WizardChem: AI-Powered Chemical Analysis and Sorting via Computer Vision

Wilbert F. Mays

Stanford University
wilfm@stanford.edu

Abstract

WizardChem presents an innovative approach to automating the identification of chemical compounds using advanced computer vision techniques. Leveraging Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Temporal Convolutional Networks (TCNs) within a Multi-Task Learning framework, this project aims to enhance the accuracy, efficiency, and safety of chemical analysis and sorting. Utilizing bulk chemical data from PubChem, WizardChem has shown promising improvements in identifying compounds using feature vectors generated from images. However, challenges remain in ensuring the accuracy of predictions due to the insufficient depth of image data per compound. Future work includes integrating compound image (2D and 3D molecular structures) datasets and Microsoft COCO, optimizing learning methods, and enhancing the user interface for more comprehensive chemical analysis.

1. Introduction

Chemical analysis and sorting are critical processes in various industries, including chemical manufacturing, pharmaceuticals, and environmental management. Traditional methods of manual analysis and sorting are labor-intensive, prone to human error, and often costly. These challenges are exacerbated by the increasing complexity and volume of chemical data that need to be processed efficiently and accurately. The WizardChem project addresses these challenges by developing an integrated computer vision system that automates the analysis and sorting of chemicals.

The primary focus areas of this project include automating the identification of chemical compounds through images and, in the future, videos, to enhance tracking accuracy. WizardChem uses computer vision to generate feature vectors from images and associate them with specific compounds, minimizing the need for extensive testing. The model is trained on these images so that compounds can be identified by the same or similar images due to feature vector matching. Additionally, video functionality for reaction tracking has been added but not yet tested; only the compound identification through 2D molecular structure images has been tested.

The overall plan for approaching this problem involves leveraging advanced AI techniques, including CNNs, RNNs, and TCNs within a Multi-Task Learning framework. Robust datasets, primarily from PubChem, are used to train the model. The project aims to transform chemical process monitoring and management, providing substantial improvements in accuracy, efficiency, cost savings, and regulatory compliance.

This integration of AI and computer vision is not just a technical endeavor but a personal mission aligned with my long-term professional goals and passion for chemistry.

2. Background/Related Work

The application of computer vision in chemical analysis has been explored in various studies, providing a foundation for WizardChem:

Förster and Hinze (2018): In their study, ”Towards Automated Analysis of Belousov-Zhabotinsky Reactions in a Petri Dish by Membrane Computing Using Optic Flow,” Förster and Hinze explore the automation of analyzing Belousov-Zhabotinsky (BZ) reactions using optic flow within the context of membrane computing. BZ reactions are known for their spiking oscillations and are used as a model for chemical information transmission. The study focuses on automating the identification and localization of these oscillatory spots using optic flow, which measures apparent movement velocities in image sequences. The authors extend existing algorithms to manage noise and environmental perturbations. Results include the development of a method that automates the identification of

BZ oscillatory spots, with improvements in analyzing the propagation velocities of expanding concentric rings.

Agrisuelas et al. (2020): The research titled ”Kinetics of Surface Chemical Reactions from a Digital Video” by Jerónimo Agrisuelas and colleagues investigates a cost-effective method for studying surface kinetics by analyzing color intensities from digital videos. This method leverages RGB color histograms to provide spatiotemporal information on surface chemical reactions. The study highlights how the shapes of curves, peak maxima, and half-peak widths are dependent on kinetic constants and reaction order. This approach offers a comprehensive analysis of surface reactions, demonstrating how color changes can be used to monitor reaction progress in real-time.

Daponte et al. (2020): The study ”Using an Automated Monitoring Platform for Investigations of Biphasic Reactions” by J.A. Daponte and colleagues details the development of an automated monitoring platform for selectively sampling organic phase and slurry solutions simultaneously. This platform facilitates the study of complex biphasic systems by providing detailed reaction profiles for both phases, offering deeper insights into reaction mechanisms. The system allows for continuous monitoring of reactions, aiding in the understanding of dynamic changes and interactions within biphasic systems.

Yan et al. (2023): In the paper ”Computer Vision for Non-Contact Monitoring of Catalyst Degradation and Product Formation Kinetics,” C. Yan and colleagues investigate the use of computer vision for non-contact monitoring of catalyst degradation and product formation kinetics. This study highlights the potential for real-time chemical process monitoring, emphasizing the application of computer vision techniques to enhance the accuracy and efficiency of monitoring catalytic processes.

Schell et al. (2020): The research ”Video Colorimetry of Single-Chromophore Systems Based on Vector Analysis in the 3D Color Space” by J. Schell, S. C. McCauley, and R. Glaser explores video colorimetry for analyzing chemical properties. The study demonstrates how vector analysis in 3D color space can enhance the understanding of chemical systems by providing a robust method for analyzing chromophore behavior in chemical reactions.

Glaser et al. (2019): The paper ”Video-Based Kinetic Analysis of Period Variations and Oscillation Patterns in the Ce/Fe-Catalyzed Four-Color Belousov-Zhabotinsky Oscillating Reaction” by R. Glaser and colleagues focuses on using video-based kinetic analysis to study BZ reactions. By capturing the spatiotemporal behavior of these reactions, the study aims to fine-tune reaction parameters and achieve uniform wave propagation. The methodology involves segmenting motion in video sequences to identify and analyze oscillatory spots accurately. The use of optic flow techniques, particularly the Horn and Schunck algorithm, is highlighted for its robustness in handling noise and environmental influences.

Capitán-Vallvey et al. (2015): The paper ”Recent developments in computer vision-based analytical chemistry: A tutorial review” by L. Capitán-Vallvey and colleagues reviews advancements in computer vision-based analytical chemistry. It discusses various digital image processing techniques applied to chemical sensing, emphasizing the importance of accurate image processing in obtaining reliable chemical data. The study highlights how advanced image analysis can enhance the sensitivity and specificity of chemical sensors.

These studies highlight the potential of computer vision techniques in enhancing the accuracy and efficiency of chemical analysis, providing a solid foundation for WizardChem’s methodologies. By leveraging these advancements, WizardChem aims to automate chemical identification and sorting processes, offering significant improvements in accuracy, efficiency, and safety.

3. Approach

WizardChem integrates several technical aspects to automate chemical identification processes.

3.1. Multi-Task Learning

- Framework: Utilizes Convolutional Neural Networks (CNNs) for image processing and Temporal Convolutional Networks (TCNs) for video processing within a unified framework.

- Feature Extraction: Extracts high-dimensional feature vectors from both images and videos for comprehensive analysis.

3.1.1 Pre-processing

- Normalization and Resizing: Ensures consistent input dimensions by normalizing and resizing images and videos.

- Data Augmentation: Techniques such as random cropping, flipping, and color jittering are applied during training to improve model robustness.

3.1.2 Model Architecture

UnifiedCompoundNet: This class defines the architecture, combining feature vectors from images and videos using fully connected layers with dropout and batch normalization to enhance feature learning.

- Image Model: Incorporates a pretrained ResNet50 model.

- Video Model: Utilizes a pretrained R3D-18 model.

- Integration: Combines image and video features into a unified feature vector through fully connected layers.

3.1.3 Loss Function

Triplet Loss: Employs triplet loss to enforce the desired distance relationships in the feature space, facilitating accurate compound identification. The TripletLoss class ensures that similar compounds are closer in the feature space while dissimilar ones are farther apart.

3.2. Data Collection

PubChem Data: Bulk chemical summaries and 2D molecular structures are downloaded from PubChem by entering a list of compound CIDs, resulting in a .json file containing detailed information for each compound.

3.2.1 Data Organization

- Data Organization Script: The ”data-org.py” script reads the .json file, extracting and updating each compound's properties based on its CID. Images and videos can be manually associated with specific compounds.

- Feature Extraction: Feature vectors are generated from images and videos using the ”UnifiedCompoundNet”, incorporating pretrained ResNet50 for images and R3D-18 for videos. This step is performed by the ”data-org.py” script.

3.2.2 Data Consolidation

Consolidation Script: Extracted data is consolidated into training and validation datasets using the ‘consolidate.py‘ script.

3.3. Model Training

- Training Script: The model is trained using the ‘train.py‘ (generating training and validation loss plots with early stopping) script, employing triplet loss learning to optimize the feature representation for accurate compound identification.

- Model Storage: The trained model is saved as ‘best-model.pth‘.

3.4. Model Testing

- Testing Script: The ‘wizardchem.py‘ script tests the trained model on new images and videos, identifying compounds based on the generated feature vectors.

- Feature Matching: The model loads the best weights saved during training and uses cosine similarity to match extracted features with known compounds.

- Result Storage: The prediction results for each run are stored in ‘predictions.json‘ for qualitative results.

By detailing each component of the WizardChem framework, this section provides a comprehensive overview of the methodologies and technical aspects involved in automating chemical identification processes.

4. Experiments/Analysis

4.1. Experimental Setup

WizardChem's experimental setup involves training the model on bulk chemical data from PubChem, using 2D molecular structures to generate feature vectors. The primary focus is on image data, with plans to integrate video data in the future. The dataset includes compound CIDs and associated images, extracted and organized using the ‘data-org.py‘ and ‘consolidate.py‘ scripts.

4.2. Datasets

PubChem Data: The dataset comprises bulk chemical summaries and 2D molecular structures downloaded from PubChem using compound CIDs.

Training and Validation Sets: Data is split into training and validation sets, with the training set used to optimize the model and the validation set used to evaluate its generalization performance.

4.3. Quantitative Results

- Below are the training and validation loss plots for fold zero and fold one measured on June 10th, 2024.

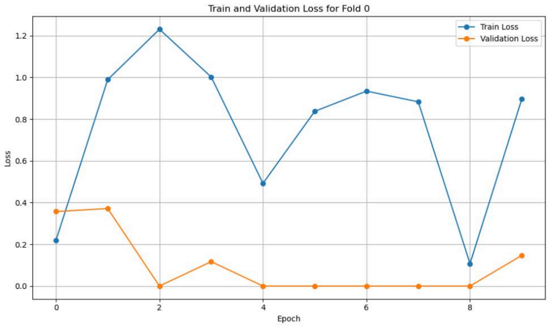


Figure 1. Fold 0

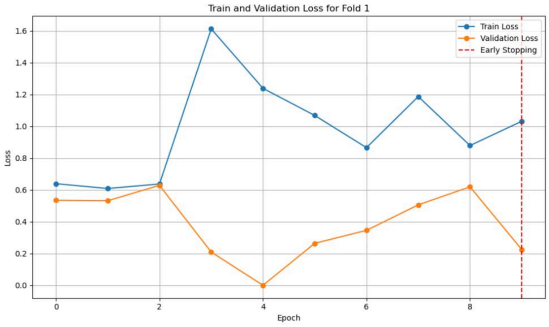


Figure 2. Fold 1

4.3.1 Evaluation Metrics

Training and Validation Loss: The primary metrics for evaluating the model's performance are the training and validation losses, measured using triplet loss. These losses indicate how well the model is learning to distinguish between similar and dissimilar compounds.

- Cosine Similarity: For testing, the model uses cosine similarity to match extracted features with known compounds.

4.3.2 Fold Zero (Figure 1)

- Training Loss: The training loss starts low, peaks at epoch 2, indicating initial overfitting, and then decreases significantly, suggesting the model is learning to generalize better over time. However, the fluctuations in training loss indicate some instability.

- Validation Loss: The validation loss starts low, slightly increases until epoch 1, then decreases significantly at epoch 2 and stabilizes at a low value. This pattern suggests good generalization despite the fluctuations in training loss.

4.3.3 Fold One (Figure 2)

- Training Loss: The training loss remains relatively stable initially, peaks dramatically at epoch 3, indicating instability, and then decreases, suggesting the model is adapting but with some fluctuations.

- Validation Loss: The validation loss increases slightly until epoch 2, then decreases sharply at epoch 4 and stabilizes at epoch 8, with a noticeable decrease at epoch 9, suggesting potential overfitting. The early stopping mechanism helps to mitigate this issue.

4.3.4 Training and Validation Loss Plots

The training and validation loss plots for both folds illustrate the model's learning process, highlighting periods of overfitting, adaptation, and generalization. These plots provide insights into areas where the model can be further optimized through additional data augmentation, alternative loss functions, and hyperparameter tuning.

4.4. Qualitative Results

The qualitative results are derived from the 'predictions.json' file, which contains the model's predictions on tested images.

4.4.1 Example Prediction

- Input Image: A 2D molecular structure image of aspirin was uploaded for testing (Figure 3).

- Prediction: The algorithm incorrectly identified the image as D-glucose (Figure 4).

- Output: Despite the incorrect identification, the model displayed the correct properties for D-glucose (Figure 5).

This example highlights the current challenge in the model's accuracy and the need for further improvements in data depth and feature extraction techniques:

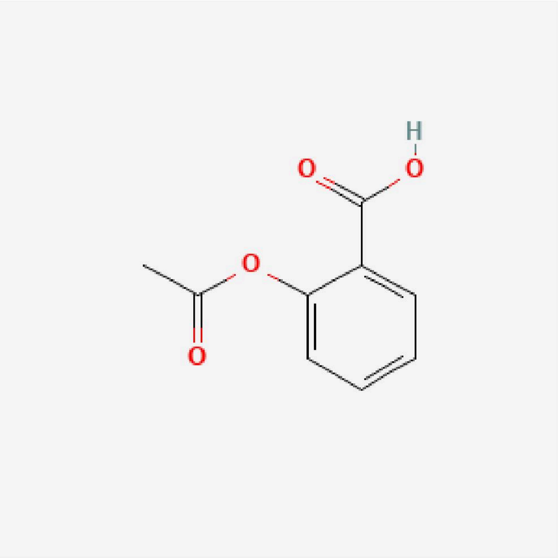


Figure 3. Aspirin (CID 2244)

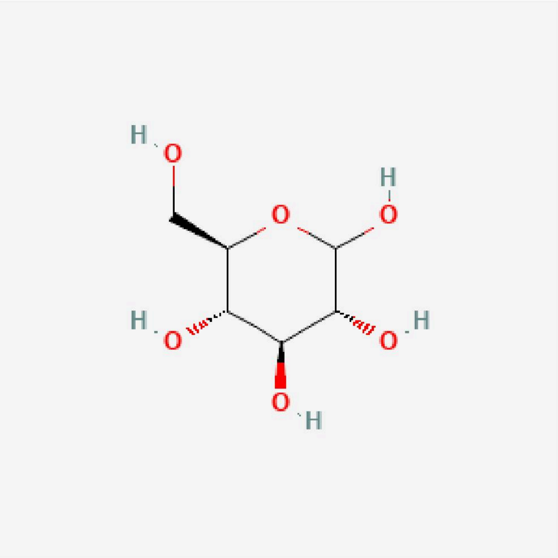
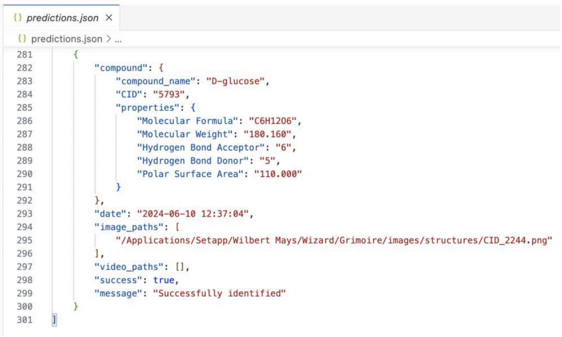


Figure 4. D-Glucose (CID 5793)

5. Conclusion and Future Work

5.1. Conclusion

The WizardChem project has made significant progress in automating chemical analysis and sorting using AI and computer vision. The preliminary results indicate promising advancements in accuracy, efficiency, and safety in chemical processing, although video functionality has not yet been tested.



```
    {
        "compound": {
            "compound_name": "D-glucose",
            "CID": "5793",
            "properties": {
                "Molecular Formula": "C6H12O6",
                "Molecular Weight": "180.160",
                "Hydrogen Bond Acceptor": "6",
                "Hydrogen Bond Donor": "5",
                "Polar Surface Area": "110.000"
            }
        },
        "date": "2024-06-10 12:37:04",
        "image_paths": [
            "/Applications/Setapp/Wilbert Mays/Wizard/Grimoire/images/structures/CID_2244.png"
        ],
        "video_paths": [],
        "success": true,
        "message": "Successfully identified"
    }
]
```

Figure 5. Failed Prediction

Key learnings from the project include:

- The importance of robust data preprocessing and augmentation techniques to enhance model performance.
- The effectiveness of using triplet loss to enforce desired distance relationships in the feature space.
- The necessity of a comprehensive dataset and the challenges posed by insufficient image data depth for individual compounds.

5.2. Future Work

5.2.1 Dataset Expansion

Microsoft COCO Integration: Implementing the Microsoft COCO dataset or finding another extensive source of image data specifically for chemistry and chemical compounds will be crucial. This expansion aims to increase the depth and variety of training data, thereby improving the model's robustness and generalization capabilities.

5.2.2 Model Optimization

- Alternative Loss Functions: To further enhance model performance and accuracy, future work will involve adjusting the learning method from the current triplet loss function to other potential loss functions such as Contrastive Loss or ArcFace Loss. These loss functions may offer better optimization for feature extraction and matching.
- Hyperparameter Tuning: Additional efforts will focus on fine-tuning hyperparameters, including learning rates, batch sizes, and dropout rates, to achieve greater model stability and performance.

5.2.3 GUI Enhancements

Enhanced Functionality: Future updates to the GUI will aim to make it more user-friendly and informative. Planned enhancements include displaying properties of identified compounds, providing summaries of important facts, information on where to purchase the compound, links to relevant research papers, and associated videos. This comprehensive tool will support chemical analysis and research, offering users a detailed and accessible interface.

By addressing these future work items, WizardChem aims to further transform chemical process monitoring and management, contributing to significant improvements in accuracy, efficiency, cost savings, and regulatory compliance.

This integration of AI and computer vision is not just a technical endeavor but a personal mission aligned with my long-term professional goals and passion for chemistry.

6. Link to Git Repository:

https://github.com/wilfm999/WizardChem.git

7. References

[1] C. Yan et al., ”Computer Vision for Non-Contact Monitoring of Catalyst Degradation and Product Formation Kinetics,” in *Chemical Science*, vol. 14, pp. 5323-5331, 2023. Available: https://www.semanticscholar.org/paper/548dfc701f13d626602eefdad5fdd31f24d247bc

[2] J. Schell, S. C. McCauley, and R. Glaser, ”Video Colorimetry of Single-Chromophore Systems Based on Vector Analysis in the 3D Color Space,” *Talanta*, vol. 220, 121303, 2020. Available: https://www.semanticscholar.org/paper/69bd78051ad391d3b3c84831f6558608bce4435b

[3] J. Agrisuelas, J. J. García-Jareño, E. Guillén, and F. Vicente, ”Kinetics of Surface Chemical Reactions from a Digital Video,” *Journal of Physical Chemistry C*, vol. 124, pp. 2050-2059, 2020. Available: https://www.semanticscholar.org/paper/c93f36275f68db3caf25a54b198350ef36cd07a2

[4] R. Glaser, M. Downing, E. Zars, J. Schell, and C. Chicone, ”Video-Based Kinetic Analysis of Period Variations and Oscillation Patterns in the Ce/Fe-Catalyzed Four-Color Belousov–Zhabotinsky Oscillating Reaction,” *ACS Symposium Series*, 2019. Available: https://www.semanticscholar.org/paper/0e050c210244aaba9bd7229ce8efec8f2600864b

[5] B. Förster and T. Hinze, ”Towards Automated Analysis of Belousov-Zhabotinsky Reactions in a Petri Dish by Membrane Computing Using Optic Flow,” *Int. Conf. on Membrane Computing*, pp. 131-141, 2018. Available: https://www.semanticscholar.org/paper/f3421128afe9af92ce1995fff00808235fefad8b

[6] Y. Shi, P. L. Prieto, T. Zepel, S. Grunert, and J. Hein, ”Automated Experimentation Powers Data Science in Chemistry,” *Accounts of Chemical Research*, 2021. Available: https://doi.org/10.1021/acs.accounts.0c00736

[7] L. Capitán-Vallvey, N. López-Ruiz, A. Martínez-Olmos, M. M. Erenas, and A. Palma, ”Recent developments in computer vision-based analytical chemistry: A tutorial review,” *Analytica Chimica Acta*, vol. 899, pp. 23-56, 2015. Available: https://doi.org/10.1016/j.aca.2015.10.009