



CHALMERS
UNIVERSITY OF TECHNOLOGY



UNIVERSITY OF GOTHENBURG

Mining Typical Phenotypes from Cell Images Using Stable Diffusion

Master's Thesis

Master's thesis in Computer science and engineering

Noah Márquez Vara
Siyu Hu

MASTER'S THESIS 2026

Mining Typical Phenotypes from Cell Images Using Stable Diffusion

Master's Thesis

Noah Márquez Vara
Siyu Hu



UNIVERSITY OF
GOTHENBURG



CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2026

Mining Typical Phenotypes from Cell Images Using Stable Diffusion
Master's Thesis
Noah Márquez Vara
Siyu Hu

© Noah Márquez Vara, 2026.
© Siyu Hu, 2026.

Supervisor at Chalmers: Télió Cropsal, Department of Computer Science and Engineering
Examiner at Chalmers: Rocío Mercado Oropeza, Department of Computer Science and Engineering
Examiner at GU: Giovanni Volpe, Department of Physics

Master's Thesis 2026
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Typeset in L^AT_EX
Gothenburg, Sweden 2026

Mining Typical Phenotypes from Cell Images Using Stable Diffusion

Master’s Thesis

Noah Márquez Vara

Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg

Siyu Hu

Department of Physics
University of Gothenburg

Abstract

This thesis studies whether diffusion-model typicality can improve cell selection for image-based morphological profiling. The project focuses on Broad Bioimage Benchmark Collection 021 (BBBC021), a Cell Painting-style benchmark with mechanism-of-action (MoA) labels and strong batch effects. The evaluation uses a controlled selector benchmark where every method starts from the same border-filtered Cellpose-SAM proposals, and only the rule for retaining proposals changes. The baseline keeps all valid proposals, random selection subsamples them, foreground fraction ranks crops by target-cell pixel coverage, and MorphoDiff-Typicality ranks the same proposals by a population-null MorphoDiff score computed over the target-cell mask. A fixed matched subset is used so that all headline comparisons operate on the same perturbations, wells, and images.

The final selector benchmark uses isolated target-cell crops, removes border-touching proposals, embeds crops with InceptionV3, normalizes the resulting cell features with cell-level typical variation normalization (TVN), and evaluates well-level profiles with copairs. The baseline retains 389,448 crops and reaches perturbation mean average precision (mAP) 0.587, perturbation fraction retrieved (FR) 57.65%, MoA mAP 0.587, and MoA FR 91.67%. Foreground-fraction ranking gives the strongest MoA mAP result, peaking at 0.635 with 20% retained crops. MorphoDiff-Typicality gives the strongest perturbation FR result, peaking at 62.35% with 45% retained crops, and reaches the MoA FR baseline at 35%. These results show that simple crop quality is a strong signal in BBBC021, while diffusion-based typicality preserves a different perturbation-aware signal that is most visible in perturbation-level detectability.

Keywords: Cell Painting, morphological profiling, diffusion models, MorphoDiff, typicality, BBBC021, copairs, Cellpose-SAM.

Acknowledgements

We would like to thank our supervisor, T lio Cropsal, for his guidance, support, and thoughtful feedback throughout this thesis. His advice helped us stay focused, improve the work, and see the project from a clearer perspective.

We are also grateful to our examiners, Roc o Mercado Oropeza and Giovanni Volpe, for their time and valuable comments on this work.

Finally, we would like to thank our friends and families for their patience, encouragement, and support throughout this period. It has meant a lot to us.

Noah M rquez Vara and Siyu Hu
G teborg, June 2026

Contents

List of Figures	xi
List of Tables	xv
Acronyms	xvii
1 Introduction	1
2 Background and Theoretical Framework	5
2.1 Phenotypic screening and Cell Painting	5
2.2 The BBBC021 benchmark	6
2.2.1 Mechanism-of-action annotation and benchmark design	6
2.3 Cell instance segmentation	7
2.3.1 Cellpose	8
2.3.2 Cellpose-SAM	8
2.3.3 Cell-level selection and crop-quality filtering	8
2.4 Deep feature extraction for morphological profiling	9
2.4.1 Transfer learning from natural images	9
2.4.2 InceptionV3 as feature extractor	9
2.5 Typical variation normalization and nuisance variation	10
2.5.1 TVN procedure	10
2.5.2 Normalization in BBBC021	10
2.6 Information retrieval evaluation with copairs	11
2.6.1 Mean average precision and its interpretation	11
2.6.2 Perturbation-level phenotypic activity	11
2.6.3 Mechanism-of-action retrieval	12
2.6.4 Using mAP and FR together	12
2.6.5 FID and KID as generative checks	13
2.7 Diffusion models, MorphoDiff, and Diff-Mining	13
2.7.1 Denoising and latent diffusion	13
2.7.2 Classifier-free guidance	13
2.7.3 MorphoDiff as a perturbation-conditioned generative model . .	14
2.7.4 Diff-Mining and spatial typicality	15
2.7.5 Null conditioning in biological typicality	16
3 Methods	19

3.1	Study design and frozen matched subset	19
3.2	Shared pipeline architecture	19
3.3	Segmentation, crop construction, and shared caches	20
3.4	Feature extraction	21
3.5	Selector definitions	22
3.5.1	Baseline	23
3.5.2	Random selection	23
3.5.3	Foreground-fraction ranking	23
3.5.4	MorphoDiff-Typicality ranking	23
3.6	Percentage-based retention sweep and seeds	24
3.7	Per-pixel typicality construction	25
3.8	Cell-level TVN	26
3.9	Evaluation protocol	27
3.9.1	copairs metrics and parameterization	27
3.9.2	Selector agreement and crop diagnostics	28
3.10	Compute environment	29
4	Results	31
4.1	Baseline	31
4.2	Selector comparison across retention budgets	31
4.3	Retention budget and crop-count distribution	33
4.4	Foreground fraction as a crop-quality signal	34
4.5	Selector agreement	35
4.6	MoA-level gains and failure modes	38
4.7	Summary	41
5	Discussion	43
5.1	Interpretation of the foreground-fraction result	43
5.2	Interpretation of MorphoDiff-Typicality results	44
5.3	Relation to previous work	45
5.4	Limitations	46
5.5	Future work	47
6	Conclusion	49
	Bibliography	51
A	Supplementary results and archived experiments	I
A.1	MorphoDiff deployment and image-quality reproduction	I
A.2	Supplementary MoA-level selector diagnostics	I
A.3	Archived exploratory experiment families	III
A.3.1	Segmentation alternatives	III
A.3.2	Score-map-guided crop alternatives	III
A.3.3	Null-condition comparison	IV

List of Figures

2.1	BBBC021 image example for each of the 12 active MoA classes used in the frozen subset.	7
2.2	Representative real and MorphoDiff-generated BBBC021 images for six compounds spanning six annotated MoA classes. Each column shows one perturbation, with a real image on the top row and a generated image on the bottom row.	15
3.1	BBBC021 image preprocessing on one representative field. The three raw assay channels are normalized separately and then merged into one RGB image. In the RGB image, R = actin, G = tubulin, and B = DAPI.	21
3.2	Cellpose-SAM proposal construction on one representative BBBC021 image. The left panel shows the processed assay image, the middle panel shows the instance masks returned by Cellpose-SAM, and the right panel shows the proposal bounding boxes used to extract candidate cell crops.	22
3.3	How one Cellpose-SAM proposal becomes one isolated single-cell crop. From left to right, the panels show the source image with the selected proposal, the raw proposal box, the padded bounding-box crop used by the pipeline, and the final crop after non-target pixels inside that box are set to zero.	22
3.4	Simple comparison of crop selection on one representative BBBC021 image. The first panel shows the source image. The remaining panels show the top five retained proposals for random selection, foreground-fraction ranking, and cell-mask typicality ranking on the same field of view.	24
3.5	MorphoDiff-Typicality scoring on one BBBC021 image. The panels show the processed image, the raw signed population-null typicality map, and the top-ranked Cellpose-SAM proposals under cell-mask scoring. The colorbar reports raw loss-difference values.	26
3.6	Normalization rationale on the frozen matched subset. The left panel compares the control counts available for fitting TVN, and the right panel shows the MoA trade-off for no normalization, week-level TVN, and cell-level TVN on the same subset before setting the final selector benchmark.	27

- 4.1 Fixed-geometry selector comparison after border filtering. The dashed line marks the baseline and open circles mark the first retained percentage that meets or exceeds that baseline. For the baseline and foreground-fraction selector, variation comes from copairs permutation seeds because the selected cell subset is fixed. For random selection, faint replicate lines and shaded regions summarize three random selector seeds. For MorphoDiff-Typicality, faint replicate lines and shaded regions summarize three independent population-null Monte Carlo draws with the same $K_{MC} = 16$ sampling rule. 32
- 4.2 Effective retained-cell budget after border filtering. (a) Box plots show the per-image retained-cell distribution at representative percentage budgets. (b) The histogram shows the full-retention detected-cell distribution across images, with reference lines marking representative percentage budgets and the baseline median. 34
- 4.3 Selected-cell property shifts relative to the full-retention baseline across retention budgets. Foreground fraction and zero fraction after masking are shown as percentage-point differences from the full-retention baseline means, while cell area and neighbor count are shown as ratios to the corresponding full-retention baseline means. Metric definitions are given in Section 3.9.2. The dashed horizontal line marks no change relative to full retention. 35
- 4.4 Pairwise selector agreement across images. The top row shows Spearman rank correlations between proposal rankings for foreground fraction versus MorphoDiff-Typicality, random selection versus foreground fraction, and random selection versus MorphoDiff-Typicality. The bottom row shows the corresponding Jaccard overlap between top- $K\%$ selected proposal sets at fixed retention budgets. Spearman correlation measures monotonic rank agreement, while Jaccard overlap measures direct selected-cell overlap. Variation reflects the distribution across images in the border-filtered proposal set. 36
- 4.5 Qualitative agreement and disagreement between foreground fraction and MorphoDiff-Typicality. Green boxes mark the top foreground-fraction proposals and plum boxes mark the top MorphoDiff-Typicality proposals on the same Cellpose-SAM proposal set. Each row uses a matched zoomed field of view across the image and heatmap panels, rather than the full microscopy image. The heatmap panel shows the raw signed typicality map with display scaling only; proposal ranking uses the raw cell-mask mean scores. 37
- 4.6 MoA-level retrieval (mAP) versus retained-cell budget K for foreground fraction and MorphoDiff-Typicality. Each curve is one MoA. Red crosses mark (selector, K) points where the class is not retrieved under the permutation-null threshold. 38

4.7	Top MorphoDiff-Typicality crops for the six visual-audit MoA classes. Crops are selected from the full border-filtered proposal pool, corresponding to the baseline selector at 100% retention, and ranked within each MoA by raw cell-mask MorphoDiff-Typicality score. Columns show ranks 1–5 after a diversity pass that prevents repeated image stems within each MoA row. The first tile of each row reports how many scored compounds were available for that MoA in the frozen subset. Repeated compounds or doses still occur because many displayed MoA rows have only one available scored compound in that pool.	40
4.8	Two distinct failure modes of the MorphoDiff-Typicality cell selector. Each row shows, from left to right, the processed RGB image, its individual fluorescence channels (R: actin, G: tubulin, B: DNA/DAPI), and the raw MorphoDiff-Typicality map. Yellow outlines denote Cellpose-SAM cell masks, and the cyan circle marks the strongest typicality response that falls outside the valid proposal set. (a) Proposal-stage failure: the hotspot lands on a genuine nucleated cell with a clear DAPI nucleus and actin/tubulin structure that Cellpose-SAM failed to segment, so the selector cannot retain it despite a high typicality score. (b) Typicality-score failure: the hotspot is an ~ 8 -pixel, single-channel (actin-only) bright speck with no nucleus and no tubulin to which MorphoDiff-Typicality assigns its strongest local response (a false positive); here Cellpose-SAM correctly rejected the region, and the failure lies in the typicality score itself.	41
A.1	Random-selection reference for the MoA-level selector audit.	II
A.2	Archived segmentation alternatives on BBBC021, including the full-image baseline, Otsu foreground mask, watershed instance mask, and random crops.	IV
A.3	Archived score-map-guided crop alternatives on BBBC021, comparing Cellpose-SAM rankings with fixed-size patch, heatmap-peak, and thresholded heatmap variants.	V
A.4	Archived null-condition comparison on BBBC021, showing Stable Diffusion, synthetic all-ones, and population-null heatmaps with the corresponding top-ranked Cellpose-SAM proposals.	V

List of Tables

3.1	Metadata fields used by copairs in the present study. The table includes both the matching role and the value structure on the frozen BBBC021 subset.	28
4.1	Selector comparison against the all-cell baseline in Figure 4.1. For each selector, the table reports the first retained percentage K at which the mean curve meets or exceeds the baseline for perturbation FR, perturbation mAP, and MoA FR, together with the best perturbation FR and best MoA mAP values observed anywhere on the curve. Lower first- K values indicate that a selector reaches the baseline with fewer retained cells, while higher values are better for the two best-value columns. The retained percentage at which each best value occurs is shown in parentheses.	33
A.1	BBBC021 image-quality comparison between the original MorphoDiff paper and the deployment check used in this thesis. Lower is better for both metrics.	I
A.2	Per-MoA support summary for the selector audit.	III

Acronyms

AP	average precision
BBBC021	Broad Bioimage Benchmark Collection 021
CORAL	correlation alignment
CPU	central processing unit
DAPI	4',6-diamidino-2-phenylindole
DINO	self-distillation with no labels
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid
FID	Fréchet Inception Distance
FR	fraction retrieved
GPU	graphics processing unit
KID	Kernel Inception Distance
mAP	mean average precision
MoA	mechanism of action
NAISS	National Academic Infrastructure for Supercomputing in Sweden
NSC	Not-Same-Compound
NSCB	Not-Same-Compound-or-Batch
PCA	principal component analysis
RGB	red–green–blue
RNA	ribonucleic acid
SAM	Segment Anything Model
Slurm	Slurm Workload Manager
SMILES	Simplified Molecular Input Line Entry System
TVN	typical variation normalization

VAE variational autoencoder

1

Introduction

Cell Painting is a high-content microscopy assay designed to capture how chemical and genetic perturbations reshape cell morphology at scale [1]. In a typical screening workflow, each field of view contains many candidate cells and each treatment is represented by several images, wells and acquisition batches. This makes the assay useful for phenotypic profiling, but it also raises a practical question for downstream analysis: which cells should contribute to the final representation of a perturbation? Here, a perturbation refers to a compound–concentration pair. Thus, the same compound applied at different doses is treated as different perturbations. We use treatment more broadly to refer to the applied experimental condition, and dosage to refer to the concentration at which a compound is applied.

Not every segmented cell in a field of view is equally informative. Some cells show a treatment phenotype clearly, while others are weakly affected, partially segmented, close to image borders, mixed with neighboring cells, or visually dominated by background. Many profiling pipelines embed all detected cells and then aggregate them into image-level or well-level profiles. The final profile may therefore combine phenotype-bearing cells with cells that are less informative for the treatment being studied. When Cell Painting images are used for retrieval, clustering, or mechanism-of-action (MoA) analysis, the choice of which cells are retained becomes part of the scientific pipeline [2], [3].

This thesis studies whether a perturbation-conditioned diffusion model can be repurposed from image generation into cell-level phenotype mining. The project started from the combination of two methods that were originally developed for different purposes. MorphoDiff is a perturbation-conditioned latent diffusion model trained on Cell Painting images, where the conditioning signal represents a chemical or genetic perturbation [4]. Diff-Mining shows that diffusion losses can be reused as localization signals, so that image regions can be scored by how specific they are to a given condition [5]. Together, these methods suggest a possible way to rank cells by how typical they are for the compound applied to their well, but they do not by themselves provide a complete biological profiling pipeline or an evaluation protocol for cell-level selection.

The central question of the thesis is therefore whether MorphoDiff and Diff-Mining can be combined to identify perturbation-specific cell-level phenotypes in Broad Bioimage Benchmark Collection 021 (BBBC021), and how such a signal can be evaluated once it appears visually plausible. The first part concerns feasibility, namely

how to compute a cell-level MorphoDiff-Typicality score from the difference between two MorphoDiff denoising losses, one conditioned on the target perturbation and one conditioned on an explicitly defined reference. The second part concerns evaluation, namely how to move from qualitative heatmaps to a controlled benchmark that tests whether typicality-ranked cells preserve perturbation-level profiling signal and broader mechanism of action (MoA) retrieval signal when fewer cells are retained.

The thesis is guided by two research questions:

RQ1 How can Diff-Mining typicality be adapted to compute cell-level MorphoDiff-Typicality scores for BBBC021 cell proposals?

RQ2 To what extent does cell-level MorphoDiff-Typicality preserve perturbation-level and MoA-level retrieval signal at reduced cell retention, and where does it provide signal beyond simpler selection baselines?

RQ1 and RQ2 address two complementary parts of the thesis. RQ1 focuses on how the typicality signal is constructed, including how MorphoDiff can compare the appearance of a cell under the applied perturbation with a suitable reference, and how this comparison can be converted into a score for each cell proposal. RQ2 then asks whether this score is useful for biological profiling once only a subset of cells is retained. Rather than relying only on visual examples, the thesis evaluates whether typicality-ranked cells preserve perturbation-level and MoA-level retrieval signal, and compares this behaviour with simpler selection strategies. In this way, the project connects a methodological question, how to define cell-level MorphoDiff-Typicality, with an evaluation question, whether this signal captures biologically meaningful information.

The project is evaluated on BBBC021, a benchmark dataset with annotated chemical perturbations and mechanism-of-action labels. The final experimental design uses a fixed-geometry comparison in which all selectors operate on the same Cellpose-SAM proposal set. The resulting crops are embedded with InceptionV3, normalized with cell-level typical variation normalization (TVN), and evaluated with copairs retrieval metrics. This design separates the ranking problem from the segmentation problem, so differences between selectors can be interpreted as differences in which proposals are retained rather than differences in the underlying crop geometry.

The baseline keeps all valid Cellpose-SAM proposals after border filtering and three selection strategies are compared against it. Random selection measures how much profiling signal remains under simple subsampling. Foreground-fraction ranking selects cells with a larger fraction of non-background pixels and is used as a perturbation-agnostic crop-quality baseline. MorphoDiff-Typicality ranks the same Cellpose-SAM proposals using a raw population-null typicality map. The null reference is estimated from a balanced set of compound conditionings that exclude both the target compound and DMSO, and each proposal is scored by the mean raw population-null typicality over the Cellpose-SAM target-cell mask. The evaluation uses a variable-DMSO control policy, so the retained cell budget is varied for both treated and control populations.

The results show that MorphoDiff-Typicality can preserve part of the biological

retrieval signal while retaining fewer cells, especially for early perturbation-level retrieval and MoA detectability. They also show that foreground-fraction ranking is a strong perturbation-agnostic baseline, particularly for MoA-level ranking quality. The thesis is therefore not framed as a search for a single universally best selector. Instead, it studies what kind of signal each selector retains, where diffusion-based typicality provides information beyond generic crop quality, and where simpler selection rules remain competitive.

2

Background and Theoretical Framework

This chapter introduces the biological and computational ideas needed to understand the thesis. It first describes phenotypic screening, Cell Painting, and the BBBC021 benchmark. It then explains how cells are segmented, embedded, normalized, and evaluated as morphological profiles. The final sections introduce MorphoDiff and Diff-Mining, which provide the basis for scoring how typical a cell is for a given perturbation.

2.1 Phenotypic screening and Cell Painting

Phenotypic drug discovery aims to identify useful perturbations by measuring how cells respond, rather than by testing compounds only against a single predefined molecular target [2], [3]. It is useful when the relevant biology spans several pathways, when the mechanism is not fully understood, or when the desired effect appears at the level of cell state rather than single-protein activity. High-content imaging is well suited to this setting because it captures several cellular structures in the same assay.

Cell Painting protocols combine fluorescent stains to mark several cellular compartments, but the exact number of dyes and imaging channels depends on the dataset and implementation. The canonical protocol described by Bray et al. uses six fluorescent dyes imaged over five channels to stain structures such as nuclei, F-actin, microtubules, endoplasmic reticulum, Golgi, plasma membrane, mitochondria, and ribonucleic acid (RNA)-rich regions [1]. In practice, the protocol is often adapted to a given dataset or screening setup. That detail matters here because BBBC021 is not a full five-channel Cell Painting dataset, but a related three-channel assay that still supports morphological profiling.

In a typical profiling pipeline, images are first segmented into cells or candidate regions, then mapped to feature vectors through handcrafted measurements or deep representations. These features are aggregated into image-level, well-level, or perturbation-level profiles and compared with a similarity measure. The aim is not only to detect visual differences, but to obtain a profile space that captures biologically meaningful relationships [2], [3].

2.2 The BBBC021 benchmark

BBBC021 is a widely used benchmark for image-based profiling because it combines biologically diverse perturbations with substantial nuisance variation across acquisition batches [6]. It contains fluorescence microscopy images of MCF-7 wild-type p53 breast cancer cells. Unlike the full Cell Painting protocol, BBBC021 uses three single-channel images corresponding to deoxyribonucleic acid (DNA), F-actin, and β -tubulin. The dataset spans 55 plates imaged over 10 weeks, so technical drift must be treated as part of the benchmark [7].

Compounds were originally tested at eight doses, with some compound-concentration pairs omitted because the dose was inactive, too toxic, or failed quality control.¹ In the inactive case, activity was defined by comparing each dose profile to the dimethyl sulfoxide (DMSO) control profiles using Mahalanobis distance. The features used for this activity call were measurements extracted by a proprietary software used at AstraZeneca. Mahalanobis distance provides one way to define whether a phenotype differs from the negative controls, but it is not the only possible definition of activity. A different set of features, control model, or activity threshold could change which compound-dose pairs enter the benchmark. BBBC021 is therefore not a random collection of treatments, since it was already filtered toward perturbations that were considered phenotypically active or biologically relevant.

We define a treatment as the experimental condition applied to a well. For active BBBC021 conditions, this treatment is a chemical compound applied at a specific concentration. We therefore use perturbation to mean a compound-concentration pair, rather than the compound name alone. The dose is the concentration of that compound. A MoA is a higher-level biological label that groups different perturbations expected to affect cells through a similar biological mechanism.

2.2.1 Mechanism-of-action annotation and benchmark design

The evaluation subset commonly used in the literature was defined by Ljosa et al. [8] and it contains 103 active compound-concentration pairs from 38 compounds grouped into 12 mechanism-of-action (MoA) classes. According to the benchmark description, only six of these classes are easily separable by visual inspection, while the other six require literature-based annotation. The benchmark therefore includes both clear and subtle phenotypes.

The MoA labels in this subset are aurora kinase inhibitors, cholesterol-lowering compounds, DNA damage agents, DNA replication inhibitors, Eg5 inhibitors, epithelial modulators, kinase inhibitors, microtubule destabilizers, microtubule stabilizers, protein degradation perturbations, protein synthesis inhibitors, and actin disruptors [8]. DMSO appears throughout the assay as a negative control, but it is not treated as an active mechanism class. Instead, it is used as a reference population for normalization and activity assessment.

¹<https://bbbc.broadinstitute.org/BBBC021>

Figure 2.1 shows one BBBC021 image example for each active MoA in the frozen subset.

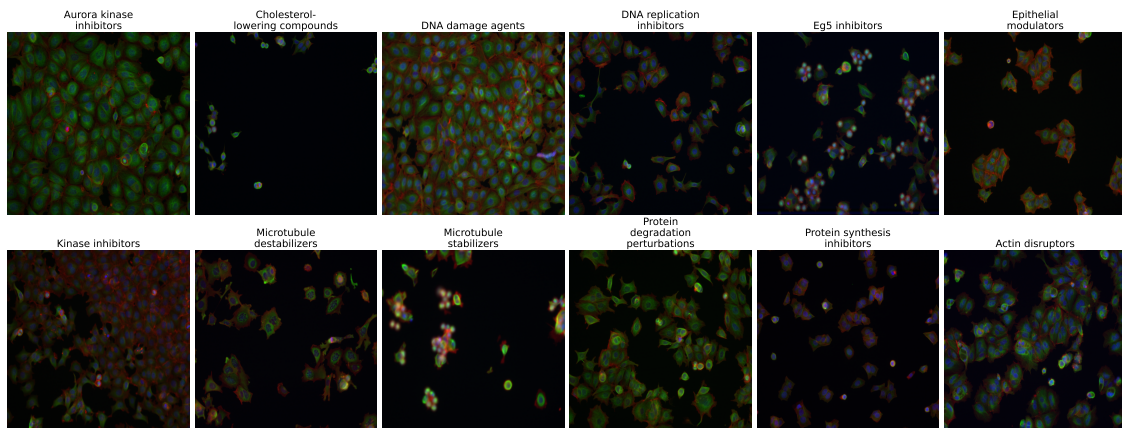


Figure 2.1: BBBC021 image example for each of the 12 active MoA classes used in the frozen subset.

BBBC021 has been often evaluated with 1-nearest-neighbor classification under restrictions such as Not-Same-Compound (NSC) and Not-Same-Compound-or-Batch (NSCB) [7]. These restrictions were designed to reduce excessive performance caused by compound identity or batch leakage. The same logic motivates the retrieval-based evaluation used here, where pair construction must separate biological similarity from replicate and batch effects.

BBBC021 is useful for this work because perturbations are usually measured in multiple replicate wells, and annotated perturbations are linked to known MoA labels, sometimes more than one. The dataset also contains structured technical variation across plates and acquisition weeks, allowing profiling methods to be tested on whether they can recover perturbation-level phenotypes despite replicate and batch effects. Because the benchmark is well established in the literature, new ideas can still be compared against strong profiling baselines [7], [8], [9].

A practical limitation is that MorphoDiff requires a perturbation representation derived from molecular identity. For compounds with public structures, BBBC021 provides Simplified Molecular Input Line Entry System (SMILES) strings, a concise text notation for molecular structure, and these are converted into numerical molecular encodings before they are used to construct the perturbation-conditioning input. Nineteen of the 103 annotated active perturbations belong to AstraZeneca compound series without public molecular identifiers or public SMILES. In the conditional MorphoDiff setting used, these perturbations cannot be scored because the required perturbation conditioning cannot be encoded.

2.3 Cell instance segmentation

Morphological profiling depends strongly on the regions selected for analysis. If a crop contains mostly background, merges multiple cells, or captures only part of a cell, the downstream embedding may not represent a clean cellular phenotype.

Segmentation is therefore more than a preprocessing step in this thesis. It defines the candidate objects that the selectors rank and that the profiling pipeline later compares and aggregates.

Traditional microscopy pipelines often relied on thresholding, watershed segmentation, and hand-tuned morphological operations [2]. These methods can work in simple settings, but they become less reliable when cells touch, overlap, vary in size, or change shape under perturbation. Those failure modes are especially relevant in phenotypic screening, where morphology is itself the signal of interest.

2.3.1 Cellpose

Cellpose addresses segmentation by predicting both a cell probability map and a two-dimensional flow field pointing each pixel toward the center of its cell [10]. Pixels that converge to the same sink are grouped into the same instance. This turns segmentation into a flow-to-center problem rather than a direct foreground-background mask prediction.

This formulation separates touching cells and handles variation in size and shape more robustly than many classical pipelines. Because Cellpose was trained on diverse microscopy images, it functions as a generalist segmentation model rather than a method tied to a single assay. For profiling, it provides a consistent way to generate instance masks and candidate regions across different imaging settings.

The original Cellpose framework can work with a cytoplasm channel, a nucleus channel, or both, and it also estimates an effective cell diameter to adapt its predictions [10]. This flexibility makes it practical for profiling pipelines in which staining strategies differ across datasets.

2.3.2 Cellpose-SAM

Cellpose-SAM extends this line of work by combining Cellpose-style mask generation with the Segment Anything Model (SAM) visual backbone [11], [12]. The output remains a set of cell proposals and masks. In this work, these proposals are used as a fixed candidate set for downstream selection experiments.

Microscopy images often contain weak, irregular, or overlapping cell boundaries, and proposal generation can affect all later profiling steps. By caching a shared set of Cellpose-SAM proposals, selector experiments can vary the ranking and filtering strategy while keeping the underlying candidate masks fixed.

2.3.3 Cell-level selection and crop-quality filtering

Cell-level profiling pipelines usually rely on quality-control decisions before aggregation, so segmentation masks, border contact, object size, foreground coverage, local crowding, and debris filtering can all affect which cells contribute to the final profile. These steps are often treated as preprocessing, but they influence the biological representation because low-quality crops can change the embedding distribution and the final well-level aggregate.

Our selector benchmark makes this usually implicit step explicit, because the baseline keeps all valid proposals after border filtering, foreground-fraction ranking acts as a simple perturbation-agnostic crop-quality rule, and MorphoDiff-Typicality tests whether a perturbation-conditioned diffusion model can provide a different ranking signal. This places the work between conventional crop-quality filtering and perturbation-aware cell selection. It does not replace segmentation, but evaluates how different rules for retaining already detected cells affect downstream profiling.

2.4 Deep feature extraction for morphological profiling

Once cells have been localized, their visual content must be converted into numerical features that preserve phenotype differences. Earlier pipelines relied on handcrafted descriptors such as area, eccentricity, intensity statistics, and texture measurements. These features remain interpretable and competitive. This thesis instead uses InceptionV3 as a practical transfer-learning baseline, where a pretrained neural network encodes cell images into feature vectors without requiring a feature extractor to be trained from scratch.

2.4.1 Transfer learning from natural images

Transfer learning from natural-image models has become common in morphological profiling because early and intermediate visual features are often reusable across domains [13]. Edges, textures, local symmetries, and multi-scale spatial structure are not unique to pictures, they also appear in microscopy. A pre-trained network therefore provides a strong feature basis without requiring a large labeled microscopy dataset for supervised training.

In profiling, the final classification layer is typically discarded and the penultimate representation is used as an embedding. This yields a fixed-dimensional feature space in which cells, images, and wells can be compared directly. It also separates representation extraction from the downstream biological task, allowing the same embeddings to be normalized, aggregated, and evaluated with retrieval metrics.

2.4.2 InceptionV3 as feature extractor

InceptionV3 is one of the most widely used transfer-learning backbones in high-content imaging [7], [14]. Its architecture combines filters with different receptive fields, which is useful in morphological profiling because relevant cues can appear at several scales, from local cytoskeletal texture to larger changes in cell shape or spreading.

The final pooled representation of InceptionV3 is 2,048-dimensional. In practice, the vector is used as a fixed summary of each isolated crop’s visual content. Recent work on Cell Painting reconstruction reports that, in its retrieval evaluations, InceptionV3 embeddings matched or exceeded a publicly available microscopy-specialized model,

supporting its use as a reasonable off-the-shelf feature space for this benchmark [15]. Other encoders, including self-supervised vision-transformer models such as self-distillation with no labels (DINO), could be evaluated in the same pipeline, but representation learning is not the experimental variable in this thesis.

Although it is less directly interpretable than handcrafted features, it often performs better for retrieval tasks. At the same time, such embeddings can capture technical variation as well as biological signal, which is why normalization remains essential.

2.5 Typical variation normalization and nuisance variation

Nuisance variation is a central issue in BBBC021 and Ando et al. show that plate, batch, and well-position effects can strongly bias profiling benchmarks, to the point that even DMSO controls can appear predictive if the evaluation does not control for these correlations [7]. A baseline is therefore only meaningful if the feature space is corrected for this technical structure.

TVN addresses this by using negative controls to define the geometry of ordinary technical variation [7]. The main assumption is that variation among control wells should mainly reflect nuisance effects rather than treatment biology. If that variation is characterized well, the embedding space can be transformed so that technical axes are reduced and biological comparisons become more reliable.

2.5.1 TVN procedure

TVN can be understood as a sequence of transformations fitted on control embeddings and then applied to all embeddings in the dataset.

First, the controls are standardized so that features with large raw variance do not dominate the space. Next, principal component analysis (PCA) with whitening rotates the control space into orthogonal axes and rescales them to unit variance. Finally, covariance alignment, often implemented with correlation alignment (CORAL), is used to match batch covariance structures to a common reference [16]. Together, these steps redefine the feature space using control variation rather than raw feature magnitude.

A key result from Ando et al. is that the principal directions with highest variance are not necessarily the most biologically useful [7]. In BBBC021, high-variance dimensions can reflect nuisance structure, while lower-variance dimensions may carry important phenotype information. TVN is therefore not just a denoising step, it also changes which dimensions matter for retrieval.

2.5.2 Normalization in BBBC021

BBBC021 makes normalization difficult because the dataset contains 55 plates, each with 6 DMSO negative-control wells and 6 Taxol positive-control wells, acquired over

10 weeks [7]. This gives enough control data to study nuisance variation, but not so much that covariance estimation becomes trivial. In addition, some MoA classes appear on relatively few batches, which increases the risk of confounding between batch structure and biology.

TVN is treated here as part of the reference profiling pipeline, not as an optional preprocessing step. Retrieval scores are only interpretable if the embedding space is not dominated by technical correlations. It is therefore part of what makes BBBC021 a meaningful profiling benchmark.

2.6 Information retrieval evaluation with copairs

Modern profiling datasets are often evaluated as retrieval problems rather than as standard classification tasks. A representation is useful if it places biologically related samples close together and unrelated samples far apart, under explicit rules for which comparisons count as valid positives and negatives. `copairs` formalizes these rules through metadata fields, making pair construction explicit rather than leaving it implicit in a class label [9].

In `copairs`, positive and negative pairs are defined through metadata rules rather than fixed labels alone. A `sameby` rule specifies which metadata fields must match, while a `diffby` rule specifies which must differ. This makes the framework flexible enough to express replicate retrieval, MoA retrieval, and other profiling tasks within a common evaluation scheme.

2.6.1 Mean average precision and its interpretation

The main continuous metric used in `copairs` is mean average precision (mAP). For each query profile, all candidates are ranked by similarity, and average precision (AP) is computed from the ranks at which true positives appear. If true positives are retrieved early, AP is high. If they are mixed with many negatives, AP is lower. The mean of AP over all queries gives mAP.

mAP is useful because it evaluates ranking quality rather than only final hits. In profiling, a good representation should place biologically relevant samples near the top of the neighbor list, not just somewhere among the nearest few. At the same time, mAP is always tied to the declared evaluation protocol, including the metadata rules, candidate pool, and similarity measure.

2.6.2 Perturbation-level phenotypic activity

At the perturbation level, the retrieval task asks whether replicate wells of the same perturbation are more similar to each other than to the rest of the dataset. In `copairs` terms, positives match on perturbation identity while differing on a replicate-level field, and negatives come from different perturbations. Therefore, high perturbation-level mAP indicates that the perturbation produces a reproducible phenotype.

copairs also provides a significance-oriented summary through a null distribution obtained by permutation. From this, one can compute the fraction of perturbations whose AP exceeds a chosen null threshold, commonly reported as fraction retrieved (FR) [9]. FR answers a different question from mAP, not how sharp the rankings are, but how many perturbations are detectably active at all.

2.6.3 Mechanism-of-action retrieval

MoA retrieval is more demanding than perturbation-level retrieval. The goal is no longer to retrieve replicate wells of the same treatment, but to retrieve distinct treatments that share a mechanism of action. Positive pairs should match on MoA while differing on treatment identity. Otherwise, performance may be inflated by compound-specific similarity rather than shared mechanism.

MoA retrieval matters because it tests whether a selector preserves biologically meaningful structure beyond replicate consistency. A selector that keeps only obvious or idiosyncratic cells may still perform reasonably at the perturbation level while failing to preserve shared morphology across compounds with the same mechanism.

2.6.4 Using mAP and FR together

Both mAP and FR are reported because they answer different questions about the profile space. mAP measures ranking quality and it is high when positive matches appear consistently near the top of the ranked neighbor list. FR measures detectability, it counts how many perturbations or MoA groups exceed a null-derived significance threshold.

This distinction matters for interpreting selector results because for example, a selector can obtain high FR but lower mAP if it preserves enough signal for many perturbations or MoA groups to be detected, while still producing rankings where the positive matches are not consistently concentrated at the very top. Conversely, a selector can obtain reasonable mAP for a smaller subset of strong phenotypes while failing to make many perturbations detectable. The two metrics therefore describe complementary properties rather than interchangeable notions of performance.

In this thesis, FR is used mainly as a retention-budget detectability metric. It asks whether a selector can keep fewer cells while still preserving a statistically detectable perturbation-level or MoA-level signal. mAP is used as a ranking-quality metric and it asks whether the retained cells produce compact and well ordered neighborhoods in the retrieval space. This distinction is important because a perturbation-aware selector may preserve detectable biological signal without producing the strongest MoA ranking, while a perturbation-agnostic crop-quality selector may improve the smoothness of MoA neighborhoods without explicitly using perturbation information.

Retrieval metrics are therefore preferable to qualitative inspection alone, but they do not remove the need for interpretation. A selector that improves perturbation-level retrieval may not improve MoA-level retrieval in the same way, because replicate

consistency and shared mechanism structure are different biological tasks.

2.6.5 FID and KID as generative checks

Fréchet Inception Distance (FID) and Kernel Inception Distance (KID) are also used in the appendix to verify MorphoDiff deployment before interpreting typicality results; the BBBC021 comparison used here is summarized in Appendix Section A.1 and Table A.1. FID compares the Gaussian statistics of embeddings from real and generated images [17], while KID measures a kernel maximum mean discrepancy in the same feature space and is often more reliable in finite samples [18].

2.7 Diffusion models, MorphoDiff, and Diff-Mining

The second part of this work uses a generative model not only to synthesize images, but also to identify which image regions are most characteristic of a perturbation. This section briefly reviews latent diffusion, then explains how MorphoDiff adapts it to Cell Painting images and how Diff-Mining turns diffusion losses into spatial scores.

2.7.1 Denoising and latent diffusion

Denoising diffusion probabilistic models learn to reverse a gradual noising process [19]. Starting from a clean sample x_0 , Gaussian noise is added over a sequence of timesteps until the sample approaches pure noise. A neural network is then trained to predict the added noise at arbitrary timesteps. During generation, the learned reverse process removes noise step by step and reconstructs a sample from an initial random input.

Latent diffusion makes this practical for larger images by performing the denoising process in the latent space of a variational autoencoder (VAE) rather than directly in pixel space [20]. The image is encoded into a compressed spatial latent, the U-Net operates in that space, and the decoded image is produced at the end. MorphoDiff follows this general design, but changes the conditioning signal from text to biological perturbations.

2.7.2 Classifier-free guidance

Classifier-free guidance is useful here because it makes explicit how conditional diffusion models can be compared against a reference prediction. In classifier guidance, the reverse diffusion process is steered by a separate classifier. At a noisy latent z_t , Bayes' rule gives

$$\nabla_{z_t} \log p_t(z_t | c) = \nabla_{z_t} \log p_t(z_t) + \nabla_{z_t} \log p_t(c | z_t), \quad (2.1)$$

since $p(c)$ does not depend on z_t . The second term is a classifier gradient that pushes the sample toward regions that are more likely under condition c .

Classifier-free guidance removes the need for this separate classifier by training the diffusion model itself with and without the condition [21]. During training, the condition is randomly dropped for part of the data, so the same network learns both a conditional prediction and a missing-condition prediction. In score notation, let

$$s_{\theta}(z_t, t, c) \approx \nabla_{z_t} \log p_t(z_t | c), \quad s_{\theta}(z_t, t, \emptyset) \approx \nabla_{z_t} \log p_t(z_t), \quad (2.2)$$

where $\emptyset$ denotes the learned missing-condition input. The difference between these two predictions approximates the conditional direction,

$$s_{\theta}(z_t, t, c) - s_{\theta}(z_t, t, \emptyset) \approx \nabla_{z_t} \log p_t(c | z_t). \quad (2.3)$$

Guidance then strengthens this direction by combining the unconditional and conditional scores,

$$\tilde{s}_{\theta}(z_t, t, c) = s_{\theta}(z_t, t, \emptyset) + \gamma (s_{\theta}(z_t, t, c) - s_{\theta}(z_t, t, \emptyset)), \quad (2.4)$$

where γ controls the guidance strength. When $\gamma = 1$, the guided score reduces to the ordinary conditional score. When $\gamma > 1$, the conditional direction is amplified.

For an ϵ -prediction diffusion model, the score and the predicted noise are proportional up to a timestep-dependent factor. The same guidance idea can therefore be written as a combination of noise predictions,

$$\tilde{\epsilon}_{\theta}(z_t, t, c) = \epsilon_{\theta}(z_t, t, \emptyset) + \gamma (\epsilon_{\theta}(z_t, t, c) - \epsilon_{\theta}(z_t, t, \emptyset)). \quad (2.5)$$

This expression matters for the present thesis because the relevant terms are both the conditional prediction and its difference from a reference prediction produced by the same denoising model. In standard classifier-free guidance, this reference is a learned missing-condition prediction. In the MorphoDiff setup used here, such a branch is not available, so the reference used for typicality scoring must be defined explicitly.

2.7.3 MorphoDiff as a perturbation-conditioned generative model

MorphoDiff adapts latent diffusion to Cell Painting images [4]. It keeps the main components of Stable Diffusion, namely a VAE, a conditional U-Net, and a diffusion scheduler, but replaces the text-conditioning module with a perturbation encoder that maps a chemical or genetic intervention to a dense representation. For chemical perturbations, this starts from a SMILES string and uses RDKit to convert the molecular structure into a numerical embedding before the model projects it into the conditioning space [4], [22]. This representation is then injected through the same cross-attention interface normally used for text conditioning.

MorphoDiff therefore learns perturbation-specific microscopy morphology rather than generic microscopy appearance. That makes the model useful not only for image generation, but also as a perturbation-aware prior. If the model has learned the morphology associated with perturbation c , then image regions that express that

morphology should be easier to explain under condition c than under an unrelated or null condition.

Figure 2.2 illustrates this idea on BBBC021. Each column shows a real image and a MorphoDiff-generated image for the same perturbation. The figure is used as a qualitative check that the deployed model captures broad perturbation-dependent morphology before its denoising losses are used for typicality scoring.

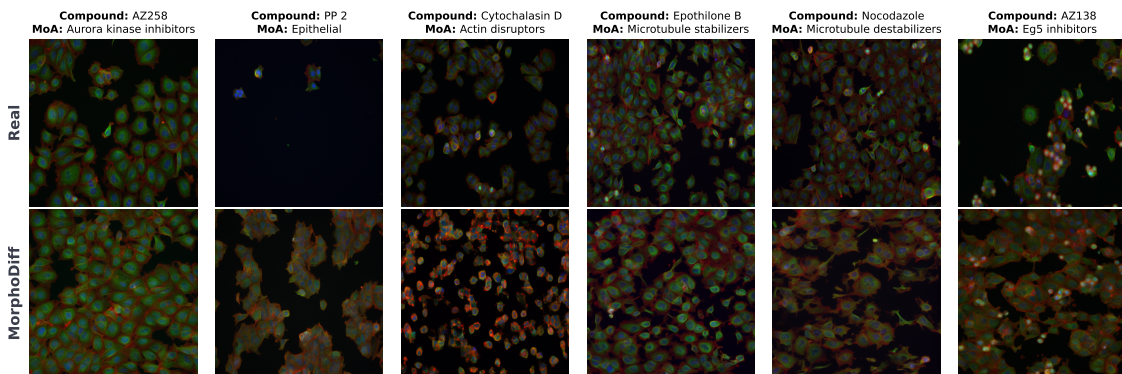


Figure 2.2: Representative real and MorphoDiff-generated BBBC021 images for six compounds spanning six annotated MoA classes. Each column shows one perturbation, with a real image on the top row and a generated image on the bottom row.

For BBBC021, the conditioning signal is derived from molecular information rather than natural language. That is a strength of the approach, but it also creates the practical limitation discussed earlier where only perturbations with available molecular embeddings can be handled through this route.

This distinction matters for the typicality score used later in the thesis. The derivation in Section 2.7.2 assumes that the model has learned both a conditional prediction and a missing-condition prediction. MorphoDiff is conditional, but the deployed model does not provide a calibrated branch trained by dropping the perturbation condition. As a result, the reference used in the loss contrast is not a technical detail, it is part of the scientific definition of what the typicality map measures.

2.7.4 Diff-Mining and spatial typicality

Diff-Mining provides the next conceptual step [5]. Its central idea is that diffusion models contain useful information not only in the final generated image, but also in their denoising losses. If a region is easier for the model to explain under one condition than under another, that difference can be interpreted as evidence that the region is specific to the first condition.

Let z_0 denote the latent representation of an observed image and let z_t be the noisy latent obtained at diffusion timestep t ,

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, \quad (2.6)$$

where ϵ is Gaussian noise and $\bar{\alpha}_t$ is determined by the diffusion noise schedule. A conditional denoising model predicts the added noise from z_t , t , and a condition c .

For a spatial position u , the local denoising loss can be written as

$$\ell_\theta(u \mid z_t, t, c) = \frac{1}{C} \sum_{k=1}^C (\epsilon_k(u) - \epsilon_{\theta,k}(z_t, t, c)(u))^2, \quad (2.7)$$

where C is the number of latent channels. Lower loss means that the model predicts the added noise more accurately under that condition.

The typicality map used in this thesis is based on a loss contrast between a target perturbation c and a reference condition c_{null} . For each spatial position u , the score is

$$T(u \mid c) = \ell_\theta(u \mid z_t, t, c_{\text{null}}) - \ell_\theta(u \mid z_t, t, c). \quad (2.8)$$

A positive value means that the target condition explains that position better than the null reference, since the target-conditioned denoising loss is lower. A value close to zero means that the target and the reference explain the region similarly. A negative value means that the reference condition explains the region better than the target condition.

This definition makes typicality a relative quantity rather than an absolute property of a cell. The same image region may receive a different score if the null reference changes. In this thesis, the pixel-level map $T(u \mid c)$ is converted into a cell-level score by averaging over the Cellpose-SAM mask of a candidate cell. For proposal j , with mask M_j , the cell score is

$$s_j(c) = \frac{1}{|M_j|} \sum_{u \in M_j} T(u \mid c). \quad (2.9)$$

The selector then ranks candidate cells by $s_j(c)$. The exact implementation of this score, including the population-null construction and the retained-cell sweep, is described in Chapter 3.

2.7.5 Null conditioning in biological typicality

The null reference is one of the main design choices in biological typicality scoring. Equation 2.8 follows the same contrastive logic as classifier-free guidance, but uses the contrast for scoring observed regions rather than for guiding generation. It does not ask whether a region is simply likely under the target perturbation. It asks whether the region is better explained by the target perturbation than by a chosen reference. The null condition therefore determines the meaning of the score.

A natural reference in diffusion models is an unconditional prediction. In classifier-free guidance, this prediction is obtained by training the same denoising model both with and without conditioning [21]. If such a missing-condition prediction were available, a target-specific score could compare the target loss against a learned unconditional loss,

$$T_\emptyset(u \mid c) = \ell_\theta(u \mid z_t, t, c_\emptyset) - \ell_\theta(u \mid z_t, t, c), \quad (2.10)$$

where $c_\emptyset$ denotes a learned missing-condition input. This would measure how much more predictable a region becomes when the target perturbation is supplied. The

deployed MorphoDiff model does not provide this missing-condition mode, so this interpretation cannot be assumed.

One possible workaround is to use a separate unconditional Stable Diffusion model trained on the same image domain. This gives a conceptually simple reference, but it also compares losses from two different models. Even if both models operate on similar images, their denoising errors are not guaranteed to be calibrated on the same scale. In the early experiments of this thesis, this option was visually and numerically weaker, so it is treated as an archived null-conditioning baseline rather than the final reference.

A second workaround is to use a synthetic vector inside the conditional MorphoDiff model, such as the all-ones vector used in the initial implementation. This keeps the target and null losses within the same model, but the vector is not a learned unconditional condition and is not a verified DMSO representation. It has the correct tensor shape, but its biological meaning is limited. It is therefore useful as an engineering baseline, not as a clean definition of untreated or generic morphology.

A biological control such as DMSO would be an attractive null because it would give the score a direct treated-versus-control interpretation,

$$T_{\text{DMSO}}(u \mid c) = \ell_{\theta}(u \mid z_t, t, c_{\text{DMSO}}) - \ell_{\theta}(u \mid z_t, t, c). \quad (2.11)$$

A positive value would then indicate that the region is better explained by the target perturbation than by the untreated control state. This requires the model to have a meaningful learned DMSO condition. In the MorphoDiff setup used here, such a calibrated DMSO condition is not available, so using DMSO as if it were a learned null would overstate what the model provides.

The final experiments therefore use a population-null reference. For a target compound c , the null pool contains non-target and non-DMSO compounds from the matched BBBC021 subset. These compounds are grouped by mechanism of action. If $\mathcal{M}_c$ is the set of eligible MoA groups for target c , and $\mathcal{C}_m$ is the set of eligible compounds in MoA group m , the full MoA-balanced weight for compound $q \in \mathcal{C}_m$ is

$$w_q = \frac{1}{|\mathcal{M}_c|} \frac{1}{|\mathcal{C}_m|}. \quad (2.12)$$

This gives each eligible MoA the same total weight, and then distributes that weight uniformly across compounds within the MoA.

In the Monte Carlo version used in the final experiments, K null compounds are sampled for each target. Each draw first samples an MoA uniformly from $\mathcal{M}_c$, and then samples a compound uniformly from the selected MoA. If compound q is sampled n_q times, its empirical null weight is

$$\hat{w}_q = \frac{n_q}{K}, \quad \sum_q \hat{w}_q = 1. \quad (2.13)$$

The final experiments use $K = 16$. The sampled null set is fixed per target compound so that all images of the same target are scored against the same population reference.

This can be viewed as replacing the unavailable missing-condition prediction in Equation 2.5 with a balanced prediction over other perturbation conditions. The analogy is limited because the population-null is not learned through condition dropout, but it preserves the important property that target and reference predictions are produced by the same MorphoDiff denoising network.

The population-null prediction is formed by averaging denoising predictions before the loss is computed,

$$\bar{\epsilon}_{\theta}^{\text{null}}(z_t, t, c) = \sum_q \hat{w}_q \epsilon_{\theta}(z_t, t, q). \quad (2.14)$$

The corresponding null loss is

$$\ell_{\theta}^{\text{pop}}(u \mid z_t, t, c) = \frac{1}{C} \sum_{k=1}^C \left(\epsilon_k(u) - \bar{\epsilon}_{\theta, k}^{\text{null}}(z_t, t, c)(u) \right)^2. \quad (2.15)$$

The final population-null typicality score is then

$$T_{\text{pop}}(u \mid c) = \ell_{\theta}^{\text{pop}}(u \mid z_t, t, c) - \ell_{\theta}(u \mid z_t, t, c). \quad (2.16)$$

This score should be interpreted as target specificity relative to a balanced population of other treatments, not as treated-versus-untreated DMSO specificity. It avoids a synthetic conditioning vector and keeps the target and null predictions within the same MorphoDiff model, but it also changes the biological question being asked. Different null choices therefore define different forms of typicality. An unconditional null asks what is specific relative to generic image content, a DMSO null asks what differs from untreated cells, and the population-null used here asks what is specific to the target perturbation relative to other perturbations in the benchmark.

3

Methods

This chapter describes the BBBC021 pipeline used to evaluate cell-level MorphoDiff-Typicality beyond visual inspection. The guiding principle is a controlled comparison where segmentation, crop construction, feature extraction, normalization, and evaluation are kept fixed as far as possible, while cell selection is treated as the experimental variable. The focus is the frozen matched subset and the fixed-geometry selector study, where all methods rank the same Cellpose-SAM proposal set.

3.1 Study design and frozen matched subset

All comparisons use a fixed, precomputed matched subset of BBBC021 so that the baseline and the MorphoDiff-guided selectors are evaluated on exactly the same perturbations, wells, and image sites. Here, *frozen* means that the subset inclusion is decided once before selector evaluation and is then kept unchanged across all selectors, retention percentages, random seeds, and metrics. This is necessary because the typicality pipeline requires a perturbation embedding for every treated condition it scores, whereas the published BBBC021 evaluation set is larger than the subset for which such embeddings are available.

The starting point is the BBBC021 evaluation set used in the project, which contains 103 active perturbation keys plus DMSO, giving 104 keys including the control. Intersecting that set with the perturbations present in the molecular embedding table removes 19 perturbations that cannot be scored by MorphoDiff. The matched subset therefore contains 84 active perturbations plus DMSO, giving 85 perturbation keys in total. Because BBBC021 contributes four image sites per retained well, the same frozen manifest corresponds to 905 unique wells and 3,620 image rows. The subset spans all 10 acquisition weeks and preserves the 12 active MoA classes used for headline retrieval.

Freezing this subset avoids a common comparison error because downstream differences could partly reflect a different benchmark rather than a different selector.

3.2 Shared pipeline architecture

The experimental pipeline has five stages:

1. **Segmentation and proposal generation.** Cellpose-SAM processes every image in the frozen subset and produces a set of cell proposals (bounding boxes with associated binary masks). Proposals and masks are cached and shared across all experimental conditions.
2. **Crop construction and feature extraction.** Each proposal is converted into an isolated single-cell crop, which is then embedded by InceptionV3 into a 2,048-dimensional feature vector. These per-crop embeddings are stored in a shared cache.
3. **Crop selection.** A selection strategy chooses a subset of crops per image. The compared selector conditions differ only in this step and are defined in Section 3.5. Each strategy indexes into the shared embedding cache rather than re-extracting features.
4. **Normalization and aggregation.** Selected crop embeddings are normalized with cell-level TVN and aggregated into well-level profiles.
5. **Evaluation.** Well-level profiles are evaluated using the copairs framework.

The final selector benchmark uses isolated single-cell crops. Each Cellpose-SAM proposal defines a target-cell mask and a bounding box, but only the target-cell pixels are retained in the saved crop. Neighboring cells and local background inside the same box are set to zero. This choice keeps the downstream representation aligned with a single-cell profiling interpretation and avoids mixing the effect of cell selection with the effect of retaining local context.

The key controlled-comparison choice is that the proposal set is shared. All selector conditions begin from the same Cellpose-SAM detections and use the same crop embeddings, so the selector curves reflect ranking and retention decisions rather than changes in upstream segmentation geometry or feature extraction.

3.3 Segmentation, crop construction, and shared caches

The segmentation input uses the same two-channel Cellpose convention as the implementation, with one channel emphasizing cytoplasmic structure and the other emphasizing nuclear signal.

For each image, Cellpose-SAM produces one proposal per detected cell. Each proposal contains a padded bounding box, a binary target-cell mask inside that box, the segmented cell area, the bounding-box area, and the resulting foreground fraction defined as cell area divided by bounding-box area. Cellpose-SAM was run with the `cpsam` model, channels [2, 3] for green cytoplasm and blue nucleus, automatic diameter estimation, minimum cell area 200 pixels, and 5-pixel bounding-box padding, and the proposal crops stayed at native resolution until InceptionV3 applied its internal resize during feature extraction. The segmentation outputs are cached once and then reused across downstream selector sweeps so that later experiments do not rerun GPU segmentation.

The final benchmark uses segmented single-cell crops. The crop is extracted from the Cellpose-SAM bounding box, but all pixels outside the target-cell mask are set to zero before feature extraction. The resulting image therefore preserves the target cell while removing neighboring cells and background inside the same proposal box.

A proposal is treated as border-touching when its padded proposal bounding box reaches any edge of the image. This is a crop-quality control because border-touching proposals can correspond to partial cells, truncated masks, or crops where the bounding box does not contain the full object. Removing them keeps the main comparison focused on ranking complete proposals rather than on whether a selector avoids incomplete cells.

BBBC021 preprocessing follows the same image-preparation convention used by the MorphoDiff pipeline. The raw 4',6-diamidino-2-phenylindole (DAPI), tubulin, and actin channels are independently min-max normalized, combined into one red–green–blue (RGB) image with actin in red, tubulin in green, and DAPI in blue, and then resized to the processed-image resolution used throughout the downstream pipeline. Figure 3.1 shows one concrete example of that step on a BBBC021 field.

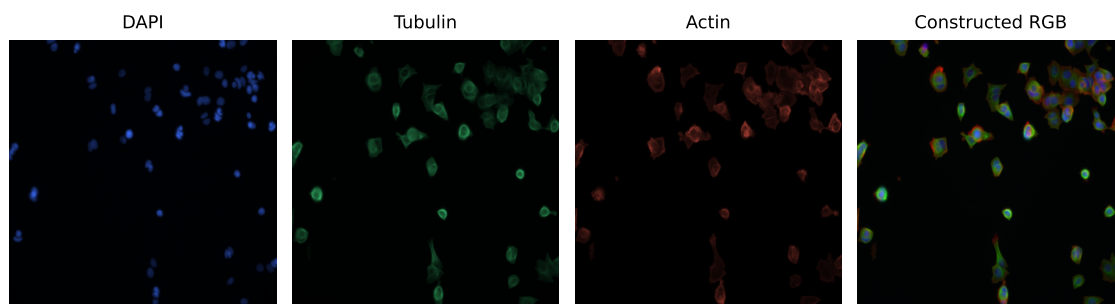


Figure 3.1: BBBC021 image preprocessing on one representative field. The three raw assay channels are normalized separately and then merged into one RGB image. In the RGB image, R = actin, G = tubulin, and B = DAPI.

Figure 3.2 shows an example of this proposal stage on one BBBC021 field. One instance mask becomes one candidate crop, and the selector only decides which of those candidate crops are kept.

Figure 3.3 makes the crop construction explicit, starting from one Cellpose-SAM proposal, the figure shows the source image, the selected proposal box, the padded bounding-box crop, and the final isolated crop. The padded box can still contain neighboring structure, but the saved crop in the final benchmark keeps only the target cell and zeros the remaining pixels inside that same box.

3.4 Feature extraction

Each retained crop is embedded with an ImageNet-pretrained InceptionV3 network whose classification head is replaced by an identity mapping [13], [14]. The resulting penultimate-layer representation is a 2,048-dimensional feature vector. The extractor handles the resizing to the native 299×299 input resolution internally, so the

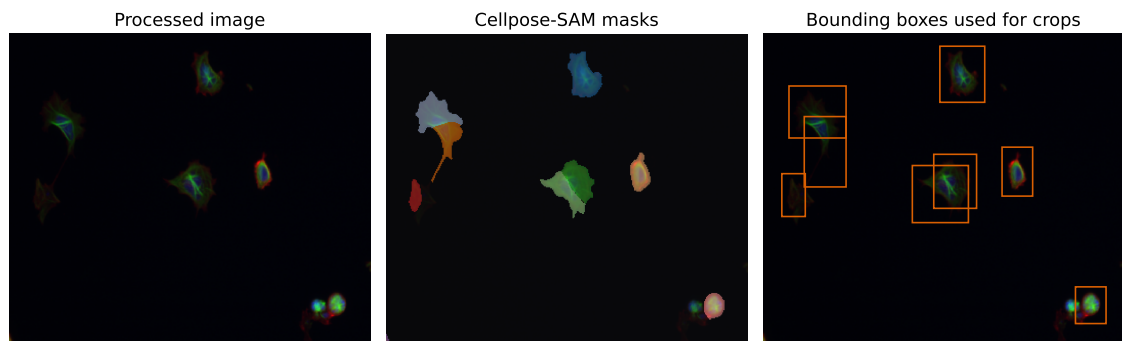


Figure 3.2: Cellpose-SAM proposal construction on one representative BBBC021 image. The left panel shows the processed assay image, the middle panel shows the instance masks returned by Cellpose-SAM, and the right panel shows the proposal bounding boxes used to extract candidate cell crops.

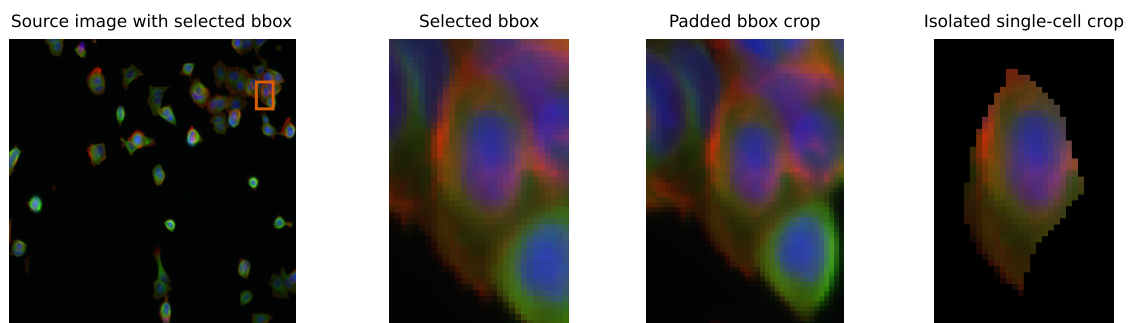


Figure 3.3: How one Cellpose-SAM proposal becomes one isolated single-cell crop. From left to right, the panels show the source image with the selected proposal, the raw proposal box, the padded bounding-box crop used by the pipeline, and the final crop after non-target pixels inside that box are set to zero.

selector study can compare different crop sets without introducing an additional custom resizing stage outside the feature extractor.

Embeddings are extracted once for the shared proposal cache and are then reused by all selector conditions. The selectors therefore index into the same embedding table rather than re-running InceptionV3 for every retention level. This keeps the comparison focused on which cells are retained, not on stochastic or implementation differences in feature extraction.

3.5 Selector definitions

The selector study compares four conditions that share the same Cellpose-SAM proposal cache and the same InceptionV3 crop embeddings. The selectors differ only in how they choose which proposals to retain from each image.

3.5.1 Baseline

The baseline keeps every valid Cellpose-SAM proposal from the shared proposal cache. It represents the all-cell reference for the fixed-geometry study. In the segmented single-cell setup, this means that all isolated target-cell crops are retained before normalization and aggregation. The baseline is not treated as an ideal end-point, but as the reference profile obtained when no cell-selection budget is imposed.

3.5.2 Random selection

Random selection keeps a fixed percentage of proposals from each image. For an image with n_{cells} detected cells and a retention level K , the retained count is

$$n_{\text{keep}} = \max \left(1, \left\lceil n_{\text{cells}} \frac{K}{100} \right\rceil \right). \quad (3.1)$$

The final random selector is evaluated with three selector seeds. For each selector seed, the selected subset is then evaluated across copairs permutation seeds. This separates variability caused by choosing different random cells from variability caused by the null permutation process. This condition measures how much profiling signal remains under simple subsampling.

3.5.3 Foreground-fraction ranking

Foreground-fraction ranking sorts the same proposals by the fraction of the bounding box occupied by the segmented target cell,

$$f_j = \frac{|M_j|}{|B_j|}, \quad (3.2)$$

where M_j is the target-cell mask of proposal j and B_j is its bounding box. Proposals with larger f_j contain less zeroed background after masking and are retained first. This selector is perturbation-agnostic and is used as a crop-quality baseline. It tests whether a simple geometric rule can preserve or improve profiling performance without using MorphoDiff.

The final selector benchmark keeps the proposal geometry fixed so that only the ranking rule changes. Representative segmentation and extraction alternatives that were considered before that design was fixed are summarized in Appendix Section A.3.1.

3.5.4 MorphoDiff-Typicality ranking

MorphoDiff-Typicality ranks proposals by the mean raw population-null typicality score inside the Cellpose-SAM target-cell mask. For proposal j , the score is

$$s_j(c) = \frac{1}{|M_j|} \sum_{u \in M_j} T_{\text{pop}}(u \mid c), \quad (3.3)$$

where $T_{\text{pop}}(u \mid c)$ is the raw population-null typicality map for the target compound c . Proposals are ranked in descending order, so cells with larger target-specific loss

contrast are retained first. The mean is used here because it matches the intended cell-level quantity, which is the average target-specific evidence over the pixels assigned to that cell. A median score would be more robust to isolated pixels, but it would also downweight localized subcellular regions where the typicality contrast is concentrated. Since the final selector is meant to rank whole-cell proposals, not to detect only the most extreme pixel inside them, the area-normalized mean is kept as the primary score. Unlike foreground-fraction ranking, this selector is perturbation-aware because the score depends on the compound conditioning used by MorphoDiff.

The final experiments use cell-mask scoring on raw heatmaps. Earlier bounding-box and transformed-heatmap variants are treated as exploratory experiments and are not part of the main comparison. The score-map-guided alternatives from that exploratory branch are summarized in Appendix Section A.3.2.

In the percentage-retention sweeps, DMSO control images are handled through a deterministic control-subset manifest. For each retention level, the manifest specifies how many DMSO cells to keep from each control image. This keeps the control-cell budget aligned with the treated-cell budget and avoids comparing downsampled treatment profiles against full-size control profiles.

3.6 Percentage-based retention sweep and seeds

Random selection, foreground-fraction ranking, and MorphoDiff-Typicality are evaluated over the retention levels $\{5, 10, 15, \dots, 100\}$. The aim is to test whether biological retrieval performance can be preserved when only a fraction of the detected cells is retained. Retention is defined as a percentage per image using Equation 3.1, so each field contributes a subset proportional to its own number of detected cells. Since BBBC021 images vary substantially in cell count, the effective number of retained cells is also reported as a diagnostic rather than assuming that a fixed percentage corresponds to the same absolute budget in every image.

Figure 3.4 gives a compact visual example of that selector step on one image. The same proposal set is shown three times, but each selector keeps a different subset of cells.

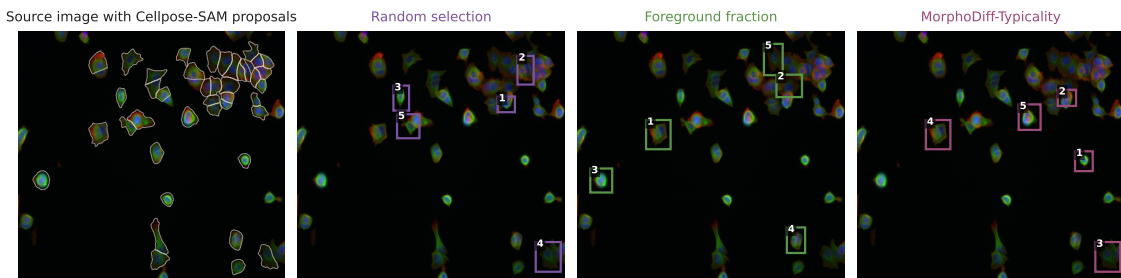


Figure 3.4: Simple comparison of crop selection on one representative BBBC021 image. The first panel shows the source image. The remaining panels show the top five retained proposals for random selection, foreground-fraction ranking, and cell-mask typicality ranking on the same field of view.

The aggregate curves use copairs seeds 42, 1995, and 2002. For the baseline and foreground-fraction sweeps, those seeds change only the copairs permutation process because the selected proposal set is deterministic. For random selection, three selector seeds (2026, 2027 and 2028) are evaluated and each is paired with the copairs permutation seeds. For MorphoDiff-Typicality, the three variants correspond to independent Monte Carlo population-null draws with the same $K_{\text{MC}} = 16$ sampling rule, using seeds 42, 2027 and 2028. The uncertainty in the final selector curves therefore has different sources for different selectors: evaluation-seed variability for deterministic selectors, selection-seed variability for random selection, and population-null sampling variability for MorphoDiff-Typicality.

3.7 Per-pixel typicality construction

The MorphoDiff-Typicality selector begins from cached MorphoDiff loss tensors rather than from precomputed cell scores. For each treated image, MorphoDiff is evaluated under the target compound conditioning and under a target-specific population-null reference. The raw cached tensor stores the null and target losses before they are reduced to an image-level map.

The final population-null reference is built from the matched BBBC021 subset. For each target compound, DMSO and the target compound itself are excluded from the null pool. The remaining eligible compounds are grouped by MoA, and the Monte Carlo null samples are drawn by first sampling an MoA uniformly and then sampling a compound uniformly within that MoA. The final experiments use $K_{\text{MC}} = 16$ null samples. Earlier checks with $K_{\text{MC}} = 8$ and $K_{\text{MC}} = 32$ made $K_{\text{MC}} = 16$ the empirical balance between null stability and computational cost. The sampled set is fixed per target compound, so all images from the same target are scored against the same population reference. Changing the sampled null compounds changes the averaged reference prediction and can therefore move the resulting typicality map and proposal ranking. This is why the final benchmark reports independent population-null draws as selector-side variability rather than treating one sampled null set as exact. A qualitative comparison between this reference and the earlier Stable Diffusion and synthetic-null alternatives is shown in Appendix Section A.3.3.

The null prediction is averaged before the mean-squared error loss is computed. If $\hat{w}_q$ is the Monte Carlo weight of null compound q , the population-null prediction is

$$\bar{\epsilon}_{\theta}^{\text{null}}(z_t, t, c) = \sum_q \hat{w}_q \epsilon_{\theta}(z_t, t, q). \quad (3.4)$$

The null loss is then computed between this weighted prediction and the diffusion noise. This matters because averaging predictions before the loss is not the same operation as averaging losses after they have been computed.

The raw loss tensor has one component for the population-null loss and one component for the target-conditioned loss. The implementation averages over the internal latent stack, upsamples the loss maps bilinearly to image resolution, and computes

the signed difference

$$T_{\text{pop}}(u \mid c) = \ell_{\theta}^{\text{pop}}(u \mid z_t, t, c) - \ell_{\theta}(u \mid z_t, t, c). \quad (3.5)$$

The repeated stochastic samples are then aggregated by the mean. In the final experiments, the resulting map is kept raw. No smoothing, centering, min-max scaling, or robust normalization is applied before cell scoring.

Proposal-level scores are obtained by averaging the raw signed map over the target-cell mask of each Cellpose-SAM proposal:

$$s_j(c) = \frac{1}{|M_j|} \sum_{u \in M_j} T_{\text{pop}}(u \mid c). \quad (3.6)$$

The selected proposals are those with the highest $s_j(c)$ values under the requested retention level. DMSO controls are not ranked by MorphoDiff-Typicality in the final experiments, they are instead selected through the shared control-subset manifest described in Section 3.5. Figure 3.5 shows the cell-mask scoring rule on one BBBC021 image.

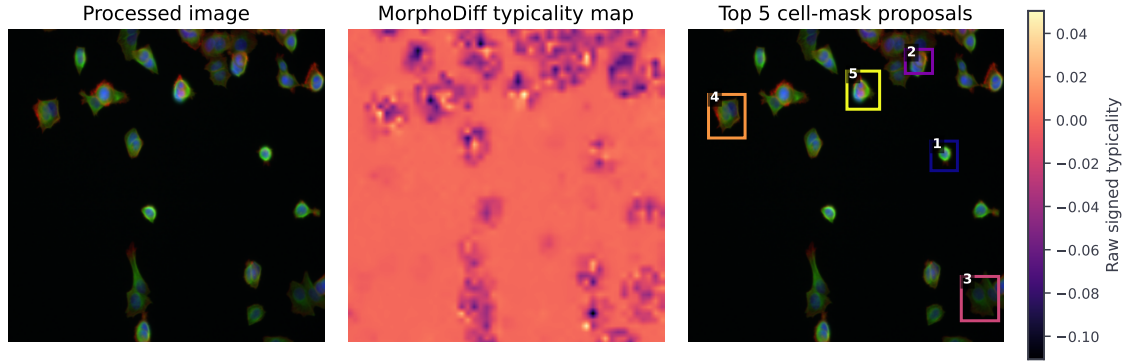


Figure 3.5: MorphoDiff-Typicality scoring on one BBBC021 image. The panels show the processed image, the raw signed population-null typicality map, and the top-ranked Cellpose-SAM proposals under cell-mask scoring. The colorbar reports raw loss-difference values.

3.8 Cell-level TVN

The chosen normalization baseline uses cell-level TVN rather than well-level TVN because, from a statistical perspective, the raw Inception representation has 2,048 dimensions, while the number of DMSO control wells available within each acquisition week is much smaller than the number of cells in those same wells. Estimating week-specific covariance structure from well profiles therefore gives a fragile basis for whitening and CORAL alignment. At cell level, the DMSO population is large enough to fit the same operations before well aggregation.

The same TVN procedure and hyperparameters are used for every selector condition. Let $x_i \in \mathbb{R}^{2048}$ be a selected crop embedding. The implementation first standardizes

embeddings using the DMSO control mean and variance, then fits a 200-component principal component analysis (PCA) model on the standardized controls. With U_{200} denoting the retained loading matrix and Λ_{200} the retained eigenvalues, the whitened representation is

$$z_i = \Lambda_{200}^{-1/2} U_{200}^\top \tilde{x}_i. \quad (3.7)$$

For each acquisition week b , the DMSO controls in this whitened space define a week mean μ_b and covariance Σ_b . They are aligned to the pooled DMSO target mean $\mu_\star$ and covariance $\Sigma_\star$ with a correlation alignment (CORAL) transform,

$$z_i^{\text{TVN}} = \Sigma_\star^{1/2} \Sigma_b^{-1/2} (z_i - \mu_b) + \mu_\star. \quad (3.8)$$

The fitted normalization artifacts are saved with each run so that aggregation and evaluation can be reproduced from the selected embeddings.

The current baseline retains 200 whitened PCA components and this value is treated as a fixed experimental setting for the selector benchmark. The role of TVN here is to reduce nuisance variation before retrieval evaluation, not to optimize the retained dimensionality of the feature space. Figure 3.6 records the practical reason for using cell-level TVN, because cell-level controls provide many more observations for fitting the normalization, and the chosen setting gives the strongest MoA profile quality among the tested normalization options.

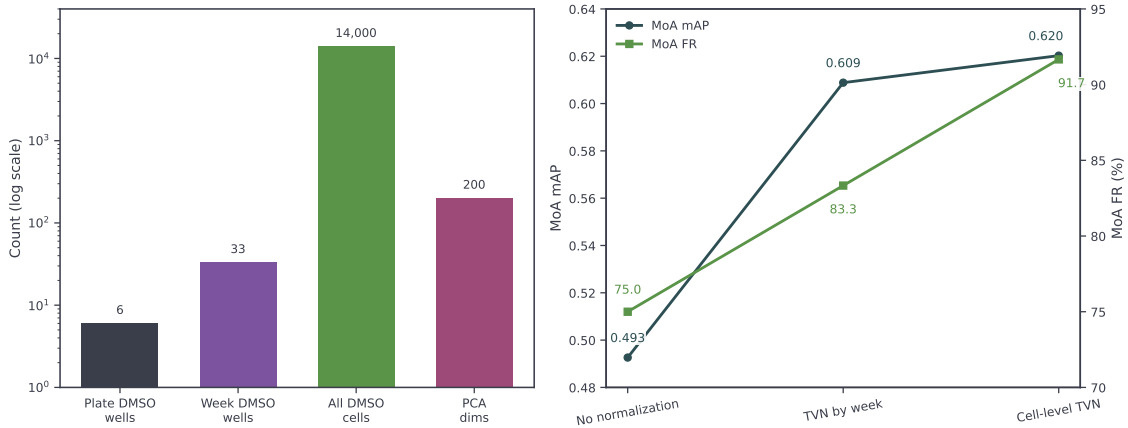


Figure 3.6: Normalization rationale on the frozen matched subset. The left panel compares the control counts available for fitting TVN, and the right panel shows the MoA trade-off for no normalization, week-level TVN, and cell-level TVN on the same subset before setting the final selector benchmark.

3.9 Evaluation protocol

3.9.1 copairs metrics and parameterization

All headline results use copairs with cosine similarity between well profiles and `n_permutations = 100`.

The concrete copairs tasks are summarized in Table 3.1. At perturbation level, positives share `Metadata_Perturbation` and differ on `Metadata_Replicate`. At MoA level, positives share `Metadata_MoA` and differ on `Metadata_Treatment`, where `Metadata_Treatment` is the perturbation key used to prevent replicate matching from inflating MoA retrieval. DMSO is retained for perturbation-level evaluation and normalization, but removed before MoA-level reporting.

Table 3.1: Metadata fields used by copairs in the present study. The table includes both the matching role and the value structure on the frozen BBBC021 subset.

Field	Role and value structure
<code>Metadata_Perturbation</code>	Compound-concentration key used to define perturbation identity. The frozen subset contains 85 retained values: 84 active perturbation keys plus <code>DMSO_0.0</code> .
<code>Metadata_Replicate</code>	Well-level replicate identifier used to require perturbation matches across different replicate wells. On the frozen subset it resolves to 905 retained well-level values.
<code>Metadata_MoA</code>	Expert mechanism-of-action label used for cross-compound biological retrieval. The retained subset contains 12 active MoA classes; DMSO is removed before MoA-level reporting.
<code>Metadata_Treatment</code>	Treatment identity used in MoA retrieval to forbid trivial same-treatment matches. In practice this is the perturbation key restricted to active perturbations.

These settings implement the copairs framework introduced in Section 2.6. They are used because the selector should preserve both within-perturbation reproducibility and broader cross-compound MoA structure. The evaluation reports both mAP and FR. mAP is interpreted as a ranking-quality metric, while FR is interpreted as a detectability metric under the retained-cell budget. This distinction is important because a selector can make many perturbations or MoA classes detectable without producing the strongest ranked neighborhoods.

The repeated selector sweeps use three copairs seeds to check whether the reported thresholds depend on the null permutation draw. The deterministic selectors keep the same selected cells across those seeds, so this check isolates evaluation-seed variability from selection variability.

3.9.2 Selector agreement and crop diagnostics

copairs metrics summarize downstream profile quality, but they do not explain why two selectors behave differently. The selector study therefore includes diagnostics at three levels: agreement between selectors, properties of the selected cells, and qualitative inspection of selected crops.

Selector agreement is measured on each image by comparing proposal rankings and selected sets for three selector pairs: foreground fraction versus MorphoDiff-Typicality, random selection versus foreground fraction, and random selection versus MorphoDiff-Typicality. Rank agreement is summarized with Spearman’s rank correlation,

$$\rho_s = \text{corr}(\text{rank}(x_i), \text{rank}(y_i)), \quad (3.9)$$

computed over proposals scored by both selectors in the same image. A value of +1 indicates identical monotonic rank order, 0 indicates no monotonic rank association, and −1 indicates a perfectly reversed rank order.

Selected-set agreement is summarized with Jaccard overlap,

$$J_K(A, B) = \frac{|S_A(K) \cap S_B(K)|}{|S_A(K) \cup S_B(K)|}, \quad (3.10)$$

where $S_A(K)$ and $S_B(K)$ are the top- $K\%$ proposal sets retained by selectors A and B in the same image. A Jaccard value of 0 means that the two selectors retain no common proposals at that budget, while a value of 1 means that they retain exactly the same proposal set. These diagnostics test whether MorphoDiff-Typicality reproduces foreground-fraction ranking, behaves like random selection, or retains a distinct subset of cells.

For each selector and retention level, the saved crop-level diagnostics include the number of selected cells per image, cell area, foreground fraction, zero fraction after masking, neighbor count within a fixed radius, and whether the proposal touches the image border. Cell area is the number of pixels in the Cellpose-SAM target-cell mask. Bounding-box area is the number of pixels in the padded proposal box. Foreground fraction is cell area divided by bounding-box area, and zero fraction after masking is the complementary fraction of the proposal crop that is set to zero when non-target pixels are removed. Neighbor count is computed from proposal centroids as the number of other detected proposals whose centroid lies within 64 pixels of the selected proposal centroid. These diagnostics are used to test whether a selector mainly changes biological signal, crop quality, local crowding, or proposal completeness.

The analysis also reports per-MoA changes in retrieval performance relative to the baseline. This is because a selector can improve the aggregate metric while harming specific MoA classes, or preserve detectability for many MoAs while reducing ranking quality for others. Qualitative galleries are then used to inspect representative agreement cases, foreground-only cases, typicality-only cases, and failure cases with the Cellpose-SAM polygons visible.

3.10 Compute environment

All experiments are executed on the Alvis cluster through National Academic Infrastructure for Supercomputing in Sweden (NAISS). The pipeline is split into reusable stages to minimize redundant graphics processing unit (GPU) work:

- **GPU stages** (executed once): Cellpose-SAM proposal generation, MorphoDiff typicality scoring, and InceptionV3 embedding extraction. These are submitted as Slurm Workload Manager (Slurm) jobs with GPU allocation.
- **central processing unit (CPU) stages** (executed per condition and per K value): crop selection, cell-level TVN normalization, well-level aggregation, and copairs evaluation.

By caching the expensive GPU outputs (segmentation masks, typicality maps, crop embeddings) and sharing them across all conditions, the full K sweep for a single condition reduces to lightweight CPU operations on pre-computed arrays. This reduces queue pressure, shortens reruns after failures, and makes the 20-point K sweep across four conditions feasible within the project timeline.

4

Results

This chapter reports the final selector comparison on the frozen BBBC021 subset. All selector conditions use the same Cellpose-SAM proposal cache, remove proposals touching image borders, isolate the target-cell pixels inside each crop, embed crops with InceptionV3, normalize with cell-level TVN, and evaluate the resulting well profiles with copairs. Under this setup, movement in the curves reflects which complete cell proposals are retained rather than changes in segmentation geometry, crop context, feature extraction, or the benchmark subset.

4.1 Baseline

The baseline keeps every valid Cellpose-SAM proposal after border filtering. It retains 389,448 cell crops across the frozen subset, corresponding to a mean of 108.2 retained crops per image and a median of 77.0. The resulting profiles reach perturbation mAP 0.5865, perturbation FR 57.65%, MoA mAP 0.5873, and MoA FR 91.67%. These values are used as the reference for the selector curves in the rest of the chapter.

Removing border-touching proposals changes the scale of the experiment in a visible way, because the baseline retention count drops from 494,749 to 389,448 crops, and the mean image-level count drops from 137.4 to 108.2 cells. The selector trends are still comparable, but the retained-cell budget is different enough that the rest of the chapter uses the same proposal cache consistently.

In the selector analysis, a selector is said to *reach* the baseline at the first retained percentage whose mean curve meets or exceeds the baseline value.

4.2 Selector comparison across retention budgets

Figure 4.1 shows the main selector comparison. Random selection is the subsampling reference, foreground fraction is the perturbation-agnostic crop-quality selector, and MorphoDiff-Typicality is the perturbation-aware selector.

Random selection only reaches the perturbation mAP and perturbation FR baselines at 85%, reaches MoA FR at 85%, and reaches MoA mAP only at 100%. This behavior is useful as a reference because it shows that the baseline profile is not recovered automatically at low budgets by keeping arbitrary cells.

4. Results

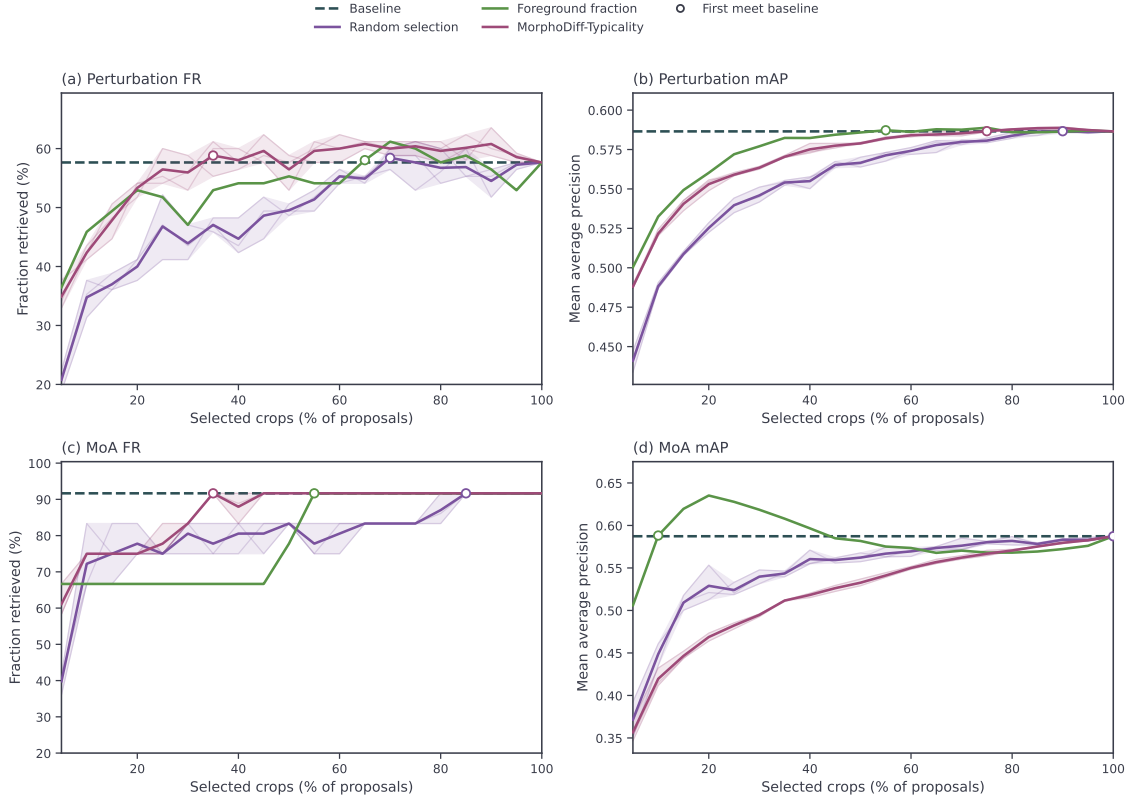


Figure 4.1: Fixed-geometry selector comparison after border filtering. The dashed line marks the baseline and open circles mark the first retained percentage that meets or exceeds that baseline. For the baseline and foreground-fraction selector, variation comes from copairs permutation seeds because the selected cell subset is fixed. For random selection, faint replicate lines and shaded regions summarize three random selector seeds. For MorphoDiff-Typicality, faint replicate lines and shaded regions summarize three independent population-null Monte Carlo draws with the same $K_{MC} = 16$ sampling rule.

Foreground-fraction ranking is strongest for MoA mAP. It reaches the MoA mAP baseline at 10% retained crops and peaks at 20% with MoA mAP 0.6352, which is 0.0479 above the baseline. It also reaches the perturbation mAP baseline at 55%, perturbation FR at 65%, and MoA FR at 55%. The MoA mAP gain is not a small late-budget effect. It appears when the selector keeps a relatively small subset of crops with high target-cell occupancy.

MorphoDiff-Typicality gives the strongest perturbation FR behavior. It reaches the perturbation FR baseline at 35% and peaks at 45% with perturbation FR 62.35%, which is 4.71 percentage points above the baseline. It reaches perturbation mAP at 75%. It also reaches the MoA FR baseline at 35%, although it does not reach the MoA mAP baseline. This separation matters because typicality is not simply reproducing the foreground-fraction curve. It preserves perturbation-level detectability better than the other selectors, while foreground fraction produces the clearest MoA mAP gain.

The copairs permutation seeds change the exact metric estimates but not the threshold conclusions for the baseline, foreground-fraction ranking, or the default MorphoDiff-Typicality sweep. In those cases, the selected cell subsets are fixed and only the copairs null permutations change. Random selection and MorphoDiff-Typicality add selector-side variability for different reasons, random selection intentionally chooses different cell subsets, while the MorphoDiff-Typicality replicates recompute the selector from independent population-null Monte Carlo draws with the same $K_{MC} = 16$ rule. The threshold statements above are therefore based on the selector behavior that remains stable under these repeated evaluations, while the visible spread in Figure 4.1 shows how much the repeated selector runs move around the mean curve.

Table 4.1 lists for each selector, the first retained percentage at which the baseline is reached and the best value obtained on the corresponding curve. Lower values are better for the first- K columns because they indicate that the selector reaches the baseline at a smaller retained budget.

Table 4.1: Selector comparison against the all-cell baseline in Figure 4.1. For each selector, the table reports the first retained percentage K at which the mean curve meets or exceeds the baseline for perturbation FR, perturbation mAP, and MoA FR, together with the best perturbation FR and best MoA mAP values observed anywhere on the curve. Lower first- K values indicate that a selector reaches the baseline with fewer retained cells, while higher values are better for the two best-value columns. The retained percentage at which each best value occurs is shown in parentheses.

Selector	First K Pert. FR	Best Pert. FR	First K Pert. mAP	First K MoA FR	Best MoA mAP
Random selection	85%	60.00% (85%)	85%	85%	0.5873 (100%)
Foreground fraction	65%	61.18% (70%)	55%	55%	0.6352 (20%)
MorphoDiff-Typicality	35%	62.35% (45%)	75%	35%	0.5873 (100%)

Foreground fraction is the strongest selector for compact MoA ranking quality, while MorphoDiff-Typicality is the strongest selector for perturbation-level fraction retrieved. Which selector is preferable depends on whether the downstream goal is to sharpen mechanism-level neighborhoods or to preserve detectable perturbation profiles under reduced cell budgets.

4.3 Retention budget and crop-count distribution

Percentage retention is applied per image, but BBBC021 images do not contain the same number of detected cells. Figure 4.2 reports the effective retained-cell budget after border filtering.

At 25%, each selector keeps 98,717 cells, with a mean of 27.4 cells per image and a median of 20.0. At 50%, each selector keeps 195,643 cells, with a mean of 54.4 cells per image and a median of 39.0. These values matter because the low-budget MoA mAP peak for foreground fraction occurs when most images contribute only

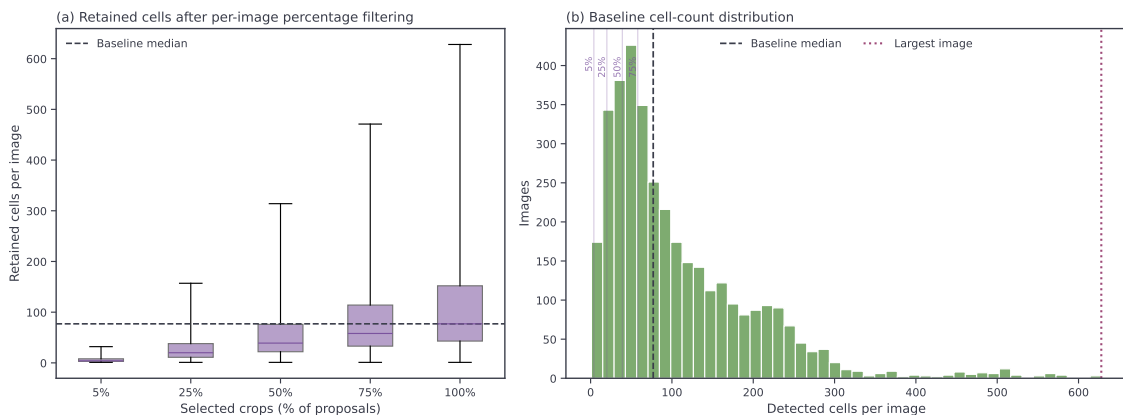


Figure 4.2: Effective retained-cell budget after border filtering. (a) Box plots show the per-image retained-cell distribution at representative percentage budgets. (b) The histogram shows the full-retention detected-cell distribution across images, with reference lines marking representative percentage budgets and the baseline median.

a few dozen cells. The selector is therefore changing both which cells are retained and how much each image contributes to the well-level profile.

The crop-count distribution also clarifies the moderate-budget typicality result, since MorphoDiff-Typicality reaches its best perturbation FR at 45%, where the retained budget is 176,977 cells and the mean retained count is 49.2 cells per image. This suggests that the typicality signal benefits from retaining enough cells for perturbation-level aggregation, rather than from selecting only the smallest possible subset.

4.4 Foreground fraction as a crop-quality signal

Foreground-fraction ranking is deliberately simple since it does not use perturbation identity, MorphoDiff losses, or molecular conditioning. It ranks each proposal by the fraction of its bounding box occupied by the segmented target cell. Its MoA mAP behavior therefore needs to be interpreted as a crop-quality result before it is interpreted biologically.

The selected-cell diagnostics in Figure 4.3 use the crop definitions from Section 3.9.2 and are reported as shifts relative to the full-retention baseline rather than as absolute crop properties. Foreground fraction and zero fraction after masking are shown as percentage-point differences from the baseline means, while cell area and neighbor count are shown as ratios to the corresponding baseline means. The baseline means are 0.352 for foreground fraction, 0.648 for zero fraction after masking, 839.0 pixels for cell area, and 9.57 neighbors for local crowding. In absolute terms, at 25% retention, foreground-fraction ranking raises the mean foreground fraction to 0.423 and lowers the zero fraction after masking to 57.7%. Random selection at 25% stays close to the baseline with mean foreground fraction 0.353, while MorphoDiff-Typicality at 25% is also much closer to the baseline at 0.362. Foreground fraction is therefore selecting a different crop population, not only a different ordering of similar-looking cells.

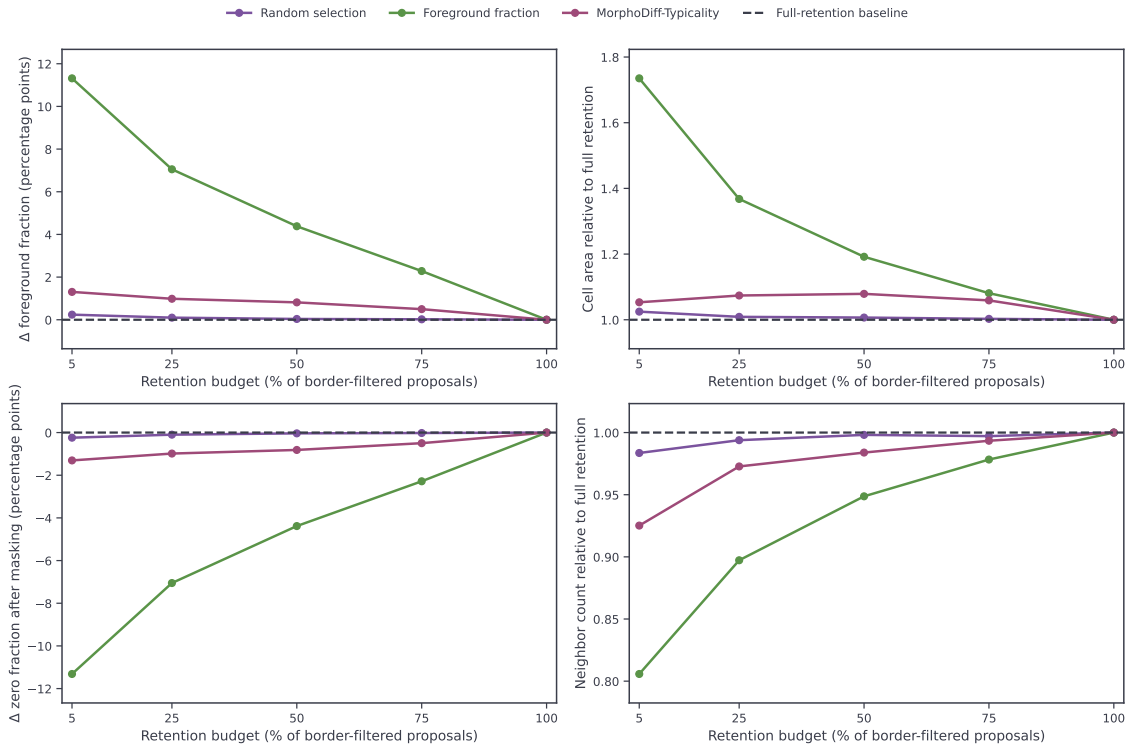


Figure 4.3: Selected-cell property shifts relative to the full-retention baseline across retention budgets. Foreground fraction and zero fraction after masking are shown as percentage-point differences from the full-retention baseline means, while cell area and neighbor count are shown as ratios to the corresponding full-retention baseline means. Metric definitions are given in Section 3.9.2. The dashed horizontal line marks no change relative to full retention.

These diagnostics support crop quality as one explanation for the foreground-fraction result. Crops with less zeroed background may give InceptionV3 and cell-level TVN a cleaner representation, which can improve well-level MoA ranking even without perturbation-aware selection. A second explanation is that foreground fraction may correlate with phenotype in this dataset. Some perturbations change cell size, spreading, cytoskeletal organization, or segmentation completeness, and those changes can affect how much of the crop box the target cell occupies. The present experiments do not separate these explanations but they do show that foreground fraction is a competitive perturbation-agnostic baseline for any claim that MorphoDiff-Typicality alone explains selector behavior.

4.5 Selector agreement

Figure 4.4 compares selector agreement for three selector pairs. Foreground fraction and MorphoDiff-Typicality show weak but positive rank agreement, with median Spearman correlation $\rho_s = 0.173$ across images. In contrast, random selection is essentially uncorrelated with both foreground fraction and MorphoDiff-Typicality, with median correlations close to zero. The Jaccard overlaps show the same pattern

4. Results

at the level of selected cells where the overlap increases with the retention budget, as expected, but foreground fraction and MorphoDiff-Typicality remain only partially overlapping. This supports the interpretation that MorphoDiff-Typicality is not simply a noisy version of the foreground-fraction selector.

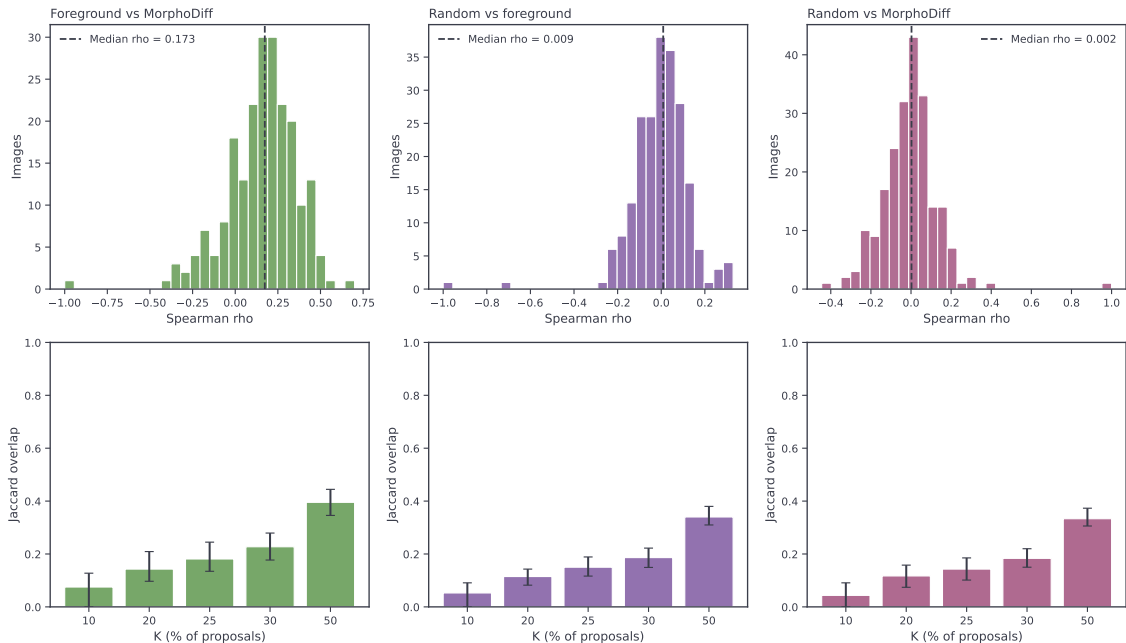


Figure 4.4: Pairwise selector agreement across images. The top row shows Spearman rank correlations between proposal rankings for foreground fraction versus MorphoDiff-Typicality, random selection versus foreground fraction, and random selection versus MorphoDiff-Typicality. The bottom row shows the corresponding Jaccard overlap between top-K% selected proposal sets at fixed retention budgets. Spearman correlation measures monotonic rank agreement, while Jaccard overlap measures direct selected-cell overlap. Variation reflects the distribution across images in the border-filtered proposal set.

The agreement diagnostics are consistent with the different metric profiles. Across images, the median Spearman correlation between foreground-fraction and typicality ranks is 0.173. The median Jaccard overlap is 0.152 at 25% retention and 0.375 at 50% retention. If MorphoDiff-Typicality were only a noisy version of foreground fraction, the typicality curve would be expected to follow the same MoA mAP pattern more closely. Instead, foreground fraction strongly enriches for high-foreground crops, while MorphoDiff-Typicality stays closer to the baseline crop geometry and gives its strongest gain in perturbation FR. This supports the interpretation that typicality carries a different signal, even though that signal is not the one that maximizes MoA mAP in this benchmark.

The same pattern is visible when agreement is summarized by MoA. Aurora kinase inhibitors have the strongest foreground–typicality agreement, with median Spearman correlation 0.423 and median Jaccard overlap 0.297 at 25%. Cholesterol-lowering compounds and Eg5 inhibitors have weak or negative median rank agreement, with Spearman correlations -0.111 and -0.025 , respectively. These values are useful

because they show that disagreement is not uniformly distributed across the benchmark. Some MoA classes produce cells that both selectors rank similarly, while others make the diffusion-based and geometry-based rules look at different parts of the same fields of view.

Figure 4.5 shows the same agreement structure at image level. In the high-agreement example, both selectors prioritize the same clean target-cell proposals. In the partial-overlap example, the two selectors share some local regions but disagree on the remaining top-ranked cells. In the no-overlap example, the highest-ranked foreground-fraction proposal and the highest-ranked MorphoDiff-Typicality proposal are different cells in the same zoomed field of view. These examples make the aggregate agreement metrics easier to interpret where we can see that the selectors can sometimes converge on the same obvious cells, but they can also prioritize different local evidence within the same proposal set.

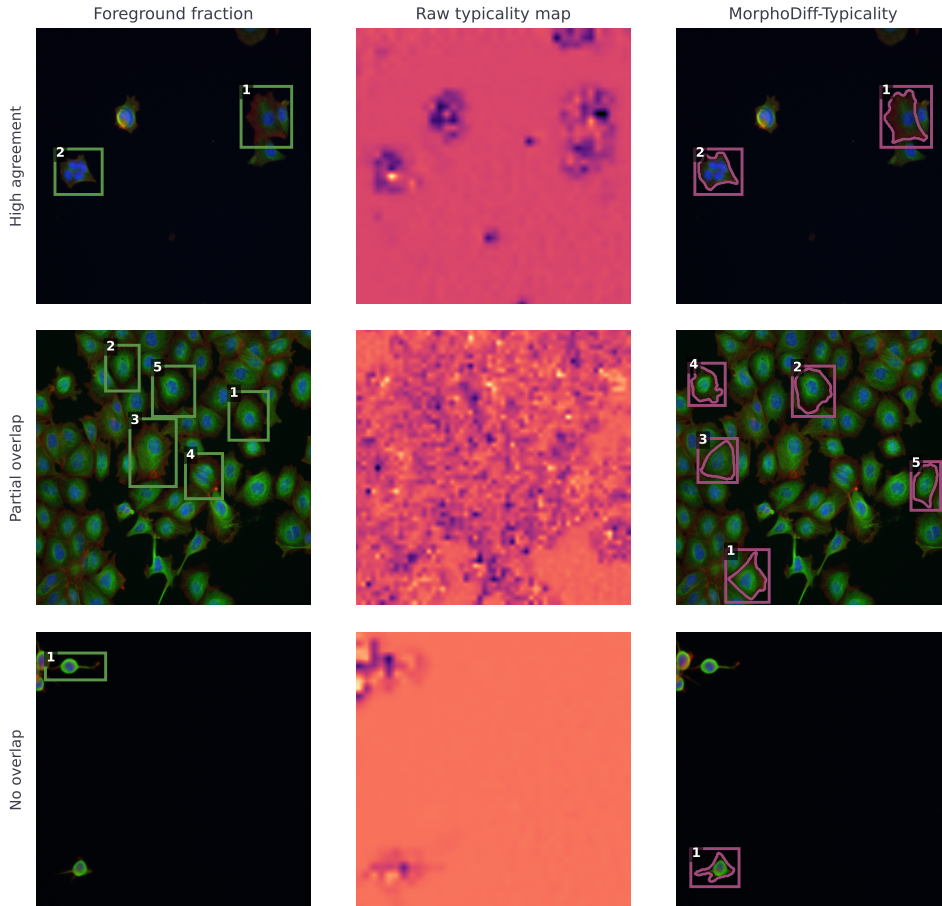


Figure 4.5: Qualitative agreement and disagreement between foreground fraction and MorphoDiff-Typicality. Green boxes mark the top foreground-fraction proposals and plum boxes mark the top MorphoDiff-Typicality proposals on the same Cellpose-SAM proposal set. Each row uses a matched zoomed field of view across the image and heatmap panels, rather than the full microscopy image. The heatmap panel shows the raw signed typicality map with display scaling only; proposal ranking uses the raw cell-mask mean scores.

4.6 MoA-level gains and failure modes

Figure 4.6 resolves MoA-level retrieval as a function of the retained-cell budget K for the two selectors that matter most in the main comparison: foreground fraction and MorphoDiff-Typicality (random-selection reference is in Appendix Figure A.1). The sweep shows that aggregate MoA mAP hides a large class-level split. The selectors do not move all mechanisms in the same direction, and the same retained-cell budget can sharpen some MoA neighborhoods while weakening others.

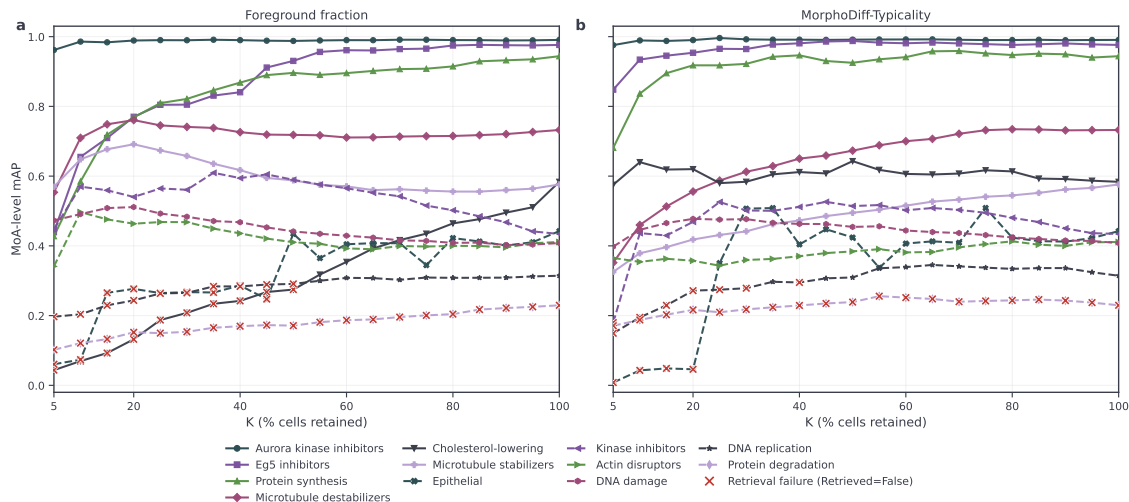


Figure 4.6: MoA-level retrieval (mAP) versus retained-cell budget K for foreground fraction and MorphoDiff-Typicality. Each curve is one MoA. Red crosses mark (selector, K) points where the class is not retrieved under the permutation-null threshold.

At 25% retention, foreground fraction is positive for five of the twelve MoA classes. Its largest gains are for kinase inhibitors (+0.131), microtubule stabilizers (+0.098), DNA damage (+0.085), and actin disruptors (+0.057) relative to the full-retention baseline. The same selector sharply hurts cholesterol-lowering compounds (−0.396), epithelial compounds (−0.178), Eg5 inhibitors (−0.172), and protein synthesis inhibitors (−0.135). The low-budget foreground-fraction peak in aggregate MoA mAP therefore does not mean that the selector is uniformly improving mechanism-level retrieval. It is selectively helping some classes while removing cells that matter for others.

MorphoDiff-Typicality has a different class profile. At 25%, it is slightly positive for Aurora kinase inhibitors (+0.005) and stays close to the baseline for cholesterol-lowering compounds (−0.004) and Eg5 inhibitors (−0.011), but it reduces microtubule destabilizers and microtubule stabilizers (−0.145) each, epithelial compounds (−0.091), and actin disruptors (−0.068). By 50%, it is positive for six of the twelve MoAs, including cholesterol-lowering compounds (+0.059), DNA damage (+0.046), kinase inhibitors (+0.080), Eg5 inhibitors (+0.011), protein degradation (+0.009), and Aurora kinase inhibitors (+0.001). DNA replication illustrates the difference in failure profile most clearly, where MorphoDiff-Typicality recovers the class by 45%,

foreground fraction does not recover it until 55%, and random selection not until 85% (Appendix Figure A.1).

Some failures are clearly not selector-specific since protein degradation is never retrieved by any selector, including at 100% retention, so that class should be interpreted as a limitation of the representation or evaluation setup rather than as a consequence of the ranking rule alone.

The crop-property audit in Appendix Table A.2 shows why simple geometry is not a complete explanation for these curves. At 25%, foreground fraction raises mean foreground occupancy for every MoA by 6.1 to 8.2 percentage points and increases mean selected-cell area by $1.24\times$ to $1.69\times$ relative to full retention, yet only five MoAs improve. MorphoDiff-Typicality stays much closer to the baseline crop geometry for most classes, usually within -1.2 to $+3.4$ percentage points in foreground occupancy and $0.55\times$ to $1.23\times$ in mean cell area, but it still produces its own class-specific gains and losses. These results are consistent with selector differences that go beyond a single shape or size statistic.

The local-neighborhood diagnostic leads to the same conclusion since at 25%, foreground fraction lowers the mean neighbor count for every MoA, with per-MoA ratios from $0.81\times$ to $0.98\times$ relative to full retention, while also selecting larger and higher-occupancy cells in every class. This means that the foreground-fraction selector is consistently moving toward larger, less crowded, less zero-padded crops. However, the MoA outcome still splits into gains and losses. The physical interpretation is therefore constrained because cell size, crop occupancy, and local crowding are all affected by the selector, but none of them alone explains why one mechanism improves while another degrades.

The same appendix table also makes the coverage limitation explicit because Eg5 inhibitors and epithelial compounds contribute only one compound each in the frozen subset, and several other MoAs are represented by only two or three compounds. Dose and compound identity may therefore contribute to some of the MoA curves, but the present benchmark is too small to separate those effects cleanly from the MoA label itself. Taken together, these results show that selectors change the MoA profile in a class-specific way, and those changes are only partly explained by simple crop-geometry shifts.

Figure 4.7 isolates the top-ranked MorphoDiff-Typicality crops for the six visual-audit MoA classes. It is drawn from the full border-filtered proposal pool and ranks crops by their raw cell-mask typicality score. A diversity pass prevents repeated image stems within each MoA row, but it does not force distinct compounds or doses. In the frozen subset used here, five of the six displayed MoA classes contribute only one scored compound, usually across two doses, so repeated compound labels are a limitation of the available pool rather than evidence of within-MoA diversity. The gallery should therefore be interpreted as a sanity check of high-scoring crops, not as evidence that the top-ranked cells are diverse within each MoA. Some high-scoring crops are clean single-cell examples, while others contain crowded or partial local structure that still receives a high typicality score.

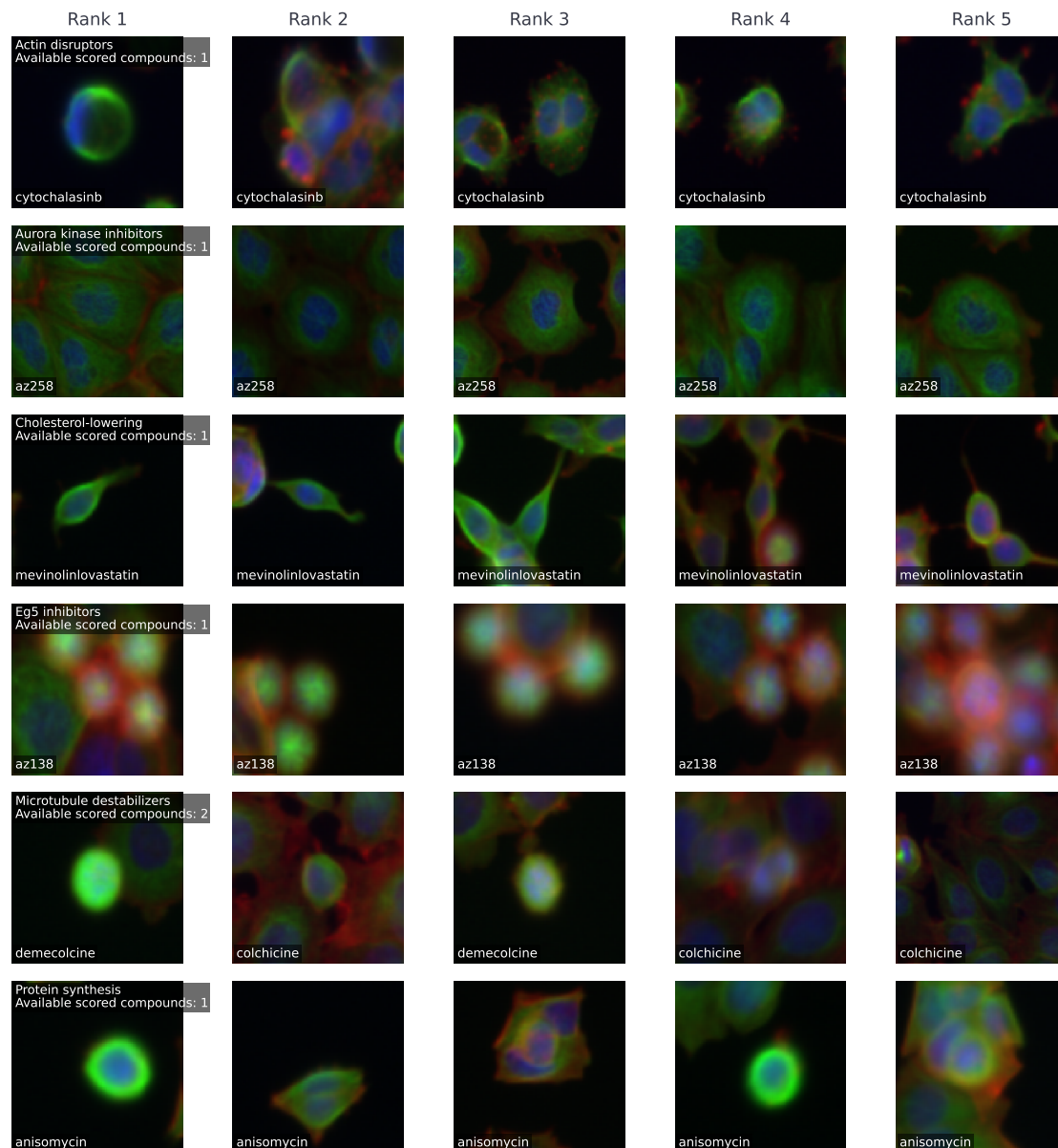


Figure 4.7: Top MorphoDiff-Typicality crops for the six visual-audit MoA classes. Crops are selected from the full border-filtered proposal pool, corresponding to the baseline selector at 100% retention, and ranked within each MoA by raw cell-mask MorphoDiff-Typicality score. Columns show ranks 1–5 after a diversity pass that prevents repeated image stems within each MoA row. The first tile of each row reports how many scored compounds were available for that MoA in the frozen subset. Repeated compounds or doses still occur because many displayed MoA rows have only one available scored compound in that pool.

A second qualitative failure mode appears when MorphoDiff-Typicality produces a strong response in a region that is not covered by any valid Cellpose-SAM proposal. Because the final selector can only retain regions that are covered by valid proposals, these out-of-proposal responses are never turned into cell crops. More generally, any systematic bias in the upstream proposal cache would propagate into all downstream

selectors, because the ranking stage can only choose among detected and retained proposals, so it cannot recover cells that were missed, merged, or filtered before scoring.

In Figure 4.8(a), the typicality hotspot sits on a real cell, with a clear DNA/DAPI nucleus and visible actin and tubulin structure, that Cellpose-SAM simply failed to segment. Here the typicality map is highlighting visually or biologically relevant structure, but the fixed proposal set prevents that cell from being represented cleanly as an independent cell crop, so the failure belongs to the segmentation-and-proposal stage. In Figure 4.8(b), the hotspot is a small, single-channel bright speck with no nucleus and no tubulin, that is, a staining or imaging artifact rather than a cell. In this case Cellpose-SAM was right not to propose a cell there, and the high typicality value is a false positive of the MorphoDiff score itself.

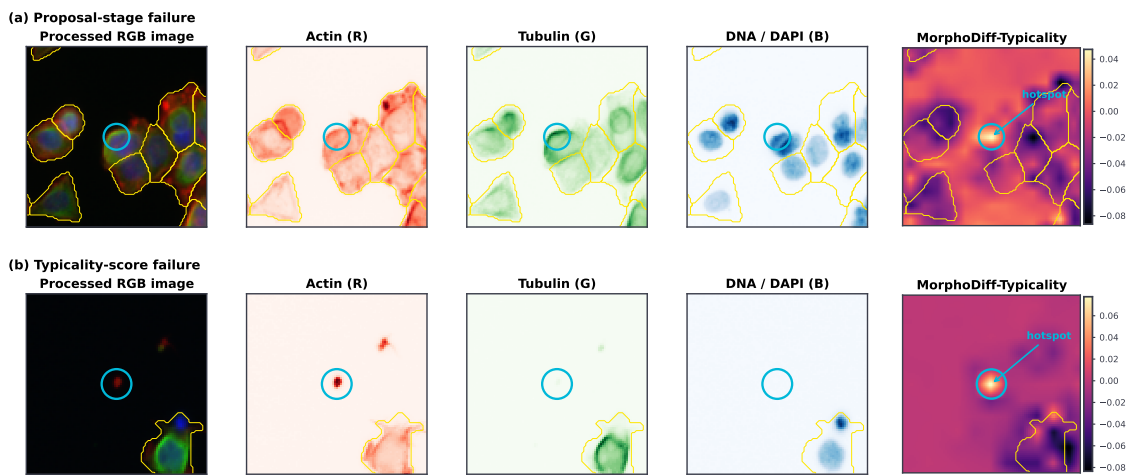


Figure 4.8: Two distinct failure modes of the MorphoDiff-Typicality cell selector. Each row shows, from left to right, the processed RGB image, its individual fluorescence channels (R: actin, G: tubulin, B: DNA/DAPI), and the raw MorphoDiff-Typicality map. Yellow outlines denote Cellpose-SAM cell masks, and the cyan circle marks the strongest typicality response that falls outside the valid proposal set. (a) Proposal-stage failure: the hotspot lands on a genuine nucleated cell with a clear DAPI nucleus and actin/tubulin structure that Cellpose-SAM failed to segment, so the selector cannot retain it despite a high typicality score. (b) Typicality-score failure: the hotspot is an ~ 8 -pixel, single-channel (actin-only) bright speck with no nucleus and no tubulin to which MorphoDiff-Typicality assigns its strongest local response (a false positive); here Cellpose-SAM correctly rejected the region, and the failure lies in the typicality score itself.

4.7 Summary

The selector comparison gives a controlled but mixed result. Foreground-fraction ranking is the clearest selector for MoA mAP and reaches its best value at low retention. MorphoDiff-Typicality is the clearest selector for perturbation FR and

reaches that baseline much earlier than random selection. Random selection only recovers the baseline at high retention.

The result does not reduce to one selector replacing the others. Foreground fraction shows that generic crop quality is a strong factor in BBBC021 profiling, especially for mechanism-level ranking. MorphoDiff-Typicality shows that perturbation-conditioned diffusion losses can still preserve useful biological signal beyond random subsampling, but that signal is most visible in perturbation-level detectability rather than MoA mAP. The selector should therefore be chosen according to the profiling question rather than by collapsing all metrics into one ranking.

5

Discussion

The final experiments suggest that MorphoDiff-Typicality can be used as a cell-level selector for image-based profiling, but not as a general replacement for simpler crop-quality rules. MorphoDiff-Typicality preserves perturbation-level signal under reduced cell budgets and performs better than random selection for perturbation fraction retrieved. Foreground-fraction ranking is stronger for MoA mAP. The two rules therefore retain different parts of the profiling signal.

5.1 Interpretation of the foreground-fraction result

The foreground-fraction result is one of the most important findings in the thesis. It is also the result that most needs interpretation because the selector is deliberately simple. It uses no perturbation labels, no diffusion model, and no molecular conditioning. It only ranks proposals by how much of the crop box is occupied by the target-cell mask, but it reaches the MoA mAP baseline at low retention and gives the highest MoA mAP observed in the final selector comparison.

The most direct explanation is that crop quality is a strong nuisance variable in the BBBC021 setup used here. The final crops are masked so that non-target pixels inside the bounding box are set to zero. A crop with a low foreground fraction therefore contains a large amount of zeroed area relative to the cell. This can affect the InceptionV3 embedding even when the biological cell mask is correct, because the feature extractor was not trained specifically for zero-masked microscopy crops. Selecting high-foreground-fraction cells may give the feature extractor a cleaner and more stable visual input, which then improves well-level aggregation and MoA ranking.

A second explanation is that foreground fraction may not only capture crop quality but may also correlate with phenotype. Some perturbations change cell size, spreading, cytoskeletal organization, or segmentation completeness. A rule that prefers cells occupying more of their crop can therefore select cells that are both technically cleaner and biologically more coherent. The present experiments cannot fully separate these two explanations. However, they show that foreground fraction must be treated as a strong baseline. Any claim about perturbation-aware selection has to be made relative to this simple crop-quality rule, not only relative to random

subsampling.

The per-MoA diagnostics make this interpretation more concrete because at the 25% budget, foreground-fraction ranking increases selected-cell occupancy and mean cell area for every MoA and also shifts the selected cells toward lower local crowding. These are exactly the kinds of physical and neighborhood properties that could plausibly change under perturbation or affect segmentation quality. However, the same directional shifts occur for MoAs that improve and for MoAs that degrade. The data therefore support that foreground fraction is tied to cell size, zero-padded crop area, and local neighborhood, but the present benchmark does not identify a single physical cause for the MoA-specific differences.

5.2 Interpretation of MorphoDiff-Typicality results

MorphoDiff-Typicality contributes mainly at the perturbation level. It does not improve MoA mAP, but it achieves the strongest perturbation FR and reaches that baseline earlier than random selection. The distinction is important because perturbation FR and MoA mAP measure different properties of the profile space. Perturbation FR asks how many perturbations remain detectably reproducible after selection. MoA mAP asks whether cross-compound mechanism neighborhoods become more compact and well ranked.

The typicality result suggests that population-null MorphoDiff losses can identify cells that preserve perturbation identity, even when those cells are not the same cells preferred by crop geometry. This is consistent with the design of the typicality score, which compares how well the target conditioning explains each spatial region relative to a population-null reference. The score is therefore expected to be sensitive to target-specific morphology, not necessarily to the generic crop properties that improve MoA-level ranking after aggregation. This is also consistent with the training setup of MorphoDiff where the model learns perturbation-conditioned image structure, but it is not trained with a downstream profiling objective.

This also explains why the typicality result is clearer at the perturbation level than at the MoA level. The target-conditioned loss is tied to the compound used for the image, whereas MoA retrieval requires generalization across different compounds that share a mechanism. A cell can therefore be useful for preserving perturbation identity without improving the cross-compound neighborhood structure used by MoA mAP. This interpretation is consistent with the metric split observed in the final benchmark, but it should not be read as proof that the selected cells are mechanistically specific at the single-cell level.

The fact that typicality performs best at moderate retention also matters because its strongest perturbation FR appears around 45%, not at the smallest budgets. This suggests that the score is more useful for enriching a profile with perturbation-relevant cells than for replacing the profile with only a very small set of top-ranked cells. In a screening pipeline, MorphoDiff-Typicality may therefore be more useful as a controlled pruning or weighting signal than as an aggressive top-cell filter.

These metric differences correspond to different downstream uses. If the goal is early hit-finding or triage, a selector that retrieves more perturbations can be valuable even when the resulting neighborhoods are less clean, because the priority is to keep a broad set of reproducibly active perturbations available for follow-up. In that setting, missing a weak but real hit can be worse than carrying forward a less compact neighborhood. If the goal is MoA annotation, compound clustering, or building compact mechanism-level neighborhoods, then MoA mAP is more important because it rewards well ordered cross-compound structure. A generative use of MorphoDiff would be, for example, perturbation-conditioned image simulation or counterfactual inspection of morphology, whereas the selector studied here asks which observed cells should contribute to downstream profiles.

5.3 Relation to previous work

This work follows the image-based profiling literature in which Cell Painting images are converted into profiles and evaluated through perturbation or MoA retrieval [1], [2], [7], [9]. The thesis uses the same kind of retrieval logic by reporting mAP and FR instead of relying on visual inspection alone. This is important for typicality-based selection because plausible heatmaps do not by themselves show that selected cells preserve useful well-level profiles.

The contribution of this thesis is not a new profiling backbone or a new segmentation model. Instead, it treats cell selection itself as the experimental object. Standard profiling pipelines often handle crop quality as a preprocessing step before profile construction. Here, selection is evaluated after fixing the proposal geometry, crop representation, embedding model, normalization, and retrieval protocol. This makes it possible to ask a narrower question: how does the choice of retained cells change the biological signal available to the final well profiles?

The selector benchmark also connects to single-cell and crop-level profiling. Tools such as CellProfiler and SPACe extract cell-level measurements before aggregation, and recent resources such as JUMP make it possible to study profiling behavior across many chemical and genetic perturbations [3], [23], [24].

The diffusion component places the work next to diffusion-based interpretability rather than only next to conventional profiling. Diff-Mining uses diffusion losses as spatial evidence for condition-specific regions, and MorphoDiff provides a perturbation-conditioned diffusion model for Cell Painting images [4], [5]. The present thesis therefore evaluates a repurposed generative model, not a diffusion model trained directly for cell selection or profiling retrieval. In parallel, self-supervised vision models such as self-distillation with no labels (DINO) have shown that learned image representations can capture single-cell morphology without manual labels [25], [26]. This thesis does not compare against those feature extractors directly, but they define an important direction for future work because the selector may behave differently when the embedding model is trained on microscopy rather than natural images.

5.4 Limitations

The conclusions are limited by the experimental scope because the final benchmark uses one dataset, one frozen matched subset of BBBC021, one Cellpose-SAM proposal cache, one MorphoDiff checkpoint, one population-null construction, one crop embedding model, and one normalization family. The result should therefore be interpreted as evidence for this controlled setting, not as a general statement about all Cell Painting screens.

A second limitation is that MorphoDiff was not trained for the task evaluated here. The model was trained to generate perturbation-conditioned Cell Painting images, with chemical or genetic perturbation encodings used as conditioning inputs. In this thesis, the same denoising model is repurposed as a source of spatial scores for cell selection. The resulting loss contrast can be useful, but it is not optimized for single-cell typicality, crop ranking, well-level retrieval, or MoA separation. This limits how strongly the typicality score can be interpreted, and it reflects what the deployed MorphoDiff model finds easier to explain under one conditioning than another, not a supervised estimate of biological cell relevance.

The population-null score also limits the interpretation of typicality since it measures target specificity relative to a sampled population of other active compounds, not specificity relative to untreated DMSO morphology. This was a deliberate choice because the deployed MorphoDiff model does not provide a calibrated unconditional or DMSO condition, but it means that the score should not be described as a direct treated-versus-control effect. A different null construction could change which cells are ranked as typical.

The molecular-conditioning requirement also limits the benchmark. MorphoDiff-Typicality can only be computed for perturbations with available molecular representations, which removes part of the original BBBC021 evaluation set. This is handled by freezing a matched subset, but it still means that the result applies to the retained perturbations.

The evaluation is also well-level after aggregation. copairs reports whether selected cells produce useful well profiles, but it does not directly prove that each selected cell is biologically typical in isolation. A selector can improve a well-level metric by changing crop quality, reducing noisy cells, changing cell-count distributions, or selecting true phenotype-bearing cells. This also explains why single-cell responses can appear as broad distributions rather than as one clean perturbation state. Fields contain mixtures of cell-cycle state, segmentation quality, local crowding, dosage response, technical variation, and cells that express the treatment phenotype to different degrees.

The isolated-crop design also assumes that the main profiling signal should come from the segmented target cell. This is appropriate for a controlled single-cell selector study, but it may discard useful local context in assays where background, extracellular structure, neighboring cells, or colony-level organization carry biological information. The results should therefore not be read as evidence that background is always uninformative, only that this thesis fixes a cell-centered representation to iso-

late the effect of cell selection. This leaves open the possibility that perturbation- or MoA-relevant evidence in truncated cells, partially segmented objects, or damaged cells is not fully represented by the pipeline. Some MorphoDiff-Typicality failure modes therefore require biological and visual inspection before being interpreted as failures of the typicality score itself.

The fixed-geometry design also means that MorphoDiff-Typicality is constrained by Cellpose-SAM. The spatial typicality map may highlight regions that are not available as clean proposals because they were missed, merged, truncated, or filtered out by the segmentation stage. The benchmark therefore evaluates typicality as a ranking rule over existing proposals, not as a complete heatmap-to-cell discovery method.

The percentage-based retention rule is another limitation. It keeps the same fraction of proposals per image, but images differ in how many proposals are available after filtering. The same nominal retention percentage can therefore correspond to different absolute cell counts across fields of view. This design is useful for a controlled selector sweep, but it does not prove that the same operating points would be optimal under a fixed-cell-count rule, a well-level budget, or a continuous weighting scheme.

The per-MoA analysis also shows a practical failure mode for MorphoDiff-Typicality, since at aggressive budgets the selector can remove cells needed by some mechanism classes even while improving perturbation FR overall. This means that a visually plausible typicality map is not enough for deployment. The selector has to be monitored by class, by crop-quality diagnostics, and by the downstream profiling metric that matters for the screen.

5.5 Future work

A deployable selector would need more than the present benchmark. Segmentation quality would need to be monitored explicitly, including border contact, merged cells, small masks, and crop foreground fraction. Artifact hotspots, missed proposals, per-MoA degradation, crop-geometry shifts, local crowding, and validation-plate performance should be checked together before trusting a selector in a new screen. In practice, this choice should be made on validation plates or historical screens, not by selecting the retention budget that gives the best MoA result on the final benchmark. A screen focused on reproducible perturbation activity may choose a different retention budget than a screen focused on MoA clustering.

A second direction is to improve the null reference itself. The population-null used here avoids synthetic conditioning vectors, but it is not equivalent to a learned unconditional or DMSO condition. A MorphoDiff model trained with condition dropout, or with an explicit DMSO or control conditioning branch, would make it possible to interpret typicality more directly as target-versus-generic or target-versus-control specificity. Another direction is to use the typicality map before proposal selection, for example through heatmap-guided proposal generation or hybrid segmentation,

rather than restricting the selector to regions that were already detected by Cellpose-SAM.

The method should also be tested beyond BBBC021, so a possible next step would be a subset of JUMP Cell Painting, where more perturbations, batches, and cell types are available. The feature extractor should also be varied because InceptionV3 is useful as a traceable baseline, but microscopy-trained self-supervised models such as DINO-style encoders may respond differently to masked crops and foreground fraction. Deployment would also need per-class monitoring. A selector that improves aggregate mAP or FR can still harm specific MoA classes, so per-MoA failure analysis should be part of the pipeline rather than an afterthought.

6

Conclusion

For RQ1, the thesis shows that Diff-Mining can be adapted to BBBC021 by comparing target-conditioned MorphoDiff denoising losses with a population-null reference loss, averaging the resulting spatial score over Cellpose-SAM target-cell masks, and using the resulting value to rank shared cell proposals.

For RQ2, the selector benchmark shows that this ranking preserves perturbation-level detectability at reduced retention, but does not dominate simpler crop-quality selection for MoA ranking. Foreground fraction remains the strongest MoA mAP selector in this setting, while MorphoDiff-Typicality contributes most clearly through perturbation FR.

The final thesis result is therefore a controlled comparison rather than a single-selector recommendation. Foreground fraction shows that simple crop quality can strongly affect BBBC021 profiling, especially MoA mAP. MorphoDiff-Typicality shows that perturbation-conditioned diffusion losses preserve a different signal, most clearly perturbation-level detectability at reduced cell budgets. Together, these results support the central framing of the thesis: the value of cell selection depends on what kind of biological signal the profile is meant to retain, and diffusion-based typicality should be evaluated against strong simple selectors before it is treated as a profiling improvement.

Bibliography

- [1] M.-A. Bray et al., “Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes,” *Nature Protocols*, vol. 11, no. 9, pp. 1757–1774, Sep. 2016. DOI: 10.1038/nprot.2016.105. [Online]. Available: <https://www.nature.com/articles/nprot.2016.105>.
- [2] J. C. Caicedo et al., “Data-analysis strategies for image-based cell profiling,” *Nature Methods*, vol. 14, no. 9, pp. 849–863, Sep. 2017. DOI: 10.1038/nmeth.4397. [Online]. Available: <https://www.nature.com/articles/nmeth.4397>.
- [3] D. R. Stirling, M. J. Swain-Bowden, A. M. Lucas, A. E. Carpenter, B. A. Cimini, and A. Goodman, “CellProfiler 4: improvements in speed, utility and usability,” *BMC Bioinformatics*, vol. 22, no. 1, p. 433, Sep. 2021. DOI: 10.1186/s12859-021-04344-9. [Online]. Available: <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-021-04344-9>.
- [4] Z. Navidi et al., “MorphoDiff: Cellular Morphology Painting with Diffusion Models,” in *International Conference on Learning Representations*, ICLR 2025 spotlight. bioRxiv preprint DOI: 10.1101/2024.12.19.629451, 2025. [Online]. Available: <https://openreview.net/forum?id=PstM8YfhvI>.
- [5] I. Siglidis, A. Holynski, A. A. Efros, M. Aubry, and S. Ginosar, “Diffusion Models as Data Mining Tools,” in *Computer Vision – ECCV 2024*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds., Cham: Springer Nature Switzerland, 2025, pp. 393–409, ISBN: 978-3-031-73030-6. DOI: 10.1007/978-3-031-73030-6_22. arXiv: 2408.02752 [cs.CV]. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-031-73030-6_22.
- [6] P. D. Caie et al., “High-content phenotypic profiling of drug response signatures across distinct cancer cells,” *Molecular Cancer Therapeutics*, vol. 9, no. 6, pp. 1913–1926, Jun. 2010. DOI: 10.1158/1535-7163.MCT-09-1148. [Online]. Available: <https://aacrjournals.org/mct/article/9/6/1913/93798/High-Content-Phenotypic-Profiling-of-Drug-Response>.
- [7] D. M. Ando, C. Y. McLean, and M. Berndl, “Improving Phenotypic Measurements in High-Content Imaging Screens,” *bioRxiv*, p. 161422, 2017. DOI: 10.1101/161422. eprint: <https://www.biorxiv.org/content/10.1101/161422v1.full.pdf>. [Online]. Available: <https://www.biorxiv.org/content/10.1101/161422v1>.
- [8] V. Ljosa, K. L. Sokolnicki, and A. E. Carpenter, “Annotated high-throughput microscopy image sets for validation,” *Nature Methods*, vol. 9, no. 7, p. 637, Jul. 2012, Correction: *Nature Methods* 10, 445 (2013). DOI: 10.1038/nmeth.2083. [Online]. Available: <https://www.nature.com/articles/nmeth.2083>.

- [9] A. A. Kalinin et al., “A versatile information retrieval framework for evaluating profile strength and similarity,” *Nature Communications*, vol. 16, no. 1, p. 5181, Jun. 2025. DOI: 10.1038/s41467-025-60306-2. [Online]. Available: <https://www.nature.com/articles/s41467-025-60306-2>.
- [10] C. Stringer, T. Wang, M. Michaelos, and M. Pachitariu, “Cellpose: a generalist algorithm for cellular segmentation,” *Nature Methods*, vol. 18, no. 1, pp. 100–106, Jan. 2021. DOI: 10.1038/s41592-020-01018-x. [Online]. Available: <https://www.nature.com/articles/s41592-020-01018-x>.
- [11] M. Pachitariu, M. Rariden, and C. Stringer, “Cellpose-SAM: superhuman generalization for cellular segmentation,” *bioRxiv*, p. 2025.04.28.651001, 2025, Preprint. DOI: 10.1101/2025.04.28.651001. eprint: <https://www.biorxiv.org/content/10.1101/2025.04.28.651001v1.full.pdf>. [Online]. Available: <https://www.biorxiv.org/content/10.1101/2025.04.28.651001v1>.
- [12] A. Kirillov et al., “Segment Anything,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Oct. 2023, pp. 4015–4026. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2023/html/Kirillov_SegmentAnything_ICCV_2023_paper.html.
- [13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848. [Online]. Available: <https://ieeexplore.ieee.org/document/5206848>.
- [14] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the Inception Architecture for Computer Vision,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2016, pp. 2818–2826. DOI: 10.1109/CVPR.2016.308. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2016/html/Szegedy_Rethinking_the_Inception_CVPR_2016_paper.html.
- [15] T. Cropsal and R. Mercado, “Compressing Biology: Evaluating the Stable Diffusion VAE for Phenotypic Drug Discovery,” in *NeurIPS 2025 Workshop for Imageomics: Discovering Biological Knowledge from Images Using AI*, 2025. arXiv: 2510.19887 [q-bio.QM]. [Online]. Available: <https://openreview.net/forum?id=ZgVGpQzmWy>.
- [16] B. Sun, J. Feng, and K. Saenko, “Return of Frustratingly Easy Domain Adaptation,” in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, ser. AAAI’16, Phoenix, Arizona: AAAI Press, 2016, pp. 2058–2065. DOI: 10.1609/aaai.v30i1.10306. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/10306>.
- [17] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 6629–6640. [Online]. Available: <https://papers.nips.cc/paper/7240-gans-trained-by-a-two-time-scale-update-rule-converge-to-a-local-nash-equilibrium>.
- [18] M. Bikowski, D. J. Sutherland, M. Arbel, and A. Gretton, “Demystifying MMD GANs,” in *International Conference on Learning Representations*, 2018.

- arXiv: 1801.01401 [stat.ML]. [Online]. Available: <https://openreview.net/forum?id=r1lU0zWCW>.
- [19] J. Ho, A. N. Jain, and P. Abbeel, “Denoising Diffusion Probabilistic Models,” in *Advances in Neural Information Processing Systems*, vol. 33, Curran Associates, Inc., 2020, pp. 6840–6851. arXiv: 2006.11239 [cs.LG]. [Online]. Available: <https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html>.
 - [20] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-Resolution Image Synthesis with Latent Diffusion Models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2022, pp. 10 684–10 695. arXiv: 2112.10752 [cs.CV]. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.html.
 - [21] J. Ho and T. Salimans, *Classifier-Free Diffusion Guidance*, Short version appeared in the NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications, 2022. DOI: 10.48550/arXiv.2207.12598. arXiv: 2207.12598 [cs.LG]. [Online]. Available: <https://arxiv.org/abs/2207.12598>.
 - [22] “RDKit: Open-source cheminformatics.” [Online]. Available: <https://www.rdkit.org>.
 - [23] F. Stossi et al., “SPACe: an open-source, single-cell analysis of Cell Painting data,” *Nature Communications*, vol. 15, no. 1, p. 10 170, Nov. 2024. DOI: 10.1038/s41467-024-54264-4. [Online]. Available: <https://www.nature.com/articles/s41467-024-54264-4>.
 - [24] S. N. Chandrasekaran et al., “Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations,” *Nature Methods*, vol. 21, no. 6, pp. 1114–1121, Jun. 2024, Published online 9 April 2024. DOI: 10.1038/s41592-024-02241-6. [Online]. Available: <https://www.nature.com/articles/s41592-024-02241-6>.
 - [25] M. Oquab et al., “DINOv2: Learning Robust Visual Features without Supervision,” *Transactions on Machine Learning Research*, 2024. arXiv: 2304.07193 [cs.CV]. [Online]. Available: <https://openreview.net/forum?id=a68SUt6zFt>.
 - [26] M. Doron et al., “Unbiased single-cell morphology with self-supervised vision transformers,” *bioRxiv*, p. 2023.06.16.545359, Jun. 2023, Preprint. DOI: 10.1101/2023.06.16.545359. [Online]. Available: <https://www.biorxiv.org/content/10.1101/2023.06.16.545359v1>.

A

Supplementary results and archived experiments

A.1 MorphoDiff deployment and image-quality reproduction

Before MorphoDiff was used for selector experiments, we carried out a short BBBC021 deployment check against the image-quality setup reported in the MorphoDiff paper. The aim was to verify that the implementation preserved the expected ordering between the conditional MorphoDiff model and the Stable Diffusion baseline before typicality scoring was used in the main benchmark. Unlike the original paper, which reports BBBC021 image-quality results from 500 generated images per perturbation, this check used only 50 generated images per perturbation. The purpose was therefore to confirm that MorphoDiff had been integrated correctly rather than to present a full reproduction attempt.

Table A.1 compares the BBBC021 values reported in the MorphoDiff paper with the corresponding values from this deployment check. To match the original presentation, the FID is shown on the same $\times 10^{-2}$ display scale as in the paper figure, together with the KID.

Table A.1: BBBC021 image-quality comparison between the original MorphoDiff paper and the deployment check used in this thesis. Lower is better for both metrics.

Experiment	Method	FID (500)	KID (500)	FID (50)	KID (50)
All compounds	MorphoDiff	1.99	0.21	2.37	0.23
All compounds	Stable Diffusion	3.84	0.47	3.44	0.35
14 compounds	MorphoDiff	2.26	0.30	2.30	0.25
14 compounds	Stable Diffusion	3.22	0.42	3.66	0.44

A.2 Supplementary MoA-level selector diagnostics

Figure A.1 shows the random-selection reference that is omitted from the main text. Because random selection does not impose a structured crop-quality or typicality

ranking, it acts as a low-information baseline for how much each MoA degrades under plain subsampling. Some classes remain comparatively robust under random retention, such as Aurora kinase inhibitors and Eg5 inhibitors, while DNA replication fails through most of the low-budget regime and protein degradation is never retrieved at any budget.

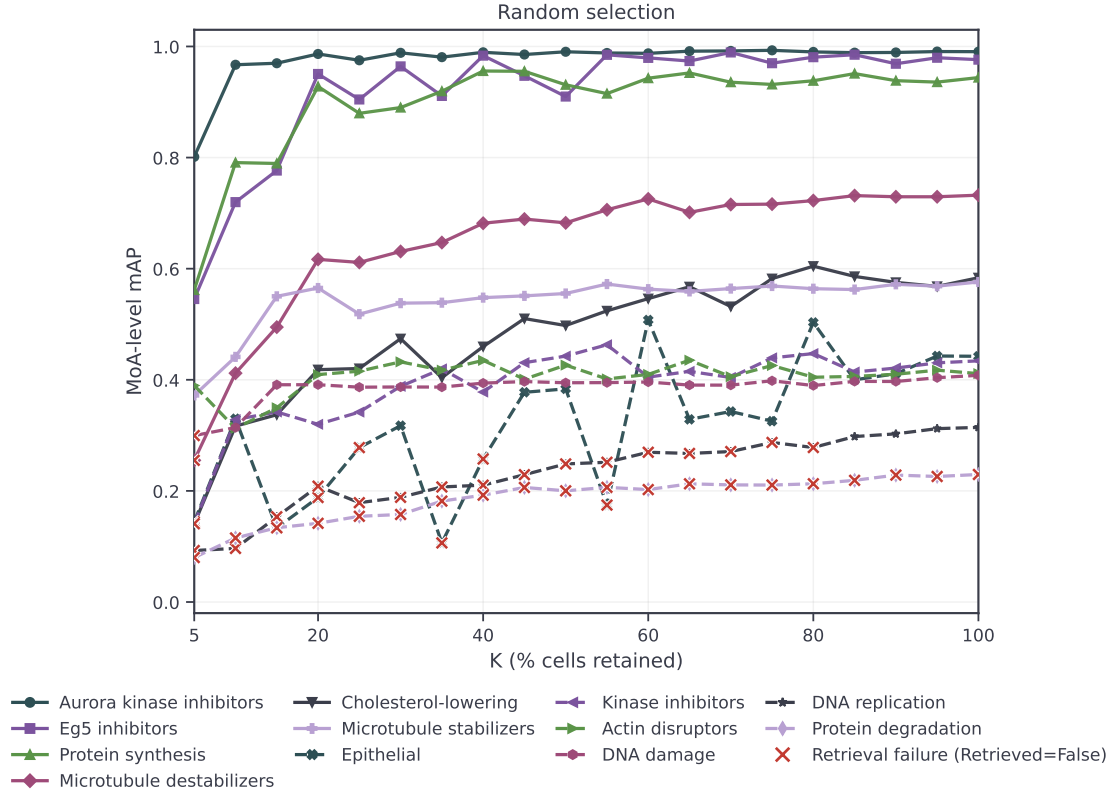


Figure A.1: Random-selection reference for the MoA-level selector audit. Each curve is one MoA, and red crosses mark (selector, K) points where the class is not retrieved under the permutation-null threshold.

Table A.2 summarizes the fixed-budget support values used in Section 4.6. The 25% columns capture the aggressive-filtering regime where the class split is strongest, while the 50% columns show the moderate-budget regime where MorphoDiff-Typicality partly recovers several classes. The crop-property columns use the definitions from Section 3.9.2 and report per-MoA means after selection, expressed either as percentage-point shifts or ratios relative to the full-retention baseline for the same MoA.

Coverage is thin for some MoAs, for example, Eg5 inhibitors and epithelial compounds contribute one compound each, and several others only two or three. Foreground fraction shifts crop geometry in the same direction for every MoA at 25% retention, but the MoA-level deltas still range from strong gains to strong losses. A larger average cell is therefore not enough to explain the split. The same selector also lowers the mean local-neighbor count for every MoA, with ratios from 0.81 to 0.98, so lower crowding is likewise not enough to explain which mechanisms improve. MorphoDiff-Typicality stays closer to the full-retention crop geometry

for most classes, yet its class-level effects are still mixed, so occupancy, area, and local-neighborhood diagnostics alone do not explain the selector split.

Table A.2: Per-MoA support summary for the selector audit. The Comp. column counts unique matched-subset compounds per MoA, and the Pert. column counts unique matched-subset compound-dose conditions. FG and Typ report the MoA-level mAP shift relative to the full-retention baseline at 25% and 50% retention. $\Delta\text{occ.}$ at 25% reports the change in mean selected-cell foreground occupancy in percentage points relative to the full-retention baseline. Area and Neigh. at 25% report ratios relative to the corresponding full-retention mean for selected-cell area and local neighbor count.

MoA	Comp.	Pert.	FG @25	FG @50	Typ @25	Typ @50	$\Delta\text{occ.}$ FG@25	$\Delta\text{occ.}$ Typ@25	Area FG@25	Area Typ@25	Neigh. FG@25	Neigh. Typ@25
Actin disruptors	3	5	+0.057	-0.001	-0.068	-0.027	+7.3	+1.7	1.46	1.08	0.81	0.85
Aurora kinase inhibitors	2	6	-0.001	-0.003	+0.005	+0.001	+8.2	+3.4	1.28	1.23	0.90	0.94
Cholesterol-lowering	2	6	-0.396	-0.309	-0.004	+0.059	+6.1	+0.5	1.27	1.04	0.98	1.06
DNA damage	4	9	+0.085	+0.033	+0.067	+0.046	+7.8	+2.2	1.30	1.12	0.90	0.95
DNA replication	4	8	-0.051	-0.023	-0.040	-0.005	+8.2	+2.2	1.33	1.17	0.89	0.99
Eg5 inhibitors	1	5	-0.172	-0.046	-0.011	+0.011	+6.7	-1.2	1.69	0.55	0.83	1.04
Epithelial	1	2	-0.178	-0.008	-0.091	-0.018	+6.6	+0.0	1.32	1.00	0.89	1.05
Kinase inhibitors	3	5	+0.131	+0.155	+0.092	+0.080	+6.1	+0.3	1.36	1.04	0.91	0.98
Microtubule destabilizers	4	14	+0.013	-0.014	-0.145	-0.059	+7.0	+1.6	1.38	1.12	0.91	0.95
Microtubule stabilizers	3	9	+0.098	+0.012	-0.145	-0.081	+8.0	+1.0	1.59	1.13	0.97	0.93
Protein degradation	4	7	-0.080	-0.058	-0.020	+0.009	+7.7	+0.2	1.24	1.04	0.89	0.90
Protein synthesis	3	8	-0.135	-0.048	-0.026	-0.019	+7.4	+0.3	1.42	1.03	0.88	0.97

A.3 Archived exploratory experiment families

A.3.1 Segmentation alternatives

Several extraction strategies were tested before the final crop pipeline was fixed. Figure A.2 shows one example for the alternatives that were discussed most often during development. The processed field in panel (a) is also the full-image baseline, since that variant passed the full BBBC021 image forward as one crop. Panel (b) shows the Otsu foreground mask. Panel (c) shows the full-field watershed instance mask, which used the green channel for a cell-body mask, the DAPI channel for nuclei seeds, and watershed separation before dynamic per-cell crops were extracted. Panel (d) shows four random crops. The Otsu and random variants both used 128×128 crops, which is useful to keep in mind because the random baseline depended on both the sampling seed and the chosen crop size.

A.3.2 Score-map-guided crop alternatives

Before the final Cellpose-SAM selector study was fixed, we also tested crop-generation strategies that relied directly on the MorphoDiff score map instead of on explicit cell proposals. Figure A.3 keeps one archived example of that branch together with the two selector families that were eventually kept in the thesis. Panel (b) shows the foreground-fraction ranking on Cellpose-SAM proposals, while panel (c) shows the final population-null typicality ranking on the same proposal set. Panel (d) shows the fixed-size patch setup, where 64×64 windows from a sliding grid were scored on

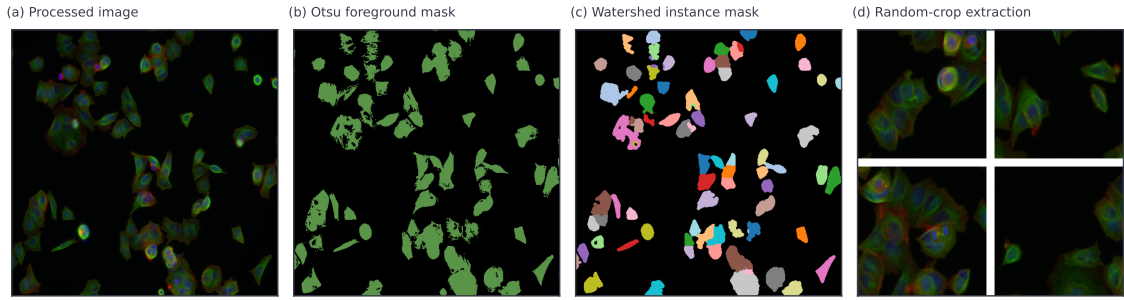


Figure A.2: Archived segmentation alternatives on BBBC021 for `anisomycin_0.3_Week4_27542_B06_497`. Panel (a) shows the processed image, which also corresponds to the full-image baseline because the full field was used as one crop. Panel (b) shows the binary Otsu foreground mask obtained from the green channel. Panel (c) shows the corresponding full-field watershed instance mask, where Otsu foreground masks and DAPI-seeded watershed were used to separate cells before crop extraction. Panel (d) shows four of the five archived 128×128 random crops.

the DMSO typicality map and ranked post hoc. Panel (e) shows the heatmap-peak setup, where local maxima on the same score map were turned into fixed 64×64 boxes with a minimum peak distance of 32 pixels. Panel (f) shows the thresholded heatmap setup, where the 90th percentile of the score map defined connected high-score regions whose bounding boxes were cropped and resized. This comparison makes it easier to see how much the selected regions depended on whether the detector started from explicit cell masks or only from score-map geometry.

A.3.3 Null-condition comparison

The null condition used to interpret MorphoDiff heatmaps was also explored before the final population-null benchmark was fixed. Figure A.4 compares the raw heatmaps for the three archived null choices that were kept in the result tree: the Stable Diffusion null used in the MorphoDiff setup, the synthetic all-ones null, and the population-null `mc16` reference that was used in the final thesis pipeline. All three heatmaps are shown on one shared diverging scale so the raw loss-difference magnitudes can be compared directly, and each heatmap also shows the top five Cellpose-SAM proposals that would be selected when the raw cell-mask score is computed from that null condition.

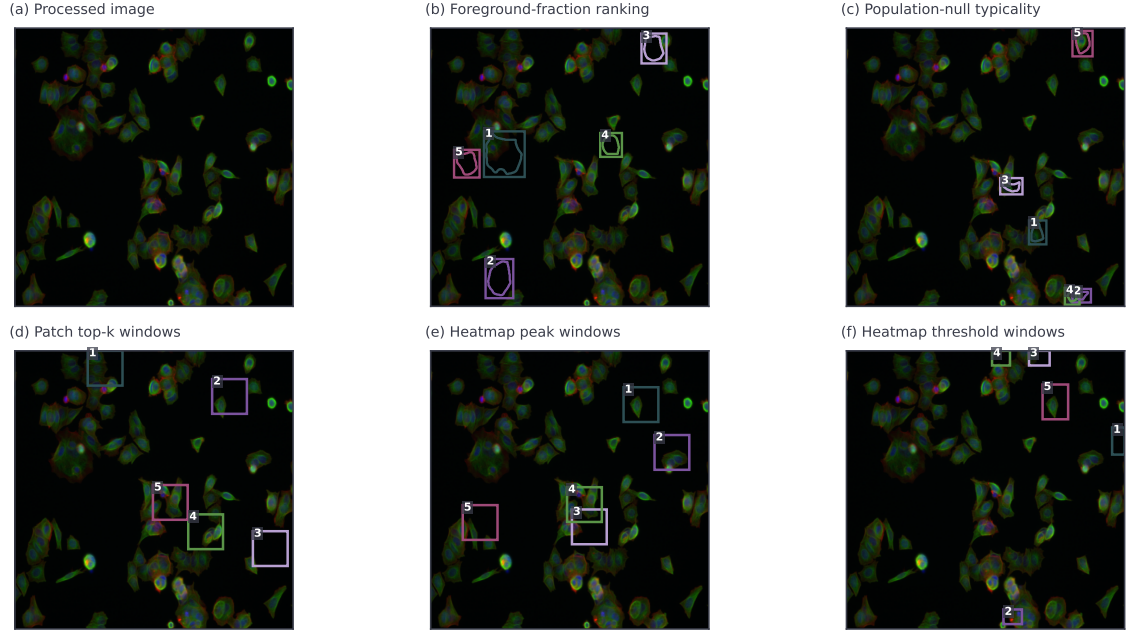


Figure A.3: Archived score-map-guided crop alternatives on BBBC021 for `anisomycin_0.3_Week4_27542_B06_497`. Panel (a) shows the processed image. Panel (b) overlays the top five Cellpose-SAM proposals ranked by foreground fraction. Panel (c) overlays the top five Cellpose-SAM proposals ranked by the final population-null typicality score. Panel (d) overlays the top five fixed-size patch windows, where 64×64 windows were ranked on the DMSO typicality map. Panel (e) overlays the top five windows from the heatmap-peak variant, where local maxima of the same score map defined fixed 64×64 boxes. Panel (f) overlays the top five windows from the thresholded heatmap variant, where score-map regions above the 90th percentile produced connected bounding boxes that were later resized.

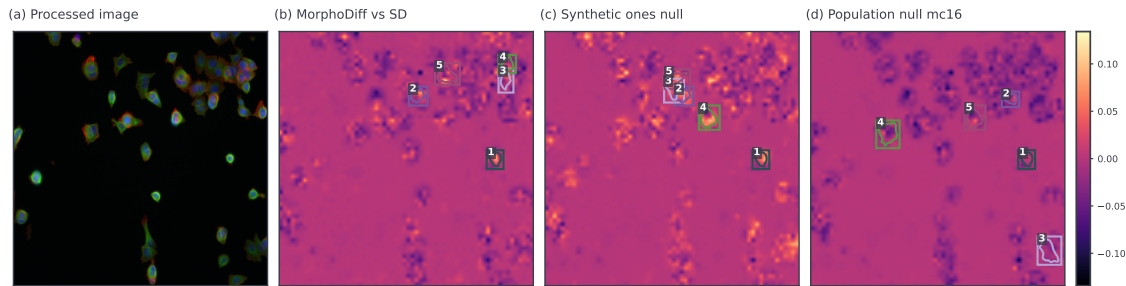


Figure A.4: Archived null-condition comparison on BBBC021 for `anisomycin_0.3_Week4_27542_B06_499`. Panel (a) shows the processed image. Panels (b)–(d) show the raw MorphoDiff heatmaps for the Stable Diffusion null, the synthetic all-ones null, and the population-null `mc16` reference, respectively, using a shared diverging color scale. In each heatmap panel, the top five Cellpose-SAM proposals are overlaid with their crop boxes and cell-mask polygons, ranked by the corresponding raw cell-mask score.