AUDITING SIZE DEPENDENCE IN ZERO-COST NAS PROXIES: CONFOUND-CONTROLLED DISCOVERY AND HELD-OUT EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Zero-cost proxy evaluation and automated discovery rely on score–accuracy correlation, but this correlation does not distinguish size-related association from association remaining after size control. In our audit, log parameter count has the largest absolute unadjusted Kendall correlation among 15 baselines in four of the original five cell-based search spaces from Network Design Spaces (NDS), using the finite support of each score. This strong size baseline makes it important to establish what predictive information a proxy retains after size control. Our audit of 24 scores shows that joint parameter-count and FLOPs controls change rankings and reveal obscured associations. We therefore use absolute residual association as the fitness objective for genetic programming (GP), yielding three candidates. The formulas, signs, and control analyses of these three candidates were frozen before NDS access. We designated the focal score ϕ_{frz} and registered the 15-baseline comparison after opening. On architectures where all compared methods produce finite scores in these five spaces, ϕ_{frz} exceeds the strongest eligible baseline in absolute association under residual and stratified controls. All ten margin confidence intervals remain positive after adjustment for the choice among the three candidates. Separately, a post-hoc comparison of 100 matched GP pairs finds higher mean absolute residual association on the five NDS spaces under controlled than unadjusted fitness. This association advantage alone does not establish better selection. Separate retrospective simulations using stored benchmark accuracies on these populations yield positive mean selected-accuracy gains over the strongest of the 15 baselines in 49 of 50 conditions. A post-hoc Bonferroni analysis retains 41 positive gain intervals. The evidence supports controlled discovery of stronger residual associations and, separately, selection gains within the evaluated scope.

1 INTRODUCTION

Zero-cost proxies rank neural architectures without training (Mellor et al., 2021; Abdelfattah et al., 2021) and are commonly evaluated by rank correlation with trained accuracy (Ying et al., 2019; Dong & Yang, 2020). However, model size can itself be predictive. In our audit of the original five cell-based search spaces from Network Design Spaces (NDS) (Radosavovic et al., 2019), log parameter count ranks first among the 15 baselines by absolute unadjusted Kendall correlation with accuracy in four spaces (Table 6). This strong size baseline makes it important to establish what predictive information a proxy retains after size control. Overall correlation alone leaves that question unresolved. The same distinction matters for automated proxy discovery, which uses such correlations as optimization objectives (Akhauri et al., 2022; Dong et al., 2024; Wei et al., 2024). An objective based on overall correlation does not distinguish gains associated with size from gains that remain after size control.

Building on evidence of size dependence (Yang et al., 2023; Krishnakumar et al., 2022), we audit 24 proxy scores across search spaces and task families using residual Kendall correlation and two-dimensional stratification over parameter count and FLOPs. The audit shows that controlling these axes changes proxy rankings and can reveal associations obscured in overall correlation. Figure 2 shows that some negative residual associations reflect sign reversals, while others retain an already

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.

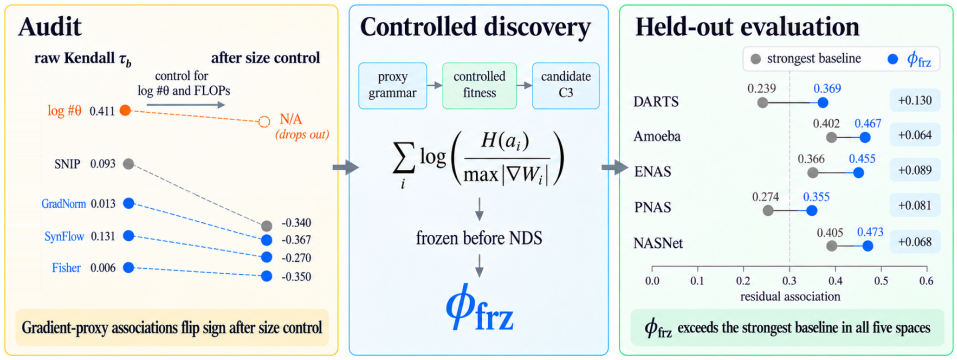

Figure 1: **Audit, controlled discovery, and held-out evaluation.** The left panel illustrates post-opening results on ENAS, where controlling for parameter count and FLOPs reverses several gradient-based proxies from positive unadjusted to negative residual association. The controlled discovery objective and three candidate formulas were fixed before NDS access. Following the sealed opening, one candidate was designated as the focal score ϕ_{frz}. Across all five held-out cell-based NDS spaces, ϕ_{frz} has stronger absolute residual association than any eligible baseline in the post-opening registered 15-baseline roster.

negative sign. The relevant control axes are space dependent. In the TransNAS-Bench-101 macro space, parameter-only control leaves FLOPs-related variation unaccounted for (Section 4).

We address this distinction with a discovery objective based on absolute residual association with accuracy. We use genetic programming (GP) to search formulas over layerwise initialization statistics on NAS-Bench-201, NAS-Bench-301, and NAS-Bench-101. To test whether this objective changes what GP discovers, we compare unadjusted and controlled fitness in a post-hoc study of 100 matched GP seed pairs on NAS-Bench-101. Controlled-fitness champions have higher mean absolute residual association on the five cell-based NDS spaces than unadjusted-fitness champions, using the common-finite support for each pair. All five pointwise paired 95% intervals for these gains are positive (Table 10).

The original discovery stage yielded three candidates, whose formulas, signs, and control analyses were frozen before NDS access. After the sealed opening, we designated one as ϕ_{frz} and registered a strengthened 15-baseline comparison before computing any added baseline score. This score sums layerwise log ratios of activation entropy to maximum gradient magnitude. On common-finite populations, where all methods produce finite scores, ϕ_{frz} exceeds the strongest eligible baseline in absolute association under both controls in all five spaces (Table 2). All ten max-aware margin intervals remain positive after adjustment for the choice among the three frozen candidates (Appendix M.1).

Stronger controlled association does not by itself establish better architecture selection. We therefore evaluate selection separately through retrospective simulations using stored benchmark accuracies. On the common-finite populations of the original five spaces, ϕ_{frz} achieves higher mean selected accuracy than the strongest of the 15 baselines in 49 of 50 combinations of space, candidate pool, and number of candidates scored. A post-hoc paired-bootstrap analysis retains 41 positive gain intervals after Bonferroni correction (Table 22).

Our contributions are threefold.

- **Audit.** We characterize how joint parameter-count and FLOPs controls change proxy evaluation across spaces and tasks.
- **Controlled discovery.** We make residual association a search objective and test its effect through matched GP comparisons under a fixed search setup.
- **Candidate evaluation.** We establish stronger controlled association and separately measured retrospective selection gains for ϕ_{frz} in the specified NDS comparisons.

2 Related Work

Zero-cost proxies and their evaluation. Zero-cost proxies score architectures using statistics computed from untrained networks (Mellor et al., 2021; Abdelfattah et al., 2021; Tanaka et al., 2020; Li et al., 2023; Lin et al., 2021). Their evaluation has consolidated around rank correlation with trained accuracy on tabular and surrogate benchmarks (Ying et al., 2019; Dong & Yang, 2020; Zela et al., 2022; Duan et al., 2021), and NAS-Bench-Suite-Zero systematizes this practice by precomputing 13 proxies on 28 tasks (Krishnakumar et al., 2022). Recent proxies such as AZ-NAS, MeCo, SWAP, and NEAR report strong rank correlations under this evaluation paradigm (Lee & Ham, 2024; Jiang et al., 2023; Peng et al., 2024; Husistein et al., 2025). Our audit asks how much association remains after controlling for size-related variation.

Size bias and debiasing. The size dependence of proxy scores has been documented previously. Ning et al. (2021) documented estimator biases, competitive parameter-count and FLOPs baselines, and search-space-dependent behavior across multiple NAS benchmarks. Yang et al. (2023) showed that popular training-free metrics correlate strongly with parameter count and proposed an efficient training-based alternative. Krishnakumar et al. (2022) reported biases toward cell size and skip connections, together with bias-reduction procedures that can trade correlation for debiasedness. Recent work has also reported negative-correlation regimes for activation-based training-free proxies as convolutional structure varies (Kang et al., 2025). GRAF showed that simple neural-graph features, including operation-count and topology summaries, can themselves be strong performance predictors (Kadlecová et al., 2024). These findings motivate distinguishing size-related association from association remaining after size control. We extend this line through joint controls over multiple size axes, a controlled discovery objective, and separate retrospective evaluation of architecture selection.

Separating size by design. A complementary line builds the separation into the proxy. Entropic Score searches topology with an entropy metric while delegating model size to a separate search dimension (Cavagnero et al., 2023b), and FreeREA rescales SynFlow logarithmically to temper raw score magnitudes (Cavagnero et al., 2023a). These approaches motivate treating size-related variation explicitly rather than assuming it away. VKDNW combines a normalized Fisher-spectrum entropy score with a separate layer-count term for ranking (Tybl & Neumann, 2025). We complement these proxy designs by controlling measured size axes in evaluation and discovery fitness.

Automated proxy discovery. EZNAS evolves proxies over a statistic grammar with raw-correlation fitness (Akhauri et al., 2022), while Auto-Prox and LPZero extend automated proxy discovery to vision transformers and language models (Wei et al., 2024; Dong et al., 2024). Phan & Luong (2025) also use genetic programming to synthesize training-free predictors from existing hand-crafted proxies across multiple NAS search spaces and tasks. EZNAS describes its best evolved proxy as a form of weighted parameter counting. Our procedure instead compares raw and size-controlled discovery objectives over the same primitive space, isolating the effect of confound control in the discovery objective itself.

3 Spaces, Statistics, and Sealing

Benchmarks and roles. Our study separates development from evaluation by construction. Table 1 lists the benchmarks used in the original audit, discovery, and sealed evaluation, with their sizes, targets, and roles. The development spaces support the audit and the discovery engine, while the sealed suite was archived under SHA-256 hashes before evaluation. Here, *held out* means that candidate formulas, signs, and size-control analyses were fixed before NDS access. The candidate labeled C3 was designated as the focal score ϕ_{frz} after opening, and the subsequent 15-baseline comparison was then registered, as labeled below.

Frozen candidate set and post-opening registration. Discovery closed with three candidates frozen in a registration commit, together with the size-control analyses, incumbent baselines, and pre-opening judgment and stopping rules. The sealed benchmarks were then opened in a single scripted run that verifies the archive hashes before reading any target. Every subsequent analysis

Table 1: Benchmark inventory. Architecture counts include only rows with finite targets. Sealed rows were hashed before opening and opened once.

Benchmark	Target	#Archs	Role
NAS-Bench-201	test accuracy (3 datasets)	15,625	audit, discovery
NATS-Bench-SSS	test accuracy (3 datasets)	32,768	transfer gate (§5)
NAS-Bench-101	validation accuracy	9,781	discovery (C3)
NAS-Bench-301	surrogate validation accuracy	11,221	discovery validation
TransNAS-Bench-101 micro	7 task metrics	4,096	sealed (C1 transfer)
TransNAS-Bench-101 macro	7 task metrics	3,256	sealed (C1 transfer)
NDS DARTS / AmoebaNet / ENAS / PNAS / NASNet	test accuracy (CIFAR-10)	4,846–5,000 per space	sealed (ϕ_{frz}, held-out)

is labeled post-opening, and no candidate formula, sign, control analysis, or hyperparameter was changed after opening. To strengthen the comparison, we registered a 15-baseline bar, including nine published proxies, after opening but before computing any added baseline score. Appendix H gives the full timeline and stopping rule.

Analysis populations and comparators. The target-finite rows of each benchmark define its available population. Every task target is oriented in its registered favorable direction before analysis. The main NDS comparison, registered after opening, uses one common-finite set

$$\mathcal{I}_{\mathcal{B}} = \{i : y_i,\ \phi_{\mathrm{frz}}(i),\ s_{ij}\ (j \in \mathcal{B}) \text{ are all finite}\},$$

where $\mathcal{B} = \mathcal{B}_{15}$ is the registered 15-score roster of six statistic baselines and nine published proxies. Thus all scores are evaluated on the same architectures, and the n used in the transforms below is $|\mathcal{I}_{\mathcal{B}}|$. Alternative row and coverage policies are treated as post-hoc sensitivity analyses, whereas the later 21-score extension was registered before its six additional baseline scores were computed. Neither replaces the main comparison population or the 15-baseline bar (Appendices C, K, and L).

Association statistics. For a length-n vector v, let $\mathcal{N}_n(v_i) = \Phi^{-1}(r_i/(n+1))$, where r_i is its average rank. We denote the unadjusted rank association by $T_{\mathrm{raw}}(s, y) = \tau_b(s, y)$. Both the raw and residual analyses use Kendall τ_b, whereas the stratified analysis uses a separate pair-count-weighted within-stratum concordance statistic defined below. Unless stated otherwise, Z contains log parameter count and log FLOPs. With $\mathcal{N}_n$ applied columnwise to Z, let $X_Z = [\mathbf{1}, \mathcal{N}_n(Z)]$ and $P_Z = X_Z X_Z^{\dagger}$. The residual analysis is

$$T_{\mathrm{res}}(s, y; Z) = \tau_b((I - P_Z)\mathcal{N}_n(s), (I - P_Z)\mathcal{N}_n(y))\,. \tag{1}$$

Because log parameter count is itself included in Z, its residual association is undefined and it is excluded from residual comparisons. The stratified analysis creates five parameter-quantile bins and, within each, four FLOPs-quantile bins. For $\mathcal{G} = \{g : n_g \geq 3\}$, it reports

$$T_{\mathrm{strat}} = \frac{\sum_{g \in \mathcal{G}} \binom{n_g}{2} \tau_{a,g}}{\sum_{g \in \mathcal{G}} \binom{n_g}{2}},$$

where $\tau_{a,g}$ is the within-stratum Kendall pair-concordance statistic, so each within-stratum architecture pair receives equal weight.

Size-explained diagnostic. The ratio $M(S) = \rho_Z/\rho$ uses the orthogonal split $\rho = \rho_Z + \rho_\perp$ of the Pearson normal-score association (Appendix C). It is a noncausal diagnostic rather than a proportion. Cancellation can put it outside $[0, 1]$, and it is undefined when $\rho = 0$. The residual association of log parameter count under its own control is also undefined, giving the open endpoint in Figure 1.

Main contrast and inference. For analysis $\ell \in \{\mathrm{res}, \mathrm{strat}\}$, let $\mathcal{B}_{\mathrm{res}} = \mathcal{B}_{15} \setminus \{\log \#\theta\}$ and $\mathcal{B}_{\mathrm{strat}} = \mathcal{B}_{15}$. The strengthened bar and candidate margin are

$$B_\ell = \max_{j \in \mathcal{B}_\ell} |T_\ell(s_j, y; Z)|, \qquad D_\ell = |T_\ell(\phi_{\mathrm{frz}}, y; Z)| - B_\ell, \quad \ell \in \{\mathrm{res}, \mathrm{strat}\}. \tag{2}$$

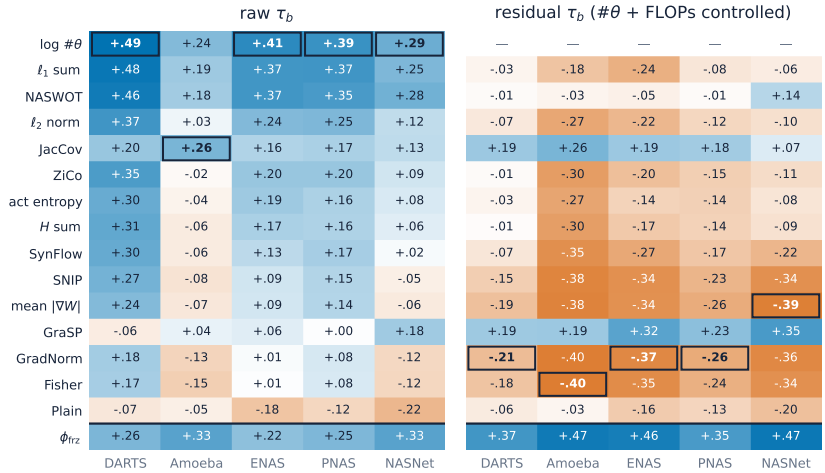

Figure 2: **The leaderboard reorders under size control.** Raw and residual Kendall τ_b associations (left and right), with each estimate computed on the finite support of the corresponding score, for the post-opening registered 15-baseline roster and ϕ_{frz} on the five sealed NDS spaces. Within each panel, boxes mark the largest absolute baseline association shown in each space. ϕ_{frz} appears below the rule for comparison. Cell values retain their signs.

We jointly resample architecture rows for the candidate and all eligible baselines, refit the residual analysis, and re-select the maximizing baseline within every BCa bootstrap replicate and jackknife group. Stratification labels are fixed from the target-finite benchmark population before the common-finite restriction and remain fixed during resampling. Under the post-opening strengthened-bar registration, an analysis is a win only when both the point margin and its 95% BCa lower endpoint exceed zero. A space is a loss if either analysis fails. Based on these pointwise intervals, the stopping rule retains the five-space headline at zero losses, scales it down at one or two, and halts it at three or more. A separate post-hoc Bonferroni pass checks the ten strengthened-bar space–analysis claims, and all ten remain above zero (Appendix L). These intervals quantify resampling stability of each fixed benchmark population, not superpopulation uncertainty. They condition on the focal designation of C3, and a post-hoc family over all three frozen candidates keeps every C3 margin positive (Appendix M.1). Exact resample counts, streams, roster membership, and the distinct Holm family for H-S1 are given in Appendices C, J, H, and L.

Noise ceilings. Benchmark accuracies are noisy estimates of architecture quality, so we do not treat $\tau = 1$ as an attainable reference. Under the replicate-noise assumptions of Section 7, reliability R implies the bound $R \leq C_\tau \leq \sqrt{R}$. On NAS-Bench-201, replicate measurements come from independent training seeds. The original five cell-based NDS spaces provide only one training seed per architecture, so we instead use jitter over the final converged epochs as a heuristic upper-bound reference. This reference is conservative for claims of proximity to the achievable ceiling but is not a direct application of the replicate-noise bound.

4 Auditing Size Dependence

The NAS-Bench-201 audit. We first audit 18 score implementations on NAS-Bench-201, including MeCo, NEAR, LogSynflow, and Entropic Score. Size control affects proxies unevenly. MeCo remains the strongest incumbent under control, with $T_{\text{res}} = 0.710$ versus $T_{\text{raw}} = 0.742$, whereas NASWOT falls from 0.464 to 0.371 with $M(S) = 0.52$. Several other proxies, including JacCov, EPE-NAS, a regularized SWAP variant, and weight norms, have size-explained ratios near or above 0.5. The wide variation in $M(S)$ across the roster shows that size dependence is proxy-specific and should be measured rather than assumed, even within NAS-Bench-201. Full tables, the roster, and a reproducibility note on the Zen score, which we recompute because public implementations disagree, appear in Appendix B.

Across spaces. The picture sharpens away from NAS-Bench-201. On the NAS-Bench-301 surrogate space, the strongest residual associations among the audited published proxies are substantially weaker than on NAS-Bench-201, and the proxy ordering changes markedly. On NAS-Bench-101, the ordering changes again, and signals that are weak or obscured in the raw ranking become visible under control.

The confound axis is space dependent. Controlling an incomplete scale-related axis can leave substantial confounding. On the TransNAS-Bench-101 macro tasks, parameter-only stratification often remains aligned with the raw association, while the residual and two-dimensional stratified analyses can yield substantially different conclusions. For mean gradient magnitude on semantic segmentation, the parameter-only analysis gives -0.289, whereas the residual and two-dimensional stratified analyses give $+0.426$ and $+0.141$, respectively. The macro space varies depth and stride, so controlling for parameters alone leaves FLOPs-related variation unaccounted for. Parameter-only stratification is therefore excluded from macro judgments under the prespecified control scheme.

On NAS-Bench-201, parameter-only and FLOPs-only controls give similar audit rankings to the two-axis control (Spearman $\rho = 0.975$–0.977), while adding convolution-edge count gives $\rho = 0.845$ (Appendix B).

Published proxies on NDS. Figure 2 shows the NDS audit on the five sealed spaces for the nine published proxies and six statistic baselines that form the strengthened 15-baseline bar. In the raw panel, log parameter count is strongest among the baselines on DARTS, ENAS, PNAS, and NASNet, while JacCov is strongest on AmoebaNet. In the residual panel, the gradient family (SNIP, GradNorm, Fisher, SynFlow, ZiCo) is negative in all five spaces, with GradNorm on ENAS moving from +0.013 raw to -0.367 residual, while NASWOT falls between -0.045 and +0.144. The baselines with positive residual association in all five spaces are GraSP and JacCov. ϕ_{frz} is shown in Figure 2 only for comparison. Its formal strengthened-bar evaluation uses the joint common-finite intersection and is reported in Section 6.

Finding 1

Overall proxy–accuracy association does not distinguish size-related association from association remaining after size control. Controlling for parameter count and FLOPs can substantially change both proxy rankings and their association with accuracy.

5 Confound-Controlled Discovery

Engine. The engine searches programs over layerwise initialization statistics to explicitly optimize association remaining after size control. For every architecture, we precompute nine per-layer statistics for each of the weight, gradient, and activation tensors, in a pipeline that never reads accuracy. Programs compose these statistics through a prefix-coded grammar. Fitness averages absolute residual Kendall association across the development datasets and subtracts a program-node penalty (Appendix D), so the objective itself is size controlled. Fitness is evaluated on the full architecture population (15,625 on NAS-Bench-201) rather than a small subsample, which eliminates fitness variation due to architecture subsampling.

Discovery outcomes across spaces. On NAS-Bench-201, all ten runs, five seeds under each of the controlled and raw objectives, converged to a gradient-entropy family with normal form $\mathrm{C1} = \sum_l \|W_l\|_1 H(|\nabla W_l|)^4$. Its residual association exceeds the strongest incumbent MeCo on all three datasets. The raw runs converged to the same family, so our preregistered prediction that the raw objective would drift toward a size proxy did not hold. Appendix D reports the detailed diagnostics.

Rerunning discovery independently across spaces produces space-specific champions on NAS-Bench-201, NAS-Bench-301, and NAS-Bench-101. The NAS-Bench-301 champion is a sign-reversed gradient mean (C2), whose magnitude is already represented by an incumbent baseline, whereas the NAS-Bench-101 run produces a layerwise log-ratio program. Before NDS access, we froze its literal form as $\mathrm{C3} = \sum_l \log\left(H(a_l)/\max|\nabla W_l|\right)$. This literal form differs from the protected operator semantics used during search (Appendix D). A registered transfer gate on NATS-Bench-SSS kept the residual association of C1 positive on all three datasets despite a size-explained

Table 2: Registered evaluation of the frozen score ϕ_{frz} on the five sealed NDS spaces. Main results use the common-finite intersection rows. For $\ell \in \{\mathrm{res}, \mathrm{strat}\}$, D is the max-aware margin over the registered 15-baseline roster, with the strongest baseline re-selected inside each BCa resample.

Space	Analysis	$T_\ell(\phi_{\mathrm{frz}})$	bar (holder)	D [BCa 95%]	switch	$n_\cap$
DARTS	residual	0.369	0.239 (mean $\lvert\nabla W\rvert$)	$+0.130$ [0.117, 0.141]	0.2%	3,660
	stratified	0.326	0.250 (mean $\lvert\nabla W\rvert$)	$+0.076$ [0.064, 0.090]	0.0%	
AmoebaNet	residual	0.467	0.402 (Fisher)	$+0.064$ [0.053, 0.076]	1.3%	4,463
	stratified	0.417	0.385 (mean $\lvert\nabla W\rvert$)	$+0.032$ [0.023, 0.040]	0.0%	
ENAS	residual	0.455	0.366 (GradNorm)	$+0.089$ [0.078, 0.100]	0.0%	4,506
	stratified	0.421	0.367 (GradNorm)	$+0.054$ [0.043, 0.066]	4.0%	
PNAS	residual	0.355	0.274 (mean $\lvert\nabla W\rvert$)	$+0.081$ [0.070, 0.092]	8.2%	4,478
	stratified	0.331	0.270 (mean $\lvert\nabla W\rvert$)	$+0.061$ [0.051, 0.072]	0.0%	
NASNet	residual	0.473	0.405 (mean $\lvert\nabla W\rvert$)	$+0.068$ [0.061, 0.075]	0.0%	4,322
	stratified	0.466	0.429 (mean $\lvert\nabla W\rvert$)	$+0.038$ [0.031, 0.045]	0.0%	

diagnostic above 0.94. These outcomes motivate studying the controlled objective across spaces, without assuming a universal fixed-sign formula.

Objective comparison. On NAS-Bench-101, mean gradient magnitude has raw correlation -0.138 and residual correlation -0.450, motivating an objective that targets association obscured in overall correlation. In a post-hoc extension with 100 matched GP seed pairs on NAS-Bench-101, controlled-fitness champions have higher mean external absolute residual association than raw-fitness champions in the original five cell-based NDS spaces by 0.025–0.058 on pair-specific common-finite support. All five pointwise paired 95% intervals are above zero (Table 10). This supports the objective under the evaluated search setup. The equal-budget random-search comparison on NAS-Bench-201 is descriptive only (Appendix D), and these results do not establish an advantage of GP as an optimizer.

Freezing the candidates. Discovery closed by freezing C1, C2, and C3, together with their signs, the two size-control analyses, and hyperparameters, in the registration commit before NDS access. The pre-opening sealed-evaluation plan included all three candidates and the five NDS spaces. At the scripted opening, C1 had residual associations from -0.13 to $+0.14$, whereas C3 had the largest residual association among the three candidates in all five spaces (Table 14). We then designated C3 as the focal score ϕ_{frz}. Subsequent comparisons are conditional on this post-opening designation and do not independently validate candidate selection. We contrast its frozen literal form with the protected search-time semantics in Appendix K.

Finding 2

In the post-hoc 100-pair comparison, controlled fitness yields higher mean external absolute residual association on pair-specific common-finite support in all five NDS spaces. This supports the objective under the evaluated setup.

6 Evaluation of the Frozen Candidate

The frozen candidate against the strengthened bar. Table 2 reports the post-opening registered comparison in the original five cell-based NDS spaces. Under both controls, ϕ_{frz} exceeds the strongest eligible member of the 15-baseline roster in absolute association. All ten max-aware BCa margin intervals, which re-select the baseline in each resample, lie above zero. Point margins range from $+0.032$ to $+0.130$ and lower limits from $+0.023$ to $+0.117$. The strongest baseline changes in at most 8.2 percent of resamples. All ten intervals remain positive after adjustment for the choice among the three frozen candidates (Appendix M.1). The registered 21-baseline extension preserves the population and all ten positive margins (Table 21). Separately, adding AZ-NAS and three VKDNW forms to the original roster also keeps all ten adjusted intervals positive in a post-hoc analysis (Appendix M.4).

Sensitivity and transfer limits. Linear layer-count, depth, and width controls retain adjusted positive margins in all five cell-based spaces, but only five of ten split nonlinear comparisons do so

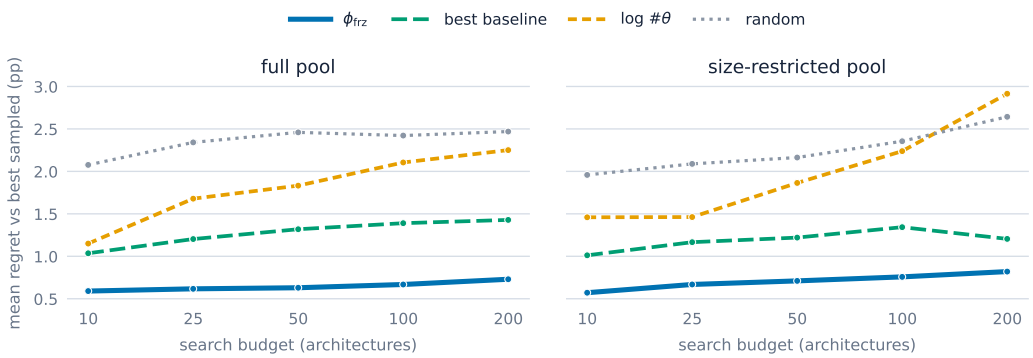

Figure 3: **Registered retrospective budgeted search.** Results use the original pools before common-finite restriction. Mean regret against the best architecture inside the sampled budget, averaged over the original five NDS spaces, with 1,000 draws per condition. The best-baseline curve re-selects the strongest member of the 15-baseline roster in every condition. ϕ_{frz} attains the lowest regret throughout the registered full and size-restricted pools, and in the size-restricted pool the parameter-count proxy incurs higher regret than random selection at large budgets.

(Table 24). Five further families were evaluated under a protocol fixed before target access. All eight adjusted intervals for the four residual families lie below zero. Vanilla's adjusted residual interval is positive, but its stratified interval includes zero (Table 28). The advantage thus depends on controls and design family.

Missing values. ϕ_{frz} is undefined on 9.9 to 26.8 percent of architectures in the original five spaces. Every observed non-finite layer term has zero maximum gradient. In 40 affected architectures per space, random reweighting of these layers left the logits unchanged for a fixed input (Appendix M.3). The selector ranks finite scores above undefined scores under the registered policy. Undefined scores are associated with lower accuracy (rank-biserial -0.26 to -0.33). The main association comparisons use common-finite rows. Since all 15 baselines are finite wherever ϕ_{frz} is finite, this intersection equals its finite support. Association claims condition on this systematically stronger subset.

Two post-hoc NDS sensitivity checks address support asymmetry without changing ϕ_{frz}. Both a failure-as-worst treatment of non-finite scores and a matched-coverage analysis yield positive margins across the ten space–analysis evaluations. These checks address support and sample-count asymmetry rather than informative missingness itself. The corresponding TransNAS missingness summary and full NDS procedures and results are reported in Appendix K.

Retrospective budgeted search. We separately evaluate selection using stored benchmark accuracies. Each draw scores $b \in \{10, 25, 50, 100, 200\}$ candidates and selects the highest-scoring architecture. Registered simulations use 1,000 draws per condition in full and size-restricted pools, the latter at or below the space median parameter count. On common-finite populations, ϕ_{frz} has higher mean selected accuracy than the strongest of the 15 baselines in 49 of 50 conditions. Post-hoc paired-bootstrap intervals are positive in 45 conditions nominally and 41 after Bonferroni correction (Table 22). They describe candidate-set draw variability within fixed benchmark conditions, not independent training runs or tasks.

Without common-finite restriction, mean selected accuracy favors ϕ_{frz} in all 50 conditions. Averaged across the five spaces, it has the lowest regret at every displayed budget in both pools (Figure 3).

Finding 3

On common-finite populations in the original five cell-based NDS spaces, ϕ_{frz} exceeds the post-opening registered 15-baseline bar in absolute association under both controls, with all ten max-aware BCa margin intervals positive. Separately, retrospective selection yields positive mean accuracy gain estimates in 49 of 50 conditions, with 41 positive intervals after post-hoc Bonferroni correction.

7 THEORETICAL CONTEXT

Four results clarify mechanisms and scope limits, with formal statements and proofs in Appendix A. The τ_a quantities appearing in T2 and T4 are theorem-specific pair-concordance statistics and are distinct from the default τ_b-based raw and residual estimators in Section 3.

T1 (Size carrier in extensive sums). A layerwise sum whose conditional means align with effective size inherits a size carrier. Under the stated Bernstein condition, misordering probability decays exponentially with the effective-size margin. This is a sufficient mechanism rather than a claim about every layerwise sum. Per-layer normalization can remove a leading carrier without eliminating all size dependence.

T2 (Occam radius for adaptive discovery). Prefix-code length controls a uniform concentration radius for raw τ_a under i.i.d. sampling, including adaptive search. An analogous result holds for fixed-stratum pair statistics under the stated conditions, but neither result certifies same-sample T_{res}. Exhaustive NAS-Bench-201 evaluation therefore provides only a description-length diagnostic.

T3 (Linear shrinkage diagnostic). The population Pearson-correlation optimum for $S - \lambda P_Z S$ has closed form. Interpreting this optimum for Kendall correlation requires a Gaussian-copula approximation, so λ^* is a diagnostic rather than a rank-optimality result. For C1 on NAS-Bench-201, $\lambda^* = 0$ indicates that no additional linear subtraction improves Pearson correlation under this model (Appendix I), without a corresponding claim for C3.

T4 (Noise ceiling). Under conditionally i.i.d. symmetric pair noise across independent seeds, $R \leq C_\tau \leq \sqrt{R}$ bounds the attainable τ_a ceiling for evaluation-seed-independent scores. NDS epoch jitter does not satisfy the independent-seed premise and is therefore only a heuristic reference. The main evidence is empirical.

8 DISCUSSION, LIMITATIONS, AND CONCLUSION

Retrospective selection utility. Our simulations evaluate selection in full pools and pools restricted to architectures at or below the space median parameter count. In the latter, parameter count can perform worse than random selection at large candidate budgets (Section 6).

Practical value of the signal. The controlled correlations of ϕ_{frz} (0.33–0.47) remain below the noise-ceiling references. The score requires one forward and backward pass per architecture and takes a median 105 ms on one GPU. Table 27 reports costs under the respective batch sizes and implementation settings.

Limitations. NAS-Bench-301 uses surrogate targets. The original five NDS spaces share cell-based designs and CIFAR-10, with noise-ceiling references based on epoch jitter rather than independent training seeds. They are not five independent domains. Row-bootstrap intervals describe perturbations within the observed benchmarks, not uncertainty over broader architecture populations. The discovery grammar uses per-layer scalar statistics. In addition, ϕ_{frz} has informative non-finite support, so its association claims remain conditional on the finite domain. On that domain, literal and protected evaluations rank architectures identically (Appendix M.3). The entropy sum remains closely tied to layer count, and the nonlinear results limit robustness claims (Appendix M.2). Measured controls do not exclude unmeasured structural confounding. Selection is evaluated using stored accuracies rather than new training runs. Common-finite gains are not uniformly positive across conditions. Discovery requires labels and lacks a matched-label comparison to learned predictors. The objective-isolation runs do not establish reliable rediscovery of C3 (Appendix E).

Conclusion. Log parameter count has the largest absolute unadjusted Kendall correlation among 15 baselines in four of the original five cell-based NDS spaces. Controlled fitness yields higher mean external absolute residual association in the post-hoc matched GP comparison. On common-finite populations in these spaces, the candidate designated after opening exceeds the 15-baseline bar under both controls, with separate retrospective evidence of selection gains. Mixed nonlinear results and additional residual-family failures bound the association advantage.

Reproducibility Statement

The supplementary repository provides the analysis code, configuration files, machine-readable results, and scripts and commands used to generate the reported tables and figures. It also contains the registration records, sealed-opening record, and SHA-256 manifests corresponding to the experimental stages described in the paper. Separate records are provided for the sealed evaluation, the strengthened 15-baseline comparison, and the subsequent sensitivity analyses. Public benchmark archives and large regenerable intermediate files are not redistributed. The repository instead provides acquisition or restoration instructions, expected input hashes, benchmark row-alignment rules, finite-support policies, random seeds, and RNG stream conventions. The statistic banks can be regenerated from the public benchmark data using the released configurations and extraction scripts. The supplementary repository is distributed as a reproducibility snapshot. It maps the reported computations and outputs to the chronology and analysis labels given in the paper, but the snapshot alone is not intended as independent evidence for the timing of every research decision. The pre-opening, registered post-opening, and post-hoc designations therefore follow the chronology reported in the main text and appendix.

Ethics Statement

This work uses publicly released architecture benchmarks only, involves no human subjects and no personally identifiable data, and raises no foreseeable ethical concern beyond the general energy cost of the underlying benchmarks, which were trained by their original authors.

AI Use Statement

We used generative AI for writing and editing, literature retrieval and discovery, research ideation and execution (experimental design, coding, statistical analysis and interpretation, and figure preparation), and developing and verifying mathematical claims and proofs. It did not generate or alter datasets or produce benchmark scores. The authors reviewed all AI-assisted work and take responsibility for the final content.

References

Mohamed S. Abdelfattah, Abhinav Mehrotra, Łukasz Dudziak, and Nicholas D. Lane. Zero-cost proxies for lightweight NAS. In *International Conference on Learning Representations (ICLR)*, 2021.

Yash Akhauri, Juan Pablo Muñoz, Nilesh Jain, and Ravishankar Iyer. EZNAS: Evolving zero-cost proxies for neural architecture scoring. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Niccolò Cavagnero, Luca Robbiano, Barbara Caputo, and Giuseppe Averta. FreeREA: Training-free evolution-based architecture search. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023a.

Niccolò Cavagnero, Luca Robbiano, Francesca Pistilli Mercurio, Barbara Caputo, and Giuseppe Averta. Entropic score metric: Decoupling topology and size in training-free NAS. In *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2023b.

Peijie Dong, Lujun Li, Xiang Liu, Zhenheng Tang, Xuebo Liu, Qiang Wang, and Xiaowen Chu. LPZero: Language model zero-cost proxy search from zero. In *Findings of the Association for Computational Linguistics: EMNLP*, 2024.

Peijie Dong, Lujun Li, Zhenheng Tang, Xiang Liu, Zimian Wei, Qiang Wang, and Xiaowen Chu. ParZC: Parametric zero-cost proxies for efficient NAS. In *AAAI Conference on Artificial Intelligence*, 2025.

Xuanyi Dong and Yi Yang. NAS-Bench-201: Extending the scope of reproducible neural architecture search. In *International Conference on Learning Representations (ICLR)*, 2020.

Yawen Duan, Xin Chen, Hang Xu, Zewei Chen, Xiaodan Liang, Tong Zhang, and Zhenguo Li. TransNAS-Bench-101: Improving transferability and generalizability of cross-task neural architecture search. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Raphael Husistein, Markus Reiher, and Marco Eckhoff. NEAR: A training-free pre-estimator of machine learning model performance. In *International Conference on Learning Representations (ICLR)*, 2025.

Tangyu Jiang, Haodi Wang, and Rongfang Bie. MeCo: Zero-shot NAS with one data and single forward pass via minimum eigenvalue of correlation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Gabriela Kadlecová, Jovita Lukasik, Martin Pilát, Petra Vidnerová, Mahmoud Safari, Roman Neruda, and Frank Hutter. Surprisingly strong performance prediction with neural graph features. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 22771–22816, 2024.

Haidong Kang, Lianbo Ma, Pengjun Chen, Guo Yu, Xingwei Wang, and Min Huang. Beyond the limits: Overcoming negative correlation of activation-based training-free nas. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 796–805. IEEE, 2025.

Arjun Krishnakumar, Colin White, Arber Zela, Renbo Tu, Mahmoud Safari, and Frank Hutter. NAS-Bench-Suite-Zero: Accelerating research on zero cost proxies. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2022.

Junghyup Lee and Bumsub Ham. AZ-NAS: Assembling zero-cost proxies for network architecture search. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Guihong Li, Yuedong Yang, Kartikeya Bhardwaj, and Radu Marculescu. ZiCo: Zero-shot NAS via inverse coefficient of variation on gradients. In *International Conference on Learning Representations (ICLR)*, 2023.

Ming Lin, Pichao Wang, Zhenhong Sun, Hesen Chen, Xiuyu Sun, Qi Qian, Hao Li, and Rong Jin. Zen-NAS: A zero-shot NAS for high-performance image recognition. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Joe Mellor, Jack Turner, Amos Storkey, and Elliot J. Crowley. Neural architecture search without training. In *International Conference on Machine Learning (ICML)*, 2021.

Xuefei Ning, Changcheng Tang, Wenshuo Li, Zixuan Zhou, Shuang Liang, Huazhong Yang, and Yu Wang. Evaluating efficient performance estimators of neural architectures. *Advances in Neural Information Processing Systems*, 34:12265–12277, 2021.

Yameng Peng, Andy Song, Haytham M. Fayek, Vic Ciesielski, and Xiaojun Chang. SWAP-NAS: Sample-wise activation patterns for ultra-fast NAS. In *International Conference on Learning Representations (ICLR)*, 2024.

Quan Minh Phan and Ngoc Hoang Luong. From hand-crafted metrics to evolved training-free performance predictors for neural architecture search via symbolic regression. *Neurocomputing*, pp. 132440, 2025.

Ilija Radosavovic, Justin Johnson, Saining Xie, Wan-Yen Lo, and Piotr Dollár. On network design spaces for visual recognition. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Hidenori Tanaka, Daniel Kunin, Daniel L. K. Yamins, and Surya Ganguli. Pruning neural networks without any data by iteratively conserving synaptic flow. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Ondrej Tybl and Lukas Neumann. Training-free neural architecture search through variance of knowledge of deep network weights. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14881–14890. IEEE, 2025.

Zimian Wei, Peijie Dong, Zheng Hui, Anggeng Li, Lujun Li, Menglong Lu, Hengyue Pan, and Dongsheng Li. Auto-prox: Training-free vision transformer architecture search via automatic proxy discovery. In *AAAI Conference on Artificial Intelligence*, 2024.

Taojiannan Yang, Linjie Yang, Xiaojie Jin, and Chen Chen. Revisiting training-free NAS metrics: An efficient training-based method. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023.

Chris Ying, Aaron Klein, Eric Christiansen, Esteban Real, Kevin Murphy, and Frank Hutter. NAS-Bench-101: Towards reproducible neural architecture search. In *International Conference on Machine Learning (ICML)*, 2019.

Arber Zela, Julien Siems, Lucas Zimmer, Jovita Lukasik, Margret Keuper, and Frank Hutter. Surrogate NAS benchmarks: Going beyond the limited search spaces of tabular NAS benchmarks. In *International Conference on Learning Representations (ICLR)*, 2022.

A THEORY

This appendix states and proves the four results summarized in Section 7. The statements below also clarify which empirical estimators are and are not covered by these guarantees.

T1 (size carrier in extensive sums). For architecture a, let

$$S(a) = \sum_{r \in \mathcal{I}(a)} X_{ar}, \qquad c(a) = \sum_{r \in \mathcal{I}(a)} \mathbb{E}[X_{ar} \mid a],$$

and

$$\xi_{ar} = X_{ar} - \mathbb{E}[X_{ar} \mid a].$$

Assume the centered terms are conditionally independent within each architecture and that scoring randomness is independent across architectures. Suppose further that

$$\log \mathbb{E}[e^{u\xi_{ar}} \mid a] \le \frac{u^2 \nu_{ar}^2}{2(1 - b_{ar}|u|)} \quad (|u| < b_{ar}^{-1}). \tag{3}$$

Write

$$V(a) = \sum_{r \in \mathcal{I}(a)} \nu_{ar}^2, \qquad B(a) = \max_{r \in \mathcal{I}(a)} b_{ar},$$

with

$$V_{ab} = V(a) + V(b), \qquad B_{ab} = \max\{B(a), B(b)\}.$$

If $\Delta_{ab} = c(a) - c(b) > 0$, then

$$\Pr\{S(a) \le S(b) \mid a, b\} \le \exp\left[-\frac{\Delta_{ab}^2}{2(V_{ab} + B_{ab}\Delta_{ab})}\right]. \tag{4}$$

Define the conditional Kendall concordance over resolved carrier pairs as

$$\bar{\tau}_c(S) = \mathbb{E}[\operatorname{sgn}\{S(A) - S(B)\} \operatorname{sgn}\{c(A) - c(B)\} \mid c(A) \ne c(B)]\,. \tag{5}$$

Then

$$\bar{\tau}_c(S) \ge 1 - 2\mathbb{E}\left[\exp\left\{-\frac{|c(A) - c(B)|^2}{2[V_{AB} + B_{AB}|c(A) - c(B)|]}\right\} \middle| c(A) \ne c(B)\right]. \tag{6}$$

Consequently, along any sequence of architecture-pair distributions for which

$$\frac{|c(A) - c(B)|^2}{V_{AB} + B_{AB}|c(A) - c(B)|} \to \infty$$

in probability conditional on $c(A) \ne c(B)$, we have $\bar{\tau}_c(S) \to 1$.

Under Gaussian He initialization, for a layer with d_ℓ inputs and o_ℓ outputs,

$$W_{\ell,ij} \sim \mathcal{N}(0, 2/d_\ell),$$

so

$$\mathbb{E}|W_{\ell,ij}| = \frac{2}{\sqrt{\pi d_\ell}}$$

and therefore

$$\mathbb{E}\|W_\ell\|_1 = \frac{2}{\sqrt{\pi}} o_\ell \sqrt{d_\ell}. \tag{7}$$

This illustrates how an extensive layerwise statistic can inherit a leading size-dependent component.

Proof. Summing Equation 3 for the centered score difference gives a Bernstein moment-generating-function bound with parameters (V_{ab}, B_{ab}). Chernoff's method with $u = \Delta_{ab}/(V_{ab} + B_{ab}\Delta_{ab})$ gives Equation 4. A resolved pair contributes at least one minus twice its misordering probability to the conditional Kendall concordance, so averaging the pairwise bound gives Equation 6. Under the stated margin-to-noise condition, the bounded exponential term converges to zero in probability, and hence its expectation converges to zero.

Finally, under Gaussian He initialization, $W_{\ell,ij} \sim \mathcal{N}(0, 2/d_\ell)$, so $\mathbb{E}|W_{\ell,ij}| = 2/\sqrt{\pi d_\ell}$. Multiplying by the $o_\ell d_\ell$ entries yields Equation 7. □

T2 (adaptive Occam radius). Let $O_1, \ldots, O_n$ be i.i.d. For each program f in the predeclared grammar, let $h_f : \mathcal{O} \times \mathcal{O} \to [-1, 1]$ denote its fixed Kendall pair kernel, let

$$\widehat{\tau}_f = \binom{n}{2}^{-1} \sum_{i<j} h_f(O_i, O_j), \qquad \tau_f = \mathbb{E}[h_f(O, O')],$$

where O and O' are independent draws, and let the grammar assign f a prefix-free code of length $L(f)$ bits. Then, for every $\delta \in (0, 1)$, with probability at least $1-\delta$, simultaneously for all programs,

$$|\widehat{\tau}_f - \tau_f| \le \min\left\{2, \sqrt{\frac{8}{n}\left[L(f)\log 2 + \log\frac{2}{\delta}\right]}\right\}. \tag{8}$$

For fixed cell memberships, let

$$W = \sum_g \binom{n_g}{2}, \qquad T_f = W^{-1} \sum_g \sum_{i<j\in g} h_f(O_i, O_j).$$

Conditional on the fixed cells, provided the observations remain independent under this conditioning, with probability at least $1 - \delta$, simultaneously for all programs,

$$|T_f - \mathbb{E}[T_f \mid G]| \le \sqrt{\frac{1}{2}\sum_i c_i^2 \left[L(f)\log 2 + \log\frac{2}{\delta}\right]}, \qquad c_i = \frac{2(n_{g(i)} - 1)}{W}. \tag{9}$$

The NAS-Bench-201 calculation below was originally intended as a certificate. However, it evaluates the complete fixed benchmark rather than an i.i.d. sample, so the sampling assumptions of T2 do not apply. We retain the numerical subtraction only as a description-length diagnostic, without a confidence-coverage interpretation.

Proof. Replacing one observation changes raw τ_a by at most $4/n$. McDiarmid's inequality therefore gives

$$\Pr(|\widehat{\tau}_f - \tau_f| > t) \le 2e^{-nt^2/8}$$

for fixed f. Assigning failure probability $\delta 2^{-L(f)}$ to program f and applying a union bound over the grammar, together with Kraft's inequality $\sum_f 2^{-L(f)} \le 1$, gives Equation 8 uniformly over the grammar.

For the fixed-cell statistic, replacing observation i changes $n_{g(i)} - 1$ pair kernels by at most two each, so its bounded-difference constant is $c_i = 2(n_{g(i)} - 1)/W$. The same McDiarmid and prefix-code argument gives Equation 9. These bounds do not cover same-sample T_{res}, because its rank transforms and fitted projections change globally when the sample changes. Residual certification therefore requires independent calibration data on which those maps are frozen. □

T3 (linear shrinkage diagnostic). Let S, Q, and Z be centered, and let P_Z denote the population L_2 orthogonal projection onto the linear span of Z. Write

$$X = P_Z S, \qquad U = S - X, \qquad Y = P_Z Q, \qquad V = Q - Y.$$

Assume $v_x = \mathrm{Var}(X) > 0$, $v_u = \mathrm{Var}(U) > 0$, and $v_Q = \mathrm{Var}(Q) > 0$, and define

$$c_z = \mathrm{Cov}(X, Y), \qquad c_r = \mathrm{Cov}(U, V).$$

Retaining a fraction $t = 1 - \lambda \in [0, 1]$ of the projected component gives

$$\rho(t) = \frac{c_r + t c_z}{\sqrt{v_Q (v_u + t^2 v_x)}}, \qquad \rho'(t) = \frac{c_z v_u - c_r v_x t}{\sqrt{v_Q}(v_u + t^2 v_x)^{3/2}}. \tag{10}$$

The global maximum over $t \in [0, 1]$ is attained at an endpoint or an interior stationary point. If $c_r > 0$ and $c_z \geq 0$, the unique maximizer is

$$t^* = \mathrm{clip}_{[0,1]} \left(\frac{c_z v_u}{c_r v_x} \right), \qquad \lambda^* = 1 - t^*. \tag{11}$$

Proof. Projection orthogonality gives $\mathrm{Cov}(U, X) = 0$, $\mathrm{Cov}(U, Y) = 0$, and $\mathrm{Cov}(X, V) = 0$. Hence

$$\mathrm{Cov}(U + tX, Q) = c_r + t c_z$$

and

$$\mathrm{Var}(U + tX) = v_u + t^2 v_x,$$

which gives $\rho(t)$ in Equation 10. Differentiation yields the displayed derivative. Its sign is determined by $c_z v_u - c_r v_x t$. Thus an interior stationary point, when present, is at $c_z v_u / (c_r v_x)$, and otherwise the maximum lies at an endpoint. For $c_r > 0$ and $c_z \geq 0$, clipping this stationary point to $[0, 1]$ gives the unique maximizer in Equation 11. □

This result concerns population Pearson correlation. Interpreting λ^* for Kendall correlation requires an additional Gaussian-copula approximation, so λ^* should be treated as a diagnostic rather than a Kendall rank-optimality result.

T4 (noise ceiling). Let A and A' be independent draws from the architecture population, and let

$$Y_s(A) = \mu(A) + \varepsilon_s(A).$$

For a pair (A, A'), define

$$\eta_s(A, A') = \varepsilon_s(A) - \varepsilon_s(A').$$

Conditional on (A, A'), assume that the pair noises $\eta_s(A, A')$ are identically distributed and independent across seeds and have distributions symmetric about zero. Let $T(A)$ be a score that is independent of the evaluation-seed noise conditional on the architecture. Define

$$q(A, A') = \left| \mathbb{E}[\mathrm{sgn}\{Y_s(A) - Y_s(A')\} \mid A, A'] \right|,$$
$$C_\tau = \mathbb{E}[q(A, A')],$$

and, for independent seeds $s \neq t$,

$$R = \tau_a(Y_s, Y_t).$$

Then

$$|\tau_a(T, Y_s)| \leq C_\tau, \qquad R = \mathbb{E}[q^2], \qquad R \leq C_\tau \leq \sqrt{R}. \tag{12}$$

The interval $[R, \sqrt{R}]$ is sharp as a moment bound for C_τ over $q \in [0, 1]$.

Proof. Let

$$m(A, A') = \mathbb{E}[\mathrm{sgn}\{Y_s(A) - Y_s(A')\} \mid A, A'].$$

By symmetry of the conditional pair-noise distribution, $m(A, A')$ has the same sign as $\mu(A) - \mu(A')$, with $|m(A, A')| = q(A, A')$. Therefore

$$\tau_a(T, Y_s) = \mathbb{E}[\mathrm{sgn}\{T(A) - T(A')\} m(A, A')],$$

and taking absolute values gives

$$|\tau_a(T, Y_s)| \leq \mathbb{E}[q] = C_\tau.$$

Table 3: Description-length-adjusted matched-Kendall diagnostic on NAS-Bench-201 (CIFAR-10, $n = 15{,}625$). The reference adjustment ϵ uses the T2 expression at $\delta = 0.05$. Adjusted values subtract ϵ from the observed matched τ_a. They are not certified confidence bounds for this exhaustive evaluation.

Champion	L (bits)	matched τ_a	reference ϵ	adjusted value	bar
seed 0	44	0.608	0.132	0.475	0.600
seed 1	37	0.581	0.123	0.459	0.600

Table 4: Noise-ceiling references $[R, \sqrt{R}]$ for raw and residual targets. NDS ships one seed per architecture, so its replicates are converged-epoch jitter and the resulting ranges are used as heuristic ceiling references rather than direct applications of T4.

Benchmark	Replicates	raw ceiling	residual ceiling
NDS DARTS	epoch jitter	[0.930, 0.965]	[0.898, 0.948]
NDS AmoebaNet	epoch jitter	[0.926, 0.962]	[0.918, 0.958]
NDS ENAS	epoch jitter	[0.931, 0.965]	[0.906, 0.952]
NDS PNAS	epoch jitter	[0.913, 0.956]	[0.888, 0.943]
NDS NASNet	epoch jitter	[0.935, 0.967]	[0.925, 0.962]
NB-201 CIFAR-10	3 seeds	[0.906, 0.952]	[0.890, 0.943]
NB-201 CIFAR-100	3 seeds	[0.896, 0.947]	[0.868, 0.931]
NB-201 ImageNet16-120	3 seeds	[0.912, 0.955]	[0.888, 0.942]

Equality is attained by any score whose pairwise ordering agrees with that of μ almost surely.

For independent seeds $s \neq t$, conditional independence gives

$$R = \mathbb{E}\big[m(A, A')^2\big] = \mathbb{E}[q^2].$$

Since $0 \leq q \leq 1$, $q^2 \leq q$ gives $R \leq C_\tau$, while Cauchy–Schwarz gives $C_\tau^2 \leq R$ and hence $C_\tau \leq \sqrt{R}$. The lower endpoint is attained when q is Bernoulli with mean R, and the upper endpoint when $q = \sqrt{R}$ almost surely. □

The formal result requires independent training-seed replicates. The NDS references use converged-epoch jitter instead and are therefore heuristic upper-bound estimates rather than direct applications of T4.

Table 3 reports the finite-benchmark diagnostic, while Table 4 reports the noise ceilings used in Sections 3 and 6. The separate registered C2 comparison could not exceed its stated bar because C2 is $-\operatorname{mean}_\ell \|\nabla W_\ell\|_1$, whereas the incumbent roster already contains $+\operatorname{mean}_\ell \|\nabla W_\ell\|_1$ and the bar uses absolute association. Under this comparison, C2 is therefore equivalent in magnitude to an existing incumbent. We report this as a design flaw. It does not affect the empirical max-aware comparison of ϕ_{frz}.

B Audit Details

Table 5 reports the NAS-Bench-201 audit, and Table 6 reports the raw NDS leaderboard behind Figure 2. The audited inventory comprises 24 distinct proxy scores. These include the 18 score implementations in Table 5, the Plain gradient proxy scored on the sealed spaces, and five additional statistic baselines from the strengthened bar. Two robustness details complement the main audit. First, relative to the two-axis control, the audit rankings have Spearman correlations of approximately 0.975–0.977 under the parameter-only and FLOPs-only controls, while adding convolution-edge count gives $\rho \approx 0.845$. Second, a separate depth analysis on NAS-Bench-201 shows especially strong coupling for a few proxies. Entropic Score has $\tau = 0.785$ against depth, compared with 0.317 for the original Zen score, and adding depth to the control set lowers its residual association from 0.600 to 0.334.

C Size-Control Analysis Comparison

Estimator details. Equation 1 is implemented using average ties, Van der Waerden scores $\Phi^{-1}(r/(n+1))$, an intercept in the OLS design, and the SciPy implementation of Kendall τ_b on the

Table 5: NAS-Bench-201 audit on CIFAR-10 (NAS-Bench-201 roster, sorted by $|T_{\text{res}}|$). $M(S)$ is the size-explained association ratio. The final column is the legacy parameter-only development-space diagnostic, not the main two-dimensional stratified analysis. Both the original and recomputed Zen scores are shown because public implementations disagree.

Proxy	raw τ_b	residual τ_b	$M(S)$	param.-only strat. τ
MeCo	0.742	0.710	0.22	0.517
NEAR	0.709	0.673	0.36	0.478
LogSynflow	0.613	0.608	0.21	0.182
Entropic	0.621	0.600	0.37	0.331
ZiCo	0.590	0.560	0.31	0.170
SWAP	0.548	0.485	0.34	0.278
ℓ_2 norm	0.510	0.458	0.51	0.008
SWAP-reg	0.463	0.400	0.58	0.278
NASWOT	0.464	0.371	0.52	-0.025
Zen (original)	0.265	0.278	0.15	0.025
JacCov	0.330	0.213	0.49	0.062
EPE-NAS	0.306	0.181	0.54	0.064
GradNorm	0.293	0.140	0.62	-0.049
SNIP	0.293	0.138	0.63	-0.040
GraSP	0.260	0.134	0.60	-0.009
Fisher	0.263	0.125	0.61	-0.019
SynFlow	0.281	0.095	0.80	-0.068
Zen (recomputed)	0.114	-0.007	1.25	-0.127

Table 6: Raw Kendall τ_b on the five sealed NDS spaces for the 15-baseline roster, with ϕ_{frz} shown for comparison. Bold marks the baseline with the largest absolute raw association among the 15 baselines in each space.

Proxy	DARTS	AmoebaNet	ENAS	PNAS	NASNet
ℓ_1 sum	+0.480	+0.190	+0.369	+0.372	+0.251
ℓ_2 norm	+0.373	+0.035	+0.245	+0.251	+0.124
mean $\lvert\nabla W\rvert$	+0.240	-0.073	+0.087	+0.145	-0.064
H sum	+0.306	-0.059	+0.167	+0.158	+0.060
act entropy	+0.296	-0.036	+0.190	+0.163	+0.075
log $\#\theta$	**+0.494**	+0.240	**+0.411**	**+0.387**	**+0.289**
SNIP	+0.270	-0.085	+0.093	+0.149	-0.049
Plain	-0.072	-0.047	-0.180	-0.124	-0.220
GradNorm	+0.176	-0.131	+0.013	+0.081	-0.121
Fisher	+0.175	-0.146	+0.006	+0.078	-0.123
GraSP	-0.057	+0.044	+0.060	+0.005	+0.185
NASWOT	+0.464	+0.183	+0.371	+0.354	+0.277
JacCov	+0.201	**+0.260**	+0.156	+0.166	+0.128
ZiCo	+0.350	-0.016	+0.201	+0.200	+0.090
SynFlow	+0.298	-0.060	+0.131	+0.174	+0.018
ϕ_{frz}	+0.256	+0.329	+0.223	+0.254	+0.328

two fitted residual vectors. For the main two-dimensional stratified analysis, parameter quintiles are formed first and FLOPs quartiles are then formed within each parameter bin, giving at most 20 cells. Cells with $n_g < 3$ are omitted and the remaining $\tau_{a,g}$ values receive pair-count weights $\binom{n_g}{2}$. This use of τ_a is specific to the fixed-stratum pair-concordance statistic. The raw and residual analyses use Kendall τ_b. The "strat." column of Table 5 is an earlier NAS-Bench-201 development-space diagnostic. It uses parameter-only groups, omits groups smaller than ten, and weights by group size. It is not the main two-dimensional analysis and is not used for sealed judgments.

To make the interpretation of $M(S)$ explicit, let $x = \mathcal{N}_n(s) - \overline{\mathcal{N}_n(s)}$ and $y = \mathcal{N}_n(y) - \overline{\mathcal{N}_n(y)}$, let $r_x = (I - P_Z)\mathcal{N}_n(s)$ and $r_y = (I - P_Z)\mathcal{N}_n(y)$, and set $z_x = x - r_x$ and $z_y = y - r_y$. Orthogonality gives

$$\rho = \frac{\langle x, y\rangle}{\|x\|\|y\|} = \underbrace{\frac{\langle z_x, z_y\rangle}{\|x\|\|y\|}}_{\rho_Z} + \underbrace{\frac{\langle r_x, r_y\rangle}{\|x\|\|y\|}}_{\rho_\perp}, \qquad M(S) = \frac{\rho_Z}{\rho}. \tag{13}$$

Thus $M(S) > 1$ is possible when ρ_Z and $\rho_\perp$ have opposite signs and partially cancel. It must not be read as a causal fraction of performance explained by size.

Population and resampling details. The target-finite benchmark population first removes only non-finite targets and fixes the favorable direction of each task metric. Parameter quintiles and within-quintile FLOPs quartiles are constructed on that population. For the main strengthened-bar NDS analysis, rows are then restricted once to the intersection where ϕ_{frz} and all members of $\mathcal{B}_{15}$ are finite. Normal scores and residual regressions are computed on this common-finite set. The cell labels constructed before the restriction are carried into the common-finite set.

The max-aware interval for Equation 2 uses paired architecture-row resampling, with a single index vector applied to the target, controls, ϕ_{frz}, and every baseline. In each replicate the normal-score transforms and residual projections are refitted, while the precomputed cell labels remain attached to their rows. The strongest baseline is re-selected within each replicate for both analyses. The registered calculation uses 600 BCa replicates, 100 grouped jackknife blocks for acceleration, and documented per-space and per-analysis random streams. A pairwise-row point-estimate sensitivity uses a separate finite intersection for each candidate–baseline pair, but it is not the main estimand.

Target extraction and orientation. The target contract is fixed at the loader boundary. NAS-Bench-201 uses API `test_accuracy` at 200 epochs averaged over its training seeds, and NATS-Bench SSS uses the seed-averaged test accuracy at `hp=90`. NAS-Bench-101 and NAS-Bench-301 use the Suite-Zero `val_accuracy` field, with the latter interpreted as surrogate validation accuracy. For TransNAS-Bench-101, per-epoch arrays contribute their final entry: `valid_top1` for scene classification, object classification, and jigsaw, `valid_neg_loss` for room layout, `valid_mIoU` for semantic segmentation, and `valid_ssim` for surface-normal prediction and autoencoding. NDS accuracy is $100 - \texttt{min_test_top1}$. These fields are already oriented so that larger is favorable. Size controls come from the benchmark parameter and FLOP fields (`params`/`flops` for NDS and `model_params`/`model_FLOPs` for TransNAS), and the released manifests record the exact input hashes and row order.

Agreement between the control analyses. Across the 152 combinations of 19 evaluation slices and eight residual-eligible scores (C1–C3 plus the five non-parameter opening-roster statistics), the residual and two-dimensional stratified analyses have opposite signs with both magnitudes exceeding 0.05 only for C2 and its sign-reversed incumbent on micro room layout, and for `l2_norm_bank` on micro semantic segmentation. In all three cases both magnitudes are below 0.14. Additional sign changes occur only when at least one magnitude is at most 0.05. By contrast, parameter-only stratification can depart visibly from both size-control analyses on the macro tasks (Figure 4). It is therefore not used for macro judgments under the prespecified control scheme.

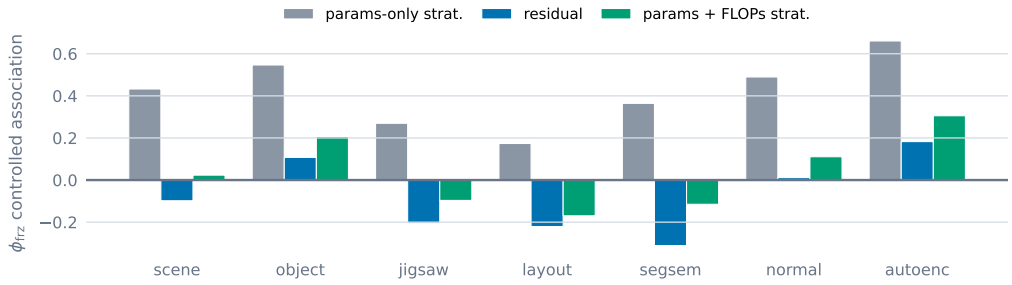

Figure 4: ϕ_{frz} on the seven TransNAS-Bench-101 macro tasks under the parameter-only, residual, and two-dimensional stratified analyses. Parameter-only stratification agrees in sign with the raw association on all seven tasks. The residual and two-dimensional stratified analyses agree in sign on six tasks and both have the opposite sign from parameter-only stratification on three.

D Discovery Engine Details

The statistic bank holds nine per-layer statistics (mean, absolute mean, standard deviation, ℓ_1 mass, squared ℓ_2 mass, maximum absolute value, positive fraction, element count, and normalized en-

tropy) for each of the weight, gradient, and activation tensors. The grammar composes bank entries with elementwise products, powers, logs, and layer aggregations, and every program carries a prefix code length. The initial searches use a population of 40 for 25 generations (940 evaluations per seed), and the refinement run uses 64 by 60 (3,664 evaluations). For a program p with node count $|p|$, the initial-search fitness is

$$F_\lambda(p) = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} |T_{\text{res}}(p; d)| - \lambda |p| \tag{14}$$

evaluated on each full architecture population rather than a noisy subsample. The refined NAS-Bench-201 champion has the normal form $\sum_l \|W_l\|_1 \cdot H(|\nabla W_l|)^4$.

Across the three NAS-Bench-201 datasets, the residual associations of C1 are 0.719, 0.695, and 0.669, compared with 0.710, 0.678, and 0.644 for MeCo on CIFAR-10, CIFAR-100, and ImageNet16-120, respectively. On the registered NATS-Bench-SSS transfer gate, C1 gives residual associations of 0.453, 0.086, and 0.303, with corresponding $M(S)$ values of 0.947, 0.985, and 0.961.

For reproducibility, the entropy statistic used in candidate C3 is max-normalized elementwise entropy. For a tensor x with N elements,

$$u_k = \frac{|x_k|}{\max_j |x_j|}, \qquad H(x) = -\frac{1}{N} \sum_{k=1}^{N} u_k \log(u_k + 10^{-9}).$$

When $\max_j |x_j| = 0$, we define $H(x) = 0$. For the statistic baselines in Figure 2 and Table 6, "H sum" denotes $\sum_\ell H(\nabla W_\ell)$, while "act entropy" denotes $\sum_\ell H(a_\ell)$ for the post-forward activation tensors.

Statistics are collected from `Conv2d` and `Linear` layers in forward order. For the NAS-Bench and NDS evaluations, activations are post-forward outputs, and weight gradients are collected after cross-entropy backpropagation in training mode using a batch of 32 Gaussian $3 \times 32 \times 32$ inputs and random 10-class labels. TransNAS-Bench-101 uses the registered $3 \times 256 \times 256$ input resolution. Initialization and inputs are regenerated from the same architecture-level seed, using seed 42 for the primary evaluation and seeds 43 and 777 in the seed campaign.

The GP grammar uses numerically protected log and division with 10^{-12} during search. Before NDS access, candidate C3 was frozen as the literal score

$$\phi_{\text{frz}} = \sum_l \log \frac{H(a_l)}{\max |\nabla W_l|}.$$

The frozen score is evaluated without those protected operators. Consequently, a zero activation-entropy term or zero gradient maximum produces a non-finite layer contribution, which gives the finite-domain behavior analyzed in Appendix K.

In the registered fitness-swap comparison, none of the five raw-fitness runs drifted toward a size proxy, and their size-explained ratios remained below 0.43. The gradient-entropy family also attained the strongest raw association within the searched primitive space. An exploratory rerun in the EZNAS regime, with fitness computed on fixed subsamples of 20 architectures, showed optimism of at most 0.07. These results support treating the risk of raw-fitness discovery as conditional rather than universal.

Cross-space primitive behavior. Figure 5 provides primitive-level context for the space-dependent discovery outcomes summarized in Finding 2. Representative bank statistics need not retain the same residual-association sign across search spaces and tasks. The corresponding NAS-Bench-201 and NAS-Bench-301 discovery results are reported in Section 5.

Post-hoc search diagnostics. We froze the historical anchor (population 40, elite 4, parent pool 8, mutation probability .6, initial depth 5, $\lambda = .002$), the 17-scenario sensitivity grid, and the 940-evaluation accounting rule before executing these post-hoc diagnostics. Every typed proposal consumes one evaluation, including duplicates. The legacy checkpoints record exactly 940 evaluations but predate proposal-level unique/duplicate provenance.

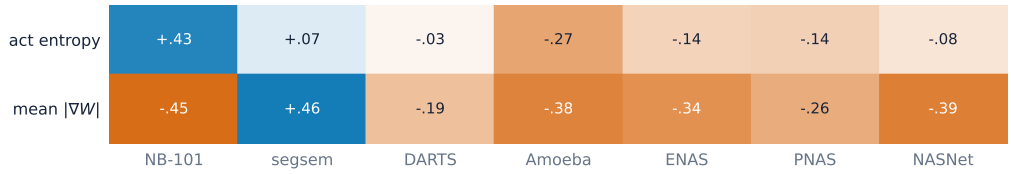

Figure 5: Residual associations of two representative bank statistics across the shown evaluation slices. Activation entropy is positive on NAS-Bench-101, slightly positive on TransNAS semantic segmentation, and near zero to negative across NDS. Mean gradient magnitude is negative on every shown slice except TransNAS semantic segmentation, where it is strongly positive.

Table 7: GP sensitivity over five seeds per setting. Entries report the mean best penalized fitness ± the standard deviation across seeds. The population settings 20 and 80 are the two protocol-fixed population-schedule settings rather than an additional duplicated factor.

Factor	Low setting	High setting
Population	20 (.6438 ± .0274)	80 (.6487 ± .0247)
Mutation probability	.2 (.6393 ± .0232)	.8 (.6412 ± .0246)
Elite count	1 (.6482 ± .0241)	8 (.6446 ± .0285)
Parent pool	4 (.6576 ± .0246)	16 (.6577 ± .0248)
Initial depth	3 (.6368 ± .0225)	7 (.6461 ± .0199)
Enforced depth cap	3 (.6306 ± .0152)	5/7 (.6468 ± .0268 / .6576 ± .0246)

The anchor is $.6470 \pm .0267$. Across the 17 protocol-fixed scenarios, the mean best penalized fitness ranges from .6306 to .6620. Penalty settings require a separate decomposition because they change the objective itself. For $\lambda = 0, .001, .002, .005$, respectively, the unpenalized mean absolute residual correlations are .6620, .6657, .6582, .6550, the mean node counts are 36.6, 7.2, 5.6, 4.4, and the penalized fitnesses are .6620, .6585, .6470, .6330. Thus the penalty substantially reduces expression size while the mean absolute residual correlation remains between .6550 and .6657.

Fifty independent random-search runs used the same grammar, validity, penalty, duplicate accounting, and 940-evaluation budget. Their mean best fitness is $.6294 \pm .0118$, compared with $.6470 \pm .0267$ for the five anchor GP seeds. Two of five GP seeds exceed the random-search 95th percentile (.6504). We treat this only as a descriptive equal-budget comparison.

For a separate adaptive-search null, the proxy columns and their missingness are fixed while accuracy is permuted within the frozen 5×4 parameter–FLOPs cells. Each of 99 null datasets reruns the complete GP search under five seeds and 940 evaluations. With T defined as the mean of the five run maxima, the observed $T_{\text{obs}} = .6470$, whereas the 99 null statistics have mean $.2836 \pm .0019$ and range $[.2788, .2872]$. None reaches the observed value, giving the finite-randomization estimate

$$p_{\text{emp}} = \frac{1 + \sum_{k=1}^{99} \mathbb{1}\{T_k^{\text{null}} \geq T_{\text{obs}}\}}{99 + 1} = .01. \tag{15}$$

This is evidence against that post-hoc but execution-time-fixed conditional permutation null. It is not a general test of GP superiority over alternative search procedures or an estimate of the probability that the exact C3 expression is rediscovered.

E Objective-Isolation Ablation on NAS-Bench-101

This post-opening auxiliary ablation isolates the association used as GP fitness on the fixed Suite-Zero NAS-Bench-101 population. Its protocol and analysis plan were fixed before execution, after the NDS suite had been opened. It therefore does not revise the original registration or provide an additional held-out test.

Design. We ran 20 paired seeds (1000–1019), with one raw-fitness and one controlled-fitness run per seed. Within each pair, the two arms used the same 40-program initial population, proposal-

engine seed, grammar, complexity penalty, and evaluation budget. Execution order was counterbalanced, with even seeds running the controlled arm first and odd seeds running the raw arm first. Each run used a population of 40, 25 generations, elite size 4, parent pool 8, mutation probability 0.6, crossover probability 0.4, depth 5, and exactly 940 program evaluations. The two objectives were

$$F_{\mathrm{raw}}(p) = |T_{\mathrm{raw}}(s_p, y)| - 0.002|p|, \tag{16}$$

$$F_{\mathrm{ctrl}}(p) = |T_{\mathrm{res}}(s_p, y; Z)| - 0.002|p|. \tag{17}$$

Pairing fixes the stochastic starting condition rather than the subsequent candidate sequence. Once the objectives induce different parent rankings, their search trajectories may diverge by design. All 40 runs completed 940 evaluations, and the manifest and input hashes matched the registered experiment specification.

For seed pair r, let $\widehat{f}_r^{\mathrm{raw}}$ and $\widehat{f}_r^{\mathrm{ctrl}}$ denote the two final champions. The primary comparison re-evaluates both champions on their pair-specific common-finite support $\mathcal{I}_r$ and computes

$$\Delta_r = \left|T_{\mathrm{res}}(\widehat{f}_r^{\mathrm{ctrl}}, y; Z; \mathcal{I}_r)\right| - \left|T_{\mathrm{res}}(\widehat{f}_r^{\mathrm{raw}}, y; Z; \mathcal{I}_r)\right|. \tag{18}$$

All 20 common-finite supports contained the complete set of 9,781 architectures. The prespecified 95% interval is obtained from a 20,000-replicate paired BCa bootstrap over seed pairs, using RNG seed 314159 and leave-one-pair-out acceleration. The resulting uncertainty reflects GP search randomness on this fixed population, not sampling over architectures or new search spaces.

Table 8: Primary NAS-Bench-101 objective-isolation result. The paired difference is the absolute residual association of the controlled-fitness champion minus that of the raw-fitness champion under the common residual evaluation in Equation 18.

Quantity	Result
Mean paired difference	$+0.04056$
Median paired difference	$+0.03915$
SD of paired differences	0.02694
Paired BCa 95% interval	$[0.03047, 0.05384]$
Controlled-arm wins	$19/20$
Wilson 95% interval for the win fraction	$[0.76387, 0.99112]$
Exact two-sided sign-flip p-value	3.81×10^{-6}

The controlled arm has the larger absolute residual association in 19 of the 20 seed pairs, and the lower endpoint of the primary interval is above zero. Thus, under the fixed NAS-Bench-101 population, grammar, budget, penalty, and paired starting conditions, the controlled fitness objective selects programs with higher absolute residual association than the raw fitness objective. The exact sign-flip test is a secondary sensitivity analysis on the paired differences. It is not a binomial sign test and does not replace the prespecified interval criterion.

Table 9: Champion distributions in the NAS-Bench-101 objective-isolation ablation. Associations and node counts are means $\pm$ standard deviations across the 20 runs in each arm.

Metric	Raw fitness	Controlled fitness
$\lvert T_{\mathrm{res}}\rvert$	0.49214 ± 0.02567	0.53271 ± 0.01375
$\lvert T_{\mathrm{raw}}\rvert$	0.53873 ± 0.02032	0.26723 ± 0.05611
$\lvert T_{\mathrm{strat}}\rvert$	0.45614 ± 0.02018	0.50759 ± 0.01309
Program nodes	4.05 ± 1.19	5.45 ± 2.31
Exact C3 rediscovery	$0/20$	$0/20$

Both arms exceed the existing NAS-Bench-101 descriptive bar of 0.349 in all 20 runs. This in-sample comparison is descriptive rather than an adaptive generalization certificate. Neither arm reproduces the exact serialized C3 expression in these new runs. The ablation therefore supports an effect of the controlled objective on the absolute residual association of the selected programs, not the uniqueness or seed-level rediscovery stability of C3.

Scope. Both arms use the NAS-Bench-101 target during adaptive search and are evaluated on the same population. The absolute associations of the selected champions therefore retain adaptive

selection optimism. The paired comparison isolates the objective under the stated search conditions, but it does not establish universal superiority of controlled fitness, generalization beyond NAS-Bench-101, or an independent validation of ϕ_{frz}.

External precision extension. A subsequent post-hoc extension increased the matched experiment to 100 GP seed pairs and evaluated the resulting champions on the five NDS spaces. The controlled-fitness champions have higher external absolute residual association in all five spaces, and all five pointwise paired BCa intervals have positive lower endpoints (Table 10). This supports a consistent external objective effect under the evaluated NDS settings, not universal superiority of controlled fitness across discovery grammars or domains.

Table 10: Post-hoc external evaluation of the 100-pair objective-isolation extension. For each paired GP seed, the raw- and controlled-fitness champions are evaluated on their pair-specific common-finite support. The reported contrast is $\Delta_{\mathrm{ext}} = |T_{\mathrm{res}}(\widehat{f}^{\mathrm{ctrl}})| - |T_{\mathrm{res}}(\widehat{f}^{\mathrm{raw}})|$. Intervals are paired BCa 95% intervals over the 100 seed pairs and are pointwise across spaces.

Space	Δ_{ext}	BCa 95% interval
DARTS	$+0.0254$	$[+0.0043, +0.0468]$
AmoebaNet	$+0.0475$	$[+0.0171, +0.0768]$
ENAS	$+0.0575$	$[+0.0337, +0.0824]$
PNAS	$+0.0386$	$[+0.0202, +0.0582]$
NASNet	$+0.0302$	$[+0.0062, +0.0560]$

F Candidate Component and Aggregation Sensitivity

The frozen score can be decomposed as

$$C3_{\mathrm{sum}} \equiv \phi_{\mathrm{frz}} = A_{\mathrm{sum}} + G_{\mathrm{sum}}, \qquad A_{\mathrm{sum}} = \sum_l \log H(a_l), \qquad G_{\mathrm{sum}} = -\sum_l \log \max |\nabla W_l|.$$

For an architecture with L_a included layers, define

$$C3_{\mathrm{mean}} = \frac{C3_{\mathrm{sum}}}{L_a}, \qquad G_{\mathrm{mean}} = \frac{G_{\mathrm{sum}}}{L_a}.$$

Table 11 compares the frozen score with its exact gradient-only component under the same layer-wise aggregation and pair-finite support. Including the activation-entropy term yields higher T_{res} in all five spaces by 0.058–0.146, with all five paired BCa intervals above zero. This incremental association largely disappears under layer-mean aggregation. The difference between $C3_{\mathrm{mean}}$ and G_{mean} is approximately -0.001 to $+0.007$. After linear projection on log parameters and log FLOPs, the entropy sum is almost collinear with the included-layer count, and the margins of ϕ_{frz} remain positive when the layer count is added as a control (Appendix M.2).

Table 11: Post-hoc component and aggregation ablation on the five NDS spaces. The exact gradient-only counterpart is $G_{\mathrm{sum}} = -\sum_l \log \max |\nabla W_l|$ and uses the same layers and pair-finite support as $C3_{\mathrm{sum}} = \phi_{\mathrm{frz}}$. Component directions are fixed on the NAS-Bench-101 development population. The contrast $\Delta_{\mathrm{ent}} = T_{\mathrm{res}}(C3_{\mathrm{sum}}) - T_{\mathrm{res}}(G_{\mathrm{sum}})$ is followed by its paired BCa 95% interval from 600 row-bootstrap replicates. The final two columns report the corresponding layer-mean variants.

Space	$T_{\mathrm{res}}(C3_{\mathrm{sum}})$	$T_{\mathrm{res}}(G_{\mathrm{sum}})$	Δ_{ent} [BCa 95%]	$T_{\mathrm{res}}(C3_{\mathrm{mean}})$	$T_{\mathrm{res}}(G_{\mathrm{mean}})$
DARTS	0.369	0.244	$+0.125$ [0.111, 0.142]	0.219	0.219
AmoebaNet	0.467	0.320	$+0.146$ [0.132, 0.160]	0.434	0.433
ENAS	0.455	0.314	$+0.141$ [0.126, 0.157]	0.380	0.374
PNAS	0.355	0.230	$+0.125$ [0.110, 0.140]	0.282	0.283
NASNet	0.473	0.415	$+0.058$ [0.042, 0.072]	0.403	0.403

In the corresponding shared-candidate budget experiment, $C3_{\mathrm{sum}}$ has a higher point estimate than the strongest single-component alternative in 69 of 75 evaluated space–pool–budget conditions, with a positive interval lower endpoint in 63 of 75. The exceptions are concentrated in several small-budget ENAS conditions and a small number of DARTS conditions. Thus the entropy term provides

incremental controlled association under the frozen sum, but does not improve selection in every budget condition.

On the finite domain, rescaling the loss by a positive constant c changes the sum as $C3_{\mathrm{sum}}(c) = C3_{\mathrm{sum}} - L_a \log c$. It can therefore change rankings when L_a varies across architectures. By contrast, $C3_{\mathrm{mean}}(c) = C3_{\mathrm{mean}} - \log c$, so its ranking is invariant. We additionally verified the corresponding ranking behavior numerically for all 35 combinations of five spaces and seven fixed scales from 10^{-3} to 10^{3}. The loss-scale convention of the frozen sum is therefore part of the method definition.

G Additional Size-Control Sensitivity

Table 12: Post-hoc sensitivity to richer measured size controls on the five NDS spaces. Extended linear T_{res} additionally controls for configuration depth, configuration width, and the number of StatBank layers. The next two columns report Kendall τ_b between cross-fitted score and accuracy residuals from the fixed spline and histogram-gradient models. The final column gives the range of the matched pair-concordance statistic τ_a across calipers 0.02, 0.05, and 0.10, both with and without exact depth–width matching. These are sensitivity point estimates under different estimands rather than directly comparable confidence intervals.

Space	Extended linear T_{res}	Spline cross-fit	Histogram cross-fit	Matched τ_a range
DARTS	0.337	0.167	0.194	0.221–0.321
AmoebaNet	0.413	0.275	0.303	0.394–0.458
ENAS	0.360	0.079	0.139	0.355–0.432
PNAS	0.347	0.092	0.106	0.271–0.338
NASNet	0.435	0.345	0.349	0.388–0.481

The main residual analysis removes linear normal-score components of log parameter count and log FLOPs. Post-hoc analyses probe whether the remaining association persists after accounting for measured depth, width, included-layer count, and nonlinear size relationships. As summarized in Table 12, the association remains positive in every space after jointly adding all three measured axes, under two fixed five-fold cross-fitted nonlinear nuisance models, and across six local matching specifications. All 30 matching conditions yielded finite supports ranging from 1,193 to 2,037. These analyses report candidate associations only. The corresponding max-aware margins over the baseline roster are reported in Appendix M.2. All remain positive under the linear controls and local matching, although adjusted support is weaker for some matching settings and margins are not uniform under flexible nonlinear controls. None of these analyses establishes that every size-related confound has been removed. Controlling included-layer count also changes the estimand by removing variation that enters the definition of the layerwise sum.

H Sealing Registry

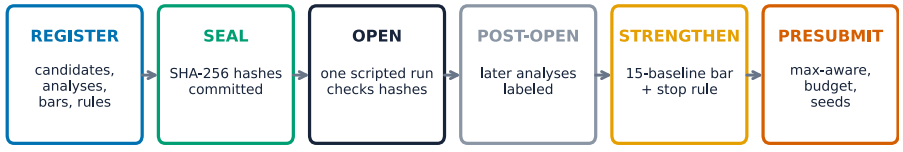

Figure 6: Registration and sealing timeline. Judgment rules and stopping conditions exist in version control before the results they govern.

Registration chronology. Figure 6 summarizes the chain. The registration commit freezing the candidates, size-control analyses, opening-stage incumbent bars, hypotheses H-S1 to H-S3, and

Table 13: H-S1 verdicts at opening. T_{res} of the frozen C1 score on the seven TransNAS-Bench-101 micro tasks, with Holm-corrected one-sided significance.

Task	T_{res}	Holm significant
class_scene	+0.319	yes
class_object	+0.207	yes
jigsaw	+0.205	yes
room_layout	+0.151	yes
segmentsemantic	+0.247	yes
normal	+0.316	yes
autoencoder	+0.126	yes

opening-stage stopping conditions predates the opening commit. H-S1 evaluates transfer of the frozen C1 signal to the sealed TransNAS micro tasks, H-S2 evaluates whether per-space GP discovery exceeds the registered space bar after the fixed correction for winner bias, and H-S3 concerns the registered shrinkage-correction criterion. The single scripted opening run verifies the SHA-256 hash of every sealed archive before reading any target.

The strengthened 15-baseline bar and its stopping rule (three or more lost spaces withdraw the five-space headline) were registered after opening and before computing any added baseline score. The registered max-aware retest, budget simulation, and seed campaign likewise predate their respective results. The later multiplicity, failure-as-worst, matched-coverage, and GP diagnostics are explicitly post-hoc.

The H-S1 family consists of the seven TransNAS micro tasks and uses Holm-corrected one-sided tests. The strengthened-bar stopping rule is based on the ten pointwise BCa intervals across five NDS spaces and two analyses. The later ten-claim Bonferroni calculation is explicitly post-hoc. The 21-score extension belongs to the registered seed campaign. For the later post-hoc diagnostics, machine-readable contracts, inputs, random streams, and interpretation limits were fixed before execution. None redefines an opening hypothesis or the 15-score strengthened bar.

The original version history, whose commit identifiers the supplement lists, supports two verifiable statements. The registration commits predate the commits that introduce the corresponding result files, and the archives read at opening match the hashes committed before opening. For the strengthened bar, which was registered after opening, the chain cannot show that the bar design was uninfluenced by the already observed opening numbers, and we do not claim otherwise. It shows only that the added baseline scores had not been computed when the bar definition, analysis rules, and stopping rule were fixed.

Opening candidate-set results. The scripted opening evaluated all three candidates frozen before NDS access. Table 14 reports their residual associations on the five NDS spaces. Because the focal designation followed this opening, the subsequent strengthened-bar analyses are interpreted as conditional on that designation rather than as an independent validation of candidate selection.

Table 14: Opening-stage residual Kendall τ_b associations on the five sealed NDS spaces for the three candidates frozen before NDS access.

Space	C1	C2	C3
DARTS	+0.140	+0.190	+0.369
AmoebaNet	-0.130	+0.380	+0.467
ENAS	+0.000	+0.340	+0.455
PNAS	-0.010	+0.260	+0.355
NASNet	+0.060	+0.390	+0.473

Opening target verification. The TransNAS target extraction was checked against the published target arrays in the NASLib NAS-Bench-Suite-Zero archive after matching all 4,096 micro architectures. Kendall τ_b is 1.0 with zero maximum absolute error for class scene, class object, jigsaw, room layout, normal prediction, and autoencoding. Semantic segmentation gives $\tau_b = .871$ because our registered target is `valid_mIoU`, whereas the generic `val_accuracy` field in the archive stores a different quantity. Using either target leaves the relevant candidate conclusions unchanged.

Opening hypothesis outcomes. Table 13 lists the H-S1 verdicts from the opening run. At opening, H-S1 produced residual correlations from 0.13 to 0.32 on the seven TransNAS micro tasks, all significant under its registered Holm correction. The C1 residual point estimate exceeded the contemporaneous incumbent point estimate on four of seven tasks.

H-S2, per-space discovery above the registered space bar after its fixed correction for winner bias, passed on two of the three registered slices and missed on the macro slice by -0.0024 (Table 15).

Table 15: H-S2 opening results. "Best" is the larger of the two registered 940-evaluation GP runs. "Corrected" subtracts the registered diagnostic adjustment for winner bias before comparison with the opening-roster bar. The verdict records the outcome of the registered diagnostic rule rather than confidence-certified superiority.

Slice	Best	Subtraction	Corrected	Bar	Verdict
micro/class_scene	.4522	.1213	.3309	.2872	pass
macro/class_scene	.2620	.1361	.1259	.1283	fail
NDS/DARTS	.4332	.1098	.3234	.1927	pass

I Shrinkage Correction and H-S3

The registered H-S3 test applies the T3 correction to `l2_norm_bank` with the coefficient fixed before opening at $\lambda = .40$. At opening, residual associations changed only slightly, the micro stratified statistic improved on six of seven tasks, and the macro stratified statistic increased on all seven tasks, mostly by becoming less negative. However, top-decile association worsened on every NDS space by .02–.05, so the registered Pareto criterion was not established. We therefore report the micro and macro stratified changes only as a scoped diagnostic rather than as a validated Pareto improvement.

Separately, fitting T3 to C1 on NAS-Bench-201 returns $\lambda^* = 0$. This is the H4-on-candidate diagnostic and is not H-S3. An exploratory C1 sweep illustrates the distinct trade-off. At $\lambda = .5$, global correlation rises from .744 to .766, while top-decile correlation falls from .412 to .384 and selection regret worsens from .42 to .61.

J Bar-Strengthening Registration

The strengthened bar is the maximum absolute controlled correlation over the union of the six frozen statistic baselines and the nine published proxies (SNIP, Plain, GradNorm, Fisher, GraSP, NASWOT, JacCov, ZiCo, and SynFlow), evaluated under both size-control analyses, with paired BCa intervals for the candidate–bar comparison. Under the registered aggregation rule, zero lost spaces retain the five-space headline. One or two lost spaces scale down the claim with the losses tabulated. Three or more lost spaces withdraw the headline and trigger redesign. The outcome was zero lost spaces.

K Missingness Analysis

ϕ_{frz} requires finite layerwise activation entropy and gradient statistics, and it is undefined for 9.9 to 26.8 percent of NDS architectures and 43.2 percent of TransNAS micro architectures. Dropped architectures tend to have lower accuracy, with rank-biserial correlations between accuracy and being dropped of -0.26 to -0.33 on NDS and -0.47 on micro.

We additionally ran three post-hoc support-policy checks on the five NDS spaces against the expanded 21-score roster. These checks comprise failure-as-worst association, matched coverage, and failure-as-worst budgeted search. After orienting every score so that larger is better, failure-as-worst replaces a non-finite score by a method-specific common value below its finite range.

$$s_{im}^{\text{FAW}} = \begin{cases} s_{im}, & s_{im} \in \mathbb{R}, \\ c_m, & \text{otherwise}, \end{cases} \qquad c_m < \min_{j : s_{jm} \in \mathbb{R}} s_{jm}. \tag{19}$$

Thus any choice of c_m below the finite range yields the same rank ordering, with all non-finite cases tied below the finite scores. The same rule is applied to every method, with no sign fitting.

Under both the residual and fixed 5×4 stratified analyses, the strongest eligible roster member is re-selected inside every bootstrap resample.

Table 16: Post-hoc NDS missingness diagnostics. n_f is the number of architectures with finite ϕ_{frz}. AUC is the ROC AUC obtained by using the finite indicator to discriminate above- from below-median accuracy, and r_{rb} is the rank-biserial association between accuracy and the dropped indicator. Negative r_{rb} values indicate lower accuracy among dropped architectures. The final columns report failure-as-worst max-aware margins over the expanded 21-score rival roster, with grouped-jackknife BCa 95% intervals from 600 bootstrap replicates.

Space	n_f/n	miss.	AUC	r_{rb}	D_{res}	D_{strat}
DARTS	3660/5000	26.8%	0.582	-0.327	+0.147 [0.129,0.167]	+0.147 [0.130,0.171]
AmoebaNet	4463/4983	10.4%	0.532	-0.257	+0.020 [0.005,0.034]	+0.018 [0.005,0.030]
ENAS	4506/4999	9.9%	0.540	-0.333	+0.057 [0.043,0.068]	+0.045 [0.033,0.060]
PNAS	4478/4999	10.4%	0.538	-0.300	+0.086 [0.069,0.104]	+0.084 [0.070,0.101]
NASNet	4322/4846	10.8%	0.535	-0.270	+0.031 [0.017,0.043]	+0.015 [0.001,0.029]

Every one of these ten pointwise intervals has a positive lower endpoint. Because the finite-indicator AUCs are only .532–.582, missingness is associated with lower accuracy but is not by itself a strong discriminator. In the matched-coverage check, each method is independently subsampled 1,000 times from its own finite rows to the finite count of ϕ_{frz}, or to the largest common feasible count if any rival has fewer finite rows. This is an unpaired sample-count sensitivity analysis. The strongest eligible roster member is re-selected in every repetition, but methods need not see the same architectures. Table 17 reports the resulting mean margins and percentile sampling intervals. All ten intervals are positive, and the win fraction of ϕ_{frz} is 1.0 in every analysis. The check isolates sample-count effects but does not remove informative-support bias.

Table 17: Post-hoc matched-coverage sensitivity against the expanded 21-score roster. Each entry reports the mean margin over 1,000 independent subsampling repetitions, followed by the 2.5th–97.5th percentile sampling interval. These intervals are empirical sampling intervals, not BCa intervals. The target sample size n equals the number of architectures with finite ϕ_{frz} in each space.

Space	n	D_{res}	D_{strat}
DARTS	3,660	$+0.163\,[0.153, 0.173]$	$+0.130\,[0.119, 0.139]$
AmoebaNet	4,463	$+0.067\,[0.062, 0.072]$	$+0.041\,[0.035, 0.047]$
ENAS	4,506	$+0.089\,[0.084, 0.093]$	$+0.056\,[0.051, 0.062]$
PNAS	4,478	$+0.093\,[0.088, 0.099]$	$+0.081\,[0.075, 0.087]$
NASNet	4,322	$+0.082\,[0.076, 0.087]$	$+0.053\,[0.047, 0.059]$

The same failure-as-worst rule was applied to retrospective search at budgets $10, 25, 50, 100, 200$ in the full and median-size-restricted pools. All methods share each candidate set. Maximum-score ties use a seeded uniform draw. The strongest baseline is the method with the largest mean selected accuracy and is re-selected in each paired bootstrap resample rather than chosen as a trial-wise oracle. Under the seed-42 score realization, ϕ_{frz} has a positive point margin in all 50 conditions, and all 50 pointwise percentile bootstrap intervals have positive lower endpoints. The 50 intervals are pointwise and receive no simultaneous or family-wise-error interpretation.

The coverage-completing variant, denoted ϕ_{mask}, uses the post-opening fixed rule of omitting non-finite layer terms. It covers all 19 evaluated NDS and TransNAS slices and exceeds the stratified bar on 14 of them. It was defined after opening and is therefore reported as exploratory only. Promoting it to a primary method would require a new untouched holdout with rules registered in advance.

The broader numerical-policy sensitivity in Table 18 shows that the frozen literal score is not invariant to how zero layer terms are treated. On the architectures where the literal score is finite, however, the protected search-time program, both ϵ variants, all clipping floors, and the layer mask induce exactly the same ranking as the literal score (Kendall $\tau = 1$). The policies differ only on architectures with non-finite literal terms, which the protected and ϵ variants score with large positive layer contributions (Appendix M.3). The complete-coverage layer-mask variant has a slightly higher residual association than the literal score in every space, although this comparison also reflects the different support.

Table 18: Post-hoc numerical-policy sensitivity on the five NDS spaces. Entries are residual Kendall associations on policy-specific finite supports. The protected row exactly reproduces the search-time protected operators. The ϵ rows evaluate $\sum_l \log[(H(a_l)+\epsilon)/(\max|\nabla W_l|+\epsilon)]$. The clipping row gives the range over development-fixed gradient floors at percentiles 0.1, 1, and 5. The layer-mask policy omits non-finite layer terms. Because supports differ, the rows are policy sensitivities rather than paired estimates on a common population. On the literal-finite architectures every row induces the same ranking as the literal score (Appendix M.3).

Policy	DARTS	AmoebaNet	ENAS	PNAS	NASNet
Literal	0.369	0.467	0.455	0.355	0.473
Protected search	0.189	0.415	0.386	0.286	0.406
$\epsilon = 10^{-12}$	0.173	0.415	0.386	0.286	0.406
$\epsilon = 10^{-8}$	0.207	0.422	0.399	0.297	0.420
Clipping range	0.280–0.292	0.436–0.439	0.421–0.425	0.316–0.320	0.439–0.442
Layer mask	0.402	0.471	0.462	0.361	0.474

Table 19: Registered budgeted search. Entries report the gain in selected accuracy, in percentage points, of ϕ_{frz} over the strongest of the 15 baselines for each space, pool, and budget. All 25 full-pool and 25 restricted-pool point estimates are positive under the seed-42 score realizations.

Space	Pool	b=10	b=25	b=50	b=100	b=200
DARTS	full	+0.21	+0.45	+0.51	+0.52	+0.53
	restricted	+0.29	+0.36	+0.28	+0.34	+0.24
AmoebaNet	full	+0.41	+0.45	+0.56	+0.66	+0.71
	restricted	+0.28	+0.18	+0.09	+0.09	+0.09
ENAS	full	+0.75	+0.45	+0.64	+0.77	+0.51
	restricted	+0.61	+0.50	+0.64	+0.61	+0.38
PNAS	full	+0.25	+0.46	+0.53	+0.53	+0.52
	restricted	+0.53	+0.74	+0.82	+1.12	+0.68
NASNet	full	+0.61	+1.12	+1.19	+1.13	+1.22
	restricted	+0.49	+0.72	+0.70	+0.77	+0.53

In 25 shared-candidate selection conditions (five spaces by five budgets), the layer-mask policy is the strongest alternative numerical variant in every condition. The literal score has a positive point margin over that variant in 19 of 25 conditions and a positive interval lower endpoint in 13 of 25. Four intervals lie entirely below zero. The point margins range from approximately -0.034 to $+0.083$ accuracy points. Under the shared ϕ_{frz}-finite filter, the literal score also has a higher point estimate than the development-oriented gradient-only score in 24 of 25 conditions. It has a higher point estimate than both indicator-only and finite-pool random selection in all 25. These results help distinguish the information carried by the finite mask from the ranking signal without establishing the literal numerical policy as uniquely optimal.

Finally, every observed non-finite layer term has an exactly zero gradient maximum. No observed non-finite case has zero activation entropy with a nonzero gradient maximum. DARTS also contains layers where both quantities are zero. Operation-level associations vary by space. We therefore treat these patterns as exploratory and do not assign a common architectural cause.

L Presubmission Experiments and the Seed Campaign

Score direction in budgeted selection. The correlation analyses compare *absolute* controlled association (Section 3), whereas budgeted search uses each score in its frozen larger-is-better deployment orientation and selects the sampled architecture with the largest score. No sign is fitted separately by space, pool, or budget, and every method receives the same sampled candidate indices within a repeat. Under failure-as-worst, a non-finite candidate remains in the pool but is tied below all finite scores of that method. Under common-finite, it is excluded when the shared pool is constructed.

Three registrations postdate the opening and predate their respective results. First, the max-aware retest behind Table 2 fixes the common-finite intersection rows, recomputes all 15 baseline correlations inside every bootstrap resample, and re-selects the strongest baseline in each resample. Grouped jackknife blocks provide the BCa acceleration.

Table 20: Seed-stability retests on the five sealed NDS spaces for the registered 15-baseline roster. Each D column reports the max-aware margin with the statistic bank and all baselines recomputed under the indicated initialization and extraction seed using the RNG streams fixed in the registration. The average board rank-averages every score across seeds 42, 43, and 777 before applying the same max-aware retest. All 30 BCa intervals have positive lower endpoints. $\min \tau_a$ is the smallest cross-seed Kendall correlation between ϕ_{frz} score vectors over the three seed pairs, and J is the smallest Jaccard agreement between their finite-row masks.

Space	Analysis	D_{43} [BCa 95%]	D_{777} [BCa 95%]	D_{avg} [BCa 95%]	$\min \tau_a$	J
DARTS	residual	$+0.130$ [0.116, 0.143]	$+0.133$ [0.119, 0.146]	$+0.131$ [0.117, 0.143]	0.946	1.000
	stratified	$+0.078$ [0.064, 0.091]	$+0.078$ [0.066, 0.091]	$+0.078$ [0.067, 0.092]		
AmoebaNet	residual	$+0.066$ [0.055, 0.076]	$+0.066$ [0.056, 0.077]	$+0.065$ [0.053, 0.075]	0.954	1.000
	stratified	$+0.030$ [0.022, 0.039]	$+0.034$ [0.026, 0.042]	$+0.032$ [0.024, 0.040]		
ENAS	residual	$+0.089$ [0.079, 0.100]	$+0.087$ [0.077, 0.097]	$+0.087$ [0.076, 0.098]	0.945	1.000
	stratified	$+0.057$ [0.046, 0.068]	$+0.053$ [0.041, 0.064]	$+0.054$ [0.042, 0.065]		
PNAS	residual	$+0.080$ [0.069, 0.090]	$+0.082$ [0.070, 0.092]	$+0.081$ [0.068, 0.091]	0.952	1.000
	stratified	$+0.061$ [0.051, 0.073]	$+0.062$ [0.051, 0.073]	$+0.061$ [0.052, 0.073]		
NASNet	residual	$+0.070$ [0.063, 0.078]	$+0.069$ [0.062, 0.077]	$+0.069$ [0.061, 0.076]	0.954	1.000
	stratified	$+0.039$ [0.034, 0.047]	$+0.039$ [0.032, 0.045]	$+0.038$ [0.032, 0.045]		

Second, the budgeted-search simulation of Table 19 and Figure 3 fixes budgets, repeat counts, pools, and random-number streams in advance. The simulation includes three registered follow-up analyses. A common-finite rerun remains favorable in a majority of conditions, but the gains are not uniform across spaces, pools, and budgets.

The imputed selector replaces non-finite scores with independent draws from the empirical distribution of finite ϕ_{frz} scores in each space, with the imputations fixed across pools, budgets, and repeats. For the ranking–coverage decomposition, the coverage component is defined as the selected-accuracy difference between the original ϕ_{frz} selector, which treats non-finite scores as worst except for the registered all-non-finite fallback, and its imputed counterpart. The ranking component is the selected-accuracy difference between the imputed selector and the strongest eligible rival, selected condition-wise from the 15-baseline roster by mean selected accuracy. The coverage component is positive in every condition, ranging from $+0.06$ to $+1.70$ points, while the ranking component is non-positive in 17 of 50 conditions.

A further registered diagnostic uses an indicator-only selector that picks uniformly among the sampled architectures where ϕ_{frz} is finite. It exceeds random selection in 41 of 50 conditions by up to $+1.77$ points but trails ϕ_{frz} in all 50 by $+0.56$ to $+2.90$ points. These results indicate that the abstention signal accounts for part, but not all, of the observed gain.

Third, the seed campaign recomputed the statistic bank and post-opening registered 15-score roster under seeds 43 and 777, each jointly controlling model initialization and the synthetic statistic-extraction batch, on all 24,827 NDS architectures, with the campaign judgment rules fixed before any campaign result was computed. Table 20 reports the outcome. Every per-seed and rank-averaged max-aware interval excludes zero, the cross-seed Kendall correlations are at least 0.945, and the finite-row masks coincide in every space and seed pair. The finite-row pattern is therefore stable over the three tested seeds, which does not establish seed invariance.

One registration deviation is recorded. The registration pins the bootstrap streams of the per-seed retests but not of the rank-average board, which was therefore fixed to campaign token 111 under the same stream formula before any number was computed and recorded in the result file.

The campaign further adds six recent baselines (MeCo, SWAP, NEAR, the original Zen-NAS formula, LogSynflow, Entropic Score) to the registered 15-baseline roster, giving 21 rival scores. Eligibility for the extension was recorded in the campaign registration before any added score was computed. An eligible addition had to be a fixed training-free score computed from initialization statistics, have a verified implementation in the frozen audit codebase, and require no per-space hyperparameter tuning. These criteria excluded the regularized SWAP variant, whose regularization prior has no published NDS values, and also learned predictors such as ParZC (Dong et al., 2025) and discovery frameworks such as EZNAS and Auto-Prox (Akhauri et al., 2022; Wei et al., 2024), which are discovery engines parallel to Section 5 rather than fixed formulas. The extended comparison therefore covers 21 rival scores within this class rather than every published method.

Table 21: Extended-roster retest on the five sealed NDS spaces under the registered campaign. The final roster adds six baselines—MeCo, SWAP, NEAR, the original Zen-NAS formula, LogSynflow, and Entropic Score—to the registered 15-baseline roster, giving 21 rival scores. The common-finite intersection is recomputed over ϕ_{frz} and all 21 rivals, and the max-aware margin D re-selects the strongest rival within each BCa resample using the RNG streams fixed in the registration. This retest uses the registration-seed score realization for the six additions. All ten BCa intervals have positive lower endpoints. The switch column is the fraction of resamples in which the bar holder changes, and $n_{\cap}$ is the common-finite intersection size.

Space	Analysis	$T_\ell(\phi_{\mathrm{frz}})$	bar (holder)	D [BCa 95%]	switch	$n_{\cap}$
DARTS	residual	0.369	0.239 (mean $\lvert\nabla W\rvert$)	$+0.130$ [0.117, 0.144]	0.2%	3,660
	stratified	0.326	0.250 (mean $\lvert\nabla W\rvert$)	$+0.076$ [0.064, 0.089]	0.0%	
AmoebaNet	residual	0.467	0.402 (Fisher)	$+0.064$ [0.054, 0.075]	1.3%	4,463
	stratified	0.417	0.385 (mean $\lvert\nabla W\rvert$)	$+0.032$ [0.024, 0.039]	0.0%	
ENAS	residual	0.455	0.366 (GradNorm)	$+0.089$ [0.080, 0.100]	0.0%	4,506
	stratified	0.421	0.367 (GradNorm)	$+0.054$ [0.044, 0.066]	5.2%	
PNAS	residual	0.355	0.274 (mean $\lvert\nabla W\rvert$)	$+0.081$ [0.071, 0.091]	5.7%	4,478
	stratified	0.331	0.270 (mean $\lvert\nabla W\rvert$)	$+0.061$ [0.051, 0.072]	0.0%	
NASNet	residual	0.473	0.405 (mean $\lvert\nabla W\rvert$)	$+0.068$ [0.061, 0.076]	0.0%	4,322
	stratified	0.466	0.429 (mean $\lvert\nabla W\rvert$)	$+0.038$ [0.032, 0.045]	0.0%	

The registration assigns the additional initialization seeds a diagnostic role for the frozen roster. The six added baselines are therefore scored under the registration seed only and are not replicated across seeds. Table 21 reports the retest. The additions leave the common-finite intersection rows unchanged and never become the max-aware bar holder. Under the independent RNG streams frozen in the registration, the ten margins match the values of Table 2 to three decimals. Under the residual analysis, at least four of the six additions have negative correlations with accuracy in every space, with the original Zen-NAS formula reaching -0.372 on AmoebaNet. This pattern is consistent with the negative size-controlled associations observed in the audit.

A post-hoc multiplicity analysis, prompted by presubmission review and labeled as such, uses fresh documented random streams. Paired bootstrap intervals for the fifty common-finite conditions, with the strongest eligible baseline re-selected inside every resample, are nominally positive in 45 of 50 conditions. Five intervals include zero and none lie entirely below zero. After Bonferroni correction, 41 of 50 intervals remain positive (Table 22). The ten max-aware margins of Table 2 also remain above zero after Bonferroni adjustment for the ten claims.

A Gaussian calibration simulation of the max-aware statistic was matched to the DARTS intersection under the residual analysis and run at the production interval settings. Two-sided coverage is 96.8 percent at the observed margin and 95.5 percent under a boundary null with the true margin fixed at zero. Under the boundary null, the win rule fires in 2.8 percent of draws, compared with its one-sided nominal rate of 2.5 percent. Each scenario uses 400 replications.

M Additional Post-hoc Analyses

The analyses in this appendix were added after the NDS suite had been opened, so each is post-hoc with respect to that opening and changes no registered result. For each, the estimator, population, roster, controls, resampling scheme, and multiplicity family were fixed in a written analysis contract before any of its results were computed, and the supplement records their version history. Unless stated otherwise, the population is the common-finite intersection of Table 2, the roster is the 15-baseline roster, intervals are BCa intervals from architecture-row resampling of the fixed benchmark population, and adjusted intervals are two-sided Bonferroni intervals within the stated family with $B \geq \max(10^4, 2000m)$ resamples.

M.1 Adjusting for the choice among the frozen candidates

C1, C2, and C3 were frozen together before NDS access, and C3 was designated the focal score after the opening. Table 23 treats the three candidates as one family of 30 margin comparisons. Resamples are drawn once from the target-finite population and restricted to each candidate's finite support, so the candidates share row weights. All ten ϕ_{frz} margins retain adjusted lower limits above

Table 22: Paired bootstrap intervals for the common-finite budgeted-search gain of ϕ_{frz} over the strongest of the 15 baselines. The common-finite intersection is formed over ϕ_{frz} and all 15 baselines within each pool, and the strongest baseline is re-selected inside every resample. Intervals use 20,000 paired bootstrap resamples of the 1,000 candidate-set draws per condition and quantify simulation uncertainty over those draws within each fixed benchmark condition. A dagger marks an interval that includes zero at the nominal 95% level. A double dagger marks a nominally positive interval that does not remain positive under the fifty-condition Bonferroni correction.

Space	b	full D [95% CI]	restricted D [95% CI]
DARTS	10	$+0.02$ [-0.03, 0.07]†	$+0.12$ [-0.05, 0.16]†
	25	$+0.14$ [0.09, 0.19]	$+0.27$ [0.12, 0.31]
	50	$+0.28$ [0.23, 0.33]	$+0.23$ [0.03, 0.30]‡
	100	$+0.45$ [0.40, 0.50]	$+0.04$ [-0.01, 0.10]†
	200	$+0.48$ [0.43, 0.52]	-0.04 [-0.10, 0.01]†
AmoebaNet	10	$+0.43$ [0.23, 0.72]	$+0.23$ [0.10, 0.41]
	25	$+0.52$ [0.43, 0.66]	$+0.07$ [0.02, 0.11]‡
	50	$+0.56$ [0.51, 0.61]	$+0.07$ [0.02, 0.12]‡
	100	$+0.67$ [0.63, 0.72]	$+0.10$ [0.06, 0.14]
	200	$+0.68$ [0.64, 0.72]	$+0.10$ [0.06, 0.14]
ENAS	10	$+0.17$ [0.02, 0.35]‡	$+0.42$ [0.27, 0.59]
	25	$+0.61$ [0.35, 0.91]	$+0.02$ [-0.04, 0.08]†
	50	$+0.47$ [0.33, 0.66]	$+0.22$ [0.08, 0.42]
	100	$+0.54$ [0.35, 0.78]	$+0.19$ [0.13, 0.25]
	200	$+0.53$ [0.43, 0.67]	$+0.27$ [0.22, 0.32]
PNAS	10	$+0.41$ [0.19, 0.62]	$+0.47$ [0.42, 0.52]
	25	$+0.35$ [0.25, 0.39]	$+0.67$ [0.61, 0.73]
	50	$+0.44$ [0.32, 0.52]	$+0.76$ [0.69, 0.83]
	100	$+0.48$ [0.36, 0.59]	$+1.03$ [0.88, 1.10]
	200	$+0.42$ [0.38, 0.46]	$+0.55$ [0.27, 0.90]
NASNet	10	$+0.49$ [0.37, 0.57]	$+0.60$ [0.43, 0.75]
	25	$+0.75$ [0.61, 0.92]	$+0.59$ [0.51, 0.66]
	50	$+1.04$ [0.79, 1.11]	$+0.77$ [0.68, 0.85]
	100	$+1.09$ [0.89, 1.25]	$+0.84$ [0.75, 0.93]
	200	$+0.94$ [0.81, 1.07]	$+0.57$ [0.49, 0.65]

Table 23: Selection-adjusted max-aware margins over the 15-baseline roster for the three candidates frozen before NDS access. Brackets are BCa intervals Bonferroni-adjusted over all $3 \times 5 \times 2 = 30$ comparisons ($B = 60{,}000$), with common row weights across candidates. C2 is the negated incumbent mean $|\nabla W|$, so its margin cannot be positive and no interval is reported. Post-hoc analysis.

Space	Analysis	C1 D [adj.]	C2 D	C3 D [adj.]
DARTS	residual	-0.067 [-0.117, -0.020]	-0.013	$+0.130$ [0.108, 0.152]
	stratified	$+0.005$ [-0.044, 0.059]	$+0.000$	$+0.076$ [0.056, 0.098]
AmoebaNet	residual	-0.270 [-0.298, -0.244]	-0.018	$+0.064$ [0.046, 0.082]
	stratified	-0.283 [-0.316, -0.252]	$+0.000$	$+0.032$ [0.019, 0.044]
ENAS	residual	-0.366 [-0.391, -0.342]	-0.028	$+0.089$ [0.072, 0.106]
	stratified	-0.316 [-0.366, -0.267]	-0.019	$+0.054$ [0.036, 0.071]
PNAS	residual	-0.254 [-0.282, -0.225]	-0.006	$+0.081$ [0.062, 0.098]
	stratified	-0.198 [-0.253, -0.145]	$+0.000$	$+0.061$ [0.045, 0.079]
NASNet	residual	-0.330 [-0.372, -0.286]	$+0.000$	$+0.068$ [0.056, 0.081]
	stratified	-0.338 [-0.385, -0.292]	$+0.000$	$+0.038$ [0.027, 0.049]

zero, the smallest being $+0.019$ (AmoebaNet, stratified). The adjustment covers the choice among the three frozen candidates. It does not correct for the post-opening design of the roster or make the NDS spaces an untouched holdout.

M.2 Richer measured size controls

After linear projection on log parameters and log FLOPs, the entropy sum $\sum_l \log H(a_l)$ of ϕ_{frz} is almost perfectly negatively correlated with the included-layer count (correlation -0.98 to -0.99 in the five spaces), so the layer count is a natural further control. Table 24 adds it (ZL) and, in addition, configuration depth and width (Z+). Under linear rank-normal control, ϕ_{frz} exceeds every eligible roster member in all five spaces, and all ten adjusted intervals exclude zero. The same holds when the linear nuisance model is fitted on a hash-selected calibration half and evaluated on the

Table 24: Max-aware margins of ϕ_{frz} over the 15-baseline roster under richer measured size controls (post-hoc). ZL adds the included-layer count and Z+ additionally adds configuration depth and width to log parameters and log FLOPs. Linear columns refit the rank-normal residual analysis on all common-finite rows. Split columns fit the Z+ nuisance model on a fixed hash-selected half and evaluate on the other half. Matching reports the range over six caliper settings and the number of settings whose adjusted interval excludes zero. *Adjusted BCa interval above zero. †Adjusted interval below zero. Adjustments are Bonferroni within each family (10 linear, 15 split, 30 matching comparisons).

	Linear, all rows		Split, Z+ nuisance			
Space	ZL	Z+	linear	spline	boosted trees	Matching
DARTS	$+0.142^{*}$	$+0.158^{*}$	$+0.143^{*}$	$+0.034^{*}$	$+0.037^{*}$	+0.007–+0.028 (2/6)
AmoebaNet	$+0.067^{*}$	$+0.091^{*}$	$+0.092^{*}$	-0.018	-0.044	+0.041–+0.045 (6/6)
ENAS	$+0.118^{*}$	$+0.118^{*}$	$+0.100^{*}$	$+0.012$	$+0.033^{*}$	+0.025–+0.046 (6/6)
PNAS	$+0.104^{*}$	$+0.121^{*}$	$+0.137^{*}$	$+0.032^{*}$	$+0.030^{*}$	+0.030–+0.046 (6/6)
NASNet	$+0.082^{*}$	$+0.039^{*}$	$+0.079^{*}$	$+0.014$	$+0.003$	+0.032–+0.046 (6/6)

complementary half. Local matching on log parameters and log FLOPs gives positive margins in all 30 settings, 26 of which keep adjusted intervals above zero. With flexible nonlinear nuisance models (additive splines or gradient-boosted trees) fitted on the calibration half, the margins are smaller and not uniform. Adjusted intervals remain above zero in DARTS and PNAS, and in ENAS with boosted trees, but not in AmoebaNet or NASNet. These split analyses first apply the rank-normal transform used by the residual estimator. The nonlinear sensitivity in Table 12 instead regresses the raw scores, which is sensitive to heavy-tailed score distributions.

Thus five of the ten split nonlinear comparisons have adjusted intervals above zero. The remaining five do not establish superiority, and the negative AmoebaNet point margins do not establish inferiority after adjustment. The nonlinear comparisons belong to the 15-comparison split family, which also includes the five linear comparisons. The common-finite population sizes for the all-row linear analyses are reported in Table 2. Table 24 uses a different nonlinear procedure from Table 12, so differences between these tables cannot be attributed to additional size controls alone.

M.3 Numerical semantics and finite support

Table 25: Post-hoc comparison of literal and search-time protected semantics on the full target-finite population. Non-finite scores are ranked worst. Margins are max-aware over the 15-baseline roster with nominal BCa intervals ($B = 10{,}000$). Selection gives the number of 10 budget-by-pool conditions with a positive paired interval against the strongest roster member for the literal and protected semantics, respectively. The last column counts non-finite layers in 40 affected architectures per space whose random-weight replacement (standard deviation 10) leaves the logits unchanged for a fixed input. On literal-finite architectures, the two semantics induce identical rankings.

	Literal D		Protected D			Perturbation
Space	residual	stratified	residual	stratified	Selection	check
DARTS	$+0.147$ [0.127, 0.168]	$+0.147$ [0.131, 0.171]	-0.017 [-0.037, 0.004]	-0.051 [-0.068, -0.028]	10/10, 0/10	639/639
AmoebaNet	$+0.020$ [0.005, 0.034]	$+0.017$ [0.004, 0.031]	$+0.016$ [0.002, 0.028]	-0.012 [-0.022, -0.002]	10/10, 0/10	1237/1237
ENAS	$+0.057$ [0.043, 0.070]	$+0.046$ [0.032, 0.060]	$+0.019$ [0.005, 0.033]	-0.017 [-0.032, -0.003]	10/10, 0/10	494/494
PNAS	$+0.086$ [0.069, 0.102]	$+0.084$ [0.069, 0.100]	$+0.025$ [0.008, 0.039]	$+0.012$ [-0.000, 0.024]	10/10, 0/10	720/720
NASNet	$+0.031$ [0.017, 0.044]	$+0.015$ [0.002, 0.029]	$+0.014$ [0.003, 0.026]	-0.018 [-0.029, -0.007]	10/10, 0/10	651/651

The GP search evaluated the serialized program $\sum_l \log(H(a_l)/\max|\nabla W_l|)$ with protected division and logarithm, whereas ϕ_{frz} uses the literal operations. On the architectures where ϕ_{frz} is finite, the protected program, both ϵ variants, all clipping floors, and the layer mask give exactly the same ranking (Kendall $\tau = 1$ in every space), so the change of operations alone does not alter any association, margin, or selection there. The policies differ only on architectures with a non-finite layer term, all of which have a zero maximum gradient. In 40 affected architectures per space, replacing the weights of every such layer by large random values left the logits bit-identical (Table 25), so these layers did not affect the logits under this perturbation. The protected program scores them with large positive contributions, whereas the literal policy yields a non-finite score, which is ranked last un-

der failure-as-worst. Under that rule the literal policy keeps positive margins in all ten evaluations. These checks are descriptive and cover the sampled architectures and a fixed input.

M.4 Comparison with AZ-NAS and VKDNW, and cost

Table 26: Extended comparison with AZ-NAS and VKDNW (post-hoc). Left: residual Kendall association on the common-finite rows. VKDNW$_H$ is the Fisher-spectrum entropy, VKDNW$_s$ adds the parameter-tensor count, and VKDNW$_{agg}$ is the published log-rank aggregate with JacCov, AZ-NAS expressivity and trainability, and FLOPs. Right: max-aware margin of ϕ_{frz} over the 15-baseline roster plus these four scores, with BCa intervals Bonferroni-adjusted over 10 comparisons ($B = 20{,}000$).

	Residual τ_b					D vs. extended roster [adj.]	
Space	ϕ_{frz}	AZ-NAS	VKDNW$_H$	VKDNW$_s$	VKDNW$_{agg}$	residual	stratified
DARTS	+0.369	+0.112	+0.146	-0.091	+0.211	$+0.130$ [0.111, 0.149]	$+0.076$ [0.058, 0.095]
AmoebaNet	+0.467	+0.271	+0.152	-0.329	+0.278	$+0.064$ [0.048, 0.080]	$+0.032$ [0.021, 0.043]
ENAS	+0.455	+0.211	+0.221	-0.216	+0.271	$+0.089$ [0.074, 0.104]	$+0.054$ [0.038, 0.070]
PNAS	+0.355	+0.138	+0.149	-0.167	+0.188	$+0.081$ [0.064, 0.096]	$+0.061$ [0.047, 0.078]
NASNet	+0.473	+0.181	+0.201	-0.129	+0.206	$+0.068$ [0.057, 0.080]	$+0.038$ [0.029, 0.047]

We added AZ-NAS (Lee & Ham, 2024) and three forms of VKDNW (Tybl & Neumann, 2025) to the roster. The three VKDNW forms are the Fisher-spectrum entropy, the entropy augmented with the parameter-tensor count, and the published log-rank aggregate with JacCov, AZ-NAS expressivity and trainability, and FLOPs. VKDNW uses the official scorer with a reformulated Jacobian routine that matches the official routine in float64. With these scores added, all ten adjusted intervals for the margins of ϕ_{frz} remain above zero (Table 26), and no added score becomes the bar holder under the two original analyses. This comparison concerns association under size control. The budgeted-selection statements in Section 6 refer to the 15-baseline roster.

Table 27: Per-architecture cost on one RTX 5090 GPU over 200 hash-selected NDS architectures (40 per space). Two warm-up runs were discarded. Model construction is excluded, and CUDA is synchronized around each measurement.

Method	median ms	p95 ms	median MiB	p95 MiB	Computation
ϕ_{frz} (needed statistics)	105	238	870	2040	1 fwd + 1 bwd, bs 32
Full statistic bank	410	974	879	2075	1 fwd + 1 bwd, bs 32 (+ statistics)
Nine published proxies (combined)	1887	5156	4061	10169	grad family: 3 fwd/bwd, incl. one double-bwd (GraSP) NASWOT: 1 fwd; JacCov: 1 fwd/bwd SynFlow: 1 fwd/bwd fp64; bs 64
VKDNW	476	663	7120	11860	1 fwd + 10 bwd (exact fast path) bs 64; + 128×128 SVD
AZ-NAS	570	1099	1872	4362	1 fwd + per-feature bwd bs 64; cuDNN off

Computing only the statistics that ϕ_{frz} needs takes a median of 105 ms per architecture on one GPU (Table 27). Under the measured settings, this is lower than the median times for the full statistic bank, VKDNW, AZ-NAS, and the nine published proxies combined. AZ-NAS was run with cuDNN disabled, matching the adapter used to produce its NDS scores. The comparison excludes the offline cost of discovery.

M.5 Five further NDS design families

The NDS archive also contains CIFAR-10 design families whose targets had not previously been read in this study: ResNet, ResNet-B, ResNeXt-A, ResNeXt-B, and Vanilla (plain networks without residual connections). Before reading any of their targets, we fixed a hash-selected sample of 5,000 architectures per family, the frozen score, the 15-baseline roster, both analyses, and the win rule of Table 2. Our reimplementation of the family generator reproduces the benchmark parameter counts exactly, and ϕ_{frz} is finite for every sampled architecture. All eight Bonferroni-adjusted intervals for the four residual families lie below zero. Vanilla has positive nominal intervals under both analyses,

Table 28: Evaluation on five NDS design families whose targets had not previously been read in this study (post-opening, with the protocol fixed before reading any target from these families). Each family is a fixed hash-selected sample of 5,000 architectures. T is the association of ϕ_{frz} on the common-finite rows, and the bar is the largest absolute association among the 15-baseline roster, re-selected in every resample. Brackets give nominal and Bonferroni-adjusted ($m = 10$) BCa intervals ($B = 20{,}000$). The last column gives the common-finite count and the fraction of target-finite rows where ϕ_{frz} is undefined.

Family	Analysis	$T(\phi_{\mathrm{frz}})$	bar (holder)	D [95%]	[adj.]	$n_\cap$ (undef.)
ResNet	residual	+0.058	0.210 (NASWOT)	-0.151 [-0.168, -0.135]	[-0.175, -0.128]	5,000 (0.0%)
	stratified	+0.060	0.243 (ℓ_1 sum)	-0.183 [-0.203, -0.163]	[-0.211, -0.155]	
ResNet-B	residual	+0.123	0.541 (ℓ_2 norm)	-0.418 [-0.443, -0.394]	[-0.454, -0.384]	5,000 (0.0%)
	stratified	+0.135	0.551 (ℓ_2 norm)	-0.416 [-0.443, -0.394]	[-0.453, -0.384]	
ResNeXt-A	residual	+0.018	0.480 (ℓ_1 sum)	-0.461 [-0.488, -0.434]	[-0.497, -0.422]	5,000 (0.0%)
	stratified	+0.006	0.579 (ℓ_1 sum)	-0.573 [-0.595, -0.558]	[-0.600, -0.548]	
ResNeXt-B	residual	+0.122	0.508 (ℓ_2 norm)	-0.387 [-0.411, -0.362]	[-0.421, -0.352]	5,000 (0.0%)
	stratified	+0.132	0.522 (ℓ_2 norm)	-0.390 [-0.414, -0.368]	[-0.422, -0.357]	
Vanilla	residual	+0.670	0.653 (H sum)	$+0.016$ [0.008, 0.024]	[0.004, 0.027]	5,000 (0.0%)
	stratified	+0.671	0.661 (H sum)	$+0.010$ [0.001, 0.018]	[-0.003, 0.022]	

but only its residual interval remains above zero after adjustment. Its adjusted stratified interval is $[-0.003, +0.022]$ (Table 28). A post-hoc diagnostic examines layer count as a possible correlate of this difference. At fixed parameters and FLOPs, depth is positively associated with accuracy in the residual families but negatively associated with accuracy in Vanilla and the five cell-based spaces, while the entropy sum of ϕ_{frz} decreases with layer count in every family. In the cell-based spaces, the margins remain positive when layer count is controlled (Appendix M.2). Three independent training seeds are available for ResNet and Vanilla, satisfying the independent-seed premise of T4. Under its remaining assumptions, the corresponding ceilings are $[0.804, 0.897]$ and $[0.890, 0.943]$, respectively.