Discovering organoid motility variation directly from video

Krishna Srinivasan[1,2], Kameron Bielawski[1,2], Douglas Blackiston[3,4], Michael Levin[3,4] and Joshua Bongard[1,2]

[1]Department of Computer Science, University of Vermont, Burlington VT, USA
[2]Vermont Complex Systems Institute, Burlington VT, USA
[3]Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston MA, USA
[4]Department of Biology, Tufts University, Medford MA, USA
Krishna.Kannan-Srinivasan@uvm.edu

Abstract

Artificial Life investigators are interested in new forms of life. One route to achieve this is to reconfigure existing life forms, such as organoids. Motile organoids are of particular interest as they demonstrate that cellular coordination can arise within novel multicellular configurations — or survive reconfigurations — conferring, among other things, non-random motion upon the organoid as a whole. However, characterizing the diversity of behavior within such constructs is challenging. Current approaches require several manual steps, including feature engineering. Whether reliable motility differences across organoids can be recovered without presupposing features remains an open question. Here we show that an automated method can indeed discover features that most reliably distinguish organoids in their motility, directly from videos, as evidenced by agreement with the classical pipeline's discovered features. The video prediction model driving the method generalizes to unseen organoid videos, establishing that organoid motility dynamics share common structure. Our approach thus enables scalable exploration of behavioral diversity across large video collections of motile organoids or other synthetic constructs, with little manual effort and without presupposing what to look for.

Data and code available at https://github.com/kkannans/DiscoveringMotilityVariationDirectlyFromVideo

Introduction

One central aim of Artificial Life is to understand the principles underlying complex behavior in living systems by studying lifelike phenomena across biological and non-biological substrates (Langton, 1997). One route to this understanding is to build new forms of life by reconfiguring existing biological substrates and studying the behavioral patterns that emerge (Blackiston et al., 2021; Kriegman et al., 2020). Motility is of particular interest both as a timescale of living organization that theories of life have historically overlooked (Froese et al., 2014), and as a behavioral readout of a system's internal state (Berg and Brown, 1972). Xenobots, assembled from *Xenopus* embryonic cells, demonstrate that novel multicellular configurations produce coordinated locomotion, with individuals exhibiting trajectories ranging from linear translation to tight circular paths, correlated with differences in cilia distribution and body size (Blackiston et al., 2021; Kriegman et al., 2020). This extends to human-derived systems: Anthrobots self-constructed from adult lung cells exhibit the same correlated relationship between form and motility (Gumuskaya et al., 2024). Across these systems, discovering what makes organoids different in their motility requires several manual steps.

Quantifying motility differences across a population of organoids begins with choosing a behavior of interest (here, motility) and then extracting relevant features from centroid trajectories obtained through segmentation and frame-by-frame tracking. Each step carries its own challenges (Maška et al., 2023). Segmentation parameters must be tuned to the specific imaging platform, magnification, and organoid morphology; even recent deep learning approaches require supervised training on manually annotated images and were developed for specific tissue types and optical configurations, precluding their use across different experimental settings without retraining (Matthews et al., 2022). Centroid tracking is sensitive to segmentation quality and accumulates errors over long sequences; manual tracking requires sustained human judgment and scales poorly, limiting reproducibility (Holme et al., 2023).

Beyond the engineering problem of feature extraction lies an issue of bias: the choice of which motility feature to compute encodes implicit assumptions about which aspects of motility matter. In principle, one could devise many motility features (Kimmel et al., 2018), yet in practice researchers select a small number that best describe the behavior of interest — trajectory shape, persistence, or some property of the dynamics — encoding a prior belief about biological relevance before any data has been examined. When experimental datasets are large and the relevant axes of variation are genuinely unknown, predefined features may miss

the dominant structure entirely. In *C. elegans*, Stephens et al. (2008) stands as a notable exception: allowing structure to emerge directly from body shape data revealed a low-dimensional behavioral space containing subtle behaviors invisible to predefined movement categories. Even this approach, however, operated on pre-extracted eigenworm shape coordinates rather than raw video. Similar approaches that discover structure without predefined features remain rare in organoid research and, more broadly, in the study of living systems.

Unexpected patterns and behaviors are common in both artificial life systems and living organisms (Gershenson, 2023), yet current methods require researchers to specify in advance which features to extract. A method that recovers the primary axes of motility variation directly from video, without committing to any features in advance, would sidestep both the engineering burden of the classical pipeline and the epistemic bias of predefined feature selection.

Video prediction models (Oprea et al., 2022) are well suited for learning dynamics from video and have been used to discover low-dimensional representations of dynamical systems without prior knowledge of the system's physics (Chen et al., 2022). Motivated by their ability to learn structure without predefined features, we employ a video prediction model to recover motility variation directly from organoid videos, bypassing the classical pipeline entirely (Fig. 1). A model trained across videos of many organoids must implicitly learn shared patterns in how they move; when it encounters a video whose motion deviates from these learned patterns, prediction error rises. We define **prediction surprise** (Eq. 4) as this per-organoid motion prediction error of a population-trained model, a candidate measure of variation in motility.

For prediction surprise to serve as a valid measure, two conditions must hold: the model must generalize to unseen organoid videos, and prediction surprise must agree with the ranking produced by the most reliable discriminating feature identified through the classical pipeline. We show that both conditions hold: (1) we identify mean speed as the most reliable discriminator of organoid motility among eight classical features, establishing an independent reference ranking, and (2) per-organoid prediction surprise correlates strongly with this reference.

Related Work

Describing motility in cells and organoids

Motility serves as a key phenotypic readout in cell biology and, more recently, in synthetic multicellular systems. The standard approach quantifies motion through trajectory-based features computed from trajectories extracted from centroid tracking. Commonly used features include velocity, mean-squared displacement, persistence time, turning angle, and outreach ratio, as many as 79 features have been computed from a single trajectory (Wu et al., 2014; Svensson et al., 2018; Kimmel et al., 2018).

Motile organoids inherit this framework. In xenobots, Blackiston et al. (2021) used cross-entropy clustering on tracked trajectories to identify discrete movement states (linear, circling, intermediate, and inactive), measured average velocity across the lifespan, and applied particle image velocimetry to probe the causal role of cilia-generated flow in locomotion. The motility of Anthrobots, self-assembled from human lung cells, was quantified by computing straightness and curvature indices from tracked centroids (Gumuskaya et al., 2024). OrganoID automates organoid tracking and analysis using deep learning, but still requires segmentation as a prerequisite (Matthews et al., 2022). Across these systems, quantifying motility relies on the classical pipeline of segmentation, tracking, and predefined feature extraction.

Video prediction models and their use in biological datasets

Video prediction, the task of forecasting future frames from past observations, has emerged as a self-supervised approach for learning spatiotemporal structure (Oprea et al., 2022). Recurrent architectures such as ConvLSTM (SHI et al., 2015) and PredRNN (Wang et al., 2017) capture temporal dependencies through memory states that propagate information across frames. SimVP replaces recurrence with a fully convolutional architecture: a spatial encoder extracts per-frame features, a temporal module aggregates information across frames via stacked inception blocks, and a spatial decoder reconstructs future frames (Gao et al., 2022). SimVP-TAU substitutes the inception blocks with temporal attention units that decompose attention into intra-frame static and inter-frame dynamic components (Tan et al., 2023a). OpenSTL provides a way to compare these architectures and benchmark them on standard video datasets (Tan et al., 2023b). Beyond forecasting, video prediction can recover latent structure of dynamical systems. Chen et al. (2022) showed that low-dimensional representations analogous to state variables can be extracted from video of physical systems such as pendulums and flames, without prior knowledge of the underlying physics. Applications to biological video remain limited. Lorimer et al. (2021) showed that deviation from a reference behavioral attractor can track variation in C. elegans motility, and PredRNN has been applied to forecast bacterial colony growth (Robertson et al., 2023); however, the Lorimer et al. method operates on pre-extracted eigenworm coordinates rather than raw video, requiring the classical feature extraction step as a prerequisite. Our method removes this prerequisite entirely by operating on raw pixel frames. To our knowledge, using prediction surprise from a video prediction model to quantify motility variation directly from raw video, without presupposing motility features, has not been demonstrated.

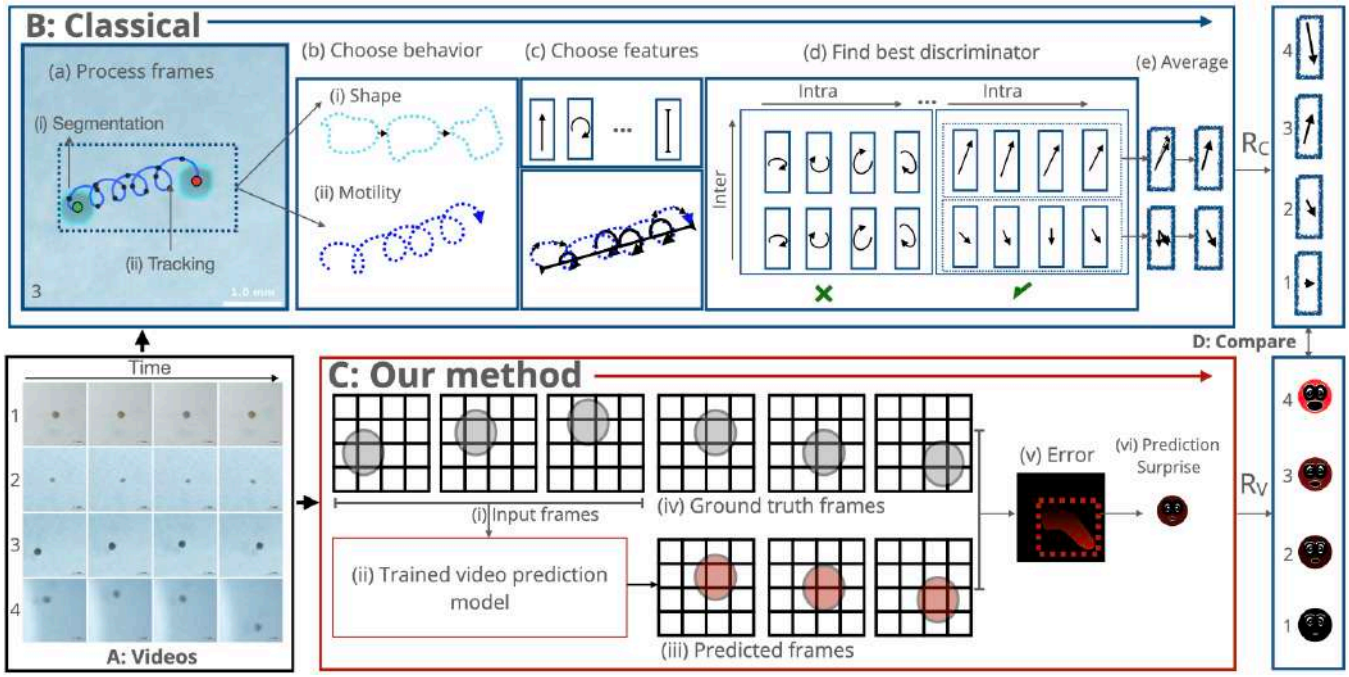

Figure 1: **Discovering organoid motility variation directly from video. A.** Raw video frames from four organoids (rows) across time, representative of different motility behaviors with varying speeds sorted from slowest (top) to fastest (bottom). **B.** The classical pipeline (illustrated for organoid 3) begins with **(a) processing frames**: (i) segmentation of the organoid from background and (ii) tracking the organoid centroid through time. The investigator then **(b) chooses a behavior** of interest ((i) shape or (ii) motility) and **(c) chooses features** for that behavior (e.g., speed, turning angle, displacement, shown as icons). **(d) The best discriminator** is identified by comparing intra- and inter-organoid variance across features: a feature with high inter- and low intra-organoid variance (Good) is selected over one with poor separation (Bad). The selected feature is then **(e) averaged** per organoid and used to rank organoids (R_C, 1–4). **C.** Our method bypasses this pipeline entirely: **(i) input frames** are fed to a **(ii) trained video prediction model**, which produces **(iii) predicted frames**; the **(v) error** between **(iv) ground-truth frames** and predicted frames is computed per organoid, and its mean constitutes the **(vi) prediction surprise** score, used directly to rank organoids (R_V, 1–4). **D. Compare.** Rank correlation between R_C (classical pipeline) and R_V (prediction surprise) shows that the two rankings agree, confirming that prediction surprise recovers the dominant axis of motility variation without predefined features, directly from video.

Methods

Dataset

Mucociliary organoids were constructed from *Xenopus laevis* embryonic tissue following standard animal cap protocols (Sive et al., 2000). Each organoid was imaged individually with a 64MP camera (9152×6944 pixels) capturing frames at 5-second intervals. From a larger collection of recordings, we selected 108 videos in which the organoid remained within a 706×706 pixel region throughout the observation period, ensuring the organoid's motion remained fully within frame without resolution loss. Each video in this set of 108 comprises 120 frames representing 10 minutes of organoid locomotion. Organoid diameters range from 0.35 to 0.6 mm. See the repository for a supplementary video of all organoids comprising the dataset, showing how imaging conditions vary across videos, with various artifacts making segmentation of the organoid from the background non-trivial.

Extracting motility features from video

To establish an independent reference for motility variation, we extracted centroid trajectories from each video using a manual pipeline with search over hyperparameters affecting segmentation and centroid detection. For each video with its own distinct imaging background, we searched over color channels and pipeline hyperparameters to maximize foreground-background contrast. These hyperparameters include Gaussian blur, contrast enhancement, binarization via adaptive or percentile-based thresholding, and morphological cleanup to isolate the largest connected component (Bradski, 2000). Parameters were refined manually with visual feedback on the resulting segmentation and tracking output. As can be validated visually from the supplementary video (in the repository), the pipeline success-

fully tracks organoid centroids throughout each recording. From the resulting centroid trajectories, we computed eight scalar features spanning commonly used measures for quantifying cell and organoid motility (Wu et al., 2014; Kimmel et al., 2018): mean speed, path length, mean-squared displacement (MSD) slope, mean turning angle, tortuosity, outreach ratio (maximum distance from the starting position divided by path length) (Svensson et al., 2018), total displacement, and acceleration. All eight features were computed from the raw centroid trajectories with identical processing, within the same 30-second windows used by the video prediction model.

Identifying the most reliable discriminator

A reference ranking of organoid motility requires a single feature that separates organoids reliably: variation across organoids should dominate variation within an organoid. We treat each organoid's non-overlapping 30-second windows as repeated measurements and compute the one-way ANOVA F-statistic (ratio of between-organoid to within-organoid variance) for each of the eight features. A high F indicates that a feature varies more across organoids than within them, making it a reliable discriminator. The highest-F feature is then averaged per organoid to produce the reference ranking against which prediction surprise is evaluated (Fig. 3A).

Video prediction model

We train a video prediction model to forecast future frames from past observations. For each organoid video, after downsampling to 128×128, we assemble training samples by extracting windows of $K + N$ consecutive frames with a stride of 2 frames between window start positions. The model receives $K = 6$ input frames and predicts the subsequent $N = 6$ frames; at 5-second intervals, this corresponds to 30 seconds of past motion used to predict 30 seconds of future motion, a window chosen to capture characteristic motility patterns while retaining sufficient training samples per organoid. Given input frames $\{I_1, I_2, \ldots, I_K\}$, the model predicts frames $\{\hat{I}_{K+1}, \hat{I}_{K+2}, \ldots, \hat{I}_{K+N}\}$.

We use a residual prediction formulation (He et al., 2016) instead of direct pixel prediction, where the model predicts per-frame differences rather than absolute pixel values. These differences are cumulatively summed from the last input frame I_K to reconstruct predicted frames:

$$\hat{I}_{K+i} = I_K + \sum_{j=1}^{i} \delta_j \tag{1}$$

where δ_j denotes the predicted residual for frame j. The model is trained to minimize mean squared error between predicted and ground truth frames:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N \cdot C \cdot H \cdot W} \sum_{i=1}^{N} \|\hat{I}_{K+i} - I_{K+i}\|_2^2 \tag{2}$$

where $C = 3$ color channels and $H = W = 128$ pixels.

We use the SimVP architecture with temporal attention units (TAU) (Tan et al., 2023a), which decomposes video prediction into spatial encoding, temporal modeling of encoded latents with TAU, and spatial decoding. We refer to this model with the residual formulation above as *SimVP-TAU+R*. For comparison, we also evaluated PredRNN (Wang et al., 2017), a recurrent architecture with spatiotemporal memory cells, with the same residual formulation (*PredRNN+R*). Architectural details are provided in the cited references and in the accompanying code.

The dataset of 108 organoid videos was split 80/10/10 into training, validation, and test sets (Goodfellow et al., 2016). We created 10 such splits, each with approximately 10–11 organoids in the test set, such that across all splits every organoid appears in at least one test set (Fig. 2A,B). This ensures that prediction error is computed on held-out data for all 108 videos. Models were trained for 100 epochs using the Adam optimizer with learning rate 10^{-3} and a one-cycle learning rate schedule (Smith and Topin, 2019), selecting the checkpoint with lowest validation loss.

Defining prediction surprise

Standard MSE loss (Eq. 2) does not differentiate between errors in predicting static appearance (organoid texture, background) and errors in predicting motion. To isolate the motion component, we define prediction surprise as the MSE between predicted and ground truth frame differences:

$$\mathcal{L}_{\text{motion}} = \frac{1}{N \cdot C \cdot H \cdot W} \sum_{i=1}^{N} \|\Delta\hat{I}_{K+i} - \Delta I_{K+i}\|_2^2 \tag{3}$$

where $\Delta I_t = I_t - I_{t-1}$ and $\Delta\hat{I}_t = \hat{I}_t - \hat{I}_{t-1}$ denote temporal differences between consecutive ground-truth and predicted frames, respectively. By comparing frame differences rather than absolute frames, static appearance cancels out by construction, and only the predicted motion contributes to the loss, assuming background does not vary within the video.

To generate motility rankings, we compute motion prediction error at two levels: per-window $\mathcal{L}_{\text{motion}}$ for analyzing within-organoid variation, and per-organoid error for comparing between-organoid variation. At test time, we extract sliding windows of $K + N = 12$ frames advancing by $N = 6$ frames (30 seconds), yielding $W_o = 19$ windows per organoid (Fig. 2C,D). For each organoid o, we define prediction surprise as the average motion loss across all windows from that organoid's held-out test video:

$$S_o = \frac{1}{W_o} \sum_{w=1}^{W_o} \mathcal{L}_{\text{motion}}^{(w)} \tag{4}$$

This per-organoid prediction surprise is used to rank organoids against the most reliable discriminating feature.

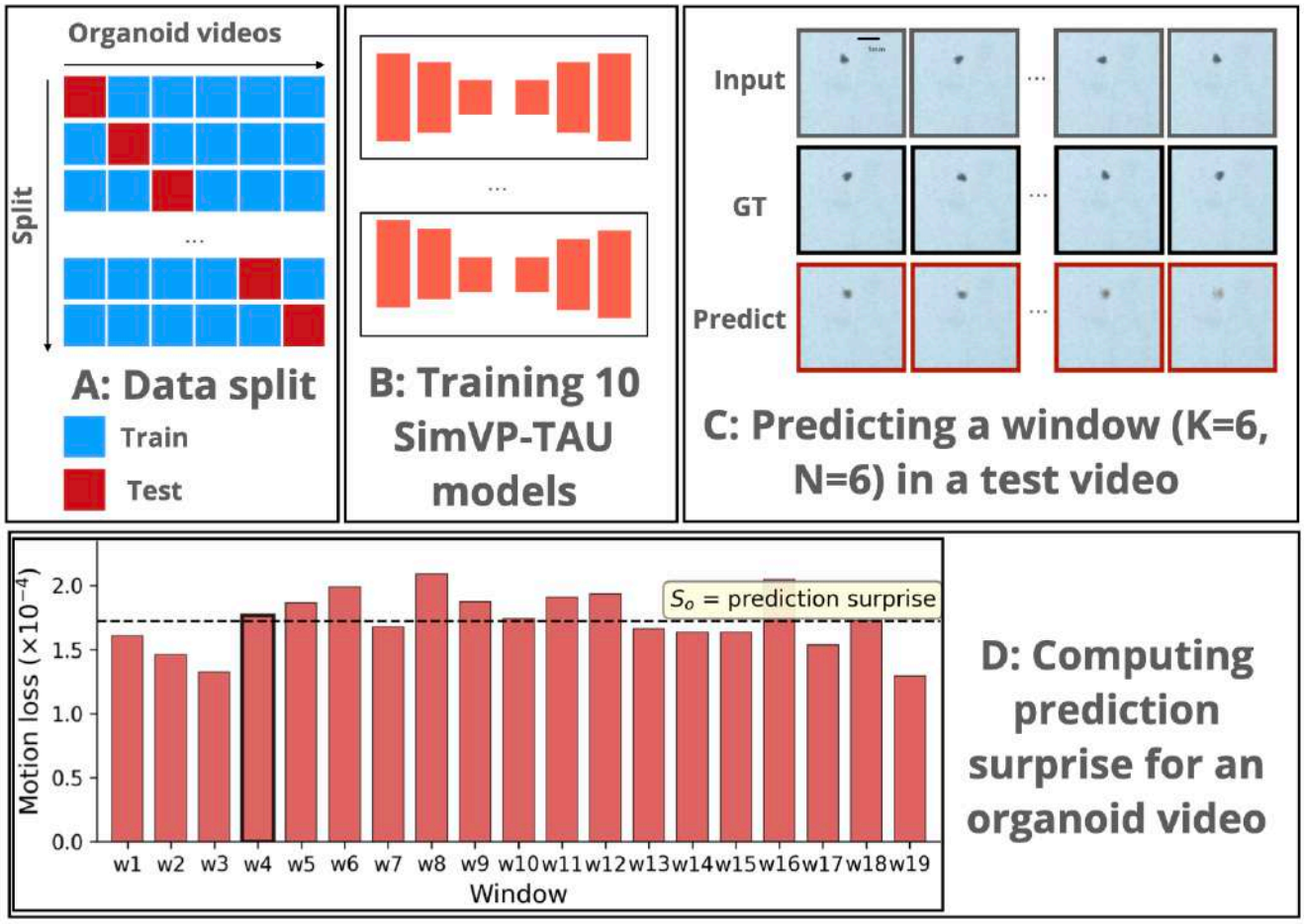

Figure 2: Training a video prediction model to compute prediction surprise. **(A)** The 108 organoids are split 10 ways so that every organoid is held out (red) in at least one split, while the remaining organoids are used for training (blue). **(B)** One video prediction model (SimVP-TAU, a residual next-frame predictor) is trained per split, yielding 10 models. **(C)** Each model is evaluated only on its held-out videos: a $K = 6$ context window predicts the next $N = 6$ frames. Input, ground-truth (GT), and predicted frames are shown for one of the 19 sliding windows (window w4). **(D)** Per-window motion loss $\mathcal{L}_{\text{motion}}(w)$ (Eq. 3) for a held-out organoid; prediction surprise S_o (dashed line) is its mean across all 19 windows (Eq. 4). The outlined bar marks w4, the window shown in (C).

Validation design

Our central claim is that prediction surprise discovers reliable differences in organoid motility directly from video, without computing classical motion features. Two conditions exist upon which this claim rests:

Condition 1: The model generalizes to unseen videos of organoids. A video prediction model that has successfully learned organoid motion dynamics should outperform simple heuristics on unseen organoids. We compare against two baselines: copy-last, which repeats the final input frame and represents the null hypothesis of no motion, and optical flow (Farnebäck, 2003), which extrapolates velocity by averaging flow from the last two context-frame pairs and iteratively warping for each prediction step. For each training seed, we average $\mathcal{L}_{\text{MSE}}$ (Eq. 2) across all windows of all held-out organoids (Fig. 3B). If the model fails to beat both baselines, it has failed to capture the non-trivial motility patterns, and prediction surprise cannot be employed as a valid measure of motility variation between organoids.

Condition 2: Agreement with the independent reference ranking. To identify the most reliable discriminating classical feature, we treat each organoid's non-overlapping 30-second windows as repeated measurements and compute the one-way ANOVA F-statistic for each of eight motion features. A high F indicates that a feature varies more across organoids than within them. We rank all 108 organoids by

the highest-F feature and by prediction surprise S_o (Eq. 4), then compute Spearman's ρ between the two rankings. The claim is falsified if $\rho \approx 0$.

Results

Result 1: Mean speed best discriminates individual organoids. Among eight classical features, mean speed yielded the highest F-statistic ($F = 576$, Fig. 3A). Path length ranked a close second ($F = 559$), but for fixed-length windows it is proportional to mean speed ($\rho \approx 1$) rather than independent; MSD slope followed ($F = 71$). Mean speed thus serves as the independent reference ranking for evaluating prediction surprise.

Result 2: Video prediction model outperforms baselines. *SimVP-TAU+R* achieved lower $\mathcal{L}_{\text{MSE}}$ (Eq. 2) than both baselines across all 10 splits ($p < 0.05$, Wilcoxon signed-rank with Bonferroni correction; Fig. 3B), satisfying Condition 1.

Result 3: Prediction surprise agrees with classical motility ranking. Spearman $\rho = 0.90$ ($p < 10^{-4}$, Fig. 3C), satisfying Condition 2.

Discussion

In this work, we address two questions: (1) Are organoid motility dynamics predictable, and can we learn these dynamics directly from videos? (2) What differentiates organoids in their motility, and can we discover the variation that reliably differentiates them without presupposing features? We find that there are common patterns in motility across organoids: a video prediction model generalizes to unseen organoids, learning non-trivial structure not explained by trivial baselines: copy-last and optical flow. Variation in organoid motility dynamics can be reliably captured without presupposing motility features, bypassing the classical multi-step pipeline: the Spearman rank correlation between prediction surprise and the classical pipeline's reference ranking is $\rho = 0.90$.

Achieving prediction generalization in high-dimensional video space is non-trivial, and four compounding challenges make the result meaningful. First, if organoid motility were random, there would be no patterns to learn. Second, even if patterns exist, it is not clear whether they are recoverable directly from videos without explicitly encoding kinematic states such as position, velocity, and acceleration. Third, Mean Squared Error (MSE), a canonical loss function in video prediction (Mathieu et al., 2015), has its own problems: if the organoid occupies a small portion of the frame, predicting the background achieves a low loss and can cause premature convergence to local optima. Fourth, the minimum neural network capacity required to capture non-linear motility dynamics is unknown; prior attempts with smaller models (Chen et al., 2022) failed to generalize in this dataset. SimVP-TAU+R (Tan et al., 2023a), with sufficient capacity (approx. $10\times$ parameters relative to (Chen et al., 2022)), overcomes each of these failure modes, demonstrating that organoid motility is not random, that its dynamics are learnable directly from videos, and that the model attends to organoid dynamics rather than exploiting background statistics to minimize loss during training.

The model captures non-trivial, non-linear motion patterns exhibited by organoids. Rather than exploiting background statistics, the model focuses on the organoid dynamics, as shown by the spatial error maps (Fig. 6B), discussed below. Its advantage over copy-last grows with organoid speed (Fig. 4), revealing that the model captures both the location and magnitude of motion more accurately than a static frame copy. Its advantage over the constant-velocity baseline of optical flow is shown in Fig. 5: the model predicts changes in organoid direction more accurately than optical flow, which assumes motion continues unchanged from the immediate past and cannot account for non-linear dynamics, particularly as motion patterns evolve over time. This advantage has one exception: for slow-moving organoids (below $\sim$0.1 px/frame), the model underperforms copy-last (Fig. 4), indicating that prediction surprise is most informative for organoids with detectable motion. This is expected: copy-last is the optimal predictor when motion is negligible.

Our method is robust to the choice of architecture and potential confounds such as organoid size. Prediction surprise rankings from SimVP-TAU+R and PredRNN+R agree closely ($\rho = 0.938$), confirming that the ranking is not attributable to a particular architectural choice. Organoid size is a potential confound: larger organoids occupy more of the frame and may generate higher prediction error simply due to their greater pixel footprint, independent of their motility. If size were correlated with prediction surprise, the ranking would reflect organoid geometry rather than dynamics. Organoid size is not correlated with prediction surprise (Fig. 6 A), ruling out size as a confounding factor in the observed relationship between prediction surprise and motility.

Prediction surprise captures aspects of organoid motility behavior that classical motility features do not. Describing motility through centroid trajectories inevitably discards information; morphodynamics and other organoid characteristics such as rotation about the organoid's axis are not easily extractable from the classical pipeline. For instance, tracking subtle shape changes requires accurate segmentation masks, and extracting features to define pose and infer rotational dynamics adds further complexity to the pipeline. Prediction surprise does not share this limitation: as the average spatial prediction error accumulated over the prediction horizon, it provides a direct measure of how dynamically distinct an organoid is.

This distinctness is most meaningful when contextualized within a reference population. Prediction surprise is inherently a population-relative measure: it describes how differently an organoid moves relative to the population dynamics the model was trained on, a comparative quantity that copy-last or optical flow cannot provide since they do not have a

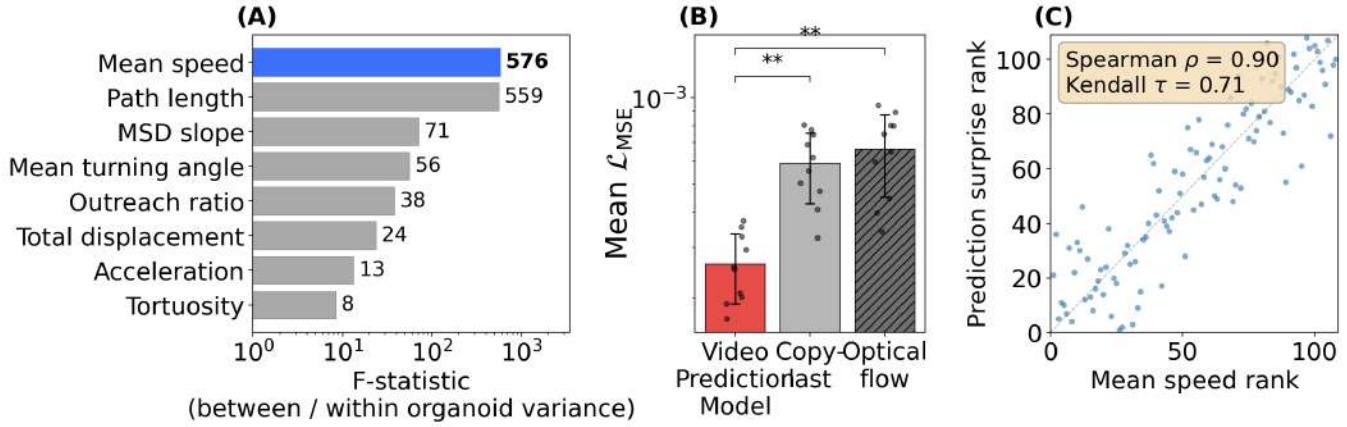

Figure 3: Results. **(A)** F-statistic (between-organoid / within-organoid variance) for eight classical features. Mean speed ($F = 576$, blue) most reliably distinguishes individual organoids, establishing the independent reference ranking. **(B)** Mean $\mathcal{L}_{\text{MSE}}$ (Eq. 2) for the video prediction model and both baselines, averaged across all windows and held-out organoids per seed. Each dot is one seed; bars show means $\pm$ 1 SD. The model (*SimVP-TAU+R*) outperforms both baselines across all seeds ($p < 0.05$, Wilcoxon signed-rank with Bonferroni correction). **(C)** Per-organoid mean speed rank (blue) and prediction surprise rank (red), sorted by mean speed ($n = 108$). Prediction surprise is highly correlated with the classical reference (Spearman $\rho = 0.90$, $p < 10^{-4}$).

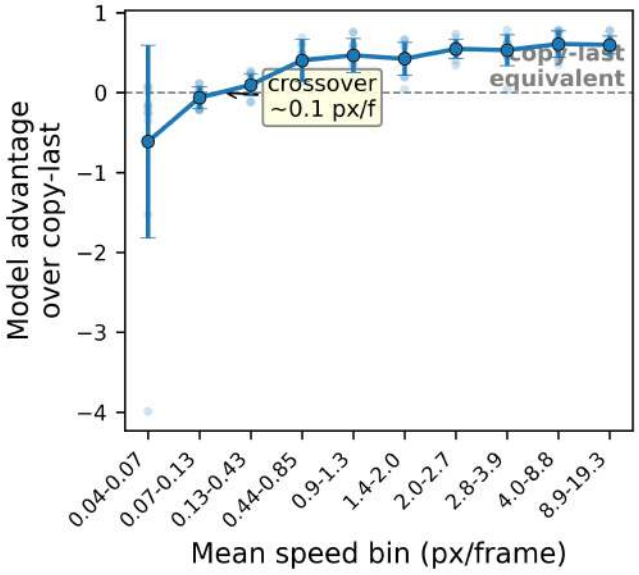

Figure 4: Model advantage over copy-last ($1 - \text{MSE}_{\text{model}}$ / $\text{MSE}_{\text{copy-last}}$) vs. mean speed bin (px/frame). Below $\sim$0.1 px/frame the model underperforms copy-last; above this crossover the advantage grows with speed. Error bars show 95% CI across seeds.

model of population dynamics. Characterizing motility relative to mean population behavior provides an immediately interpretable frame of reference: knowing that a particular organoid is a strong outlier focuses attention on its distinguishing properties, whether it is faster, more directed, or more confined than typical, and motivates downstream questions about what drives this divergence, from macro-scale video observations down to micro-scale properties such as cilia distribution, cell shape, and cytoskeletal organization.

Prediction surprise quantifies how much an organoid's motion deviates from the population norm, but it does not decompose this deviation into interpretable behavioral components. Whether this difference arises from speed, directionality, or morphodynamic shape change is not directly decoded from the scalar score alone. The spatial prediction error maps (Fig. 6B) provide partial insight by localizing where prediction error concentrates on each organoid, but a systematic decomposition of surprise into independent behavioral axes remains an open problem.

Several directions could extend this work. Longer observation windows would test whether prediction surprise captures sustained behavioral modes or is dominated by transient events. Interpreting the learned latent representations could help quantify the complexity of different organoid dynamics and reveal what the model encodes beyond speed. Applying the method to organoids before and after pharmacological or environmental interventions would test whether prediction surprise can detect treatment effects without presupposing which features the intervention would affect. Finally, applying the framework to datasets where speed is not the dominant axis of variation, for instance systems where trajectory geometry or morphodynamic shape changes differentiate individuals, would test the generality of prediction surprise beyond the motility regime studied here.

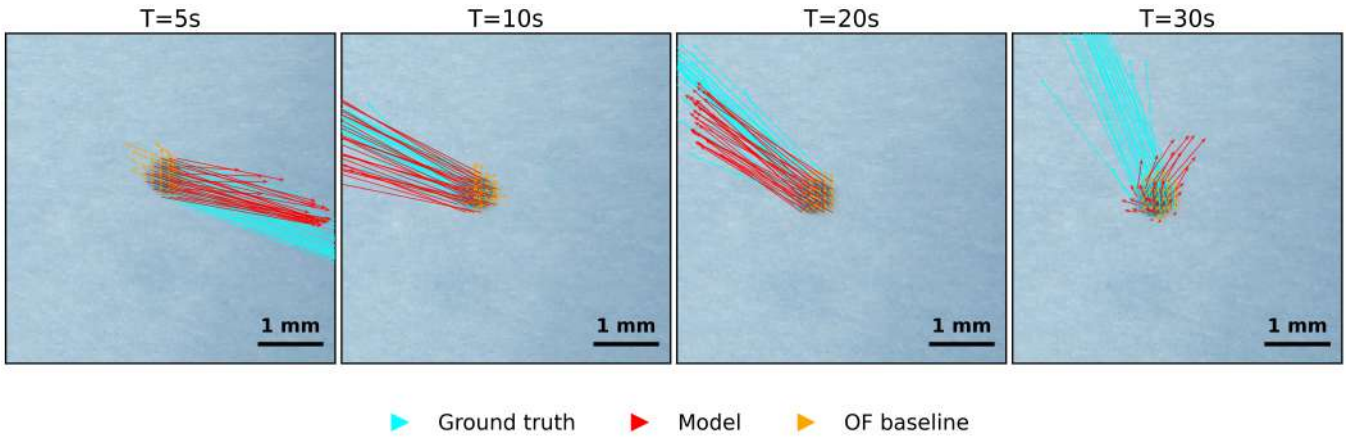

Figure 5: Visualizing model advantage over optical flow. Velocity flow vectors overlaid on a representative organoid at four time points (T = 5s, 10s, 20s, 30s). Cyan encodes ground-truth flow, red represents model prediction, and orange represents the optical flow baseline.

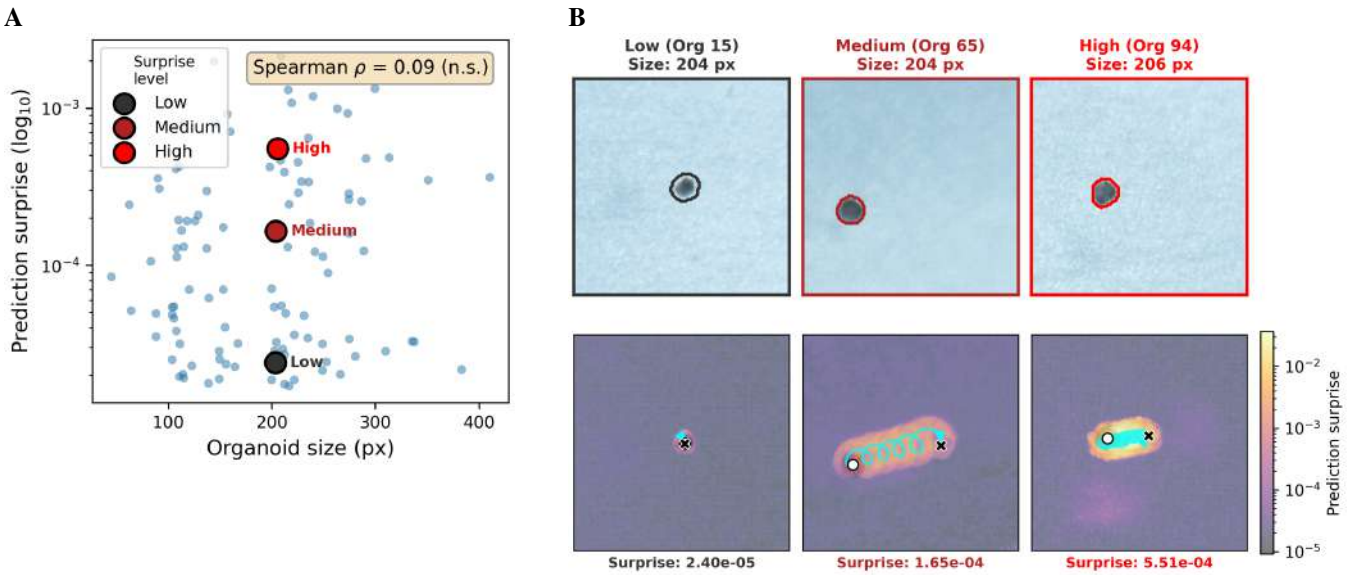

Figure 6: Prediction surprise is independent of organoid size. **(A)** Prediction surprise vs. organoid size (px). Spearman $\rho = 0.09$ (n.s.), confirming that the model's ranking is not confounded by organoid size. Three labeled organoids are shown in panel B. **(B)** Raw video frames (top) and spatial prediction error maps (bottom) for low, medium, and high values of prediction surprise, for organoids of similar size.

Conclusion

Characterizing behavioral diversity in novel living systems, of which motility variation is one instance, does not require knowing in advance what to look for. The classical pipeline's dependence on segmentation, tracking, and predefined features constrains it to a particular imaging condition and a prior assumption about which behaviors matter, requiring sustained human judgment that does not scale across large video collections with varying imaging conditions. Our method carries no such dependency: a population-trained video prediction model produces a ranking of behavioral distinctiveness from raw videos alone. Any motile construct imaged under varying backgrounds and imaging conditions is amenable to the same framework, including systems where the relevant axes of behavioral variation are genuinely unknown. For Artificial Life investigators studying novel living systems, a tool that characterizes behavioral diversity without presupposing what to measure offers a capability that predefined feature pipelines cannot provide.

Acknowledgements

This research was supported by the Engineer Research and Development Center – Cold Regions Research and Engineering Laboratory (ERDC-CRREL) under Contract No. W913E524C0012. The content of the information does not necessarily reflect the position or policy of the government, and no official endorsement should be inferred.

The authors used generative AI tools to assist with grammar and language editing of the manuscript and with code development. All schematic figures were created manually by the authors. All AI-assisted content, including code, was reviewed and verified by the authors, who bear sole responsibility for the content of this paper.

References

Berg, H. C. and Brown, D. A. (1972). Chemotaxis in *Escherichia coli* analysed by three-dimensional tracking. *Nature*, 239(5374):500–504.

Blackiston, D., Lederer, E., Kriegman, S., Garnier, S., Bongard, J., and Levin, M. (2021). A cellular platform for the development of synthetic living machines. *Science Robotics*, 6(52):eabf1571.

Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

Chen, B., Huang, K., Raghupathi, S., Chandratreya, I., Du, Q., and Lipson, H. (2022). Automated discovery of fundamental variables hidden in experimental data. *Nature Computational Science*, 2(7):433–442.

Farnebäck, G. (2003). Two-Frame Motion Estimation Based on Polynomial Expansion. In Goos, G., Hartmanis, J., Van Leeuwen, J., Bigun, J., and Gustavsson, T., editors, *Image Analysis*, volume 2749, pages 363–370. Springer Berlin Heidelberg, Berlin, Heidelberg. Series Title: Lecture Notes in Computer Science.

Froese, T., Virgo, N., and Ikegami, T. (2014). Motility at the origin of life: Its characterization and a model. *Artificial Life*, 20(1):55–76.

Gao, Z., Tan, C., Wu, L., and Li, S. Z. (2022). SimVP: Simpler yet Better Video Prediction. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3160–3170, New Orleans, LA, USA. IEEE.

Gershenson, C. (2023). Emergence in Artificial Life. *Artificial Life*, 29(2):153–167.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. Adaptive computation and machine learning. The MIT press, Cambridge, Mass.

Gumuskaya, G., Srivastava, P., Cooper, B. G., Lesser, H., Semegran, B., Garnier, S., and Levin, M. (2024). Motile living biobots self-construct from adult human somatic progenitor seed cells. *Advanced Science*, 11(4):e2303575.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, USA. IEEE.

Holme, B., Bjørnerud, B., Pedersen, N. M., Rodriguez de la Ballina, L., Wesche, J., and Haugsten, E. M. (2023). Automated tracking of cell migration in phase contrast images with CellTraxx. *Scientific Reports*, 13:22982.

Kimmel, J. C., Chang, A. Y., Brack, A. S., and Marshall, W. F. (2018). Inferring cell state by quantitative motility analysis reveals a dynamic state system and broken detailed balance. *PLOS Computational Biology*, 14(1):e1005927.

Kriegman, S., Blackiston, D., Levin, M., and Bongard, J. (2020). A scalable pipeline for designing reconfigurable organisms. *Proceedings of the National Academy of Sciences*, 117(4):1853–1859.

Langton, C. G., editor (1997). *Artificial Life: An Overview*. MIT Press, Cambridge, MA.

Lorimer, T., Goodridge, R., Bock, A. K., Agarwal, V., Saberski, E., Sugihara, G., and Rifkin, S. A. (2021). Tracking changes in behavioural dynamics using prediction error. *PLOS ONE*, 16(5):e0251053.

Maška, M., Ulman, V., Delgado-Rodriguez, P., Gómez-de Mariscal, E., Nečasová, T., Guerrero Peña, F. A., Ren, T. I., Tsaris, A., Wiesmann, V., et al. (2023). The cell tracking challenge: 10 years of objective benchmarking. *Nature Methods*, 20(7):1010–1020.

Mathieu, M., Couprie, C., and LeCun, Y. (2015). Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*.

Matthews, J. M., Schuster, B., Kashaf, S. S., Liu, P., Ben-Yishay, R., Ishay-Ronen, D., Izumchenko, E., Shen, L., Weber, C. R., Bielski, M., et al. (2022). OrganoID: A versatile deep learning platform for tracking and analysis of single-organoid dynamics. *PLOS Computational Biology*, 18(11):e1010584.

Oprea, S., Martinez-Gonzalez, P., Garcia-Garcia, A., Castro-Vargas, J. A., Orts-Escolano, S., Garcia-Rodriguez, J., and Argyros, A. (2022). A Review on Deep Learning Techniques for Video Prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):2806–2826.

Robertson, C., Wilmoth, J. L., Retterer, S., and Fuentes-Cabrera, M. (2023). Video frame prediction of microbial growth with a recurrent neural network. *Frontiers in Microbiology*, 13:1034586.

SHI, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-k., and WOO, W.-c. (2015). Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Sive, H. L., Grainger, R. M., and Harland, R. M. (2000). *Early development of Xenopus laevis: a laboratory manual*. Cold Spring Harbor Laboratory Pr, Cold Spring Harbor, New York.

Smith, L. N. and Topin, N. (2019). Super-convergence: very fast training of neural networks using large learning rates. In Pham, T., editor, *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, page 36, Baltimore, United States. SPIE.

Stephens, G. J., Johnson-Kerner, B., Bialek, W., and Ryu, W. S. (2008). Dimensionality and dynamics in the behavior of C. elegans. *PLoS computational biology*, 4(4):e1000028.

Svensson, C., Medyukhina, A., Belyaev, I., Al-Zaben, N., and Figge, M. T. (2018). Untangling cell tracks: Quantifying cell migration by time lapse image data analysis. *Cytometry Part A*, 93(3):357–370.

Tan, C., Gao, Z., Wu, L., Xu, Y., Xia, J., Li, S., and Li, S. Z. (2023a). Temporal Attention Unit: Towards Efficient Spatiotemporal Predictive Learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18770–18782, Vancouver, BC, Canada. IEEE.

Tan, C., Li, S., Gao, Z., Guan, W., Wang, Z., Liu, Z., Wu, L., and Li, S. Z. (2023b). OpenSTL: A Comprehensive Benchmark of Spatio-Temporal Predictive Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Wang, Y., Long, M., Wang, J., Gao, Z., and Yu, P. S. (2017). PredRNN: Recurrent Neural Networks for Predictive Learning using Spatiotemporal LSTMs. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wu, P.-H., Giri, A., Sun, S. X., and Wirtz, D. (2014). Three-dimensional cell migration does not follow a random walk. *Proceedings of the National Academy of Sciences*, 111(11):3949–3954.