HIGH SCHOOL EDITION
Journal of Student Research
Volume 14 Issue 1 (2025)

A Machine Learning Approach for Plant-based Drug Discovery: High-Throughput Prediction of Biological Activities and Enzyme Commission Numbers from Phytochemicals and Amino Acid Sequences of Plants

Leoni Kim[1], Ryan Oh[1] and Christopher Koester[#]

[1]Chadwick International School Songdo, Republic of Korea
[#]Advisor

ABSTRACT

The success of many plant-based drugs and the acknowledgement of the limitations of synthetic drugs has again sparked interest in plant-derived natural products (NP) as a valuable source for novel drug development. Researchers have traditionally used the knowledge-based approach, which relies on traditional medicines to identify candidate plants and extracts. However, NP-based drug development comes with many limitations during the screening stage. First, NP extracts are mostly incompatible with target-based or high-throughput screening. Furthermore, in the case of phenotypic assays, the deconvolution of the mechanism of action of the compound is costly and time-consuming. This study proposes a novel machine learning framework for the high-throughput identification and characterization of plant-derived NPs. This framework consists of two independent models. The first model is a neural network designed to predict phytochemicals' bioactivities through multi-label classification in four categories: antioxidant activity, anti-inflammatory, neurotoxicity, and lipid metabolism. The second model is a convolutional neural network (CNN) that predicts the Enzyme Commission (EC) numbers of enzymes present in the plant. The proposed framework showed robust performance with the Bioactivity Prediction Model achieving 97.62% accuracy and the EC-number Prediction Model achieving 81.97% accuracy. The framework facilitates a more efficient NP-based drug development by providing important insights applicable to the screening, isolation, and deconvolution of NPs.

Introduction

The use of plants for medicine dates back to ancient times. The oldest evidence is a Sumerian clay slab made in 5000 BC that gave recipes for preparing drugs out of various plants (Petrovska 2012). Plant-derived medicines have been extensively used in numerous civilizations throughout history for the prevention and treatment of disease and still are the source of primary healthcare for almost 80% of the world's population (Licciardi and Underwood 2011).

Plants continue to be an important source for novel drugs targeting various diseases (Newman and Cragg 2020). These plant-based drugs utilize plant-based natural products (NP) or NP-based derivatives, which includes a wide range of bioactive compounds. NP-based drugs have multiple advantages over traditional synthetic drugs. First, NPs often have less side-effects. In addition, plant-based drugs are found to be more acceptable to different cultures (Nasim et al. 2022). Furthermore, the advent of the high-throughput screening (HTS) of pure synthetic compound libraries in the 1990s soon revealed their limited chemical diversity. Since then, the total number of FDA-approved drugs has been declining. This trend led to the recent renewal of interest in NP-based drugs, because the plant kingdom has a much greater chemical diversity (Atanasov et al. 2015). Reflecting the success of NP-based drugs, NPs or their derivatives represent more than one-third of all FDA-approved new molecular entities (Patridge et al. 2016). Notably successful examples include the flu treatment Tamiflu, based on the compounds found in *Illicium verum* (Licciardi

and Underwood 2011). Such success can be partially attributed to e*thnopharmacology*, a knowledge-based approach classically taken by researchers. The ethnopharmacological approach allows researchers to select the starting test material and the corresponding assays based on plants therapeutically used by traditional medicines of different ethnic groups. Knowledge gathered and verified by traditional medicine provides insight and guidance for transforming plants into pharmaceutical products (Mushtaq et al. 2018).

However, multiple challenges remain due to the nature of plant-based drug discovery. In the context of drug discovery, screening refers to the process of identifying candidate molecules that produce the desired effect from a large set of compounds. Since NP screens often involve multiple crude extracts from natural sources, hypothesizing biological targets prior to the screening, also called the target-based approach, is not compatible. Identifying the bioactive compounds of interest is often challenging, with the rediscovery of NPs that are already known or do not have drug-like properties (Atanasov et al. 2021). Moreover, plant extracts are commonly incompatible with HTS assays, hindering the screening of large NP libraries (Atanasov et al. 2015). Phenotypic assays, which identify the specific mechanism of action of a phenotype, are considered the standard screening approach for NP-based drug discovery. However, such deconvolution often requires significant time and effort (Corson and Crews 2007). The combination of these limitations led to a decline in the absolute number of NP-based drug approvals by the FDA since the 1990s (Patridge et al. 2016).

Aiming to address some of these chronic limitations, this study proposes a machine-learning (ML) based framework consisting of two independent models for the high-throughput identification and characterization of plant-derived NPs. The first model, the Bioactivity Prediction Model (BPM), analyzes the phytochemicals in the extract to multi-label classify each phytochemical into the following bioactivity categories: antioxidant activity, anti-inflammatory, neurotoxicity, and lipid metabolism. The second model is a convolutional neural network (CNN) that processes the amino acid sequences of proteins to predict their Enzyme Commission (EC) numbers, which is a numerical classification system for the type of chemical reactions the enzymes catalyze. Henceforth, this model will be referred to as the EC Number Prediction Model (ECNPM). The information gathered from the prediction of these NP characteristics can be flexibly utilized in many areas of the drug development, such as a more efficient and effective screening, isolation of relevant constituents from the extract, and deconvolution of the mechanism of action.

The contents of this paper are organized in the following order. Chapter 2 explains the necessary background knowledge, including bioactive metabolites. Next, chapter 3 describes the architecture of the algorithms used for the framework. Then, chapter 4 analyzes and discusses the experiment results. Finally, chapter 5 provides a summary and conclusion of the research.

Background Knowledge

Bioactive Metabolites

Natural products are defined as chemicals produced by living organisms, and can be classified into primary and secondary metabolites. Plants produce a wide variety of metabolites that are bioactive, meaning they produce specific effects when registered to an organism (Bano et al. 2023). Bioactive metabolites are important resources in drug discovery and development due to their diverse chemical structures and biological activities. Scientists often isolate these compounds from natural products or synthesize analogs inspired by their structures to develop new medicines. However, despite their advantages, scientists have struggled to enhance further bioactive metabolites because of their complexity and diversity. The bioactive metabolites of natural products can be highly complex and diverse in their structure, making the synthesis, isolation, and characterization of these compounds difficult, time-consuming, and inefficient. The extraction and purification processes to isolate bioactive metabolites from natural sources would also be resource-intensive, requiring large amounts of raw materials and solvents.

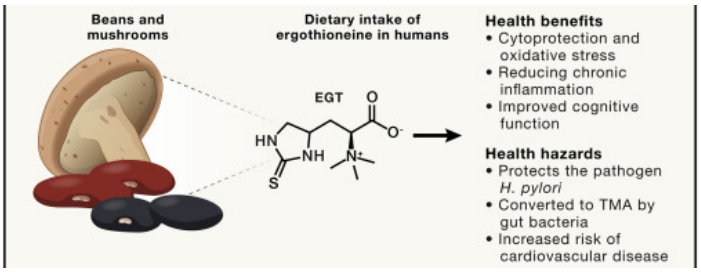


Figure 1. Bioactive Metabolites of Beans and Mushrooms (Nielsen 2022)

An example of a bioactive metabolite is Ergothioneine obtained from beans and Mushrooms as shown in Figure 1. Ergothioneine can have health benefits such as cytoprotection, reducing chronic inflammation, and improving cognitive functions upon human dietary intake. However, Ergothioneine also has health hazards including increased risk of cardiovascular disease. As such, identifying the benefits and risks of metabolites is important when obtaining metabolites for drug discovery. The first model of the framework will predict the potential benefits and hazards of metabolites in natural product extracts during the screening stage.

Object Classification

The second model of the framework utilizes the Convolutional Neural Network (CNN) to classify bioactive metabolites based on their structural data to undergo a multilabel classification test. The model identifies patterns associated with their benefits and potential as a drug. Datasets such as bioactive metabolites represented as SMILES (Toropov et al. 2005), InChl, molecular graphs, or gene ontology are used to process the data in various ways. Later, these structures will be converted into a suitable input format for our CNN model, including generating 2D molecular images or creating adjacency matrices for graph-based representations. The input data is then augmented through various transformations to train and, further on, enhance the model.

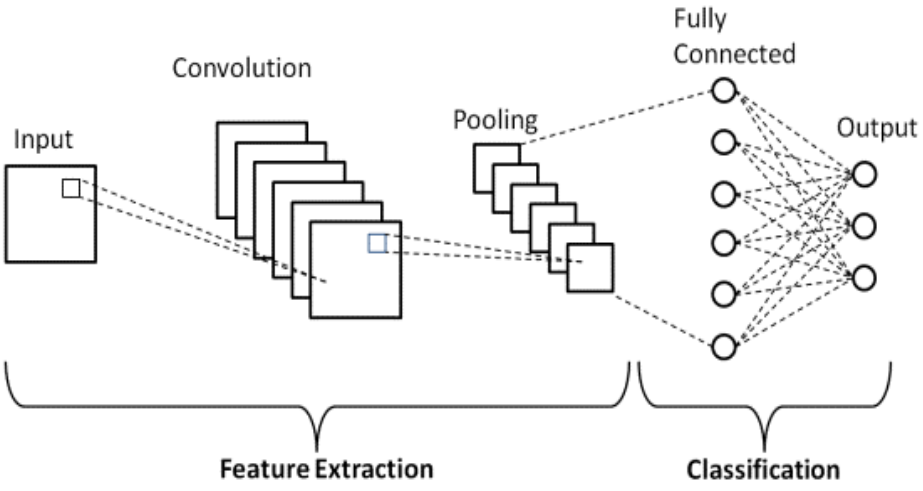


Figure 2. Architecture of convolutional neural networks

The overall structure of the CNN model compromises input layers for processed molecular data, multiple convolutional layers to extract high-level features, pooling layers for dimensionality reduction, and fully connected

layers acting as classifiers. The output layer, which usually uses a softmax activation function, classifies the metabolites into four classes: anti-oxidant, protection toxicity, anti-inflammation immunity, and lipid metabolism.

During training, a categorical cross-entropy loss function is optimized using algorithms like Adam. Performance is assessed using k-fold cross-validation and measures including accuracy, precision, recall, F1-score, and AUC-ROC.

Proposed Method

To predict the effects of natural products, we analyze them using two metrics and two models. The first model analyzes phytochemicals to predict their potential biological activities. The second model predicts the EC number of natural products by analyzing their amino-acid sequences. With the help of these two networks, we construct a high-throughput system capable of predicting the biochemical and pharmaceutical characteristics of plants and their natural products.

Biological Activity Prediction Network

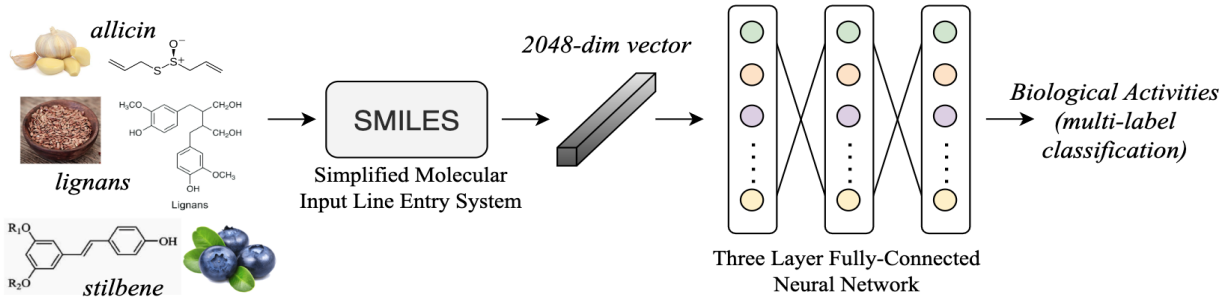


Figure 3. Architecture of biological activity prediction network

Figure 3 illustrates the proposed biological activity prediction network. The proposed network takes phytochemical structures of natural products (molecular structures) as input. Next, these phytochemical structures go through a Simplified Molecular Input Line Entry System (SMILES) to convert their term into a 2048-dim mathematical vector. This conversion process is necessary to utilize our machine learning network because it requires a number value input to classify the data.

After being classified through the three-layer fully connected neural networks, the biological activity prediction of those phytochemical structures is divided into four labels as a multi-label classification: antioxidant, protection toxicity, anti-inflammation immunity, and lipid metabolism. Antioxidant refers to the ability to prevent cell damage by neutralizing free radicals. Protection toxicity indicates the ability to protect cells from toxic substances. Anti-inflammation immunity involves reducing inflammation and enhancing the immune response. Lastly, lipid metabolism relates to regulating the fats and lipids in the body.

To train the proposed network, we utilize the Binary Cross Entropy Loss Function, which is often used in binary outcome prediction tasks. Equation 1 demonstrates the function employed in the method.

Equation 1: Binary Cross Entropy Loss Function

$$L_{BCE} = -\frac{1}{C}\sum_{i \in C} y_i log_e(\hat{p}_i) + (1 - y_i)\log(1 - \hat{p}_i)$$

Here, $\boldsymbol{C}$ denotes the number of categories of biological activities. The variables $\boldsymbol{y}$ and $\boldsymbol{p}$ represent the ground truth and predicted values, respectively. The first term in the sigma notation calculates the loss of the prediction when the ground truth value is 1, meaning the phytochemical causes the biological activity. The second term in the sigma notation calculates the loss when the ground truth value is 0, indicating the phytochemical does not incite biological activity. The fraction $\frac{1}{C}$ calculates the average loss of the prediction for each of the categories.

Enzyme Commission Number Prediction Network

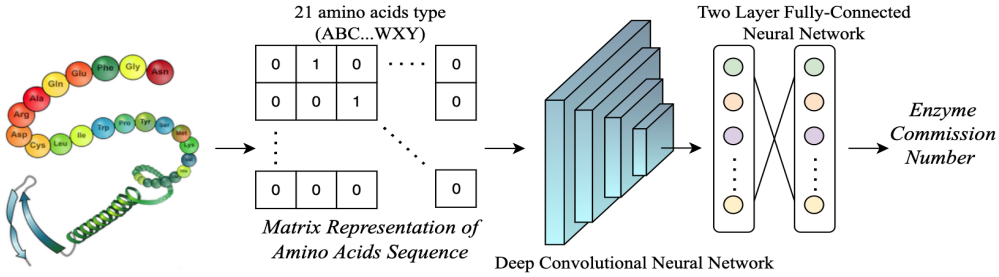


Figure 4. Architecture of enzyme commission number prediction network

Figure 4 illustrates the proposed enzyme commission number prediction network. The proposed network takes the amino acid sequences of natural products as input. These amino sequences are converted into matrix representations representing all 21 amino acid types. This conversion process must utilize the proposed deep convolutional layers because they require a matrix input to classify the data. After classification through the two later fully connected neural networks, the enzyme commission number is according to the following amino acid input results.

Each EC number has four numerical indexes that classify the enzyme's function. The enzyme is first classified into one of the seven main classes of enzymes. The second index, called the sub-class, often indicates what type of compound or bond the enzyme acts on. The sub-sub-class is the third index and further classifies the enzyme's catalysis. Finally, the fourth number, the serial number, is a unique number given to each enzyme, often in the order that the enzyme was added to the list.

We utilized the Cross-Entropy Loss Function to train the EC Number prediction network. Equation 2 provides the said function.

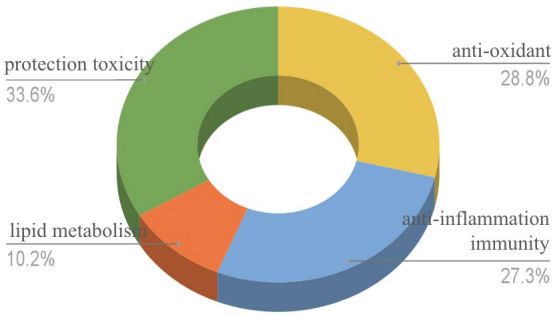
Equation 2: Cross-Entropy Loss Function

$$L_{CE} = -log_e(\hat{p})$$

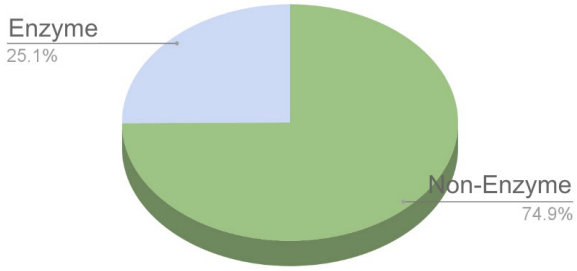
Experimental Results

Datasets

We utilized two types of natural plant datasets to train and test the proposed networks (AI Hub 2024). Figure 5 demonstrates two distribution analysis charts for the biological activity and EC number prediction datasets. Figure 5 (a) presents the distribution of the four biological activities after categorization: anti-oxidant, protection toxicity, anti-inflammation immunity, and lipid metabolism. Figure 5 (b) presents the distributions of the enzymes and non-enzymes among the amino acid sequences dataset used for the first division process of the enzyme commission number prediction network.



(Dataset 1. Bioactivity Distribution of Phytochemicals)



(Dataset 2. Distribution of Genes)

Figure 5. Distribution of Dataset 1 and 2

Inference Metrics

To assess the proposed method's performance, we used four inference metrics often used in classification tasks: accuracy, precision, recall, and F1-score. These inference metrics utilize the following terms: true positive, true negative, false positive, and false negative.

A true positive is an outcome when the prediction model correctly predicts the positive class, and a true negative is when the model correctly predicts the negative class. A false positive is an outcome when the prediction model falsely predicts the positive class, and a false negative is when the model fails to predict the negative class.

Equation 3: Accuracy

$$Accuracy = \frac{True\ Positive + True\ Negative}{Total}$$

In classification, Accuracy is an evaluation metric that measures the ratio/proportion of the correct predictions made by a classification model on a given dataset. As written in the equation, Accuracy is defined as the ratio of the proper proportions to the total number of predictions made by the model.

Equation 4: Precision

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

In classification, Precision is an evaluation metric that measures the ratio/proportion of what is positive (True Positive) among the total positive (True Positive + False Positive) the model predicted. It is used to evaluate how precise the model is in predicting positive data.

Equation 5: Recall

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

The recall metric indicates the percentage of True Positive values within the actual positive sample, which consists of the sum of True Positive and False Negative values.

Equation 6: F1-Score

$$F1score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

The metric F1-Score is the middle value between the Recall and Precision values. It is often used to check the balance between precision and recall.

Table 1. Evaluation results of biological activity predictionAF

Neural Network	Accuracy	Recall	Precision	F1-Score
Two Layer (hidden size=128)	96.02	96.53	95.53	96.02
Two Layer (hidden size=256)	97.54	96.55	95.56	96.05
Two Layer (hidden size=512)	97.55	96.55	95.57	96.05
Three Layer (hidden size=128)	97.59	96.61	95.78	96.19
Three Layer (hidden size=256)	97.62	96.63	95.81	96.21
Three Layer (hidden size=512)	97.59	96.61	95.81	96.21

The data table on the top represents the inference metrics' scores of the biological activity prediction network that takes phytochemical structures as input. The neural network varies with the difference in the number of layers: two or three and the number of hidden layers: 128, 256, and 512.

The data table shows that the difference between the two-layer neural networks and three-layer neural networks was most evident. The three latter neural networks generally showed an increase of (#) from all inference metrics categories (accuracy, recall, precision, F-1 score). The hidden size increase from 128 to 256 or 512 also showed a general increase of (#). In contrast, the rise in hidden size from 256 to 512 showed equal or slightly increased or decreased inference metric scores, showing an overfitting problem of the network.

Throughout the three-layer neural networks, the network with the hidden size of 256 showed the highest results with the four inference metrics tests: 97.62 for accuracy, 96.63 for recall, 95.81 for precision, and 96.21 for F-1 score. Our team could conclude that a three-layer hidden size 256 model should be used for the biological activity prediction network.

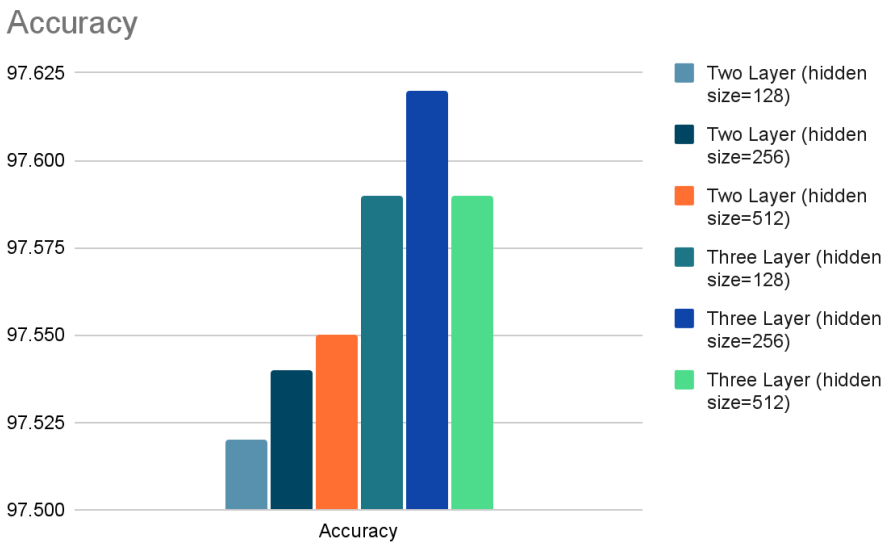


Figure 6. Evaluation result of biological activity prediction (accuracy)

The bar graph illustrates the accuracy score of the neural networks with different architectures. The two-layer and three-layer models are compared with hidden layers varying to 128, 256, and 512 sizes. The bar graph indicates that the three-layer network models surpass the two-layer models' accuracy throughout all the hidden layer sizes. However, for the three-layer models, the 256 hidden size showed the best fit and performance in accuracy, suggesting that the 512 hidden size causes an overfitting problem, decreasing the accuracy inference metric score.

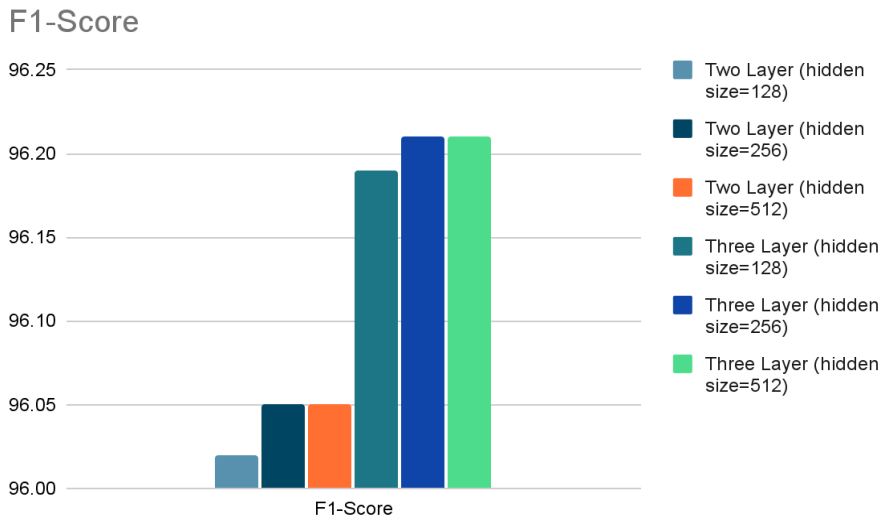
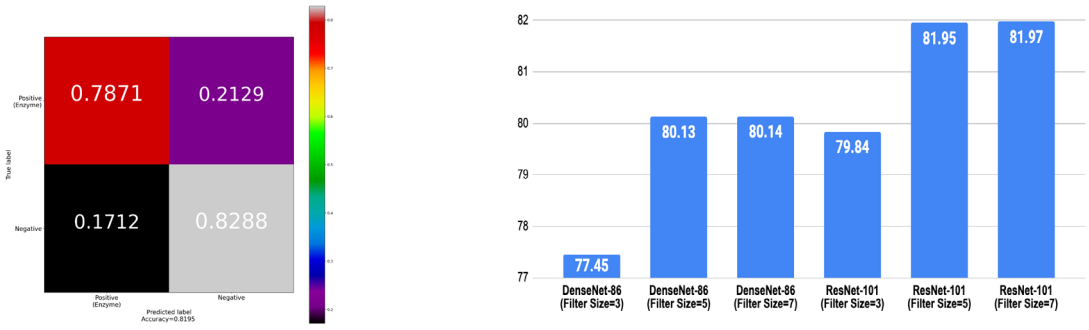


Figure 7. Evaluation result of biological activity prediction (F1-Score)

The bar graph illustrates the F-1 score of the neural networks with different architectures. The two-layer and three-layer models are compared with hidden layers varying to 128, 256, and 512 sizes. The bar graph indicates that the three-layer network models surpass the two-layer models in F-1 score throughout all the hidden layer sizes. However, for the three-layer models, the 256 and 512 hidden sizes showed the best fit and performance in accuracy.



(a) (b)

Figure 7. Evaluation results
(a): confusion matrix (enzyme prediction) and (b): Enzyme commission number prediction

As can be shown in Figure 7(a), ECNPM correctly determined whether a protein was an enzyme or not with an average accuracy of 81.95%. This accuracy is a promising result considering that the dataset has a disproportionately lower number of samples that were enzymes (25.1%). Figure 7(b) demonstrates the experimentation results of ECNPM with different CNN architectures. Generally, CNNs with greater filter sizes performed better than those with smaller filter sizes. For example, despite having a deeper layer, the ResNet-101 CNN with a filter size of 3 had slightly lower accuracy than the DenseNet-86 CNN with a filter size of 7. Models with larger filter sizes had higher accuracy because the models had a greater receptive field and were able to better capture the physical structure of the amino-acid sequence in the form of matrix representation. For the same reason, Resnet-101 with a filter size of 7 performed the best with 81.97% accuracy in EC number prediction. Generally, the accuracy reflected a robust performance, considering that the model had to make a prediction from the 8423 existing EC numbers registered in BRENDA (BRENDA). However, this does not mean that the relatively low accuracy of ECNPM should be left unaddressed. Further research should focus on a more effective model architecture suitable for the complexity of the EC number system. Variants of Recurrent Neural Network (RNN) could potentially improve the accuracy by analyzing the amino-acid sequence of the proteins, since RNNs are specifically trained on sequential data (IBM).

Conclusion

This research proposes a novel method that can effectively complement NP-based drug development. The framework addresses traditional challenges in the screening process by utilizing different machine learning networks to predict the characteristics of NPs present in an extract. First, BPM was able to make a four-class prediction of phytochemicals' bioactivities with a 97.6% accuracy. Second, ECNPM could predict, with 81.9% accuracy, the EC number of enzymes by analyzing their amino-acid sequences. There were also several limitations. The small size of the phytochemicals dataset might raise questions about BPM's efficacy in actual development settings. The limitation can be easily addressed by training BPM with a greater number of samples. In addition, due to the complexity of the EC number classification system, ECNPM achieved a relatively lower accuracy. Further research should focus on a more effective network structure, potentially using RNNs. In conclusion, the innovative framework exemplifies the effective use of machine learning networks in novel drug development. Rather than focusing on specific challenges, the framework provides important characteristics of NPs that can assist a wide variety of areas, such as isolating candidate compounds from the extract and identifying molecular targets for a phenotype.

Acknowledgments

I would like to thank my advisor for the valuable insight provided to me on this topic.

References

A. G. Atanasov *et al.*, "Discovery and resupply of pharmacologically active plant-derived natural products: A review," *Biotechnology Advances*, vol. 33, no. 8, pp. 1582–1614, Dec. 2015, doi: 10.1016/j.biotechadv.2015.08.001.

A. G. Atanasov, S. B. Zotchev, V. M. Dirsch, and C. T. Supuran, "Natural products in drug discovery: advances and opportunities," *Nature Reviews Drug Discovery*, vol. 20, no. 3, pp. 200–216, Jan. 2021, doi: 10.1038/s41573-020-00114-z.

AI Hub. (2024, Aug 21). "*Plant functionality prediction genomic data*": AI Hub. https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=71316

Bano, Asghari, et al. "Bioactive Metabolites of Plants and Microbes and Their Role in Agricultural Sustainability and Mitigation of Plant Stress." *South African Journal of Botany*, vol. 159, Aug. 2023, pp. 98–109, doi:10.1016/j.sajb.2023.05.049.

BRENDA, "All enzymes," BRENDA Enzyme Database. Accessed: Sep. 22, 2024. [Online]. Available: https://www.brenda-enzymes.org/all_enzymes.php

Bruce, Stella Omokhefe. "Secondary Metabolites from Natural Products." *IntechOpen*, 16 Feb. 2022, https://www.intechopen.com/chapters/80477.

D. J. Newman and G. M. Cragg, "Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019," *Journal of Natural Products*, vol. 83, no. 3, pp. 770–803, Mar. 2020, doi: 10.1021/acs.jnatprod.9b01285.

E. Patridge, P. Gareiss, M. S. Kinch, and D. Hoyer, "An analysis of FDA-approved drugs: natural products and their derivatives," *Drug Discovery Today*, vol. 21, no. 2, pp. 204–207, Feb. 2016, doi: 10.1016/j.drudis.2015.01.009.

IBM, "Recurrent neural network (RNN)," IBM. Accessed: Sep. 22, 2024. [Online]. Available: https://www.ibm.com/topics/recurrent-neural-networks

Licciardi, P. V., & Underwood, J. R. (2011). Plant-derived medicines: a novel class of immunological adjuvants. International immunopharmacology, 11(3), 390-398.

Mushtaq, S., Abbasi, B. H., Uzair, B., & Abbasi, R. (2018). Natural products as reservoirs of novel therapeutic agents. EXCLI journal, 17, 420.

Nasim, N., Sandeep, I. S., & Mohanty, S. (2022). Plant-derived natural products for drug discovery: current approaches and prospects. The Nucleus, 65(3), 399-411.

Nielsen, J. (2022). Bioactive metabolites: The double-edged sword in your food. Cell, 185(24), 4469-4471.

Petrovska, B. B. (2012). Historical review of medicinal plants' usage. Pharmacognosy reviews, 6(11), 1.

Toropov, A. A., Toropova, A. P., Mukhamedzhanoval, D. V., & Gutman, I. (2005). Simplified molecular input line entry system (SMILES) as an alternative for constructing quantitative structure-property relationships (QSPR).

T. W. Corson and C. M. Crews, "Molecular Understanding and Modern Application of Traditional Medicines: Triumphs and Trials," *Cell*, vol. 130, no. 5, pp. 769–774, Sep. 2007, doi: 10.1016/j.cell.2007.08.021.