Bridging the Phenotype-Proteome Gap: A Multi-Modal AI Framework for analysis of Cell Painting images

Maurycy Chronowski[3], Aleksandra Nowak[3], Stephan Eckert[1], Yun-Chien Chang[1], Nicola Berner[1], Amirhos-sein Sakhteman[1], Manuel Lau-benheimer[2], Georg Tascher[2] , Jan Carsten Pieck[2], Bernhard Kuster[1], Magdalena Otrocka[3], Michał Warchoł[3], Adriana Borowa[3], Sakshi Garg[2], Bolek Zapiec[2]

1. School of Life Sciences, Technical University of Munich, 85354 Freising, Germany
2. Discovery Pharmacology, Discovery & Development Technologies, Merck Healthcare KGaA, 64293 Darmstadt, Germany
3. Ardigen, Leona Henryka Sternbacha 1, 30-394 Kraków, Poland

ABSTRACT

Cell Painting assay captures a vast range of morphological information. However, translating these visual phenotypes into biological insight remains a challenge. This study investigates the capacity of AI architectures to reconstruct proteomic profiles directly from morphological features.

We developed a **multi-modal AI framework for proteomic profile prediction** by integrating Cell Painting images with corresponding mass spectrometry data from cells treated with ~2000 reference compounds. We compared CellProfiler features against various Deep Learning embeddings: Masked Autoencoder (MAE), self-distillation with no labels (DINO), and CLOOME. We focused on MAE with a ViT-B/8/224 backbone and optimized the model for microscopic images through high masking ratios and Fourier domain reconstruction loss. Using a Multilayer Perceptron (MLP) and nested cross-validation, we evaluated the models on two primary tasks: the classification of protein up/down regulation and the regression of normalized protein abundance.

Our findings demonstrate that classifying protein expression regulation is more robust than direct abundance regression. A substantial fraction of investigated proteins was predicted with high accuracy, with performance scaling in response to compound-induced perturbations. In a focused analysis of chemical treatments, the **system successfully identified a large proportion of regulated proteins, showing strong dose-dependency for top-performing markers**.

Results suggest a latent but measurable correspondence between cellular morphology and proteomic states. While challenges remain in achieving high-resolution reconstruction for all protein classes, we show that phenotypic profiling can serve as a proxy for capturing broader biological shifts, offering a potential bridge between morphological changes and proteomics cell state to support drug discovery processes.

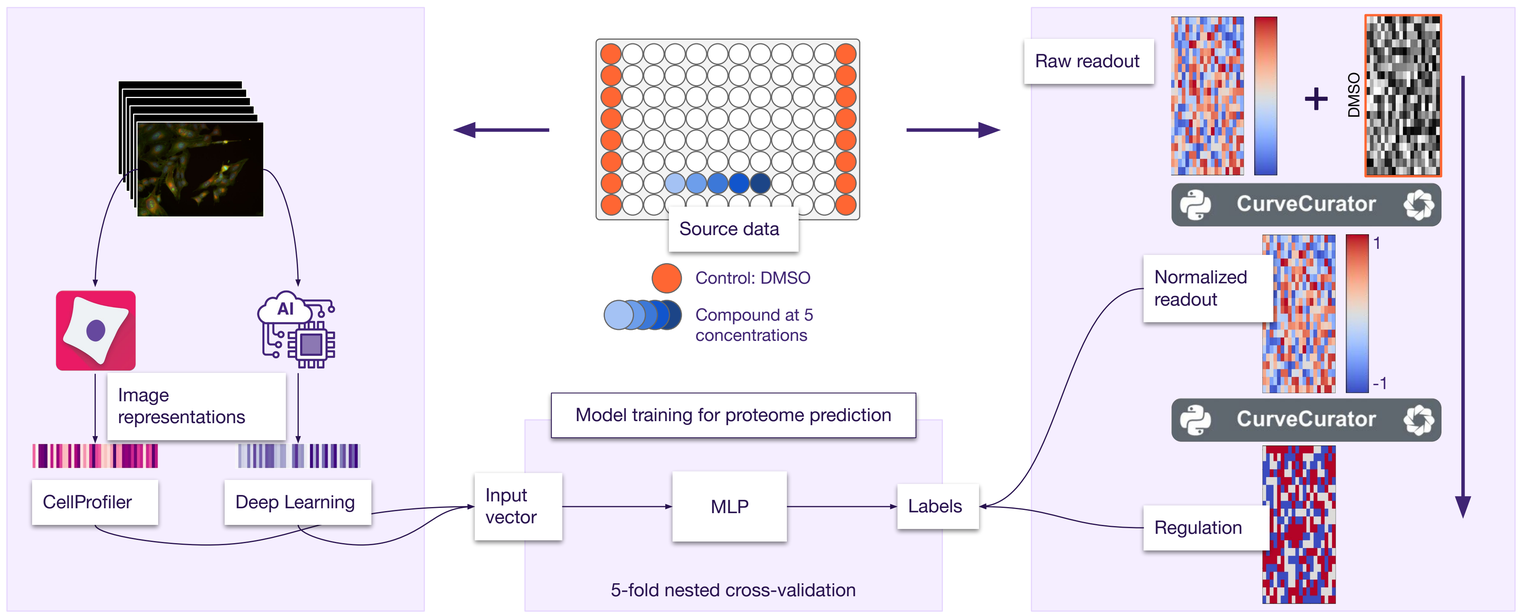
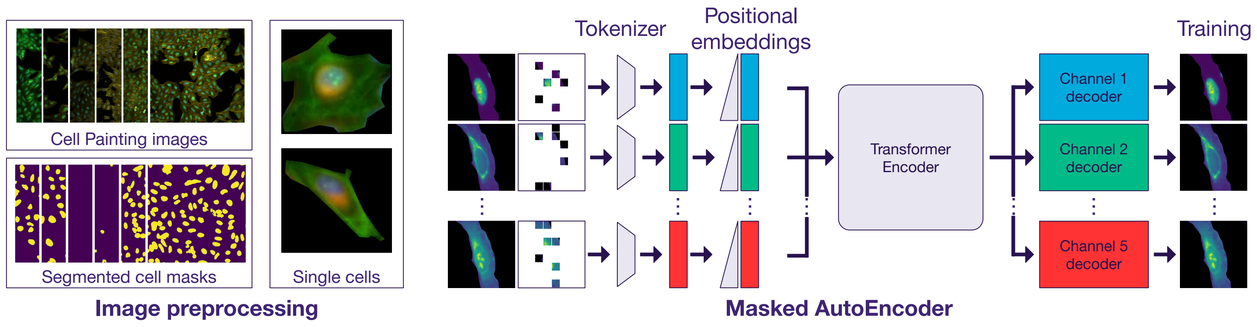


Figure 1. Overview of the experimental and computational pipeline. Compounds (5 concentrations) and DMSO controls were screened concurrently in Cell Painting and mass spectrometry-based proteomics. For the image modality, both CellProfiler and Deep Learning features are extracted. In parallel, proteomics data is processed using CurveCurator to generate normalized readouts and binary regulation statuses. Finally, these image features serve as inputs to train a Multi-Layer Perceptron (MLP) within a 5-fold nested cross-validation framework to predict the corresponding proteomic readouts.



Loss functions

ViT-MAE

$$\mathcal{L}_{MAE} = \frac{1}{P}\sum_{p=1}^{P} L_2(y_p, y'_p)$$

Fourier reconstruction

$$\mathcal{L}_{FT} = \frac{1}{P}\sum_{p=1}^{P} L_1(|\mathcal{F}(y_p)|, |\mathcal{F}(y'_p)|)$$

SC-MAE

$$\mathcal{L}_{MAE+} = (1-\alpha)\mathcal{L}_{MAE} + \alpha\mathcal{L}_{FT}$$

Figure 2. Single-cell Masked Autoencoder architecture. Cell Painting images are segmented to isolate individual cell and remove background. Then, each channel is masked and tokenized and passed through Transformer Encoder which aims to reconstruct a cell image. Three loss functions are used jointly: Vit-MAE, Fourier reconstruction, and SC-MAE loss.

DATA & PREPROCESSING

- **Image dataset:** Merck's proprietary dataset of ~2.8 M 5-channel Cell Painting images of U2OS cells, treated with various chemical perturbations across 5 concentrations. The dataset covers a wide range of modes of action (MoA) to capture diverse phenotypic changes. This entire set was used to train the MAE model (see: Methods). Compounds with matching proteomic readouts were used for prediction.
- **Proteomics dataset:** consisting of mass spectrometry readouts generated for ~2000 perturbations included in the image dataset. Data was processed with CurveCurator [1] to obtain normalized readouts with respect to negative control samples.

Two sets of prediction tasks were derived from the proteomics data:

- **Classification task:** predicting up/down regulation of each gene under a given chemical perturbation as a whole (curve fitted across the concentrations), as generated by the CurveCurator pipeline. This task was split into two subtasks: UP vs. NOT-regulated & DOWN vs. NOT-regulated, as it should not be assumed that these two types of regulation are biologically opposite. Number of genes selected: 6,035 (UP-NOT) and 6,782 (DOWN-NOT).
- **Regression task:** predicting the normalised fold change for each gene under a given chemical perturbation, at each concentration separately. Ratios were normalised to mean=0, std=1. Number of genes selected: 2,944 (based on the best-performing ones in the classification task).

METHODS

The entire processing pipeline of our method is presented in Figure 1.

Image feature extraction - 4 morphological profiling methods were tested:

- **CellProfiler (CP):** 4,459 human-engineered features,
- **Deep Learning Embeddings:**
  - Contrastive Learning and leave-One-Out-boost for Molecule Encoders (CLOOME) [2] (512 features),
  - Self-distillation with no labels (DINO) [3, 4] (768),
  - Masked Autoencoder (MAE) [5] (768).

Normalization:

- CellProfiler features were standardized per-plate to respective negative control wells.
- For Deep Learning models, image pixel intensities were standardized to mean and standard deviation calculated across all the negative control wells.

Aggregation: Image features were extracted at the Field of View (FOV) level and averaged to well-level representations before feeding classifiers.

Our **Masked Autoencoder (MAE) model** was trained following the methodology described by [5]. We utilized a ViT-Base/8 architecture with a 224px image size, training the model from a randomly initialized state using a 5-channel input. To optimize performance, we incorporated a high masking ratio (0.9) alongside a single-layer decoder, as recommended in [6]. Additionally, a Fourier domain reconstruction loss was introduced to enhance performance on microscopic data, as demonstrated by [5]. **The training dataset consisted of single-cell crops with masked backgrounds;** these segmentations were initially derived from CellProfiler outlines and later substituted with Cellpose-SAM [7] masks, which yielded no noticeable difference in quality or feature replicability. Finally, the model was trained for 100k steps using the Lion optimizer [8] with a cosine-annealed learning rate over the first 10% of training. This was executed with a batch size configuration of 32 FOVs × 8 cells per FOV across 4 GPUs with a 2-step gradient aggregation, resulting in an effective batch size of 2048.

Features from CLOOME and DINO were generated using the reproducibility instructions described by the respective authors.

For **proteomic prediction tasks**, we trained a MLP architecture with four hidden layers using Adam optimizer (learning rate = 1e-4), with dropout (0.2).

- **Classification task** was trained with Binary Cross Entropy loss per target with masking of missing annotations.
- **Regression task** optimised the Mean Squared Error (MSE). As an alternative cost function, SMAPE was tested, but failed to converge towards good performance. Tanh nonlinearity was substituted on MLP layers upon grid-search.

Additionally, we evaluated a **compound target prediction task** using DrugBank and Therapeutic Target Database (TTD) annotations. MLP classifiers were trained to **predict binary bioactivity on a panel of 252 compound targets**. The task served as an additional proxy to assess whether proteomic information can support the prediction of compound activity compared to the image-derived features alone. In the multi-modal approach (image+proteome), image features were concatenated with the normalized gene readouts.

For **model validation**, we employed a 5-fold cross-validation framework, incorporating nested sub-sampling within the validation folds during the hyperparameter search phase. The performance metrics reported herein represent the average scores computed across the 5 independent test folds.

RESULTS AND DISCUSSION

Table 1 shows the ROC AUC scores and the number of tasks above respective thresholds.

- CellProfiler is clearly the best performing across all the compared image feature extraction methods.
- Custom-trained **MAE model comes second best, outperforming public pretrained vision models**. It shows, however, a significant performance gap from the CellProfiler.
- In the most significant predictability bracket (AUC ≥ 0.8), a much higher percentage of up-regulated genes is correctly modelled.

	CellProfiler		MAE		CLOOME		DINO	
ROC AUC	UP-NOT	DOWN-NOT	UP-NOT	DOWN-NOT	UP-NOT	DOWN-NOT	UP-NOT	DOWN-NOT
Average	**0.64**	**0.66**	0.62	0.63	0.58	0.60	0.58	0.59
#tasks ≥ 0.6	**4303 (71%)**	**5730 (84%)**	3802 (63%)	4892 (72%)	2555 (42%)	3955 (58%)	2597 (43%)	3144 (46%)
#tasks ≥ 0.7	**1428 (24%)**	**1881 (28%)**	869 (14%)	679 (10%)	196 (3%)	203 (3%)	198 (3%)	115 (2%)
#tasks ≥ 0.8	**96 (1.6%)**	**37 (0.6%)**	48 (0.8%)	14 (0.2%)	4 (0.07%)	5 (0.07%)	6 (0.10%)	5 (0.07%)

Table 1. ROC AUC scores on the classification task: average and number of tasks achieving ROC AUC above threshold (0.6, 0.7, 0.8). Comparison between image features. CellProfiler clearly outperforms other methods, while our custom MAE outperforms public pretrained models.

For the regression task, a panel of 2944 genes was selected, based on the results from the classification task. The goal was to show if a direct perturbation response (measured as the normalized protein abundance fold change) at each dose can be robustly predicted for these genes. Our findings, see Table 2, show that:

- Despite iterative refinement (MLP parameter search, standardizing protein ratios, tanh non-linearities on layers) our **best-performing model achieved a maximum R^2 of only ~0.45** (for the gene MGP). Moreover, we found the SMAPE (Symmetric Mean Absolute Percentage Error) metric to be entirely uncorrelated with R^2 and particularly high in value, further supporting the conclusion that the **regression models fail to predict the normalized gene readouts robustly**.
- Our scores are substantially lower than the benchmarks reported by [9], where top-performing proteins reached R2 values as high as 0.73 (e.g. CXCL11).
- **Regression performance exhibits a steep dose-dependency**, with predictability dropping near zero at lower compound concentrations.

Given these limitations and the significantly higher reliability of our binary classification models, **it was decided to pivot the framework entirely toward predicting discrete up/down regulation**.

Gene	R^2 ↑	SMAPE ↓
MGP	0.4482	48.8926%
LRRN2	0.4282	61.6972%
CD44	0.3703	57.3468%
FBXL14	0.3685	57.6521%
SMTN	0.3628	54.2414%
P4HA2	0.3509	62.6770%
CEP131	0.3428	66.5592%
PODXL	0.3385	61.3162%
DLGAP5	0.3242	64.4138%
CD63	0.3238	55.8664%

Table 2. Results of the regression task. Top 10 best-performing tasks, sorted by R^2. Metrics were average across all concentrations.

ROC AUC	CellProfiler	MAE	Proteome	Proteome + CellProfiler
Average	0.63	0.60	0.49	**0.64**
#tasks ≥ 0.6	**171 (68%)**	133 (53%)	67 (27%)	154 (61%)
#tasks ≥ 0.7	49 (19%)	19 (8%)	10 (4%)	**78 (31%)**
#tasks ≥ 0.8	7 (3%)	2 (1%)	0	**27 (11%)**
#tasks ≥ 0.9	0	0	0	**8 (3%)**

Table 3. ROC AUC scores on the compound target prediction task. Combining image features and normalized proteomic readouts significantly boosts the classification performance.

In the **compound target prediction task**, presented in Table 3, our findings show that CellProfiler again performs better than MAE, which is why it was selected for further evaluation in the multli-modal training setup. Interestingly, it can be observed that combining the image and proteomic data results in a profound performance boost. The **multi-modal approach scores much higher in the upper ROC AUC brackets than either of its components alone.** This points to a potential link between the proteome and the compounds' bioactivity, that can be unlocked using the information contained in Cell Painting images.

CONCLUSIONS

- We demonstrated that high-content Cell Painting images contain information sufficient to predict the gene regulation induced by the chemical perturbations for a significant part of the proteome.
- Predicting discrete protein up/down regulation is significantly more robust and reliable than direct abundance regression. Our models struggled to obtain satisfactory performance on the regression task.
- Further experiments that would include a separate activity detection in Cell Painting images could improve the predictive performance. We showcased a drastic dose-dependency of the models.
- CellProfiler features remain the most robust overall baseline. Custom-trained MAE model shows superior performance over public embeddings like CLOOME and DINO, however, there's still a significant performance gap to be matched.
- High scores reported on the compound target prediction task using a combination of image and proteomic data highlight a link between the two modalities, which could be further explored in future studies.
- Our multi-modal framework offers a baseline for a scalable proxy for capturing proteomic shifts from inexpensive imaging assays, potentially reducing the need for late-stage mass spectrometry in early phenotypic screening pipelines in Drug Discovery projects.

REFERENCES

1. Bayer, Florian P., et al. "CurveCurator: a recalibrated F-statistic to assess, classify, and explore significance of dose–response curves." Nature Communications 14.1 (2023): 7902.
2. Sanchez-Fernandez, Ana, et al. "CLOOME: contrastive learning unlocks bioimaging databases for queries with chemical structures." Nature Communications 14.1 (2023): 7339.
3. Caron, Mathilde, et al. "Emerging properties in self-supervised vision transformers." Proceedings of the IEEE/CVF international conference on computer vision. 2021.
4. Doron, Michael, et al. "Unbiased single-cell morphology with self-supervised vision transformers." bioRxiv (2023).
5. Kraus, Oren, et al. "Masked autoencoders for microscopy are scalable learners of cellular biology." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.
6. Wei, Zihao, et al. "Masked autoencoders are secretly efficient learners." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.
7. Pachitariu, Marius, Michael Rariden, and Carsen Stringer. "Cellpose-SAM: superhuman generalization for cellular segmentation." BioRxiv (2025): 2025-04.
8. Chen, Xiangning, et al. "Symbolic discovery of optimization algorithms." Advances in neural information processing systems 36 (2023): 49205-49233.
9. Mehrizi, Rahil, et al. "Multi-omics prediction from high-content cellular imaging with deep learning." arXiv preprint arXiv:2306.09391 (2023).

Download the poster here: