EQUIVARIANT NEURAL FORECASTING: IRREDUCIBLE REPRESENTATIONS FOR MULTIVARIATE TIME SERIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Multivariate forecasters increasingly exploit cross-variable dependencies to improve prediction, yet they typically treat these relationships as ephemeral computational pathways that dissolve after inference, leaving a divide between predictive accuracy and physical understanding. We introduce the Equivariant Neural Forecaster (ENF), an architecture that unifies coordinate consistency with explicit, disentangled relational modeling. Operating within an invertible normalization frame that guarantees joint equivariance across channel permutations and affine unit changes, ENF orthogonally decomposes channel states into global consensus and local deviation subspaces, while deploying algebraic double-centering to isolate bilateral interactions from unilateral baseline drifts. These disentangled interactions supply target-dependent messages to cross-channel attention and retain an explicit relational substrate for post-hoc analysis. By integrating equivariant sensor identity conditioning with relational subspace routing, the ENF framework achieves an overall 17.5% MSE reduction over iTransformer on four traffic benchmarks. Controlled tasks further confirm that the retained pair field renders multi-hop relational composition directly accessible to linear readouts, without requiring higher-order decoders. Beyond forecasting, diagnostic probes indicate that the learned field captures partially transferable coupling signals in chaotic circuits and guides symbolic distillation into analytical expressions that transfer to wind-tunnel measurements.

1 INTRODUCTION

Complex multivariate time series—from atmospheric flows and power grids to coupled dynamical circuits—are driven by structured interactions among variables. Modern deep forecasters use these relationships as computational pathways to reduce predictive loss: models either enforce channel independence (Nie et al., 2023; Zeng et al., 2023), or exchange information across channels through attention (Liu et al., 2024; Han et al., 2024). The relationships themselves are not returned: models output forecasts, while the interactions that produced them remain internal. Exposing these interactions as first-class objects prompts a foundational question: *can a forecaster simultaneously deliver accurate predictions and structured representations for physical discovery?*

Realizing this goal requires addressing two intrinsic characteristics of multivariate time series. First, neural representations must maintain coordinate consistency: a data-driven forecaster is tied to the sensor ordering and measurement units chosen by the experimenter. If channel histories, heterogeneous sensor identities (Shao et al., 2022), and inter-variable relations do not transform in lockstep under channel permutations and affine unit changes—a property of *joint equivariance* (Cohen & Welling, 2016; Maron et al., 2019)—the extracted relationships depend on the arbitrary measurement frame rather than the physical system. Second, relational representations must separate additive endpoint effects from the interaction: raw pairwise correlations and attention weights are dominated by individual channel dynamics (Jain & Wallace, 2019), such as baseline drift or local amplitude swings. Unless the bilateral interaction is algebraically isolated from these single-variable components (Hartford et al., 2018), the retained relationships conflate unilateral endpoint effects with the physical coupling they are meant to describe.

These characteristics directly motivate two design requirements: ensuring coordinate consistency across arbitrary measurement frames and isolating bilateral coupling from unilateral endpoint effects. Synthesizing these principles, we introduce ENF (Figure 1), which organizes representations across two explicit levels. At the channel level, ENF obeys a joint transformation law carrying dynamic

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

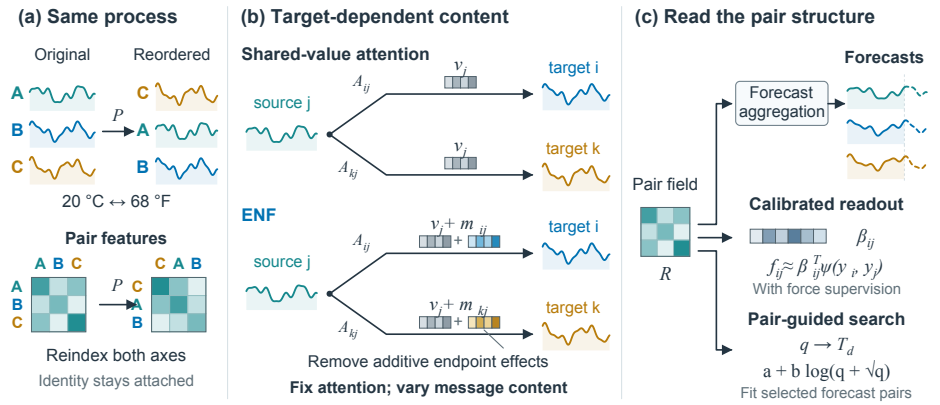


Figure 1: **Overview of ENF.** (a) Equivariance under sensor permutations and affine unit changes. (b) Double-centering isolates bilateral interactions m_{ij} to augment attention values. (c) Preserved pair fields enable forecasting, continuous force calibration, and symbolic formula discovery.

observations and sensor identities (Shao et al., 2022) together, decomposing states into a shared consensus and zero-mean deviations to preserve coordinate symmetry under channel permutations and affine unit changes. At the pair level, ENF constructs an explicit pair field and applies algebraic double-centering (Hartford et al., 2018): single-variable components are routed back to update channel states, while the orthogonal residual $R^{\perp}$ isolates the bilateral interaction. This residual augments attention values with target-dependent messages and directly provides an explicit object for relational composition and physical analysis. Our main contributions are:

- **Theoretical Equivariant and Subspace Framework:** We establish a principled foundation proving joint permutation and affine unit equivariance across heterogeneous channels in the invertible normalization frame (Theorem 1). Furthermore, we prove an orthogonal decomposition of pair spaces, showing double-centering isolates bilateral interactions from additive single-variable baselines, with the default bilinear field yielding symmetric interactions of bounded rank.
- **Two-Level Architecture and Exact Factorization:** We develop a hierarchical forecaster that decouples global consensus from local deviations, augmenting attention with target-dependent interaction messages. We prove this pairwise aggregation admits an *exact computational factorization* (Proposition 1), eliminating dense $C^2 \times d_r$ pair-tensor storage and substantially accelerating training without sacrificing the explicit relational substrate needed for post-hoc analysis.
- **Predictive Benchmarks, Compositional Accessibility, and Physical Transfer:** On matched multi-horizon traffic benchmarks, ENF achieves an overall 17.5% mean squared error reduction over iTransformer across 48 comparisons, supported by the synergy of equivariant identity conditioning and subspace routing. Controlled experiments confirm that the retained pair field makes multi-hop relational composition through shared variables linearly accessible, where first-order node states remain at chance. Beyond forecasting, the learned field enables calibrated continuous force estimation in chaotic circuits and guides symbolic search to discover analytical laws that transfer directly to wind-tunnel instrumentation.

2 Related Work

Implicit temporal and cross-channel forecasting. Early multivariate forecasters often favored channel-independent designs, modeling temporal dynamics via self-attention (Nie et al., 2023) or linear and frequency projections (Zeng et al., 2023; Yi et al., 2024). To capture inter-variable dependencies, recent architectures introduce cross-channel attention (Liu et al., 2024) and multi-scale feature mixing (Chen et al., 2023), while instance normalization addresses temporal distribution shifts (Kim et al., 2022; Liu et al., 2022). In all these approaches, however, cross-variable relationships serve merely as transient, implicit computational weights to update channel representations (Han et al., 2024). Once forward prediction concludes, these relational pathways dissolve without leaving any readable or reusable interaction structure for post-hoc analysis.

Explicit structures, spatial identities, and equivariance. To incorporate structured priors, spatial-temporal graph networks model traffic dynamics using predefined or learned topologies (Li et al.,

2018; Wu et al., 2019), while spatial identity models capture persistent sensor distinctions through learned embeddings (Shao et al., 2022). In parallel, equivariant networks formalize coordinate symmetries across permutation groups on sets (Zaheer et al., 2017) and graph tensors (Hartford et al., 2018; Maron et al., 2019). Recently, CPiRi (Xu et al., 2026) incorporated permutation invariance to improve robustness against channel reordering and transfer on anonymous channels. Nonetheless, existing spatial models either rely on static graphs or treat identities as unconstrained parameters detached from transformation laws, leaving a gap in guaranteeing joint coordinate consistency when sensor arrangements and measurement units undergo permutation and affine changes.

Relational inference, pairwise attention, and physical distillation. Relational inference methods, e.g., NRI (Kipf et al., 2018), infer latent edge types via variational relaxations. Within attention architectures, relation-aware mechanisms (Shaw et al., 2018) and edge-augmented Transformers (Hussain et al., 2022) inject pairwise biases into attention values, but typically materialize dense $O(C^2d)$ pair tensors and conflate unilateral baseline drifts with bilateral interactions (Jain & Wallace, 2019). Concurrently, Dynamic Relational Priming (Lee & Clark, 2025) addresses the shared-value limitation of attention by dynamically modulating representations across token pairs. While symbolic distillation can extract physical equations from networks, it predominantly presupposes known interaction topologies (Cranmer et al., 2020). Extracting disentangled, coordinate-consistent pair representations directly from unmapped time series without dense memory overhead remains challenging.

3 METHODS

3.1 PROBLEM FORMULATION AND SYMMETRY FRAMEWORK

Problem formulation and dual objectives. Let $X = [x_1, \dots, x_T]^\top \in \mathbb{R}^{T\times C}$ denote a multivariate time series containing T historical observations across C continuous channels, and $Y = [x_{T+1}, \dots, x_{T+H}]^\top \in \mathbb{R}^{H\times C}$ the future H steps to predict, estimated by $\widehat{Y}$. Each channel corresponds to a spatial sensor characterized by its dynamic trajectory $X_{:,c} \in \mathbb{R}^T$ and a persistent identity vector $E_c \in \mathbb{R}^{d_e}$, forming an identity table $E \in \mathbb{R}^{C\times d_e}$ (Shao et al., 2022). In addition to channel-level representations, we define a second-order *pair field* $R \in \mathbb{R}^{C\times C\times d_r}$, where each vector $R_{ij} \in \mathbb{R}^{d_r}$ represents symmetric bilateral interaction field between j and i.

Rather than treating inter-variable dependencies merely as ephemeral computational pathways to predict $\widehat{Y}$, our framework pursues a dual objective: (1) *Predictive forecasting*: Generate calibrated future trajectories $\widehat{Y} \in \mathbb{R}^{H\times C}$ by capturing temporal dynamics alongside spatial couplings; (2) *Relational extraction*: Simultaneously infer and preserve a disentangled interaction field $R^\perp \in \mathbb{R}^{C\times C\times d_r}$ (the orthogonal residual of R stripped of unilateral baseline drifts) as an explicit, coordinate-consistent object for downstream physical analysis.

Observation symmetries. In multivariate recording, two coordinate conventions are physically arbitrary: (1) sensor channels have no canonical ordering, so permuting observation columns via a permutation matrix $P \in \{0,1\}^{C\times C}$ transforms inputs as $X \mapsto XP^\top$; and (2) measurement units (e.g., Celsius versus Fahrenheit) are chosen independently per channel, inducing positive affine transformations $X \mapsto a \odot X + b$ with channel-wise scales $a \in \mathbb{R}^C_{>0}$ and offsets $b \in \mathbb{R}^C$. Because permuting sensor channels simultaneously reindexes their respective physical unit parameters, these coordinate transformations combine to form the semi-direct product group $G = \mathrm{Aff}_+(1)^C \rtimes S_C$.

Architectural constraint. A naive neural network is coordinate-dependent: permuting or rescaling inputs causes representations to drift and assigns learned sensor identities to the wrong channels. To ensure coordinate consistency, we therefore impose an explicit **architectural constraint**: the forecaster must be *jointly equivariant* under G. Let $\theta_C = (\gamma, \beta, E)$ denote the complete channel-indexed state of the model, comprising the normalization parameters (γ, β) and the sensor identity table E. Under any coordinate transformation $g = (a, b, P) \in G$, an equivariant forecaster $F(X; \theta_C)$ must satisfy the joint transformation law:

$$F\big(a \odot (XP^\top) + b;\, P\theta_C\big) = a \odot \big(F(X;\theta_C)P^\top\big) + b, \tag{1}$$

where $P\theta_C = (P\gamma, P\beta, PE)$. Equation (1) is a rigorous design requirement: transforming the observational coordinates must alter only the description of the dynamical process, while preserving the underlying physical process and relational correspondence without representation drift.

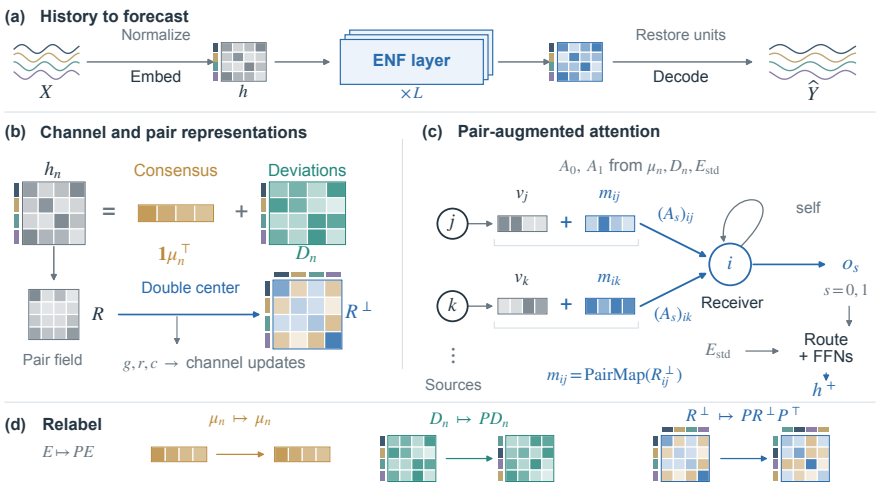


Figure 2: **ENF layer architecture.** (a) Encoding, interaction, and decoding pipeline. (b) State decomposition into consensus/deviations and double-centered pair field $R^{\perp}$. (c) Attention aggregating target-conditioned messages $v_j + m_{ij}$ through equivariant routing. (d) Permutation equivariance across states, identities, and pair axes. (Arrays are schematic.)

3.2 ENF Architecture and Computational Mechanics

Guided by this symmetry principle, the Equivariant Neural Forecaster (ENF) realizes Eq. (1) through a structured hierarchy (Figure 2): 1) Subspace and pair decomposition: first-order states are orthogonally split into a global consensus and zero-mean deviations, while second-order pair features undergo algebraic double-centering to purge additive unilateral drifts (g, r_i, c_j), isolating the bilateral interaction residual $R^{\perp}$. 2) Pair-augmented attention and routing: overcoming the shared-value constraint of conventional attention (Shaw et al., 2018), $R^{\perp}$ supplies target-dependent messages m_{ij} that augment source values $v_j + m_{ij}$, routing updates back to consensus and deviation accumulators. 3) Joint coordinate equivariance: all operations commute with simultaneous permutations and affine unit changes across observations, sensor identities, and pair indices under G.

3.2.1 First-order channel subspace representations

A common challenge in multivariate forecasting is that shared macroscopic trends can easily dominate localized, sensor-specific dynamics. Monolithically updating channel states risks allowing dominant common-mode signals to drown out subtle relative shifts. To decouple these dynamics under the permutation symmetry, ENF organizes first-order states via canonical invariant subspace projections.

Given observations X, each channel's history is mapped by a shared temporal embedder into a d-dimensional latent state $h_i \in \mathbb{R}^d$ (optionally preceded by channel-wise normalization (Kim et al., 2022)), forming the channel state matrix $h \in \mathbb{R}^{C\times d}$. According to the representation theory of S_C, the channel space $\mathbb{R}^C$ decomposes orthogonally into the 1-dimensional *trivial representation* and the $(C-1)$-dimensional *standard representation* via averaging and centering projectors $P_{\text{triv}} = \frac{1}{C}\mathbf{1}\mathbf{1}^{\top}$ and $P_{\text{std}} = I - P_{\text{triv}}$, decomposing state $h = \mathbf{1}\mu^{\top} + D$ with zero-sum constraint $\mathbf{1}^{\top}D = 0$. Here, $\mathbf{1} \in \mathbb{R}^C$ is the all-ones vector and I is the identity matrix. This projects the channel state into two decoupled components:

1) **Global consensus** $\mu = \frac{1}{C}\mathbf{1}^{\top}h = \text{mean}_i\, h_i \in \mathbb{R}^d$, which is permutation-invariant $(\mu(Ph) = \mu(h))$ and tracks shared macroscopic shifts; and

2) **Relative deviations** $D = P_{\text{std}}h \in \mathbb{R}^{C\times d}$, which are strictly zero-mean across channels $(\mathbf{1}^{\top}D = 0)$ and permute equivariantly $(D(Ph) = PD(h))$ to isolate localized dynamics. This separation allows the network to process aggregate trends independently from relative variations, while providing zero-mean deviations D to drive the pair field in §3.2.2.

3.2.2 Decomposing the pair field for interaction messaging

While first-order states h_i capture individual channel dynamics, modeling coupled systems requires evaluating relationships across variable pairs. To incorporate these dependencies, ENF projects

channel representations into relational features $z_i = W_h h_i + b_h \in \mathbb{R}^{d_r}$ and constructs a pairwise representation $R_{ij} \in \mathbb{R}^{d_r}$ for every source–target pair:

$$R_{ij} = W_a z_i + W_b z_j + W_\times (z_i \odot z_j) + b_R. \tag{2}$$

Under channel permutations, this pair field transforms by conjugation as $R \mapsto PRP^\top$ (Hartford et al., 2018; Maron et al., 2019).

Separating additive endpoint effects via double-centering. A common characteristic of raw pairwise features is that they frequently entangle single-variable marginal statistics with coupled joint dependencies: a prominent value in R_{ij} may largely reflect that source j has an elevated baseline amplitude, rather than a specific dynamic coupling between i and j. To isolate the component that depends strictly on both endpoints, we decompose R_{ij} according to its endpoint dependencies:

$$\begin{aligned} g &= \mathrm{mean}_{ij}\, R_{ij}, & r_i &= \mathrm{mean}_j\, R_{ij} - g, \\ c_j &= \mathrm{mean}_i\, R_{ij} - g, & R^\perp_{ij} &= R_{ij} - g - r_i - c_j. \end{aligned} \tag{3}$$

Here, vector-valued $g, r_i, c_j \in \mathbb{R}^{d_r}$ capture global, target, and source deviations. Subtracting these additive components corresponds to algebraic *double-centering*, which projects the pair field onto its two-variable subspace $R^\perp = P_{\mathrm{std}} R P_{\mathrm{std}}$.

By construction, $R^\perp$ has zero row and column means ($\sum_j R^\perp_{ij} = 0$ and $\sum_i R^\perp_{ij} = 0$), rendering it orthogonal to all additive single-variable fields. For the bilinear formulation of R_{ij}, double-centering algebraically eliminates all linear terms and bias, reducing to the centered-product interaction:

$$R^\perp_{ij} = W_\times \big((z_i - \bar{z}) \odot (z_j - \bar{z}) \big), \tag{4}$$

where $\bar{z} = \mathrm{mean}_c\, z_c$. Formal subspace properties and rank bounds of $R^\perp$ appear in §3.3.

Connecting the decomposed field to attention. This orthogonal decomposition governs how information is utilized within the layer: (1) the single-variable terms (g, r, c) are routed into channel state accumulators to steer individual updates; (2) the centered interaction residual $R^\perp$ is separated into symmetric part $R^{\mathrm{sym}} = \frac{1}{2}(R^\perp + R^{\perp\top})$ and antisymmetric part $R^{\mathrm{anti}} = \frac{1}{2}(R^\perp - R^{\perp\top})$ to synthesize target-dependent interaction messages $m_{ij} = W_{\mathrm{pair}}[R^{\mathrm{sym}}_{ij}, R^{\mathrm{anti}}_{ij}] + b_{\mathrm{pair}}$ (where $R^{\mathrm{anti}} \equiv 0$ for the bilinear field, with attention weights providing directional routing).

Unlike standard cross-channel attention, where all targets receive an identical source value v_j, these messages m_{ij} augment attention values as $v_j + m_{ij}$ in §3.2.3, allowing the transmitted message to vary with the receiving target's state.

3.2.3 Pair-augmented attention and equivariant routing

In standard multi-head dot-product attention (Vaswani et al., 2017), each source channel j is projected into a value vector $v_j \in \mathbb{R}^{d_v}$. While attention weights A_{ij} differ across targets, the transmitted content v_j remains identical for all receivers i. This shared-value formulation limits the capacity of attention to adapt message content to specific channel pairings. To alleviate this constraint, ENF integrates the interaction message m_{ij} directly into the value aggregation.

Identity-conditioned attention patterns. For each attention head, let Q_D, K_D denote query and key projections of the centered deviations D, and q_μ, k_μ denote projections of the global consensus μ. Normalized attention logits are computed by combining deviation-to-deviation interactions with consensus-conditioned terms:

$$\Lambda_{ij} = \frac{1}{\sqrt{d_v}} \left(Q^\top_{D,i} K_{D,j} + q^\top_\mu K_{D,j} \right), \tag{5}$$

where the consensus–consensus term is constant across each row and naturally cancels out in row-wise softmax. To incorporate persistent spatial geometry, we introduce an identity-conditioned bias $B(E_{\mathrm{std}}, D)$ formed by projecting centered sensor identities $E_{\mathrm{std}} = P_{\mathrm{std}} E$ alongside deviations D. This yields two complementary attention patterns via row-wise softmax: dynamic attention $A_0 = \mathrm{softmax}(\Lambda)$ and identity-biased attention $A_1 = \mathrm{softmax}(\Lambda + B)$. Here, A_0 captures dynamic temporal dependencies across histories, while A_1 incorporates persistent spatial sensor relationships.

Pair-augmented aggregation and subspace routing. For each attention pattern $s \in \{0, 1\}$, ENF aggregates the augmented values $v_j + m_{ij}$ across source channels and projects the multi-head representation for each target i:

$$(o_s)_i = W_o \sum_{j=1}^{C} (A_s)_{ij}(v_j + m_{ij}), \qquad \delta = P_{\text{std}}(o_1 - o_0). \tag{6}$$

Here, m_{ij} enriches message content while identities modulate routing weights via A_1. Updates are routed through canonical projectors: the invariant mean $P_{\text{triv}}o_0$ updates consensus μ, while centered $P_{\text{std}}o_0 + \delta$ updates deviations D. Because both P_{triv} and P_{std} commute with permutations, this routing strictly preserves permutation equivariance. Shared and identity-conditioned feed-forward networks (FFNs) then complete the layer. Stacking L layers iteratively alternates between subspace routing and second-order interactions before prediction.

3.3 Theoretical Guarantees and Exact Factorization

This subsection formally establishes the coordinate consistency guarantees of ENF and derives an exact computational factorization that resolves the quadratic memory bottleneck.

Joint coordinate equivariance. Recall that an equivariant forecaster must respect the joint transformation law under the semi-direct product group $G = \text{Aff}_+(1)^C \rtimes S_C$ (Eq. (1)). We formalize this guarantee for the complete ENF architecture (Full proof in Appendix A).

Theorem 1 (Joint coordinate equivariance)**.** *Let $F(\cdot; \theta_C)$ denote the complete ENF forecaster with channel-indexed state $\theta_C = (\gamma, \beta, E)$. (1)* ***Permutation****: In evaluation mode, $F(XP^\top; P\theta_C) = F(X; \theta_C)P^\top$ for any permutation P. (2)* ***Affine frame****: Under RevIN in the continuous limit $\epsilon \to 0$ ($\gamma_c \neq 0$)[1] with non-zero channel variance, $F(a \odot X + b; \theta_C) = a \odot F(X; \theta_C) + b$ for all $a \in \mathbb{R}^C_{>0}, b \in \mathbb{R}^C$. (3)* ***Joint group action****: The family $\{F(\cdot; \theta_C)\}$ is jointly equivariant under $G = \text{Aff}_+(1)^C \rtimes S_C$, satisfying Eq.* (1) *for all $g \in G$.*

Target-conditioned context and exact factorization. In shared-value attention, identical attention rows ($A_{i\cdot} = A_{k\cdot} = a$) produce identical aggregated outputs aV. In contrast, ENF's pair field provides target-dependent message content while avoiding dense pair-tensor storage.

Proposition 1 (Target-conditioned context and exact factorization)**.** *Under the default bilinear content field (Eq.* (4)*) with attention matrix A, let $\tilde{z}_i = z_i - \bar{z}$. The relational update $u_i^{\text{rel}} = \sum_j A_{ij} W_m(\tilde{z}_i \odot \tilde{z}_j)$ admits the* ***exact factorization****:*

$$u_i^{\text{rel}} = W_m\left(\tilde{z}_i \odot \eta_i\right), \qquad \textit{where} \quad \eta_i = \sum_{j=1}^{C} A_{ij}\tilde{z}_j. \tag{7}$$

If two targets i and k share identical attention rows $A_{i\cdot} = A_{k\cdot} = a$, their updates differ strictly by target states: $u_i^{\text{rel}} - u_k^{\text{rel}} = W_m((\tilde{z}_i - \tilde{z}_k) \odot \eta)$. (Proof in Appendix A.)

Significance of the factorization. Proposition 1 establishes two core properties: (1) *Expressive target differentiation*: Even when attention rows coincide, Proposition 1 shows that the pair field still differentiates targets via their centered states $\tilde{z}_i - \tilde{z}_k$, provided the attention-weighted source context η_i is non-zero; (2) *Eliminating dense pair-tensor storage*: Because the aggregated source context $\eta_i = \sum_j A_{ij}\tilde{z}_j$ is computed prior to target modulation, the model never needs to instantiate or store the $C \times C \times d_r$ pairwise tensor during prediction. Consequently, forward forecasting scales with $O(Cd_r)$ relation-feature memory without dense pair-tensor materialization, while the mathematically defined pair field $R^\perp$ remains fully preserved and retrievable for post-hoc physical inquiry.

4 Experiments

We evaluate ENF across four key questions: **(Q1)** Does ENF improve multi-horizon forecasting across channel scales? (§4.2) **(Q2)** How do identities, routing, and pair fields drive the gains? (§4.3) **(Q3)** Can pair fields compose relations through shared variables? (§4.4) **(Q4)** Do learned fields translate into physical forces and transferable laws? (§4.5)

[1]Exact affine equivariance holds in the theoretical limit $\epsilon \to 0$; practical implementations adopt $\epsilon = 10^{-5}$ for numerical stability, incurring negligible numerical error.

4.1 Experimental Setup

Datasets and physical testbeds. Our evaluation spans 13 benchmark datasets ($C = 7$ to 883) alongside two instrumented physical systems: (1) *Traffic networks*: PEMS03, PEMS04, PEMS07, and PEMS08 record 5-minute highway flows, while Traffic monitors hourly freeway occupancy (Song et al., 2020; Lai et al., 2018); (2) *Energy and meteorology*: Electricity (hourly load), Solar (10-minute solar power), and Weather (10-minute meteorological indicators) (Lai et al., 2018; Wu et al., 2021); (3) *Low-channel series ($C < 16$)*: ETTh1/2 (hourly), ETTm1/2 (15-minute) transformer temperatures, and daily currency Exchange rates (Zhou et al., 2021; Lai et al., 2018); (4) *Coupled physical testbeds*: a 28-node chaotic *Rössler circuit network* (Vera-Ávila et al., 2020) across 101 physical coupling strengths, and the *Causal Chambers wind tunnel* (Gamella et al., 2025).

Baselines. We compare across four paradigms: channel-independent models (PatchTST (Nie et al., 2023), DLinear (Zeng et al., 2023)), frequency/low-resource baselines (PaiFilter (Yi et al., 2024), MixLinear (Ma et al., 2026)), cross-channel attention (iTransformer (Liu et al., 2024)), and spatial-temporal embedding methods (STAEformer (Liu et al., 2023), STID (Shao et al., 2022)).

Training and evaluation protocols. Datasets follow chronological 60/20/20 splits (70/10/20 on Weather/Traffic). With input $T = 96$, PEMS evaluates multi-horizon forecasting $H \in \{12, 24, 48, 96\}$ under matched dimensions ($d = 128, L = 2$), while the 13-dataset study adopts a compact setup ($d = 64, L = 1$). Models use Adam with cosine-annealed learning rate from 5×10^{-4} and early stopping. We report averaged MSE and MAE across random seeds $\{1, 2, 42\}$.

4.2 Multi-Horizon and Cross-Domain Forecasting Performance

Multi-horizon forecasting on traffic benchmarks. Under matched capacity and $T = 96$, ENF achieves the lowest MSE and MAE across all 48 evaluated dataset–horizon–seed comparisons (Table 1). Overall, ENF reduces mean squared error by 17.5% relative to iTransformer, with the performance advantage progressively widening at longer forecast horizons (e.g., reaching a 28.3% reduction on PEMS08 at $h = 96$). This behavior highlights a key architectural property: while short-term forecasting relies heavily on local temporal auto-correlation, multi-step predictions suffer from error propagation, where coordinate-consistent spatial structures and isolated pair interactions become essential to arrest error compounding.

Performance across diverse domains and channel scales. Figure 3a compares the compact model ($d64/L1$) across seven representative benchmarks. ENF consistently excels on high-channel systems with rich spatial heterogeneity (e.g., PEMS, Electricity, and Weather), achieving up to 15.4% error reductions. Conversely, on low-channel series ($C < 16$, ETT) or domains dominated by a single macroscopic periodic mode (Solar and Traffic), ENF yields competitive but neutral performance, remaining within 1.1% of iTransformer. These results indicate that second-order relational modeling provides the greatest utility in distributed systems with rich cross-variable interactions, whereas first-order dynamics suffice when systems are heavily governed by low-rank global modes.

4.3 Dissecting Architectural Gains: Component Attribution and Ablation

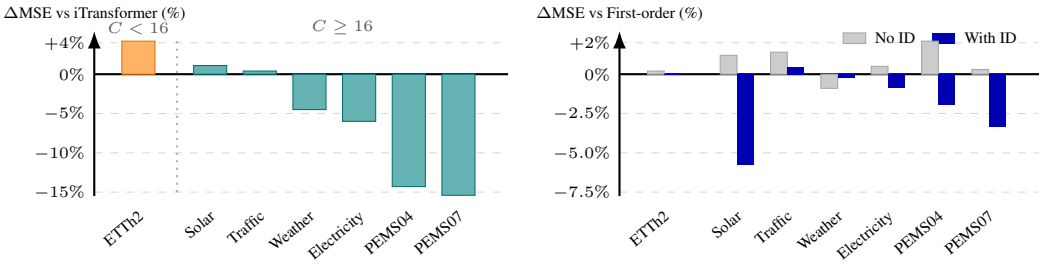
We dissect the distinct contributions of different designs to address **(Q2)**.

Staged performance attribution. A staged ablation on the matched PEMS benchmark isolates the cumulative source of ENF's gains (Appendix Table 5): adding learned sensor identities (ID) reduces error by 12.4%, establishing spatial differentiation is critical in sensor networks; consensus/deviation subspace routing provides an additional 2.0% reduction by decoupling global trends from local shifts; and the relational pair configuration contributes a further 3.9% reduction. These results confirm that the overall 17.5% gain arises from structural contributions across the hierarchy.

Interplay between sensor identity and relations. Figure 3b evaluates the interaction between sensor identities and relational modeling across seven benchmarks. First-order EQNN updates channel states via attention without second-order pair fields, while ENF incorporates the full pair representation. Here, second-order relational modeling complements spatial identities: with learned IDs, ENF consistently outperforms First-order EQNN across heterogeneous networks (reducing MSE by up to 5.75% on Solar and 3.33% on PEMS07). Without identities, channels lack spatial distinction, rendering pairwise differences uninformative and yielding neutral performance.

Table 1: Averaged PEMS results over four horizons with input length 96 (three seeds). Bold indicates the best mean, and underline indicates the second-best.

		ENF		iTransformer		STAEformer		PatchTST		DLinear		PaiFilter		MixLinear	
Dataset	h	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
PEMS03	12	**0.062**	**0.166**	0.066	0.172	0.076	0.185	0.086	0.203	0.117	0.226	0.086	0.196	0.199	0.307
	24	**0.079**	**0.186**	0.087	0.195	0.112	0.222	0.120	0.237	0.233	0.321	0.140	0.252	0.434	0.477
	48	**0.113**	**0.222**	0.135	0.243	0.173	0.271	0.205	0.305	0.532	0.511	0.271	0.357	0.781	0.663
	96	**0.172**	**0.276**	0.202	0.305	0.235	0.325	0.292	0.367	1.025	0.760	0.490	0.496	1.354	0.928
	Avg	**0.107**	**0.212**	0.123	0.229	0.149	0.251	0.176	0.278	0.477	0.455	0.247	0.325	0.692	0.594
PEMS04	12	**0.069**	**0.172**	0.081	0.188	0.086	0.195	0.100	0.215	0.128	0.238	0.101	0.211	0.211	0.322
	24	**0.083**	**0.189**	0.097	0.208	0.105	0.218	0.137	0.251	0.243	0.334	0.164	0.273	0.441	0.491
	48	**0.105**	**0.217**	0.132	0.248	0.152	0.266	0.213	0.318	0.548	0.525	0.316	0.387	0.808	0.682
	96	**0.139**	**0.252**	0.186	0.302	0.187	0.298	0.336	0.410	1.102	0.798	0.561	0.540	1.434	0.960
	Avg	**0.099**	**0.207**	0.124	0.237	0.133	0.244	0.197	0.299	0.505	0.474	0.286	0.353	0.724	0.614
PEMS07	12	**0.055**	**0.149**	0.061	0.158	0.078	0.185	0.073	0.181	0.109	0.221	0.078	0.184	0.190	0.308
	24	**0.073**	**0.168**	0.083	0.184	0.094	0.200	0.115	0.226	0.227	0.324	0.138	0.249	0.441	0.492
	48	**0.097**	**0.195**	0.120	0.227	0.115	0.218	0.179	0.278	0.539	0.522	0.267	0.352	0.813	0.685
	96	**0.130**	**0.232**	0.162	0.273	0.177	0.282	0.280	0.351	1.066	0.774	0.489	0.490	1.395	0.947
	Avg	**0.089**	**0.186**	0.106	0.210	0.116	0.221	0.161	0.259	0.485	0.460	0.243	0.319	0.710	0.608
PEMS08	12	**0.072**	**0.173**	0.075	0.176	0.095	0.202	0.100	0.212	0.121	0.232	0.097	0.203	0.201	0.314
	24	**0.095**	**0.196**	0.104	0.206	0.115	0.222	0.156	0.263	0.232	0.328	0.165	0.270	0.429	0.486
	48	**0.138**	**0.231**	0.165	0.259	0.182	0.283	0.248	0.327	0.540	0.524	0.322	0.384	0.817	0.688
	96	**0.233**	**0.289**	0.325	0.373	0.284	0.346	0.355	0.390	1.120	0.787	0.630	0.540	1.486	0.965
	Avg	**0.134**	**0.222**	0.167	0.253	0.169	0.263	0.215	0.298	0.503	0.468	0.304	0.349	0.733	0.613
Overall		**0.107**	**0.207**	0.130	0.232	0.142	0.245	0.187	0.283	0.493	0.464	0.270	0.337	0.715	0.607



(a) MSE change.

(b) Factorial ablation (ΔMSE vs. First-order EQNN).

Figure 3: **Benchmarking and ablation across representative domains.** (a) ΔMSE relative to iTransformer. (b) Factorial ablation against First-order EQNN with and without learned identities. Negative values favor ENF (Appendix Table 10).

Role of the interaction field. Ablating individual routing pathways indicates that additive single-variable routes (g, r, c) steer primary channel updates, whereas disabling interaction messages (m_{ij}) has a marginal impact on forecast MSE (Appendix Table 6 and Table 17). Because time-series forecasting is heavily governed by temporal auto-correlation and first-order spatial trends, ENF's double-centering strategically purges these dominant unilateral drifts from $R^{\perp}$. This leaves the isolated bilateral interaction residual to serve as a pure relational substrate for higher-level relation operations (§4.4) and continuous physical force discovery (§4.5).

Scalability and computational efficiency. Channel-count scaling across nested subsets of PEMS07 ($C = 8$ to 883) demonstrates that ENF's relative advantage over iTransformer widens from 1.5% to 15.4% as spatial dimensionality expands (Appendix Table 9). Empirical profiling confirms that the exact factorization (Proposition 1) matches dense double-precision outputs to below 10^{-15}, eliminating dense $C^2 \times d_r$ tensor storage and reducing training step latency by over 14x at $C = 256$.

4.4 Composing Relations under a Common Coordinate Action

To address **(Q3)**, we test whether the pair field makes multi-hop relations accessible across shared intermediate variables k. In controlled signed-link and common-neighbor benchmarks (Appendix E), endpoints share identical histories across classes, isolating evidence entirely within intermediates. Applying the common-node transform $Z_{ij} = |\mathcal{K}|^{-1} \sum_{k \in \mathcal{K}} S_{ik} S_{jk}^{\top}$ to $R^{\perp}$ achieves 0.9146 and 0.9051

Table 2: **Discovered formulas across Weather and Wind Tunnel.** Weather tests source extrapolation (R^2_{ext}); wind tunnel tests transfer to physical sensor measurements (R^2_{meas}). Details in Appendix G.

System	Direction	Discovered Analytical Law	Accuracy (R^2)
Weather (extrapolative)	$e \to T_d$	$T_d = -31.3 + 15.55(k/\log e + \log e)$	$\mathbf{0.9968}$ (poly: < 0)
	$c_v \to T_d$	$T_d = -35.07 + 16.27\log(c_v + \sqrt{c_v})$	$\mathbf{0.9938}$ (poly: < 0)
	$T_d \to e$	$e = -1.057 + 7.15\exp((k-1)T_d)$	$\mathbf{0.9973}$ (poly: 0.996)
Wind Tunnel (measured)	$u_{\text{in}} \to n_{\text{in}}$	$n_{\text{in}} = 498.7 + 821.2\lvert z + 1.342\rvert$	$\mathbf{0.9714}$
	$u_{\text{out}} \to n_{\text{out}}$	$n_{\text{out}} = 4337.1 - 1583.8\sqrt{\lvert z + 2.706\rvert}e^{-z/2}$	$\mathbf{0.9545}$
	$p_{\text{up}} \to p_{\text{down}}$	$p_{\text{down}} = 96300.3 + 74.63z$	$\mathbf{0.8436}$

AUC (layers 1/2), whereas additive fields remain at chance (0.5024 / 0.5058) and the model's first-order query head fails (0.5025; Appendix Table 23). Misaligning node alignment collapses AUC to 0.4991. This confirms that $R^{\perp}$ preserves multi-hop relational structure that is linearly accessible even when first-order node states fail.

4.5 Coupling Information Readout and Pair-Guided Symbolic Distillation

For **Q4**, we examine if pair representations capture transferable coupling for symbolic distillation.

Supervised coupling readout in chaotic circuits. On a 28-node Rössler oscillator network across 101 coupling strengths, coupling forces $F_{ij}(t) = g_{ij}(y_j(t) - y_i(t))$ are governed by circuit equations (Appendix Eq. (8)). Using frozen ENF representations, context-pooled pair descriptors ρ_{ij} are mapped through a supervised quadratic readout to estimate coefficients of a degree-two state dictionary (Appendix G.4). On independent test recordings, while estimating continuous forces from unmapped dynamics remains challenging, the $R^{\perp}$ readout achieves $R^2 = 0.0996$, exceeding the additive baseline ($R^2 = 0.0808$) and shuffled controls ($R^2 = 0.0359$; Appendix Table 28). Evaluated on 76 reserved pairs excluded from fitting, the residual map achieves $R^2 = 0.1176$ (versus 0.0971 for additive and 0.0429 for shuffled controls), confirming that supervised readout detects partially transferable coupling information beyond single-variable baselines.

Pair-guided symbolic distillation on forecasts. On Weather, we test whether relationally selected variables yield compact analytical distillations of modeled dynamics (Table 2, top). The six strongest pairs ranked by raw pair amplitude define twelve directed symbolic regressions on model forecasts, fitted on the lower 60% of source values and evaluated on the upper 20% extrapolation range. Three searches discover logarithmic dew-point formulas achieving $R^2_{\text{ext}} > 0.99$, whereas polynomial baselines fail ($R^2_{\text{ext}} < 0$). While static rank correlations provide an equally effective baseline for variable screening (Appendix Table 25), endogenous pair amplitudes identify dynamically coupled coordinates whose modeled dependencies extrapolate robustly under shifts.

Transfer of forecast-derived relations to physical sensors. We further examine whether equations distilled from forecasts generalize to physical instruments in the Causal Chambers wind tunnel (Gamella et al., 2025) (Table 2, bottom). Twelve selected pairs define 24 directed searches per seed; expressions fitted on one-step forecasts are frozen and evaluated directly against physical sensor measurements across three unseen runs. Across 48 directed fits, 45 achieve positive measured R^2 (median $R^2 = 0.8381$), suggesting that relational representations can help extract empirical coupling relationships that transfer meaningfully to real hardware measurements.

5 Conclusion

This work bridges the gap between black-box forecasting and coordinate-consistent relational analysis in complex multivariate systems. We introduce ENF, an architecture that grounds heterogeneous sensor dynamics within a unified joint symmetry framework. By decomposing channel representations into global consensus and local deviation subspaces and applying algebraic double-centering to isolate pure bilateral interactions from unilateral baseline drifts, ENF resolves both coordinate sensitivity and relational confounding. Crucially, an exact computational factorization eliminates quadratic tensor storage, enabling efficient forecasting while preserving explicit pair fields for post-hoc analysis. Across diverse forecasting benchmarks and physical testbeds, ENF not only delivers robust multi-horizon prediction improvements over channel-attention baselines, but also demonstrates genuine physical fidelity—enabling calibrated readouts of continuous coupling forces in chaotic circuits and guiding symbolic formulas that transfer directly to real-world physical instrumentation.

AI USE STATEMENT

Generative AI tools assisted manuscript drafting and editing, literature and citation checks, interpretation of experimental results, and LaTeX formatting. The authors take responsibility for the final content, including all text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Si-An Chen, Chun-Liang Li, Nathanael C. Yoder, Sercan Ö. Arık, and Tomas Pfister. TSMixer: An all-MLP architecture for time series forecasting. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=wbpxTuXgm0.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International Conference on Machine Learning (ICML)*, 2016. URL https://proceedings.mlr.press/v48/cohenc16.html.

Miles Cranmer, Alvaro Sanchez Gonzalez, Peter Battaglia, Rui Xu, Kyle Cranmer, David Spergel, and Shirley Ho. Discovering symbolic models from deep learning with inductive biases. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 17429–17442, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/c9f2f917078bd2db12f23c3b413d9cba-Abstract.html.

Juan L. Gamella, Jonas Peters, and Peter Bühlmann. Causal chambers as a real-world physical testbed for ai methodology. *Nature Machine Intelligence*, 7(1):107–118, 2025. doi: 10.1038/s42256-024-00964-x. URL https://doi.org/10.1038/s42256-024-00964-x.

Lu Han, Han-Jia Ye, and De-Chuan Zhan. The capacity and robustness trade-off: Revisiting the channel independent strategy for multivariate time series forecasting. *IEEE Transactions on Knowledge and Data Engineering*, 36(11):7129–7142, 2024. doi: 10.1109/TKDE.2024.3400008. URL https://doi.org/10.1109/TKDE.2024.3400008.

Jason Hartford, Devon Graham, Kevin Leyton-Brown, and Siamak Ravanbakhsh. Deep models of interactions across sets. In *International Conference on Machine Learning (ICML)*, pp. 1909–1918, 2018. URL https://proceedings.mlr.press/v80/hartford18a.html.

Md Shamim Hussain, Mohammed J. Zaki, and Dharmashankar Subramanian. Global self-attention as a replacement for graph convolution. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 655–665, 2022. doi: 10.1145/3534678.3539296. URL https://doi.org/10.1145/3534678.3539296.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of NAACL-HLT, Volume 1 (Long and Short Papers)*, pp. 3543–3556, 2019. doi: 10.18653/v1/N19-1357. URL https://aclanthology.org/N19-1357/.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International Conference on Learning Representations (ICLR)*, 2022.

Thomas Kipf, Ethan Fetaya, Kuan-Chieh Wang, Max Welling, and Richard Zemel. Neural relational inference for interacting systems. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 2688–2697. PMLR, 2018.

Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long- and short-term temporal patterns with deep neural networks. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 95–104, 2018. doi: 10.1145/3209978.3210006.

Hunjae Lee and Corey Clark. Dynamic relational priming improves transformer in multivariate time series. *arXiv preprint arXiv:2509.12196*, 2025.

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *International Conference on Learning Representations (ICLR)*, 2018.

Hangchen Liu, Zheng Dong, Renhe Jiang, Jiewen Deng, Jinliang Deng, Quanjun Chen, and Xuan Song. Spatio-temporal adaptive embedding makes vanilla transformer sota for traffic forecasting. In *ACM International Conference on Information and Knowledge Management (CIKM)*, 2023.

Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: Exploring the stationarity in time series forecasting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2310.06625.

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

Haggai Maron, Heli Ben-Hamu, Nadav Shamir, and Yaron Lipman. Invariant and equivariant graph networks. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://openreview.net/forum?id=Syx72jC9tm.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=Jbdc0vTOcol.

Zezhi Shao, Zhao Zhang, Fei Wang, Wei Wei, and Yongjun Xu. Spatial-temporal identity: A simple yet effective baseline for multivariate time series forecasting. In *ACM International Conference on Information and Knowledge Management (CIKM)*, 2022. URL https://arxiv.org/abs/2208.05233.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *Proceedings of NAACL-HLT, Volume 2 (Short Papers)*, pp. 464–468. Association for Computational Linguistics, 2018. doi: 10.18653/v1/N18-2074. URL https://aclanthology.org/N18-2074/.

Chao Song, Youfang Lin, Shengnan Guo, and Huaiyu Wan. Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 914–921, 2020. doi: 10.1609/aaai.v34i01.5438.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

V. P. Vera-Ávila, Ricardo Sevilla-Escoboza, A. A. Lozano-Sánchez, R. R. Rivera-Durón, and Javier M. Buldú. Experimental datasets of networks of nonlinear oscillators: Structure and dynamics during the path to synchronization. *Data in Brief*, 28:105012, 2020. doi: 10.1016/j.dib.2019.105012. URL https://doi.org/10.1016/j.dib.2019.105012.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 22419–22430, 2021.

Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. Graph wavenet for deep spatial-temporal graph modeling. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2019.

Jiyuan Xu, Wenyu Zhang, Xin Jing, Jiahao Nie, Shuai Chen, and Shuai Zhang. Cpiri: Channel permutation-invariant relational interaction for multivariate time series forecasting. *arXiv preprint arXiv:2601.20318*, 2026.

Kun Yi, Jingru Fei, Qi Zhang, Hui He, Shufeng Hao, Defu Lian, and Wei Fan. FilterNet: Harnessing frequency filters for time series forecasting. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 55115–55140, 2024. doi: 10.52202/079017-1749. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/6323d96f79d5d49e0d3fc88835c082cd-Abstract-Conference.html.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

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

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

SUPPORTING MATERIAL

The supporting material develops the same progression from coordinate guarantees to operations on relations. Appendix A proves the equivariance and decomposition identities; Appendix B specifies implementation and evaluation. Forecasting controls and scaling appear in Appendix C. Controlled route and representation tests (Appendix D) distinguish predictive use from relational information, and Appendix E evaluates composition through shared variables. Topology readouts and physical analyses appear in Appendices F and G.

A MATHEMATICAL FOUNDATIONS

A.1 PERMUTATION-EQUIVARIANT COMPUTATION

The coordinate guarantee follows by tracking each object's channel indices through the layer. A permutation $\pi \in S_C$ acts by $(\pi \cdot h)_i = h_{\pi^{-1}(i)}$, where π^{-1} reverses the reordering. Its matrix P acts as $h \mapsto Ph$, $X \mapsto XP^\top$, and $R \mapsto PRP^\top$. The last action, called conjugation, reorders both pair indices together. Each feature coordinate carries a copy of this action. We first prove the layer identity with training randomness disabled, then combine it with the normalization frame's unit-change law.

Table 3: Transformation types under channel permutations. Zero channel mean and equivariance place different conditions on a representation.

Object	Action	Representation
h, identity E	Ph, PE	natural permutation
μ, relational g	unchanged	trivial
D, E_{std}, r, c	one permuted axis	standard, when centered
R	$PRP^\top$	second-order tensor
$R^\perp$	$PR^\perp P^\top$	$\mathrm{Std} \otimes \mathrm{Std}$, reducible
γ, β	$P\gamma, P\beta$	channel-indexed affine state

Theorem 2 (Permutation equivariance)**.** *An ENF encoder layer in evaluation mode satisfies* $T(Ph, PE) = PT(h, E)$. *When supplied with a coherence field* Γ*, it satisfies* $T(Ph, PE, P\Gamma P^\top) = PT(h, E, \Gamma)$.

Proof. Shared feature maps commute with channel permutations. Channel means are invariant and centering commutes with P, so $\mu(Ph) = \mu(h)$ and $D(Ph) = PD(h)$.

For content features, the same function is applied to each ordered pair: $R'_{ij} = \Phi(h_{\pi^{-1}(i)}, h_{\pi^{-1}(j)})$. Thus $R' = PRP^\top$. A channel-wise Fourier transform followed by cross-products gives the same covariance for coherence. The grand mean is invariant, r, c transform as Pr, Pc, and $R^\perp$ transforms by conjugation. Transposition commutes with conjugation, so the symmetric and antisymmetric parts do as well. Shared routing maps preserve these laws; a shared bias preserves covariance, but need not keep the updates centered.

Consensus-conditioned logits satisfy $\Lambda' = P\Lambda P^\top$. Centered identity satisfies $E'_{\mathrm{std}} = PE_{\mathrm{std}}$, so both static and dynamic identity biases transform by conjugation. Row-wise softmax obeys

$$\mathrm{softmax}(P\Lambda P^\top) = P\,\mathrm{softmax}(\Lambda)P^\top.$$

For each attention pattern A_s, $s \in \{0, 1\}$, the values obey $v'_j = v_{\pi^{-1}(j)}$ and messages satisfy $m'_{ij} = m_{\pi^{-1}(i),\pi^{-1}(j)}$. Reindexing its aggregation then yields the same channel transformation:

$$\sum_j (A'_s)_{ij}(v'_j + m'_{ij}) = \sum_k (A_s)_{\pi^{-1}(i),k}(v_k + m_{\pi^{-1}(i),k}).$$

Both o_0 and o_1 therefore transform by P, as does $\delta = P_{\mathrm{std}}(o_1 - o_0)$.

Shared FFNs, consensus concatenation, and identity-conditioned products are equivariant. A sigmoid gate transforms by permutation but is not generally zero-mean; final centering supplies its standard-sector output. Split-and-route commutes with P by the proposition below. Composition proves the layer identity. The shared embedding, final normalization, and decoder preserve this identity under composition. □

Proposition 2 (Split-and-route on invariant subspaces)**.** *Let ϕ be an equivariant map into a representation space, and Q a projection commuting with every channel permutation. Then $Q\phi(Px) = PQ\phi(x)$ and $Q\phi(x) \in \operatorname{im} Q$, the set of vectors obtainable by applying Q. The complementary output $(I - Q)\phi(x)$ is also equivariant, with sum $\phi(x)$, so this routing preserves the sum of the nonlinear update.*

Proof. $Q\phi(Px) = QP\phi(x) = PQ\phi(x)$ by equivariance and commutation. The image claim follows from $Q^2 = Q$; replacing Q by $I - Q$ proves the complement claim. Their sum is $\phi(x)$. □

Remark 1 (Equivariance and zero channel mean)**.** *For $Q = P_{\mathrm{std}}$, routing separates an equivariant nonlinear update into its mean and a zero-mean component. Both outputs remain equivariant. Appendix B.5 describes how the complete layer combines these projected outputs with shared affine updates.*

A.2 Affine frames and the joint coordinate action

Reordering changes the placement of channels, whereas an affine unit change modifies their numerical values. RevIN makes the normalized input identical under positive affine changes when its statistics transform exactly. This property supplies the affine factor of the joint law.

Theorem 3 (Affine equivariance)**.** *With the instance-dependent affine frame (RevIN) using $\epsilon = 0$, nonzero per-channel variance, and nonzero learned gains γ_c, ENF satisfies $F(a \odot X + b) = a \odot F(X) + b$ for any per-channel positive affine (a, b), $a > 0$.*

Proof. RevIN computes per-channel (m_c, s_c) and normalizes $\tilde{x}_{t,c} = (x_{t,c} - m_c)/s_c$; the network operates on $\tilde{X}$ and the output is de-normalized $\hat{y} = s \odot \tilde{y} + m$. For an affine-transformed input $X' = a \odot X + b$: $m'_c = a_c m_c + b_c$, $s'_c = a_c s_c$ (positive affine of the statistics), so $\tilde{x}'_{t,c} = (a_c x_{t,c} + b_c - (a_c m_c + b_c))/(a_c s_c) = \tilde{x}_{t,c}$. A fixed learned affine transform $\gamma \odot \tilde{X} + \beta$ is also invariant, and its inverse is defined when every $\gamma_c \neq 0$. Thus the normalized prediction after undoing this learned affine transform satisfies $\tilde{y}' = \tilde{y}$, and $\hat{y}' = s' \odot \tilde{y} + m' = a \odot (s \odot \tilde{y} + m) + b = a \odot \hat{y} + b$. □

Corollary 1 (Equivariance of the parameterized family under $G = \mathrm{Aff}_+(1)^C \rtimes S_C$)**.** *Let $F(\,\cdot\,; \gamma, \beta, E)$ denote the deployed ENF forecaster with RevIN affine parameters $\gamma, \beta \in \mathbb{R}^C$ and channel-identity field $E \in \mathbb{R}^{C \times d_e}$ (the channel-indexed state defined in §3). With exact instance statistics ($\epsilon = 0$, nonzero channel variance and $\gamma_c \neq 0$), the* family $\{F(\,\cdot\,; \gamma, \beta, E)\}_{\gamma,\beta,E}$ *is equivariant under $G = \mathrm{Aff}_+(1)^C \rtimes S_C$: for every $g = (a, b, \pi)$ with $a \in \mathbb{R}^C_{>0}$, $b \in \mathbb{R}^C$, $\pi \in S_C$, acting on input and per-channel state by $g \cdot (X, \gamma, \beta, E) = \big(a \odot (\pi \cdot X) + b,\ \pi \cdot \gamma,\ \pi \cdot \beta,\ \pi \cdot E\big)$,*

$$F\big(a \odot (\pi \cdot X) + b;\ \pi \cdot \gamma, \pi \cdot \beta, \pi \cdot E\big) = a \odot \big(\pi \cdot F(X; \gamma, \beta, E)\big) + b,$$

or in matrix form for the column-channel input $X \in \mathbb{R}^{T \times C}$, $F(a \odot XP^\top + b;\ P\gamma, P\beta, PE) = a \odot \big(F(X; \gamma, \beta, E)\, P^\top\big) + b$. Channel-indexed state follows the same relabeling as its observations. If $P\gamma = \gamma$, $P\beta = \beta$, and $PE = E$, the identity specializes to input-only permutation equivariance for that fixed model. Distinct learned identities instead give the joint action verified in Appendix B.4.

Proof. **Step 1 (permutation factor, joint).** $F(\pi \cdot X; \pi \cdot \gamma, \pi \cdot \beta, \pi \cdot E) = \pi \cdot F(X; \gamma, \beta, E)$. The RevIN statistics are computed per channel over time, so they relabel with the channels ($m(\pi \cdot X) = \pi \cdot m(X)$, likewise s); the channel embedding and output projection are shared maps applied per channel; every ENF layer is equivariant by Theorem 2 with the identity reindexed $\pi \cdot E$; and the RevIN affine (γ, β) enters per channel, so reindexing it composes with the frame. Composing embedding, layers, projection, and the frame, the equality follows.

Step 2 (affine factor). Let $Y := (\pi^{-1} \cdot a) \odot X + (\pi^{-1} \cdot b)$, so that $\pi \cdot Y = a \odot (\pi \cdot X) + b$ (both sides place $a_i \odot X_{:,\pi^{-1}(i)} + b_i$ in channel i, using $(\pi \cdot (\pi^{-1} \cdot a)) = a$). Then

$$\begin{aligned} F\big(a \odot (\pi \cdot X) + b;\ \pi \cdot \gamma, \pi \cdot \beta, \pi \cdot E\big) &= F\big(\pi \cdot Y;\ \pi \cdot \gamma, \pi \cdot \beta, \pi \cdot E\big) \\ &\overset{\text{Step 1}}{=} \pi \cdot F(Y; \gamma, \beta, E) \\ &\overset{\text{Thm. 3}}{=} \pi \cdot \big((\pi^{-1} \cdot a) \odot F(X; \gamma, \beta, E) + (\pi^{-1} \cdot b)\big) \\ &= a \odot \big(\pi \cdot F(X; \gamma, \beta, E)\big) + b, \end{aligned}$$

where Theorem 3 applies with fixed parameters (γ, β, E) under the input-side affine transform. □

A.3 Pair decomposition and content structure

The pair decomposition must isolate additive endpoint effects without changing the coordinate action. The following projections identify the interaction component while preserving this action.

Lemma 1 (Invariant decomposition of the pair space)**.** *Under simultaneous channel permutation, the pair space decomposes orthogonally as*

$$\begin{aligned}\mathbb{R}^{C\times C} = \operatorname{span}(\mathbf{1}\mathbf{1}^\top) \oplus \{r\mathbf{1}^\top : \mathbf{1}^\top r = 0\} \\ \oplus \{\mathbf{1}c^\top : \mathbf{1}^\top c = 0\} \oplus \{M : M\mathbf{1} = 0,\ \mathbf{1}^\top M = 0\}.\end{aligned}$$

The ENF sectors are projections onto these invariant summands.

Proof. Set $J = \mathbf{1}\mathbf{1}^\top/C$ and $P_{\mathrm{std}} = I - J$. Expanding $(J + P_{\mathrm{std}})R(J + P_{\mathrm{std}})$ gives JRJ, $P_{\mathrm{std}}RJ$, JRP_{std}, and $P_{\mathrm{std}}RP_{\mathrm{std}}$. They are mutually orthogonal because $JP_{\mathrm{std}} = 0$, and both projectors commute with every permutation. Their dimensions $1, C-1, C-1, (C-1)^2$ sum to C^2. □

The four-way split is organized by endpoint dependence. Its interaction subspace can be decomposed further: P_{std} itself lies in that subspace and satisfies $PP_{\mathrm{std}}P^\top = P_{\mathrm{std}}$ for every permutation, spanning an invariant one-dimensional component. Thus the routing decomposition removes additive fields, while irreducible representation structure is a finer mathematical distinction.

Lemma 2 (Default content-field structure)**.** *For $R_{ij} = W_a z_i + W_b z_j + W_\times(z_i \odot z_j) + b_R$,*

$$R^\perp_{ij} = W_\times\big((z_i - \bar z) \odot (z_j - \bar z)\big).$$

Each feature matrix of $R^\perp$ is symmetric and has rank at most d_r, the dimension of z_i. Hence the antisymmetric part vanishes in exact arithmetic.

Proof. Double centering annihilates terms depending only on i, only on j, or neither. For each product feature, $P_{\mathrm{std}}(z_{\cdot k}z_{\cdot k}^\top)P_{\mathrm{std}} = \tilde z_{\cdot k}\tilde z_{\cdot k}^\top$, where $\tilde z = z - \bar z$. Each output feature is a weighted sum of at most d_r symmetric rank-one matrices. □

For the shared affine message $m_{ij} = W_{\mathrm{pair}}R^\perp_{ij} + b_{\mathrm{pair}}$, only the first term is double-centered. Uniform attention annihilates the centered term and leaves the shared bias. Nonuniform attention can still produce different updates across receiving targets: the same symmetric pair field is evaluated at each target and aggregated with that target's attention row.

B Experimental protocols and implementation scope

The comparisons hold forecasting tasks fixed while varying architecture, identity, or pair routes. This appendix specifies the data and optimization shared within each comparison, the configurations that differ, and the numerical checks linking the implemented model to Appendix A.

B.1 Experimental setup

Datasets and horizons. PEMS03, PEMS04, PEMS07, and PEMS08 record traffic every five minutes with $C = 358, 307, 883, 170$ channels. The compact study predicts 12 future steps; the matched study predicts 12, 24, 48, and 96. Other compact benchmarks predict 96 steps: Electricity records electricity demand ($C = 321$), Weather meteorological variables ($C = 21$), Traffic road occupancy ($C = 862$), Solar solar power ($C = 137$), ETT transformer temperature and load variables ($C = 7$), and Exchange currency exchange rates ($C = 8$). All use 96 input steps; RevIN is used on every dataset except Solar.

Training and reporting. ENF and iTransformer use Adam, an adaptive gradient optimizer, with zero weight decay and MSE loss. Learning rates follow cosine decay. Training uses bfloat16 (bf16) mixed precision, at most 50 passes through the training data (epochs), and early stopping after five epochs without validation improvement. Compact models have width 64, one layer, FFN hidden width $d_{ff} = 128$, four heads ($d_v = 16$). The initial learning rate is 5×10^{-4} on all compact datasets, including ETT and Exchange.

Table 10 uses five seeds $\{1, 2, 3, 4, 42\}$ on PEMS03, PEMS04, and PEMS08, and three seeds $\{1, 2, 42\}$ elsewhere, matched across models. Each mean includes runs at the full-channel configuration in Table 4, including batch 16 for ENF on PEMS07 and Electricity. Validation selects the

checkpoint within each run. All reported SDs are sample SDs, with denominator $n - 1$. The matched multi-horizon PEMS study uses seeds $\{1, 2, 42\}$.

Table 7 reports a batch-matched identity–relation panel: seven datasets, four configurations, and five seeds $\{1, 2, 3, 4, 42\}$. All 140 runs use batch 32, $d64/L1$, FFN width 128, input length 96, and the common compact optimization schedule; ENF uses $d_r = 8$. Each architecture is evaluated with a trainable identity table and with that table frozen at zero.

Variants and regularization. First-order EQNN combines identity, consensus/deviation decomposition, consensus-conditioned attention, and split-and-route. ENF additionally constructs and routes the pair field; the compact content relation width is $d_r = 8$. The staged comparison measures these complete configurations. Within-ENF controls in Appendix C.1 isolate individual routes.

B.2 Complete ENF layer

Let h_n be the channel state after normalization across its feature dimension, with $\mu_n = \mathrm{mean}_c\, h_n$ and $D_n = h_n - \mathbf{1}\mu_n^\top$. The layer computes pair content from this shared input, combines it with two identity-conditioned attention patterns, and routes their updates as follows.

1. Construct R from h_n. Route g to the consensus accumulator and a shared map of $[r, c]$ to the channel accumulator. Map the interaction component to target–source messages, $m_{ij} = W_{\mathrm{pair}}[R_{ij}^{\mathrm{sym}}, R_{ij}^{\mathrm{anti}}] + b_{\mathrm{pair}}$. The shared bias is channel-constant and is separate from the pair features produced by double centering.
2. Form attention logits by combining deviation-to-deviation terms with consensus-conditioned terms, $\Lambda_{ij} = (Q_{D,i}^\top K_{D,j} + q_\mu^\top K_{D,j} + Q_{D,i}^\top k_\mu)/\sqrt{d_v}$. Here Q_D, K_D are query and key projections of D_n, while q_μ, k_μ are projections of μ_n. The omitted consensus–consensus term is constant within a row and is absorbed by softmax.
3. Center the identity as $E_{\mathrm{std}} = P_{\mathrm{std}}E$, use it with D_n to form the static and dynamic bias $B(E_{\mathrm{std}}, D_n)$, and obtain two attention patterns, $A_0 = \mathrm{softmax}(\Lambda)$ and $A_1 = \mathrm{softmax}(\Lambda + B)$. For $s \in \{0, 1\}$, compute

$$o_s = W_o \left[\sum_j (A_s)_{ij}(v_j + m_{ij}) \right]_{i=1}^{C}, \qquad \delta = P_{\mathrm{std}}(o_1 - o_0).$$

   Both patterns aggregate $v_j + m_{ij}$: the first-order identity field changes the routing weights, and the pair object changes the transmitted content, which is where the relation itself enters the update. Send the mean of o_0 to consensus and its centered component to deviations, together with the identity-dependent correction δ.
4. Apply the shared deviation FFN, optionally conditioned on μ_n, and split its output into mean and centered updates. A shared identity-conditioned FFN supplies another centered update. Recombine the accumulators to produce the next channel state.

After the final layer, feature normalization and a shared linear decoder produce each channel's normalized future. Inverse RevIN restores its units.

The channel-indexed state consists of identity rows E and RevIN parameters γ, β. Jointly reindexing these quantities gives the evaluation-mode permutation identity proved in Appendix A.1. During training, jointly reindexing dropout masks preserves sample-wise equivariance; independent masks preserve equivariance in distribution.

B.3 Per-dataset hyperparameters

Table 4 summarizes the compact single-horizon study and selected larger configurations. The matched multi-horizon PEMS study uses $d128/L2$. With four heads, $d_v = d/4$, giving $d_v = 16$ for the compact backbone and $d_v = 32$ for the matched PEMS backbone.

Compact ENF uses batch 16 on PEMS07, Electricity, and Traffic because of its pair-activation memory, while compact iTransformer uses batch 32. A batch-size check on PEMS07/Electricity gave MSE 0.0585/0.1430 at batch 32 and 0.0584/0.1426 at batch 16. The difference is below 0.3% in both cases, so the compact comparison is insensitive to batch size.

Aggregate estimands and dependence. Let e_{dhs} and b_{dhs} be paired MSEs for dataset d, horizon h, and seed s. Headline percentages use $\overline{e}/\overline{b} - 1$. Averaging individual ratios gives $\overline{e/b - 1}$, while log-ratio aggregation gives $\exp(\overline{\log(e/b)}) - 1$. These are different estimands. For ENF versus iTransformer on PEMS they are -17.5%, -14.8%, and -15.1%, respectively.

We average log-ratios within each of the four datasets before a one-sample t-test (df 3). The cluster bootstrap resamples datasets, averages their log-ratios, and exponentiates the endpoints; its interval therefore targets the geometric relative change. Per-horizon runs share a test series, so significance is established at the dataset level rather than by pooling cells.

Input length and learning rate for DLinear. We vary learning rate over $\{0.005, 0.05\}$ and input length over $\{96, 336\}$ with three seeds. At the protocol length 96, neither learning rate improves the Table 1 entry: geometric mean changes are $+0.8\%$ and $+8.0\%$.

Using input length 336 reduces DLinear's MSE by 56.6% relative to its length-96 entry. Against ENF at length 96, its geometric mean MSE is higher by 32.3% on PEMS03, 54.0% on PEMS04, and 54.1% on PEMS08, with no wins among 33 shared cells. Input duration materially affects this baseline, while the main table retains the common length-96 task. PEMS07 at length 336 is omitted because the full-split GPU preload exceeds available memory.

STAEformer baseline. STAEformer uses three temporal attention layers followed by three spatial attention layers, with learned adaptive channel embeddings and no supplied graph. Its original training recipe uses learning rate 0.001, Huber loss (quadratic for small errors and linear for large errors), scheduled stepwise learning-rate drops, at most 100 epochs, and early stopping patience 10. Time-stamp features match the original pipeline exactly; normalization and data splits remain the same across the compared models.

We evaluate input lengths 12 and 96. At length 96, STAEformer's geometric mean MSE per dataset is 27–38% above ENF, between iTransformer and PatchTST on all four datasets. It wins none of the 48 cells in either MSE or MAE.

At its native input length 12, the per-dataset gap is 57–117% and grows with the horizon on each dataset, reaching 261% at $h = 96$ on PEMS07. Extending the input to 96 improves STAEformer substantially, while increasing the temporal extent processed by spatial attention eightfold. The main table uses this stronger length-96 result.

PaiFilter and MixLinear baselines. PaiFilter uses FilterNet's shared temporal filter and two-layer prediction MLP with hidden width 256 (Yi et al., 2024). Its native affine RevIN supplies the sole instance-normalization layer, with the external wrapper disabled. MixLinear uses its temporal and frequency paths with period length 24, low-pass width 4, and mixing coefficient 0.9 (Ma et al., 2026), together with external affine RevIN and the backbone's mean centering. These native architectures are evaluated under the common forecasting protocol below.

Both use all channels, input length 96, chronological 60/20/20 splits, and a dataset standardizer fitted only on training observations. Adam uses learning rate 5×10^{-4}, zero weight decay, MSE loss, bf16 precision, at most 50 epochs, and early-stopping patience five. Cosine annealing uses $T_{\max} = 100$ epochs. Batch size is 16 on PEMS07 and 32 elsewhere. Validation selects each checkpoint; test MSE and MAE are measured in training-standardized dataset units after reversing instance normalization. Each dataset–horizon entry averages seeds $\{1, 2, 42\}$.

B.4 Equivariance verification

The numerical tests follow the objects transformed in the proof: encoder states first, then the complete model with its channel-indexed parameters. They disable training randomness and retain nonzero static and dynamic identity-bias coefficients. Full-model tests also randomize RevIN's γ, β away from their defaults. Relative maximum error compares the transformed output with the original output after the same reordering. An input-only permutation provides a control that changes the association between histories and sensor-specific state.

The 32-bit floating-point (fp32) tests report approximately 1.8×10^{-7} for the encoder with identity disabled, 3.8×10^{-7} for the encoder with input and identity jointly permuted, and 1.8×10^{-7} for the full model with input, identity, and RevIN parameters jointly permuted. Holding all channel-indexed parameters fixed produces an input-only error of approximately 0.70. The deployed coordinate identity is

$$F(XP^\top; P\gamma, P\beta, PE) = F(X; \gamma, \beta, E)P^\top.$$

Table 4: Compact single-horizon configurations and validation-selected alternatives. Entries list d/L/batch size. Compact means use the seed counts in Table 10; alternative seed counts are listed. The matched multi-horizon PEMS study uses $d128/L2$ (Section 4.1).

Dataset	C	H	Compact (d/L/bs)			Tuned (d/L/bs)		
			ENF	iTf	MSE	ENF	MSE	seeds
PEMS03	358	12	64/1/32	64/1/32	0.06429	128/2/32	0.0636	3
PEMS04	307	12	64/1/32	64/1/32	0.07391	128/2/32	0.0689	1
PEMS07	883	12	64/1/**16**	64/1/32	0.05870	128/2/**16**	0.0562	3
PEMS08	170	12	64/1/32	64/1/32	0.07446	256/2/32	0.0729	5
Electricity	321	96	64/1/**16**	64/1/32	0.14292	64/1/**16**	0.1429	5
Weather	21	96	64/1/32	64/1/32	0.15779	64/1/32	0.1578	5
Traffic	862	96	64/1/**16**	64/1/32	0.45199	64/3/**16**	0.4264	1
Solar†	137	96	64/1/32	64/1/32	0.19704	256/1/32	0.1874	3
ETTh1/h2/m1/m2	7	96	64/1/32	64/1/32	–	64/1/32	–	3
Exchange	8	96	64/1/32	64/1/32	–	64/1/32	–	3

Bold **16** = ENF batch size reduced for C^2 memory (iTransformer stays at 32). ETT/Exchange provide low-channel comparisons (Appendix C.10). FFN widths are 128, 256, and 512 for backbone widths 64, 128, and 256.

The joint action carries each channel's history, normalization parameters, and identity together. Holding the parameters fixed instead assigns reordered histories to different channel-specific states.

Affine equivariance is checked separately with the implemented RevIN stabilizer $\epsilon = 10^{-5}$. Writing $v_c = \mathrm{var}_c(X)$, normalization after a positive scale change gives

$$\frac{a_c(x_c - m_c)}{\sqrt{a_c^2 v_c + \epsilon}} = \frac{x_c - m_c}{\sqrt{v_c + \epsilon/a_c^2}}.$$

The exact theorem follows at $\epsilon = 0$; with stabilization, the discrepancy depends on the scale relative to channel variance. Under the tested transformations, ENF, iTransformer, PatchTST, and DLinear with the shared RevIN frame give affine-test errors between 5×10^{-6} and 4×10^{-5}.

Relation-field scope. The relation-field harness checks $R \mapsto PRP^\top$ and the corresponding transformations of $R^\perp$, R^{sym}, and R^{anti} to approximately 10^{-7} in a one-layer coherence configuration. Coherence is computed directly from normalized input. Content features are recomputed from each layer's hidden state, which can depend on identity, so their covariance follows the joint action on observations and channel-indexed state. These checks establish coordinate consistency; Appendix F.1 separately evaluates topology readouts against known coupling matrices.

B.5 Projected updates and shared affine components

Centering assigns a proposed channel update to the deviation subspace: $\mathbf{1}^\top P_{\mathrm{std}} u = 0$. The attention/FFN probe measures channel-mean deviations of order 10^{-2} before projection and 10^{-8} after projection in evaluation mode.

The complete layer also routes endpoint features through a shared affine map:

$$u_i = W_{\mathrm{rc}}[r_i, c_i] + b_{\mathrm{rc}}, \qquad \mathrm{mean}_i\, u_i = b_{\mathrm{rc}}.$$

The bias is a channel-constant component, so the complete affine update and its centered part both preserve permutation equivariance. Independent dropout masks preserve equivariance in distribution while allowing a nonzero channel mean in an individual realization.

Thus channel centering specifies the subspace of each projected output, while joint equivariance governs every update in the complete layer.

C Forecasting controls and scaling

The forecasting controls identify the effects of sensor identity and pair routes before varying channel count, backbone size, and training history. Resource measurements and the content factorization then distinguish the cost of storing pair features from the cost of computing their contribution.

C.1 COMPONENT-LEVEL ABLATIONS

The staged comparison tests how sensor identity and representation organization contribute to prediction (Table 5). First-order computation already exchanges information between channels; the relational configuration keeps each pair as an object that routes messages and later supports relational readouts. Compact, batch-matched controls isolate forecast routes (Table 6), while Appendix E tests computation on the retained relationships.

Table 5: Staged PEMS comparison (4-horizon mean MSE, 3 seeds). iTf: iTransformer; ID: learned identity; Org.: consensus/deviation organization; Rel.: relational configuration. Each Δ is the adjacent-model ratio from unrounded means minus one.

Dataset	iTf	iTf+ID	First-order	ENF	ID Δ	Org. Δ	Rel. Δ
PEMS03	0.123	0.109	0.110	0.107	-11.2%	$+1.2\%$	-3.2%
PEMS04	0.124	0.101	0.101	0.099	-18.7%	$+0.6\%$	-2.1%
PEMS07	0.106	0.096	0.094	0.089	-9.8%	-2.0%	-5.6%
PEMS08	0.167	0.150	0.141	0.134	-10.3%	-6.1%	-4.6%
Avg	0.130	0.114	0.112	0.107	-12.4%	-2.0%	-3.9%

Figure 4 displays the staged $d128/L2$ comparison, averaging four horizons within each seed. Table 6 varies individual ENF choices at $d64/L1$ with three seeds and batch 32, including a batch-matched Electricity control rather than the batch-16 model in Table 10.

Removing the additive $g/r/c$ route increases mean MSE on PEMS04, PEMS08, and Electricity by 0.8%, 1.3%, and 0.5%, respectively. Endpoint components therefore supply useful predictive updates. Interaction-value ablations quantify their contribution to forecasting, while frozen-field probes separately test what can be computed from the retained interaction object (Appendices E and G). Together these controls connect the decomposition to distinct predictive and analytical operations.

The default interaction field is symmetric (Lemma 2), so Table 6 tests its active message and endpoint routes. A zero-output control that disables the algebraically zero antisymmetric branch gives independently trained MSE changes of -0.9%, $+1.5\%$, and $+0.0\%$ on PEMS04, PEMS08, and Electricity; these are run-to-run differences, not a directional-feature ablation. Directional content is evaluated with spectral coherence in Appendix F.1.

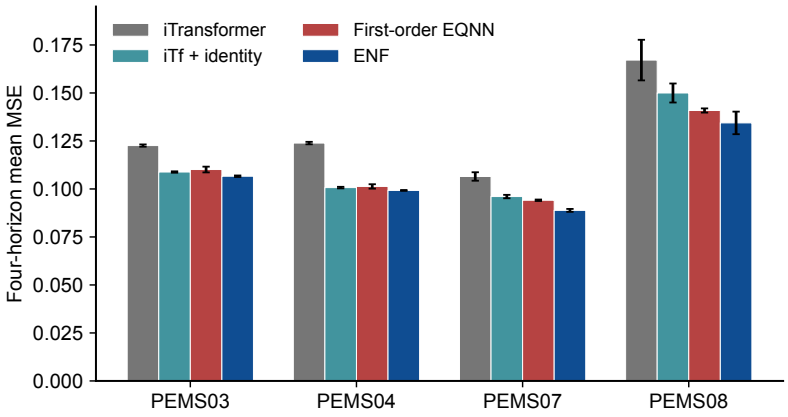


Figure 4: Matched PEMS configurations, $d128/L2$, three seeds and four horizons. Bars show mean MSE; error bars show the sample SD across the three seed-specific horizon averages. Each configuration is trained independently, as detailed in Table 5.

C.2 IDENTITY AND RELATIONAL REPRESENTATION

The five-seed factorial comparison crosses persistent variable meta-representations with the pair-field configuration (Table 7). Each first-order model already exchanges information across variables. ID conditions those exchanges on a stable variable representation, while ENF additionally constructs input-dependent relation objects. On Solar, PEMS04, and PEMS07, the pair configuration gives lower

Table 6: Component comparisons against full-ENF controls ($d64/L1$, 3 seeds, batch 32). Entries are MSE(modified)/MSE(full)-1; positive means higher error. Electricity uses a batch-matched control. The symmetric content field has an identically zero antisymmetric branch.

Ablation	PEMS04	PEMS08	Electricity
Identity disabled	$+12.5\%$	$+8.0\%$	$+5.2\%$
Unprojected updates	-1.7%	$+0.4\%$	-0.3%
Spectral-coherence features	-0.1%	-0.7%	$+0.4\%$
$d_r = 4$ (vs 8)	-0.7%	$+2.5\%$	$+0.1\%$
$d_r = 16$ (vs 8)	-0.4%	$+2.1\%$	-0.5%
No interaction-value message	-0.8%	$+1.6\%$	-0.1%
no $g/r/c$ routing	$+0.8\%$	$+1.3\%$	$+0.5\%$
zero symmetric $R^{\perp}$	-0.8%	-0.1%	-0.1%

MSE with ID than its first-order counterpart. Weather also gives a pair-configuration improvement without ID, illustrating that histories alone can support useful relations. Figure 3b reports all seven dataset contrasts under the same convention.

Table 7: Identity–relation factorial comparison ($d64/L1$, batch 32, five seeds). Entries are test MSE $\pm$ sample SD. ID is the learned variable-identity table; no ID freezes it at zero. Each architecture retains its deployed dropout placement. Horizons are 12 on PEMS and 96 elsewhere.

	First-order EQNN		ENF	
Dataset	No ID	ID	No ID	ID
ETTh2	$0.30425 \pm .00296$	$0.31091 \pm .00284$	$0.30496 \pm .00317$	$0.31098 \pm .00610$
Weather	$0.16567 \pm .00148$	$0.15886 \pm .00125$	$0.16418 \pm .00124$	$0.15850 \pm .00125$
Solar	$0.19827 \pm .00332$	$0.19930 \pm .00451$	$0.20061 \pm .01094$	$0.18783 \pm .01319$
PEMS04	$0.08300 \pm .00099$	$0.07442 \pm .00130$	$0.08477 \pm .00132$	$0.07299 \pm .00157$
PEMS07	$0.06659 \pm .00098$	$0.06032 \pm .00067$	$0.06681 \pm .00034$	$0.05831 \pm .00036$
Electricity	$0.15264 \pm .00098$	$0.14381 \pm .00094$	$0.15337 \pm .00083$	$0.14504 \pm .00026$
Traffic	$0.44412 \pm .00196$	$0.45216 \pm .00223$	$0.45046 \pm .00228$	$0.45417 \pm .00259$

The identity-removal test complements the staged identity addition. Disabling and freezing ENF's identity field removes its attention biases and identity-conditioned FFN modulation. Table 8 compares this configuration with full ENF on five high-channel benchmarks at $d64/L1$ with three seeds. Unlike inference-time relabeling, this comparison removes identity information during training.

Table 8: Identity ablation ($d64/L1$, 3-seed mean MSE). Identity is a first-order channel property; Δ =MSE(no-ID)/MSE(full)-1, so positive values indicate increased error after disabling it. PEMS07 also differs in batch size (see below).

Dataset	C	ENF (full)	ENF (no-id)	Δ
PEMS07	883	0.05870	0.06638	$+13.1\%$
PEMS04	307	0.07404	0.08330	$+12.5\%$
PEMS03	358	0.06450	0.06962	$+8.0\%$
PEMS08	170	0.07366	0.07953	$+8.0\%$
Electricity	321	0.14366	0.15107	$+5.2\%$

Disabling identity increases mean MSE by 5.2–12.5% on the four datasets with batch-matched controls (batch 32). PEMS07 is excluded from this summary: its full model uses batch 16 while the no-identity control uses batch 32, so that row varies both factors. Table 6 collects the batch-matched PEMS04, PEMS08, and Electricity controls. The results support learned identity as a useful part of ENF, consistent with its use in other traffic forecasters (Shao et al., 2022; Liu et al., 2023).

C.3 Channel-count sweep against forecasting baselines

The channel-count sweep tests each architecture on the same nested PEMS07 subsets (Table 9, three seeds, compact channel-attention backbones). ENF has lower mean MSE than iTransformer at every C. DLinear, FITS, and MixLinear have higher errors throughout and change less beyond $C = 64$. Each row is a within-subset comparison; successive rows also change which channels are predicted.

Table 9: C-scaling on PEMS07 (3-seed mean MSE). ENF has the lowest mean error at every channel count; the margin over channel attention is largest beyond $C = 16$.

C	iTransformer	ENF	DLinear	FITS	MixLinear
8	0.10401	0.10249	0.16050	0.16093	0.27207
16	0.07340	0.06646	0.11725	0.11743	0.20502
64	0.06904	0.05867	0.11407	0.11427	0.20184
256	0.07197	0.06257	0.11154	0.11165	0.19690
883	0.06935	0.05870	0.10896	0.10905	0.19046

Table 10: Compact full-channel comparison ($d64/L1$; input 96). PEMS uses horizon 12; others use 96. All entries report mean test MSE±sample SD; n is the seed count for each model. ΔMSE is ENF/iTransformer minus one. Appendix B.1 specifies configurations and seed coverage.

Dataset	C	n	iTransformer	ENF	ΔMSE
Low-channel datasets ($C < 16$)					
ETTh1	7	3	0.38735±.0028	0.38757±.0011	+0.1%
ETTh2	7	3	0.29788±.0017	0.31033±.0009	+4.2%
ETTm1	7	3	0.32752±.0028	0.32902±.0036	+0.5%
ETTm2	7	3	0.18246±.0022	0.18026±.0021	−1.2%
Exchange	8	3	0.09245±.0045	0.09090±.0036	−1.7%
Higher-channel datasets ($C \geq 16$)					
Weather	21	3	0.16524±.0020	0.15779±.0006	−4.5%
Solar	137	3	0.19482±.0052	0.19704±.0027	+1.1%
Traffic	862	3	0.45008±.0009	0.45199±.0007	+0.4%
Electricity	321	3	0.15205±.0005	0.14292±.0004	−6.0%
PEMS03	358	5	0.07210±.0023	0.06429±.0007	−10.8%
PEMS04	307	5	0.08623±.0005	0.07391±.0005	−14.3%
PEMS07	883	3	0.06935±.0003	0.05870±.0004	−15.4%
PEMS08	170	5	0.08293±.0010	0.07446±.0016	−10.2%

C.4 Performance across diverse domains and channel scales.

To assess generalizability beyond traffic networks, Table 10 compares the compact model ($d64/L1$) across 13 diverse benchmarks. The compact ENF configuration reduces mean MSE by 10.2–15.4% on PEMS, 6.0% on Electricity, and 4.5% on Weather. The table evaluates the complete architectures across all 13 datasets. The five-seed factorial comparison (Table 7) separately identifies how identity and pair representations contribute within seven of these domains.

C.5 Channel-count sweep: relational and first-order configurations

Table 11 isolates the pair-field configuration while retaining identity and consensus/deviation organization in both models. Each comparison uses the same nested channel subset and three seeds. On PEMS07, the ENF advantage grows from -0.6% at $C = 8$ to 3.2% at $C = 883$. On PEMS03, ENF improves mean MSE at every tested count, with gains from 0.4% to 2.3%. These within-subset comparisons establish how the pair configuration contributes across channel populations.

Table 11: Relational versus first-order configurations (3-seed means). Gain is first-order minus ENF, over first-order; positive values indicate lower error for ENF.

PEMS07				PEMS03			
C	first-order	ENF	gain	C	first-order	ENF	gain
8	0.10183	0.10249	−0.6%	8	0.04181	0.04085	+2.3%
16	0.06710	0.06646	+0.9%	16	0.07485	0.07378	+1.4%
64	0.05952	0.05867	+1.4%	64	0.06854	0.06826	+0.4%
256	0.06411	0.06257	+2.4%	128	0.06910	0.06843	+1.0%
883	0.06061	0.05870	+3.2%	256	0.06510	0.06374	+2.1%
				358	0.06517	0.06450	+1.0%

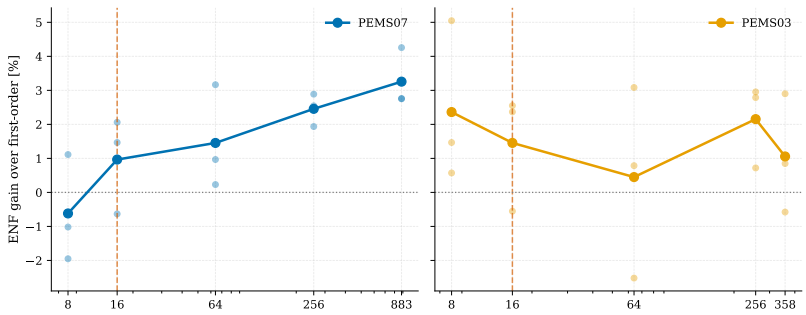


Figure 5: Relative MSE change (first-order minus ENF, over first-order) across nested channel subsets, three seeds. Positive favors ENF. The vertical line marks $C = 16$ as a channel-count reference. Table 11 gives the corresponding within-subset comparisons.

C.6 Full width and depth sweep

Increasing backbone width and depth tests whether ENF's predictive advantage persists across model sizes (Table 12, PEMS08). ENF has lower mean MSE at all four shared sizes, with the largest margin at $d64/L1$ and a 0.5% margin at $d256/L3$. Comparisons match width and depth with $d_{ff} = 2d$; parameter counts and resource use are characterized in Appendix C.7. The $d256/L3$ cells use common seeds $\{3, 4\}$, and the $d512/L2$ entry reports the additional iTransformer configuration.

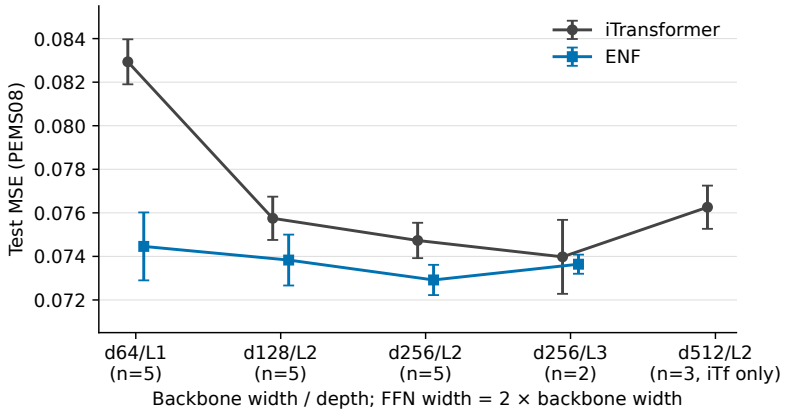


Figure 6: PEMS08 width/depth sweep. ENF improves on iTransformer at every shared size, with the largest margin in the compact regime. Points show means and sample SD; seed counts and FFN widths follow Table 12. Shared configurations extend to $d256/L3$, with an additional $d512/L2$ iTransformer point.

Table 12: Model-capacity scaling on PEMS08 (mean MSE; n per cell, $d_{ff} = 2d$ throughout).

Model	d64/L1	d128/L2	d256/L2	d256/L3	d512/L2
iTransformer	0.08293_{n5}	0.07575_{n5}	0.07473_{n5}	0.07398_{n2}	0.07626_{n3}
ENF	0.07446_{n5}	0.07383_{n5}	0.07292_{n5}	0.07364_{n2}	–
Δ	-10.2%	-2.5%	-2.4%	-0.5%	–

C.7 Resource cost

Exact factorization evaluates the same deterministic content update without storing a vector for every channel pair. Table 13 gives the dense implementation's resource profile against iTransformer; Table 14 measures the reduction in training-step cost from factorization. In the dense profile, ENF uses approximately $2.3\times$ the parameters, 3–$11\times$ the peak forward memory, and 3.4–$7.1\times$ the forward latency over the tested channel counts. The factorization below identifies the pair-feature and message tensors responsible for the removable activation cost.

Table 13: Resource measurements at $d = 128/L = 2$: batch 1, bf16, one forward pass. Pairwise activations scale with C^2, while learned feature projections are shared across channel pairs.

	Params (M)		Peak memory (MB)		Forward (ms)	
C	iTransformer	ENF	iTransformer	ENF	iTransformer	ENF
170	0.28	0.65	13	39	1.0	3.4
307	0.28	0.65	17	99	1.0	3.4
883	0.28	0.66	62	705	1.0	7.1

C.8 Exact factorization of the content computation

The default content field need not be materialized to compute its routed updates. Let $z_i = W_h h_i + b_h$, $\bar{z} = C^{-1} \sum_i z_i$, and $\tilde{z}_i = z_i - \bar{z}$. Averaging its additive and product terms gives

$$\begin{aligned} g &= (W_a + W_b)\bar{z} + W_\times(\bar{z} \odot \bar{z}) + b_R, \\ r_i &= W_a \tilde{z}_i + W_\times(\tilde{z}_i \odot \bar{z}), \\ c_i &= W_b \tilde{z}_i + W_\times(\bar{z} \odot \tilde{z}_i). \end{aligned}$$

For each head, the interaction contribution, including its shared bias, is

$$u_i^{\mathrm{rel}} = W_m \left(\tilde{z}_i \odot \sum_j A_{ij} \tilde{z}_j \right) + b_{\mathrm{pair}} \sum_j A_{ij}.$$

Together with Proposition 1, these expressions evaluate both attention patterns A_0 and A_1 with the original identity-biased update, $g/r/c$ routes, FFNs, and decoder. Ordinary attention still stores a $C \times C$ weight matrix, but the $C \times C \times d_r$ interaction field and the pair-value tensor are unnecessary for forecasting. The same formula still defines pair features for subsequent analysis.

Dense and factorized content models agree numerically with nonzero route weights and identity gates. With dropout zero, maximum output/parameter-gradient errors are $5.55 \times 10^{-16}/3.33 \times 10^{-16}$ in double precision and $1.49 \times 10^{-7}/1.49 \times 10^{-7}$ in single precision. Three paired Adam steps give output discrepancy 2.68×10^{-7}; evaluation of all 21 trained full-model checkpoints gives a maximum discrepancy of 3.58×10^{-7}. These checks establish deterministic equivalence. The benchmark training recipe additionally uses independent pair-message dropout, so its stochastic updates differ from the dropout-free factorized profile.

Table 14: Dense and factorized content computation: median CPU training-step time in milliseconds with dropout zero, four threads, batch eight, input length 32, and $d64/L1$. Each measured step includes forward propagation, backward propagation, and Adam. Three warmup steps precede ten measured steps.

Channels C	Dense (ms)	Factorized (ms)
16	8.33	7.19
64	37.34	13.64
256	934.88	62.91

C.9 Forecasting with shorter recent training histories

We retain the most recent N training windows, using fractions $\{10\%, 25\%, 50\%, 100\%\}$, and evaluate all models on the same test set. Each entry averages three seeds. Because the retained tail changes both sample count and temporal coverage, these results concern training on a shorter recent history.

Increasing recent training coverage improves both architectures. ENF's relative advantage grows from 0.9% to 10.1% on PEMS08 and from 1.4% to 5.5% on Electricity between the shortest and full histories, with Electricity's largest gap at 50%. The advantage persists with more history; tail-N couples sample count to temporal coverage.

C.10 Low-channel ETT comparisons

The ETT datasets evaluate forecasting with seven measured variables, complementing the larger channel populations in the traffic and energy studies. Table 10 supplies the matched compact comparison.

Table 15: Recent-history training subsets (tail-N, 3-seed mean test MSE). ENF has lower point estimates at all fractions. Its relative advantage is larger at 25–100% than at 10%, with a non-monotone pattern on Electricity.

Dataset	Model	10%	25%	50%	100%
PEMS08	iTransformer	0.10653	0.09405	0.08724	0.08319
	ENF	0.10559	0.08591	0.07849	0.07476
	Δ	-0.9%	-8.7%	-10.0%	-10.1%
Electricity	iTransformer	0.19151	0.16599	0.15748	0.15205
	ENF	0.18878	0.15598	0.14715	0.14366
	Δ	-1.4%	-6.0%	-6.6%	-5.5%

The compact comparison gives ETT relative MSE changes from -1.2% to $+4.2\%$. Table 16 additionally reports larger ENF configurations and published iTransformer reference values. Its final column compares those larger models with the published references; the common-protocol comparison remains Table 10.

Table 16: ETT at horizon 96. Compact and larger ENF configurations are shown beside published iTransformer reference values (Liu et al., 2024). The last column compares the larger ENF entry with that reference; the matched compact comparison is Table 10.

Dataset	ENF (d64/L1)	ENF (larger)	iTf (published)	Δ (larger/ref.)
ETTh1	0.38757	0.38622 (d128/L4)	0.384	$+0.6\%$
ETTh2	0.31033	0.30949 (d128/L2)	0.302	$+2.5\%$
ETTm1	0.32902	0.33111 (d256/L2)	0.337	-1.7%
ETTm2	0.18026	0.17935 (d128/L2)	0.184	-2.5%

These comparisons characterize complete-model prediction in low-channel settings. Table 7 provides the separate identity–relation controls on ETTh2.

D Predictive use and readout of controlled interactions

A relational forecaster supports two testable operations: using cross-variable information for prediction and exposing relations for subsequent readout. We evaluate route usage by varying nonlinear coupling and measuring mixed responses, then evaluate relational information on processes with identical univariate laws. The studies comprise 72 interaction-strength/sample-size fits and 12 signed-community fits, using seeds $\{1, 2, 42\}$ and four model configurations.

Matched model configurations. All runs use the ENF content backbone with $C = 16$, $d = 64$, one layer, four heads, $d_r = d_e = 16$, and FFN width 128. Full ENF is compared with interaction-off, which zeros the feature-dependent input to the pair-message map while retaining its shared bias, and pair-off, which zeros both the $g/r/c$ and interaction-message routes. These configurations preserve modules, initialization, and minibatch order, isolating the active routes. A wide first-order control uses pair-off with FFN width 256 and its own wider-layer initialization. Disabled paths retain dormant parameters; resource comparisons instead use the implementations in Appendix C.7.

The simulator units are retained without RevIN or estimated per-channel scaling. Every configuration has dropout zero. Adam uses learning rate 5×10^{-4}, zero weight decay, batch 64, gradient-norm clipping at one, and cosine decay over 40 epochs, with early stopping after eight epochs without validation improvement. The validation-MSE checkpoint is used for all analyses. Fits run on the same CPU with four threads in single precision. Protocols and seed sets are shared within each comparison, independently of the observed test results.

D.1 Interaction strength and sample size

For a channel state x_t, define $z_t = \tanh(x_t)$ elementwise and generate

$$x_{t+1} = 0.5z_t + 0.35Wz_t + \lambda\, z_t \odot (Kz_t) + \epsilon_t, \qquad \epsilon_t \sim \mathcal{N}(0, 0.25I).$$

Both W and K have signed weights on the four ring neighbors at offsets $\{-2, -1, 1, 2\}$, zero diagonal, and unit row ℓ_1 norm. The seed fixes these matrices, initial states, and innovation draws

across strengths $\lambda \in \{0, 0.4, 0.8\}$. The deterministic drift is bounded. At zero strength, each output is additive in the current channel values; nonzero strength introduces target–source products. Innovation variance stays fixed, while total signal variance may change.

Each example is an independently simulated 32-step history with a one-step target. The 256-example training set is a prefix of the 1,024-example set; validation and test use 256 and 512 separate histories. Increasing training size reduces error for every configuration (Table 17). Full ENF and interaction-off give nearly identical forecast errors across strengths. With 1,024 examples, the wider first-order control gives the lowest means. These controls show that cross-variable nonlinear forecasting in this process can be learned through first-order updates; stronger generator coupling alone does not establish a larger contribution from the explicit interaction-message route.

Table 17: Controlled nonlinear forecasting: test MSE $\pm$ sample SD over three seeds. Int.-off retains the shared pair-message bias; pair-off removes all pair-routing outputs; wide uses pair-off with wider FFNs. Noise variance is fixed across strengths.

Train n	λ	Full ENF	Int.-off	Pair-off	Wide first-order
256	0.0	$0.265105 \pm .002378$	$0.265097 \pm .002372$	$0.265899 \pm .002720$	$0.265444 \pm .001116$
256	0.4	$0.267367 \pm .001465$	$0.267356 \pm .001468$	$0.268245 \pm .002131$	$0.267516 \pm .000690$
256	0.8	$0.275095 \pm .000648$	$0.275090 \pm .000647$	$0.275873 \pm .001313$	$0.274987 \pm .000989$
1024	0.0	$0.259631 \pm .001657$	$0.259622 \pm .001653$	$0.260088 \pm .001871$	$0.259302 \pm .001558$
1024	0.4	$0.261896 \pm .001097$	$0.261894 \pm .001095$	$0.262277 \pm .001180$	$0.261366 \pm .001125$
1024	0.8	$0.269694 \pm .000884$	$0.269681 \pm .000863$	$0.269854 \pm .000298$	$0.268864 \pm .000559$

D.2 Four-condition finite differences

We hold each test history fixed except for its last observed values on channels 0 and 1. Setting those two entries to $(+a, +a), (+a, -a), (-a, +a), (-a, -a)$ gives four deterministic conditional means for target 0. Their mixed contrast is

$$\Delta = f_{++} - f_{+-} - f_{-+} + f_{--} = 4\lambda K_{01} \tanh^2(a).$$

The additive generator has zero contrast. We evaluate $a \in \{0.5, 1.5\}$ on the first 128 test histories for every seed, without adding target noise. These prescribed interventions measure mixed responses at two amplitudes; training does not exclude any particular sign combination.

Retraining and frozen-checkpoint interventions agree on the interaction route's predictive contribution in this process. Full ENF and interaction-off give nearly identical mixed-contrast RMSE at both amplitudes (Table 18); disabling the feature-dependent route in a trained checkpoint changes condition-averaged test MSE by at most 1.14×10^{-5}. All four models underestimate the true contrast at nonzero strength. Thus this forecasting objective yields a small measured contribution from interaction messages. The next experiment asks the complementary question: which relational information remains accessible in the learned states and pair fields?

Table 18: Four-condition contrast RMSE (three-seed mean). Each cell lists amplitudes $a = 0.5 \,/\, 1.5$ in that order. The simulator's noise-free conditional mean defines the true contrast.

Train n	λ	Full ENF	Int.-off	Pair-off	Wide first-order
256	0.0	.0033 / .0189	.0033 / .0189	.0034 / .0197	.0041 / .0232
256	0.4	.0776 / .2988	.0776 / .2987	.0776 / .2986	.0785 / .3049
256	0.8	.1546 / .5941	.1546 / .5941	.1545 / .5934	.1554 / .5997
1024	0.0	.0029 / .0163	.0029 / .0163	.0028 / .0158	.0033 / .0188
1024	0.4	.0779 / .3002	.0779 / .3002	.0778 / .3002	.0786 / .3058
1024	0.8	.1550 / .5964	.1550 / .5964	.1548 / .5954	.1555 / .6006

D.3 Relations with identical univariate process laws

To test relational information separately from route usage, the signed-community process fixes every channel's marginal dynamics while varying its joint correlations. A probe must then use information shared across channels rather than a distinctive univariate process law.

Balanced signs $s_i \in \{-1, 1\}$ define $U = ss^\top / C$ and the stationary Gaussian process

$$x_{t+1} = (\rho I + \kappa U)x_t + \epsilon_t, \qquad \rho = 0.2, \quad \kappa = 0.65, \quad \epsilon_t \sim \mathcal{N}(0, 0.25I).$$

Here $U^2 = U$, and the two transition eigenvalues are 0.2 and 0.85. Its covariance is $\Sigma = v_0 I + (v_1 - v_0)U$, where $v_0 = 0.25/(1-\rho^2)$ and $v_1 = 0.25/[1-(\rho+\kappa)^2]$. Since every $U_{ii} = 1/C$, channel variances and all univariate lag covariances agree across channels and sign assignments. The joint correlations, however, distinguish the two sign communities. All off-diagonal couplings are nonzero, so the target is common community membership rather than nonzero edge support.

Each seed provides 512 training graphs, 128 validation graphs, and 256 test graphs, with independent stationary trajectories, 96-step inputs, and one-step targets. Graphs are distinct across splits, including the equivalence of s and $-s$. Forecasting is trained without graph labels. We freeze the model, fit supervised same-community probes on training graphs, and select their hyperparameters or checkpoints on validation graphs.

The pair inputs are taken at the actual input to the layer's content construction. Ridge probes use $R^{\perp}_{ij}$, the additive field $g + r_i + c_j$, or the centered product $\tilde{z}_i \odot \tilde{z}_j$. Features are standardized with training-graph statistics, and ridge penalties $\{10^{-3}, 10^{-2}, 10^{-1}, 1\}$ are selected by validation squared label error. A node-pair decoder receives $[h_i + h_j, |h_i - h_j|, h_i \odot h_j]$ from that same layer input, using a 64-unit hidden layer, GELU, and a scalar output. Training uses 200 Adam steps at rate 10^{-3}, 2,048-pair batches, and validation selection every 20 steps. Raw signed correlation supplies the parameter-free baseline.

We compute ROC AUC over the 120 upper-triangle pairs separately for each graph, average over graphs, and then average the three model seeds. For clustering, we center each symmetric score matrix and threshold the sign of the eigenvector with largest absolute eigenvalue, reporting adjusted Rand index (ARI). This rule is fixed for all readouts; no test-dependent sign or layer selection is used.

Table 19: Signed-community recovery from the full ENF checkpoint: three-seed mean ± sample SD of graph-wise metrics. All readouts are probes on frozen first-order states or on the pair object, supervised on separate training graphs. Forecast test MSE is 0.276116 ± 0.008514.

Readout	AUC ↑	ARI ↑
Interaction $R^{\perp}$	$0.7501 \pm .0090$	$0.4335 \pm .0326$
Additive field	$0.5016 \pm .0062$	$0.0187 \pm .0140$
Centered feature product	$0.7528 \pm .0055$	$0.4417 \pm .0132$
Same-state node-pair MLP	$0.8937 \pm .0069$	$0.8511 \pm .0074$
Raw signed correlation	$0.8903 \pm .0091$	$0.8974 \pm .0216$

The interaction field exposes community membership to a linear readout, with AUC 0.7501 versus 0.5016 for the additive field. Centered feature products give similar performance and reconstruct $R^{\perp}$ within 6.2×10^{-7} in single precision, matching Lemma 2's algebraic form. The same-state nonlinear node-pair decoder and raw correlation give higher recovery scores (Table 19). These comparisons locate relational information in both node states and interaction features. Pair indexing additionally makes a common variable available as an explicit composition index, the operation evaluated in Appendix E.

Table 20: Signed-community route controls (three-seed means). Forecasting and frozen-field AUCs are nearly unchanged by disabling interaction messages. Pair-off and wide retain an unused pair lift; its readout evaluates a feature construction from learned node states. Node-pair MLPs use the same layer input as each pair lift.

Configuration	Forecast MSE	$R^{\perp}$ AUC	Additive AUC	Node-pair AUC
Full ENF	0.276116	0.7501	0.5016	0.8937
Interaction-off	0.276103	0.7501	0.5016	0.8937
Pair-off	0.276728	0.7526	0.5016	0.8938
Wide first-order	0.277341	0.7346	0.5026	0.8931

E PAIR CAPTURE AND COMPOSITION WITH OBSERVABLE INTERMEDIATES

Composition operates on relations sharing an intermediate to answer a query about their endpoints. These tasks hold query-endpoint histories fixed within positive/negative sequence pairs and place distinguishing evidence in observed intermediates. Basic-link supervision establishes acquisition; frozen-field tests then locate the links and combine them through common nodes. The encoder uses

the two-layer ENF content architecture with zero identity embeddings and the auxiliary supervision below. This protocol tests relations between relations while retaining the joint permutation action of the architecture.

E.1 Observable intermediate-node tasks

Each sequence has $C = 16$ channels and $T = 256$ observations. Query endpoints occupy positions 0 and 1. The remaining 14 channels are randomly permuted within each positive/negative pair, with the same permutation for both sequences. Four of these channels provide intermediate-node evidence. Endpoints have identical histories across the pair; intermediates distinguish their relation labels.

For Signed, let u_0, u_1 and ϵ_k be independent standard Gaussian innovation sequences. Four intermediates have innovations

$$v_k = \frac{(a_k u_0 + b_k u_1)/\sqrt{2} + 0.25\epsilon_k}{\sqrt{1 + 0.25^2}}, \qquad (a_1, a_2, a_3, a_4) \in \mathrm{Perm}(-1, -1, 1, 1).$$

Set $b_k = a_k$ for a positive sequence and $b_k = -a_k$ for its negative twin. The label is the sign of $\sum_k a_k b_k$. Both endpoints have balanced positive and negative incident signs, so their separate sign counts do not determine this label. Eight endpoint–intermediate sign labels supply the basic relation task.

For the common-neighbor task, four independent roots have innovations $u_0, \ldots, u_3$. Each of four intermediates observes the normalized noisy sum of two roots, using the same mixing and noise scales as above. Positive sequences use root pairs $(0, 1), (0, 1), (2, 3), (2, 3)$; negative sequences use $(0, 2), (0, 3), (1, 2), (1, 3)$. Every root and intermediate has degree two, but query roots 0 and 1 share two intermediates only in the positive class. All 16 root–intermediate edge/nonedge labels supply the basic task. Appendix F.1 instead studies whole-ring VAR topology.

Background channels have independent Gaussian innovations. Applying the same stationary AR(0.2) filter to every channel gives identical zero-mean, unit-variance univariate Gaussian process laws across roles and classes. The input contains all histories, without edge labels or intermediate-node identities. Labels select training losses and evaluation queries. Independent sequence, noise, and permutation draws form the splits; the motif definitions are shared across splits.

E.2 Basic acquisition and intermediate visibility

The backbone uses width 64, two layers, four heads, FFN width 128, relation width 16, dropout zero, frozen-zero identity fields, and no RevIN. Two symmetric heads map $[h_i + h_j, |h_i - h_j|, h_i \odot h_j]$ through a 64-unit hidden layer to a binary logit. One predicts basic relations; the other predicts the query from contextualized endpoint states. This query head tests what the encoder's node outputs use, before introducing a readout that explicitly composes pair fields.

Training first establishes basic-link representations, then jointly optimizes basic and query predictions. Stage A minimizes basic-link binary cross-entropy (BCE), selects by validation BCE, and uses validation AUC ≥ 0.90 as its acquisition criterion. Stage B initializes from this checkpoint and minimizes the equally weighted sum of basic and query BCE, selecting by the summed validation loss. Both stages use Adam at 5×10^{-4}, batch 64, zero weight decay, gradient clipping at one, a 40-epoch cosine schedule, at most 40 epochs, and patience eight. Training, validation, and test sets contain 1,024/256/512 sequences per seed, or 512/128/256 positive/negative pairs.

Basic-link predictions are accurate in both stages across seeds $\{1, 2, 42\}$. Stage-A test AUC is 1.000000 ± 0.000000 on Signed and 0.999863 ± 0.000022 on common-neighbor edges; stage-B values are 0.999609 ± 0.000499 and 0.999882 ± 0.000033. All three seeds satisfy the acquisition criterion. These results establish the availability of basic relations for the query and composition evaluations.

We then freeze each stage-B model and retain 4, 2, 1, or 0 intermediates in a nested random order. Removed histories are replaced by independent marginal-matched AR noise, shared by the paired sequences. No internal message is masked. At zero visibility the entire positive and negative inputs coincide, giving AUC 0.5 for any deterministic predictor. Table 21 reports the original query head, separately from the field readouts below.

E.3 Matched controls of the actual pair field

The field controls test whether composition depends on centered interaction and correct node correspondence. From the six frozen stage-B checkpoints, we extract each layer's actual pair projection R

Table 21: Intermediate visibility for the frozen ENF query head: AUC, mean±SD over three seeds. Removed histories are replaced by marginal-matched noise. At zero visibility, paired inputs coincide. Pair-field composition is evaluated separately in Table 23.

Task	4 visible	2 visible	1 visible	0 visible
Signed	$.9221 \pm .0204$	$.7877 \pm .0285$	$.6617 \pm .0207$	$.5000 \pm .0000$
Common neighbor	$.5025 \pm .0018$	$.5011 \pm .0022$	$.4988 \pm .0012$	$.5000 \pm .0000$

and decompose it as $L_{ij} = g + r_i + c_j$ and $R^{\perp} = R - L$. The four conditions are R, L, $R^{\perp}$, and $PR^{\perp}P^{\top}$, scored against the original node labels. A uniformly random P is drawn independently for each positive/negative pair and shared by its sequences and layers, independently of labels and hidden roles. This conjugation preserves field entries, norms, zero row/column sums, and global structure while changing which features correspond to the queried nodes.

All targets are undirected, so every field F is symmetrized as $S_{ij} = (F_{ij} + F_{ji})/2$. Basic relation recovery uses the 16-dimensional vector S_{ij}. Composition pairs the relations through the same intermediate:

$$Z_{ij} = \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} S_{ik} S_{jk}^{\top}, \qquad \mathcal{K} = \{0, \ldots, 15\} \setminus \{i, j\}.$$

The common-intermediate transform is applied to every field and uses all 14 nonquery nodes, without intermediate-role information or predicted edge labels. We symmetrize Z_{ij} and pack its upper triangle, scaling off-diagonal entries by $\sqrt{2}$, into 136 features. This frozen-field composition readout supplements the encoder's contextualized-node query head.

Each condition uses training-only feature standardization and logistic regression with inverse regularization $C = 1$, the L-BFGS solver, and at most 3,000 iterations. Features with training standard deviation at most 10^{-6} are excluded. There is no hyperparameter search. Basic heads fit training edge labels; composition heads fit training query labels. The nominal feature dimensions and fitting settings match across all four fields within each objective.

Readouts are fitted using the model-training sequences, with the original validation and probe-test sequences used for development. Final evaluation uses 512 independently generated sequences (256 pairs) per task and seed, generated after all readouts are fixed. Both layers are reported. The design comprises $6 \times 2 \times 4 \times 2 = 96$ fitted linear readouts on frozen backbones, with uncertainty reported as the three-seed sample SD.

E.4 Pair capture and compositional accessibility

Table 22 locates basic relational information in the pair residual. Signed relation AUC remains above 0.999 after removing the additive field; common-neighbor edge AUC reaches 0.988–0.989, while its additive readout is at chance. Misaligning pair correspondence brings both tasks close to chance.

Table 22: Basic relation recovery from frozen ENF fields on the fresh test sequences: AUC ± SD across three seeds. All conditions use matched 16-feature linear readouts. $L = g + r + c$; shuffled refers to the node-misaligned residual.

Task	Layer	R	L	$R^{\perp}$	Shuffled $R^{\perp}$
Signed	1	$.9999 \pm .0000$	$.6976 \pm .0224$	$.9999 \pm .0000$	$.4953 \pm .0046$
Signed	2	$.9991 \pm .0005$	$.6816 \pm .0139$	$.9994 \pm .0003$	$.4963 \pm .0095$
Common neighbor	1	$.8905 \pm .0092$	$.5000 \pm .0000$	$.9892 \pm .0005$	$.5037 \pm .0096$
Common neighbor	2	$.8925 \pm .0445$	$.5000 \pm .0000$	$.9882 \pm .0018$	$.5038 \pm .0026$

Double-centering improves compositional access under the matched common-intermediate readout. Residual AUCs are 0.9986/0.9886 on Signed and 0.9146/0.9051 on common-neighbor queries at layers 1/2 (Table 23). On common-neighbor queries, the full-field scores are 0.7106/0.7182 and the additive scores are 0.5024/0.5058. Shuffling node correspondence brings the residual readouts to chance. The contrast identifies centered interaction content and correct shared-node alignment as the ingredients used by the composition readout.

The contextualized-node query head gives common-neighbor AUC 0.5025 with all intermediates visible (Table 21), whereas the frozen-field transform exposes the query relation on the independent

composition test set. The two evaluations use their specified readout architectures and training procedures. Together they distinguish learning basic relations from accessing their shared-intermediate composition, and connect the second-order representation to a concrete operation on relationships.

Table 23: Relational composition from frozen fields: AUC ± sample SD over three seeds on fresh test sequences. All fields use the same common-intermediate transform and logistic readout. $L = g+r+c$; shuffled residuals preserve field structure but misalign node correspondence with query labels.

Task	Layer	R	L	$R^{\perp}$	Shuffled $R^{\perp}$
Signed	1	$.9963 \pm .0021$	$.5320 \pm .0045$	$.9986 \pm .0006$	$.5012 \pm .0267$
Signed	2	$.9825 \pm .0077$	$.7180 \pm .0067$	$.9886 \pm .0017$	$.5127 \pm .0325$
Common neighbor	1	$.7106 \pm .0240$	$.5024 \pm .0015$	$.9146 \pm .0079$	$.4991 \pm .0208$
Common neighbor	2	$.7182 \pm .0815$	$.5058 \pm .0058$	$.9051 \pm .0111$	$.4950 \pm .0083$

F TOPOLOGY READOUTS FROM EQUIVARIANT PAIR FIELDS

Known simulated couplings provide reference targets for evaluating pair-field readouts. The experiments compare interaction and additive scores on ring and community structures, then test directional features supplied by spectral coherence.

F.1 SYNTHETIC TOPOLOGY READOUTS

Known coupling matrices let the same scalar readout be tested on local ring edges and two-community membership. This study asks which structures are exposed by averaged field scores, complementing the supervised, same-marginal probes of Appendix D.3.

Simulation and readout. The first-order vector autoregressive process, VAR(1), generates each channel vector from the previous one:

$$x(t) = Ax(t-1) + \varepsilon(t), \qquad \varepsilon(t) \sim \mathcal{N}(0, 0.1^2 I).$$

Here A is the coupling matrix and ε independent Gaussian noise. Weights start at scale 0.8 and are rescaled to keep the spectral radius, the largest absolute eigenvalue, at most 0.95. Each series has 5,000 time points, with 96-step inputs and one-step targets. The multiseed study uses rings linking neighboring channels and two-community graphs linking channels within two groups, with $C \in \{8, 16, 32, 64, 128, 256, 384, 512\}$ and four attention heads ($d_v = 16$).

All models share the benchmark preprocessing and chronological evaluation pipeline.

Each edge score is the absolute value of its field averaged over 64 test windows and relation features. The area under the receiver-operating-characteristic curve (ROC-AUC) is the probability that a true edge receives a higher score than a nonedge, with half credit for ties. Edges satisfy $|A_{ij}| > 0$; evaluation includes all C^2 entries and diagonal nonedges. AUC measures ranking, not precision at a chosen edge threshold. The additive comparison uses $g + r + c = R - R^{\perp}$.

Ring and community results. The fixed scalar interaction readout ranks ring edges above the additive readout at all eight channel counts (Table 24). Three-seed ring AUCs range from 0.57 to 0.72; at $C = 64$ and 256, the interaction/additive means are 0.66/0.44 and 0.69/0.46. Community scores are generally near or below chance with this absolute pooled score. The following decomposition explains how community structure can reside in the interaction space, while Appendix D.3 evaluates supervised recovery from its vector features.

Community membership can lie within the centered interaction space even when the scalar readout does not recover it. For a balanced two-block matrix with membership signs $s_i \in \{-1, 1\}$ and $\mathbf{1}^{\top} s = 0$,

$$B = \tfrac{1}{2}(\mathbf{1}\mathbf{1}^{\top} + ss^{\top}), \qquad P_{\mathrm{std}} B P_{\mathrm{std}} = \tfrac{1}{2} ss^{\top} \neq 0.$$

The interaction retains $ss^{\top}$, which encodes whether two channels share a block, while row and column averages cannot distinguish their memberships. The interaction space represents block membership; the table measures its extraction by trained scalar readouts.

Table 24: Interaction and additive topology readouts (ROC-AUC, three-seed mean±SD; chance 0.5). Identical windows and pooling are used for both fields. Interaction means exceed additive means at all eight ring sizes.

	Ring		Community	
C	$R^{\perp}$	$g+r+c$	$R^{\perp}$	$g+r+c$
8	0.72 ± 0.07	0.48 ± 0.03	0.28 ± 0.18	0.55 ± 0.11
16	0.57 ± 0.21	0.52 ± 0.01	0.39 ± 0.17	0.37 ± 0.24
32	0.60 ± 0.04	0.50 ± 0.01	0.47 ± 0.07	0.50 ± 0.00
64	0.66 ± 0.07	0.44 ± 0.11	0.49 ± 0.02	0.38 ± 0.17
128	0.61 ± 0.04	0.60 ± 0.04	0.49 ± 0.00	0.50 ± 0.01
256	0.69 ± 0.11	0.46 ± 0.06	0.49 ± 0.01	0.56 ± 0.10
384	0.57 ± 0.06	0.46 ± 0.10	0.50 ± 0.01	0.50 ± 0.00
512	0.62 ± 0.05	0.47 ± 0.11	0.50 ± 0.00	0.50 ± 0.00

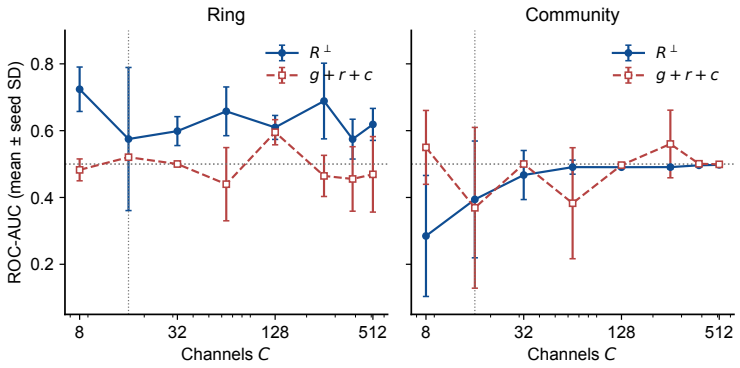


Figure 7: Topology readout across channel counts (three-seed mean±SD). Filled markers show interaction $R^{\perp}$ and open markers additive $g+r+c$. Panels compare ring and community coupling; the vertical line marks $C = 16$ as a channel-count reference.

Forecasting and representation readouts. These simulation runs measure both forecasting error and topology readouts. Ring forecasting changes relative to first-order EQNN include $+10.3\%$, $+17.8\%$, and -12.3% at $C = 32, 256, 512$, respectively, with substantial seed variation. The topology comparison addresses the information exposed by each readout; the matched real-data experiments in Section 4.2 evaluate the forecaster's predictive performance.

Spectral alternative and directional structure. On the tested directed ring, spectral coherence supplies directional pair features. We evaluate this variant with $A_{(i+1) \bmod C,i} = 0.8$, $C = 32$, 50-epoch maximum training, relation width 16, and seeds $\{1, 2, 42\}$. If Z_i is a channel's Fourier transform, normalized cross-products $\Gamma_{ij} = Z_i \overline{Z_j} / [(|Z_i| + \epsilon)(|Z_j| + \epsilon)]$ have a symmetric real part and an antisymmetric imaginary part. The complex feature itself is conjugate-symmetric, not antisymmetric.

The averaged antisymmetric readout correlates with $A - A^{\top}$ at $+0.701$, -0.666, and $+0.735$ for seeds 1, 2, and 42. This gives alignment of magnitude 0.666–0.735 with the directed structure, up to a seed-dependent global sign. An oriented edge predictor additionally requires a sign convention fixed independently of the evaluation graph.

Symmetric and antisymmetric matrices are orthogonal. The coherence construction therefore separates features for undirected association from features that can align with $A - A^{\top}$, complementing the content field's symmetric target-conditioned messages.

G From predictive dependencies to physical descriptions

Pair features can be evaluated against different kinds of relationship. We first compare fixed scores with statistical dependence and predictive support. Physical interpretation then uses either calibrated forces to fit a coefficient map or relation-selected forecasts to fit symbolic expressions. Rössler, Weather, and wind-tunnel protocols follow Section 4.5's progression.

Table 25: Dependency recovery: Spearman rank correlation with F (↑), one seed per model. Amplitude sums featurewise root-mean-square (RMS) values; static norm is the length of the mean pair vector; projection measures alignment with the global mean vector. Signs are fixed independently of F.

Readout	Weather	Solar	Traffic
ENF amplitude	0.806	0.801	0.321
ENF static norm	0.614	0.868	0.397
ENF global projection	0.223	0.863	0.545
ENF attention A_0	-0.082	-0.201	-0.178
ENF attention A_1	0.144	-0.031	-0.029
iTransformer attention	0.269	-0.445	-0.051
Raw $\lvert\text{Spearman}\rvert$	0.869	0.920	0.710

G.1 Discovery checkpoints and dependency targets

The discovery studies use trained ENF checkpoints with input length 96, horizon 96, seed 1, and relation width 16. Weather uses $d64/L1$, batch size 32; Solar uses $d256/L1$, batch size 32; Traffic uses $d64/L3$, batch size 16. These discovery configurations are separate from the compact and matched PEMS comparisons. The iTransformer attention controls use corresponding capacity, normalization, and training settings. Readouts pool all test windows, with layers also pooled for Traffic; attention scores are averaged over all heads.

The dependency target F_{ij} measures how well one contemporaneous variable predicts another. A quantile-binned median estimator divides the source range into approximately equal-count bins and predicts each bin's median target value; held-out folds give out-of-fold R^2. The first 70% of each dataset supplies these folds: Weather uses 256 bins and five folds, Solar 128 bins and three folds, and Traffic 64 bins and two folds. Weather and Solar retain the larger directional score as F_{ij}; Traffic fits the $i < j$ direction and mirrors it. The evaluations contain 210, 9,316, and 371,491 pairs. This random-fold target describes contemporaneous statistical dependence within the training period, distinct from physical adjacency or future-time generalization.

G.2 Fixed relation and attention readouts

Write $\bar{R}_{ij} = \mathbb{E}[R_{ij}]$ for the raw pair field averaged over test windows and pooled layers, and $\bar{g} = \text{mean}_{ij}\, \bar{R}_{ij}$. The three ENF scores in Table 25 are

$$S_{ij}^{\text{amp}} = \sum_{q=1}^{d_r} \sqrt{\mathbb{E}[R_{ijq}^2]}, \qquad S_{ij}^{\text{stat}} = \|\bar{R}_{ij}\|_2, \qquad S_{ij}^{\text{proj}} = \frac{\langle \bar{R}_{ij}, \bar{g} \rangle}{\max(\|\bar{g}\|_2, 10^{-12})}.$$

Each score reduces a pair object to one number, so it ranks which variables are related rather than describing the relation itself. Amplitude sums featurewise RMS, distinct from the RMS of a vector norm. All three scores use the raw field before subtracting additive endpoint effects. Attention controls average A_0, A_1, or vanilla iTransformer attention over the same windows. Spearman correlations use fixed $i < j$ entries with their original sign, independently of evaluation labels. Absolute training-data Spearman correlation provides the data-only baseline.

For Weather, the seven pairs with $F_{ij} > 0.999$ are dew point–vapor pressure, specific humidity–water-vapor concentration, temperature–saturation vapor pressure, vapor pressure–water-vapor concentration, vapor pressure–specific humidity, dew point–water-vapor concentration, and dew point–specific humidity. As reported in Table 25, the raw data Spearman rank correlation serves as a strong empirical baseline (reaching 0.869), and both raw rank correlations and ENF pair amplitudes recover all top seven pairs, whereas standard attention baselines recover none. We transparently note that while offline pairwise statistics can effectively identify high-correlation pairs in tabular settings, ENF extracts these relations endogenously within a unified, end-to-end forecasting framework without requiring separate data-mining stages or external statistical pre-computation. Furthermore, the learned pair field directly provides an internal computational substrate that simultaneously drives cross-channel attention and downstream distillation.

G.3 WIND-TUNNEL FORECASTING AND PREDICTIVE-SUPPORT READOUTS

Causal Chambers provides instrumented physical systems with controlled actuators and measured responses (Gamella et al., 2025). We use 16 time-varying wind-tunnel channels from `wt_walks_v1`: hatch opening, two potentiometers, two fan loads, fan currents and speeds, four pressures, microphone level, and two electrical signals. Random-walk recordings 2–11 form the training set, 12–13 validation, and 14–16 the test set, with 10,000 observations per recording. Inputs and forecasts follow the standard training-set data preprocessing used throughout this work (Appendix B.1), with window strides of two for training and four for evaluation within each recording. Slow and fast run-5 recordings with deterministic hatch/load excitation provide the changed-excitation test pool.

The forecasting comparison crosses the first-order and relational architectures with zero/frozen or trainable identity fields. Both use $d128/L2$, FFN width 256, four heads ; ENF has $d_r = 16$. Training uses batch 256, AdamW with learning rate 5×10^{-4} and zero weight decay, cosine decay over at most 40 epochs, validation patience six, and seeds 1 and 2. Table 26 reports the two-seed means. Identity improves both architectures on both evaluation pools, showing the predictive value of persistent variable meta-representations across changing excitation. ENF with identity attains the lowest changed-excitation mean, while First-order EQNN with identity attains the lowest random-walk mean. Three external forecasters use the identical recipe: channel-independent DLinear, vanilla iTransformer at the same width and depth, and parameter-free persistence, which repeats the final input sample over the horizon. Both identity-conditioned variants stay below every baseline on both pools. These results distinguish ID's contribution to prediction from the zero-ID pair readouts below.

Table 26: Wind-tunnel forecasting (two-seed mean MSE, $C = 16$, $d128/L2$, $T = H = 96$). Changed excitation uses held-out deterministic hatch/load sequences. These settings are separate from the benchmark protocol.

Architecture	Identity	Random-walk test	Changed excitation
First-order EQNN	Zero	0.25072	0.30774
First-order EQNN	Learned	0.24024	0.28858
ENF	Zero	0.24959	0.30589
ENF	Learned	0.24108	0.28739
DLinear	–	0.29289	0.29069
iTransformer	–	0.24323	0.33365
Persistence	–	0.44746	0.34106

Two distinct dependency targets. We compare pair scores with the 26 undirected physical-edge pairs among the 120 distinct channel pairs, and with an empirical predictive-support target. For each ordered source–target pair, a ridge predictor augments the target's eight lags, selected quadratic terms, and increment with the source's current value, increment, and last-eight mean and SD. Predictors are fitted on training recordings and evaluated on the random-walk test recordings at horizons $\{1, 2, 4, 8, 16\}$. A pair enters the support target when adding either endpoint improves held-out R^2 by more than 0.003 at any of these horizons, giving 13 pairs. This target measures incremental predictive support under the specified regression basis.

For ENF with its identity field fixed at zero, we extract the raw pair field separately at both layers. The time-mean norm is $\mathbb{E}\|R_{ij}\|_2$; the mean-field norm is $\|\mathbb{E}R_{ij}\|_2$. iTransformer scores average attention heads before window averaging. All scores are evaluated on the fixed upper triangle, with diagonals excluded. Table 27 gives both layers without choosing a layer per reference. Its SDs are across the two model seeds, not resampled channel pairs.

The two targets distinguish physical adjacency from incremental predictive support. Second-layer ENF pair norms align with predictive support: AUCs rise from 0.458 to 0.713 and from 0.495 to 0.750 across layers, versus 0.431 for raw correlation. First-layer iTransformer attention reaches 0.734. These scores measure access to predictive dependencies under each scalar readout. The coefficient and formula studies below characterize the relationships through calibrated physical targets and transfer to measurements.

G.4 PHYSICAL COUPLING READOUT IN A RÖSSLER OSCILLATOR NETWORK

A calibrated coupling-force target makes the physical-readout question explicit: does the pair residual predict force through coefficients that transfer beyond the fitted recording and pairs? We use the R1

Table 27: Wind-tunnel dependency readouts: AUC ± sample SD across two seeds. The physical graph and empirical predictive-support target are distinct. Raw correlation averages absolute within-window Pearson correlation and has no model seed.

Readout	Layer	Physical edges	Predictive support
ENF time-mean norm	1	0.611 ± 0.043	0.458 ± 0.024
ENF mean-field norm	1	0.526 ± 0.049	0.495 ± 0.021
ENF time-mean norm	2	0.605 ± 0.003	0.713 ± 0.007
ENF mean-field norm	2	0.515 ± 0.028	0.750 ± 0.051
iTransformer attention	1	0.550 ± 0.008	0.734 ± 0.001
iTransformer attention	2	0.537 ± 0.010	0.581 ± 0.006
Raw correlation	–	0.633	0.431

network from the electronic-oscillator dataset of Vera-Ávila et al. (2020), comprising 28 oscillators with fixed adjacency A. Each recording contains 30,000 samples of the second circuit voltage y_i, acquired at 30 kHz. Released signals have undergone third-order Butterworth filtering at 1,500 Hz; we retain every sample without further filtering or resampling. Three independent recordings are available at each of 101 coupling settings $\kappa = k/100$, $k = 0, \ldots, 100$.

The forecasting backbone uses 96 consecutive voltage samples to predict the next 96. For each coupling setting, we distribute 128 window starts across each of recordings 1 and 2 using an evenly spaced grid with uniform jitter of up to 0.3 grid intervals (sampling seed zero). The first 115 starts per recording supply training windows and the remaining 13 validation windows; recording 3 supplies 32 evaluation windows. This yields 23,230/2,626/3,232 forecasting windows. Inputs and forecasts follow the same standard data preprocessing (Appendix B.1).

We train zero-identity ENF with $d = 128$, two layers, FFN width 256, four heads, $d_r = 16$, and model seed one. AdamW uses batch size 256, learning rate 10^{-3}, and zero weight decay; cosine decay reaches 2×10^{-5} over at most 200 epochs, with validation-MSE patience 20. The selected forecasting checkpoint is frozen before fitting the physical readout.

The circuit calibration provides two complementary targets. For the contribution of node j to node i, the published component values $R_{15} = 500\,\mathrm{k\Omega}$ and $C_2 = 1\,\mathrm{nF}$ give

$$g_{ij} = \frac{\kappa A_{ij}}{R_{15} C_2} = 2000 \kappa A_{ij}\ \mathrm{s}^{-1}, \qquad F_{ij}(t) = g_{ij}\big(y_j(t) - y_i(t)\big). \tag{8}$$

Here g_{ij} is the nominal coupling coefficient and F_{ij} its signed instantaneous contribution to $\dot{y}_i$, in $\mathrm{V\,s^{-1}}$. The readout is fitted only through force targets constructed from this circuit law; nominal coefficients are used for subsequent evaluation.

Relations generate coefficients, states evaluate the polynomial. For each recording and setting, 32 nonoverlapping 96-sample context windows are evenly distributed over the first 10,000 samples. At each layer we symmetrize a field as $(R_{ij} + R_{ji})/2$, then concatenate its coordinate-wise means and population standard deviations over these windows. Two layers with $d_r = 16$ give a 64-dimensional descriptor ρ_{ij}. The same construction is applied to $R^{\perp}$, the additive field $L = g + r + c$, the full field R, and a control that permutes residual descriptors across the 378 unordered pairs independently for each recording and setting (seed 20260911). The coefficient map receives only this descriptor, without node indices, A_{ij}, κ, or evaluation-state voltages.

We generate all six coefficients of a degree-two physical-state dictionary,

$$\begin{aligned} \widehat{F}_{ij}(t) &= c_0 + c_i y_i(t) + c_j y_j(t) + c_{ii} y_i(t)^2 + c_{ij} y_i(t) y_j(t) + c_{jj} y_j(t)^2, \\ \mathbf{c}_{ij} &= B \phi_d(\rho_{ij}), \qquad d \in \{1, 2\}. \end{aligned} \tag{9}$$

The coefficient-map features ϕ_d contain an intercept and all monomials through degree d, giving 65 or 2,145 features. This degree concerns the map from relations to coefficients; the physical-state dictionary has six terms in both cases. Each coefficient vector is fixed across 64 subsequent states, taken at the last samples of nonoverlapping 96-sample windows evenly distributed over samples 10,000–29,999. Both pair orientations supply force supervision, with the two state voltages exchanged and the target sign reversed. No coefficient zero pattern, sign constraint, or equality $c_i = -c_j$ is imposed. This paired supervision favors antisymmetric force functions for every field.

Matched fitting and transfer. Readout settings use $k \in \{0, 1, 2, 5, 10, 20, 30, 50, 70, 90\}$. Recording 1 supplies fitting examples, recording 2 selection examples, and recording 3 test examples. We reserve 76 of the 378 unordered pairs, selected by a fixed permutation with seed 20260911, from both fitting and selection. The remaining 302 pairs, including nonedges, provide 3,020 descriptor groups per fitting or selection split, each evaluated on 64 states in both orientations. All scores exclude diagonal pairs. The all-setting protocol holds out recording 3 at the same ten settings. A second protocol fits at $\kappa \in \{0, 0.02, 0.10, 0.30, 0.70\}$, selects at $\{0.01, 0.20\}$, and tests at $\{0.05, 0.50, 0.90\}$ in recording 3. The latter measures transfer of the physical readout across settings; the frozen backbone was trained across all 101 settings on R1.

All fields use identical descriptor standardization, force loss, regularization grid, and selection rules. Descriptor coordinates and expanded monomials are standardized on fitting examples; physical-state columns and force targets are divided by their fitting-set RMS without centering, preserving a diagonal conversion back to physical coefficients. Ridge penalties are selected from $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$. A matrix-free conjugate-gradient solver uses relative tolerance 10^{-7} and at most 600 iterations; only converged candidates enter selection. Starting from the selected unrestricted six-term fit, coefficients whose fitting-set RMS is below $\tau \in \{0.01, 0.03, 0.1\}$ times the largest coefficient RMS are removed and the remaining terms refitted. Selection chooses the sparsest converged candidate within 1% of the lowest validation force MSE. All 16 field/degree/protocol combinations retain y_i and y_j.

Table 28: **Physical information in frozen Rössler relations.** Test R^2 for force F, force on the 76 reserved pairs F_{pair}, and inferred gain $\hat{g} = (c_j - c_i)/2$. Degree d specifies the relation-to-coefficient map. All-setting tests use an independent recording; held-out-setting tests additionally exclude test settings from readout fitting and selection, with F_{pair} testing both transfers jointly. All entries use the same frozen ENF checkpoint (seed one).

		All settings			Held-out settings		
Field	d	F	F_{pair}	$\hat{g}$	F	F_{pair}	$\hat{g}$
$R^{\perp}$	1	.0728	.0893	.0440	.0616	.0718	−.0036
	2	.0996	.1176	.0657	.0994	.1173	.0420
$L = g + r + c$	1	.0650	.0805	.0329	.0511	.0601	−.0098
	2	.0808	.0971	.0521	.0668	.0877	.0340
Shuffled $R^{\perp}$	1	.0408	.0472	.0009	.0288	.0309	−.0482
	2	.0359	.0429	−.0068	.0359	.0407	−.0320
R	1	.0750	.0917	.0516	.0554	.0630	−.0179
	2	.1061	.1261	.0686	.1000	.1220	.0364

Exploratory diagnostic: qualitative coupling detection beyond additive fields. Table 28 reports force prediction R^2 across different representations. We emphasize that continuous force estimation from unmapped dynamical observations remains fundamentally difficult, as reflected by the modest overall test performance ($R^2 = 0.0996$). This experiment should therefore be viewed strictly as an exploratory diagnostic rather than an accurate quantitative reconstruction of the full continuous force field. Nonetheless, the supervised quadratic readout on $R^{\perp}$ consistently surpasses both the matched quadratic additive field ($R^2 = 0.0808$) and shuffled-residual controls ($R^2 = 0.0359$), with this relative ordering consistently maintained on reserved oscillator pairs ($R^2 = 0.1176$) and under held-out parameter settings ($R^2 = 0.0994$). These results indicate that despite low overall explained variance, the double-centered residual substrate $R^{\perp}$ retains a genuine, partially transferable qualitative coupling signal that is not accessible from single-variable additive dynamics alone.

Coefficient structure and stability. Before term removal, the quadratic residual map has force $R^2 = 0.0997$. Its constant and cross-product coefficients have RMS $3.01 \times 10^{-7}\,\mathrm{V\,s^{-1}}$ and $8.81 \times 10^{-8}\,\mathrm{V^{-1}s^{-1}}$, respectively, and $c_i + c_j$ has RMS $1.52 \times 10^{-7}\,\mathrm{s^{-1}}$. The two squared-voltage coefficients each have RMS $1.36\,\mathrm{V^{-1}s^{-1}}$; the constant and quadratic terms together contribute RMS $2.77\,\mathrm{V\,s^{-1}}$. Term selection retains the two linear-voltage terms, recovering the voltage-difference form through paired force supervision. The residual readout improves on additive and shuffled-residual controls under the transfer protocols in Table 28.

Keeping the coefficients fixed, force R^2 is 0.1011, 0.0992, 0.0996, and 0.0986 over four consecutive evaluation-state blocks. Recomputing descriptors from the first and second 16-window halves of

the context gives a $17.01\,\mathrm{s}^{-1}$ RMS difference in linear coefficients, whose primary estimates have RMS $48.54\,\mathrm{s}^{-1}$. At zero coupling, gain and force RMSE are $9.87\,\mathrm{s}^{-1}$ and $19.65\,\mathrm{V\,s}^{-1}$; R^2 is undefined because the targets vanish. These measurements characterize coefficient accuracy and context sensitivity alongside transfer to subsequent states.

G.5 Symbolic search and source-range extrapolation

Weather examines the second route to a physical description: the model selects related variables, while symbolic regression identifies a functional form from their forecasts. The six strongest unordered pairs are selected from the first-layer amplitude field by sorting ordered entries and retaining each unordered pair once. They cover the humidity, water-vapor, and dew-point relationships in Table 29. Each pair is fitted in both directions, then tested on source values withheld from fitting and selection.

For an input window ending at time t, a fitting sample is $(x, y) = (\hat{y}_{t+h,i}, \hat{y}_{t+h,j})$: both coordinates are ENF's own predicted channel values from the same window and the same lead time $h \in \{1, \ldots, 96\}$. Forecasts are pooled across test windows and lead times, and returned to physical units using the training-data scaler. ENF supplies the pair selection and forecast values; symbolic regression supplies the expression and its fitted constants. Variable names are used only for reporting, not as search features. No known physical expression is provided as a candidate template.

For each direction, a fixed random sample of 4,000 forecast pairs is sorted by its source value. The lowest 60% supplies the fitting segment; the next 20% selects expressions; the highest 20% is used only for the reported R^2_{ext}. Thus extrapolation refers to the range held out from the symbolic fit, not to inputs outside the forecasting model's data support. Forecast samples may share overlapping input windows; they are not treated as independent experimental replicates.

The search enumerates candidate subexpressions, called atoms, from x, fitted constants, arithmetic, $\exp$, $\log$, $\sqrt{\cdot}$, and negation. Each expression tree has at most nine nodes and two constant leaves. It considers one atom with a fitted scale and offset, or a greedy two-atom combination. Fitting uses up to 300 quantile-bin medians of the fitting segment, a coarse two-restart search, and a four-restart refinement. Candidate shortlists are stratified by parameter count. Selection maximizes the next-segment R^2 minus 0.005 per fitted parameter beyond two, with parameter-count tie-breaking. The polynomial control fits degrees one through five on the raw fitting segment and chooses its degree using the same selection segment.

Operators are protected numerically: $\ell(x) = \log(|x| + 10^{-9})$, $s(x) = \sqrt{|x|}$, and exponentials clip arguments to $[-60, 60]$. Division replaces denominators with magnitude below 10^{-9} by 10^{-9}. The positive-input expressions in Table 2 use ordinary notation for these protected operators. Table 29 gives all twelve directed fits and their validation-selected polynomial controls.

Table 29: All twelve Weather symbolic fits. Scores use the upper 20% source-value segment; polynomial degree is chosen on the preceding 20%. Symbols follow Table 2; additionally, q is specific humidity (g/kg) and x the source forecast in recorded physical units. ℓ and s are protected operators; k is a fitted constant specific to each expression. Displayed coefficients are rounded.

Pair	Selected expression	Symbolic R^2_{ext}	Poly. degree	Poly. R^2_{ext}
$c_v \to e$	$-0.03261 + 0.9971\ell(\exp x - s(x))$	0.9960	1	0.9930
$e \to c_v$	$0.1441 + 0.9872\ell(\ell(x) - \exp x)$	0.9883	1	0.9886
$q \to e$	$-0.03878 + 0.7991(x + x)$	0.9943	1	0.9919
$e \to q$	$0.05582 + 0.6209x$	0.9904	1	0.9878
$q \to c_v$	$0.01563 + 1.599x$	0.9997	1	0.9998
$c_v \to q$	$-0.01121 + 0.6255s(x)s(x)$	0.9998	1	0.9998
$e \to T_d$	$-31.3 + 15.55(k/\ell(x) + \ell(x))$	0.9968	5	-149.6377
$T_d \to e$	$-1.057 + 7.15\exp((k-1)x)$	0.9973	3	0.9969
$c_v \to T_d$	$-35.07 + 16.27\ell(x + s(x))$	0.9938	2	-0.6572
$T_d \to c_v$	$-1.464 + 7.623\exp((k-1)x)$	0.9561	2	0.8014
$q \to T_d$	$-29.32 + 16.69\ell(x + s(x))$	0.9945	2	-0.5761
$T_d \to q$	$-0.9019 + 4.744\exp((k-1)x)$	0.9559	2	0.8112

The six weakest pairs test whether the search also finds useful formulas where ENF assigns low relation scores. All twelve extrapolation scores are negative, ranging from -9.2863 to -0.0046. Two fits return symbolic expressions and the others a linear fallback; none predicts better than the evaluation segment's mean.

The three logarithmic dew-point fits retain accuracy beyond their fitting range, while humidity conversions are effectively linear. This establishes compact descriptions of selected forecast relationships. The wind-tunnel test next evaluates frozen formulas on physical measurements.

G.6 Wind-tunnel formulas and transfer to physical measurements

Transfer to measured responses tests whether a forecast-derived expression describes the instrumented system beyond the predictor's outputs. The 16 wind-tunnel channels and recording-level split follow Appendix G.3: recordings 2–11 supply training data, 12–13 validation, and 14–16 external evaluation, with 10,000 observations each. We retain the recorded grid without additional smoothing or interpolation. Forecasting inputs follow the same standard data preprocessing as the forecasting runs above.

ENF uses the same zero-identity architecture, optimizer, and early-stopping recipe as Appendix G.3, with a 96-step input and a one-step prediction target. Seeds 1 and 2 are trained separately. We retain the original horizon-96 window-start grid when shortening the target: starts run from 0 through 9,808 within each recording, at stride two for training and four for validation and evaluation. This gives 49,050 training, 4,906 validation, and 7,359 evaluation windows, with no window crossing a recording boundary. Forecasts are first inverse-transformed using the forecasting scaler. The current channels are then converted from ADC values to amperes by $I = \mathrm{ADC} \times 1.16/2046$; speed, pressure, and load retain their recorded rpm, Pa, and fan-control units.

Pair selection is fixed before short-horizon fitting. For each seed, we use the twelve strongest unordered pairs under the original horizon-96 model's training-set mean relation amplitude, averaged over both directions and layers. These comprise all six pairs among the four pressure channels and all three pairs among load, speed, and current for each fan. Each pair is searched in both directions. Thus the same relationally selected sensor set is used for both seeds, although its ranking differs.

For a window ending at time t, a fitting sample is $(x, y) = (\hat{y}_{t+1,i}, \hat{y}_{t+1,j})$: both coordinates are forecasts from the same window and lead, in the units above. We sample 4,000 windows without replacement from the training pool and another 4,000 from the validation pool, using sampling seeds zero and one, respectively; the same indices serve every directed pair. The training sample fits the expression, and the validation sample selects it. The complete training source range is available for fitting. A second, symbolic-fit normalization centers and scales both coordinates using only these 4,000 training forecasts. After sorting by the source coordinate, we divide the sample into 300 equal-count bins and fit to the source and target medians in each bin. These symbolic-fit statistics are distinct from the forecasting model's training-recording scaler.

The grammar uses arithmetic, negation, exponentials, logarithms, and square roots, with at most nine nodes and two internal constants per atom. For each candidate, linear least squares fits the output offset and atom weights on a centered and scaled design matrix, while nonlinear least squares fits internal constants. A two-restart coarse search retains 50 atoms per internal-constant count, followed by four-restart refinement. The leading atoms are evaluated individually and in greedy two-atom combinations fitted to the residual. Selection maximizes validation $R^2 - 0.005 \max(p - 2, 0)$, where p counts fitted parameters; ties favor fewer parameters. A polynomial control fits degrees one through five in a Chebyshev basis on the same normalized bin medians and uses the same validation criterion to select degree. No physical response template is included in the grammar.

For numerical protection, $E(z) = \exp(\mathrm{clip}(z, -12, 12))$, $S(z) = \sqrt{|z|}$, and $\ell(z) = \log(|z| + 10^{-7})$. Division floors denominator magnitudes at 10^{-7} while preserving their sign (zero uses the positive floor). Once selected, an expression and both sets of normalization statistics are frozen. On recordings 14–16, we first score the formula between contemporaneous source and target forecasts, then apply the identical physical-unit formula to the measured source and score it against the measured target at the same time. Neither its coefficients nor its offset or calibration is re-estimated on these measurements. This recording-level transfer test complements Weather's source-range split.

For compact reporting, define $\Phi(z) = z\sqrt{2E(z)}$ and $\Psi(z) = zS(E(z)/S(z))$, with protected division as above. Table 31 gives both seeds' expressions for six selected relations, including the three displayed in the main text. Each row specifies a normalized expression F; its physical-unit prediction is

$$y = \mu_y + s_y F(z), \qquad z = (x - \mu_x)/s_x. \tag{10}$$

Table 30 supplies the symbolic-fit source and target constants, computed after inverse forecasting normalization and current conversion. Thus the same source variable has the same normalization in every

relation for a given seed. Equation 10 restores the formula's physical units; the main-text expressions absorb the output scale and offset and simplify within the evaluated range, where clipping is inactive.

Table 30: Wind-tunnel source and target constants for Eq. 10. Load u is the recorded fan-control value, speed n is in rpm, current I in A, and pressure p in Pa. All constants come from training-recording forecasts and remain fixed on validation and external measurements.

Variable	μ, seed 1	s, seed 1	μ, seed 2	s, seed 2
u_{in}	0.4985927	0.2950043	0.5013030	0.3035744
u_{out}	0.5279351	0.2840624	0.5305409	0.2909133
n_{in}	1632.2298	784.6960	1638.9325	807.2391
n_{out}	1743.4369	736.3975	1749.3702	753.9769
I_{in}	0.1994633	0.0309756	0.1995106	0.0319099
I_{out}	0.1949068	0.0284260	0.1949477	0.0291576
p_{up}	96318.2993	75.1251	96318.5995	75.5471
p_{down}	96299.7384	74.8175	96299.9995	75.1937

Table 31: Both seeds' wind-tunnel expressions for six selected relations. $F(z)$ is the normalized target in Eq. 10; E, S, ℓ, Φ, and Ψ use the protected definitions above. Forecast and measured R^2 evaluate the same frozen expression on 7,359 external windows. Coefficients are rounded.

Source → target	Seed	$F(z)$	Forecast R^2	Measured R^2
$u_{\text{in}} \to n_{\text{in}}$	1	$-1.444540 + 1.046458\lvert z + 1.341687\rvert$	0.9758	0.9714
	2	$-1.428618 + 1.042529\lvert z + 1.334215\rvert$	0.9761	0.9719
$u_{\text{out}} \to n_{\text{out}}$	1	$3.522133 - 2.150800S((z + 2.706191)/E(z))$	0.9595	0.9545
	2	$-0.021114 + 1.033247zS(\ell(z) - S(z))$	0.9524	0.9458
$u_{\text{in}} \to I_{\text{in}}$	1	$-0.422050 + 0.812648\Psi(z)$	0.9804	0.8875
	2	$-0.413442 + 0.798212\Psi(z)$	0.9791	0.8875
$n_{\text{in}} \to I_{\text{in}}$	1	$-0.417120 + 0.546072\Phi(z)$	0.9779	0.8799
	2	$-0.422031 + 0.807379\Psi(z)$	0.9795	0.8805
$n_{\text{out}} \to I_{\text{out}}$	1	$-0.410469 + 0.848842\Psi(z)$	0.9211	0.8331
	2	$-0.403659 + 0.838277\Psi(z)$	0.9229	0.8335
$p_{\text{up}} \to p_{\text{down}}$	1	$0.006952 + 0.997553z$	0.8475	0.8436
	2	$0.005348 + 0.995924z$	0.8470	0.8427

The measured external inlet-speed range is 56.0–2929.7 rpm and the outlet range 42.2–3004.8 rpm; over 99.8% of these speeds lie within the fitting ranges. This is primarily recording-level and forecast-to-measurement transfer, complementing Weather's source-range extrapolation. Across all 48 directed fits, 45 have positive measured-pair R^2, with median 0.8381. These load–speed, current, and pressure relations transfer from forecasts to measured responses.

H LIMITATIONS

Table 13 quantifies the quadratic activation cost of the dense implementation. With dropout disabled, factorization removes dense pair-feature storage (Appendix C.8); attention still has quadratic cost.

The forecasting evidence covers the stated datasets and configurations. Four PEMS datasets provide the independent units for aggregate inference. The staged relational increment includes changes in dropout placement, and the channel-count sweeps vary channel composition across nested subsets. These conditions define the comparisons behind the reported forecasting gains.

The observable-intermediate tasks use auxiliary edge supervision and test new sequences from the same motif definitions. They establish discrete relation capture and compositional access, with frozen-field readouts evaluated separately from the original query head.

The physical studies use additional readout targets or transfer tests. Rössler coefficients are recovered through calibrated force supervision on a forecasting checkpoint and transfer to a held-out recording. Weather evaluates formulas on held-out source ranges of forecasts, while the wind tunnel transfers frozen formulas to measured responses in independent recordings. Together these protocols establish calibrated coefficient recovery and formula transfer beyond the fitting data.