Chapter 1

Introduction

1.1 Problem definition

The need for novel medicines has become increasingly urgent due to the failure and insufficiency of conventional form of medication. Also, some novel sources of medicines have been found that are perceived to have fewer side effects than their conventional counterparts. The utilization of artificial intelligence (AI) has gained considerable significance within the domain of drug discovery due to the high cost and complexity associated with this process. The addition of AI to the drug discovery process has the potential to bring about a significant transformation within the pharmaceutical industry. This integration can facilitate the expedited, streamlined, and cost-effective creation of novel medicines for various medical conditions in the following ways.

1. AI possesses the capability to analyze extensive biological datasets, resulting in the identification of patterns in existing drug molecules to enable the prediction of novel drug candidates. This aids in the identification of medicinal compounds that possess the highest likelihood of achieving success.
2. AI has the potential to optimize clinical trial designs through the determination of suitable patient populations, optimal dosages of the given drugs, and patient's responses to drug administration.

3. AI can identify potential applications of already-approved medications for treating other diseases, reducing both time and cost incurred during traditional drug development.

Therapeutic peptides are short amino acid (AA) sequences with good potential to treat various disorders. They have garnered considerable traction in the medical field owing to their high target specificity, minimal toxicity, and comparatively fewer side effects in comparison to conventional medications. Rapid progress in peptide synthesis, modification, and structural optimization has significantly enabled the creation of stable and potent peptide-based treatments. Some of the growing avenues for peptide-based therapeutics are as follows.

1. **Antimicrobial Peptides**: These peptides can kill or inhibit the growth of harmful microbes (bacteria, viruses, and fungi). They can protect against diseases caused by multi-drug-resistant (MDR) pathogens, which have developed high levels of antimicrobial resistance (AMR) against existing medicines.
2. **Neurological peptides**: They play key roles in the operation and control of the nervous system. Furthermore, these can be used to treat various neurological illnesses, such as Alzheimer's disease (AD) and Parkinson's disease (PD).
3. **The blood-brain barrier-penetrating peptides**: They can cross the blood-brain barrier, which is not penetrable by most of the small molecules. So, these peptides can be used as drug delivery vehicles for delivering therapeutic molecules to the brain, thus enabling the treatment of a range of neurological illnesses and brain-related ailments.

1.2 Existing works

The world faces numerous challenges with a "post-antibiotic future" clearly in sight. Also, the rise in cases of neurological diseases is being witnessed with few or no solutions in hand. The drive to discover newer and safer medicines has received acceleration since

the advent of AI. Thus, numerous *in silico* tools have been developed to complement the efforts of biomedical scientists in discovering novel medicines, like the ones derived from therapeutic peptides.

The identification of novel antimicrobial peptides (AMPs) accelerated in response to the current SARS-CoV-2 pandemic [1]. As a result, various tools and techniques that employ machine and deep learning (ML and DL) to classify and discover new therapeutic AMPs have been proposed. They can be categorized into four main groups: anti-bacterial peptides (ABPs), anti-fungal peptides (AFPs), anti-parasitic peptides (APPs), and anti-viral peptides (AVPs). Several classifiers have been developed to identify these peptides, including AntiBP [2], Deep-ABPpred [3], iAMPpred [4], AniAMPpred [5], AMPer [6], CAMP [7], AVPpred [8], AVPIden [9], AmPEP [10], ClassAMP [11], AntiVPP 1.0 [12], ENNAVIA [13], FIRM-AVP model [14], PreTP-Stack [15], etc. Most of these works use handcrafted features and classic machine learning (ML) algorithms, like random forests (RFs), gradient boosting (GB), support vector machines (SVM), logistic regression (LR), etc.

Furthermore, another popular area in therapeutic drug discovery is the design and development of neurological medicines using neurological peptides (NPs), along with a good mechanism for their delivery to the brain across the blood-brain barrier (BBB) in the form of BBB penetrating peptides (B3P2s). In several studies, models like NeuroPIpred [16], NeuroPred-FRL [17], NeuroPred-PLM [18], B3Ppredict [19], BBPpred [20], and B3Pred [21] have been created in this regard.

The shortcomings of the aforementioned tools are as follows.

1. Most of the models are based on ML algorithms, which are known to witness a significant drop in performance with an increase in the size of the dataset.
2. Many models are not available online as web or mobile applications to aid in screening novel molecules.
3. Even when the web or mobile app has been made publicly available, the model

at the back-end is not built to adapt to or learn on new data points. In other words, the models are static, and their performance would drop over the years with the development of newer models that are trained on larger amounts of data compared to them.

4. Some state-of-the-art deep learning (DL)-based tools have been proposed, but they have a greater number of trainable parameters whose training takes a lot of computing power. The final result is a bulky model, which results in a slow and/or bulky mobile or web application.
5. The existing models and tools are focused on only identifying therapeutic peptides with the help of a classifier. However, no thought was given to changing the structure of the peptide to improve its desirable properties and make it more efficient as a drug.
6. The existing state-of-the-art models are not explainable. A major application of explainability could be to identify the desirable characteristics of therapeutic peptides that can be optimized to increase their efficacy by structural modification.

1.3 Motivation

The work done in this thesis has been primarily focused on antibacterial peptides (ABPs), antiviral peptides (AVPs), neurological peptides (NPs), and blood-brain barrier penetrating peptides (B3P2s). Using AI to analyze complex hidden patterns in huge datasets comprising medicinal peptides and optimizing them based on those patterns (as shown in stages 2 and 3 of Figure 1.1) might lead to better preliminary screening of novel peptide-based drug molecules, which is otherwise costly and tedious. In this dissertation, numerous tools and techniques have been proposed to overcome the computational challenges mentioned before. In summary, the main objectives that were tackled and achieved are as follows:

1. Building a tool that identifies therapeutic peptides and takes less time to get

trained despite being based on sequential deep learning algorithms like recurrent neural networks (RNNs).

2. Building a light and efficient model to identify therapeutic peptides to make it easily deployable as an app on resource-constrained devices.
3. Identifying therapeutic peptides using a DL-based model with the ability to re-train fast so that it can continuously learn from new data and hence prevent going into obsolescence.
4. Proposing a multi-objective optimization (MOO) framework using explainable AI (XAI) to modify the composition of therapeutic peptides while optimizing some of their desirable characteristics.
5. Proposing a novel MOO framework that uses an XAI-based model to find the desirable characteristics of therapeutic peptides to modify and optimize the structure of existing therapeutic peptides.
6. Deploying dedicated online web applications based on each proposed framework to help the medical research community.

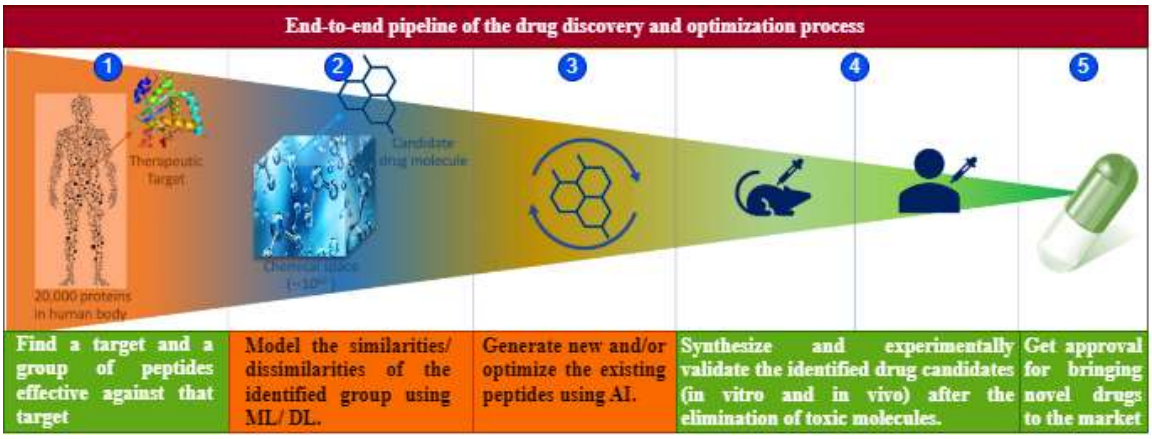

Figure 1.1: Therapeutic peptide-based drug discovery

1.4 Preliminaries

This section contains a concise description of some tools, techniques, and methodologies used to propose the computational frameworks described in this dissertation.

1.4.1 Bidirectional long short-term memory networks (biLSTM)

The bidirectional long short-term memory (biLSTM) networks are improved variants of RNNs for modeling sequential data. In this, two LSTM layers process any given sequence in both forward and backward directions to capture information from past and future contexts. The outputs of both the layers are concatenated at each time step (Figure 1.2). This bidirectional approach allows the network to better capture the dependencies between the components of the input sequence.

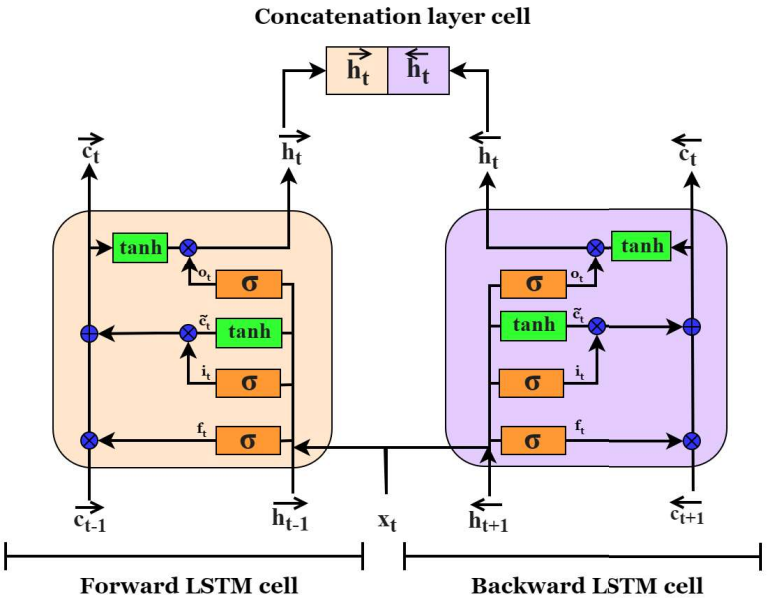

Figure 1.2: Working of a biLSTM cell

1.4.2 Temporal Convolutional Networks (TCNs)

Temporal convolutional networks (TCNs) are one-dimensional convolutional neural networks (CNNs) specially designed for sequence modeling. TCNs are good at capturing

temporal dependencies because they can easily map long-range dependencies. They employ dilated convolutions to capture temporal patterns over long distances without significantly increasing the number of model parameters. Figure 1.3 shows an example of dilated convolutions. TCNs are of two types, enumerated as follows.

1. Causal TCNs: This architecture strictly enforces the causality constraint, which means that the predictions can only depend on the past data and not the future.
2. Acausal TCNs: This architecture relaxes the causality constraint and lets the network obtain information from past and future time steps.

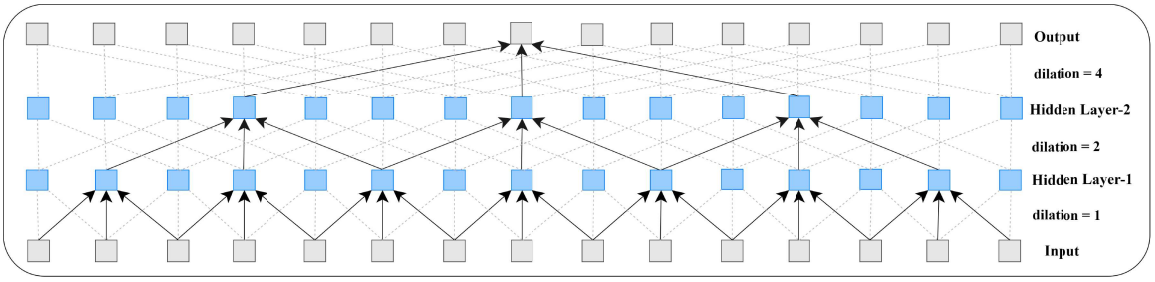

Figure 1.3: A TCN block

1.4.3 Bidirectional encoding representations from transformers (BERT)

The architecture of Bidirectional encoding representations from transformers (BERT) is developed using the transformers, renowned for their proficient handling of interdependencies in sequential data. The model incorporates several layers of self-attention mechanisms, enabling it to assess the importance of various words in a sentence in relation to one another. It is pre-trained on substantial textual data by implementing two unsupervised tasks: the masked language model (MLM) and next sentence prediction (NSP). In contrast to preceding language models, BERT considers both the preceding and succeeding contexts while processing sequences. Its bidirectional nature enables it to capture a more comprehensive context and extract deeper meanings effectively. BERT can be fine-tuned on the end tasks, including but not limited to sequence classification.

1.4.4 Non-dominated sorting genetic algorithms (NSGA)-II

It is an evolutionary algorithm utilized for multi-objective optimization (MOO), particularly those involving several conflicting objectives where a good trade-off is sought between them. Key features of Non-dominated sorting genetic algorithms (NSGA-II) include non-dominated sorting (categorizing solutions into different fronts based on their dominance over each other), crowding distance (a measure that encourages diversity in the population of solutions), and genetic operators (selection, crossover, and mutation that enable the creation of new solutions).

1.4.5 Gravitational Search Algorithm (GSA)

The Gravitational Search Algorithm (GSA) is a metaheuristic MOO algorithm inspired by the laws of physics. It imitates the behavior of celestial bodies in a solution search space, where each solution is considered a celestial body that is influenced by the gravitational forces applied to it by other bodies. In GSA, two solutions exert gravitational force on each other based on their masses and the distance between them, which are used to calculate their fitness value. The algorithm iteratively updates the positions of these solutions, allowing them to maintain a reasonable balance between the exploration and exploitation of the search space and converge to the regions of near-optimal solutions.

1.4.6 Continual Learning

Continual learning pertains to the capacity of a model to acquire knowledge and adjust its behavior based on newly presented data while retaining previously acquired information without experiencing any loss. Conventional machine learning models frequently have difficulties adapting to novel data, as they may erase or entirely disregard previously acquired knowledge, also called catastrophic forgetting. Various methodologies based on regularization, replay, etc., are employed to alleviate this phenomenon and

facilitate the acquisition of new knowledge by the model while preserving the previously acquired knowledge. Regularization techniques impose penalties on changes to learned parameters, promoting gradual rather than abrupt model adaptation. Two examples of such techniques are elastic weight consolidation (EWC) and synaptic intelligence (SI). Replay methods store and regularly replay previous data to the model to mitigate the phenomenon of forgetting. This can be accomplished by methods such as generative replay or experience replay. Lastly, dynamic architectures can be employed, which involve models with dynamic structures capable of expanding and contracting in response to new data. Progressive neural networks (PNNs) incorporate more neural networks to learn novel data while retaining the knowledge of the older ones.

1.4.7 Transfer Learning

Transfer learning is a machine learning methodology that involves the reuse or adaptation of a pre-existing model, first constructed for one specific task, as a foundation for training a second closely related task. Instead of commencing a model's training from its initial state on the second task, transfer learning builds on the knowledge acquired from the first task. This methodology can significantly enhance the efficiency of training the models, particularly in scenarios where the amount of data available for the new task is limited.

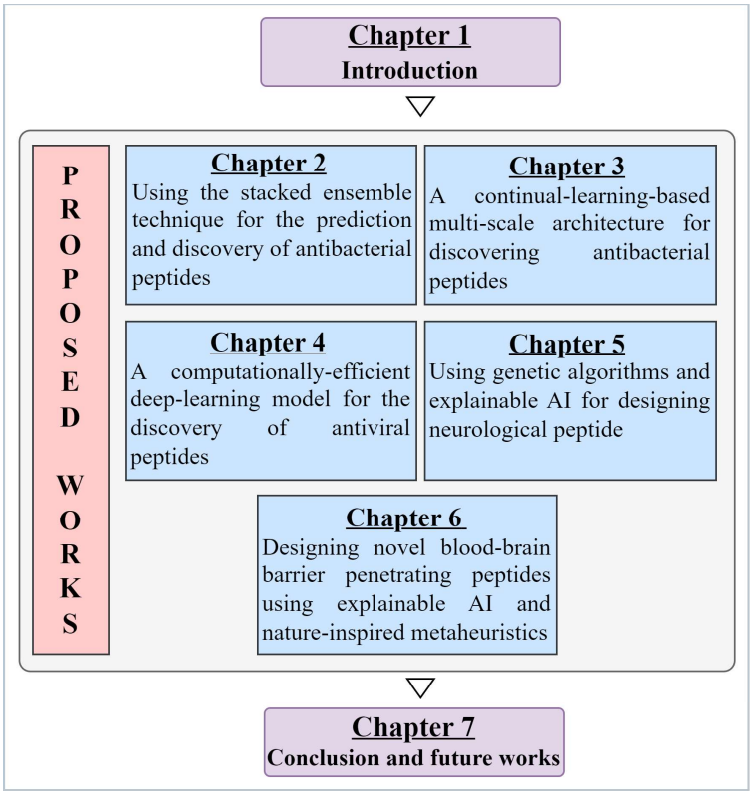

Figure 1.4: Layout of the thesis

1.5 Contributions

The layout of this thesis is illustrated in Figure 1.4. It has five contributing chapters, which are briefly described as follows.

Chapter-2: A stacked ensemble model has been proposed to classify antibacterial and non-antibacterial peptides (ABPs and non-ABPs), which takes less time to get trained despite being based on a biLSTM-based architecture. In this model, biLSTM was applied at the base level whose output was input to the meta-level in conjunction with a few handcrafted features. At the meta-level, an ensemble of three ML algorithms was used to perform the final classification. The entire model was termed the StaBLe-ABPpred (**Sta**cked ensem**ble** based **ABP pred**ictor), and its corresponding app has been deployed at `https://stable-abppred.anvil.app`. In this chapter, the computational challenge of decreasing the time to train a sequential deep learning model has

been overcome.

Chapter-3: This chapter contains details about the Deep-AVPiden model, which has been built using depth-wise separable TCNs to build an efficient, low-memory-consuming model to identify antiviral peptides (AVPs). This was done in order to deploy the model easily at the backend of an AVP identification application so that it could run on resource-constrained devices without using much of their resources. This app was deployed at `https://deep-avpiden.anvil.app`. In this chapter, the computational challenge of decreasing the training time and storage space consumed by a massive deep-learning model has been fulfilled.

Chapter-4: This chapter details the construction of a continual learning-based re-trainable model called MSTCN-ABPpred that identifies ABPs using multi-scale TCNs and is re-trainable on new data (a feature that prevents this model from going into obsolescence). Thus, the computational challenge tackled in this work was to propose a fast, re-trainable model that uses continual learning while preventing catastrophic forgetting so that the app based on the model would remain relevant for many years. The app corresponding to this work has been deployed at `https://mstcn-abppred.anvil.app`.

Chapter-5: In this chapter, a multi-objective optimization (MOO) framework has been proposed that optimizes the structure of existing and uncharacterized neurological peptides (NPs) using non-dominated sorting genetic algorithms (NSGA-II). This proposed framework also uses an explainable AI (XAI) to identify and modify the NPs based on some desirable features to increase their efficacy. A web app corresponding to this work has been deployed at `https://neuropred.anvil.app`.

Chapter-6: A novel MOO framework based on the gravitational search algorithm (GSA) has been proposed in this chapter. It uses XAI to find the desirable characteristics of blood-brain barrier penetrating peptides (B3P2s) and modify their structure by optimizing them based on them. Moreover, it also uses an XAI-based DL tool

to find the AAs of the B3P2s that are essential for their penetrating nature to prevent their modification. A web app corresponding to this work has been deployed at `https://b3p2design.anvil.app`.