INTRINSIC GOALS FOR AUTONOMOUS AGENTS: HIERARCHICAL EXPLORATION CAPTURES FUTURE VOLITIONAL ACTION IN DROSOPHILA

Anonymous authors
Paper under double-blind review

ABSTRACT

Animals have a remarkable ability to act autonomously in the absence of external sensory cues or extrinsic objectives, unlike most frontier artificial agents today. The computational mechanisms behind this ability largely remain a mystery. We leverage recent whole-brain recordings in the fruit fly (*Drosophila melanogaster*), in which neural activity predicts spontaneous turns more than 10 seconds in advance, to test a computational substrate of future volitional action. Specifically, we study whether intrinsically motivated artificial agents embodied in a biomechanically realistic fly-body can develop behavioral and internal representations quantitatively similar to those observed in biological flies during exploration. Our policies operate over a discrete set of high-level motor primitives while low-level central pattern generators actuate joint dynamics, closely matching current ethological theories of behavioral syllables. We train agents with no external reward using 20 intrinsic motivation algorithms spanning curiosity, entropy, skill-discovery, and empowerment methods. We then evaluate frozen policies in a virtual tethered-ball assay modeled after the fruit-fly experiment, comparing the extent to which spontaneous locomotor statistics and internal policy representations predict future turn directions on the same timescales observed in the fly brain. Using these criteria, we find that DIAYN—a skill-discovery algorithm that learns a discrete set of policies to maximize both the entropy of behavior and the separability in sensory observations those behaviors induce—achieves the strongest behavioral and neural alignment with the fruit-fly, enabled by operating hierarchically over motor primitives. Notably, the set of policies DIAYN discovers closely matches the wide individual variability observed in fly exploration strategies, thus suggesting this behavioral diversity optimizes a concrete computational goal. Together, these results establish hierarchical, intrinsically motivated agents as a powerful framework for reverse-engineering animal autonomy.

1 INTRODUCTION

Animals continuously acquire structured experience even in the absence of an explicit task, external reward, or instructive sensory cue. They explore their environment and learn about how their own actions affect the world before those experiences are necessarily tied to a particular objective. Such self-directed experience may contribute to the flexibility with which animals later learn new tasks, adapt to changing environments, and acquire new behaviors over their lifetime. In contrast, artificial agents are commonly optimized for extrinsic objectives and task-specific reward functions, and achieving similarly broad and persistent autonomy remains challenging. Understanding the computational objectives that explain intrinsically-driven animal behavior may therefore provide principles for building artificial agents that autonomously learn useful representations and behaviors without extrinsic goals.

A wide range of intrinsic-motivation objectives have been proposed as computational mechanisms for generating autonomous behavior in the context of exploration and reinforcement learning. These include objectives such as novelty (Burda et al., 2019), prediction error (Pathak et al., 2017), uncertainty (Pathak et al., 2019), state-space coverage (Liu & Abbeel, 2021), empowerment (Klyubin et al., 2008), and unsupervised skill discovery (Eysenbach et al., 2019). However, these methods

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.

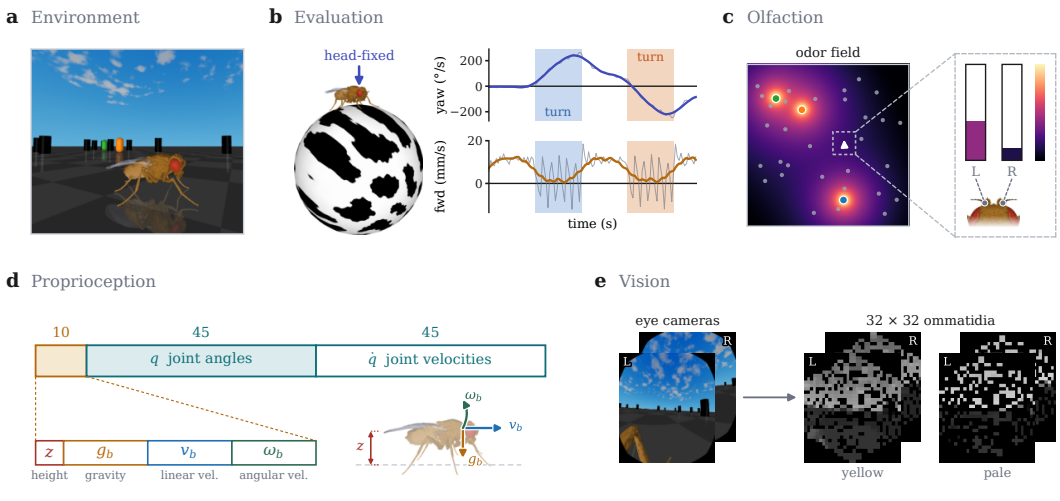


Figure 1: **Environment Overview. a,** The fly policy is trained in an environment with obstacles and multiple odor sources. **b,** The policy is then evaluated in a tethered-ball environment where the fly is head-fixed. The angular and forward velocities are read off the ball just as done in the biological experiment of Brezovec et al. (2023). **c,** The policy receives olfaction using a weighted sum from readings off the fly's antenna and palps. **d,** The policy proprioceptive input also comes directly from the NeuroMechFly simulation including joint angle, velocities and body frame gravity, height, and velocities. **e,** Vision is also provided to the policy by simulating the ommatidia mapping of pale and yellow type cells to pixels.

are typically evaluated according to criteria designed for artificial agents, such as downstream task performance. Such evaluations do not directly test for autonomy, nor do they establish whether the resulting behavior or internal representations resemble those of biological agents that *already* exhibit robust autonomous behavior. Biological systems therefore provide an additional and independently motivated constraint on the space of possible intrinsic objectives. Hence, a candidate objective should not only produce useful exploration, but ideally reproduce measurable properties of animal behavior and neural activity.

Recent whole-brain recordings in *Drosophila* provide a particularly useful "ground-truth" target for these kinds of comparisons. Brezovec et al. (2023) recorded neural activity while head-fixed flies explored spontaneously on a spherical treadmill in darkness. They found that interhemispheric activity in a subregion of the Inferior Posterior Slope (IPS) was highly predictive of future left-versus-right turn direction approximately 15 to 8 seconds *before* a turn. This activity emerged well before the onset of widespread locomotion-related activity across the brain. This provides two complementary constraints on computational hypotheses of internally generated behavior; namely, the spatial and temporal dynamics of the agent's behavior and internal representations of *future* actions.

Methods in computational ethology that describe spontaneous behavior of recurring behavioral syllables (Wiltschko et al., 2015; Mathis et al., 2018; Hsu & Yttri, 2021; Pereira et al., 2022) motivate us to consider the level of behavioral abstraction at which the intrinsic-drive methods should operate. These syllables exhibit organizational structure across multiple time scales. These findings motivate our implementation of a *hierarchical* model policy where the policy trained with intrinsic reward is over a categorical distribution of actions. These high-level actions, or primitives, are then implemented at the joint-torque level by a low-level central pattern generator (CPG) (Mantziaris et al., 2020; Chun et al., 2021) and kinematic replays of real fly locomotion (Fig. 2b). This separates exploratory action selection from the mechanics of locomotion and allows intrinsic objectives to operate over behaviorally meaningful actions rather than individual motor commands.

We extend the goal-driven modeling approach of computational neuroscience (Yamins & DiCarlo, 2016; Richards et al., 2019; Feather et al., 2025), in which candidate objectives are instantiated in artificial networks and compared against neural recordings, to intrinsic motivation. We require that a candidate objective account for behavior as well as neural activity. Our agents are embodied

in NeuroMechFly v2, a biomechanical model of *Drosophila* simulated in FlyGym, with vision, olfaction, proprioception, and articulated body dynamics (Lobato-Rios et al., 2022; Wang-Chen et al., 2024) (Fig. 1). We train recurrent policies with 20 intrinsic drive methods spanning curiosity, state-space coverage, skill discovery, and empowerment (Fig. 2a). All methods share the same embodiment, architecture, training environment, and training budget. We then freeze each policy and evaluate it in a virtual tethered-ball assay that reproduces the biological experiment (Fig. 1b) from Brezovec et al. (2023). For each agent, we compare turning behavior with that of the flies, test whether internal representations predict future turn direction on the timescale of the IPS signal and what ongoing behavior predicts, and ask whether its low-dimensional activity shares the IPS's relationship to angular and forward velocity.

Contributions To summarize, our contributions are as follows:

- **We implement hierarchical, intrinsically motivated agents embodied in a biomechanically-realistic fly body, in which 20 intrinsic objectives operate over seven ethologically motivated behavioral primitives.**
- **We show that prospective representations of future action can emerge from intrinsic motivation alone, without an externally specified task or reward.** In particular, several intrinsically motivated agents develop internal states that predict turn direction seconds in advance, paralleling the prospective signal observed in the fly IPS.
- **Among the intrinsic objectives tested, unsupervised skill discovery provides the strongest computational account of spontaneous fly behavior and its neural dynamics.** DIAYN, hierarchically operating on low-level behavioral primitives, jointly reproduces future-action predictivity, the relationship between neural and behavioral predictability, inter-animal variability in turning behavior, and low-dimensional activity structure related to forward and angular velocity.
- **We demonstrate that behavioral predictability and average locomotor statistics alone are insufficient to distinguish candidate explanations of spontaneous action.** Objectives that generate predictable behavior or match mean turning bias can fail to reproduce the fly's prospective internal signal; jointly evaluating behavioral variability and internal future-action predictivity distinguishes DIAYN from these alternatives.

2 Background and Related Work

2.1 Intrinsically generated and hierarchically structured animal behavior

Animals can generate spontaneous behavior in the absence of explicit external reward or immediate instructive cues. Such behavior cannot be straightforwardly attributed to a specified external reward or cue and therefore provides a setting for studying internally generated action selection. The field of computational ethology has shown that the behavior of animals such as the fly (Berman et al., 2014; 2016; Katsov et al., 2017), mouse (Wiltschko et al., 2015), zebrafish (Marques et al., 2018) and *C. elegans* (Kaplan et al., 2020) can be decomposed into "behavioral syllables" (Anderson & Perona, 2014). A "behavioral syllable" is a recurring, stereotyped pattern of movement that can be reused and combined with other syllables to form more complex behavior. Recent studies have also shown behavior exhibits hierarchical organization across syllables and across timescales. Specifically, across species, animal behavior can be represented at multiple nested levels of abstraction, with fine grained behavioral states carrying predictive information over short timescales and coarser grouping being informative over longer timescales (Berman et al., 2016; Kaplan et al., 2020; Mathis et al., 2018; Hsu & Yttri, 2021; Pereira et al., 2022). Utilizing advances in *whole-brain* neuroimaging of animals during awake behavior, we can get a deeper understanding of how neural dynamics lead to behavior (Mann et al., 2021; Aimon et al., 2023; Brezovec et al., 2024). Recent whole brain recordings of fly brains during autonomous behavior by Brezovec et al. (2023) have shown that across bilaterally matched, locomotion-related superclusters distributed around the brain, summed bilateral activity was associated with forward velocity, whereas interhemispherical activity difference were associated with angular velocity. Even further, the interhemispheric asymmetry emerges much earlier in IPS region of the brain, around 15 to 8 seconds before a turn. Using this region's activity, they show you can predict future actions above chance. These findings provide measurable

behavioral and representational signatures of internally generated action against which computational models can be evaluated. The multiscale organization of animal behavior motivates us to test a hierarchical architecture in which intrinsic drive operates over higher-level locomotor primitives while lower-level controllers execute joint dynamics. This allows us to then compare our trained agents' behavior as well as internal representations.

2.2 Computational theories of intrinsic motivation

While the intrinsic motivations that drive autonomous behavior in animals remain incompletely understood, multiple computational theories of intrinsic motivation have been proposed. These methods can be loosely organized into four functional, non-exclusive families that illustrate different mechanisms for generating intrinsically motivated behavior.

Curiosity-based methods reward properties such as novelty, prediction error, model disagreement, or learning progress. These intrinsic drives therefore encourage agents to seek unfamiliar, surprising, or poorly predicted experiences. From this family, we implemented disagreement (Pathak et al., 2019), γ-progress (Kim et al., 2020), ICM (Pathak et al., 2017), MAX (Shyam et al., 2019), NGU (Badia et al., 2020), and RND (Burda et al., 2019).

Entropy-based exploration methods encourage broad coverage of the state or representation space by rewarding agents for visiting diverse states. Methods in this family include APT (Liu & Abbeel, 2021), C-TeC (Mohamed et al., 2026), Proto-RL (Yarats et al., 2021), and RE3 (Seo et al., 2021).

A third family consists of unsupervised skill-discovery methods. These algorithms seek to learn repertoires of distinguishable latent behaviors or skills without an external reward, often by maximizing mutual information between a latent skill variable and visited states or transitions. This family includes CSF (Zheng et al., 2025), DADS (Sharma et al., 2020), DIAYN (Eysenbach et al., 2019), VALOR (Achiam et al., 2018), and VIC (Gregor et al., 2017). We additionally implement rDADS, a custom reverse-DADS variant developed in this work.

Finally, empowerment-based methods reward an agent for maintaining or increasing its ability to influence future states through its actions. Empowerment is commonly formulated in information-theoretic terms as the mutual information, or channel capacity, between actions and resulting future states conditioned on the current state. We implement several variants. Forward empowerment is a custom variant developed in this work. Empowerment gain is a custom implementation inspired by prior formulations of empowerment gain (Becker-Ehmck et al., 2021; Yiu et al., 2026), but is not an exact implementation of either. Potential empowerment draws on previous formulations and estimators of empowerment (Mohamed & Rezende, 2015; Zhao et al., 2021; Klyubin et al., 2008), while process empowerment follows Tiomkin et al. (2025).

2.3 Artificial neural networks as hypotheses of biological computation

A major approach in computational neuroscience, known as NeuroAI, is to treat task optimized artificial neural networks as candidate models of biological computation (Yamins & DiCarlo, 2016; Richards et al., 2019). Studies have shown similarities in representations formed by ANNs can be compared to *single* brain areas, such as the primate visual cortex (Yamins et al., 2013), mouse visual cortex (Nayebi et al., 2023), the mouse entorhinal/navigation system (Nayebi et al., 2021), auditory (Kell et al., 2018) and the motor system (Sussillo et al., 2015). More recently, NeuroAI has leveraged embodied AI to bridge computational models with neuroscience, including virtual animal models such as the virtual rodent (Merel* et al., 2020; Aldarondo et al., 2024), the virtual fruit fly (Vaxenburg et al., 2025; Lobato-Rios et al., 2022; Wang-Chen et al., 2024), and the virtual zebrafish (Keller et al., 2025). With the exception of Keller et al. (2025), our model is trained entirely via intrinsically motivated exploration, rather than imitation trained on animal behavior, thereby providing a *normative* account of autonomous behavior by replicating neural and behavioral dynamics at test time. Additionally, owing to our intrinsic motivation objectives operating hierarchically over over a range of the fly's behavioral primitives, we advance the scope of Keller et al. (2025) virtual zebrafish beyond two behavioral states (activity and passivity) to a much larger behavioral repertoire.

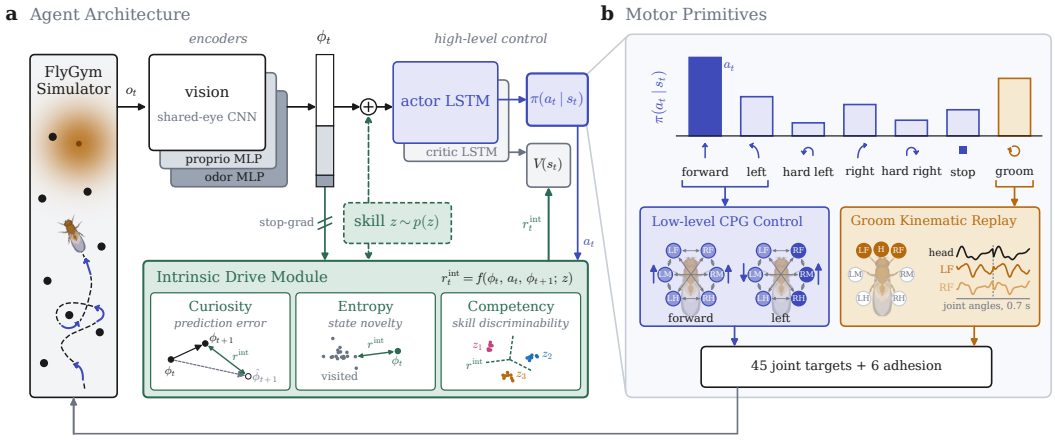


Figure 2: **Agent Architecture Overview. (a)** The FlyGym simulator provides visual, proprioceptive, and olfactory observations from a biomechanically realistic fly body. Modality-specific CNN and MLP encoders embed these observations into a shared representation ϕ_t, which is provided to the actor–critic and intrinsic drive module. When applicable, the policy is additionally conditioned on a sampled skill z. The intrinsic drive module computes an intrinsic reward from the encoded state transition and selected action. The actor outputs a categorical distribution over discrete motor primitives, while the critic estimates the state value. The policy selects a new motor primitive every 100 ms. **(b)** Locomotor primitives are translated into joint-level commands by a low-level central pattern generator (CPG) controller, while grooming is executed using a kinematic replay. These motor commands are applied within the FlyGym physics simulation, which is integrated with a timestep of 100 μs. Thus, each high-level policy action is executed for 1,000 physics steps before the next action is selected.

3 Methods

3.1 Biomechanically Realistic Drosophila Agent With Hierarchical Control

Our agents are embodied in NeuroMechFly v2, a biomechanical model of an adult female *Drosophila* simulated in FlyGym with MuJoCo at a 10^{-4} s timestep (Lobato-Rios et al., 2022; Wang-Chen et al., 2024; Todorov et al., 2012). The policy receives compound-eye vision from pale- and yellow-type ommatidia, a bilateral olfactory signal, and a proprioceptive state of joint angles, joint velocities, and body kinematics (Fig. 1c–e; App. A.1). Control is divided hierarchically (Fig. 2). Every 100 ms, the policy selects one of seven behavioral primitives: forward, left, hard left, right, hard right, stop, and groom. The six locomotor primitives set left and right descending signals that modulate a central pattern generator (CPG). The CPG phase indexes step trajectories recorded from real walking flies (Mantziaris et al., 2020; Chun et al., 2021). The groom primitive executes a kinematic replay of recorded fly grooming (App. A.2). Behavior therefore arises from the policy chaining primitives across many decisions, while joint-level mechanics are fixed across all intrinsic objectives. Modality-specific encoders produce a 310-dimensional feature vector that feeds separate actor and critic LSTMs (256 units each), followed by MLP heads. The actor outputs a categorical distribution over the seven primitives (Fig. 2a; Table 2). For skill-conditioned methods, the skill vector is appended to the input of both LSTMs but withheld from the intrinsic reward module.

3.2 Intrinsic Drive Methods and Training

We train policies with 20 intrinsic drive methods from three main families: curiosity, entropy, and skill-discovery (Sec. 2.2; App. A.4). The intrinsic reward is the only learning signal the policy receives. We use Recurrent PPO (Raffin et al., 2021; Schulman et al., 2017) in 2,048 parallel environments, procedurally generated arenas with three odor sources and 12 to 50 obstacles which vary the visual, olfactory, and proprioceptive input (Fig. 1a). All methods share the same architecture, environment, hyperparameters, and budget of 50 million policy decisions (Table 5), and we

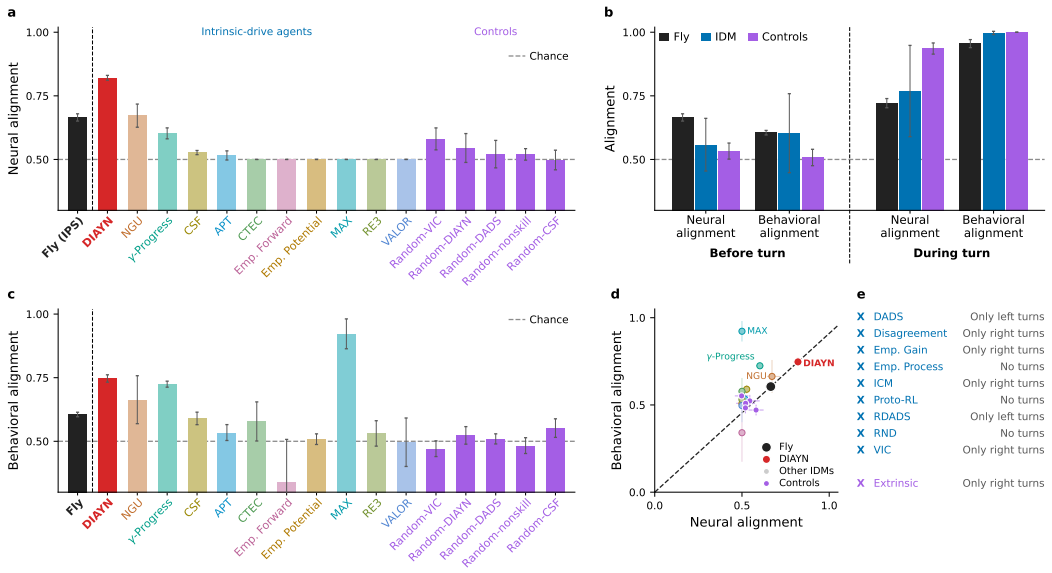


Figure 3: **Decoding from Neural and Behavioral data.** **(a)** The neural alignment of the models and fly decoding future turns in the -15 to $-8\,\mathrm{s}$ window using ElasticNet. **(c)** The behavioral alignment of the models and fly decoding from the behavioral data in the -15 to $-8\,\mathrm{s}$ window using ElasticNet. **(b)** The difference in performance for each group, Fly, IDM and Control from -15 to $-8\,\mathrm{s}$ window with neural and behavioral data to -0.4 to $+0.4\,\mathrm{s}$ window. **(d)** Examines the ratio of behavior alignment to neural alignment. DIAYN captures the precise inter-animal ratio between behavioral and neural accuracy (lying on the dashed line), while other algorithms result in a bias towards behavioral accuracy. **(e)** List of models that do not result in any turns or only turn one direction which excludes them from being scored for decoding as there are no stratified folds for cross validation.

evaluate the final checkpoint. As controls, we evaluate randomly initialized policies with the same architectures, and a policy trained on extrinsic odor navigation task (App. A.7).

3.3 Tethered-Ball Evaluation

Frozen policies are evaluated in a virtual version of the assay of Brezovec et al. (2023), where the fly is head-fixed on a $9\,\mathrm{mm}$ air-supported ball, and angular and forward velocity are read from the ball simulator state (Fig. 1b). To match the darkness and absence of odor in the biological experiment, visual and olfactory inputs are set to zero. For each policy we collect nine 30-minute rollouts, matching the nine flies and 270 minutes of the biological dataset. When applicable, skills are sampled once at the start of each episode and then every time the policy decision reaches the skill horizon of the IDM. Actions are sampled, not greedily selected. The LSTM hidden state is not reset during the evaluation. Details are given in App. A.6.

4 Results

4.1 Behavioral and Representational Analysis

Behavioral data is processed identically for agents and flies matching the procedure of Brezovec et al. (2023). First, velocities are smoothed with a $500\,\mathrm{ms}$ Savitzky-Golay filter (Savitzky & Golay, 1964). Then, turns are detected as peaks in angular velocity above $200°/\mathrm{s}$ (Brezovec et al., 2023). Each evaluation rollout or fly experiment is summarized by its turning bias, $\log(N_R/N_L)$. To test for a prototypical planning representation, we decode future turn direction from the activity in the -15 to $-8\,\mathrm{s}$ window using elastic-net penalized logistic regression (Zou & Hastie, 2005; Pedregosa et al., 2011) with five-fold stratified cross-validation and report balanced accuracy (Brodersen et al.,

2010). Decoder hyperparameters are tuned once on the fly data and fixed for all agents. In our agents, we decode from four recorded layers and three concatenations of related layers pooled to match the imaging rate of fly neural data. For the biological fly, performance is decoded from the IPS activity (App. A.8). Because future turns may be predictable from current behavior alone, we fit the same decoder to mean angular and forward velocity in the same window to identify policies that converged to a relatively stable, predictable behavioral strategy (App. A.9). Finally, to compare the representation geometry without assuming an anatomical correspondence, we apply PCA to each agent's activity in its highest performing neural layer for decoding. We test whether pairs of components related to angular and forward velocity as the IPS hemispheres do (App. A.11).

Neural and Behavioral Decoding. Using the network activations from the models in the tethered-ball evaluation, we follow the same conceptual test used for the biological fly, asking whether internal activity before a turn predicts its eventual direction and plot the alignment in Fig. 3a. The biological fly IPS has a neural alignment of $66.5\% \pm 0.014$ while DIAYN and NGU of the IDM trained models meet or exceed this alignment with $82.1\% \pm 0.009$ and $67.2\% \pm 0.045$ respectively (Fig. 3a). Many IDM-trained policies are near 50% alignment or had an undefined alignment due to never having a detected turn or only having detected turns of a single direction. Both of these conditions do not allow us to do stratified folds for the accuracy calculation. Similarly to neural decoding, we perform a behavioral decoding. Using the mean angular (ω) and forward (v) velocity of the ball in the before turn window, we fit a logistic regression classifier to decode the future direction of the turn and plot the alignment in Fig. 3c. For the flies, their behavior is predictive of future turns with an alignment of $60.5\% \pm 0.009$. For the IDM-trained policies, many exhibit behavior that is predictive of future turn direction and decode-able including MAX ($92.2\% \pm 0.059$), DIAYN ($74.7\% \pm 0.014$), γ-progress ($72.4\% \pm 0.012$) and NGU ($66.3\% \pm 0.094$). Nine of the 20 methods produce no turns or turns in only one direction during evaluation (Fig. 3e), a behavioral mismatch with the fly in itself.

Performance Ratio. These results distinguish IDMs that produce internally predictive representations from simple behavioral patterns that are easily predictable from past behavior. Critically, an IDM that has similar behavior to that of the fly, should have the same relative ratio of neural decoding accuracy to behavioral decoding accuracy as this means the internal representations are predictive of future activity and the behavior itself is stereotypical but not trivially predicted. Therefore, Fig. 3d considers neural and behavioral predictability jointly. Among the evaluated IDMs, DIAYN and NGU are the only IDMs that match or exceed the neural and behavioral predictability of the fly. Unlike NGU, DIAYN also maintains the same ratio of predictability as the fly.

Turning Bias. Models that match the mean statistics of the flies do not match the flies' internal predictability. In Fig. 4b, the turning bias distribution of each model is represented, models that capture the mean of the flies, such as APT and the controls, do not match the distribution of the fly turning biases and have a worse R^2 with the flies (Fig. 4a). Compare with DIAYN, that captures a larger swath of turning biases with its skills allow it to have a greatly improved R^2 with the flies. DIAYN holds one skill per 30-minute rollout, so its nine rollouts sample nine skills, paralleling the nine flies. CSF and Random-DIAYN use the same schedule yet show far narrower bias distributions (Fig. 4b), indicating that the fly-like spread arises from the learned objective rather than the sampling scheme. **Internal Representations.** The internal activation dynamics of the models were also compared with the fly neural dynamics. As shown in Figure 5a, the fly IPS exhibits separable structure in the network activations that jointly correlate with the angular and forward velocity. Figure 5b shows that trained DIAYN recovers a similar relationship, whereas this organization is absent in the random-initialized control (Figure 5c). The trained DIAYN maintains a similar relationship between principal components as the IPS interhemispheric activity with correlation r with ω of $+0.153$ and $+0.413$ with v compared with the Fly IPS that has $+0.191$ with ω and $+0.429$ with v. This relationship does not appear in the randomly initialized model of DIAYN with $+0.021$ for ω and $+0.113$ with v. DIAYN is the only IDM that simultaneously captures fly-like behavioral variability, future-action predictability, and the relationship of internal representations to angular (ω) and forward (v) velocity. Other trained IDMs recover only subsets of these properties.

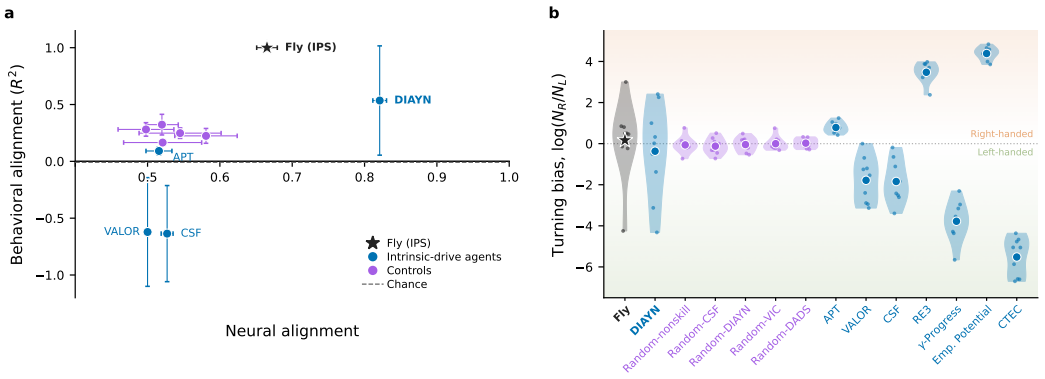


Figure 4: **Behavioral and Neural Alignment. (a)** The R^2 relationship between the turning bias of each episode for each IDM to each fly sorted by turning bias vs. the neural alignment in the before turn window. Marker is the mean R^2 while error bars are the $\pm$ standard deviation when sampling episodes with replacement. **(b)** The skills of DIAYN allow it to capture more of the variability of the fly behavior as opposed to other IDMs that capture only a single mode (bias) or are overly biased to either direction. IDMs that produced episodes with no turns, only left or right turns are excluded from the figure panels as they have undefined or infinite turning bias.

5 Discussion

5.1 Hierarchical Skill Discovery as a Model of Fly Autonomy

The central implication of our results is that behavioral diversity and prospective internal activity may reflect a shared computational objective. Methods that match average turning bias or generate highly predictable behavior need not reproduce the accompanying prospective signal (Figs. 3 and 4). DIAYN's joint correspondence with these properties therefore motivates an explanation of autonomy that addresses both the diversity of spontaneous behavior and the internal organization that makes future actions predictable.

DIAYN's skill-discovery objective provides a concrete candidate for this organization: it rewards skills that are distinguishable through the states they induce while maintaining a high-entropy policy. In our hierarchical implementation, these skills are not additional motor primitives, but different strategies for selecting among the same seven primitives. The learned diversity therefore concerns the organization of behavior above the level of individual motor commands, and even across individual flies. Specifically, individual flies exhibit persistent locomotor handedness (Buchanan et al., 2015; Ayroles et al., 2015; de Bivort et al., 2022). DIAYN's learned skill repertoire captures a broader range of turning biases than models concentrated near the population mean (Fig. 4). This correspondence suggests that different exploratory strategies need not require different underlying objectives: they can arise within a repertoire shaped by a single intrinsic goal. Thus, our results suggest intrinsic motivation serves the role of producing behavioral strategies, that across *individuals of the species*, induce distinguishable sensory consequences which maximally cover sensory space.

Skill discovery may also help explain why internal activity becomes predictive of future turns. Maintaining distinguishable sensorimotor strategies over time (where the temporal aspect is implemented by the RL optimizer) encourages the policy to learn a recurrent state that sustains different tendencies to select particular motor primitives depending on the skill. Consequently, the latent state carries information about later actions over the trial by virtue of optimizing the policy to produce different actions for different skills. Under this hypothesis, prospective activity reflects the continuing organization of a behavioral strategy over time, rather than only the execution of an imminent movement on a moment-to-moment basis. This remains our current best *quantitative* explanation of the joint behavioral and representational findings. Hierarchical skill discovery therefore provides a concrete computational framework in which animal-like behavioral diversity and future-action representations emerge together under a common intrinsic objective.

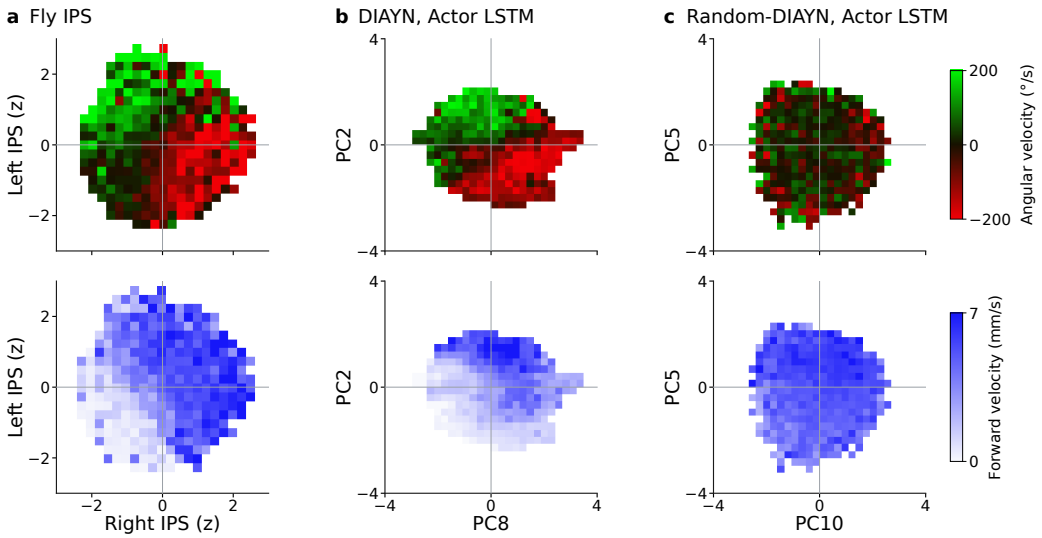


Figure 5: **Low-dimensional internal structures relate with ω and v.** **(a)** Reconstructed from Brezovec et al. (2023), the IPS of the fly brain has an interhemispheric relationship of activity that is correlated with angular (interhemispheric difference) and forward velocity (interhemispheric sum). **(b)** Utilizing PCA as an analogous method to the Brezovec et al. (2023) superclusters, we find principal components in the network activation of the LSTM layer of the Actor in the policy whose inverse activity correlates with angular velocity and combined activity correlates with forward velocity. This structural relationship does not appear in **(c)** the randomly initialized DIAYN model.

5.2 Limitations and Future Work

The main limitation of our training is the simple ecological environment. While it has randomized sensory and locomotor variation and was sufficient to explain the neural and behavioral variability of our fly data, it is substantially simpler than the ecological environments in which flies develop. Richer environments with interactive objects, variable terrain, resources and textures may favor different intrinsic objectives or generate a broader repertoire of learned behaviors. Future work could leverage our algorithms and agents to train them in more ecological environments, and compare them to whole-brain data from *Drosophila* in natural environments once such brain data is available. Finally, the space of behaviors available to each policy/IDM is constrained by the predefined primitives and low-level controllers. While this reflects the hierarchical organization of animal behavior, it limits the learned behaviors implicitly. Future work could model the behavioral primitives and their high-level selection, allowing the IDM to shape its low-level repertoire.

6 Conclusion

Our results support hierarchical skill discovery as a candidate computational account of fly autonomy. Among the objectives tested, DIAYN most consistently reproduced inter-fly behavioral variability together with prospective internal activity and low-dimensional structure related to locomotion. The neuroscientific implication is that individuality and future-action representations may be connected: an intrinsic objective that favors distinguishable sensorimotor strategies could generate both diverse ways of behaving and internal states informative about subsequent actions. Different flies could thus express different exploratory strategies while sharing an underlying computational goal. By testing candidate objectives against both behavioral variation and neural dynamics, our framework moves beyond describing spontaneous behavior toward explaining the principles that organize it. These findings suggest a view of autonomy in which the absence of an external task does not imply an absence of purpose: behavior may instead be organized by intrinsic goals for developing distinct ways of acting on and experiencing the world.

AI USE STATEMENT

Generative AI tools were used to aid and polish writing, retrieve and discover related work, and assist with research ideation and execution. In preparing this manuscript, ChatGPT and Claude were used to assist with creating and formatting plots.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide an anonymized code supplement containing the complete code used for model training, evaluation, and analysis. The repository includes the scripts and configuration files required to reproduce the experiments and generate the reported results. We also provide instructions for installing the required dependencies and running each stage of the experiments. Random seeds and experimental settings are specified in the provided configurations where applicable. The repository will be made publicly available upon publication.

REFERENCES

Joshua Achiam, Harrison Edwards, Dario Amodei, and Pieter Abbeel. Variational option discovery algorithms. *arXiv preprint arXiv:1807.10299*, 2018.

Sophie Aimon, Karen Y. Cheng, Julijana Gjorgjieva, and Ilona C. Grunwald Kadow. Global change in brain state during spontaneous and forced walk in Drosophila is composed of combined activity patterns of different neuron classes. *eLife*, 12:e85202, 2023. doi: 10.7554/eLife.85202.

Diego Aldarondo, Josh Merel, Jesse D Marshall, Leonard Hasenclever, Ugne Klibaite, Amanda Gellis, Yuval Tassa, Greg Wayne, Matthew Botvinick, and Bence P Ölveczky. A virtual rodent predicts the structure of neural activity across behaviors. *Nature*, pp. 1–3, 2024.

David J. Anderson and Pietro Perona. Toward a science of computational ethology. *Neuron*, 84(1): 18–31, 2014. doi: 10.1016/j.neuron.2014.09.005.

Brian B. Avants, Nicholas J. Tustison, and Gang Song. Advanced normalization tools (ANTS). *The Insight Journal*, 2(365):1–35, 2009. doi: 10.54294/uvnhin.

Brian B. Avants, Nicholas J. Tustison, Gang Song, Philip A. Cook, Arno Klein, and James C. Gee. A reproducible evaluation of ANTs similarity metric performance in brain image registration. *NeuroImage*, 54(3):2033–2044, 2011. doi: 10.1016/j.neuroimage.2010.09.025.

Julien F. Ayroles, Sean M. Buchanan, Chelsea O'Leary, Kyobi Skutt-Kakaria, Jennifer K. Grenier, Andrew G. Clark, Daniel L. Hartl, and Benjamin L. de Bivort. Behavioral idiosyncrasy reveals genetic control of phenotypic variability. *Proceedings of the National Academy of Sciences*, 112 (21):6706–6711, 2015. doi: 10.1073/pnas.1503830112.

Adrià Puigdomènech Badia, Pablo Sprechmann, Alex Vitvitskyi, Daniel Guo, Bilal Piot, Steven Kapturowski, Olivier Tieleman, Martín Arjovsky, Alexander Pritzel, Andrew Bolt, and Charles Blundell. Never give up: Learning directed exploration strategies. In *International Conference on Learning Representations*, 2020.

Philip Becker-Ehmck, Maximilian Karl, Jan Peters, and Patrick van der Smagt. Exploration via empowerment gain: Combining novelty, surprise and learning progress. In *ICML 2021 Workshop on Unsupervised Reinforcement Learning*, 2021. URL `https://openreview.net/forum?id=pmN5h1McUWV`.

Gordon J. Berman, Daniel M. Choi, William Bialek, and Joshua W. Shaevitz. Mapping the stereotyped behaviour of freely moving fruit flies. *Journal of the Royal Society Interface*, 11(99): 20140672, 2014. doi: 10.1098/rsif.2014.0672.

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.

John A. Bogovic, Hideo Otsuna, Larissa Heinrich, Masayoshi Ito, Jennifer Jeter, Geoffrey W. Meissner, Aljoscha Nern, Jennifer Colonell, Oz Malkesman, Kei Ito, and Stephan Saalfeld. An unbiased template of the Drosophila brain and ventral nerve cord. *PLOS ONE*, 15(12):e0236495, 2020. doi: 10.1371/journal.pone.0236495.

Bella E. Brezovec, Andrew B. Berger, Yukun A. Hao, Feng Chen, Shaul Druckmann, and Thomas R. Clandinin. Mapping the neural dynamics of locomotion across the Drosophila brain. *Current Biology*, 34(4):710–726.e4, 2024. doi: 10.1016/j.cub.2023.12.063.

Luke E. Brezovec, Andrew B. Berger, Shaul Druckmann, and Thomas R. Clandinin. Neural correlates of future volitional action in Drosophila. *bioRxiv*, 2023. doi: 10.1101/2023.09.08.556917.

Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M. Buhmann. The balanced accuracy and its posterior distribution. In *2010 20th International Conference on Pattern Recognition*, pp. 3121–3124. IEEE, 2010. doi: 10.1109/ICPR.2010.764.

Sean M. Buchanan, Jamey S. Kain, and Benjamin L. de Bivort. Neuronal control of locomotor handedness in drosophila. *Proceedings of the National Academy of Sciences*, 112(21):6700–6705, 2015. doi: 10.1073/pnas.1500804112.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. In *International Conference on Learning Representations*, 2019.

Chanwoo Chun, Tirthabir Biswas, and Vikas Bhandawat. Drosophila uses a tripod gait across all walking speeds, and the geometry of the tripod is important for speed control. *eLife*, 10:e65878, 2021. doi: 10.7554/eLife.65878.

Benjamin de Bivort, Sean Buchanan, Kyobi Skutt-Kakaria, Erika Gajda, Julien Ayroles, Chelsea O'Leary, Pablo Reimers, Jamilla Akhund-Zade, Rebecca Senft, Ryan Maloney, Sandra Ho, Zach Werkhoven, and Matthew A.-Y. Smith. Precise quantification of behavioral individuality from 80 million decisions across 183,000 flies. *Frontiers in Behavioral Neuroscience*, 16:836626, 2022. doi: 10.3389/fnbeh.2022.836626.

Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. In *International Conference on Learning Representations*, 2019.

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.

Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202, 1980. doi: 10.1007/BF00344251.

General Plastics Manufacturing Company. FR-4600 High-Density Fine Cell Polyurethane Foam. `https://www.generalplastics.com/products/fr-4600`. FR-4615 density: 15 lb/ft^3 (approximately 240 kg/m^3). Accessed 2026-09-25.

Karol Gregor, Danilo Jimenez Rezende, and Daan Wierstra. Variational intrinsic control. In *International Conference on Learning Representations Workshop*, 2017.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

Alexander I Hsu and Eric A Yttri. B-soid, an open-source unsupervised algorithm for identification and fast prediction of behaviors. *Nature communications*, 12(1):5188, 2021.

Arnim Jenett, Gerald M. Rubin, Teri-T. B. Ngo, David Shepherd, Christine Murphy, Heather Dionne, Barret D. Pfeiffer, Amanda Cavallaro, Donald Hall, Jennifer Jeter, Nirmala Iyer, Dona Fetter, Joanna H. Hausenfluck, Hanchuan Peng, Eric T. Trautman, Robert R. Svirskas, Eugene W. Myers, Zbigniew R. Iwinski, Yoshinori Aso, Gina M. DePasquale, Adrianne Enos, Phuson Hulamm, Shing Chun Benny Lam, Hsing-Hsi Li, Todd R. Laverty, Fuhui Long, Lei Qu, Sean D. Murphy, Konrad Rokicki, Todd Safford, Kshiti Shaw, Julie H. Simpson, Allison Sowell, Susana Tae, Yang Yu, and Christopher T. Zugates. A GAL4-driver line resource for Drosophila neurobiology. *Cell Reports*, 2(4):991–1001, 2012. doi: 10.1016/j.celrep.2012.09.011.

Harris S. Kaplan, Oriana Salazar Thula, Niklas Khoss, and Manuel Zimmer. Nested neuronal dynamics orchestrate a behavioral hierarchy across timescales. *Neuron*, 105(3):562–576.e9, 2020. doi: 10.1016/j.neuron.2019.10.037.

Alexander Y. Katsov, Limor Freifeld, Mark Horowitz, Seppe Kuehn, and Thomas R. Clandinin. Dynamic structure of locomotor behavior in walking fruit flies. *eLife*, 6:e26410, 2017. doi: 10.7554/eLife.26410.

Alexander J. E. Kell, Daniel L. K. Yamins, Erica N. Shook, Sam V. Norman-Haignere, and Josh H. McDermott. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3):630–644.e16, 2018. doi: 10.1016/j.neuron.2018.03.044.

Reece Keller, Alyn Kirsch, Felix Pei, Xaq Pitkow, Leo Kozachkov, and Aran Nayebi. Intrinsic goals for autonomous agents: Model-based exploration in virtual zebrafish predicts ethological behavior and whole-brain dynamics. In *Advances in Neural Information Processing Systems*, volume 38, pp. 40321–40352, 2025. doi: 10.52202/085713-1205.

Kuno Kim, Megumi Sano, Julian De Freitas, Nick Haber, and Daniel Yamins. Active world model learning with progress curiosity, 2020. URL `https://arxiv.org/abs/2007.07853`.

Alexander S. Klyubin, Daniel Polani, and Chrystopher L. Nehaniv. Keep your options open: An information-based driving principle for sensorimotor systems. *PLoS ONE*, 3(12):e4018, 2008. doi: 10.1371/journal.pone.0004018.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Hao Liu and Pieter Abbeel. Behavior from the void: Unsupervised active pre-training. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18459–18473, 2021.

Victor Lobato-Rios, Shravan Tata Ramalingasetty, Pembe Gizem Özdil, Jonathan Arreguit, Auke Jan Ijspeert, and Pavan Ramdya. NeuroMechFly, a neuromechanical model of adult Drosophila melanogaster. *Nature Methods*, 19(5):620–627, May 2022. doi: 10.1038/s41592-022-01466-7. URL `https://doi.org/10.1038/s41592-022-01466-7`.

Kevin Mann, Stephane Deny, Surya Ganguli, and Thomas R. Clandinin. Coupling of activity, metabolism and behaviour across the Drosophila brain. *Nature*, 593(7858):244–248, 2021. doi: 10.1038/s41586-021-03497-0.

Charalampos Mantziaris, Till Bockemühl, and Ansgar Büschges. Central pattern generating networks in insect locomotion. *Developmental Neurobiology*, 80(1-2):16–30, 2020. doi: 10.1002/dneu.22738.

João C. Marques, Simone Lackner, Rita Félix, and Michael B. Orger. Structure of the zebrafish locomotor repertoire revealed with unsupervised behavioral clustering. *Current Biology*, 28(2):181–195.e5, 2018. doi: 10.1016/j.cub.2017.12.002.

Alexander Mathis, Pranav Mamidanna, Kevin M Cury, Taiga Abe, Venkatesh N Murthy, Mackenzie Weygandt Mathis, and Matthias Bethge. Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nature neuroscience*, 21(9):1281–1289, 2018.

Josh Merel*, Diego Aldarondo*, Jesse Marshall*, Yuval Tassa, Greg Wayne, and Bence Ölveczky. Deep neuroethology of a virtual rodent. *International Conference on Learning Representations*, 2020.

Faisal Mohamed, Catherine Ji, Benjamin Eysenbach, and Glen Berseth. Temporal representations for exploration: Learning complex exploratory behavior without extrinsic rewards. In *International Conference on Learning Representations*, 2026.

Shakir Mohamed and Danilo Jimenez Rezende. Variational information maximisation for intrinsically motivated reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 28, pp. 2125–2133, 2015.

Richard J. D. Moore, Gavin J. Taylor, Angelique C. Paulk, Thomas Pearson, Bruno van Swinderen, and Mandyam V. Srinivasan. FicTrac: A visual method for tracking spherical motion and generating fictive animal paths. *Journal of Neuroscience Methods*, 225:106–119, 2014. doi: 10.1016/j.jneumeth.2014.01.010.

Aran Nayebi, Alexander Attinger, Malcolm G. Campbell, Kiah Hardcastle, Isabel I. C. Low, Caitlin S. Mallory, Gabriel C. Mel, Ben Sorscher, Alex H. Williams, Surya Ganguli, Lisa M. Giocomo, and Daniel L. K. Yamins. Explaining heterogeneity in medial entorhinal cortex with task-driven neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 12167–12179, 2021.

Aran Nayebi, Nathan C. L. Kong, Chengxu Zhuang, Justin L. Gardner, Anthony M. Norcia, and Daniel L. K. Yamins. Mouse visual cortex as a limited resource system that self-learns an ecologically-general representation. *PLOS Computational Biology*, 19(10):e1011506, 2023. doi: 10.1371/journal.pcbi.1011506.

Pembe Gizem Özdil, Jonathan Arreguit, Clara Scherrer, Femke Hurtak, Auke Ijspeert, and Pavan Ramdya. Centralized brain networks controlling antennal grooming coordination. *Nature Communications*, 2026.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2778–2787. PMLR, 2017.

Deepak Pathak, Dhiraj Gandhi, and Abhinav Gupta. Self-supervised exploration via disagreement. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5062–5071. PMLR, 2019.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85): 2825–2830, 2011.

Talmo D Pereira, Nathaniel Tabris, Arie Matsliah, David M Turner, Junyu Li, Shruthi Ravindranath, Eleni S Papadoyannis, Edna Normand, David S Deutsch, Z Yan Wang, et al. Sleap: A deep learning system for multi-animal pose tracking. *Nature methods*, 19(4):486–495, 2022.

Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021.

Blake A. Richards, Timothy P. Lillicrap, Philippe Beaudoin, Yoshua Bengio, Rafal Bogacz, Amelia Christensen, Claudia Clopath, Rui Ponte Costa, Archy de Berker, Surya Ganguli, Colleen J. Gillon, Danijar Hafner, Adam Kepecs, Nikolaus Kriegeskorte, Peter Latham, Grace W. Lindsay, Kenneth D. Miller, Richard Naud, Christopher C. Pack, Panayiota Poirazi, Pieter Roelfsema, João Sacramento, Andrew Saxe, Benjamin Scellier, Anna C. Schapiro, Walter Senn, Greg Wayne, Daniel Yamins, Friedemann Zenke, Joel Zylberberg, Denis Therien, and Konrad P. Kording. A deep learning framework for neuroscience. *Nature Neuroscience*, 22(11):1761–1770, 2019. doi: 10.1038/s41593-019-0520-2.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986. doi: 10.1038/323533a0.

Abraham Savitzky and Marcel J. E. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36(8):1627–1639, 1964. doi: 10.1021/ac60214a047.

John Schulman, Philipp Moritz, Sergey Levine, Michael I. Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations*, 2016.

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

Younggyo Seo, Lili Chen, Jinwoo Shin, Honglak Lee, Pieter Abbeel, and Kimin Lee. State entropy maximization with random encoders for efficient exploration. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9443–9454. PMLR, 2021.

Archit Sharma, Shixiang Gu, Sergey Levine, Vikash Kumar, and Karol Hausman. Dynamics-aware unsupervised discovery of skills. In *International Conference on Learning Representations*, 2020.

Pranav Shyam, Wojciech Jaśkowski, and Faustino Gomez. Model-based active exploration. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5779–5788. PMLR, 2019.

David Sussillo, Mark M. Churchland, Matthew T. Kaufman, and Krishna V. Shenoy. A neural network that finds a naturalistic solution for the production of muscle activity. *Nature Neuroscience*, 18(7):1025–1033, 2015. doi: 10.1038/nn.4042.

Stas Tiomkin, Christoph Salge, and Daniel Polani. Process empowerment for robust intrinsic motivation. *Journal of Physics: Complexity*, 6(3):035011, 2025. doi: 10.1088/2632-072X/adf2ec.

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*, pp. 5026–5033. IEEE, 2012. doi: 10.1109/IROS.2012.6386109.

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 Alfred Stimpfling, Thomas Ka Chung Lam, Pembe Gizem Özdil, Louise Genoud, Femke Hurtak, and Pavan Ramdya. Neuromechfly v2: simulating embodied sensorimotor control in adult drosophila. *Nature Methods*, 21(12):2353–2362, November 2024. ISSN 1548-7105. doi: 10.1038/s41592-024-02497-y. URL http://dx.doi.org/10.1038/s41592-024-02497-y.

Alexander B. Wiltschko, Matthew J. Johnson, Giuliano Iurilli, Ralph E. Peterson, Jesse M. Katon, Stan L. Pashkovski, Victoria E. Abraira, Ryan P. Adams, and Sandeep Robert Datta. Mapping sub-second structure in mouse behavior. *Neuron*, 88(6):1121–1135, 2015. doi: 10.1016/j.neuron.2015.11.031.

Daniel L. Yamins, Ha Hong, Charles Cadieu, and James J. DiCarlo. Hierarchical modular optimization of convolutional networks achieves representations similar to macaque IT and human ventral stream. In *Advances in Neural Information Processing Systems*, volume 26, pp. 3093–3101, 2013.

Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, 2016.

Denis Yarats, Rob Fergus, Alessandro Lazaric, and Lerrel Pinto. Reinforcement learning with prototypical representations. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11920–11931. PMLR, 2021.

Eunice Yiu, Kelsey Allen, Shiry Ginosar, and Alison Gopnik. Empowerment gain and causal model construction: Children and adults are sensitive to controllability and variability in their causal interventions. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 384(2320):20250003, 2026. doi: 10.1098/rsta.2025.0003.

Ruihan Zhao, Kevin Lu, Pieter Abbeel, and Stas Tiomkin. Efficient empowerment estimation for unsupervised stabilization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=u2YNJPcQlwq.

Chongyi Zheng, Jens Tuyls, Joanne Peng, and Benjamin Eysenbach. Can a MISL fly? analysis and ingredients for mutual information skill learning. In *International Conference on Learning Representations*, 2025.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005. doi: 10.1111/j.1467-9868.2005.00503.x.

Index	Primitive	d_L	d_R	Controller
0	Left	-1.0	1.2	CPG turning
1	Forward	1.0	1.0	CPG walking
2	Right	1.2	-1.0	CPG turning
3	Stop	0	0	CPG, amplitude decays to zero
4	Groom	0	0	Joint replay overrides CPG
5	Hard left	-1.2	1.2	CPG turning
6	Hard right	1.2	-1.2	CPG turning

Table 1: Seven discrete behavioral primitives. The sign of d sets the stepping direction of that side's legs (negative is backward) and $|d|$ sets the stepping amplitude, so left and right turns are pivot turns.

A Additional Experimental Details

A.1 Sensory Observations

Proprioception. The proprioceptive observation is

$$\mathbf{o}_t^{\text{prop}} = [\mathbf{q},\ \dot{\mathbf{q}},\ h,\ \mathbf{g}_{\text{body}},\ \mathbf{v}_{\text{body}},\ \boldsymbol{\omega}_{\text{body}}] \in \mathbb{R}^{100}, \tag{1}$$

where $\mathbf{q}, \dot{\mathbf{q}} \in \mathbb{R}^{45}$ are the angles (rad) and angular velocities (rad/s) of the 45 position-actuated degrees of freedom, h is the height of the body root (mm), $\mathbf{g}_{\text{body}} \in \mathbb{R}^3$ is gravity in the body frame, and $\mathbf{v}_{\text{body}}, \boldsymbol{\omega}_{\text{body}} \in \mathbb{R}^3$ are the linear (mm/s) and angular (rad/s) velocities of the body root. The 45 degrees of freedom are seven per leg (thorax–coxa yaw, pitch, and roll; coxa–trochanter pitch and roll; femur–tibia pitch; tibia–tarsus pitch), ordered LF, LM, LH, RF, RM, RH, followed by neck pitch, roll, and yaw (Fig. 1d). No contact or torque sensors are used, and non-finite values are set to zero.

Vision. Each compound eye is rendered by a fisheye camera (512×450 pixels, 157° field of view) and read out as 721 ommatidia, 505 yellow-type and 216 pale-type (Lobato-Rios et al., 2022; Wang-Chen et al., 2024). Each ommatidium takes the mean of the green (yellow-type) or blue (pale-type) channel over its pixels, scaled to $[0, 1]$. For computational efficiency, the ommatidial map is resampled from 512×450 to a 32×32 grid by nearest-neighbour sampling at each grid-pixel centre; grid pixels outside the lattice are zero, and 623 of the 721 ommatidia are represented. The visual input is a $2 \times 2 \times 32 \times 32$ tensor (eye $\times$ ommatidium type $\times$ height $\times$ width; Fig. 1e).

Olfaction. Odor intensity at position $\mathbf{x}$ is $I(\mathbf{x}) = \sum_k A_k \|\mathbf{x} - \mathbf{x}_k\|^{-2}$, summed over sources k with peak intensity $A_k = 1$. Intensity is sampled at the left and right maxillary palps and antennae and combined on each side as $0.1\, I_{\text{palp}} + 0.9\, I_{\text{antenna}}$, giving a two-dimensional left/right observation (Fig. 1c).

Normalization. Odor and proprioceptive observations are normalized with running means and variances and clipped at ± 10. Visual input is not normalized.

A.2 Hierarchical Locomotor Control

Every 0.1 s, the policy selects one of seven behavioral primitives (Table 1; Fig. 2b). Six set a pair of descending signals (d_L, d_R), applied to the three left and three right legs respectively; the seventh triggers a grooming replay.

Central pattern generator. Locomotor primitives drive a six-oscillator CPG, a GPU reimplementation of the FlyGym hybrid turning controller (Lobato-Rios et al., 2022; Wang-Chen et al., 2024) without its stumbling and retraction correction rules. For leg i with side signal $d_i \in \{d_L, d_R\}$, the phase θ_i and amplitude r_i evolve as

$$\dot{\theta}_i = 2\pi\nu\, s(d_i) + \sum_j r_j\, w_{ij} \sin(\theta_j - \theta_i - \varphi_{ij}), \qquad \dot{r}_i = \alpha\,(|d_i| - r_i), \tag{2}$$

Component	Architecture
Visual encoder (shared across eyes)	Conv($2 \to 32$, 5×5, stride 2) – Conv($32 \to 64$, 3×3, stride 2) – Conv($64 \to 64$, 3×3, stride 2) – FC($1024 \to 64$)
Visual embedding	$2 \times 64 = 128$
Olfactory encoder	MLP $2 \to 32 \to 16$, concatenated with raw input (18)
Proprioceptive encoder	MLP $100 \to 128 \to 64$, concatenated with raw input (164)
Base features ϕ_t	310
Actor and critic LSTMs	One layer, 256 units each
Actor head	$256 \to 128 \to 128 \to 7$, categorical
Critic head	$256 \to 128 \to 128 \to 1$
Activations	ReLU (encoders), tanh (heads)
Optimizer	Adam, learning rate 3×10^{-4}, $\epsilon = 10^{-5}$

Table 2: Policy network architecture.

where $s(d) = -1$ if $d < 0$ and $+1$ otherwise, $\nu = 12\,\text{Hz}$, $\alpha = 20$, $w_{ij} = 10$ for coupled legs, and the phase biases $\varphi_{ij} \in \{0, \pi\}$ place $\{\text{LF}, \text{LH}, \text{RM}\}$ and $\{\text{LM}, \text{RF}, \text{RH}\}$ in antiphase tripods. The equations are integrated with forward Euler at the physics timestep (10^{-4} s). Joint targets are

$$\mathbf{q}_i^* = \bar{\mathbf{q}}_i + r_i \left(\boldsymbol{\psi}_i(\theta_i) - \bar{\mathbf{q}}_i \right), \tag{3}$$

where $\boldsymbol{\psi}_i$ is a periodic cubic spline through a single recorded step of untethered fly walking for leg i, and $\bar{\mathbf{q}}_i = \boldsymbol{\psi}_i(\pi)$ is the neutral pose. Joints are position-controlled ($k_p = 45$, force range ± 65). Leg adhesion (gain 40) is on during stance and off during swing, with the swing interval extended by $\pi/4$. At each reset, phases are drawn from $\mathcal{U}[0, 2\pi)$ and amplitudes are zero. Under the stop primitive, amplitudes decay to zero with time constant $1/\alpha = 50\,\text{ms}$ while phases continue to advance.

Grooming. The groom primitive replays a 35-frame (0.35 s) clip of a recorded, optogenetically evoked grooming bout (aJO-GAL4 $>$ CsChrimson) at 100 Hz, driving 17 degrees of freedom: the three neck joints and the seven joints of each front leg (Özdil et al., 2026). The CPG receives a stop signal, the middle and hind legs are held at their pose at groom onset with adhesion on for stability, and the body root is held fixed for the duration of the bout. To avoid abrupt movement, the replay blends in linearly from the onset pose over 0.05 s; on the next non-groom action, control switches back to the CPG. Grooming is not a macro-action: each groom action lasts one 0.1 s decision, consecutive groom actions continue the clip (looping every 35 frames), and the policy can end a bout at any decision.

A.3 Policy Architecture

Policies use the recurrent actor–critic of sb3-contrib RecurrentPPO (Raffin et al., 2021), with a shared multimodal feature extractor and separate actor and critic LSTMs (Hochreiter & Schmidhuber, 1997) (Fig. 2a; Table 2). A CNN with weights shared across the two eyes (Fukushima, 1980; LeCun et al., 1998) produces a 64-dimensional embedding per eye, for a 128-dimensional visual representation. MLPs (Rumelhart et al., 1986) embed odor (16 dimensions) and proprioception (64 dimensions). The odor embedding is concatenated with the raw 2-dimensional odor observation and the proprioceptive embedding with the raw 100-dimensional proprioceptive state, yielding modality representations of 128, 18, and 164 dimensions and a 310-dimensional base feature vector ϕ_t. This is passed to the actor and critic LSTMs (256 hidden units each) and to the intrinsic drive module (IDM). Separate actor and critic MLP heads follow the LSTMs, and the actor outputs a categorical distribution over the seven primitives.

For skill-conditioned methods, the skill z is appended to ϕ_t, without normalization, at the input of both LSTMs. The IDM receives only ϕ_t, never z, which prevents the skill discriminator from being solved trivially. Linear and convolutional layers use orthogonal initialization (gain $\sqrt{2}$; 0.01 for the action head and 1 for the value head), and the LSTMs use PyTorch's default initialization. The policy without a skill input has 1,407,576 parameters; a skill of dimension d_z adds $2{,}048\,d_z$.

Method	Family	Skill prior (resampling)	Trainable IDM parameters
ICM	Curiosity	–	543,495
RND	Curiosity	–	276,992
Disagreement	Curiosity	–	1,134,350
γ-Progress	Curiosity	–	226,870
MAX	Curiosity	–	40,603,008
NGU	Curiosity	–	449,063
C-TeC	Coverage	–	325,505
RE3	Coverage	–	0 (fixed random encoder)
APT	Coverage	–	609,287
Proto-RL	Coverage	–	507,136
DIAYN	Skill	$\mathrm{Cat}(50)$ (per episode)	198,650
VIC	Skill	$\mathrm{Cat}(12)$ (50 decisions)	227,852
DADS	Skill	$\mathcal{U}([-1,1]^2)$ (per episode)	740,972
rDADS	Skill	$\mathcal{U}([-1,1]^2)$ (per episode)	1,006,600
VALOR	Skill	$\mathrm{Cat}(12)$ (50 decisions)	1,067,020
CSF	Skill	$\mathcal{U}(S^3)$ (per episode)	1,372,164
Forward empowerment	Empowerment	–	617,420
Empowerment gain	Empowerment	–	3,124,150
Potential empowerment	Empowerment	–	620,470
Process empowerment	Empowerment	–	620,470

Table 3: Intrinsic drive methods. Parameter counts exclude frozen targets, lagged copies, and memories.

A.4 Intrinsic Drive Methods

Shared implementation. All 20 methods share one pipeline. The IDM receives the base features ϕ_t, computed by the policy's feature extractor without gradient flow into the extractor, along with the one-hot action a_t and the next features ϕ_{t+1}. Observations are normalized with statistics frozen at the start of each rollout. Intrinsic rewards are computed online at every decision using the IDM parameters from the previous rollout. After each PPO rollout, the IDM takes one Adam step (learning rate 10^{-4}) on a minibatch sampled without replacement from that rollout (256 transitions unless stated otherwise), giving approximately 244 IDM updates over training; VIC and VALOR instead train on completed skill segments, which may span rollouts. The PPO reward is the intrinsic reward alone, with no extrinsic term and no PPO reward normalization; where noted, the intrinsic reward is divided by its running root-mean-square (RMS). Unless stated otherwise, networks use ReLU activations and Xavier-uniform initialization. Table 3 summarizes the methods. The descriptions below state each method as implemented and note departures from the original publication where we identified them.

Curiosity and prediction error

ICM (Pathak et al., 2017). An encoder g ($310 \to 256 \to 256$) is trained jointly with a forward model f ($263 \to 256 \to 256 \to 256$) and an inverse model ($512 \to 256 \to 256 \to 7$), with loss $0.8\,\mathcal{L}_{\mathrm{inv}} + 0.2\,\mathcal{L}_{\mathrm{fwd}}$. The reward is $r_t = \frac{\eta}{2}\|f(g(\phi_t), a_t) - g(\phi_{t+1})\|^2$ with $\eta = 0.01$. Unlike the original, the encoder operates on the policy's features rather than learning from pixels end to end.

RND (Burda et al., 2019). A predictor $\hat{f}$ regresses a frozen random target f (both $310 \to 256 \to 256 \to 512$). The reward is the squared error $\|\hat{f}(\phi_t) - f(\phi_t)\|^2$, averaged over dimensions and divided by a running mean of the absolute error. The original normalizes by the standard deviation of intrinsic returns and uses separate intrinsic and extrinsic value heads with non-episodic intrinsic returns; here there is one value head and returns are episodic.

Disagreement (Pathak et al., 2019). Five forward models ($317 \to 256 \to 256 \to 310$), each trained with bootstrap-resampled mean squared error to predict ϕ_{t+1} from (ϕ_t, a_t). The reward is the variance of their predictions, averaged over dimensions, without normalization.

γ-Progress (Kim et al., 2020). A forward model f ($317 \to 256 \to 256 \to 310$) and a lagged copy f_{old}, updated as an exponential moving average with rate 0.01 per IDM update ($\gamma = 0.99$). The reward is the reduction in prediction error of ϕ_{t+1}, $r_t = \mathcal{L}(f_{\text{old}}) - \mathcal{L}(f)$, and can be negative.

MAX (Shyam et al., 2019). Thirty-two diagonal-Gaussian forward models (four hidden layers of 512 units, swish activations, $\log \sigma^2$ bounded to $[-5, -1]$), trained by negative log-likelihood on bootstrap samples. The reward is the Jensen–Rényi (quadratic) divergence between the ensemble's predictions, with variance tempering $\lambda = 0.1$, clamped at zero. The original plans exploration inside the learned model; here the divergence is used as a model-free reward.

NGU (Badia et al., 2020). The reward is $r_t = 0.3\, r_t^{\text{epi}} \cdot \text{clip}\big(1 + (e_t^{\text{RND}} - \mu)/\sigma,\ 1,\ 5\big)$, where e_t^{RND} is an RND error with running mean μ and standard deviation σ. The episodic term is $r_t^{\text{epi}} = 1/\big(\sqrt{\sum_k K(e_t, e_k)} + 10^{-3}\big)$ over the 10 nearest neighbours in an episodic memory reset at each episode, set to zero above a similarity threshold of 8 as in the original, with kernel $K(x, y) = \epsilon/\big(\max(d^2(x, y)/\bar{d}^2 - \xi,\ 0) + \epsilon\big)$, $\epsilon = 10^{-4}$, $\xi = 8 \times 10^{-3}$. Embeddings e_t ($310 \to 256 \to 256 \to 32$) are trained with an inverse model ($64 \to 256 \to 256 \to 7$); the RND networks are $310 \to 256 \to 256 \to 256$. We use a single mixing coefficient rather than the original family of policies.

State-space coverage

C-TeC (Mohamed et al., 2026). A state–action encoder ψ ($317 \to 256 \to 256 \to 64$) and a future-state encoder ν ($310 \to 256 \to 256 \to 64$), both with ℓ_2-normalized outputs, are trained with an InfoNCE loss using negative ℓ_1 distance as the logit, 256 in-batch negatives, and a learned temperature. The reward is $r_t = \|\psi(\phi_t, a_t) - \nu(\phi_{t+1+\Delta})\|_1$ for a single offset Δ drawn from a geometric distribution ($\gamma = 0.99$) truncated to the episode. Rewards and advantages are recomputed at the end of each rollout.

RE3 (Seo et al., 2021). A fixed, randomly initialized encoder h ($310 \to 256 \to 256 \to 50$) embeds ϕ_t. The reward is the distance to the 53rd nearest neighbour in a memory of 2,048 embeddings sampled from recent training data (re-encoded every rollout), divided by its running RMS. The original applies a random convolutional encoder to pixels with a $\log(1 + d)$ reward and a decaying coefficient; here the random encoder acts on the policy's features, with no log and a constant coefficient of 1.

APT (Liu & Abbeel, 2021). An encoder g ($310 \to 256 \to 256 \to 256$) is trained with forward ($263 \to 256 \to 256 \to 256$) and inverse ($512 \to 256 \to 256 \to 7$) losses. The reward is $r_t = \log\big(1 + \max(\bar{d}_{16}/\bar{d} - 5 \times 10^{-4},\ 0)\big)$, where $\bar{d}_{16}$ is the mean distance from the ℓ_2-normalized embedding of ϕ_{t+1} to its 16 nearest neighbours in a memory of 2,048 embeddings (re-embedded after every update), and $\bar{d}$ is a running mean distance.

Proto-RL (Yarats et al., 2021). An encoder ($310 \to 256 \to 256 \to 256$), projector ($256 \to 256 \to 128$), and predictor ($128 \to 512 \to 128$) are trained with a swapped-prediction loss against 512 prototypes (Sinkhorn with 3 iterations, temperature 0.1) and an exponential-moving-average target network (rate 0.95). The reward is the distance from the embedding of ϕ_{t+1} to its third-nearest neighbour in a queue of 2,048 candidate embeddings. Unlike the original, the representation is trained online alongside PPO rather than in a separate pre-training phase, without data augmentation, and the queue is not re-embedded when the encoder changes.

Skill discovery

DIAYN (Eysenbach et al., 2019). Skills are one-hot, $z \sim \text{Cat}(50)$ uniform, sampled once per training episode. A discriminator $q(z \mid \phi)$ ($310 \to 300 \to 300 \to 50$) is trained with cross-entropy (batch 128), and the reward is $r_t = \log q(z \mid \phi_t) - \log p(z) = \log q(z \mid \phi_t) + \log 50$. The original maximizes action entropy through SAC; here the only action-entropy term is PPO's entropy bonus (coefficient 0.001), and the discriminator scores the pre-transition features ϕ_t.

VIC (Gregor et al., 2017). Skills $z \sim \mathrm{Cat}(12)$, uniform, fixed over 50-decision segments. A posterior $q(z \mid \phi_0, \phi_T)$ ($620 \to 256 \to 256 \to 12$) is trained with cross-entropy on 512 completed segments, and the reward $\log q(z \mid \phi_0, \phi_T) + \log 12$ is given only at a segment's final step. The prior is fixed and segments have fixed length rather than learned termination.

DADS (Sharma et al., 2020). Skills $z \sim \mathcal{U}([-1, 1]^2)$, sampled once per episode. A skill-dynamics model $q(\phi' \mid \phi, z)$, a single diagonal Gaussian ($312 \to 512 \to 512 \to 620$), predicts the next features. The reward is

$$r_t = \log q(\phi_{t+1} \mid \phi_t, z) - \log \frac{1}{L} \sum_{i=1}^{L} q(\phi_{t+1} \mid \phi_t, z_i), \qquad z_i \sim p(z),\ L = 100, \tag{4}$$

divided by its running RMS. The original uses a mixture-of-experts dynamics model that predicts state differences; here the model is a single Gaussian over absolute next features.

rDADS. A reverse formulation of DADS that infers the skill from transitions. The reward is $r_t = \log q_\psi(z \mid \phi_t, \phi_{t+1}) - \log q_\omega(z \mid \phi_t)$, with Gaussian posteriors q_ψ ($620 \to 512 \to 512 \to 4$) and q_ω ($310 \to 512 \to 512 \to 4$) over $z \sim \mathcal{U}([-1, 1]^2)$, trained by negative log-likelihood; the reward is divided by its running RMS.

VALOR (Achiam et al., 2018). Skills $z \sim \mathrm{Cat}(12)$, fixed over 50-decision segments. The decoder embeds the ten feature differences between eleven evenly spaced states of a segment ($310 \to 256 \to 256$ each) and maps their concatenation ($2560 \to 256 \to 512 \to 256 \to 12$) to a posterior over z. The reward $\log q(z \mid \tau) + \log 12$ is given at a segment's final step.

CSF (Zheng et al., 2025). Skills are drawn uniformly from the unit sphere in $\mathbb{R}^4$, once per episode. A representation f ($310 \to 1024 \to 1024 \to 4$) is trained to minimize $-\mathbb{E}\big[\Delta f^\top z - \xi \log \sum_{j=1}^{256} \exp(\Delta f^\top z_j)\big]$, where $\Delta f = f(\phi_{t+1}) - f(\phi_t)$, $\xi = 5$, and the negatives z_j are drawn from the prior. The reward is $r_t = \Delta f^\top z$. The original trains a successor-feature critic; here PPO's value function is used.

EMPOWERMENT

Forward empowerment. The reward is $r_t = \log q_\theta(\phi_{t+1} \mid \phi_t, a_t) - \log q_\omega(\phi_{t+1} \mid \phi_t)$, with diagonal-Gaussian models (two hidden layers of 256 units; the action is re-concatenated at every layer) trained by maximum likelihood. This estimates the one-step mutual information $I_\pi(A; \Phi' \mid \Phi)$ under the current policy, without maximizing over action distributions.

Potential empowerment. Following Klyubin et al. (2008); Mohamed & Rezende (2015); Zhao et al. (2021), the method learns a latent Gaussian channel. An encoder f maps features to $y \in \mathbb{R}^{64}$, and $y' = f(\phi_{t+1})$ is modeled as Gaussian with a mean $\mu_a(y)$ formed from a drift term and a per-action gain, a shared learned log-variance ℓ, and a decoder $64 \to 310$ for reconstruction (two hidden layers of 256 units). The loss is $\frac{1}{2}\,\mathbb{E}\big[(y' - \mu_a)^2 e^{-\ell} + \ell\big] + \|\mathrm{dec}(f(\phi_t)) - \phi_t\|^2/310$, with gradients flowing through both y' and μ_a. The reward is the channel capacity $C(\phi_t) = \max_{p(a)} I(A; Y' \mid \phi_t)$ over the seven actions, computed with 32 log-space Blahut–Arimoto iterations and 64 antithetic samples.

Empowerment gain. Uses the channel of potential empowerment with a larger network (five hidden layers of 400 units) and a slowly updated copy (exponential moving average with rate 0.01 per IDM update). The reward is the increase in estimated capacity, $r_t = C_{\mathrm{fast}}(\phi_t) - C_{\mathrm{slow}}(\phi_t)$, divided by its running RMS.

Process empowerment (Tiomkin et al., 2025). Uses the channel of potential empowerment, with reward $r_t = \log q_{a_t}(y_{t+1}) - \log \sum_b \pi(b \mid h_t)\, q_b(y_{t+1})$, where $\pi(\cdot \mid h_t)$ are the policy's action probabilities from the pre-step LSTM state. The reward can be negative.

Method	Hours	Method	Hours
ICM	121.3	DIAYN	140.3
RND	141.7	VIC	121.8
Disagreement	117.8	DADS	150.7
γ-Progress	157.9	rDADS	106.0
MAX	136.6	VALOR	119.9
NGU	161.8	CSF	140.7
C-TeC	112.2	Forward empowerment	159.3
RE3	154.1	Empowerment gain	148.9
APT	146.0	Potential empowerment	111.3
Proto-RL	145.8	Process empowerment	105.4

Table 4: Recorded training wall hours for the 50-million-decision runs, rounded to 0.1 hour. Values sum the final rollout-end timer reading from each resumed process and are lower bounds on full job time.

A.5 Training Protocol

Environment. We train each policy in a procedurally randomized FlyGym environment (Fig. 1a) that exposes the agent to variation in its visual, olfactory, and proprioceptive inputs. Each arena contains three odor sources and 12 to 50 obstacles, regenerated at every episode. Odor sources are placed uniformly in $[-35, 35]^2$ mm at a height of 2 mm, at least 15 mm from the spawn point and 12 mm from each other. Obstacles are placed uniformly in $[-45, 45]^2$ mm, at least 4 mm from the spawn point and 5 mm from odor sources and from each other. The fly spawns at the origin with zero yaw. The odor sources and obstacles vary the visual and olfactory observations, and physical contact with obstacles further diversifies proprioceptive input. Episodes last 5 s (50 decisions) and terminate early on a physics error or non-finite state, when the body falls below a height threshold for five consecutive decisions, or when the fly leaves $|x|, |y| \leq 60$ mm.

Optimization. Policies are trained with Recurrent PPO (Raffin et al., 2021; Schulman et al., 2017) in $2,048$ parallel environments, each with a distinct random seed, using the settings in Table 5 and generalized advantage estimation (Schulman et al., 2016). Each method is trained for the same budget of 50 million policy decisions, summed across environments (5×10^6 s of simulated experience, approximately 244 PPO updates). The scalar reward used by PPO is supplied entirely by the intrinsic drive module, so differences between trained policies arise from the intrinsic objective rather than from external task reward. LSTM states are reset at episode starts and carried across rollout boundaries; backpropagation through time runs over sequences split at episode boundaries (at most 50 decisions). Truncated episodes are bootstrapped with the value of the final observation, and optimization epochs stop early when the approximate KL divergence exceeds 1.5 times the target. Each method is trained once, with seed 0. We evaluate the checkpoint at the end of the training budget, without selecting checkpoints on any evaluation metric.

Compute and software. Each run used one GPU (NVIDIA L40S or comparable) and 64 CPU cores. Throughput ranged from roughly 40 to 260 policy decisions per second depending on the method. Recorded training wall time was 105–162 hours per method (Table 4). We summed the elapsed times logged at the end of each rollout across resumed processes. These times exclude setup, the final PPO update, interrupted work without a rollout log, and time between jobs, so they are lower bounds on full job time. Software: Python 3.12, FlyGym 2.1.0, MuJoCo 3.9.0 with MuJoCo-Warp 3.9.0.1, Warp 1.14.0, Stable-Baselines3 and sb3-contrib 2.8.0, PyTorch, and JAX 0.4.33.

A.6 Tethered-Ball Evaluation

Assay. In the biological experiment, head-fixed flies walk on a 9 mm air-suspended foam ball in darkness, and FicTrac (Moore et al., 2014) recovers the ball's angular and forward velocity from an infrared-visible painted pattern (Brezovec et al., 2023). In simulation, we read these velocities directly from the ball. The ball is a 9 mm-diameter sphere with the density of the experimental foam (0.240 mg/mm^3; mass ≈ 0.092 g) (General Plastics Manufacturing Company) on an undamped ball

Setting	Value
Recurrent PPO	
Parallel environments	2,048
Decisions per environment per rollout	100
Transitions per rollout	204,800
Minibatch size	4,096 (50 minibatches per epoch)
Epochs per rollout	10 (early stop at $1.5\times$ target KL)
Learning rate	3×10^{-4}
Discount γ	0.99
GAE λ	0.95
Clip range	0.2
Target KL	0.03
Entropy coefficient	0.001
Value-function coefficient	0.5
Maximum gradient norm	0.5
Training budget	50×10^6 policy decisions (≈ 244 rollouts)
Observation normalization	Odor and proprioception, clipped at 10
Reward normalization	None
Training seeds	1 per method
Simulation	
Physics timestep	10^{-4} s
Decision interval	0.1 s (1,000 physics steps)
Episode length	5 s (50 decisions)
Extrinsic reward scale	0
Intrinsic reward scale	1

Table 5: PPO training and environment hyperparameters.

joint. The fly's body root is fixed in place above the ball, and only the tibiae and tarsi contact it (friction coefficient 2.0). Because the root is fixed, the body linear- and angular-velocity entries of the proprioceptive observation are zero. Forward velocity is $v = -R\,\omega_y^{\mathrm{ball}}$ (mm/s), where R is the ball radius, and angular velocity is the ball's yaw rate (deg/s, positive for left turns); both are recorded at 50 Hz after an unrecorded settling period.

Sensory masking. To match the darkness and control for odor in the biological experiment, the raw visual and olfactory observations are set to zero.

Rollouts. Each frozen policy runs nine 30-minute rollouts (18,000 decisions each), matching the nine flies and 270 minutes of the biological dataset. Rollouts differ in their initial CPG phases (seeds 0–8) and in the actions sampled from the policy; actions are sampled rather than chosen greedily. The LSTM state is initialized once and never reset. For DIAYN and CSF, one skill is sampled per rollout and held for the full 30 minutes; VIC and VALOR resample every 5 s (50 decisions), and DADS and rDADS every 10 s (100 decisions).

Recorded activations. At every decision we record the input embedding ϕ_t, the actor and critic LSTM hidden states, and the policy and value MLP hidden representations. Activations are averaged over consecutive 0.532 s bins from rollout onset, matching the fly volume rate (1.88 Hz), giving 3,384 bins per rollout.

A.7 Controls

Random initialization. Each control policy is a random initialization of a source policy so the LSTM input has the same size as in the trained policy. Random-DIAYN, Random-CSF, Random-VIC (the reference for VIC and VALOR), and Random-DADS (the reference for DADS and rDADS) retain the corresponding skill input; Random-nonskill is derived from a run of a method without a skill input. Controls are evaluated exactly as trained policies.

Fly	N_L	N_R	$\log(N_R/N_L)$
087	222	519	0.85
089	38	33	−0.14
094	140	2	−4.25
097	47	105	0.80
098	178	291	0.49
099	27	543	3.00
100	364	601	0.50
101	276	219	−0.23
105	70	111	0.46
Total	1,362	2,424	0.17 ± 1.90 (mean $\pm$ s.d.)

Table 6: Turns per fly in the biological dataset.

Extrinsic reward. The extrinsic baseline uses the same architecture and PPO settings, but is trained on an odor-seeking task instead of an intrinsic reward. Each arena has one odor source at $x \in [20, 30]$ mm, $y \in [-6, 6]$ mm and 3 to 5 obstacles, and the reward is $r_t = (d_{t-1} - d_t) - 0.01 + 10\mathbb{1}[d_t < 2\,\mathrm{mm}]$, where d_t is the distance to the source. Episodes terminate on reaching the source, and reward normalization is enabled. The baseline was trained for 10-M decision steps as it had saturated the success rate for the task. This final checkpoint was used for evaluation.

A.8 Biological Data

We use the whole-brain recordings of Brezovec et al. (2023), released on the DANDI Archive as dataset 000727 (Brezovec et al., 2024): nine flies (IDs 087, 089, 094, 097, 098, 099, 100, 101, 105) and the processed functional GCaMP6f signal. Brain regions are defined by the JFRC2010 (Jenett et al., 2012) anatomical template, warped to the JRC2018F (Bogovic et al., 2020) template with ANTs SyN (Avants et al., 2011; 2009). The supercluster that Brezovec et al. (2023) identify as the pre-motor center is not available in the public release, so we use the whole anatomically defined IPS in each hemisphere.

Turns are re-detected from the FicTrac yaw rate as described in Appendix A.9, discarding the first and last 40 s of each 30-minute session. This yields 3,786 turns (1,362 left, 2,424 right), matching the count reported by Brezovec et al. (2023). Turn direction is strongly imbalanced within some flies (Table 6); Brezovec et al. (2023) attribute part of the per-fly rotational skew to small asymmetries in how each fly was mounted on the ball. Turns are pooled across flies. The stationary subset consists of the 27 left and 27 right turns with the least movement (mean $|v|$ from -15 to -1 s), pooled across flies.

A.9 Turn Detection and Behavioral Statistics

For agents and flies, angular and forward velocity at 50 Hz are smoothed with a Savitzky–Golay filter (Savitzky & Golay, 1964) (25 samples, i.e. 0.5 s; polynomial order 3). Turns are local maxima of $|\omega|$ exceeding $200°$/s, following Brezovec et al. (2023); the turn time is the time of the peak. For flies, peaks in the first and last 40 s of each session are discarded. To compare behavior, we compute the number of left and right turns, N_L and N_R, for each rollout or fly, and the turning bias $\log(N_R/N_L)$. Turning bias is reported only when $N_L > 0$ and $N_R > 0$; otherwise it is undefined, and such rollouts are omitted from bias summaries but retained in turn counts.

A.10 Decoding

Features. For each turn, features are averaged over a window relative to the turn: -15 to -8 s (before the turn) or -0.4 to $+0.4$ s (during the turn). For agents, the average is over the 0.532 s activation bins whose left edge falls within the window. Candidate internal layers are the actor LSTM hidden state, the critic LSTM hidden state, the policy MLP hidden representation, the value MLP hidden representation, and three concatenations (actor LSTM with policy MLP; critic LSTM with value MLP; all five recorded representations). The proprioceptive observation, with vision and odor zeroed, is decoded in the same way. For flies, layers are the window-averaged activity of every

voxel in the left and right IPS, concatenated; unlike Brezovec et al. (2023), we neither restrict the analysis to the pre-motor supercluster nor use the interhemispheric difference. Behavioral features are the window means of angular and forward velocity.

Classifier. We use logistic regression with an elastic-net penalty (Zou & Hastie, 2005; Pedregosa et al., 2011) and the saga solver. Hyperparameters were chosen by nested cross-validation on the fly IPS data over $C \in \{10^{-4}, 10^{-3}, \ldots, 10^{4}\}$ and ℓ_1 ratio $\in \{0.25, 0.5, 0.75\}$, and the most frequently selected configuration ($C = 0.1$, ℓ_1 ratio 0.75) is used for all decoders of agents and flies. Features are standardized with statistics from the training fold only.

Evaluation. Accuracy is estimated by stratified k-fold cross-validation over individual turns, with $k = \min(5, N_{\min})$, where $N_{\min}$ is the smaller class count; methods with fewer than two turns in either class are not scored. We refer to the balanced accuracy (Brodersen et al., 2010) of a neural or behavioral decoder as its alignment,

$$\mathrm{BAcc} = \frac{1}{2}\left(\frac{TP_L}{N_L} + \frac{TP_R}{N_R}\right), \tag{5}$$

where TP_L and TP_R are the numbers of correctly classified left and right turns and N_L and N_R are the numbers of left and right turns. We report the mean $\pm$ s.d. across folds. For each method, the selected layer is the candidate layer with the highest before-turn alignment, and the same layer is used for the during-turn window.

A.11 Representational Analysis

Because the policy has no anatomical analogue of the fly's bilaterally matched superclusters, we cannot reproduce the spatial pairing analysis of Brezovec et al. (2023) directly. Instead, we use PCA to identify low-dimensional axes of model activity and ask whether pairs of components show the same functional relationship to behavior: opposing activity associated with angular velocity and shared activity associated with forward velocity. This is a functional comparison of representational geometry, not an anatomical correspondence between model components and fly hemispheres.

For DIAYN and Random-DIAYN, we fit a 10-component PCA (randomized SVD, seed 0) to the actor LSTM activations pooled across all nine rollouts ($9 \times 3{,}384$ samples; units centered but not standardized), and scale each component to unit variance. Angular and forward velocity are interpolated to the activation timestamps. For every unordered pair (i, j) of components, we compute

$$S_{ij} = |r(\mathrm{PC}_i - \mathrm{PC}_j,\ \omega)| + |r(\mathrm{PC}_i + \mathrm{PC}_j,\ v)|\,, \tag{6}$$

where r is the Pearson correlation over all timepoints, and select the maximizing pair, breaking ties by the smaller of the two correlations. The pair is oriented so that the difference increases with angular velocity. The same search is applied to the randomly initialized policy as a reference. The selected pairs are PCs 2 and 8 for DIAYN and PCs 5 and 10 for Random-DIAYN (Fig. 5). The biological panel (Fig. 5a) is reconstructed from Fig. 1e–f of Brezovec et al. (2023).

B Extended Results

Table 7 reports turn statistics and alignment for all 20 intrinsic drive methods, the randomly initialized controls, the extrinsic-reward control, and the flies. Table 8 gives the before-turn alignment of every candidate layer; "Obs" denotes the proprioceptive observation.

Method	N_L	N_R	Turns/min	Turn bias $\log(N_R/N_L)$	Sel. layer	Internal acc. ($-15..-8$ s)	Behav. acc. ($-15..-8$ s)	Margin	Internal (± 0.4 s)	Behav. (± 0.4 s)	Unscored reason
ICM	0	14997	55.55 ± 1.30	–	–	–	–	–	–	–	only right turns
RND	0	0	0.00 ± 0.00	–	–	–	–	–	–	–	no turns
Disagreement	0	62288	230.70 ± 12.13	–	–	–	–	–	–	–	only right turns
γ-Progress	10320	295	39.32 ± 9.03	-3.78 ± 0.96	A-LSTM+Pol	0.603 ± 0.022	0.724 ± 0.012	-0.122	0.864 ± 0.028	1.000 ± 0.000	–
RE3	179	5379	20.59 ± 1.90	3.47 ± 0.49	A-LSTM (8 tied)‡	0.500 ± 0.000	0.531 ± 0.050	-0.031	0.756 ± 0.019	1.000 ± 0.000	–
APT	252	534	2.91 ± 0.51	0.79 ± 0.27	All	0.516 ± 0.018	0.534 ± 0.031	-0.019	0.964 ± 0.013	1.000 ± 0.000	–
Proto-RL	0	0	0.00 ± 0.00	–	–	–	–	–	–	–	no turns
MAX	4	18486	68.48 ± 33.12	$5.14 \pm 0.00^{\dagger}$(1/9)	A-LSTM (8 tied)‡	0.500 ± 0.000	0.922 ± 0.059	-0.422	0.500 ± 0.000	0.985 ± 0.025	–
Emp. Forward	10212	4	37.84 ± 7.35	$-6.88 \pm 0.14^{\dagger}$(4/9)	A-LSTM (8 tied)‡	0.500 ± 0.000	0.341 ± 0.167	+0.159	0.500 ± 0.000	0.983 ± 0.027	–
Emp. Gain	0	134808	499.29 ± 6.28	–	–	–	–	–	–	–	only right turns
Emp. Potential	63	4850	18.20 ± 0.85	4.39 ± 0.31	A-LSTM (8 tied)‡	0.500 ± 0.000	0.508 ± 0.021	-0.008	0.548 ± 0.030	1.000 ± 0.000	–
Emp. Process	0	0	0.00 ± 0.00	–	–	–	–	–	–	–	no turns
DIAYN	894	785	6.22 ± 2.04	-0.37 ± 2.27	A-LSTM	0.821 ± 0.009	0.747 ± 0.014	+0.074	0.985 ± 0.005	1.000 ± 0.000	–
VIC	0	160	0.59 ± 0.41	–	–	–	–	–	–	–	only right turns
DADS	23	0	0.09 ± 0.05	–	–	–	–	–	–	–	only left turns
RDADS	79348	0	293.88 ± 13.52	–	–	–	–	–	–	–	only left turns
VALOR	125	19	0.53 ± 0.25	-1.78 ± 1.11	A-LSTM (8 tied)‡	0.500 ± 0.000	0.496 ± 0.095	+0.004	0.875 ± 0.112	1.000 ± 0.000	–
CSF	1830	280	7.81 ± 3.79	-1.84 ± 1.03	All	0.527 ± 0.008	0.590 ± 0.025	-0.063	0.921 ± 0.014	1.000 ± 0.000	–
NGU	76	121	0.73 ± 0.70	$0.38 \pm 1.46^{\dagger}$(6/9)	Obs	0.672 ± 0.045	0.663 ± 0.094	+0.009	0.778 ± 0.069	1.000 ± 0.000	–
CTEC	6703	37	24.96 ± 1.29	-5.52 ± 0.93	A-LSTM (8 tied)‡	0.500 ± 0.000	0.578 ± 0.077	-0.078	0.767 ± 0.045	1.000 ± 0.000	–
Random-DIAYN	195	189	1.42 ± 0.23	-0.05 ± 0.34	Obs	0.545 ± 0.057	0.523 ± 0.034	+0.022	0.914 ± 0.039	1.000 ± 0.000	–
Random-CSF	211	181	1.45 ± 0.23	-0.12 ± 0.38	C-LSTM	0.498 ± 0.039	0.552 ± 0.036	-0.054	0.944 ± 0.011	1.000 ± 0.000	–
Random-DADS	184	185	1.37 ± 0.21	0.02 ± 0.23	C-LSTM	0.521 ± 0.054	0.509 ± 0.020	+0.011	0.957 ± 0.010	1.000 ± 0.000	–
Random-VIC	178	183	1.34 ± 0.24	-0.00 ± 0.33	Obs	0.581 ± 0.043	0.471 ± 0.031	+0.110	0.911 ± 0.044	1.000 ± 0.000	–
Random-nonskill	194	183	1.40 ± 0.15	-0.06 ± 0.39	Obs	0.520 ± 0.023	0.483 ± 0.031	+0.037	0.954 ± 0.029	1.000 ± 0.000	–
Extrinsic	0	10541	39.04 ± 19.43	–	–	–	–	–	–	–	only right turns
Fly (IPS)	1362	2424	–	0.17 ± 1.90	IPS	0.665 ± 0.014	0.605 ± 0.009	+0.060	0.721 ± 0.018	0.955 ± 0.016	–

Table 7: Extended per-method results (all_bouts). Accuracy is balanced accuracy; uncertainty is fold s.d. where shown. † denotes bias from the indicated number of finite per-rollout values. ‡ marks a tie at maximum accuracy. See accompanying evidence notes for definitions and caveats.

Method	A-LSTM	C-LSTM	Pol	Val	A-LSTM+Pol	C-LSTM+Val	All	Obs
ICM	–	–	–	–	–	–	–	–
RND	–	–	–	–	–	–	–	–
Disagreement	–	–	–	–	–	–	–	–
γ-Progress	0.579	0.529	0.592	0.500	**0.603**	0.541	0.584	0.508
RE3	**0.500**	0.500	0.500	0.500	0.500	0.500	0.500	0.500
APT	0.500	0.503	0.500	0.500	0.498	0.502	**0.516**	0.505
Proto-RL	–	–	–	–	–	–	–	–
MAX	**0.500**	0.500	0.500	0.500	0.500	0.500	0.500	0.500
Emp. Forward	**0.500**	0.500	0.500	0.500	0.500	0.500	0.500	0.500
Emp. Gain	–	–	–	–	–	–	–	–
Emp. Potential	**0.500**	0.500	0.500	0.500	0.500	0.500	0.500	0.500
Emp. Process	–	–	–	–	–	–	–	–
DIAYN	**0.821**	0.814	0.819	0.813	0.820	0.816	0.820	0.791
VIC	–	–	–	–	–	–	–	–
DADS	–	–	–	–	–	–	–	–
RDADS	–	–	–	–	–	–	–	–
VALOR	**0.500**	0.500	0.500	0.500	0.500	0.500	0.500	0.500
CSF	0.512	0.507	0.503	0.504	0.514	0.507	**0.527**	0.500
NGU	0.645	0.605	0.621	0.653	0.654	0.598	0.633	**0.672**
CTEC	**0.500**	0.500	0.500	0.500	0.500	0.500	0.500	0.500
Random-DIAYN	0.516	0.506	0.521	0.539	0.513	0.534	0.529	**0.545**
Random-CSF	0.474	**0.498**	0.496	0.477	0.477	0.485	0.489	0.478
Random-DADS	0.502	**0.521**	0.502	0.496	0.488	0.505	0.483	0.498
Random-VIC	0.519	0.558	0.520	0.541	0.508	0.545	0.503	**0.581**
Random-nonskill	0.468	0.472	0.503	0.512	0.485	0.480	0.510	**0.520**
Extrinsic	–	–	–	–	–	–	–	–

Table 8: Internal (layer) early-window (-15 to -8 s) balanced accuracy per candidate layer; bold = selected layer (argmax). – = not scored (need $\geq$2 bouts per class).