MODULAR DECOMPOSITION OF LEARNED CONTROLLED DYNAMICS FROM ONE FIT

Anonymous authors
Paper under double-blind review

ABSTRACT

Learned models of controlled systems are often trained and deployed as monolithic predictors, even when the underlying physical system is organized into interacting subsystems that are built, operated, and maintained separately. Decomposing such a predictor usually requires choosing module boundaries and retraining a new model for each candidate structure. We study coupling-additive predictors, in which the prediction is a sum of contributions, one for each pair of a sensor group and a control channel; the class contains bilinear Koopman models with grouped dictionaries and any network with a linear head over per-group features. We introduce MODEC, which evaluates every candidate decomposition of a ridge-fitted predictor in this class from that single fit: the weights that restricted retraining would produce for any set of removed couplings, and the resulting increase in training loss, follow in closed form from the cached fit, and a covariance matrix of the fitted coupling contributions scores all candidate boundaries at once. On independent held-out episodes, a finite-sample bound on the cost of a restriction holds simultaneously over all candidate boundaries, with their number entering only through a logarithm. A selected decomposition runs as separate modules that reproduce the restricted predictor exactly, and a finite-horizon recursion bounds how its deviation from the original model accumulates over multi-step rollouts. On the Tennessee Eastman process and the IEEE 39-bus system, the scores rank the exact retraining cost of candidate modules with Spearman correlation 0.91 and 0.89, and the selected decompositions remove 67% and 84% of the removable couplings at 10% and 38% relative increases in one-step NRMSE; on a six-reactor cascade, MODEC recovers the planted subsystem structure at every coupling strength, with a restriction cost that becomes negligible when the subsystems are decoupled.

1 INTRODUCTION

Learned models of physical systems are widely used for prediction, monitoring, and control, including lifted Koopman surrogates (Williams et al., 2015; Korda & Mezić, 2018; Brunton et al., 2021; Nüske et al., 2023). These models are usually trained as monolithic predictors even when the underlying system is organized into interacting subsystems. A modular predictor can be updated, replaced, or run unit by unit, as a process plant or a power network is operated, but obtaining one requires deciding which state groups and control channels each module may read, and every such choice costs prediction accuracy.

A common way to evaluate a proposed decomposition is to impose its structure, refit the model, and measure the resulting prediction error. This becomes expensive when many alternatives must be compared, because each candidate may require another training or constrained fitting step. Sparse identification methods trace model families by refitting across sparsity levels (Jovanović et al., 2014; Brunton et al., 2016; Yuan & Lin, 2006), while modular Koopman approaches typically assume a prescribed interconnection or module structure before fitting (Guo et al., 2025; Schlosser & Korda, 2022; Li et al., 2019; Jiang et al., 2024). We ask instead which module boundaries a trained controlled predictor can support, at what cost in accuracy, and whether both can be read from a single fit with a guarantee. We call the resulting procedure MODEC, modular decomposition from one fit.

Our starting point is a *coupling-additive* model organized by state group and control channel, so each state–control coupling is a removable coefficient block and a module boundary is a set of

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.

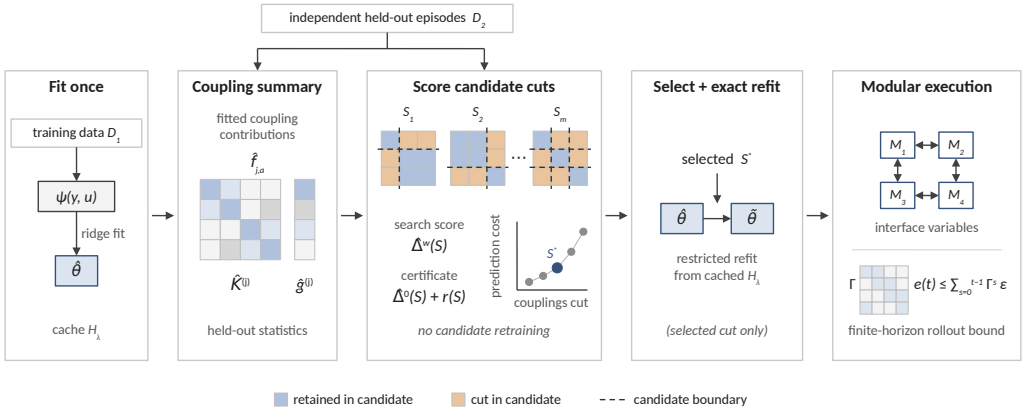


Figure 1: **MODEC workflow.** One coupling-additive predictor is fit on $\mathcal{D}_1$. Held-out episodes $\mathcal{D}_2$ provide coupling and residual covariances used to score and simultaneously bound candidate cuts. The selected cut is exactly refit from cached ridge statistics, executed module by module, and equipped with a finite-horizon rollout bound.

removed couplings, or a cut. For ridge regression, block elimination on the cached regularized Gram matrix yields, for any cut, the restricted retraining weights and the increase in the training objective without another pass over the training samples. The fitted coupling contributions also yield a covariance matrix that scores cuts in closed form, with surviving contributions either frozen or scalar-reweighted, allowing many cuts to be ranked before the exact refit of the selected one.

On independent held-out episodes, a concentration argument simultaneously bounds the frozen cost of every cut in a predeclared family, with only logarithmic dependence on the family size and without requiring identifiable coupling components or a population-optimal fitted predictor. The selected cut is then refit exactly and can be evaluated centrally or module by module; for state-increment predictors, a finite-horizon recursion bounds how its deviation from the full model accumulates during rollout. Figure 1 summarizes the workflow.

Contributions.

- **Exact restriction from one fit.** For ridge regression, block elimination gives the restricted-retraining weights and training-objective increase for any cut from the cached monolithic fit (Proposition 1), applying the block form of second-order pruning (Frantar & Alistarh, 2022; Kurtić et al., 2023) to physical state–control couplings.
- **One-fit scores with a held-out guarantee.** We derive closed-form scores for every cut from a covariance matrix of the fitted coupling contributions, with the surviving contributions frozen or rescaled by one scalar each (Proposition 2), and prove a finite-sample bound on the held-out frozen cost that holds simultaneously over a finite family of cuts whose size enters only through a logarithm (Theorem 1).
- **Modular execution.** We show that the selected restriction runs exactly as separate modules and bound its finite-horizon rollout deviation from the monolith (Proposition 3).
- **Evaluation.** On the Tennessee Eastman process and the IEEE 39-bus system, MODEC ranks exact retraining costs with Spearman correlation 0.91 and 0.89 and removes 67% and 84% of the removable couplings at 10% and 38% relative increases in one-step NRMSE, and on a six-reactor cascade it recovers the planted subsystem structure at every coupling strength (Section 7).

2 Related Work

Koopman and structured dynamical models. Extended dynamic mode decomposition and controlled Koopman methods fit linear operators in lifted coordinates (Williams et al., 2015; Proctor

et al., 2016; Korda & Mezić, 2018), including bilinear control-affine forms (Surana, 2016; Bruder et al., 2021) and finite-data guarantees for operator estimation (Nüske et al., 2023; Kostic et al., 2022). Several works impose modular or network structure on Koopman models: Schlosser & Korda (2022) use a prescribed sparsity graph, Guo et al. (2025) learn subsystem generators for a known interconnection, and Li et al. (2019); Mukherjee et al. (2022); Jiang et al. (2024) build models around predefined objects, clusters, or microgrids. These approaches specify or learn structure as part of model construction. Instead, we fit a single coupling-additive predictor and evaluate alternative restrictions afterward.

Structure discovery and one-fit restriction. Sparse identification and graph-learning methods recover model structure jointly with the fitted parameters (Brunton et al., 2016; Jovanović et al., 2014; Yuan & Lin, 2006; Anantharaman & Mauroy, 2025; Tank et al., 2021; Marcinkevičs & Vogt, 2021), while modular neural architectures encourage structured computation during training (Goyal et al., 2019; Lei et al., 2024; Boyd et al., 2025). Our exact restricted update is the partitioned-regression identity (Puntanen & Styan, 2005) in the block form used by second-order pruning (LeCun et al., 1989; Hassibi & Stork, 1992; Frantar & Alistarh, 2022; Kurtić et al., 2023; Frantar & Alistarh, 2023), which prices many structured removals from one cached inverse Hessian; magnitude-based scores (Sun et al., 2024) omit the update. Supernets and zero-cost proxies also reuse a single training run to compare architectures (Cai et al., 2019; Mellor et al., 2021; Abdelfattah et al., 2021). Here the restricted coordinates have a physical interpretation as state–control couplings, and the same fitted model supplies both exact restricted ridge weights and covariance-based scores for comparing candidate cuts.

System decomposition. Classical decentralized control, model reduction, compositional analysis, and distributed predictive control generally start from a known interconnection or a fixed partition (Siljak, 2011; Cheng & Scherpen, 2021; Chow, 2013; Arcak et al., 2016; Zhang et al., 2023; Christofides et al., 2013). Other methods choose partitions using graph structure, clustering, or goal-oriented search (Tang & Daoutidis, 2018; Yin et al., 2022; Kazm & Taha, 2026; Riccardi et al., 2026). MODEC addresses a different stage of this problem: given one trained controlled predictor, it scores candidate restrictions, recovers the exact restricted ridge solution for a selected cut, and evaluates the resulting predictor as a collection of modules.

3 SETUP

We consider one-step prediction for a controlled system with state $y_t \in \mathbb{R}^{n_y}$ and input $u_t \in \mathcal{U} \subset \mathbb{R}^{n_u}$. A sample is $Z_t = (Y_t, U_t, Y_t^+)$ with law ρ, where Y_t^+ is either the next state or, as in our experiments, the increment $Y_{t+1} - Y_t$. For output j, $R_j(f) = \mathbb{E}[(f(Y,U) - Y_j^+)^2]$.

In the episodic experiments, ρ is the law of a transition at a uniformly random position of an independent episode.

Let $\mathcal{G}_{\mathrm{dep}}$ be a partition of the measured variables into M deployment groups, and let $\mathcal{G}_{\mathrm{feat}}$ denote the feature blocks. A block $g \in \mathcal{G}_{\mathrm{feat}}$ reads variables y_g and produces $\varphi_g(y_g) \in \mathbb{R}^{D_g}$. A feature block may span several deployment groups; in IEEE, for example, a machine-pair block is retained only when both machines are available to the same module. Controls enter through $q(u) = (1, u_1, \ldots, u_{n_u})^\top$.

A *coupling* $a = (i, g)$ pairs control channel i with feature block g, with $i = 0$ denoting a state-only term. We also use a constant block $*$, with $\varphi_* \equiv 1$, for the intercept and direct control terms. Let $\mathcal{A}$ be the full coupling set and $\mathcal{A}_0 = \mathcal{A} \setminus \{(0,*)\}$ the removable couplings, with $N = |\mathcal{A}_0|$. For output j,

$$f_{\theta_j}(y,u) = \sum_{a=(i,g)\in\mathcal{A}} q_i(u)\varphi_g(y_g)^\top \theta_{j,a} = \langle \theta_j, \psi(y,u) \rangle \,. \tag{1}$$

Here $\psi(y,u)$ stacks the vectors $q_i(u)\varphi_g(y_g)$ over all couplings $a = (i,g) \in \mathcal{A}$, and for $A \subseteq \mathcal{A}$, $\mathcal{F}_A$ denotes the subclass retaining only couplings in A. We call (1) *coupling-additive*; the feature maps may be nonlinear, but the final model is linear in the coupling blocks.

A decomposition $D = (P, \{\mathcal{I}_k\}, \{\mathcal{U}_k\})$ partitions the deployment groups into $|P|$ modules $P = \{\mathcal{B}_k\}$. Module k predicts $\mathcal{B}_k$, may read groups $\mathcal{I}_k$ with $\mathcal{B}_k \subseteq \mathcal{I}_k$, and may use controls $\mathcal{U}_k$. Let

$\mathcal{G}_{\mathrm{feat}}(\mathcal{I}_k)$ contain feature blocks whose source variables all lie in $\mathcal{I}_k$. For module k, define the retained coupling set

$$A_k(D) = \{(0,*)\} \cup \{(0,g) : g \in \mathcal{G}_{\mathrm{feat}}(\mathcal{I}_k)\} \cup \{(i,g) : i \in \mathcal{U}_k,\ g \in \mathcal{G}_{\mathrm{feat}}(\mathcal{I}_k) \cup \{*\}\}.$$

For every output $j \in \mathcal{B}_k$, let $A_j(D) = A_k(D)$ and $S_j(D) = \mathcal{A}_0 \setminus A_j(D)$. We call D closed if $\mathcal{I}_k = \mathcal{B}_k$ for every k, so that no module reads another module's groups. In our experiments, the controls $\mathcal{U}_k$ are fixed by an actuator rule (Appendix E.1), so a closed decomposition is determined by its partition P.

Let $m_j = \mathbb{E}[Y_j^+ \mid Y, U]$ and let $f_j^\star = \Pi_{\mathcal{F}} m_j$ be its $L^2(\rho)$ projection onto the full class $\mathcal{F}$ in (1). The population cost of decomposition D is

$$\Delta R(D) = \sum_{j=1}^{n_y} \left[\min_{f \in \mathcal{F}_{A_j(D)}} R_j(f) - R_j(f_j^\star) \right]. \tag{2}$$

We measure composability by $c_{\mathrm{b}}(D) = 1 - \mu_{\mathrm{b}}(D)/N$, where $\mu_{\mathrm{b}}(D) = |(\bigcup_k A_k(D)) \cap \mathcal{A}_0|$. Because c_{b} is defined from the union of retained coupling blocks across modules, it measures couplings eliminated globally; output-specific removals are represented by $S_j(D)$.

Assumption 1 (Boundedness)**.** For every feature block g, $\|\varphi_g(Y_g)\|_2 \le b_\varphi$ ρ-almost surely, $\sup_{u \in \mathcal{U}} \max_i |q_i(u)| < \infty$, and $|Y_j^+| \le b_y$ almost surely.

Assumption 2 (Compatibility)**.** There exists $\vartheta > 0$ such that

$$\left\| \sum_{a \in \mathcal{A}} c_a \right\|_{L^2(\rho)}^2 \ge \vartheta \sum_{a \in \mathcal{A}} \|c_a\|_{L^2(\rho)}^2$$

for every $c_a \in \mathcal{H}_a$, where $\mathcal{H}_a = \{q_i(u)\varphi_g(y_g)^\top \eta : \eta \in \mathbb{R}^{D_g}\}$ for $a = (i,g)$.

Assumption 2 makes the decomposition $f_j^\star = \sum_{a \in \mathcal{A}} f_{j,a}^\star$ with $f_{j,a}^\star \in \mathcal{H}_a$ unique. It is used only for the population scores in Section 4; the fitted-model certificate does not require it.

4 Evaluating Restrictions from One Fit

Using fitting data $\mathcal{D}_1$, with $n_1 = |\mathcal{D}_1|$, and fixed $\lambda > 0$, for each output we fit

$$\hat{\theta}_j = \arg\min_\theta \left\{ \frac{1}{n_1} \sum_{t \in \mathcal{D}_1} (Y_{j,t}^+ - \langle \theta, \psi_t \rangle)^2 + \lambda \|\theta\|_2^2 \right\}. \tag{3}$$

Denote the objective in (3) by $J_j(\theta)$, and set $\hat{G} = n_1^{-1} \sum_t \psi_t \psi_t^\top$, $\hat{b}_j = n_1^{-1} \sum_t Y_{j,t}^+ \psi_t$, and $H_\lambda = \hat{G} + \lambda I$, so $\hat{\theta}_j = H_\lambda^{-1} \hat{b}_j$. We write $\hat{f}_j = f_{\hat{\theta}_j}$ for the fitted predictor and $\hat{f}_{j,a}(y,u) = q_i(u)\varphi_g(y_g)^\top \hat{\theta}_{j,a}$ for its contribution from coupling $a = (i,g)$.

4.1 Exact restricted refit

Removing fitted coefficients is not equivalent to retraining because the surviving coefficients generally move. For ridge regression, that movement follows exactly from the full fit.

Proposition 1 (Restricted refit)**.** *Fix one output j, let S contain the coefficient coordinates of the removed couplings and A the remaining coordinates, write $H = H_\lambda$, and let $\tilde{\theta}$ denote the minimizer of J_j over $\{\theta : \theta_S = 0\}$. Then*

$$\tilde{\theta}_A = \hat{\theta}_A + H_{AA}^{-1} H_{AS} \hat{\theta}_S, \qquad \tilde{\theta}_S = 0, \tag{4}$$

$$\tilde{\theta} = \hat{\theta} - (H^{-1})_{\cdot S} [(H^{-1})_{SS}]^{-1} \hat{\theta}_S. \tag{5}$$

Its regularized-objective increase is

$$J_j(\tilde{\theta}) - J_j(\hat{\theta}) = \hat{\theta}_S^\top \left(H_{SS} - H_{SA} H_{AA}^{-1} H_{AS} \right) \hat{\theta}_S = \hat{\theta}_S^\top [(H^{-1})_{SS}]^{-1} \hat{\theta}_S. \tag{6}$$

These identities reproduce constrained retraining on $\mathcal{D}_1$ exactly. Since $\hat{b}_j = H\hat{\theta}_j$, (4) is the reduced ridge solve $H_{AA}^{-1}(\hat{b}_j)_A$ on the cached statistics, as in partitioned regression (Puntanen & Styan, 2005); the downdate form (5) can be preferable for a small cut, and (6) prices a cut without forming the restricted weights.

4.2 SCORING CANDIDATE CUTS

Let

$$K^{(j)}_{ab} = \mathrm{Cov}(f^\star_{j,a}, f^\star_{j,b}), \qquad a, b \in \mathcal{A}_0, \tag{7}$$

and consider a removed set S with survivors $A' = \mathcal{A}_0 \setminus S$. Freeze removes S while keeping the surviving nonconstant contributions fixed and reoptimizing the intercept, reweight additionally permits one scalar per survivor, and full refit optimizes the entire reduced class. We write $\Delta^0(S)$, $\Delta^{\mathrm{w}}(S)$, and $\Delta(S)$ for their corresponding population excess risks relative to $f^\star_j$.

Proposition 2 (Freeze and reweight scores)**.** *Under Assumption 2, with $\mathbf{1}_S$ the indicator of S, $K_{\cdot A'}$ and $K_{A'\cdot}$ the column and row blocks of K indexed by A', and $(K_{A'A'})^+$ the Moore–Penrose pseudoinverse of $K_{A'A'}$,*

$$\Delta^0(S) = \mathbf{1}_S^\top K \mathbf{1}_S, \tag{8}$$

$$\Delta^{\mathrm{w}}(S) = \mathbf{1}_S^\top \left(K - K_{\cdot A'}(K_{A'A'})^+ K_{A'\cdot}\right) \mathbf{1}_S, \tag{9}$$

$$\Delta(S) \le \Delta^{\mathrm{w}}(S) \le \Delta^0(S). \tag{10}$$

The ordering follows because the three predictor classes are nested: freezing changes only the intercept, reweighting can also rescale the surviving contributions, and full refitting can change all surviving coefficients. We use reweighting for search because it is tighter than freezing and monotone as removed couplings are restored. Proposition 1 concerns the empirical ridge problem on $\mathcal{D}_1$, whereas Proposition 2 concerns population risk relative to $f^\star$. The covariance K determines the freeze and reweight costs, but not the full-refit cost in general; Appendix A.3 gives a counterexample. Under Assumption 2 it does bracket that cost, $\vartheta \sum_{a\in S} K_{aa} \le \Delta(S) \le \Delta^0(S) \le |S| \sum_{a\in S} K_{aa}$ (Lemma A.1), so the diagonal of K identifies exactly the cuts of zero population cost. MODEC starts from singleton modules and greedily merges the pair that most decreases the total fitted reweight score defined in Section 5, producing one decomposition for each number of modules $|P|$, denoted K in Section 7; Algorithm 1 in the appendix gives the full procedure.

5 CERTIFYING CANDIDATE RESTRICTIONS

Proposition 2 describes population restriction costs relative to the unknown $f^\star$. We now define analogous quantities for the fitted predictor $\hat{f}$, which can be estimated on an independent fold $\mathcal{D}_2$. Fix the model fitted on $\mathcal{D}_1$ and define $\hat{r}_j = Y^+_j - \hat{f}_j(Y, U)$ and $\hat{c}_S = \sum_{a\in S} \hat{f}_{j,a}$. We measure restriction cost relative to the re-centered fitted model, $\bar{R}_j = \min_\beta R_j(\hat{f}_j + \beta) = \mathrm{Var}(\hat{r}_j)$. Here $\Delta^0_{\hat{f},j}(S)$ is the population cost of removing $\hat{c}_S$ while keeping the surviving contributions fixed, $\Delta^{\mathrm{w}}_{\hat{f},j}(S)$ additionally allows one scalar reweighting of each survivor, and $\Delta_{\hat{f},j}(S)$ is the minimum population cost over the full reduced model class.

Define

$$\tilde{K}^{(j)}_{ab} = \mathrm{Cov}(\hat{f}_{j,a}, \hat{f}_{j,b}), \qquad g^{(j)}_a = \mathrm{Cov}(\hat{r}_j, \hat{f}_{j,a}). \tag{11}$$

For $A' = \mathcal{A}_0 \setminus S$ and $w_S = (\tilde{K}^{(j)}\mathbf{1}_S)_{A'} + g^{(j)}_{A'}$, the frozen and reweight costs are

$$\Delta^0_{\hat{f},j}(S) = \mathbf{1}_S^\top \tilde{K}^{(j)} \mathbf{1}_S + 2 g^{(j)\top} \mathbf{1}_S, \tag{12}$$

$$\Delta^{\mathrm{w}}_{\hat{f},j}(S) = \Delta^0_{\hat{f},j}(S) - w_S^\top (\tilde{K}^{(j)}_{A'A'})^+ w_S. \tag{13}$$

As before,

$$\Delta_{\hat{f},j}(S) \le \Delta^{\mathrm{w}}_{\hat{f},j}(S) \le \Delta^0_{\hat{f},j}(S).$$

The residual term $g^{(j)}$ appears because $\hat{f}_j$ need not be population optimal. If $\hat{f}_j = f^\star_j$, then $g^{(j)} = 0$, and uniqueness under Assumption 2 gives $\tilde{K}^{(j)} = K^{(j)}$, recovering Proposition 2. Since the reference is the fitted monolith, these excess costs may be negative. Appendix B gives the derivation.

We estimate (11) on the independent fold $\mathcal{D}_2$, writing $\hat{K}^{(j)}$ and $\hat{g}^{(j)}$ for the sample covariances and $\hat{\Delta}^0_{\hat{f},j}(S)$ and $\hat{\Delta}^{\mathrm{w}}_{\hat{f},j}(S)$ for the corresponding plug-in scores. MODEC uses the reweight score for search. The frozen score admits the following simultaneous finite-sample guarantee.

Theorem 1 (Simultaneous held-out bound)**.** *Grant Assumption 1. Let $\mathcal{D}_2$ contain $n_{\mathrm{ep}} \geq 2$ i.i.d. equal-length episodes, independent of $\mathcal{D}_0$ and $\mathcal{D}_1$, and let $\mathfrak{S}$ be a finite cut family fixed before $\mathcal{D}_2$ is observed. Then, conditionally on $\mathcal{D}_0$ and $\mathcal{D}_1$, with probability at least $1 - \delta_{\mathrm{eval}}$, simultaneously for every output j and $S \in \mathfrak{S}$,*

$$\left|\hat{\Delta}^0_{\hat{f},j}(S) - \Delta^0_{\hat{f},j}(S)\right| \leq r_j(S), \qquad \Delta_{\hat{f},j}(S) \leq \hat{\Delta}^0_{\hat{f},j}(S) + r_j(S), \tag{14}$$

where $r_j(S)$ is the episode-level empirical Bernstein radius given in Appendix B; its dependence on the candidate family is logarithmic in $|\mathfrak{S}|$.

The first inequality controls the frozen score, while the second upper-bounds the best reduced-class population cost. Since the event is simultaneous over $\mathfrak{S}$, any $\mathcal{D}_2$-based selection over this family, including the greedy path and fixed-K local search, preserves the bound. Predeclaring the $2^M - 1$ nonempty blocks covers all closed-partition candidates under the fixed actuator rule, with at most one induced cut per block and output. We use the reweight score for search, the frozen score for certification, and Proposition 1 for the final exact refit; Corollary B.1 covers the deployed predictor.

6 Modular Execution

After selecting a decomposition, each module stores the restricted weights for its outputs and reads only retained states and controls; retained cross-module state values are supplied synchronously at the current step.

In our experiments, the learned model predicts the state increment. Let $\tilde{f}$ denote the restricted increment predictor and $\hat{f}$ the original monolithic increment predictor. Starting from the same initial condition and control sequence, their rollouts satisfy $\tilde{y}(t+1) = \tilde{y}(t) + \tilde{f}(\tilde{y}(t), u_t)$ and $\hat{y}(t+1) = \hat{y}(t) + \hat{f}(\hat{y}(t), u_t)$.

Fix a convex operating region $\mathcal{Y} \subseteq \mathbb{R}^{n_y}$. For a partition P, let $\Gamma_\Delta \in \mathbb{R}^{|P|\times|P|}_{\geq 0}$ bound the sensitivity of the restricted increment predictor between modules on $\mathcal{Y}$, so that

$$\|\tilde{f}_k(x,u) - \tilde{f}_k(x',u)\|_2 \leq \sum_{k'} (\Gamma_\Delta)_{kk'} \|x_{k'} - x'_{k'}\|_2.$$

For each module k, let $\varepsilon_k = \sup_{(y,u)\in\mathcal{Y}\times\mathcal{U}} \|\hat{f}_k(y,u) - \tilde{f}_k(y,u)\|_2$ denote the worst-case one-step discrepancy between the monolithic and restricted increment predictors, and define $\Gamma = I + \Gamma_\Delta$. Appendix D gives computable bounds for Γ_Δ and ε from the fitted weights.

Proposition 3 (Modular execution and finite-horizon error)**.** *Fix a decomposition with restricted weights $\tilde{\theta}$, and let all modules update synchronously.*

- *(i)* ***Exact modular execution.*** *Module-by-module evaluation gives the same state update as centralized evaluation of the restricted predictor.*
- *(ii)* ***Finite-horizon error.*** *Let $e_k(t) = \|\tilde{y}_k(t) - \hat{y}_k(t)\|_2$, $e(t) = (e_k(t))_k$, $\varepsilon = (\varepsilon_k)_k$, and $e(0) = 0$. While both rollouts remain in $\mathcal{Y}$, componentwise,*

$$e(t+1) \leq \Gamma e(t) + \varepsilon, \qquad e(t) \leq \sum_{s=0}^{t-1} \Gamma^s \varepsilon. \tag{15}$$

Part (i) shows that distributing a fixed restricted predictor across modules introduces no additional approximation. Part (ii) bounds how the discrepancy between the restricted and monolithic predictors accumulates during rollout. Since $\Gamma = I + \Gamma_\Delta$ with $\Gamma_\Delta \geq 0$, we have $\varrho(\Gamma) \geq 1$, so the bound does not imply infinite-horizon contraction; it provides only the finite-horizon guarantee in (15), without requiring asymptotic stability of the learned dynamics. Rollout diagnostics are reported in the appendix.

7 Numerical Experiments

We evaluate MODEC on three controlled-system benchmarks with complementary roles. The six-reactor continuous stirred-tank reactor (CSTR) benchmark has a known three-module structure and

tunable inter-module coupling, allowing direct tests of physical structure recovery and restriction cost. The Tennessee Eastman process (TEP) provides a realistic seven-unit process for which all 877 closed partitions can be evaluated, allowing comparison with exhaustive selection. The IEEE 39-bus system gives a larger ten-group problem with $115{,}975$ closed partitions, testing the method on a substantially larger model and decomposition space.

Episodes are split into four disjoint folds. We use $\mathcal{D}_0$ for preprocessing and hyperparameter selection, $\mathcal{D}_1$ to fit the coupling-additive predictor, $\mathcal{D}_2$ to compute restriction scores and select decompositions, and $\mathcal{D}_3$ only for final reporting. Thus, neither the decomposition nor its hyperparameters are selected using the reported test errors. We predict standardized state increments $Y^+ = Y_{t+1} - Y_t$ and report $\mathrm{NRMSE} = \sqrt{\sum_j \mathrm{MSE}_j / \sum_j \mathrm{Var}(Y_j^+)}$. Full simulator, excitation, dictionary, grouping, and data-generation details are given in Appendix E.1.

We report the composability $c_{\mathrm{b}}(D)$ of Section 3, the fraction of removable couplings that no module retains; $c_{\mathrm{b}} = 0$ corresponds to the monolithic predictor, and larger values indicate more aggressive decomposition. Prediction errors use the exact restricted ridge refit of Proposition 1. At each K, the score-optimal reference minimizes the total $\mathcal{D}_2$ reweight score, while the refit-selected reference minimizes the exact restricted-refit error on $\mathcal{D}_2$. Influence clustering groups units according to their learned predictive influence, correlation clustering groups units whose measurements vary together, and spectral clustering groups units using a learned similarity matrix. We also compare with a fixed-K local search that directly evaluates exact-refit $\mathcal{D}_2$ cost under the same per-K time budget as MODEC's full decomposition path. Implementation and timing details are given in the appendix.

7.1 Composability–fidelity trade-off

This experiment asks how much cross-module connectivity can be removed before prediction accuracy degrades, and how closely MODEC tracks the best available decompositions. Figure 2 shows the resulting composability–fidelity trade-offs. On CSTR, the planted three-module partition at $\epsilon_{\mathrm{inter}} = 0.1$ removes about 44% of the removable connectivity and yields an NRMSE of 0.16. The same physical partition is recovered throughout the coupling sweep, while the prediction cost increases as the true interaction across module boundaries becomes stronger. On TEP, MODEC reaches $c_{\mathrm{b}} = 0.67$ at $K = 6$ with NRMSE 0.46, matching the refit-selected partition at that operating point; its greedy merge path also reaches the global reweight-score optimum at every K. On the IEEE 39-bus system, the representative $K = 6$ decomposition removes about 84% of the removable connectivity and yields an NRMSE of 0.80. The greedy path reaches the global score optimum for five of the eight nontrivial values $K = 2, \ldots, 9$ and remains within 1.5% of the optimum for the other three. At $K = 6$, the exhaustive score optimum gives essentially the same test error, whereas the refit-selected reference obtains slightly lower NRMSE by retaining more connectivity. On TEP and IEEE, the fixed-K local search reaches the refit-selected solution at every tested $K = 2, \ldots, 6$; on IEEE it matches or slightly improves on the greedy MODEC path over these values (Appendix E.1).

7.2 One-fit ranking of restriction cost

MODEC uses the reweight score to rank candidate restrictions before performing the final exact refit. The key question is therefore whether restrictions assigned low cost from the single fitted model also tend to remain inexpensive after exact refitting. Figure 3 compares the $\mathcal{D}_2$ reweight score with the independently evaluated $\mathcal{D}_3$ cost of exact restricted refitting. The Spearman correlation is $\rho_s = 0.908$ over all 127 TEP candidate blocks, including the full block, and $\rho_s = 0.888$ over the $1{,}022$ nontrivial IEEE blocks. Although the score and exact-refit cost are different quantities, these correlations show that the one-fit score preserves their ranking well enough to provide a useful search signal across both systems. Both the scores and exact refits reuse cached statistics, so no candidate requires a new fit; the score evaluates restrictions in coupling space and is monotone as couplings are restored, so every greedy merge can only reduce it (Remark A.2). Matched score diagnostics, exactness checks, and computational details are reported in the appendix.

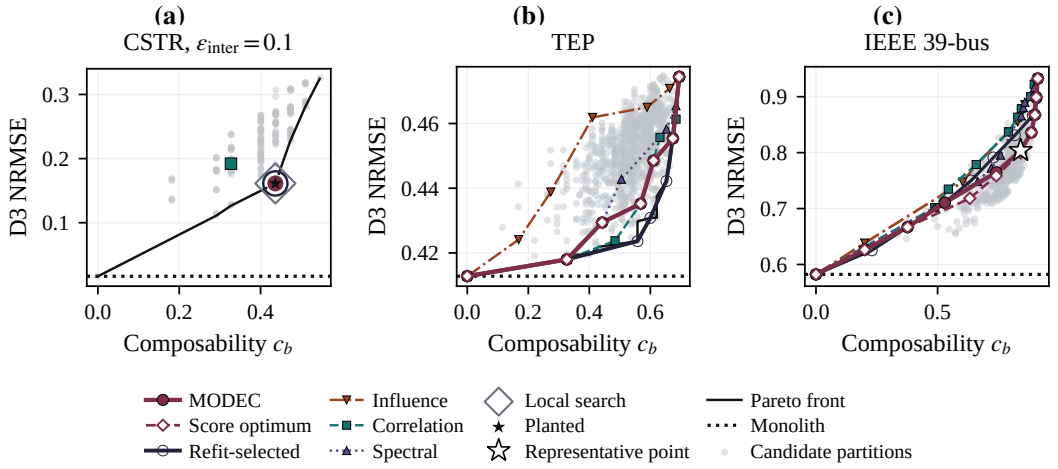


Figure 2: **Composability–fidelity trade-off.** $\mathcal{D}_3$ NRMSE after exact restricted refitting is plotted against composability c_{b}. **(a)** CSTR at $\epsilon_{\text{inter}} = 0.1$, with all 203 partitions shown in gray and the planted partition highlighted. **(b)** TEP, with all 877 closed partitions available for comparison. **(c)** IEEE 39-bus; the score optimum and refit-selected reference are exhaustive over all 115,975 closed partitions, while gray points show a fixed sample for context. All decomposition decisions use $\mathcal{D}_2$; $\mathcal{D}_3$ is used only for reporting.

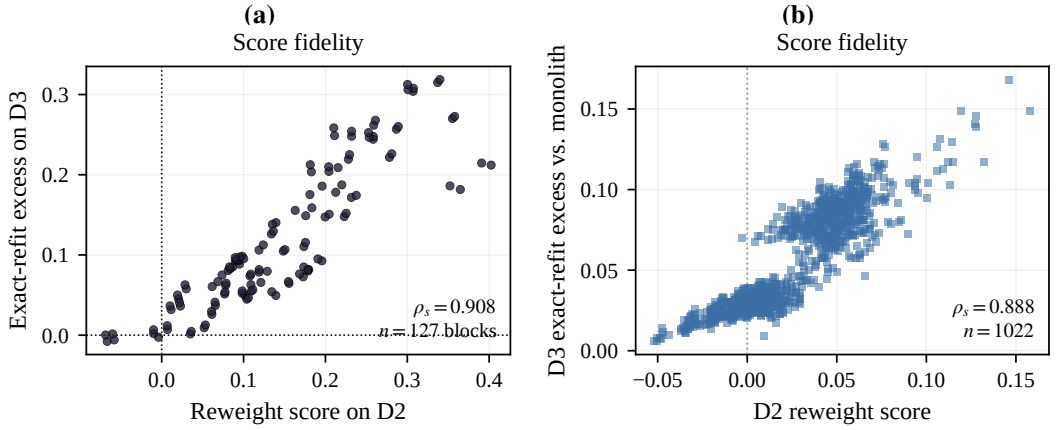


Figure 3: **One-fit ranking of exact-refit restriction cost.** The $\mathcal{D}_2$ reweight score is compared with the independently evaluated $\mathcal{D}_3$ cost of the exact restricted ridge refit. **(a)** TEP: $\rho_s = 0.908$ over all 127 candidate blocks, including the full block. **(b)** IEEE 39-bus: $\rho_s = 0.888$ over all 1,022 nontrivial candidate blocks.

7.3 Recovery of Physical Subsystem Structure

This experiment asks whether MODEC's low-cost decompositions correspond to meaningful physical subsystems rather than merely favorable prediction error. CSTR provides direct ground truth: the planted modules are the reactor pairs $\{1, 2\}$, $\{3, 4\}$, and $\{5, 6\}$. Figure 4 compares the exact one-step plant dependence, MODEC's learned coupling attribution, an unrestricted-MLP gradient baseline, and recovery of the planted $K = 3$ partition over the coupling sweep. MODEC recovers the planted partition at every tested $\epsilon_{\text{inter}} \in \{0, 0.05, 0.10, 0.20, 0.30\}$, giving adjusted Rand index (ARI) $= 1$ throughout. At $\epsilon_{\text{inter}} = 0.1$, its attribution is strongly concentrated within the planted reactor pairs: the cross-to-within-module magnitude ratio is 0.038, compared with 0.548 for the MLP-gradient baseline, while the corresponding within-versus-cross discrimination AUC is 0.972 versus 0.806. The refit-selected reference and spectral clustering also recover the planted partition across the sweep; influence clustering also recovers it at $\epsilon_{\text{inter}} = 0.1$, whereas correlation clustering does not. Additional baseline and pairwise-agreement results are reported in the appendix.

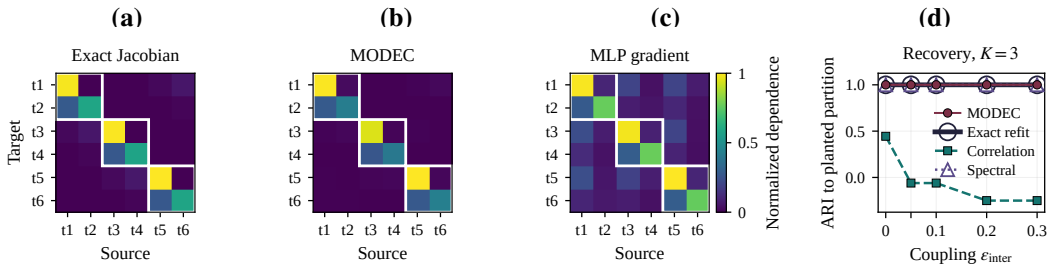


Figure 4: **Recovery of physical modular structure on CSTR. (a)** Exact one-step plant Jacobian, **(b)** MODEC coupling attribution, obtained by aggregating absolute fitted freeze-score Shapley attributions from each source reactor to each target reactor, **(c)** unrestricted-MLP gradient sensitivity at $\epsilon_{\text{inter}} = 0.1$, and **(d)** ARI of the recovered $K = 3$ partition over the coupling sweep. White boxes mark the planted reactor pairs.

Table 1: **Representative operating points.** NRMSE is measured on $\mathcal{D}_3$ after exact restricted refitting. Larger c_{b} means more removable connectivity has been eliminated.

System	Method	K	c_{b}	$\mathcal{D}_3$ NRMSE	ARI
CSTR	Monolith	1	0.000	0.016	–
	MODEC	3	0.436	0.161	1.000
	Refit-selected	3	0.436	0.161	1.000
TEP	Monolith	1	0.000	0.413	–
	MODEC	6	0.674	0.455	–
	Refit-selected	6	0.674	0.455	–
IEEE 39-bus	Monolith	1	0.000	0.582	–
	MODEC	6	0.841	0.804	–
	Score optimum	6	0.841	0.800	–
	Refit-selected	6	0.729	0.791	–

7.4 Representative operating points

Table 1 summarizes representative operating points from the three benchmarks. On CSTR, MODEC recovers the planted $K = 3$ structure exactly (ARI $= 1$) while removing 43.6% of the removable connectivity; when $\epsilon_{\text{inter}} = 0$, the same restriction is essentially lossless (NRMSE 0.0169 versus 0.0167). On TEP, the $K = 6$ decomposition removes 67.4% of the removable connectivity while NRMSE increases from 0.413 for the monolith to 0.455, and MODEC coincides with the refit-selected reference. IEEE presents a harder fidelity trade-off: MODEC removes 84.1% of the removable connectivity at $K = 6$ with NRMSE 0.804, close to the exhaustive score optimum of 0.800 at the same composability; the refit-selected reference obtains NRMSE 0.791 by retaining more connectivity ($c_{\text{b}} = 0.729$). The full comparison with additional structure-selection baselines is reported in the appendix.

8 Conclusion

We introduced MODEC for extracting modular restrictions from a single fitted coupling-additive controlled predictor. For ridge regression, cached statistics give exact restricted solutions, while coupling-level scores rank cuts before exact refitting. Across CSTR, TEP, and IEEE 39-bus, MODEC traces prediction–composability trade-offs, and CSTR recovers the planted subsystem structure across the tested coupling sweep.

Beyond identifying module boundaries, MODEC produces the corresponding restricted predictor rather than only a structural decomposition. Selected restrictions execute exactly as separate modules, with a finite-horizon bound that describes how deviations from the monolithic model accumulate during rollout. This provides a direct path from a monolithic fit to modular deployment.

REPRODUCIBILITY STATEMENT

Proposition 1 uses the ridge objective with $\lambda > 0$, Proposition 2 uses Assumption 2 to define the population coupling components, and Theorem 1 uses Assumption 1 together with independent evaluation episodes. The gain construction for Proposition 3 additionally assumes bounded dictionary Jacobians on the operating region, as stated in Appendix D. Complete proofs are given in Appendices A, B, and D. The estimator is specified by (3) and Algorithm 1, and the supplement provides the CSTR, TEP, and IEEE 39-bus configurations, hyperparameters, groupings, candidate generation, fold assignments, and code used to regenerate the reported figures and tables.

AI USE STATEMENT

In this work, we used generative AI tools for language editing and drafting assistance, literature search and reference checking, coding assistance, checking and explaining algebraic steps and proof arguments, and feedback on experimental design, implementation, result interpretation, figures, and manuscript organization. AI-assisted suggestions were treated as proposals subject to verification: the authors verified references, tested code and numerical outputs, re-derived mathematical arguments, and reviewed all AI-assisted material before inclusion. The authors made all final research decisions and take responsibility for the final content of this work, including all text, claims, code, experiments, results, and artifacts produced with AI assistance.

REFERENCES

Mohamed S Abdelfattah, Abhinav Mehrotra, Łukasz Dudziak, and Nicholas D Lane. Zero-cost proxies for lightweight nas. *arXiv preprint arXiv:2101.08134*, 2021.

Ramachandran Anantharaman and Alexandre Mauroy. Koopman operator based identification of nonlinear networks. *Nonlinear Dynamics*, 113(20):27031–27051, 2025. doi: 10.1007/s11071-025-11560-y.

Murat Arcak, Chris Meissen, and Andrew Packard. *Networks of dissipative systems: compositional certification of stability, performance, and safety*. Springer, 2016.

Alexander Boyd, Franz Nowak, David Hyland, Manuel Baltieri, and Fernando E Rosas. From monoliths to modules: Decomposing transducers for efficient world modelling. *arXiv preprint arXiv:2512.02193*, 2025.

Daniel Bruder, Xun Fu, and Ram Vasudevan. Advantages of bilinear koopman realizations for the modeling and control of systems with unknown dynamics. *IEEE Robotics and Automation Letters*, 6(3):4369–4376, 2021.

Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the national academy of sciences*, 113(15):3932–3937, 2016.

Steven L Brunton, Marko Budišić, Eurika Kaiser, and J Nathan Kutz. Modern koopman theory for dynamical systems. *arXiv preprint arXiv:2102.12086*, 2021.

Han Cai, Chuang Gan, Tianzhe Wang, Zhekai Zhang, and Song Han. Once-for-all: Train one network and specialize it for efficient deployment. *arXiv preprint arXiv:1908.09791*, 2019.

Xiaodong Cheng and Jacquelien MA Scherpen. Model reduction methods for complex network systems. *Annual Review of Control, Robotics, and Autonomous Systems*, 4(1):425–453, 2021.

Joe H Chow. *Power system coherency and model reduction*, volume 84. Springer, 2013.

Panagiotis D Christofides, Riccardo Scattolini, David Munoz De La Pena, and Jinfeng Liu. Distributed model predictive control: A tutorial review and future research directions. *Computers & Chemical Engineering*, 51:21–41, 2013.

Hantao Cui, Fangxing Li, and Kevin Tomsovic. Hybrid symbolic-numeric framework for power system modeling and analysis. *IEEE Transactions on Power Systems*, 36(2):1373–1384, 2020.

Mahyar Fazlyab, Alexander Robey, Hamed Hassani, Manfred Morari, and George Pappas. Efficient and accurate estimation of lipschitz constants for deep neural networks. *Advances in neural information processing systems*, 32, 2019.

Elias Frantar and Dan Alistarh. Optimal brain compression: A framework for accurate post-training quantization and pruning. *Advances in Neural Information Processing Systems*, 35:4475–4488, 2022.

Elias Frantar and Dan Alistarh. Sparsegpt: Massive language models can be accurately pruned in one-shot. In *International conference on machine learning*, pp. 10323–10337. PMLR, 2023.

Anirudh Goyal, Alex Lamb, Jordan Hoffmann, Shagun Sodhani, Sergey Levine, Yoshua Bengio, and Bernhard Schölkopf. Recurrent independent mechanisms. *arXiv preprint arXiv:1909.10893*, 2019.

Yang Guo, Manuel Schaller, Karl Worthmann, and Stefan Streif. Modularized data-driven approximation of the koopman operator and generator. *Physica D: Nonlinear Phenomena*, 476:134651, 2025.

Babak Hassibi and David Stork. Second order derivatives for network pruning: Optimal brain surgeon. *Advances in neural information processing systems*, 5, 1992.

Xinyuan Jiang, Yan Li, and Daning Huang. Modularized bilinear koopman operator for modeling and predicting transients of microgrids. *IEEE Transactions on Smart Grid*, 15(5):5219–5231, 2024.

Mihailo R Jovanović, Peter J Schmid, and Joseph W Nichols. Sparsity-promoting dynamic mode decomposition. *Physics of Fluids*, 26(2), 2014.

Mohamad H Kazm and Ahmad F Taha. Partitioning and observability in linear systems via submodular optimization. *IEEE Transactions on Automatic Control*, 2026.

Milan Korda and Igor Mezić. Linear predictors for nonlinear dynamical systems: Koopman operator meets model predictive control. *Automatica*, 93:149–160, 2018.

Vladimir Kostic, Pietro Novelli, Andreas Maurer, Carlo Ciliberto, Lorenzo Rosasco, and Massimiliano Pontil. Learning dynamical systems via koopman operator regression in reproducing kernel hilbert spaces. *Advances in Neural Information Processing Systems*, 35:4017–4031, 2022.

Eldar Kurtić, Elias Frantar, and Dan Alistarh. Ziplm: Inference-aware structured pruning of language models. *Advances in Neural Information Processing Systems*, 36:65597–65617, 2023.

Yann LeCun, John Denker, and Sara Solla. Optimal brain damage. *Advances in neural information processing systems*, 2, 1989.

Anson Lei, Frederik Nolte, Bernhard Schölkopf, and Ingmar Posner. Compete and compose: Learning independent mechanisms for modular world models. *arXiv preprint arXiv:2404.15109*, 2024.

Yunzhu Li, Hao He, Jiajun Wu, Dina Katabi, and Antonio Torralba. Learning compositional koopman operators for model-based control. *arXiv preprint arXiv:1910.08264*, 2019.

Ričards Marcinkevičs and Julia E Vogt. Interpretable models for granger causality using self-explaining neural networks. *arXiv preprint arXiv:2101.07600*, 2021.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

Joe Mellor, Jack Turner, Amos Storkey, and Elliot J Crowley. Neural architecture search without training. In *International conference on machine learning*, pp. 7588–7598. PMLR, 2021.

Sayak Mukherjee, Sai Pushpak Nandanoori, Sheng Guan, Khushbu Agarwal, Subhrajit Sinha, Soumya Kundu, Seemita Pal, Yinghui Wu, Draguna L Vrabie, and Sutanay Choudhury. Learning distributed geometric koopman operator for sparse networked dynamical systems. In *Learning on Graphs Conference*, pp. 45–1. PMLR, 2022.

Feliks Nüske, Sebastian Peitz, Friedrich Philipp, Manuel Schaller, and Karl Worthmann. Finite-data error bounds for koopman-based prediction and control. *Journal of Nonlinear Science*, 33(1):14, 2023.

Joshua L Proctor, Steven L Brunton, and J Nathan Kutz. Dynamic mode decomposition with control. *SIAM Journal on Applied Dynamical Systems*, 15(1):142–161, 2016.

Simo Puntanen and George PH Styan. Schur complements in statistics and probability. In *The Schur Complement and Its Applications*, pp. 163–226. Springer, 2005.

Alessandro Riccardi, Luca Laurenti, and Bart De Schutter. Partitioning techniques for non-centralized predictive control: A systematic review and novel theoretical insights. *Annual Reviews in Control*, 61:101046, 2026.

Corbinian Schlosser and Milan Korda. Sparsity structures for koopman and perron–frobenius operators. *SIAM Journal on Applied Dynamical Systems*, 21(3):2187–2214, 2022.

Dragoslav D Siljak. *Decentralized control of complex systems*. Courier Corporation, 2011.

Mingjie Sun, Zhuang Liu, Anna Bair, and Zico Kolter. A simple and effective pruning approach for large language models. In *International Conference on Learning Representations*, volume 2024, pp. 4942–4964, 2024.

Amit Surana. Koopman operator based observer synthesis for control-affine nonlinear systems. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pp. 6492–6499. IEEE, 2016.

Wentao Tang and Prodromos Daoutidis. Network decomposition for distributed control through community detection in input–output bipartite graphs. *Journal of Process Control*, 64:7–14, 2018.

Alex Tank, Ian Covert, Nicholas Foti, Ali Shojaie, and Emily B Fox. Neural granger causality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4267–4279, 2021.

Matthew O Williams, Ioannis G Kevrekidis, and Clarence W Rowley. A data–driven approximation of the koopman operator: Extending dynamic mode decomposition. *Journal of Nonlinear Science*, 25(6):1307–1346, 2015.

Xunyuan Yin, Yan Qin, Hongtian Chen, Wenli Du, Jinfeng Liu, and Biao Huang. Community detection based process decomposition and distributed monitoring for large-scale processes. *AIChE Journal*, 68(11):e17826, 2022.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(1):49–67, 2006.

Songyuan Zhang, Yumeng Xiu, Guannan Qu, and Chuchu Fan. Compositional neural certificates for networked dynamical systems. In *Learning for Dynamics and Control Conference*, pp. 272–285. PMLR, 2023.

Overview of the Appendix

The appendix follows the organization of the main text and provides the technical details, implementation material, and supporting experiments omitted there due to space constraints. Appendix A contains the proofs and supporting results for the restriction and scoring results of Section 4, including score properties, attribution, and model-class limitations. Appendix B derives the fitted-model restriction costs and finite-sample bounds of Section 5. Appendix C gives the full MODEC procedure and its computational details, while Appendix D provides the gain construction and proofs underlying the modular-execution result of Section 6.

Appendix E provides the benchmark configurations, baselines, and additional results supporting Section 7. It includes extended CSTR composability and structure-recovery studies, score and finite-horizon diagnostics, modular execution and rollout results for TEP and IEEE 39-bus, additional score checks and modeling ablations, and computation time measurements.

A Proofs for Restriction and Scoring

A.1 Exact restricted refit

Proof of Proposition 1. Proof idea. Since $\hat{\theta}$ minimizes the quadratic J_j, the extra objective of any θ is a quadratic form in $\theta - \hat{\theta}$, and restricted retraining minimizes this form with the removed coordinates forced from $\hat{\theta}_S$ to zero; solving directly gives (4), and solving with a Lagrange multiplier gives (5).

Since $\lambda > 0$, $H = \hat{G} + \lambda I \succ 0$, so H_{AA} and $(H^{-1})_{SS}$ are positive definite and invertible, and

$$J_j(\theta) - J_j(\hat{\theta}) = (\theta - \hat{\theta})^\top H(\theta - \hat{\theta}).$$

Let $\delta = \theta - \hat{\theta}$. The restriction $\theta_S = 0$ fixes $\delta_S = -\hat{\theta}_S$. Minimizing the quadratic form over the surviving coordinates gives

$$\delta_A = -H_{AA}^{-1} H_{AS} \delta_S = H_{AA}^{-1} H_{AS} \hat{\theta}_S,$$

and hence

$$\tilde{\theta}_A = \hat{\theta}_A + H_{AA}^{-1} H_{AS} \hat{\theta}_S.$$

Alternatively, let E_S select the coordinates in S. Minimizing $\delta^\top H \delta$ subject to $E_S^\top \delta = -\hat{\theta}_S$ gives the first-order condition $H\delta + E_S \mu = 0$, hence $\delta = -H^{-1} E_S \mu$. Enforcing the constraint gives

$$\mu = \left[(H^{-1})_{SS}\right]^{-1} \hat{\theta}_S,$$

and therefore

$$\tilde{\theta} = \hat{\theta} - (H^{-1})_{\cdot S} \left[(H^{-1})_{SS}\right]^{-1} \hat{\theta}_S,$$

which is (5). Substitution gives

$$J_j(\tilde{\theta}) - J_j(\hat{\theta}) = \hat{\theta}_S^\top \left(H_{SS} - H_{SA} H_{AA}^{-1} H_{AS}\right) \hat{\theta}_S.$$

Finally, the block-inverse identity

$$H_{SS} - H_{SA} H_{AA}^{-1} H_{AS} = \left[(H^{-1})_{SS}\right]^{-1}$$

gives the second expression in (6). □

Remark A.1 (Empirical refit versus population scores). Proposition 1 concerns the empirical ridge objective on $\mathcal{D}_1$ and exactly reproduces constrained retraining on that fold. The freeze and reweight quantities instead describe prediction risk under the data distribution. Their orderings therefore must not be combined with the empirical criterion increase of Proposition 1.

A.2 Freeze, reweight, and interaction

Proof of Proposition 2. Proof idea. After the error of $f^\star$ is subtracted, the excess risk of a predictor in $\mathcal{F}$ is its squared distance from $f^\star$. Removing S displaces the predictor by c_S; freezing can remove only the mean of c_S, and reweighting can also remove its linear regression on the surviving contributions.

Fix one output and write

$$r = m - f^\star, \qquad c_S = \sum_{a \in S} f_a^\star.$$

Because $f^\star$ is the $L^2(\rho)$ projection of m onto $\mathcal{F}$, r is orthogonal to every function in $\mathcal{F}$; the constants belong to $\mathcal{F}$ through the block $(0, *)$. For any square-integrable predictor f of (Y, U), $R(f) = \mathbb{E}[(f - m)^2] + \mathbb{E}[(Y^+ - m)^2]$, because the cross term $2\mathbb{E}[(f - m)(m - Y^+)]$ vanishes after conditioning on (Y, U). The noise term $\mathbb{E}[(Y^+ - m)^2]$ does not depend on f, so it cancels in every excess risk, which can therefore be computed with m in place of Y^+.

For the frozen restriction,

$$f^\star - c_S + \beta,$$

the excess risk is therefore

$$\mathbb{E}[(r + c_S - \beta)^2] - \mathbb{E}[r^2] = \mathbb{E}[(c_S - \beta)^2].$$

Optimizing over the intercept gives

$$\Delta^0(S) = \mathrm{Var}(c_S) = \mathbf{1}_S^\top K \mathbf{1}_S.$$

For reweighting, the predictor $f^\star_{(0,*)} + \sum_{a\in A'} \delta_a f^\star_a + \beta$ has excess risk $\mathbb{E}[(c_S + \sum_{a\in A'} \eta_a f^\star_a - \beta)^2]$ with $\eta_a = 1 - \delta_a$. Minimizing over β leaves $\mathrm{Var}(c_S + \eta^\top f^\star_{A'})$ with $f^\star_{A'} = (f^\star_a)_{a\in A'}$, a least-squares problem with normal equations $K_{A'A'}\eta = -K_{A'\cdot}\mathbf{1}_S$. These are solvable even if $K_{A'A'}$ is singular: if $K_{A'A'}v = 0$, then $v^\top f^\star_{A'}$ is almost surely constant, so $v^\top K_{A'\cdot}\mathbf{1}_S = \mathrm{Cov}(v^\top f^\star_{A'}, c_S) = 0$. Taking $\eta = -(K_{A'A'})^+ K_{A'\cdot}\mathbf{1}_S$ gives

$$\Delta^{\mathrm{w}}(S) = \mathbf{1}_S^\top \big(K - K_{\cdot A'} K^+_{A'A'} K_{A'\cdot}\big)\mathbf{1}_S.$$

Finally, the admissible predictor classes are nested: freezing is a special case of reweighting, and every reweighted predictor belongs to the full reduced class. Hence

$$\Delta(S) \le \Delta^{\mathrm{w}}(S) \le \Delta^0(S).$$

□

Lemma A.1 (Bracketing by the diagonal of K)**.** *Under Assumption 2, $\vartheta \sum_{a\in S} K_{aa} \le \Delta(S) \le \Delta^0(S) \le |S| \sum_{a\in S} K_{aa}$ for every $S \subseteq \mathcal{A}_0$.*

Proof. Proof idea. Every reduced predictor must drop each $f^\star_a$ with $a \in S$, and Assumption 2 prevents the remaining blocks from canceling these components by more than the factor ϑ; the upper bound is the freeze cost, which is the variance of a sum.

With $r = m - f^\star$ as in the preceding proof, $r \perp \mathcal{F}$ gives $\Delta(S) = \min_{f\in\mathcal{F}_A} \|f^\star - f\|^2_{L^2(\rho)}$ with $A = \mathcal{A} \setminus S$. Writing $f = \sum_{a\in A} f_a$ with $f_a \in \mathcal{H}_a$ gives $f^\star - f = \sum_{a\in\mathcal{A}} c_a$ with $c_a = f^\star_a$ for $a \in S$ and $c_a = f^\star_a - f_a \in \mathcal{H}_a$ for $a \in A$. Applying Assumption 2 to $(c_a)_{a\in\mathcal{A}}$ and dropping the terms with $a \in A$ yields $\|f^\star - f\|^2 \ge \vartheta \sum_{a\in S} \mathbb{E}[(f^\star_a)^2] \ge \vartheta \sum_{a\in S} K_{aa}$ for every $f \in \mathcal{F}_A$, hence for the minimum. For the upper bound, $\Delta(S) \le \Delta^0(S)$ by Proposition 2, and $\Delta^0(S) = \mathrm{Var}(\sum_{a\in S} f^\star_a) \le \big(\sum_{a\in S} \sqrt{K_{aa}}\big)^2 \le |S| \sum_{a\in S} K_{aa}$, using $\mathrm{Cov}(f^\star_a, f^\star_b) \le \sqrt{K_{aa}K_{bb}}$ and then the Cauchy–Schwarz inequality. □

Remark A.2 (Monotonicity). If $S' \subseteq S$, the feasible reweighted and fully refit predictor classes for S' contain those for S. Hence $\Delta^{\mathrm{w}}(S') \le \Delta^{\mathrm{w}}(S)$ and $\Delta(S') \le \Delta(S)$. The same nesting argument applies to the $\mathcal{D}_2$ plug-in reweight criterion, so the empirical merge objective is also nonincreasing as removed couplings are restored. The freeze score need not be monotonic because negatively correlated contributions can cancel out. We therefore use reweighting as the merge objective in Algorithm 1.

A.3 Why the coupling covariance does not determine full-refit cost

Let X, X', Y be independent standard normal variables and define

$$Z = X\cos\alpha + X'\sin\alpha, \qquad \alpha \in (0, \pi/2].$$

Take the prediction target

$$m = X + Y.$$

Let coupling 1 contain the feature X, and let coupling 2 contain the features (Y, Z). Since m lies in the full class, the population components are

$$f^\star_1 = X, \qquad f^\star_2 = Y,$$

and they are unique. Indeed, for $c_1 = aX$ and $c_2 = bY + cZ$ we have $\|c_1\|^2 = a^2$, $\|c_2\|^2 = b^2 + c^2$, and $|\langle c_1, c_2\rangle| = |ac|\cos\alpha \le \cos\alpha\ \|c_1\|\,\|c_2\|$, so

$$\begin{aligned}\|c_1 + c_2\|^2 &\ge \|c_1\|^2 + \|c_2\|^2 - 2\cos\alpha\ \|c_1\|\,\|c_2\| \\ &\ge (1 - \cos\alpha)\big(\|c_1\|^2 + \|c_2\|^2\big),\end{aligned}$$

where the second step uses $\cos\alpha \geq 0$ and $2\|c_1\|\|c_2\| \leq \|c_1\|^2 + \|c_2\|^2$. The intercept block is orthogonal to both couplings because all variables are centered, so Assumption 2 holds with $\vartheta = 1 - \cos\alpha > 0$. The coupling covariance is $K = I$ for every α, so cutting coupling 1 always has freeze cost 1; the reweight cost is also 1 because the survivor $f_2^\star = Y$ is uncorrelated with X. After the cut, coupling 2 still contains Z, so the full reduced refit predicts

$$Y + \cos\alpha\, Z$$

and has excess risk $\sin^2\alpha$. Thus the same coupling covariance K is compatible with different full-refit costs, showing that K determines the freeze and reweight scores but not the full-refit cost.

A.4 Shapley attribution

Corollary A.1 (Exact Shapley attribution of the freeze score)**.** *For the cooperative game $v(S) = \Delta^0(S)$ on the removable couplings $\mathcal{A}_0$, the Shapley value of coupling a is*

$$\phi_a = \sum_{b\in\mathcal{A}_0} K_{ab} = \mathrm{Cov}(f_a^\star, f^\star).$$

Thus all coupling attributions are obtained directly from K without permutation sampling, and they satisfy

$$\sum_{a\in\mathcal{A}_0} \phi_a = v(\mathcal{A}_0).$$

Proof. From $v(S) = \mathbf{1}_S^\top K \mathbf{1}_S$, adding coupling a to S changes the value by

$$v(S\cup\{a\}) - v(S) = K_{aa} + 2\sum_{b\in S} K_{ab}.$$

In a uniformly random ordering, every $b \neq a$ precedes a with probability $1/2$. Averaging the marginal contribution therefore gives

$$\phi_a = K_{aa} + \sum_{b\neq a} K_{ab} = \sum_b K_{ab}.$$

Since $f^\star = \sum_b f_b^\star$ up to the constant intercept, this equals $\mathrm{Cov}(f_a^\star, f^\star)$. Summing over a gives $\mathbf{1}^\top K \mathbf{1} = v(\mathcal{A}_0)$. □

The attribution applies to the freeze cost and not to the full-refit risk. Because off-diagonal entries of K may be negative, the Shapley values can also be negative, indicating that a coupling participates in cancellation with other contributions. For the fitted freeze game $\hat{v}_j(S) = \hat{\Delta}^0_{\hat{f},j}(S) = \mathbf{1}_S^\top \hat{K}^{(j)} \mathbf{1}_S + 2\hat{g}^{(j)\top}\mathbf{1}_S$ computed on $\mathcal{D}_2$, the linear term is additive, so the same argument gives $\hat{\phi}_{j,a} = \sum_{b\in\mathcal{A}_0} \hat{K}^{(j)}_{ab} + 2\hat{g}^{(j)}_a$, again with $\sum_{a\in\mathcal{A}_0}\hat{\phi}_{j,a} = \hat{v}_j(\mathcal{A}_0)$.

A.5 Misspecification and interaction couplings

The restriction scores describe the best predictor within the chosen coupling-additive class. They need not reveal mechanisms absent from that class. For example, suppose y_g and $y_{g'}$ are scalar and centered, y_g, $y_{g'}$, and all remaining inputs are mutually independent, no coupling reads both groups, and an output depends on them only through $m = y_g y_{g'}$. Each coupling function depends on at most one of $y_g, y_{g'}$ together with the remaining inputs, so by independence its product with m has zero mean. Hence m is orthogonal to the coupling-additive class and its population projection is zero: this mechanism contributes nothing to $f^\star$ or to any population restriction cost, although it couples the two groups. Fitted plug-in scores on finite data can still be nonzero because of estimation error.

Remark A.3 (Interaction couplings). We therefore compare the structured monolith with a more flexible, unrestricted predictor to serve as a misspecification diagnostic. When needed, pairwise terms $q_i(u)c(y_g, y_{g'})$ can be introduced as additional coupling blocks. All restriction and scoring identities then apply unchanged to the enlarged feature set.

B HELD-OUT CERTIFICATION

This section gives the fitted-model identities and the explicit finite-sample radius used in Theorem 1.

B.1 FITTED-MODEL RESTRICTION COSTS

Throughout this subsection $\mathcal{D}_0$ and $\mathcal{D}_1$ are fixed, and all expectations are under ρ. For output j write $\hat{f} = \hat{f}_j$, $\hat{f}_a = \hat{f}_{j,a}$, $\hat{r} = Y_j^+ - \hat{f}(Y,U)$, and $\hat{c}_S = \sum_{a\in S} \hat{f}_a$, and recall $\tilde{K}$ and g from (11) and the monolith-relative costs of Section 5, whose reference risk is $\bar{R}_j = \mathrm{Var}(\hat{r})$.

Proposition B.1 (Monolith-relative freeze and reweight costs)**.** *For every output j and $S \subseteq \mathcal{A}_0$, the identities* (12)–(13) *hold with $w_S = (\tilde{K}^{(j)}\mathbf{1}_S)_{A'} + g^{(j)}_{A'}$, and $\Delta_{\hat{f},j}(S) \le \Delta^{\mathrm{w}}_{\hat{f},j}(S) \le \Delta^{0}_{\hat{f},j}(S)$. Under Assumption 2, when $\hat{f}_j = f^\star_j$, uniqueness gives $\hat{f}_{j,a} = f^\star_{j,a}$ for every a; the residual covariance then vanishes and the identities reduce to Proposition 2. The plug-in of* (12)–(13) *with sample covariances on a fold is the criterion increase on that fold of the corresponding restriction with its intercept and scalars refit there.*

Proof. Proof idea. The argument is that of Proposition 2 with $\hat{f}$ in place of $f^\star$; because $\hat{r}$ need not be orthogonal to the fitted contributions, its covariance g with them appears as an extra term.

$\min_\beta R_j(\hat{f} - \hat{c}_S + \beta) = \min_\beta \mathbb{E}[(\hat{r} + \hat{c}_S - \beta)^2] = \mathrm{Var}(\hat{r} + \hat{c}_S)$; subtracting $\bar{R}_j = \mathrm{Var}(\hat{r})$ and expanding by bilinearity gives the first identity. For reweight write $\sum_{a\in A'} \delta_a \hat{f}_a + \beta = \hat{f} - \hat{c}_S - \sum_{a\in A'} \eta_a \hat{f}_a + \beta'$ with $\eta_a = 1 - \delta_a$; then the risk is $\mathbb{E}[(\hat{r} + \hat{c}_S + \sum_{a\in A'} \eta_a \hat{f}_a - \beta')^2]$. After centering the surviving contributions, the normal equations give

$$\eta = -(\tilde{K}^{(j)}_{A'A'})^+ w_S.$$

Substituting this minimizer gives

$$\mathrm{Var}(\hat{r} + \hat{c}_S) - w_S^\top (\tilde{K}^{(j)}_{A'A'})^+ w_S,$$

because $w_S = \mathrm{Cov}(\hat{f}_{A'}, \hat{r} + \hat{c}_S)$, and subtracting $\mathrm{Var}(\hat{r})$ gives the second identity. The normal equations are consistent because w_S lies in the range of $\tilde{K}^{(j)}_{A'A'}$: if $\tilde{K}^{(j)}_{A'A'} v = 0$, then $v^\top \hat{f}_{A'}$ is almost surely constant, so $v^\top w_S = 0$. The freeze predictors are the reweight predictors with $\eta = 0$, and the reweight predictors lie in $\mathcal{F}_A$, which gives the ordering. Under Assumption 2, $\hat{f} = f^\star_j$ also gives $\hat{f}_a = f^\star_{j,a}$ for every a by uniqueness, so $\tilde{K} = K^{(j)}$. Moreover $\hat{r} = (Y_j^+ - m_j) + (m_j - f^\star_j)$ is orthogonal to $\mathcal{F}$, the first term because it has zero conditional mean given (Y,U) and the second by the definition of $f^\star_j$; since $\mathcal{F}$ contains the constants, $\mathbb{E}\hat{r} = 0$, hence $g = 0$ and $\bar{R}_j = R_j(f^\star_j)$. The last claim follows by repeating the argument with ρ replaced by the empirical distribution of the fold, whose covariances are the $1/n$-normalized sample covariances. □

The radius. Let $\mathcal{D}_2$ consist of n_{ep} i.i.d. episodes of T transitions each. For a function h of one transition taking values in an interval of width W_h, let $\bar{a}_\ell(h) = T^{-1}\sum_{t\in\mathcal{T}_\ell} h(Z_t)$ be its average over episode ℓ, where $\mathcal{T}_\ell$ indexes the transitions of episode ℓ, and let $\hat{m}(h)$ and $\hat{V}(h)$ be the mean and the Bessel-corrected sample variance of these n_{ep} episode averages. For $a = (i,g)$ set $\hat{B}_{j,a} = \bar{q}_i \max\{1, b_\varphi\}\|\hat{\theta}_{j,a}\|_2$ with $\bar{q}_i = \sup_{u\in\mathcal{U}} |q_i(u)|$, so that $\sup\left|\hat{f}_{j,a}\right| \le \hat{B}_{j,a}$, and put

$$R_c = \sum_{a\in S} \hat{B}_{j,a}, \qquad R_r = b_y + \left|\hat{\theta}_{j,(0,*)}\right| + \sum_{a\in\mathcal{A}_0} \hat{B}_{j,a}, \qquad \phi_S = \hat{c}_S(\hat{c}_S + 2\hat{r}_j),$$

so that $W_{\hat{c}_S} = 2R_c$, $W_{\hat{r}_j} = 2R_r$, and $W_{\phi_S} = R_c^2 + 4R_rR_c$ are valid widths. With

$$\tau(h) = \sqrt{\frac{2\hat{V}(h)L}{n_{\mathrm{ep}}}} + \frac{7W_hL}{3(n_{\mathrm{ep}}-1)}, \qquad L = \log\frac{4n_y(2|\mathfrak{S}|+1)}{\delta_{\mathrm{eval}}},$$

the radius of Theorem 1 is

$$r_j(S) = \tau(\phi_S) + \tau(\hat{c}_S)\big(2|\hat{m}(\hat{c}_S)| + \tau(\hat{c}_S)\big) + 2\big(|\hat{m}(\hat{c}_S)|\,\tau(\hat{r}_j) + |\hat{m}(\hat{r}_j)|\,\tau(\hat{c}_S) + \tau(\hat{r}_j)\tau(\hat{c}_S)\big). \tag{16}$$

Every quantity in (16) is computable from $\mathcal{D}_1$, $\mathcal{D}_2$, and b_y, and $\tau(\hat{r}_j)$ is shared by all candidates.

Proof of Theorem 1. Proof idea. The frozen cost is a fixed combination of three means, $\mathbb{E}\phi_S$, $\mathbb{E}\hat{c}_S$, and $\mathbb{E}\hat{r}_j$, and the plug-in is the same combination of sample means. Each sample mean is an average of independent episode averages, so the empirical Bernstein inequality and a union bound control all of them simultaneously, and the errors are then propagated through the combination.

By (12),$\Delta^0_{\hat{f},j}(S) = \mathbb{E}\phi_S - \mathbb{E}\hat{c}_S(\mathbb{E}\hat{c}_S + 2\mathbb{E}\hat{r}_j)$. Because episodes have equal length, the plug-in $\hat{\Delta}^0_{\hat{f},j}(S)$, computed with $1/|\mathcal{D}_2|$-normalized sample covariances, equals $\hat{m}(\phi_S) - \hat{m}(\hat{c}_S)(\hat{m}(\hat{c}_S) + 2\hat{m}(\hat{r}_j))$. Conditionally on $\mathcal{D}_0$ and $\mathcal{D}_1$, the episode averages $\bar{a}_\ell(h)$ are independent and identically distributed with mean $\mathbb{E}_\rho h$ and values in an interval of width W_h, so the empirical Bernstein inequality (Maurer & Pontil, 2009, Theorem 4, applied two-sidedly) gives $|\hat{m}(h) - \mathbb{E}h| \le \tau(h)$ with probability at least $1-\delta'$ when $L = \log(4/\delta')$. Apply it to the $n_y(2\,|\mathfrak{S}|+1)$ functions ϕ_S and $\hat{c}_S$ for $S \in \mathfrak{S}$ and $\hat{r}_j$, for every output, at $\delta' = \delta_{\mathrm{eval}}/(n_y(2\,|\mathfrak{S}|+1))$, and take a union bound. On the intersection, with $\mu_c = \mathbb{E}\hat{c}_S$, $\mu_r = \mathbb{E}\hat{r}_j$, $\tau_c = \tau(\hat{c}_S)$, $\tau_r = \tau(\hat{r}_j)$, we have $\left|\hat{m}(\hat{c}_S)^2 - \mu_c^2\right| \le \tau_c(2\,|\hat{m}(\hat{c}_S)| + \tau_c)$ and $|\hat{m}(\hat{c}_S)\hat{m}(\hat{r}_j) - \mu_c\mu_r| \le |\hat{m}(\hat{c}_S)|\,\tau_r + (|\hat{m}(\hat{r}_j)| + \tau_r)\tau_c$; adding $|\hat{m}(\phi_S) - \mathbb{E}\phi_S| \le \tau(\phi_S)$ gives (16). The second inequality in (14) follows from the ordering in Proposition B.1. The widths follow from $\left|\hat{f}_{j,a}\right| \le \hat{B}_{j,a}$ and $\left|Y_j^+\right| \le b_y$ under Assumption 1, which give $|\hat{c}_S| \le R_c$ and $|\hat{r}_j| \le R_r$, hence $\phi_S = \hat{c}_S^2 + 2\hat{c}_S\hat{r}_j \in [-2R_rR_c,\, R_c^2 + 2R_rR_c]$. They depend only on $\mathcal{D}_0$, $\mathcal{D}_1$, and the predeclared constants of Assumption 1, so they are fixed before $\mathcal{D}_2$ is observed. □

Corollary B.1 (Bound for a deployed restriction)**.** *In the setting of Theorem 1, let $p_{j,S} = \langle \theta'_{j,S}, \psi \rangle$ be any predictor whose weights are determined by $\mathcal{D}_0$ and $\mathcal{D}_1$, and hence fixed before $\mathcal{D}_2$ is observed; in particular, $p_{j,S}$ may be the exactly refit predictor of Proposition 1. Define*

$$d_{j,S} = \hat{f}_j - p_{j,S}, \qquad \chi_{j,S} = d_{j,S}(d_{j,S} + 2\hat{r}_j).$$

Then

$$R_j(p_{j,S}) - R_j(\hat{f}_j) = \mathbb{E}[\chi_{j,S}].$$

Moreover, let b_ψ be any finite bound satisfying

$$\|\psi(Y,U)\|_2 \le b_\psi \qquad \rho\text{-almost surely},$$

which is computable from Assumption 1 and the chosen dictionary, and define

$$R_{d,j,S} = b_\psi \|\hat{\theta}_j - \theta'_{j,S}\|_2.$$

Then $\chi_{j,S}$ has range width at most

$$W_{j,S} = R_{d,j,S}^2 + 4R_rR_{d,j,S}.$$

For any finite family $\mathfrak{S}$ fixed before $\mathcal{D}_2$ is observed, let

$$L_{\mathrm{dep}} = \log \frac{4n_y|\mathfrak{S}|}{\delta_{\mathrm{dep}}},$$

and define

$$r_j^{\mathrm{dep}}(S) = \sqrt{\frac{2\hat{V}(\chi_{j,S})L_{\mathrm{dep}}}{n_{\mathrm{ep}}}} + \frac{7W_{j,S}L_{\mathrm{dep}}}{3(n_{\mathrm{ep}}-1)}.$$

Conditionally on $\mathcal{D}_0$ and $\mathcal{D}_1$, with probability at least $1 - \delta_{\mathrm{dep}}$, simultaneously for every output j and $S \in \mathfrak{S}$,

$$\left|\hat{m}(\chi_{j,S}) - \big(R_j(p_{j,S}) - R_j(\hat{f}_j)\big)\right| \le r_j^{\mathrm{dep}}(S).$$

Proof. Proof idea. The deployed excess risk is the mean of one fixed bounded function $\chi_{j,S}$, so the concentration argument of Theorem 1 applies directly, without error propagation.

Since $Y_j^+ - p_{j,S} = \hat{r}_j + d_{j,S}$,

$$R_j(p_{j,S}) - R_j(\hat{f}_j) = \mathbb{E}[(\hat{r}_j + d_{j,S})^2 - \hat{r}_j^2] = \mathbb{E}[\chi_{j,S}].$$

Since $|d_{j,S}| \le R_{d,j,S}$ by Cauchy–Schwarz and $|\hat{r}_j| \le R_r$, $\chi_{j,S} = d_{j,S}^2 + 2d_{j,S}\hat{r}_j$ lies in $[-2R_rR_{d,j,S},\, R_{d,j,S}^2 + 2R_rR_{d,j,S}]$, an interval of width $W_{j,S}$. Once $\mathcal{D}_0$ and $\mathcal{D}_1$ are fixed, $\chi_{j,S}$ is a fixed bounded function of one transition. The stated result therefore follows from the same episode-level empirical Bernstein inequality used in the proof of Theorem 1, followed by a union bound over outputs and candidates. □

Algorithm 1 MODEC from one fitted controlled predictor

Require: fitting fold $\mathcal{D}_1$, independent evaluation episodes $\mathcal{D}_2$, deployment grouping $\mathcal{G}_{\mathrm{dep}}$, feature blocks $\mathcal{G}_{\mathrm{feat}}$, dictionaries $\{\varphi_g\}$, ridge parameter λ, candidate cut family $\mathfrak{S}$, confidence δ_{eval}
Ensure: scored decomposition path and restricted module weights

Stage 1: fit once
build ψ on $\mathcal{D}_1$ and solve (3) for all outputs
cache a factorization of $H_\lambda = \hat{G} + \lambda I$
Stage 2: summarize on held-out episodes
evaluate $\hat{f}_{j,a}$ and residuals $\hat{r}_j$ on $\mathcal{D}_2$
form $\hat{K}^{(j)}$ and $\hat{g}^{(j)}$
compute the freeze radii $r_j(S)$ for every $S \in \mathfrak{S}$
Stage 3: score candidate restrictions
FREEZE$(j, S) \leftarrow \hat{\Delta}^0_{\hat{f},j}(S)$ and $r_j(S)$
REWEIGHT$(j, S) \leftarrow \hat{\Delta}^{\mathrm{w}}_{\hat{f},j}(S)$
Stage 4: trace a decomposition path
initialize P as the singleton partition and let D be its decomposition
while $|P| > 1$ **do**
    evaluate the decrease in the total reweight score for every candidate merge
    merge the pair giving the largest decrease
    record the resulting decomposition, connectivity, scores, and freeze certificate
end while
retain the decompositions that are nondominated in composability and total reweight score
Stage 5: deploy
for each selected decomposition **do**
    compute its restricted weights using Proposition 1
    form the gain matrix Γ using Lemma D.1
end for

The radius is data-driven through the between-episode variance $\hat{V}(\cdot)$, so arbitrary dependence within each episode is allowed, and no mixing coefficient is required. The effective sample size is therefore the number of independent episodes. Larger cuts have wider ranges and hence looser bounds. If only one continuous trajectory is available, the theorem does not apply as stated.

Corollary B.2 (Decomposition-level bound)**.** *Let $\mathfrak{D}$ be any family of decompositions whose output-wise cut sets belong to $\mathfrak{S}$, and define $\Delta R_{\hat{f}}(D) = \sum_{j=1}^{n_y} \Delta_{\hat{f},j}(S_j(D))$. On the event of Theorem 1, simultaneously for every $D \in \mathfrak{D}$,*

$$\Delta R_{\hat{f}}(D) \le \sum_{j=1}^{n_y} \left[\hat{\Delta}^0_{\hat{f},j}(S_j(D)) + r_j(S_j(D)) \right].$$

Thus the candidate-wise bounds from Theorem 1 also give a simultaneous upper envelope for the monolith-relative decomposition costs.

Proof. Sum (14) over outputs and use $\Delta_{\hat{f},j}(S) \le \Delta^0_{\hat{f},j}(S)$ from Proposition B.1. □

Corollary B.2 bounds the evaluation of candidates and leaves search optimality open; Algorithm 1 explores the full candidate family only when enumeration is feasible.

C SEARCH AND IMPLEMENTATION

Merge objective. We use the reweight score as the agglomerative objective because its feasible predictor class expands whenever a removed coupling is restored, so every merge can only decrease its cost (Remark A.2). The freeze score lacks this monotonicity when removed contributions are negatively correlated.

Complexity. The covariance scores operate on the N coupling blocks, whereas an exact refit operates on feature coordinates. For a cut S, let $d_S = \sum_{a \in S} \dim(\theta_{j,a})$ denote the number of removed feature coordinates, which can be much larger than the number $|S|$ of removed coupling blocks. A freeze score requires $O(|S|^2)$ operations once $\hat{K}$ is available. Reweighting requires solving a system over the surviving couplings, with a worst-case cost of $O(N^3)$. Proposition 1 instead solves over feature coordinates and can approach the cost of a direct reduced ridge solve for large cuts. We therefore use the covariance scores to search many candidates and apply the exact refit only to selected decompositions.

D Recomposition Theory

Lemma D.1 (Module gains)**.** *Assume each feature block is continuously differentiable on the convex operating region $\mathcal{Y}$. For a feature block $g \in \mathcal{G}_{\mathrm{feat}}$, let $\mathrm{reads}(g) \subseteq \mathcal{G}_{\mathrm{dep}}$ denote the deployment groups whose variables are used by φ_g. For every $h \in \mathrm{reads}(g)$, assume*

$$\sup_{y \in \mathcal{Y}} \left\| \frac{\partial \varphi_g}{\partial y_h}(y_g) \right\|_{\mathrm{op}} \le L_g^{(h)},$$

and assume

$$\sup_{u \in \mathcal{U}} \max_i |q_i(u)| \le \bar{q}.$$

For output j and deployment group h, define

$$\gamma_{j,h} = \bar{q} \sum_{g:\, h \in \mathrm{reads}(g)} L_g^{(h)} \sum_{i:(i,g) \in A_j} \|\tilde{\theta}_{j,(i,g)}\|_2.$$

For modules k, k', define

$$(\Gamma_\Delta)_{kk'} = \left(\sum_{j \in \mathcal{B}_k} \sum_{h \in \mathcal{B}_{k'}} \gamma_{j,h}^2 \right)^{1/2}.$$

Then, for all $x, x' \in \mathcal{Y}$ and $u \in \mathcal{U}$,

$$\|\tilde{f}_k(x,u) - \tilde{f}_k(x',u)\|_2 \le \sum_{k'} (\Gamma_\Delta)_{kk'} \|x_{k'} - x'_{k'}\|_2.$$

Moreover, if $\|\psi(y,u)\|_2 \le b_\psi$ for all $(y,u) \in \mathcal{Y} \times \mathcal{U}$, define the output-wise one-step discrepancy by

$$\varepsilon_j = \sup_{(y,u) \in \mathcal{Y} \times \mathcal{U}} |\hat{f}_j(y,u) - \tilde{f}_j(y,u)|.$$

Then

$$\varepsilon_j \le b_\psi \|\hat{\theta}_j - \tilde{\theta}_j\|_2.$$

For module k, the discrepancy defined in Section 6 satisfies

$$\varepsilon_k \le \left(\sum_{j \in \mathcal{B}_k} \varepsilon_j^2 \right)^{1/2}.$$

Proof. Proof idea. Each output changes by at most its retained weights times the change in the feature blocks it reads, and each block changes by at most its Jacobian bounds times the change in its inputs; Cauchy–Schwarz and Minkowski then aggregate these output-wise bounds into Γ_Δ.

For a feature block g and $x, x' \in \mathcal{Y}$, the integral form of the mean-value theorem along the segment from x to x', which lies in $\mathcal{Y}$ by convexity, together with the Jacobian bounds at every point of that segment, gives

$$\|\varphi_g(x_g) - \varphi_g(x'_g)\|_2 \le \sum_{h \in \mathrm{reads}(g)} L_g^{(h)} \|x_h - x'_h\|_2.$$

Using the triangle and Cauchy–Schwarz inequalities over retained feature blocks and control channels, together with $|q_i(u)| \le \bar{q}$, therefore gives, for each output j,

$$|\tilde{f}_j(x,u) - \tilde{f}_j(x',u)| \le \sum_{h \in \mathcal{G}_{\mathrm{dep}}} \gamma_{j,h} \|x_h - x'_h\|_2.$$

For a fixed source module k', Cauchy–Schwarz over its deployment groups gives

$$\left(\sum_{j \in \mathcal{B}_k} \left[\sum_{h \in \mathcal{B}_{k'}} \gamma_{j,h} \|x_h - x'_h\|_2 \right]^2 \right)^{1/2} \le (\Gamma_\Delta)_{kk'} \|x_{k'} - x'_{k'}\|_2.$$

Applying Minkowski's inequality over the source modules k' yields the stated blockwise Lipschitz inequality. The one-step discrepancy follows directly from Cauchy–Schwarz:

$$|\hat{f}_j(y,u) - \tilde{f}_j(y,u)| \le \|\psi(y,u)\|_2 \|\hat{\theta}_j - \tilde{\theta}_j\|_2.$$

Taking the module norm and then the supremum gives

$$\varepsilon_k \le \left(\sum_{j \in \mathcal{B}_k} \varepsilon_j^2 \right)^{1/2}.$$

□

The lemma holds for any differentiable dictionary with bounded features and bounded Jacobians on the operating region; frozen neural encoders inherit it with encoder Lipschitz constants estimated as in Fazlyab et al. (2019).

Proof of Proposition 3. Proof idea. Part (i) holds because the restricted predictor has zero weight on every coupling a module cannot read. Part (ii) tracks the gap between the two rollouts, which grows each step by at most the sensitivity of the restricted predictor applied to the current gap plus the one-step discrepancy.

Exact modular execution follows because the restricted weights vanish outside $A_j(D)$, so each output of module k depends only on the retained state and control values that module k reads, and under synchronous updates these values coincide with those used by centralized evaluation at the same step.

For the rollout bound, let

$$e_k(t) = \|\tilde{y}_k(t) - \hat{y}_k(t)\|_2.$$

Using the increment dynamics, the definition of ε_k, and the sensitivity bound that defines Γ_Δ (for example, the construction of Lemma D.1),

$$\begin{aligned} e_k(t+1) &\le e_k(t) + \|\tilde{f}_k(\tilde{y}(t), u_t) - \tilde{f}_k(\hat{y}(t), u_t)\|_2 + \varepsilon_k \\ &\le e_k(t) + \sum_{k'} (\Gamma_\Delta)_{kk'} e_{k'}(t) + \varepsilon_k. \end{aligned}$$

Thus

$$e(t+1) \le (I + \Gamma_\Delta) e(t) + \varepsilon = \Gamma e(t) + \varepsilon.$$

With $e(0) = 0$, induction gives, since Γ has nonnegative entries and therefore $a \le b$ componentwise implies $\Gamma a \le \Gamma b$,

$$e(t) \le \sum_{s=0}^{t-1} \Gamma^s \varepsilon.$$

The inequalities hold while both rollouts remain in $\mathcal{Y}$. □

E Additional Experimental Details

This appendix provides the experimental details and supporting results behind Section 7. The main paper is organized around three questions: the composability–fidelity trade-off, the fidelity of the one-fit restriction score, and recovery of physically meaningful subsystem structure. Here we provide the benchmark configurations, extended CSTR sweeps, additional structure-recovery diagnostics, deployed-restriction and finite-horizon checks, modular execution and rollout results on TEP and IEEE 39-bus, and additional robustness diagnostics. All reported decomposition decisions are made without access to the final reporting fold $\mathcal{D}_3$.

E.1 Benchmark configuration and evaluation protocol

All three benchmarks use the same four-fold experimental protocol. Table 2 summarizes the three configurations. The selection fold $\mathcal{D}_0$ is used for preprocessing, dictionary construction where applicable, and ridge-parameter selection; $\mathcal{D}_1$ is used once to fit the coupling-additive monolithic predictor; $\mathcal{D}_2$ is used to construct the coupling statistics, evaluate restriction scores, and select decompositions; and $\mathcal{D}_3$ is reserved for final evaluation. Rather than assigning individual transitions to folds, we assign episodes so that temporally adjacent samples from the same trajectory cannot cross between selection, fitting, scoring, and reporting. An episode is accepted only if it stays inside an operating region fixed before $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_3$ are used, and rejected episodes are replaced, so the law ρ and the constants of Assumption 1 refer to this conditioned episode law. The later-fold rejection rate is 0% on CSTR, at most 0.25% on TEP, and at most 2.3% on IEEE; TEP also discards three simulator shutdowns in each of $\mathcal{D}_2$ and $\mathcal{D}_3$. The reported NRMSE values are computed on the standardized state increments $Y^+ = Y_{t+1} - Y_t$.

The actuator assignment is fixed before scoring: on CSTR, every module retains the shared feed F and the jacket duties Q_r of the reactors it contains; on TEP, a module retains the manipulated variables assigned to its process units; and on IEEE, a module retains the governor-input channel of each machine it contains. The constant channel is always retained. On CSTR, for example, the shared feed is retained by every module, so a closed partition removes only couplings between a jacket duty and the features of a reactor outside its module; for the planted partition, these are the 24 couplings $(Q_r, g_{r'})$ with r and r' in different pairs, and $c_{\mathrm{b}} = 24/55 \approx 0.436$.

The CSTR benchmark consists of six reactors organized into the planted pairs $\{1,2\}$, $\{3,4\}$, and $\{5,6\}$. Within-pair coupling is fixed, while the coupling between neighboring pairs is scaled by $\epsilon_{\mathrm{inter}} \in \{0, 0.05, 0.10, 0.20, 0.30\}$. A shared feed signal and reactor-specific jacket duties provide the excitation. Each episode contains 200 transitions at a sampling interval of $0.02\,\mathrm{h}$, and the four folds contain $50/200/1000/2000$ episodes. The coupling-additive predictor uses six reactor feature groups with linear and random Fourier features (RFF); the representative $\epsilon_{\mathrm{inter}} = 0.1$ model has $d = 472$ coefficients, $N = 55$ removable couplings, and $\lambda = 10^{-6}$.

TEP is simulated in MANUAL mode under independent open-loop excitation of the eleven manipulated variables. We apply pseudorandom binary sequence (PRBS) excitation with amplitude 3% of each nominal manipulated-variable value plus a small Gaussian AR(1) jitter, and the controls are bounded by the actuator saturation limits, which give the control bound of Assumption 1. We retain the 22 continuously measured state channels. These channels are grouped into seven process units: feed, recycle, reactor, purge, separator, stripper, and compressor. The four folds contain $100/800/3000/2000$ accepted episodes, each contributing 14 transitions at a $60\,\mathrm{s}$ recording interval. The final coupling-additive predictor has $d = 1044$ coefficients and $N = 95$ removable couplings, with $\lambda = 10^{-2}$ selected on $\mathcal{D}_0$. The seven-unit decomposition family contains 877 closed partitions and 127 distinct nonempty candidate blocks.

The IEEE 39-bus benchmark is simulated using ANDES 2.0.0 (Cui et al., 2020) with the ten synchronous-machine groups as the deployment units. The control signal is a bounded piecewise-constant perturbation of `TGOV1N.paux0`; the final amplitude is 3.0, with all ten input directions excited. Each accepted episode contributes 60 transitions at a $0.05\,\mathrm{s}$ recording interval after burn-in. The folds contain $40/200/800/2000$ episodes, giving $2400/12000/48000/120000$ transitions. The feature representation contains ten individual-machine blocks together with twenty topology-screened machine-pair blocks, yielding $d = 2013$ coefficients and $N = 340$ removable couplings. Ridge regularization is selected by episode-level cross-validation inside $\mathcal{D}_0$, giving $\lambda = 0.1$. Ten

Table 2: **Benchmark configuration.** Episodes are independently generated and split by episode into selection ($\mathcal{D}_0$), fitting ($\mathcal{D}_1$), scoring ($\mathcal{D}_2$), and reporting ($\mathcal{D}_3$) folds. d is the fitted coefficient dimension and N is the number of removable couplings in the monolithic predictor.

	CSTR	TEP	IEEE 39-bus
groups	6 reactors	7 process units	10 machines
outputs n_y	24	22	20
inputs n_u	7	11	10
sampling interval	0.02 h	60 s	0.05 s
episodes $\mathcal{D}_0/\mathcal{D}_1/\mathcal{D}_2/\mathcal{D}_3$	50/200/1000/2000	100/800/3000/2000	40/200/800/2000
transitions $\mathcal{D}_0/\mathcal{D}_1/\mathcal{D}_2/\mathcal{D}_3$	10,000/40,000/200,000/400,000	1,400/11,200/42,000/28,000	2,400/12,000/48,000/120,000
feature blocks	6	7	$10 + 20$ pair blocks
dictionary	linear + RFF	linear + RFF	linear + RFF
d	472	1044	2013
N	55	95	340
λ	10^{-6}	10^{-2}	10^{-1}
candidate blocks	63	127	1023
closed partitions	203	877	115,975

deployment groups produce $115{,}975$ possible closed partitions and 1023 distinct candidate blocks. Because every output depends only on its own module, the $\mathcal{D}_2$ cost of a partition is the sum of its block costs, so the refit-selected reference and the score optimum are found by enumerating all $115{,}975$ partitions over cached block quantities. The gray points in Figure 2(c) are a fixed sample of 412 partitions evaluated on $\mathcal{D}_3$ for context only; for $2 \le K \le 9$ the exhaustive refit optimum has 3.4% to 6.1% lower $\mathcal{D}_2$ cost than the best partition in that sample.

The structure baselines used in the main paper are computed from the same folds and grouping. Influence clustering uses aggregate learned inter-group influence from the fitted predictor, correlation clustering uses measurement correlation without access to the fitted model, and spectral clustering operates on a learned affinity matrix. The refit-selected reference evaluates the exact restricted ridge refit of every candidate block on $\mathcal{D}_2$ and selects, at each K, the partition with the lowest summed $\mathcal{D}_2$ prediction error; on IEEE this takes 77.4 s for the 1023 block refits and 0.32 s to rank all partitions (Appendix E.11). Random partitions are used only as a contextual baseline; at $K = 6$ the mean $\mathcal{D}_3$ NRMSE over 20 random partitions is 0.4683 on TEP and 0.8367 on IEEE, with c_{b} not matched to the other selectors. We also run a goal-oriented local search at fixed K that evaluates the exact-refit $\mathcal{D}_2$ cost directly from the cached $\mathcal{D}_1$ statistics. Its nominal wall-clock budget, applied separately at each K, is set to the time MODEC spends on coupling statistics, scoring, and search for its whole path (12.9 s on TEP and 276.0 s on IEEE). The budget is checked before each new partition evaluation, so an evaluation already in progress is allowed to finish. On TEP this gives a maximum elapsed time of 13.7 s; the remaining runs finish below the nominal budget. The same fixed-K search is also shown for CSTR; the timing comparison below focuses on TEP and IEEE. On both benchmarks, the search reaches the exhaustive $\mathcal{D}_2$ refit optimum at every tested $K = 2, \dots, 6$. On TEP its $\mathcal{D}_3$ NRMSE therefore equals the refit-selected reference; on IEEE it is 0.6252, 0.6667, 0.7104, 0.7503, and 0.7914 for $K = 2, \dots, 6$, against 0.6262, 0.6667, 0.7104, 0.7646, and 0.8037 for the agglomerative MODEC path. The local search prices only the partitions it visits; because the candidate cut family is fixed before $\mathcal{D}_2$ is observed, the simultaneous frozen-cost bound also applies to restrictions selected by this search. After any partition is selected, all reported MODEC prediction errors use the exact restricted ridge solution of Proposition 1; the frozen and scalar-reweighted predictors are used only during score evaluation.

E.2 Comparison with additional baselines

Table 3 extends Table 1 with the influence, correlation, and spectral baselines at the same operating points.

E.3 Extended CSTR composability–fidelity sweep

The main paper shows the representative $\epsilon_{\mathrm{inter}} = 0.1$ CSTR front. Figure 5 shows the two ends of the coupling sweep. At $\epsilon_{\mathrm{inter}} = 0$, the planted reactor pairs are dynamically decoupled conditional on the shared feed, and the three-module restriction is essentially lossless: the monolith has NRMSE 0.0167 and the planted restricted model has NRMSE 0.0169 at $c_{\mathrm{b}} = 0.436$. Increasing

Table 3: **Full comparison at the representative operating points.** Results corresponding to Table 1, including the additional structure-selection baselines.

System	Method	K	c_{b}	$\mathcal{D}_3$ NRMSE	ARI
CSTR, $\epsilon_{\mathrm{inter}} = 0.1$	Monolith	1	0.000	0.016	–
	MODEC	3	0.436	0.1613	1.000
	Refit-selected	3	0.436	0.1613	1.000
	Influence	3	0.436	0.1613	1.000
	Correlation	3	0.327	0.1921	−0.061
	Spectral	3	0.436	0.1613	1.000
TEP	Monolith	1	0.000	0.4128	–
	MODEC	6	0.674	0.4554	–
	Refit-selected	6	0.674	0.4554	–
	Influence	6	0.663	0.4708	–
	Correlation	6	0.684	0.4614	–
	Spectral	6	0.684	0.4655	–
IEEE 39-bus	Monolith	1	0.000	0.5822	–
	MODEC	6	0.841	0.8037	–
	Score optimum	6	0.841	0.7995	–
	Refit-selected	6	0.729	0.7914	–
	Influence	6	0.829	0.8552	–
	Correlation	6	0.829	0.8634	–
	Spectral	6	0.853	0.8805	–

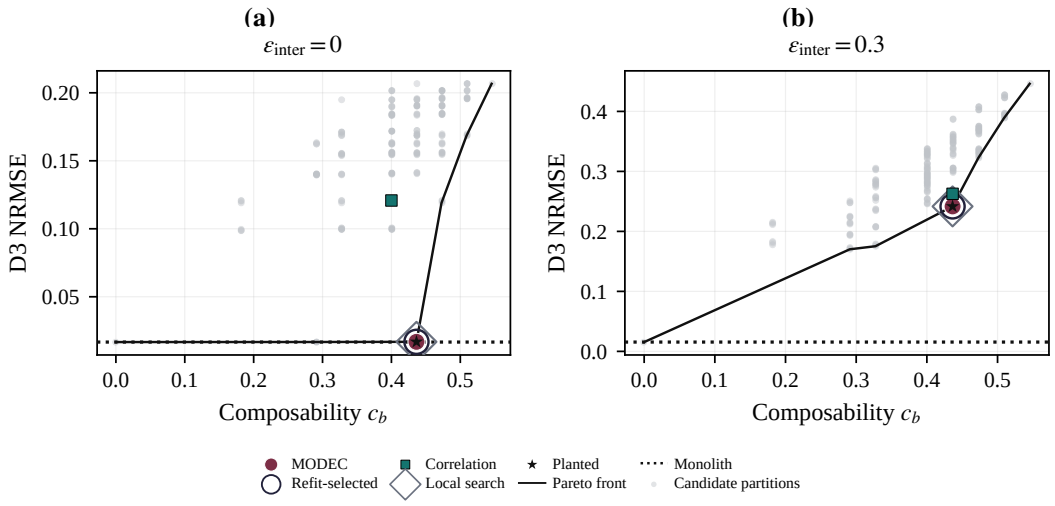


Figure 5: **Extended CSTR composability–fidelity fronts.** The left panel shows the nearly perfectly decomposable case $\epsilon_{\mathrm{inter}} = 0$, while the right panel shows the more strongly coupled case $\epsilon_{\mathrm{inter}} = 0.3$. Gray points denote all 203 closed partitions. The planted three-module decomposition is highlighted in both panels. Together with the $\epsilon_{\mathrm{inter}} = 0.1$ panel in the main paper, the sweep shows that the same physically meaningful module boundary becomes progressively more expensive as the interactions crossing it become stronger.

the true inter-module interaction moves the entire high-composability portion of the front upward. At $\epsilon_{\mathrm{inter}} = 0.3$, the same planted boundary still gives $c_{\mathrm{b}} = 0.436$, but its NRMSE rises to 0.2416. The decomposition therefore continues to reflect the known physical organization while MODEC correctly assigns a higher predictive cost to enforcing that organization as the omitted interactions become stronger.

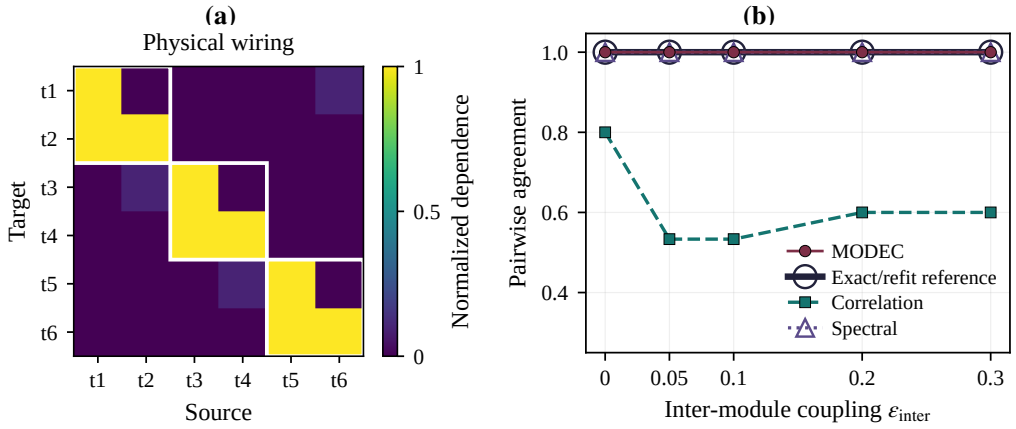


Figure 6: **Additional CSTR structure-recovery diagnostics.** **(a)** Known physical wiring of the six-reactor benchmark. **(b)** Pairwise agreement between recovered and planted partitions across the inter-module coupling sweep. These diagnostics complement the Jacobian, MODEC-attribution, MLP-gradient, and ARI panels in the main paper.

E.4 Additional CSTR structure-recovery diagnostics

Figure 6 supplements the structure-recovery figure in the main paper with the known physical wiring and a second quantitative agreement measure. The wiring map makes the planted architecture explicit: strong dependencies occur inside the reactor pairs, and only the prescribed neighboring-pair connections cross their boundaries. MODEC recovers the planted $K = 3$ partition for every tested coupling strength, yielding ARI $= 1$ throughout the sweep. Pairwise partition agreement provides the same conclusion from a complementary metric and confirms that the result is not an artifact of the ARI normalization.

The MODEC attribution in Figure 4(b) assigns to target reactor r and source reactor r' the sum of $|\hat{\phi}_{j,a}|$ over the outputs j of reactor r and the couplings $a = (i, g)$ whose feature block g belongs to reactor r', where $\hat{\phi}_{j,a}$ are the fitted freeze-score Shapley values defined after Corollary A.1, normalized to a maximum of one. The structure experiment also illustrates the difference between predictive dependence and raw statistical association. At $\epsilon_{\text{inter}} = 0.1$, MODEC's learned attribution has a cross-to-within-module magnitude ratio of only 0.038, compared with 0.548 for the unrestricted MLP-gradient baseline. The corresponding within-versus-cross discrimination AUC is 0.972 for MODEC and 0.806 for the MLP gradient. Correlation clustering does not recover the planted partition because the shared excitation induces association between physically distinct modules. MODEC instead identifies the contribution of those groups to prediction, giving a substantially cleaner representation of the known modular structure.

E.5 CSTR restriction-score validation across physical coupling

The CSTR sweep provides a controlled test of whether the score tracks changes in the cost of a fixed physical restriction. Figure 7 follows the planted three-module cut as ϵ_{inter} varies. The reweight score computed on $\mathcal{D}_2$ increases from 0.002803 at $\epsilon_{\text{inter}} = 0$ to 0.291475 at $\epsilon_{\text{inter}} = 0.3$, while the centered excess of the same scalar-reweighted restriction, measured independently on $\mathcal{D}_3$, increases from 0.002804 to 0.290916. At the intermediate values $\epsilon_{\text{inter}} = 0.05, 0.10, 0.20$, the $\mathcal{D}_2$-score/$\mathcal{D}_3$-evaluation pairs are $0.047012/0.046661$, $0.124727/0.124166$, and $0.229275/0.228761$, respectively. The relative discrepancy remains below 0.8% throughout the sweep. This is the matched comparison of Appendix E.9: the score is compared with the scalar-reweighted restriction it directly prices rather than with the exact restricted refit.

The deployed restricted model shows the same monotone trend. When the modules are physically independent, the exact restricted refit is nearly indistinguishable from the monolith. As ϵ_{inter} increases, the measured excess error increases because more predictive interaction crosses the fixed

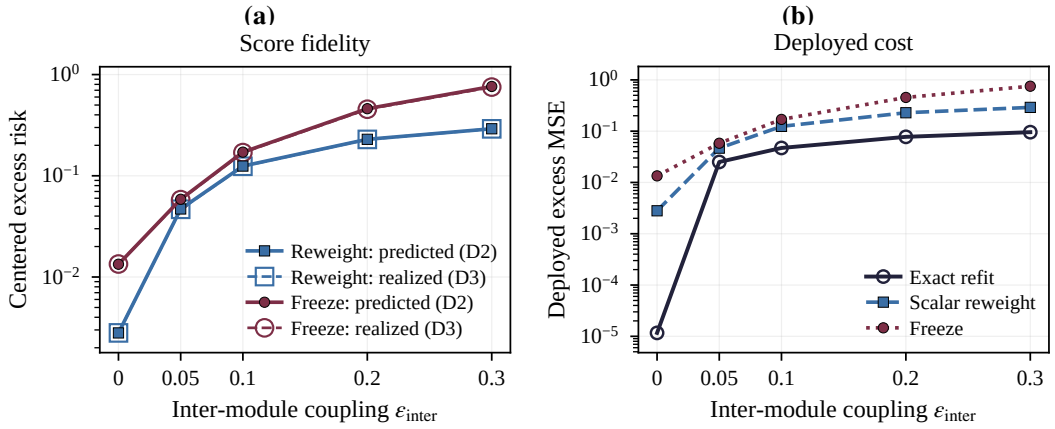


Figure 7: **CSTR score and deployed-restriction diagnostics.** The planted three-module restriction is held fixed while the true physical inter-module coupling is increased. The $\mathcal{D}_2$ reweight score closely tracks the independently evaluated $\mathcal{D}_3$ cost of the scalar-reweighted restriction it prices, while the deployed restricted predictor shows the expected increase in error as more dynamical dependence crosses the enforced module boundary.

planted boundary. This is the behavior MODEC is intended to expose: the subsystem identity can remain stable while the numerical cost of enforcing its boundary varies with the operating physics.

E.6 Finite-horizon diagnostic on CSTR

Figure 8 evaluates the finite-horizon tracking statement associated with the modular predictor. Module-by-module execution of a fixed restricted predictor is algebraically equivalent to centralized evaluation of that same predictor, and the numerical implementation satisfies this identity to machine precision. The nontrivial finite-horizon comparison instead asks how one-step restriction error propagates when predictions are recursively rolled forward. Because $\Gamma = I + \Gamma_\Delta$ with $\Gamma_\Delta \geq 0$, the blockwise construction has $\varrho(\Gamma) \geq 1$ by construction and therefore cannot certify horizon-uniform contraction. At the tested CSTR operating points, $\varrho(\Gamma) = 5.70$, 22.95, and 23.32 at $\epsilon_{\text{inter}} = 0$, 0.1, and 0.3, so the finite-horizon envelope grows rapidly with the rollout horizon. The measured trajectories remain far below this conservative envelope. A sampled full-map Jacobian diagnostic based on $I + \partial \tilde{f}/\partial y$ also has spectral radius above one on the evaluated models (1.76 to 11.33 on CSTR, 1.49 to 1.72 on TEP, and 1.45 to 1.96 on IEEE). Thus this blockwise test does not yield a horizon-uniform small-gain certificate, and we use the finite-horizon recursion (15). On IEEE, the learned trajectories leave the declared operating region at $t = 8$, so the envelope there is licensed only through $t = 7$.

This experiment is included as a theory diagnostic and is separate from the primary empirical evidence for MODEC. Its useful implication is that the one-step restriction quantities used by the decomposition procedure remain directly interpretable, while repeated worst-case gain composition is substantially more conservative than the observed rollout behavior. We therefore use the one-step composability–fidelity and score-fidelity experiments as the main empirical tests and report the longer-horizon envelope here for completeness.

E.7 TEP deployment footprint and rollout behavior

The TEP experiment provides a useful deployment interpretation of the learned decomposition. The monolithic predictor contains 1044 fitted coefficients in a single computational block. Once a decomposition is selected, each module evaluates only its own retained features and actuators. For the $K = 2$ MODEC decomposition, for example, the largest local feature dimension drops from 1044 to 568, while the sum of the two local dimensions is 653. Increasing the number of modules further reduces the maximum model footprint seen by any individual module. In every tested decomposition, each module reads only the state variables assigned to its own process units, so no state messages pass between modules, and centralized evaluation of the restricted model agrees with

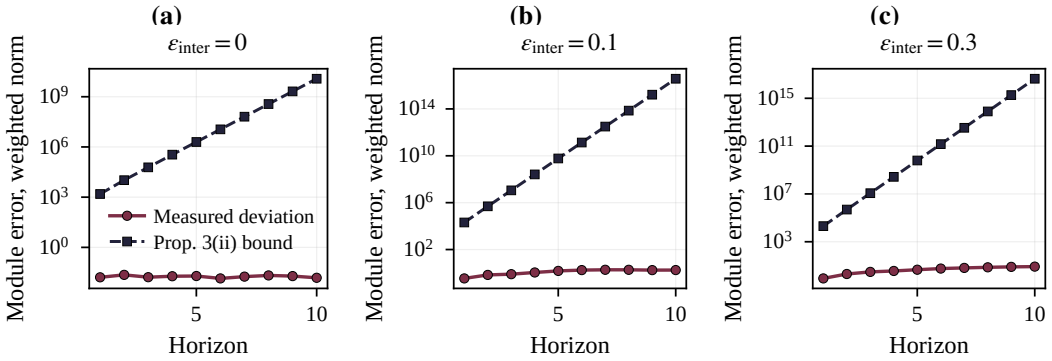


Figure 8: **Finite-horizon tracking diagnostic on CSTR.** Measured modular-versus-monolithic rollout deviation is compared with the finite-horizon gain envelope for multiple coupling strengths. The one-step quantities are informative, while repeated worst-case gain composition becomes increasingly conservative with horizon.

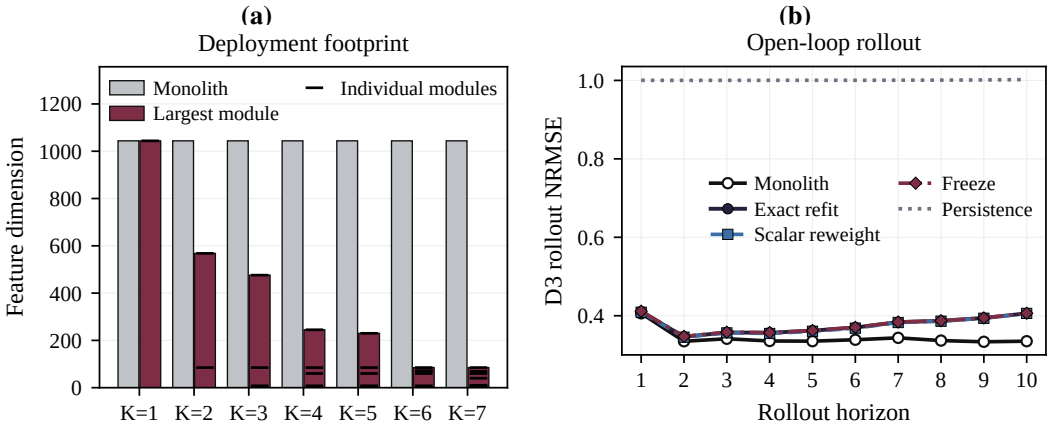


Figure 9: **TEP deployment and rollout behavior.** **(a)** Maximum and aggregate local model footprint as the selected decomposition is refined. **(b)** Recursive open-loop prediction error for the monolith and the $K = 2$ MODEC restriction under exact refit, scalar reweighting, and freezing. The decomposition reduces the model handled by each module, while the rollout curves show how the one-step cost of this restriction grows with horizon.

module-by-module evaluation to within 7.5×10^{-15}. Thus the algebraic decomposition translates directly into an executable collection of smaller predictors, beyond a sparsity pattern in a centralized model.

Figure 9 also reports open-loop rollout behavior. The one-step error of the selected restrictions remains close to the monolith for the low- and moderate-K decompositions, while the gap naturally grows over longer recursive horizons. At $K = 2$, the exact restricted refit has only about 2.2% more mean-squared error than the monolith at one step; the relative difference in mean-squared error is approximately 16.7% at horizon 5 and 46.9% at horizon 10. Figure 9(b) shows the $K = 2$ restriction under exact refit, scalar reweighting, and freezing. The rollout therefore illustrates how the one-step cost of this restriction accumulates with prediction horizon.

E.8 IEEE 39-BUS MODULAR EXECUTION AND ROLLOUT BEHAVIOR

The IEEE experiment verifies the same deployment property in the largest benchmark. For each selected partition, the restricted model can be evaluated either centrally or as separate modules that read only their retained machine-local information. Because the partitions are closed and a pair block is retained only when both of its machines lie in the same module, no module reads another module's machines, so this experiment checks exact modular evaluation without inter-module state

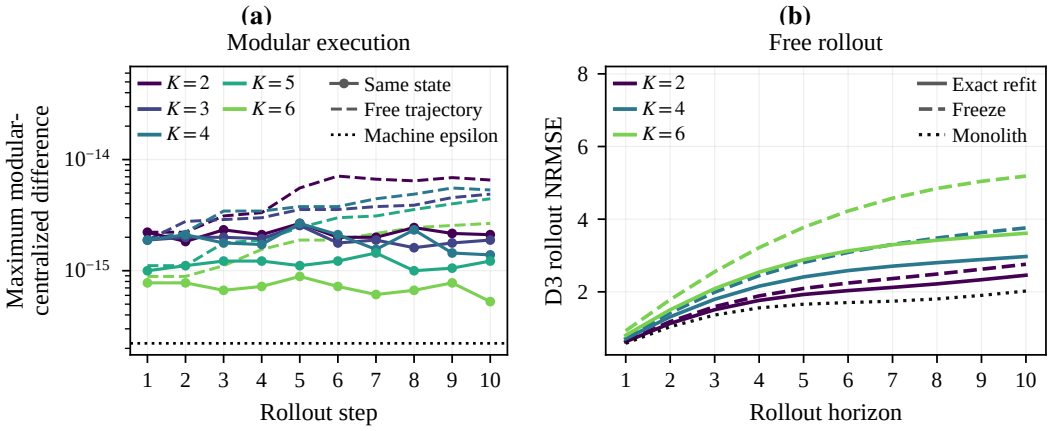


Figure 10: **IEEE 39-bus modular execution and rollout behavior. (a)** Numerical difference between centralized and module-by-module evaluation of the same restricted predictor; the discrepancy remains at machine precision. **(b)** Recursive prediction error of the monolith and selected decompositions, showing how the one-step composability trade-off propagates over longer horizons.

messages. Across the tested $K = 2, \ldots, 6$ decompositions, the maximum per-step discrepancy between the two implementations remains at the level of floating-point roundoff. For example, at $K = 2$ the largest one-step disagreement is 2.7×10^{-15}, and the maximum discrepancy accumulated over the tested ten-step trajectory is 7.1×10^{-15}. This confirms in a nontrivial 2013-parameter model that the decomposition changes the computational organization of the predictor without introducing an additional approximation beyond the selected restriction itself.

The right panel of Figure 10 shows the corresponding free-rollout behavior. The monolithic one-step NRMSE is 0.5822, while the $K = 2$ restricted model begins at 0.6262; both errors increase when predictions are recursively propagated. Increasing K produces the expected progression: more connectivity is removed, and the one-step predictor becomes more modular, while long-horizon error accumulates faster. The rollout experiment therefore reinforces the role of the main composability front as an operating curve from which a decomposition is chosen for the horizon of interest.

E.9 Additional score-fidelity and exactness checks

The main score-fidelity experiment compares the $\mathcal{D}_2$ reweight score with the $\mathcal{D}_3$ cost of the exact restricted refit, which gives Spearman correlations of 0.908 over all 127 TEP candidate blocks, including the full block, and 0.888 over all $1{,}022$ nontrivial IEEE candidate blocks. We also retain matched-cost diagnostics in which each score is compared with the specific restricted predictor that it directly prices. These comparisons answer a different, easier question: whether the same freeze or scalar-reweight cost is stable from $\mathcal{D}_2$ to an independent $\mathcal{D}_3$ fold.

On TEP, these matched diagnostics give Spearman correlations of 0.9988 for freeze and 0.9993 for scalar reweighting. On IEEE, freeze gives $\rho_s = 0.9991$ over the nontrivial candidate blocks, while scalar reweighting gives $\rho_s = 0.9977$ on 200 sampled blocks. The near-perfect matched-cost correlations show that the corresponding plug-in quantities are stable across independent folds, whereas the lower but still strong score-to-exact-refit correlations in the main paper quantify the harder question relevant to search: whether the inexpensive reweight score ranks the cost of the final fully refit restriction.

We separately verify the exact algebra used after a partition is selected. On TEP, the weights produced by the restricted-solution formula and an explicit constrained ridge solve agree to relative error 8.8×10^{-15}. The corresponding discrepancy on IEEE is 9.0×10^{-15}. Direct evaluations of the freeze and reweight formulas also agree with their brute-force same-fold counterparts to numerical precision. These checks distinguish two roles of the framework: the score provides a statistical search signal for restriction cost, while the final restricted ridge solution is an algebraic identity.

E.10 Robustness to modeling choices

We also vary the dictionary construction, feature width, regularization, and fitting-set size, and report the resulting monolith NRMSE. The variants change predictive accuracy in the expected bias-variance directions. On CSTR at $\epsilon_{\text{inter}} = 0.1$, the base linear-plus-RFF model has NRMSE 0.016. A purely linear dictionary increases NRMSE to 0.0701, while increasing the RFF width from the base configuration improves the monolithic NRMSE to 0.0138. Reducing the fitting set to 25% still yields an NRMSE of 0.0171, indicating that the controlled CSTR benchmark is comparatively data-rich.

TEP is less sensitive to nonlinear feature expansion: its final coupling-additive model yields an NRMSE of 0.4128, while several simpler variants achieve similar one-step errors. MODEC does not require the most accurate possible monolith, only a fitted coupling-additive representation whose removable components can be scored. For the base TEP model, the exhaustive 877-partition experiment shows that the agglomerative path reaches the global reweight-score optimum at every K.

IEEE is the most data-limited benchmark relative to model dimension, with $n_1/d \approx 5.96$, and consequently shows greater sensitivity to regularization and fitting-set size than to feature width. The final configuration uses $\lambda = 0.1$ selected exclusively inside $\mathcal{D}_0$, where $n_0/d \approx 1.2$; holding $\lambda = 10^{-2}$ fixed instead gives monolith NRMSE 0.5573, and fitting on 25% of $\mathcal{D}_1$ gives 0.6363, while the two feature-width variants stay within 0.001 of the base value. $\mathcal{D}_3$ is not used to choose among these variants. Removing the pair-interaction blocks reduces the representation to $d = 748$ and, with its own $\mathcal{D}_0$-selected $\lambda = 10^{-2}$, gives NRMSE 0.5750, compared with 0.5822 for the pair-augmented predictor at $\lambda = 0.1$. At the same $\lambda = 10^{-2}$ the pair-augmented model gives 0.5573, so the pair blocks improve one-step accuracy at matched regularization, and the base model is worse only because $\mathcal{D}_0$ selects the larger λ. Across these variants, the decomposition identities and one-fit scoring construction remain unchanged.

E.11 Computation and search scaling

Both the scores and the exact refits reuse statistics from a single fit across all candidate restrictions; this subsection reports how the cost is divided between forming those statistics and evaluating candidates. On TEP, fitting the monolith takes approximately 3.70 s, forming the $\mathcal{D}_2$ coupling statistics takes 12.40 s, scoring all 127 candidate blocks takes approximately 0.47 s, and the agglomerative search itself takes about 1.3 ms. Exhaustively ranking all 877 TEP partitions by the already-computed score takes only 0.066 s. The exact restricted refits of the same 127 blocks, both computed from the single $\mathcal{D}_1$ fit, take 111.4 s through the downdate form (5) and 11.1 s through the reduced solve (4); the downdate is the faster route on 30% of the blocks, mostly those that remove less than half of the feature coordinates. This separation is useful because the same fitted statistics support the entire decomposition family once formed.

IEEE provides the larger-scale comparison. Forming the core reusable score moments takes 261.2 s in the current implementation, after which all 1023 candidate blocks are scored in 14.8 s and the agglomerative partition search takes less than 1 ms. The exact restricted refits of the same 1023 blocks, computed by explicit reduced solves on the cached $\mathcal{D}_1$ statistics and equal to Proposition 1 to round-off, take 77.4 s, of which 5.7 s is spent in the linear solves, and ranking all $115{,}975$ closed partitions from the cached block costs then takes 0.32 s. At this scale, exhaustive exact-refit selection therefore costs less than forming the score moments; the value of the score in these experiments is its ranking fidelity (Section 7.2) and a per-candidate cost that scales with the number of couplings instead of the number of feature coordinates (Appendix C); wall-clock savings are absent at this scale. These numbers describe an implementation-level decomposition of computational cost for these benchmarks. Both procedures reuse a single set of fitted sufficient statistics rather than constructing an independently trained predictor for each possible decomposition.