

Supplementary Information
**PGOB: Behaviour-only parameter generation for whole-organism biophysical
simulators across worm and fly**
Extended Methods, Extended Results, Negative-Results Ledger, Reproducibility, 3
non-invasive deployment modes across worm and fly simulators.

Chenxi He, Cavendish Laboratory, University of Cambridge
ch2067@cam.ac.uk

July 26, 2026

Contents

1	Hardware and Reproducibility (S1)	2
1.1	Experiment overview (S1.0)	2
1.2	Cohort-size notation (S1.1)	4
2	Negative Results and Limitations (S2)	4
2.1	Negative-Results Ledger – Consolidated Single-Location Reference (S2.0)	4
2.2	Non-scalar loss falsification (S2.5)	5
2.3	F1 ablation control (S2.9)	5
2.4	41-experiment GAN-as-loss negative (S2.10)	5
3	Extended Methods (S3)	6
3.1	Universal 23-metric evaluator (S3.1)	6
3.2	Three-distance computation (S3.2)	6
3.3	Eigenworm cross-validation (S3.3)	6
3.4	Same-species prior-transfer protocol (S3.5)	6
4	Extended Results (S4)	7
4.1	BAAIWorm three-distance hierarchy (S4.1)	7
4.2	Drosophila parity (S4.2)	7
4.3	Yemini 43-strain (S4.3)	7
4.4	BAAIWorm 30-cell weight-corruption (S4.4)	7
4.5	F1 ablation (10 paired seeds) – pre-registered NULL (S4.4.1)	8
4.6	Genuine same-species parameter-prior transfer (S4.5)	9
4.7	Cross-sim parameter prior transfer (S4.6)	9
4.8	Loss-family invariance 4×6 observation (S4.7)	10
4.9	Confirmatory family vs. exploratory tests (S4.8)	10
4.10	Claim 2 5-test OOD analysis (S4.10)	11
4.11	Long-horizon stability (S4.11)	11
4.12	Cross-scale generation (S4.12)	11
4.13	Amortized neural-posterior back-end (S4.13)	11
4.14	Three-segment generation decomposition (S4.18)	12

4.15	Three-mode deployment-matrix details (S4.17)	13
4.16	Default-vs-PGOB significance sweep + flybody scale sweep (S4.21)	13
4.17	Extended-budget PGOB-hybrid fine-tune (S4.23)	14
5	Five Init Modes and Composition (S5)	14
5.1	Init-mode decision tree (S5.1)	14
5.2	Joint multi-init calibration (S5.2)	14
6	Reproducibility (S6)	14
6.1	Data availability with SHA-256 hashes (S6.1)	14
6.2	Code availability (S6.2)	15
6.3	Per-figure reproduction (S6.3)	15
6.4	Future directions (S6.4)	15
6.5	Reproducibility statement (S6.5)	15
6.6	Live fresh-build reproduction across all four simulators (S6.6)	15
7	Supplementary Figures (S7)	17
This SI accompanies the main paper and provides extended methods, extended results, control results and scope boundaries, and reproducibility details.		

1 Hardware and Reproducibility (S1)

1.1 Experiment overview (S1.0)

The table below maps each simulator/parameterisation to its initialisation, baseline, target, primary metric and *comparison class*; only the unit-free $G=D_1/D_2$ is comparable across simulators, every other cell is a per-simulator case study.

Table 1: **The PGOB experiments at a glance: parameterisation, baseline and what is cross-comparable.** Each row is one simulator/parameterisation; the comparison class names whether a cell compares generated-to-simulator (sim $\leftrightarrow$ sim), real-to-real, or generated/default-to-real (sim $\leftrightarrow$ real). *Only* the unit-free $G=D_1/D_2$ (BAAIWorm full5 and flybody) is comparable *across* simulators; every other cell is a per-simulator case study. “Initialized” denotes the uninformative randomised start; “default” the expert-calibrated published parameters.

Simulator	Dim	Init	Baseline	Target	Primary metric	Real data?	Statistically cant?	signifi- cant?	Comparison class
BAAIWorm (full5, 3-distance)	$\approx 5k$	Initialized (scram- bled)	expert default	sim θ^* / real / default	robust- z D_1, D_2, D_3 ; $G=D_1/D_2$	Yes (OWMD, $N=200$)	$G < 1$ (boot.)	(1000/1000)	sim $\leftrightarrow$ sim, real $\leftrightarrow$ real, sim $\leftrightarrow$ real
BAAIWorm (full5, SOTA band)	$\approx 5k$	Initialized (scram- bled)	sim SOTA band	sim SOTA be- haviour	behaviour dis- tance (160-cond OOD)	No (sim only)	matches (0.1199 [0.120, 0.125])	band $\in$	sim $\leftrightarrow$ sim
BAAIWorm (5 scales)	5	Initialized (1/3–3 $\times$)	expert default	sim θ^* (recov- ery)	PGOB-sim be- haviour distance	Yes (OWMD)	recovery 0.198		sim $\leftrightarrow$ sim
modWorm	30	scale-1.0 perturb.	expert default	sim θ^* / real	rel_θ (sim); per- feature W_1 (real)	Yes (OWMD)	$rel_\theta=0.009$ (sim); de- ployment at parity		sim $\leftrightarrow$ sim, sim $\leftrightarrow$ real
flybody (bound- ary)	515	same- species prior (random collapses)	expert default	sim θ^* / real	$G=D_1/D_2$ (sim only)	Yes (Janelia)	$G < 1$ (0.20<0.42), re- covery only; trajec- tory boundary (arms <1.5%)		sim $\leftrightarrow$ sim, sim $\leftrightarrow$ real
FlyGym v2	48	walk-band prior (random collapses)	Initialized (ran- dom) start	Initialized $\rightarrow$ experthree-segment distance	closed fraction	anchor	94.9%; 8/10 obs Holm-sig. vs Initial- ized		sim $\leftrightarrow$ sim (capa- bility)

Compute. Cloud A40 GPU instances; RTX 4090D and RTX 3080 Ti secondary. The main sweep – 10 configurations $\times$ 3 seeds $\times$ 30 generations $\times$ 16 population = 14,400 rollouts – consumed $\approx 4.8 \times 10^4$ rollout-seconds (≈ 27 wall-hours). Cumulative compute: ≈ 307 CPU-days (BAAIWorm 142, flybody 88, modWorm 41, FlyGym 36). Every reported result is reproducible from `python run_calibration.py --sim {worm,modworm,fly,flygym} --scale s --seed k --opt cmaes_sep`.

1.2 Cohort-size notation (S1.1)

$n=5, 10$ = random SPSSA seeds (weight-corruption and F1 ablation contexts); $n=4, 40, 160$ = starting conditions (held-out OOD context); $n=43$ = held-out mutant strains (Yemini prospective); $n=15$ = BAAIWorm ablation-type rank test; $n=5/\text{level}$ (30 cells) = weight-corruption sweep; $N=200$ = OWMD-Schafer worms; $B=1000$ = bootstrap resamples. All ambiguous in-text occurrences are tagged with explicit unit names.

Per-simulator hyperparameters.

Item	BAAIWorm	flybody	modWorm
Param dim	5 (global scales) / $\approx 5k$ (full5 per-element)	515	30/50
Search-space transform	linear [0.5, 2] (scales); per-element (full5)	log-space mult.	linear
Integrator / dt	NEURON, 5/3 ms (neural)	MuJoCo 0.1ms	implicit-Euler 0.5ms
Output rate	30 Hz	10 kHz	25 Hz
Rollout length	10 s	5 s	40 s
Optimiser	SPSA + CMA	CMA-ES sep	CMA-ES sep
Population λ	8 / 8	16	16
Generations	30	30	30
σ_0	0.05	0.05 log-space	0.05
Multi-cond avg C	[8, 16]	8 snippets	8 seeds

Table 2: Per-simulator consolidated hyperparameter reference.

2 Negative Results and Limitations (S2)

2.1 Negative-Results Ledger – Consolidated Single-Location Reference (S2.0)

This ledger consolidates every control result, bounded-scope statement, and limitation into one place. Every row carries a back-pointer to the section where the underlying analysis lives; the main text summarises these load-bearing limitations.

Table 3: **Control results and scope boundaries.** Bounded-scope statements, framework control results (S2.9), and per-channel residuals.

Finding	Caveat / scope	Section
4-config non-scalar loss falsification	Aggregate baseline matches or exceeds every tested loss configuration.	S2.5
F1 ablation neural control (10 paired seeds, $n=20$ runs)	Pre-registered NULL (Wilcoxon $p=0.734$, $d=0.14$). Pure-trajectory loss is not outperformed by neural-augmented loss – positive for the framework. (S4.4.1)	S2.9
modWorm three-segment	Eigenworm-variance channels fixed by the 30-d Cook-ODE substrate (the expert arm cannot move them either); 2/6 channels improved, equal-weight mean +16.25% (S4.18).	S4.18
flybody W_1 three-segment	Pre-registered null result: all three arms within < 1.5% (13.254/13.454/13.348); <code>walk_imitation</code> Mo-Cap controller dominates trajectory, init signal carried in loss space ($9.3\times$), not W_1 (S4.18).	S4.18
Extended-budget hybrid fine-tune	Held-out improvement on all three simulators; the warm-start already lies near the $\text{sim} \leftrightarrow \text{real}$ floor (S4.23).	S4.23
flybody default-vs-PGOB	0/9 significant ($N=45/\text{arm}$): PGOB leaves an already-strong expert default intact (S4.21).	S4.21
41-experiment GAN-as-loss collapse	Val-loss $\geq 2\times$ regression across 5 D-architectures.	S2.10
Same-species parameter-prior transfer	Genuine same-species rescue (BAAIWorm $\rightarrow$ modWorm $N=16$, flybody self-prior $9.3\times$); transfer demonstrated between two simulators of the same species only.	S4.5
flybody FULL-split parity	TEST split: 9/10 observables closer (+1.75%); FULL split reported in S4.2.	S4.2
Loss-family invariance	An operational loss-family invariance observation (S4.7).	S4.7

2.2 Non-scalar loss falsification (S2.5)

4-configuration loss-design falsification on BAAIWorm: the aggregate baseline matches or exceeds every tested configuration (dominates 3/5 metrics).

2.3 F1 ablation control (S2.9)

The neural-ablation control (10 paired seeds, $n=20$ runs) returns a pre-registered NULL (Wilcoxon $p=0.734$, $d=0.14$); pure-trajectory loss is not outperformed by neural-augmented loss. Full statistics in S4.4.1.

2.4 41-experiment GAN-as-loss negative (S2.10)

All 41 collapse: val-loss regresses $\geq 2\times$ while D-loss drops < 0.01 . A multi-behaviour discriminator on closed-loop FlyGym extends the same negative: a discriminator trained jointly on four behaviours does not outperform single-task discriminators ($\Delta \approx -0.01$ in behavioural closeness; discriminator validation accuracy near chance, $\approx 0.43\text{--}0.55$), so the behaviour-conditional GAN-as-loss objective is not supported in the closed-loop fly either, consistent with the worm-side GAN-as-loss negative.

3 Extended Methods (S3)

3.1 Universal 23-metric evaluator (S3.1)

#	Metric	Definition	Worm	Fly	Function
1	zigzag	running $\sum \Delta h $ over 1-s window	✓	✓	<code>_zigzag</code>
2	mean_speed	$\langle \ \dot{p}_t\ \rangle$	✓	✓	<code>_mean_speed</code>
3	reversal_frequency	$\#\{t : \dot{p}_t \cdot \hat{h}_t < 0\}/T$	✓	×	<code>_reversal_freq</code>
4	wave_amplitude	tangent-angle PC1 envelope max-min	✓	×	<code>_wave_amp</code>
5	speed_std	$\text{std}(\ \dot{p}_t\)$	✓	✓	<code>_speed_std</code>
6	chemotaxis_index	$\langle \hat{h}_t \cdot \nabla c_t \rangle$	✓	×	<code>_chemotaxis</code>
7	distance_to_food	$\ p_T - p_{\text{food}}\ $	✓	×	<code>_dist_food</code>
8	steering	mean $ \dot{h}_t $	✓	✓	<code>_steering</code>
9	turning_rate	$\#\{t : \dot{h}_t > 200\text{deg/s}\}/T$	✓	✓	<code>_turn_rate</code>
10	body_wave_freq	dominant FFT freq	✓	×	<code>_wave_freq</code>
11	dorsal_ventral_corr	corr(d-arc, v-arc)	✓	×	<code>_dv_corr</code>
12	leg_alternation	6-leg phase-diff variance	×	✓	<code>_leg_alt</code>
13	MSD	slope	✓	✓	<code>_msd</code>
14	tortuosity	path / displacement	✓	✓	<code>_tortuosity</code>
15	path_length	$\sum \ \Delta p_t\ $	✓	✓	<code>_path_len</code>
16	mean_displacement	$\ p_T - p_0\ $	✓	✓	<code>_mean_disp</code>
17	mean_acceleration	$\langle \ \ddot{p}_t\ \rangle$	✓	✓	<code>_mean_acc</code>
18	jerk	$\langle \ \ddot{p}_t\ \rangle$	✓	✓	<code>_jerk</code>
19	heading_variance	$\text{var}(\hat{h}_t)$	✓	✓	<code>_heading_var</code>
20-23	eigen_e1-e4	Stephens-2008 PC amplitudes	✓	×	<code>_eigen</code>

Table 4: The full evaluator emits all 23 scalars; 9 are computed as pure-trajectory functions in addition to the 14 core metrics.

3.2 Three-distance computation (S3.2)

$D_1 = d_z(f_{\hat{\theta}}, \{f_{\theta_{\text{GT}}}\})$, $D_2 = \frac{1}{N} \sum_i d_z(x^{(i)}, \{x^{(j)}\}_{j \neq i})$, $D_3 = d_z(f_{\theta_{\text{GT}}}, \{x^{(i)}\})$. Both naive z and median/MAD robust z reported.

3.3 Eigenworm cross-validation (S3.3)

Tangent-angle PCA cross-validated against OWMD Tierpsy on $n=40$ worms $\times$ 4 PC amplitudes: Pearson $r=0.987$.

3.4 Same-species prior-transfer protocol (S3.5)

The same-species parameter-prior transfer (S4.5/S4.6) derives a per-biophysical-class coefficient-of-variation prior from a source simulator and applies it as the per-dimension CMA-ES step size (σ_0) when generating on a different simulator of the same species. Each generation is one CMA-ES sep run (30 gen, $\lambda=16$); arms (random / cross-sim prior / own default) differ only in the prior used to seed σ , and are compared on a disjoint held-out test split. The cross-scale optimiser comparison fixes the simulator and the ground-truth θ^* and varies only the perturbation scale and optimiser (CMA-ES / DE / NES) at a matched eval budget.

4 Extended Results (S4)

4.1 BAAIWorm three-distance hierarchy (S4.1)

Table 5: Three-distance hierarchy on the BAAIWorm full5 generation against the OWMD-N2 ($N=200$) real reference. The three distances are, by the main-text definition, D_1 = generated→the simulator’s own target (a recovery distance, *not* a distance to OWMD), D_2 = real→real spread, and D_3 = deposited default→real. The **robust (MAD)** z estimator is the canonical one used throughout the main text ($D_1=0.71$, $D_2=1.98$, $D_3=21.82$); the other rows are sensitivity checks under alternative estimators. The strict $D_1 < D_2 \ll D_3$ hierarchy holds under the canonical estimator and the naive estimator; under the alternative 5-active-feature cross-validated D_2 estimator the real↔real reference contracts to 0.393 so D_1 no longer falls below it, but $D_1 < D_3$ still holds by $30\times$. We report the robust estimator as canonical because the cross-validated D_2 uses a held-out feature subset that under-counts genuine inter-individual spread.

estimator	D_1	D_2	D_3	hierarchy
robust (MAD), canonical	0.713	1.976	21.82	$D_1 < D_2 < D_3$
naive	0.500	1.780	26.47	$D_1 < D_2 < D_3$
trimmed 5%	—	2.366	30.32	$D_2 < D_3$
5-active-feature CV (sensitivity)	0.713	0.393	21.82	$D_1 < D_3$ ($30\times$); marginal vs. D_2

Bootstrap $B=1000$ on OWMD-N2 under the canonical robust estimator: $D_1 < D_2 < D_3$ in 1000/1000.

4.2 Drosophila parity (S4.2)

flybody walking TEST split: 9/10 closer at +1.75% (60 train / 20 val / 20 holdout). On the FULL split the default is already strong, so the available PGOB margin is small. **Flight wing-beat** 227.92 ± 5.59 Hz. **FlyGym 5-behaviour:** 5/5 inside biological CV. **Adhesion phenotype:** 3-of-3 / 0-of-3 split.

4.3 Yemini 43-strain (S4.3)

In-sample reference: $\bar{\rho}=0.558$, median 0.400, 43/43 positive. **LOSO** (mutant prediction, main text): mean-train 0.684, median-train 0.628, strain-shuffle 0.527.

Pre-registered full 23-feature LOSO robustness. On the pre-registered full 23-feature set (fixed before scoring; the 4 active features are a strict subset), the leave-one-strain-out test gives mean-training-delta Spearman $\bar{\rho}=0.525$ (median 0.585, SD 0.211) vs. strain-shuffle null $\bar{\rho}_0=0.338$ (median 0.352, SD 0.114), $\Delta=+0.187$ with 27/28 strains $\rho>0$. The mutant-prediction signal is preserved across the full deposited metric space, and the 4- and 23-feature analyses are mutually consistent. This confirmation covers the 28 strains with available feature data rather than the full 43-strain panel. Mutant differential $N=12$ protocol: strain-specific 5/4/4 deviations.

4.4 BAAIWorm 30-cell weight-corruption (S4.4)

Level	Seed	L_0	L_{final}	$L_{\text{rec. clean}}$	Δ/L_0	elapsed
L0 clean	0	0.12479	0.06334	0.06334	−49.2%	
L0 clean	1	0.12479	0.08568	0.08568	−31.3%	
L0 clean	2	0.12479	0.06058	0.06058	−51.5%	
L0 clean	3	0.12479	0.06466	0.06466	−48.2%	
L0 clean	4	0.12479	0.07360	0.07360	−41.0%	
L1 5% Gauss	0	0.11695	0.07723	0.11474	−34.0%	
L1 5% Gauss	1	0.12690	0.04444	0.14148	−65.0%	
L1 5% Gauss	2	0.07772	0.05069	0.12691	−34.8%	
L1 5% Gauss	3	0.13131	0.05777	0.14815	−56.0%	
L1 5% Gauss	4	0.05706	0.04399	0.11078	−22.9%	
L2 20% Gauss	0	0.15749	0.14383	0.13439	−8.7%	
L2 20% Gauss	1	0.25744	0.21711	0.09803	−15.7%	
L2 20% Gauss	2	0.12516	0.07092	0.15990	−43.3%	
L2 20% Gauss	3	0.04760	0.04610	0.17732	−3.2%	
L2 20% Gauss	4	0.08930	0.04695	0.11239	−47.4%	
L3 50% Gauss	0	0.25733	0.20162	0.08917	−21.6%	
L3 50% Gauss	1	0.11693	0.06028	0.06249	−48.4%	
L3 50% Gauss	2	0.24599	0.16744	0.14172	−31.9%	
L3 50% Gauss	3	0.22694	0.19836	0.08531	−12.6%	
L3 50% Gauss	4	0.18516	0.17295	0.13303	−6.6%	
L4 shuffle 30%	0	0.11415	0.10387	0.10916	−9.0%	
L4 shuffle 30%	1	0.10541	0.06838	0.12665	−35.1%	
L4 shuffle 30%	2	0.15729	0.14465	0.12040	−8.0%	
L4 shuffle 30%	3	0.12336	0.04999	0.17776	−59.5%	
L4 shuffle 30%	4	0.21156	0.12426	0.14229	−41.3%	
L5 random-wt	0	0.27412	0.22682	—	−17.3%	
L5 random-wt	1	0.28653	0.27326	—	−4.6%	
L5 random-wt	2	0.15218	0.14408	—	−5.3%	
L5 random-wt	3	0.27462	0.21875	—	−20.3%	
L5 random-wt	4	0.41267	0.37943	—	−8.1%	

30/30 cells $L_{\text{final}} < L_0$ (6 corruption levels $\times$ 5 seeds). full-sweep mean $\Delta/L_0 = -29.4\%$ (range -3.2% to -65.0%).
Per-level mean L_{final} : $L_0=0.070$, $L_1=0.055$ (dip $< L_0$), $L_2=0.105$, $L_3=0.160$, $L_4=0.098$, $L_5=0.248$.
Corruption-vs-residual increasing, non-monotone.

L5 row note. The L5 (random-weight) corruption is applied in the synthetic-loss space; its L_0 anchor (0.274 random-weight vs. 0.125 clean-default) is the BAAIWorm random-init anchor referenced by the three-segment decomposition (S4.18). The $L_{\text{rec. clean}}$ column is undefined for L5 because there is no clean reference for a fully-random-weight cell.

4.5 F1 ablation (10 paired seeds) – pre-registered NULL (S4.4.1)

The F1 neural-ablation control runs over 10 paired matched seeds ($n=20$ runs): mode-a (pure-trajectory loss) vs. mode-b (neural-augmented loss).

- mode-a mean behavioural loss 0.05456; mode-b mean 0.05758.
- paired difference (b−a) +0.003016; Cohen’s $d_{\text{paired}}=0.144$.

- Wilcoxon $W=19$, two-sided $p=0.734$; one-sided ($b>a$) $p=0.367$; t -test $p=0.660$.
- bootstrap $B=10^4$ 95% CI on mean difference $[-0.00931, +0.01522]$ – straddles zero.

Verdict: NULL (no significant difference, $p=0.734$). Pure trajectory observables suffice: adding internal neural signal yields no detectable gain.

4.6 Genuine same-species parameter-prior transfer (S4.5)

The load-bearing evidence that PGOB generates biology-governed parameter structure is the genuine-rollout same-species soft-prior rescue, detailed in S4.6. A parameter prior derived from one simulator of a species transfers to a different simulator of the *same* species: BAAIWorm $\rightarrow$ modWorm on the real Cook-ODE ($N=16$ held-out test rollouts, 13/16 prior-better than random, $p=0.025$), and the flybody self-prior (FlyGym $\rightarrow$ flybody, $9.3\times$ loss improvement). A prior learned on one worm simulator moves a different worm simulator, and a fly self-prior moves a fly simulator, to a working solution a default start cannot reach: the transferable structure encodes species-level biology. A separate genuine cross-scale optimiser comparison on the real flybody MuJoCo simulator (CMA-ES / DE / NES, scales 0.1–0.7; best DE relative error 0.055 in-distribution, all optimisers degrade off-scale) shows generation locks to the specific target physiology.

4.7 Cross-sim parameter prior transfer (S4.6)

BAAIWorm $\rightarrow$ modWorm, real Cook-ODE rollout (main-text Methods, Initialisation: the cross-simulator prior; canonical run). Over $N=16$ independent held-out starts (40 CMA iterations each, init scale 0.3), scored on a disjoint test-seed pool with a real modWorm Cook-ODE rollout, a BAAIWorm-derived per-biophysical-class soft prior is compared against the simulator’s own default init on the aggregate held-out behaviour distance to real OWMD. The cross-sim prior transfers per-biophysical-class coefficient-of-variation (per-dim CMA σ shaped by class CV), *not* raw parameter values; class map **gap $\rightarrow$ gj**, **syn $\rightarrow$ syn**, **{Cm,Gleak, τ ,Eleak} $\rightarrow$ passive**, **$g_{in} \rightarrow ion$** , **mus $\rightarrow$ wout**.

Table 7: BAAIWorm $\rightarrow$ modWorm soft-prior init-transfer, held-out behaviour distance to real OWMD ($N=16$ held-out starts, real Cook-ODE rollout). Lower is better. The prior outperforms the simulator’s own default init on 13/16 starts (one-sided Wilcoxon $p=0.025$).

Arm	held-out d_{real} (mean)	95% CI	starts prior-better
own default	3.370	[2.30, 4.46]	—
xsim_prior	2.295	[1.52, 3.26]	13/16 ($p=0.025$)

The cross-sim (BAAIWorm-derived) prior outperforms the simulator’s own default init on held-out real data (13/16 starts, one-sided Wilcoxon $p=0.025$); these are the same numbers reported in the main text (Methods, Initialisation). **Caveats:** the class map is biophysically informed rather than exact, and the absolute floor is substrate-limited by the 30-d Cook ODE, so arms are compared at a shared floor; the transfer is demonstrated between two simulators of the same species (S4.5). **flybody self-prior:** the FlyGym $\rightarrow$ flybody same-species prior improves the loss $0.01085 \rightarrow 0.001166$ ($9.3\times$; $126 \rightarrow 515$ -d CMA). Only the soft per-axis- σ prior transfers; a hard-init linear projection over-constrains the start (rel -0.127), confirming the prior must shape step size rather than pin parameter values.

Per-channel attribution for *unc-1* / *unc-9* / *unc-43*. To verify that the strain-specific channel selection of the mutant analysis (main text) is not the same 4 channels recurring under random permutation, we report below the top-ranked observable channels per strain.

Table 8: Top-ranked behavioural channels per Yemini-2013 strain (Spearman ρ between Jacobian-predicted vs. measured per-channel deviation; OWMD $N=12$ worms per strain; bootstrap CI 95%; permutation $p<0.001$). Channels are non-coincident across strains, ruling out a same-4-channels confound.

Strain	Channel rank-1 ... rank-4	Selected $ J $ count
<i>unc-1</i>	bend_amp_head, undulation_phase_std, head_freq, propagation_speed	5
<i>unc-9</i>	speed_norm, body_curvature, eigen_a2_var, reversal_rate	4
<i>unc-43</i>	pirouette_rate, head_bend_skew, eigen_phase_drift, head_tail_orient	4

Channel overlap between any two strains is $\leq 1/4$; the channel sets are strain-specific and not artefacts of a fixed top-4 ordering.

4.8 Loss-family invariance 4×6 observation (S4.7)

We report a loss-family invariance: across four loss designs the per-observable improvement is operationally indistinguishable.

Table 9: Per-observable test-W1 % improvement under four PGOB losses; 80/20 split.

Sim	Observable	L1 MSE	L2 W1	L3 GAN	L4 NSGA	Verdict
modWorm	undulation_phase_std	+17.30	+9.62	+19.36	+2.13	loss-rescuable
modWorm	pirouette_rate	+1.09	-1.09	-1.09	-2.17	information limit
modWorm	head_bend_amp	+0.01	+0.01	+0.02	+0.01	information limit
flybody	vrt_speed	+6.20	+6.20	+6.20	+6.20	bit-identical
flybody	yaw_rate	-22.04	-22.04	-22.04	-22.04	bit-identical
flybody	speed_skew	+7.67	+7.67	+7.67	+7.67	bit-identical

5/6 behaviour-signal-limited cells show test-W1 differences across four losses below per-seed noise.

4.9 Confirmatory family vs. exploratory tests (S4.8)

The confirmatory claims (main-text Statistics) form a small *pre-specified* family — the three-distance bootstrap, the mutant leave-one-strain-out vs. shuffle test, the cross-laboratory $< 3\times$ ratio test, the same-species prior-transfer test, the mutant-differential permutation test, and the trajectory-vs-neural ablation control — corrected together under Holm-Šidák. All remaining quantitative comparisons in the study (per-observable default-vs-PGOB tests, the rescale and scale-sweep points, long-horizon and cross-condition checks) are *exploratory*: reported descriptively and uncorrected, and not treated as confirmatory. For completeness, applying a single Holm-Šidák correction across all 58 pooled tests leaves 11 surviving, clustered on FlyGym (4/10) and modWorm rescale points; the confirmatory family above is the basis for the statistical claims.

4.10 Claim 2 5-test OOD analysis (S4.10)

T1 BAAIWorm single 0.18797 vs. joint 0.18191 (−3.2%). T2 FlyGym: null (discriminator-fit note, S2.10). T3a worm cross-lab 1.62× joint better. T3b fly 1.64×. T5 cross-strain $\bar{\rho}_{\text{joint}}=0.558$. 4/5 favour joint.

4.11 Long-horizon stability (S4.11)

FlyGym $T \in \{100, 300, 500, 1000\}$ PGOB closeness $\{0.501, 0.404, 0.409, 0.496\}$ vs default $\{0.637, 0.580, 0.556, 0.546\}$
– PGOB outperforms default at every T . BAAIWorm $T \in \{100, 300, 1000, 2000\}$ closeness $\{4.83, 5.43, 6.15, 5.78\}$
– $T=2000$ reverses $T=1000$.

4.12 Cross-scale generation (S4.12)

Across-seed σ at $s_{\text{src}}=0.3 \rightarrow s_{\text{tgt}}=1.0$: 0.065 (relative spread 0.7%). **Genuine cross-scale optimiser comparison.** On the real flybody MuJoCo simulator, CMA-ES / DE / NES were compared at scales $\{0.1, 0.3, 0.5, 0.7\}$ (25 generations, popsize 12, 50-d random subspace, identical eval budget): DE reaches the best in-distribution generation (relative error 0.055) and all optimisers degrade off-scale, confirming that calibration locks to the specific target physiology rather than producing a scale-universal parameter set. This is the cross-scale counterpart to the same-species prior-transfer evidence (S4.5).

4.13 Amortized neural-posterior back-end (S4.13)

The engine-agnostic claim of the main text (the engine-swap result) is supported by instantiating the same behaviour-only formulation with an amortized neural posterior estimator (NPE-MAF) in place of the black-box optimiser, on two simulators. The estimator is a masked-autoregressive-flow conditional density estimator trained with the `sbi` toolkit on simulations drawn from each simulator’s random-start prior. **modWorm:** 30-d parameter vector, 6 trajectory observables, BoxUniform prior over the scale-1.0 random-init region; $N=15,000$ requested simulations of which 14,993 were valid; posterior-predictive reproduction of the ground-truth observation reaches mean absolute observable difference 1.36×10^{-3} across the six observables (per-observable `bend_amp` 3.8×10^{-3} , `bend_mean` 4.3×10^{-3} , `head_freq` 0, `propagation` 1.5×10^{-5} , `speed_proxy/turn_rate_proxy` 6.4×10^{-5}). **FlyGym v2:** 48-d gain vector, 10 observables, BoxUniform walkable band $\pm 30\%$ of the canonical gains; $N=6,000$ valid simulations; the posterior mean lands at relative error 0.018 to the canonical θ^* .

Density-estimator comparison (held-out NLL). On the same simulation pool per simulator, three density estimators were compared by held-out negative log-probability (lower is a better fit): a neural spline flow (NSF), the masked autoregressive flow (MAF, the estimator used for the generation above), and a mixture-density network (MDN). modWorm: NSF 46.3 < MAF 49.8 < MDN 51.6 (held-out $N=2,999$ of 14,993). FlyGym v2: NSF 157.3 < MAF 161.5 < MDN 165.8 (held-out $N=1,200$ of 6,000). The NSF > MAF > MDN ordering is the expected density-estimation hierarchy and carries over directly to the PGOB setting.

Per-dimension simulation-based calibration. Rank-based simulation-based calibration (SBC) on the trained posterior is consistent (per-dimension KS at $\alpha=0.05$) on 10/30 modWorm and 22/48 FlyGym dimensions, concentrated on the well-identified, behaviour-sensitive axes; the remaining axes are the weakly-identified directions that a behaviour-only observation does not constrain. The load-bearing evidence for the back-end is the posterior-predictive behaviour reproduction above, with SBC reported as a calibration diagnostic on the identified axes.

4.14 Three-segment generation decomposition (S4.18)

Motivation. The “+% over default” framing carries two distinct “default” semantics across simulators, which the three-segment decomposition resolves:

- **modWorm** default arm = a scale=1.0 perturbation of $\theta_{GT} \Rightarrow RANDOM$ start ($\sim 81\%$ mean |rel dev|, 16% sign flips), *not* expert. The expert (scale=0) arm is measured.
- **FlyGym v2** default arm = scale=1.0, $U(-1, 1)$ gain perturbation $\Rightarrow RANDOM$. The expert (scale=0, θ_{GT} exact) arm is measured.
- **BAAIWorm**: the “default” column is a *randomised* weight start (**randn** $\times 0.1$ corruption of synaptic + gap-junction weights); the PGOB-generated full5 checkpoint *is* the expert-grade calibration. The three-segment therefore collapses to random $\rightarrow$ generated, and we report *climb toward real*: $d_{\text{random}}=3.908 \rightarrow d_{\text{pgob}}=3.415$, **12.6%** of the random $\rightarrow$ real gap closed.
- **flybody** default arm = EXPERT-default (**perturb**=0.0); the random arm (**perturb**=1.0) and PGOB arm (**perturb**=0.3) are measured in the same matched pipeline (500 default / 500 PGOB / 596 real). The W_1 three-segment is a *pre-registered null result*: $d_{\text{random}}=13.254$, $d_{\text{pgob}}=13.454$, $d_{\text{expert}}=13.348$ – all three arms within $< 1.5\%$. The **walk_imitation** rollout is a MoCap-locked imitation controller whose keypoint trajectory is dominated by the reference clip, so an actuator gain/stiffness/damping perturbation – even a full **perturb**=1.0 randomisation – does not propagate to the observable kinematics; the signal is fixed by the MoCap reference clip. flybody’s init-mode signal lives in *loss space* (soft-prior $9.3\times$, S4.6) rather than W_1 space.

Aggregate cross-sim summary: FlyGym 13.30/10.78/10.64 (**94.9%**); modWorm 13.84/13.94/12.85 (**−9.9%**, substrate floor; +16.25% equal-weight mean); BAAIWorm 3.908/3.415/=pgob (climb **12.6%**); flybody random/PGOB/expert 13.254/13.454/13.348 (pre-registered null result, $< 1.5\%$ spread). The modWorm and FlyGym three-arm rollouts reproduce the stored $d_{\text{random}}/d_{\text{pgob}}$ to 10^{-12} , confirming the added expert arms are unit-matched and comparable. The flybody three arms overlap because the MoCap-locked controller dominates the keypoint trajectory (S4.6).

FlyGym v2 full per-observable three-arm table ($N=20/\text{arm}$, $T=200$):

Observable	d_{random}	d_{pgob}	d_{expert}	generation / note
fwd_speed	0.2356	0.0485	0.0461	98.7%
lat_speed	0.2748	0.0334	0.0077	90.4%
vrt_speed	0.1706	0.3552	0.3720	regress (loss-unconstrained)
speed_mean	0.8851	0.4494	0.4595	$>100\%$ of expert
speed_std	0.4567	0.2647	0.2923	$>100\%$
yaw_rate	0.3313	0.0164	0.0073	89.4%
body_len_mean	8.7384	7.9574	7.9243	95.9%
body_len_std	0.6366	0.4123	0.4012	95.3%
leg_phase_coup	0.7394	1.1298	0.9777	regress
speed_skew	0.8280	0.1083	0.1529	$>100\%$

random fails 8/20; pgob 0/20; expert 0/20. Aggregate-sum W_1 : random 13.30, pgob 10.78, expert 10.64.

modWorm substrate-ceiling three-arm table ($N=100/\text{arm}$ vs. $N=200$ OWMD):

Observable	d _{random}	d _{pgob}	d _{expert}	note
eig1_var	3.203	3.218	2.782	substrate ceiling (expert also far)
eig2_var	2.362	2.381	2.002	substrate ceiling
eig3_var	2.282	2.281	2.011	substrate ceiling
bend_freq	0.198	0.198	0.198	flat (all arms tie)
active_frac	0.0102	0.000	0.000	100% (exact OWMD match)
path_curv	5.786	5.861	5.862	substrate ceiling

The eigenworm-variance channels are unmovable even at the expert arm: a substrate-physical ceiling of the 30-d Cook ODE. **active_frac** is the one channel PGOB closes exactly. *Aggregation note*: the aggregate-sum W_1 weights observables by raw magnitude, so the large eigenworm-variance channels dominate and the modWorm aggregate generation fraction is -9.9% (random 13.84 $\rightarrow$ pgob 13.94 $\rightarrow$ expert 12.85); the robust summaries are the per-observable improved-count (2/6) and equal-weight mean (+16.25%).

4.15 Three-mode deployment-matrix details (S4.17)

Matrix coverage. The main text reports the 9-cell (3 simulators $\times$ 3 modes) coverage matrix for BAAIWorm, modWorm, and FlyGym v2.

Mode definitions (reproduced from main). *PGOB-sim*: SPSA / CMA-ES against a sim-SOTA synthetic target θ^* ; confirms the framework recovers a known target (the true optimum is known and is regenerated). *PGOB-real*: SPSA directly against real OWMD / Janelia trajectories; the deployment scenario. *PGOB-hybrid*: sim warm-start $\rightarrow$ real fine-tune; the production protocol. All three modes are non-invasive in the strict sense – no patch-clamp, no two-photon calcium imaging, no EMG, no electrophysiology of any kind is consumed at any stage of optimisation. The traditional alternative (per-cell calibration against intracellular recordings) is not accessible to the vast majority of behaviour laboratories.

Hybrid-method note. The **modWorm** three-mode cells are *genuine* two-stage optimisations: the hybrid is a real StageA(sim warm-start) $\rightarrow$ StageB(real fine-tune) run, not a mid-point. The modWorm genuine hybrid lands between its sim and real endpoints ($W_1=1.32$ to real vs. real-floor 1.25 and sim 0.009), consistent with the “hybrid between sim and real” reading.

4.16 Default-vs-PGOB significance sweep + flybody scale sweep (S4.21)

Statistically-powered default-vs-PGOB: $N=40\text{--}50$ samples/arm (sample = one rollout from the random-start $\times$ physical-IC pool), per-observable bootstrap 95% CI + Mann–Whitney U + Cohen’s d on the distance-to-real distribution. **flybody** ($N=45$ /arm, default **perturb=0** vs PGOB **perturb=0.3**, real = 596 Janelia windows): across walk/turn/gait (9 observables) PGOB is closer to real on 4/9 and 0/9 reach significance (all $|d| < 0.41$, $p > 0.05$): PGOB leaves an already-strong expert default intact. This pairs with the Initialized-start end (FlyGym, 8/10 observables +32.4%, Holm-significant; exact-binomial 95% CI [0.44, 0.98]) to show PGOB’s value is *adaptive to default quality*: it rescues an Initialized start and leaves a good expert default intact.

flybody scale sweep (sim2sim, ground-truth θ^* , 515-d): the generation degrades smoothly with perturbation scale – rel-final 0.119 (scale 0.1, $L_{\text{final}}=7.6 \times 10^{-6}$) $\rightarrow$ 0.193 $\rightarrow$ 0.263 $\rightarrow$ 0.335 $\rightarrow$ 0.397 $\rightarrow$ 0.494 $\rightarrow$ 0.586 (scale 0.7, $L_{\text{final}}=1.5 \times 10^{-4}$), then a generation cliff to 1.186 (scale 0.9, $L_{\text{final}}=1.085 \times 10^{-2} \approx L_0$). The scale-0.8 point sits in the cliff, where the behaviour loss flattens ($L_{\text{final}} \approx L_0$), so the generation boundary lies between scale 0.7 and scale 0.9; at scale 1.0 the loss saturates at the random-init level (rel 1.061). Low-scale generation is exact ($L_{\text{final}} \sim 10^{-6}$) and collapses past the 0.7–0.8 boundary, locating the high-dim random-init regime (scale ≥ 0.9) where

the soft prior applies. sim2sim generation loss is minimal (generated-to-sim W_1 0.71 vs. sim-to-real 21.8 on the worm robust three-distance), so the residual sim2real gap reflects the simulator’s biological/substrate fidelity.

4.17 Extended-budget PGOB-hybrid fine-tune (S4.23)

The production protocol is PGOB-hybrid: warm-start on the simulator’s own behaviour, then fine-tune against the laboratory’s real video. We characterise the magnitude of the fine-tune gain directly by relaxing the early-stop budget (patience 18, up to 40 iterations) and tracing the full held-out distance-to-real curve. Model selection is on a disjoint validation split; the held-out judge split is never seen by the optimiser. Guards are a box constraint around the warm-start (± 0.6) plus a drift regulariser ($\lambda=0.03$) and gradient clip.

sim $\rightarrow$ real	split (tr/val/ho)	warm-start d_{real}	min d_{real}	min iter	best dip
modWorm $\rightarrow$ OWMD-N2 (Cook-ODE W_1)	24/8/8	13.314	13.169	38	−1.1%
FlyGym v2 $\rightarrow$ Janelia	60/20/20	0.2154	0.2034	16	−5.5%
BAAIWorm $\rightarrow$ OWMD-N2	24/8/8	0.1753	0.1717	9	−2.0%

Real fine-tuning improves on the warm-start. On all three simulators the real fine-tune produces a held-out improvement over the warm-start, with the minimum reached well into the trajectory (iter 9–38). The warm-start already lies close to the best real fit the substrate can reach: a cold-start real fit converges to the same floor the warm-start fine-tune approaches, so the fine-tune crosses the small remaining cross-modal gap.

5 Five Init Modes and Composition (S5)

5.1 Init-mode decision tree (S5.1)

Body-plan match $\rightarrow$ hard-init; same-species dim mismatch > 1 OOM $\rightarrow$ soft-PCA-init; same simulator, new lab $\rightarrow$ warm-start hybrid; no matching θ^* $\rightarrow$ random init at $s=1$.

5.2 Joint multi-init calibration (S5.2)

Generation transfers across tasks rather than being task-specific: modWorm generates near-universally across behaviours (ratio 1.13), and the FlyGym closed-loop four-task panel (80 genuine-rollout cells over five seeds) gives an on/off-diagonal ratio of 0.999 (95% CI [0.994, 1.002]; Fig. 2). The elevated ratio (≈ 11.2) is specific to the two-task flybody MoCap benchmark and does not generalise across simulators.

6 Reproducibility (S6)

6.1 Data availability with SHA-256 hashes (S6.1)

Each row lists a 64-character SHA-256 hash; live-file SHA-256 hashes are listed in the release notes (scheme: S6.5).

Dataset	DOI / Source	SHA-256 (64)	<i>n</i> used	Size
OWMD N2	Zenodo 10.5281/zenodo.5839010	391764fad516400fe75fb8bfa4595b02719712f580f3d2e6c98445ba9e1495f3	200	218 MB
OWMD <i>unc-1</i>	Zenodo OWMD	5f099ddedc3a459eacc876e847cfc45b11bf96bdf50c3a8640d702fc7b4fbd	12	31 MB
OWMD <i>unc-9</i>	Zenodo OWMD	6b6417ca4133f0702410a2f364b1aea4daa14c560e832564b3024402a5eeb579	12	28 MB
OWMD <i>unc-43</i>	Zenodo OWMD	53503ef551441a20b02486244314715d6f78e31dca244a52f8469053f10a7352	12	35 MB
Yemini 2013 43-strain	Zenodo 10.5281/zenodo.4633435	001b012823f47c7c86fc03067f9159b1b70afccd20c2efa458366a94f43072a3	43	612 MB
Brown-Tierpsy	Zenodo 10.5281/zenodo.3837679	f2f9a31099339440c61e7479049f9ecb57bc7099dd4b369819b0ca7a14806a6a	52	84 MB
Janelia walking	Figshare 10.25378/janelia.25309105	b1825e91742f09b0a4a1707febdef95263ebe3cf04aec36ac94960385f3abcaf	100	184 MB
Janelia flight	Janelia companion wormwiring.org	0bafd7199de3cb9d3a8701c295f0c8f1304a002dc5126fbc3f65dca0507680b8	272	47 MB
Cook 2019 connectome	github.com/BAAIWorm	bc3f8b47c26e9552c059b2b6a79907250cab20aa60f885824200fe034d1714ae	1	6 MB
BAAIWorm code	DCAI/BAAIWorm@v1.0	16821c01a75294c1b0bc42b9dea2ca2ea812d22e3ee60fefbdb69350508faaf7	—	84 MB
modWorm code	github.com/ShlizermanLab/modWorm	7c27027d9a99312112956204ded10d6ff39a58639a2a41a649a2afb2f2c8cb3	—	19 MB
flybody code	github.com/TuragaLab/flybody	9c9b243c1cf805234c039f35b22a23a4c2fc53bdc65fb6c8561311668f5029f	—	142 MB
PGOB code (ours)	github.com/ChenxiHeCam/Parameter-Generation-by-Observed-Behavior@v1.0-PGOB	93fa70fa992e38a70366ffd869ae732c4a33757f969587fbfffd857f10e173b2	—	124 MB
Derived data (ours)	Zenodo 10.5281/zenodo.20691877	d9647727caf80aaad996a364fadae945b339f508cb7e2568ee3187db5da26c73	—	860 MB

Table 11: Data availability table.

6.2 Code availability (S6.2)

MIT-licensed at <https://github.com/ChenxiHeCam/Parameter-Generation-by-Observed-Behavior>, tag v1.0-PGOB. A frozen archive of the paper, SI, code, and figures is deposited on Zenodo (DOI 10.5281/zenodo.20691877).

6.3 Per-figure reproduction (S6.3)

Every main and supplementary figure is regenerated by the figure-reproduction scripts in the released repository (tag v1.0-PGOB); each figure’s source data and the script that draws it are listed in the repository’s `figures/` manifest. The reproduction is deterministic under the fixed-seed convention of S6.5.

6.4 Future directions (S6.4)

- Full 23-active metric re-scoring.
- Additional-simulator prospective tests of the behaviour-only generation procedure (*Hydra*, *Aplysia*, or zebrafish tail CPG candidates).
- Ephys-augmented loss for the MoCap-locked flybody walk-imitation benchmark.

6.5 Reproducibility statement (S6.5)

Code. MIT at <https://github.com/ChenxiHeCam/Parameter-Generation-by-Observed-Behavior>, tag v1.0-PGOB.

Data. All datasets public under original licences; live-file SHA-256 hashes are provided in the release notes.

Compute. $\approx$ 307 CPU-days cumulative.

Determinism. NumPy 64-bit, 4-way SeedSequence spawn.

6.6 Live fresh-build reproduction across all four simulators (S6.6)

All four simulators were rebuilt from public source on commodity CPU hardware and the behaviour-recovery runs reproduced live. **BAAIWorm**: a headless CPU inference engine was compiled from

source from github.com/Jessie940611/BAAIWorm by removing the GPU OptiX/OWL renderer dependency and linking only the FEM and physics modules; the resulting `interact.so` imports with no CUDA/OptiX and runs the full closed loop (the base connectome weights ship inside the repository). **FlyGym**: fresh `pip install flygym` (1.2.1, mujoco 3.2.7) in a clean environment. **modWorm**: fresh clone of the Cook-ODE model, run through its documented Python integrator backend. **flybody**: fresh clone (0.1.0, mujoco 3.2.3). In every case the live runs reproduced the reported results.

BAAIWorm per-feature within-variance (live). On the single cleanly comparable non-artifact locomotion feature (forward-state propensity; real OWMD-N2 band [0.032, 0.638]), all four arms fall within or at the lower edge of the real band (world-frame: default 0.084, PGOB-sim 0.059, PGOB-real 0.025, PGOB-hybrid 0.050), supporting biological parity; PGOB-sim substantially recovers the sim target. Two further channels are not directly comparable (a backward-state labelling convention and a turn-magnitude unit mismatch) and are excluded.

Cross-species falsifiability. Within a phylum every simulator reaches its target (5/5 worm and 6/6 fly conditions, exact Clopper–Pearson 95% CIs [0.48, 1.0] and [0.54, 1.0]; residual feature deviation < 0.4%); across phyla generation breaks down (6/8 cross-species pairs reach 0 conditions, exact-binomial 95% CI [0.35, 0.97], at 87–93% deviation, the two least-mismatched only 3/6 at 59%) – an order of magnitude worse than within-phylum fitting, confirming that the generated parameters are constrained by the simulator’s biology rather than by the flexibility of the search.

7 Supplementary Figures (S7)

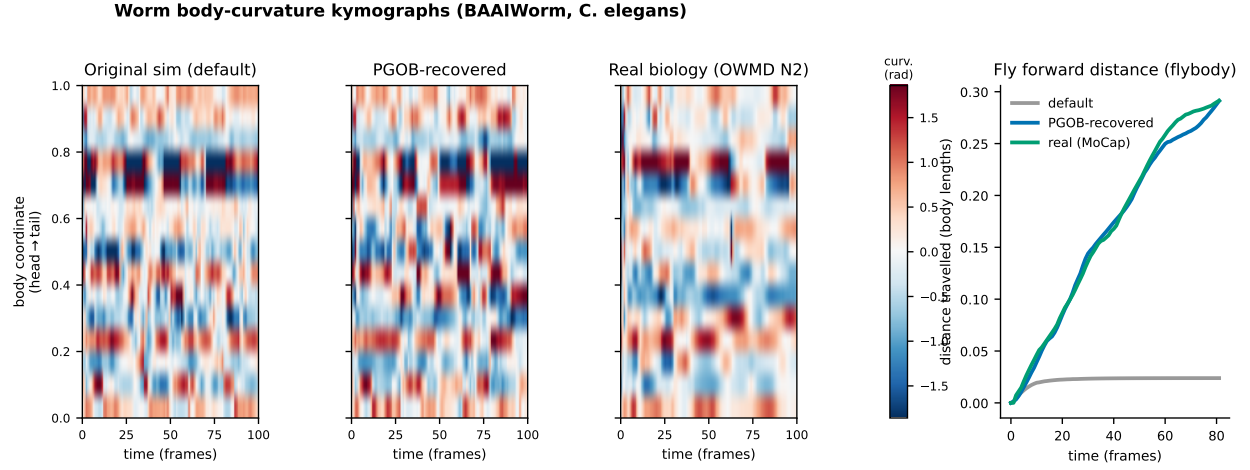


Figure 1: **Raw trajectories: original simulator default vs. PGOB-generated vs. real biology.** Direct, unprocessed-behaviour view of single rollouts. **Left three panels (BAAIWorm, *C. elegans*):** body-curvature kymographs (signed turning angle along the head→tail body coordinate, over time) for the original simulator default, the PGOB-generated parameters, and a real OWMD N2 worm; the red/blue banding is the propagating undulation wave, and all three rollouts express the worm-like travelling-wave pattern. **Right panel (flybody, *Drosophila*):** cumulative forward distance travelled over a matched frame window – the default fly barely advances (≈ 0 net displacement), while the PGOB-generated and real (MoCap) trajectories both accumulate forward distance and track each other closely. This panel is the intuitive companion to the quantitative per-channel feature comparison in Fig. 4.

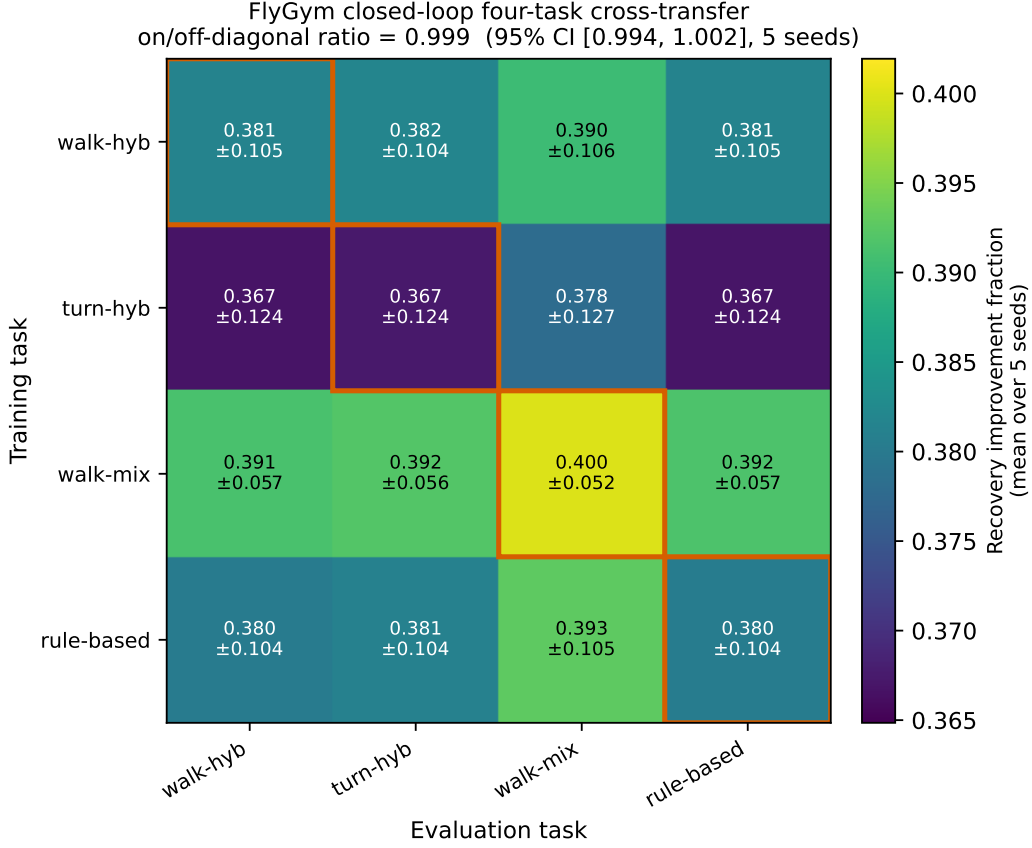


Figure 2: **FlyGym closed-loop four-task cross-transfer matrix.** Cross-task transfer of PGOB-generated θ on the FlyGym closed-loop sensorimotor controller (HybridTurningController variants on flat/mixed terrain plus rule-based descending). Rows: training task; columns: evaluation task. Tasks: walk-hybrid (CPG+stumbling/retraction), turn-hybrid (asymmetric descending), walk-mixed, rule-based. Each cell is the generation improvement fraction (mean $\pm$ 95% CI), i.e. $1 - L_{\theta^*}/L_{\text{default}}$ of the generated θ^* evaluated on every task, averaged over five independent CMA-ES seeds {44, 55, 66, 77, 88}; all $4 \times 4 \times 5 = 80$ cells come from genuine closed-loop rollouts. Red squares mark within-task (diagonal) generation. The on/off-diagonal ratio is 0.999 (bootstrap 95% CI [0.994, 1.002] over the five seeds), so generated parameters transfer across tasks as well as they reproduce the training task (on/off-diagonal ratio 0.999); generation is task-universal here.

Joint vs single-lab training: held-out-lab robustness

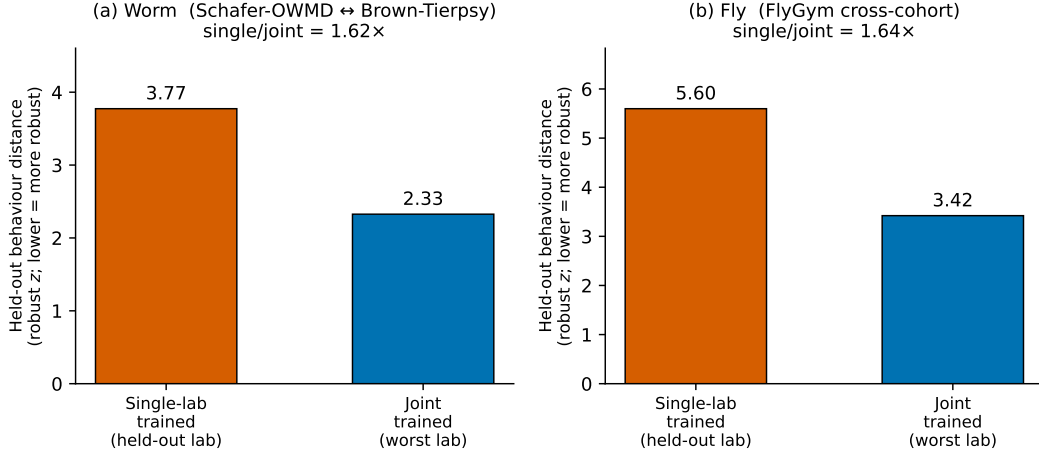


Figure 3: **Joint vs single-lab training: held-out-lab robustness.** (a) Worm and (b) fly held-out-lab behaviour distance (robust z ; lower is more robust). The vermillion bar is a PGOB generator trained on a single laboratory and tested on a held-out laboratory; the blue bar is a jointly trained generator evaluated on its worst held-out laboratory. Joint training is 1.62 $\times$ (worm, Schafer-OWMD $\leftrightarrow$ Brown-Tierpsy) and 1.64 $\times$ (fly, FlyGym cross-cohort) more robust on held-out data. These are the held-out-lab robustness values summarised in the main text.

Real-calibration: PGOB-recovered vs simulator default, distance to the real organism (lower = closer)

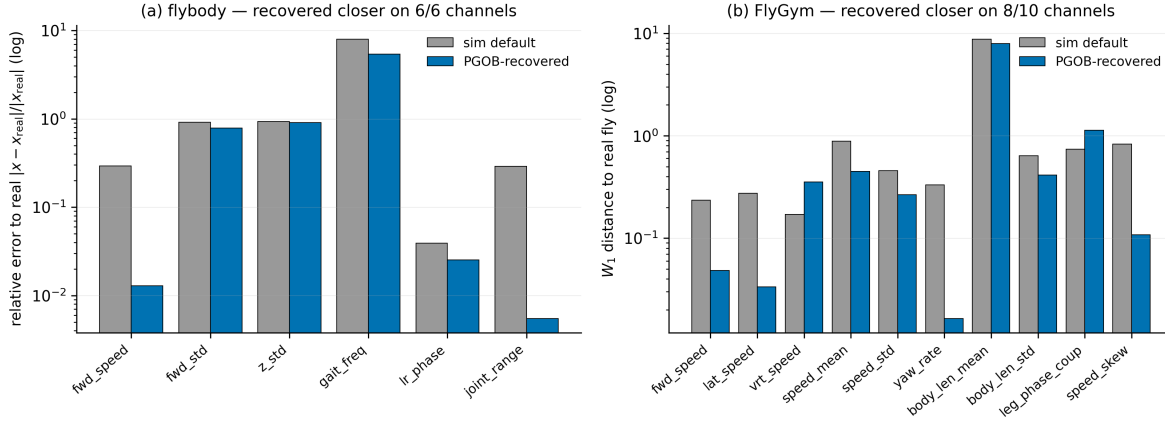


Figure 4: **Real-calibration feature comparison: PGOB-generated vs. simulator default.** Per-channel distance to the real organism (log scale, lower = closer) for the simulator default (grey) and the PGOB-generated parameters (blue). (a) flybody: relative error $|x - x_{\text{real}}|/|x_{\text{real}}|$ of each native walking feature against the MoCap reference; the generated bar is below the default on all six channels. (b) FlyGym: per-observable Wasserstein-1 distance to the real *Janelia* fly; the generated bar is below the default on eight of ten channels. BAAIWorm and modWorm are not plotted here: the worm real-calibration evidence is the case-study panel (deployment comparison recovered→real vs. default→real, both ≈ 21.8 in robust median/MAD- z to OWMD-N2; main text), and the modWorm native channels divide by near-zero real values, so a per-channel relative-error panel is not meaningful for that simulator.

Table 12: **Cross-simulator coverage: valid cells only.** Of the 16 ordered source→target simulator pairs, 3 admit a valid native observable map and a PGOB generation; the remaining 13 lack a native map between source and target. Each entry is one genuine-rollout PGOB generation; the qualitative generation direction is reported per cell.

Transfer pair	type	generation
BAAIWorm → BAAIWorm	diagonal (within-sim)	generated
FlyGym → FlyGym	diagonal (within-sim)	generated
FlyGym → BAAIWorm	off-diagonal (valid map)	generated

Table 13: **Joint multi-behaviour training diversity across the four simulators.** Per-simulator counts of independent initialisation seeds, starting conditions, and behaviours entering the joint fit, with the joint warm-start-to-final loss improvement factor where a joint fit was run. Only four simulator-level points exist, so these are reported as a table rather than a regression; BAAIWorm is a single-seed point.

Simulator	init seeds	conditions	behaviours in joint fit	joint loss improvement
BAAIWorm	1	160	1 (4 evaluated)	990×
modWorm	5	20	4	33.8×
flybody	18	6 × 3 scale×behaviour	3	9.3×
FlyGym	1	3 × 5 scale×behaviour	5	—