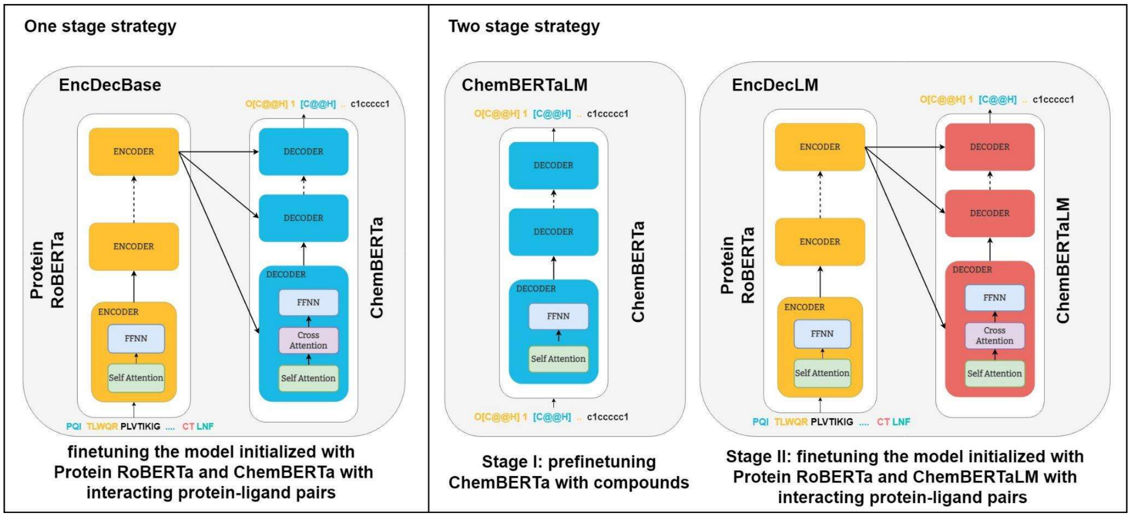
Exploiting Pretrained Biochemical Language Models for Targeted Drug Design

Motivation: The development of novel compounds targeting proteins of interest is one of the most important tasks in the pharmaceutical industry. Deep generative models have been applied to targeted molecular design and have shown promising results. Recently, target-specific molecule generation has been viewed as a translation between the protein language and the chemical language [1]. However, such a model is limited by the availability of interacting protein–ligand pairs. On the other hand, large amounts of unlabelled protein sequences and chemical compounds are available and have been used to train language models that learn useful representations [2, 3, 4]. In this study, we propose exploiting pretrained biochemical language models to initialize (i.e. warm start) targeted molecule generation models. We investigate two warm start strategies: (i) a one-stage strategy where the initialized model is trained on targeted molecule generation and (ii) a two-stage strategy containing a pre-finetuning on molecular generation followed by target-specific training. An overview of these strategies is presented in Figure 1. We also compare two decoding strategies to generate compounds: beam search and sampling.

Results: The results show that the warm-started models perform better than a baseline model trained from scratch. The two proposed warm-start strategies achieve similar results to each other with respect to widely used metrics from benchmarks. However, docking evaluation of the generated compounds for a number of novel proteins suggests that the one-stage strategy generalizes better than the two-stage strategy. Additionally, we observe that beam search outperforms sampling in both docking evaluation and benchmark metrics for assessing compound quality.

Availability and implementation: The source code is available at https://github.com/boun-tabi/biochemical-lms-for-drug-design and the materials (i.e., data, models, and outputs) are archived in Zenodo at https://doi.org/10.5281/zenodo.6832145. The application that allows users to instantly generate molecules with the trained models is available at https://huggingface.co/spaces/gokceuludogan/WarmMolGen.

Figure 1: Warm start strategies exploiting the pretrained protein language model Protein RoBERTa [3] and the chemical language model ChemBERTa [4].



finetuning the model initialized with Protein RoBERTa and ChemBERTa with interacting protein-ligand pairs

Stage I: prefinetuning ChemBERTa with compounds

Stage II: finetuning the model initialized with Protein RoBERTa and ChemBERTaLM with interacting protein-ligand pairs

References:

1. Grechishnikova, D. (2021) Transformer neural network for protein-specific de novo drug generation as a machine translation problem. Sci. Rep., 11, 1–13.
2. Rives,A. et al. (2021) Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc. Natl. Acad. Sci. USA, 118, e2016239118
3. Filipavicius, M. et al. (2020) Pre-training protein language models with label-agnostic binding pairs enhances performance in downstream tasks. arXiv, preprint arXiv:2012.03084.
4. Chithrananda,S. et al. (2020) Chemberta: large-scale self-supervised pretraining for molecular property prediction. arXiv, preprint arXiv:2010.09885.