False)  
          (resid_dropout): Dropout(p=0.1, inplace=False)  
        )  
        (ln_2): LayerNorm((768,), eps=1e-05, elementwise_affine=True)  
        (mlp): GPT2MLP(  
          (c_fc): Conv1D()  
          (c_proj): Conv1D()  
          (act): NewGELUActivation()  
          (dropout): Dropout(p=0.1, inplace=False)  
        )  
      )  
    )  
    (1): GPT2Block(  
      (ln_1): LayerNorm((768,), eps=1e-05, elementwise_affine=True)  
      (attn): GPT2Attention(  
        (c_attn): Conv1D()  
        (c_proj): Conv1D()
```

```

        (attn_dropout): Dropout(p=0.1, inplace=False)
        (resid_dropout): Dropout(p=0.1, inplace=False)
    )
    (ln_2): LayerNorm((768,), eps=1e-05, elementwise_affine=True)
    (mlp): GPT2MLP(
        (c_fc): Conv1D()
        (c_proj): Conv1D()
        (act): NewGELUActivation()
        (dropout): Dropout(p=0.1, inplace=False)
    )
)
(2): GPT2Block(
    (ln_1): LayerNorm((768,), eps=1e-05, elementwise_affine=True)
    (attn): GPT2Attention(
        (c_attn): Conv1D()
        (c_proj): Conv1D()
        (attn_dropout): Dropout(p=0.1, inplace=False)
        (resid_dropout): Dropout(p=0.1, inplace=False)
    )
    (ln_2): LayerNorm((768,), eps=1e-05, elementwise_affine=True)
    (mlp): GPT2MLP(
        (c_fc): Conv1D()
        (c_proj): Conv1D()
        (act): NewGELUActivation()
        (dropout): Dropout(p=0.1, inplace=False)
    )
)
(3): GPT2Block(
    (ln_1): LayerNorm((768,), eps=1e-05, elementwise_affine=True)
    (attn): GPT2Attention(
        (c_attn): Conv1D()
        (c_proj): Conv1D()
        (attn_dropout): Dropout(p=0.1, inplace=False)
        (resid_dropout): Dropout(p=0.1, inplace=False)
    )
    (ln_2): LayerNorm((768,), eps=1e-05, elementwise_affine=True)
    (mlp): GPT2MLP(
        (c_fc): Conv1D()
        (c_proj): Conv1D()
        (act): NewGELUActivation()
        (dropout): Dropout(p=0.1, inplace=False)
    )
)
(4): GPT2Block(

```

```

(ln_1): LayerNorm((768,), eps=1e-05, elementwise_affine=True)
(attn): GPT2Attention(
  (c_attn): Conv1D()
  (c_proj): Conv1D()
  (attn_dropout): Dropout(p=0.1, inplace=False)
  (resid_dropout): Dropout(p=0.1, inplace=False)
)
(ln_2): LayerNorm((768,), eps=1e-05, elementwise_affine=True)
(mlp): GPT2MLP(
  (c_fc): Conv1D()
  (c_proj): Conv1D()
  (act): NewGELUActivation()
  (dropout): Dropout(p=0.1, inplace=False)
)
)
(5): GPT2Block(
  (ln_1): LayerNorm((768,), eps=1e-05, elementwise_affine=True)
  (attn): GPT2Attention(
    (c_attn): Conv1D()
    (c_proj): Conv1D()
    (attn_dropout): Dropout(p=0.1, inplace=False)
    (resid_dropout): Dropout(p=0.1, inplace=False)
  )
  (ln_2): LayerNorm((768,), eps=1e-05, elementwise_affine=True)
  (mlp): GPT2MLP(
    (c_fc): Conv1D()
    (c_proj): Conv1D()
    (act): NewGELUActivation()
    (dropout): Dropout(p=0.1, inplace=False)
  )
)
(6): GPT2Block(
  (ln_1): LayerNorm((768,), eps=1e-05, elementwise_affine=True)
  (attn): GPT2Attention(
    (c_attn): Conv1D()
    (c_proj): Conv1D()
    (attn_dropout): Dropout(p=0.1, inplace=False)
    (resid_dropout): Dropout(p=0.1, inplace=False)
  )
  (ln_2): LayerNorm((768,), eps=1e-05, elementwise_affine=True)
  (mlp): GPT2MLP(
    (c_fc): Conv1D()
    (c_proj): Conv1D()
    (act): NewGELUActivation()

```

```

        (dropout): Dropout(p=0.1, inplace=False)
    )
)
(7): GPT2Block(
  (ln_1): LayerNorm((768,), eps=1e-05, elementwise_affine=True)
  (attn): GPT2Attention(
    (c_attn): Conv1D()
    (c_proj): Conv1D()
    (attn_dropout): Dropout(p=0.1, inplace=False)
    (resid_dropout): Dropout(p=0.1, inplace=False)
  )
  (ln_2): LayerNorm((768,), eps=1e-05, elementwise_affine=True)
  (mlp): GPT2MLP(
    (c_fc): Conv1D()
    (c_proj): Conv1D()
    (act): NewGELUActivation()
    (dropout): Dropout(p=0.1, inplace=False)
  )
)
(8): GPT2Block(
  (ln_1): LayerNorm((768,), eps=1e-05, elementwise_affine=True)
  (attn): GPT2Attention(
    (c_attn): Conv1D()
    (c_proj): Conv1D()
    (attn_dropout): Dropout(p=0.1, inplace=False)
    (resid_dropout): Dropout(p=0.1, inplace=False)
  )
  (ln_2): LayerNorm((768,), eps=1e-05, elementwise_affine=True)
  (mlp): GPT2MLP(
    (c_fc): Conv1D()
    (c_proj): Conv1D()
    (act): NewGELUActivation()
    (dropout): Dropout(p=0.1, inplace=False)
  )
)
(9): GPT2Block(
  (ln_1): LayerNorm((768,), eps=1e-05, elementwise_affine=True)
  (attn): GPT2Attention(
    (c_attn): Conv1D()
    (c_proj): Conv1D()
    (attn_dropout): Dropout(p=0.1, inplace=False)
    (resid_dropout): Dropout(p=0.1, inplace=False)
  )
  (ln_2): LayerNorm((768,), eps=1e-05, elementwise_affine=True)

```

```

        (mlp): GPT2MLP(
          (c_fc): Conv1D()
          (c_proj): Conv1D()
          (act): NewGELUActivation()
          (dropout): Dropout(p=0.1, inplace=False)
        )
      )
    (10): GPT2Block(
      (ln_1): LayerNorm((768,), eps=1e-05, elementwise_affine=True)
      (attn): GPT2Attention(
        (c_attn): Conv1D()
        (c_proj): Conv1D()
        (attn_dropout): Dropout(p=0.1, inplace=False)
        (resid_dropout): Dropout(p=0.1, inplace=False)
      )
      (ln_2): LayerNorm((768,), eps=1e-05, elementwise_affine=True)
      (mlp): GPT2MLP(
        (c_fc): Conv1D()
        (c_proj): Conv1D()
        (act): NewGELUActivation()
        (dropout): Dropout(p=0.1, inplace=False)
      )
    )
  )
  (11): GPT2Block(
    (ln_1): LayerNorm((768,), eps=1e-05, elementwise_affine=True)
    (attn): GPT2Attention(
      (c_attn): Conv1D()
      (c_proj): Conv1D()
      (attn_dropout): Dropout(p=0.1, inplace=False)
      (resid_dropout): Dropout(p=0.1, inplace=False)
    )
    (ln_2): LayerNorm((768,), eps=1e-05, elementwise_affine=True)
    (mlp): GPT2MLP(
      (c_fc): Conv1D()
      (c_proj): Conv1D()
      (act): NewGELUActivation()
      (dropout): Dropout(p=0.1, inplace=False)
    )
  )
)
)
)
  (ln_f): LayerNorm((768,), eps=1e-05, elementwise_affine=True)
)
  (lm_head): Linear(in_features=768, out_features=50257, bias=False)
)

```

- **AMPSorter and BioToxiPept**

```
GPT2ForSequenceClassification(
  (transformer): GPT2Model(
    (wte): Embedding(50257, 768)
    (wpe): Embedding(1024, 768)
    (drop): Dropout(p=0.1, inplace=False)
    (h): ModuleList(
      (0-11): 12 x GPT2Block(
        (ln_1): LayerNorm((768,), eps=1e-05, elementwise_affine=True)
        (attn): GPT2Attention(
          (c_attn): Conv1D()
          (c_proj): Conv1D()
          (attn_dropout): Dropout(p=0.1, inplace=False)
          (resid_dropout): Dropout(p=0.1, inplace=False)
        )
        (ln_2): LayerNorm((768,), eps=1e-05, elementwise_affine=True)
        (mlp): GPT2MLP(
          (c_fc): Conv1D()
          (c_proj): Conv1D()
          (act): NewGELUActivation()
          (dropout): Dropout(p=0.1, inplace=False)
        )
      )
    )
    (ln_f): LayerNorm((768,), eps=1e-05, elementwise_affine=True)
  )
  (score): Linear(in_features=768, out_features=2, bias=False)
)
```

Supplementary Table

Supplementary Table 1 Model metrics of AMPSorter on the test set and external validation set.

Data set	TP	FP	TN	FN	Precision (%)	F1-score (%)	MCC (%)
Test set							
All sequences	3026	118	3202	177	96.25	95.35	90.96
Sequences with UAAs	108	4	81	1	96.43	97.74	94.79
External validation set							
All sequences	344	22	348	26	93.99	93.48	87.03

Supplementary Table 2 The first amino acid distribution of known AMPs collected.

Amino acid	Frequency	Amino acid	Frequency
G	3092	H	166
K	1971	r	33
F	1630	l	30
R	1576	k	28
A	1046	a	16
L	1038	f	11
I	902	O	8
V	759	w	5
S	651	Z	5
W	457	c	5
T	387	v	5
D	337	y	5
M	333	B	4
C	319	i	3
E	272	t	1
Y	258	J	1
N	246	e	1
Q	244	n	1
P	215	g	1

Supplementary Table 6 Catalogue of reagents used.

Reagents	Company names	Catalog numbers
dimethylsulfoxide (DMSO)	Mcklin	D806645
cyclophosphamide	Mcklin	C804838
3,3'-dipropylthiadicarbocyanine iodide [DiSC ₃ -(5)]	Mcklin	D864435
vancomycin	Mcklin	V820413
HEPES	Solarbio	H8090
polymyxin B	Solarbio	P8350
MTT Cell Proliferation and Cytotoxicity Assay Kit	Solarbio	M8180
Fresh collected sheep red blood cells	Solarbio	TX0030
propidium iodide (PI)	Yuanye	R20287
PBS	cytiva	SH30850.03
TritonX-100	Sigma-Aldrich	93443
DMEM/High glucose medium	Shanghai Zhong Qiao Xin Zhou Biotechnology Co., Ltd.	ZQ-100
penicillin-streptomycin mixed solution	Shanghai Zhong Qiao Xin Zhou Biotechnology Co., Ltd.	0513
0.25% pancreatic enzymes	Shanghai Zhong Qiao Xin Zhou Biotechnology Co., Ltd.	0103
fetal bovine serum (FBS)	Shanghai Zhong Qiao Xin Zhou Biotechnology Co., Ltd.	AU0600
D-Glucopyranose	Hushi	63005518
potassium chloride (KCl)	Hushi	10016318
mannitol salt agar (MSA)	Hopebiol	HB8562
macconkey agar (MCA)	Hopebiol	10016318
mueller hinton broth (MHB)	Hopebiol	HB6231
tryptone soybean broth (TSB)	Hopebiol	HB4114
tryptone soybean agar (TSA)	Hopebiol	HB4138
mueller hinton agar (MHA)	Hopebiol	HB6232
human embryonic kidney (HEK293) cells	FUNHENG BIOLOGY	FH0242
EasyPure Stool Genomic DNA kit	TransGene	EE301-01

Supplementary Table 7 Time gradient for mobile phase composition.

Time	Flow Rate (mL min⁻¹)	A%	B%
Initial	0.35	90	10
2	0.35	90	10
5	0.35	10	90
10	0.35	10	90
11	0.35	90	10
15	0.35	90	10