Cell2Structure: Identification of chemical compound properties in CNN-generated embeddings of cell microscopy images

Matthias Hübscher, Florian Gillet

Abstract

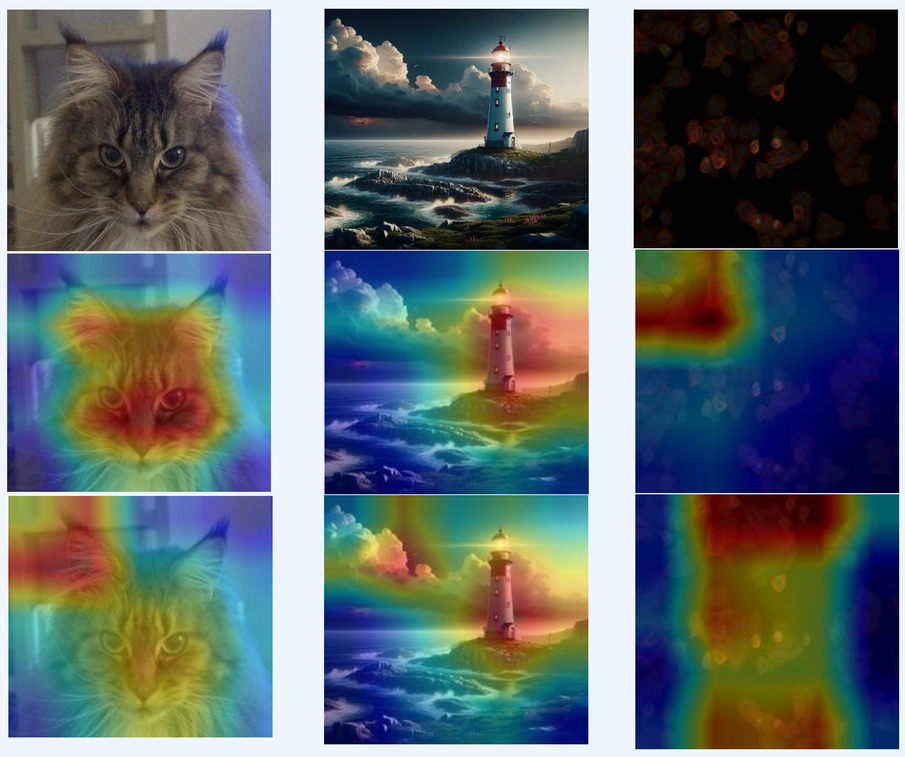
High content screening (HCS) is a well established cell-based screening technique in drug discovery. In this type of screen cultured cells are treated with a large set of chemical compounds. After the treatment the cells are imaged using high-throughput microscopy. The morphological changes of the cells induced by the chemical treatment are a signature of the mechanism of action (MoA) of the chemical compounds.

The Cell2Structure project explores the potential of using convolutional neural networks (CNN) to generate embeddings of the cell images to test our hypothesis that these embeddings contain features of the of the chemical structure of the compound used to treat the cells. This would open the way to generative models, taking cell images as input and outputting chemical structures that would induce the changes in the morphology of cells observed in the images.

We trained a CNN on the BBBC021 cell image dataset (Broad Bioimage Benchmark Collection. BBBC021 dataset: https://bbbc.broadinstitute.org/BBBC021), and explored the embeddings. We were able to demonstrate that the embeddings resulting from compounds sharing an MoA are similar, suggesting that they contain featrues of the compound structure.

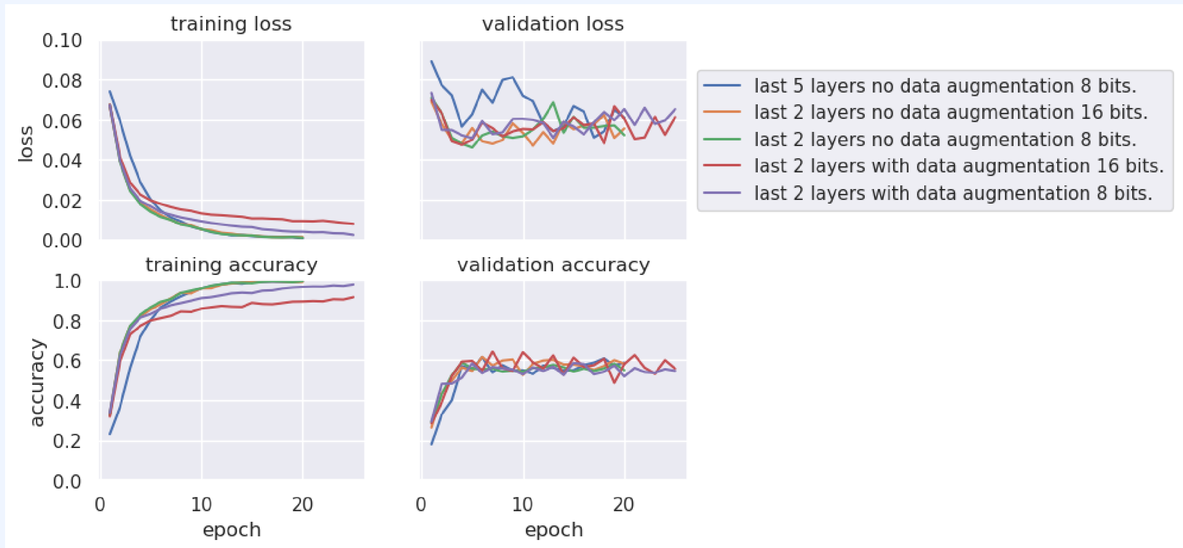
CNN Finetuning

Convolutional layers detecting low-level features were identified using activation maps overlaid with images of a cat, a lighthouse, and cell images from the dataset.



The figure above shows the average activation maps of layers Mixed_7b (second row) and Mixed_7c (third row). These layers detected low-level features in the cat and lighthouse images but didn't seem to detect anything relevant in cell images. Those two layers were unfrozen and opened for further training.

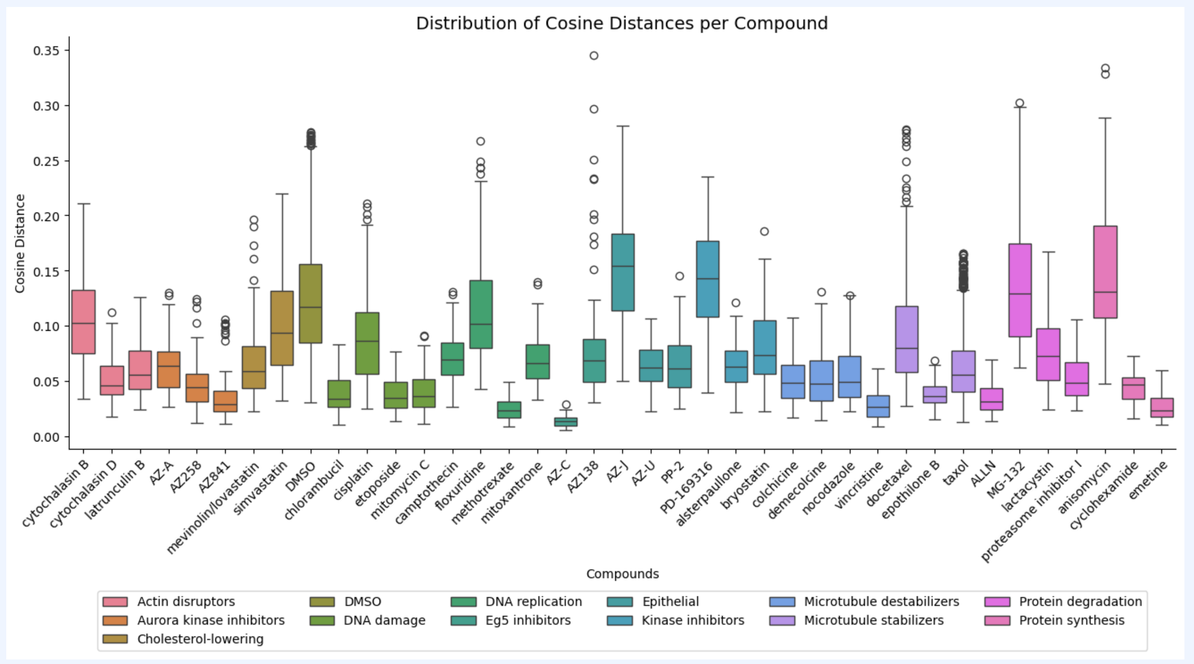
Despite a training accuracy beyond 95%, the maximum validation accuracy obtained was ~60%, indicating heavy overfitting. This was true regardless of the techniques we applied to improve the validation accuracy (data augmentation, L2 regularization, early stopping, training on 16 bit images instead of 8 bit ones). Even retraining the last five convolutional layers did not improve performance.



Because our main objective was not to produce a performant classifier, but to get a model able to generate embedding vectors that are representative of MoAs, we decided to not push the model optimization further. The training set accuracy was beyond 95%, which indicated that generated embeddings would be representative of MoAs, despite not being generalizable to other datasets. We therefore proceeded with our overfitting model.

Image embedding vector analysis

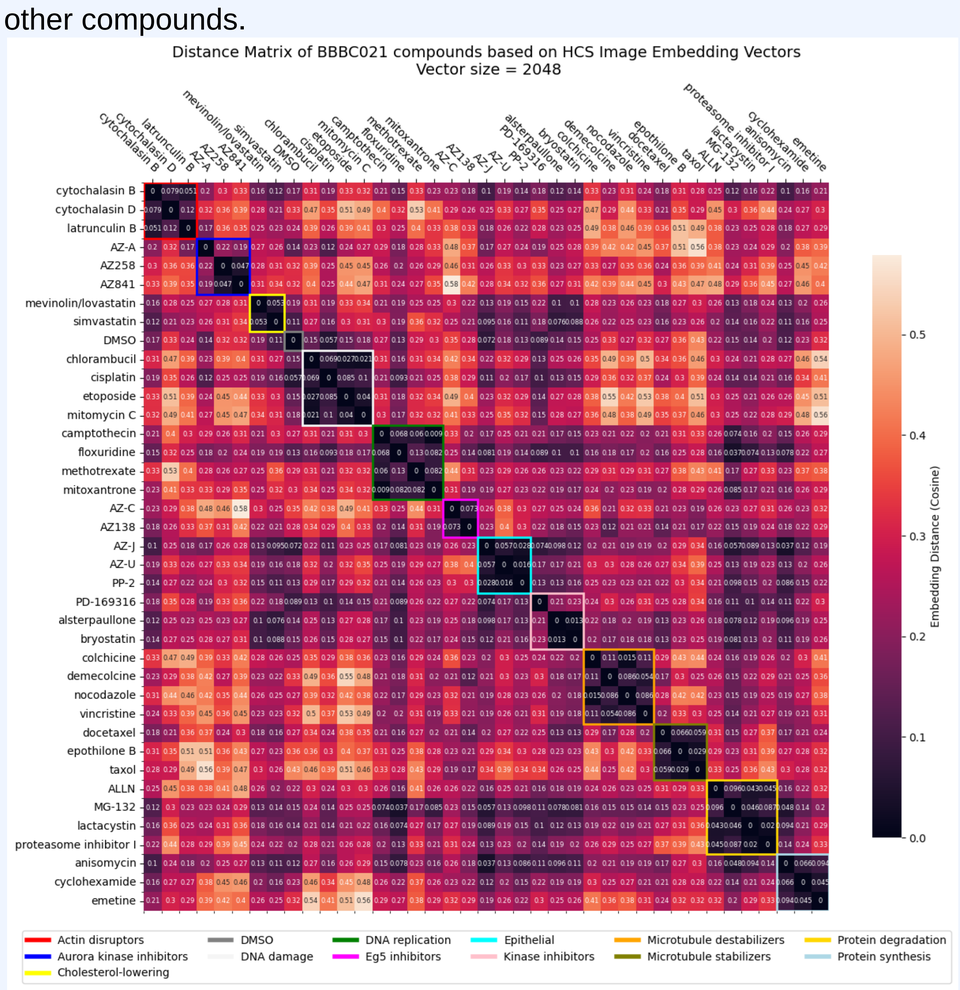
Embedding vectors were grouped by compound. For each group, a mean vector was computed. Then, within each group, the cosine distance between individual vectors and the respective mean vector was calculated. Distribution of cosine distances for embeddings grouped by compound is shown below.



Embedding vectors from compounds having a similar MoA showed similarity to each other. This is an indication that the embedding vectors we generated with a classifier trained only to recognize MoA, also contained features of the compounds themselves.

Embedding distance matrix

Distance matrix based on cosine similarity between embedding vectors was computed and visualized as a heatmap. The heatmap showed dark squares along the diagonal, showing that compounds sharing a similar MoA have more similar embeddings than with other compounds.



Conclusion

The CNN we trained achieved good accuracy on the training set, but did not generalize well to our validation set. As the classification task was not our objective, we proceeded with this overfitting model, but the results showed are only specific to this dataset. However, we obtained evidence that the embeddings generated by a model trained to recognize MoA carry additional information on the chemical compounds, that allow compound sharing a same MoA to cluster together. These observations answer our question: embedding vectors of cell images seem to contain information on the compound that was used to produce the cell images, opening the way to research on generative models.