TOWARDS SCALING SEARCH IN ROBOT CO-DESIGN

Anonymous authors
Paper under double-blind review

ABSTRACT

Robot co-design has begun to scale its search to millions of candidate bodies. Search methods that train a controller for every body can scale by screening candidates with brief training, keeping the highest-scoring body and retraining it fully. We measure the screen's exact gain over a random body as the screen budget, candidate count (up to 128) and total compute vary. Each body in our pools has separately trained screening and final controllers. In three of four design spaces, keeping the best of four after a quarter of the full training budget improves the final score by 0.63 to 0.73 standard deviations over a random body. More candidates add to the gain, and at the same total compute the screen beats training two bodies fully and retraining the better one by 0.27 to 0.29 standard deviations. In a Walker2d variant whose reward pays only for forward travel, the same screen picks bodies that travel about 5% less than a random body. A preregistered test confirmed this harm, and it replicated with new seeds and grew with longer final training. Brief training favors heavy walkers that are weak for their weight, which full training ranks low. The harm fades with more candidates. Continuing this screen's controllers costs less and beats retraining in every pool we trained. In the walker, keeping the best of four after half the training budget instead improves the final score by 0.09 to 0.11 standard deviations over a random body. We release all our pools and code at `anonymous.4open.science/r/scaling-codesign-search`, and any best-of-N screen at our recorded budgets can be scored on the pools without training.

1 INTRODUCTION

Scaling has transformed language modeling and other fields of machine learning (Kaplan et al., 2020; Hoffmann et al., 2022), and robot co-design has begun to scale its search. Deep evolutionary reinforcement learning (DERL) tried about 4,000 bodies, each with its own trained controller (Gupta et al., 2021), and LOKI tried about 3.12 million by sharing one controller per cluster of similar bodies (Jeon et al., 2025). For methods that train a controller for each body, the compute budget must be split between trying more bodies and training each for longer. We ask how much a larger search buys under brief training, and when such cheap search picks worse bodies than a random draw.

A common shortcut is to *screen* candidates. Controllers are trained briefly on N candidate bodies, and the body that scores best gets a fresh controller trained with the full budget (Arza et al., 2025). Arza et al. (2025) found that such a screen beats full training at equal compute. We measure its exact gain over a random body at every N and recorded screen budget on released pools, confirm the gain on new bodies, and find where screening fails. Screening pays only if the body that does best after brief training also does well after full training.

That assumption can fail. In neural architecture search, proxies that correlate with final accuracy can still pick worse than random at the top of their ranking (Abdelfattah et al., 2021; Ning et al., 2021). Best-of-N sampling against a proxy reward model eventually lowers the true reward (Gao et al., 2023), and short training can favor candidates that learn fast over those that finish well (Wu et al., 2018; Mertan & Cheney, 2026; Song et al., 2026).

We measure the *selection lift* of a screen, the amount by which the body it keeps beats a random body from the same pool once both are fully trained. Our pools of 256 to 2,048 bodies span four design spaces, with four controller seeds per body and returns recorded at twelve budgets. On them we evaluate every best-of-N screen exactly with an established order-statistic identity. Screening

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.

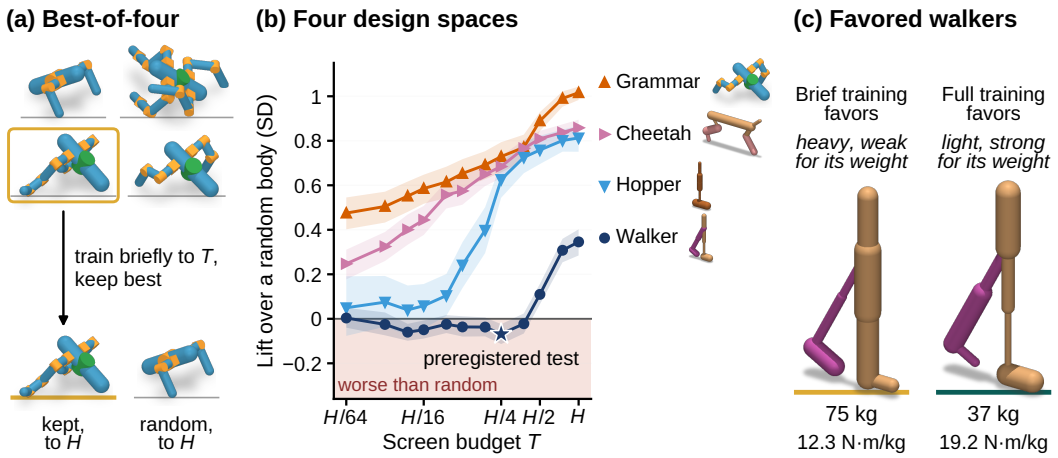


Figure 1: Screening robot bodies and what it buys. (a) Best-of-four screening with screen budget T and full budget H. (b) Lift of best-of-four over a random body against the screen budget, with 95% intervals. The star marks the walker's test on new bodies. (c) Walkers favored by brief and by full training. **Takeaway:** The quarter-budget screen that helps in three design spaces picks worse walkers than a random draw, because brief training favors heavy walkers that are weak for their weight.

and final training use separate seeds, so the lift describes a pipeline that retrains its winner from scratch. We preregistered the central tests.

Screening pays in three of the four spaces. In hopper, cheetah and a grammar-built space, best-of-four after a quarter of the full budget gains 0.63 to 0.73 standard deviations (SD) of the final score over a random body (Fig. 1). The gain grows with N and usually with the screen budget (Figs. 1b and 3). At the same total compute, the quarter-budget screen beats training two bodies to the full budget and retraining the better one by 0.27 to 0.29 SD. A test on new cheetah bodies confirmed that a screen of four at an eighth of the full budget helps. In an architecture-search analysis, screens using validation accuracy picked significantly better architectures than random at every setting in the plan. In a public archive of 1.3 million EvoGym bodies whose evolved controllers are continued, not retrained (Mertan & Cheney, 2026), screening on fitness later in evolution gains at every N we evaluate, while screening at the first generation picks worse than random over a range of N.

Among the spaces where the winner is retrained, the walker is the exception. In our variant of Walker2d, a preregistered test on new bodies found that keeping the best of four after a quarter of the full budget picks bodies that travel about 0.2 m less than a random body's 4 m once retrained. A second, independent set of controller seeds showed the same harm. The harm grew when the chosen bodies were trained twice as long. Brief training favors heavy walkers that are weak for their weight, and full training favors light walkers that are strong for their weight (Figs. 6 and 7). The harm fades with more candidates, continuing the screened controllers beats retraining, and a screen twice as long picks better walkers than random.

Our contributions are as follows:

- **Released pools.** Seven fully trained pools with separate screening and final controllers, and an exact order-statistic evaluation that scores any best-of-N screen at our recorded budgets on them without training (Secs. 2 and 3.1).
- **What a larger search buys.** In three of four design spaces the screen's gain keeps growing with the number of candidates, and at equal compute the screen beats training fewer bodies fully (Sec. 4).
- **When cheap search fails.** In the walker, brief training favors heavy bodies that are weak for their weight, so the screen picks worse than a random draw, a harm that was preregistered and replicated. Screening with at least half the final budget, or continuing the screened controllers, avoids it (Sec. 5).

2 PROBLEM FORMULATION

2.1 SCREENING AND RETRAINING

A design space defines a distribution over bodies, and a pool is a set of D bodies drawn from it. For offline evaluation, we train every body's controllers with the same learner to a full budget of H environment steps and use earlier checkpoints as screens. A body's *screen score* x is the mean return of k_s runs at $T \leq H$ steps, and its *final score* y is the mean return of $k_f = 2$ separate, independent runs at H steps. A best-of-N screen draws N bodies at random from the pool, keeps the one with the highest screen score, and trains fresh controllers only for that body to obtain its final score.

2.2 SELECTION LIFT

Definition 1 (Finite-pool selection lift)**.** *Given a pool $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{D}$ with $D \geq 2$, draw A uniformly from the subsets of size N and let $I(A)$ maximize x_i over $i \in A$, with ties broken uniformly at random. The selection lift is $L_{\mathcal{D}}(N,T) = \big(\mathbb{E}[y_{I(A)}] - \bar{y}\big)/s_y$, where the expectation is over A and the tie-break, and $\bar{y}$ and $s_y > 0$ are the mean and sample SD of the pool's final scores. A negative lift is* harm*.*

Proposition 1 (Exact evaluation)**.** *Order the pool by increasing screen score, $y_{(r)}$ being the final score of the body at rank r, and let $w_r = \binom{r-1}{N-1}/\binom{D}{N}$. For $1 \leq N \leq D$,*

$$L_{\mathcal{D}}(N,T) = \frac{1}{s_y} \sum_{r=N}^{D} w_r \left(y_{(r)} - \bar{y}\right), \tag{1}$$

with the final scores of each block of tied screen scores replaced by the block mean.

Proof. Rank r is kept exactly when it is drawn and the other $N-1$ draws rank below it, which holds for $\binom{r-1}{N-1}$ of the $\binom{D}{N}$ subsets. With ties, the kept body lies in a tied block with probability equal to the block's summed weight and is equally likely to be any of its members, so each member's final score is replaced by the block mean. Because the weights sum to one, $\bar{y}$ moves inside the sum. □

The evaluation is exact given the recorded scores, and the intervals of Sec. 2.3 describe the uncertainty over new bodies and new training runs.

Corollary 1. *$L_{\mathcal{D}}(N,T) = D\,\mathrm{Cov}_r\big(w_r, y_{(r)}\big)/s_y$, with ranks r weighted uniformly. The kept body's expected screen rank is $N(D+1)/(N+1)$, with ranks within tied blocks ordered by the uniform tie-break.*

The covariance form is Price's equation with each rank's probability of being kept as its fitness (Price, 1970). The lift therefore weights the top of the screen ranking, and its sign can differ from that of the pool-wide correlation between screen and final scores. If the screen and final score pairs were drawn independently from a bivariate normal distribution with correlation ρ, the lift at fixed N would approach $\rho\, e_N$ as D grows, where e_N is the expected maximum of N standard normals. This limit is the expected induced selection differential (Nagaraja, 1982; David & Nagaraja, 2003).

Equation 1 is the expected concomitant of a sample maximum (David & Nagaraja, 2003) and the finite-population selection differential of quantitative genetics (Robertson, 1966; Price, 1970; Walsh & Lynch, 2018). Its weights are the distribution of the largest rank in a sample drawn without replacement (O'Neill, 2025). In machine learning it is the best-of-N estimator of Nakano et al. (2021, App. I) and Gao et al. (2023). When the score is binary and serves as its own screen, the kept body's expected score before standardizing is pass@k (Chen et al., 2021). We claim no new estimator.

2.3 UNCERTAINTY AND SEARCH SIZE

With four seeds per body, a *seed split* uses two seeds for the screen score and the other two for the final score. We average the lift over the six possible splits. They reuse the same four runs, so together they give one estimate rather than six independent ones. We compute 95% bootstrap

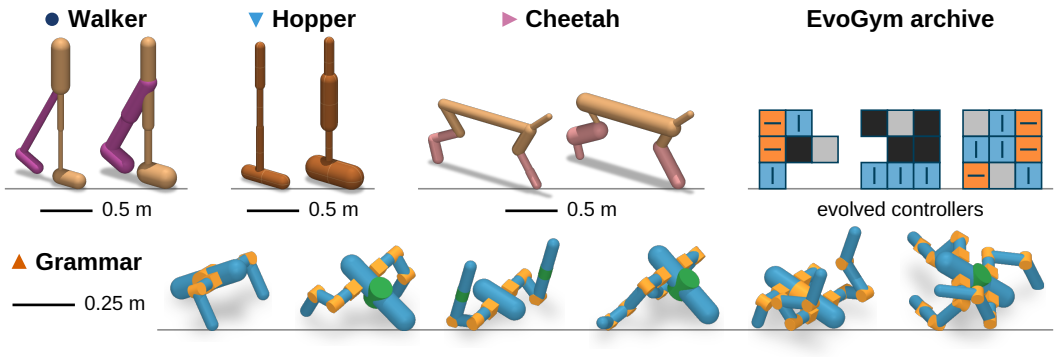


Figure 2: Bodies from the four design spaces we train and from the public EvoGym archive differ widely in shape, size and actuation. Walker, hopper and cheetah bodies at the 10th and 90th percentiles of body volume share one scale, archive bodies sit at the 10th, 50th and 90th percentiles of voxel count, and grammar bodies span the number of actuated joints at their own scale.

intervals (Agarwal et al., 2021) by resampling bodies together with their runs (App. A), and call a lift *significant* when its interval excludes zero. A *setting* is one screen budget and one value of N, and a claim that covers many settings at once uses a simultaneous 95% band over them. As N grows, the weights of Eq. 1 concentrate on a few top ranks, so we report a value of N only while the effective number of bodies, $1/\sum_r w_r^2$, is at least 30. This rule allows $N \leq 16$, 65 and 131 for pools of 256, 1,024 and 2,048 bodies, about $D/16$.

2.4 Compute Cost

We count cost in environment steps, in units of H. Screening N bodies with k_s seeds for T steps costs $k_s NT$, retraining the winner with two seeds costs $2H$, and the total is $k_s NT + 2H$. A random body costs $2H$. The quarter-budget screen of four, with two seeds and $T = H/4$, costs $2H$ to screen and $2H$ to retrain. Its total of $4H$ equals that of best-of-two at the full budget, which trains two bodies with one seed each for H and retrains the better one. Continuing the winner's k_s screened controllers to H instead of retraining it costs $k_s NT + k_s(H - T)$, or $3.5H$ for the quarter-budget screen of four.

3 Experimental Setup

3.1 Design Spaces

Walker, hopper and cheetah (Fig. 2) rescale one Gymnasium body's limb radii, actuator gears and density (total mass for cheetah), keeping limb lengths fixed. The grammar space uses RoboGrammar's grammar (Zhao et al., 2020). A motor's gear is the largest torque it can apply, so a body's summed gear divided by its mass is its peak joint torque per kilogram, which we call its *strength for weight* (N·m/kg). Every task rewards forward travel on flat ground. The original pools hold 996 walkers (the pilot pool), 256 hoppers, 512 cheetahs and 512 grammar bodies. After scoring them we trained 1,024 new walkers (the test pool), 2,048 new cheetahs and 2,048 new grammar bodies. In all seven pools, screen and final scores come from separately trained controllers, so we call them the retrained pools. The test walkers were trained again with four new controller seeds, and we call the two groups of four seeds the first and second seed sets. Curves against the number of candidates use the 2,048-body cheetah and grammar pools, and all other results use the original pools and the test walkers' first seed set unless stated otherwise. In the grammar space nearly half the bodies fall over even with no action, so its earliest screens gain partly by rejecting them. Among the bodies that stay up, the quarter-budget screen of four still gains $+0.47$ SD in the new pool and $+0.39$ SD in the original (App. E).

Our walker is a variant of Gymnasium's Walker2d with DERL's locomotion reward (Gupta et al., 2021). It rewards forward travel and subtracts a small action cost. An episode ends when the walker

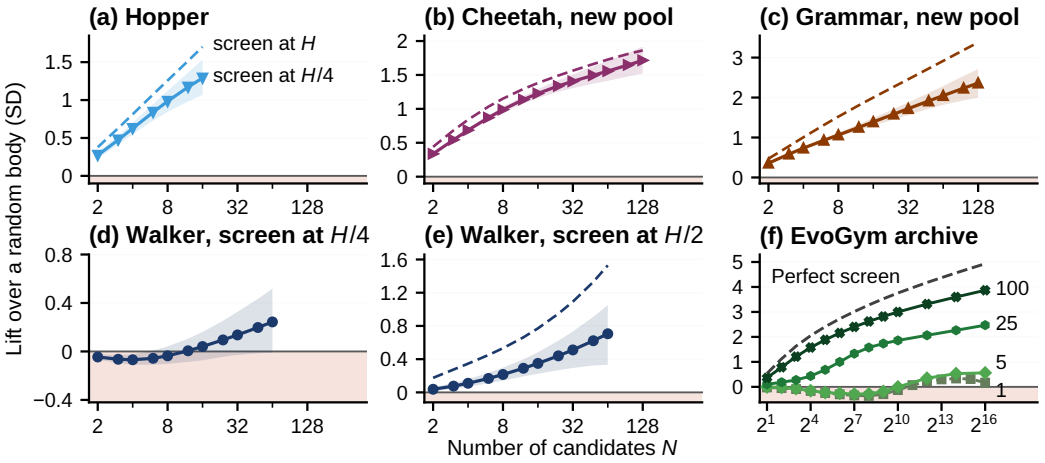


Figure 3: Lift of best-of-N over a random body against the number of candidates N. Shading shows 95% intervals, except for the dashed full-budget screens and the archive, and pink marks where a screen does worse than a random body. (a–c) The screen at a quarter of the full budget and at the full budget (dashed). (d, e) The walker with the screen at a quarter and at half of the full budget, and in (e) at the full budget (dashed). (f) The EvoGym archive, with the screened controller continued, for screens at the generation printed at each line's end and for a perfect screen. **Takeaway:** In hopper, cheetah and grammar more candidates add to the gain, and in the walker the harm of the quarter-budget screen first deepens, then fades as more candidates are added.

falls or after 8 s. We also report walker returns in meters of travel, where one SD of the test pool's final return is about 2.7 m (App. F gives the conversion).

3.2 Controller Training and Evaluation

Every controller in our pools is trained with the same PPO hyperparameters in MuJoCo MJX. The full budget H is 64 updates of 131,072 steps, 8.4 million steps in all, and returns are recorded at twelve budgets from $H/64$ (one update) to H (App. B lists every setting). Each return is one deterministic episode. A body's final score is the mean over two seeds, and so is its screen score unless stated otherwise. Final scores are read at a fixed budget rather than at convergence. At that budget the mean walker travels about 4 m, its returns are still rising, and 98% of final episodes end in a fall, so a walker's final score measures how far it gets by the end of the episode. The walker's final scores are also noisier than those of cheetah and grammar. Differences between bodies explain at most 0.34 of the test pool's two-seed final-score variance, against 0.78 in cheetah and 0.90 in grammar (App. E). Noisier final scores shrink a lift toward zero.

3.3 Preregistered Tests

A preregistered test had its prediction and decision rule fixed and recorded before its result was seen. Every other analysis is post hoc, not registered in advance, and figures and tables mark preregistered results with †. Our claims rest on such tests of the walker, its follow-ups on new final controllers and on held-out walkers, a cheetah permutation test and an analysis of architecture-search benchmarks (App. D). Our anonymous repository lists every registered test, its outcome and any deviation from its registration.

4 How Screening Scales

4.1 Screen Budget

Screening helps even at small budgets. Simultaneous bands over the twelve budgets show that best-of-four beats random from $H/64$ to H in cheetah and grammar, and from $H/8$ in hopper (Figs. 1

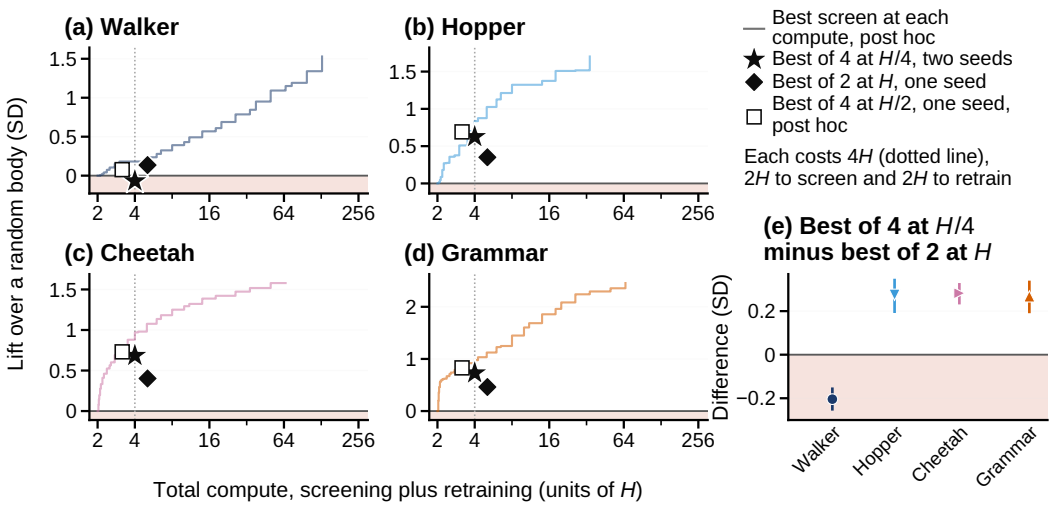


Figure 4: Compare the star with the diamond at $4H$. Lift over a random body against the total compute of screening and retraining the kept body, in units of the full budget H. The line is the best lift of any screen on our grid at or below each level of compute, chosen post hoc, and Fig. 11 shows every screen. The black markers each cost $4H$, with the diamond and square offset horizontally for legibility. (e) The star minus the diamond, with 95% intervals. **Takeaway:** Outside the walker, the quarter-budget screen of four beats training two bodies fully and retraining the better one, at the same compute.

and 12). Longer screens usually improve selection, and at $H/4$ the gains are $+0.63$, $+0.68$ and $+0.73$ SD in hopper, cheetah and grammar. In the five hopper, cheetah and grammar pools, 528 of the 561 settings with screens from $H/64$ to $3H/4$ have significantly positive lift under per-pool simultaneous bands, and none has significantly negative lift (App. C). A preregistered permutation test on 2,048 new cheetah bodies supported positive lift for best-of-four at $H/8$, estimated post hoc at $+0.58$ SD (App. D).

Screens also pick better than random outside robotics. In a preregistered analysis of architecture-search benchmarks, validation-accuracy screens based on short or partial training had significantly positive lift under per-pool simultaneous bands at all 394 settings (App. C).

4.2 Number of Candidates

With a quarter-budget screen, the lift in hopper, cheetah and grammar grows up to the largest N each pool supports, reaching $+2.36$ SD at $N = 128$ in the new grammar pool.

To follow the lift beyond our pools, up to $N = 65{,}536$ (Fig. 3f), we use the public archive of Mertan & Cheney (2026), which holds all 1,305,840 connected 3×3 EvoGym voxel bodies with at least three actuated voxels (Bhatia et al., 2021), each with an evolved controller. Because screen and final fitness come from one controller run, these curves describe continuing the screened controller rather than retraining it (App. G). Screens from generation 25 of evolution onward gain at every N, and the gain keeps rising with N. Generation 25 is the first plotted generation where this holds. At generation 100, best-of-256 gains $+2.62$ SD, 80% of what a perfect screen would gain. A generation-1 screen picks worse than random over a range of N, a second case of screening harm.

4.3 Equal Total Compute

At equal total compute, screening beats training fewer bodies fully in every retrained space except the walker (Fig. 4). Every pipeline retrains its winner from scratch, so the comparison isolates what the screen's choice of body is worth. In all five hopper, cheetah and grammar pools the quarter-budget screen of four beats best-of-two at the full budget by 0.27 to 0.29 SD (Fig. 4e and App. C).

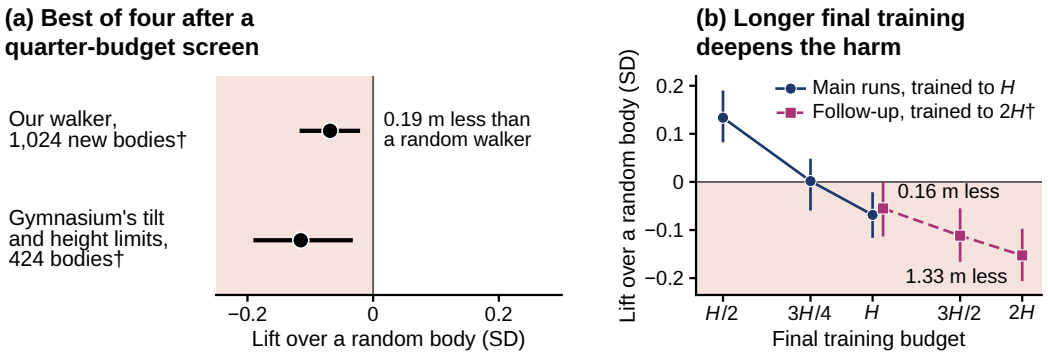


Figure 5: Lift of best-of-four in the walker, with the screen at $H/4$ and the kept body retrained from scratch. (a) At the full final budget, for our walker on 1,024 new bodies and for 424 held-out walkers whose episodes end under Gymnasium's tilt and height limits. (b) Against the final budget, for the main runs and for new final controllers trained to $2H$ (purple, dashed, offset at H for legibility). Each lift is in SD of the scores at its final budget. Bars show 95% intervals, and † marks preregistered tests. **Takeaway:** The quarter-budget screen picks worse walkers than a random draw, and longer final training deepens the harm.

In the walker the order reverses. Best-of-two at H beats the quarter-budget screen by about 0.20 SD, and a one-seed screen of four at $H/2$ beats it by about 0.15 SD.

Retraining evaluates the kept body with fresh controller runs, while continuation keeps training its screened runs. For the quarter-budget screen of four, continuing the winner's two controllers to H costs $3.5H$ rather than $4H$ and gives a significantly higher lift than retraining in every retrained pool and seed set, by $+0.03$ to $+0.13$ SD (post hoc, Table 6).

5 When Screening Fails in the Walker

The 996-body pilot pool first showed the harm for best-of-four at $H/4$, and a preregistered test on new bodies confirmed it (App. D). The test trained four controller seeds on each of 1,024 new bodies. The test required a one-sided 95% upper limit below zero and a point estimate of at most -0.03 SD for the harm. It passed with a lift of -0.07 $[-0.11, -0.02]$ SD and a one-sided upper limit of -0.03 (Fig. 5a). The same rule applied to four new controller seeds on the same bodies gave -0.08 $[-0.12, -0.04]$ SD (upper limit -0.05). The pilot pool gave -0.06 $[-0.11, -0.02]$. The harm also appears on 424 held-out walkers whose episodes end under Gymnasium's tilt and height limits (Fig. 5a). In the test pool, the harm holds under a simultaneous band over all twelve budgets, in all 70 four-seed subsets of the eight seeds, and when seeds and bodies are resampled (App. F).

In the two seed sets, the kept body traveled about 0.19 and 0.22 m less than a random body's 4 m. This deficit is 20% and 27% of what a screen of four at the full budget gains ($+0.35$ and $+0.31$ SD).

The harm is a best-of-few effect. Simultaneous per-pool bands over every reported N and all eleven screen budgets below H still show significant harm at $H/4$ for N up to 3 in the first seed set and up to 6 in the second. With more candidates the lift rises, and its point estimate is positive from $N = 12$ in the first seed set and the pilot pool and from $N = 48$ in the second (Figs. 3d and 9d). A screen twice as long, at $H/2$, has a positive lift at every N, $+0.11$ $[+0.07, +0.16]$ and $+0.09$ $[+0.04, +0.14]$ SD at $N = 4$ against final scores at H, and the lift grows with N (Fig. 3e).

5.1 Longer Final Training

The sign of the quarter-budget screen's lift depends on how long the chosen body is trained. It helps when the chosen body's final training stops at $H/2$ and harms when it continues to H (Fig. 5b). A preregistered follow-up trained new final controllers for the test bodies to $2H$ and scored the same screens against them. Against these controllers, and in each budget's own SD, the harm grew from -0.06 SD at H to -0.15 SD at $2H$, and the paired increase met the condition fixed in advance

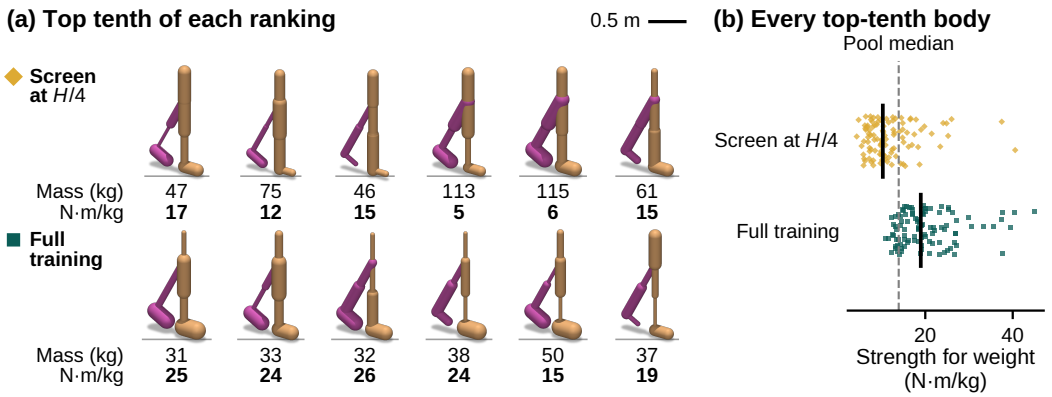


Figure 6: The walkers each ranking favors. (a) Six bodies evenly spaced through the top tenth of test walkers, counting from the best, ranked by the quarter-budget screen (top row) and by full training (bottom row). (b) Strength for weight of every top-tenth body, with medians as black ticks. **Takeaway:** Brief training favors heavy walkers that are weak for their weight, and full training favors light walkers that are strong for their weight.

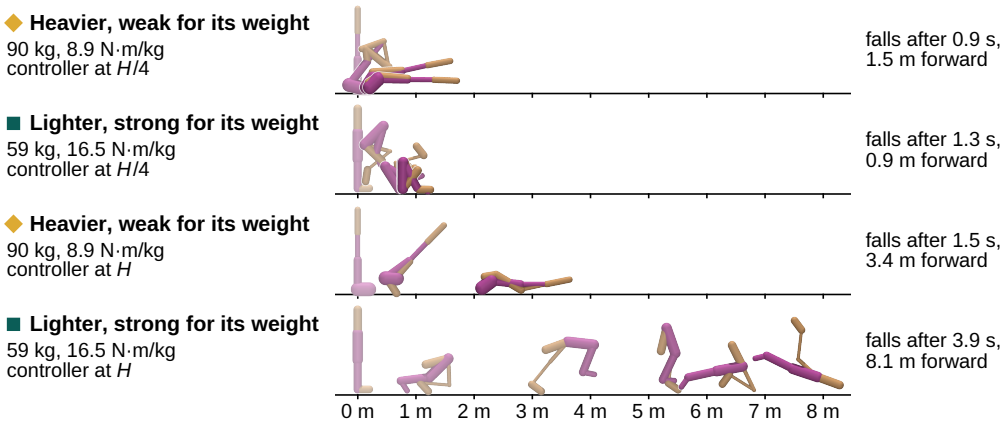


Figure 7: One rollout per row of two pilot-pool walkers, a heavier one that is weak for its weight and a lighter one that is strong for its weight, each after a quarter of the training and after full training, with travel and fall time labeled. **Takeaway:** After brief training both walkers tumble and fall within about a second. The heavier one travels farther in that episode, much of it while falling. After full training the lighter walker travels much farther.

for calling it deeper. In meters, the kept body's deficit grew from 0.16 m at H to 1.33 m at $2H$, or from 3.8% to 11% of a random body's travel. The SD scale understates this growth because the SD tripled.

Across our recorded settings, screens using between half and the full final-training budget had a non-negative estimated lift (App. C). At $N = 4$, we checked every recorded pair of screen and final budgets with the screen no longer than the final training, in the retrained pools and the additional walker analyses listed in App. C. With final training fixed at H, we also checked every reported N. In the retrained pools, every significantly harmful screen no longer than the final training was a walker screen of at most $3H/8$. Across all our recorded settings, every significantly harmful screen no longer than the final training used at most three eighths of the final budget (App. C).

5.2 Morphology and Motion

The quarter-budget screen measures something other than the final score. Its ranking is slightly anti-correlated with the final ranking (Spearman -0.08 and -0.12 in the two seed sets) and positively

correlated with final scores taken at $3H/8$ ($+0.28$ and $+0.26$). In replays of 512 pilot bodies, 98.7% of the quarter-budget controllers fall, after a median of 0.93 s and 1.16 m, and most falls end pitched forward. Each return counts travel through the fall until the episode ends, so the screen ranks how far a body gets by the end of an early fall (Fig. 7).

Heavy bodies do well at the quarter-budget screen and badly after full training. On four-seed means, the Spearman correlation of body mass with return is $+0.48$ and $+0.49$ at the quarter-budget screen and -0.45 and -0.51 at the full budget, while that of summed actuator gear changes about a third as much. Strength for weight shows the mirror image, correlating negatively with return at the screen and positively at the full budget (Figs. 6 and 17). In the hopper, whose reward also pays only for forward travel, the screen and full training both favor bodies that are strong for their weight (Spearman $+0.63$ and $+0.75$), and screening helps. In the walker, the rank lift, Eq. 1 applied to screen and final ranks instead of scores, is negative at $H/4$, like the rank correlation. Removing a linear fit on the rank of strength for weight from both rankings turns the rank lift positive, $+0.11$ and $+0.09$ SD of the residual final rank. Best-of-four drawn within quintiles of strength for weight, a different selection rule, has a significantly positive lift in both seed sets (post hoc, App. F). These are associations under our learner, whose actions and exploration noise both scale with actuator gear.

6 Related Work

Many co-design systems avoid training a controller for every body with shared, pretrained or body-conditioned controllers (Yuan et al., 2022; Gupta et al., 2022; Lu et al., 2025), and inherited or briefly trained controllers undervalue new bodies unless those bodies get time to adapt (Cheney et al., 2018; Mertan & Cheney, 2026). DERL trains a fresh controller for each mutated offspring and selects parents by tournaments of four, the size that motivates our $N = 4$ (Gupta et al., 2021). Closest to our question, Arza et al. (2025) found within evolutionary and tree searches that screening four times as many designs at a quarter of the budget and retraining the best design beats full training at equal compute. Our equal-compute comparison agrees outside the walker, and our pools score the screen itself exactly, as a lift over a random body at every N and recorded screen budget. Multi-fidelity morphology optimization models how scores relate across training budgets (Hu et al., 2021), and recent pipelines screen designs by brief training and keep the best, with successive halving (Nagiredla et al., 2025; Schneider et al., 2025) or at half the final budget (Fang et al., 2025), a length that avoided harm in all our pools. Successive halving, Hyperband and freeze-thaw methods screen hyperparameter configurations at small budgets (Jamieson & Talwalkar, 2016; Li et al., 2018; Swersky et al., 2014), and learning curves that cross can degrade such screens (Egele et al., 2024). With an imperfect reward model, best-of-N sampling can degrade as N grows (Gao et al., 2023; Huang et al., 2025; Khalaf et al., 2025). We measure the sign of the screening lift in co-design, with the winner retrained on separate seeds and tested on new bodies, and App. I covers further related work.

7 Conclusion and Limitations

Conclusion. In three of four design spaces a quarter-budget screen of four beats a random body by 0.63 to 0.73 SD and, at equal compute, also beats training two bodies fully and retraining the better one. Where the screen ranks bodies well, more candidates keep adding to the gain up to the largest N we evaluate, in our largest pools and in the EvoGym archive, so co-design search can scale by screening. In a Walker2d variant it picks worse-than-random bodies that brief training favors and full training ranks low, a harm that replicated and grew with longer final training. In the walker, the same quarter-budget screen helped when final training stopped at $H/2$ and harmed when it continued to H, so a screen budget should be chosen as a fraction of the final budget. With the quarter-budget screen of four, continuing the screened controllers instead of retraining the winner was cheaper and did better in every pool.

Limitations. Every design space rewards forward travel on flat ground, and all our trained pools share one set of PPO hyperparameters and one simulator. Pool sizes bound the number of candidates we can evaluate with retraining, up to $N = 128$, and the EvoGym archive reaches larger N only with continued controllers. We study one best-of-N step on fixed pools rather than iterative search, so whether design search follows scaling laws remains open.

REPRODUCIBILITY STATEMENT

Our anonymous repository holds the per-body, per-seed returns behind our results on the trained pools, from which the lift of any best-of-N screen at our recorded budgets follows without training. These returns cover all seven retrained pools at all twelve budgets (the walker test pool with both seed sets, the walker pilot, hopper, both cheetah pools and both grammar pools), the new final controllers of the $2H$ follow-up and the two groups of walkers run under Gymnasium's termination rule. It also holds the preregistrations with their amendments and scored outcomes, the estimator and analysis code with their tests, a script that checks each printed number against the result file it comes from, and the code that trained the controllers, but no script that reruns the training. The repository is at `anonymous.4open.science/r/scaling-codesign-search`.

ETHICS STATEMENT

The study concerns simulated robot selection and does not establish hardware safety or deployment performance.

AI USE STATEMENT

We conceived and directed this research. We used generative AI tools, including coding agents working under our direction, to implement methods, clean and reformat outcome data, help interpret results, give feedback on methodology and experimental design, propose or refine hypotheses, help derive the exact selection identity, search the literature, and help draft and edit the text. We did not use them to generate synthetic data, to alter any recorded outcome, or for translation. We verified all AI-assisted work. A script checks every printed number against its result file, and we checked the selection identity by brute-force enumeration. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Mohamed S. Abdelfattah, Abhinav Mehrotra, Łukasz Dudziak, and Nicholas D. Lane. Zero-cost proxies for lightweight NAS. In *International Conference on Learning Representations*, 2021. URL `https://arxiv.org/abs/2101.08134`.

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 29304–29320. Curran Associates, Inc., 2021.

Etor Arza, Frank Veenstra, Tønnes F. Nygaard, and Kyrre Glette. An empirical study on the computation budget of co-optimization of robot design and control in simulation, 2025. URL `https://arxiv.org/abs/2409.08621v2`. arXiv preprint, version 2.

Jagdeep Singh Bhatia, Holly Jackson, Yunsheng Tian, Jie Xu, and Wojciech Matusik. Evolution Gym: A large-scale benchmark for evolving soft robots. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 2201–2214. Curran Associates, Inc., 2021.

Lingjiao Chen, Jared Davis, Boris Hanin, Peter Bailis, Ion Stoica, Matei Zaharia, and James Zou. Are more LM calls all you need? Towards the scaling properties of compound AI systems. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 45767–45790. Curran Associates, Inc., 2024. doi: 10.52202/079017-1455.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders,

Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code, 2021. URL https://arxiv.org/abs/2107.03374. arXiv preprint arXiv:2107.03374.

Nick Cheney, Josh Bongard, Vytas SunSpiral, and Hod Lipson. Scalable co-optimization of morphology and control in embodied machines. *Journal of The Royal Society Interface*, 15 (143):20170937, June 2018. ISSN 1742-5662. doi: 10.1098/rsif.2017.0937. URL http://dx.doi.org/10.1098/rsif.2017.0937.

Valentin Dalibard and Max Jaderberg. Faster improvement rate population based training, 2021. URL https://arxiv.org/abs/2109.13800. arXiv preprint arXiv:2109.13800.

H. A. David and H. N. Nagaraja. *Order Statistics*. Wiley, July 2003. ISBN 9780471722168. doi: 10.1002/0471722162. URL http://dx.doi.org/10.1002/0471722162.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. Expected validation performance and estimation of a random variable's maximum. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 4066–4073, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.342. URL https://aclanthology.org/2021.findings-emnlp.342/.

Xuanyi Dong and Yi Yang. NAS-Bench-201: Extending the scope of reproducible neural architecture search. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/2001.00326.

Xuanyi Dong, Lu Liu, Katarzyna Musial, and Bogdan Gabrys. NATS-Bench: Benchmarking NAS algorithms for architecture topology and size. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3634–3646, 2022. doi: 10.1109/TPAMI.2021.3054824.

Florian Eddie Dorner, Yatong Chen, André F. Cruz, and Fanny Yang. ROC-n-reroll: How verifier imperfection affects test-time scaling. In *International Conference on Learning Representations*, volume 2026, pp. 150895–150936, 2026.

Romain Egele, Felix Mohr, Tom Viering, and Prasanna Balaprakash. The unreasonable effectiveness of early discarding after one epoch in neural network hyperparameter optimization. *Neurocomputing*, 597:127964, September 2024. ISSN 0925-2312. doi: 10.1016/j.neucom.2024.127964. URL http://dx.doi.org/10.1016/j.neucom.2024.127964. arXiv 2404.04111.

Stefan Falkner, Aaron Klein, and Frank Hutter. BOHB: Robust and efficient hyperparameter optimization at scale. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1437–1446. PMLR, 10–15 Jul 2018.

Jiawei Fang, Yuxuan Sun, Chengtian Ma, Qiuyu Lu, and Lining Yao. RoboMoRe: LLM-based robot co-design via joint optimization of morphology and reward, 2025. URL https://arxiv.org/abs/2506.00276. arXiv preprint arXiv:2506.00276.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10835–10866. PMLR, 23–29 Jul 2023.

Agrim Gupta, Silvio Savarese, Surya Ganguli, and Li Fei-Fei. Embodied intelligence via learning and evolution. *Nature Communications*, 12(1), October 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-25874-z. URL http://dx.doi.org/10.1038/s41467-021-25874-z.

Agrim Gupta, Linxi Fan, Surya Ganguli, and Li Fei-Fei. MetaMorph: Learning universal controllers with transformers. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=Opmqtk_GvYL.

Donald J. Hejna, III, Pieter Abbeel, and Lerrel Pinto. Task-agnostic morphology evolution. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=CGQ6ENUMX6.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack W. Rae, and Laurent Sifre. Training compute-optimal large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 30016–30030. Curran Associates, Inc., 2022. doi: 10.52202/068431-2176.

Sha Hu, Zeshi Yang, and Greg Mori. Neural fidelity warping for efficient robot morphology design. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7079–7086. IEEE, May 2021. doi: 10.1109/icra48506.2021.9561733. URL http://dx.doi.org/10.1109/icra48506.2021.9561733.

Audrey Huang, Adam Block, Qinghua Liu, Nan Jiang, Akshay Krishnamurthy, and Dylan J. Foster. Is Best-of-N the best of them? Coverage, scaling, and optimality in inference-time alignment. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 25075–25126. PMLR, 13–19 Jul 2025.

Max Jaderberg, Valentin Dalibard, Simon Osindero, Wojciech M. Czarnecki, Jeff Donahue, Ali Razavi, Oriol Vinyals, Tim Green, Iain Dunning, Karen Simonyan, Chrisantha Fernando, and Koray Kavukcuoglu. Population based training of neural networks, 2017. URL https://arxiv.org/abs/1711.09846. arXiv preprint arXiv:1711.09846.

Kevin Jamieson and Ameet Talwalkar. Non-stochastic best arm identification and hyperparameter optimization. In Arthur Gretton and Christian C. Robert (eds.), *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pp. 240–248, Cadiz, Spain, 09–11 May 2016. PMLR.

Hyeonseong Jeon, Ainaz Eftekhar, Aaron Walsman, Kuo-Hao Zeng, Ali Farhadi, and Ranjay Krishna. Convergent functions, divergent forms. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 46049–46076. Curran Associates, Inc., 2025. doi: 10.52202/085713-1534.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Hadi Khalaf, Claudio Mayrink Verdun, Alex Oesterling, Himabindu Lakkaraju, and Flavio du Pin Calmon. Inference-time reward hacking in large language models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 61720–61760. Curran Associates, Inc., 2025. doi: 10.52202/085713-2064.

Liam Li, Kevin Jamieson, Afshin Rostamizadeh, Ekaterina Gonina, Jonathan Ben-Tzur, Moritz Hardt, Benjamin Recht, and Ameet Talwalkar. A system for massively parallel hyperparameter tuning. In I. Dhillon, D. Papailiopoulos, and V. Sze (eds.), *Proceedings of Machine Learning and Systems*, volume 2, pp. 230–246, 2020.

Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, 18(185):1–52, 2018.

Haofei Lu, Zhe Wu, Junliang Xing, Jianshu Li, Ruoyu Li, Zhe Li, and Yuanchun Shi. BodyGen: Advancing towards efficient embodiment co-design. In *International Conference on Learning Representations*, volume 2025, pp. 41559–41580, 2025.

Kevin Sebastian Luck, Heni Ben Amor, and Roberto Calandra. Data-efficient co-adaptation of morphology and behaviour with deep reinforcement learning. In Leslie Pack Kaelbling, Danica Kragic, and Komei Sugiura (eds.), *Proceedings of the Conference on Robot Learning*, volume 100 of *Proceedings of Machine Learning Research*, pp. 854–869. PMLR, 2020.

Alican Mertan and Nick Cheney. Investigating premature convergence in co-optimization of morphology and control in evolved virtual soft robots. In *Genetic Programming: 27th European Conference, EuroGP 2024, Held as Part of EvoStar 2024*, volume 14631 of *Lecture Notes in Computer Science*, pp. 38–55. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-56957-9_3. URL https://doi.org/10.1007/978-3-031-56957-9_3.

Alican Mertan and Nick Cheney. Evolutionary brain-body co-optimization consistently fails to select for morphological potential. *Artificial Life*, pp. 1–20, 2026. doi: 10.1162/artl.a.476. URL http://dx.doi.org/10.1162/artl.a.476. arXiv:2508.17464.

H. N. Nagaraja. Some asymptotic results for the induced selection differential. *Journal of Applied Probability*, 19(2):253–261, June 1982. ISSN 1475-6072. doi: 10.2307/3213478. URL http://dx.doi.org/10.2307/3213478.

Kishan Reddy Nagiredla, Buddhika Laknath Semage, Arun Kumar Anjanapura Venkatesh, Thommen George Karimpanal, and Santu Rana. ECoDe: A sample-efficient method for co-design of robotic agents. In *AI 2024: Advances in Artificial Intelligence*, volume 15443 of *Lecture Notes in Computer Science*, pp. 3–15. Springer Nature Singapore, 2025. doi: 10.1007/978-981-96-0351-0_1.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. WebGPT: Browser-assisted question-answering with human feedback, 2021. URL https://arxiv.org/abs/2112.09332. arXiv preprint arXiv:2112.09332.

Xuefei Ning, Changcheng Tang, Wenshuo Li, Zixuan Zhou, Shuang Liang, Huazhong Yang, and Yu Wang. Evaluating efficient performance estimators of neural architectures. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 12265–12277. Curran Associates, Inc., 2021.

B. O'Neill. The distribution of order statistics under sampling without replacement. *Journal of Statistical Theory and Applications*, 24(3):663–698, 2025. doi: 10.1007/s44199-025-00125-y.

George R. Price. Selection and covariance. *Nature*, 227(5257):520–521, August 1970. ISSN 1476-4687. doi: 10.1038/227520a0. URL http://dx.doi.org/10.1038/227520a0.

Herilalaina Rakotoarison, Steven Adriaensen, Neeratyoy Mallik, Samir Garibov, Edward Bergman, and Frank Hutter. In-context freeze-thaw Bayesian optimization for hyperparameter optimization. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 41982–42008. PMLR, 21–27 Jul 2024.

Alan Robertson. A mathematical model of the culling process in dairy cattle. *Animal Production*, 8 (1):95–108, February 1966. doi: 10.1017/s0003356100037752. URL http://dx.doi.org/10.1017/s0003356100037752.

Rylan Schaeffer, Joshua Kazdan, John Hughes, Jordan Juravsky, Sara Price, Aengus Lynch, Erik Jones, Robert Kirk, Azalia Mirhoseini, and Sanmi Koyejo. How do large language monkeys get their power (Laws)? In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 53132–53176. PMLR, 13–19 Jul 2025.

Raphael Schneider, Daniel Honerkamp, Tim Welschehold, and Abhinav Valada. Task-driven co-design of mobile manipulators. *IEEE Robotics and Automation Letters*, 10(7):7158–7165, 2025. doi: 10.1109/LRA.2025.3573622.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations*, volume 2025, pp. 10131–10165, 2025.

Junru Song, Yang Yang, Yaqing Xu, Ying Wen, Wei Peng, Guozhen Li, Wei'en Zhou, and Wen Yao. Shaping the evolutionary dynamics of robot morphology via adaptive control learning, 2026. URL https://arxiv.org/abs/2608.23100v1. arXiv preprint arXiv:2608.23100.

Luke Strgar and Sam Kriegman. Accelerated co-design of robots through morphological pretraining. In *International Conference on Learning Representations*, volume 2026, pp. 7450–7470, 2026.

Kevin Swersky, Jasper Snoek, and Ryan Prescott Adams. Freeze-thaw Bayesian optimization, 2014. URL https://arxiv.org/abs/1406.3896. arXiv preprint arXiv:1406.3896.

Raphael Tang, Jaejun Lee, Ji Xin, Xinyu Liu, Yaoliang Yu, and Jimmy Lin. Showing your work doesn't always work. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2766–2772, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.246. URL https://aclanthology.org/2020.acl-main.246/.

Bruce Walsh and Michael Lynch. *Evolution and Selection of Quantitative Traits*. Oxford University Press, September 2018. ISBN 9780198830870. doi: 10.1093/oso/9780198830870.001.0001. URL http://dx.doi.org/10.1093/oso/9780198830870.001.0001.

Tsun-Hsuan Wang, Pingchuan Ma, Andrew Everett Spielberg, Zhou Xian, Hao Zhang, Joshua B. Tenenbaum, Daniela Rus, and Chuang Gan. SoftZoo: A soft robot co-design benchmark for locomotion in diverse environments. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Xyme9p1rpZw.

Colin White, Arber Zela, Robin Ru, Yang Liu, and Frank Hutter. How powerful are performance predictors in neural architecture search? In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 28454–28469. Curran Associates, Inc., 2021.

Martin Wistuba, Arlind Kadra, and Josif Grabocka. Supervising the multi-fidelity race of hyperparameter configurations. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 13470–13484. Curran Associates, Inc., 2022. doi: 10.52202/068431-0979.

Yuhuai Wu, Mengye Ren, Renjie Liao, and Roger B. Grosse. Understanding short-horizon bias in stochastic meta-optimization. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1803.02021.

Cheng Yan, Felix Mohr, and Tom Viering. LCDB 1.1: A database illustrating learning curves are more ill-behaved than previously thought. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, volume 38. Curran Associates, Inc., 2025. doi: 10.52202/085713-0487.

W. B. Yeo and H. A. David. Selection through an associated characteristic, with applications to the random effects model. *Journal of the American Statistical Association*, 79(386):399–405, June 1984. ISSN 1537-274X. doi: 10.1080/01621459.1984.10478063. URL http://dx.doi.org/10.1080/01621459.1984.10478063.

Chris Ying, Aaron Klein, Esteban Real, Eric Christiansen, Kevin Murphy, and Frank Hutter. NAS-Bench-101: Towards reproducible neural architecture search. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7105–7114. PMLR, 09–15 Jun 2019.

Kaicheng Yu, Christian Sciuto, Martin Jaggi, Claudiu Musat, and Mathieu Salzmann. Evaluating the search phase of neural architecture search. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=H1loF2NFwr.

Ye Yuan, Yuda Song, Zhengyi Luo, Wen Sun, and Kris M. Kitani. Transform2Act: Learning a transform-and-control policy for efficient agent design. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=UcDUxjPYWSr.

Allan Zhao, Jie Xu, Mina Konaković-Luković, Josephine Hughes, Andrew Spielberg, Daniela Rus, and Wojciech Matusik. RoboGrammar: Graph grammar for terrain-optimized robot design. *ACM Transactions on Graphics*, 39(6):188:1–188:16, December 2020. doi: 10.1145/3414685.3417831. URL https://doi.org/10.1145/3414685.3417831.

Appendix contents

A Estimator and Sampling Model . . . 16
B Training and Design Spaces . . . 16
C Results by Design Space . . . 17
D Preregistered Tests . . . 23
E Task Checks . . . 24
F Walker Details . . . 25
G The EvoGym Archive . . . 28
H Body Galleries . . . 29
I Further Related Work . . . 31

A Estimator and Sampling Model

Ties and scale. Order a pool by increasing screen score, with rank $r = 1$ the worst and $r = D$ the best. The mean $\bar{y}$ and sample SD s_y use the original per-body final scores, before averaging within blocks of tied screen scores.

Exactness. At $N = 1$ every w_r is $1/D$ and the lift is zero. We compute the binomial ratio in log space. Given a pool's scores the lift has no Monte Carlo error, and unit tests check it against brute-force enumeration of all subsets at small D, with ties and reversed rankings. When the screen score is the final score, the lift is the standardized unbiased expected-maximum estimator (Tang et al., 2020; Dodge et al., 2021).

Normal limit. Under the bivariate normal model of Sec. 2.2, a positive correlation gives a positive lift at every $N \geq 2$, and a small N only makes it small.

Sampling model. Bodies are drawn independently from a design space's generator, except that a grammar pool rejects duplicates and bodies of the earlier grammar pool. Each controller run is an independent draw of the training procedure. The evaluation of Eq. 1 is exact given each pool. Each controller run uses a random-number stream determined by both its seed and its body. Bodies with the same seed index therefore share neither an initial policy nor noise, so controller runs are nested within bodies. The target is the lift for a new pool sampled from the same body generator and trained with the same procedure. We compute bootstrap percentile intervals by resampling bodies together with their controller runs, approximating a new draw of both bodies and runs. Our anonymous repository gives the number of replicates with each result. A pointwise interval covers one setting, and a simultaneous band takes the largest standardized deviation over a stated family of settings, such as the twelve budgets at $N = 4$ or every budget and N of a pool.

Reported values of N. In the retrained pools, N runs over 2, 3, 4, 6, 8, 12, 16, 24, 32, 48, 64, 96 and 128, up to the limit on the effective number of bodies in Sec. 2.3.

Body generators. In the walker, hopper and cheetah spaces each scale factor is drawn independently and uniformly from $[0.5, 2]$, one for each limb radius, one for each actuator gear and one for density, with limb lengths fixed. The cheetah's model sets its total mass directly, so its density factor scales that mass, from 7 to 28 kg. In the grammar space each nonterminal expands uniformly among its alternatives. The body grows by a fair coin to at most 8 segments, each segment carries a mirrored pair of legs with probability $1/2$, and a leg has 1, 2 or 3 links with probabilities $1/2$, $1/4$ and $1/4$. A grammar body is kept if it has at most 20 geometries counting the floor, at least one actuator, no self-contact at rest and no duplicate in the pool. Each pool is a fixed random permutation of bodies drawn from one seed, and each new pool shares no body with the earlier pool of its space (Table 1).

B Training and Design Spaces

Table 2 lists the PPO settings and what each design space varies. We also analyze a public archive of EvoGym voxel bodies whose controllers were evolved rather than trained with PPO (App. G). The walker test pool, with its two seed sets, and the new cheetah and grammar pools were trained from scratch after their tests were preregistered. Training the walker pilot, cheetah and grammar pools took about 846 GPU-hours, including walker bodies that were trained more than once. The walker test's first seed set took about 170 GPU-hours and the new grammar pool about 598 GPU-hours over 8,192 runs. These sums exclude scheduler overhead and the project's other runs.

Table 1: Bodies, seeds and budgets of each analysis. Unless noted, every body has four PPO seeds. A screen averages two of them at the screen budget and the final score averages the other two at H, so the kept body is retrained from scratch. Results average the six such splits. †Preregistered.

Analysis	Bodies	Screen	Final score	Budgets
Walker pilot pool	996	2 of 4 seeds	other 2 seeds	12, $H/64$ to H
Walker test, first seed set†	1,024, none in the pilot	2 of 4 seeds	other 2 seeds	12, $H/64$ to H
Walker test, second seed set	the same 1,024	2 of 4 further seeds	other 2 seeds	12, $H/64$ to H
Twice the final training†	the same 1,024	recorded screens, 2 seeds of each seed set, at $H/4$ (and $H/2$)[a]	2 new seeds at $2H$ (and H, $3H/2$)	14, $H/64$ to $2H$
Gymnasium rule, first 600 bodies	595 test bodies[b]	2 of 4 new seeds	other 2 seeds	12, $H/64$ to H
Gymnasium rule, test†	the other 424 test bodies[b]	2 of 4 new seeds	other 2 seeds	12, $H/64$ to H
Hopper and the original cheetah and grammar pools	256, 512 and 512	2 of 4 seeds	other 2 seeds	12, $H/64$ to H
Cheetah, new	2,048, none in the original	2 of 4 seeds, and a permutation test† at $H/8$	other 2 seeds	12, $H/64$ to H
Grammar, new	2,048, none in the original	2 of 4 seeds	other 2 seeds	12, $H/64$ to H
EvoGym archive	1,305,840	the body's one evolution run at generation T	the same run at generation 299, controller continued	$T = 1$, 5, 10, 25, 50, 100, 200, and every T in App. G
Architecture search†	423,624 in NAS-Bench-101 and 8,169 per NATS-Bench dataset	1 of 3 runs after short or partial training[c]	final test accuracy of another run	7 and 9 screens[c]

[a]Two fixed splits, one per seed set. [b]The first 600 test bodies less five whose partial records had been read, and the other 424, with the same four new seeds. [c]Six ordered pairs of three runs. NAS-Bench-101 screens at 2, 4, 6, 12, 18 and 36 epochs from separate 4-, 12- and 36-epoch runs and at 54 epochs from a 108-epoch run, NATS-Bench at 1, 2, 3, 6, 12, 25, 50, 100 and 150 of 200 epochs.

C Results by Design Space

The lift at four candidates. Table 4 gives the lift of best-of-four in every pool at two screen budgets, and Figure 12 gives it at all twelve, with each pool's simultaneous band. Across all 913 settings with screens from $H/64$ to $3H/4$ in the seven retrained pools, with both walker seed sets, 630 have a pointwise significant positive lift and 61 a pointwise significant negative one (Table 5). All 61 settings with a pointwise significant negative lift are in the walker test pool's two seed sets and the pilot pool, at $N \leq 16$. Under per-pool simultaneous bands the counts are 589 and 12. Settings share bodies and seeds, so these counts are not independent. In the five hopper, cheetah and grammar pools, 528 of the 561 settings with screens from $H/64$ to $3H/4$ have a significant positive lift under per-pool simultaneous bands, and none has a significant negative one.

More candidates. Figure 9 gives the lift against the number of candidates in every pool. The bodies in Fig. 8 come from the 2,048 new grammar bodies and one split of the seeds, two for the screen and two for the final score, while its curve averages the six seed splits. Figure 10 repeats Fig. 1b with the EvoGym archive added.

Screen and final budgets. No screen at least half as long as the final training and no longer than it had a negative estimated lift. The same held for screens of at least $H/2$ that were no longer than the final training. The first condition covers all 355 recorded pairs of screen and final budgets at $N = 4$ with the screen between half and all of the final training, in every retrained pool, the

Table 2: Controller training and design spaces. PPO settings apply to every pool we trained. Archive controllers were evolved, not trained with PPO. Scale factors multiply the base body's value.

Setting	Value
Optimization (PPO)	
Optimizer	Adam
Learning rate / schedule	`3e-4` / constant
Adam epsilon	`1e-5`
Gradient-norm clip	0.5
Discount / GAE lambda	0.99 / 0.95
PPO clip epsilon	0.2
Value / entropy coefficient	0.5 / 0
Epochs per update / minibatch size	4 / 512
Advantage standardization	Per minibatch
Network	
Policy / value networks	Separate MLPs
Hidden layers / activation	2×64 / tanh
Policy distribution	Gaussian
Log standard deviation	Learned, state-independent
Environment and evaluation	
Observations	Raw joint positions and velocities[a]
Observation normalization / clipping	None / none
Parallel environments / rollout steps	2,048 / 64
Environment steps per update	131,072
Full budget H	8,388,608 steps (64 updates)
Score checkpoints (update numbers)	1, 2, 3, 4, 6, 8, 12, 16, 24, 32, 48, 64
Action repeat	4
Control step: walker, hopper, grammar	0.008 s
Control step: cheetah / EvoGym	0.04 s / n/a
Design spaces	
Walker pools: pilot / test	996 / 1,024
Walker scale factors[b]	14: 7 limb radii, 6 gears, density
Hopper pool	256
Hopper scale factors[b]	8: 4 limb radii, 3 gears, density
Cheetah pools: original / new	512 / 2,048
Cheetah scale factors[b]	15: 8 limb radii, 6 gears, total mass
Grammar pools: original / new	512 / 2,048
Grammar generator	Derivations of the 20-rule RoboGrammar grammar
Grammar limits	8 segments, 20 geometries (maxima)[c]
EvoGym archive pool	1,305,840
EvoGym grid	3×3, five cell states[d]
EvoGym inclusion	Connected, ≥ 3 voxels, ≥ 3 actuators
EvoGym controller evolution	
Method	Age-fitness Pareto optimization
Population / generations	20 / 300
Mutation standard deviation	0.1
MLP hidden layers / width	1 / 32
Controller query interval	5 simulator steps
Episode length	500 steps

[a]Horizontal position is excluded. [b]Every scale factor is in $[0.5, 2]$. [c]The geometry count includes the floor. [d]Each cell is empty, rigid, soft, a horizontal actuator, or a vertical actuator.

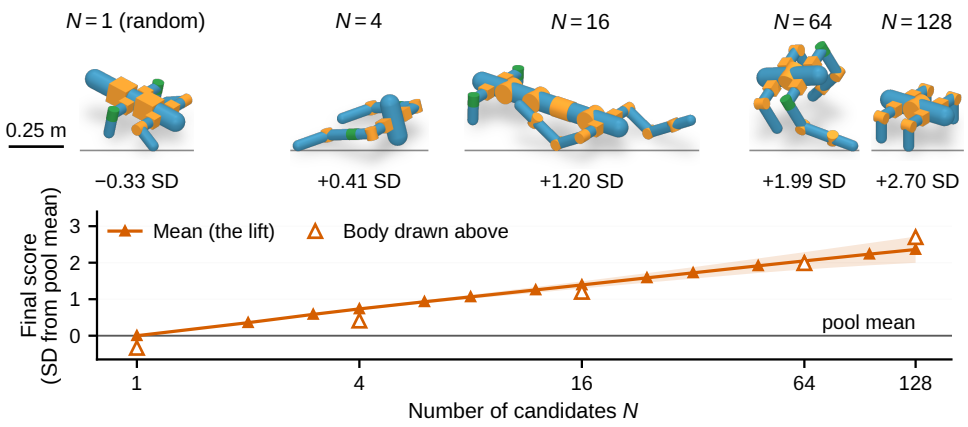


Figure 8: What a larger search selects in the new grammar pool, using a quarter-budget screen. Each drawn body has the median final score among the bodies selected over all draws of N candidates for one seed split, and labels give final scores in SD from the pool mean. The curve shows the selection lift averaged over the six seed splits, with its 95% interval.

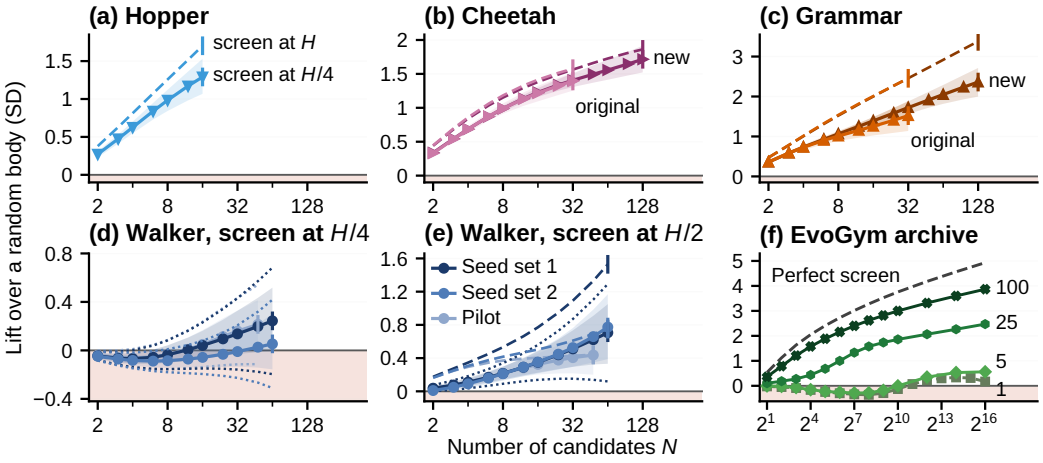


Figure 9: Lift of best-of-N over a random body against the number of candidates N, with pointwise 95% intervals shaded (not shown for the archive in (f)), a short bar at the largest N a pool supports, and pink shading below zero where a search does worse than a random body. (a–c) Hopper and the original and new cheetah and grammar pools, with the screen at a quarter of the full budget and at the full budget (dashed). (d, e) The walker's two test seed sets and pilot pool, with the screen at a quarter of the full budget in (d) and at half of it in (e), where the full-budget screen is dashed. Simultaneous 95% bands over every screen budget below H and every N are dotted, in (e) for the first seed set only. (f) The EvoGym archive, with the screened controller continued, for a perfect screen and for screens at the generation printed at each line's end. **Takeaway:** In hopper, cheetah and grammar more candidates add to the gain, and in the walker the harm of the quarter-budget screen first deepens, then fades as more candidates are added.

follow-up and both Gymnasium-rule analyses. The result also holds at every reported N at the full final budget. Among all recorded settings with the screen no longer than the final training, every significantly harmful one is a walker screen of at most $3H/8$, under each of our three sets of bootstrap replicates. All of them have a final budget of at least H except two narrowly significant ones under Gymnasium's termination rule at a final budget of $3H/4$, with screens of $3H/16$ on the 595 bodies and $H/16$ on the 424 bodies. In the archive and the architecture-search benchmarks, no estimate with a screen of at least half the final training is negative. This statement excludes screens longer than the final training. Of the 726 recorded pairs with such screens, 55 have a negative

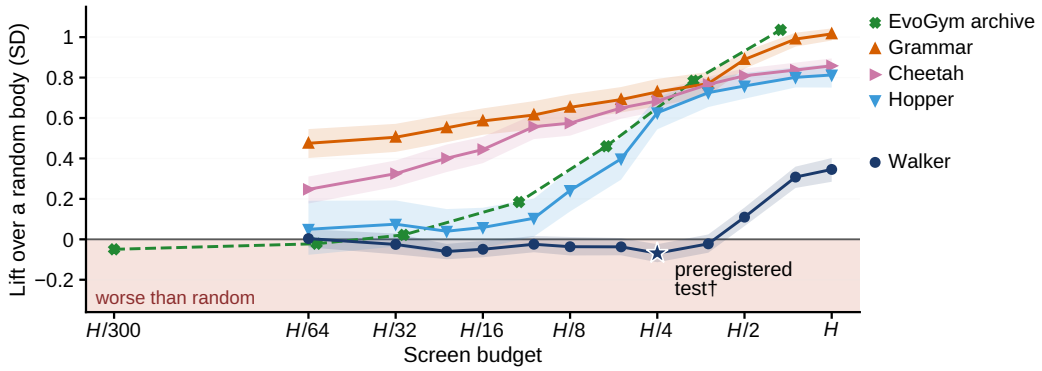


Figure 10: Lift of best-of-four over a random body against the screen budget in hopper, the original cheetah and grammar pools and the walker test pool's first seed set, with pointwise 95% intervals, and for the EvoGym archive (dashed), whose screen generation is plotted as a fraction of its 300-generation run from generation 1 at $H/300$ and whose screened controller is continued. The star marks the walker's preregistered test. **Takeaway:** Every screen beats random in cheetah and grammar, screens from $H/8$ do in hopper and from $H/2$ in the walker, and archive screens do once controllers have evolved for 10 generations.

estimate and 23 are significantly harmful under both sets of bootstrap replicates that cover every pair. All 23 are walker pairs that screen at H against a final budget of $H/64$ to $3H/8$.

Final budget. Table 3 varies the final budget while the screen stays at $H/4$. In both walker seed sets and the pilot pool the lift falls as the final training runs longer, while the other spaces stay positive at every final budget.

Table 3: Lift of best-of-four, averaged over the six seed splits, with the screen at $H/4$ and the final score taken at later budgets. All values are post hoc except the preregistered one marked †. **Takeaway:** In the walker the sign of the lift depends on the final budget.

Pool	$3H/8$	$H/2$	H
Walker test, first seed set	$+0.22$	$+0.13$	$-0.07^{\dagger}$
Walker test, second seed set	$+0.15$	$+0.07$	-0.08
Walker pilot pool	$+0.25$	$+0.10$	-0.06
Hopper	$+0.69$	$+0.67$	$+0.63$
Cheetah, original	$+0.75$	$+0.74$	$+0.68$
Grammar, original	$+0.82$	$+0.77$	$+0.73$

Equal compute. Figure 11 places every screen on our grid at its total compute. In all five hopper, cheetah and grammar pools the quarter-budget screen of four beats best-of-two at the full budget, by 0.27 to 0.29 SD (Fig. 11d). Among screens shorter than H at a screening cost of $2H$, only the one-seed screen of four at $H/2$ has a significantly positive lift in every pool of Fig. 11 (post hoc).

Continuing the screened controllers. Continuing the screened controllers trains the kept body's k_s controllers from T to H rather than training new controllers from scratch. The screen score and final score then come from the same runs, and the total cost is $k_s NT + k_s(H - T)$. Table 6 compares continuing the screened controllers with retraining for the quarter-budget screen of four. For continuing, each of the six pairs drawn from the four seeds supplies both the screen score and the final score.

Architecture-search benchmarks (preregistered). To test whether screening harm is a generic property of low-fidelity selection, we applied the same estimator to two architecture-search benchmarks. NAS-Bench-101 (Ying et al., 2019) has 423,624 architectures with three training runs each. For the NATS-Bench topology space (Dong et al., 2022; Dong & Yang, 2020) on CIFAR-10, CIFAR-100 and ImageNet16-120, we use the 8,169 of its 15,625 architectures that have all three training runs recorded. For NATS-Bench the screen is accuracy after 1 to 150 of one run's 200 epochs, on

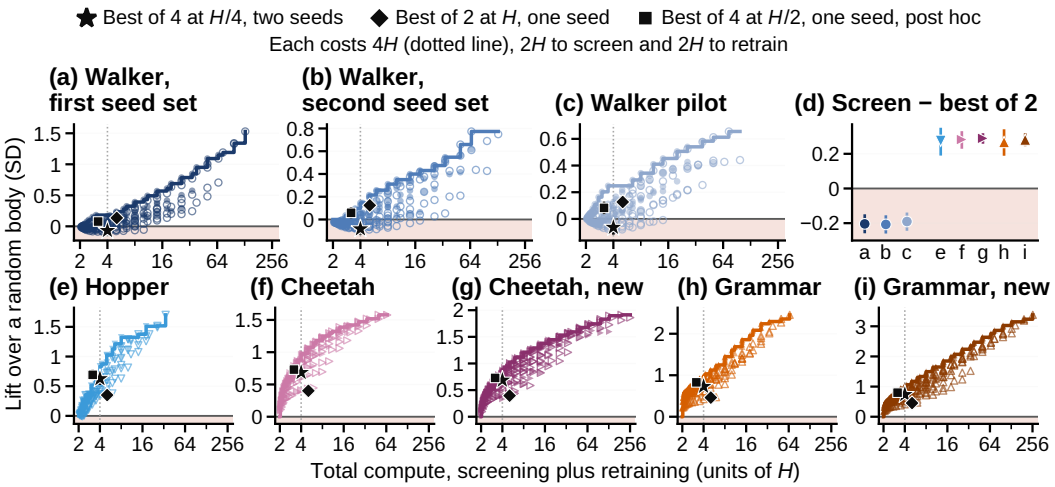


Figure 11: Lift over a random body against the total compute of screening and retraining the kept body with two seeds, in units of the full budget H, in every retrained pool. Small markers show every screen on our grid, filled for one screen seed and open for two. The black markers each cost $4H$ and carry pointwise 95% intervals, and two are offset from the dotted $4H$ line for legibility. The line is the best lift at or below each level of compute, chosen post hoc. (d) The lift of the quarter-budget screen of four minus that of best-of-two at H in each pool, with its paired 95% interval, labeled by the pool's panel letter. **Takeaway:** Outside the walker, the quarter-budget screen of four beats training two bodies fully and retraining the better one, at the same compute.

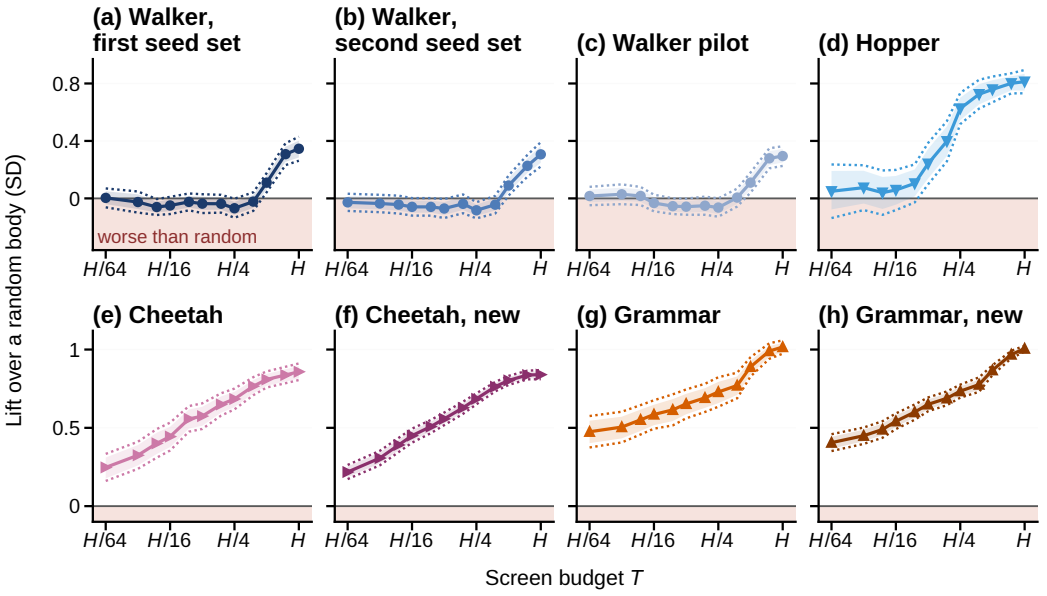


Figure 12: Lift of best-of-four over a random body against the screen budget in each retrained pool, with the kept body retrained and scored at H. Pointwise 95% intervals are shaded, and the pool's simultaneous 95% band over the twelve budgets is dotted (post hoc). **Takeaway:** The simultaneous band lies above zero at every budget in both cheetah and both grammar pools and from $H/8$ in hopper, while in the walker it lies below zero at $H/4$ and above zero from $H/2$.

the validation split for CIFAR-10. For NAS-Bench-101 it is the halfway or final validation accuracy of separate 4-, 12- and 36-epoch runs, each with its own schedule, or the halfway accuracy of a 108-epoch run, which gives screens at 2 to 54 epochs. Each final score is the final test accuracy of a different run from the screen, and we average the selection lift over the six ordered pairs of runs.

Table 4: Lift of best-of-four at two of the twelve screen budgets, with the kept body retrained from scratch, in SD of the final score with its 95% interval. All values are post hoc except the preregistered one marked †. **Takeaway:** Outside the walker the lift is positive at every screen budget.

Pool	Bodies	Screen $H/4$	Screen $H/2$
Walker test, first seed set	1,024	$-0.07^{\dagger}$ $[-0.11, -0.02]$	$+0.11$ $[+0.07, +0.16]$
Walker test, second seed set	1,024	-0.08 $[-0.12, -0.04]$	$+0.09$ $[+0.04, +0.14]$
Walker pilot pool	996	-0.06 $[-0.11, -0.02]$	$+0.11$ $[+0.06, +0.16]$
Hopper	256	$+0.63$ $[+0.54, +0.70]$	$+0.76$ $[+0.70, +0.82]$
Cheetah, original	512	$+0.68$ $[+0.63, +0.73]$	$+0.81$ $[+0.77, +0.84]$
Cheetah, new	2,048	$+0.69$ $[+0.66, +0.71]$	$+0.80$ $[+0.78, +0.82]$
Grammar, original	512	$+0.73$ $[+0.66, +0.79]$	$+0.89$ $[+0.84, +0.93]$
Grammar, new	2,048	$+0.73$ $[+0.70, +0.76]$	$+0.87$ $[+0.84, +0.89]$

Table 5: Settings of each retrained pool whose pointwise 95% interval lies above or below zero (post hoc). A setting is one screen budget from $H/64$ to $3H/4$ and one reported N. **Takeaway:** Significant harm appears only in the walker.

Pool	Bodies	Settings	Above zero	Below zero
Walker test, first seed set	1,024	121	37	14
Walker test, second seed set	1,024	121	25	29
Walker pilot pool	996	110	30	18
Hopper	256	77	54	0
Cheetah, original	512	99	99	0
Cheetah, new	2,048	143	143	0
Grammar, original	512	99	99	0
Grammar, new	2,048	143	143	0

Table 6: Paired difference in the lift of best-of-four with a two-seed screen at $H/4$ between continuing the kept body's two screened controllers to H, at $3.5H$ in total, and retraining the kept body, at $4H$, with 95% intervals (post hoc). **Takeaway:** Continuing the screened controllers beats retraining in every pool at a lower cost.

Pool	Bodies	Continued minus retrained
Walker test, first seed set	1,024	$+0.12$ $[+0.09, +0.15]$
Walker test, second seed set	1,024	$+0.11$ $[+0.07, +0.14]$
Walker pilot pool	996	$+0.13$ $[+0.10, +0.17]$
Hopper	256	$+0.10$ $[+0.07, +0.14]$
Cheetah, original	512	$+0.10$ $[+0.08, +0.12]$
Cheetah, new	2,048	$+0.11$ $[+0.10, +0.12]$
Grammar, original	512	$+0.03$ $[+0.01, +0.04]$
Grammar, new	2,048	$+0.03$ $[+0.02, +0.04]$

The analysis plan, prediction and scoring rule were fixed before any lift was computed. In the three pools whose screens follow the plan (NAS-Bench-101 and both NATS-Bench CIFAR-10 pools), all 394 settings have a positive lift whose interval lies above zero, both pointwise and under per-pool simultaneous bands (Table 7). For CIFAR-100 and ImageNet16-120 the benchmark records per-epoch accuracy only on the test split, so there the screen reads test images, which the final score also uses. In these 522 settings every lift is positive and 521 intervals lie above zero.

Table 7: Settings of the preregistered architecture-search analysis, whose 95% interval lies above or below zero. A setting pairs one screen, accuracy after a short training run or part of a full one, with one reported N. The 394 settings in NAS-Bench-101 and the two NATS-Bench CIFAR-10 rows follow the screening plan. The other 522, on CIFAR-100 and ImageNet16-120, use test-split screens instead, because there the benchmark records per-epoch accuracy only on the test split. Every point estimate is positive. Connected cells drop architectures with no path from input to output. **Takeaway:** Screening on short training picks better architectures than random at every setting.

Benchmark and dataset	Architectures	Settings	Above zero	Below zero
NAS-Bench-101, CIFAR-10	423,624	133	133	0
NATS-Bench, CIFAR-10	8,169	135	135	0
connected cells	7,828	126	126	0
NATS-Bench, CIFAR-100	8,169	135	135	0
connected cells	7,828	126	125	0
NATS-Bench, ImageNet16-120	8,169	135	135	0
connected cells	7,828	126	126	0
Total		916	915	0

D PREREGISTERED TESTS

Table 8 lists the preregistered tests we report, with each rule's decision limit and outcome. Every amendment described below was made before the test's outcome was seen.

Table 8: The preregistered tests we report. One-sided limits are 95% bootstrap limits, and each outcome is the label its rule defines.

Test	Bodies	Endpoint	Result	Outcome
Walker test, harm	1,024 new	lift of best-of-four at $H/4$, one-sided 95% upper limit below zero and point estimate at most -0.03 SD	-0.07, upper limit -0.03	supported
Twice the final training	the same, new final seeds	lift at $H/4$ below zero against new final controllers at $2H$	-0.15, upper limit -0.11	supported
Gymnasium termination	424 held out	lift at $H/4$ below zero	-0.12, upper limit -0.05	supported
Cheetah permutation test	2,048 new	positive lift at $H/8$, $N = 4$	$p = 1/1001$	supported
Architecture search	seven benchmark pools	no negative lift	none negative in the 394 settings that follow the plan or the 522 others	as predicted

The walker test. The test point came from the pilot pool, where the lift at $N = 4$ was negative under every seed split at three of the eleven budgets below H, the latest being $H/4$. No new body repeats a pilot body. For the harm at $H/4$ (Table 8), the test required a one-sided 95% upper limit below zero and a point estimate of at most -0.03 SD. Amendments fixed the jobs' time limits and the interval construction. The second seed set was trained after the first was scored, under a condition fixed in advance that did not depend on the outcome, and was scored with the unchanged scoring code.

Twice the final training. Preregistered before any of its controllers was trained, this test reuses the test bodies and their recorded two-seed screens at $H/4$ and trains new final controllers from two further seeds to $2H$ (128 updates). The primary endpoint is the mean lift over two fixed screens, one per seed set, evaluated against final scores from the new controllers trained to $2H$. The endpoint is supported if its one-sided 95% upper limit is below zero. The result is -0.15, with one-sided 95% lower and upper limits of -0.20 and -0.11. Against the new seeds at H the lift is -0.06 (upper limit -0.01) and at $3H/2$ it is -0.11 (upper limit -0.07). The paired difference between $2H$ and

H is -0.10 (upper limit -0.05), which the preregistration requires before calling the harm deeper. With the screen fixed at $H/2$, the lift against the new final controllers is $+0.09$ at H (one-sided lower limit $+0.04$) and $+0.05$ at $3H/2$ (lower limit just above zero).

The cheetah test. We preregistered this test before training any body. Amendments made before any outcome was seen fixed it as a paired, studentized permutation test of the lift's sign for best-of-four at $H/8$. Each permutation randomly reassigns the bodies' final scores, each with its seed variance, to the screen's ranks. Its primary endpoint gave a studentized statistic of 34.48 and one-sided $p = 1/1001$, the floor for 1,000 permutations. The lift itself is $+0.58$ SD $[+0.54, +0.62]$ (post hoc).

E TASK CHECKS

We compared each task and its scoring code with the task it was ported from, replaying bodies with no action and with random actions. Table 9 gives each task's rule for ending an episode and how often episodes end early, and Table 10 shows how noisy each pool's final scores are. The walker's final scores are noisier than those of cheetah and grammar. The reliability of a two-seed final score is 0.34 in the walker test pool's first seed set and 0.23 in the second, compared with 0.78 in the original cheetah pool and 0.90 in the original grammar pool (Table 10). Noisier final scores shrink a lift toward zero.

Table 9: The rule that ends an episode in each task, and the share of episodes that end before the time limit under no action and under random actions. *No action* holds every actuator at zero, and under *random actions*, each action is drawn uniformly at every control step.

Space (bodies)	Episode ends when	No action	Random actions
Walker test (1,024)	torso below 0.25 m or above 3.75 m, with no angle limit	99.0%	95.2%
Hopper (256)	as the walker	98.4%	100%
Grammar, new (2,048)	root below 0.07 m (a fixed fraction of spawn height) or above 2 m	46.7%	48.3%
Grammar, original (512)	as the new pool	47.7%	51.4%
Cheetah, original (512)	effectively never	0%	0%
Cheetah, new (2,048)	as the original pool	0%	0%
EvoGym archive (240)	simulation instability, goal or 500 steps	0%	0.4%

Table 10: Ratio of between-seed variance σ_e^2 to between-body variance σ_b^2 in full-budget returns, from a one-way decomposition. The reliability of a two-seed mean, $\sigma_b^2/(\sigma_b^2 + \sigma_e^2/2)$, is the share of its variance that comes from differences between bodies. The last column gives the number of runs a body's mean needs for a reliability of 0.8. **Takeaway:** Walker final scores are noisy, and cheetah and grammar final scores are reliable.

Pool	Runs per body	σ_e^2/σ_b^2	Reliability, two runs	Runs for 0.8
Walker test, first seed set	4	3.88	0.34	16
Walker test, second seed set	4	6.57	0.23	27
Walker pilot pool	4	7.80	0.20	32
Cheetah, original	4	0.56	0.78	3
Grammar, original	4	0.22	0.90	1

Grammar bodies that fall with no action. Every grammar body spawns with its root at 0.35 m, and an episode ends when the root drops below a fifth of that height, a limit set by the spawn height and not by the body's own standing height. With no action, 937 of the 2,048 new bodies (46%) and 240 of the 512 original bodies (47%) fall below this limit within the first 250 control steps (2 s). In the new pool, these bodies have a mean return after final training of 13.5, compared with 195.9 for bodies that do not fall below the limit within that time. Among the 1,111 new and 272 original bodies that do not fall below the limit within those 250 control steps, best-of-four with the screen at $H/4$ gains $+0.47$ $[+0.42, +0.52]$ and $+0.39$ $[+0.27, +0.50]$, respectively, in SD of each subset's final scores (post hoc). The corresponding gain is $+0.73$ in both full pools.

F WALKER DETAILS

Task and scale. The walker's reward is the locomotion reward of Gupta et al. (2021), and Table 11 gives its termination rule and compares the walker with Gymnasium's Walker2d.

Table 11: Gymnasium's Walker2d (v5, Gymnasium 1.3.0), our walker, and our walker under Gymnasium's termination rule. x, z and θ are the torso's horizontal position, height and pitch, and a is the action. The nominal torso height before reset noise is 1.25 m.

	Gymnasium Walker2d-v5	Our walker	Ours, Gymnasium termination
Reward per step	$\dot{x} - 10^{-3}\|a\|^2$, plus 1 on every healthy step	$\dot{x} - 10^{-3}\|a\|^2$, a clipped to $[-1, 1]$, no alive bonus	as ours
Episode ends	$z \notin (0.8, 2)$ m or $\theta \notin (-1, 1)$ rad	$z < 0.25$ or $z > 3.75$ m (0.2 and 3 times spawn height), or a non-finite state, with no angle limit	$z \notin [0.64, 1.6]$ times spawn height, that is $[0.8, 2]$ m, $\|\theta\| > 1$ rad, or a non-finite state
Control step, length	$4 \times 0.002 = 0.008$ s, 1,000 steps (8 s)	the same	the same
Observation	joint positions without x, velocities clipped to ± 10 (17 values)	joint positions without x, velocities unclipped (17 values)	as ours
Start state	uniform noise of ± 0.005 on positions and velocities	fixed, no noise	as ours
Simulator	MuJoCo, RK4	MuJoCo MJX, semi-implicit Euler	as ours
Body	fixed, both feet with friction 1.9	7 limb radii, 6 gears and density, each scaled by an independent uniform factor in $[0.5, 2]$, lengths fixed, right foot friction 0.9	as ours

Each recorded value is one episode of the policy's deterministic mean action from an unperturbed start. The mean return is still rising over the last three budgets (209, 341 and 509 in the first seed set). Return divided by 125 is the distance the torso travels minus an action cost of at most 0.048 m. In return units the harm at $H/4$ is -23.6 and -27.2 in the two seed sets, and Fig. 13 compares the five walker analyses.

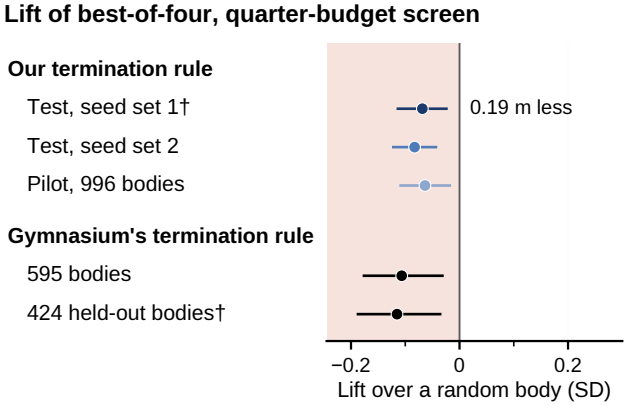


Figure 13: Lift of best-of-four with the quarter-budget screen in five walker analyses, with pointwise 95% intervals: the test pool's two seed sets and the pilot pool under our termination rule, and two groups of test bodies under Gymnasium's termination rule. The label converts the first test's lift to meters. †Preregistered.

The reversal between $H/4$ and $H/2$. The paired change in lift from $H/4$ to $H/2$ is $+0.18$ $[+0.13, +0.23]$ and $+0.17$ $[+0.12, +0.23]$ in the two seed sets, and $+0.17$ $[+0.12, +0.22]$ in the pilot. The six seed-split estimates at $H/4$ span $[-0.12, -0.03]$ and $[-0.14, -0.03]$, and at $H/2$ $[+0.07, +0.14]$ and $[+0.03, +0.12]$. A band over all twelve budgets, with critical values 2.84 and 2.80, lies below zero at $H/4$ (its upper edge is just below zero in the first seed set and -0.03 in the second) and above zero at $H/2$ (lower edges $+0.04$ and $+0.02$). Pointwise, the harm is also significant at two other short screens in the first seed set and three in the second (Fig. 15a).

Robustness. In 2,000 random halves of the test bodies, 512 bodies each, the lift at $H/4$ is negative in 99.9% and 100% of halves for the two seed sets (Fig. 14).

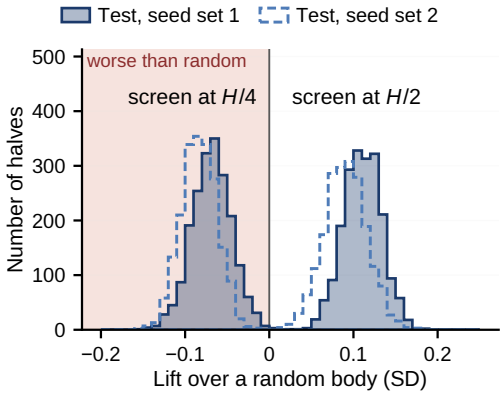


Figure 14: Lift of best-of-four with the screen at $H/4$ and at $H/2$ in 2,000 random halves of the 1,024 test bodies, with the first seed set filled and the second as a dashed outline (post hoc). **Takeaway:** The harm is stable across random halves of the test bodies.

All 70 four-seed subsets of the eight seeds give a negative lift (-0.10 to -0.03). Resampling bodies and the eight seeds independently gives -0.07 $[-0.11, -0.02]$ at $H/4$ and $+0.10$ $[+0.05, +0.16]$ at $H/2$. Screening on one seed and scoring on the other three gives -0.07 $[-0.11, -0.04]$ and -0.07 $[-0.10, -0.03]$ at $H/4$, and the benefit at $H/2$ remains. Restricted to the 956 and 951 bodies whose four final episodes all end in a fall, the lift at $H/4$ is -0.09 $[-0.14, -0.05]$ and -0.09 $[-0.13, -0.05]$.

Mass and gear. Body mass ranges from 9 to 138 kg. In the test pool, the lowest and highest tenths by strength for weight have median masses of about 100 and 22 kg. In the first seed set, the top tenth ranked by the quarter-budget screen has a median mass of 69 kg and a median strength for weight of 10.3 N·m/kg, against 41 kg and 19.0 N·m/kg for the top tenth ranked by final training, and Fig. 16 shows every body in these top tenths. Figure 1c draws one walker from each of these top tenths, and Fig. 17 shows the reversal across tenths of strength for weight, with the hopper for comparison. On four-seed means, the rank correlation between strength for weight and return is most negative at the quarter-budget screen (-0.51 in both seed sets), crosses zero between $H/2$ and $3H/4$, and reaches $+0.49$ and $+0.53$ at H (Fig. 15a). On the pilot pool it is -0.56 at $H/4$ and $+0.50$ at H. The return at the full budget rises with strength for weight up to the ninth decile, while the quarter-budget screen's return falls from 163 in the lowest decile to 83 in the highest in the first seed set (Fig. 15b, c). Summed gear's correlation with return changes about a third as much as mass's, so the reversal is carried mostly by mass. Measured on ranks, the lift of best-of-four at $H/4$ is -0.08 and -0.12 SD of the final rank in the two seed sets, close to the screen's rank correlations with the final ranking in Sec. 5.2, as expected for ranks. Best-of-four drawn within quintiles of strength for weight, a different selection rule, helps in both seed sets ($+0.05$ $[+0.01, +0.09]$ and $+0.06$ $[+0.02, +0.10]$, in SD of each quintile's final scores).

Replays of saved controllers. We replayed the saved controllers of 512 pilot bodies, all four seeds, on the CPU simulator. Replayed and recorded returns agree at Spearman 0.83 at $H/4$ and 0.93 at H. At $H/4$, 98.7% of the controllers fall, the median episode lasts 0.93 s and covers 1.16 m, and 83% of falls end pitched forward. Figure 7 shows the median body of the quarter-budget screen's top tenth and of final training's top tenth among these 512 bodies, chosen by a rule fixed in advance,

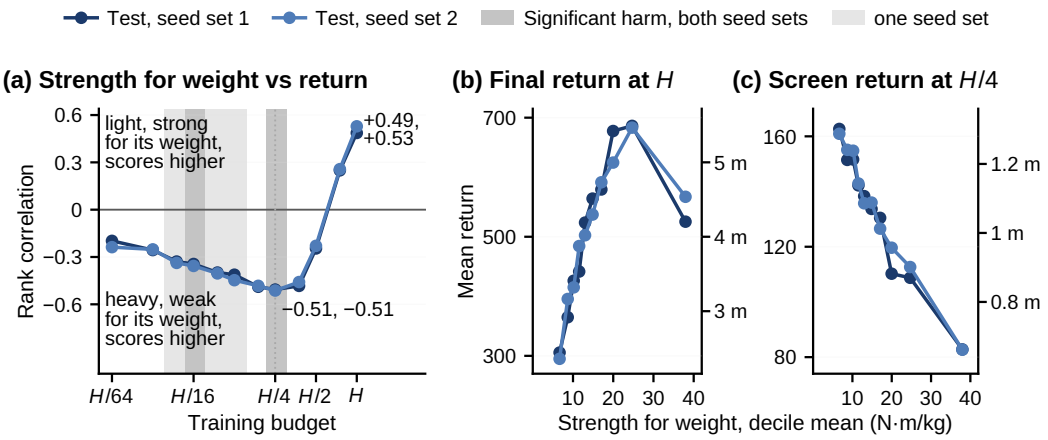


Figure 15: Strength for weight and return in the walker test pool (post hoc). (a) Rank correlation between strength for weight and four-seed mean return against the training budget in both seed sets, with printed values giving the first seed set first. Shaded columns mark budgets where best-of-four is significantly harmful, darker in both seed sets and lighter in one, and a dotted line marks the quarter-budget screen. (b, c) Mean four-seed return by decile of strength for weight, after final training and at the quarter-budget screen, with right axes giving the return divided by 125, the torso's travel in meters less a small action cost. **Takeaway:** The quarter-budget screen favors walkers that are weak for their weight, and final training favors strong ones.

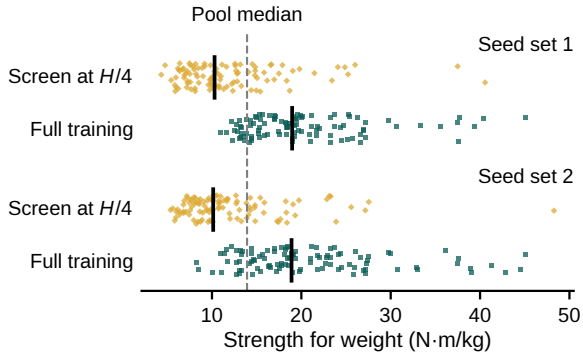


Figure 16: Strength for weight of each body in the top tenth of the test pool ranked by the quarter-budget screen and by full training, in both seed sets, with medians as black ticks and the pool median dashed (post hoc). **Takeaway:** In both seed sets the screen's top tenth is weak for its weight and full training's is strong.

each with the seed, chosen separately at each budget, whose replay is closest to the body's mean return. Its frames, every 0.4 s after brief training and every 0.8 s after full training, and at the fall, sit at their true positions on one scale, with earlier frames paler. Because walker rollouts are chaotic, a single replay can differ from the recorded return.

Gymnasium termination. Gymnasium's Walker2d ends an episode when torso height leaves $(0.8, 2)$ m or pitch leaves $(-1, 1)$ rad. Under Gymnasium's termination rule, our walker uses the same bands with their end points included. An episode ends when torso height leaves $[0.8, 2]$ m, which is $[0.64, 1.6]$ times the spawn height of 1.25 m, or when the pitch exceeds 1 rad in magnitude. Under this rule and our reward, with four new seeds, we analyzed 595 of the first 600 test bodies once, with the analysis fixed before any record was opened (post hoc). The quarter-budget screen of four again picked worse-than-random walkers, -0.11 $[-0.18, -0.03]$ SD, significant at the four screens from $3H/32$ to $H/4$, while from $3H/8$ on the screen helps ($+0.36$ $[+0.27, +0.44]$ at $H/2$). Under this rule 94% of final episodes end in termination, and returns at the full budget agree more

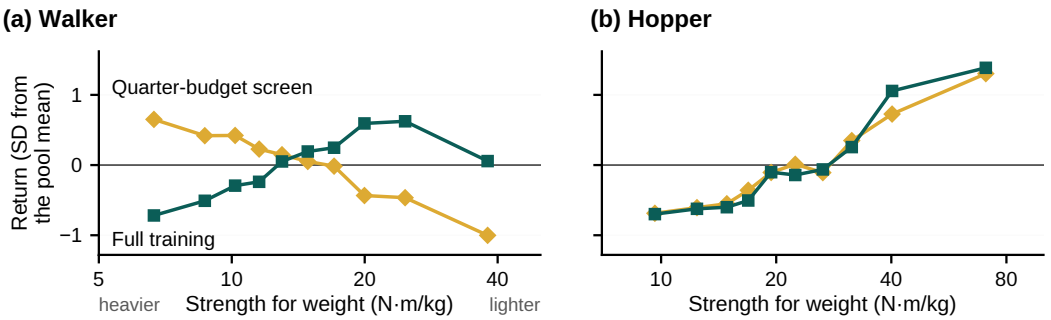


Figure 17: The quarter-budget screen and full training favor opposite walkers, while in the hopper both favor bodies that are strong for their weight. Each point averages one tenth of the pool, sorted by strength for weight, with returns in SD from the pool mean at each budget. **Takeaway:** The walker's quarter-budget screen rewards bodies that are heavy and weak for their weight, and full training ranks those bodies low.

across seeds (mean single-seed correlation 0.48, against 0.22 and 0.12 for the test on these bodies). The Spearman correlation of mass with return is $+0.51$ at $H/4$ and -0.44 at H, and that of strength for weight is -0.54 and $+0.45$, the same reversal.

The analysis of 595 bodies under Gymnasium's termination rule motivated a new preregistered test on the other 424 test bodies, numbered 600 to 1,023. The test protocol was fixed before training began, and each body was trained from scratch with four new seeds. The test used the same estimator and decision rule, with a random draw from these 424 bodies as the baseline. Before scoring the test, the protocol required a mean single-seed correlation of returns after final training of at least 0.07. All 424 bodies completed, and the correlation was 0.36. Best-of-four at $H/4$ has a lift of -0.12 $[-0.19, -0.04]$ SD, with one-sided 95% lower and upper limits of -0.18 and -0.05, so the test is supported, and 94% of its final episodes end in termination. On these bodies mass correlates with return at $+0.54$ at $H/4$ and -0.45 at H, and strength for weight at -0.56 and $+0.49$. The remaining five of the first 600 bodies are in neither set, because their partial records had been read. Pooled with the 595, the 1,019 bodies give -0.11 $[-0.16, -0.06]$, a descriptive estimate, since the new test was designed after the 595 were read. The top tenth of final scores accounts for 85% of the harm in the 595-body analysis and 77% in the preregistered 424-body test. The selection lift detects the harm because it gives more weight to the top of the screen ranking and uses final scores, not only their ranks (Corollary 1).

Where we detected the harm. We detected it when the reward pays for forward travel alone, under both termination rules, at final budgets from the full budget up (and, under Gymnasium's termination rule, narrowly at $3H/4$ for two briefer screens), and when the kept body is retrained. The lift was positive with a screen of half the final budget or, under our termination rule, when the screened controllers were continued. We preregistered the test on new bodies, the follow-up at twice the full budget and the Gymnasium termination test on held-out bodies. The other analyses are post hoc.

G The EvoGym Archive

Mertan & Cheney (2026) released an archive of every connected EvoGym voxel body (Bhatia et al., 2021) with at least three actuated voxels on a 3×3 grid, 1,305,840 in all, where each cell is empty, rigid, soft, or a horizontal or vertical actuator. Each body has one controller-evolution run of 300 generations, with fitness recorded during the run. The fitness after an early generation T is the screen score, and the fitness at generation 299 is the final score. Because the screen score and final score come from the same run, the lift evaluates continuing the screened controllers rather than retraining them. The generation-1 score reflects an effectively untrained controller.

With a generation-1 screen, best-of-N picks worse than random over a broad range of N, reaching -0.36 SD $[-0.38, -0.35]$ at $N = 256$, and recovers to $+0.34$ at $N = 16{,}384$ (Fig. 18a), although that screen is nearly uncorrelated with the final score, as Corollary 1 allows (rank correlation $+0.05$, Pearson -0.03). At generation 5 the lift is negative from $N = 2$ to $1{,}024$, with a minimum of -0.29

SD at $N = 128$, and at generation 10 only from $N = 8$ to 64, with a minimum of -0.08 at $N = 32$. Recomputed at every generation from the archive's fitness records, the lift is positive at every N we evaluate, from 2 to 65,536, for every screen from generation 13 on. Generation 25 is the first plotted screen with this property. At generation 100 best-of-256 gains $+2.62$ SD, 80% of the 3.28 that a perfect screen would gain. For a fixed screening budget of $N \times T$ generations, the generation-200 screen, the longest we compare, gives the largest lift at 24 of the 49 distinct budgets. The shortest screen gives the largest lift only at the three smallest budgets, where no other screen is affordable (Fig. 18b). Figure 19 shows the bodies that each screen keeps in four pools of 256.

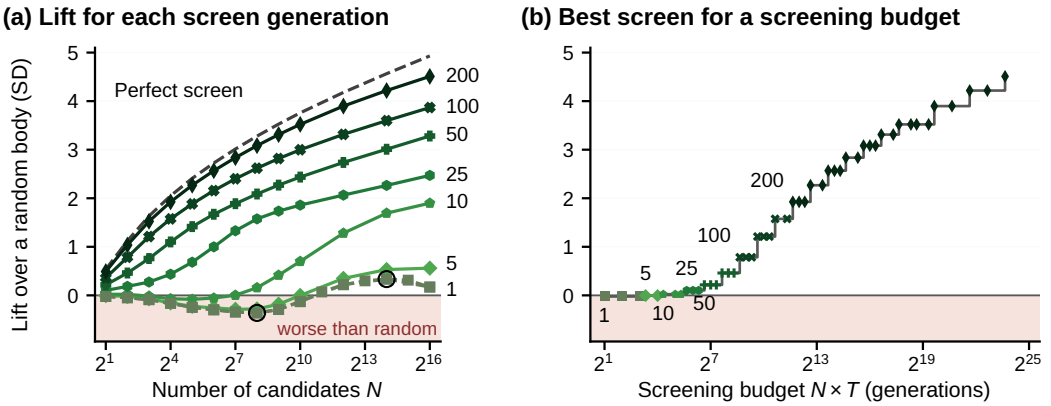


Figure 18: Lift in the public EvoGym archive, with the screened controller continued (post hoc). (a) Lift against N for screens at the generation printed at each line's end, with a perfect screen for reference and the generation-1 minimum and peak ringed. (b) The largest lift among the recorded screens (generations 1 to 200, $N \geq 2$) for a screening budget of $N \times T$ generations, labeled with the screen generation where it takes over, on the same lift axis as (a). **Takeaway:** A generation-1 screen picks worse than random up to about $N = 1{,}024$, while screens from generation 25 onward gain at every N.

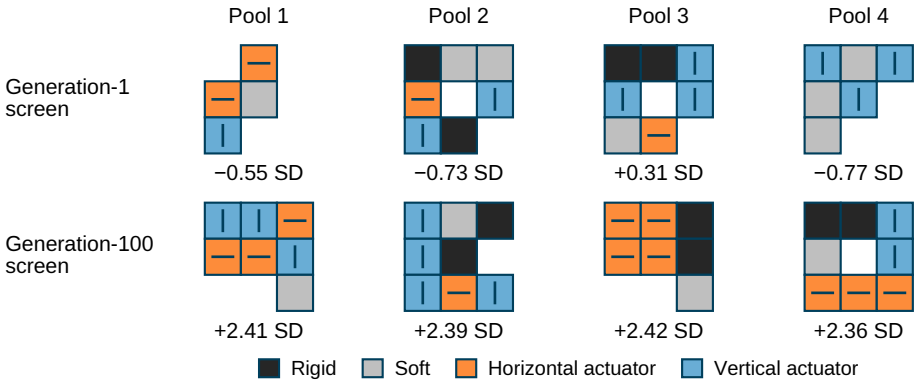


Figure 19: The body that a generation-1 and a generation-100 screen keep in each of four random pools of 256 archive bodies, labeled with its final score in SD from the archive mean. **Takeaway:** Three of the four bodies kept by the generation-1 screen sit below the archive mean, and all four kept by the generation-100 screen sit far above it.

H Body Galleries

Figures 20 to 24 show twelve bodies of each design space and of the archive in the colors of Fig. 2. Bodies of the four spaces we train are drawn through the camera of Fig. 2, sit at evenly spaced ranks of total mass in one pool, and are labeled with their mass. Grammar bodies that settle tipped over are replaced by their neighbor in that ranking. Figure 2 shows walker, hopper and cheetah bodies

at the 10th and 90th percentiles of body volume in their pools, archive bodies at the 10th, 50th and 90th percentiles of voxel count, and six grammar bodies spread over the number of actuated joints, with walkers in a stride pose so both legs show.

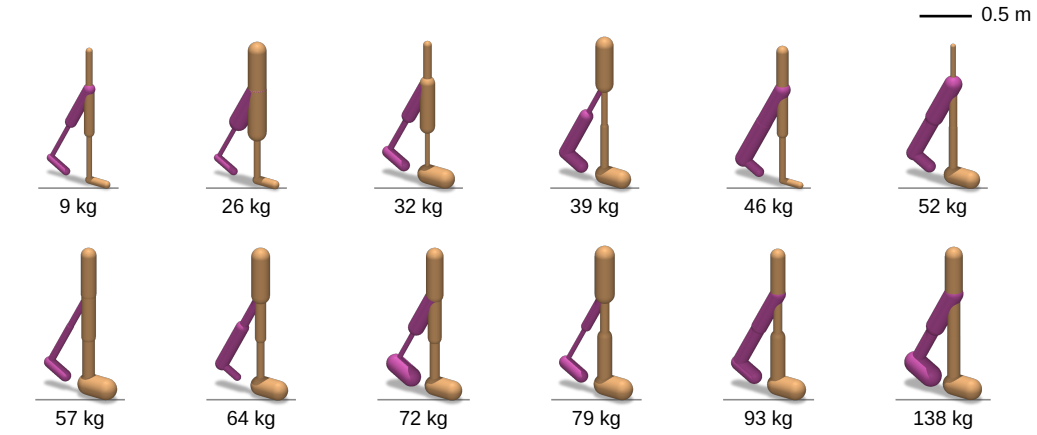


Figure 20: Twelve walkers of the 1,024-body test pool, at one scale, in a stride pose so both legs show. **Takeaway:** With limb lengths fixed, limb thickness and mass vary widely, from 9 to 138 kg.

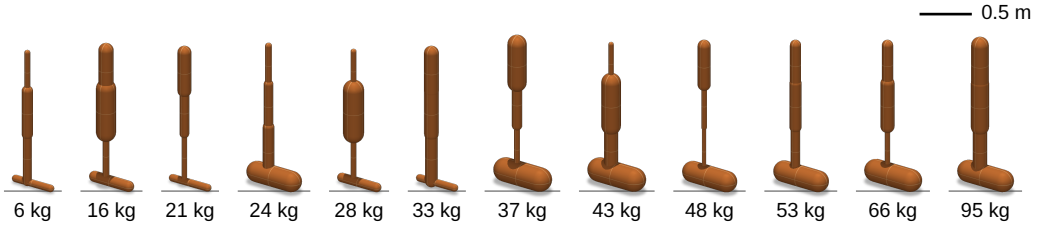


Figure 21: Twelve hoppers of the 256-body pool, at one scale, with masses from 6 to 95 kg.

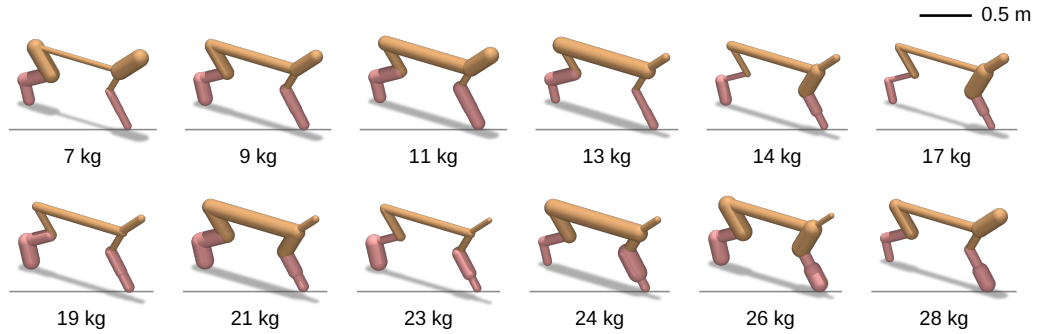


Figure 22: Twelve cheetahs of the original 512-body pool, at one scale. **Takeaway:** Total mass ranges from 7 to 28 kg, and limb radii vary independently of it.

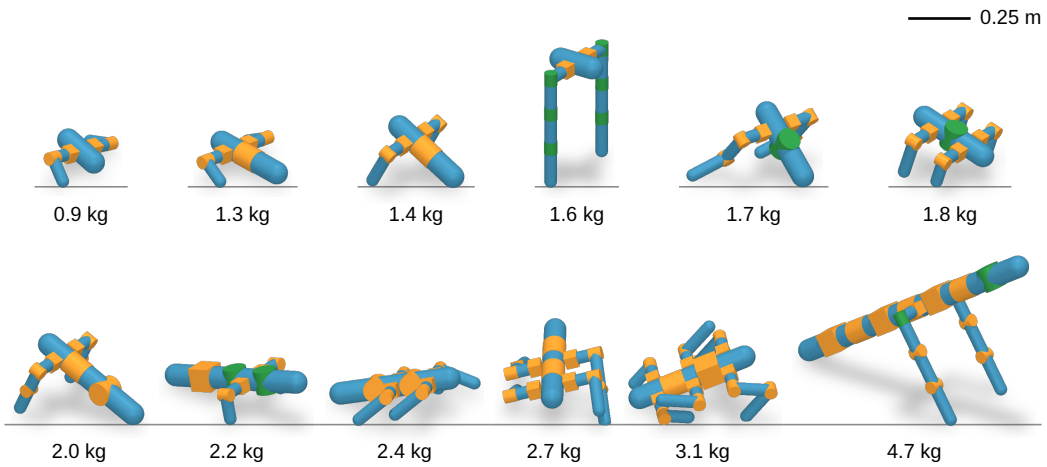


Figure 23: Twelve grammar bodies of the new 2,048-body pool, at one scale. **Takeaway:** The grammar builds bodies with different numbers of segments, legs and links, and the bodies shown weigh 0.9 to 4.7 kg.

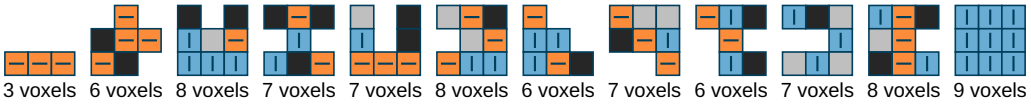


Figure 24: Twelve bodies of the public EvoGym archive at evenly spaced archive indices, labeled with their number of voxels, whose controllers were evolved in the archive. **Takeaway:** These examples span three to nine voxels and together include all four voxel types.

I Further Related Work

Co-design. Co-design benchmarks such as EvoGym and SoftZoo score each body through a controller trained or optimized for it (Bhatia et al., 2021; Wang et al., 2023). Co-optimization of body and controller can converge before it reaches the best regions of body space (Mertan & Cheney, 2024). Learned critics reduce the need for body-specific training by ranking proposed designs from experience collected on other bodies (Luck et al., 2020). Other co-design methods avoid per-body training with task-agnostic scores that need no task training (Hejna et al., 2021) or with shared or universal controllers (Jeon et al., 2025; Strgar & Kriegman, 2026).

Low-fidelity search in AutoML. Asynchronous, model-based and learning-curve methods extend low-fidelity screening of hyperparameters (Li et al., 2020; Falkner et al., 2018; Rakotoarison et al., 2024; Wistuba et al., 2022), and a large database of learning curves finds them more ill-behaved than previously thought (Yan et al., 2025). In hyperparameter search, discarding candidates after one epoch does about as well as successive halving and learning-curve extrapolation on standard benchmarks, and its authors note that it cannot perform well when learning curves often cross (Egele et al., 2024). Population-based training also selects by partial training, and its greedy selection can favor short-term gains (Jaderberg et al., 2017; Dalibard & Jaderberg, 2021). In architecture search, performance predictors have been compared at scale (White et al., 2021), and search policies on average do no better than random selection (Yu et al., 2020).

Best-of-N and test-time compute. For language-model sampling, performance can be non-monotone in the number of calls (Chen et al., 2024), and scaling in the number of attempts follows from the distribution of per-item success (Schaeffer et al., 2025). Compute-optimal test-time scaling trades sampled candidates against sequential revisions (Snell et al., 2025), as our equal-compute comparison trades candidates against training per body. Best-of-N accuracy on each prompt is set by the base accuracy and the verifier's ROC curve, so small-N behavior cannot in general identify large-N behavior (Dorner et al., 2026).

Selection theory. Tang et al. (2020) and Dodge et al. (2021) give the expected maximum of a metric over random subsets of size N with the weights of Eq. 1. Quantitative genetics predicts the response to selection on a trait correlated with the target, here the screen score (Robertson, 1966; Price, 1970; Walsh & Lynch, 2018), and order statistics give the selected individual's expected target value (Nagaraja, 1982) and the chance that such selection keeps the best individuals (Yeo & David, 1984).