EMERGENT FORAGING STRATEGIES AND FUNCTIONAL NEURAL MOTIFS IN HOMEOSTATIC DEEP RL AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Adaptive behavior in natural environments depends on a dynamic coupling between an organism's internal physiological state and its ongoing motor decisions. However, the computational mechanisms underlying the integration of interoceptive signals with motor control in embodied agents remain poorly understood. We train embodied artificial agents on a continuous two-resource foraging task using a homeostatic reinforcement-learning (HRL) objective, in which the agent minimizes deviation of its internal state from a set point rather than maximizing extrinsic reward. The resulting agents develop state-dependent search strategies that parallel biological foraging principles. Both exploration extent and movement velocity scale with homeostatic state, a dynamic analogous to hunger-modulated behavior. Clustering population activity around foraging events, such as food picking, reveals recurring temporal motifs: pre-event, peri-event, post-event, ramping, and sustained. Systematic cluster-level ablation demonstrates that these motifs contribute distinctly to behavior; for instance, ablating ramping clusters yields hyperactive, excessive-exploration policies. To assess the importance of learning under physiological constraints, we also train agents with an identical neural architecture, body, and environment, but using conventional reinforcement learning (RL). On the behavioral side, the RL agents lack the general state-dependent foraging strategies; and on the neural side, most of the functional neural motifs found in HRL agents become less prevalent or disappear. These results provide a neuroethologically grounded computational framework that links behavior to neural circuit dynamics and embodied constraints.

1 INTRODUCTION

Adaptive decision-making in animals and machines within naturalistic environments arises from the dynamic integration of (1) environmental state, (2) biomechanical affordances, and (3) internal physiological (homeostatic) state. Foraging, one of the most fundamental adaptive behaviors (Grima et al., 2025), depends on all three factors. Although environmental properties (e.g., average reward rate, Stephens & Krebs, 1986) influence foraging, physiological need strongly modulates exploration (Kamath et al., 2025). Additionally, body mechanics constrain movement costs (Halsey, 2016). Reinforcement learning (RL) (Dayan & Daw, 2008; Sutton et al., 1998) models, a widely used framework for studying adaptive decision-making, have primarily focused on environmental state and paid less attention to biomechanical and physiological factors. In this study, we examine foraging decisions by incorporating all three components.

To account for biomechanical affordances and realistic environmental dynamics, and to investigate their influence on adaptive decision processes, we use the well-established MuJoCo (Todorov et al., 2012) physics simulation framework. MuJoCo is widely adopted in robotics (Zakka et al., 2025) and computational neuroscience, supporting a growing set of virtual animals, including embodied models of fly (Vaxenburg et al., 2025; Wang-Chen et al., 2024) and rodents (Aldarondo et al., 2024; Merel et al., 2019), which reproduce sensorimotor loops in realistic bodies. We use this framework to place a quadruped in a continuous two-resource foraging environment. Although these virtual-animal models are valuable, they share a limitation: the agent possesses a body and an environment, but lacks an internal physiological state to regulate.

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.

To go beyond conventional biomechanical affordances, we equip the embodied agents with an internal physiological state and employ homeostatic reinforcement learning (HRL) as the primary optimization objective. In HRL, the agent minimizes deviations of its internal state from a specified set point rather than maximizing an abstract external reward (Yoshida et al., 2024a;b). Most RL agents are trained to maximize an abstract reward signal, resulting in policies that provide limited insight into the interoceptive mechanisms underlying animal decision-making. Determining whether an internal physiological signal is represented remains challenging, and identifying the specific coupling between such signals and the motor outputs that execute decisions is even more complex. Benchmark scores commonly used for behavioral assessment are insufficient for addressing these issues, as agents matched on reward may achieve similar performance through fundamentally different internal mechanisms (Simmons-Edler et al., 2025). Therefore, understanding interoceptive computation in artificial systems and evaluating their conceptual alignment with biological frameworks requires a biologically grounded definition of reward (Weber et al., 2025) and a mechanistic analysis of the trained policy and neural dynamics, rather than reliance on aggregate performance metrics. Thus, integrating HRL and MuJoCo unifies all three components of adaptive decision-making within a single framework.

To investigate the neural mechanisms underlying decision-making, we implement HRL in an embodied artificial neural network (ANN) trained for foraging. Virtual animals replicate the physics and sensorimotor loops of flies and rodents (Vaxenburg et al., 2025; Wang-Chen et al., 2024; Merel et al., 2019; Aldarondo et al., 2024). However, these models typically generate only simple behaviors, and researchers seldom map their neural dynamics to complex actions (see Appendix A for details). For example, Aldarondo et al. (2024) demonstrated that real rat neural activity could be predicted by the ANN controlling a virtual MuJoCo rodent, surpassing predictions based solely on raw kinematics. Such studies remain uncommon and are primarily correlational. While correlation indicates that neural circuits encode behavior, establishing causal contributions necessitates experimental intervention. We train actor–critic ANNs using proximal policy optimization (PPO) (Schulman et al., 2017) on a continuous two-resource problem (TRP) foraging task and analyze the trained policy across five physiological states, ranging from fully deprived to fully sated. This methodology facilitates the investigation of the neural and computational basis of adaptive decision-making in embodied agents.

We conduct a diverse set of neuroethologically grounded analyses of agents' behavior and neural dynamics, which reveal the emergence of adaptive foraging strategies and specialized functional subpopulations. The homeostatic objective leads to state-dependent foraging, including a torpor-like, energy-conserving mode during periods of severe deprivation. Underlying these rich behaviors, the agent's neural network consistently develops distinct functional motifs defined by neural dynamics, such as ramping and various other patterns associated with food-consumption events. Ablation analysis demonstrates the causal significance of these clusters. For example, removing ramping motifs yields faster but more confined movement, whereas ablating post-event motifs lowers both food intake and arena coverage. A conventional RL control without the homeostatic objective does not produce these specific behavioral and neural dynamics (Appendix K). These findings demonstrate that homeostatic regulation is the primary factor shaping adaptive behaviors and their corresponding neural signatures. Collectively, these results establish a neuroethologically grounded framework for investigating the neural and computational bases of adaptive decision-making in embodied agents.

2 METHODS

2.1 ENVIRONMENT: THE TWO-RESOURCE PROBLEM

The agent is instantiated in a continuous TRP foraging task using the MuJoCo engine (Todorov et al., 2012). The agent's body is a quadrupedal Ant robot that operates in a continuous two-dimensional arena. At the start of each episode, the arena is randomly populated with two distinct food types (red and blue). Each food type replenishes a specific internal variable the agent must regulate. Locomotion depletes the agent's internal reserves, whereas consuming food replenishes the corresponding internal variable. This coupling creates a two-resource trade-off, requiring the agent to balance acquiring both food types rather than optimizing a single resource. To assess adaptation to physiological need, the agent is evaluated under five fixed internal states ($h \in \{-0.7, -0.3, 0, +0.3, +0.7\}$), ranging from fully deprived to fully sated (Figure 1a). At the beginning of each episode, there are 6 blue and 4 red

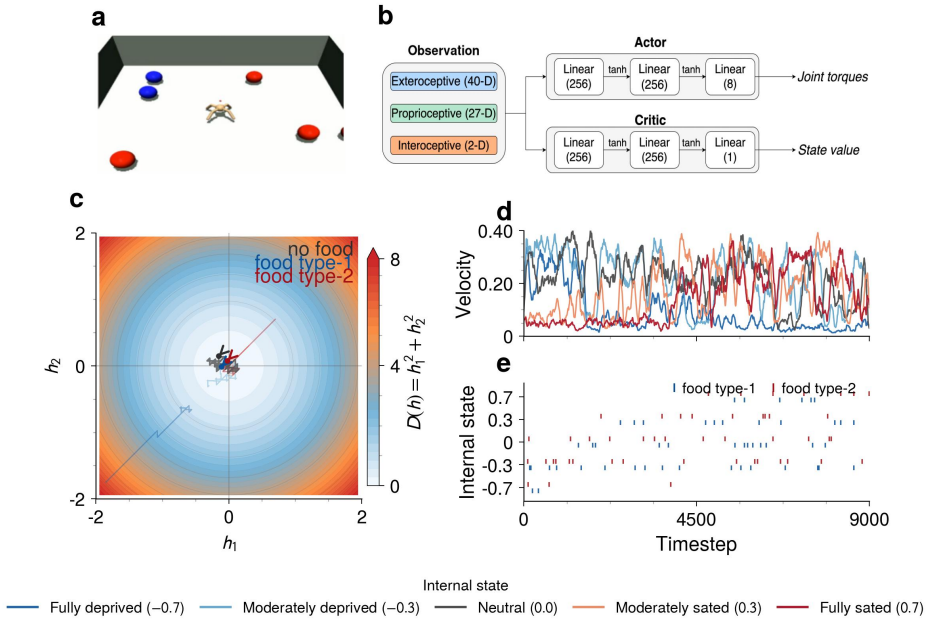


Figure 1: **Embodied homeostatic forager in the Two-Resource Problem.** (a) Quadrupedal Ant body within a continuous arena containing two food types (blue: type-1, red: type-2), each replenishing a distinct internal physiological variable. (b) Actor and critic MLPs, each consisting of two hidden layers with 256 tanh-activated units. (c) Drive landscape $D(H) = h_1^2 + h_2^2$ defined over the two internal variables. Arrows indicate one-step transitions resulting from representative food-consumption events (no food, type-1, type-2). (d) Locomotor velocity measured over a 9,000-step evaluation episode across all five internal states. (e) Food-consumption event raster for the same episode; ticks mark consumption of food type-1 and type-2.

food items, which respawn at random locations after consumption. Both internal variables decay at a constant metabolic rate of 1.5×10^{-4} per simulation step.

2.2 Agent: PPO Actor–Critic with a Homeostatic Reward

The agents employ a stochastic policy $\pi_\theta(a_t \mid o_t)$ and a value function $V_\phi(o_t)$, both implemented as two-layer MLPs (256 tanh units per layer). The actor represents a Beta distribution over the eight-dimensional joint-torque action space, while the critic outputs a scalar value. Observations o_t (69 dimensions) integrate exteroceptive information (40 dimensions: proximity to food in 20 angular bins for each type), proprioceptive data (27 dimensions: joint positions and velocities), and interoceptive signals (2 dimensions: internal variables h_1 and h_2). This architecture provides the policy with direct access to internal state and sensory input relevant to foraging (Figure 1b).

The agents are trained to minimize deviations from a physiological set point rather than maximize an extrinsic task reward. Let $H_t \in \mathbb{R}^n$ denote the agent's internal state vector at time t, where each entry represents a homeostatic variable, and let $H^\star$ denote the homeostatic set point. The drive is defined as follows: $D(H_t) = \|H_t - H^\star\|_2^2$. The per-step reward is defined as the temporal reduction in drive (Yoshida et al., 2024a), $r_t = c_{\text{drive}}\big(D(H_{t-1}) - D(H_t)\big) - c_{\text{ctrl}}\,\|u_t\|_2^2 - c_{\text{head}}\,\theta_{\text{head},t}$, with $c_{\text{drive}} = 100$, $c_{\text{ctrl}} = 10^{-3}$, and $c_{\text{head}} = 5 \times 10^{-3}$. Telescoping yields $\sum_{t=1}^{T} r_t = D(H_0) - D(H_T)$, showing that the cumulative reward over an episode equals the net cumulative drive reduction achieved by the agent. This cumulative drive reduction provides a quantitative behavioral measure of homeostatic stability (Figure 1c).

We optimize the policy with PPO for 1.5×10^8 environment steps. We train 10 independently initialized agents that could successfully collect food across different internal state conditions. We evaluate each network under 5 stochastic rollouts (with the environment seed held fixed), yielding 50 rollouts per condition; we assess cross-condition variability through stochastic rollouts (Appendix C). We also evaluated the impact of environmental layout across 21 realizations (Appendix L). Appendix B

lists the hyperparameters. Locomotor velocity profiles (Figure 1d) and food-consumption patterns across the internal states (Figure 1e) illustrate how the internal-state initialization reorganizes both movement and feeding rates.

2.3 Event-Centered Functional Clustering

We apply event-centered clustering to layer 1 activations to identify functional neural motifs in the actor network (also see Appendix M for analysis of layer 2). For each rollout, we identify the timesteps corresponding to food consumption events by detecting positive increments in either food-type counter. We aggregate consumption of either food type into a single event series. We then extract neural layer 1 activations from steps around each event ($T_w = 201$). For each internal state and rollout, we concatenate all event windows along the time axis. We then perform K-means clustering on the neuron-by-time matrix, treating each neuron as a sample with features corresponding to its concatenated event-window activations. We fix the number of clusters at $K{=}8$ throughout, based on silhouette, elbow, gap statistic, and BIC analyses, as detailed in Appendix D.

To evaluate whether similar functional repertoires emerge across independent agent rollouts, we project event-window activations from a representative rollout into a two-dimensional UMAP embedding (McInnes et al., 2018), computed separately for each internal state. Point colors indicate cluster assignments from the per-seed K-means pipeline to facilitate visual cross-seed correspondence (Figure 3a). Appendix D details the UMAP hyperparameters.

2.4 Cluster-Level Ablation and Behavioral-Effect Grouping

To determine each cluster's causal contribution to the agent's behavior, we perform a zero ablation. At each forward pass through the actor, we set the activations of all neurons in the target cluster to zero using a forward hook on the layer-1 activation outputs, while keeping all network weights unchanged. We run all combinations of five internal states and eight layer-1 clusters, for a total of forty ablation conditions. We repeat each condition across 50 rollouts and record it under identical environment initializations as the baseline rollouts. To bound the effect attributable to any specific cluster, we include random-population controls consisting of randomly selected layer-1 neurons. We use six fixed control sizes ($n_{\text{rand}} \in \{10, 20, 30, 40, 50, 60\}$), spanning the empirical per-cluster neuron count distribution (Appendix E). For each size and evaluation condition, we perform thirty independent random draws to estimate the null distribution. Because random-population ablation characterizes the network, we compute the random null once per network and reuse it across that network's five rollouts. We quantify the behavioral effect of each ablation condition as the signed difference $\Delta = m_{\text{abl}} - m_{\text{base}}$ between the metric value under ablation and its baseline counterpart, computed for each rollout and each metric.

Behavioral-effect grouping. We organize the high-dimensional ablation-effect signatures by hierarchically clustering the ablation-effect matrix (seed $\times$ origin state $\times$ cluster rows), with each ablation condition's 10-dimensional behavioral-effect vector serving as its feature representation. We apply three complementary hierarchical agglomerative clustering (HAC) dissimilarity measures: a pattern-cosine measure (sensitive to the relative direction of the effect across metrics), a pattern-correlation-shape measure (invariant to both location and scale), and a magnitude-aware Euclidean measure (jointly sensitive to pattern and magnitude). Appendix E provides the complete metric list and formal dissimilarity definitions. We apply HAC separately for each evaluation state and also pool across all five states. Each measure assigns seed–cluster pairs to behavioral-effect meta-clusters. We retain meta-clusters with *seed support* ≥ 7 out of 10 trained network seeds, and report the resulting motifs (Figure 4). To facilitate visual inspection of the effect-vector geometry, the ablation effects are embedded into two dimensions using UMAP (McInnes et al., 2018), separately for each internal state and pooled (Figure 4b). We report all aggregated quantities as mean $\pm$ standard error of the mean (SEM) across an $n = 50$ rollout population. Appendix F provides further discussion.

2.5 Conventional Reinforcement Learning Control

To determine whether the observed signatures depend on the homeostatic objective or solely on embodied foraging, we train a non-homeostatic control agent (conventional RL). The control agent retains the same body, environment, observation space (including the two interoceptive channels),

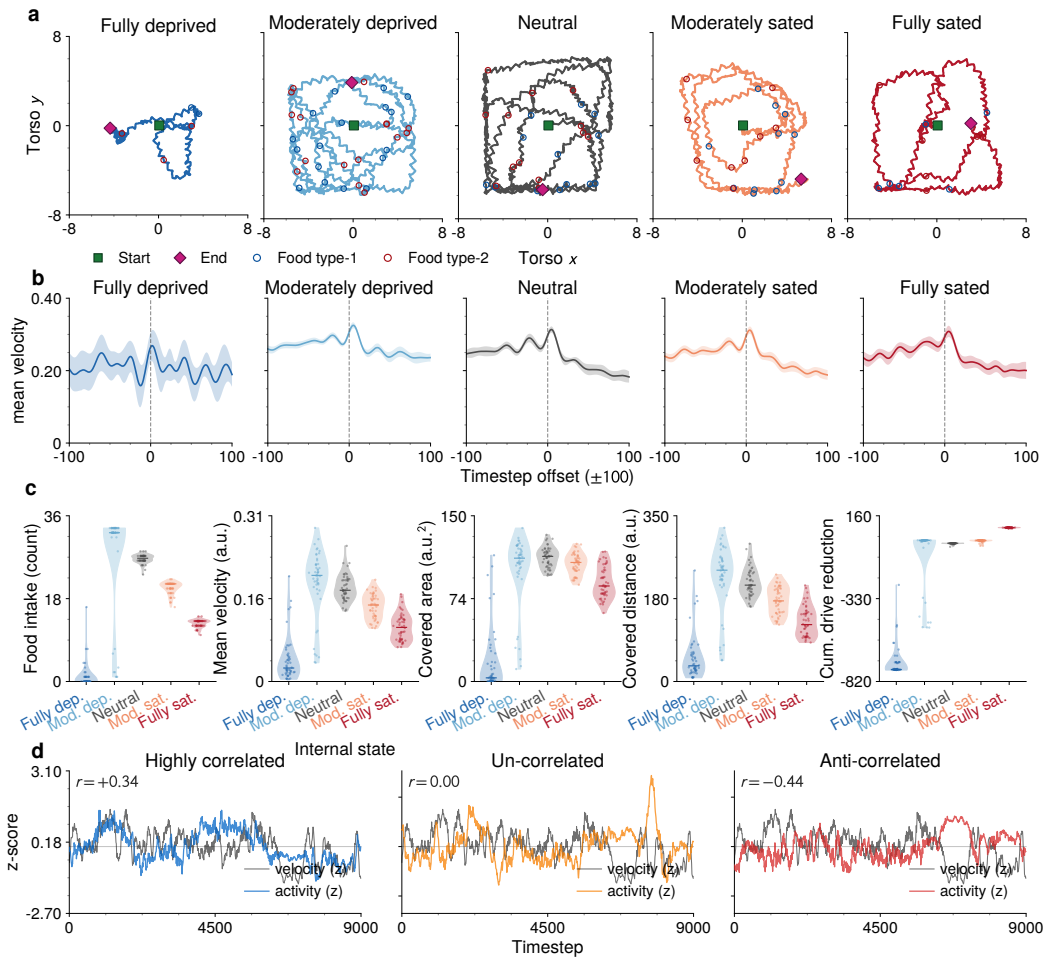


Figure 2: **Adaptive foraging is modulated by internal physiological state.** (a) Representative XY trajectories at five fixed internal states. A green square marks the starting point, and a pink diamond marks the endpoint. Open circles indicate food-consumption locations. One representative rollout is presented. (b) Event-aligned mean velocity within a ± 100-step window surrounding food-consumption events. (c) Distributions of behavioral metrics across the five internal states and all rollouts: food consumed (count), mean velocity, covered area, covered distance, and cumulative drive reduction. (d) Three example layer-1 actor neurons illustrating the range of velocity coupling: highly correlated (left), uncorrelated (middle), and anti-correlated (right) with z-scored locomotor velocity.

policy and value architectures, and PPO hyperparameters; only the reward function is altered. Specifically, the homeostatic drive-reduction reward is replaced with an extrinsic reward that combines a per-food intake signal and two regularization penalties:

$$r_{\text{Conventional RL},t} \;=\; c_{\text{food}}\,(n_{1,t}+n_{2,t}) \;-\; c_{\text{ctrl}}\,\|u_t\|_2^2 \;-\; c_{\text{head}}\,\theta_{\text{head},t}, \tag{1}$$

where $n_{k,t} \in \{0,1\}$ indicates whether food type k was consumed at step t, $\|u_t\|_2^2$ denotes the squared action norm, $\theta_{\text{head},t}$ represents the heading-angle deviation, and the coefficients are $c_{\text{food}} = 100$, $c_{\text{ctrl}} = 0.5$, and $c_{\text{head}} = 0.005$. The value of c_{ctrl} was selected from a sweep over $\{0.05, 0.10, 0.50\}$ (Appendix G). This control condition fixes the body, sensory input, and optimization algorithm, varying only the reward function. Therefore, we attribute any differences between the two agents to the homeostatic objective. We train the control agent using the same 10-network protocol and analyze it using the same extraction, clustering, and ablation pipelines described above (Appendix G).

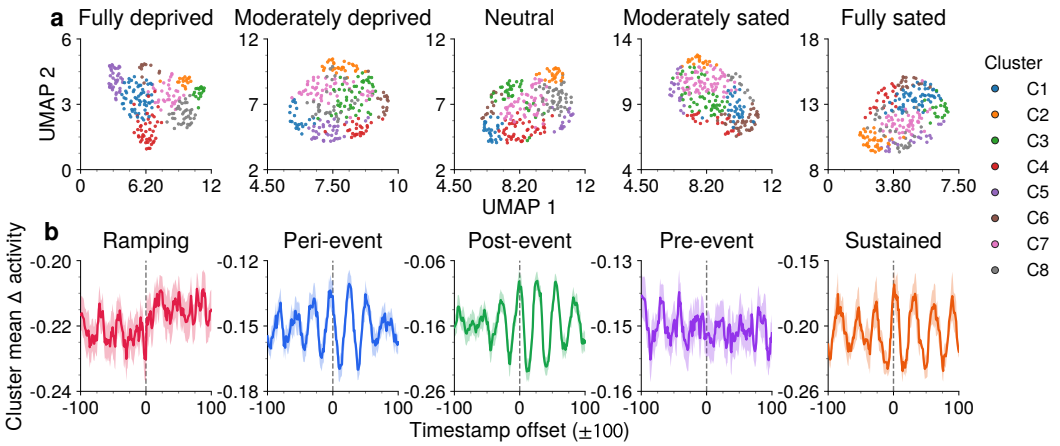


Figure 3: **Event-centered clustering identifies a canonical set of functional neural motifs.** (a) UMAP visualization of layer-1 actor-network event-window activations, displayed separately for each internal state. Each point corresponds to a neuron and is colored by K-means cluster label ($K = 8$; C1–C8). The analysis is based on a single representative rollout. (b) Cluster-mean activity (mean $\pm$ SEM across food-consumption events, $\pm$100-step window) for five representative functional motifs: ramping (monotonic pre-event rise), peri-event, post-event, pre-event, and sustained oscillation.

3 Results

3.1 Homeostatic objectives produce state-dependent foraging

Trained agents optimized for homeostatic objectives exhibit adaptive, state-dependent foraging dynamics that closely resemble those observed in rodents (Kamath et al., 2025). We assessed foraging behavior across five physiological states using five quantitative metrics and an event-aligned velocity profile, extracted from a population of $n = 50$ rollouts (10 networks $\times$ 5 rollouts) (Figure 2).

State-dependent modulation is both pronounced and non-monotonic (Figure 2a,c). Under the fully deprived condition ($h = -0.7$), the agents restrict their trajectory to a small region of the arena, consume significantly fewer food items, and maintain the lowest mean velocity. This low-activity, energy-conserving pattern is conceptually analogous to biological torpor (Hrvatin et al., 2020) (Appendix I). However, in most rollouts, agents collect more food, move faster, and traverse the arena more extensively at moderate deprivation than at full deprivation. Moreover, both food consumption and velocity reach their peak near moderate deprivation and decline sharply under full deprivation. Covered area and covered distance exhibit the same non-monotonic profile (mean $\pm$ SEM across rollouts; Figure 2c). Cumulative drive reduction is most negative at full deprivation and approaches zero at satiety. These results are replicable across different environmental layouts (Appendix L). Across all states, food consumption is consistently preceded by a stereotyped approach, with mean velocity increasing approximately 100 steps before each event and decelerating at the point of consumption (Figure 2b). This behavioral signature is embedded within broader state modulation. Layer-1 actor neurons encode this structure heterogeneously (velocity–activity correlations $r = +0.34$ to $r = -0.44$; Figure 2d), indicating the presence of functionally distinct subpopulations. We further show that interoceptive signals causally drive these behavioral signatures (Appendix H).

3.2 The actor network develops canonical functional neural motifs

The actor network represents physiological state by organizing activity into structured functional neural groups. Event-centered K-means clustering (K=8; layer 1) across $n = 50$-rollout population recovers a canonical repertoire of functional neural groups (Figure 3). A UMAP of event-window activations from a representative rollout (Figure 3a) shows distinct cluster regions within each physiological state. Quantitative analyses confirm consistent cluster assignments across independently trained networks (Appendix D).

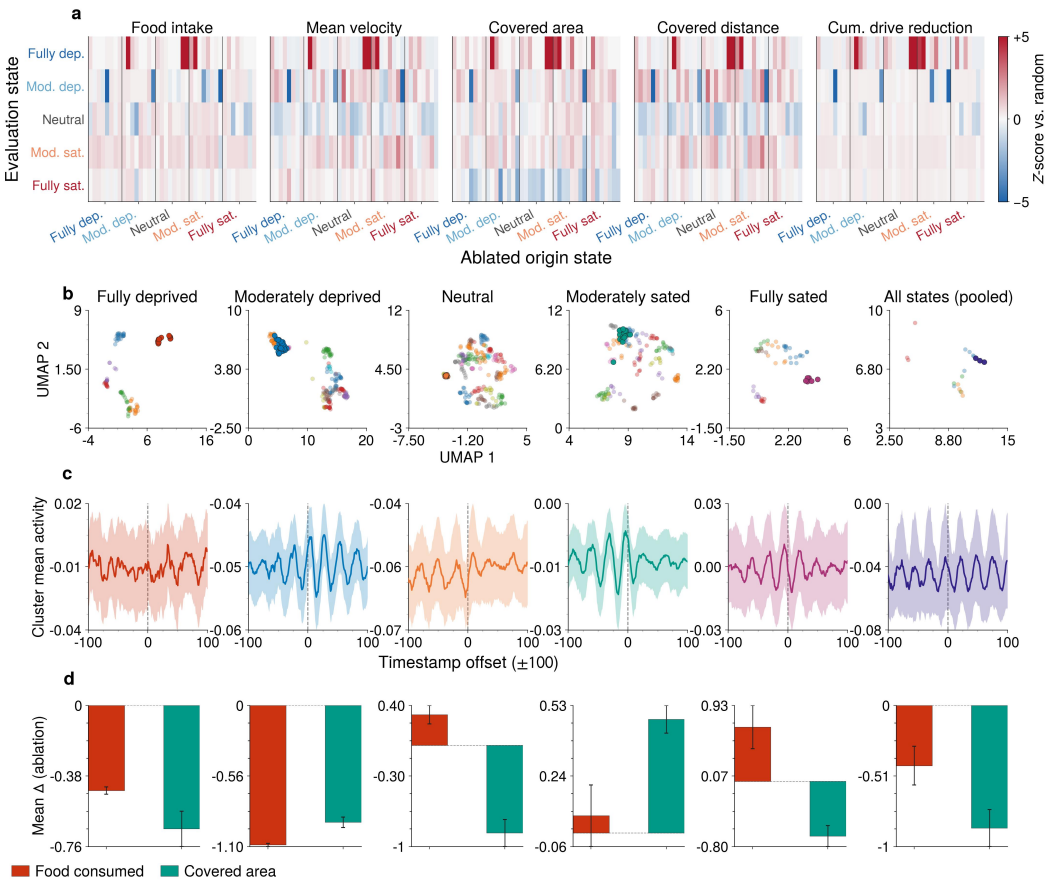


Figure 4: **Cluster-level ablation reveals functional motifs with distinct causal roles in foraging.** (a) Behavioral ablation heatmap. Cell color represents the Z-score of the cluster ablation effect relative to the random-population control distribution. (b) UMAP visualization of Δ-from-baseline behavioral-effect vectors for identified meta-clusters, computed for each evaluation internal state and pooled. (c) Event-centered cluster-mean activity (mean $\pm$ SEM, $\pm$100-step window) for representative members of each meta-cluster (also see Figure 3b for the corresponding figure in the HRL case). (d) Mean $\pm$ SEM behavioral $\pm$ (ablation minus baseline) for two primary metrics: food consumed and covered area, for each representative meta-cluster in (c).

Among the K=8 clusters per state, five distinct waveform motifs emerge, each defined by temporal dynamics in a $\pm$100-step window before and after food-consumption events. All clusters show a shared oscillation frequency around 0.03 cycles/step, consistent with oscillations in joint movement and velocity. However, the amplitude of oscillations within each cluster followed characteristic waveforms: 1) *ramping*: exhibit a monotonic rise that peak around $t = 0$. 2) *peri-event*: show almost symmetric increased amplitude around the event. 3) *post-event* and 4) *pre-event*: show elevated amplitude after/before food consumption. 5) *sustained*: roughly maintain constant amplitude throughout the window. Figure 3b shows exemplars (mean $\pm$ SEM), and Appendix D confirms cross-network stability.

3.3 Cluster ablation reveals functional groups with distinct causal roles

To establish causal necessity, we ablate up to 40 cluster conditions per rollout (5 states multiplied by K=8 clusters) across a population of $n = 50$ rollouts and compare each condition against random-population controls ($n_{\text{rand}} \in \{10, 20, 30, 40, 50, 60\}$; 30 independent draws per size, computed once per network; Figure 4). Significant behavioral effects are concentrated at specific origin and evaluation state combinations, with the largest effects observed in fully and moderately deprived

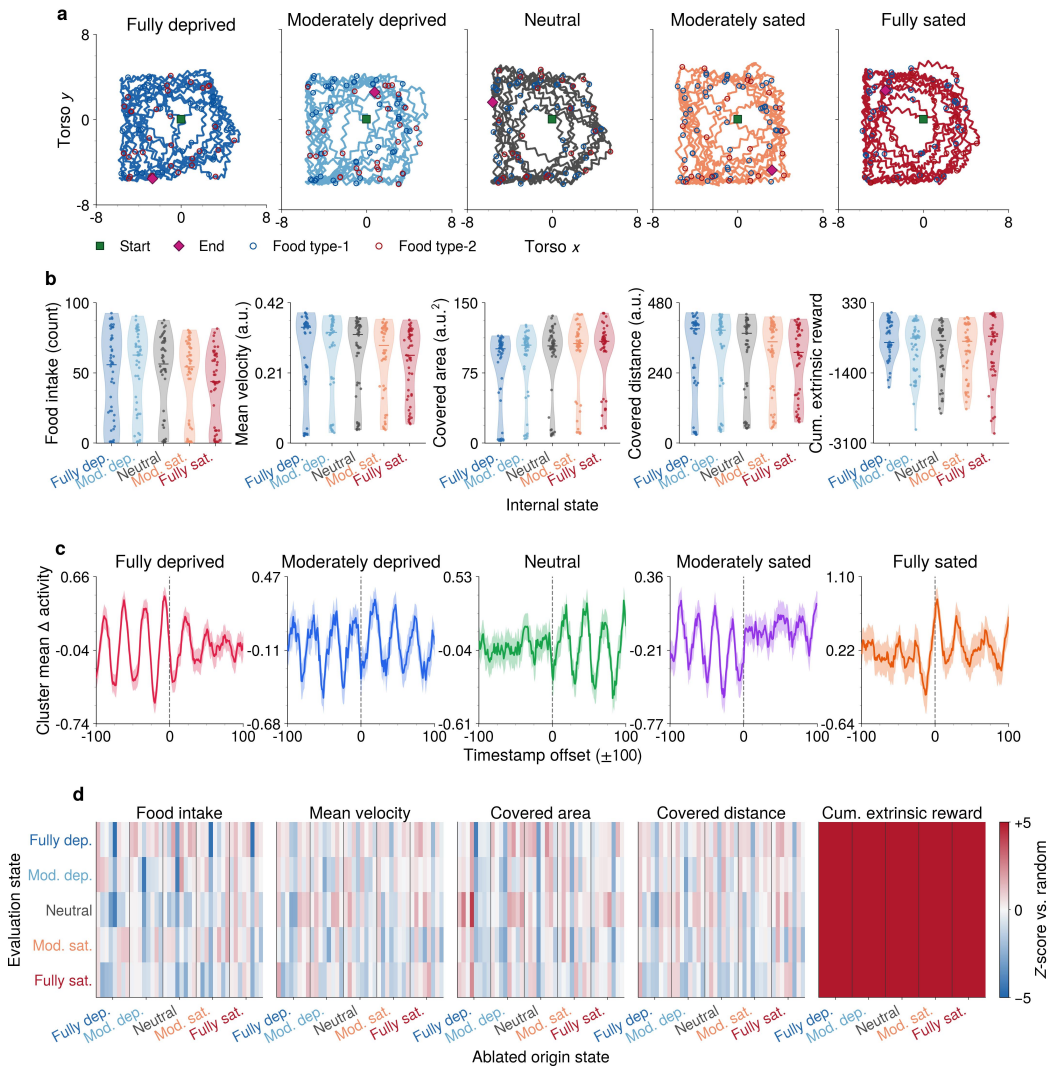


Figure 5: **Conventional RL control fails to reproduce behavioral and neural patterns seen with HRL.** (a) XY trajectories of the conventional RL agent across all five internal states. Each state shows uniform full-arena coverage, with no evidence of a low-activity, torpor-like state. (b) Distributions of behavioral metrics across internal states. Overlapping distributions indicate state-dependent behavioral modulation relative to the HRL agent (Figure 2c). (c) Cluster-mean activity (mean ± SEM, ±100-step window) for five representative conventional RL layer-1 functional groups. (d) Behavioral heatmap of conventional RL ablation (same format as Figure 4a). Uniformly positive Z-scores for cumulative extrinsic reward indicate systematic motor-effort reduction under cluster silencing, rather than cluster-specific foraging improvement.

agents (Figure 4a). The behavioral-effect UMAP demonstrates that meta-cluster ablations are distinct from both random controls and excluded ablations across all physiological states, thereby establishing functional specificity (Figure 4b; full behavioral-effect profiles in Appendix J).

Our results reveal two distinct causal signatures (Figure 4c,d; full ten-metric profiles are provided in Figure 8). The effects shown in Figure 4d are z-scored relative to all clustered ablations of the same representation (mean ± SEM across members). Each neural motif alters behavior: covered area deviates by more than five standard errors in all six columns, and food consumption deviates by more than two in all but the pre-event column. During deprivation, ablating the sustained cluster at full deprivation and the post-event cluster at moderate deprivation reduces both food consumption and covered area (food -0.46 ± 0.02 and -1.10 ± 0.01; area -0.66 ± 0.09 and -0.92 ± 0.04;

columns 1–2). The sustained cluster, when pooled across states, exhibits a similar concordant decrease (column 6). Under neutral and sated conditions, the two metrics diverge (columns 3–5). Specifically, ablating the ramping cluster at neutral increases food consumption ($+0.30 \pm 0.09$) but decreases covered area (-0.87 ± 0.13). Ablating the pre-event cluster at moderate satiety does not alter food consumption ($+0.07 \pm 0.13$) but increases covered area ($+0.47 \pm 0.06$), indicating greater space coverage without additional consumption. Ablating the peri-event cluster at full satiety increases food consumption ($+0.66 \pm 0.27$) while decreasing covered area (-0.67 ± 0.13).

3.4 Homeostatic objective is key for neural and behavioral signatures

To determine whether the previously identified signatures require the homeostatic objective, we analyzed the conventional RL control agent using the same 10-network protocol and identical extraction, clustering, and ablation pipelines (Figure 5). The control agents acquire an effective foraging policy, but their behavior is largely independent of internal state. All trajectories demonstrate full-arena coverage without a torpor-like state at full deprivation (compare left-most panels of Figure 5a and Figure 2a). Likewise, the distributions of behavioral metrics are broad and largely overlapping across all conditions (Figure 5b vs. Figure 2c). Moreover, RL agents reach their highest food intake at full deprivation, whereas HRL agents peak at moderate deprivation and forage minimally under full deprivation to conserve energy (torpor-like behavior). These results reflect the RL reward structure that incentivizes acquisition regardless of physiological state.

Cluster ablation in the control agent produces substantially weaker behavioral effects compared to HRL agents. When we apply the same meta-clustering procedure in our control population, none of the meta-clusters in control (RL) agents span seven or more seeds, whereas 15 such meta-clusters are observed for HRL. Furthermore, ablation in control RL agents reflects a general reduced motor-effort penalty (Equation 1) rather than cluster-specific changes in foraging behaviors (food intake, velocity, and coverage) as they show no cluster-specific effects. However, HRL agents develop neural clusters with distinct functional roles in behavioral control (compare Figure 4a with Figure 5d and Appendix K). These results suggest that the state-dependent behavioral repertoire and the recurrence of functional groups across networks are determined by the homeostatic objective rather than by embodied foraging alone. This underscores the significance of the biological principle of *homeostasis* for adaptive agents (Weber et al., 2025).

4 Discussion

The results demonstrate that training embodied deep RL agents under a homeostatic objective produces structured, state-dependent behavior and characteristic neural organization. The agent exhibits distinct behavioral regimes across physiological states, including reduced activity under extreme deprivation and increased exploration as physiological need increases. At the neural level, event-centered clustering reveals diverse functional motifs with distinct temporal response profiles recurring across networks. Ablation of these clusters demonstrates their causal roles in shaping behavior. These findings suggest that a homeostatic objective induces structured internal representations that couple physiological state to action selection. Notably, these experiments provide insights into a potential neural substrate for foraging decisions under realistic homeostatic constraints, which are challenging to manipulate in animal experiments. In contrast, the conventional RL agent lacked state-dependent behavioral modulation. Furthermore, its functional groups show no or weaker cluster-specific causal effects, highlighting the importance of homeostatic principles for neural and behavioral organization in adaptive biological agents (see Appendix N for extended context). More broadly, the analysis pipeline (combining behavioral quantification, event-centered clustering, and causal ablation) provides a general framework for mechanistic analysis of embodied agents with internal objectives.

Limitations. By holding body and environmental conditions constant, the analysis isolates the homeostatic objective's impact. However, the roles of environmental complexity remain unresolved. Although we measure compensatory responses, zero ablation yields only a coarse estimate of causal contributions. Lastly, other architectures, for instance, recurrent neural networks trained to comparable performance on the same task can implement it through substantially different internal dynamics (Maheswaranathan et al., 2019).

AI USE STATEMENT

Generative AI tools were used for manuscript editing, including grammar, spelling, and word choice, as well as for visualization scripting and analysis. These tools were not part of the core scientific methodology and did not influence this study's scientific results or conclusions. The authors reviewed all AI-assisted work and assume responsibility for the paper's final content, including text, claims, and artifacts produced with generative AI.

ETHICS STATEMENT

All experiments were conducted using simulated agents within a MuJoCo physics environment. The study does not involve human subjects, sensitive personal data, or real-world deployment, and therefore raises no privacy, fairness, or dual-use concerns. This research is foundational and focuses on the internal organization of simulated embodied agents trained with a homeostatic objective. No evident pathway to harmful applications exists, and we anticipate no direct negative societal impacts.

REPRODUCIBILITY STATEMENT

All information required to reproduce the main results is provided in the paper and appendix: PPO hyperparameters (Appendix B), agent architecture and homeostatic reward function (§2.2), environment parameters including food counts and metabolic decay (§2.1), the evaluation protocol with fixed seeds (Appendix C), event-centered clustering implementation and the choice of K (Appendix D), the ablation protocol including random-control sizes, draw counts, and the meta-cluster agreement and network-support criteria (Appendix E), and the conventional RL reward specification with all coefficient values (§2.5, Appendix G). Aggregate behavioral quantities are reported as mean $\pm$ SEM across the $n = 50$ rollout population (10 networks $\times$ 5 rollouts; Appendix F); event-aligned velocity profiles use mean $\pm$ 95% CI from 1,000-iteration bootstrap resampling; individual cluster-ablation effects are reported as Z-scores relative to a random-population null distribution, whereas meta-cluster effect bars are z-scored against all grouped clusters of the same representation. Compute resources are detailed in Appendix B. Code and trained agent checkpoints will be released upon acceptance.

REFERENCES

Diego Aldarondo, Josh Merel, Jesse D. Marshall, et al. A virtual rodent predicts the structure of neural activity across behaviours. *Nature*, 632:594–602, 2024. doi: 10.1038/s41586-024-07633-4.

Gordon J. Berman, William Bialek, and Joshua W. Shaevitz. Predictability and hierarchy in Drosophila behavior. *Proceedings of the National Academy of Sciences*, 113(42):11943–11948, 2016. doi: 10.1073/pnas.1607601113.

Sandeep Robert Datta, David J. Anderson, Kristin Branson, Pietro Perona, and Andrew Leifer. Computational neuroethology: A call to action. *Neuron*, 104(1):11–24, 2019. doi: 10.1016/j.neuron.2019.09.038.

Peter Dayan. "liking" as an early and editable draft of long-run affective value. *PLOS Biology*, 20(1):e3001476, 2022. doi: 10.1371/journal.pbio.3001476.

Peter Dayan and Nathaniel D Daw. Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience*, 8(4):429–453, 2008.

Jennifer D. Deem, Chelsea L. Faber, and Gregory J. Morton. AgRP neurons: Regulators of feeding, energy expenditure, and behavior. *The FEBS Journal*, 289(8):2362–2381, 2022. doi: 10.1111/febs.16176.

Jenelle Feather, Meenakshi Khosla, N. Apurva Ratan Murty, and Aran Nayebi. Brain-model evaluations need the NeuroAI Turing test. *arXiv preprint arXiv:2502.16238*, 2025.

Laura L Grima, Hannah Haberkern, Rishika Mohanta, Mai M Morimoto, Adithya E Rajagopalan, and Emma V Scholey. Foraging as an ethological framework for neuroscience. *Trends in Neurosciences*, 2025.

Lewis G Halsey. Terrestrial movement energetics: current knowledge and its application to the optimising animal. *Journal of Experimental Biology*, 219(10):1424–1431, 2016.

Sinisa Hrvatin, Senmiao Sun, Oren F. Wilcox, Hanqi Yao, Aurora J. Lavin-Peter, Marcelo Cicconet, Elena G. Assad, Michaela E. Palmer, Sage Aronson, Alexander S. Banks, Eric C. Griffith, and Michael E. Greenberg. Neurons that regulate mouse torpor. *Nature*, 583(7814):115–121, 2020. doi: 10.1038/s41586-020-2387-5.

Tarun Kamath, Bart Lodder, Eliana Bilsel, Isobel Green, Rochelin Dalangin, Michelle Raghubardayal, Wengang Wang, Paolo Capelli, Jessie Legister, Joshua Timmins, et al. Hunger modulates exploration through suppression of dopamine signaling in the tail of the striatum. *Neuron*, 113(23):4055–4068, 2025.

Mehdi Keramati and Boris Gutkin. Homeostatic reinforcement learning for integrating reward collection and physiological stability. *eLife*, 3:e04811, 2014. doi: 10.7554/eLife.04811.

Niru Maheswaranathan, Alex H. Williams, Matthew D. Golub, Surya Ganguli, and David Sussillo. Universality and individuality in neural dynamics across large populations of recurrent networks. *arXiv preprint arXiv:1907.08549*, 2019. URL https://arxiv.org/abs/1907.08549.

Raaghav Malik, Satpreet H. Singh, Sonja Johnson-Yu, Nathan Wu, Roy Harpaz, Florian Engert, and Kanaka Rajan. Dissecting larval zebrafish hunting using deep reinforcement learning trained RNN agents. *arXiv preprint arXiv:2510.03699*, 2025.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Josh Merel, Diego Aldarondo, Jesse Marshall, Yuval Tassa, Greg Wayne, and Bence Ölveczky. Deep neuroethology of a virtual rodent. *arXiv*, 2019. doi: 10.48550/arxiv.1911.09451.

Richard Meyes, Melanie Lu, Constantin Waubert de Puiseau, and Tobias Meisen. Ablation studies in artificial neural networks. *arXiv preprint arXiv:1901.08644*, 2019.

Frederike H. Petzschner, Sarah N. Garfinkel, Martin P. Paulus, Christof Koch, and Sahib S. Khalsa. Computational models of interoception and body regulation. *Trends in Neurosciences*, 44(1):63–76, 2021. doi: 10.1016/j.tins.2020.09.012.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Riley Simmons-Edler, Ryan P. Badman, Felix Baastad Berg, Raymond Chua, John J. Vastola, Joshua Lunger, William Qian, and Kanaka Rajan. Deep RL needs deep behavior analysis: Exploring implicit planning by model-free agents in open-ended environments. In *Advances in Neural Information Processing Systems*, 2025. NeurIPS 2025 Poster.

Satpreet H. Singh, Sonja Johnson-Yu, Zhouyang Lu, Aaron Walsman, Federico Pedraja, Denis Turcu, Pratyusha Sharma, Naomi Saphra, Nathaniel B. Sawtell, and Kanaka Rajan. Understanding electro-communication and electro-sensing in weakly electric fish using multi-agent deep reinforcement learning. *arXiv preprint arXiv:2511.08436*, 2025.

David W Stephens and John R Krebs. *Foraging theory*, volume 6. Princeton university press, 1986.

Richard S Sutton, Andrew G Barto, and Andrew Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5026–5033. IEEE, 2012. doi: 10.1109/IROS.2012.6386109.

Tristan Trim and Triston Grayston. Mechanistic interpretability of reinforcement learning agents. *arXiv preprint arXiv:2411.00867*, 2024.

Roman Vaxenburg, Igor Siwanowicz, Josh Merel, Alice A Robie, Carmen Morrow, Guido Novati, Zinovia Stefanidi, Gert-Jan Both, Gwyneth M Card, Michael B Reiser, et al. Whole-body physics simulation of fruit fly locomotion. *Nature*, 643(8074):1312–1320, 2025.

Sibo Wang-Chen, Victor A. Stimpfling, Thomas K. C. Lam, et al. Neuromechfly v2: simulating embodied sensorimotor control in adult drosophila. *Nature Methods*, 21:2353–2362, 2024. doi: 10.1038/s41592-024-02497-y.

Lilian A. Weber, Debbie M. Yee, Dana M. Small, and Frederike H. Petzschner. The interoceptive origin of reinforcement learning. *Trends Cogn Sci.*, 29(9):840–854, September 2025. ISSN 13646613. doi: 10.1016/j.tics.2025.05.008. URL https://linkinghub.elsevier.com/retrieve/pii/S1364661325001202.

Naoto Yoshida, Etsushi Arikawa, Hoshinori Kanazawa, and Yasuo Kuniyoshi. Modeling long-term nutritional behaviors using deep homeostatic reinforcement learning. *PNAS Nexus*, 3(12):pgae540, 2024a. doi: 10.1093/pnasnexus/pgae540.

Naoto Yoshida, Tatsuya Daikoku, Yukie Nagai, and Yasuo Kuniyoshi. Emergence of integrated behaviors through direct optimization for homeostasis. *Neural Networks*, 177:106379, 2024b.

Kevin Zakka, Baruch Tabanpour, Qiayuan Liao, Mustafa Haiderbhai, Samuel Holt, Jing Yuan Luo, Arthur Allshire, Erik Frey, Koushil Sreenath, Lueder A Kahrs, et al. Mujoco playground. *arXiv preprint arXiv:2502.08844*, 2025.

Liang Zhang, Justin Lieffers, and Adarsh Pyarelal. Exploring semantic clustering in deep reinforcement learning for video games. *arXiv preprint arXiv:2409.17411*, 2024.

A Background Supplement

A.1 Homeostatic Reinforcement Learning as Drive Reduction

This approach is based on HRL, as formalized by (Keramati & Gutkin, 2014). In this framework, an agent learns motivated behavior by defining primary reward as the reduction of deviation from a physiological set point. A key formal result is that reward maximization is mathematically equivalent to long-term physiological stability, thereby formalizing a modern computational version of Hull's classical drive-reduction theory.

Dayan (2022) extends this connection to affective value, proposing that immediate hedonic "liking" serves as an early, modifiable draft of long-term utility. This potential-based shaping signal accelerates temporal credit assignment while remaining consistent with the agent's physiological objectives. More broadly, Petzschner et al. (2021) situate homeostatic computation within a unified framework encompassing interoception, body regulation, and predictive control across health and disease.

Recent research has extended HRL to high-dimensional continuous motor control. Yoshida et al. (2024a) demonstrate that HRL agents trained solely on interoceptive reward can reproduce long-term foraging characteristics observed in real animals. Additionally, Yoshida et al. (2024b) show that integrated, multi-objective behaviors emerge directly from optimizing for homeostasis in embodied agents.

A.2 Mechanistic Analysis of Deep RL Policies

Neuroscience-inspired ablation provides a practical method for examining knowledge representations and structural robustness in artificial neural networks (Meyes et al., 2019). Trim & Grayston (2024) analyze procedurally trained agents and demonstrate that individual neurons in the policy network selectively respond to environmental features. Targeted ablation reveals these neurons' specific contributions to decision-making. In a complementary approach, (Zhang et al., 2024) embed a semantic clustering module within a deep reinforcement learning (DRL) architecture to expose the network's internal organization during training.

From a NeuroAI perspective, Feather et al. (2025) argue that current brain-likeness benchmarks, which rely on representational similarity metrics, are insufficient. They propose a "NeuroAI Turing Test" that requires convergence in both behavior and internal representations.

A.3 Neurotechnological Analysis of Deep RL Agents

Recent research on DRL agents has incorporated tools from neuroscience and ethology, moving beyond reliance on reward curves alone. This methodology draws inspiration from behavioral state discovery in real animals (Berman et al., 2016) and seeks to integrate computational neurotechnology (Datta et al., 2019). (Simmons-Edler et al., 2025) employ behavioral segmentation and recurrent state decoding to reveal planning-like behavior in model-free RNN agents during foraging tasks. Similarly, Malik et al. (2025) dissect larval zebrafish hunting with RNN agents, and Singh et al. (2025) investigate electro-communication in a multi-agent setting. These studies demonstrate that recurrent agents can reproduce biologically plausible behavioral signatures, enabling meaningful comparison to real animal data.

B Training Hyperparameters

Table 1 presents the PPO hyperparameters used to train the homeostatic agent (subsection 2.2). We use the same hyperparameters for the conventional RL control (subsection 2.5), so the only difference between the two conditions is the reward function.

Compute resources. We ran all training on a CPU cluster (8 CPU cores, 64 GB RAM per job; no GPU). Each complete training run of 1.5×10^8 environment steps required approximately 77 hours of wall-clock time (range 70–83. We trained the 10 networks per condition as independent jobs. We distributed cluster-ablation rollouts as separate jobs, with up to 40 conditions per rollout and 50 rollouts per condition, each with a 24-hour wall-clock limit.

Table 1: PPO hyperparameters used for homeostatic agent training. Identical settings apply to the conventional RL control; only the reward function varies.

Hyperparameter	Value
Total environment steps	1.5×10^8
Learning rate (Adam)	3×10^{-4}
Rollout horizon	3×10^5 steps
Number of parallel envs	1
Number of minibatches	6
Update epochs per rollout	30
Discount factor γ	0.99
generalized advantage estimation (GAE) coefficient λ	0.95
Clip range	0.3
Value-loss coefficient	0.5
Entropy coefficient	0.001
Max gradient norm	0.5
Hidden layer width	256
Hidden activations	tanh
Action distribution	Beta (per joint)
Training seeds	10 independent runs
Max episode steps (training)	6×10^4
Test rollout interval	every 10 updates
Environment id	`SmallLowGearAntTRP-v0`

C Data Extraction and Behavioral Quantification

After training, we fix each policy and evaluate the trained agent in the same environment across each of the five predetermined internal states. The environment layout remains constant (`env.seed` $= 100$), while only the action-sampling random number generator differs across the five policy seeds (100, 42, 142, 242, 342) for each network. Each rollout consists of $T = 9{,}000$ steps, resulting in an analysis dataset of 5 states $\times$ 10 networks $\times$ 5 rollouts $\times$ 9,000 steps per experimental condition. At each step, we record position, velocity, action, internal state, food-consumption events, and the activations of all units in both the actor and critic layers.

For each rollout, ten behavioral metrics are extracted: total food consumed, mean velocity, covered area (defined as the area of the convex hull of the trajectory), covered distance (total Euclidean distance traversed), cumulative drive reduction ($\sum_t [D(H_{t-1}) - D(H_t)]$; §2.2), event-aligned mean velocity (mean velocity within a ± 100-step window around each food-consumption event), path tortuosity (path length divided by net start-to-end displacement), convex-hull growth rate (mean per-step gain in hull area, measured at 500-step intervals), drive variance (variance of the internal-need magnitude $|h_t|$ over the rollout; defined for both agents, since the internal variables evolve in the control even though they are not rewarded), and mean action magnitude $\|u_t\|_2^2$ averaged over time (used as a control-effort readout for subsequent ablation analyses; Figure 2). For each rollout, we also compute the Pearson correlation between mean velocity and the magnitude of internal need $|h|$.

Cumulative drive reduction and the terminal condition. The drive term of the homeostatic reward (§2.2) telescopes exactly, $\sum_{t=1}^{T} [D(H_{t-1}) - D(H_t)] = D(H_0) - D(H_T)$, and the evaluation rollouts run their full 9,000-step horizon without termination, so this identity holds over the whole recorded episode. Two qualifications apply. First, the episode return is not the cumulative drive reduction, because the implemented reward scales the drive term and adds motor and posture penalties that do not telescope; we therefore report the drive term itself. Second, training does enforce a terminal condition on the internal state, ending an episode once the deficit exceeds the survival boundary ($|h_i| > 1$); at evaluation, this boundary is recorded as an event but does not end the rollout.

D CLUSTERING SUPPLEMENTARY

Clustering implementation. We normalize event-window activations per neuron using `StandardScaler` (zero mean, unit variance across the time dimension) before clustering. K-means is initialized with K-means++, run with `n_init` $= 10$ independent restarts, `max_iter` $=$ 300, and `random_state` $= 42$. Event windows that fall within $W = 100$ steps of an episode boundary are edge-padded to the full $T_w = 201$ sample length before clustering.

Choice of number of clusters. $K = 8$ was selected for layer-1 event-centered clustering based on four data-driven criteria evaluated over $K \in \{3, \ldots, 12\}$: silhouette score, within-cluster sum-of-squares elbow (second derivative of inertia), gap statistic ($n_{\text{ref}} = 5$ uniform reference samplings), and BIC of a Gaussian Mixture Model. We determined the final K for each rollout as the median of the four criteria; $K = 8$ represented the consensus across rollouts, and all five values of K merged functionally distinct groups, while larger values produced singleton clusters that did not recur across rollouts. $K = 8$ is not favored by every individual clustering-quality criterion. The selection also considers the interpretability of the resulting waveform motifs and the specificity of their causal effects, both detailed below.

Cross-network consistency. K-means cluster identities (C1–C8) are arbitrary within each rollout and are not aligned across rollouts. Correspondence therefore emerges only indirectly, from the meta-clustering of ablation-effect vectors (Appendix E): a meta-cluster spanning many networks indicates that groups with similar ablation effects recur. Because each cluster is tagged with its trained network and rollout, we count recurrence over distinct trained networks, not over rollouts of a single network, and interpret it relative to the conventional-RL control analyzed with the same procedure. The single-rollout UMAP (Figure 3a) qualitatively illustrates the within-rollout organization on which this aggregation operates.

UMAP hyperparameters. The two-dimensional UMAP embedding (Figure 3a) is computed from a single representative rollout, separately for each of the five internal states (one embedding per state; no pooling across states). Hyperparameters: `n_neighbors` $= \min(25, \max(5, n_{\text{assigned}} - 1))$ where n_{assigned} is the number of neurons with a valid cluster label; `min_dist` $= 0.15$; distance metric $=$ Euclidean; `random_state` $= 42$.

E ABLATION SUPPLEMENTARY

Random-population control sizes. The six random-population control sizes ($n_{\text{rand}} \in \{10, 20, 30, 40, 50, 60\}$) span the empirical per-cluster neuron count distribution in layer 1 (median ≈ 32, range $\approx$ 12–60). This design tests the hypothesis that the observed per-cluster effects could result from removing any sufficiently large random subset of neurons. For each control size and evaluation condition, 30 independent random draws are performed (seeded deterministically as *seed* $+ 100{,}000 \cdot$ *layer_offset* $+ 1{,}000 \cdot$ *draw_index*). This procedure yields a per-condition null distribution of 30 samples per size, which we use to compare against cluster ablation effects. We implement zero ablation using a PyTorch forward hook registered on the layer-1 activation outputs. We set target neuron activations to 0 at each forward pass, without modifying any network weights.

Behavioral-effect dissimilarity measures. Let $\mathbf{e}_i \in \mathbb{R}^{10}$ denote the behavioral-effect vector for ablation condition i. Each entry represents a signed difference $\Delta = m_{\text{abl}} - m_{\text{base}}$, z-scored across conditions, for the following ten metrics: total food consumed (Δn_{food}), mean locomotor velocity ($\Delta \bar{v}$), event-aligned mean velocity ($\Delta \bar{v}_{\text{event}}$), covered area ($\Delta$area), covered distance (Δd), path tortuosity ($\Delta \tau$), convex-hull growth rate ($\Delta \dot{\text{area}}$), cumulative reward (Δr_{cum}), drive variance ($\Delta \sigma_H^2$), and mean action magnitude ($\Delta \|u\|_2^2$). For the conventional RL control, the drive-variance entry is undefined. Cross-agent clustering therefore uses the remaining 9-dimensional subvector, and we recompute the dissimilarity matrices accordingly.

The three HAC dissimilarity measures used in subsection 2.4 are as follows:

1. *Pattern-cosine*: $1 - \cos(\mathbf{e}_i, \mathbf{e}_j)$, which is sensitive to the relative direction of the effect across metrics but invariant to magnitude.

2. *Pattern-correlation-shape*: $1 - \rho(\mathbf{e}_i, \mathbf{e}_j)$, where ρ is the Pearson correlation across metrics. This measure is invariant to both location and scale.

3. *Pattern-plus-magnitude Euclidean*: $\|\mathbf{e}_i - \mathbf{e}_j\|_2$, which is jointly sensitive to both pattern and magnitude.

Pattern-cosine and pattern-correlation-shape measures use average-linkage HAC. The pattern-plus-magnitude Euclidean measure uses Ward linkage, which minimizes within-cluster variance.

Meta-cluster Retention and Combination Across Methods. For each of the three dissimilarity measures, we apply HAC to the behavioral-effect vectors, separately for each evaluation state and for all states pooled. We identify a cluster by its (network, rollout, origin state, cluster id) tuple. Link two clusters only if all three measures place them in the same HAC group; meta-clusters are the connected components of the resulting agreement graph, with singletons discarded. We retain a meta-cluster if its network support, the number of distinct trained networks contributing at least one member, is at least 7. We report the retained members in Figure 4b–d.

The two-of-three rule is derived from empirical observation. When agreement is required from only one measure, or equivalently, when we take the union of the three measures, groups merge transitively through single shared conditions. This process produces a small number of large components with flat mean event-aligned waveforms. Conversely, requiring agreement across all three measures yields almost no components exceeding the support threshold. The two-of-three rule provides an intermediate criterion, ensuring that components are both well supported and maintain waveform coherence. This method replaces the union rule employed in a previous version of the analysis.

F Statistical Analysis

Unless otherwise specified, we report all aggregated quantities as the mean $\pm$ SEM across the 50-rollout population (10 networks $\times$ 5 rollouts). We report per-rollout Pearson correlations between mean velocity and $|h|$ with 95% confidence intervals, estimated using $1{,}000$-iteration bootstrap resampling.

When a quantity is tested against internal state, each rollout contributes one value at each of the five states, resulting in a paired, repeated-measures design. Accordingly, the appropriate omnibus test is the Friedman test. For each (agent, protocol), we correct the five metrics' p values using the Benjamini–Hochberg false discovery rate procedure. The χ^2 statistic (df $= 4$) and corrected p values are reported throughout. Rollouts of the same network are not independent, so these tests and the SEM above quantify rollout-level variability, not variability across networks.

Ablation effects of individual clusters (Figure 4a, Figure 5d) are z-scored against the pooled effects of random-neuron ablations (populations of 10–60 neurons) from the same rollout and clipped at ± 5. The behavioral-effect bars in Figure 4d are z-scored against all grouped clusters of the same representation and show the mean $\pm$ SEM across the members of each meta-cluster, as defined by the three-of-three agreement rule (Appendix E).

G Non-Homeostatic Control Sweep

The motor-energy penalty coefficient $c_{extctrl}$ for conventional RL control was determined through a brief parameter sweep over $c_{extctrl} \in \{0.05, 0.10, 0.50\}$ using a reduced training horizon of 3×10^7 environment steps and a single training seed (seed 100). The selection criteria were: (i) non-zero food consumed across all five evaluation internal states during post-training rollout, and (ii) mean action magnitude $\|u_t\|_2^2$ comparable to that of the homeostatic agent. The optimal value was $c_{extctrl} = 0.50$, as it was the only coefficient that yielded non-zero food consumed across all physiological states while maintaining comparable action magnitude. We then trained the full conventional RL agent for 1.5×10^8 steps using the hyperparameters listed in Table 1.

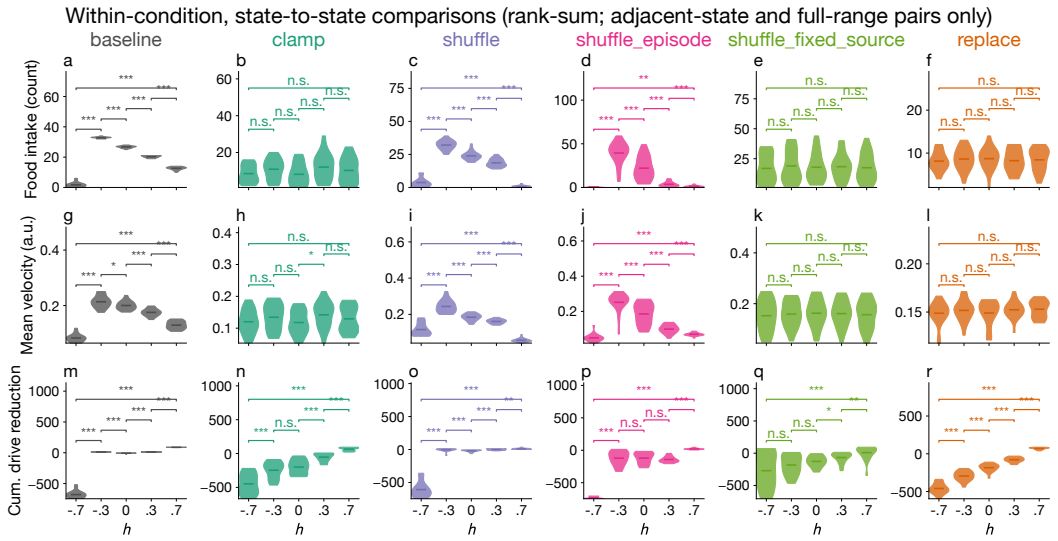


Figure 6: **Behavioral state-dependence survives only when the interoceptive signal carries true internal-state information.** Within each of 6 interoceptive-observation conditions (columns; see §Methods), distributions of food intake (a–f), mean velocity (g–l), and cumulative drive reduction (m–r) across all evaluation seeds, shown separately for each of the 5 true internal states h ($h = -0.7$ = most deprived, $h = +0.7$ = most sated). Brackets show two-sided rank-sum tests between adjacent internal states and between the two extremes ($h = -0.7$ vs. $h = +0.7$); $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$, n.s. not significant. Under baseline (unperturbed interoception), shuffle (fed a true reading from an earlier step, same episode), and shuffle_episode (fed a true reading from anywhere in the same episode), behavior varies significantly and monotonically with h. Under clamp (interoceptive channels fixed to a constant), shuffle_fixed_source (fed readings from a single, fixed reference episode regardless of true h), and replace (fed i.i.d. uniform noise), food intake and velocity no longer differ significantly across internal states — the policy's observed interoceptive signal carries no information about the true state, and behavior stops tracking it. Cumulative drive reduction remains state-dependent in every condition because it is computed from the environment's true (unperturbed) internal state regardless of what the policy observes, serving as a manipulation check.

H Causal Necessity of the Interoceptive Channels

Although the interoceptive stream is accessible to the policy and behavior is state-dependent, mere availability and correlation do not demonstrate necessity. To address this, we disrupted the two interoceptive observation channels in the already-trained agent exclusively during evaluation using three manipulations: additive uniform noise on $[-1.0, 1.0]$; clamping both channels to the homeostatic set point; and shuffling the channel values temporally (Figure 6).

In the unmanipulated baseline, behavioral statistics across rollouts, including food consumed, velocity, and covered distance, differ significantly among the fully deprived, neutral, and fully sated conditions (Kruskal–Wallis, $p < 0.001$). However, under each disruption, these differences disappear. This finding indicates that the interoceptive signal is not only present and correlated with behavior, but that the policy's state-dependent modulation is causally dependent on this signal. Removing the interoceptive signal abolishes the modulation while leaving the body, environment, and network weights unchanged.

I Torpor Under Complete Deprivation: Food Approach, Not Food Use

The main text describes a low-activity state under complete deprivation. Behavior alone does not show what fails: the agent may be unable to use food in this state, or it may not take available food.

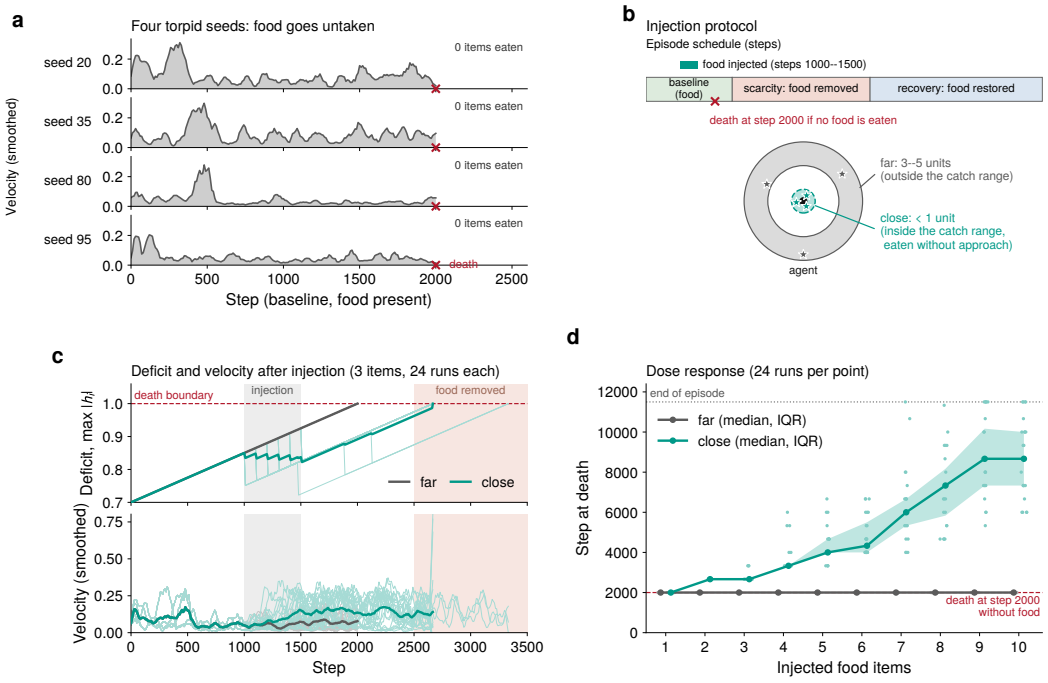


Figure 7: **Food approach limits recovery in torpid seeds.** One trained network, started fully deprived ($h_1 = h_2 = -0.7$), four evaluation seeds (20, 35, 80, 95: no food eaten, lowest late-baseline velocity). (a) Without intervention, none of the four seeds ate any of the ten available food items; velocity (50-step moving average) stays low, and each seed reaches the death boundary at step 2,000 ($\times$). (b) Injection protocol: during a 2,500-step baseline, food is injected at steps 1,000–1,500 either 3–5 units from the agent (far, outside the one-unit catch range) or within the catch range (close); the protocol removes all food from step 2,500. (c) Deficit (max $|h_i|$) and smoothed velocity after injecting three items. Thin lines: the 24 runs per condition (4 seeds $\times$ 6 injection steps); bold lines: their mean (close, while at least 20 runs are alive). The far deficit traces coincide because no food is eaten. Grey band: injection window; red band: food removed. (d) Step at death against the number of injected items (1–10; 24 runs per point): median and interquartile range, with individual runs (points offset horizontally for readability). Far food never prolongs survival; close food does so in a dose-dependent way, with no delay for one item. Dotted line: end of the 11,500-step episode.

We test this by placing food at controlled distances from the agent. All runs use one trained network started fully deprived ($h_1 = h_2 = -0.7$) with ten food objects present.

Figure 7 summarizes these experiments. Without intervention, all four seeds ate nothing during the baseline, although food was present throughout, and reached the death boundary at step 2,000. Their velocity is low but not zero (mean 0.03–0.10 over the last 300 steps).

In each seed, we placed three food items around the agent's current position at one of six injection steps (1,000–1,500), well before death, giving 24 runs per condition. Items placed 3–5 units away, outside the 1-unit catch range, were never eaten, and no run survived beyond step 2,000; the closest approach was 2.3 units. Items placed within the catch range were eaten in all 24 runs, and all of these runs survived beyond step 2,000 (death at step 2,666 in 21 runs and 3,333 in 3; the protocol removes all food from step 2,500). In 7 runs, the agent then ate 1 to 3 additional items on its own. Varying the number of injected items from 1 to 10 (480 runs) confirmed the contrast. At a distance, no run survived beyond step 2,000, and only 3 of 240 ate any food. In contact, all 240 runs ate food, and survival grew with dose: the median death step was 2,000 for one item (no delay), 2,666 for two, 4,000 for five, and 8,666 for nine or ten, and 10 runs (at seven, nine, or ten items) survived the entire 11,500-step episode. Thereby, under complete deprivation, the limitation is not that food cannot be used, but that the agent does not approach food outside its reach, even at 3–5 units, within its usual foraging range.

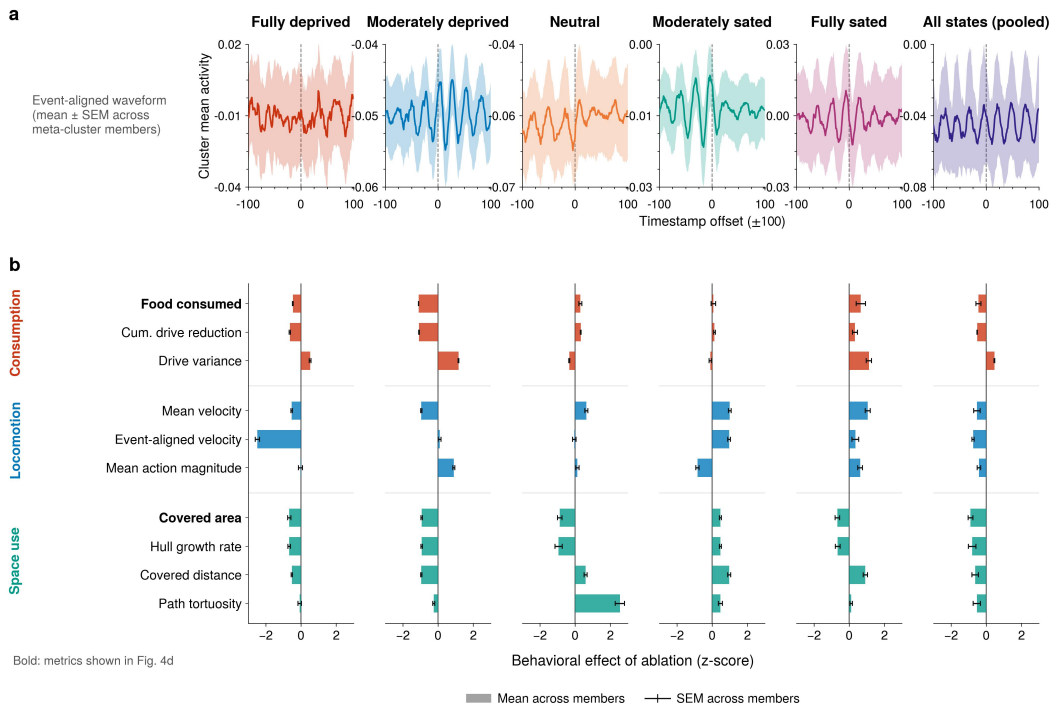


Figure 8: **Waveforms and full behavioral-effect profiles of the exported meta-clusters.** One column per meta-cluster shown in Figure 4b–d (fully deprived, moderately deprived, neutral, moderately sated, fully sated, and all states pooled). (a) Event-aligned mean activity of each meta-cluster (mean $\pm$ SEM across members, ± 100 steps around food consumption; colors as in Figure 4b,c). (b) Effect of ablating the member clusters on all ten behavioral metrics, grouped into consumption, locomotion, and space use. Effects are z-scored against all members of the same representation, as in Figure 4d.

J Full behavioral-effect profiles of the exported meta-clusters

Figure 4d presents two of the ten behavioral metrics (food consumed and covered area) for the selected meta-clusters. Figure 8 displays all ten for the same six meta-clusters, along with their event-aligned waveforms (Figure 4c). Bars are z-scores relative to all grouped clusters of the same representation (one evaluation state, or all states pooled); zero therefore denotes the average grouped cluster, not the unablated agent. The profiles vary across meta-clusters in both direction and metric family. The moderately deprived meta-cluster has the broadest profile: lower food consumed and cumulative drive reduction (≈ -1.1), lower velocity, covered area, and covered distance (-0.9 to -1.0), and higher drive variance ($+1.2$) and mean action magnitude ($+0.9$). The fully deprived meta-cluster mainly lowers event-aligned velocity (-2.5); the changes in food consumed (-0.5) and cumulative drive reduction (-0.6) are small, consistent with the floor set by near-zero baseline intake (≈ 1 item per rollout). The neutral meta-cluster dissociates movement from coverage: path tortuosity ($+2.6$), velocity, and covered distance ($\approx +0.6$) increase while covered area decreases (-0.9). The moderately sated meta-cluster raises velocity, event-aligned velocity, and covered distance ($\approx +1.0$) and lowers mean action magnitude (-0.8), with consumption metrics near zero ($|z| \leq 0.14$). The fully sated meta-cluster raises drive variance ($+1.1$), velocity ($+1.1$), and covered distance ($+0.9$) and lowers covered area (-0.7). The pooled meta-cluster lowers every metric except drive variance (-0.4 to -0.9) and raises drive variance ($+0.5$). Hull growth rate tracks covered area to within 0.1 z, as expected since it is derived from hull area, and path tortuosity is otherwise small ($|z| < 0.6$).

K Conventional RL agent developed less functional clusters

Under ablation, the uniformly positive Z-scores for the control agent's cumulative extrinsic reward do not reflect cluster-specific foraging improvements. Silencing any cluster reduces motor output, which

in turn decreases the motor-effort penalty term in Equation 1, regardless of which cluster is silenced. Consequently, this metric does not reliably assess the control agent's foraging competence. Instead, food intake, velocity, and coverage metrics (which do not exhibit cluster-specific effects) offer more informative measures. Whereas HRL agents develop neural clusters with distinct functional roles in controlling behavior (note distinct pixels in Figure 4a in comparison to the corresponding RL agent in Figure 5d).

L Environment Robustness

The main text addresses training-seed variability on 10 independently trained networks $\times$ 5 rollouts. Here we vary a second source of stochasticity, the environment layout, while holding the trained network and the policy seed fixed: each agent's network is evaluated on 21 layouts (environment seeds $0, 5, \ldots, 100$) with the policy seed fixed at 100. Figure 9 shows the homeostatic agent and Figure 10 the conventional-RL control.

Behavior is state-dependent under layout variation. For the homeostatic agent, the median food consumed is 1, 32, 27, 21, and 13 items from full deprivation to full satiety (Figure 9b). In the example layout, its trajectory at full deprivation stays close to the start, whereas it covers the arena at the other states (Figure 9a). Internal state explains 91–97% of the variance of each headline metric, whereas layout explains at most 2.3% (two-way decomposition over 21 layouts $\times$ 5 states; Table 2), and all five metrics depend on state after correction. Sixteen of the 21 layouts have at least one food item at full deprivation. For the conventional control (Figure 10), every layout eats at every state, and the trajectory covers the arena at every state (Figure 10a); median food consumed decreases from 67 to 41 items (Figure 10b). Internal state explains 66–69% of the variance of food consumed, velocity, and distance (layout 5–12%) and 24% and 10% for area and reward (layout 36% and 33%). After correction, four metrics are significant, and reward is weakly so (Table 2).

Table 2: State-dependence under environment-layout variation (21 layouts $\times$ 5 internal states, one network per agent). Variance explained by internal state and by layout (two-way decomposition), Friedman χ^2 (df $= 4$), and Benjamini–Hochberg-corrected p values across the five metrics, per agent.

	Homeostatic				Control			
Metric	State (%)	Layout (%)	χ^2	p_{fdr}	State (%)	Layout (%)	χ^2	p_{fdr}
Food consumed	91.3	0.8	76.1	2.0×10^{-15}	69.3	5.1	66.4	3.2×10^{-13}
Velocity	91.3	2.2	79.6	5.3×10^{-16}	67.8	11.5	64.9	4.6×10^{-13}
Area covered	91.4	0.8	53.3	7.3×10^{-11}	23.6	35.6	32.9	1.6×10^{-6}
Distance covered	91.5	2.3	79.6	5.3×10^{-16}	66.3	11.7	67.5	3.2×10^{-13}
Reward	97.1	0.2	75.5	2.0×10^{-15}	9.6	33.1	10.2	0.037

The deprivation-specific drop also separates the two agents in this protocol. The ratio of a metric at full deprivation to its value at moderate deprivation, computed per layout, is far below one for the homeostatic agent (0.098 ± 0.042 for food consumed and 0.376 ± 0.018 for velocity; mean $\pm$ SEM over 21 layouts) and close to one for the control (1.065 ± 0.025 and 1.085 ± 0.016), reproducing the contrast observed across independently trained networks in the main text.

Neural organization recurs across layouts in both agents. We applied the meta-clustering pipeline to the layouts using HAC with consistent thresholds and filters. We required agreement across all three dissimilarity measures, and defined support as the number of distinct layouts (out of 21) spanned by a meta-cluster. Consistent with the main text, we required a minimum support of 70% of the units, corresponding to 15 out of 21 layouts. We observed meta-clusters in every state for the HRL agent. In contrast, no meta-cluster reached this threshold for the conventional RL control agent; the highest support observed was 10 layouts.

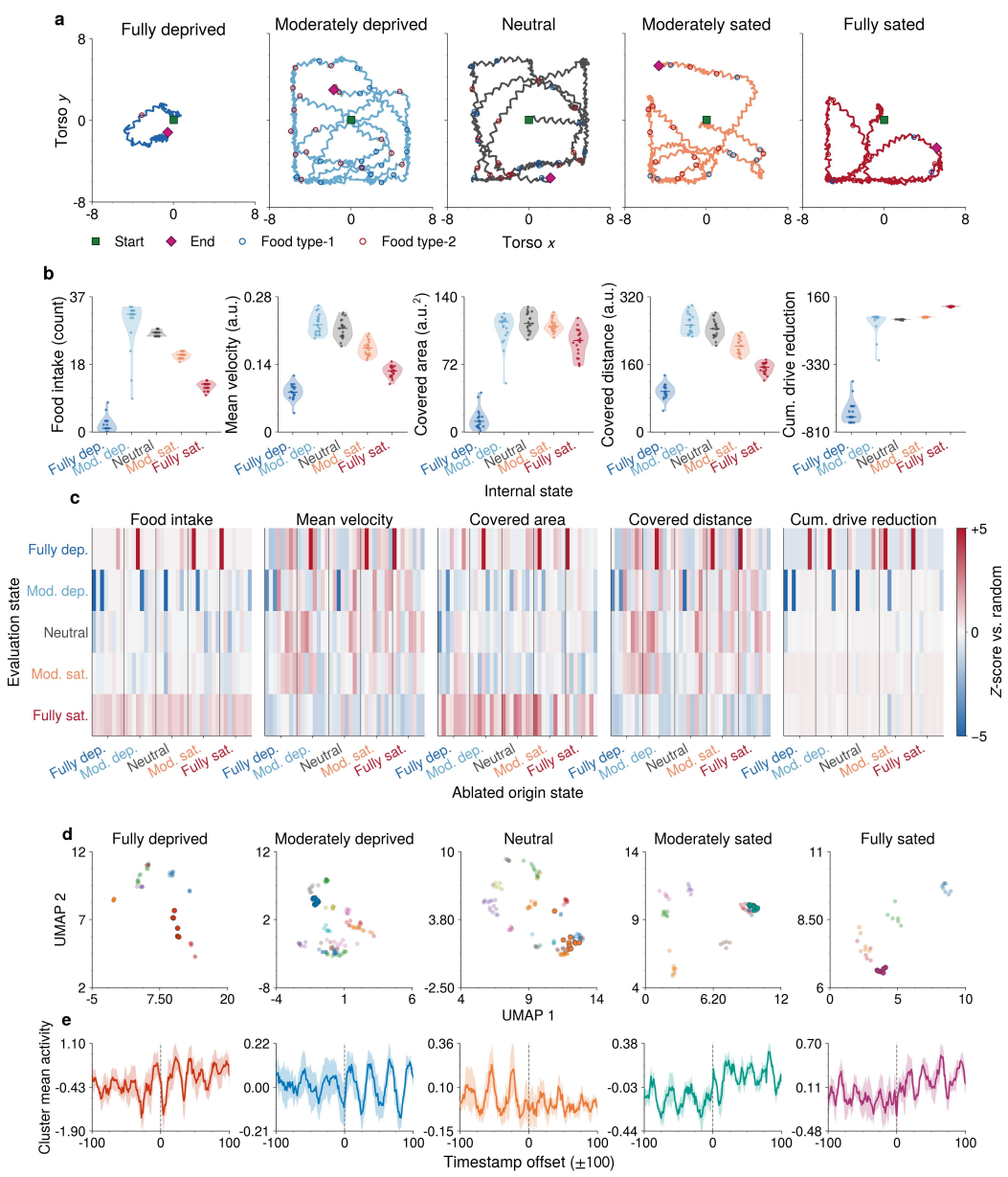


Figure 9: **Environment-layout robustness of the homeostatic agent.** One trained network evaluated on 21 environment layouts (environment seed $0, 5, \ldots, 100$) with the policy seed fixed at 100. (a) XY trajectories at the five internal states (start, end, and food events marked). (b) Behavioral profile over all 21 layouts, as in Figure 2c: violins with one point per layout, bars indicate medians. (c) Effect of ablating each L1 cluster (columns, grouped by the state in which it was identified) on five behavioral metrics at each evaluation state (rows), as Z-scores against that rollout's random-ablation null. (d) UMAP of ablation-effect vectors at each evaluation state; points are clusters, colored by meta-cluster. (e) Event-aligned mean activity ($\pm$ SEM across members) of the highlighted meta-cluster in each column; dashed line: food consumption.

M Layer-2 Organization

The main text analyzes layer 1, the first hidden layer (256 units). Layer 2 is the hidden layer immediately before the action head (256 units). We repeated the analysis pipeline on layer 2 in the same homeostatic networks (10 trained networks, 5 rollout seeds, environment seed 100): event-

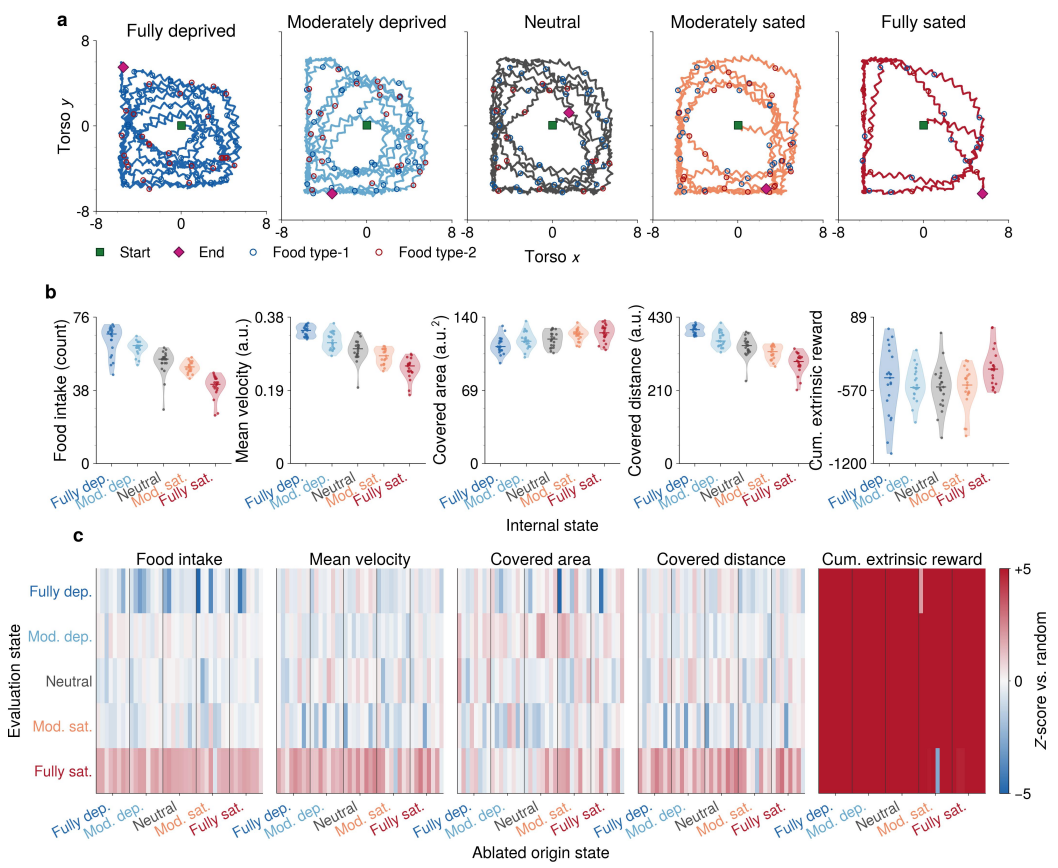


Figure 10: **Environment-layout robustness of the conventional RL control.** Panels as in Figure 9a–c, for one trained control network evaluated on 21 environment layouts with the policy seed fixed at 100: (a) trajectories of the agent, (b) behavioral profile over all 21 layouts, and (c) ablation effects of L1 clusters as Z-scores against the random-ablation null.

centered clustering ($K = 8$), whole-episode cluster ablation with a random-ablation null, and HAC meta-clustering.

The observed clusters in layer 2 analysis resemble the layer 1 neural clusters (Figure 11). Ablating layer-2 clusters produces state-dependent effects (Figure 12a), and meta-clusters recur across seeds (Figure 12b,c).

We compared the whole-episode effect of ablating a cluster in layer 1 and layer 2, using one unit per (network, state) pair: for each rollout, the median absolute change from baseline over the eight clusters identified at that state and evaluated at that state, averaged over the network's five rollouts (46 units; the four networks without fully-deprived food events contribute no fully-deprived unit). Effect sizes do not differ significantly for any metric (paired Wilcoxon, Benjamini–Hochberg corrected $p_{\text{fdr}} \geq 0.16$; Figure 13), and the ratio of the layer-2 to the layer-1 median lies between 0.92 and 1.13.

N Extended Discussion and Biological Context

The observed behavioral and neural signatures align with established biological mechanisms that regulate physiological needs. The low-activity state observed at deep deficit resembles energy-conserving survival strategies in mammals, in which hypothalamic circuits regulate transitions between active foraging and quiescence (Hrvatin et al., 2020). Causal tests presented in Appendix I support this interpretation: the state is consistently exited once resources are restored, up to a deficit threshold beyond which recovery becomes unreliable. Furthermore, the state-dependent modulation

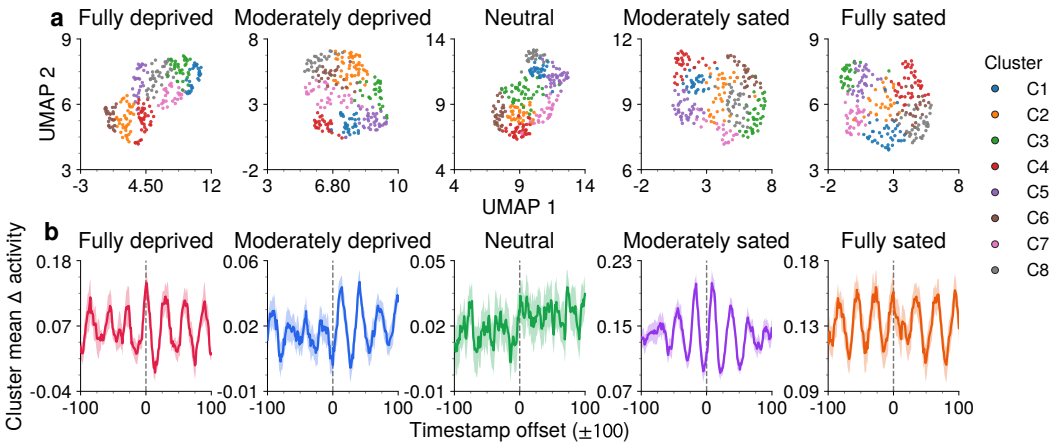


Figure 11: **Layer-2 observed neural clusters.** Layer-2 analog of Figure 3. (a) UMAP of layer-2 neurons at each internal state, colored by cluster (C1–C8). (b) Event-aligned mean activity (± SEM across food events;

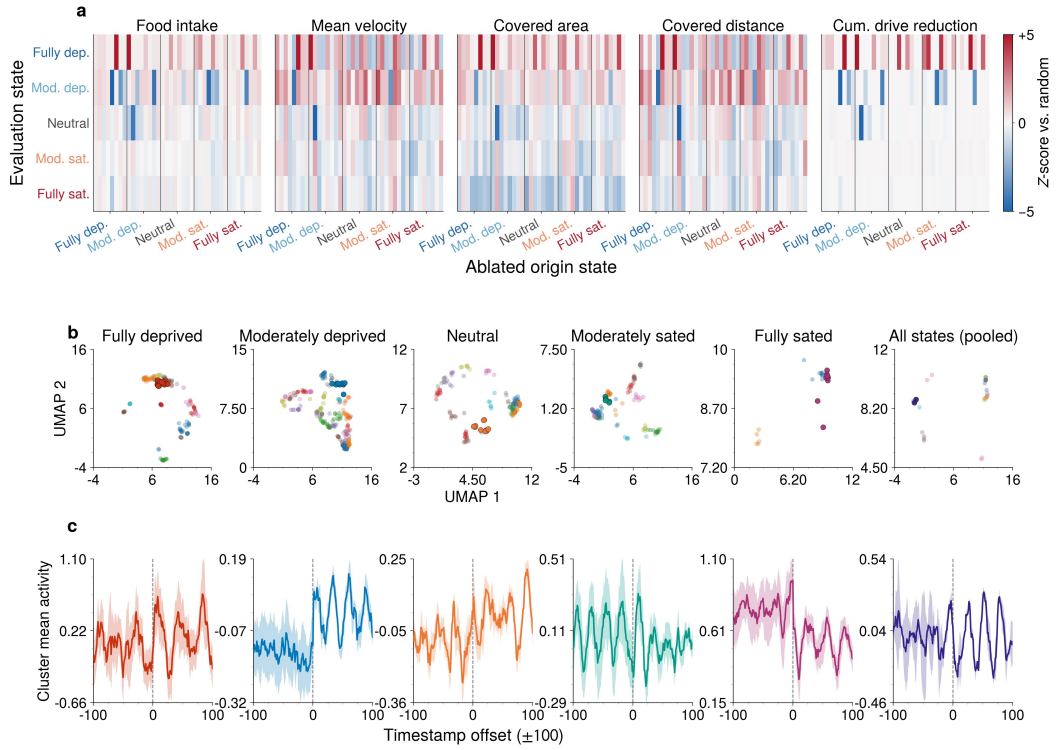


Figure 12: **Layer-2 ablation effects and meta-clusters.** Layer-2 analog of Figure 4. (a) Effect of ablating each layer-2 cluster on five behavioral metrics. (b) UMAP of ablation-effect vectors per evaluation state and pooled, colored by meta-cluster. (c) Event-aligned mean activity (± SEM across members) of the highlighted meta-cluster.

of exploration and foraging intensity parallels hunger-dependent risk-taking behaviors observed in rodents (Kamath et al., 2025; Deem et al., 2022).

At the neural level, the identified waveform motifs, such as ramping pre-event activity and sustained dynamics, resemble temporal coding patterns found in biological motor and interoceptive circuits (Datta et al., 2019). The observed dissociation between cluster types parallels the functional special-

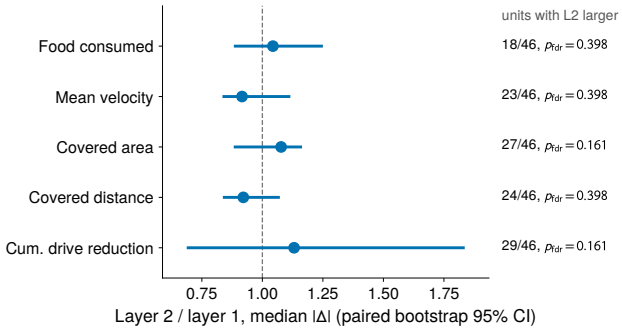


Figure 13: **Layer-2 and layer-1 ablation effect sizes do not differ across trained networks.** Ratio of the layer-2 to the layer-1 median absolute change from baseline (points) with paired bootstrap 95% intervals (10,000 resamples), for 46 (network, state) units of the 10 networks $\times$ 5 rollouts; each unit is the median over the eight clusters identified at that state, evaluated at that state, averaged over the network's rollouts. Right: number of units in which layer 2 is larger and Benjamini–Hochberg-corrected paired Wilcoxon p values.

ization present in biological circuits responsible for motivation and motor control (Kamath et al., 2025; Deem et al., 2022). Although these similarities are qualitative, they indicate that optimization under a homeostatic objective may produce organizational principles analogous to those found in biological systems.

The neuroethological analysis pipeline used in this study, which integrates behavioral quantification, event-centered clustering, and causal ablation, bridges conventional RL and neuroscience-inspired analytical approaches. In particular, applying controlled ablations against random-population baselines enables causal attribution at the level of functional neural groups.