Accelerating Antimicrobial Peptide Discovery with Latent Sequence-Structure Model

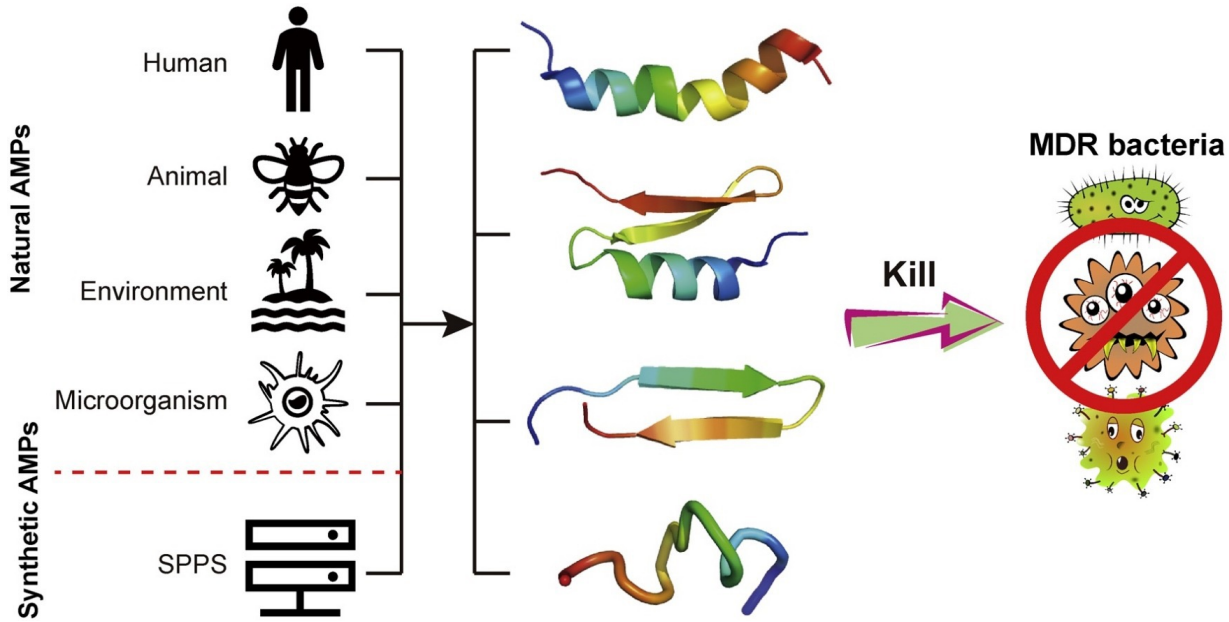
Danqing Wang, Zeyu Wen, Fei Ye, Hao Zhou, Lei Li

ByteDance AI Lab
字节跳动人工智能实验室

AIR 清華大學 智能产业研究院
Institute for AI Industry Research, Tsinghua University

What is Antimicrobial Peptide ?

❖ Peptide: short protein

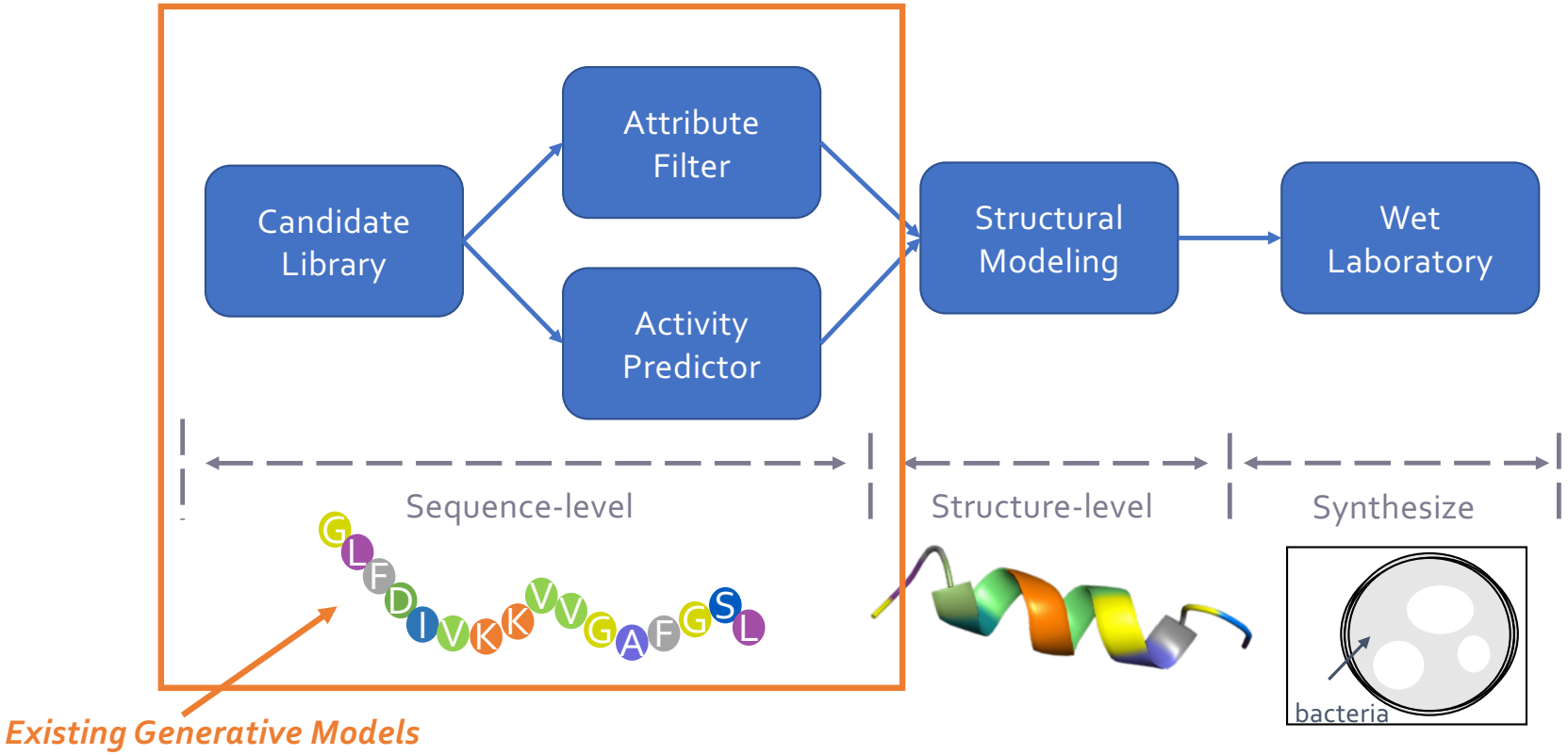


How can AMP kill bacteria?

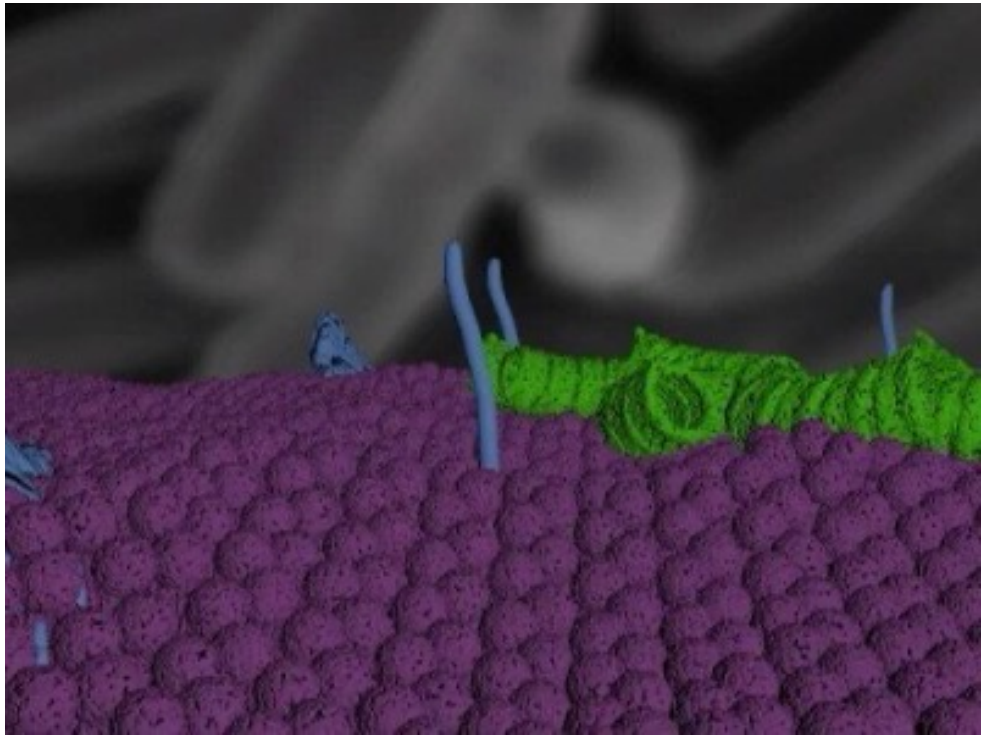
So, we now know that

- Peptide is short protein (< 50 amino acid)
- Antimicrobial Peptide: kill bacteria
  - for example, insert into the bacteria membrane and destroy it
- The main challenge:
  - The unknow mechanism
  - The cost to discover new AMP

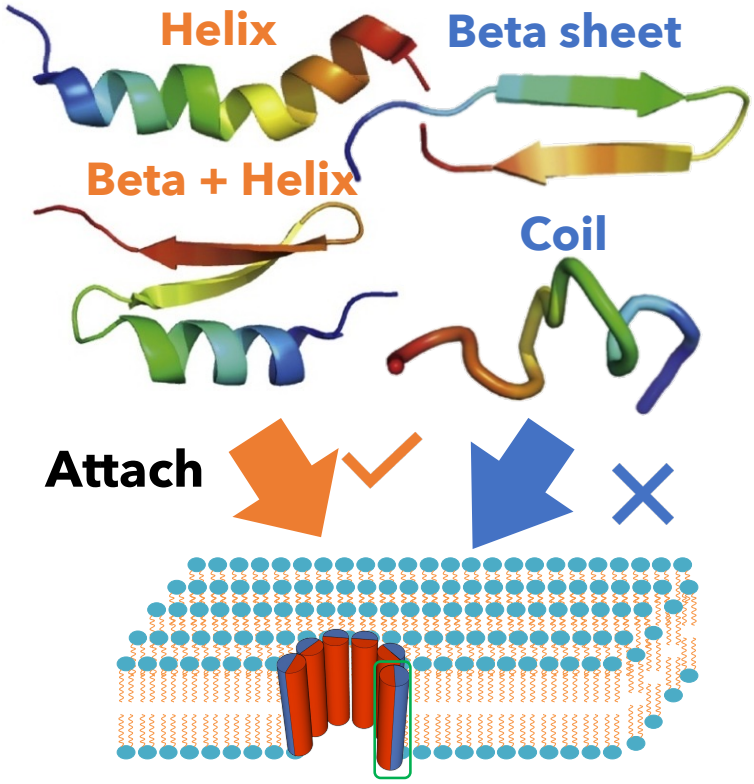
Currently, the AMP discovery is usually



Can we ignore the structure ?



Structure plays an important role in biological functionality



Without right structure

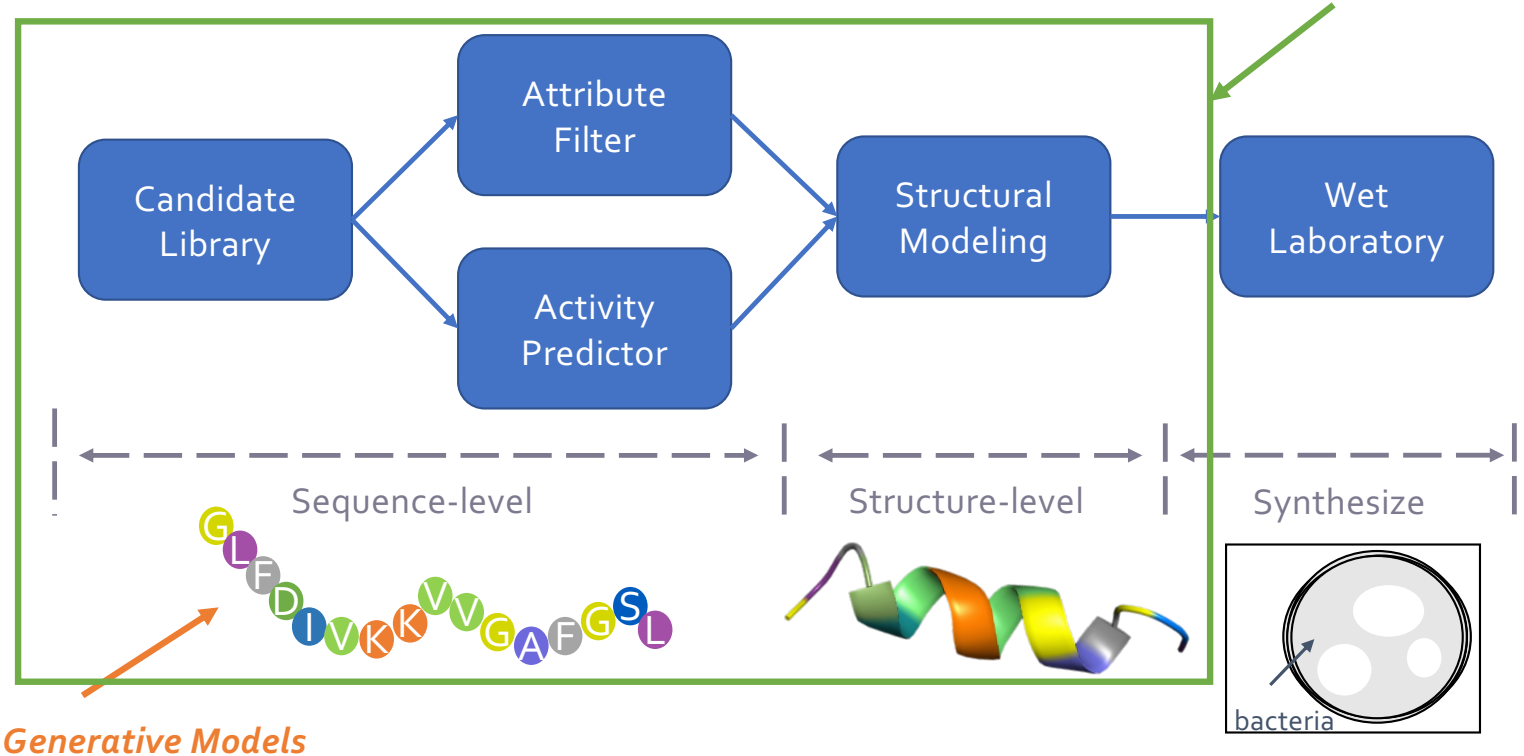
Cannot insert and destroy

Fold into appropriate 3D structure

Insert and maintain hole

From *Sequence-then-Structure* to *Sequence-Structure*

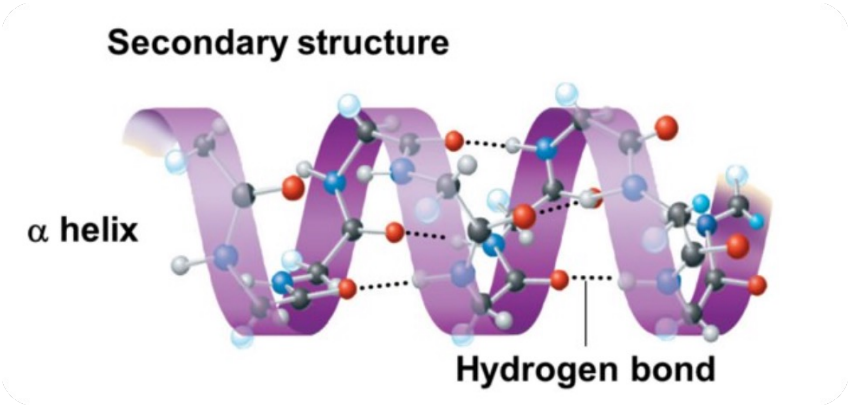
Latent Sequence-Structure Model for AMP (LSSAMP)



Existing Generative Models

Secondary Structure

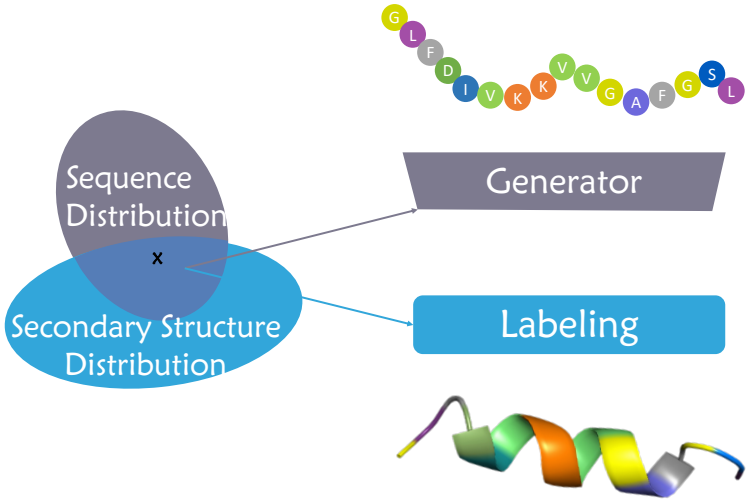
- 3D structure is complex: relationship between residue
  - relative distance, direction, dihedral angle, …
- Divide into local segments
  - annotate each residue with a label
  - indicate the structural element it locates



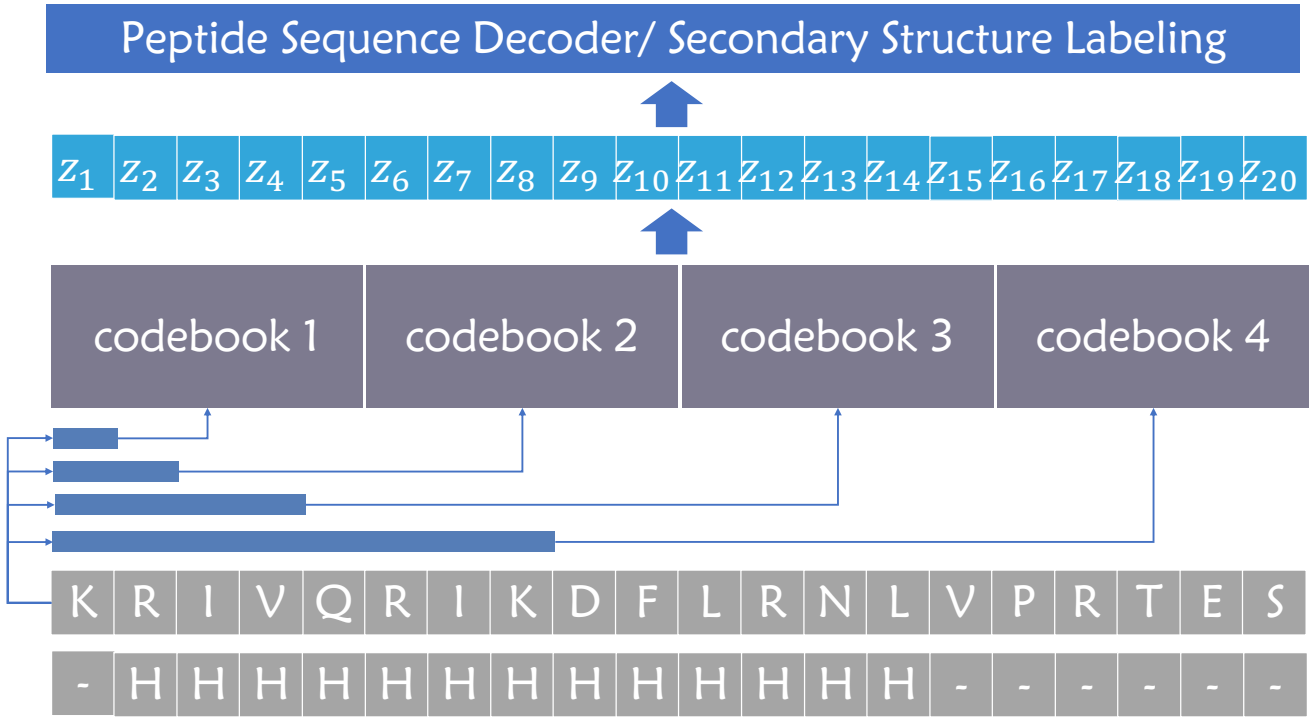
---HHHHHHHHHH---

The problem is …

- How to generate ideal peptide sequences with ideal secondary structures simultaneously
  - sequence: amino acid / residue, 20 vocabulary size
  - secondary structure: labels, 8 category
- A distribution for sequence
- A distribution for secondary structure

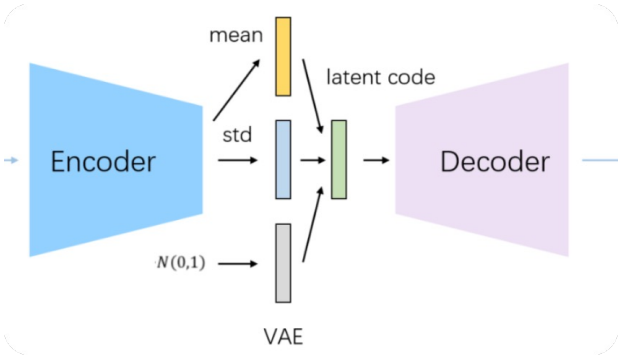


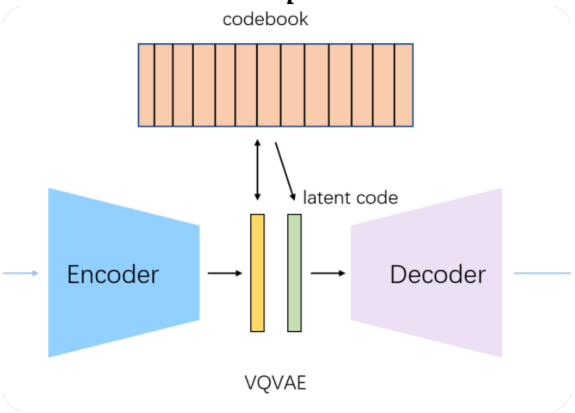
LSSAMP: overview



LSSAMP: latent variable per position

- fine-grained control
  - model the residue & ss distribution on this position
- intractable to sum over the whole sequence
  - use VQ-VAE instead of VAE
  - a continuous z -> look up a discrete representation z_q from the codebook





LSSAMP: multi-scale features

- Sequence and structure have various feature scale
  - sequence pattern: relative short, e.g. 1~4
  - structure motif: much longer, e.g. at least 4 for alpha-helix
- Different feature selectors
  - with different codebooks

Codebook	PPL ↓	Loss ↓	AA Acc.↑	SS Acc.↑
[1]	19.04 ± 2.84	2.94 ± 0.14	65.49 ± 3.49	83.41 ± 2.34
[1, 2]	3.84 ± 0.09	1.35 ± 0.02	99.40 ± 0.45	85.39 ± 0.26
[1, 2, 4]	3.32 ± 0.03	1.20 ± 0.01	**100.00** ± 0.00	85.95 ± 0.42
[1, 2, 4, 8]	**3.24** ± 0.16	**1.17** ± 0.05	99.79 ± 0.20	**87.20** ± 0.62

LSSAMP: training phase

- Main challenge: no enough AMP data
  - only 3k+ known AMPs !
  - structure data is very limited !
- AMP -> special peptide -> short protein
  - from protein database (Uniprot): limit length to 100
    - ✓ D_r: 57k -> pretrain for sequence reconstruction
  - from alphafold: predict secondary structure from protein sequence
    - ✓ D_s: 46k -> further pretrain for structure labeling
  - from AMP dataset (APD):
    - ✓ D_{AMP}: 3222 (positive)
    - ✓ Decoy: 2021 (negative)

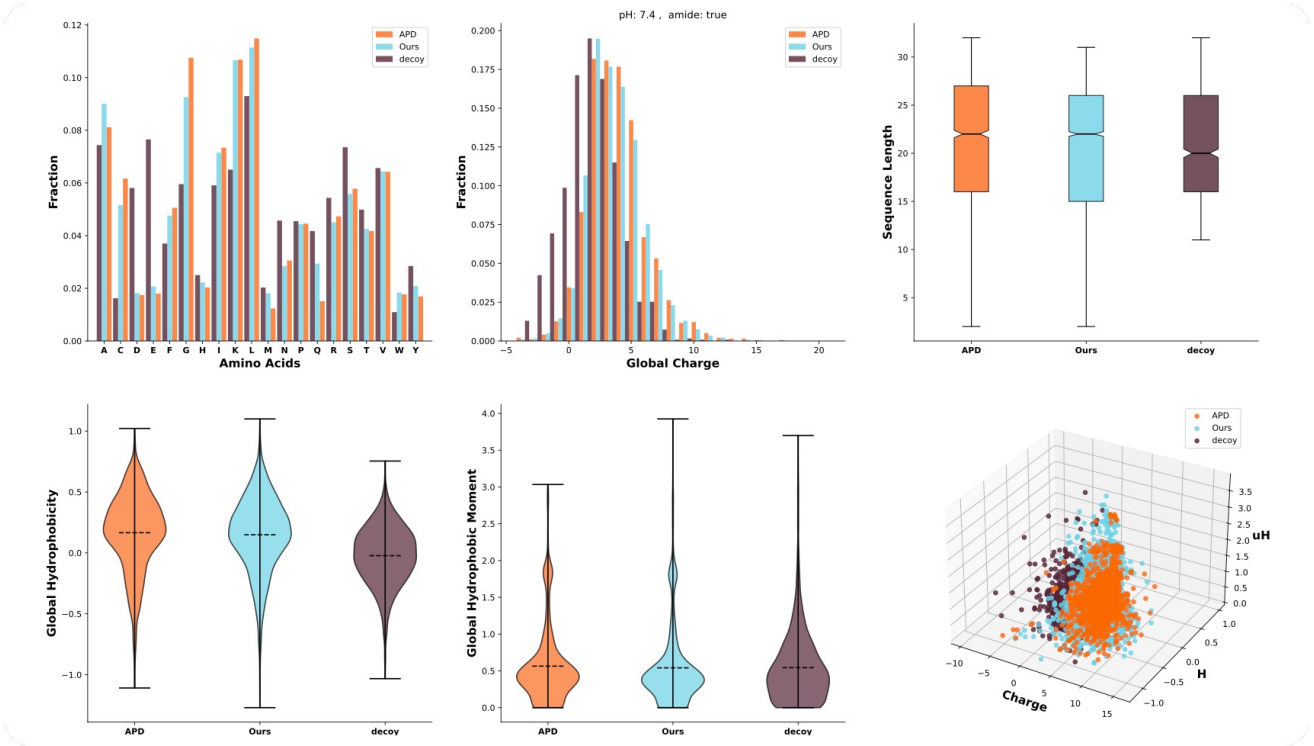
How to evaluate the generated peptide?

- Computational metric
  - charge, hydrophobicity, hydrophobic moment (amphipath)
- Public classifiers
  - score the probability of AMP
- Wet laboratory experiment
  - minimal inhibitory concentration (MIC)

Performance: outperform on combination of three attributes

	Uniq	C	H	uH	Combination
APD	3222	68.75%	27.96%	4.72%	6.15%
Decoy	2020	21.83%	8.81%	1.98%	0.10%
Random $p = 0.1$	4978	65.86% ± 0.19%	**26.80**% ± 0.23%	23.10% ± 0.58%	4.38% ± 0.16%
Random $p = 0.2$	5000	62.13% ± 0.39%	24.87% ± 0.29%	20.79% ± 0.76%	2.47% ± 0.17%
VAE	4988	38.00% ± 0.36%	21.07% ± 0.58%	12.43% ± 0.66%	0.34% ± 0.11%
AMP-GAN	4976	**87.66**% ± 0.45%	17.31% ± 0.74%	23.45% ± 0.73%	1.92% ± 0.05%
PepCVAE	1346	15.61% ± 0.06%	14.54% ± 0.55%	11.65% ± 0.23%	2.75% ± 0.25%
MLPeptide	4486	77.95% ± 0.72%	8.11% ± 0.27%	32.91% ± 0.60%	2.90% ± 0.16%
LSSAMP	4876	81.88% ± 0.31%	25.06% ± 0.45%	**37.10**% ± 0.33%	**6.26**% ± 0.07%
LSSAMP w/o cond	4903	82.04% ± 0.42%	21.32% ± 0.34%	30.51% ± 0.51%	4.46% ± 0.20%

Performance: generation has the similar distribution of existing AMP



Performance: outperform on the average score of 7 classifiers

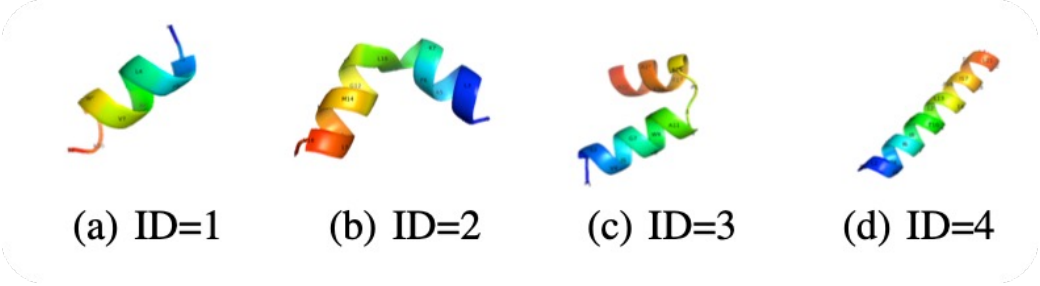
	SVM	RF	DA	Scanner	AMPMIC	IAMPE	amPEP	Average
APD	87.78%	91.24%	86.24%	94.66%	98.42%	97.83%	91.50%	92.52%
Decoy	17.43%	13.71%	16.04%	0.25%	18.07%	23.53%	52.92%	20.28%
Random $p = 0.1$	86.06%	86.12%	84.01%	93.23%	79.14%	95.60%	91.74%	87.99%
Random $p = 0.2$	76.66%	76.64%	74.83%	86.95%	68.57%	91.14%	87.89%	80.38%
VAE (Dean and Walper 2020)	24.90%	15.30%	13.83%	15.12%	15.25%	40.31%	24.30%	21.29%
AMP-GAN (Van Oort et al. 2021)	78.62%	87.29%	83.82%	82.17%	89.58%	93.88%	80.52%	85.13%
PepCVAE (Das et al. 2018)	82.84%	85.96%	93.33%	85.44%	**98.44%**	**98.14%**	80.77%	89.27%
MLPeptide (Capecchi et al. 2021)	90.43%	92.55%	93.08%	**93.72%**	96.34%	97.05%	91.37%	93.51%
LSSAMP	**92.03%**	**92.60%**	**93.45%**	91.52%	95.84%	96.64%	**93.23%**	**93.62%**
LSSAMP w/o cond	78.98%	80.24%	80.01%	86.73%	83.81%	93.80%	85.32%	84.13%

Performance: real AMPs found!

- 2/21 have been verified with high antimicrobic activity (<128)

No	Sequence	Activity (ug/mL) ↓ A. Baumannii	P. aeruginose	E. coli	Sequence identity ↓	Hemolysis/Toxicity ↓
P1	GAFGNFLKNVAKKAGIYLLSIAQCKLFGTP	16-32	/	32-64	83.30%	Low
P2	FIGFLFKLAKKIIPSLFQTKTE	8	32	/	75.00%	Low

- Our generated peptides have ideal alpha-helix structures



(a) ID=1 (b) ID=2 (c) ID=3 (d) ID=4

Sum up

- AMP is promising treatment to replace antibiotic
  - challenge: unknown mechanism & costly discovery process
- Accelerate the discovery by creating more effective candidates
  - current generative works only focus on sequence-level
  - still need to check the structure
- Model the sequence-structure distribution
  - fine-grained control on position
  - multi-scale features
- Evaluate from different aspects to verify effectiveness