Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Data-Driven Phase-Aware Knowledge Distillation for Stiff Chemical Reaction Networks

Wuyue Yang

yangwuyue@bimsa.cn

Beijing Institute of Mathematical Sciences and Applications https://orcid.org/0000-0002-8270-8840

Sheng Ran

Liu Hong

Article

Keywords: Chemical reaction network, Stiff dynamical system, Model reduction, Knowledge distillation

Posted Date: August 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10546497/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Data-Driven Phase-Aware Knowledge Distillation for Stiff Chemical Reaction Networks

Sheng Ran[1,2], Liu Hong[3*], Wuyue Yang[2*]

[1*]Institute of Statistics and Big Data, Renmin University of China, Beijing, 100872, China.
[2]Beijing Institute of Mathematical Sciences and Applications, Beijing, 101408, China.
[3]School of Mathematics, Sun Yat-sen University, Guangzhou, 510275, China.

*Corresponding author(s). E-mail(s): hongliu@sysu.edu.cn; yangwuyue@bimsa.cn;

Abstract

Many natural processes involve dynamics that span multiple widely separated timescales. Classical simplification strategies such as the Quasi-Steady-State Approximation and Partial-Equilibrium Approximation remain limited by their reliance on expert-guided assumptions and manual timescale identification. Here we introduce Phase-Aware Knowledge Distillation (PAKD), a fully data-driven framework that unifies surrogate teacher learning, unsupervised dynamical-phase discovery, and knowledge distillation. A high-fidelity teacher surrogate captures the full stiff dynamics. An unsupervised Hidden Markov Model then autonomously infers latent dynamical phases and their characteristic lifetimes from the teacher's trajectories. A compact student model subsequently distills the teacher's essential behaviors at substantially reduced computational cost. Validated on Michaelis–Menten enzyme kinetics, POLLU atmospheric chemistry network, Fisher–KPP reaction-diffusion equation, and a real-world MCF7 breast cancer signaling dataset, PAKD accurately reproduces key dynamics where classical approximations struggle, offering a principled, assumption-free route to model reduction across computational chemistry, system biology, and beyond.

Keywords: Chemical reaction network, Stiff dynamical system, Model reduction, Knowledge distillation

Introduction

Many physical, chemical, and biological processes involve dynamics that evolve across multiple different timescales. In chemical reaction networks, for example, some species concentrations change within microseconds while others take seconds or longer to reach equilibrium. Modeling and simulating such systems remains among the most computationally demanding tasks in scientific computing because their dynamics typically span a wide spectrum of characteristic timescales. This wide gap in timescales makes the equations stiff and forces numerical solvers to take very small steps throughout the simulation. These intrinsic multiscale and nonlinear effects motivate model-reduction strategies[1–6] that preserve the essential behaviors of the system while substantially reducing the computational cost, with complex reaction networks (CRNs) serving as a canonical testbed.

Classical reduction techniques typically require explicit identification of fast–slow structure and careful, regime-dependent tuning. In chemical kinetics and systems biology, this is exemplified by the Quasi-Steady-State Approximation (QSSA) [7] and the Partial-Equilibrium Approximation (PEA) [8–10]. QSSA assumes that certain reactive intermediates rapidly attain a quasi-steady state relative to other species. PEA, by contrast, treats selected reversible reactions as rapidly equilibrating and removes them from the system to reduce the stiffness. While effective when a clear separation of timescales exists, these approaches become fragile in large, highly coupled networks where fast and slow processes are tightly interwoven and may interchange roles across operating conditions. Related methods, such as computational singular perturbation [11], likewise depend on chemical intuition, parameter tuning, and explicit timescale identification. Projection-based methods such as proper orthogonal decomposition, balanced truncation, and dynamic mode decomposition [2, 12, 13] work well for linear or mildly nonlinear systems but generally need access to the governing equations and lose accuracy for the strongly nonlinear, stiff dynamics found in complex reaction networks. Consequently, many classical techniques lack robustness across regimes, cannot adapt to evolving kinetic structure, and fail when only data are available.

Recent advances in data-driven modeling and operator learning offer an alternative pathway that approximates complex kinetics directly from high-fidelity simulations. In principle, learning-based reduction can automate the identification of influential species and reactions, capture strong nonlinear couplings through latent representations, and combine mechanistic constraints with statistical flexibility in hybrid models, enabling fast, adaptive, and uncertainty-aware surrogates for reactive systems. Sparse identification of governing equations (SINDy) [14] has demonstrated that compact reduced models can be recovered directly from trajectory data, without access to the underlying equations. Peherstorfer and Willcox [15] proposed a nonintrusive operator inference approach that infers the reduced operators of a projection-based model directly from state trajectories via a least-squares problem. Yet most existing surrogates remain largely black-box regressors, with limited interpretability and persistent difficulty in reproducing the intrinsic stiffness and multiscale characteristics of CRNs. Moreover, both classical reduction and learning-based surrogates face a common bottleneck. Beyond numerically generated trajectories, reliable supervision is

scarce, limiting transfer and generalization to experimental or out-of-distribution conditions. Along these lines, recent neural architectures have begun to address stiffness more directly. Ji et al. [16] incorporated QSSA into physics-informed neural networks (PINNs) to promote chemically consistent dynamics and alleviate stiffness, while Kim et al. [17] proposed stiff neural ODE formulations using implicit components and exponential integrators to stabilize both forward simulation and backpropagation. In [18], a deep learning framework for high-dimensional multi-scale dynamic systems was proposed. Data-driven mechanism reduction methods for combustion kinetics indicate that compact surrogates can be directly learned from trajectory data [19, 20]. Caldana and Hesthaven [21] proposed a data-driven time reparametrization for neural ODEs that embeds the adaptive time-stepping of an implicit solver. These approaches, however, typically rely on expert-driven inductive biases and target simulation rather than real-world settings, leaving open the question of how to achieve accurate, interpretable, and stiffness-aware reductions in a purely data-driven, regime-adaptive fashion.

Knowledge distillation (KD) [22–24] offers a complementary perspective. KD has emerged as a powerful paradigm for transferring information from large, computationally intensive models to smaller, more efficient ones without substantial performance loss. Originally popularized in natural language processing (NLP) [25–28] and computer vision [29, 30], especially alongside the rise of foundation models, KD techniques have shown that rich representations can be compressed into lightweight surrogates. Motivated by this success, we extend KD to scientific computing and chemical kinetics, where compression must also preserve the key macroscopic dynamics and physical consistency of the model.

To address the limitations of traditional reduction approaches, we propose Phase-Aware Knowledge Distillation (PAKD). This unified framework integrates teacher surrogate modeling, unsupervised phase discovery, and KD-based model reduction within a single pipeline (see Algorithm 1 and Figure 1). Table 1 summarizes how PAKD compares with representative classical and learning-based reduction methods. PAKD proceeds in three stages. First, a high-capacity teacher network is trained on the original reaction data to capture the full stiff dynamics and generate dense temporal trajectories for all chemical species. Second, an unsupervised Hidden Markov Model (HMM) is applied to the teacher's trajectories to infer latent kinetic phases and compute their posterior probabilities over time. The resulting phase sequence reveals the system's intrinsic multiscale organization, separating fast transient bursts from slow, relaxation- and transport-dominated regimes. Third, we incorporate the phase posterior probabilities into the distillation objective as time-dependent weights. These phase-aware weights modulate the training loss to emphasize long-time-scale segments while retaining essential contributions from fast events. A lightweight student model trained under this weighted objective serves as an efficient reduced-order surrogate that reproduces dominant slow-scale behavior and maintains physically essential dynamics in stiff settings.

In ***Results***, we evaluate PAKD on a number of classical multiscale benchmarks (Michaelis–Menten, POLLU, Fisher–KPP) and a real-world MCF7 signaling dataset, demonstrating that our reduced-order surrogates are both accurate and numerically smooth. ***Methods*** details our approach, which combines a high-fidelity teacher

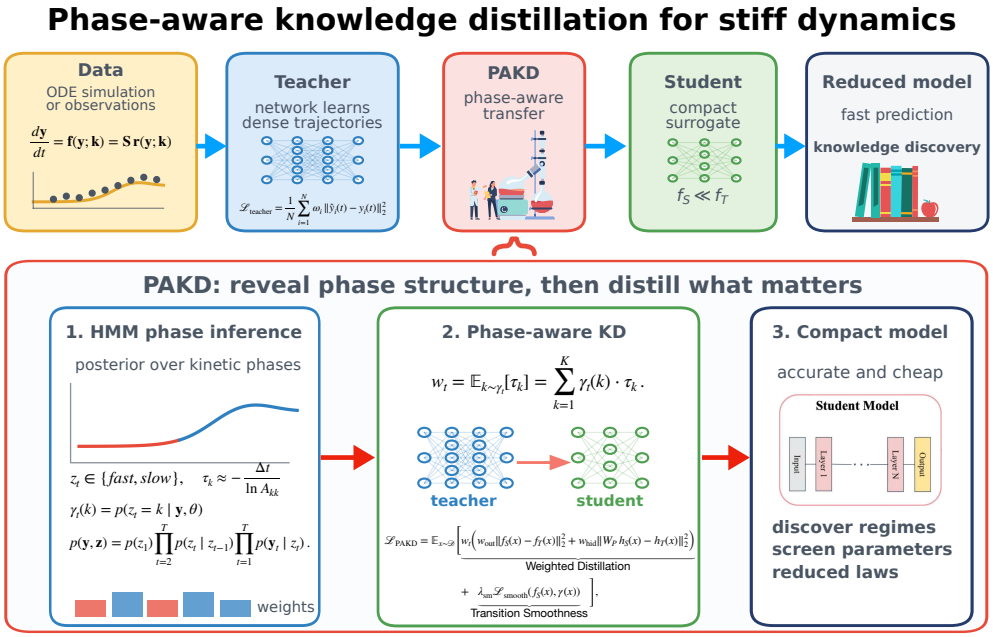

Fig. 1: Overview of the PAKD architecture. The student model is trained under two phase-aware supervision signals from the teacher, namely (i) feature distillation, which aligns intermediate representations, and (ii) output distillation, which matches the student's predictions to the teacher's. Time-dependent weights derived from HMM phase posteriors modulate both channels to emphasize persistent slow regimes while retaining informative transient dynamics, producing an accurate and numerically smooth reduced-order surrogate.

surrogate, unsupervised HMM phase discovery yielding phase posteriors $\gamma_t(k)$ and lifetimes τ_k, the construction of phase-aware distillation weights $w_t = \sum_k \gamma_t(k)\tau_k$, and the full PAKD objective. For MCF7, we additionally perform interpretable reaction mechanism discovery via gated Hill-type ODEs with differentiable topology.

Results

Michaelis–Menten enzyme kinetics: from teacher fidelity to stiffness reduction

The Michaelis–Menten enzyme kinetics model is characterized by a stark timescale separation, coupling a fast enzyme–substrate association rate ($k_1 = 100$) with slower dissociation ($k_{-1} = 10$) and catalytic turnover ($k_2 = 1$), spanning two orders of magnitude in reaction rates (Supplementary Section S1.1). This inherent stiffness makes the system a stringent benchmark for any model reduction strategy. While classical analytical reductions require the fast–slow boundary to be identified a priori, PAKD autonomously recovers it from data. The teacher surrogate faithfully recapitulates

Table 1: Comparison of PAKD with classical and learning-based model-reduction methods. ✓: fully satisfied; ✗: not satisfied; "Partial": partially or conditionally satisfied.

Method	No governing equations	No priori fast-slow split	Strongly nonlinear stiffness	Noisy/experimental data	Stiffness reduction
QSSA [7]	✗	✗	Partial	✗	✓
PEA [8]	✗	✗	Partial	✗	✓
CSP [11]	✗	Partial	✓	✗	✓
Projection ROM (POD/DMD) [2]	✗	—	✗	✗	✗
SINDy [14]	✓	—	Partial	Partial	✗
Operator inference [15]	✓	—	Partial	✗	✗
Stiff-PINN [16]	Partial	✗	✓	✗	✓
Stiff neural ODE [17]	✓	—	✓	✗	✓
PAKD (this work)	✓	✓	✓	✓	✓

"No governing equations" indicates whether the method operates purely from trajectory data without access to the underlying equations. "No priori fast-slow split" indicates whether the timescale separation is discovered automatically rather than imposed manually.

the full dynamics; predicted trajectories for the four species (S, E, $\overline{ES}$, and P) agree closely with the stiff-solver reference across eight decades in time (Fig. 2a).

Leveraging the teacher's trajectories, the unsupervised HMM partitions the dynamics into two distinct phases without any prior biochemical knowledge. The first is a short-lived, fast-transient phase dominated by the rapid accumulation of the enzyme–substrate complex $\overline{ES}$, followed by a prolonged slow-relaxation phase that captures the subsequent quasi-steady catalytic regime (Fig. 2b). This data-driven phase boundary precisely coincides with the theoretical timescale separation that the QSSA must manually impose.

This phase-resolved structure enables the distilled student model to succeed precisely where classical QSSA fails (Fig. 2d). Although both methods aim to reduce complexity, QSSA enforces the quasi-steady-state assumption from $t = 0$, leading to systematic deviations in the fast-transient regime, most visible in the dynamics of S and $\overline{ES}$. In contrast, the PAKD student eliminates these transient artifacts by preserving the teacher's phase-aware topology, accurately tracking the ground truth

through the slow relaxation. Furthermore, the student model maintains numerical stability across the stiff-to-non-stiff transition. Quantitatively, as shown in Fig. 2e, the student achieves a dramatic reduction in both the mean and peak effective stiffness, lowering it by orders of magnitude compared to both the teacher surrogate and the original ground-truth system.

POLLU air-chemistry benchmark

To evaluate the scalability of PAKD to realistic atmospheric chemistry, we applied it to the POLLU benchmark, a highly stiff air-pollution model comprising 20 coupled chemical species and 25 irreversible mass-action reactions originally developed by the Dutch National Institute of Public Health and Environmental Protection [6]. The system spans active timescales over fifteen orders of magnitude, from fast radical initiation at $t \sim 10^{-12}$ s to slow reservoir interconversion beyond 10^3 s, which places it among the harder benchmarks in the standard ODE literature.

Despite this extreme stiffness, the teacher surrogate faithfully reproduced the trajectories of all 20 species, closely tracking the BDF reference solution across the entire fifteen-decade time horizon (Fig. 3a). Subsequently, the unsupervised HMM autonomously partitioned the dynamics into a fast initial phase dominated by explosive radical chemistry and a slow relaxation phase governed by diffusion-limited reservoir exchange (Fig. 3b). Crucially, this two-phase structure emerged purely from the data, without any prior knowledge of the underlying reaction network.

The advantage of PAKD over classical model-reduction techniques is most apparent during the initial radical burst. As shown in Fig. 3d, the Quasi-Steady-State Approximation (QSSA) produces spurious oscillations in y_1 and y_5 and significantly overshoots y_3 and y_6. In contrast, the PAKD student model successfully avoids these transient artifacts, maintaining stable and accurate trajectories across both dynamical regimes (Fig. 3d). Furthermore, the student model achieves this high fidelity in slow dynamics with a computational cost orders of magnitude lower than that of the traditional BDF reference integrator and teacher surrogate inference (Fig. 3e). The phase-aware distillation process converges stably within approximately 900 epochs (Fig. 3c).

Fisher–KPP reaction-diffusion equation

We consider the Fisher–KPP (Kolmogorov–Petrovsky–Piskunov) equation, a prototypical reaction-diffusion partial differential equation (PDE) widely used to model the spatial spread of biological populations and chemical waves:

$$\frac{\partial u}{\partial t} = D\,\frac{\partial^2 u}{\partial x^2} + \alpha\, u(1-u), \qquad x \in (0, L), \quad t > 0, \tag{1}$$

This benchmark extends the PAKD framework from well-mixed kinetics to spatially extended dynamics, introducing stiffness that arises from the multiscale competition between rapid diffusive smoothing and slower logistic growth. Discretization on $N = 100$ interior grid points with a diffusion coefficient $D = 10^{-2}$ yields a high effective stiffness ratio ($D/h_x^2 \gg 1$). Furthermore, a sharp, step-like initial front ($x_s = 0.3$,

$\delta = 0.05$) concentrates steep spatial and temporal gradients, posing a stringent test for data-driven reduction.

In this spatially extended regime, the teacher surrogate successfully captured the spatiotemporal dynamics of the BDF reference solution, and the student model effectively distilled the reduced-order slow dynamics from the teacher. As shown in Fig. 4a, the log-scale error map between the teacher and the ground truth reveals that residuals are confined to a narrow band near the propagating front and the earliest time slices, indicating high-fidelity alignment. Concurrently, the student's error maps against both the ground truth and the teacher surrogate demonstrate its validity as a reduced-order slow dynamical model. Building on this, the unsupervised HMM resolved a sharp two-phase partition on a logarithmic time scale (Fig. 4b), comprising a fast, diffusion-dominated regime ($t < t^*_{\mathrm{HMM}} \approx 0.21$) and a slow, reaction-diffusion regime, separated by a near-step posterior transition. Guided by this phase-aware objective, the PAKD distillation converged stably over approximately 900 epochs (Fig. 4c). Spatial snapshots at four representative time points (Fig. 4d) confirm that the student model tracks the BDF reference faithfully after the separation point.

Beyond merely reproducing trajectories, the student model provides interpretable insights into the underlying physical mechanisms. By regressing an effective Fisher–KPP law from the student's trajectories at each time slice,

$$\partial_t u \;\approx\; D(t)\,\partial_{xx} u + r(t)\,u(1-u), \tag{2}$$

we uncovered a clear two-regime structure in the effective parameters (Fig. 4e). During the fast transient phase, both the effective diffusion coefficient $D(t)$ and reaction rate $r(t)$ collapsed to near zero. This indicates that the student model initially acts as an initial-condition projector, filtering out high-frequency spatial modes consistent with the rapid decay of stiff diffusive transients. At the HMM-inferred transition time, both coefficients stepped sharply to their true physical values with relative errors below 2.5%($D_{\mathrm{slow}} \approx 9.76 \times 10^{-3}$, $r_{\mathrm{slow}} \approx 0.989$). Ultimately, PAKD autonomously discovers a slow-manifold reduction: it projects the initial condition onto the slow manifold and subsequently activates a reduced reaction-diffusion law with physically meaningful parameters recovered directly from data, demonstrating that it does more than merely compress computation.

MCF7 breast cancer signaling

We evaluated PAKD on the HPN-DREAM breast cancer network dataset [31], which characterizes the phosphorylation dynamics of 41 signaling proteins in the MCF7 cell line across 36 combinatorial perturbation conditions (8 stimuli $\times$ 3 inhibitor backgrounds) measured at 7 time points spanning 0 to 240 minutes. This dataset faithfully represents real experimental conditions, including sparse temporal sampling, significant measurement noise, and crosstalk between the PI3K/AKT and MAPK/ERK signaling pathways. These challenging characteristics render it an ideal benchmark for assessing the robustness of any data-driven model reduction framework.

To further evaluate the generalizability of our approach, we applied it to noisy, sparsely sampled experimental data beyond conventional numerical benchmarks. As

Table 2: **Consensus network composition across signaling modules.** The teacher model contains 445 consensus edges active across all 36 perturbation conditions, whereas the student distills a sparser set of 256. Inter-module edges are reduced by 46% (366 → 197), whereas core signaling modules (PI3K-AKT, MAPK-ERK, mTOR) are largely preserved. The most aggressive compression occurs in the Cell-cycle module (9 → 4 edges).

	Teacher model			Student model		
Module	**Total**	**Unique**	**Shared**	**Total**	**Unique**	**Shared**
RTK	4	2	2	2	0	2
PI3K-AKT	14	4	10	11	1	10
MAPK-ERK	10	6	4	6	2	4
mTOR	6	2	4	6	2	4
Stress	9	1	8	10	2	8
Cell cycle	9	5	4	4	0	4
Context	27	15	12	20	8	12
Inter-module	366	269	97	197	100	97
Total	**445**	**304**	**141**	**256**	**115**	**141**

demonstrated in Fig. 5a, the teacher model exhibited high fidelity in capturing the phosphorylation dynamics of critical signaling nodes (including ERK, Src, STAT3, and EGFR) across diverse perturbation conditions, with excellent concordance to the experimental measurements. The student model distilled the essential slow dynamics from the teacher model, preserving the key temporal characteristics of the signaling pathways while achieving significant computational efficiency. From these trajectories, the HMM successfully identified a two-phase organizational structure, characterized by an early signaling burst phase followed by a late adaptive relaxation phase. The transition between these phases occurred at approximately $t^*_{\text{HMM}} \approx 35.6$ minutes, with 1,801 samples assigned to the early phase and 1,655 samples to the late phase out of the total 3,456 observations (Fig. 5b). Furthermore, the PAKD algorithm demonstrated efficient convergence of the target dynamics within 500 training epochs, as evidenced in Fig. 5c.

Furthermore, PAKD successfully distilled a sparse, biologically interpretable consensus regulatory network from the noisy experimental data. As illustrated in Fig. 6a, a topological comparison of the consensus gate matrices reveals that while the teacher model encodes 445 condition-invariant edges, the student model achieves a highly compact representation with only 256 edges (Table 2). This structural compression was highly selective rather than uniform. Inter-module couplings were pruned by 46% (from 366 to 197), whereas the intra-module topologies of core signaling cascades—namely PI3K/AKT, MAPK/ERK, and mTOR—were predominantly preserved. The largest reduction occurred within the cell-cycle module (9 → 4 edges), consistent

with the biological premise that cell-cycle regulation is context-dependent and thus excluded from the robust, condition-agnostic consensus backbone.

This selective pruning translates into a dramatically simplified macroscopic topology, as visualized in Fig. 6b. While the teacher network forms a dense, tangled web comprising 226 activation and 219 inhibition edges, the student network resolves into a streamlined, hierarchical signal flow from membrane receptors through adaptor and kinase layers to downstream effectors and nuclear targets. Notably, this compression is highly asymmetric with respect to edge polarity: the student model retains 200 out of 226 activation edges (safeguarding the primary excitatory cascade) but drastically prunes the inhibition edges from 219 down to just 56. This indicates that the student filters out spurious, noise-driven inhibitory feedback while preserving the core forward signaling pathways.

Within the 141 edges constituting the shared consensus backbone, the most prominent interactions include the canonical MEK1 $\rightarrow$ ERK cascade and the bidirectional feedback loop between AKT T308 and S473 (student gate $g = 0.999$). Notably, the 115 student-unique edges capture biologically coherent stress-pathway crosstalk, such as the p38 $\rightarrow$ MAPK interaction ($g = 0.976$). Ultimately, this demonstrates that the distillation process effectively acts as a topological denoiser, discarding spurious cross-module correlations and redundant inhibitory loops induced by experimental noise, while retaining mechanistically meaningful crosstalk [32, 33]. Overall, the extracted consensus architecture demonstrates strong concordance with established pathway topologies documented in the literature [34–36].

To translate the topological findings into quantitative mechanisms, we fit a Hill-type ODE model to the student's predicted trajectories:

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = -\gamma_i x_i + b_i + \sum_{j \neq i} g_{ji}\, \beta_{ji}\, H_{ji}(x_j), \tag{3}$$

where $H^{\mathrm{act}}(x) = x^n/(K^n + x^n)$ and $H^{\mathrm{inh}}(x) = K^n/(K^n + x^n)$ are Hill-type activation and inhibition terms, respectively. In Eq. (3), x_i denotes the activity of phosphoprotein i, γ_i is the first-order decay rate, and b_i is the basal activation level, while the summation runs over upstream regulators j. Each regulatory link is controlled by a consensus gate $g_{ji} \in [0, 1]$, encoding whether the interaction is robustly active across all 36 perturbation conditions, and a condition-specific rate coefficient β_{ji} that captures the interaction strength. The regulatory term is further shaped by a Hill function with half-saturation K_{ji} and cooperativity exponent n_{ji}.

The Hill-ODE parameters recovered for the top ten consensus edges are listed in Table 3, revealing a highly biologically coherent regulatory architecture. The strongest interaction is the bidirectional AKT T308↔S473 feedback loop (gate $g \approx 0.999$ in both directions, with Hill coefficients $n = 4.22$ and 6.01), which perfectly matches the known positive-feedback architecture of the PI3K/AKT node. Other highly ranked edges include the GSK3 cross-regulatory pair and the ERK→p90RSK and ERK→MEK1 arms of the MAPK cascade. Notably, the Hill coefficients exceed 2 for all ten top interactions, indicating strong cooperative binding dynamics that a linear surrogate model would inherently miss. Consequently, the distilled student model

Table 3: **Hill-ODE parameters for the top ten consensus edges.** For each interaction, the table reports the recovered consensus gate g_{ji}, rate coefficient β_{ji}, half-saturation K_{ji}, and Hill coefficient n_{ji} from Eq. (3). The coupling strength β_{ji} was evaluated across all 36 perturbation conditions to ensure robust consensus.

Rank	Interaction	Sign	g	β	K	n
1	AKT S473 → AKT T308	act	0.999	0.075	0.775	4.22
2	AKT T308 → AKT S473	act	0.999	0.066	0.915	6.01
3	GSK3ab → GSK3ab S9	act	0.988	0.051	0.630	2.92
4	ERK → p90RSK	act	0.957	0.047	0.736	3.01
5	GSK3ab S9 → GSK3ab	act	0.981	0.049	0.721	3.16
6	ERK → MEK1	act	0.983	0.049	0.523	2.93
7	CHK2 → p27 T198	act	0.956	0.048	0.348	3.16
8	PKCa → Src Y527	act	0.959	0.045	0.850	4.89
9	PRAS40 → p70S6K	act	0.969	0.046	0.696	4.92
10	EGFR Y992 → ERα	act	0.964	0.042	0.682	3.83

is not only topologically sparser but also mechanistically interpretable at the kinetic level.

Discussion

Stiff multiscale dynamics challenge surrogate modeling in two ways. Sharp early-time transients must be resolved without sacrificing long-time accuracy, yet widely separated timescales often render compact models numerically unstable or overly sensitive to regime changes. PAKD addresses this challenge by coupling knowledge distillation with unsupervised timescale discovery. Unlike classical reductions that impose mechanistic fast–slow assumptions, such as quasi-steady-state or partial-equilibrium approximations, PAKD first learns a high-fidelity teacher and then infers latent kinetic regimes from its dense trajectories. These inferred phases guide the distillation of a compact student through phase-aware importance weights that prioritize persistent slow dynamics while retaining informative transients.

The HMM posteriors provide a soft, time-resolved regime description, and the phase lifetimes translate regime persistence into an interpretable measure of importance, linking statistical segmentation to physical relevance. Long-lived regimes correspond to slow-manifold dynamics that dominate macroscopic behaviors and guide mechanistic interpretation, whereas short-lived regimes capture stiff transients that are essential but typically not the primary modeling target. Empirically, this coupling stabilizes distillation, so that the resulting student models reproduce long-time behaviors more reliably without sacrificing essential early-time structure.

Across Michaelis–Menten, POLLU, and Fisher–KPP, the distilled student is both accurate and less stiff than the teacher surrogate, as measured by reduced mean stiffness ratios. PAKD thus does more than reduce the parameter count. By enforcing phase-weighted matching and transition smoothness, it distils a numerically smoother

representation of the underlying dynamics. This smoothness matters in stiff dynamical systems, where small modeling errors can trigger unstable solver behaviors or spurious oscillations.

On the MCF7 signaling dataset, we further show that distillation produces models that are mechanistically interpretable, in addition to being predictive. The differentiable gated Hill-type ODEs yield signed, directed interactions and nonlinear kinetic parameters, enabling the extraction of sparse, condition-consistent regulatory graphs. Notably, the student achieves comparable fit quality with fewer high-confidence edges than the teacher. This suggests that phase-aware distillation promotes a parsimonious mechanistic summary, suppressing peripheral interactions that lack consistent support across the inferred kinetic regimes.

Several limitations motivate future work. First, the discrete phase model may be highly sensitive to the number of states and emission assumptions, and richer unsupervised regime models could capture gradual transitions and hierarchical timescales. Second, because distillation optimizes the student against the teacher, teacher bias can propagate, so improving teacher calibration and uncertainty-aware weighting are important directions. Finally, although we empirically demonstrate stiffness reduction, a deeper theoretical characterization of when and why distillation yields numerically smoother surrogates remains open. More broadly, PAKD provides a general template for combining probabilistic regime discovery with model compression, extensible to other stiff physical systems where dense supervision is costly but multiscale structure is intrinsic.

Methods

Surrogate teacher model

The teacher model f_T is a residual MLP with three hidden layers of width 128, GELU activations, and skip connections

$$\mathbf{h}_{\ell+1} = \mathbf{h}_\ell + \sigma(\mathbf{W}_\ell \mathbf{h}_\ell + \mathbf{b}_\ell). \tag{4}$$

Given an initial condition $\mathbf{y}_0$ and time t, the teacher predicts $\hat{\mathbf{y}}(t) = f_T(\mathbf{y}_0, t)$. Training minimizes the weighted mean-squared-error objective

$$\mathcal{L}_{\text{teacher}} = \frac{1}{N} \sum_{i=1}^{N} \omega_i \|\hat{y}_i(t) - y_i(t)\|_2^2, \tag{5}$$

where the per-species weights ω_i are set inversely proportional to the typical magnitude of species i so that errors are balanced across widely varying scales. We optimize using Adam with weight decay 10^{-5} and a learning-rate scheduler that reduces the rate when validation loss plateaus. Inputs are log-transformed and normalized, and outputs are normalized as well. The trained teacher generates dense, log-spaced temporal trajectories used for downstream phase discovery.

Unsupervised Phase Discovery via Hidden Markov Models

To uncover latent kinetic regimes directly from trajectories without manual annotation, we fit an unsupervised HMM to the teacher-generated temporal sequence $\{\mathbf{y}_t\}_{t=1}^{T}$. The system is assumed to occupy a discrete latent state $z_t \in \{1, \ldots, K\}$ at each time step, representing a distinct kinetic phase (e.g., rapid transient initiation versus long-time slow relaxation). The joint distribution of observed dynamics $\mathbf{y}$ and latent phases $\mathbf{z}$ factorizes as[37]

$$p(\mathbf{y}, \mathbf{z}) = p(z_1) \prod_{t=2}^{T} p(z_t \mid z_{t-1}) \prod_{t=1}^{T} p(\mathbf{y}_t \mid z_t), \tag{6}$$

where Gaussian emissions are $p(\mathbf{y}_t \mid z_t = k) = \mathcal{N}(\mathbf{y}_t; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ and $\mathbf{A}$ is the state transition matrix [38]. Parameters are estimated via the Baum–Welch (EM) algorithm [39, 40], and posterior responsibilities $\gamma_t(k) = p(z_t = k \mid \mathbf{y}, \theta)$ are obtained via the forward–backward recursion. Phase lifetimes are estimated as

$$\tau_k \approx -\frac{\Delta t}{\ln A_{kk}}, \tag{7}$$

where $\Delta t = \text{median}(t_{i+1} - t_i)$ is the effective sampling interval, corresponding to the expected dwell time under a continuous-time exponential decay model. All experiments used $K = 2$ phases (fast transient and slow relaxation).

Phase-Aware Knowledge Distillation (PAKD)

We distill the high-fidelity teacher model f_T into a compact student model f_S using the dynamical insights extracted by the HMM. Unlike standard distillation, which treats all time points uniformly, PAKD introduces a phase-aware weighting scheme that prioritizes the most informative kinetic regimes. To explicitly couple the statistical segmentation with physical relevance, we construct a time-dependent importance weight w_t. Since $\gamma_t(k)$ is the posterior probability of phase k at time t and τ_k is the characteristic lifetime of that phase, we define w_t as the expected phase lifetime at time t, given by

$$w_t = \mathbb{E}_{k \sim \gamma_t}[\tau_k] = \sum_{k=1}^{K} \gamma_t(k) \cdot \tau_k. \tag{8}$$

This formulation assigns higher weights to samples with persistent, slow-manifold dynamics (large τ_k), which are often the primary targets of mechanistic modeling, while downweighting short-lived transient regimes. The phase-aware weight w_t is normalized across the batch and incorporated into the distillation loss, thereby directing the student's limited capacity toward the dominant dynamical trends. The PAKD loss integrates output-level matching, representation alignment, and temporal smoothness

regularization, and is defined as

$$\mathcal{L}_{\text{PAKD}} = \mathbb{E}_{(\mathbf{y}_0,t)\sim\mathcal{D}}\Bigg[\underbrace{w_t\Big(w_{\text{out}}\|f_S(\mathbf{y}_0,t) - f_T(\mathbf{y}_0,t)\|_2^2 + w_{\text{hid}}\|W_P\, h_S(\mathbf{y}_0,t) - h_T(\mathbf{y}_0,t)\|_2^2\Big)}_{\text{Weighted distillation}} + \underbrace{\lambda_{\text{sm}}\,\mathcal{L}_{\text{smooth}}\big(f_S(\mathbf{y}_0,t),\gamma_t\big)}_{\text{Transition smoothness}}\Bigg], \tag{9}$$

where $h_T(\mathbf{y}_0,t)$ and $h_S(\mathbf{y}_0,t)$ are the hidden-layer outputs of the teacher and student, respectively, and W_P is a learnable linear projection that maps the student's latent representation into the teacher's representation space. This projection accommodates both dimensional and structural mismatches between the two networks, enabling flexible alignment of their internal feature geometries during distillation [41]. The first term enforces trajectory fidelity and transfers the teacher's internal geometry, weighted by the phase-importance score w_t. The second term, $\mathcal{L}_{\text{smooth}}$, penalizes abrupt output fluctuations in phase-transition regions, which are identified by high posterior entropy $H(\gamma_t) = -\sum_k \gamma_t(k)\log\gamma_t(k)$, through a first-order temporal-derivative penalty defined as

$$\mathcal{L}_{\text{smooth}} = \frac{1}{T-1}\sum_{i=1}^{T-1} H(\gamma_{t_i})\left\|\frac{f_S(\mathbf{y}_0,t_{i+1}) - f_S(\mathbf{y}_0,t_i)}{t_{i+1}-t_i}\right\|_2^2. \tag{10}$$

This ensures that the strongest smoothing is applied where the HMM is most uncertain about the phase assignment, thereby promoting coherent transitions between early and late response modes.

Experimental Settings

Teacher networks are trained on either numerically simulated trajectories or experimentally collected data, contingent upon empirical availability. All numerical simulations utilize adaptive-step stiff solvers (e.g., `BDF` or `LSODA`) to generate high-fidelity temporal trajectories. To mitigate the wide dynamic range characteristic of stiff kinetics, raw input states undergo a log transformation followed by normalization; output targets are similarly normalized. Unless otherwise specified, teacher models are implemented as residual multi-layer perceptrons (MLPs) comprising three hidden layers of width 128. In contrast, student models adopt an analogous residual MLP architecture but are restricted to a single 128-unit hidden layer to enforce a compact, reduced-order representation. Across all experiments, models are optimized using the Adam optimizer with an initial learning rate of 10^{-3} and a weight decay of 10^{-5}. Furthermore, a `ReduceLROnPlateau` scheduler is employed to halve the learning rate if the validation loss exhibits no improvement over a patience window of 50 epochs.

Supplementary Information. Details on the mass-action mechanisms, rate constants and initial conditions of each benchmark problem, the classical model reduction

Algorithm 1 Phase-Aware Knowledge Distillation (PAKD)

Input: Sparse dataset $\mathcal{D}_{\text{exp}}$, Teacher architecture f_T, Student architecture f_S
Output: Optimized Student parameters θ_S^*

// Stage 1: High-Fidelity Surrogate
Train Teacher f_T on $\mathcal{D}_{\text{exp}}$ until convergence
Generate dense dataset $\hat{\mathcal{D}}$ using f_T with log-spaced time sampling

// Stage 2: Unsupervised Timescale Discovery
Fit HMM to $\hat{\mathcal{D}}$ to identify latent phases $k \in \{0, 1\}$
Compute phase posteriors $\gamma_t(k)$ for each sample $x_t \in \hat{\mathcal{D}}$
Estimate effective sampling interval $\Delta t \leftarrow \text{median}(t_{i+1} - t_i)$
Estimate phase lifetimes: $\tau_k \approx -\Delta t / \ln A_{kk}$
Compute phase-aware importance weights: $w_t \leftarrow \sum_k \gamma_t(k) \cdot \tau_k$

// Stage 3: Phase-Aware Distillation
Initialize Student f_S and projection matrix W_P
while not converged **do**
    Sample batch $\mathcal{B} = \{(x, y_T, w, \gamma)\} \sim \hat{\mathcal{D}}$
    Compute Student predictions: $y_S = f_S(x)$, $h_S = \text{hidden}(f_S(x))$
    // Compute Weighted Distillation Loss
    $\mathcal{L}_{\text{distill}} \leftarrow \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} w_i \left(w_{\text{out}} \| y_S^{(i)} - y_T^{(i)} \|_2^2 + w_{\text{hid}} \| W_P h_S^{(i)} - h_T^{(i)} \|_2^2 \right)$
    // Compute Smoothness Regularization
    $\mathcal{L}_{\text{smooth}} \leftarrow \text{SmoothnessLoss}(y_S, \gamma)$
    Update θ_S, W_P via $\nabla(\mathcal{L}_{\text{distill}} + \lambda_{\text{sm}} \mathcal{L}_{\text{smooth}})$
end while
return Optimized Student f_S

approaches, statistical analysis and derivations of Hidden Markov Model, statistical consistency justification for the proposed PAKD pipeline, as well as supplementary figures.

Acknowledgements. This work was supported by the National Key R&D Program of China (Grant No. 2021YFA0719200), the National Natural Science Foundation of China (12301617), and the Guangdong Provincial Key Laboratory of Mathematical and Neural Dynamical Systems (2024B1212010004).

Author contributions

S.R. conceived the study, developed the PAKD framework, implemented the software, and performed the numerical experiments and formal analysis. L.H. and W.Y. conceived the study, supervised the project, and acquired funding. S.R. wrote the original draft. L.H. and W.Y. revised the manuscript.

Competing interests

The authors declare no competing interests.

Data availability

The MCF7 breast cancer data analyzed in this study are publicly available from the HPN-DREAM breast cancer network inference challenge through the Synapse repository under accession (https://www.synapse.org/#!Synapse:syn1720047) [31]. The trajectory datasets generated for the Michaelis–Menten, POLLU, and Fisher–KPP benchmarks, together with all other data supporting the findings of this study, are available at https://github.com/Kiwi-Orange/PAKD[42]. Source data for the figures are provided with this paper.

Code availability

The custom code implementing the PAKD framework, including the teacher surrogate, the unsupervised HMM phase-discovery module, and the phase-aware knowledge-distillation training pipeline, together with scripts to reproduce all results in this study, is available at https://github.com/Kiwi-Orange/PAKD[42].

References

[1] Radulescu, O., Gorban, A.N., Zinovyev, A., Noel, V.: Reduction of dynamical biochemical reactions networks in computational biology. Frontiers in genetics **3**, 131 (2012)

[2] Benner, P., Gugercin, S., Willcox, K.: A survey of projection-based model reduction methods for parametric dynamical systems. SIAM review **57**(4), 483–531 (2015)

[3] Antoulas, A.C.: Approximation of Large-Scale Dynamical Systems. Advances in Design and Control. SIAM, Philadelphia, PA (2005)

[4] Epstein, I.R., Pojman, J.A.: An Introduction to Nonlinear Chemical Dynamics: Oscillations, Waves, Patterns, and Chaos. Topics in Physical Chemistry. Oxford University Press, New York (1998)

[5] Voit, E.O.: Computational Analysis of Biochemical Systems: a Practical Guide for Biochemists and Molecular Biologists. Cambridge University Press, Cambridge (2000)

[6] Verwer, J.G.: Gauss–seidel iteration for stiff odes from chemical kinetics. SIAM Journal on Scientific Computing **15**(5), 1243–1250 (1994)

[7] Segel, L.A., Slemrod, M.: The quasi-steady-state assumption: a case study in perturbation. SIAM review **31**(3), 446–477 (1989)

[8] Gorban, A.N., Karlin, I.V.: Invariant Manifolds for Physical and Chemical Kinetics vol. 660. Springer, Berlin, Heidelberg (2005)

[9] Huang, Y.-J., Hong, L., Yong, W.-A.: Partial equilibrium approximations in apoptosis. ii. the death-inducing signaling complex subsystem. Mathematical biosciences **270**, 126–134 (2015)

[10] Peng, L., Hong, L.: Thermodynamics for reduced models of chemical reactions by pea and qssa. Physical Review Research **6**(1), 013296 (2024)

[11] Zagaris, A., Kaper, H.G., Kaper, T.J.: Fast and slow dynamics for the computational singular perturbation method. Multiscale Modeling & Simulation **2**(4), 613–638 (2004)

[12] Rowley, C.W., Dawson, S.T.: Model reduction for flow analysis and control. Annual Review of Fluid Mechanics **49**, 387–417 (2017)

[13] Schmid, P.J.: Dynamic mode decomposition of numerical and experimental data. Journal of fluid mechanics **656**, 5–28 (2010)

[14] Brunton, S.L., Proctor, J.L., Kutz, J.N.: Discovering governing equations from data by sparse identification of nonlinear dynamical systems. Proceedings of the national academy of sciences **113**(15), 3932–3937 (2016)

[15] Peherstorfer, B., Willcox, K.: Data-driven operator inference for nonintrusive projection-based model reduction. Computer Methods in Applied Mechanics and Engineering **306**, 196–215 (2016)

[16] Ji, W., Qiu, W., Shi, Z., Pan, S., Deng, S.: Stiff-pinn: Physics-informed neural network for stiff chemical kinetics. The Journal of Physical Chemistry A **125**(36), 8098–8106 (2021)

[17] Kim, S., Ji, W., Deng, S., Ma, Y., Rackauckas, C.: Stiff neural ordinary differential equations. Chaos: An Interdisciplinary Journal of Nonlinear Science **31**(9) (2021)

[18] Yao, J., Yi, Y., Hang, L., Wang, W., Zhang, Y., Zhang, T., Xu, Z.-Q.J., et al.: Solving multiscale dynamical systems by deep learning. Computer Physics Communications, 109802 (2025)

[19] Wang, Z., Zhang, Y., Lin, P., Zhao, E., Zhang, T., Xu, Z.-Q.J., *et al.*: Deep mechanism reduction (deepmr) method for fuel chemical kinetics. Combustion and Flame **261**, 113286 (2024)

[20] Zhang, T., Yi, Y., Xu, Y., Chen, Z.X., Zhang, Y., Xu, Z.-Q.J., *et al.*: A multi-scale sampling method for accurate and robust deep neural network to predict combustion chemical kinetics. Combustion and Flame **245**, 112319 (2022)

[21] Caldana, M., Hesthaven, J.S.: Neural ordinary differential equations for model order reduction of stiff systems. International Journal for Numerical Methods in Engineering **126**(12), 70060 (2025)

[22] Buciluǎ, C., Caruana, R., Niculescu-Mizil, A.: Model compression. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 535–541 (2006)

[23] Hinton, G.E., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. CoRR **abs/1503.02531** (2015)

[24] Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: A survey. Int. J. Comput. Vision **129**(6), 1789–1819 (2021) https://doi.org/10.1007/s11263-021-01453-z

[25] Kim, Y., Rush, A.M.: Sequence-level knowledge distillation. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 1317–1327 (2016)

[26] Hsieh, C.-Y., Li, C.-L., Yeh, C.-k., Nakhost, H., Fujii, Y., Ratner, A., Krishna, R., Lee, C.-Y., Pfister, T.: Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Findings of the Association for Computational Linguistics: ACL 2023, pp. 8003–8017. Association for Computational Linguistics, Toronto, Canada (2023). https://doi.org/10.18653/v1/2023.findings-acl.507 . https://aclanthology.org/2023.findings-acl.507/

[27] Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., Liu, Q.: TinyBERT: Distilling BERT for natural language understanding. In: Cohn, T., He, Y., Liu, Y. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 4163–4174. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.findings-emnlp.372 . https://aclanthology.org/2020.findings-emnlp.372/

[28] Gu, Y., Dong, L., Wei, F., Huang, M.: Minillm: Knowledge distillation of large language models. In: International Conference on Learning Representations, vol. 2024, pp. 32694–32717 (2024)

[29] Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In: International Conference on Learning Representations (2017). https://openreview.net/forum?id=Sks9_ajex

[30] Zhao, B., Cui, Q., Song, R., Qiu, Y., Liang, J.: Decoupled knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11953–11962 (2022)

[31] Hill, S.M., Heiser, L.M., Cokelaer, T., Unger, M., Nesser, N.K., Carlin, D.E., Zhang, Y., Sokolov, A., Paull, E.O., Wong, C.K., *et al.*: Inferring causal molecular networks: empirical assessment through a community-based effort. Nature methods **13**(4), 310–318 (2016)

[32] Manning, B.D., Cantley, L.C.: Akt/pkb signaling: navigating downstream. cell **129**(7), 1261–1274 (2007)

[33] Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., Sidiropoulos, K., Cook, J., Gillespie, M., Haw, R., *et al.*: The reactome pathway knowledgebase. Nucleic acids research **48**(D1), 498–503 (2020)

[34] Cross, D.A., Alessi, D.R., Cohen, P., Andjelkovich, M., Hemmings, B.A.: Inhibition of glycogen synthase kinase-3 by insulin mediated by protein kinase b. Nature **378**(6559), 785–789 (1995)

[35] Roskoski Jr, R.: Erk1/2 map kinases: structure, function, and regulation. Pharmacological research **66**(2), 105–143 (2012)

[36] Rodchenkov, I., Babur, O., Luna, A., Aksoy, B.A., Wong, J.V., Fong, D., Franz, M., Siper, M.C., Cheung, M., Wrana, M., *et al.*: Pathway commons 2019 update: integration, analysis and exploration of pathway data. Nucleic acids research **48**(D1), 489–497 (2020)

[37] Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE **77**(2), 257–286 (1989)

[38] Bishop, C.M.: Pattern Recognition and Machine Learning, Softcover reprint of the original 1st edition 2006 (corrected at 8th printing 2009) edn. Information science and statistics. Springer, New York, NY (2016)

[39] Baum, L.E., Petrie, T., Soules, G., Weiss, N.: A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. The annals of mathematical statistics **41**(1), 164–171 (1970)

[40] Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. Journal of the royal statistical society: series B (methodological) **39**(1), 1–22 (1977)

[41] Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015). http://arxiv.org/abs/1412.6550

[42] Ran, S., Hong, L., Yang, W.: PAKD: Phase-Aware Knowledge Distillation for Stiff Dynamical Systems. https://github.com/Kiwi-Orange/PAKD. GitHub repository (2026)

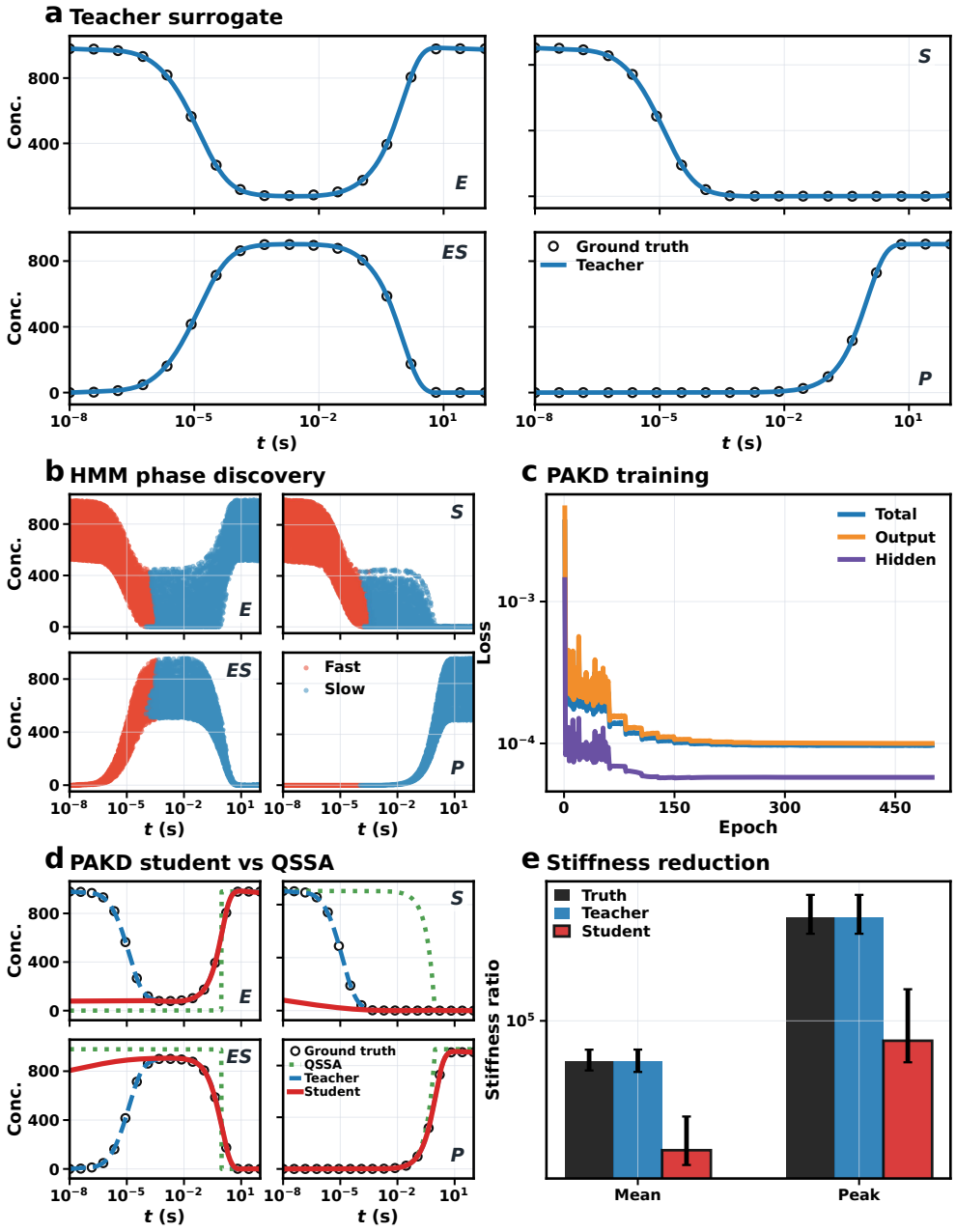

Fig. 2: Phase-aware distillation resolves stiff Michaelis–Menten dynamics. **a**, Trajectories of the teacher surrogate model compared with the stiff ODE solver reference solution for all four reacting species (E, S, $\overline{ES}$, and P). **b**, HMM-discovered dynamical phases comprising fast transient and slow relaxation regimes, identified entirely from data without prior knowledge. **c**, PAKD training loss curves converging over 450 epochs. **d**, Trajectory comparison of the ground-truth solution, QSSA approximation, teacher surrogate, and PAKD student model. **e**, Mean and peak effective stiffness ratios for the ground-truth reference, teacher surrogate, and student model.

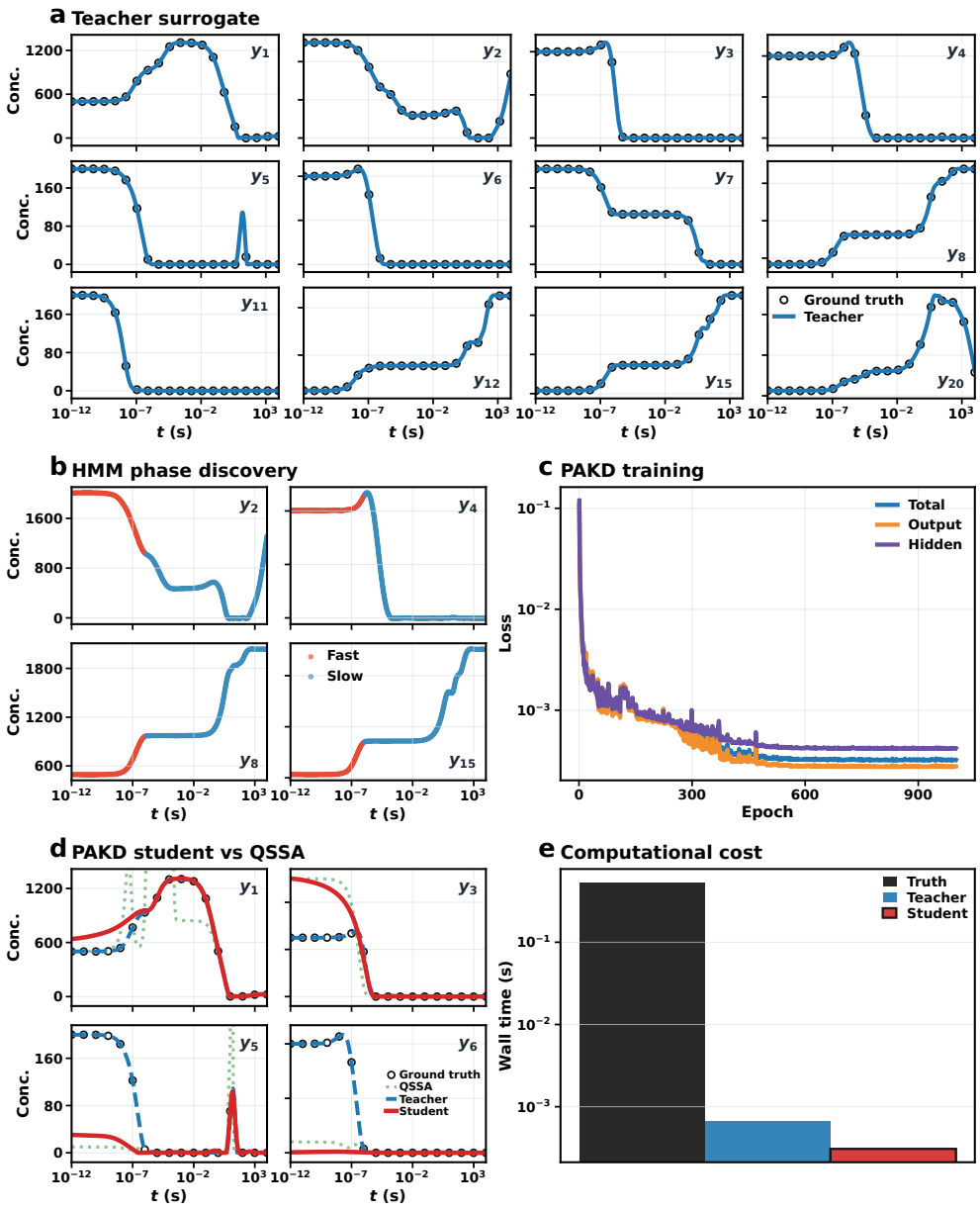

Fig. 3: **Phase-aware distillation compresses stiff POLLU chemistry. a**, Trajectories of the teacher surrogate model compared with the BDF reference solution for 12 representative species. **b**, Dynamical phases discovered via unsupervised HMM—comprising a fast radical-burst regime and a slow relaxation regime—inferred entirely from data without prior knowledge of the reaction network. **c**, PAKD training-loss curves converging over approximately 900 epochs. **d**, Trajectory comparison of the ground-truth solution, QSSA approximation, teacher surrogate, and PAKD student model for representative species y_1, y_3, y_5, and y_6. **e**, Computational costs of the BDF numerical solving, teacher surrogate inference, and student model inference.

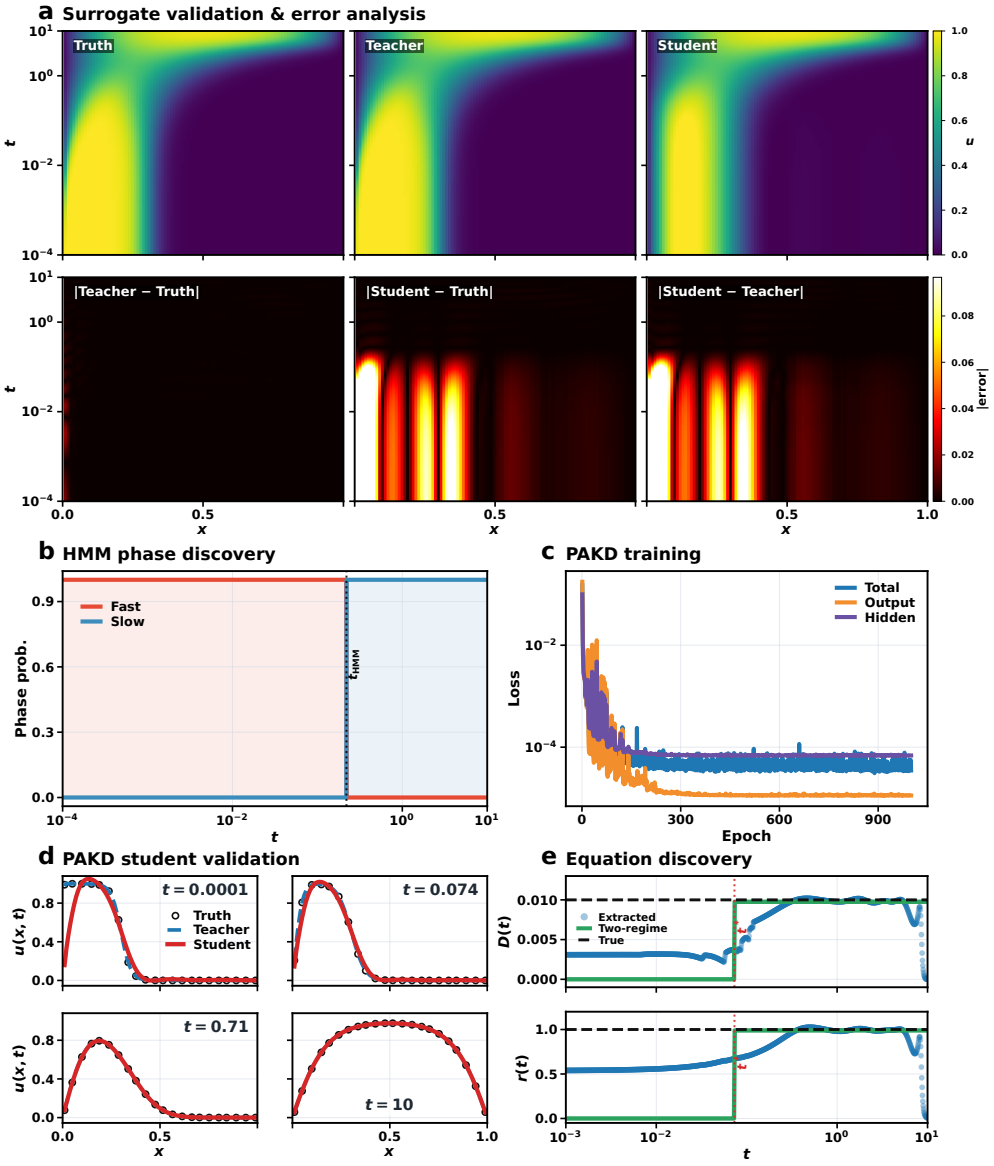

Fig. 4: **Phase-aware distillation reveals Fisher–KPP slow dynamics. a**, Spatiotemporal heatmaps of the solution $u(x,t)$ for the BDF ground truth, the teacher surrogate, and the PAKD student model, along with their corresponding error maps. **b**, HMM-inferred phase probabilities revealing a sharp transition at $t^*_{\mathrm{HMM}} \approx 0.21$ between the fast diffusion-dominated and slow reaction-diffusion regimes. **c**, PAKD training loss curves converging stably over approximately 900 epochs. **d**, Spatial profiles at four representative times comparing the BDF reference solution, the teacher surrogate, and the PAKD student model. **e**, Equation discovery results, illustrating the time-dependent diffusion coefficient $D(t)$ and reaction rate $r(t)$ recovered from student trajectories, compared against the two-regime step-function fit and ground-truth values.

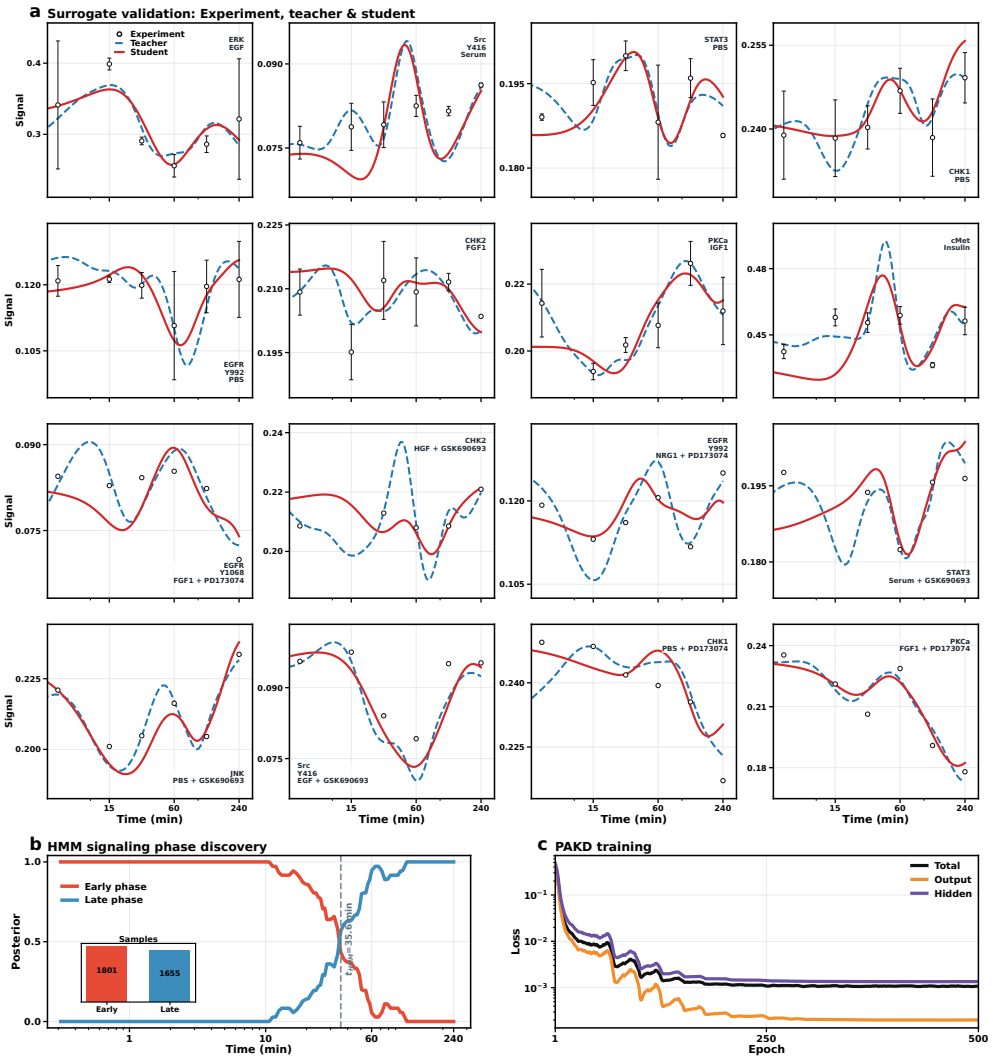

Fig. 5: **PAKD on MCF7 breast cancer signaling. a**, Comparison of teacher surrogate model fits and student model predictions against experimental measurements across 16 representative phosphoproteins. **b**, HMM-inferred signaling phases, comprising an early phase and a late phase, switching at $t^*_{\mathrm{HMM}} \approx 35.6$ min. **c**, PAKD training loss curves converging over 500 epochs.

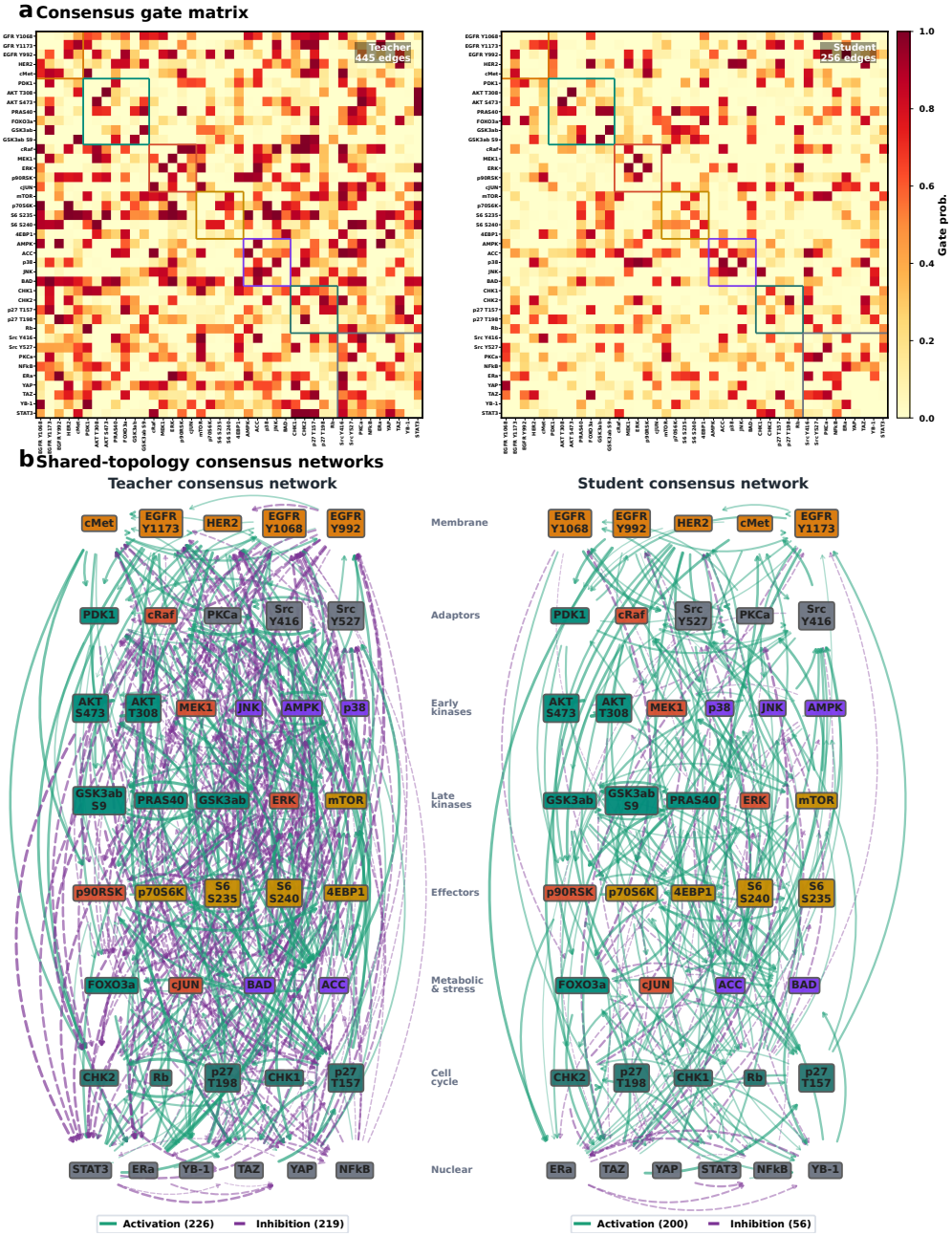

Fig. 6: **Regulatory network discovery from the PAKD student. a**, Consensus gate matrices for the teacher (left, 445 edges) and distilled student (right, 256 edges) models under 36 perturbation conditions. The student matrix exhibits marked sparsity, pruning inter-vertex connections typical of the early phase. **b**, Causal networks of the teacher and student models showing the topological geometry of their dynamics. Compared to the teacher model, the student's topology is markedly reduced across the inner and outer protein layers of the cellular reaction cycle.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SI.pdf
- MLreportingsummary17871909481.pdf