SAME FORECAST, DIFFERENT CAPACITY: ACQUISITION PATHS IN TIME-SERIES FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

How much structure does a forecasting model need to retain its prediction accuracy? The answer depends on how the smaller model is obtained. We study capacity as the first structural budget that reaches a fixed accuracy target, comparing models learned at that budget with restrictions of a learned full model. Across six public forecasting model families, the two routes can require substantially different capacities. More surprisingly, the complete forecasting function need not change for its restricted capacity to change: transferring functionally invisible components between trained linear representations changes the required budget in both directions, consistently across three seeds on four external model–task pairs. We identify two distinct sources of path dependence. A restriction can expose components that cancel in the full function; even when this source is eliminated, retained and discarded features remain coupled through the training covariance, producing different restricted and refitted solutions. Exact linear controls separate these sources from incomplete optimization. An exact risk response connects function-preserving interventions to capacity thresholds. Residual alignment explains why closer approximation of the full model need not improve task accuracy. Capacity is therefore an operational property of a forecasting task and its acquisition path, not a consequence of the full predictor alone.

1 INTRODUCTION

How much structure does a forecasting model need to retain its accuracy? A compact model can be learned directly or obtained by restricting a trained full model. Direct training lets the retained components adapt to what is absent; restriction inherits a solution learned with the discarded components. In long-horizon forecasting, where simple linear models compete with larger architectures (Zeng et al., 2023), this distinction affects how we interpret the structure needed for accurate prediction.

We study *acquisition-conditioned capacity*: the first structural budget at which a specified acquisition path recovers the accuracy of a selected full forecaster. Across six forecasting families, direct learning and restriction produce different capacities while sharing the same full endpoint. Our central finding goes further: *a fixed complete forecasting function can have different restricted capacities.*

In NLinear and DLinear, components of a trained representation can cancel in the full forecast. Removing structure can break that cancellation and expose these otherwise invisible components. We transfer such components between trained representations while keeping each host's full forecasting function fixed. Its restricted capacity changes in both directions (Figure 1). The internal representation thus carries information about the restricted path that full predictions alone leave unresolved.

We identify two distinct sources of path dependence in these linear models. An algebraic condition characterizes when restriction preserves functional equivalence. Restrictions satisfying this condition eliminate the representation effect, yet exact ridge controls still exhibit a learned–restricted gap. This remaining gap arises from coupling between retained and discarded features through the training data. An exact risk response then connects prediction changes to capacity thresholds through their magnitude and alignment with the target residual. This connection explains how closer approximation of a full predictor can worsen task accuracy.

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.

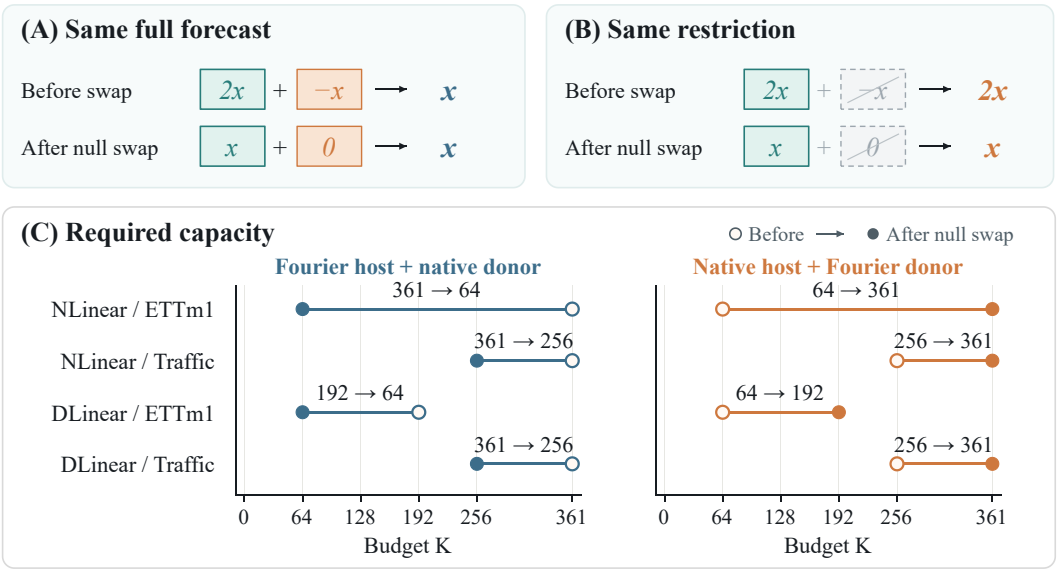


Figure 1: **The same full forecast can have different restricted capacities.** (A,B) In a two-coordinate example, changing contributions $(2x, -x)$ to $(x, 0)$ preserves their sum but changes the result of retaining the first coordinate. (C) Transferring trained null components preserves each host's full prediction function while changing its capacity at $\epsilon = 1\%$ under the same Fourier restriction. All three seeds give the displayed transitions ($H = 336$). "Fourier" and "native" identify training coordinates.

Our contributions are:

- **Capacity along acquisition paths.** We establish capacity separations across forecasting families and show how equal capacities can conceal different risk profiles.
- **Two sources of path dependence.** We separate representation exposure from training coupling through same-function interventions, predicted negative controls, and exact solutions.
- **Risk and capacity transitions.** We connect prediction changes to task-specific capacity through an exact risk response and show how teacher fidelity and task accuracy can move in opposite directions.

2 Capacity along a common structural ladder

To compare acquisition paths, we hold the forecasting task, structural ladder, and full-model accuracy target fixed. At each intermediate budget, we vary how the predictor is acquired.

Forecasting risk and the full reference. For an input history $x_t \in \mathbb{R}^{L\times d}$ and future target $y_t \in \mathbb{R}^{H\times d}$, test MSE on a fixed chronological sequence of n forecast origins is

$$\widehat{R}(f) = \frac{1}{nHd}\sum_{t=1}^{n} \|f(x_t) - y_t\|_F^2. \tag{1}$$

MAE uses the corresponding absolute loss and serves as a secondary metric. For each training seed, we select a full predictor f_{F} using validation data. Its test risk provides the task-specific accuracy reference. Capacity asks how far one can move down the structural ladder while retaining this achieved level of forecasting accuracy.

One ladder, two acquisition paths. Let $\mathcal{K} = \{K_0 < \cdots < K_m\}$ be an ordered set of budgets for a nested family of retained structures. At budget K, a restriction T_K acts on the implemented representation of the full predictor. Depending on the model family, K indexes retained frequency

coordinates, harmonic terms, or residual blocks. Each family uses its own structural units and a common ladder for the two paths:

$$f_{\mathrm{L},K} = \mathrm{Train}(\mathcal{F}_K), \qquad f_{\mathrm{R},K} = T_K(f_{\mathrm{F}}), \qquad f_{\mathrm{L},K_m} = f_{\mathrm{R},K_m} = f_{\mathrm{F}}. \tag{2}$$

The *learned* path trains directly with the structure available at budget K. The *restricted* path acquires the full model and removes the structure excluded by T_K. Within each paired comparison, the data, structural ladder, inactive-coordinate convention, and full endpoint are matched. Retained parameters start from matched initial values: learning optimizes them in the absence of the discarded structure, while restriction inherits values fitted in its presence. The two paths therefore compare acquisition histories for the same intermediate structure and meet at the same selected full predictor.

Acquisition-conditioned capacity. For a relative tolerance $\epsilon \geq 0$ and path $a \in \{\mathrm{L}, \mathrm{R}\}$, define

$$\mathcal{A}_a(\epsilon) = \{K \in \mathcal{K} : \widehat{R}(f_{a,K}) \leq (1+\epsilon)\widehat{R}(f_{\mathrm{F}})\}, \qquad C_a(\epsilon) = \min \mathcal{A}_a(\epsilon). \tag{3}$$

The *acquisition-conditioned capacity* $C_a(\epsilon)$ is the first measured budget that reaches the shared accuracy target. A difference between $C_{\mathrm{L}}(\epsilon)$ and $C_{\mathrm{R}}(\epsilon)$ means that recovering the same achieved forecast accuracy requires different structural budgets along the two acquisition paths.

The acceptable set $\mathcal{A}_a(\epsilon)$ records every qualifying budget. Risk can rise and fall as structure increases, and two paths can share a first adequate budget while differing substantially at smaller budgets. A path can also qualify at one budget and fail again at a larger one. We therefore analyze capacity together with the full acceptable set and complete risk profiles.

Primary readout and aggregation. The main text uses $\epsilon = 1\%$. Sensitivity on the core tasks uses 51 tolerances from 0% to 5%; alternative capacity readouts and grid-resolution checks appear in the appendix. Repeated acquisitions vary the training seed, with each risk aggregating the overlapping windows and channels of the same test segment. For task-level summaries, we average the three seed-wise risk profiles and their full-model risks, then apply Equation (3). This defines the *pooled capacity*. Exact linear controls are deterministic.

With the comparison fixed, the next question is whether the two paths actually assign different capacities to the same forecasting task. We first establish that separation across forecasting families, then identify two sources of path dependence through controlled linear interventions.

3 Acquisition changes capacity across forecasting structures

We evaluate FITS, MixLinear, NLinear, DLinear, N-BEATS, and N-HiTS (Xu et al., 2024; Ma et al., 2026; Zeng et al., 2023; Oreshkin et al., 2020; Challu et al., 2023). The core comparison covers ETTh2, Weather, Electricity, and Exchange at $H \in \{96, 336, 720\}$, using $L = 720$ and three training seeds. The N-BEATS family contributes an interpretable configuration on this full matrix and a generic configuration on ETTh2, yielding 75 task–configuration comparisons evaluated with both MSE and MAE. Architectures retain their own forecasting components; restrictions target frequency coordinates, native harmonic expansions, or native residual blocks. Appendix A specifies the implementation and training recipes.

Large, repeatable separations. Acquisition produces large capacity separations across frequency-based and residual forecasting structures (Figure 2A). For FITS on Weather at $H = 336$, the pooled MSE capacities are 34 and 128 frequency bins for learning and restriction. For MixLinear on Weather at $H = 720$, every seed gives 5 versus 12 at both MSE and MAE. For interpretable N-BEATS on Weather at $H = 336$, the three seed pairs are $4/167$, $4/167$, and $8/167$, also identical under the two metrics. Its learning path reaches the target earlier at every tested MSE tolerance from 0% to 5%. Each pair measures the structural budget needed to reach a shared accuracy within its forecasting family.

Equal capacities conceal different risk profiles. N-HiTS has equal pooled MSE capacities in all 12 core tasks, while its paths can favor different procedures at intermediate budgets. On Weather at $H = 720$, the pooled capacity is 2 for both paths, yet learning is better at block $K = 1$ and restriction is better at $K = 2$ in every seed. Interpretable N-BEATS on Electricity at $H = 336$ and 720 gives

Table 1: Capacity relations across the 75 core tasks at $\epsilon = 1\%$. Each triple counts tasks with $C_L < C_R$, $C_L = C_R$, and $C_L > C_R$, respectively. Capacities are computed from mean three-seed risks; the final column records whether MSE and MAE give the same relation. The two N-BEATS rows are configurations of the same model family.

Configuration	Tasks	MSE $(<, =, >)$	MAE $(<, =, >)$	Agreement
FITS	12	(8, 3, 1)	(7, 4, 1)	11/12
MixLinear	12	(9, 3, 0)	(9, 3, 0)	8/12
N-BEATS (generic)	3	(0, 3, 0)	(0, 3, 0)	3/3
N-BEATS (interpretable)	12	(7, 5, 0)	(6, 6, 0)	11/12
N-HiTS	12	(0, 12, 0)	(1, 11, 0)	11/12
NLinear	12	(6, 6, 0)	(6, 6, 0)	12/12
DLinear	12	(8, 4, 0)	(9, 3, 0)	11/12
All configurations	75	(38, 36, 1)	(38, 36, 1)	67/75

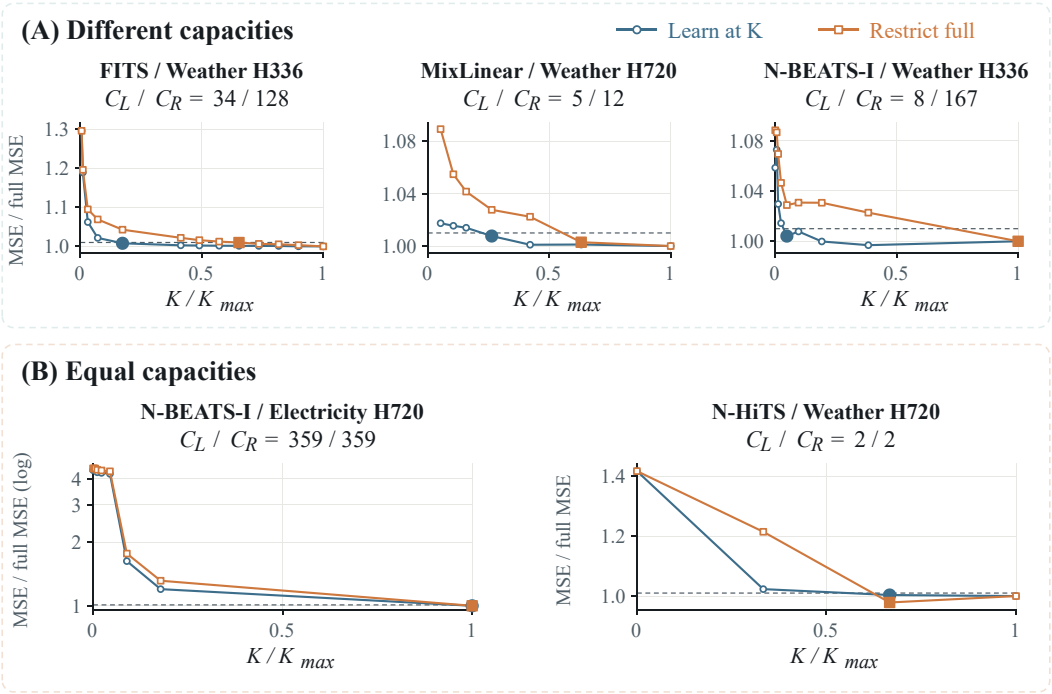


Figure 2: **Capacity reads the first adequate point of a risk profile.** Pooled MSE is divided by pooled full-model MSE. Dashed lines mark the 1.01 target; filled markers mark the first adequate measured budget. (A) Capacity separations across forecasting structures. (B) Equal capacities with different intermediate-budget errors. Lines connect measured grid points; capacities are not interpolated. N-BEATS-I denotes interpretable N-BEATS. The Electricity panel uses a logarithmic vertical axis.

another pattern: the two paths first qualify at the full endpoint throughout the tolerance scan, while learning has lower MSE at every non-full budget (Figure 2B). The shared first hit locates adequate accuracy, while the risk profiles resolve behavior at the other budgets.

Capacity directions across metrics and seeds. At the primary tolerance, MSE and MAE preserve the direction of the pooled capacity relation in $67/75$ core tasks; the remaining eight change between equality and separation. Three seeds agree on the direction in $59/75$ tasks under each metric; the complete seed results appear in Appendix C. FITS Exchange at $H = 720$ illustrates the role of aggregation: it has pooled MSE capacities $196/176$, even though two seeds favor learning and the

third is equal: averaging risks and voting on first hits are different operations. Across tasks, the relative capacity of learning and restriction depends on the task, metric, and aggregation rule.

The capacity gap also appears on ETTh1 across NLinear, DLinear, FITS, and MixLinear at $H \in \{96, 192, 336, 720\}$. At the 1% tolerance, learning requires a smaller structural budget in $11/16$ MSE comparisons and $15/16$ MAE comparisons; the remaining comparisons have equal capacities (Table 11).

To identify the source of these capacity separations, we hold the full forecasting function fixed and intervene on its representation. This isolates the role of the learned representation; exact-solution controls then distinguish the contribution of refitting from incomplete optimization.

4 Same full function, different restricted capacity

What a complete forecast cannot reveal. Consider a linear forecasting architecture, including its fixed centering or decomposition operations. Write its trainable coefficient matrix as B, and let A be the linear realization map from those coefficients to the complete affine prediction operator. A fixed residual forecast, when present, is held constant. Two coefficient matrices are functionally equivalent when $AB = AB'$. This equivalence is over inputs, not merely equality of test MSE. With the declared Euclidean coefficient metric, let

$$Q = A^{\dagger}A, \qquad B = QB + (I - Q)B =: B_{\mathrm{vis}} + B_{\mathrm{null}}. \tag{4}$$

Here $A^{\dagger}$ denotes the Moore–Penrose pseudoinverse. The first component determines the full function. The second belongs to $\ker A$ and is invisible until an operation breaks its cancellation. For a coefficient restriction P_K, its contribution becomes $AP_K B_{\mathrm{null}}$. Thus a component that contributes nothing at full budget can affect all intermediate forecasts.

Proposition 1 (When restriction depends only on the full function)**.** *For a linear realization A and a linear restriction P_K, all coefficient representations of the same full function have the same restricted function if and only if*

$$P_K \ker(A) \subseteq \ker(A). \tag{5}$$

Equivalently, there is a linear map $\widetilde{T}_K$ on $\mathrm{range}(A)$ such that $AP_K = \widetilde{T}_K A$.

When every restriction satisfies this criterion, the complete function determines the entire restricted prediction path and, with a common reference, its capacity at every tolerance. An incompatible restriction can expose differences between equivalent representations through $AP_K \ker(A)$. The criterion therefore specifies where a function-preserving intervention can change the forecast and where its response must be zero. These predictions guide the paired interventions below; proofs appear in Appendix B.

Transferring the invisible component. For a fixed trained host B and a separately trained donor B', construct

$$B^{\mathrm{swap}} = QB + (I - Q)B'. \tag{6}$$

Then $AB^{\mathrm{swap}} = AB$ exactly, while its restricted change is $AP_K(I - Q)(B' - B)$. The intervention preserves the host's full predictor, its accuracy target, and the restriction family. It changes only which functionally invisible component that representation carries.

We apply this test to NLinear and DLinear on ETTm1 and Traffic at $H = 336$, using three paired seeds and two training-coordinate systems: native lag coordinates and orthonormal real Fourier coordinates. The two training arms start from function-matched initializations and use full-batch Adam on the same ridge objective. After training, both hosts are expressed in the same coefficient system and evaluated under the same Fourier-prefix restriction on each architecture's processed features. Native centering, decomposition, and biases are preserved throughout, fixing the preprocessing and the restriction rule across the paired hosts.

Capacity follows the transferred representation. All 12 paired seeds show the bidirectional transitions in Figure 1C. For NLinear on ETTm1, the Fourier-trained host moves from capacity 361 to 64 after receiving the native-trained donor's invisible component; the reverse intervention moves the native-trained host from 64 to 361. DLinear on ETTm1 moves $192 \to 64$ and $64 \to 192$. On

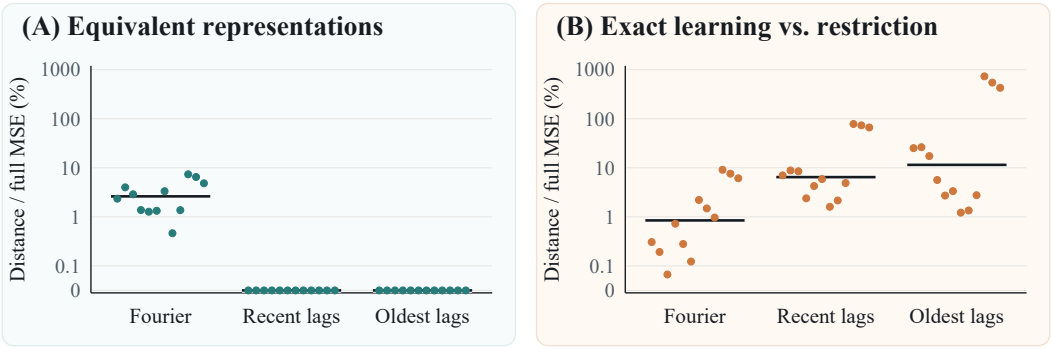


Figure 3: **Closing the representation channel does not close the acquisition gap.** NLinear across 12 tasks and three restriction families. (A) Original versus minimum-norm equivalent representations after restriction; points average three seeds. (B) Exact ridge learning versus restriction of the full ridge solution; points are deterministic task results. Distance is the integral of absolute MSE differences over normalized budget, divided by full MSE, in percent. Horizontal bars show medians. The scale is linear below 0.1% and logarithmic above it.

Traffic, both architectures move $361 \to 256$ in one direction and $256 \to 361$ in the other. Across interventions, the maximum absolute full affine-operator difference in Frobenius norm is 5.8×10^{-14} and the maximum full-MSE difference is 2.1×10^{-15}.

Each donor supplies its trained component at its original magnitude. The causal comparison is within each host against its unchanged full-function target. The two independently trained hosts need not have identical full functions to each other. The experiment therefore isolates an effect that a full-model accuracy comparison cannot explain: invisible training representation can determine how soon the restricted model reaches the accuracy target.

5 Separating Representation from Training Coupling

The kernel criterion predicts where the effect vanishes. Equation (5) predicts that NLinear's recent-lag and oldest-lag masks preserve its nullspace, whereas its intermediate Fourier masks can expose it. Removing the invisible component should consequently leave NLinear's temporal restrictions unchanged. This prediction holds in all 72 temporal seed–restriction comparisons: the maximum restricted-function difference is 4.93×10^{-14} (Figure 3A). DLinear's two-branch decomposition behaves differently; its null component remains visible under all three restriction families. The observed pattern across architectures and restriction families matches the operator criterion's prediction of representation exposure.

Exact refitting remains different after representation exposure vanishes. On the same 12 NLinear tasks, exact ridge solutions still differ from restricted full ridge solutions under both temporal masks. Each of the 24 task–restriction comparisons has an integrated absolute profile difference larger than 1% of full-model MSE (Figure 3B). These are deterministic solutions, so unfinished Adam optimization and minibatch noise cannot cause this discrepancy. For NLinear on ETTh2 at $H = 336$ with recent-lag restriction, the exact capacities are 192 for learning and 361 for restriction, while the invisible-component intervention has zero effect.

The remaining difference is training covariance. Let G be the training feature second-moment matrix and C the feature–target cross moment after subtracting any fixed residual forecast. The full ridge solution satisfies $(G + \lambda I)B_{\mathrm{F}} = C$. Let U_K embed the retained orthonormal coordinates, $P_K = U_K U_K^\top$, and $G_K = U_K^\top G U_K$. The independently learned retained coefficients $W_{\mathrm{L},K}$ solve $(G_K + \lambda I)W_{\mathrm{L},K} = U_K^\top C$.

Proposition 2 (Exact acquisition discrepancy)**.** *For $\lambda > 0$ and the two exact solutions above,*

$$U_K^\top B_{\mathrm{F}} - W_{\mathrm{L},K} = -(G_K + \lambda I)^{-1} U_K^\top G (I - P_K) B_{\mathrm{F}}. \tag{7}$$

Consequently, restriction and refitting have identical retained coefficients and complete restricted functions exactly when the rightmost coupling term vanishes.

The full fit allocates predictive responsibility between retained and discarded features. Once the latter are removed, the retained coefficients generally need to change. Orthogonal coordinate axes do not prevent this: training data can correlate those axes through G. Equation (7) specifies the target-weighted coupling that matters, rather than treating nonzero covariance alone as sufficient. Our baseline exact controls use $\lambda = 10^{-6}$; the maximum relative residual of the identity is 3.17×10^{-13}. Across seven regularization settings, including unregularized minimum-norm fitting, the exact linear controls retain all 24 baseline capacity separations and every NLinear temporal profile gap (Appendix D.2).

The two propositions address different operations. Proposition 1 asks whether two implementations of one function restrict identically. Proposition 2 asks whether optimizing with all features and then restricting agrees with optimizing only the retained ones. NLinear's temporal controls satisfy the first property but violate the second. This separation is the mechanism insight: representation exposure is a source of capacity change, but it is not the only source.

6 From prediction changes to capacity changes

The direction depends on the target residual. A nonzero prediction difference does not by itself specify which path is more accurate. For predictions f and $g = f + \delta$, define $e = f - y$. Below, squared norms and inner products are summed over all evaluation entries and divided by nHd, as in Equation (1). Squared loss gives

$$\widehat{R}(g) - \widehat{R}(f) = \underbrace{\widehat{\mathbb{E}}\|\delta\|^2}_{\text{perturbation energy}} + 2 \underbrace{\widehat{\mathbb{E}}\langle e, \delta\rangle}_{\text{residual alignment}} . \tag{8}$$

Norm measures how much the forecast changed; alignment determines whether that change corrects or reinforces an existing error. This identity connects the operator mechanisms to task risk. Equal-norm, opposite-direction origin interventions in FITS and MixLinear supply architecture-level controls for this distinction (Appendix E).

A fixed function can cross a capacity threshold. For the host–donor pair in Equation (6), let $B(\alpha) = B + \alpha(I - Q)(B' - B)$ and denote its restricted predictor by $f_{K,\alpha}$. The full function is unchanged; $\alpha = 0$ is the host and $\alpha = 1$ is the swap. With $\delta_K = f_{K,1} - f_{K,0}$, Equation (8) gives the exact response

$$\widehat{R}(f_{K,\alpha}) = \widehat{R}(f_{K,0}) + b_K\alpha + a_K\alpha^2, \quad a_K = \widehat{\mathbb{E}}\|\delta_K\|^2 \geq 0, \tag{9}$$

where $b_K = 2\widehat{\mathbb{E}}\langle f_{K,0} - y, \delta_K\rangle$. Capacity is the first budget whose quadratic response meets the unchanged target $(1 + \epsilon)\widehat{R}(f_{\mathrm{F}})$. Exposure determines the available prediction change, residual alignment determines its risk effect, and the threshold determines whether that effect moves capacity.

A capacity is a first hit, not a complete ordering. Once a profile is known, changing ϵ only changes where it is read. N-HiTS Weather at $H = 720$ illustrates the consequence: its pooled MSE capacity is $2/2$ at 1%, but all three seeds favor learning at every tested tolerance from 3.1% to 5%. Under MAE at 1%, the pooled pair is already $1/2$.

Grid resolution determines which adequate budgets are observed. With the target and retained risks fixed, thinning the grid shifts the first observed adequate budget later or leaves it unchanged. The shift depends on which points are removed as well as the largest gap. Our FITS study evaluates all 2,048 endpoint-preserving subgrids per horizon of a 13-point ladder (Appendix D). Together with the disconnected accepted sets of generic N-BEATS, these results motivate reporting the measured first hit alongside the observed adequate set.

Requiring adequacy at every subsequent measured budget gives a sustained-capacity readout. At the primary 1% MSE tolerance, all 24 same-function host interventions retain their original capacity transitions. Across the 75 core tasks, the MSE counts for $C_{\mathrm{L}} < C_{\mathrm{R}}$, $C_{\mathrm{L}} = C_{\mathrm{R}}$, and $C_{\mathrm{L}} > C_{\mathrm{R}}$ change from $(38, 36, 1)$ to $(37, 31, 7)$ under this readout (Appendix D.5). First attainment and sustained adequacy describe different properties of the acquired risk profiles.

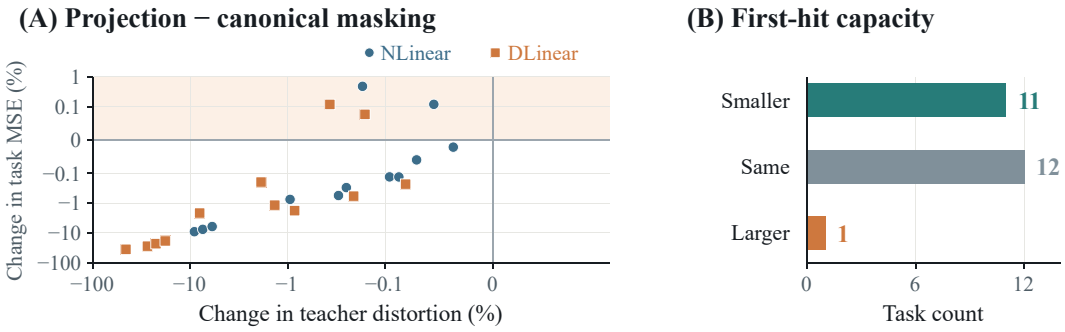


Figure 4: **Better teacher fidelity and better task retention are different outcomes.** Function projection versus canonical masking of the minimum-norm representation QB. (A) Projection-minus-masking changes in teacher-distortion and task-MSE areas, integrated over normalized budget and divided by host full MSE. Each point averages three seeds for one task; negative values improve the corresponding objective. Axes are linear within $\pm 0.1\%$ and logarithmic outside. (B) Task counts for smaller, equal, and larger first-hit capacity at $\epsilon = 1\%$. The primary Gram cutoff is 10^{-10}.

Optimization and robustness have distinct roles. Longer training preserves FITS's empirical capacity comparison and generic N-BEATS's first adequate trained depth on its sentinel grid; some intermediate N-BEATS holes change. Exact linear controls separate acquisition effects from incomplete convergence, while seed, tolerance, and metric analyses characterize variation in the capacity readout.

7 A CONSTRUCTIVE IMPLICATION: PRESERVE FUNCTIONS, EVALUATE TASKS

If coefficient restriction can expose invisible representation, one natural alternative is to restrict the *function*. For a fixed full teacher f_B and retained function class $\mathcal{F}_K$, the ideal projection fits

$$g_K = \arg\min_{g \in \mathcal{F}_K} \widehat{\mathbb{E}}_{\text{train}} \|g(x) - f_B(x)\|^2. \tag{10}$$

Our numerical implementation minimizes this objective over the empirical Gram eigenspace above the relative cutoff 10^{-10}, using its minimum-norm solution. This operation depends on the teacher's function on training inputs, not on its invisible coefficient component. We use function projection (Piironen et al., 2020; Hinton et al., 2015) to test whether removing representation dependence also preserves task capacity. The full endpoint is the teacher itself.

We compare this function projection with masking the minimum-norm equivalent representation QB (canonical masking). Both remove dependence on the original invisible component; projection additionally adapts the retained function to the training input distribution. Across the 24 NLinear/DLinear core tasks, projection reduces test teacher-distortion area in all 24 tasks. Task-MSE area improves in 20 and worsens in four (Figure 4A). Relative to each teacher's full-risk target, capacity becomes smaller in 11 tasks, stays equal in 12, and becomes larger in one (Figure 4B); the directions are consistent across seeds.

Across five relative Gram cutoffs from 10^{-12} to 10^{-8}, projection improves teacher-distortion area in all 24 tasks and worsens task-MSE area in the same four. Intermediate-budget risks can change substantially; at 1% tolerance, projection capacity differs from the 10^{-10} result in one task (Appendix F.4).

Equation (8) explains the divergence. Teacher distortion is the energy term around the full predictor; task error also includes alignment with the teacher's residual. On DLinear Exchange at $H = 336$, for example, projection at $K = 1$ decreases teacher distortion by 20.52% of full MSE while increasing task MSE by 34.91%, averaged over seeds. Conversely, DLinear on ETTh2 at $H = 720$ moves from capacity 32 to 1 even though its integrated task MSE increases slightly, by 0.077% of full MSE. A favorable first hit, a favorable average over budgets, and faithful teacher approximation are three different statements. Full results appear in Appendix F.

8 Related work and implications

Capacity, compression, and retraining. Task-relative threshold measurements connect to intrinsic-dimension studies, which measure the size of a subspace sufficient to reach a target performance (Li et al., 2018). Pruning research has long distinguished removing parameters, selecting subnetworks, and training compact models (LeCun et al., 1989; Hassibi & Stork, 1992; Han et al., 2015; Liu et al., 2019; Frankle & Carbin, 2019; Renda et al., 2020; Zhang & Papyan, 2024). Time-series foundation models also benefit from structured pruning before fine-tuning (Zhao et al., 2025). Rank-based analyses study both post-training compression and pretraining low-rank architectures (Yu et al., 2026). Our object is the first adequate budget on two explicitly matched forecasting paths. Holding the full function fixed, we identify capacity changes caused by its trained invisible representation; kernel-preserving restrictions then isolate a separate exact training-coupling mechanism. Sets of near-optimal models provide another perspective on the existence of simpler predictors (Semenova et al., 2022); a path capacity asks which of these predictors a specified acquisition procedure actually reaches.

Function equivalence and reduction. Function-preserving transformations are useful for constructing networks and studying their geometry (Chen et al., 2016; Dinh et al., 2017). Functional dimension and hidden-symmetry studies distinguish model functions from their parameterizations (Grigsby et al., 2023; 2025; Phuong & Lampert, 2020; Shen, 2024). SliceGPT exploits computational invariance to rotate representations before dimensionality reduction (Ashkboos et al., 2024). Function-based reduction studies whether selection scores and retained sets are invariant to equivalent implementations (Asiaee, 2026). Here the structural ladder and restriction rule remain fixed: transferring actual trained invisible components changes restricted risk and capacity. A kernel-preserving control then separates this representation effect from exact refitting. This connects function equivalence to the structural budget required to reach a forecasting task's accuracy target.

Forecasting structure and the full-function view. Linear forecasting can be competitive despite architectural differences (Zeng et al., 2023), and several linear architectures share function classes that admit closed-form optimization (Toner & Darlow, 2024). Frequency-based predictors (Xu et al., 2024; Ma et al., 2026) and structured residual predictors (Oreshkin et al., 2020; Challu et al., 2023) offer interpretable ways to vary the retained structure. We use those structures to ask a different question from full-model benchmarking. Full-function equivalence is a valuable control, but it does not imply equivalent restricted behavior unless the restriction respects that equivalence.

What should change in capacity evaluation? Three pieces of information make a capacity interpretable: the target accuracy, the structural ladder, and how the model at each budget is acquired. Report the first adequate budget together with the nearby risk profile; distinguish learning a compact predictor from retaining a full predictor; and evaluate task error separately from teacher fidelity. These choices identify the scientific quantity being compared and make its acquisition path explicit. Our controlled characterization is exact for linear realizations; the broader public-model results establish the operational phenomenon in nonlinear forecasting structures. Extending the same intervention-level identification to nonlinear symmetries is a natural next step.

9 Conclusion

The structure needed to retain forecasting accuracy is not fixed by the full predictor alone. Acquisition changes capacity through two distinguishable channels: a restriction can expose functionally invisible representation, and refitting can compensate for training-covariance coupling with discarded features. Same-function transfers and exact negative controls make these channels experimentally separable. The fixed-function risk response connects representation changes to capacity thresholds, while residual alignment explains why preserving a teacher and retaining task accuracy can diverge.

AI USE STATEMENT

Generative AI tools assisted with manuscript drafting and language revision, literature retrieval and discovery, experimental code development and execution, and analysis of experimental results. The authors take responsibility for the final manuscript, including its claims, results, citations, and code.

REFERENCES

Saleh Ashkboos, Maximilian L. Croci, Marcelo Gennari do Nascimento, Torsten Hoefler, and James Hensman. SliceGPT: Compress large language models by deleting rows and columns. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=vXxardq6db.

Amir Asiaee. Causal mechanism reduction: Mechanism replacement for neural network pruning and abstraction, 2026. URL https://arxiv.org/abs/2602.24266v2. Version 2.

Cristian Challu, Kin G. Olivares, Boris N. Oreshkin, Federico Garza Ramirez, Max Mergenthaler-Canseco, and Artur Dubrawski. NHITS: Neural hierarchical interpolation for time series forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6989–6997, 2023. doi: 10.1609/aaai.v37i6.25854. URL https://ojs.aaai.org/index.php/AAAI/article/view/25854.

Tianqi Chen, Ian Goodfellow, and Jonathon Shlens. Net2Net: Accelerating learning via knowledge transfer. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1511.05641v4.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1019–1028. PMLR, 2017. URL https://proceedings.mlr.press/v70/dinh17b.html.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJl-b3RcF7.

Elias Frantar, Sidak Pal Singh, and Dan Alistarh. Optimal brain compression: A framework for accurate post-training quantization and pruning. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/1caf09c9f4e6b0150b06a07e77f2710c-Abstract-Conference.html.

Elisenda Grigsby, Kathryn Lindsey, and David Rolnick. Hidden symmetries of ReLU networks. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 11734–11760. PMLR, 2023. URL https://proceedings.mlr.press/v202/grigsby23a.html.

J. Elisenda Grigsby, Kathryn Lindsey, Robert Meyerhoff, and Chenxi Wu. Functional dimension of feedforward ReLU neural networks. *Advances in Mathematics*, 482:110636, 2025. doi: 10.1016/j.aim.2025.110636. URL https://doi.org/10.1016/j.aim.2025.110636.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1832–1841. PMLR, 2018. URL https://proceedings.mlr.press/v80/gunasekar18a.html.

Song Han, Jeff Pool, John Tran, and William J. Dally. Learning both weights and connections for efficient neural networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/ae0eb3eed39d2bcef4622b2499a05fe6-Abstract.html.

Babak Hassibi and David G. Stork. Second order derivatives for network pruning: Optimal brain surgeon. In *Advances in Neural Information Processing Systems*, volume 5, pp. 164–171, 1992. URL https://proceedings.neurips.cc/paper/1992/hash/303ed4c69846ab36c2904d3ba8573050-Abstract.html.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long- and short-term temporal patterns with deep neural networks. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 95–104. Association for Computing Machinery, 2018. doi: 10.1145/3209978.3210006. URL https://doi.org/10.1145/3209978.3210006.

Yann LeCun, John S. Denker, and Sara A. Solla. Optimal brain damage. In *Advances in Neural Information Processing Systems*, volume 2, pp. 598–605, 1989. URL https://proceedings.neurips.cc/paper/1989/hash/6c9882bbac1c7093bd25041881277658-Abstract.html.

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=ryup8-WCW.

Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, and Trevor Darrell. Rethinking the value of network pruning. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJlnB3C5Ym.

Aitian Ma, Dongsheng Luo, and Mo Sha. MixLinear: Extreme low resource multivariate time series forecasting with 0.1K parameters. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/7db2348b5bfeca620aa7327df815adcc-Abstract-Conference.html.

Boris N. Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1ecqn4YwB.

Mary Phuong and Christoph H. Lampert. Functional vs. parametric equivalence of ReLU networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Bylx-TNKvH.

Juho Piironen, Markus Paasiniemi, and Aki Vehtari. Projective inference in high-dimensional problems: Prediction and feature selection. *Electronic Journal of Statistics*, 14(1):2155–2197, 2020. doi: 10.1214/20-EJS1711. URL https://doi.org/10.1214/20-EJS1711.

Alex Renda, Jonathan Frankle, and Michael Carbin. Comparing rewinding and fine-tuning in neural network pruning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1gSj0NKvB.

Lesia Semenova, Cynthia Rudin, and Ronald Parr. On the existence of simpler machine learning models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1827–1858. Association for Computing Machinery, 2022. doi: 10.1145/3531146.3533232. URL https://doi.org/10.1145/3531146.3533232.

Guohao Shen. Exploring the complexity of deep neural networks through functional equivalence. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 44643–44664. PMLR, 2024. URL https://proceedings.mlr.press/v235/shen24a.html.

William Toner and Luke Nicholas Darlow. An analysis of linear time series forecasting models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 48404–48427. PMLR, 2024. URL https://proceedings.mlr.press/v235/toner24a.html.

Zhijian Xu, Ailing Zeng, and Qiang Xu. FITS: Modeling time series with 10k parameters. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/701251e1db4a2e4dd2ef23f5265d5936-Abstract-Conference.html.

Annan Yu, Danielle Maddix, Boran Han, Xiyuan Zhang, Abdul Fatir Ansari, Oleksandr Shchur, Christos Faloutsos, Andrew Gordon Wilson, Michael W. Mahoney, and Bernie Wang. Understanding transformers for time series: Rank structure, flow-of-ranks, and compressibility. In *International Conference on Learning Representations*, pp. 42449–42489, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/467655d26fcc207bca08915dc91964c6-Abstract-Conference.html.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9):11121–11128, 2023. doi: 10.1609/aaai.v37i9.26317. URL https://ojs.aaai.org/index.php/AAAI/article/view/26317.

Stephen Zhang and Vardan Papyan. Sparsest models elude pruning: An exposé of pruning's current capabilities. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 59576–59600. PMLR, 2024. URL https://proceedings.mlr.press/v235/zhang24av.html.

Lifan Zhao, Yanyan Shen, Zhaoyang Liu, Xue Wang, and Jiaji Deng. Less is more: Unlocking specialization of time series foundation models via structured pruning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 171877–171902, 2025. doi: 10.52202/085713-5720. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/facaf651f417d3e4afaf23c579e701e6-Abstract-Conference.html.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):11106–11115, 2021. doi: 10.1609/aaai.v35i12.17325. URL https://ojs.aaai.org/index.php/AAAI/article/view/17325.

A EXPERIMENTAL DEFINITIONS AND IMPLEMENTATION

A.1 DATASETS, PREDICTION TARGETS, AND AGGREGATION

The core matrix uses ETTh2, Weather, Electricity, and Exchange at forecast horizons 96, 336, and 720, with lookback 720. All numeric channels are forecast, excluding the timestamp; the target channel OT is last when present. Each channel is standardized using only its training rows. For ETTh2, target segments are rows $[0, 8640)$, $[8640, 11520)$, and $[11520, 14400)$ for training, validation, and testing. For the other core datasets with N rows, training ends at $\lfloor 0.7N \rfloor$, testing comprises the final $\lfloor 0.2N \rfloor$ rows, and validation fills the interval between them. The ETT benchmark and its month-based split originate with Informer (Zhou et al., 2021); Electricity and Exchange are also used by LSTNet (Lai et al., 2018). The $70/10/20$ splits here specify our evaluation, not LSTNet's original split. Weather is the 21-variable, 10-minute Jena weather dataset.[1]

The ETTh1 extension evaluates NLinear, DLinear, FITS, and MixLinear at $H \in \{96, 192, 336, 720\}$ with lookback 720. ETTh1 uses the same training, validation, and test row intervals as ETTh2, with training-only standardization and all numeric channels as prediction targets.

Forecast origins have stride one. A target window is used only when all H future rows lie in its target segment; validation and test inputs may use earlier observations as context. Metrics average all eligible origins, channels, and forecast leads. MSE and MAE are on the training-standardized scale. A pooled risk averages the three seed risks at each budget; the pooled full reference is averaged in the same way. Neither a pooled first hit nor its direction is obtained by voting on seeds.

The external same-function interventions use ETTm1 and Traffic at $H = 336$. ETTm1 uses the corresponding $12/4/4$-month segments at 15-minute resolution: $[0, 34560)$, $[34560, 46080)$, and $[46080, 57600)$. Traffic uses the same chronological $70/10/20$ convention as the non-ETT core tasks. These tasks extend the controlled linear mechanism test beyond its original four datasets.

A.2 RESTRICTION FAMILIES AND ORIGINS

Each comparison fixes the implemented architecture and prediction target while varying specified coordinates or executed blocks. The following grids define the complete evaluated restriction families.

FITS. The shared-channel public FITS model retains its normalization and complex frequency interpolation. K keeps the first K input real-FFT bins and the induced first $\lfloor K(720 + H)/720 \rfloor$ output bins. Its grid is

$$\{1, 2, 6, 14, 34, 81, 96, 112, 128, 144, 160, 176, 196\}.$$

The primary restriction sets inactive complex weights to zero, matching the learning path. A separate origin control restores inactive coefficients to their matched full-model initialization. No $K = 0$ measurement is introduced. The full endpoint is $K = 196$ for all evaluated horizons.

MixLinear. The grid is $\{1, 2, 3, 5, 8, 12, 19\}$ with the research configuration $\mathrm{lpf} = 19$ and $\alpha = 0.5$. K restricts frequency-prefix columns; the complete time branch, convolution, and rank-two output factor remain trainable. It is not a dimension of the complete function class. The primary restriction zeros inactive columns; an additional path restores their matched initializer. Period is 4 on Weather and 24 on the other tasks. The segment DFT length is consequently 180 on Weather and 30 elsewhere, so a common K does not denote an identical physical frequency across datasets.

NLinear and DLinear. The models share coefficients across channels. NLinear keeps its last-value forecast and centers inputs by the last value. DLinear keeps its two-branch, kernel-25 moving-average decomposition with endpoint replication. Orthonormal real Fourier coordinates are ordered DC, complete cosine–sine pairs, and Nyquist. The grid is

$$\{0, 1, 2, 4, 8, 16, 32, 64, 96, 128, 192, 256, 361\}.$$

For $K > 0$, each branch keeps $d_K = \min(720, 2K - 1)$ real lag coordinates and its bias. Both DLinear branches use the same mask; their biases remain distinct. $K = 0$ is the native untrained

[1] Station and data archive: `https://www.bgc-jena.mpg.de/wetter/`.

origin: last-value persistence for NLinear and zero for DLinear. Restrictions act *after* native centering/decomposition. Recent-lag and oldest-lag controls retain the corresponding d_K processed coordinates at the same grid indices. They do not crop the raw series and recompute preprocessing.

Interpretable N-BEATS. The public interpretable design has a degree-three polynomial trend block (width 256) and a Fourier seasonality block (width 2048), each with four ReLU layers and three applications of shared weights. The original reversed residual inputs, backcast updates, input mask, and last-value origin are kept. K retains DC and the first K positive harmonics in all four seasonality head groups: forecast/backcast, cosine/sine. All six residual steps execute; the trend block is unrestricted. The grid is the sorted unique clipping of $\{0, 1, 2, 4, 8, 16, 32, 64\}$ to the maximum, with that maximum included: 47, 167, and 359 for the three horizons. Here $K = 0$ is a trained trend-plus-DC model, not persistence alone. Restriction reruns the residual chain with the masked heads rather than projecting its final forecast.

Generic N-BEATS. Thirty generic blocks each have distinct weights and four ReLU layers of width 512. The executed-prefix grid on ETTh2 is $\{0, 1, 2, 3, 4, 5, 6, 8, 10, 14, 19, 24, 30\}$. A prefix forecast executes blocks 1 through K. $K = 0$ is native last-value persistence; $K = 30$ reuses the selected full model.

N-HiTS. The public three-block coarse-to-fine model uses pooling kernels $(2, 2, 1)$, forecast downsampling $(4, 2, 1)$, the source-defined $[512, 512]$ MLP in each stack, ReLU, and linear interpolation. There is no dropout or weight sharing. The executed-block grid is $\{0, 1, 2, 3\}$. Its zero budget is the same native last-value origin on both paths; the full endpoint is the identical selected three-block model.

A.3 Training and selection

At every fitted budget, the retained parameters start from a common full-shaped initializer for that seed. Inactive coefficients remain zero in the primary coefficient-mask comparison; block-prefix fits optimize only executed blocks. Validation governs early stopping and checkpoint selection, with test errors used for evaluation. Each architecture and task uses a fixed training rule across fitted budgets and acquisition paths. For restriction, the selected full-model weights are held fixed while the declared coefficient masks or block prefixes are applied.

FITS uses Adam with learning rate 5×10^{-4}, batch size 128, maximum 50 epochs per stage, and validation patience 20. At $H = 96$, whole-window reconstruction is followed by forecast-only fine-tuning; the other horizons use forecast-only training.

MixLinear uses its epochwise learning-rate adjustment and initialization convention, Adam, no weight decay, and no gradient clipping. The fixed dataset recipes are in Table 2. Its original scheduler construction sets the initial rate to the nominal rate divided by 25 and the first-moment coefficient to 0.95; that scheduler is not stepped. Training and validation use shuffled complete batches. Validation averages batch MSE and updates the selected state on ties. Test evaluation includes every complete target window. The original ETTh2 $H = 96/720$ results and the completed core extension use the same frequency-prefix definition.

Table 2: Fixed MixLinear recipes for the core comparison and ETTh1 extension. The frequency configuration is $\mathrm{lpf} = 19$, $\alpha = 0.5$ throughout.

Dataset	Period	Batch	Nominal LR	Epoch limit	Patience
ETTh1	24	128	0.02	30	5
ETTh2, $H = 96$	24	256	0.02	30	5
ETTh2, $H = 336$	24	128	0.02	30	5
ETTh2, $H = 720$	24	310	0.02	30	5
Weather	4	64	0.02	30	10
Electricity	24	64	0.03	30	10
Exchange	24	32	0.02	30	5

NLinear and DLinear use full-batch Adam on the same quadratic ridge objective as their baseline exact controls, with $\lambda = 10^{-6}$ including biases. Training has learning rate 10^{-3}, at most 5000 updates, halving at 1700 and 3400, and gradient-norm clipping at 1. Validation occurs every 100 updates with patience 20 checks. Moments include all training window–channel pairs. Exact ridge solves use the same objective; minimum-norm unregularized controls use an eigendecomposition with relative cutoff 10^{-10}. Exact controls have no checkpoint search or replicated optimizer seeds.

Interpretable N-BEATS uses Adam at 10^{-3}, no weight decay, gradient clipping at 1, and a maximum of 5000 steps. Each step samples 256 training window–channel pairs with replacement; the same sample stream is used across budgets within a seed. Learning rate halves at steps 1700 and 3400. Complete validation is evaluated every 100 steps, with patience 20. Generic N-BEATS and N-HiTS use Adam at 10^{-3}, no weight decay, clipping at 1, and batches of 128 windows; the maximum is 50 epochs, patience 20, and learning rate halves after epochs 17 and 34.

The full-model risk defines each accuracy target. A capacity of zero means that the declared zero-budget model already meets that target. The full tables include such cases in several ETTh2/Exchange nonlinear comparisons. The native origins and interpretable N-BEATS's trained trend-plus-DC zero rung are specified above.

B Proofs and mathematical details

B.1 Realizations, empirical risk, and capacity

We make explicit the distinction between an architectural realization and an empirical design matrix. Let $B \in \mathbb{R}^{D\times H}$ contain the trainable coefficients, and let $A \in \mathbb{R}^{p\times D}$ map each output column to its complete correction function in a fixed raw feature basis. The prediction is

$$f_B = r + ZAB, \tag{11}$$

where r is a fixed, nonlearned origin. In the affine forecasting examples, $p = L + 1$ and a row of Z is $[x, 1]$. Equality $AB = AB'$ means equality of the complete correction functions, not merely agreement on an observed sample. The origin is the same for all compared realizations.

For a split s, stack its window–channel pairs into n_s rows and set

$$X_s = Z_s A, \qquad \widetilde{Y}_s = Y_s - r_s, \qquad G_s = \frac{X_s^\top X_s}{n_s}, \qquad C_s = \frac{X_s^\top \widetilde{Y}_s}{n_s}, \qquad y_s^2 = \frac{\|\widetilde{Y}_s\|_F^2}{n_s}. \tag{12}$$

Thus $G_s \in \mathbb{R}^{D\times D}$, $C_s \in \mathbb{R}^{D\times H}$, and

$$\widehat{R}_s(B) = \frac{\|X_s B - \widetilde{Y}_s\|_F^2}{n_s H} = \frac{\operatorname{tr}(B^\top G_s B) - 2\langle B, C_s\rangle_F + y_s^2}{H}. \tag{13}$$

The Frobenius inner product is $\langle V, W\rangle_F = \operatorname{tr}(V^\top W)$. These are empirical risks on a fixed split. Window overlap and channel stacking do not make the n_s rows independent statistical replicates.

A retained coefficient space is represented by an isometric embedding $U_K \in \mathbb{R}^{D\times m_K}$, with $U_K^\top U_K = I$ and $P_K = U_K U_K^\top$. A compact coefficient matrix W realizes $U_K W$ in the ambient coordinates. We use $P_0 = 0$ for an origin rung when present and $P_{K_{\max}} = I$ for the full rung. The index K is a declared budget index; m_K, $\operatorname{rank}(AU_K)$, and deployment cost are distinct quantities.

For a fixed finite ordered grid $\mathcal{K}$, path B_K^a, and positive reference risk R_{ref}, define

$$\widehat{\mathcal{C}}_{\epsilon,s}^{\,a} = \min\big\{K \in \mathcal{K} : \widehat{R}_s(B_K^a) \le (1+\epsilon)R_{\mathrm{ref}}\big\}. \tag{14}$$

An empty accepted set is reported as not attained. This definition does not assume monotonicity of an achieved-risk path. In a within-host intervention, $R_{\mathrm{ref}} = \widehat{R}_s(B)$ is the unchanged full-host risk.

B.2 When restriction descends to the function quotient

Let $Q = A^\dagger A$ be the Euclidean orthogonal projector onto $\operatorname{range}(A^\top)$. Each coefficient matrix has the decomposition $B = QB + (I - Q)B$, and $A(I - Q)B = 0$ columnwise.

Proposition: function-consistent restriction. For a fixed coefficient projection P, the following conditions are equivalent:

1. $AB = AB'$ implies $APB = APB'$ for all coefficient matrices B, B'.
2. $P\ker(A) \subseteq \ker(A)$.
3. $AP(I - Q) = 0$.
4. There is a linear map $T_P : \mathrm{im}(A) \to \mathrm{im}(A)$ such that $AP = T_P A$.

Proof. It suffices to argue columnwise. Applying the first condition to $b = n$ and $b' = 0$ for $n \in \ker(A)$ gives $APn = 0$, proving the second condition. Conversely, $Ab = Ab'$ implies $b - b' \in \ker(A)$, and hence $AP(b - b') = 0$ under the second condition. Since $I - Q$ projects onto $\ker(A)$, the second and third conditions are equivalent. Under the second condition, define $T_P(Ab) = APb$. If $Ab = Ab'$, the same condition gives $APb = APb'$; therefore the definition is independent of the representative and is linear. Its output belongs to $\mathrm{im}(A)$. Conversely, the factorization implies $APn = T_P An = 0$ for every $n \in \ker(A)$. One explicit expression on $\mathrm{im}(A)$ is $T_P = APA^\dagger$. This map is unique on $\mathrm{im}(A)$.

The criterion determines whether restriction is well defined on function equivalence classes and specifies compatible and incompatible controls. When it fails, the realized risk and capacity effects depend on the trained coefficients, input distribution, and threshold; these effects can be zero.

Corollary: capacities under compatible restriction. If every P_K satisfies the proposition, then equivalent complete predictors have identical restricted predictions on every split. With the same loss and full reference, their accepted-budget sets and empirical capacities are equal. The converse need not hold for a fixed split and threshold: a scalar first hit can remain unchanged even when restricted predictions differ.

B.3 Function-preserving transplantation and its capacity consequences

Fix a host B_h and donor B_d for the same architectural map and coefficient metric. Define

$$B_{h\leftarrow d} = QB_h + (I - Q)B_d. \tag{15}$$

Then

$$AB_{h\leftarrow d} = AB_h, \tag{16}$$

$$AP_K(B_{h\leftarrow d} - B_h) = AP_K(I - Q)(B_d - B_h). \tag{17}$$

These equalities follow from $AQ = A$ and $A(I - Q) = 0$. They preserve the host function, without requiring the donor and host to have equal complete functions. Therefore the full-host risk and the threshold in Equation (14) are unchanged.

For clarity about the empirical implication, let the original host capacity be c. If a transplanted path has an accepted rung $K' < c$, its capacity is at most $K' < c$. If every rung $K \le c$ is rejected after transplantation, its capacity is larger than c, or is not attained. These statements require the indicated actual risk inequalities, not only a nonzero operator in Equation (17). In the boundary-preserving path, the full rung remains accepted for $\epsilon \ge 0$, so the latter capacity is attained. No monotonicity assumption is needed.

Proposition: fixed-function risk response. Let $B_c = QB$, $N = (I-Q)B$, and $B(\alpha) = B_c + \alpha N$. For each fixed split and budget, write

$$a_{s,K} = \frac{\langle P_K N, G_s P_K N\rangle_F}{H}, \tag{18}$$

$$b_{s,K} = \frac{2\langle P_K N, G_s P_K B_c - C_s\rangle_F}{H}, \tag{19}$$

$$c_{s,K} = \widehat{R}_s(P_K B_c). \tag{20}$$

Then

$$\widehat{R}_s(P_K B(\alpha)) = a_{s,K}\alpha^2 + b_{s,K}\alpha + c_{s,K}, \qquad a_{s,K} \ge 0. \tag{21}$$

The complete function and its full risk are independent of α.

Proof. The function statement follows from $AN = 0$. Substitute $P_K B_c + \alpha P_K N$ into Equation (13) and collect quadratic, linear, and constant terms. Since $G_s \succeq 0$, the quadratic coefficient is nonnegative. Equivalently, it is $\|X_s P_K N\|_F^2/(n_s H)$. If this coefficient is zero, the empirical prediction perturbation is zero and the linear coefficient is also zero.

For a nonzero empirical perturbation, the sign of $b_{s,K}$ is unrestricted. Thus risk can decrease and then increase along α, despite convexity. Accepted rungs are determined by the quadratic inequalities

$$a_{s,K}\alpha^2 + b_{s,K}\alpha + c_{s,K} \le (1+\epsilon)\widehat{R}_s(B). \tag{22}$$

The first accepted rung can change while the reference stays fixed. A nonzero architectural perturbation can nevertheless have zero effect on a particular sample, or an effect too small to change any accepted rung.

For a host B_h and donor B_d, set $N = (I - Q)(B_d - B_h)$ instead and take $B(\alpha) = B_h + \alpha N$. The same expansion yields Equation (9), with $\delta_K = X_s P_K N$ and empirical averages normalized by $n_s H$. The linear coefficient is the alignment with the restricted host residual; the quadratic coefficient is the prediction-change energy. The cases $\alpha = 0$ and $\alpha = 1$ are the host and its transplantation. All values of α share the same full function and accuracy target.

B.4 Architectural examples of the compatibility condition

The following formulas use row-vector inputs. Let $F \in \mathbb{R}^{L\times L}$ be the real orthogonal Fourier basis used for the coefficient atlas, and let e_L be the last standard basis vector.

NLinear. Last-value centering is xT, where $T = I - e_L \mathbf{1}^\top$. Including its learned bias, the correction map is

$$A = \begin{bmatrix} TF & 0 \\ 0^\top & 1 \end{bmatrix}, \qquad q = \begin{bmatrix} F^\top e_L \\ 0 \end{bmatrix}. \tag{23}$$

Here $\ker(A) = \operatorname{span}\{q\}$, $\|q\|_2 = 1$, and $\operatorname{rank}(A) = L$. The persistence origin is unchanged and is not a trainable component of A. For a temporal mask D diagonal in the centered-lag coordinates, the Fourier coefficient projection is $F^\top D F$. Since De_L is either e_L or zero, the masked q is either q or zero. Both recent-lag and oldest-lag masks therefore preserve the architectural kernel, independently of training data.

For a nonzero Fourier prefix containing DC and complete sine/cosine pairs, let d be its number of retained real lag coefficients, and let $\Pi = F P_{\text{lag}} F^\top$. The complete-pair ordering gives $e_L^\top \Pi e_L = d/L$ and $\mathbf{1}^\top \Pi e_L = 1$. Because Π is an orthogonal projector,

$$\|T\Pi e_L\|_2^2 = \|\Pi e_L - e_L\|_2^2 = 1 - d/L. \tag{24}$$

Consequently, for a multi-output null component $N = qa^\top$,

$$\|AP_K N\|_F^2 = (1 - d/L)\|a\|_2^2. \tag{25}$$

This is an all-input operator norm, not an empirical prediction-energy or task-risk formula. The origin rung is zero separately, and the full rung has $d = L$ and zero leakage.

DLinear. Let xM be the native moving-average trend, and let W_s, W_t be the seasonal and trend coefficient matrices in lag coordinates. Its effective lag map is $W_{\text{eff}} = (I-M)W_s + MW_t$. Applying the same lag projection Π to each branch gives

$$W_{\text{mask}} = (I - M)\Pi W_s + M\Pi W_t \tag{26}$$

$$= \Pi W_{\text{eff}} + (M\Pi - \Pi M)(W_t - W_s). \tag{27}$$

The second equality is obtained by expanding both terms on the right. A complete-function-null branch perturbation can therefore be exposed by the commutator. Opposing bias perturbations still cancel when both biases are retained. A circulant moving average commutes with the complete-frequency Fourier projectors; it need not commute with a temporal diagonal mask. Accordingly, the zero prediction is specific to an operator–restriction pair, not to the name of the architecture alone.

B.5 Exact ridge acquisition and retained–discarded coupling

In this subsection $G = G_{\mathrm{tr}}$, $C = C_{\mathrm{tr}}$, and

$$J(B) = \widehat{R}_{\mathrm{tr}}(B) + \frac{\lambda}{H}\|B\|_F^2, \qquad \lambda > 0. \tag{28}$$

The penalty is on all retained coefficients, including the declared biases. The orthogonal embedding preserves this penalty since $\|UW\|_F = \|W\|_F$. Suppressing the budget subscript, define

$$B_F = (G + \lambda I)^{-1}C, \quad g = U^\top G U, \quad W_L = (g + \lambda I)^{-1}U^\top C, \quad W_U = U^\top B_F, \quad B_L = UW_L. \tag{29}$$

Proposition: acquisition–restriction identity. The two compact coefficient paths obey

$$W_L - W_U = (g + \lambda I)^{-1}U^\top G(I - P)B_F. \tag{30}$$

They are equal if and only if $U^\top G(I - P)B_F = 0$. If V is an orthogonal complement of U, this condition is $(U^\top GV)(V^\top B_F) = 0$. The stronger condition $U^\top GV = 0$ is sufficient for all right-hand sides; a nonzero cross block alone does not force a difference for the present task.

Proof. The retained block of $(G + \lambda I)B_F = C$ is

$$(g + \lambda I)W_U + U^\top G(I - P)B_F = U^\top C. \tag{31}$$

Subtract $(g + \lambda I)W_L = U^\top C$ and invert the positive definite retained matrix. The penalty creates no off-diagonal block because $U^\top V = 0$.

For these exact isotropically regularized least-squares solutions, the criterion also characterizes equality of training predictions and complete restricted functions. To see this, $C \in \mathrm{range}(G)$ columnwise implies $B_F \in \mathrm{range}(X_{\mathrm{tr}}^\top)$. Consequently, $W_U \in \mathrm{range}((X_{\mathrm{tr}}U)^\top)$. The restricted normal equation similarly places W_L in that same row space. Their difference can therefore be annihilated by $X_{\mathrm{tr}}U$ only when it is zero. Complete function equality implies this empirical equality, and coefficient equality implies both. For arbitrary or inexact coefficients the general conditions $AU(W_U - W_L) = 0$ and $X_{\mathrm{tr}}U(W_U - W_L) = 0$ need not imply coefficient equality. Equality of scalar risks, in either setting, is weaker than equality of predictions. Compatibility with the architectural kernel and coupling under G concern different operators.

Finite-solution residuals. For fixed, possibly inexact coefficients, set

$$r_F = (G + \lambda I)B_F - C, \qquad r_L = (g + \lambda I)W_L - U^\top C. \tag{32}$$

With $W_U = U^\top B_F$ the exact algebraic identity is

$$(g + \lambda I)(W_U - W_L) = -U^\top G(I - P)B_F + U^\top r_F - r_L. \tag{33}$$

The identity separates retained–discarded coupling from finite-solution residual contributions for the fixed coefficients.

Objective and risk are different quantities. For $\delta = W_U - W_L$, direct expansion gives

$$J(UW_U) - J(UW_L) = \frac{\mathrm{tr}\{\delta^\top(g + \lambda I)\delta\} + 2\langle \delta, r_L\rangle_F}{H}. \tag{34}$$

For the exact restricted optimum $r_L = 0$, the gap is nonnegative. Even then, the unpenalized training-risk difference subtracts $\lambda(\|W_U\|_F^2 - \|W_L\|_F^2)/H$ from that objective gap and need not have the same sign. No sign for validation or test risk follows from training optimality. The exact ridge paths have the same origin and full endpoint; changing an isometric coordinate basis does not change the full ridge solution.

B.6 Prediction components, interaction, and task risk

For any baseline coefficient B_0 and perturbation Δ, Equation (13) yields

$$\widehat{R}_s(B_0 + \Delta) - \widehat{R}_s(B_0) = \frac{\langle \Delta, G_s \Delta \rangle_F}{H} + \frac{2\langle \Delta, G_s B_0 - C_s \rangle_F}{H}. \tag{35}$$

The first term is squared prediction-change energy; the second is alignment with the baseline residual, defined as prediction minus target.

For a full acquisition B_A compared with a full ridge reference B_F and restricted ridge B_L, write $D = B_A - B_F$ and

$$d_c = PB_F - B_L, \qquad d_v = PQD, \qquad d_n = P(I - Q)D. \tag{36}$$

They sum exactly to $PB_A - B_L$. The first is the fixed ridge acquisition–restriction difference. The other two are obtained by splitting the full-acquisition difference $D = B_A - B_F$ into structurally visible and hidden components and then applying P.

Define $E_{ij} = \langle d_i, G_s d_j \rangle_F / H$ and $a_i = 2\langle d_i, G_s B_L - C_s \rangle_F / H$, with $i, j \in \{c, v, n\}$. Substituting $\Delta = \sum_i d_i$ in Equation (35) proves

$$\widehat{R}_s(PB_A) - \widehat{R}_s(B_L) = \sum_i E_{ii} + 2\sum_{i<j} E_{ij} + \sum_i a_i. \tag{37}$$

The 3×3 matrix E is a Gram matrix and is positive semidefinite, but its off-diagonal entries and the residual alignments can have either sign. In particular, the diagonal energies are not additive causal shares of task risk. With $E_{\mathrm{diag}} = \sum_i E_{ii}$ and $E_{\mathrm{total}} = \sum_{ij} E_{ij}$,

$$0 \le E_{\mathrm{total}} \le 3E_{\mathrm{diag}}. \tag{38}$$

The upper bound follows from $\|v_c + v_v + v_n\|^2 \le 3\sum_i \|v_i\|^2$ for the empirical prediction perturbations. If $E_{\mathrm{diag}} > 0$, their ratio measures cancellation or reinforcement and need not lie below one. If the denominator is zero, the ratio is undefined, even though the total energy is also zero.

At the origin all three components vanish. At the full rung $d_c = 0$ and $Ad_n = 0$, while Ad_v can remain nonzero because the two full acquisitions need not have the same function. Dividing these restricted-path comparison terms by the host's full risk expresses them on a dimensionless scale relative to that reference.

B.7 Teacher projection: invariance, optimization, and limits

Fix a complete teacher B and the training-input geometry. For a retained space, set $g = U^\top G_{\mathrm{tr}} U$, $h = U^\top G_{\mathrm{tr}} B$ and $c = U^\top C_{\mathrm{tr}}$. Teacher distortion on a split is

$$D_s(W; B) = \frac{\|X_s(UW - B)\|_F^2}{n_s H} = \frac{\langle UW - B, G_s(UW - B) \rangle_F}{H}. \tag{39}$$

An unpenalized minimum-norm training projection is $W_T = g^\dagger h$. Indeed, h is in $\mathrm{range}(g)$: every $v \in \ker(g)$ has $X_{\mathrm{tr}} U v = 0$, and therefore $v^\top h = 0$. This also holds columnwise for c. The matrix h is constructed from teacher predictions and training inputs.

Representation invariance. If $AN = 0$, then $X_s N = Z_s AN = 0$ and $G_s N = 0$ on every split. Replacing the teacher by $B + N$ leaves h unchanged. Consequently every fixed linear solver applied to this right-hand side, including minimum-norm and ridge projection, returns the same interior coefficients. Empirical invisibility alone is not enough for this all-input statement: $G_{\mathrm{tr}} N = 0$ does not imply $AN = 0$.

Numerical rank and positive regularization. For a spectral cutoff τ, let V_τ contain the eigenvectors of g with eigenvalues above τ, and define

$$Q_\tau = V_\tau V_\tau^\top, \qquad g_\tau^\dagger = V_\tau \,\mathrm{diag}(\sigma_i^{-1}) V_\tau^\top, \qquad W_{T,\tau} = g_\tau^\dagger h. \tag{40}$$

This is the minimum-norm optimum in the declared retained eigenspace. Its ideal-arithmetic normal equation is $gW_{T,\tau} = Q_\tau h$, not necessarily $gW_{T,\tau} = h$. The architectural projector $Q = A^\dagger A$ and

empirical Q_τ are different objects. Independent numerical rank cutoffs need not preserve exact kernel inclusions or ideal nested-OLS identities.

The fixed ridge alternative $W_{T,\lambda} = (g + \lambda I)^{-1}h$ minimizes $D_{\mathrm{tr}}(W; B) + \lambda\|W\|_F^2/H$. For each positive empirical eigenvalue σ_i, unpenalized and ridge projection use multipliers $1/\sigma_i$ and $1/(\sigma_i+\lambda)$, respectively, so positive regularization changes the spectral response even on a common support.

For any computed teacher coefficients W_T, competitor W, and $\delta = W - W_T$, expansion of the distortion objective gives

$$D_{\mathrm{tr}}(W; B) - D_{\mathrm{tr}}(W_T; B) = \frac{\langle \delta, g\delta\rangle_F + 2\langle \delta, gW_T - h\rangle_F}{H}. \tag{41}$$

For exact unpenalized projection the residual is zero. For ideal spectral projection the residual is orthogonal to the retained eigenspace, so the nonnegative quadratic guarantee applies to competitors in that same space. It need not apply to an untruncated competitor. For a finite numerical solve, the explicit residual term remains part of the identity; a small relative residual alone does not justify dropping it when coefficient differences are large.

Relation to supervised restricted solutions. For the same fixed solver S, let $W_T = Sh$ and $W_Y = Sc$. With the teacher's unpenalized normal residual $E_B = C_{\mathrm{tr}} - G_{\mathrm{tr}}B$,

$$W_T - W_Y = -SU^\top E_B. \tag{42}$$

This follows immediately from $h = c - U^\top E_B$. For spectral minimum-norm solutions, equality holds if and only if $V_\tau^\top U^\top E_B = 0$. For a positive ridge inverse, equality holds if and only if $U^\top E_B = 0$. For an exact full supervised OLS teacher, $E_B = 0$, so teacher projection and the corresponding restricted supervised minimum-norm fit coincide. For a full ridge teacher,

$$E_B = \lambda B - r_\lambda, \qquad r_\lambda = (G_{\mathrm{tr}} + \lambda I)B - C_{\mathrm{tr}}. \tag{43}$$

Even when $r_\lambda = 0$ and the same λ is used for both restricted ridge objectives, their difference is $-\lambda(g + \lambda I)^{-1}U^\top B$, which is generally nonzero.

Boundary-preserving paths. The teacher-based paths use the declared origin at $K = 0$ and explicitly reuse the unchanged host at $K_{\max}$. A full-budget ridge projection can yield a different predictor. Full empirical minimum-norm projection can also alter the all-input function when its empirical kernel is larger than the architectural kernel or when numerical truncation discards nonzero directions. Supervised paths retain their own independently acquired full endpoints and reference risks.

Fidelity does not order task risk or capacity. For two students on the same split, set $d_i = X_s(UW_i - B)$ and $e = X_sB - \widetilde{Y}_s$. Since their task residuals are $e + d_i$, expansion proves

$$\widehat{R}_s(UW_2) - \widehat{R}_s(UW_1) = D_s(W_2; B) - D_s(W_1; B) + \frac{2\langle d_2 - d_1, e\rangle_F}{n_s H}. \tag{44}$$

The last term has no fixed sign. Therefore a reduction in teacher distortion alone supplies no ordering of target risk. It also supplies no ordering of empirical capacities, which depend on fixed task-risk inequalities at all earlier rungs. As a scalar counterexample to an inference from fidelity alone, specify two arbitrary restricted paths directly by their forecasts. Take target 1, full teacher 0, an origin prediction 3, and a grid with one interior rung. Stipulate interior predictions 0.5 and -0.1 to test the fidelity-only implication. They have teacher distortions 0.25 and 0.01, but task risks 0.25 and 1.21. With a common full risk of 1 and $\epsilon = 0.01$, the first path qualifies at the interior rung and the second only at the full rung.

Similarly, a lower integral of risk over budget does not imply an earlier first accepted rung, nor does an earlier first hit imply a lower integral. One is an average over the declared budget measure; the other is a threshold functional. Same-split comparisons in Equation (44) separate fidelity and task risk within a common evaluation geometry. Across splits, alignment changes depend jointly on input moments and input–target moments, leaving the form of distribution shift unidentified from alignment alone.

B.8 Canonical starts and nested compact restriction

Let $A_K = AU_K$ and $Q_K = A_K^{\dagger} A_K$, using exact Moore–Penrose projectors and the same Euclidean coefficient metric. If $b = Qb$, then $b = A^{\top} z$ for some z, so

$$U_K^{\top} b = (AU_K)^{\top} z \in \mathrm{range}(A_K^{\top}), \qquad Q_K U_K^{\top} Q = U_K^{\top} Q. \tag{45}$$

For nested supports $U_J = U_K V_{JK}$, this implies

$$V_{JK}^{\top} Q_K U_K^{\top} Q = U_J^{\top} Q. \tag{46}$$

Thus compact canonicalization is redundant after restricting a full canonical representative, and any nested compact chain has the same final coefficients as direct restriction, including for redundant compact models. The identity concerns exact compact restriction under fixed preprocessing and coefficient metrics; ambient re-expansion $QP_K Qb$, changes to preprocessing or metrics, and independently truncated approximate projectors are separate operations.

C Complete capacity results

C.1 Seventy-five core task comparisons

Each table gives $C_{\mathrm{L}}/C_{\mathrm{R}}$ at $\epsilon = 1\%$ for each seed and for the risk-pooled result. All capacities are grid indices in the restriction family of Appendix A.2. A zero, when present, is the actual evaluated zero-budget model, not a missing value. The primary comparison uses zero inactive coordinates for FITS/MixLinear and native restrictions for the other architectures. Both metrics use their own corresponding full-model risk.

Table 3: FITS: complete first-hit capacities at $\epsilon = 1\%$. Entries give the capacity pair C_L/C_R: L trains the restricted model and R restricts the full model. Pooled capacities are computed after averaging risks. K counts retained input frequencies, with the output cutoff induced by the forecast length.

		MSE				MAE			
Dataset	H	Seed 0	Seed 1	Seed 2	Pooled	Seed 0	Seed 1	Seed 2	Pooled
ETTh2	96	112/176	112/176	112/176	112/176	81/176	81/176	81/176	81/176
ETTh2	336	81/96	81/96	81/96	81/96	81/96	81/96	81/96	81/96
ETTh2	720	81/96	81/96	81/96	81/96	81/81	81/81	81/81	81/81
Weather	96	81/196	81/196	81/176	81/196	81/196	81/196	81/196	81/196
Weather	336	34/128	34/128	34/144	34/128	34/144	34/144	34/144	34/144
Weather	720	14/81	14/81	14/81	14/81	34/81	34/81	34/81	34/81
Electricity	96	196/196	196/196	196/196	196/196	196/196	196/196	196/196	196/196
Electricity	336	196/196	196/196	196/196	196/196	196/196	196/196	196/196	196/196
Electricity	720	196/196	196/196	196/196	196/196	196/196	196/196	196/196	196/196
Exchange	96	96/196	81/196	96/196	81/196	81/196	81/196	81/196	81/196
Exchange	336	112/196	112/196	112/196	112/196	96/196	96/196	96/196	96/196
Exchange	720	112/176	34/176	196/196	196/176	96/160	34/144	196/160	196/160

Table 4: MixLinear: complete first-hit capacities at $\epsilon = 1\%$. Entries give the capacity pair C_L/C_R: L trains the restricted model and R restricts the full model. Pooled capacities are computed after averaging risks. K counts retained input-cycle Fourier frequencies.

		MSE				MAE			
Dataset	H	Seed 0	Seed 1	Seed 2	Pooled	Seed 0	Seed 1	Seed 2	Pooled
ETTh2	96	3/19	1/19	5/19	3/19	3/19	1/19	1/19	1/19
ETTh2	336	2/1	2/1	2/12	2/8	1/12	1/8	2/12	1/12
ETTh2	720	1/3	1/8	1/8	1/5	1/1	1/1	1/8	1/1
Weather	96	12/12	12/12	8/12	8/12	12/12	12/12	8/12	12/12
Weather	336	8/12	8/12	8/12	8/12	8/12	12/12	8/12	8/12
Weather	720	5/12	5/12	5/12	5/12	5/12	5/12	5/12	5/12
Electricity	96	19/19	19/19	19/19	19/19	19/19	19/19	19/19	19/19
Electricity	336	19/19	19/19	12/19	19/19	12/19	12/19	12/19	12/19
Electricity	720	12/12	12/12	12/12	12/12	8/12	12/12	8/12	8/12
Exchange	96	2/19	1/19	1/19	1/19	2/19	1/19	1/19	1/19
Exchange	336	8/19	1/19	1/19	3/19	2/19	1/19	1/19	1/19
Exchange	720	1/19	5/12	1/19	2/19	1/19	3/8	1/19	1/19

Table 5: N-BEATS (generic): complete first-hit capacities at $\epsilon = 1\%$. Entries give the capacity pair C_L/C_R: L trains the restricted model and R restricts the full model. Pooled capacities are computed after averaging risks. K is the retained residual-block prefix length; K=0 is persistence.

		MSE				MAE			
Dataset	H	Seed 0	Seed 1	Seed 2	Pooled	Seed 0	Seed 1	Seed 2	Pooled
ETTh2	96	1/19	1/24	0/0	0/0	0/0	1/19	0/0	0/0
ETTh2	336	0/0	0/0	1/30	0/0	0/0	0/0	0/0	0/0
ETTh2	720	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0

Table 6: N-BEATS (interpretable): complete first-hit capacities at $\epsilon = 1\%$. Entries give the capacity pair C_L/C_R: L trains the restricted model and R restricts the full model. Pooled capacities are computed after averaging risks. K counts retained harmonics; K=0 retains the trained trend and DC components.

		MSE				MAE			
Dataset	H	Seed 0	Seed 1	Seed 2	Pooled	Seed 0	Seed 1	Seed 2	Pooled
ETTh2	96	1/4	0/0	16/4	2/4	1/4	0/4	16/4	2/4
ETTh2	336	0/16	0/0	8/167	0/16	0/64	0/0	64/167	0/16
ETTh2	720	1/359	1/64	0/359	0/359	1/359	1/64	0/359	1/359
Weather	96	1/32	0/47	2/32	1/32	1/32	0/4	4/32	1/32
Weather	336	4/167	4/167	8/167	8/167	4/167	4/167	8/167	8/167
Weather	720	359/359	359/359	4/359	8/359	359/359	16/359	4/64	32/359
Electricity	96	32/47	47/47	32/47	32/47	47/47	47/47	32/47	47/47
Electricity	336	167/167	167/167	167/167	167/167	167/167	167/167	167/167	167/167
Electricity	720	359/359	359/359	359/359	359/359	359/359	359/359	359/359	359/359
Exchange	96	0/0	0/0	0/0	0/0	0/1	0/0	0/0	0/0
Exchange	336	8/0	2/0	0/0	0/0	8/0	1/0	0/0	0/0
Exchange	720	16/0	0/0	0/0	0/0	16/0	0/0	0/0	0/0

Table 7: N-HiTS: complete first-hit capacities at $\epsilon = 1\%$. Entries give the capacity pair C_L/C_R: L trains the restricted model and R restricts the full model. Pooled capacities are computed after averaging risks. K is the retained coarse-to-fine block prefix length; K=0 is persistence.

		MSE				MAE			
Dataset	H	Seed 0	Seed 1	Seed 2	Pooled	Seed 0	Seed 1	Seed 2	Pooled
ETTh2	96	$3/2$	$0/0$	$3/3$	$0/0$	$1/2$	$0/0$	$3/3$	$0/0$
ETTh2	336	$2/3$	$0/0$	$0/0$	$0/0$	$2/2$	$0/0$	$0/0$	$0/0$
ETTh2	720	$0/0$	$0/0$	$0/0$	$0/0$	$0/0$	$0/0$	$0/0$	$0/0$
Weather	96	$2/2$	$2/2$	$2/2$	$2/2$	$2/2$	$2/2$	$2/2$	$2/2$
Weather	336	$1/2$	$2/2$	$2/2$	$2/2$	$1/1$	$2/1$	$1/2$	$1/1$
Weather	720	$3/2$	$2/2$	$2/2$	$2/2$	$1/2$	$1/2$	$2/2$	$1/2$
Electricity	96	$3/3$	$3/3$	$3/3$	$3/3$	$3/3$	$3/3$	$3/3$	$3/3$
Electricity	336	$3/3$	$3/3$	$3/3$	$3/3$	$3/3$	$3/3$	$3/3$	$3/3$
Electricity	720	$3/3$	$3/3$	$3/3$	$3/3$	$3/3$	$3/3$	$3/3$	$3/3$
Exchange	96	$0/0$	$0/0$	$0/0$	$0/0$	$0/0$	$0/0$	$0/0$	$0/0$
Exchange	336	$0/0$	$0/0$	$0/0$	$0/0$	$0/0$	$0/0$	$0/0$	$0/0$
Exchange	720	$0/0$	$0/0$	$0/0$	$0/0$	$0/0$	$0/0$	$0/0$	$0/0$

Table 8: NLinear: complete first-hit capacities at $\epsilon = 1\%$. Entries give the capacity pair C_L/C_R: L trains the restricted model and R restricts the full model. Pooled capacities are computed after averaging risks. K counts retained real-Fourier frequency groups; K=0 is the native persistence origin.

		MSE				MAE			
Dataset	H	Seed 0	Seed 1	Seed 2	Pooled	Seed 0	Seed 1	Seed 2	Pooled
ETTh2	96	$64/361$	$64/361$	$64/361$	$64/361$	$64/192$	$64/192$	$64/192$	$64/192$
ETTh2	336	$64/361$	$64/361$	$64/361$	$64/361$	$64/256$	$64/256$	$64/256$	$64/256$
ETTh2	720	$64/361$	$64/361$	$64/361$	$64/361$	$32/192$	$32/192$	$32/192$	$32/192$
Weather	96	$32/192$	$32/256$	$32/192$	$32/192$	$32/128$	$32/192$	$32/128$	$32/128$
Weather	336	$16/192$	$16/192$	$16/192$	$16/192$	$16/96$	$16/96$	$16/96$	$16/96$
Weather	720	$16/192$	$16/192$	$16/192$	$16/192$	$16/96$	$16/96$	$16/96$	$16/96$
Electricity	96	$361/361$	$361/361$	$361/361$	$361/361$	$361/361$	$361/361$	$361/361$	$361/361$
Electricity	336	$361/361$	$361/361$	$361/361$	$361/361$	$361/361$	$361/361$	$361/361$	$361/361$
Electricity	720	$361/361$	$361/361$	$361/361$	$361/361$	$361/361$	$361/361$	$361/361$	$361/361$
Exchange	96	$0/0$	$0/0$	$0/0$	$0/0$	$0/0$	$0/0$	$0/0$	$0/0$
Exchange	336	$0/0$	$0/0$	$0/0$	$0/0$	$0/0$	$0/0$	$0/0$	$0/0$
Exchange	720	$0/0$	$0/0$	$0/0$	$0/0$	$0/0$	$0/0$	$0/0$	$0/0$

Table 9: DLinear: complete first-hit capacities at $\epsilon = 1\%$. Entries give the capacity pair C_L/C_R: L trains the restricted model and R restricts the full model. Pooled capacities are computed after averaging risks. K counts retained real-Fourier frequency groups in both branches; K=0 has zero branch coefficients.

		MSE				MAE			
Dataset	H	Seed 0	Seed 1	Seed 2	Pooled	Seed 0	Seed 1	Seed 2	Pooled
ETTh2	96	$64/256$	$64/256$	$64/256$	$64/256$	$64/256$	$64/256$	$64/256$	$64/256$
ETTh2	336	$64/64$	$64/64$	$64/64$	$64/64$	$64/96$	$64/96$	$64/96$	$64/96$
ETTh2	720	$1/32$	$1/32$	$1/32$	$1/32$	$1/64$	$1/64$	$1/64$	$1/64$
Weather	96	$16/361$	$16/256$	$16/361$	$16/361$	$16/361$	$16/361$	$16/361$	$16/361$
Weather	336	$16/256$	$16/256$	$16/256$	$16/256$	$16/256$	$16/256$	$16/256$	$16/256$
Weather	720	$16/192$	$16/192$	$16/192$	$16/192$	$16/192$	$16/192$	$16/192$	$16/192$
Electricity	96	$361/361$	$361/361$	$361/361$	$361/361$	$361/361$	$361/361$	$361/361$	$361/361$
Electricity	336	$361/361$	$361/361$	$361/361$	$361/361$	$361/361$	$361/361$	$361/361$	$361/361$
Electricity	720	$361/361$	$361/361$	$361/361$	$361/361$	$361/361$	$361/361$	$361/361$	$361/361$
Exchange	96	$16/256$	$16/256$	$16/256$	$16/256$	$16/361$	$16/361$	$16/361$	$16/361$
Exchange	336	$16/96$	$16/96$	$16/96$	$16/96$	$16/192$	$16/192$	$16/192$	$16/192$
Exchange	720	$8/64$	$8/64$	$8/64$	$8/64$	$8/64$	$8/64$	$8/64$	$8/64$

C.2 THE FULL PREDICTORS DEFINING THE ACCURACY TARGETS

Table 10 reports the full-model errors defining the capacity targets under the acquisition recipes in Appendix A.3. Both paths within each task and seed share the same accuracy reference.

Table 10: Full-reference prediction error for the 75 core acquisition comparisons. Each entry is mean three-seed MSE or MAE on the training-standardized scale. These are the full predictors defining the accuracy targets of the corresponding core capacity tables. Each configuration retains its stated training procedure.

		H96		H336		H720	
Configuration	Dataset	MSE	MAE	MSE	MAE	MSE	MAE
FITS	ETTh2	0.2710	0.3360	0.3540	0.3956	0.3775	0.4229
	Weather	0.1686	0.2226	0.2574	0.2938	0.3205	0.3390
	Electricity	0.1412	0.2427	0.1715	0.2706	0.2098	0.3019
	Exchange	0.0878	0.2094	0.3699	0.4450	1.0650	0.7783
MixLinear	ETTh2	0.2860	0.3413	0.3568	0.3949	0.3814	0.4239
	Weather	0.1704	0.2253	0.2597	0.2973	0.3236	0.3440
	Electricity	0.1385	0.2331	0.1700	0.2651	0.2102	0.2995
	Exchange	0.1019	0.2285	0.3368	0.4216	1.0896	0.7805
N-BEATS (generic)	ETTh2	0.4778	0.4728	0.8424	0.6440	1.8214	1.0238
N-BEATS (interpretable)	ETTh2	0.3885	0.4292	0.7927	0.6344	1.0539	0.7547
	Weather	0.1624	0.2226	0.2695	0.3182	0.3354	0.3600
	Electricity	0.1343	0.2347	0.1668	0.2690	0.1987	0.2990
	Exchange	0.3418	0.4128	2.1284	1.1240	4.0281	1.5595
N-HiTS	ETTh2	0.4278	0.4397	0.7144	0.5859	1.2124	0.7997
	Weather	0.1590	0.2159	0.2643	0.3166	0.3283	0.3604
	Electricity	0.1325	0.2290	0.1721	0.2674	0.2022	0.2968
	Exchange	0.1431	0.2662	1.8630	1.0119	4.4219	1.6500
NLinear	ETTh2	0.2699	0.3359	0.3553	0.4022	0.3913	0.4409
	Weather	0.1680	0.2218	0.2574	0.2941	0.3207	0.3400
	Electricity	0.1330	0.2282	0.1638	0.2579	0.2030	0.2911
	Exchange	0.0881	0.2061	0.3688	0.4353	1.0937	0.7818
DLinear	ETTh2	0.3092	0.3729	0.5390	0.5116	0.8997	0.6725
	Weather	0.1678	0.2258	0.2551	0.3032	0.3158	0.3536
	Electricity	0.1327	0.2295	0.1623	0.2610	0.1967	0.2945
	Exchange	0.0914	0.2176	0.4497	0.5001	1.4181	0.9015

C.3 ETTh1 CAPACITY EXTENSION

Table 11 gives the complete four-configuration, four-horizon extension with training seeds 0, 1, and 2. Capacities are computed after averaging the three risk profiles and full-model risks. The extension uses the restriction grids and zero inactive-coordinate convention in Appendix A.2, with the training rules in Appendix A.3. Within each comparison, both paths share the same full-model reference for the corresponding metric.

Table 11: ETTh1 capacities at $\epsilon = 1\%$, computed from mean three-seed risks. Each entry gives $C_{\mathrm{L}}/C_{\mathrm{R}}$ using the corresponding full-model MSE or MAE as the accuracy reference. Structural budgets use each model's own units.

Configuration	H	MSE $C_{\mathrm{L}}/C_{\mathrm{R}}$	MAE $C_{\mathrm{L}}/C_{\mathrm{R}}$
NLinear	96	192/256	192/256
	192	192/256	192/256
	336	192/256	192/256
	720	256/256	192/192
DLinear	96	192/256	192/256
	192	192/256	192/256
	336	256/256	192/256
	720	192/256	192/256
FITS	96	160/160	128/160
	192	128/176	128/176
	336	128/160	96/160
	720	128/176	96/160
MixLinear	96	5/8	5/8
	192	1/1	1/5
	336	1/1	1/5
	720	1/2	1/2

C.4 DETERMINISTIC LINEAR CONTROLS

The following 48 solutions comprise ridge and minimum-norm least squares on the 24 core linear tasks. They are separate acquisition procedures, each with its own full reference. They are not extra random seeds for the Adam tables.

Table 12: NLinear: deterministic exact-training control capacities at $\epsilon = 1\%$. Entries are C_L/C_R on the Fourier atlas. Ridge uses $\lambda = 10^{-6}$; numerical minimum-norm OLS uses a relative Gram-eigenvalue cutoff of 10^{-10}. These solutions have no random-seed replication.

		Ridge		Minimum-norm OLS	
Dataset	H	MSE	MAE	MSE	MAE
ETTh2	96	64/96	64/64	64/96	64/64
ETTh2	336	64/64	64/64	64/64	64/64
ETTh2	720	64/64	32/32	64/64	32/32
Weather	96	32/64	32/96	32/64	32/96
Weather	336	16/32	16/64	16/32	16/64
Weather	720	16/16	16/16	16/16	16/16
Electricity	96	361/361	361/361	361/361	361/361
Electricity	336	361/361	361/361	361/361	361/361
Electricity	720	361/361	361/361	361/361	361/361
Exchange	96	0/0	0/0	0/0	0/0
Exchange	336	0/0	0/0	0/0	0/0
Exchange	720	0/0	0/0	0/0	0/0

Table 13: DLinear: deterministic exact-training control capacities at $\epsilon = 1\%$. Entries are C_L/C_R on the Fourier atlas. Ridge uses $\lambda = 10^{-6}$; numerical minimum-norm OLS uses a relative Gram-eigenvalue cutoff of 10^{-10}. These solutions have no random-seed replication.

		Ridge		Minimum-norm OLS	
Dataset	H	MSE	MAE	MSE	MAE
ETTh2	96	$64/96$	$64/128$	$64/96$	$64/128$
ETTh2	336	$64/32$	$64/64$	$64/32$	$64/64$
ETTh2	720	$1/32$	$1/32$	$1/32$	$1/32$
Weather	96	$16/128$	$16/128$	$16/128$	$16/128$
Weather	336	$16/64$	$16/64$	$16/64$	$16/64$
Weather	720	$16/32$	$16/32$	$16/32$	$16/32$
Electricity	96	$361/361$	$361/361$	$361/361$	$361/361$
Electricity	336	$361/361$	$361/361$	$361/361$	$361/361$
Electricity	720	$361/361$	$361/361$	$361/361$	$361/361$
Exchange	96	$16/128$	$16/192$	$16/128$	$16/192$
Exchange	336	$16/64$	$16/64$	$16/64$	$16/64$
Exchange	720	$8/64$	$8/64$	$8/64$	$8/64$

C.5 Additional restriction and external-task controls

The next tables evaluate the same 24 linear tasks under two temporal restriction families, yielding 48 exact controls, followed by four external Fourier-ridge task results. The temporal index K is translated to d_K lag coordinates as in Appendix A.2.

Table 14: NLinear: exact-ridge capacities on temporal restriction atlases, MSE and $\epsilon = 1\%$. Entries are C_L/C_R. At positive K the atlas retains $\min(720, 2K-1)$ processed-feature lag coefficients per branch and all branch biases; this is not a change of input lookback. The full reference is shared across these orthogonal coordinate atlases.

Dataset	H	Recent-lag atlas	Oldest-lag atlas
ETTh2	96	$192/256$	$361/361$
ETTh2	336	$192/361$	$361/361$
ETTh2	720	$192/361$	$361/361$
Weather	96	$361/256$	$361/361$
Weather	336	$361/361$	$361/361$
Weather	720	$361/361$	$361/361$
Electricity	96	$361/361$	$361/361$
Electricity	336	$361/361$	$361/361$
Electricity	720	$256/361$	$361/361$
Exchange	96	$0/0$	$0/0$
Exchange	336	$0/0$	$0/0$
Exchange	720	$0/0$	$0/0$

Table 15: DLinear: exact-ridge capacities on temporal restriction atlases, MSE and $\epsilon = 1\%$. Entries are C_L/C_R. At positive K the atlas retains $\min(720, 2K-1)$ processed-feature lag coefficients per branch and all branch biases; this is not a change of input lookback. The full reference is shared across these orthogonal coordinate atlases.

Dataset	H	Recent-lag atlas	Oldest-lag atlas
ETTh2	96	$192/256$	$361/361$
ETTh2	336	$96/192$	$361/361$
ETTh2	720	$32/96$	$361/361$
Weather	96	$361/361$	$361/361$
Weather	336	$256/361$	$361/361$
Weather	720	$361/361$	$361/361$
Electricity	96	$361/361$	$361/361$
Electricity	336	$361/361$	$361/361$
Electricity	720	$361/361$	$361/361$
Exchange	96	$1/16$	$361/361$
Exchange	336	$1/4$	$361/361$
Exchange	720	$1/2$	$361/361$

Table 16: External exact-ridge capacity controls on the Fourier atlas, MSE and $\epsilon = 1\%$. Entries are C_L/C_R. These H336 tasks accompany the paired native/Fourier full-model interventions.

Dataset	H	NLinear	DLinear
ETTm1	336	$64/64$	$64/64$
Traffic	336	$256/256$	$256/256$

D Tolerance, regularization, grid resolution, and optimization budget

D.1 Tolerance and metric sensitivity

The primary tolerance is fixed at 1%. With the selected predictors held fixed, every core MSE and MAE risk profile is evaluated at $\epsilon \in \{0, 0.001, \ldots, 0.05\}$ to characterize tolerance sensitivity. All extended exact controls receive the same MSE scan. The accompanying numerical files provide the per-seed and pooled first hits at each tolerance.

The two metrics preserve the primary capacity direction in $67/75$ core tasks. Exact capacity pairs agree in $46/75$; agreement at every one of the 51 tolerance points is a stronger condition and holds for the direction in $30/75$. Seed unanimity at the primary tolerance is $59/75$ for MSE and also $59/75$ for MAE; unanimity across the entire tolerance scan holds in $34/75$ MSE tasks and $31/75$ MAE tasks. Metric agreement and seed unanimity are reported separately at the primary tolerance and across the full tolerance scan.

The eight primary metric-direction changes are: DLinear ETTh2 $H = 336$, FITS ETTh2 $H = 720$, MixLinear ETTh2 $H = 720$, MixLinear Electricity $H = 336/720$, MixLinear Weather $H = 96$, interpretable N-BEATS Electricity $H = 96$, and N-HiTS Weather $H = 720$. Each changes between equality and learning-first separation. Their exact values remain in the preceding tables.

D.2 Regularization sensitivity of the exact linear controls

The acquisition discrepancy persists across regularization strengths. We recompute the exact NLinear and DLinear paths at $\lambda \in \{0, 10^{-8}, 10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}\}$ on the four core datasets and three horizons, under all three linear restriction families. The 13-point budget grid, preprocessing, origins, and training moments remain fixed. Each value of λ defines its own full-model accuracy reference, shared by both paths. Positive values use direct ridge solves; zero uses the minimum-norm solution with the original relative Gram-eigenvalue cutoff 10^{-10}. This gives 72 model–task–restriction settings per value of λ, evaluated using MSE.

All 24 capacity separations at the baseline $\lambda = 10^{-6}$ retain their exact capacity pairs throughout the sweep. Across the six non-baseline settings of λ, $431/432$ capacity pairs match the baseline at $\epsilon = 1\%$; the unregularized solution matches all 72. Table 17 also reports agreement over the 51-point tolerance grid. In the NLinear temporal zero-effect controls, all $24 \times 7 = 168$ profile pairs remain different, with a minimum normalized integrated absolute MSE difference of 1.178% (Figure 5, left). Thus training coupling continues to separate the acquisition paths when representation exposure vanishes, including without regularization.

Table 17: Regularization sensitivity of exact linear acquisition. Each row contains 72 model–task–restriction settings. "Same pair" compares $C_{\mathrm{L}}/C_{\mathrm{R}}$ with $\lambda = 10^{-6}$, using each setting's own full reference. The last column repeats this comparison over all 51 tolerances in $[0, 5\%]$ ($72 \times 51 = 3672$ readouts per row).

	$\epsilon = 1\%$		$\epsilon \in [0, 5\%]$
λ	Same pair	$C_{\mathrm{L}} \neq C_{\mathrm{R}}$	Same pair
0	72/72	24/72	3639/3672
10^{-8}	71/72	25/72	3613/3672
10^{-7}	72/72	24/72	3645/3672
10^{-6}	72/72	24/72	3672/3672
10^{-5}	72/72	24/72	3649/3672
10^{-4}	72/72	24/72	3623/3672
10^{-3}	72/72	24/72	3579/3672

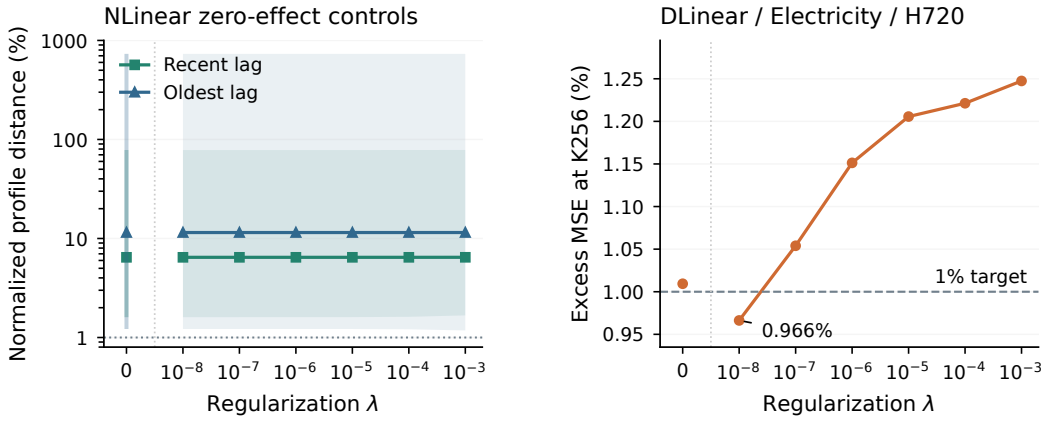


Figure 5: **Path separation survives changes in regularization.** Left: NLinear temporal controls across the 12 core tasks. Lines show task medians of the integrated absolute path-risk difference divided by full MSE; shading spans the task minimum and maximum. Every curve pair has a distance above 1%. Right: the sole changing capacity pair, DLinear on Electricity at $H = 720$ under Fourier restriction. At $K = 256$, the learned predictor crosses the 1% accuracy threshold at $\lambda = 10^{-8}$. The zero setting is plotted separately because it uses the fixed minimum-norm spectral cutoff.

The single change is a threshold crossing. For DLinear on Electricity at $H = 720$ under Fourier restriction, the capacity pair changes from $361/361$ at $\lambda = 10^{-6}$ to $256/361$ at $\lambda = 10^{-8}$. At $K = 256$, the learned predictor's excess MSE above its full reference falls from 1.1513% to 0.9663%, crossing the fixed 1% target. Direct, Cholesky, and spectral solves agree to within 1.5×10^{-13} MSE at the latter setting. At $\lambda = 0$, the excess is 1.0094% and the pair is $361/361$: this solution discards eigenmodes below the fixed cutoff (2.01×10^{-8} at this budget), whereas positive ridge weights them by $(s + \lambda)^{-1}$ for eigenvalue s. The change illustrates the threshold nature of capacity: a risk shift near the accuracy target can move the first adequate budget while the acquisition discrepancy persists.

D.3 Empirical resolution without new predictions

For FITS ETTh2 we retain the first and last points of the 13-point grid and enumerate all $2^{11} = 2048$ subsets of its interior points, separately at each horizon. Risks and the full reference remain fixed. We then compute the pooled empirical first hit on each subset, using exactly the same definition as the primary analysis. This deterministic subgrid analysis describes empirical grid resolution using the saved risks.

Let $\mathcal{K}' \subseteq \mathcal{K}$ contain the full endpoint. With fixed risks and target, $\mathcal{A}_a(\epsilon) \cap \mathcal{K}' \subseteq \mathcal{A}_a(\epsilon)$, hence $C_a^{\mathcal{K}'}(\epsilon) \geq C_a^{\mathcal{K}}(\epsilon)$. Thinning can delay an observed capacity, but cannot make it earlier. It can remove different witnesses on the two paths, so it can also erase or alter their relation. Figure 6 shows all subgrids grouped by their number of retained points.

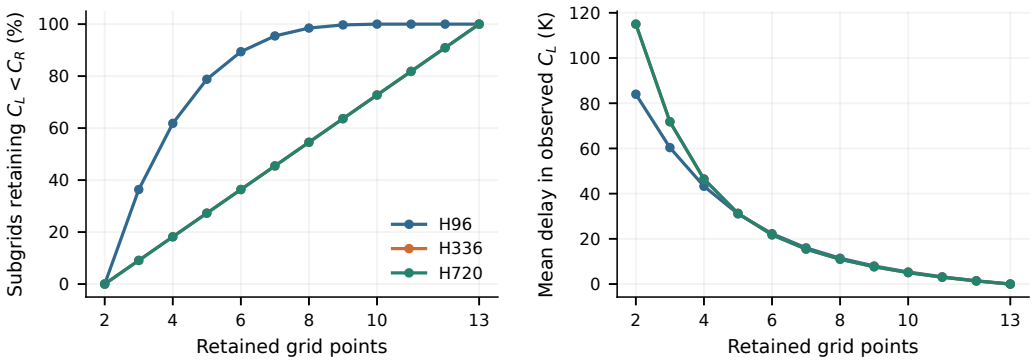


Figure 6: **A sparser grid can censor capacity separation.** All 2048 endpoint-preserving subgrids of the FITS ETTh2 ladder, at each horizon. Left: fraction retaining learning-first separation of pooled empirical capacities. Right: average delay of the learning first hit relative to the dense grid. Outcomes depend on which transition witnesses remain, not just how many points or how large a gap the subgrid contains.

D.4 Training-budget controls

FITS on Exchange $H = 720$ and generic N-BEATS on ETTh2 $H = 336$ were trained for extended continuous runs. Validation-best models from the first 50, 100, and 200 epochs define three nested budget prefixes. There are six FITS structural budgets and nine N-BEATS depths per seed, respectively. The N-BEATS sentinel grid starts at $K = 1$; its first observed trained depth is distinct from capacity on the main ladder that also includes the untrained $K = 0$ origin. Prefix results are repeated observations of a trajectory, not independent retrainings. The FITS comparison retains both restriction-origin choices; the N-BEATS budget control measures its learning curve only. Table 18 reports the empirical results from their saved risks.

Table 18: Capacity under longer optimization. The same training trajectories are continued to 200 epochs; validation chooses a model within each budget. Capacities use the experiment's sentinel grid and the full model selected at that budget, at $\epsilon = 1\%$. For FITS, C_{R} uses zero-origin restriction and $C_{\mathrm{R,init}}$ uses matched-initializer restriction. N-BEATS measures the learned curve. Components are contiguous on the observed grid.

Configuration / task	Epochs	C_{L}	C_{R}	$C_{\mathrm{R,init}}$	Full MSE	Learn components
FITS, Exchange H720	50	196	176	196	1.065042	1
	100	196	176	196	1.064774	1
	200	196	176	196	1.064774	1
N-BEATS, ETTh2 H336	50	1	–	–	0.842413	3
	100	1	–	–	0.842413	3
	200	1	–	–	0.920487	2

On FITS, the selected $K = 160, 176, 196$ states are unchanged from 100 to 200 epochs for all seeds. Longer training therefore does not remove their local reversal. In generic N-BEATS, intermediate accepted components change with training duration, even where the first adequate sampled budget is unchanged. These observations motivate treating the optimizer budget as part of acquisition rather than assuming it affects only an error bar.

Nine additional DLinear Electricity continuations at $K = 16$ improved the training objective by extending Adam to 10000 updates but did not reach the exact solution. They are fixed-budget convergence diagnostics, not full capacity profiles. Optimization geometry can itself select different representatives of a function class (Gunasekar et al., 2018). The exact ridge comparisons in Section 5 provide the optimizer-independent mechanism control.

D.5 First attainment and sustained adequacy

A profile can meet the target and fail again at a larger budget. To distinguish first attainment from adequacy throughout the remaining measured ladder, define

$$C_a^{\mathrm{sustain}}(\epsilon) = \min\{K \in \mathcal{K} : \widehat{R}(f_{a,K'}) \le (1+\epsilon)\widehat{R}(f_{\mathrm{F}}) \text{ for every } K' \in \mathcal{K},\ K' \ge K\}. \tag{47}$$

This is the grid point following the last failure, or K_0 if every point qualifies. The shared full endpoint makes it well-defined, and $C_a^{\mathrm{sustain}} \ge C_a$. It describes the observed ladder, with no interpolation or smoothing. Risks, full references, and grids are fixed; task summaries average seed risks before applying either definition.

Table 19 shows how the two readouts compare at 1%. For MSE, 15 of 75 core capacity pairs change, including 13 changes of ordering; for MAE, the corresponding counts are 16 and 12. In the 225 seed-level comparisons, the MSE counts are 55 and 42, and the MAE counts are 50 and 34. For example, generic N-BEATS on ETTh2 at $H = 336$ moves from $0/0$ to $30/0$ under both metrics: its learning path qualifies at the origin but fails again before the full endpoint. Appendix E.3 discusses these disconnected adequate sets.

Table 19: **First attainment and sustained adequacy change some capacity comparisons.** Counts among 75 trained tasks at $\epsilon = 1\%$, after averaging the three seed risks.

Metric	Readout	$C_L < C_R$	$C_L = C_R$	$C_L > C_R$
MSE	First attainment	38	36	1
MSE	Sustained adequacy	37	31	7
MAE	First attainment	38	36	1
MAE	Sustained adequacy	37	29	9

Table 20: **The same-function capacity transitions survive sustained adequacy.** Raw-to-swapped capacities at the 1% MSE target and $H = 336$. Each transition is identical under both readouts and in all three seeds: four tasks, two host directions, 24 interventions.

Model	Dataset	Fourier host + native donor	Native host + Fourier donor
NLinear	ETTm1	$361 \to 64$	$64 \to 361$
DLinear	ETTm1	$192 \to 64$	$64 \to 192$
NLinear	Traffic	$361 \to 256$	$256 \to 361$
DLinear	Traffic	$361 \to 256$	$256 \to 361$

The same-function intervention retains every original capacity transition at the primary 1% MSE tolerance (Table 20). Across all 51 tolerances, the raw-to-swapped capacity pair differs between the two definitions in 2 of 1224 host–tolerance entries. These entries reuse the same 24 risk-profile pairs. Table 21 gives all 75 pooled trained-task comparisons. The accompanying numerical records provide both readouts at every tolerance for all seeds, deterministic controls, and representation interventions, keeping these groups separate.

Table 21: **Complete pooled comparison of the two capacity readouts.** Every entry is C_L/C_R at $\epsilon = 1\%$. Bold sustained pairs differ from first-hit pairs. N-BEATS-G and -I denote generic and interpretable configurations.

			MSE		MAE	
Model	Dataset	H	First	Sustained	First	Sustained
FITS	ETTh2	96	112/176	112/176	81/176	81/176
FITS	ETTh2	336	81/96	81/96	81/96	81/96
FITS	ETTh2	720	81/96	81/96	81/81	81/81
FITS	Weather	96	81/196	81/196	81/196	81/196
FITS	Weather	336	34/128	34/128	34/144	34/144
FITS	Weather	720	14/81	14/81	34/81	34/81
FITS	Electricity	96	196/196	196/196	196/196	196/196
FITS	Electricity	336	196/196	196/196	196/196	196/196
FITS	Electricity	720	196/196	196/196	196/196	196/196
FITS	Exchange	96	81/196	81/196	81/196	81/196
FITS	Exchange	336	112/196	112/196	96/196	96/196
FITS	Exchange	720	196/176	196/176	196/160	196/160
MixLinear	ETTh2	96	3/19	3/19	1/19	**3/19**
MixLinear	ETTh2	336	2/8	2/8	1/12	1/12
MixLinear	ETTh2	720	1/5	1/5	1/1	**1/8**
MixLinear	Weather	96	8/12	8/12	12/12	12/12
MixLinear	Weather	336	8/12	8/12	8/12	8/12
MixLinear	Weather	720	5/12	5/12	5/12	5/12
MixLinear	Electricity	96	19/19	19/19	19/19	19/19
MixLinear	Electricity	336	19/19	19/19	12/19	12/19
MixLinear	Electricity	720	12/12	12/12	8/12	8/12
MixLinear	Exchange	96	1/19	**19/19**	1/19	**12/19**
MixLinear	Exchange	336	3/19	3/19	1/19	1/19
MixLinear	Exchange	720	2/19	2/19	1/19	1/19
N-BEATS-G	ETTh2	96	0/0	**30/0**	0/0	**30/0**
N-BEATS-G	ETTh2	336	0/0	**30/0**	0/0	**30/0**
N-BEATS-G	ETTh2	720	0/0	0/0	0/0	0/0
N-BEATS-I	ETTh2	96	2/4	**8/4**	2/4	**8/4**
N-BEATS-I	ETTh2	336	0/16	**64/16**	0/16	**32/16**
N-BEATS-I	ETTh2	720	0/359	**4/359**	1/359	**4/359**
N-BEATS-I	Weather	96	1/32	1/32	1/32	1/32
N-BEATS-I	Weather	336	8/167	8/167	8/167	8/167
N-BEATS-I	Weather	720	8/359	**359/359**	32/359	**359/359**
N-BEATS-I	Electricity	96	32/47	32/47	47/47	47/47
N-BEATS-I	Electricity	336	167/167	167/167	167/167	167/167
N-BEATS-I	Electricity	720	359/359	359/359	359/359	359/359
N-BEATS-I	Exchange	96	0/0	0/0	0/0	0/0
N-BEATS-I	Exchange	336	0/0	0/0	0/0	0/0
N-BEATS-I	Exchange	720	0/0	**359/0**	0/0	**359/0**
N-HiTS	ETTh2	96	0/0	**3/3**	0/0	**3/0**
N-HiTS	ETTh2	336	0/0	**0/3**	0/0	0/0
N-HiTS	ETTh2	720	0/0	0/0	0/0	0/0
N-HiTS	Weather	96	2/2	2/2	2/2	2/2
N-HiTS	Weather	336	2/2	2/2	1/1	1/1
N-HiTS	Weather	720	2/2	2/2	1/2	**3/2**
N-HiTS	Electricity	96	3/3	3/3	3/3	3/3
N-HiTS	Electricity	336	3/3	3/3	3/3	3/3
N-HiTS	Electricity	720	3/3	3/3	3/3	3/3
N-HiTS	Exchange	96	0/0	0/0	0/0	0/0
N-HiTS	Exchange	336	0/0	0/0	0/0	0/0

Table 21 continued

			MSE		MAE	
Model	Dataset	H	First	Sustained	First	Sustained
N-HiTS	Exchange	720	$0/0$	$\mathbf{2/0}$	$0/0$	$\mathbf{2/0}$
NLinear	ETTh2	96	$64/361$	$64/361$	$64/192$	$64/192$
NLinear	ETTh2	336	$64/361$	$64/361$	$64/256$	$64/256$
NLinear	ETTh2	720	$64/361$	$64/361$	$32/192$	$32/192$
NLinear	Weather	96	$32/192$	$32/192$	$32/128$	$32/128$
NLinear	Weather	336	$16/192$	$16/192$	$16/96$	$16/96$
NLinear	Weather	720	$16/192$	$16/192$	$16/96$	$16/96$
NLinear	Electricity	96	$361/361$	$361/361$	$361/361$	$361/361$
NLinear	Electricity	336	$361/361$	$361/361$	$361/361$	$361/361$
NLinear	Electricity	720	$361/361$	$361/361$	$361/361$	$361/361$
NLinear	Exchange	96	$0/0$	$\mathbf{0/192}$	$0/0$	$\mathbf{0/192}$
NLinear	Exchange	336	$0/0$	$\mathbf{0/64}$	$0/0$	$\mathbf{0/64}$
NLinear	Exchange	720	$0/0$	$\mathbf{0/4}$	$0/0$	$0/0$
DLinear	ETTh2	96	$64/256$	$64/256$	$64/256$	$64/256$
DLinear	ETTh2	336	$64/64$	$64/64$	$64/96$	$64/96$
DLinear	ETTh2	720	$1/32$	$\mathbf{32/32}$	$1/64$	$\mathbf{32/64}$
DLinear	Weather	96	$16/361$	$16/361$	$16/361$	$16/361$
DLinear	Weather	336	$16/256$	$16/256$	$16/256$	$16/256$
DLinear	Weather	720	$16/192$	$16/192$	$16/192$	$16/192$
DLinear	Electricity	96	$361/361$	$361/361$	$361/361$	$361/361$
DLinear	Electricity	336	$361/361$	$361/361$	$361/361$	$361/361$
DLinear	Electricity	720	$361/361$	$361/361$	$361/361$	$361/361$
DLinear	Exchange	96	$16/256$	$16/256$	$16/361$	$16/361$
DLinear	Exchange	336	$16/96$	$16/96$	$16/192$	$16/192$
DLinear	Exchange	720	$8/64$	$8/64$	$8/64$	$8/64$

E ORIGIN, DIRECTION, AND NON-MONOTONE RISK CONTROLS

E.1 INACTIVE COORDINATES ARE PART OF THE PATH

For FITS and MixLinear, independently trained restricted models keep inactive coefficients at zero. Restoring those coefficients to their initial values when restricting a full model defines a distinct functional intervention and prediction path. Both restriction variants are reported in the numerical risk records. The primary comparison fixes the zero-origin convention across acquisition paths to isolate learning versus restriction.

Additional origin controls evaluate zero, matched-initializer, sign/phase, and dose variants using fixed trained models. For equal-norm perturbations δ and $-\delta$ around a forecast f, Equation (8) gives

$$\widehat{R}(f+\delta) - \widehat{R}(f-\delta) = 4\widehat{\mathbb{E}}\langle f - y, \delta\rangle.$$

Their energy is identical, while their residual alignment has opposite sign. This separates response direction from perturbation magnitude. Cross-band controls matched on validation norm are kept distinct from exact equal-norm sign-reversed pairs on the evaluation inputs.

E.2 SAME-FUNCTION DOSE RESPONSE

For linear hosts, the fixed sequence $B(\alpha) = QB + \alpha(I - Q)B$, with $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$, retains the original component's magnitude range. The full function stays fixed for every α. The restricted risk is quadratic in α (Appendix B), but need not be monotone. Across 72 hosts and 792 interior host–budget combinations, 243 five-point responses first decrease and then increase. The minimum-norm representation changes the 1% first hit in 39 hosts, all to smaller capacity, with 33 unchanged. At zero tolerance, three hosts instead move to larger capacity. The direction is determined by the target and residual, not by the label "canonical."

E.3 Nested structures can have disconnected adequate sets

The generic N-BEATS ETTh2 $H = 336$ learning profile has multiple adequate components on its measured depth ladder. Pooling risk and examining the three seeds need not give the same number or position of components. We preserve the original grid values and compare them directly with the accuracy target; we do not fit an isotonic risk curve. Increasing the optimization budget can change the holes, as Appendix D.4 shows. Nesting of model classes is therefore not a justification for monotonicity of validation-selected test error.

F Function projection and risk decomposition

F.1 Projection design

Each selected full linear host is held fixed. The retained teacher target is its forecast on training inputs, with the architecture's fixed origin subtracted. The primary solver is the minimum-norm empirical projection using a Gram eigenvalue cutoff of 10^{-10} relative to the largest eigenvalue. This fixes the numerical subspace in which the least-squares objective is minimized. A teacher-ridge control uses $\lambda = 10^{-6}$. At the full budget, both teacher-based paths use the identity host; at zero budget, they use the native origin.

Post-training reconstruction objectives also underlie curvature-aware compression (Frantar et al., 2022); here the linear design permits a direct comparison to the task-target normal equations. Raw coefficient masking, canonical masking, teacher projection, teacher ridge, supervised minimum-norm learning, and supervised ridge are kept separately. Teacher paths and coefficient masks use the same host reference for their capacity comparison. Supervised full solutions retain their actual predictions; own-reference capacities and comparisons to the host target are different readouts. Table 22 concerns the projection–canonical comparison with a common host reference.

Table 22: Function projection versus canonical coefficient restriction. Changes are integrated over normalized budget, in percent of full MSE. Capacities use the same host reference; each listed capacity is identical across the three seeds.

Model	Data	H	ΔD	ΔR	C_{can}	C_{proj}
NLinear	ETTh2	96	-0.250	-0.303	96	64
NLinear	ETTh2	336	-0.096	-0.133	64	64
NLinear	ETTh2	720	-0.071	-0.060	64	64
NLinear	Weather	96	-0.301	-0.559	64	32
NLinear	Weather	336	-0.087	-0.134	32	16
NLinear	Weather	720	-0.037	-0.021	16	16
NLinear	Exchange	96	-0.946	-0.754	0	0
NLinear	Exchange	336	-0.171	0.487	0	0
NLinear	Exchange	720	-0.055	0.121	0	0
NLinear	Electricity	96	-9.087	-9.109	361	361
NLinear	Electricity	336	-7.456	-7.590	361	361
NLinear	Electricity	720	-5.966	-6.118	361	361
DLinear	ETTh2	96	-1.358	-1.174	96	64
DLinear	ETTh2	336	-0.369	0.120	32	64
DLinear	ETTh2	720	-0.161	0.077	32	1
DLinear	Weather	96	-0.850	-1.799	128	16
DLinear	Weather	336	-0.211	-0.586	64	16
DLinear	Weather	720	-0.081	-0.232	32	16
DLinear	Exchange	96	-45.782	-35.361	128	16
DLinear	Exchange	336	-8.017	-2.160	64	16
DLinear	Exchange	720	-1.879	-0.198	64	8
DLinear	Electricity	96	-27.551	-27.774	361	361
DLinear	Electricity	336	-22.521	-22.991	361	361
DLinear	Electricity	720	-18.113	-18.618	361	361

F.2 TEACHER FIDELITY DOES NOT IDENTIFY TASK BENEFIT

The teacher and target residual are evaluated on the same split. Writing $d_i = g_i - f_B$, the task-risk contrast between two compact models is

$$\widehat{R}(g_2) - \widehat{R}(g_1) = \widehat{\mathbb{E}}(\|d_2\|^2 - \|d_1\|^2) + 2\widehat{\mathbb{E}}\langle d_2 - d_1, f_B - y\rangle.$$

This same-split identity separates teacher fidelity from task error through the residual-alignment term. Among 792 interior host–budget comparisons, projection is closer to the teacher and better on the task in 588, closer but worse in 195, farther but better in three, and farther and worse in six. These correlated budget observations describe coverage, not independent evidence counts.

F.3 SEPARATING THE COMPONENTS OF A FULL PATH GAP

For a full Adam host B_A, full ridge solution B_{F}, and retained ridge solution B_{L} expressed in ambient coordinates, write

$$D = B_A - B_{\mathrm{F}}, \quad d_c = P_K B_{\mathrm{F}} - B_{\mathrm{L}}, \quad d_v = P_K Q D, \quad d_n = P_K(I - Q)D.$$

Then $d_c + d_v + d_n = P_K B_A - B_{\mathrm{L}}$. This auxiliary decomposition compares the restricted Adam host with the exact restricted ridge predictor; the original acquisition paths both use Adam. The first component is exact training coupling; the others are function-visible and function-invisible components of the full Adam–ridge difference.

The risk expansion includes self-energies, pairwise interactions, and alignment with the ridge residual, as derived in Appendix B. Across the original 24 tasks, the seed-averaged, profile-integrated test self-energies for d_c, d_v, d_n are 5.7548%, 0.1694%, and 4.6614% of host full MSE, respectively; the net pairwise contribution is -1.0618%. These are energy terms, not additive causal percentages of forecast error. On DLinear ETTm1, the coupling–null interaction is positive for all three Fourier-trained hosts and negative for all three native-trained hosts. Cancellation and reinforcement can therefore change with the trained representation, consistent with the task-dependent risk response.

F.4 SPECTRAL CUTOFF CHANGES RISKS AND THRESHOLD CROSSINGS

The separation between teacher fidelity and task accuracy persists across five relative Gram cutoffs. We evaluate $\tau \in \{10^{-12}, 10^{-11}, 10^{-10}, 10^{-9}, 10^{-8}\}$ on all 24 core NLinear/DLinear tasks, retaining the same 72 Fourier-trained teachers, data, and structural grids. Each retained projection solves the training teacher-matching objective in the Gram eigenspace with eigenvalues above τ times the largest eigenvalue. The architecture-level SVD threshold used to construct Q remains 10^{-12}; it is distinct from this empirical Gram cutoff. The zero-budget origin and full-budget identity teacher are fixed. The primary setting remains 10^{-10}; test risk is used to describe sensitivity across the five settings.

Table 23: **Teacher fidelity and task accuracy remain distinct across five cutoffs.** Counts among 24 tasks, relative to canonical coefficient masking. Capacity uses the pooled 1% MSE target; area comparisons average host-normalized integrals.

Gram cutoff	Teacher area improves	Task area improves / worsens	Projection capacity smaller / equal / larger
10^{-12}	24/24	20/4	12/11/1
10^{-11}	24/24	20/4	12/11/1
10^{-10}	24/24	20/4	11/12/1
10^{-9}	24/24	20/4	11/12/1
10^{-8}	24/24	20/4	11/12/1

At every cutoff, teacher-distortion area improves in all 24 tasks, while task-MSE area improves in 20 and worsens in the same four: NLinear Exchange at $H = 336/720$ and DLinear ETTh2 at $H = 336/720$. The areas use the normalized-budget trapezoidal integral, divided by each host's full MSE and then averaged across seeds, as in Figure 4. Capacities instead pool risks before applying the common target.

The sole task whose primary capacity changes with cutoff is DLinear Electricity at $H = 720$. Its projection capacity is 256 at 10^{-12} and 10^{-11}, versus 361 at the other three settings, identically across seeds; canonical masking has capacity 361. The pooled excess MSE at $K = 256$ is 0.8958%, 0.9698%, 1.0094%, 1.1112%, and 1.2084% of full risk, respectively, crossing the fixed 1% target.

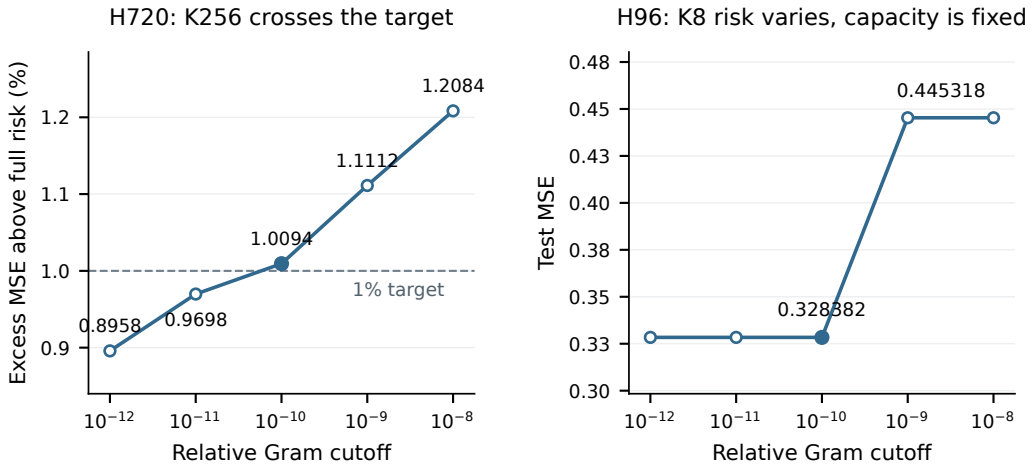


Figure 7: **Spectral truncation can move a capacity threshold or change risk within a fixed capacity.** DLinear on Electricity, with risks averaged over three seeds. Left: at $H = 720$, the retained budget $K = 256$ qualifies only at the two smallest cutoffs. Right: at $H = 96$, risk at $K = 8$ changes substantially although first-hit capacity stays fixed. Filled markers identify the primary cutoff 10^{-10}.

All NLinear tasks retain the same numerical ranks across the five cutoffs; rank changes occur within DLinear. For Electricity at $H = 96$, the pooled $K = 8$ MSE rises from 0.328382 at 10^{-10} to 0.445318 at 10^{-9} and 10^{-8}. Thus agreement of first-hit capacities need not imply agreement of risk profiles. Across the 51 tolerances, projection capacities match the primary cutoff in 1176, 1185, 1206, and 1173 of the 1224 task–tolerance entries for 10^{-12}, 10^{-11}, 10^{-9}, and 10^{-8}, respectively.

Numerical precision. Recomputing the primary cutoff reproduces saved task risks and teacher distortions to a maximum absolute difference of 2.50×10^{-12}. At 10^{-12}, the maximum retained condition number reaches 9.84×10^{11}. A second eigendecomposition implementation, using NumPy in place of PyTorch, recomputes all 3960 interior projections; the maximum absolute test-MSE difference is 1.76×10^{-7}. Two DLinear Electricity tasks exceed the recorded 2×10^{-8} absolute comparison threshold. The two implementations yield identical acceptance decisions across all 51 tolerances.