SHARED RELATIONS IN TIME-SERIES FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Do time-series foundation models organize input histories in a shared way? We compare the matrices of pairwise history similarities produced by ten variants from seven families on 17 datasets. Released pretrained weights (PT) agree more than independent architecture-matched initializations (INIT). On the full inventory, more of the additional agreement lies in the input kernels for mean pooling than for last-position pooling. The last-position residual stays positive across depth, reference expansion and family deletion. The mean-linear endpoint interval crosses zero; mean-RBF and intermediate mean-linear depths stay positive. On later Wikipedia records the mean-linear residual is larger than the last-linear residual. The last-linear remainder is a difference in residual direction: PT and INIT have similar residual norms, and permuting each released parameter tensor removes the additional agreement. Small MLP and CNN predictors that improve within-dataset forecasting also develop additional last-position agreement outside the same input span. The balance between the input span and the residual depends on the readout and the population.

1 INTRODUCTION

Consider the same collection of observed time series passed through two forecasting models. Each model assigns a feature vector to every history. Comparing pairs of those vectors produces a square matrix: entry (i, j) says how similar histories i and j appear to that model. These *relation matrices* can be compared even when the two models have different feature dimensions. Centered kernel alignment (CKA) measures agreement by taking an inner product between their centered, normalized matrices (Kornblith et al., 2019).

A high agreement score leaves an important question open: *which relationships between histories do the models share?* Two histories may have similar overall shape, recurring frequencies, or recent values. Different models may emphasize different combinations of these properties. The feature extraction choice matters too. We call the combination of recorded layer, mean or last-position pooling, and linear or RBF similarity a *readout*. Pooling here selects a representation of the observed history, before the model's forecast projection.

The Platonic Representation Hypothesis proposes that models converge toward shared representations as learning progresses (Huh et al., 2024). Time-series foundation models offer a setting in which to examine that proposal: their architectures, forecasting objectives and training corpora differ, while their inputs have measurable temporal structure. Input kernels make properties such as lagged dependence and recent behavior available as reference directions.

Independent initialization provides another reference. Even without trained weights, networks induce structured similarities between inputs (Lee et al., 2018; Jacot et al., 2018). We therefore compare released pretrained models (PT) with architecture-matched independent initializations (INIT). For each state, we separate agreement into a component in the span of specified input kernels and a remainder orthogonal to that span. The PT–INIT contrast asks how the extra agreement associated with pretrained weights is distributed between these components.

We follow three questions: how much of the PT–INIT contrast lies in the input span; whether residual agreement also develops in ordinary predictors; and whether input features recover model relations on held-out sources. The first comparison is then refined by separating residual norm from residual

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.

direction, and by permuting the released weights inside each parameter tensor. We then connect these comparisons to controlled temporal-content and adaptation experiments.

2 FROM HISTORIES TO COMPARABLE RELATIONS

2.1 HISTORIES, MODEL STATES AND READOUTS

The observational inventory contains 13,621 histories from 17 datasets grouped into ten sources. We use ten variants from seven families: Chronos, TimesFM, Moirai, Sundial, Time-MoE, Mamba4Cast and TiRex. PT denotes released pretrained weights; INIT denotes independent architecture-matched initialization. The original pretraining initializations are unavailable, so INIT uses each model's public construction recipe. Appendix A records model identities, initialization details and dataset provenance.

Histories use the trailing 256 observed values, with fixed standardization and masks. Several of the models accept longer native contexts. Native tokenization and patching map a history to a sequence of hidden vectors. Mean pooling averages the valid historical positions; last pooling takes the final valid position. Each pooled vector is normalized to unit Euclidean norm. A readout then specifies a recorded depth, one of these pooling choices, and a linear or RBF kernel on the resulting vectors. The endpoint is the final recorded representation before forecast projection. Four relative recorded depths let us follow the comparison through the model.

Same histories	Two model readouts	Compare relations
$x_1, \ldots, x_n$	$\text{features}_1 \longrightarrow K$	
	$\text{features}_2 \longrightarrow T$	$\langle K, T \rangle_F = \text{CKA}$

Input kernels define one span: $K = \mathcal{P}(K) + (I - \mathcal{P})(K), \quad T = \mathcal{P}(T) + (I - \mathcal{P})(T).$
Compare the parts within the span and the parts left over.

For a linear readout, the uncentered Gram matrix contains the inner products of the unit feature vectors. An RBF readout instead converts feature distances into similarities. Both give one row and column per history. Centering removes row and column means, and Frobenius normalization puts the matrix on a common scale. Thus models with different hidden dimensions can be compared using the same history indices. We use linear readouts to introduce the results and retain the full readout grid in the appendix.

2.2 CKA SPLIT INTO INPUT SPAN AND RESIDUAL

To locate shared structure, we also compute relation matrices directly from the input histories. Their linear span provides a set of reference directions in the space of matrices. Projecting a model relation into that span asks which part can be expressed as a linear combination of those directions. This follows the idea of adjusting representation similarity for input relations (Cui et al., 2022). We apply the same projection to PT and INIT and compare their components in original CKA units.

For n histories let $H = I - \mathbf{1}\mathbf{1}^\top/n$ and $N(G) = HGH/\|HGH\|_F$. Normalized relations are $K = N(G_K)$ and $T = N(G_T)$, with CKA $A = \langle K, T \rangle_F$. Let $\mathcal{P}$ be the orthogonal projector onto the span of centered input relations and centered identity. Orthogonality gives

$$A = \underbrace{\langle \mathcal{P}(K), \mathcal{P}(T) \rangle_F}_{A_{\text{span}}} + \underbrace{\langle (I - \mathcal{P})(K), (I - \mathcal{P})(T) \rangle_F}_{A_{\text{res}}}, \tag{1}$$
$$\Delta A = \Delta A_{\text{span}} + \Delta A_{\text{res}}, \qquad \Delta A = A^{\text{PT}} - A^{\text{INIT}}.$$

Span: agreement carried by the input-reference directions and the identity baseline.
Residual: agreement between the remaining matrices, without renormalizing them.
Difference Δ: the PT value minus the INIT value, using the same projector.

Centered identity accounts for the agreement expected after randomly permuting history identities in one matrix. Averaging over these relabelings makes all diagonal entries equal and all off-diagonal

entries equal; after centering, the average is proportional to centered identity. Appendix B gives the permutation formula. The residual retains both magnitude and directional agreement. It is small when little energy remains or the remaining directions agree weakly. Keeping the magnitudes makes the components add back to observed CKA. Appendix H separates those two contributions and compares a within-tensor permutation of the released weights.

Each model has its own projection coefficients in the shared span. Fitting them on the evaluated relations gives descriptive accounting. Features requiring scaling use other sources. Section 5 instead fits maps on training sources and tests their predictions on a held-out source.

2.3 Aggregation and uncertainty

We average variant pairs within each cross-family pair, then family pairs, datasets within source, and sources equally. Paired source bootstrap intervals keep model identities and histories fixed. Simultaneous bands cover the named families of depth/readout, family-deletion and recovery comparisons; other intervals are nominal. Appendix A gives the comparison families, resampling rules and analysis status. Each experiment below identifies its evaluation population, because changing the selected histories changes the relation matrices.

3 Input span and the last-position remainder

3.1 Broadening the input reference

We first ask how the PT–INIT difference is distributed between the input span and the residual. For example, an autocorrelation kernel compares histories by their lag dependence, while a recent-segment kernel compares their final observed values. These belong to the base and local groups in Figure 1. Catch22 adds 22 time-series summary features; random ReLU and convolution maps supply nonlinear features without model training. The bank grows from 46 descriptive kernels to 67 with random features, then to a nested expansion of 134; centered identity is included separately. Each expansion preserves the preceding columns and uses the same histories. Appendix B gives the construction. Shapley allocation divides the span contrast among overlapping groups by averaging their contributions over insertion orders (Shapley, 1952).

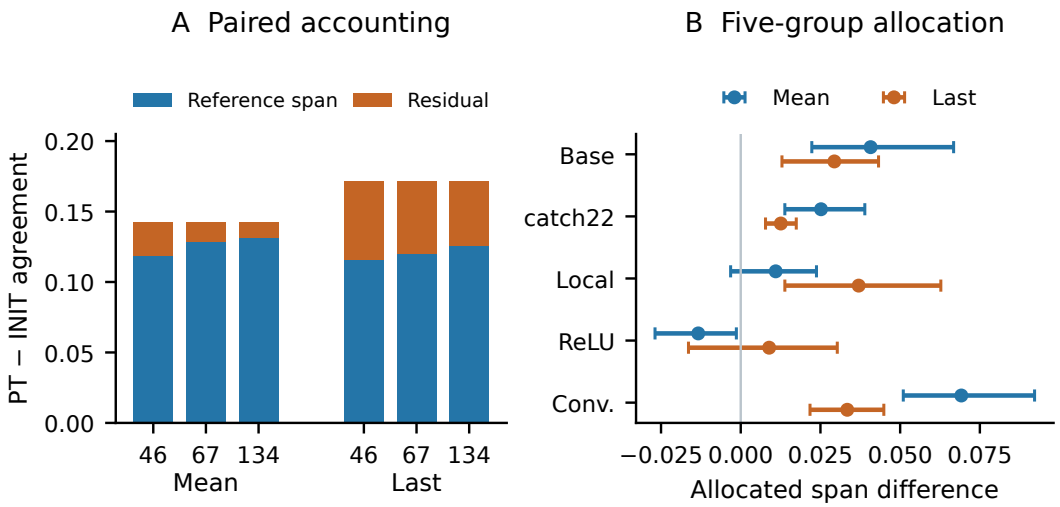

Figure 1: **Input accounting on the full 17-dataset inventory.** Left: endpoint linear PT–INIT CKA split into span and residual components. The three banks contain 46 descriptive, 67 mixed, and 134 expanded kernels, each plus identity. Right: five-group Shapley allocation of the 67-kernel span contrast beyond identity, with nominal 95% source intervals. The groups are the base temporal references, catch22, local features, random ReLU features and random convolutions.

Allocations can be negative because the quantity being divided is a PT–INIT difference in projected inner products.

3.2 Depth, family and population checks

We next hold the expanded bank fixed and examine the remainder. Depth checks ask whether it appears only at the endpoint. Family deletion asks whether one family accounts for its sign. Two later-collected populations then change the histories on which the same type of comparison is made. Figure 2 places these checks together; their bars have different meanings, identified in the caption.

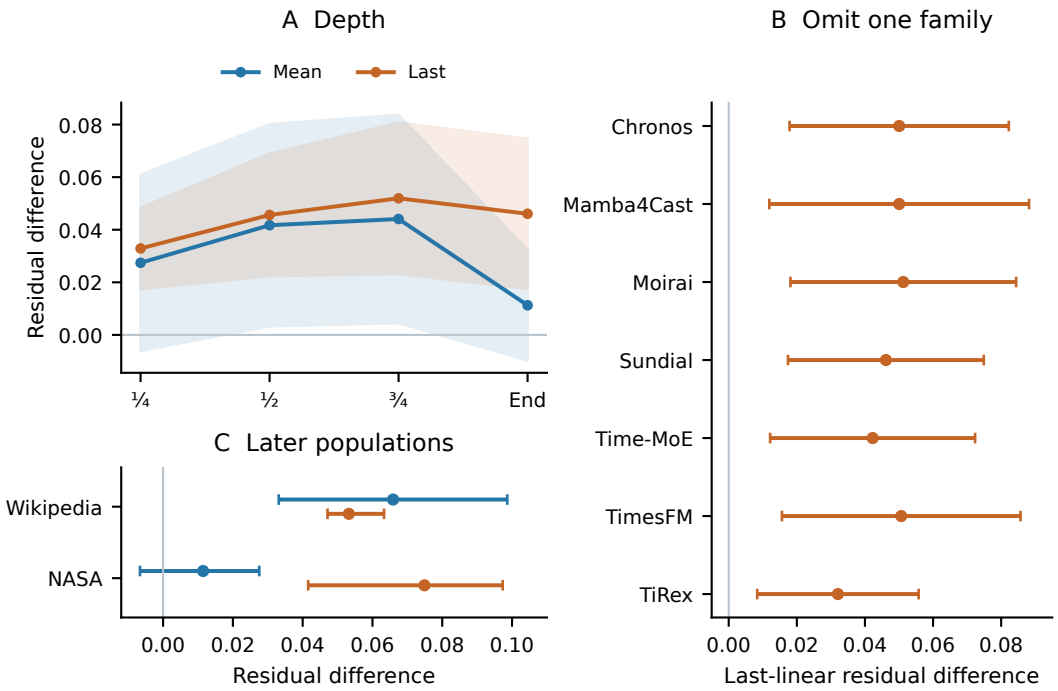

Figure 2: **Residual PT–INIT agreement with 134 input kernels plus identity.** A: four depths on the full inventory; bands jointly cover the depth/readout/bank contrasts. B: last-linear endpoint after each family deletion from that inventory; bars are joint bands across the deletion comparisons. C: the marker is the timing-verified estimate, and the bar runs from the lowest to the highest single-family deletion on that population. Appendix A specifies the bands in A and B.

On the full inventory, last-position residuals stay positive at every depth (Table 4). The mean-linear endpoint interval crosses zero. Mean-RBF at the endpoint, and mean-linear at depths $1/2$ and $3/4$, stay positive.

3.3 Residual direction and weight arrangement

Pairwise, the residual in Equation (1) is a cosine times the product of residual Frobenius norms. At the last-linear endpoint, splitting ΔA_{res} gives a direction term of 0.042 [0.025, 0.061] and an energy term of 0.004 [-0.002, 0.011]. Mean residual norms are 0.557 for PT and 0.549 for INIT. One within-tensor permutation of the released weights, seed 91026, keeps each parameter tensor's values, mean and norm. Shuffled last-linear residual agreement is 0.006 [0.004, 0.009]: -0.055 [-0.081, -0.033] relative to PT and -0.009 [-0.015, -0.004] relative to INIT. Appendix H gives the identity, the other readouts and the protocol.

The external check uses Wikipedia page views and NASA weather records collected later than the timing-verified weights. Eight variants from five families meet the timing requirements. Endpoint residuals on Wikipedia are 0.066 for mean-linear and 0.053 for last-linear, with span shares 0.30 and 0.62. On NASA they are 0.011 and 0.075, with span shares 0.97 and 0.81. The NASA mean-linear

deletion range crosses zero. NASA locations are spatially related. Appendix E records dates and weight hashes.

History-count and concatenated-token checks give additional endpoint sensitivities, with nominal intervals reported separately in Appendix B.5.

A follow-up asks whether one input combination can predict another family's relations. On the same histories, a combination fitted to six families predicts the seventh; for PT last-linear relations, the expanded combination improves on both uniform weights and a training-selected single kernel. This supplies a use for the common reference space without requiring identical model-specific coefficients (Appendix G).

4 Ordinary forecasting training and residual agreement

Can ordinary forecasting training also produce additional agreement outside the input span? We train an MLP and a small dilated CNN independently to predict a future segment. Both predict a correction to the last observed value, with zero-initialized output heads, so initial forecasts equal persistence. Their hidden features are independently initialized. Training and model selection use forecasting losses alone, without foundation-model targets or relation matching.

4.1 Establishing a forecasting comparison

The previously fixed test set contains 128 histories per dataset. Remaining original series supply training and validation, with duplicates of test series and histories excluded. Forecasting uses an observed prefix and a 32-point future suffix. Thirteen datasets from seven sources have enough eligible series. We fit a within-dataset arm and a held-source arm with three seeds, selecting learning rate and checkpoint by validation forecasting loss. Appendix C specifies eligibility and all source-level scores.

The within-dataset arm tests transfer to held-out series from the training dataset. The held-source arm excludes the entire evaluated source from fitting and validation. For geometry, small predictors and foundation models receive the same full observed histories. Forecasting uses a shorter prefix to reserve the future target.

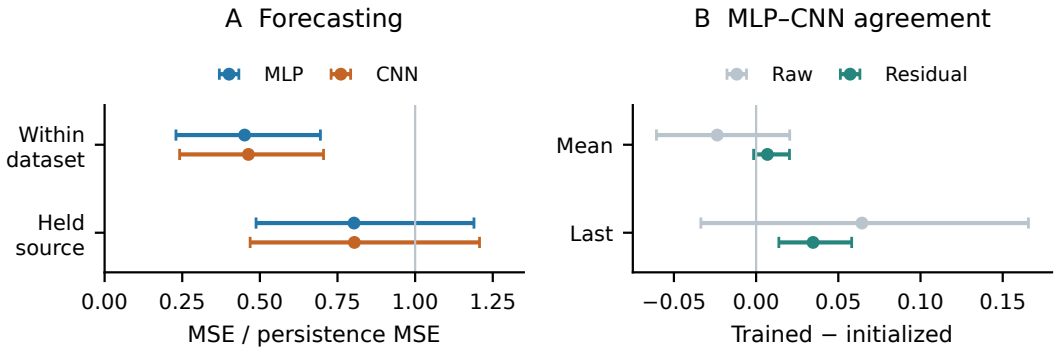

Figure 3: **Forecasting and relation changes in ordinary predictors.** Left: held-out MSE relative to persistence; one means equal error. Within-dataset training improves the aggregate for both models, while held-source intervals cross one. Right: within-dataset MLP–CNN changes on matched histories. The MLP uses the same unit 64-vector for both pooling labels; the CNN normalizes its mean and last vectors separately. Bars are nominal 95% paired intervals over seven sources, conditional on fitted models.

Within-dataset MSE is approximately 45–46% of persistence for both architectures, with favorable point estimates in every source. Every selected checkpoint is after initialization. The held-source aggregate intervals cross persistence, so the within-dataset arm supplies the primary learned-relation comparison. The forecast test establishes that this arm learned useful predictive structure on held-out series.

4.2 Comparing relations before and after training

After freezing the forecast-based choices, we apply the same 134-kernel accounting to the MLP–CNN pair before and after training. The last-linear residual increases, with a nominal source interval above zero. The foundation-model comparison uses these same histories from the 13 eligible datasets. Table 1 collects the magnitudes by evaluation population after introducing the recovery comparison.

Ordinary forecasting training therefore also develops additional last-position residual agreement. Mean-linear residual change and raw last-linear change remain uncertain.

5 Predicting model relations on held-out sources

Here a *teacher* is a model whose endpoint features are the targets of a fitted input-to-feature map. Recovery predicts those features from input histories, forms a linear kernel from the predicted vectors, and compares that kernel with the teacher's relation matrix on a held-out source. This asks whether a fitted map can generalize the relation, complementing the population-specific projection in Section 3.

5.1 Linear and nonlinear recovery

Each of ten source folds reserves one source for testing and the next for validation; the other eight supply fitting histories. Linear and RBF kernel ridge regression use the same input blocks: raw history, recent values, catch22 and random convolutions. Validation selects regularization and RBF width. Each teacher is fitted separately, with every model family supplying training targets. The held-out unit is a source, and each fold has already supervised all teacher families on other sources.

A direct-input comparator selects a kernel and width by the same validation criterion, without fitting a ridge map to teacher features. This distinguishes the effect of learning a predictor from choosing an input similarity already present in the candidate set. Input scaling and fitting exclude the test source. Predictions are normalized row by row before their linear relations are scored against each teacher using CKA; the full fitting and selection rules are in Appendix D.

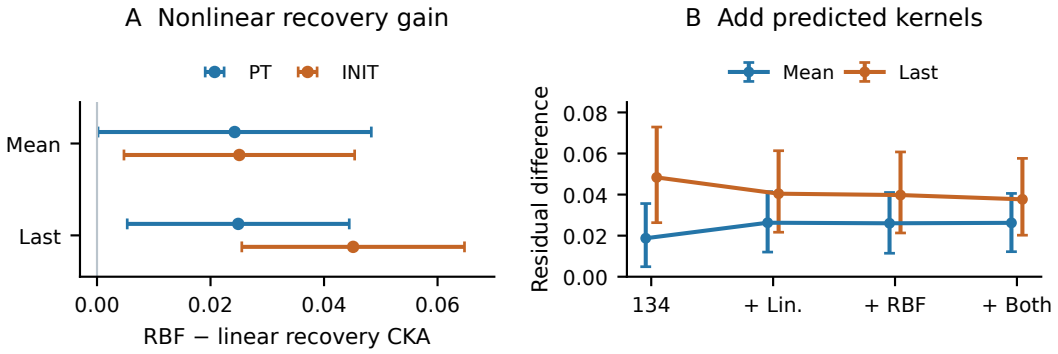

Figure 4: **Held-source recovery (left) and exploratory reference augmentation (right).** Left: RBF-minus-linear recovery with a joint band including the direct-input comparisons. Absolute RBF CKA is 0.61–0.71; the gains shown are 0.024–0.045. Right: a secondary diagnostic, added after inspecting recovery, augments the reference span with PT and INIT predicted kernels. All right-panel contrasts use the same 128 histories per dataset, a population distinct from Figures 1–3. Its bars are nominal source intervals.

RBF ridge improves over linear ridge and the validation-selected direct input kernel in all four state/pooling combinations under the joint band. The PT mean-pooled lower bound is close to zero. The gains appear for INIT as well as PT teachers: both states have relations that are better recovered by a nonlinear map of the same inputs (Figure 4, left). Absolute recovery levels and every contrast are retained in Table 10.

5.2 Adding predicted relations to the reference span

Can the source-held-out predictions account for some of the remainder? In a secondary, exploratory analysis, we add predicted kernels from all PT and INIT teachers to the 134-kernel bank and use this enlarged span for both true states. Equation (1) explains how to read this test. Expanding a span reduces each matrix's residual energy, but a signed PT–INIT difference in residual inner products can move in either direction. Its change depends on what is removed from each state.

On the recovery population, adding both learned banks reduces last-linear additional residual agreement, which remains positive (Figure 4, right). Linear and nonlinear augmentations have similar effects on this contrast. Table 1 places its magnitude alongside the earlier comparisons: all use CKA units, but their history sets differ.

Table 1: **Last-linear residual comparisons by population.** Approximate CKA differences with 134 input kernels plus identity. Foundation-model contrasts are PT–INIT; the small-predictor contrast is trained–initialized. The last row is exploratory reference augmentation.

Evaluation population	Residual comparison
Full inventory: 17 datasets	Foundation models: 0.046
Matched predictor histories: 13 eligible datasets, 128 histories each	Small predictors: increase of 0.035; foundation models on the same histories: 0.042
Recovery histories: 17 datasets, 128 histories each	Foundation models: 0.048 $\rightarrow$ 0.038 after adding both predicted-kernel banks

For mean pooling, each state's residual agreement decreases, with a larger decrease for INIT. The PT–INIT change is consequently positive in point estimate, but its interval crosses zero. The predictions exclude the evaluated source; the subsequent projection coefficients are fitted on its true relations. Recovery and augmentation therefore answer complementary questions: whether relations can be predicted across sources, and where those predicted directions lie in the paired accounting.

6 Temporal content and conditional adaptation

The comparisons above locate a remainder and test how it arises and can be predicted. Two controlled settings connect that geometry to temporal behavior. They use separate synthetic populations and outcomes. The first shows which temporal factors each pooling retains; the second adapts two models toward one shared relation.

6.1 What changes when pooling changes?

This controlled illustration asks which temporal factors pooling retains; it does not identify the real-data remainder. On a balanced panel, sinusoids cross four frequencies and four phases with repeated noisy realizations. The factor labels define orthogonal subspaces on the history axes. Splitting each centered relation by those subspaces decomposes total CKA into frequency, phase, interaction and the remaining cross/realization terms. This isolates known temporal factors directly, whereas the real-data projection uses input kernels. Appendix F gives the panel and factor definitions.

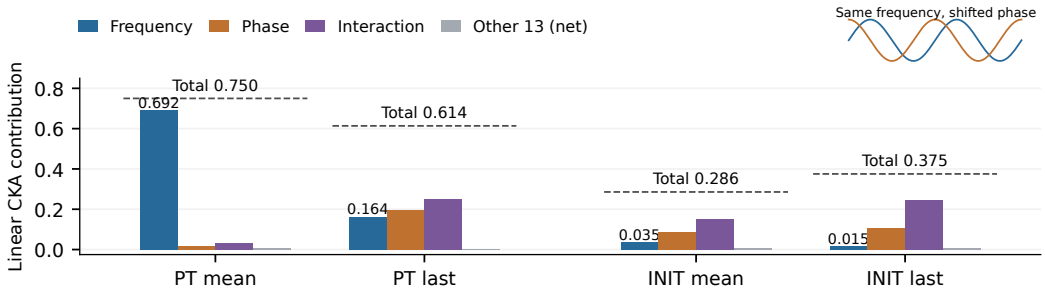

Figure 5: **Temporal content of endpoint linear CKA on controlled histories.** The six-family panel crosses frequency and phase, with noise standard deviation 0.25 and two fixed panels. PT mean and last readouts are adjacent. Bars sum to each dashed total; "other" combines 13 terms and can conceal cancellation. The waveforms illustrate equal frequency at different phases. This population and factor decomposition differ from the real-data reference accounting.

Pooling changes the temporal content behind a single alignment score. The controlled panel therefore explains why a claim about shared structure must specify its readout.

6.2 Does favoring a common relation improve transfer?

A separate intervention adapts TimesFM and TiRex toward a common autocorrelation (ACF) relation and evaluates a synthetic frequency shift. Each architecture is compared with its own forecast-only arm at matched updates and seeds; adapter locations differ across architectures. Figure 6 tracks mutual RBF *excess* (CKA minus its permutation reference) and forecast error.

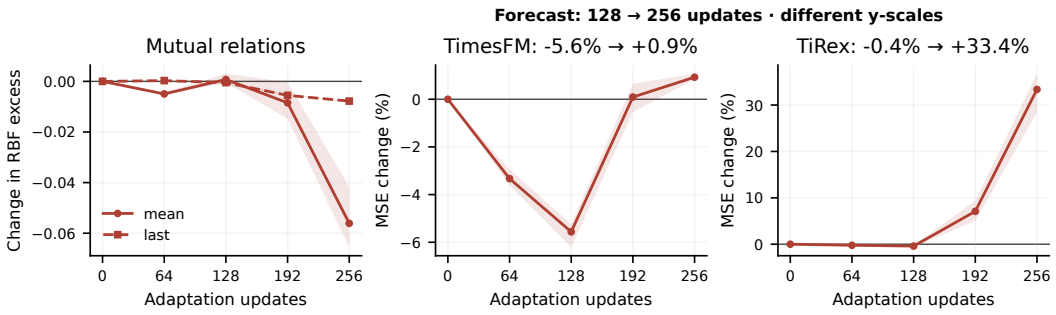

Figure 6: **Agreement and forecasting during common-target adaptation.** PT TimesFM and TiRex, synthetic frequency shift, ACF minus matched forecast-only adaptation. Negative values mean less mutual RBF excess (left) or lower forecast MSE (middle/right). Lines join saved checkpoints in a post hoc analysis; shading is the range across three seeds, not a confidence interval. The two forecast panels use different vertical scales.

Closer correspondence to the same target does not ensure greater mutual agreement or consistently better forecasting: at 256 updates both models move closer to ACF, while mutual agreement falls and TiRex forecasting worsens; TimesFM's error increase remains uncertain. At 128 updates, both forecast point estimates instead improve while the mutual-excess interval crosses zero. Geometrically, agreement in directions orthogonal to the common target can fall more than agreement along it rises. These trajectories provide context for the reference, predictor and recovery comparisons; Appendix F retains the detailed intervention results, including inconclusive real-data transfer.

7 Related work

7.1 Comparing representations

Representational similarity analysis and CKA compare relations among samples (Kriegeskorte et al., 2008; Kornblith et al., 2019); linear CKA is the RV coefficient (Robert & Escoufier, 1976). Feature-subspace comparisons address another aspect of representation similarity (Raghu et al., 2017; Bansal et al., 2021; Klabunde et al., 2025). Sensitivity to transformations and dimensionality motivates explicit control of the readout and reference (Davari et al., 2023; Murphy et al., 2024).

7.2 Input references and recovery

Our paired projection uses the input-adjustment principle described in Section 2, while preserving an additive split of the PT–INIT contrast. catch22 and random convolutions supply established input representations (Lubba et al., 2019; Dempster et al., 2020); centered-alignment kernel learning and relational distillation motivate the separate recovery question (Cortes et al., 2012; Park et al., 2019).

7.3 Shared representations and transfer

The Platonic hypothesis motivates comparisons across independently developed models and reports mutual nearest neighbors (Huh et al., 2024); the comparisons here use CKA. Temporal-concept and architectural-bias studies motivate inspecting what their agreement contains (Wiliński et al., 2025; Yu et al., 2026). Calibration work shows that global relations and local neighborhoods can behave differently (Gröger et al., 2026), while similarity evaluation studies examine connections to downstream behavior (Ding et al., 2021). Our comparisons place foundation models beside initialized teachers and ordinary trained predictors, using held-source recovery and conditional adaptation to connect descriptive agreement with tests of learning and transfer.

8 Discussion

The comparisons support a view of shared relations with two distinguishable components. On the full inventory, a substantial part of additional pretrained agreement follows the supplied input relations, with a larger span share for mean pooling than for last-position pooling. Last-position residual agreement stays positive across depth, reference expansion and family deletion. Mean-RBF and intermediate mean-linear depths stay positive as well; the mean-linear endpoint interval crosses zero. On Wikipedia the mean-linear residual is larger than the last-linear residual. The last-linear remainder is a difference in residual direction: PT and INIT have similar residual norms, and permuting each released parameter tensor removes the additional agreement. Ordinary forecasting models also develop additional last-position residual agreement. On the controlled panel, changing pooling changes which temporal factors carry agreement.

An input-aligned component has a useful interpretation in its own right. Autocorrelation, local behavior and random convolution features describe relationships that a forecasting system can learn to emphasize. The projection locates agreement with these supplied descriptions; the remaining matrices identify where those descriptions are incomplete. Nonlinear recovery adds a complementary observation: predictive maps can reproduce more of both pretrained and initialized relations on held-out sources. Recoverability alone therefore does not identify which relationships were acquired through pretraining.

The reference bank was expanded on this same source inventory. The ordinary predictors differ in capacity and training data from the foundation models, and recovery supervises every teacher family. NASA locations are spatially related. On the recovery population, the mean-pooled change after adding predicted kernels has an interval crossing zero.

Kernels from the real next horizon and from seasonal-naive, Holt–Winters and autoregressive forecasts change last-linear ΔA_{res} by at most 0.008, and the residual stays positive (Appendix I). What those residual directions contain, and which of them help forecasting, remains open. Two further tests can alter phase alone or the recent segment alone, and can refit every preprocessing step on training sources before evaluating relation prediction and forecast utility on new sources.

AI USE STATEMENT

We used generative AI tools to assist with research planning and iteration, literature search and discussion, experiment-code implementation and debugging, data-analysis and verification scripting, and manuscript drafting and editing. We also used AI agents to help organize experiments and inspect results. We obtained the reported numerical results from executable model evaluations and saved numerical artifacts. We take full responsibility for the work.

REPRODUCIBILITY STATEMENT

The numerical supplement provides frozen protocols, source/cell statistics, input-reference parameters, compact relation coordinates and standalone numerical checks. It reconstructs the expanded-bank and group-allocation results and the forecast, geometry, recovery and external summaries, including their stated intervals. Additional records document the history-count and concatenated-token checks, the controlled-content and adaptation figures, and the shared-combination follow-up. Figure and table generation uses saved numerical records. The compact release supports numerical reconstruction; rerunning training additionally requires the raw datasets, foundation-model weights, full features and training checkpoints. Appendices A–E specify model identities, exclusions, scaling, fitting and uncertainty. Appendix F separates earlier exploratory evidence from the present comparisons; Appendix G gives the combination protocol and its preprocessing boundary. Appendix H reports the residual-direction split and the within-tensor shuffle. Appendix I reports the forecast-kernel enlargement. Both replay the published endpoint cells. Their records are `analysis/opus55_followup_20260926/`.

REFERENCES

Amazon. Chronos-Bolt-Base Model Card. Hugging Face model documentation, 2024. URL https://huggingface.co/amazon/chronos-bolt-base.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the Language of Time Series. *arXiv:2403.07815*, 2024. URL https://arxiv.org/abs/2403.07815.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, Mononito Goswami, Shubham Kapoor, Danielle C. Maddix, Pablo Guerron, Tony Hu, Junming Yin, Nick Erickson, Prateek Mutalik Desai, Hao Wang, Huzefa Rangwala, George Karypis, Yuyang Wang, and Michael Bohlke-Schneider. Chronos-2: From Univariate to Universal Forecasting. *arXiv:2510.15821*, 2025. URL https://arxiv.org/abs/2510.15821.

Andreas Auer, Patrick Podest, Daniel Klotz, Sebastian Böck, Günter Klambauer, and Sepp Hochreiter. TiRex: Zero-Shot Forecasting Across Long and Short Horizons with Enhanced In-Context Learning. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2505.23719.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. In *Advances in Neural Information Processing Systems*, 2021. URL https://arxiv.org/abs/2106.07682.

Sathya Kamesh Bhethanabhotla, Omar Swelam, Julien Siems, David Salinas, and Frank Hutter. Mamba4Cast: Efficient Zero-Shot Time Series Forecasting with State Space Models. *arXiv:2410.09385*, 2024. URL https://arxiv.org/abs/2410.09385.

Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Algorithms for Learning Kernels Based on Centered Alignment. *Journal of Machine Learning Research*, 13:795–828, 2012. URL https://www.jmlr.org/papers/v13/cortes12a.html.

Tianyu Cui, Yogesh Kumar, Pekka Marttinen, and Samuel Kaski. Deconfounded representation similarity for comparison of neural networks. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://papers.nips.cc/paper_files/paper/2022/hash/79cbf4f96c2bcc67267421154da689dd-Abstract-Conference.html.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2310.10688.

MohammadReza Davari, Stefan Horoi, Amine Natik, Guillaume Lajoie, Guy Wolf, and Eugene Belilovsky. Reliability of CKA as a similarity measure in deep learning. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.16156.

Angus Dempster, François Petitjean, and Geoffrey I. Webb. ROCKET: Exceptionally Fast and Accurate Time Series Classification Using Random Convolutional Kernels. *Data Mining and Knowledge Discovery*, 34(5):1454–1495, 2020. doi: 10.1007/s10618-020-00701-z. URL https://arxiv.org/abs/1910.13051.

Frances Ding, Jean-Stanislas Denain, and Jacob Steinhardt. Grounding representation similarity through statistical testing. In *Advances in Neural Information Processing Systems*, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/0c0bf917c7942b5a08df71f9da626f97-Abstract.html.

Rakshitha Godahewa, Christoph Bergmeir, Geoffrey I. Webb, Rob J. Hyndman, and Pablo Montero-Manso. Monash Time Series Forecasting Archive. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. URL https://arxiv.org/abs/2105.06643.

Fabian Gröger, Shuo Wen, and Maria Brbić. Revisiting the Platonic Representation Hypothesis: An Aristotelian View. *arXiv:2602.14486*, 2026. URL https://arxiv.org/html/2602.14486v2.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The Platonic Representation Hypothesis. In *International Conference on Machine Learning*, pp. 20617–20642, 2024. URL https://proceedings.mlr.press/v235/huh24a.html.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://arxiv.org/abs/1806.07572.

Julie Josse, Jérôme Pagès, and François Husson. Testing the significance of the RV coefficient. *Computational Statistics & Data Analysis*, 53(1):82–91, 2008. doi: 10.1016/j.csda.2008.06.012.

Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of neural network models: A survey of functional and representational measures. *ACM Computing Surveys*, 57(9):242:1–242:52, 2025. URL https://arxiv.org/abs/2305.06329.

Dmitry Kobak, Wieland Brendel, Christos Constantinidis, Claudia E. Feierstein, Adam Kepecs, Zachary F. Mainen, Xue-Lian Qi, Ranulfo Romo, Naoshige Uchida, and Christian K. Machens. Demixed principal component analysis of neural population data. *eLife*, 5:e10989, 2016. doi: 10.7554/eLife.10989. URL https://elifesciences.org/articles/10989.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of Neural Network Representations Revisited. In *International Conference on Machine Learning*, pp. 3519–3529, 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2: 4, 2008. URL https://www.frontiersin.org/journals/systems-neuroscience/articles/10.3389/neuro.06.004.2008/full.

Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S. Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as gaussian processes. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1711.00165.

Yong Liu, Guo Qin, Zhiyuan Shi, Zhi Chen, Caiyin Yang, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Sundial: A Family of Highly Capable Time Series Foundation Models. *arXiv:2502.00816*, 2025. URL https://arxiv.org/abs/2502.00816.

Carl H. Lubba, Sarab S. Sethi, Philip Knaute, Simon R. Schultz, Ben D. Fulcher, and Nick S. Jones. catch22: CAnonical Time-series CHaracteristics. *arXiv preprint arXiv:1901.10200*, 2019. URL https://arxiv.org/abs/1901.10200.

Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The M4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1): 54–74, 2020. doi: 10.1016/j.ijforecast.2019.04.014. URL https://www.sciencedirect.com/science/article/pii/S0169207019301128.

Alex Murphy, Joel Zylberberg, and Alona Fyshe. Correcting biased centered kernel alignment measures in biological and artificial neural networks. *arXiv preprint arXiv:2405.01012*, 2024. URL https://arxiv.org/abs/2405.01012.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational Knowledge Distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. URL https://arxiv.org/abs/1904.05068.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, 2017. URL https://arxiv.org/abs/1706.05806.

P. Robert and Yves Escoufier. A unifying tool for linear multivariate statistical methods: The RV-coefficient. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 25(3): 257–265, 1976. doi: 10.2307/2347233. URL https://academic.oup.com/jrsssc/article/25/3/257/6953583.

Lloyd S. Shapley. A value for n-person games. Technical Report P-295, RAND Corporation, 1952. URL https://www.rand.org/pubs/papers/P295.html.

Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. Time-MoE: Billion-Scale Time Series Foundation Models with Mixture of Experts. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2409.16040.

Le Song, Alex Smola, Arthur Gretton, Justin Bedo, and Karsten Borgwardt. Feature selection via dependence maximization. *Journal of Machine Learning Research*, 13:1393–1434, 2012. URL https://www.jmlr.org/papers/v13/song12a.html.

Michał Wiliński, Mononito Goswami, Willa Potosnak, Nina Żukowska, and Artur Dubrawski. Exploring Representations and Interventions in Time Series Foundation Models. In *International Conference on Machine Learning*, pp. 66861–66886, 2025. URL https://proceedings.mlr.press/v267/wilinski25a.html.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2402.02592.

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 the Implicit Biases of Design Choices for Time Series Foundation Models. In *International Conference on Learning Representations*, pp. 71998–72049, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/753d9584b57ba01a10482f1ea7734a89-Abstract-Conference.html.

A MODELS, POPULATIONS AND INFERENCE

Table 2: The three comparisons use distinct fitting and evaluation units. PT–INIT always compares states on identical histories within a given comparison.

Comparison	What is fitted	Where it is evaluated
Reference accounting	Projection onto input-kernel span	Same relation matrices; full inventory
Ordinary predictors	Forecast model, selected by forecast loss	Held-out series in eligible datasets
Held-source recovery	Map from inputs to each teacher's features	Histories from a source excluded from fitting

A.1 INVENTORY AND EXTRACTION

Family	**Variants**
Chronos	Chronos-2, Chronos-Bolt-base, Chronos-T5-base (Ansari et al., 2025; Amazon, 2024; Ansari et al., 2024)
TimesFM	TimesFM-2.0-500M, TimesFM-2.5-200M (Das et al., 2024)
Moirai	Moirai-1.1-R-base (Woo et al., 2024)
Sundial	Sundial-base-128M (Liu et al., 2025)
Time-MoE	Time-MoE-200M (Shi et al., 2025)
Mamba4Cast	Mamba4Cast (Bhethanabhotla et al., 2024)
TiRex	Original TiRex-35M (Auer et al., 2025)

The canonical dataset provenance follows the Monash archive and M4 competition (Godahewa et al., 2021; Makridakis et al., 2020). Table 3 lists the selected population. Source groups determine weighting and resampling. Public training corpora can overlap these datasets, and exact window membership is generally unknown.

Table 3: Canonical inventory; source grouping governs aggregation and resampling.

Source	Dataset	Histories
m4	`m4_hourly`	414
	`m4_daily`	1024
	`m4_weekly`	359
	`m4_monthly`	1024
	`m4_quarterly`	1024
	`m4_yearly`	1024
wikipedia	`kaggle_web_traffic_weekly`	1024
	`wiki-rolling_nips`	1024
	`wiki2000_nips`	1024
traffic	`traffic`	862
london	`london_smart_meters_without_missing`	1024
car parts	`car_parts_without_missing`	1024
hospital	`hospital`	767
electricity	`electricity_nips`	370
kdd	`kdd_cup_2018_without_missing`	270
wind	`wind_farms_without_missing`	339
weather	`weather`	1024

Native extraction preserves each model's tokenization, patching, masks and normalization. The endpoint is the final recorded representation before forecast projection: encoder states for Chronos

and Moirai, so the Chronos-T5 endpoint is the encoder rather than the decoder that emits the forecast; historical decoder states for TimesFM, Sundial and Time-MoE; and historical stack states for Mamba4Cast and TiRex. Mean pooling averages valid historical-token features; last pooling uses the final valid historical position. Feature rows are normalized before constructing a linear Gram. Observational RBF relations use the within-population median-distance convention. Source-held-out input references and training targets have separate bandwidth rules, specified below.

With L recorded positions, relative depth $q \in \{1/4, 1/2, 3/4, 1\}$ selects the zero-based index $\lceil q(L-1) \rceil$. The recorded sets include architecture-specific input and final-normalization positions. Mamba4Cast has only three positions, so relative depths $1/4$ and $1/2$ select the same layer (indices 1, 1, 2 and 2). Recorded positions differ in architectural role and computational cost. The numerical release retains the exact layer identities.

A.2 Initialization and sensitivity scope

The full-depth analysis uses INIT seed 42. Two further complete endpoint inventories use seeds 43 and 44. Chronos-T5 uses its configured Transformers constructor, including dimension-dependent T5 initialization. Chronos-2/Bolt, Moirai, Mamba4Cast, Time-MoE and Sundial use their configured public constructors without learned tensors. Inference-only modules require explicit completion: TimesFM uses unit effective RMS scales and attention-scaling defaults; TiRex uses linear-layer constructors, unit effective normalization, headwise weights with standard deviation $\sqrt{2/(5 \cdot 512)}$, zero recurrent sLSTM weights and block-dependent forget-gate biases. Coverage checks first mark parameters nonfinite and verify that every parameter is filled. The released models' original pretraining initializations are unavailable.

Additional endpoint checks multiply each unique rank-at-least-two parameter tensor at seed 42 by $\sqrt{0.5}$ or $\sqrt{2}$. Scalars, one-dimensional parameters and buffers stay fixed. This shape-defined perturbation includes embeddings and higher-dimensional state parameters. The resulting 340 model/dataset cells pass finite-output and repeated-forward checks. Seed and scale checks use the earlier 22/35/67-kernel banks. The 134-kernel depth analysis uses seed 42. Each inventory retains the same ten source groups for resampling.

A.3 Distinct experimental populations

The full-inventory analysis uses all 13,621 histories. Family deletion removes each family from this same inventory. External checks use 128 Wikipedia and 128 NASA histories and the model subsets in Appendix E. Held-source ridge recovery uses a previously fixed 128 histories per canonical dataset (2,176 total), with 64 per training dataset for fitting. Small-forecaster geometry uses the same selection on the 13 eligible datasets (1,664 histories); forecasting evaluates only histories meeting its prefix/suffix eligibility rules. The full-inventory and ridge-augmentation comparisons therefore have different evaluation populations, identified in each figure.

A.4 Weighting and uncertainty

For a cross-family pair, first average all variant-pair values, then average family pairs equally. Within a source, average datasets equally; finally average sources equally. PT and INIT retain identical weights. The main source bootstrap draws ten sources with replacement, 10,000 times, seed 8606. Follow-up analyses use seed 8906; small-predictor intervals resample their seven eligible sources. Histories, fitted weights and model identities remain fixed. Recovery first averages variants within each teacher family and families equally, then datasets and sources. Small-predictor results also average the three training seeds before source resampling.

Nominal intervals are percentile intervals from paired draws. For a named family of m contrasts with estimate $\hat{\theta}_j$, bootstrap estimate θ^*_{bj} and bootstrap standard deviation s_j, the joint band is

$$\hat{\theta}_j \pm q_{.95} s_j, \qquad q_{.95} = \mathrm{quantile}_{.95}\left[\max_{j \le m} \left| \frac{\theta^*_{bj} - \hat{\theta}_j}{s_j} \right| \right]. \tag{2}$$

The families contain 32 residual contrasts (two banks, four depths, four readouts), 112 deletion contrasts (seven omitted families, four depths, four readouts), and eight recovery contrasts (two states,

two pools, two baselines). The joint bands cover these three specified comparison families; other exploratory intervals are nominal. Source resampling is conditional on the chosen inventory and fixed model releases.

A.5 Analysis status and verification

The original inventory, reference expansion, group allocations and sensitivity analyses were developed iteratively. The new small-forecaster and ridge protocols were written before their fitting and evaluation; they remain follow-ups on already studied sources. The matched addition of predicted kernels to the 134-kernel bank was added after inspecting the primary recovery results and is exploratory. External accounting and family deletion extend existing cached populations.

Checks cover original-series split disjointness, all selected validation minima, 120 saved forecast runs and independent checkpoint replays. Three replayed fits reproduce parameters exactly. An independent projector reconstructs external accounting within 1.4×10^{-15}, and the independently implemented matched FM pipelines agree within 1.1×10^{-16}. Standalone table reconstruction passes 4,281 checks within 2.3×10^{-16}. One float32 centering diagnostic on a near-constant training prefix exceeded its initial absolute tolerance; the failed diagnostic is retained, and an exact-formula/forward-error check explains the discrepancy without retraining. Archived first-pass corrections to feature normalization and external bandwidth handling are documented; all manuscript figures use the final verified outputs.

The shared-combination follow-up in Appendix G has a separate protocol. Its source exclusions apply to coefficient fitting and condition on frozen input-reference preprocessing.

B Reference construction and algebra

B.1 Permutation reference and projection

For normalized centered relations K, T, the fixed-matrix sample-permutation expectation is $\operatorname{tr}(K)\operatorname{tr}(T)/(n-1)$ (Josse et al., 2008). Centered identity $J = H/\sqrt{n-1}$ has unit Frobenius norm, and its contribution is exactly this product. Hence excess alignment is

$$E = \langle K, T\rangle_F - \langle K, J\rangle_F \langle T, J\rangle_F. \tag{3}$$

Excess subtracts the fixed-matrix permutation mean while keeping the observed spectra fixed. The separate unbiased-HSIC sensitivity is specified below.

To form a reference span, center and Frobenius-normalize each input relation, include J, and compute a rank-revealing SVD. We use relative tolerance 10^{-10} and absolute tolerance 10^{-12}. Symmetric matrices may be packed by retaining one triangle and multiplying off-diagonal coordinates by $\sqrt{2}$; this preserves Frobenius inner products. Orthogonal projection gives Equation (1) because cross-terms between the span and its orthogonal complement vanish. Some distance-derived input relations are indefinite; their scores are centered Frobenius alignment, with the same projection identity.

A signed span share is the ratio of the aggregated ΔA_{span} to the aggregated ΔA. It describes a fraction of the signed agreement contrast. Small denominators can make this ratio unstable, and signed allocations can fall outside $[0, 1]$. Enlarging a span reduces each relation's residual squared norm; residual inner products and their PT–INIT difference can increase or decrease.

B.2 Descriptive and random-feature banks

The base bank contains 22 kernels: linear kernels from 256 masked raw values, 32 autocorrelation lags, 129 power ordinates and 15 summary statistics; and three RBF widths for each of these four distances, DTW and patch-Wasserstein. Summary statistics cover location, scale, quantiles, extremes, change magnitude, sign crossings and third/fourth moments. DTW uses squared local costs on valid histories. Patch-Wasserstein uses uniform transport between sixteen length-16 patches with squared Euclidean cost. Base linear input features retain their row magnitudes; model feature rows have unit norm.

For squared cost D, the input RBF rule is $\exp[-D/(2ms^2)]$, $s \in \{1/2, 1, 2\}$, where D already contains the squared cost. Width summaries use other-source histories and the recorded source-

balanced aggregation. Features requiring learned scaling exclude the evaluated source. Catch22 uses official pycatch22 0.4.5 (Lubba et al., 2019), with missing features for histories shorter than ten or constant histories, other-source median imputation and IQR scaling, and zero IQR replaced by one. Its scaled rows are normalized for the linear kernel. Medians/IQRs pool training histories; comparison weights and bandwidth aggregation are source-balanced. Four catch22 relations add linear and three RBF kernels.

Local references use final 16/32-value tails, histories weighted with half-life 16/64, and a three-coordinate endpoint summary (last value, final up-to-16 mean, and least-squares slope). At lag ℓ, weighted values are multiplied by $2^{-\ell/(2h)}$. Tails preserve whole-history standardization and zero padding. Endpoint summaries use valid observations and other-source balanced first/second moments. Five maps times four kernels contribute 20 directions. Base, catch22 and local groups therefore form the 46-kernel descriptive bank.

Random ReLU maps use Gaussian weights scaled by $1/\sqrt{256}$ and Gaussian biases, with nested widths 128, 512 and 2,048 under seeds 8501–8503. Each row-normalized map supplies one linear kernel, giving nine directions. Random convolution maps follow the ROCKET construction (Dempster et al., 2020): three seeds 8701–8703, each with 1,024 filters of lengths 7/9/11, centered Gaussian weights, uniform biases in $[-1, 1]$, feasible log-uniform dilations and optional symmetric zero padding. Maximum activation and positive-activation proportion produce 2,048 coordinates. Other-source balanced moments standardize coordinates before row normalization. Each seed supplies one linear and three RBF kernels, adding 12 directions. Together these give 67 kernels, plus identity.

The nested expansion adds tails of lengths 8/64/128 (12 kernels), 256-filter prefixes (12), 4,096-filter extensions (12), seven extra random-convolution seeds (28), and three width-8,192 ReLU maps (3). The cumulative banks have 79, 91, 103, 131 and 134 kernels. Existing columns are preserved. Bank size counts proposed relation directions before the SVD rank reduction.

B.3 Group allocation and depth results

For a subset S of the five groups (base, catch22, local, ReLU, convolution), let $v(S)$ be its projected PT–INIT contrast, always including identity. For group g, exact Shapley allocation averages $v(S \cup \{g\}) - v(S)$ over its position in all $5!$ orders (Shapley, 1952). Allocations sum to $v(\mathrm{all}) - v(\varnothing)$; identity is accounted separately. The allocation depends on the supplied, correlated feature groups. We verify the subset-weight formula against all 120 orders. Table 4 reports all four readouts.

Table 4: Full-inventory decomposition with the expanded reference bank. Residual intervals belong to the 32-contrast joint band, including both 67 and 134 kernels.

Readout	Depth	ΔA	ΔA_{span}	ΔA_{res} [joint 95%]
mean / linear	0.250	0.367	0.340	0.027 [-0.006, 0.061]
mean / linear	0.500	0.380	0.338	0.042 [0.003, 0.080]
mean / linear	0.750	0.310	0.266	0.044 [0.004, 0.084]
mean / linear	1.000	0.142	0.131	0.011 [-0.010, 0.032]
mean / rbf	0.250	0.328	0.289	0.039 [0.003, 0.075]
mean / rbf	0.500	0.346	0.291	0.055 [0.013, 0.097]
mean / rbf	0.750	0.300	0.243	0.057 [0.015, 0.100]
mean / rbf	1.000	0.176	0.146	0.030 [0.008, 0.052]
last / linear	0.250	0.268	0.235	0.033 [0.017, 0.048]
last / linear	0.500	0.343	0.298	0.046 [0.022, 0.069]
last / linear	0.750	0.275	0.224	0.052 [0.023, 0.081]
last / linear	1.000	0.171	0.125	0.046 [0.018, 0.075]
last / rbf	0.250	0.246	0.203	0.043 [0.024, 0.062]
last / rbf	0.500	0.312	0.259	0.053 [0.027, 0.080]
last / rbf	0.750	0.263	0.205	0.058 [0.027, 0.089]
last / rbf	1.000	0.191	0.139	0.052 [0.018, 0.086]

The late 103-to-134 expansion was also checked against an exploratory resolution of ± 0.005 CKA units with a 16-contrast joint band. Ten of sixteen intervals lie within this range, including all last-position cases; six mean-pooled intervals extend beyond it. The resolution is an exploratory threshold for changes in CKA units.

B.4 Complementary bias and orientation checks

An endpoint sensitivity repeats normalized alignment with unbiased HSIC (Song et al., 2012). For zero-diagonal G^0, U-centering sets diagonal entries to zero and, for $i \neq j$, uses

$$\widetilde{G}_{ij} = G^0_{ij} - \frac{G^0_{i+} + G^0_{+j}}{n-2} + \frac{G^0_{++}}{(n-1)(n-2)}. \tag{4}$$

Frobenius normalization cancels the common HSIC factor. Identity is annihilated, so this is a different relation space from the main decomposition. Direct-formula checks verify the implementation. The numerical supplement retains results for the earlier reference banks used in this sensitivity.

A separate spectrum-preserving reference rotates every centered relation by the same Haar-orthogonal matrix. It preserves all model-to-model inner products while changing their orientation relative to the input span. For centered symmetric B, $d = n-1$, $a = \langle B, J\rangle_F$, and a rank-$r$ span containing J, the expected projected energy is

$$\mathbb{E}_U \|P(UBU^\mathsf{T})\|_F^2 = a^2 + \frac{r-1}{d(d+1)/2 - 1}(\|B\|_F^2 - a^2). \tag{5}$$

This follows from isotropy of the traceless symmetric component under conjugation. This reference measures expected projected energy at fixed spectra. The paired agreement analysis uses the source-bootstrap intervals described in Appendix A.

B.5 History-count and concatenated-token checks

These follow-ups test the dependence of Section 3's endpoint result on the number of histories and the representation of token positions. They use the same 134-kernel projector, INIT seed 42, fixed models and paired source bootstrap with 10,000 draws. They were run after the main inventory; the intervals below are nominal and are outside the simultaneous bands.

Nested subsets select 32, 64 and 128 histories per dataset with seed 8701, preserving the published other-source bandwidths. The full-count calculation reproduces the published endpoint cells within 1.46×10^{-13}. As history count grows, the last-linear span contrast stays near 0.12 while the raw contrast rises. Its span share therefore decreases, and its residual grows (Table 5). Mean pooling does not show the same share trend; all four mean residual intervals cross zero. These are observed finite-population sensitivities, with each projection refitted on the selected matrices.

Table 5: Endpoint linear history-count sensitivity with 134 reference kernels plus identity. Intervals are nominal paired 95% source intervals. The share is the ratio of the aggregated span contrast to the aggregated raw contrast. "Full" uses each dataset's original history count.

Pool	Histories	ΔA	ΔA_{span}	ΔA_{res} [95%]	Share
Last	32	0.139	0.122	0.017 [0.007, 0.027]	0.879
Last	64	0.147	0.117	0.029 [0.013, 0.047]	0.801
Last	128	0.165	0.125	0.040 [0.021, 0.061]	0.758
Last	Full	0.171	0.125	0.046 [0.025, 0.070]	0.731
Mean	32	0.116	0.108	0.007 [-0.001, 0.016]	0.936
Mean	64	0.120	0.113	0.007 [-0.006, 0.022]	0.939
Mean	128	0.137	0.130	0.007 [-0.007, 0.023]	0.950
Mean	Full	0.142	0.131	0.011 [-0.004, 0.029]	0.921

For the concatenated-token readout, masked endpoint token features are flattened into one vector per history and that complete vector is Euclidean-normalized once. The linear Gram sums matching-position inner products, followed by the same centering and Frobenius normalization as in Section 2.

This preserves differences in token energy rather than normalizing each position separately. Mean and last endpoint cells bridge to the published inventory within 2.25×10^{-11}.

Table 6: Endpoint linear accounting for concatenated and pooled tokens on the full inventory, with the same 134-kernel bank. All intervals in this table are nominal paired 95% source intervals.

Representation	ΔA [95%]	ΔA_{res} [95%]	Share [95%]
All tokens	0.119 [0.080, 0.167]	0.030 [0.012, 0.052]	0.748 [0.466, 0.916]
Last position	0.171 [0.124, 0.216]	0.046 [0.025, 0.070]	0.731 [0.537, 0.863]
Mean pooling	0.142 [0.093, 0.210]	0.011 [-0.004, 0.029]	0.921 [0.766, 1.025]

The concatenated-token residual is 0.030 [0.012, 0.052]. Its point estimate is below the matched last-position value, with overlapping intervals (Table 6). Deleting each family changes only the cross-family pair weights; the input span remains fixed. Across the seven deletions, concatenated-token residual estimates range from 0.024 to 0.038, and every nominal interval lies above zero. The result therefore extends beyond either mean pooling or the final token alone.

C Ordinary forecasting controls

C.1 Population and training

We preserve the previously selected 128 test histories per dataset and remove identical histories/original series from the remaining training pool. Forecasting requires at least 96 observed points, a prefix of at least 64 points with standard deviation at least 10^{-6}, and a 32-point suffix. Eligible datasets need at least 64 remaining training/validation series and 32 test suffixes. M4-yearly has only three eligible remaining series and one test suffix; car parts and hospital have none; wind has 48 remaining series. These four datasets are excluded before training, leaving 13 datasets in seven source groups. All 128 histories in each retained dataset still enter geometry, irrespective of forecasting eligibility.

Original series are split 80/20 into training/validation with seed 8900. The within-dataset arm fits each dataset separately; the held-source arm trains and validates only on the other six eligible sources. Test files are loaded after selection is recorded. Prefix-only mean and population standard deviation normalize the prefix and suffix, and the prefix is left-padded to 256 with a mask. Geometry instead uses the full observed standardized history, as for the FM feature caches.

The MLP has layers 512–128–64 with GELU and a 32-output head (76,000 parameters). The CNN takes value/mask channels, a width-7 stem with 32 channels, and four width-5 residual convolutions with dilations 1/2/4/8. Its concatenated mean/last states feed a 32-output head (23,168 parameters). Convolutions are noncausal within observed context; future targets are excluded. Both models predict a correction to the last value, with zero-initialized output heads, so initial forecasts equal persistence. Their hidden features remain independently initialized.

Each architecture, arm, fold and seed (8901–8903) has two fixed learning-rate candidates, 3×10^{-4} and 10^{-3}, each trained for 768 updates with batch size 64. AdamW uses weight decay 10^{-4}, gradient clipping at one and SmoothL1 loss with $\beta = 1$. Pooled batches balance sources and datasets. There are 13 within-dataset and seven held-source folds, hence 240 candidates and 120 selected runs. Validation occurs every 64 updates and at initialization. Selection minimizes source-balanced, dataset-balanced MSE/persistence MSE with a denominator floor of 10^{-6}; ties prefer earlier steps then smaller rates. No FM features or relation scores enter fitting or selection. All selected checkpoints are after step zero.

C.2 Forecasting and geometry results

The within-dataset aggregate MSE ratios are 0.451 [0.230, 0.695] for the MLP and 0.463 [0.242, 0.705] for the CNN. The corresponding held-source ratios are 0.803 [0.488, 1.189] and 0.804 [0.469, 1.207]. Table 7 retains every source, including the unfavorable held-source M4 results. Six of seven held sources have favorable point estimates; both aggregate intervals straddle one.

Table 7: Held-out MSE / persistence MSE, with equal dataset weight within each source and seed averaging. Columns separate within-dataset and held-source training.

Source	Within MLP	Within CNN	Held MLP	Held CNN
electricity	0.162	0.156	0.235	0.279
kdd	0.236	0.248	0.671	0.687
london	0.620	0.626	0.745	0.774
m4	0.762	0.801	1.808	1.875
traffic	0.078	0.100	0.367	0.243
weather	0.323	0.338	0.799	0.775
wikipedia	0.974	0.973	0.998	0.997
Aggregate	0.451	0.463	0.803	0.804

After freezing selection, the MLP contributes its unit-normalized 64-vector to both pooling labels. The CNN contributes separately normalized mean/last vectors. FM relations are reconstructed on exactly the same histories. All four endpoint readouts and the 22/46/67/134 reference banks are retained numerically; Tables 8 and 9 show the expanded-bank changes and linear levels. Small-model comparisons average paired training seeds, whereas FM comparisons average family pairs.

Table 8: Small-model trained-minus-initialized agreement. Nominal 95% paired source intervals; all four readouts are reported.

Training	Readout	Raw change [95%]	Residual change [95%]
Within	mean/linear	-0.024 [-0.061, 0.020]	0.007 [-0.002, 0.020]
Within	mean/rbf	-0.048 [-0.092, 0.005]	0.004 [-0.002, 0.014]
Within	last/linear	0.064 [-0.034, 0.166]	0.035 [0.014, 0.058]
Within	last/rbf	0.049 [-0.002, 0.103]	0.027 [0.014, 0.042]
Held source	mean/linear	0.019 [-0.041, 0.073]	-0.002 [-0.005, 0.001]
Held source	mean/rbf	-0.018 [-0.066, 0.022]	-0.001 [-0.006, 0.002]
Held source	last/linear	0.196 [0.136, 0.297]	0.064 [0.028, 0.123]
Held source	last/rbf	0.170 [0.123, 0.244]	0.050 [0.022, 0.098]

Table 9: Within-dataset matched-population linear agreement levels. The FM comparison uses the same 13 datasets and 128 histories per dataset; these are not full-inventory values.

Pool	Participants	Raw [95%]	Residual [95%]
mean	MLP–CNN initial	0.305 [0.200, 0.397]	-0.001 [-0.004, 0.002]
mean	MLP–CNN trained	0.281 [0.208, 0.354]	0.006 [-0.002, 0.018]
mean	FM INIT	0.357 [0.311, 0.408]	0.046 [0.034, 0.061]
mean	FM PT	0.499 [0.421, 0.593]	0.059 [0.039, 0.083]
last	MLP–CNN initial	0.412 [0.254, 0.553]	0.003 [0.000, 0.005]
last	MLP–CNN trained	0.476 [0.369, 0.605]	0.037 [0.016, 0.061]
last	FM INIT	0.391 [0.333, 0.447]	0.013 [0.008, 0.021]
last	FM PT	0.582 [0.530, 0.635]	0.055 [0.027, 0.090]

For the primary within-dataset last-linear residual, the small-model change is 0.0346 [0.0138, 0.0581] and the matched FM PT–INIT change is 0.0420 [0.0177, 0.0705]. The trained small models' residual level is lower than the FM PT level: their difference is -0.0177 $[-0.0377, -0.0039]$. The mean-linear change is uncertain. Training lowers the MLP's mean-linear correspondence to PT while raising the CNN's point estimate.

D HELD-SOURCE RIDGE RECOVERY AND REFERENCE AUGMENTATION

D.1 FITTING AND SELECTION

The ten outer folds use the canonical source grouping. In each fold, the test source is excluded from every training statistic; the next source in sorted cyclic order supplies validation, and the other eight supply fitting. We select 64 histories per training dataset using seed 8907 from the fixed 128-history population and evaluate all 128 histories per validation/test dataset. All folds use the previously studied canonical source inventory.

The input concatenates four row-normalized blocks: raw 256-point history, last 32 values, catch22, and the existing 1,024-filter random convolution map with seed 8701. Catch22 uses training-only median/IQR imputation and scaling; convolution coordinates use training-only source/dataset-balanced moments. Dividing the concatenated vector by two weights the blocks equally. These input maps are teacher-independent. RBF width is the positive training squared-distance median times s^2, $s \in \{1/2, 1, 2\}$.

For each PT/INIT teacher and mean/last endpoint, we fit weighted kernel ridge regression to the unit teacher feature vectors. For feature target y_i, predictor f and unpenalized intercept b, the objective is

$$\sum_i w_i \|f(x_i) + b - y_i\|_2^2 + \lambda \|f\|_{\mathcal{H}}^2, \qquad \lambda \in \{10^{-5}, 10^{-3}, 10^{-1}\}. \tag{6}$$

The weights sum to one and balance sources, datasets within source and histories within dataset. Weighted centering of the kernel and targets implements the intercept. Linear and RBF methods use exactly the same features and training histories. An independent small NumPy primal solve checks the weighted-ridge predictions.

For each state/pool, validation selects a single method-specific setting across teachers, averaging CKA by family and validation dataset. All ten teachers are supervised separately in every fold. Test predicted rows are normalized before forming linear relations. The direct-input comparator uses an input kernel selected by the same validation criterion, with kernel and width as its selectable parameters. The primary estimands compare nonlinear recovery with linear recovery and this direct baseline, separately for each state/pool.

Table 10: Per-teacher source-held-out feature recovery, evaluated as relation CKA. All teachers are supervised. The joint band includes these four nonlinear-minus-linear and four nonlinear-minus-direct comparisons. The PT mean lower bound is 0.000173, rounded below.

State	Pool	Direct	Linear	RBF	RBF − linear [joint 95%]
PT	mean	0.467	0.584	0.609	0.024 [0.000, 0.048]
PT	last	0.582	0.676	0.701	0.025 [0.005, 0.044]
INIT	mean	0.434	0.613	0.638	0.025 [0.005, 0.045]
INIT	last	0.553	0.663	0.708	0.045 [0.026, 0.065]

D.2 MATCHED REFERENCE ACCOUNTING

A secondary diagnostic augments the existing 134-kernel span with predicted relations from all ten PT and ten INIT teachers. Each method contributes 20 kernels before possible rank reduction. The union includes both linear and nonlinear predictions. Every comparison uses the same 128 histories per dataset and the same combined span for the two true model states. Input maps and predictions exclude the test source, but projection coefficients are fitted to the evaluated true relations.

For last pooling, the changes from the original bank are -0.00793 $[-0.01251, -0.00316]$ with linear predictions, -0.00859 $[-0.01374, -0.00385]$ with RBF predictions, and -0.01068 $[-0.01682, -0.00494]$ with both. Linear and RBF augmentations yield similar point reductions. Mean pooling's combined change is $+0.00750$ $[-0.00060, 0.01626]$, a positive point estimate with an interval crossing zero. Its PT and INIT residual levels both decline, with a larger INIT decline.

All source-level scores, validation choices, states and methods are retained regardless of effect direction. These scores summarize recovery of individual teacher relations on held-out sources.

Table 11: Shared reference augmentation on identical 128-history populations. Intervals are nominal, and coefficients are fitted descriptively on evaluated relations.

Pool	Bank	PT	INIT	ΔA_{res} [95%]	Change
mean	134	0.058	0.039	0.019 [0.005, 0.036]	0.000
mean	134+linear	0.046	0.020	0.026 [0.012, 0.041]	0.008
mean	134+rbf	0.046	0.020	0.026 [0.011, 0.041]	0.007
mean	134+both	0.043	0.017	0.026 [0.012, 0.041]	0.007
last	134	0.059	0.010	0.048 [0.026, 0.073]	0.000
last	134+linear	0.046	0.006	0.040 [0.022, 0.061]	-0.008
last	134+rbf	0.045	0.005	0.040 [0.021, 0.061]	-0.009
last	134+both	0.042	0.005	0.038 [0.020, 0.058]	-0.011

E External populations and family composition

The two existing external caches contain 128 Wikipedia page-view histories and 128 NASA POWER hourly temperature histories. The first observations are dated 2025-11-05 and 2026-07-20, respectively. Public release bounds precede the selected observations, with matching local parameter hashes for eight variants: the three Chronos variants, two TimesFM variants, Moirai, Sundial and Time-MoE. The audit verifies 16 model/population weight combinations. Mamba4Cast lacks the corresponding historical checkpoint evidence and is omitted from this timing subset; it remains in a nine-variant available-cache sensitivity. TiRex has no matching external cache.

The timing audit verifies cached weight identities against acquisition dates; pretraining exposure to related patterns or locations remains unknown. NASA locations are spatially related, and placeholder dates in some canonical datasets provide no window-membership information. We summarize the two external populations using descriptive family-deletion ranges.

We reconstruct the 22/46/67/134 reference banks at all four depths and readouts. NASA scaling uses the canonical sources; external Wikipedia excludes the canonical Wikipedia group. The original catch22 bandwidth convention, including zero-distance behavior, is preserved after a documented first-pass correction. Independent Gram projection checks reconstruct the final accounting. Table 12 reports endpoint results for the timing subset, with descriptive ranges obtained by omitting each of its five families.

Table 12: Later-population endpoint results, timing-verified subset. Deletion ranges give the minimum and maximum across omitted families.

Population	Readout	ΔA	ΔA_{span}	ΔA_{res}	Deletion range
Wikipedia	mean/linear	0.094	0.028	0.066	[0.033, 0.099]
Wikipedia	mean/rbf	0.112	0.020	0.092	[0.057, 0.128]
Wikipedia	last/linear	0.139	0.086	0.053	[0.047, 0.063]
Wikipedia	last/rbf	0.209	0.148	0.060	[0.047, 0.067]
NASA	mean/linear	0.467	0.455	0.011	[-0.007, 0.028]
NASA	mean/rbf	0.458	0.448	0.010	[-0.001, 0.026]
NASA	last/linear	0.398	0.323	0.075	[0.042, 0.097]
NASA	last/rbf	0.407	0.329	0.078	[0.052, 0.097]

On the full canonical inventory, every single-family deletion is evaluated over four depths and four readouts with 134 kernels. The joint band covers all 112 deletion cells. All seven endpoint last-linear intervals are positive; the complete depth/readout results are supplied numerically. Separating Chronos-2 from the Chronos family and omitting Moirai are additional historical grouping sensitivities preserved in the numerical results. These tests evaluate alternative groupings of the observed releases.

F CONTROLLED CONTENT AND ADAPTATION RECORDS

The studies in this appendix use their own populations and estimands. Their numbers are separate from the real-history input accounting, ordinary-predictor and ridge comparisons. Section 6 uses them to give temporal and training context to the main results.

F.1 CONTROLLED TEMPORAL CONTENT

Figure 5 uses the six-family core comprising Chronos, TimesFM, Moirai, Sundial, Time-MoE and Mamba4Cast. Unit-amplitude sinusoids cross cycle counts $\{2, 4, 8, 16\}$ and phases $\{0, \pi/2, \pi, 3\pi/2\}$ over 256 observations. Eight realizations per cell give 128 histories per panel. Two fixed panels use Gaussian noise standard deviation 0.25; other recorded panels have noise zero or one. The main figure averages the two noise-0.25 panels and family pairs equally.

For this balanced design, let $C = \mathbf{1}\mathbf{1}^{\mathsf{T}}/n$. Group-averaging matrices M_f, M_p, M_{fp} have entries $1/32$ for equal frequency, $1/32$ for equal phase and $1/8$ for equal frequency–phase cell, respectively, and zero otherwise. The projectors

$$Q_f = M_f - C, \quad Q_p = M_p - C, \quad Q_{fp} = M_{fp} - M_f - M_p + C, \quad Q_r = I - M_{fp} \tag{7}$$

are mutually orthogonal and sum to H. They act on history identities, following a marginalization logic related to demixed PCA (Kobak et al., 2016). For centered unit relations K, T,

$$\langle K, T \rangle_F = \sum_{a,b \in \{f,p,fp,r\}} \langle Q_a K Q_b, Q_a T Q_b \rangle_F. \tag{8}$$

The three diagonal frequency, phase and interaction terms are plotted separately; "other" sums the remaining 13 terms. This is an exact decomposition of raw linear CKA on the controlled panel. For PT mean pooling, frequency contributes 0.692 to total 0.750. PT last pooling distributes contributions across frequency (0.164), phase (0.198) and interaction (0.248), with total 0.614. INIT mean pooling has frequency contribution 0.035 and total 0.286. The numerical figure record preserves the full-precision components. TiRex extension panels further demonstrate readout dependence, including substantial last-position phase content.

F.2 COMMON-TARGET ADAPTATION

Figure 6 uses PT TimesFM-2.5-200M and TiRex-35M with three paired optimization seeds. Training sequences have 256 historical values and 32 future values, random amplitude and phase, a second harmonic and noise. Training has 4,096 sequences with 2–8 cycles per history. Evaluation uses new histories from the training generator, a shift to 12–24 cycles, and chirps; the main figure shows frequency shift.

The forecast-only arm and ACF arm share initialization and sample order. Rank-16 adapters update attention projections in TimesFM and feed-forward projections in TiRex, with the remaining weights fixed. The auxiliary objective aligns both mean- and last-pooled endpoint linear relations to an RBF target on ACF lags 1–64. A development-set distance median fixes target width. The auxiliary coefficient initially sets its gradient norm to one quarter of the forecast gradient norm, then remains fixed. Training uses AdamW, learning rate 10^{-4}, batch size 32 and 256 updates. Forecast objectives retain each model's native quantile loss and TimesFM's mean-head term; reported MSE evaluates the median prediction.

Saved checkpoints at 0, 64, 128, 192 and 256 updates were evaluated after the endpoint results were known, without retraining or checkpoint selection. Each contrast subtracts its same-step, same-seed forecast-only reference. Panel means are formed within seed, then averaged across seeds. Relative MSE averages the seed-level ratios of panel means. Shaded ranges in Figure 6 are seed ranges. Table 13 separately retains the available paired intervals; the 128-update forecast point estimates lack corresponding intervals.

At 256 updates under frequency shift, mean-linear target correspondence rises by 0.024 for TimesFM and 0.199 for TiRex relative to forecast-only adaptation. Their mutual raw linear CKA falls by 0.101. Decomposing along the ACF direction yields a target-component increase of 0.0835 and a remaining-component decrease of 0.1841. This geometric accounting locates the change in agreement.

Table 13: Synthetic frequency shift: PT ACF minus same-step forecast-only adaptation, at the initial gradient ratio $1/4$. Negative forecast changes are improvements. Brackets are paired 95% intervals; the 128-update MSE values are means only.

Updates	Mean mutual RBF excess	TimesFM MSE change	TiRex MSE change
128	$+0.001$ [-0.012, 0.010]	-5.6%	-0.4%
256	-0.056 [-0.107, -0.023]	$+0.9\%$ [-0.3, 2.2]	$+33.4\%$ [26.7, 40.9]

Other strengths, adapter sites and checkpoints have different outcomes; the model-specific forecast changes are reported separately in Table 13.

The real-data extension trains PT TimesFM, TiRex and Chronos-Bolt on electricity, traffic, London smart meters and hourly M4, then evaluates KDD, Wikipedia, wind and weather. Each source supplies 256 training or 128 evaluation examples, respectively, with disjoint historical/future blocks and history-only standardization. Three seeds pair forecast-only, ACF and shuffled-ACF arms; all runs finish before evaluation. All four primary forecast intervals use the prespecified 98.75% level and cross zero (Table 14). The model-pair comparisons share TimesFM. Source-stratified target shuffling preserves some source-block structure. Removing flat-segment copy matches leaves the primary interval conclusions unchanged.

Table 14: Real-source intervention: the four primary forecast contrasts at 256 updates, with 98.75% intervals. Negative MSE changes favor ACF training.

Model	Reference arm	Relative MSE change [98.75%]
TimesFM	Forecast-only	-1.027% [-2.886, $+0.028$]
TiRex	Forecast-only	-0.506% [-2.882, $+0.584$]
Chronos-Bolt	Forecast-only	-0.316% [-1.071, $+0.004$]
Chronos-Bolt	Shuffled ACF	$+0.102\%$ [-0.170, $+1.011$]

F.3 Other exploratory records

Consensus recovery and family transfer. A temporal encoder learned family-balanced relations with one teacher family excluded per fold, using four training sources, four evaluation sources and three learner seeds. An exploratory mean-pooled advantage did not establish reliable gains over a supervised 22-kernel mixture for excluded-family transfer, initialized teachers, alternative readouts or neighborhoods. Frozen-fit new-source evaluation and crossed refits retain the distinction between consensus reproduction and family transfer; this differs from the all-teachers-supervised ridge protocol in Section 5.

Synthetic imitation and capability associations. A separate observable-frequency benchmark found some gains from more objects or updates, without establishing longer-run superiority to the strong mixture. Its future segment exactly copied a historical segment, so its forecast comparisons were excluded. This defect belongs to that imitation benchmark, separate from the audited ACF study above. Observational capability–alignment correlations were also sensitive to readout and family deletion.

G PREDICTING A HELD-OUT FAMILY WITH ONE INPUT COMBINATION

G.1 QUESTION AND PROTOCOL

The projection in Section 2 allows different coefficients for every model. Here we ask the stronger question of whether a single coefficient vector fitted to other families predicts a held-out family's relations. This is a follow-up on the existing 17-dataset inventory. Its protocol was written before coefficient fitting, after the main reference-bank analyses. It uses endpoint mean-linear and last-linear relations in PT and INIT, with 22, 46, 67 and 134 input kernels plus centered identity.

Let B_d contain the frozen, normalized input-reference matrices for dataset d, expressed in the compact orthonormal coordinates released with the main analysis, and let t_{dm} denote the coordinates of a unit-Frobenius model relation. We fit a minimum-norm least-squares coefficient vector to the mean target from the six training families, weighting families equally and variants equally within family. The predicted relation is $B_d\alpha/\|B_d\alpha\|$, and its inner product with t_{dm} is the reported CKA. SVD retains singular values above $\max(10^{-10}s_{\max}, 10^{-12})$. This unconstrained linear combination need not be positive semidefinite; a nonnegative least-squares comparison is also recorded.

For the primary family holdout, fitting and evaluation use the same histories within each dataset; only the teacher family is held out. Each of the seven families is tested in turn. Uniform weights and the single column with highest average training-family CKA use the identical reference bank. We average test variants within family, families, datasets within source and the ten sources equally. Nominal 95% intervals use 10,000 paired source bootstrap draws with seed 8906, conditional on the fitted combinations and model inventory. They are not simultaneous bands across the follow-up grid.

Table 15: Held-family relation prediction on the same histories. Minimum-norm combinations and paired gains over a training-selected single kernel and uniform weights; nominal 95% source intervals. Banks have the indicated kernel count plus identity. Full four-bank records are in the supplement.

State / pool	Bank	Learned CKA	Δ single	Δ uniform
PT mean	46	$0.593\ [0.510, 0.679]$	$0.038\ [0.021, 0.054]$	$0.083\ [0.043, 0.122]$
PT mean	134	$0.608\ [0.529, 0.690]$	$0.041\ [0.020, 0.065]$	$0.046\ [0.020, 0.072]$
PT last	46	$0.659\ [0.590, 0.723]$	$0.050\ [0.032, 0.067]$	$0.046\ [0.027, 0.068]$
PT last	134	$0.667\ [0.605, 0.727]$	$0.056\ [0.036, 0.075]$	$0.057\ [0.037, 0.078]$
INIT mean	46	$0.482\ [0.414, 0.550]$	$0.040\ [0.016, 0.066]$	$0.057\ [0.038, 0.078]$
INIT mean	134	$0.486\ [0.420, 0.550]$	$0.045\ [0.020, 0.071]$	$0.061\ [0.041, 0.083]$
INIT last	46	$0.559\ [0.500, 0.615]$	$0.055\ [0.020, 0.093]$	$0.003\ [-0.006, 0.010]$
INIT last	134	$0.558\ [0.500, 0.613]$	$0.054\ [0.020, 0.092]$	$0.037\ [0.015, 0.058]$

The expanded PT last-linear combination reaches 0.667 CKA on held-out families. Its paired gains over the training-selected single kernel and uniform mixture are 0.056 [0.036, 0.075] and 0.057 [0.037, 0.078]. The 46-kernel result and both initialized readouts show why the comparison is conditional on the bank and model state. Nonnegative fitting reaches 0.661 for the expanded PT last-linear setting; the positive fit is therefore not confined to signed combinations. These tests measure prediction of another family's relations on common histories, without evaluating forecasting utility.

G.2 SOURCE EXCLUSION WITH FROZEN REFERENCES

The saved protocol also excludes a source from coefficient fitting, either retaining the target family or excluding it as well. Dataset weights are equal within source and sources are equally weighted. However, the reference bank was frozen before these folds: a training dataset's scaling and bandwidth parameters omit its own source, and can include input statistics from the later held-out source. Held-out model targets are excluded from fitting, but held-out input information is not fully excluded from preprocessing. These source comparisons therefore condition on the frozen reference bank and do not establish strictly inductive transfer to an unseen source. The family-only experiment above intentionally shares histories and does not require that stronger interpretation.

Table 16: Joint family/source target holdout, conditional on the frozen reference bank. Columns follow Table 15. Held-out source inputs can enter preprocessing of training matrices; this is not a strictly inductive source-transfer comparison.

State / pool	Bank	Learned CKA	Δ single	Δ uniform
PT mean	46	$0.563\ [0.488, 0.644]$	$0.007\ [-0.020, 0.028]$	$0.054\ [0.014, 0.092]$
PT mean	134	$0.579\ [0.503, 0.658]$	$0.039\ [0.023, 0.057]$	$0.016\ [0.002, 0.032]$
PT last	46	$0.626\ [0.554, 0.693]$	$0.113\ [0.082, 0.145]$	$0.013\ [-0.004, 0.028]$
PT last	134	$0.635\ [0.565, 0.701]$	$0.122\ [0.090, 0.156]$	$0.025\ [0.007, 0.044]$
INIT mean	46	$0.470\ [0.404, 0.534]$	$0.080\ [0.060, 0.101]$	$0.045\ [0.020, 0.067]$
INIT mean	134	$0.469\ [0.401, 0.534]$	$0.079\ [0.054, 0.104]$	$0.044\ [0.024, 0.064]$
INIT last	46	$0.553\ [0.495, 0.604]$	$0.030\ [0.009, 0.050]$	$-0.003\ [-0.022, 0.011]$
INIT last	134	$0.553\ [0.493, 0.607]$	$0.030\ [0.010, 0.049]$	$0.033\ [0.011, 0.054]$

The expanded PT last-linear combination reaches 0.635 in the joint family/source target holdout. The compact release retains every fold, all four bank sizes, the source-only results, inner-source comparisons, ridge fits with nine penalty candidates, and nonnegative fits. Inner-versus-outer intervals crossing zero do not establish equivalence or absence of overfitting. A strictly inductive source test would rebuild the reference preprocessing within each outer training fold.

G.3 What the projection ceiling does and does not measure

The per-model projection ceiling maximizes correspondence to that model inside the supplied linear span, using its own coefficients. For expanded PT last-linear relations it is 0.793, above the held-family fit of 0.667. The difference mixes the common-coefficient constraint with fitting and transfer error. It lies within the supplied reference space and is distinct from the pairwise residual CKA in Equation 1. It neither measures a private component of each model nor bounds every nonlinear recovery method. Components left over after removing one common direction may still agree across models.

The mean pairwise cosine between projected family-mean directions is 0.774 for expanded PT last-linear relations and 0.539 for INIT. This is a normalized directional statistic, whereas the span contribution in Equation 1 retains both magnitude and direction. Thus alignment within a shared span and successful use of a common coefficient vector are related but different observations.

H RESIDUAL DIRECTION AND WEIGHT ARRANGEMENT

H.1 SYMMETRIC SPLIT

Let $R_i = (I - \mathcal{P})(K_i)$ be the residual of model relation i, and set $g_{ij} = \|R_i\|_F \|R_j\|_F$. For $g_{ij} > 0$ the residual cosine is $c_{ij} = \langle R_i, R_j \rangle_F / g_{ij}$, so the residual inner product factors as $A_{\mathrm{res},ij} = c_{ij} g_{ij}$. Across PT and INIT, write $\bar{c} = (c^{\mathrm{PT}} + c^{\mathrm{INIT}})/2$ and $\bar{g} = (g^{\mathrm{PT}} + g^{\mathrm{INIT}})/2$. Entrywise,

$$\Delta(cg) = \bar{c}\,\Delta g + \bar{g}\,\Delta c. \tag{9}$$

The same cross-family pair weights used for A_{res} are applied to each term. The first term is the energy component and the second is the direction component. They sum to ΔA_{res}. Pair weights are zero within a family, so diagonal entries do not enter the aggregate. The mean residual norm in Table 17 averages $\|R_i\|_F$ over the ten variants. The energy component pair-weights products of those norms.

The split replays the published endpoint residual, residual cosine and mean squared residual norm from the stored relation coordinates. The largest absolute discrepancy is 5.9×10^{-14}. Intervals are paired source percentiles, 10,000 draws, seed 8606. The last-linear endpoint residual, 0.046 [0.025, 0.070], matches Table 5. Table 4 is the joint band for the original residual contrasts.

Table 17: Symmetric split of endpoint ΔA_{res} on the full inventory. Energy pair-weights the change in products of residual Frobenius norms; direction pair-weights the change in residual cosines. Paired source percentiles use seed 8606 and 10,000 draws.

Readout	ΔA_{res}	Energy	Direction
last / linear	0.046 [0.025, 0.070]	0.004 [−0.002, 0.011]	0.042 [0.025, 0.061]
last / rbf	0.052 [0.028, 0.081]	0.005 [−0.002, 0.012]	0.047 [0.028, 0.071]
mean / linear	0.011 [−0.004, 0.029]	−0.005 [−0.015, 0.005]	0.016 [0.003, 0.031]
mean / rbf	0.030 [0.013, 0.048]	−0.003 [−0.013, 0.007]	0.032 [0.018, 0.049]

Last-position readouts assign ΔA_{res} to the direction term. Their energy intervals cross zero, as do the differences in mean residual norm. Mean-linear ΔA_{res} also has an interval crossing zero, while its direction term stays positive. Mean-RBF has a positive residual contrast and a positive direction term; its energy interval crosses zero. At last-linear depths $1/4$ and $1/2$, the mean-norm difference is negative, so INIT leaves more residual norm, while the residual contrast stays positive through the direction term (Table 18).

Table 18: Last-linear residual split by relative depth. The endpoint row matches Table 17. Δ mean norm is the PT minus INIT difference in the unweighted average of $\|R_i\|_F$.

Depth	ΔA_{res}	Direction	Δ mean norm
0.250	0.033 [0.021, 0.046]	0.032 [0.022, 0.043]	−0.053 [−0.092, −0.017]
0.500	0.046 [0.028, 0.065]	0.047 [0.031, 0.065]	−0.070 [−0.107, −0.036]
0.750	0.052 [0.030, 0.076]	0.049 [0.031, 0.069]	−0.034 [−0.072, 0.001]
1.000	0.046 [0.025, 0.070]	0.042 [0.025, 0.061]	0.009 [−0.036, 0.055]

H.2 WITHIN-TENSOR PERMUTATION

Each released parameter tensor is replaced by a permutation of its own entries. One CPU stream with seed 91026 visits parameters in named order. Buffers stay fixed, so every tensor keeps its values, mean and Frobenius norm. The unshuffled endpoint is compared with the published features on a prefix of the M4 hourly population before any permutation. The largest absolute discrepancy is 3.5×10^{-5}. Mamba4Cast receives the real calendar tokens used for its published features. The shuffled relations use the same 134-kernel projector. The permutation reorders the released values. The initialization draws a different multiset from the public recipe.

Tables 19 and 20 report all four endpoint readouts. Shuffled residual agreement lies below both PT and INIT in every row. For last-linear relations the shuffle–PT energy component is -0.002 [-0.007, 0.002], so the gap from PT is carried by direction. Last-RBF also loses a small amount of residual norm: its shuffle–PT energy component is -0.006 [-0.012, -0.001]. The two mean-pooled energy components have intervals crossing zero. Pair-weighted residual cosines move with the inner products. At the last-linear endpoint they are 0.202 [0.176, 0.230] for PT, 0.069 [0.052, 0.086] for INIT and 0.025 [0.019, 0.032] after shuffling. One seed supplies the permutation.

Table 19: Endpoint residual levels after one within-tensor permutation, seed 91026. Intervals are nominal.

Readout	PT	INIT	Shuffled
last / linear	0.061 [0.037, 0.089]	0.015 [0.009, 0.022]	0.006 [0.004, 0.009]
last / rbf	0.067 [0.039, 0.102]	0.015 [0.010, 0.022]	0.007 [0.004, 0.010]
mean / linear	0.059 [0.039, 0.081]	0.048 [0.036, 0.062]	0.032 [0.022, 0.044]
mean / rbf	0.076 [0.048, 0.105]	0.046 [0.034, 0.060]	0.031 [0.022, 0.042]

Table 20: Shuffled endpoint minus PT. Direction is the direction component of that residual gap. The shuffled state also lies below INIT: the last-linear residual difference is -0.009 [-0.015, -0.004], and the other three readouts are -0.008, -0.015 and -0.014, each with a nominal interval below zero.

Readout	Shuffled$-$PT	Direction
last / linear	-0.055 [-0.081, -0.033]	-0.053 [-0.076, -0.034]
last / rbf	-0.060 [-0.093, -0.034]	-0.054 [-0.082, -0.032]
mean / linear	-0.027 [-0.043, -0.012]	-0.030 [-0.047, -0.016]
mean / rbf	-0.044 [-0.064, -0.026]	-0.044 [-0.063, -0.027]

I FORECAST KERNELS AND THE LAST-LINEAR RESIDUAL

The residual in Equation (1) is what remains after the 134 input kernels and centered identity. This appendix adds kernels built from the forecast horizon and repeats the same paired comparison. The real-future kernel uses the next dataset horizon after the history the models saw, expressed in the context z-score. Seasonal naive, additive Holt–Winters and an autoregression of order at most five are functions of that same trailing context. Holt–Winters uses an additive seasonal form when at least two seasons fit in the context, and Holt's linear trend otherwise. Smoothing constants are chosen from $\{0.2, 0.4, 0.6, 0.8\}$ by in-sample one-step error on the context. Shape kernels row-normalize the forecast path before the linear Gram. Level kernels keep the z-score magnitude. The joint kernel combines the real future with the three classical forecasts. Both states use the enlarged span.

The published residual is recovered within 3.3×10^{-12} before any column is added. Table 21 gives the last-linear changes. The largest is -0.008 for the joint shape kernel, and the residual remains 0.038 [0.019, 0.059]. Single kernels move it by 0.002 to 0.004. Level kernels, which retain forecast magnitude, move it by less.

Table 21: Change in last-linear ΔA_{res} after adding forecast kernels to the 134-kernel span. The baseline value is 0.046 [0.025, 0.070]. Intervals are paired source percentiles, seed 8606, 10,000 draws.

Added kernel	ΔA_{res}	Change
shape, real future	0.042 [0.022, 0.064]	-0.004 [-0.007, -0.002]
shape, seasonal naive	0.043 [0.023, 0.066]	-0.003 [-0.005, -0.001]
shape, Holt–Winters	0.045 [0.023, 0.069]	-0.002 [-0.003, -0.000]
shape, autoregression	0.044 [0.023, 0.067]	-0.003 [-0.004, -0.001]
shape, joint	0.038 [0.019, 0.059]	-0.008 [-0.013, -0.004]
level, joint	0.044 [0.022, 0.068]	-0.002 [-0.004, -0.001]