AMPCLGPT

With the beginning of a new school term, I've been incredibly busy and haven't had much hands-on time experimenting with transformers or eagerly checking the status of my model's loss function. Instead, I've shifted toward a more passive form of learning, one that I believe has proven incredibly useful for the future of this project. While reviewing several research papers, I discovered what I consider the holy grail: *"Harnessing Generative Pre-trained Transformer for Antimicrobial Peptide Generation and MIC Prediction with Contrastive Learning"* from Tsinghua University.

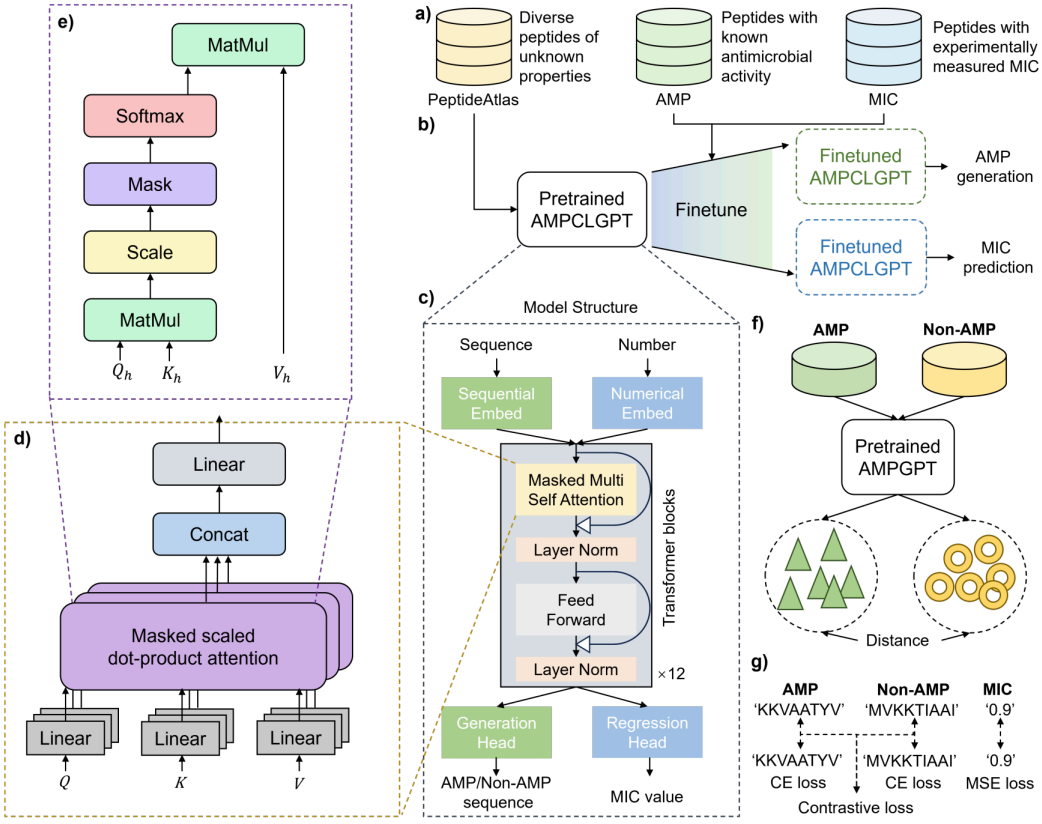

Fig. 1. The overall framework of AMPCLGPT.

Of all the papers I've read, this particular model aligns perfectly with my ambition for AMPForge. It employs an almost identical transformer structure, complete with sequential and numerical embeddings, and utilizes multi-headed self-attention in the same way my model does.

Contrastive Learning

What stood out to me most about this model is that it's trained on a wide range of peptides, not just AMP sequences. The beauty of this approach is that I can train on confirmed AMP and non-AMP sequences using contrastive learning, which helps highlight differences between them in latent space. This is achieved through an additional feature in the model called the generation head. Here, cross-entropy (CE) loss is used to separate data into positive and negative samples. Adopting this strategy will be essential for me to push my own model toward frontier performance.

MIC Values and Evaluation

Throughout the various iterations of my model, I've been able to immerse myself in deep learning and generate novel data that I'm genuinely proud of. However, these models lack practical utility without a way to empirically evaluate the significance of the sequences. AMPCLGPT addresses this issue through its ability to predict MIC values using a regression head.

MIC, or *Minimal Inhibitory Concentration*, refers to the lowest concentration of an AMP that completely inhibits the visible growth of a microorganism. Incorporating this ability will massively improve my current models, enabling them to generate novel data that is not only theoretically interesting but also practically valuable in real-world applications.