**Related Work**

# Design Sequences for Specific Properties

**Peptide 🡪 HydrAMP [Szymczak et al., 2023].**

Peptides, characterized as short chains of amino acids with lengths varying from 2 to 50 units, exhibit a wide range of biological activities. Rather than simply screening an existing library, peptides with desired properties can be designed through generative models. Szymczak et al. (2023) introduces HydrAMP, a conditional variational autoencoder (cVAE) designed to generate antimicrobial peptides (AMPs) with specific properties. HydrAMP is trained on a curated dataset of peptide sequences, employing both discrete and continuous labels to indicate antimicrobial activity and its strength via minimal inhibitory concentration (MIC) values, respectively. The model utilizes a regularization strategy, combining latent reconstruction, KL divergence, and Jacobian disentanglement, to achieve a structured latent space that enhances the interpretability of generated peptides. HydrAMP operates in three modes: reconstruction, analogue generation, and unconstrained generation, allowing for the generation of analogues from both AMPs and non-AMPs, as well as unconstrained generation of novel peptide sequences.

## Differences:

* The length of sequences and size of dataset.
* Properties in our case is treated as distribution but here both discrete and continuous labels were used.

## Similarities:

* Disentangled latent space
* Design biological sequences with desired properties

**De Novo Generation 🡪 ProGen [Madani et al., 2023]**

Madani et al., 2023 designs a decoder-only transformer for autoregressive generation and incorporates control tags (e.g., protein family, biological process, and molecular function), which are metadata prepended to the protein during training, to enable controllable protein design.

ProGen, developed by Madani et al. (2023), is a 1.2-billion-parameter transformer-based language model trained on a dataset comprising 280 million non-redundant protein sequences from over 19,000 Pfam families. The dataset was enriched with metadata control tags, including 1,100 keyword tags related to cellular components, biological processes, and molecular functions, as well as 100,000 taxonomic tags from the NCBI taxonomy. These control tags served as labels to guide the model's generation of protein sequences with specific properties. ProGen's architecture features a decoder-only transformer structure optimized through unsupervised learning to predict the next amino acid in a sequence without relying on explicit structural information or coevolutionary assumptions. The model was further fine-tuned with curated sequences and control tags for specific protein families, enhancing its ability to generate functional protein sequences with predictable functions across diverse protein families.

# Conditional Generation or Regularize based on Complex Labels

# Geometry Preserving