

# Cell2Structure – Report

Florian Gillet, Matthias Hübscher

## Introduction



One important task in drug discovery is the detection and exploration of the influence potential drugs have on cells in culture. Various techniques have been developed to study the influence of small chemical compounds on cells to discover their mechanism of action (MoA). An increasingly important technique in this field is high-content screening (HCS). In this screening method cells are treated with a large amount of different chemical compounds for a fixed period e.g. 24h. This treatment induces changes in the appearance of the cells, their phenotype. After the treatment the changes in the phenotype of the cells are imaged using a fast microscopy technique, high-throughput microscopy (Götte et al.,

2010). In one screen often 10<sup>th</sup> or even 100<sup>th</sup> of thousands of images are produced. These screens have the advantage of providing rich information about the mechanism of the treatment, including different treatment targets and cellular pathways (Steigele et al., 2020).

Various tools and techniques have been and are still developed to analyze images from high-content screens. CellProfiler for example is an open-source tool developed to address a variety of biological questions including the study of the phenotype of cells (Carpenter et al., 2006). Modern techniques apply convolutional-neural networks (CNN) to achieve this task. CNNs can be used as a supervised task (Warchal et al., 2019) or as an unsupervised task included in a deep clustering model (Janssens et al., 2021). In our project we plan to use a pre-trained CNN to analyze the images and to generate embedding vectors. These embedding vectors generated by the CNN trained to recognize the MoA, contain encoded information of the cell structure post treatment. We assume that these cell modifications are a signature of the chemical compound used to treat the cells and that they contain some information about the chemical structure. This information could be used in multiple different ways. Our idea is to use it as a filter in a generative model generating chemical structures allowing for the design of compounds for a specific MoA.

The objective of our project is to answer the following high-level question:

- Can cell image embedding vectors be used as a translator between the image space and the space of chemical structures?

## The BBBC021 Dataset

In our project we use the publicly available BBBC021 dataset from the Broad Bioimage Benchmark Collection (*Human MCF7 Cells – Compound-Profilig Experiment*, n.d.). It consists of microscopy cell images acquired using a high content screen. This dataset was used in numerous research projects. It is described in Ljosa et al. (Ljosa et al., 2012).

Example images are provided in Figure 1.

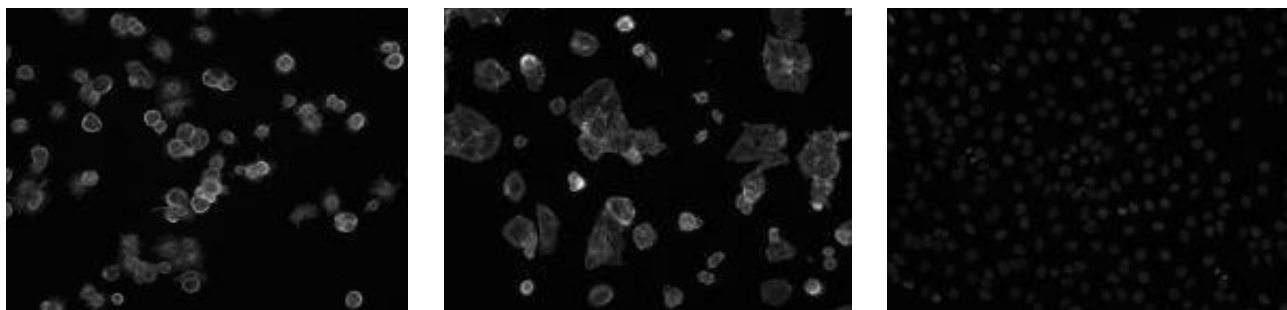


Figure 1: Examples of cell images. Left: Tubulin; Center: Actin; Right: DAPI

## Dataset description

Each cell image was acquired three times with different fluorescence channels: Actin, Tubulin, and DAPI. Actin and Tubulin, proteins of the cytoskeleton, are stained to highlight cell shape. DAPI binds to DNA, highlighting the nucleus. These channels are shown on a grey scale in Figure 1 and provide an overview of cell structural changes induced by compound treatment.

Each image comes with metadata in 3 different CSV files, including for example the compound name and, if known, the MoA. The dataset includes 12 known MoA and some unknown ones. Some images were acquired using Dimethyl sulfoxide (DMSO) as treatment, an organic solvent used to dissolve all chemical compounds in the experiment. DMSO serves as negative control.

## Origin of the cells

Cells used in the cell culture experiment are derived from the MCF-7 breast cancer cell line. The cells were originally donated by Sister Catherine Frances (Helen Marion) Mallon, a nun at Immaculate Heart of Mary convent in Monroe, Michigan. According to her convent's archive, Sister Catherine knew she had terminal breast cancer and, when asked for help by her doctor Michael Brennan, agreed to the cell donation procedure hoping to participate in the research effort against cancer (Crane & Brennan, 2013).

## Dataset filtering

In our project we used only images treated with a chemical compound of known MoA. We therefore filtered out any images which aren't linked to a MoA. The original dataset contains 39,600 images. We reduced it to approximately 6000 images due to this constraint, for 12 known MoA, and 1 negative control (no MoA).

# Methods

## Molecular Fingerprints and Descriptors Calculation

Molecular fingerprints are a way to represent chemical structures as a vector, making them easily accessible to data analysis and machine learning. Various types of fingerprints exist. In our project we use Morgan fingerprints. These encode the circular neighborhoods around each atom in a structure into a bit vector, using the Morgan algorithm (Morgan, 1965). Its main purpose is to compare structural similarity. The Morgan fingerprint of each compound structure was calculated using the

RDKit Python library (Landrum, n.d.). RDKit is an open-source cheminformatics library widely used in academia and industry.

Besides the molecular fingerprints we also calculated molecular descriptors using RDKit. Their main use is the classification of chemical structures. Like fingerprints, they are commonly used in data analysis and machine learning and include among others for example the molecular weight and polar surface area, a measure for the distribution of polar atoms in chemical structures (Ertl et al., 2000). In our project we decided to take all the RDKit descriptors into account.

## Distance Matrices Generation

The distance matrix of the molecular fingerprints of the chemical compounds was calculated using Tanimoto similarity (Bajusz et al., 2015), the standard way to compare the chemical structures represented as bit vectors. The distance matrix of the molecular descriptors was based on the cosine similarity of the descriptor vectors.

Both distance matrices are visualized using a heatmap of the pairwise distance of the chemical structures.

## Pre-Trained CNN fine-tuning

### Dataset Preparation

To avoid data leakage, we split the dataset in a way that none of the compounds appeared in both training and test sets by keeping the MoA labels in all sets. In addition, we balanced the training and test sets to ensure no overrepresentation of images per MoA.

### CNN Fine-Tuning

Based on a 2022 review by Kim et al. (Kim et al., 2022), 425 peer-reviewed articles highlighted the use of transfer learning with deep models like ResNet or Inception for medical applications. We decided to use the more recent Inception V3 (Szegedy et al., 2015) model for our project.

We first used the pre-trained instance of Inception V3 available via PyTorch and retrained the final fully connected layer. We screened parameters with different optimizers (SGD and Adam) and learning rates (0.001 to 0.1), monitoring cross-entropy loss over 10 epochs to pre-select the best conditions.

To evaluate options for fine-tuning we tried to identify which convolutional layers to include by visualizing the average activation map of each layer to find layers detecting low level object-specific features. We then unfroze these layers for further training over 20 to 25 epochs, aiming for maximum test set accuracy. The resulting models were exported for further use in the embedding vector generation.

## Embedding Vectors Generation

The image embedding vectors were generated using our trained Inception V3 CNNs. We extracted the vectors from the outputs of the last pooling layer.

## Embedding Vector Batch Correction

Due to experimental conditions such as measurements on different days or slight temperature fluctuations between measurements the embedding vectors can show statistically significant variations. To account for this, we applied a batch correction to correct batches of experimental measurements. We used a recent implementation of the PyCombat tool provided in the InMoose library (Colange, n.d.). PyCombat is widely used to correct batch effects in various biological Omics experiments.

## Embedding Vectors Analysis

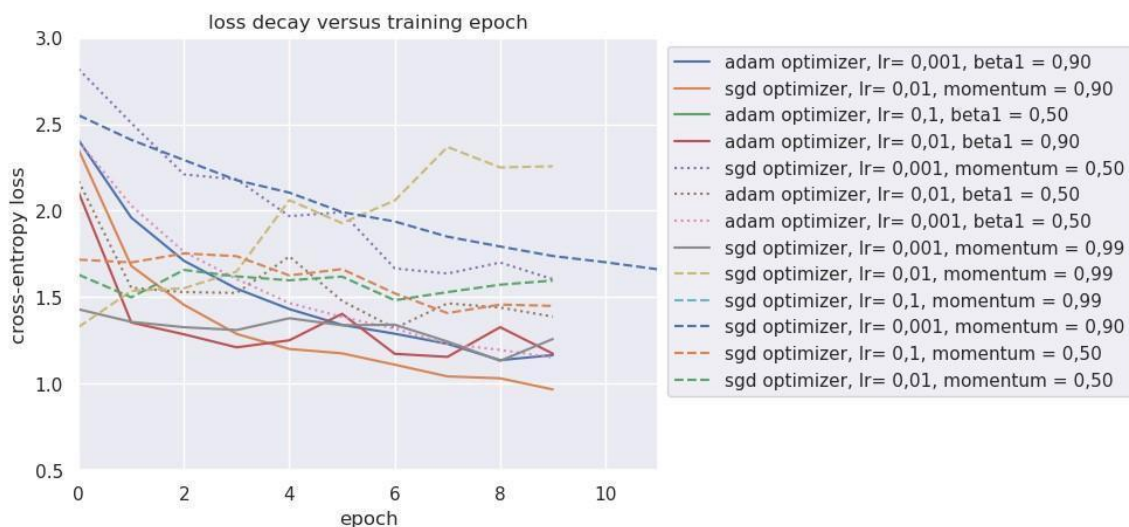
The distribution of individual embedding vectors for each compound was measured using cosine similarity. The vectors were grouped per compound, and the mean vector was calculated for each compound. The cosine similarity of the individual vectors to the mean vector for the corresponding compound was calculated. The distribution of these cosine similarities for each compound was analyzed visually using bar and box plots, to evaluate how similar vectors corresponding to each compound were to each other. In addition, we applied the simple IQR rule as a robust method to detect and remove outlier vectors.

For the final analysis a distance matrix, like the ones of the molecular fingerprints, was calculated based on the pairwise cosine distance of the mean compound vectors and visualized using a heatmap.

## Results

### CNN finetuning results

As a first step, Inception V3 was entirely frozen except for the last fully connected layer. The model was trained on our dataset, with different sets of parameters and optimizers. The training loss monitored over 10 epochs of training is plotted in Figure 2.



**Figure 2:** Cross entropy loss evolution over 10 epochs for different optimizers/parameters combinations

Based on curves displayed in Figure 2, the loss decay was slow independent of the tested parameters. Tuning only the last fully connected layer seemed to be a suboptimal approach. To optimize model training, we aimed at identifying convolutional layers which detected low level features. We used images representing objects that are recognized by the native inception v3 model: a cat, and a lighthouse. For these two images, we extracted the average activation maps of each convolutional layer and overlaid them with the original image. We did the same with an image extracted from our dataset. A representative image of the average activation map of one convolutional layer from the set of layers Mixed\_7b, and Mixed\_7c, for the cat, the lighthouse, and the cell images is provided in Figure 3.

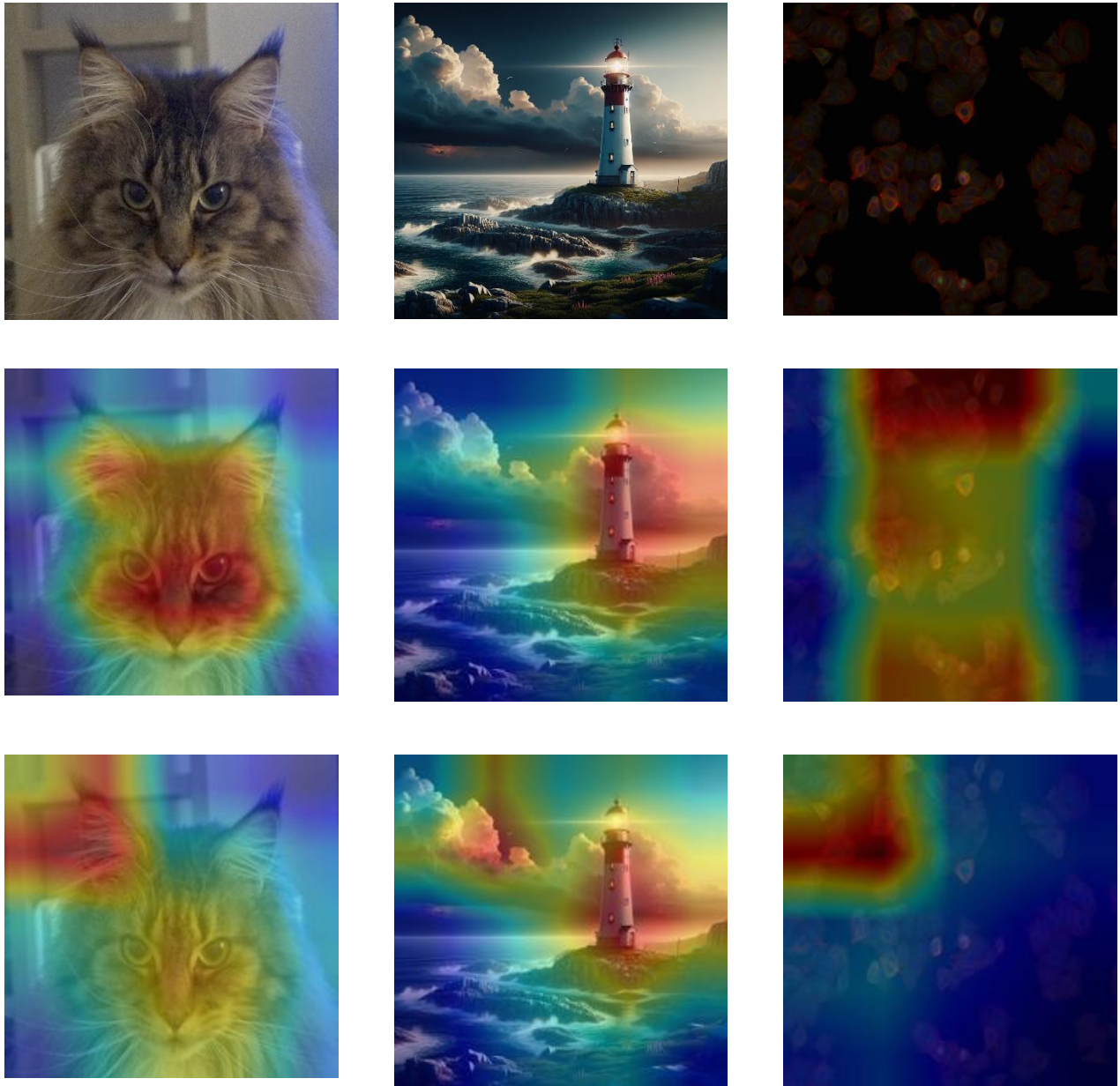
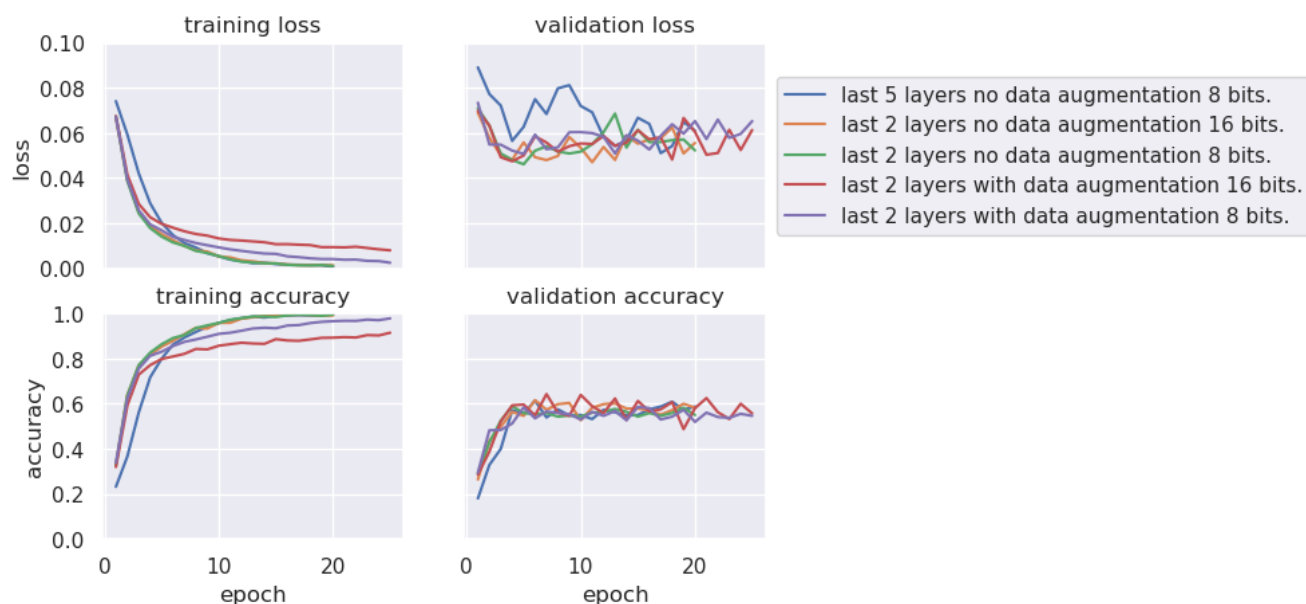


Figure 3: Images and activation maps. First row: original images. Second row: activation map of a set of convolutional layers “Mixed\_7b”. Third row: activation map of set of convolutional layers “Mixed\_7c”.

Based on Figure 3, the layers Mixed\_7b and Mixed\_7c detect low-level features in both the cat and lighthouse images, such as the cat's face and ears, and the lighthouse building and clouds. However, for cell pictures, the activation maps appeared randomly, indicating that our model does not recognize cells. While earlier convolutional layers detecting edges or general shapes can be kept, Mixed\_7b and Mixed\_7c need fine-tuning on our dataset. We unfroze these layers for the second round of training. The new training approach, under optimal conditions, yielded loss and accuracy versus epoch curves in Figure 4. All models were trained with the Adam optimizer, a learning rate of 0.0001, and a batch size of 32, with L2 regularization applied. Some models were trained with data augmentation (random image flips) and some using 16-bit images, while others were converted to 8-bit.



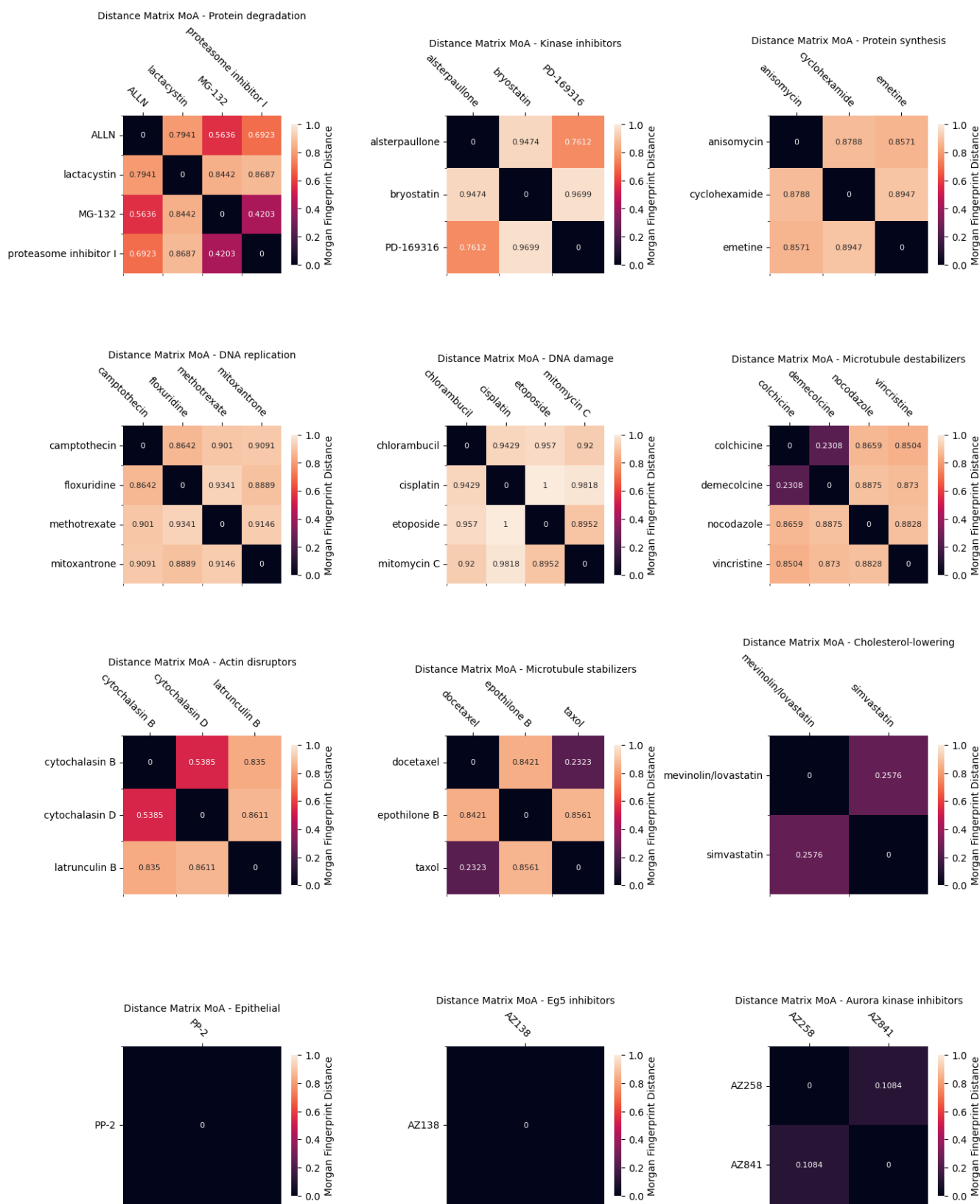
**Figure 4:** Loss versus epoch (left), and accuracy versus epoch (right)

Curves displayed in Figure 4 showed that while loss decreased and accuracy increased steadily on the training set, they both seemed to reach a plateau for the validation set. The maximum validation accuracy obtained was 60%, regardless of the model. This showed that all models were overfitting the training set.

## Chemical Structures Analysis

Examples of chemical structure depictions are available in appendix (Figure 9). Morgan molecular fingerprints were computed for each chemical compound and the pairwise Tanimoto distance between them was calculated. Distance matrices per MoA are available in Figure 5. The overall distance matrix is available in Figure 6.





**Figure 5:** Distance matrix of compounds Morgan fingerprint per MoA

Distance Matrix of BBBC021 compounds based on Morgan Fingerprints  
Radius = 2, Fingerprint size = 2048

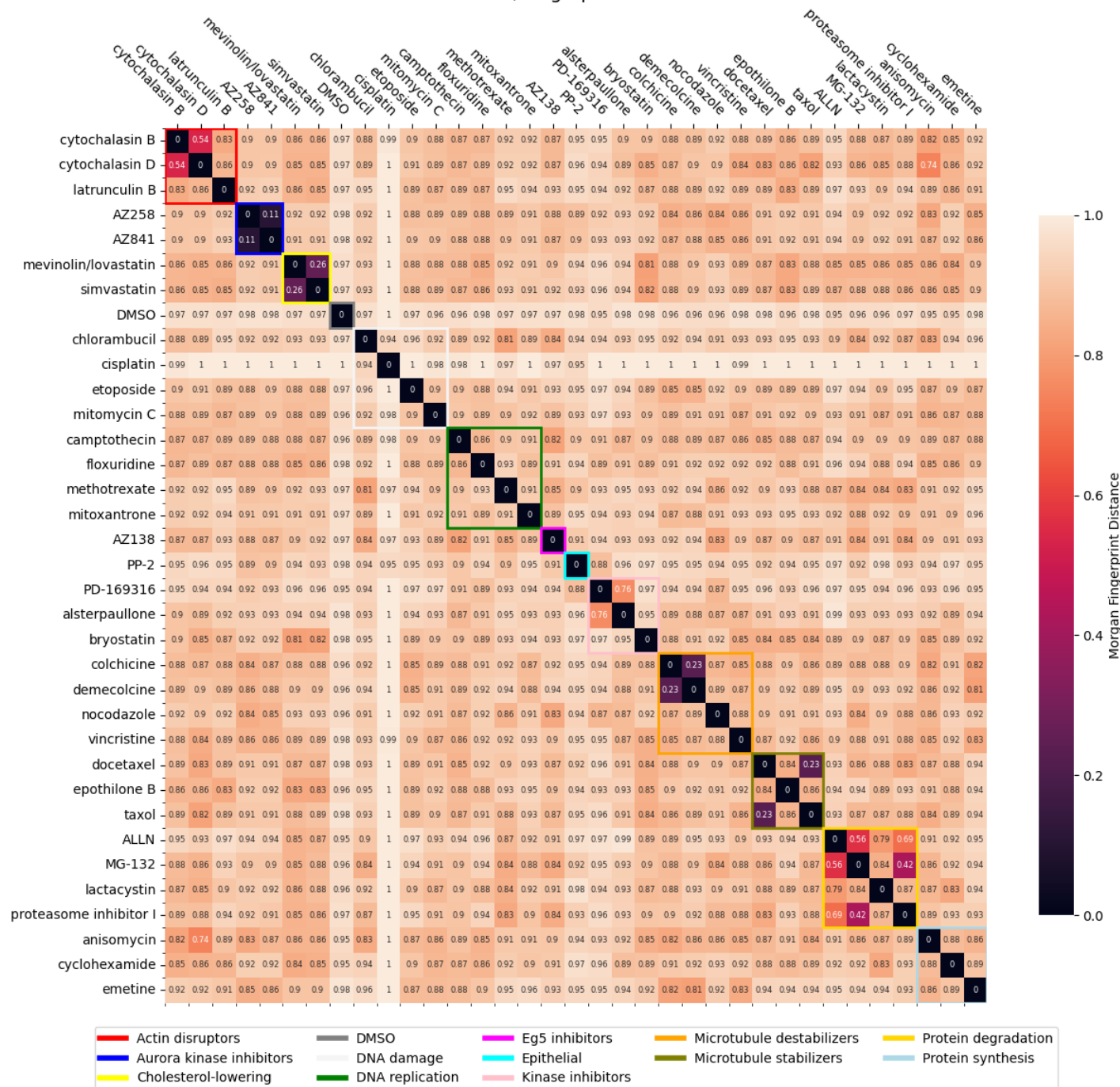


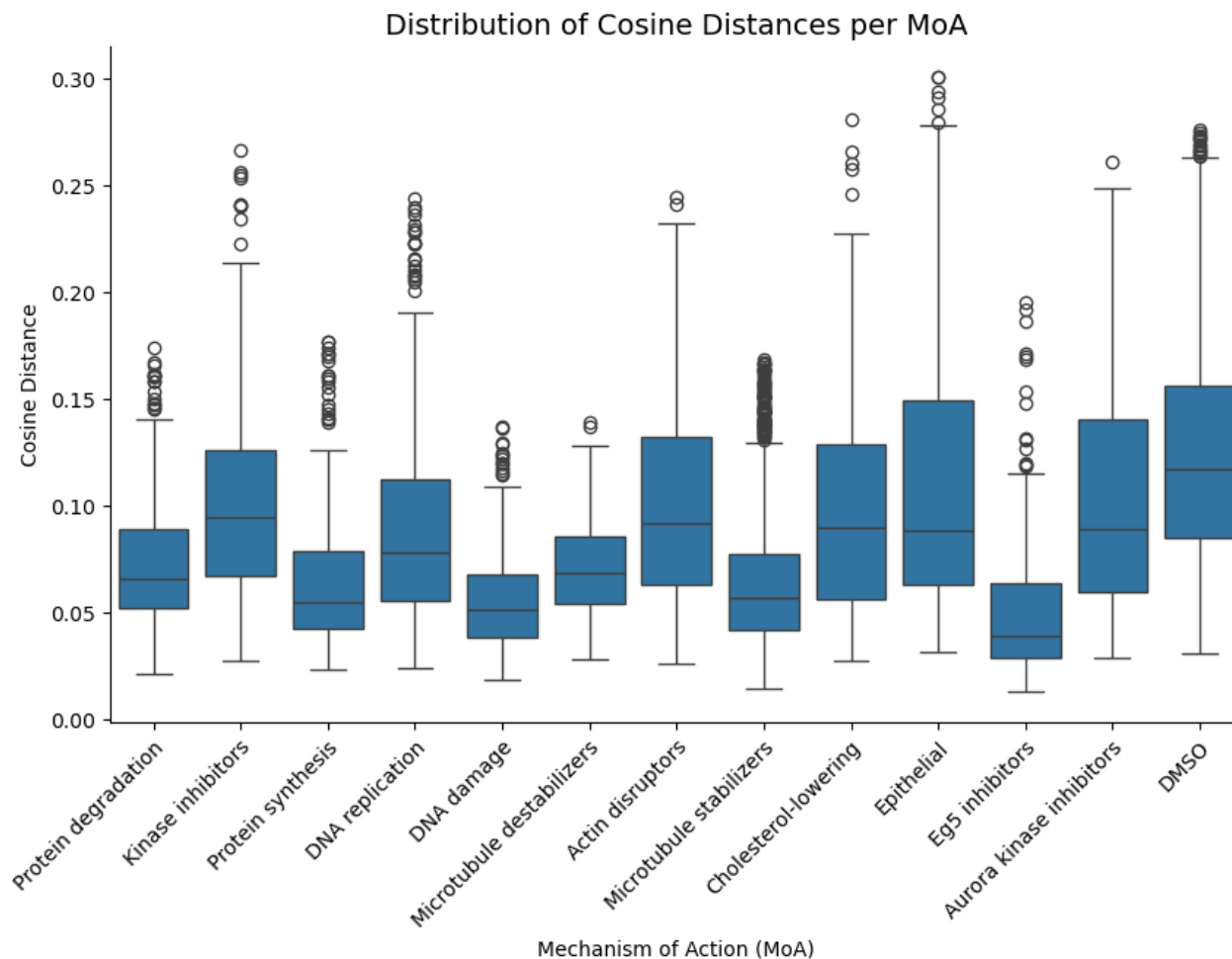
Figure 6: Distance matrix of Morgan Fingerprints of BBBC021 compounds, based on Tanimoto distance

The distance matrices showed that compounds associated with the same MoA have a low similarity in this dataset. The distribution of the compounds is very heterogeneous in the dataset even though a few MoAs such as Aurora kinase inhibitors or Cholesterol-lowering contain somehow similar ones.



## Image Embedding Vectors Analysis

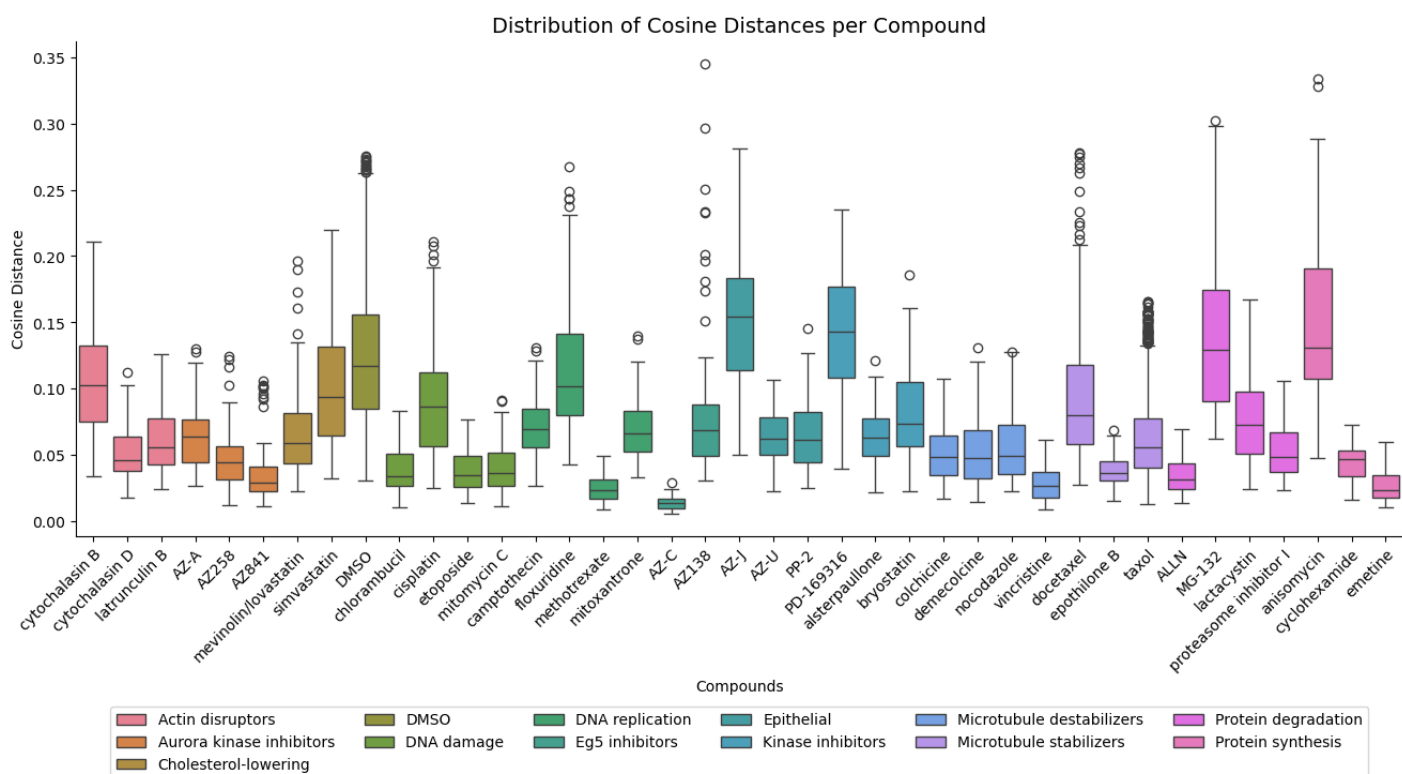
The embedding vectors were grouped by MoA. For each MoA group, a mean vector was computed. Within each group, the cosine distance between any individual vector and the corresponding mean vector was calculated. The box plots representing the distributions of the cosine distances for each MoA group are plotted in Figure 7.



**Figure 7:** Distribution of cosine distance to mean vector for each MoA.

The distributions show that in most cases the vectors are similar to each other within a MoA group. This was expected, given that the classifier used to generate these vectors was trained to recognize MoAs. However, we discovered a large number of outliers.

The embedding vectors were also grouped by compound. For each group, the mean vector was computed. Within each group, the cosine distance between any individual vector and the corresponding mean vector was calculated. The distribution of the cosine distances is plotted in Figure 8.



**Figure 8:** Distribution of cosine distance to mean vector per compound.

Results show that the distributions are skewed to the left, with most compounds having a median below 0.1. At least half of the embedding vectors for most compounds are very similar to their respective mean vectors, indicating that the generated embeddings are highly similar within each compound group.

We used T-SNE and UMAP clustering for alternative visualization, which provided similar observations. They are included in the appendix (Figures 12 and 13). We noticed that many compounds had a large proportion of vectors, more than 90%, similar to their mean vectors. Shown in the appendix (Figure 13). Attempts to eliminate outliers slightly narrowed the distributions.

# Comparison of structure distance versus embeddings distance

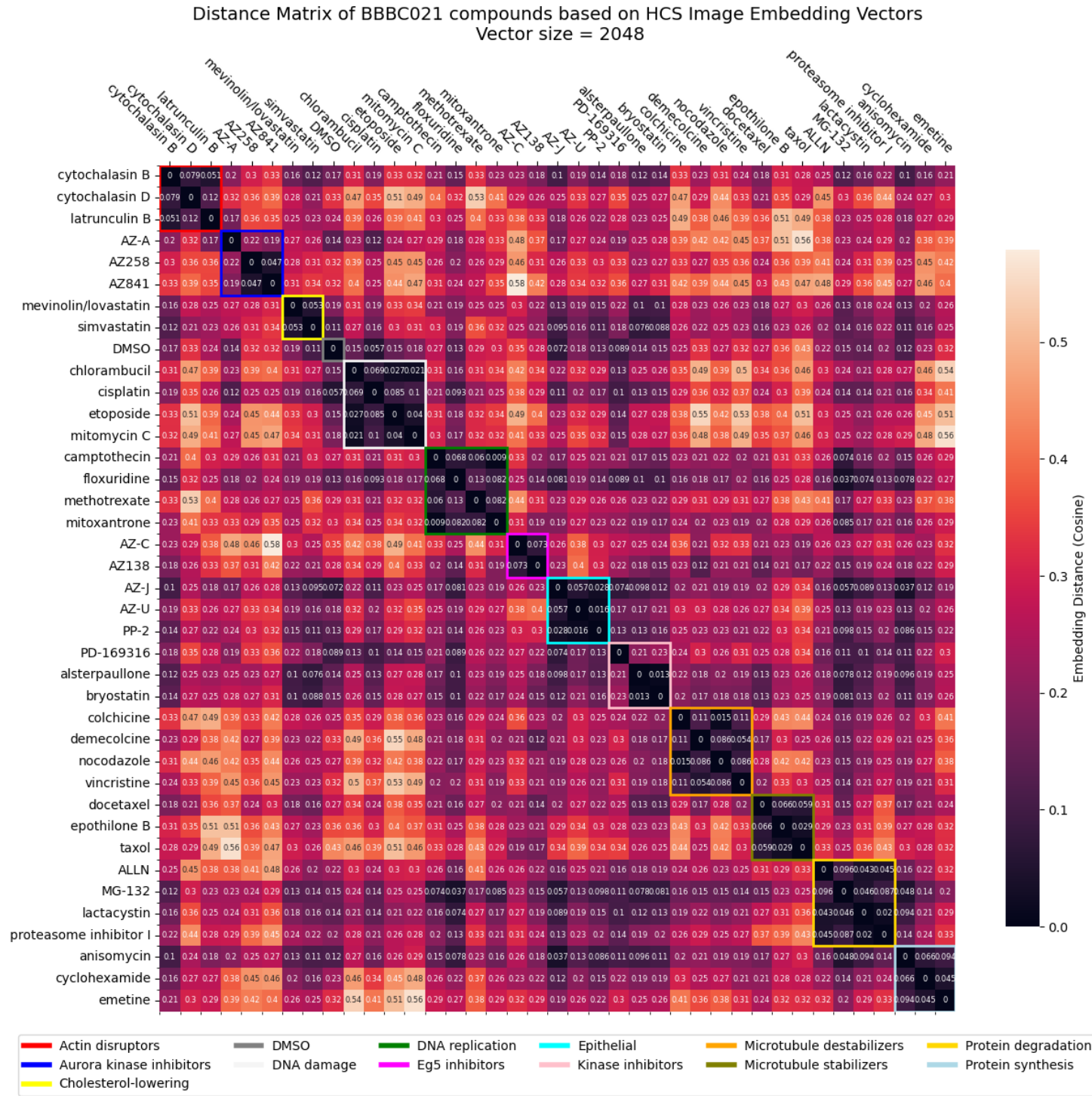


Figure 9: Distance matrix of image embedding vectors, based on Cosine distance

The heatmap shows dark squares along the diagonal, showing that compounds sharing a similar MoA have more similar embeddings. This signal is largely absent in the heatmap based on the comparison of molecular fingerprints in Figure 6.

# Discussion

The intention was to train a classifier based on the Inception V3 CNN architecture, to classify MoAs based on input images. Our idea was to use the performance of the classifier as a proxy for the quality of the underlying image embedding vectors. Despite a very good training accuracy, the maximum accuracy we obtained on the validation set was ~60%. The outliers we discovered could be one explanation for the difficulty obtaining a performant classifier. Compounds vary in their bioactivity even when assigned to the same MoA. Low bioactivity will lead to images close to the negative control (DMSO).

Because our main objective was to generate image embedding vectors that represent MoAs, we decided to not further improve the classifier. Mainly due to time constraints we chose the best model for this task. The training accuracy was beyond 95%, indicating that the generated embeddings would be representative of MoAs. We accepted that the model is not generalizable to other dataset and most likely contains quite some noise. The data leakage investigation originally planned was not performed due to the low accuracy on the validation set, which made such investigation irrelevant.

The embedding vectors were similar (based on cosine similarity) within MoA groups. This was expected because the model we used to generate them was trained to recognize MoAs. We noticed that when grouping the embeddings by compound a similar picture appears. Embeddings of one compound are very similar to each other. We interpret this as an indication that image embedding vectors generated with our approach likely contain information of compounds responsible for producing the changes in the cell's phenotype. When comparing the distance matrices between the molecular fingerprints and the image embedding vectors, we can show that embeddings better cover the bioactivity compared to fingerprints. Morgan fingerprints cover the features of the chemical structure, the structure graph whereas embedding can cover a wide range of features from the image and the chemical structure.

**Our observations can be used to answer the question:** Image embedding vectors of cell images seem to contain information about the compound that was used to produce these cell images.

We see a potential of using embeddings in generative chemistry models either to generate chemical structures directly or as another dimension to represent the bioactivity of compounds.

## Outlook

As a next step we would propose improving the quality of the embedding vectors. Our results showed that despite the general similarity between vectors obtained from the same compound, a high number of outliers was observed. Multiple reasons for this behavior might exist such as variations in bioactivities probably due to different concentrations or artifacts in the images. Our outlier detection and removal method could be changed to a more sophisticated one like density-based outlier detection. In such a setup a second training step after removing the outlier images would likely improve the quality of the embeddings.

Our objective was to provide some evidence that embedding vectors contain chemical structure information. We basically tried to build an encoder for a generative model. An extension of this project could be to build the decoder to enable the generation of chemical structures. To achieve this, we think that a thorough analysis of the embedding vectors could be necessary to discover the correlation between the embeddings and features of the chemical structure.

## Statement of work

Matthias Hübscher	Florian Gillet
Chemical structure analysis CNN finetuning Embeddings generation Embedding analysis GitHub repository Readme file Standup recording and posting	CNN activation maps computing CNN finetuning Report writing Poster writing Readme file

## Ethical statement

The images in our dataset are of cells from the MCF-7 breast cancer cell line, a popular research cell line that has been used in tens of thousands of publications (Lee et al., 2015). As mentioned above, the cells were originally donated by Sister Catherine Frances (Helen Marion) Mallon. However, public information on Sister Catherine's participation in the creation of this cell line is limited to a 2013 convent archives newsletter (Crane & Brennan, 2013), so the degree she was able to provide informed consent cannot be certain.

The models we built aimed at identifying MoAs for in vitro research purposes. Bias and fairness, as well as responsible data usage are therefore not a concern in our case. The dataset was publicly available, and the CNN we used to produce the embeddings was published. Therefore, there is no concern related to intellectual property.



# References

- Bajusz, D., Rácz, A., & Héberger, K. (2015). Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics*, 7(1), 20.  
<https://doi.org/10.1186/s13321-015-0069-3>
- Carpenter, A. E., Jones, T. R., Lamprecht, M. R., Clarke, C., Kang, I. H., Friman, O., Guertin, D. A., Chang, J. H., Lindquist, R. A., Moffat, J., Golland, P., & Sabatini, D. M. (2006). CellProfiler: Image analysis software for identifying and quantifying cell phenotypes. *Genome Biology*, 7(10), R100. <https://doi.org/10.1186/gb-2006-7-10-r100>
- Colange, M. (n.d.). InMoose (Version 0.7.2) [Python; Linux]. <https://inmoose.readthedocs.io/en/stable/>
- Crane, A., & Brennan, M. (2013). Catherine Frances Mallon's "Immortal Cells" [PDF].  
[https://ihmsisters.org/wp-content/uploads/2021/03/Archives-Notes\\_june-2013.pdf](https://ihmsisters.org/wp-content/uploads/2021/03/Archives-Notes_june-2013.pdf)
- Ertl, P., Rohde, B., & Selzer, P. (2000). Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties. *Journal of Medicinal Chemistry*, 43(20), 3714–3717.  
<https://doi.org/10.1021/jm000942e>
- Götte, M., Hofmann, G., Michou-Gallani, A.-I., Glickman, J. F., Wishart, W., & Gabriel, D. (2010). An imaging assay to analyze primary neurons for cellular neurotoxicity. *Journal of Neuroscience Methods*, 192(1), 7–16. <https://doi.org/10.1016/j.jneumeth.2010.07.003>
- Human MCF7 cells – compound-profiling experiment (BBBC021; Version 1). (n.d.). [Images]. Broad Bioimage Benchmark Collection. Retrieved October 2, 2024, from  
<https://bbbc.broadinstitute.org/BBBC021>
- Janssens, R., Zhang, X., Kauffmann, A., De Weck, A., & Durand, E. Y. (2021). Fully unsupervised deep mode of action learning for phenotyping high-content cellular images. *Bioinformatics*, 37(23), 4548–4555. <https://doi.org/10.1093/bioinformatics/btab497>

- Kim, H. E., Cosa-Linan, A., Santhanam, N., Jannesari, M., Maros, M. E., & Ganslandt, T. (2022). Transfer learning for medical image classification: A literature review. *BMC Medical Imaging*, 22(1), 69. <https://doi.org/10.1186/s12880-022-00793-7>
- Landrum, G. (n.d.). *RDKit* (Version 2024.09.1) [C++, Python; Linux]. <https://rdkit.org/docs/index.html>
- Lee, A. V., Oesterreich, S., & Davidson, N. E. (2015). MCF-7 Cells—Changing the Course of Breast Cancer Research and Care for 45 Years. *JNCI Journal of the National Cancer Institute*, 107(7), djv073–djv073. <https://doi.org/10.1093/jnci/djv073>
- Ljosa, V., Sokolnicki, K. L., & Carpenter, A. E. (2012). Annotated high-throughput microscopy image sets for validation. *Nature Methods*, 9(7), 637–637. <https://doi.org/10.1038/nmeth.2083>
- Morgan, H. L. (1965). The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *Journal of Chemical Documentation*, 5(2), 107–113. <https://doi.org/10.1021/c160017a018>
- Steiglele, S., Siegismund, D., Fassler, M., Kustec, M., Kappler, B., Hasaka, T., Yee, A., Brodte, A., & Heyse, S. (2020). Deep Learning-Based HCS Image Analysis for the Enterprise. *SLAS Discovery*, 25(7), 812–821. <https://doi.org/10.1177/2472555220918837>
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2015). *Rethinking the Inception Architecture for Computer Vision* (Version 3). arXiv. <https://doi.org/10.48550/ARXIV.1512.00567>
- Warchal, S. J., Dawson, J. C., & Carragher, N. O. (2019). Evaluation of Machine Learning Classifiers to Predict Compound Mechanism of Action When Transferred across Distinct Cell Lines. *SLAS Discovery*, 24(3), 224–233. <https://doi.org/10.1177/2472555218820805>

## Appendix

The appendix contains mainly additional visualizations we created during the project. These visualizations represent parts of our work which we decided not to include in the main report. Even though we consider them important they don't add much to the interpretation of the results and the overall story of our project.

### Example Structures From The BBBC021 Dataset


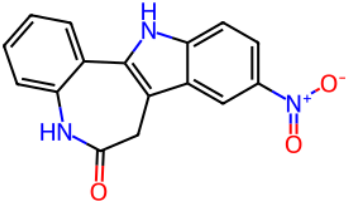
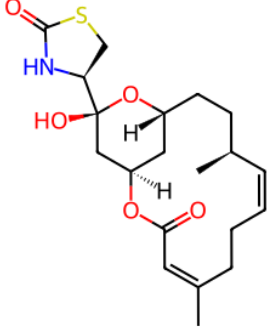
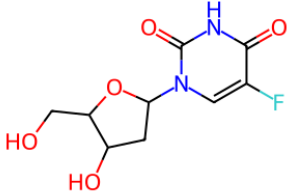
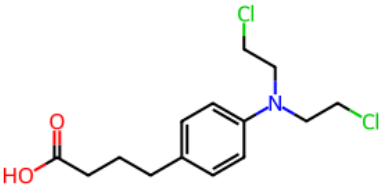
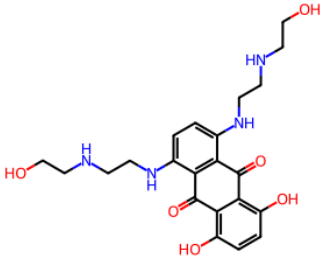
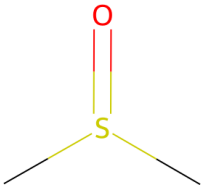
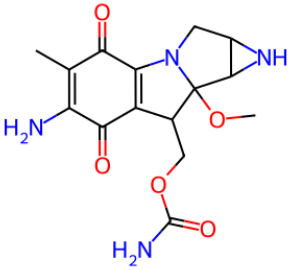
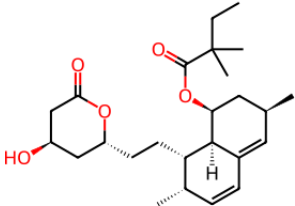
		
AZ841	Alsterpaullone	Latrunculin B
		
Floxuridine	Chlorambucil	Mitoxantrone
		
Dimethyl sulfoxide (DMSO)	Mitomycin C	Simvastatin

Figure 9: Compound structures from BBBC021

# Distance Matrix Of The Molecular Descriptors

Even though the distance matrix of the molecular descriptors shows some interesting patterns, we decided to leave it out of the main report. Instead, we used fingerprints, which can be seen as a kind of embedding vector like the image embedding vectors. Nevertheless, analyzing the descriptors and mapping them to the image embedding could be an interesting yet very complex extension of our project.

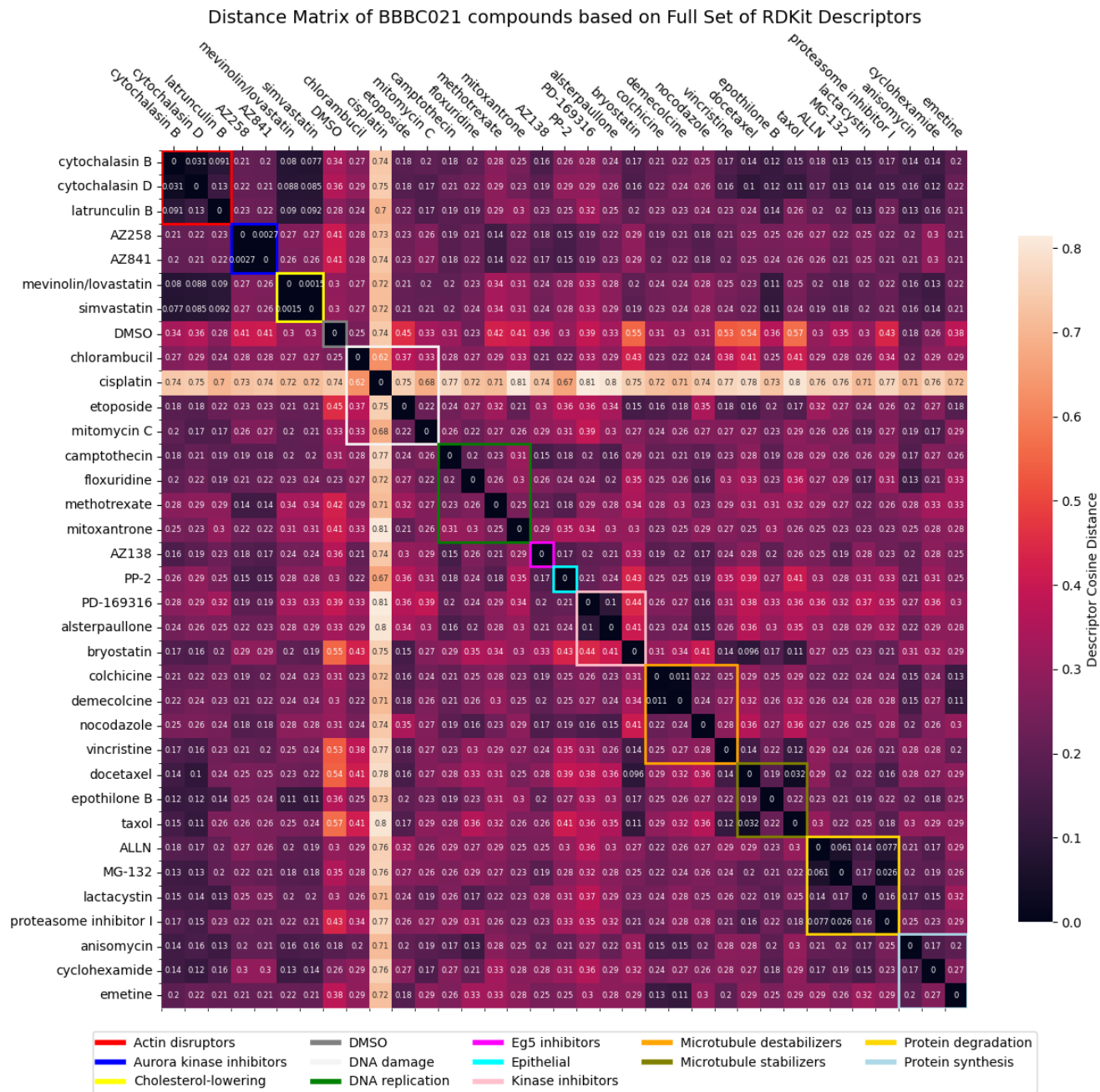


Figure 10: Cosine distance between compounds based on molecular descriptors

# Percentage Of Embedding Vectors > 90%

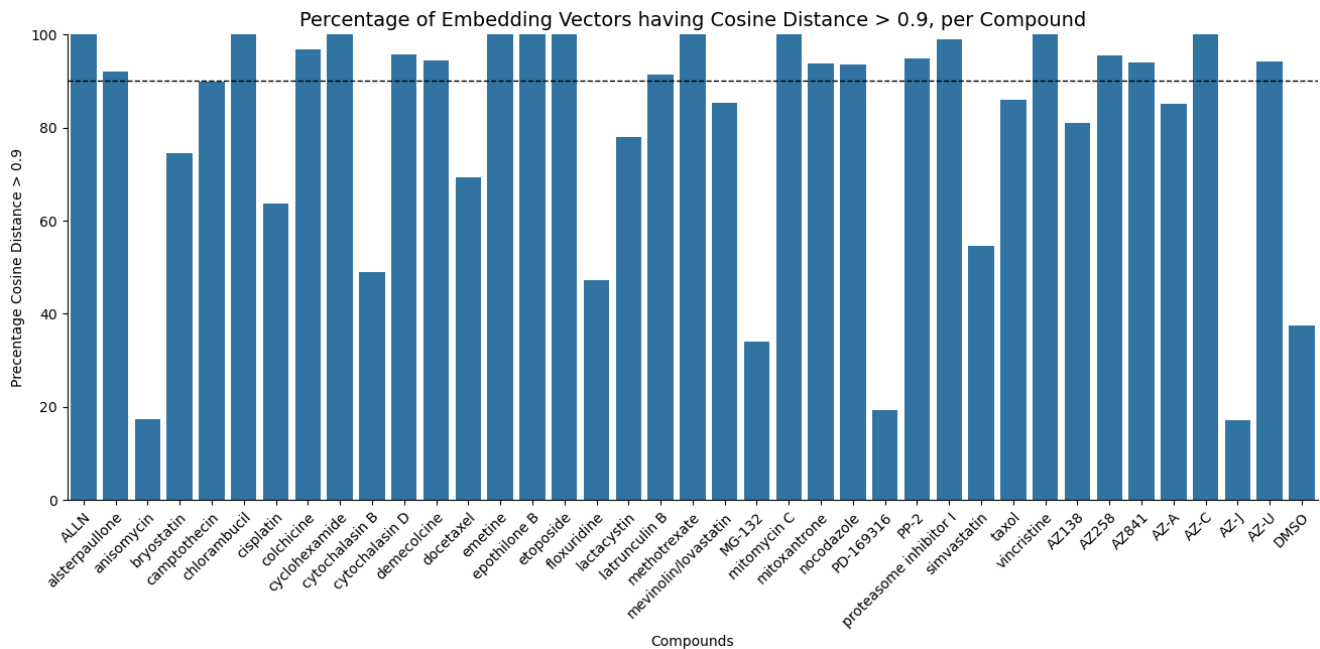


Figure 11: Percentage of embedding vectors being at least 90% similar to their respective mean vector, for each compound.

## Dimensionality Reduction And Clustering

We applied t-SNE and UMAP on the mean image embedding vectors to explore the clustering of the chemical compounds.

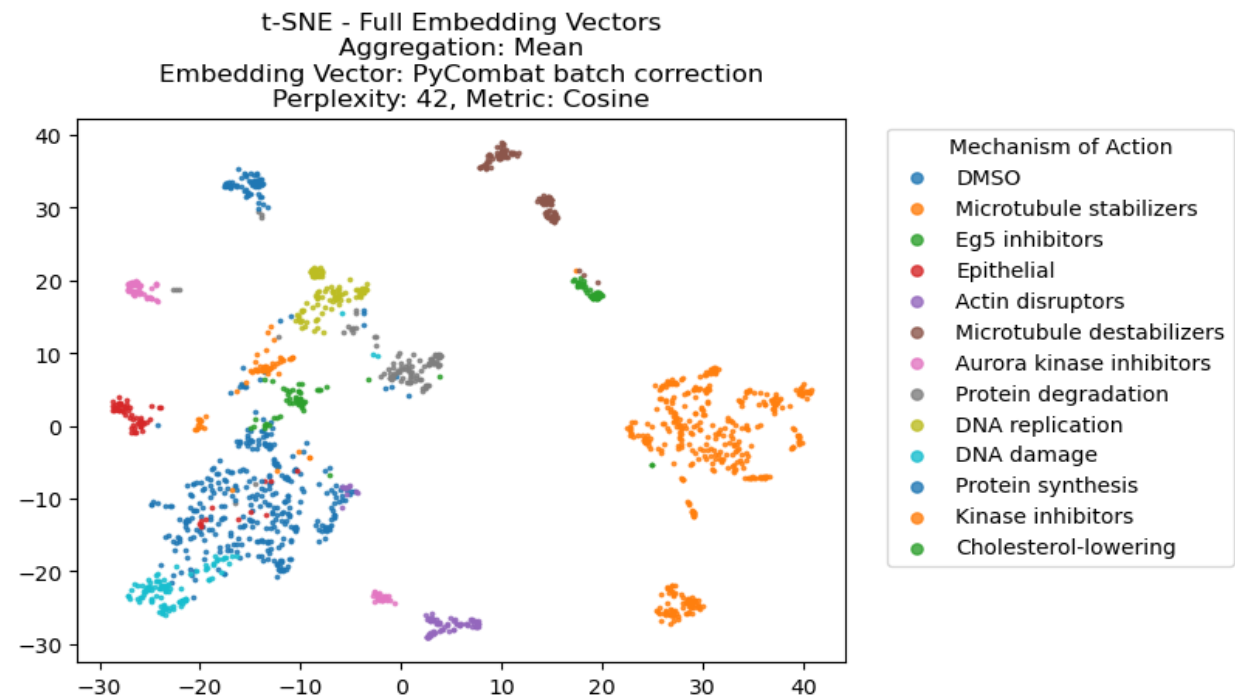


Figure 12: t-SNE clustering of embedding vectors



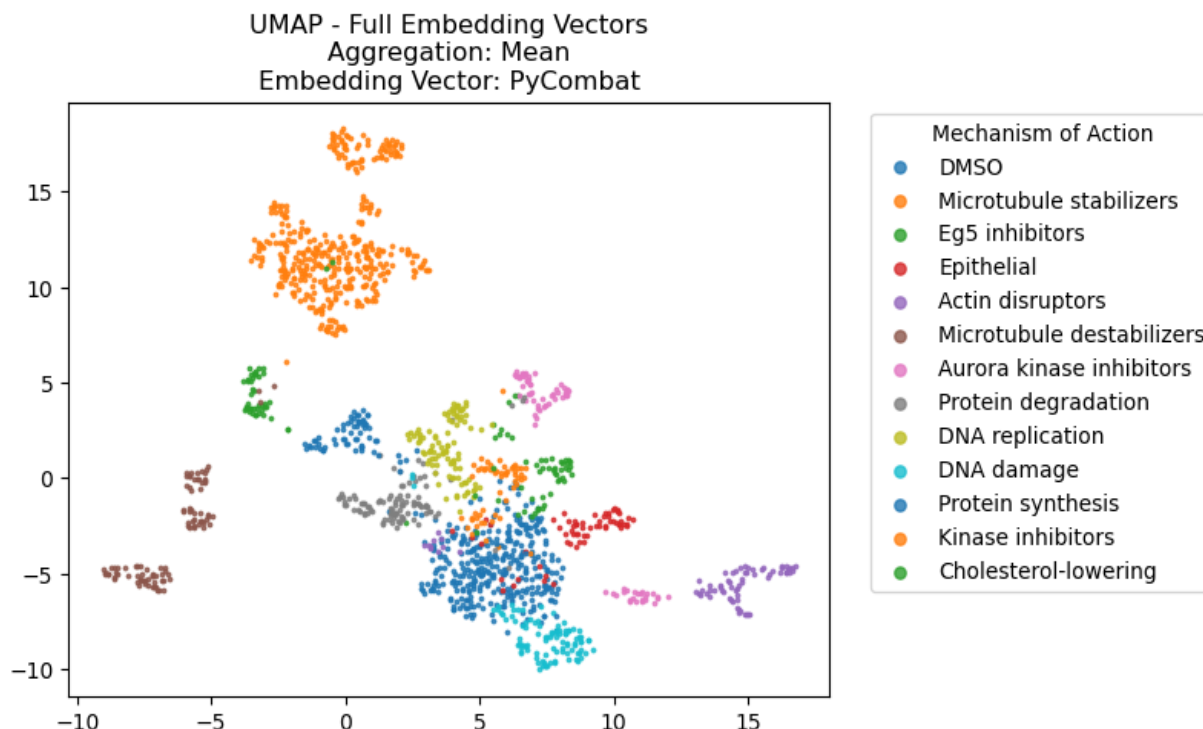


Figure 13: UMAP clustering of embedding vectors

## List Of Trained Models In The Project

Multiple models were trained and finetuned in this project. These models were used to generate image embedding vectors.

- Inception V3 base model without finetuning of convolutional layers
- Inception V3 base model with re-training the last 2 convolutional mixture layers
- Inception V3 re-training the last 2 convolutional mixture layers and adding data augmentation to the training images (random horizontal and vertical flip)

As an experiment we also used DinoV2 vision transformer model from Meta to generate image embedding vectors (DinoV2 ViT-S/14). The results are not included in the project report but are considered as potential improvement in follow-up work.