This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Inverse Protocol Prediction from Spheroid Microscopy Imaging via Morphology-Aware Structured Learning

Joohi Chauhan[1[0000−0001−7331−851X]⋆,⋆⋆], Prateek Mittal[1[0000−0003−4727−1708]⋆], and Ayush Srivastava[1,2]

[1] Vision Exploration and Data Analytics (VEDAs) Lab
Motilal Nehru National Institute of Technology Allahabad, India
[2] Vellore Institute of Technology, Chennai, India
joohi@mnnit.ac.in, prateekmittal154@gmail.com

Abstract. We introduce *Inverse Protocol Prediction (IPP)* as the task of inferring experimental culture conditions directly from a single bright-field spheroid image. We formulate IPP as a structured multi-label prediction problem and propose a protocol-aware learning framework that integrates morphology extraction, multimodal representation learning, and dependency-aware inference. Morphometric descriptors derived from automated spheroid segmentation are fused with deep visual embeddings via a morphometry vision fusion module. To capture biological and procedural dependencies, we incorporate a Hierarchical Multi-Task Transformer (HMTT) that is conditioned to predict across protocol attributes. The framework is trained with domain-adversarial supervision and morphology-preserving augmentation and tested for cross-dataset generalisation. Hybrid convolution attention encoders achieve the best performance, reaching 95.7% multi-attribute accuracy. Results demonstrate that structured, dependency-aware modeling enables reliable reconstruction of experimental protocols from imaging alone, supporting reproducibility auditing and protocol validation in 3D cell culture systems.

Keywords: Spheroids · Inverse Protocol Prediction · cell culture · multi-label prediction

1 Introduction and Background

Three-dimensional (3D) spheroids and organoids are widely used in cancer biology, drug discovery, and tissue engineering [1,2]. Unlike 2D cultures, spheroids reproduce gradients, cell–cell interactions, and necrotic cores, and can be imaged non-destructively at scale with bright-field/phase-contrast microscopy [3]. Microscopy is typically used to measure outcomes (e.g., size/morphology) but

⋆ These authors contributed equally and share first authorship.
⋆⋆ Correspondence at joohi@mnnit.ac.in

hardly to verify whether the observed morphology is consistent with the recorded experimental protocol. As a result, protocol or metadata inconsistencies (cell line, medium, seeding density, formation method, imaging setup) can remain undetected, undermining reproducibility, especially in large-scale or multi-site studies. We formalize *Inverse Protocol Prediction* (IPP): inferring experimental conditions directly from microscopy images to enable automated protocol validation, inconsistency detection, and analysis of which protocol factors leave recoverable morphological signatures. IPP is not intended to replace recorded experimental metadata. Rather, it provides an independent morphology-based estimate that can help identify protocol drift, metadata inconsistencies, incomplete annotations, and cross-site variability in large-scale imaging studies.

Recent work has explored multi-label and transformer-based metadata inference from biomedical images [4,5] and spatiotemporal models (e.g., ConvLSTM/attention) for longitudinal microscopy prediction [6,7]. IPP remains challenging due to (i) morphological ambiguity across conditions, (ii) acquisition variability and artifacts across microscopes/settings that confound biological signals [8,9], and (iii) limited datasets pairing 3D microscopy with rich protocol metadata.

We address these challenges using SLiMIA [10] dataset. We cast IPP as structured multi-label prediction, reduce acquisition bias via domain-adversarial training and morphology-preserving augmentation, and benchmark convolutional, transformer, and hybrid encoders. **Our key contributions** are: (i) Morphometry-vision fusion combining classical geometric descriptors with deep embeddings for protocol-aware representation learning, (ii) Dependency-aware inference via a hierarchical multi-task transformer for structured protocol reconstruction, (iii) Protocol-level interpretability with Grad-CAM to separate biological cues from acquisition-driven confounders and (iv) Systematic attribute recoverability analysis identifying protocol variables with stable morphological signatures.

2 Methodology and Experimental Results

Dataset Description: SLiMIA (Spheroid Light Microscopy Image Atlas) [10] is an open-access morphometric image dataset designed to support machine learning and computational modeling of three-dimensional (3D) cell culture systems. SLiMIA comprises approximately 8,000 light microscopy images of spheroids spanning a diverse range of experimental conditions, including nine microscope types, 47 distinct cell lines, eight culture media, four spheroid formation protocols, and multiple cell seeding densities, accompanied by rich metadata to facilitate reproducible analysis and benchmarking.

Implementation Details: To prevent image-level leakage across acquisition sessions, we split data by technical replicate (T1–T24): all images from a replicate were grouped and assigned exclusively to train (T1–T4), val (T5,T8), or test (T6,T7,T9–T24), balancing sample counts while maintaining replicate exclusivity (no image-level random splitting). Test performance is reported only on held-out replicates, giving a conservative estimate under cross-replicate acquisi-

Table 1: Model configurations and training settings for segmentation and IPP. Unless stated otherwise, CNN backbones use ImageNet pretraining.

Model	Backbone	Training Setup	Loss	Notes / Key design elements
Segmentation				
U-Net++ [11]	ResNet-50	Adam [12] + RLRP [13]	Focal Tversky [14] (α=0.7,β=0.3,γ=0.75)	CNN backbones use ImageNet init
DeepLabV3+ [15]	ResNet-50	Adam [12] + RLRP [13]	Focal Tversky [14]	–
DeepLabV3 [15]	ResNet-34	AdamW [16] (3e-4, wd=1e-4)	BCE + Dice	–
Attention U-Net [17]	ResNet-based	AdamW [16]	BCE + Dice	Attention gating in decoder
Swin-UNet [18]	Swin-Tiny	Adam [12] (1e-4)	BCEWithLogits	Windowed self-attention
TransUNet [19]	ViT encoder + CNN decoder	Adam [12] (1e-4)	BCEWithLogits	Transformer encoder + conv decoder
RefineNet [20]	ResNet-34	Adam [12] + RLRP [13]	Focal Tversky [14]	Multi-path refinement
SegNet [21]	VGG-style enc–dec	Adam [12] + RLRP [13]	Focal Tversky [14]	Pooling indices for upsampling
IPP				
ConvNeXt-Tiny [22]	ImageNet pretrained	Adam [12] (1e-4) + RLRP, light aug.	Cross-entropy	Convolutional locality priors; multi-head prediction
ViT-B/16 [23]	Pretrained timm weights	Adam [12] + scheduling	Cross-entropy	Global self-attention via CLS token
CoAtNet-0 [24]	Trained from scratch	Adam [12] + scheduling	Cross-entropy	Hybrid convolution + attention
Fusion Model [25]	ConvNeXt-Tiny + shape tokens	AdamW [16]	Weighted cross-entropy	9 morphometric descriptors + light Transformer (d=256, 3L, 4H)
HMTT [26]	ViT-B/16 encoder	AdamW [16] (1e-4, wd=1e-2)	Weighted CE + Focal [27] (γ=2, α=0.25)	Causal label conditioning (hierarchical prediction)

tion variability. We audited per-attribute class coverage across splits to quantify distribution shift. For 7/8 attributes (microscope, cell line, culture medium, formation method, seeding density, biological rep, magnification), all validation/test classes appear in training. For timepoint (109 classes), training contains 107; two test-only classes (000h, 032h) are unseen during training. Thus, evaluation primarily measures cross-replicate generalization rather than robustness to large unseen class sets. All codes and scripts will be released upon acceptance [3].

All segmentation models used RGB inputs with a single binary mask output. CNN backbones were ImageNet-pretrained. Training used Adam/AdamW with ReduceLROnPlateau (patience=5, factor=0.5) and standard flips with intensity/geometric augmentations. Losses were architecture-dependent to handle overlap/imbalance (e.g., Focal Tversky or BCE+Dice) (refer Table 1). CNN baselines (U-Net++, Attention U-Net, SegNet, RefineNet) span efficiency to boundary refinement; DeepLabV3/V3+ provide atrous-convolution baselines; Swin-UNet and TransUNet capture long-range context via self-attention. Results are in Table 2. We report mean Dice/IoU over test images, excluding edge cases with empty prediction and empty ground-truth masks (Dice=1, IoU=0); 17 such samples lacked masks and were also excluded to avoid skewed averages. All models

[3] The code and associated artifacts can be accessed at - `https://github.com/VEDAs-Lab/Inverse-Inference`

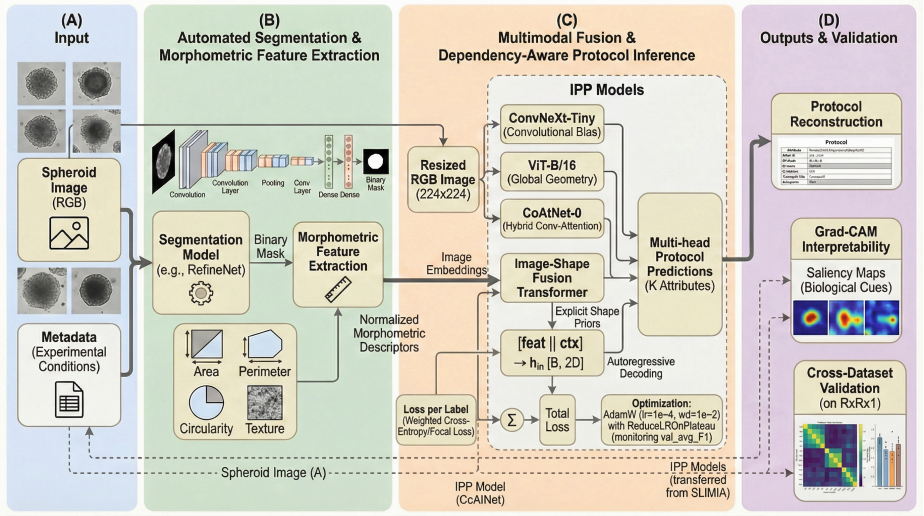

Fig. 1: **Inverse protocol prediction (IPP) workflow.** (A) Input spheroid image and metadata. (B) Automated segmentation and morphometric feature extraction. (C) Multimodal fusion and dependency-aware protocol inference using HMTT (D) Outputs including protocol reconstruction, Grad-CAM interpretability and cross data validation

achieved Dice > 0.94. RefineNet performed best (strong boundary/multi-scale refinement) but was slowest. Attention U-Net and U-Net++ were competitive with moderate inference time. Transformer models were slightly lower yet comparatively efficient. SegNet remained a fast, solid baseline. We conclude that **boundary-focused designs benefit spheroid edges**, with accuracy-speed trade-offs across architectures. For IPP, morphometric descriptors are computed only from predicted masks: RefineNet (best in Table 1) is trained on annotations, then frozen; its predictions generate all morphometrics. Ground-truth masks are never used in IPP training/evaluation, preventing annotation leakage and keeping the pipeline fully automatic.

All IPP models take RGB images resized to 224×224 and use multi-head prediction (one classifier per protocol label). ConvNeXt-Tiny [22] is the convolutional baseline (locality priors for fine-grained cues), ViT-B/16 [23] models global dependencies, CoAtNet-0 [24] hybridizes conv+attention [24]. A fusion variant augments ConvNeXt embeddings with normalized morphometric descriptors for explicit shape priors and interpretability [25]. The hierarchical prediction process factorizes the joint distribution over protocol attributes as $P(y_1, \ldots, y_K|x) = \prod_{k=1}^{K} P(y_k|x, y_{<k})$, where each prediction is conditioned on previously predicted attributes (Figure 1). During training, teacher forcing is used with ground-truth preceding labels, whereas inference proceeds autoregressively following the predefined protocol hierarchy. HMTT imposes a predefined causal label order, conditioning each attribute on previously predicted labels

Table 2: **Segmentation performance on the SLiMIA dataset**. Metrics are reported as mean ± standard deviation computed across test images. 95% confidence intervals were derived as $\mu \pm 1.96 \cdot \frac{s}{\sqrt{n}}$.

Model Name	Dice	IoU	Accuracy	Precision	Recall	F1	Inf. Time (s)
U-Net++ [11]	0.9582 ± 0.00019	0.9316 ± 0.00023	0.9929 ± 0.00002	0.9737 ± 0.00016	0.9447 ± 0.00018	0.9582 ± 0.00019	262
SegNet [21]	0.9521 ± 0.00020	0.9178 ± 0.00025	0.9913 ± 0.00001	0.9666 ± 0.00017	0.9402 ± 0.00019	0.9521 ± 0.00020	173
Swin-UNet [18]	0.9467 ± 0.00021	0.9103 ± 0.00026	0.9907 ± 0.00002	0.9428 ± 0.00016	0.9524 ± 0.00018	0.9467 ± 0.00021	208
TransUNet [19]	0.9579 ± 0.00019	0.9260 ± 0.00023	0.9929 ± 0.00002	0.9618 ± 0.00016	0.9554 ± 0.00018	0.9579 ± 0.00019	**164**
Attention U-Net [17]	0.9604 ± 0.00018	0.9361 ± 0.00022	0.9934 ± 0.00002	0.9615 ± 0.00017	**0.9609 ± 0.00018**	0.9604 ± 0.00018	166
DeepLabV3 [28]	0.9571 ± 0.00019	0.9302 ± 0.00023	0.9928 ± 0.00003	0.9596 ± 0.00016	0.9568 ± 0.00018	0.9571 ± 0.00019	184
DeepLabV3+ [28]	0.9551 ± 0.00019	0.9271 ± 0.00023	0.9925 ± 0.00002	0.9710 ± 0.00016	0.9426 ± 0.00020	0.9551 ± 0.00019	239
RefineNet [20]	**0.9665 ± 0.00016**	**0.9437 ± 0.00021**	**0.9938 ± 0.00001**	**0.9765 ± 0.00015**	0.9576 ± 0.00018	**0.9665 ± 0.00016**	286

Table 3: **IPP model performance** reported as mean ± standard deviation across three independent random seeds (42, 123, 999). Model-specific 95% confidence intervals are computed using the normal approximation ($1.96 \times \mathrm{SD}/\sqrt{S}$, with $S = 3$ seeds).

Model Name	Accuracy	Precision	Recall	F1 Score	Inference Time (s)
ImageShapeFusionTransformer [29]	0.9739 ± 0.00022	0.9319 ± 0.00048	0.9440 ± 0.00044	0.9350 ± 0.00046	220
ViT-B/16 [23]	0.9781 ± 0.00020	0.9450 ± 0.00042	0.9470 ± 0.00043	0.9440 ± 0.00042	122
ConvNeXt-Tiny [22]	0.9783 ± 0.00020	**0.9493 ± 0.00039**	**0.9542 ± 0.00037**	**0.9462 ± 0.00039**	**106**
Hierarchical Multi-Task Transformer (HMTT) [26]	0.9707 ± 0.00026	0.9126 ± 0.00066	0.9349 ± 0.00061	0.9162 ± 0.00064	275
CoAtNet-0 [24]	**0.9804 ± 0.00018**	0.9271 ± 0.00033	0.9192 ± 0.00037	0.9178 ± 0.00034	**106**

via teacher forcing in training and autoregressive decoding at test time [26]; the order goes from upstream stable factors (e.g., cell line/medium) to downstream conditional variables (e.g., density/timepoint), with possible error propagation under auto regression. IPP training uses morphology-preserving augmentation only: resize, small rotations (±10°), horizontal flip, then normalization; no elastic/crop/scale/color/intensity jitter to avoid altering morphometric cues. Resizing may reduce absolute size cues, but timepoint-linked structure (e.g., compactness/necrotic-core patterns) remains intact, improving generalization without introducing artifacts. These choices span convolutional, transformer, hybrid, feature-augmented, and dependency aware models to isolate effects of inductive bias, explicit priors, and label-structure modeling.

Given an input image x and morphometric descriptor vector m, the visual and shape representations are obtained as $z_{\text{img}} = f_\theta(x), \quad z_{\text{shape}} = Wm + b, \quad z_{\text{fusion}} = T([z_{\text{img}}; z_{\text{shape}}])$, where $f_\theta(\cdot)$ is the visual encoder, $T(\cdot)$ is the fusion Transformer, and $[\cdot;\cdot]$ denotes concatenation.

IPP infers experimental conditions from spheroid morphology, requiring local texture, global context, and label-dependency modeling. We compare five architectures (Table 3): ConvNeXt-Tiny (convolutional locality bias), ViT-B/16 (long-range/global geometry), CoAtNet (hybrid conv–attention), Image–Shape Fusion Transformer (ConvNeXt + explicit morphometric priors for robustness and interpretability), and HMTT (hierarchical, biologically motivated label conditioning for causally consistent predictions). Together, they isolate the roles of inductive bias, explicit priors, and hierarchy in inverse protocol inference. IPP is a multi-label, multi-class task with K protocol attributes per image. We report **Exact-Match (Subset) Accuracy:** $\frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left(\bigwedge_{k=1}^{K} \hat{y}_{ik} = y_{ik}\right)$, and **Mean**

Per-Attribute Accuracy: $\frac{1}{K}\sum_{k=1}^{K}\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}(\hat{y}_{ik} = y_{ik})$. Unless stated otherwise, "accuracy" in Table 3 denotes Mean Attribute Accuracy. Models show complementary strengths across protocol variables. CoAtNet attains the highest macro-averaged accuracy (98.0%), while ConvNeXt-Tiny achieves the best macro F1 (0.946) via stronger precision–recall balance. ViT-B/16 is comparable (consistent recall; competitive F1). Image–Shape Fusion remains competitive (macro F1=0.935), supporting explicit morphometric priors at higher compute cost; HMTT is lower in macro accuracy (97.1%) and F1 (0.916) but yields stable performance across interdependent labels. Despite modest gains, HMTT explicitly models protocol dependencies and biologically plausible label relationships. Per-label results (Table 4) show strong-signature attributes (cell line, medium, formation method) are reliably predicted, whereas weaker/less visually encoded factors (seeding density, timepoint, replicate) are harder; near-perfect microscope/magnification likely reflect dataset artifacts.

Biological-Only Inverse Protocol Prediction: To remove acquisition cues, we exclude microscope, magnification, and replicate labels and recompute IPP on five attributes (cell line, culture medium, seeding density, timepoint, formation method; Table 4). The macro-averaged Biological IPP F1 is 0.914: cell line and formation method are highly recoverable (F1 $\approx$ 0.99), culture medium and seeding density remain reliably inferable but with reduced separability, and timepoint is most challenging (F1 $\approx$ 0.75). Strong performance after excluding acquisition-driven variables supports biologically meaningful structure encoded in spheroid morphology.

Domain Adversarial Training: To mitigate acquisition bias, we apply domain-adversarial training to CoAtNet-0, defining domain as the technical replicate (T1–T24). The model comprises a CoAtNet-0 feature extractor, per-attribute biological heads (linear classifiers), and a domain discriminator MLP (GRL$\to$Linear $d\to256$ $\to$ReLU$\to$Dropout $p=0.3$ $\to$Linear $256\to N_{\text{domain}}$). The Gradient Reversal Layer (GRL) is identity in the forward pass and multiplies gradients by $-\lambda$ in backpropagation to promote domain-invariant features, with $\lambda(p) = \frac{2}{1+\exp(-10p)} - 1, p \in [0,1]$. Training minimizes $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{bio}} + \mathcal{L}_{\text{domain}}$, where both terms are equally weighted cross-entropy losses. We evaluated biological prediction and domain predictability on the held-out test set. CoAtNet + DANN gives Accuracy = 0.8973, Precision = 0.8267, Recall = 0.7979, F1 = 0.8010 (All are (Macro (across heads)). It gives Accuracy / F1 = 0.8973 for Micro (flattened across attributes).

Interpretability Analysis: To probe IPP decision-making, we applied Grad-CAM [30,31] to the best model, CoAtNet, to confirm reliance on biological cues and expose dataset bias. We analyzed 10 representative spheroids spanning cell lines (A549, HCT116, PANC1, SKOV3, U251MG, SW837, MCF10A and CT5.3hTERT), media, formation methods (ULA, Hanging Drop, Agarose, Microchip), seeding densities, timepoints, replicates, and magnifications. Grad-CAM maps (Figure 2) focus on interpretable morphology: cell line attends to global shape/boundary; formation method and density to compactness/internal organization; later timepoints to dense cores consistent with maturation. Replicate-

Table 4: **Per-label best performance across models.** Labels are grouped into biological protocol variables and acquisition/replicate variables. The Biological Macro F1 (0.914) is computed over cell line, culture medium, seeding density, timepoint, and formation method only, isolating biologically meaningful inference from acquisition-specific cues. Technical replicate identifiers are excluded from macro-level aggregation, as data partitioning was performed at the replicate level to prevent information leakage across acquisition sessions.

Label	**Best F1**	**Best Model**	**Average F1 (all models)**
Biological Protocol Variables			
Cell Line	0.9944	ImageShapeFusionTransformer	0.9891
Culture Medium	0.9642	ViT	0.9604
Seeding Density	0.9302	CoAtNet	0.8781
Timepoint	0.8331	ConvNeXt	0.7482
Formation Method	0.9949	ConvNeXt	0.9941
Biological Macro F1	–	–	**0.914**
Acquisition and Replicate Variables			
Biological Rep	0.9583	ConvNeXt	0.9444
Magnification	0.9997	CoAtNet	0.9983
Microscope	1.0000	ImageShapeFusionTransformer	0.9993

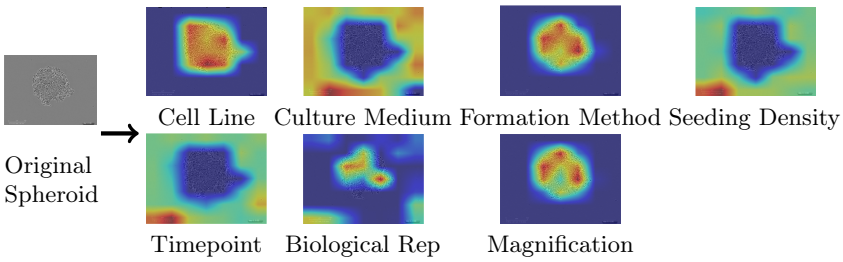

Fig. 2: Grad-CAM visualizations for inverse protocol prediction. Left: original spheroid image. Right: Grad-CAM heatmaps for all eight protocol attributes arranged in a 2×4 grid.

related predictions show diffuse/background attention, indicating sensitivity to technical artifacts. Attention also shifts with magnification, emphasizing global structure at low resolution and boundary detail at higher resolution. Overall, Grad-CAM supports biologically grounded inference for key variables while highlighting vulnerability to replicate-level confounders.

Cross-Dataset Validation on RxRx1 for Protocol Prediction (IPP): To test cross-domain generalization (Table 5), we evaluated IPP on RxRx1 [32] (2D fluorescence monolayers; strong batch effects), adapting IPP to predict metadata: *experiment* (51), *plate* (4), *site* (2), *cell_type* (4), *sirna* (1139).

Table 5: **Cross-dataset IPP results on RxRx1** with 95% confidence intervals.

Model Name	Accuracy	Precision	Recall	F1 Score
Image–Shape Fusion Transformer [29]	**0.7687 ± 0.00042**	**0.7726 ± 0.00048**	**0.7684 ± 0.00044**	**0.7680 ± 0.00046**
HMTT [26]	0.7328 ± 0.00055	0.7388 ± 0.00062	0.7325 ± 0.00057	0.7319 ± 0.00060
CoAtNet-0 [24]	0.6559 ± 0.00070	0.6707 ± 0.00076	0.6555 ± 0.00072	0.6533 ± 0.00074

Table 6: **Ablation study results** for multimodal metadata prediction. Values show mean accuracy, precision, recall, and macro-F1 score with estimated 95% confidence intervals computed using the normal approximation ($1.96 \times \mathrm{SD}/\sqrt{N}$, with $N \approx 8000$ validation samples).

Model Variant	Accuracy	Precision	Recall	Macro-F1
Full Model (Image + Shape + Transformer Fusion)	**0.8526 ± 0.00042**	**0.8672 ± 0.00046**	**0.8758 ± 0.00044**	**0.8681 ± 0.00045**
Image–Only	0.8516 ± 0.00045	0.8651 ± 0.00049	0.8734 ± 0.00047	0.8662 ± 0.00048
No–Transformer	0.8513 ± 0.00047	0.8638 ± 0.00050	0.8726 ± 0.00049	0.8650 ± 0.00049
No–Fusion	0.8478 ± 0.00053	0.8597 ± 0.00056	0.8689 ± 0.00055	0.8610 ± 0.00056
Shape–Only	0.5381 ± 0.00112	0.5515 ± 0.00118	0.5591 ± 0.00122	0.5530 ± 0.00120

Uniform $1/K$ baselines are 0.0196/0.25/0.50/0.25/0.000878; the calculated majority baselines are 0.0196/0.25/0.50/0.470/0.00392. Joint exact-match baselines are 5.38×10^{-7} (uniform) and 4.53×10^{-6} (majority), underscoring the difficulty of multi-attribute reconstruction. We used 125,511 Channel 1 images with a strict 70:15:15 split grouped by `sirna_id` (Train 87,857 / Val 18,826 / Test 18,827) to prevent leakage; models were transferred from SLiMIA with no fine-tuning to isolate representation robustness. We evaluated Image–Shape Fusion, HMTT, and CoAtNet-0; 95% CIs used $1.96\cdot\mathrm{SD}/\sqrt{N}$ on the test set (N=18,827). Image–Shape Fusion generalized best (explicit morphometric priors), HMTT remained competitive (structured dependencies; more sensitive to perturbation variability), and CoAtNet-0 degraded most (spheroid-tuned bias), indicating feature-augmented and dependency-aware designs are most robust under severe geometric/acquisition shift.

Ablation Study: To quantify each fusion component, we ablated five variants: full fusion, image-only (no shape), shape-only (no image), no-fusion (both branches without interactive fusion), and no-transformer (both modalities without transformer fusion). All variants used identical splits, augmentation, and early stopping. Table 6 reports the details of biological-only IPP performance (avg. accuracy/precision/recall/macro-F1). The full model performs best (macro-F1=0.8681) with only a small gain over image-only (0.8662; ΔF1=0.0019), indicating ConvNeXt encodes substantial shape/texture implicitly and explicit morphometrics add limited marginal benefit. Shape-only drops sharply (macro-F1=0.5530), showing morphology alone is insufficient without appearance cues. No-fusion underperforms (0.8610 vs. 0.8681), supporting the need for cross-modal interaction; removing the transformer yields a small consistent decline (0.8650), suggesting transformer fusion further refines cross-modal alignment. In conclusion, images carry most signal, morphometrics are complementary, and explicit (transformer) fusion gives the most stable gains.

3 Conclusion

We propose inverse protocol prediction (IPP) from single bright-field spheroid images, showing that morphology contains recoverable signatures of experimental conditions. On SLiMIA, we train an automatic segmentation model to derive morphometric descriptors and benchmark IPP across convolutional, transformer, hybrid, feature-augmented, and dependency-aware architectures. Hybrid encoders provide a strong accuracy–efficiency trade-off, while morphometry fusion and hierarchical conditioning improve structured protocol reconstruction and interpretability. Cross-dataset evaluation on RxRx1 demonstrates that explicit priors and label-structure modeling yield greater robustness under severe geometric and acquisition shift. Grad-CAM analysis confirms reliance on biologically meaningful cues for key variables while revealing sensitivity to acquisition-specific artifacts, consistent with per-label recoverability trends. Although morphometry fusion and hierarchical conditioning improve structured protocol reconstruction, gains over strong image-only baselines are modest, suggesting modern visual encoders already capture substantial morphological information. Timepoint prediction remains challenging due to gradual biological transitions and overlapping developmental states. Future work will expand protocol diversity, reduce technical confounding, and integrate mechanistic priors to further improve robustness and biological fidelity.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Aliya Fatehullah, Si Hui Tan, and Nick Barker. Organoids as an in vitro model of human development and disease. *Nature cell biology*, 18(3):246–254, 2016.
2. Maria Chatzinikolaidou. Cell spheroids: the new frontiers in in vitro models for cancer drug validation. *Drug discovery today*, 21(9):1553–1560, 2016.
3. Rasheena Edmondson, Jessica Jenkins Broglie, Audrey F Adcock, and Liju Yang. Three-dimensional cell culture systems and their applications in drug discovery and cell-based biosensors. *Assay and drug development technologies*, 12(4):207–218, 2014.
4. Jiequan Zhang, Qingyu Zhao, Ehsan Adeli, Adolf Pfefferbaum, Edith V Sullivan, Robert Paul, Victor Valcour, and Kilian M Pohl. Multi-label, multi-domain learning identifies compounding effects of hiv and cognitive impairment. *Medical image analysis*, 75:102246, 2022.
5. Sumit Madan, Manuel Lentzen, Johannes Brandt, Daniel Rueckert, Martin Hofmann-Apitius, and Holger Fröhlich. Transformer models in biomedicine. *BMC medical informatics and decision making*, 24(1):214, 2024.
6. Raj Dave, Kshipra Pandey, Ritu Patel, Nidhi Gour, and Dhiraj Bhatia. Leveraging 3d cell culture and ai technologies for next-generation drug discovery. *Cell Biomaterials*, 1(3), 2025.

7. Alfredo De Cillis, Valeria Garzarelli, Alessia Foscarini, Giuseppe Gigli, Antonio Turco, Elisabetta Primiceri, Maria Serena Chiriacò, and Francesco Ferrara. 3d-printed barriers with machine learning powered image analysis for enhanced wound healing assays. *Materials & Design*, page 114746, 2025.
8. Maximiliaan Huisman, Mathias Hammer, Alex Rigano, Ulrike Boehm, James J Chambers, Nathalie Gaudreault, Alison J North, Jaime A Pimentel, Damir Sudar, Peter Bajcsy, et al. A perspective on microscopy metadata: data provenance and quality control. *arXiv preprint arXiv:1910.11370*, 2019.
9. Paula Montero Llopis, Rebecca A Senft, Tim J Ross-Elliott, Ryan Stephansky, Daniel P Keeley, Preman Koshar, Guillermo Marqués, Ya-Sheng Gao, Benjamin R Carlson, Thomas Pengo, et al. Best practices and tools for reporting reproducible fluorescence microscopy methods. *Nature Methods*, 18(12):1463–1476, 2021.
10. Eva Blondeel, Arne Peirsman, Stephanie Vermeulen, Filippo Piccinini, Felix De Vuyst, Diogo Estêvão, Sayida Al-Jamei, Martina Bedeschi, Gastone Castellani, Tânia Cruz, et al. The spheroid light microscopy image atlas for morphometrical analysis of three-dimensional cell cultures. *Scientific Data*, 12(1):283, 2025.
11. Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *International workshop on deep learning in medical image analysis*, pages 3–11. Springer, 2018.
12. Kingma DP Ba J Adam et al. A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 1412(6), 2014.
13. Leslie N Smith. Cyclical learning rates for training neural networks. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pages 464–472. IEEE, 2017.
14. Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, and Ali Gholipour. Tversky loss function for image segmentation using 3d fully convolutional deep networks. In *International workshop on machine learning in medical imaging*, pages 379–387. Springer, 2017.
15. Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.
16. Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
17. Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.
18. Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pages 205–218. Springer, 2022.
19. Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.
20. Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934, 2017.

21. Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.
22. Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.
23. Abdullah A Asiri, Ahmad Shaf, Tariq Ali, Muhammad Ahmad Pasha, Muhammad Aamir, Muhammad Irfan, Saeed Alqahtani, Ahmad Joman Alghamdi, Ali H Alghamdi, Abdullah Fahad A Alshamrani, et al. Advancing brain tumor classification through fine-tuned vision transformers: A comparative study of pre-trained models. *Sensors*, 23(18):7913, 2023.
24. Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *Advances in neural information processing systems*, 34:3965–3977, 2021.
25. Jingsong Xia, Yue Yin, and Xiuhan Li. An efficient medical image classification method based on a lightweight improved convnext-tiny architecture. *arXiv preprint arXiv:2508.11532*, 2025.
26. Bardia Rafieian and Pere-Pau Vázquez. Improved multi-label hierarchical patent classification using llms. *World Patent Information*, 81:102356, 2025.
27. Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.
28. Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.
29. Fei Luo, Daoqi Wu, Luis Rojas Pino, and Weichao Ding. A novel multimodel medical image fusion framework with edge enhancement and cross-scale transformer. *Scientific Reports*, 15(1):11657, 2025.
30. Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.
31. Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018.
32. Maciej Sypetkowski, Morteza Rezanejad, Saber Saberian, Oren Kraus, John Urbanik, James Taylor, Ben Mabey, Mason Victors, Jason Yosinski, Alborz Rezazadeh Sereshkeh, et al. Rxrx1: A dataset for evaluating experimental batch correction methods. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4285–4294, 2023.