Artificial Intelligence-based Novel Techniques for Accelerating Drug Discovery

Thesis submitted in partial fulfillment
for the award of the degree of

Doctor of Philosophy

by

Vishakha Singh

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY
(BANARAS HINDU UNIVERSITY)
VARANASI - 221005

Roll No. 20071505 *Year 2024*

Chapter 7

Conclusion and Future Directions

7.1 Conclusion

This thesis comprises AI-based frameworks that enable efficient sequence modeling and optimization. The focus has been on using state-of-the-art AI algorithms for modeling the existing therapeutic peptide sequences in a way that allows for the design and optimization novel ones. Various deep and machine learning algorithms, like biLSTM, TCN, BERT, XGBoost, etc., have been used, often in conjunction with several nature-inspired multi-objective optimization techniques like GSA and NSGA-II. Some common techniques, like explainable AI, stacked ensemble learning, transfer learning, and continual learning, have also been used while building deep learning-based models. The methodologies that are used to build the proposed frameworks help tackle a multitude of computational challenges, like (i) decreasing the model training time, (ii) reducing the size of the model without compromising its performance to deploy it on resource-constrained devices, (iii) using continual learning to update the models periodically to prevent their obsolescence, (iv) using explainable AI and population-based MOO techniques for searching for optimal therapeutic sequences with desirable characteristics that enable their therapeutic action (since the classification of sequences as therapeutic or non-therapeutic is not sufficient for their quick experimental validation

and subsequent production as a potential medication), and (v) deploying the models as web applications online so that the scientific community may be able to screen and optimize the potential therapeutic peptides for further synthesis, experimental validation, and production.

The major findings of the overall research are as follows: (i) Since sequence models based on biLSTM take a lot of time and resources for training, stacking them with classic ML models can greatly reduce their training time; (ii) the web applications based on deep learning algorithms like TCNs can be made memory-efficient (so as to be deployed on resource-constrained devices) using depthwise-separable convolutions without compromising their performance; (iii) the deep learning techniques, especially the large language models like BERT, are better than the traditional sequence modeling techniques, including the classical machine learning algorithms, which later in conjunction with XAI and MOO techniques can help in optimizing the existing sequences to amplify their therapeutic action; (iv) the proposed frameworks, techniques, and methodologies can be easily extended to model any type of biological sequence with little modification and fine-tuning.

7.2 Future Directions

In the future, various potential avenues can be explored concerning the research presented in this thesis. A few of the promising ones are as follows:

1. Novel therapeutic sequences can be generated using autoencoders like variational autoencoders (VAEs) in conjunction with various MOO algorithms.
2. The concept of federated learning can be used to periodically re-train the model which has been deployed online.
3. A tool can be built for finding the toxicity and potential side effects of the discovered or designed therapeutic peptides given by the proposed frameworks.

4. State-of-the-art techniques such as reinforcement learning can be used for designing novel therapeutic molecules from scratch.