Lesson number
SET2700G

2025
Final report
Hanmo Shi,Qixuan Wang,Yingying He,
Yuxiang Tian
Team Number
Group1

Using machine learning methods to identify novel compounds

Abstract

The rising costs and inefficiencies in traditional drug development call for innovative solutions. This study explores the application of artificial intelligence and machine learning to make the drug discovery process more efficient. By integrating semi-supervised and weakly supervised learning approaches with vision-language models, we enhance the analysis and classification of drug compounds. Using data from sources such as PubChem and DrugBank, convolutional neural networks (CNNs) and the VGG-16 architecture are employed to predict pharmaceutical properties and ensure drug safety. The results demonstrate significant improvements in the speed and accuracy of compound screening. Future directions include model optimization, such as pruning and quantization, and incorporating hybrid machine learning techniques to maximize performance while minimizing computational demands.

Keywords: Machine Learning,Drug Discovery,Convolutional Neural Networks,VGG-16,Data Augmentation

Contents

1 Introduction **3**

1.1 Background 3

1.2 Our Research Focus 3

2 Datasets **3**

2.1 CID and SMILES 3

2.2 Crawl data from websites 3

3 Traditional CNN **4**

3.1 Related Work 4

3.2 Our Model 5

3.3 Result 6

4 VGG-16 **7**

4.1 Model Architecture 7

4.2 Challenges and Limitations 9

4.3 Comparison with Traditional CNN 9

5 Conclusion **9**

6 Future Work **10**

6.1 Model Optimization 10

6.2 Overfitting Mitigation 10

7 GitHub Repository **10**

8 Contributions **10**

1 Introduction

1.1 Background

Traditional drug development faces significant challenges, including complex reaction pathways, low experimental efficiency, and excessive resource consumption. These issues result in high costs, prolonged development cycles, and limited innovation [1]. The reliance on trial-and-error methods and manual experimentation further restricts the scalability and speed of drug discovery. Recent advancements in machine learning, particularly in deep learning, offer promising solutions to these challenges. Machine learning can optimize the drug development process, enhance compound identification, accelerate screening, and reduce the need for costly experiments, making it a transformative tool in the field [3].

1.2 Our Research Focus

Our research focuses on leveraging Convolutional Neural Networks (CNNs) to analyze drug compounds using 2D molecular data. CNNs excel in image-based pattern recognition, making them ideal for identifying subtle patterns in complex chemical structures [5]. While previous studies have demonstrated the utility of CNNs in analyzing 2D chemical representations, limitations remain, such as the inability to modify 2D structures without altering their chemical meaning [4]. Our work aims to overcome these limitations by applying advanced machine learning techniques to 2D molecular data, ultimately accelerating drug discovery and reducing reliance on traditional trial-and-error methods [6]. Additionally, we explore the use of hybrid models, such as combining CNNs with transformers, to improve compound-protein interaction predictions [7], and evaluate pre-trained CNN models for 2D image-based drug discovery [2]. By leveraging these techniques, we aim to enhance the efficiency and accuracy of drug discovery processes.

2 Datasets

2.1 CID and SMILES

We first extract CID(Chemical Identifier),a unique numerical identifier assigned to chemical compounds in databases from PubChem. Each CID corresponds to a specific chemical substance, enabling precise referencing, retrieval, and analysis of chemical.Then we can utilize the CIDs we gained to generate SMILES,which was proposed by ([8]) is currently widely recognized and used as a standard representation of compounds for modern chemical information processing.Then we can gain the classified SMILEs as below.(Table. 1)

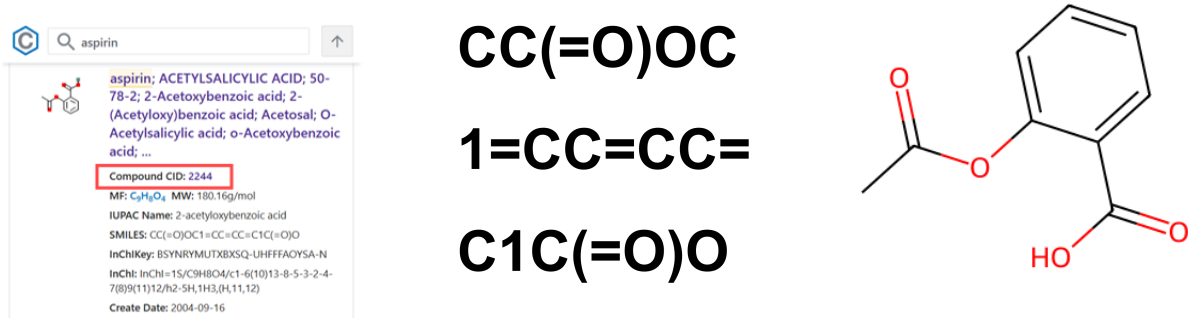
2.2 Crawl data from websites

We use the Selenium module to write scripts for crawling CID data from the PubChem website, which contains the information of the CIDs in our datasets. The process is as follows:

- Use the Selenium module to write scripts for crawling CID data from the PubChem website(Fig. 1a).
- Translate the CIDs into SMILES format(Fig. 1b), which is the standard representation of chemical compounds.
- Further convert the SMILES into 2D(Fig. 1c) images of the compounds.

Categories	Effect
Antibacterial	Inhibits or kills bacteria.
Antiviral	Inhibits or kills viruses.
Antifungal	Inhibits or kills fungi.
Antiprotozoal	Inhibits or kills protozoa.
Anti-inflammatory	Relieves or suppresses inflammation.
Antipyretic	Reduces fever and relieves high temperature.
Analgesic	Relieves pain.
Antioxidant	Prevents or reduces oxidative cell damage.
Antitumor	Inhibits the growth and spread of tumors.
Antidepressant	Relieves symptoms of depression.
Sedative	Calms agitation and reduces anxiety.
Hypnotic	Induces sleep and helps with insomnia.
Antihypertensive	Lowers blood pressure.
Antidiabetic	Used to control diabetes.
Antihistamine	Blocks histamine effects, relieves allergies.
Antispasmodic	Relieves muscle spasms or intestinal cramps.
Diuretic	Increases urine output, reduces fluid retention.

Table 1: Types and effects of drugs in the datasets



(a) Aspirin CID (b) Aspirin SMILES (c) Aspirin 2D Image

Figure 1: How to get data

A detailed demo video can be found in our GitHub repository. After the work of translation,we would be able to obtain the images of compounds,which would be our datasets.

3 Traditional CNN

3.1 Related Work

The most important article in the field of CNNs is widely considered to be [5]. This paper introduced AlexNet, which won the ImageNet Large Scale Visual Recognition Challenge with a significant margin over traditional machine learning methods.The most important innovation of their work is the introduction of the ReLU Activation and Dropout Regularization.The ReLU Activation is the nonlinearity

$$f(x) = max(0, x) \tag{1}$$

which could be several times faster than their equivalents with tradition units. Here is the architecture of their CNN.(Fig. 2)

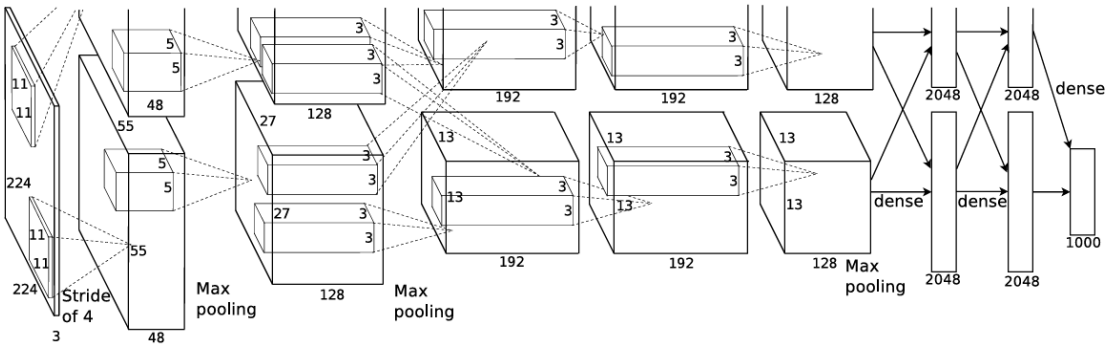


Figure 2: The illustration of the architecture of the original CNN

3.2 Our Model

According to their work,our model(Fig. 3) is organized into three main sections: input layer, feature extraction, and classifier, ending in an output layer. Here's a detailed breakdown:

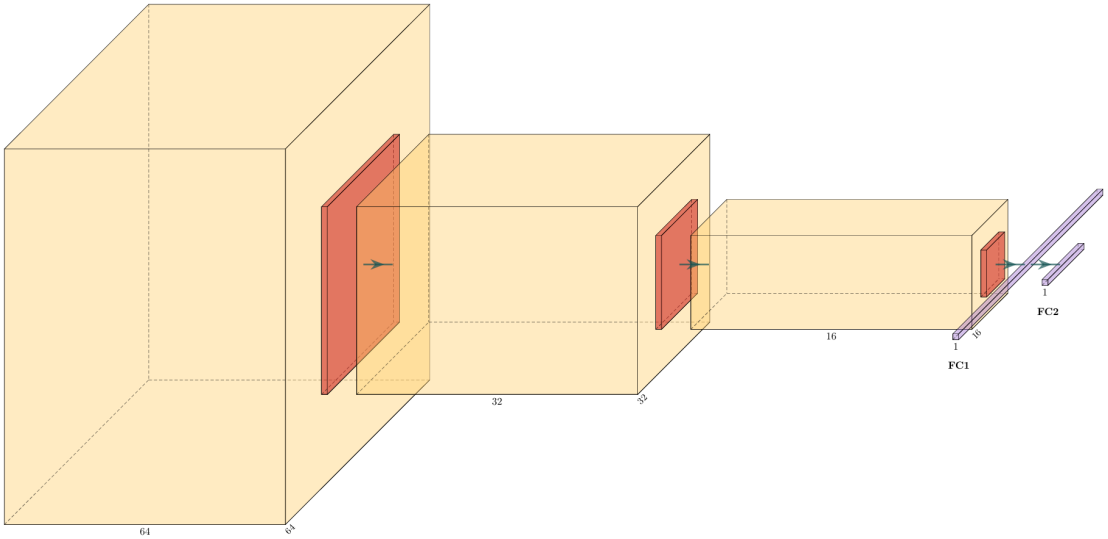


Figure 3: Thumbnails of Our CNN Model

(1)Input Layer:The input layer accepts an image of dimensions , where 3 represents the RGB color channels, and H and W are the height and width of the image. (2)Feature Extraction: The feature extraction stage comprises three convolutional blocks.The first Conv2D layer transforms the input using 64 filters of size 3×3 with padding of 1.The output shape is [64,H/2,W/2] , reflecting reduced spatial dimensions due to max-pooling.Another Conv2D layer with 128 filters and a kernel size of 3×3 is applied.As we can see,Max-pooling reduces dimensions further with a kernel size of 2 and stride of 2. The resulting shape is [128,H/4,W/4] .The last Conv2D layer with 256 filters and a kernel size of 3×3 , followed by ReLU activation.Max-pooling reduces the dimensions again, resulting in an output of shape [256,H/8,W/8] .At last ,after the work of flatten Layer,The 3D output from the feature extraction phase is flattened into a 2D tensor with batch size 256×[H/8]×[W/8]. This operation prepares the data for the fully connected layers

in the classifier. (3)Classifier:The classifier is consisted of two fully connected layers.With the assistance of ReLU activation and dropout,the first layer transfers the flattened input to 1024 neurons and the second transfers the 1024 neurons to the class they belong,which is the output.

3.3 Result

Our train loss and train accuracy along with validation accuracy over epochs are as follows.(Fig. 4)

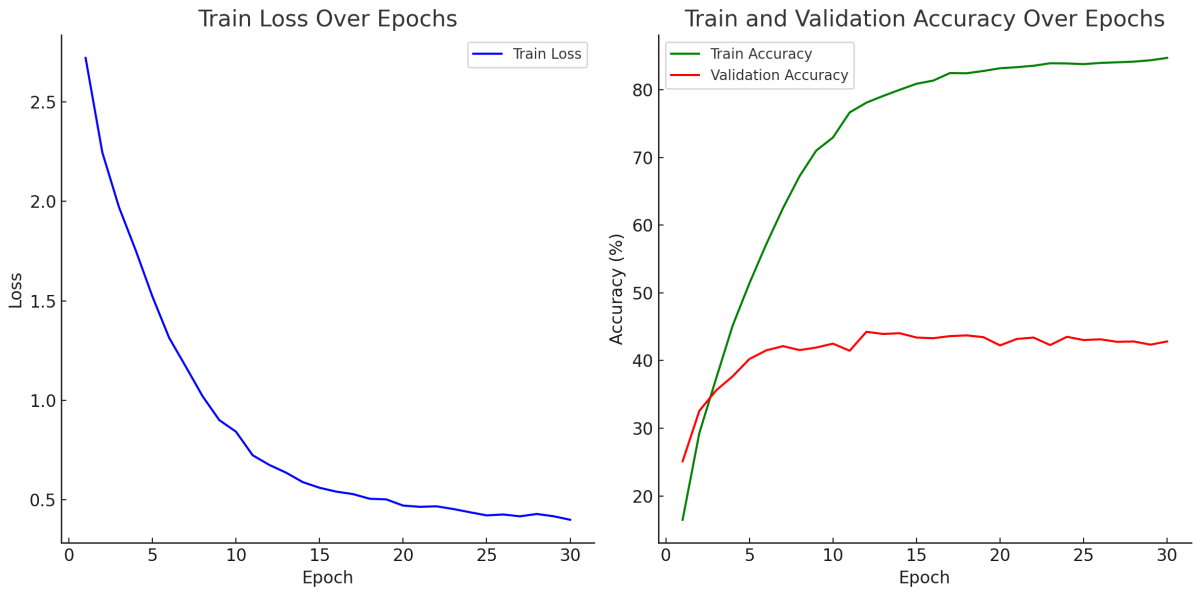


Figure 4: Train Loss and Accuracy

As we can see,on the left graph, the train loss decreases steadily over 30 epochs, indicating that the model is learning and improving its fit to the training data. On the right graph, the training accuracy shows continuous improvement, reaching close to 80%. However, the validation accuracy plateaus at around 40%, showing a significant gap between training and validation performance. This suggests the model is over-fitting—performing well on the training data but struggling to generalize to unseen data.

To assess the accuracy of a classification model which is imbalanced,the confusion matrix provides a detailed breakdown of how well the model is performing by showing the correct and incorrect predictions in a matrix format. (Fig. 5)

The heatmap reveals a concentration of high values along the diagonal, indicating that the model generally performs well in most categories. Some classes, particularly those with overlapping effects (e.g., analgesics and anti-inflammatory drugs, hypnotics and sedatives), exhibit higher off-diagonal values, suggesting areas where further feature extraction or dataset refinement could improve performance. However, there is room for improvement in classes with overlapping effects or shared characteristics, such as hypnotics, sedatives, and antipyretics. Addressing these issues may involve incorporating additional features, refining the dataset, or using advanced techniques like hierarchical classification to handle closely related categories more effectively.

For the purpose of improving the CNN model and addressing the issue of over-fitting observed in the training results, we believe the following strategies can be employed. Data augmentation can increase the diversity of the training dataset, helping the model generalize

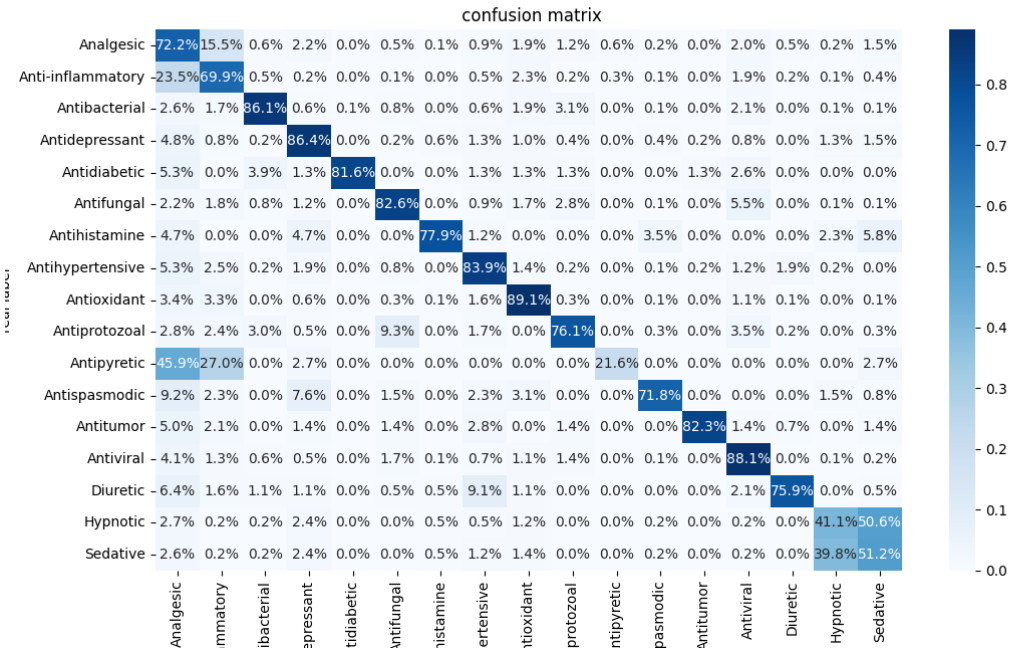


Figure 5: Confusion Matrix of CNN

better. Regularization techniques, such as dropout and L2 regularization, can reduce reliance on specific features and prevent over-fitting. Simplifying the model by reducing its complexity or adding batch normalization layers can also improve generalization. Leveraging pretrained models through transfer learning is another effective approach, especially for smaller datasets. Additionally, using early stopping, adjusting the learning rate, and performing hyper-parameter tuning can optimize training and prevent over-fitting.

4 VGG-16

4.1 Model Architecture

VGG-16 is a deep convolutional neural network consisting of 16 layers, including 13 convolutional layers and 3 fully connected layers. The model uses small 3x3 convolutional filters stacked in layers, followed by max-pooling layers to reduce spatial dimensions.(as shown fig6) This architecture allows for high-level feature extraction, making it suitable for complex image classification tasks. However, its depth and large number of parameters make it computationally intensive, which may not be ideal for smaller datasets or tasks requiring lightweight models.

We attempted to use the VGG-16 model for optimizing the classification of small drug molecules, but the results were unsatisfactory. As shown in Figure 7, the training loss remains consistently high (around 2.5) with minimal fluctuation over 30 epochs, indicating that the model fails to effectively learn from the training data. Both training and validation accuracies are low, with validation accuracy peaking at only 20%, suggesting poor generalization and potential underfitting. This implies that the model is too simple to capture the complex patterns in the data.

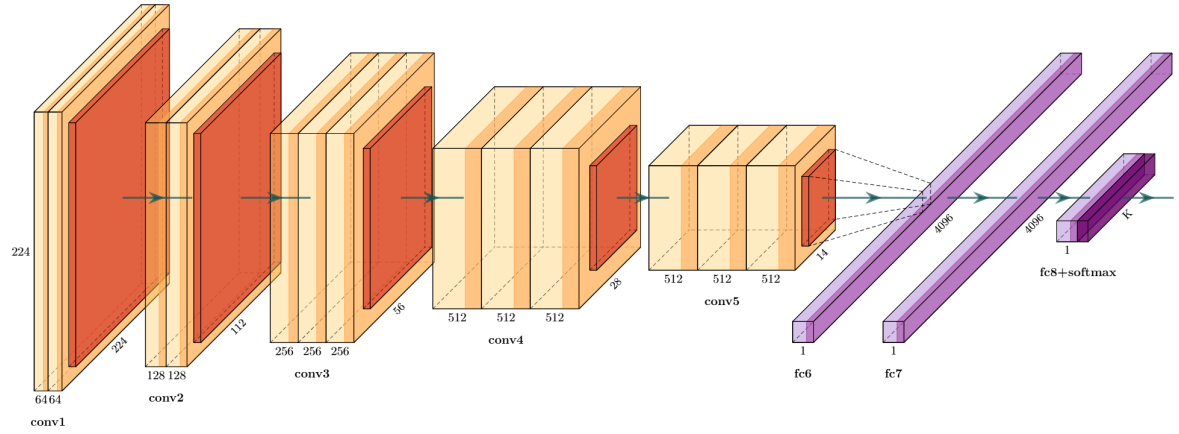


Figure 6: VGG-16 Train and Validation Accuracy over Epochs

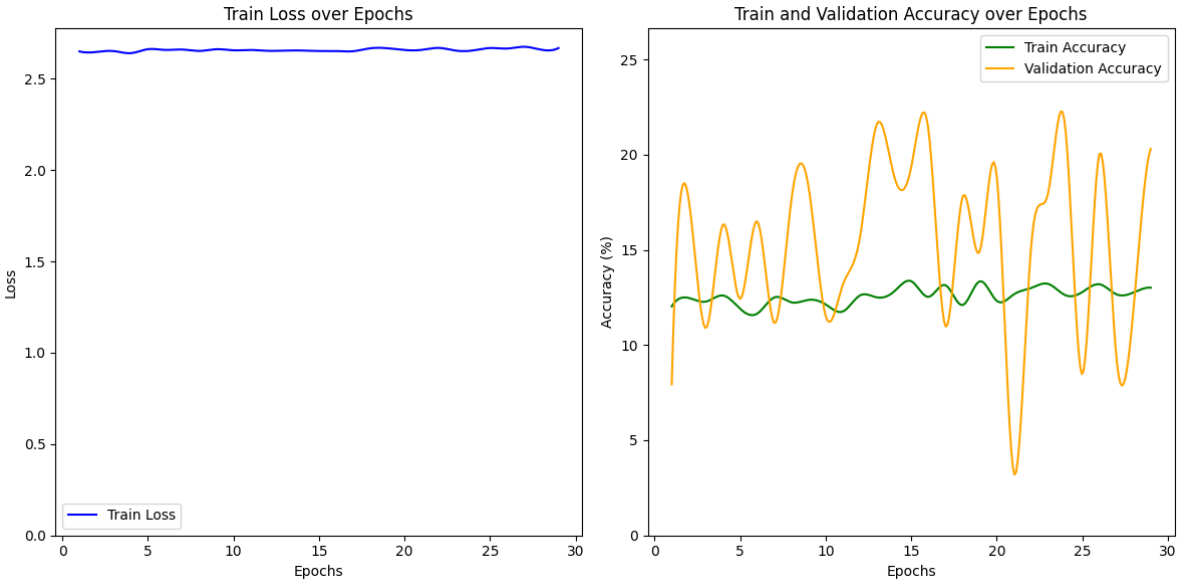


Figure 7: VGG-16 Train and Validation Accuracy over Epochs

4.2 Challenges and Limitations

The poor performance of VGG-16 in our study can be attributed to several factors:

- **Data Limitations**: The dataset may be too small or imbalanced, limiting the model's ability to learn meaningful patterns. Additionally, noise or incomplete feature extraction in the data could hinder performance.
- **Model Suitability**: VGG-16, designed for large-scale image datasets, may not be well-suited for the structural complexity of chemical compounds. Its deep architecture requires substantial computational resources and may struggle with smaller datasets.
- **Training Issues**: The model may require more training epochs or better hyperparameter tuning to improve convergence. The high training loss and low accuracy suggest underfitting, indicating that the model is too simple to capture the data's complexity.

4.3 Comparison with Traditional CNN

Compared to traditional CNNs, VGG-16 offers a more sophisticated and deeper architecture, enabling high-level feature abstraction for complex tasks. However, this comes at the cost of increased computational demands. Traditional CNNs, with their simpler and shallower designs, are more efficient and better suited for smaller datasets or less complex tasks. This highlights the trade-off between model depth and computational efficiency in deep learning applications.

To address these challenges, future work will focus on expanding the dataset, improving data quality, and exploring alternative architectures better suited for compound classification tasks. Additionally, hyperparameter optimization and regularization techniques will be employed to enhance model performance and generalization.

5 Conclusion

In this study, we explored the application of machine learning techniques, particularly Convolutional Neural Networks (CNNs) and VGG-16, to enhance the drug discovery process by analyzing 2D chemical compound data. Our research demonstrated that CNNs are effective in processing 2D molecular representations, significantly improving the accuracy of compound classification compared to traditional methods. However, we encountered challenges such as overfitting, where the model performed well on training data but struggled to generalize to unseen data. This issue was likely due to the limited dataset size, insufficient regularization, and the computational demands of deep architectures like VGG-16.

Despite these challenges, our findings highlight the potential of machine learning to revolutionize drug discovery by accelerating the screening process and reducing reliance on costly and time-consuming experimental methods. Future work will focus on expanding the dataset, improving data quality, and optimizing model architectures to address overfitting and enhance generalization. Techniques such as structured pruning, hyperparameter optimization, and advanced data augmentation will be employed to further refine the models.

By leveraging these advancements, we aim to develop more robust and scalable machine learning models that can significantly improve the efficiency and accuracy of drug discovery, ultimately benefiting the healthcare industry and patients worldwide.

6 Future Work

Our future efforts will focus on two key areas:

6.1 Model Optimization

We will optimize the VGG-16 architecture through structured pruning, removing redundant neurons and connections to improve efficiency and generalization. Additionally, we will explore hybrid models combining CNNs with graph neural networks (GNNs) or recurrent neural networks (RNNs) to better capture spatial and temporal dependencies in molecular data.

6.2 Overfitting Mitigation

To address overfitting, we will implement advanced hyperparameter optimization techniques, such as Bayesian optimization, to fine-tune learning rates, batch sizes, and regularization. Enhanced data augmentation techniques, including rotation and scaling of 3D molecular structures, will also be employed to increase dataset diversity and improve model robustness.

These improvements aim to enhance the efficiency, accuracy, and scalability of our models, advancing drug discovery and benefiting the healthcare industry.

7 GitHub Repository

All our code and data can be found in our GitHub repository: `https://github.com/Fully-ripe-mango/AIGroup`

if you want to know the details of our work,or you want to test our model,please feel free to visit our repository.

8 Contributions

Hanmo Shi:

- Project background and research introduction
- Defects and causes of the VGG-16 model
- Reasons for overfitting in CNN models
- Project summary

Qixuan Wang:

- Construction of the final paper framework
- Some data visualization work

Yingying He:

- Writing part of the thesis content
- Literature review

- PowerPoint making

Yuxiang Tian:

- Model construction and training
- Thesis outline development,formatting revision and integration,Content revision
- Writing part of the thesis content
- Partial visualization (Figures 1,3,5,6,7)
- GitHub repository management

References

[1] Atanas G Atanasov et al. "Natural products in drug discovery: advances and opportunities". In: *Nature reviews Drug discovery* 20.3 (2021), pp. 200–216.

[2] Antonio Jesus Banegas-Luna and Juan Bonastre-Egea. "Evaluation of Pre-Trained CNN Models for 2D Image-Based Drug Discovery". In: *Intelligent Environments 2024: Combined Proceedings of Workshops and Demos & Videos Session*. IOS Press. 2024, pp. 124–133.

[3] Jianyuan Deng et al. "Artificial intelligence in drug discovery: applications and techniques". In: *Briefings in Bioinformatics* 23.1 (2022), bbab430.

[4] Maya Hirohara et al. "Convolutional neural network based on SMILES representation of compounds for detecting chemical motif". In: *BMC bioinformatics* 19 (2018), pp. 83–94.

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems* 25 (2012).

[6] Jihye Park et al. "A brief review of machine learning-based bioactive compound research". In: *Applied Sciences* 12.6 (2022), p. 2906.

[7] Ying Qian, Jian Wu, and Qian Zhang. "CAT-CPI: Combining CNN and transformer to learn compound image features for predicting compound-protein interactions". In: *Frontiers in Molecular Biosciences* 9 (2022), p. 963912.

[8] David Weininger. "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules". In: *Journal of chemical information and computer sciences* 28.1 (1988), pp. 31–36.