INFORMATION-THEORETIC EMPOWERMENT ENHANCES HEALING FOLLOWING DAMAGE IN SELF-ORGANIZING NEURAL CELLULAR AUTOMATA.

Anonymous authors
Paper under double-blind review

ABSTRACT

A core problem in modern artificial intelligence research is building systems that are robust to the unexpected injuries and assaults that inevitably come from being embedded in a complex and unpredictable environment. Prior research has shown that distributed healing capabilities can be trained into ensembles of self-organizing neural cellular automata (NCAs), but the underlying mechanisms of how the system learns to heal remain unclear. Here we show that this innate healing capacity can be augmented and tuned by the inclusion of information-theoretic auxiliary loss functions: additional metrics that guide the formation of emergent structure but are independent from the actual task being performed. We explored two primary auxiliaries: the time-delayed mutual information between the state of the system before damage and after healing, and the empowerment. We report that including both auxiliaries improves the capacity of NCAs to heal following random damage above and beyond the performance achieved using the typical training regime. Furthermore, we find that this effect is most pronounced when the NCA attempts to fit complex, high-entropy targets, as well as improving the ability of individual cells to represent their position in space. This work uses efficiently and fully-differentiable estimators, and validates the potential of information-theoretic auxiliary functions as active channels with engineering potential. Finally, these results point towards a future high part-count machines and/or computers that can sustain and recover from unexpected internal damage or external surprise.

1 INTRODUCTION

A fundamental problem in both biology and artificial intelligence research is how populations of agents coordinate to produce collective patterns that are robust to a noisy world. In biology, the solution to this challenge is healing: when organisms are damaged, a cascade of processes begin that return the organism to its "expected" morphology. In extreme cases, this can extend as far as regrowing whole limbs Davidian & Levin (2021) and the core cellular machinery is believed to be involved in recovery from a wide range of ills Naviaux (2023). In artificial contexts, unexpected perturbations can leave robots crippled due to a lack of adaptive flexibility. This is a significant limitation, especially in contexts where robots may be deployed beyond the reach of help or repair Hein & Bongard (2026).

While individual robotic agents can adapt to injury via updated self-models Bongard et al. (2006); Cully et al. (2015), distributed systems have no such luxury, as no individual cell or agent orchestrates the whole body plan. Instead, some kind of representation of the "whole" must be distributed over all of the "parts". Evolution has solved this problem in living systems, but the mechanisms remain mysterious. This recognition has inspired a line of research into biologically-inspired forms of artificial self-organization and robustness, drawing inspiration from hormonal signaling Shen et al. (2004), the immune system Timmis et al. (2016), and stem cells Rubenstein et al. (2009). In recent years, neural cellular automata (NCA) have emerged as one of the most intriguing models of bio-inspired self-organization Hartl et al. (2026). The NCA model generalizes classic, discrete cellular automata by replacing deterministic update rules with a neural network that can be trained. NCAs

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

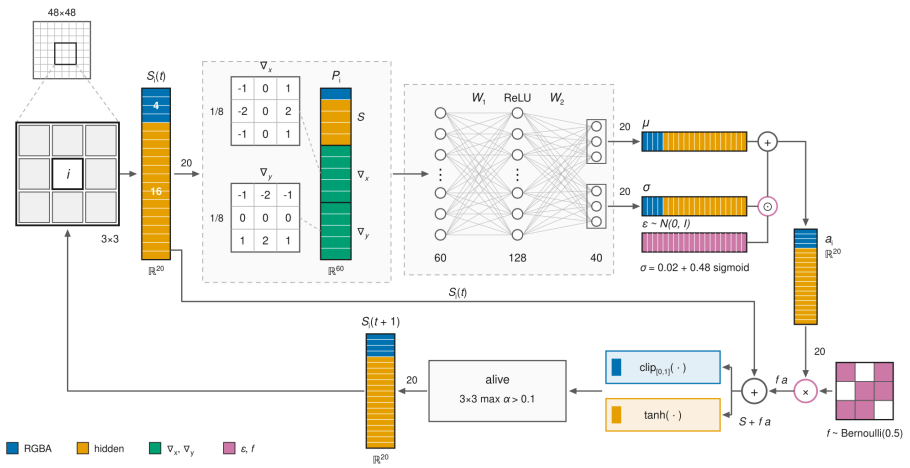

Figure 1: **Structure of the neural cellular automata.** At every timestep, the NCA perceives both its own current state (a 20-dim vector, and the Sobel-filtered state of its eight immediate neighbors. This 60-dim vector is then fed into a two-layer ANN, which outputs a 40-dim vector that is partitioned into two groups of 20. These specify the μ and σ parameters that define the distribution that the action vector is sampled from. On a given pass, there is a 50% chance that the cell will "fire" (update), drawn from a Bernoulli distribution. If the cell fires, the action is added to the current cell's state, the RGBA channels are clipped to (0,1) and the hidden channels passed through a $\tanh()$. If enough of the cells neighbors are "alive" (α-chanel ¿ 0.1), the update is accepted. This framework is adapted almost directly from Mordvintsev et al. (2020).

have proven to be remarkably flexible, not just as models of static self-organization and morphogenesis, but also for healing in response to injury Mordvintsev et al. (2020).

Recently, Grasso and Bongard Grasso & Bongard (2022; 2023) showed that, not only could NCAs learn to solve specific morophogensis and healing tasks, but that this performance could be augmented by the addition of intrinsic goal auxiliary functions to training. In their case, the auxiliary function was empowerment Klyubin et al. (2005); Salge et al. (2014), a measure of how capable an agent is of influencing its environment through its own action. Like other intrinsic motivators (curiosity, free-energy minimization, etc), empowerment does not promise to solve any particular task, but rather it is meant to produce a global competency that generally beneficial across domains.

A historical limitation of empowerment research has been reliance on discrete, non-differentiable estimators of mutual information. Studies will either use naturally discrete systems (such as Grasso & Bongard (2022; 2023); Varley & Bongard (2024)), or discretize a continuous signal Klyubin et al. to make it amenable to discrete estimators. These approaches, as well as continuous approximations based on Monte Carlo sampling Salge & Polani (2017); Ingel et al. (2022) are not generally differentiable or amenable to gradient-descent-based optimization pipelines. Discrete state-spaces also have limited biological plausibility.

In this paper, we leverage Gaussian estimators Varley (2025) of multivariate mutual information, implementing a fully differentiable analysis pipeline in Torch Paszke et al. (2019), which can be efficiently built into standard gradient-based learning pipelines. Building Gaussian empowerment (and two other auxiliary functions as references) into Mordvintsev-style NCAs, we explore how intrinsic motivations can facilitate healing following damage.

2 Methods

2.1 NCA Architecture

The architecture of our NCA was based on the model introduced in Mordvintsev et al. (2020). Cells are defined on a 48×48 lattice and each cell's state is an element in $\mathbf{S}_i \in \mathbb{R}^{20}$. The twenty

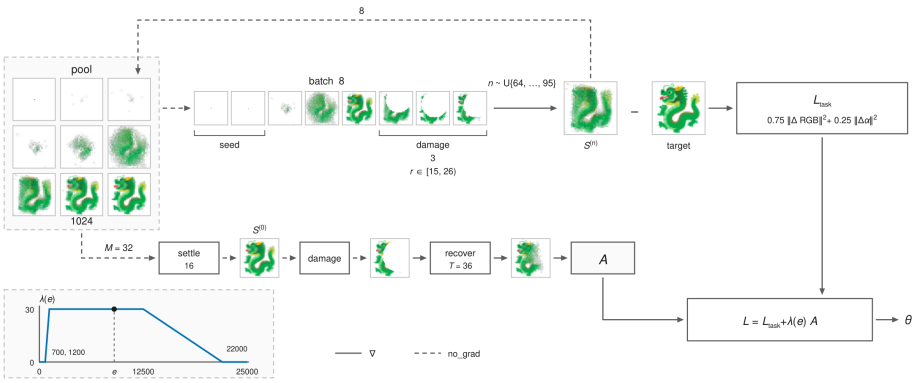

Figure 2: **The structure of each training epoch.** The system is initialized with a pool of 1024 seeds. At each epoch, a batch of eight seeds is drawn and the most well-formed (lowest-loss) are subjected to random damage. All eight are then run for between 64 and 95 steps (randomly sampled from a uniform distribution), and the final loss averaged over the batch. All seeds are then returned to the pool in their new state. In the same epoch, the 32 lowest-loss seeds are selected, allowed to run for 16 steps, damaged, and then allowed to heal for another 36 steps. From pre-damage/post-healing snapshots, the auxiliaries (L2, TDMI, or empowerment) are computed, and combined with the shape-matching loss to backpropagate.

dimensions are partitioned into subsets consisting of four visible channels (RGBA; clamped to $[0, 1]$) and sixteen hidden channels (bounded by $\tanh$). Growth begins from a single living cell (alpha $= 1$ at the center of the lattice.

Perception. Each cell perceives its own state together with the local spatial gradient of every channel. These gradients are computed using 3×3 Sobel filters (scaled by $1/8$) applied depthwise to all 20 channels.

Update network. The same network is applied at every cell. The perception vector is fed into a two-layer multilayer perceptron (MLP) that maps $\mathbb{R}^{60} \to \mathbb{R}^{128} \to \mathbb{R}^{40}$, with ReLUs between each layer. The total number of parameters is 12,968. The 40 outputs are partitioned into two 20-dimensional vectors: a mean $\boldsymbol{\mu}_i$ and raw scale vector that is normalized to:

$$\boldsymbol{\sigma}_i \;=\; 0.02 + 0.48 \;\times \mathrm{sigmoid}(\cdot) \;\in\; (0.02,\, 0.5), \tag{1}$$

This ensures that the noise scale has a hard floor.

Stochastic emission and state update. Both Mordvintsev et. al, and Grasso and Bongard used deterministic updates, however here we used a stochastic update to ensure that the Gaussian empowerment function is well-behaved. The output of the NCA is a stochastic 20-channel vector:

$$\mathbf{a}_i \;=\; \boldsymbol{\mu}_i + \boldsymbol{\sigma}_i \odot \boldsymbol{\epsilon}_i, \qquad \boldsymbol{\epsilon}_i \sim \mathcal{N}(0, I), \tag{2}$$

which we use to define the cell's action. At every step each cell independently "fires" with probability 0.5 and a firing cell applies its action to itself:

$$\mathbf{S}_i^{\text{vis}} \leftarrow \mathrm{clip}_{[0,1]}\big(\mathbf{S}_i^{\text{vis}} + f_i\, \mathbf{a}_i^{\text{vis}}\big), \qquad \mathbf{S}_i^{\text{hid}} \leftarrow \tanh\big(\mathbf{S}_i^{\text{hid}} + f_i\, \mathbf{a}_i^{\text{hid}}\big), \tag{3}$$

with $f_i \sim \mathrm{Bernoulli}(0.5)$. Finally, any cell whose 3×3 neighbourhood (itself included) contains no cell with alpha above 0.1 is zeroed and considered "dead".

2.2 Training

Targets. We trained one NCA per target, and target patterns were twelve animal emoji (Twemoji artwork, downsampled with a Lanczos filter to 48×48 RGBA): dragon, whale, phoenix, lizard, flamingo, kangaroo, worm, mammoth, lobster, wolf, honeybee, and oyster. Each configuration (target, auxiliary, seed) was trained from an independent initialization, with 32 seeds per target and auxiliary.

Task loss. The morphogenesis objective is a weighted squared error to the target image, summed over channels and pixels and averaged over the batch:

$$\mathcal{L}_{\text{task}} \;=\; \Big\langle \sum_{x,y} \Big[0.75 \sum_{c\in\{R,G,B\}} \big(S_c - T_c\big)^2 \;+\; 0.25\,\big(S_\alpha - T_\alpha\big)^2 \Big] \Big\rangle_{\text{batch}}. \tag{4}$$

Pool-based training with damage. Following Mordvintsev et al. (2020), we trained with a pool of 1024 grids, each initialized with a single live cell. At each epoch we drew a batch of 8 grids from the pool, replaced the highest-loss grid with a fresh seed and additionally reset the first slot to a fresh seed. This second swap was added so that there's always a guaranteed structure-free grid in a known slot that's excluded from the disk-damage eligibility pool (described below). Of the remaining grids, the three with the lowest loss (the best-formed patterns) were damaged by zeroing 20 channels inside of a disk randomly placed over the middle half of the lattice. The disk's radius was sampled from a uniform distribution over $[15, \ldots, 26]$ pixels). The batch was then run for a number of stochastic steps (allowed to "heal") drawn uniformly from $[64, \ldots, 95]$. The loss $\mathcal{L}_{\text{task}}$ was computed on the final states. Final states were returned to the pool for future sampling. We we used an AdamW optimizer (learning rate $= 10^{-3}$) with global gradient-norm clipping at 1.0, for $25,000$ epochs.

Auxiliary weighting. The auxiliary functions were normalized via a running root-mean-square (exponential moving average of the square, with decay $= 0.99$). This ensured that their weights were comparable across functions, and was added to the task loss:

$$\mathcal{L} \;=\; \mathcal{L}_{\text{task}} + \lambda(e)\,\frac{\mathcal{A}}{\sqrt{\text{EMA}[\mathcal{A}^2]} + 10^{-6}}. \tag{5}$$

The weight $\lambda(e)$ follows a scaffold schedule (see Fig. 2, bottom-left inset box): zero for the first 700 epochs, rises linearly to $\lambda = 30$ over the next 500 epochs, is held until 50% of the training epochs have elapsed before decaying linearly to zero at 88%, and remaining zero for the rest of training.

2.3 Auxiliary Functions

Each auxiliary is computed on a short shaping episode that is run parallel to task rollout and which is never written to the pool. The episode takes the $M{=}32$ lowest-loss members of the pool and runs them for 16 deterministic steps, giving a pre-injury state $\mathbf{S}^{(0)}$. An injury is then inflicted with the same disk damage as above, followed by 36 recovery steps that end at $\mathbf{S}^{(36)}$.

We write $\mathbf{h}_i \in \mathbb{R}^{16}$ for the hidden channels of cell i. We then collect all cells that are alive at both endpoints (subsampled to at most $16,384$ pairs) for computing the auxiliary losses.

2.3.1 Mutual Information Estimator

Both information-theoretic auxiliaries use a Gaussian (linear) estimator of mutual information Varley (2025). After z-scoring each variable over the sample axis, mutual information was defined as:

$$I(X;Y) = \tfrac{1}{2}\Big(\log|\Sigma_{XX}| + \log|\Sigma_{YY}| - \log|\Sigma_{XY}| \Big), \tag{6}$$

where Σ is a covariance matrix with a ridge 10^{-3} added. The primary benefit of this approach is that the estimator itself is differentiable (which is not typically the case for discrete mutual information estimators, such as those used by Grasso & Bongard (2022; 2023).

2.3.2 L2 Loss

As an information-free control, we give the system the answer key. The settled pre-injury hidden state $\mathbf{h}_i^{(0)}$ (held fixed, no gradient) is compared with the same cell's hidden state after recovery $\mathbf{h}_i^{(36)}$, and the mean squared distance is minimised directly:

$$\mathcal{A}_{L2} \;=\; \frac{1}{16\,N} \sum_{i=1}^{N} \big\|\mathbf{h}_i^{(T)} - \mathbf{h}_i^{(0)}\big\|_2^2, \tag{7}$$

over the N cells alive both at the end of the settle and at the end of recovery. The L2 auxiliary uses the same cell pairs as the TDMI loss, so it differs only in whether the objective is an information quantity or a squared distance to the truth.

2.3.3 TIME-DELAYED MUTUAL INFORMATION

The predictive-information auxiliary measures how much a cell's hidden state *before* injury tells us about its hidden state after injury and recovery:

$$\mathcal{A}_{TDMI} \;=\; -I\big(\{\mathbf{h}_i^{(0)}\};\{\mathbf{h}_i^{(35)}\}\big), \tag{8}$$

where the pre-injury side is held fixed (no gradient) and the post-recovery side carries the gradient. Unlike the L2 auxiliary loss, TDMI does not force the cells to match their pre-injury state, instead it only forces a correlation between them.

2.3.4 EMPOWERMENT

Following Grasso & Bongard (2022; 2023), we operationalized empowerment as the TDMI between the *actions* emitted by the system immediately before injury and the vector of Sobel-filtered percepts (plus its own state) after the recovery window:

$$E(\mathbf{X}) = I\big(\{\mathbf{a}_i\};\{\mathbf{p}_i^{(36)}\}\big). \tag{9}$$

Concretely, immediately prior to injury the system takes one step, emitting $\{\mathbf{a}_i\}$ with gradient, producing $\mathbf{S}^{(1)}$. The disk injury is then applied applied to $\mathbf{S}^{(1)}$ and the system runs the remaining 35 steps. The estimator is applied over pairs of cells alive after the action step and at the end of recovery.

2.3.5 CONTROLS

The baseline control runs the same settle/injure/recover episode as L2 and TDMI but computes nothing from it, so that its cost is matched and its loss is the task loss alone. Each information-theoretic auxiliary additionally has a shuffled null with an identical episode, cell pairs and estimator, in which the rows of the second variable ($\mathbf{h}^{(36)}$ or $\mathbf{p}^{(36)}$) are randomly permuted before the estimate, destroying the correspondence between a cell and its own future while preserving both marginals.

2.4 EVALUATION

Healing metric. Evaluation was noise-free (actions $\mathbf{a}_i = \boldsymbol{\mu}_i$; only the fire mask was stochastic). We took the 48 lowest-loss pool members of the trained model, settled them for 16 steps ($\mathbf{S}^{\mathrm{clean}}$), applied a disk injury (radius uniform over $[15, 26)$, fixed across all runs) to obtain $\mathbf{S}^{\mathrm{dmg}}$, and ran 48 recovery steps to obtain $\mathbf{S}^{\mathrm{post}}$. With $\mathrm{MSE} = \mathcal{L}_{\text{task}}$ evaluated on these states, healing is scored as the fraction of the injury-induced error that was recovered:

$$\mathrm{regen} \;=\; 1 - \frac{\mathrm{MSE}(\mathbf{S}^{\mathrm{post}}) - \mathrm{MSE}(\mathbf{S}^{\mathrm{clean}})}{\mathrm{MSE}(\mathbf{S}^{\mathrm{dmg}}) - \mathrm{MSE}(\mathbf{S}^{\mathrm{clean}})}, \tag{10}$$

so that 1 is complete recovery, 0 is no recovery, and negative values are worse than the injured state.

Statistics. For each target, each auxiliary was compared with the baseline over 32 seeds paired by seed. We report the paired mean difference with a 95% bootstrap confidence interval, the paired t-test, and Hedges' g computed with the pooled standard deviation of the two arms. A seed was excluded from a comparison if either arm's clean error exceeded $\max(300, 3\times$ the median clean error of that target's baseline runs$)$, which flags a collapsed run. This only occurred for the kangaroo emoji target, for unknown reasons, and meaning the kangaroo only had 25 of the 32 replicates.

3 RESULTS

Auxiliary	Own target (unit)	Mean Δ	Range	Targets sig.	Wrong direction
TDMI	$I_G(\mathbf{h}^{(0)}; \mathbf{h}^{(T)})$ (nats)	1.87 ± 0.37	$[1.22,\ 2.34]$	12/12, 12	0/12
Empowerment	$I_G(\mathbf{a}; \mathbf{P}^{(T)})$ (nats)	3.16 ± 0.51	$[2.54,\ 3.94]$	12, 12/12	0/12
L2	$\|\mathbf{h}^{(T)} - \mathbf{h}^{(0)}\|^2$	0.027 ± 0.008	$[0.017,\ 0.045]$	12, 12/12	0/12

Table 1: Each auxiliary reliably moves its own target quantity beyond the baseline. All $12 \times 3 = 36$ target-arm tests are significant at $p < 0.05$ and no target moves in the wrong direction for any arm.

For twelve different animal emoji, we trained 32 separate NCAs each on pure shape-matching (no auxiliaries), the three auxiliaries (L2, TDMI, and empowerment), as well as on shuffled nulls for TDMI and empowerment. We found that all twelve emojis successfully developed their target auxiliary above the baseline (see Table 3).

3.1 INFORMATION-THEORETIC AUXILIARIES IMPROVE HEALING ABOVE BASELINE.

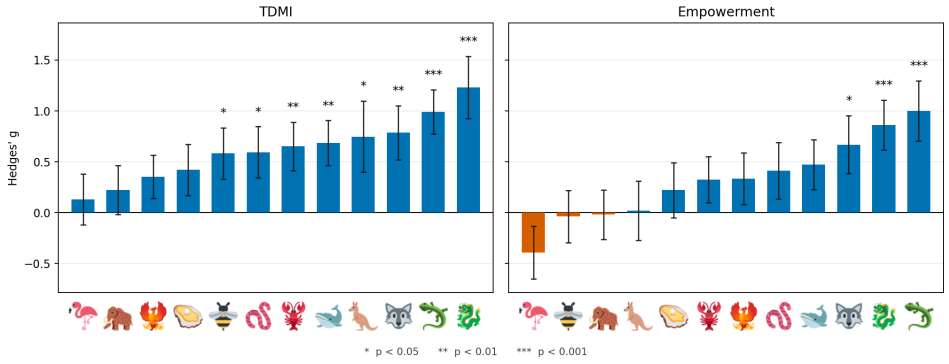

Figure 3: **TDMI and empowerment have varying impact on healing depending on target. Left:** The twelve emoji targets, ranked by the benefit to healing from the inclusion of the auxiliary in training. All twelve showed some directional benefit, and eight of the twelve showed a significant benefit (N=32, uncorrected p-value), with lizard and dragon showing the most benefit. **Right:** The benefits of empowerment were generally weaker than TDMI. Only three emoji showed a statistically significant benefit (again, dragon and lizard showed the clearest effect). Unlike TDMI, there was one target that actually suffered worse performance with the auxiliary: the flamingo.

Auxiliary	Mean g	Range	Sig. targets (uncorr.)	Targets $g < 0$	Aggregate (t, p)
TDMI	$+0.62$	$[+0.13,\ +1.23]$	8/12	0/12	$6.83,\ p < 0.0001$
Empowerment	$+0.32$	$[-0.39,\ +1.00]$	3/12	3/12	$2.79,\ p = 0.018$
L2	$+0.48$	$[-0.54,\ +1.05]$	7/12	1/12	$3.87,\ p = 0.003$

Table 2: Healing benefit of each auxiliary relative to the no-auxiliary baseline, across the 12 emoji targets (Hedges' g on regeneration, paired by seed, 32 seeds per target (25 for kangaroo)). Aggregate row is a one-sample t-test of the 12 per-target g values against zero.

There was a surprising amount of variability between emoji in regards to the amount of healing benefit they received from any auxiliary (see Fig. 3 and Table 2). For L2, seven out of the twelve emoji showed a statistically significant benefit to healing capacity over baseline. The same pattern was true

for TDMI, which showed approximately the same pattern: eight out of twelve showed benefit. Empowerment showed a weaker effect: onlyl three out of twelve showed significant benefit, while one emoji (the flamingo) actually suffered worse performance than baseline. Surprisingly, there as no correlation between the degree of auxiliary learned and the benefit to healing - the variability in healing gain between emoji was not attributable simply to the overall amount of L2/TDMI/empowerment learned. We also found no benefit to healing in any of the shuffled nulls for TDMI and empowerment. This is a good sanity check as it confirms that the healing benefit is attributable to the specific information-structure learned by the system, and not attributable to some trivial effect of the extra damage episode.

3.2 AUXILIARY BENEFIT CORRELATES WITH LEARNED LOCAL COLOR ENTROPY

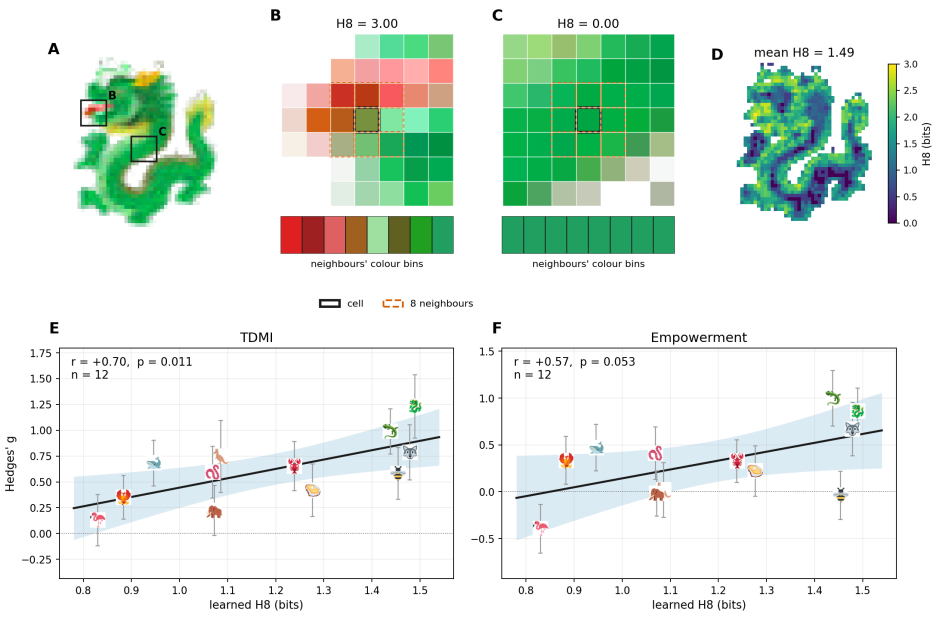

Figure 4: **Auxiliary benefit correlates with learned local entropy. A:** Different parts of the target image have differing levels of local color-complexity (the entropy of the colors of all eight neighbors). We have selected two regions, a high-entropy region (**B**) and a low-entropy region (**C**) to illustrate. **D** Every pixel in the sprite colored by its local neighborhood color entropy. **E:** There was a significant correlation between the average local color entropy of a sprite and the degree to which it benefited from the TDMI auxiliary function (Pearson's r=0.7, $p = 0.011$). **F:** The relationship between the benefit attributable to empowerment and local entropy was weaker (Pearsonr's r=0.57, $p = 0.053$) but remained directionally consistent.

We hypothesized that one possible explanation for why information-theoretic auxiliaries helped some emoji and not others is that different emoji have different levels of visual complexity, and that empowerment might help the NCA "process" more complicated percepts. For each cell, we quantized each neighbor's color into four bins per channel (64 possible colors, with "dead" channels getting their own bin), and then computed the Shannon entropy of that eight-sample distribution. Averaging this entropy over all cells in the image gave a heuristic global measure of the color-complexity of the image. For visualization, see Figure 4A-D. We found no significant correlation between healing gain and the local color entropy of the target images themselves, however, there was a correlation between healing gain and the *learned* target entropy (i.e. the color entropy of the NCA in a settled state prior to damage - for visualization see Fig. 4E-F). This discrepancy occurs because the NCA is not a perfectly high-definition match to the target image: small-scale textures get blurred, and so the learned entropy can be different from the true target entropy.

3.3 INFORMATION-THEORETIC AUXILIARIES TRAIN CELLS TO KNOW WHERE THEY ARE

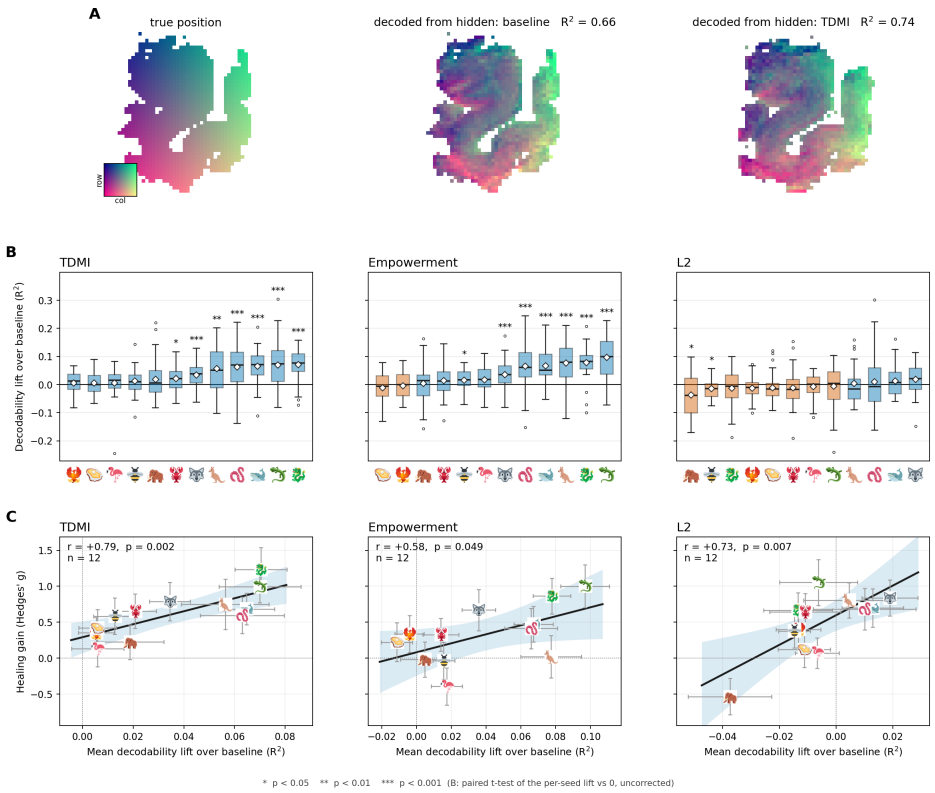

Figure 5: **Location decodability correlates with healing benefit from information-theoretic functions. A:** Every pixel is located at a (x, y) coordinates in the 48×48 grid. In the leftmost dragon, pixels are colored by exact location: the x-coordinate is given by the red channel, the y-coordinate is given by the green channel, the blue channel is fixed. The second two are the predicted (x, y) locations based on the sixteen hidden channels. **B:** Both TDMI and empowerment show show a significant growth in location decodability for some emoji, while the L2 (non-information theoretic auxiliary) shows no significant benefits. **C:** For all three nulls, the increase in decodability (R^2) correlates with the improvement in healing capacity (Hedge's g). Here we report the two information theoretic functions which showed growth in decodability (unlike L2): TDMI (Pearson's r=0.79, $p = 0.002$), followed by empowerment (Pearson's r=0.58, $p = 0.049$.)

Following damage, the NCA has to correctly regrow the lost pieces of itself (head, limbs, etc). Doing this correctly requires having some access to where in the body a given cell is located, as its location determines what features it needs to express. To test whether the auxiliaries improved the ability of cells to represent their location (known as positional information Wolpert (2016)), we tested whether the sixteen-channel hidden state stored information about location in the (x, y) plane. We used a ridge regression (λ=1) predicts a cell's own coordinates from the states of its 16 hidden channels. The regression was validated by 5-fold cross-validated on R^2. For visualization see Figure 5A.

We found that both TDMI and empowerment increased the degree to which location could be decoded from hidden state for at least some of the emoji (7/12 for TDMI and 6/12 for empowerment). However, for the L2 auxiliary, there was no significant lift in decodability. See Figure 5B. For TDMI and empowerment, there was a significant correlation between average decodability lift from baseline to auxiliary and the healing factor gain: the more the NCAs for a given emoji learned their own location, the better they were at healing following injury (see Fig. 5C).

This is an intriguing result, as the problem of cellular positional information is a long-standing one Wolpert (2008; 2016), and has previously been described in explicitly information-theoretic descriptive terms Dubuis et al. (2013); Tkačik & Gregor (2021). This is the first time, to our knowledge, that a link has been proposed between a dynamical intrinsic quality such as empowerment and spontaneous development of cellular positional information. Future work in this vein may provide new insights both into the functioning of self-organizing biological systems, as well as the engineering of embodied with emergent structure.

4 Discussion

In this paper, we have shown that the inclusion of differentiable, information-theoretic auxiliaries into the training of a neural cellular automata can improve the ability of the collective to respond to injuries. Furthermore, these benefits are not totally random, but can be associated to specific features of both the individual cell's ability to represent salient information (its location), and the nature of its local environment (neighborhood color entropy). This builds directly on prior work with NCA Mordvintsev et al. (2020); Grasso & Bongard (2022; 2023), showing that information theoretic intrinsic goal functions can synergistically interact with complex tasks to produce systems more robust than can be achieved with just task-specific training alone. By using a fully differentiable mutual information estimator, we also show that the auxiliaries can be efficiently built into standard training pipelines, without needing compute-intensive strategies like evolutionary optimization and Monte Carlo sampling.

The question of why some emoji are more amenable to auxiliary-supported healing than others remains unclear. With 32 replicates per auxiliary per emoji, we are reasonably confident that the differences are not random sampling effects, but what the relevant driver is is an open question. Future work on what statistical features of a target (or task more generally) make in amenable to TDMI or empowerment may provide more concrete insights into when, and why, these measures work the way they do. The finding that improvement is correlated with learned local entropy suggests that part of the benefit may be in helping cells cope with complex, noisier environments, although this hypothesis requires further testing.

These results also give us some insight into what kinds of systems might benefit from empowerment. Specifically, systems that have a non-trivial geometric structure or morphology. The finding that information-theoretic auxiliary gain is associated with cells ability to represent their own location in the collective suggests that homogeneous systems may not benefit the way the NCA does. Conversely, it is possible that empowerment *helps* individuals leverage the existing structure of the system - our current battery cannot distinguish between these two cases.

Collectively, these results point towards a promising future in "emergent engineering" - rather than just trying to train systems to satisfy specific tasks, controlling the collective patterns of self-organization across time and space to produce more powerful and adaptive systems.

Author Contributions

TFV and JB conceptualized the project. TFV wrote code, performed analysis, and wrote the manuscript. JB provided supervision and feedback.

Disclosure of AI use

The Claude model (Sonnet and Opus) was used by the authors for assistance with research (literature search), for assistance with coding (foundational code was written by TFV, Claude ran specific experiments as directed and interfaced with the UVM HPC cluster), and general ideation/"rubber ducking". The manuscript was written by the authors.

References

Josh Bongard, Victor Zykov, and Hod Lipson. Resilient Machines Through Continuous Self-Modeling. *Science*, 314:1118–1121, 2006. doi: {10.1126/science.1133687}.

Antoine Cully, Jeff Clune, Danesh Tarapore, and Jean-Baptiste Mouret. Robots that can adapt like animals. *Nature*, 521:503–507, 2015. doi: {10.1038/nature14422}.

Devon Davidian and Michael Levin. Inducing Vertebrate Limb Regeneration: A Review of Past Advances and Future Outlook. *Cold Spring Harbor Perspectives in Biology*, pp. a040782, 2021. doi: {10.1101/cshperspect.a040782}.

Julien O. Dubuis, Gašper Tkačik, Eric F. Wieschaus, Thomas Gregor, and William Bialek. Positional information, in bits. *Proceedings of the National Academy of Sciences*, 110:16301–16308, 2013. doi: {10.1073/pnas.1315642110}.

Caitlin Grasso and Josh Bongard. Empowered neural cellular automata. pp. 108–111, 2022. doi: {10.1145/3520304.3529067}. URL https://doi.org/10.1145/3520304.3529067.

Caitlin Grasso and Josh Bongard. Selection for short-term empowerment accelerates the evolution of homeostatic neural cellular automata. pp. 147–155, 2023. doi: {10.1145/3583131.3590469}. URL https://doi.org/10.1145/3583131.3590469.

Benedikt Hartl, Michael Levin, and Léo Pio-Lopez. Neural cellular automata: Applications to biology and beyond classical AI. *Physics of Life Reviews*, 56:94–108, 2026. doi: {10.1016/j.plrev.2025.11.010}.

Alice Hein and Josh Bongard. Environmental resilience via morphological diversity within machines. 2026. URL https://arxiv.org/abs/2608.02395.

Anti Ingel, Abdullah Makkeh, Oriol Corcoll, and Raul Vicente. Quantifying Reinforcement-Learning Agent's Autonomy, Reliance on Memory and Internalisation of the Environment. *Entropy*, 24:401–401, 2022. doi: {10.3390/e24030401}.

Alexander S. Klyubin, Daniel Polani, and Chrystopher L. Nehaniv. Keep Your Options Open: An Information-Based Driving Principle for Sensorimotor Systems. *PLoS ONE*, 3:e4018. doi: {10.1371/journal.pone.0004018}. URL http://dx.doi.org/10.1371/journal.pone.0004018.

A.S. Klyubin, D. Polani, and C.L. Nehaniv. Empowerment: A Universal Agent-Centric Measure of Control. volume 1, pp. 128–135, 2005. doi: {10.1109/cec.2005.1554676}. URL http://dx.doi.org/10.1109/cec.2005.1554676.

Alexander Mordvintsev, Ettore Randazzo, Eyvind Niklasson, and Michael Levin. Growing Neural Cellular Automata. *Distill*, 5, 2020. doi: {10.23915/distill.00023}.

Robert K. Naviaux. Mitochondrial and metabolic features of salugenesis and the healing cycle. *Mitochondrion*, 70:131–163, 2023. doi: {10.1016/j.mito.2023.04.003}. URL https://www.sciencedirect.com/science/article/pii/S1567724923000351.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. *PyTorch: an imperative style, high-performance deep learning library*. 2019.

Michael Rubenstein, Ying Sai, Cheng-Ming Chuong, and Wei-Min Shen. Regenerative patterning in Swarm Robots: mutual benefits of research in robotics and stem cell biology. *The International Journal of Developmental Biology*, 53:869–881, 2009. doi: {10.1387/ijdb.092937mr}.

Christoph Salge and Daniel Polani. Empowerment As Replacement for the Three Laws of Robotics. *Frontiers in Robotics and AI*, 4, 2017. doi: {10.3389/frobt.2017.00025}.

Christoph Salge, Cornelius Glackin, and Daniel Polani. *Empowerment–An Introduction*, pp. 67–114. 2014. doi: {10.1007/978-3-642-53734-9_4}. URL http://dx.doi.org/10.1007/978-3-642-53734-9_4.

Wei-Min Shen, Peter Will, Aram Galstyan, and Cheng-Ming Chuong. Hormone-Inspired Self-Organization and Distributed Control of Robotic Swarms. *Autonomous Robots*, 17:93–105, 2004. doi: {10.1023/B:AURO.0000032940.08116.f1}.

J. Timmis, A.R. Ismail, J.D. Bjerknes, and A.F.T. Winfield. An immune-inspired swarm aggregation algorithm for self-healing swarm robotic systems. *Biosystems*, 146:60–76, 2016. doi: {10.1016/j.biosystems.2016.04.001}.

Gašper Tkačik and Thomas Gregor. The many bits of positional information. *Development*, 148, 2021. doi: {10.1242/dev.176065}.

Thomas F. Varley. Information theory for complex systems scientists: What, why, and how. *Physics Reports*, 1148:1–55, 2025. doi: {10.1016/j.physrep.2025.09.007}. URL https://www.sciencedirect.com/science/article/pii/S037015732500256X.

Thomas F. Varley and Josh Bongard. Evolving higher-order synergies reveals a trade-off between stability and information-integration capacity in complex systems. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 34:63127–63127, 2024. doi: {10.1063/5.0200425}. URL https://doi.org/10.1063/5.0200425.

Lewis Wolpert. Positional Information, 2008.

Lewis Wolpert. Positional Information and Pattern Formation. pp. 597–608. 2016. doi: {10.1016/bs.ctdb.2015.11.008}.