Supplementary Information:
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;

S1 Model System Specifications

This section provides the complete mathematical specifications of each benchmark system studied in the main text, including their governing equations, parameter values, and initial conditions.

S1.1 Michaelis–Menten Enzyme Kinetics

Let $E(t)$, $S(t)$, $\overline{ES}(t)$, and $P(t)$ denote the enzyme, substrate, enzyme–substrate complex, and product concentrations, respectively. Under mass-action kinetics, the

governing equations are

$$\begin{aligned}
\frac{dE}{dt} &= -k_1\, ES + (k_{-1} + k_2)\, \overline{ES}, \\
\frac{dS}{dt} &= -k_1\, ES + k_{-1}\, \overline{ES}, \\
\frac{d\overline{ES}}{dt} &= k_1\, ES - (k_{-1} + k_2)\, \overline{ES}, \\
\frac{dP}{dt} &= k_2\, \overline{ES},
\end{aligned} \tag{1}$$

where $(k_1, k_{-1}, k_2) = (100, 10, 1)$. The stiffness ratio is of order $k_1/k_2 \sim 10^2$. Initial conditions are drawn from

$$E(0),\, S(0) \sim \mathcal{U}(5 \times 10^2,\, 10^3), \qquad \overline{ES}(0) = 0, \quad P(0) = 0, \tag{2}$$

exposing the model to enzyme-limited, balanced, and substrate-limited regimes.

The classical QSSA reduces (1) by assuming $d\overline{ES}/dt \approx 0$, yielding the Michaelis–Menten rate law

$$\frac{dS}{dt} \approx -\frac{V_{\max} S}{K_M + S}, \qquad V_{\max} = k_2 E_0, \quad K_M = \frac{k_{-1} + k_2}{k_1}, \tag{3}$$

where $E_0 = E(0)$ is the total enzyme concentration. This approximation is valid when $E_0 \ll K_M + S$, a condition that may fail for the initial conditions considered here.

S1.2 POLLU Air-Chemistry Benchmark

The POLLU benchmark [1] models atmospheric photo-chemistry with $n_s = 20$ species $\mathbf{y}(t) = [y_1, \ldots, y_{20}]^\top$ and $n_r = 25$ irreversible mass-action reactions. The dynamics follow

$$\frac{d\mathbf{y}}{dt} = \mathbf{N}\, \mathbf{r}(\mathbf{y}; \mathbf{k}), \tag{4}$$

where $\mathbf{N} \in \mathbb{R}^{20 \times 25}$ is the stoichiometric matrix and $\mathbf{r}$ is the reaction-rate vector.

Reaction rates

$$\begin{aligned}
r_1 &= k_1 y_1, & r_2 &= k_2 y_2 y_4, & r_3 &= k_3 y_5 y_2, \\
r_4 &= k_4 y_7, & r_5 &= k_5 y_7, & r_6 &= k_6 y_7 y_6, \\
r_7 &= k_7 y_9, & r_8 &= k_8 y_9 y_6, & r_9 &= k_9 y_{11} y_2, \\
r_{10} &= k_{10} y_{11} y_1, & r_{11} &= k_{11} y_{13}, & r_{12} &= k_{12} y_{10} y_2, \\
r_{13} &= k_{13} y_{14}, & r_{14} &= k_{14} y_1 y_6, & r_{15} &= k_{15} y_3, \\
r_{16} &= k_{16} y_4, & r_{17} &= k_{17} y_4, & r_{18} &= k_{18} y_{16}, \\
r_{19} &= k_{19} y_{16}, & r_{20} &= k_{20} y_{17} y_6, & r_{21} &= k_{21} y_{19},
\end{aligned}$$

$$r_{22} = k_{22} y_{19}, \qquad r_{23} = k_{23} y_1 y_4, \qquad r_{24} = k_{24} y_{19} y_1,$$
$$r_{25} = k_{25} y_{20}.$$

Governing ODEs

$$\begin{cases}
\dot{y}_1 = -r_1 - r_{10} - r_{14} - r_{23} - r_{24} + r_2 + r_3 + r_9 + r_{11} + r_{12} + r_{22} + r_{25}, \\
\dot{y}_2 = -r_2 - r_3 - r_9 - r_{12} + r_1 + r_{21}, \\
\dot{y}_3 = -r_{15} + r_1 + r_{17} + r_{19} + r_{22}, \\
\dot{y}_4 = -r_2 - r_{16} - r_{17} - r_{23} + r_{15}, \\
\dot{y}_5 = -r_3 + 2r_4 + r_6 + r_7 + r_{13} + r_{20}, \\
\dot{y}_6 = -r_6 - r_8 - r_{14} - r_{20} + r_3 + 2r_{18}, \\
\dot{y}_7 = -r_4 - r_5 - r_6 + r_{13}, \\
\dot{y}_8 = r_4 + r_5 + r_6 + r_7, \\
\dot{y}_9 = -r_7 - r_8, \\
\dot{y}_{10} = -r_{12} + r_7 + r_9, \\
\dot{y}_{11} = -r_9 - r_{10} + r_8 + r_{11}, \\
\dot{y}_{12} = r_9, \\
\dot{y}_{13} = -r_{11} + r_{10}, \\
\dot{y}_{14} = -r_{13} + r_{12}, \\
\dot{y}_{15} = r_{14}, \\
\dot{y}_{16} = -r_{18} - r_{19} + r_{16}, \\
\dot{y}_{17} = -r_{20}, \\
\dot{y}_{18} = r_{20}, \\
\dot{y}_{19} = -r_{21} - r_{22} - r_{24} + r_{23} + r_{25}, \\
\dot{y}_{20} = -r_{25} + r_{24}.
\end{cases} \tag{5}$$

Rate constants

$$\mathbf{k} = \begin{bmatrix}
3.5 \times 10^{0} & 5.0 \times 10^{1} & 1.0 \times 10^{4} & 5.0 \times 10^{-2} & 5.0 \times 10^{-2} \\
2.0 \times 10^{4} & 5.0 \times 10^{-3} & 5.0 \times 10^{4} & 2.0 \times 10^{4} & 1.0 \times 10^{5} \\
5.0 \times 10^{-1} & 2.0 \times 10^{4} & 3.0 \times 10^{-1} & 1.0 \times 10^{4} & 1.0 \times 10^{6} \\
1.0 \times 10^{-2} & 2.0 \times 10^{0} & 5.0 \times 10^{5} & 2.0 \times 10^{6} & 3.0 \times 10^{3} \\
3.0 \times 10^{0} & 8.0 \times 10^{0} & 5.0 \times 10^{0} & 3.0 \times 10^{3} & 5.0 \times 10^{0}
\end{bmatrix}^{\top}. \tag{6}$$

Initial conditions

$$\mathbf{y}(0) = \begin{bmatrix}
5 \times 10^{2} & 2 \times 10^{3} & 1 \times 10^{2} & 5 \times 10^{2} & 2 \times 10^{2} \\
1 \times 10^{2} & 1 \times 10^{3} & 5 \times 10^{2} & 1 \times 10^{2} & 5 \times 10^{1} \\
2 \times 10^{2} & 5 \times 10^{1} & 1 \times 10^{2} & 8 \times 10^{1} & 3 \times 10^{1} \\
1 \times 10^{2} & 5 \times 10^{1} & 2 \times 10^{2} & 1 \times 10^{2} & 5 \times 10^{1}
\end{bmatrix}^{\top}. \tag{7}$$

S1.3 Fisher–KPP Reaction-Diffusion Equation

The Fisher–KPP equation is

$$\frac{\partial u}{\partial t} = D\frac{\partial^2 u}{\partial x^2} + \alpha\, u(1-u), \qquad x \in (0, L),\ t > 0, \tag{8}$$

with homogeneous Dirichlet boundary conditions $u(t, 0) = u(t, L) = 0$. We set $L = 1$, $D = 10^{-2}$, $\alpha = 1$, and $T = 10$.

Spatial discretization

We partition $[0, L]$ into $N+1$ subintervals with mesh width $h_x = L/(N+1)$ and interior grid points $x_j = jh_x$, $j = 1, \ldots, N = 100$. Using the central-difference approximation for the Laplacian, the semi-discrete system is

$$\frac{d\mathbf{y}}{dt} = \frac{D}{h_x^2} A\mathbf{y} + \alpha\, \mathbf{y} \circ (\mathbf{1} - \mathbf{y}), \tag{9}$$

where $\mathbf{y} = (y_1, \ldots, y_N)^\top$, $\circ$ is the Hadamard product, and

$$A = \begin{pmatrix} -2 & 1 & & \\ 1 & -2 & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & -2 \end{pmatrix} \in \mathbb{R}^{N\times N} \tag{10}$$

is the discrete Laplacian. The diffusion operator introduces eigenvalues of magnitude $\mathcal{O}(D/h_x^2)$, giving an effective stiffness ratio $D/h_x^2 \approx 10^2$ for the chosen parameters.

Initial condition

The initial condition is a smoothed step profile,

$$u_0(x) = \frac{1}{2}\left(1 - \tanh\left(\frac{x - x_s}{\delta}\right)\right), \tag{11}$$

with $x_s = 0.3$ and transition width $\delta = 0.05$, chosen to generate a sharp front and emphasise the diffusion-dominated transient.

Solver and Jacobian

Integration uses an implicit BDF solver with tolerances $(\text{atol}, \text{rtol}) = (10^{-10}, 10^{-8})$. The analytical sparse Jacobian

$$J(\mathbf{y}) = \frac{D}{h_x^2} A + \operatorname{diag}(\alpha(1 - 2\mathbf{y})) \tag{12}$$

is supplied together with the Laplacian sparsity pattern to improve solver efficiency.

S1.4 MCF7 Breast Cancer Signalling Dataset

The HPN-DREAM dataset [2] captures signal transduction dynamics in the MCF7 breast cancer cell line. The experimental design is summarised in Table S1.

Table S1: Experimental design of the HPN-DREAM MCF7 signalling dataset.

Component	Description
Stimuli (8)	EGF, Insulin, FGF1, Serum, HGF, NRG1, IGF1, PBS (negative control).
Inhibitors (3)	AKT inhibitor, FGFR inhibitor, MEK/AKT dual inhibitor.
Total conditions	36 (stimulus–inhibitor combinations, excluding redundant controls).
Readouts	Phosphorylation levels of 41 key signalling proteins, including p-AKT(S473), p-ERK1/2(T202/Y204), p-S6(S235/S236), p-EGFR(Y1068), p-HER2(Y1248), p-MEK1/2(S217/S221), and others.
Time points	$t \in \{0, 5, 15, 30, 60, 120, 240\}$ minutes post-stimulation.

This dataset is particularly challenging because (i) only 7 time points are available per trajectory, (ii) the PI3K/AKT and MAPK/ERK pathways exhibit complex crosstalk, and (iii) the inhibitor conditions shift the effective network topology across different experimental conditions.

S2 Classical Reduction Methods

This section provides complete mathematical descriptions of the classical model-reduction methods used as baselines in the main text.

S2.1 Partial-Equilibrium Approximation (PEA)

PEA assumes that a subset of reactions ($r = 1, \ldots, p$, $p < R$) equilibrate rapidly relative to the time scale of interest. The net flux of each fast reaction satisfies

$$r_r(\mathbf{y}) \approx 0, \qquad r = 1, \ldots, p, \tag{13}$$

yielding local forward–reverse balance

$$k_r^f \prod_{i=1}^{N} y_i^{\nu'_{ir}} = k_r^r \prod_{i=1}^{N} y_i^{\nu''_{ir}}, \tag{14}$$

or the equilibrium constraint

$$\prod_{i=1}^{N} y_i^{\nu''_{ir} - \nu'_{ir}} = K_r, \qquad K_r = \frac{k_r^f}{k_r^r}, \qquad r = 1, \ldots, p, \tag{15}$$

which defines an equilibrium manifold in the concentration space. The remaining slow reactions then drive motion along this manifold.

S2.2 Quasi-Steady-State Approximation (QSSA)

QSSA assumes that a subset of species ($y_1, \ldots, y_p$, $p < N$) relax rapidly to quasi-steady values. The condition

$$\frac{dy_i}{dt} \approx 0, \qquad i = 1, \ldots, p, \tag{16}$$

combined with

$$\frac{dy_i}{dt} = \sum_{r=1}^{R} \nu_{ir}\, r_r(\mathbf{y}), \tag{17}$$

gives algebraic constraints

$$\sum_{r=1}^{R} \nu_{ir}\, r_r(\mathbf{y}) = 0, \qquad i = 1, \ldots, p, \tag{18}$$

which implicitly defines a quasi-steady manifold. The reduced dynamics evolve only the slow species $y_{p+1}, \ldots, y_N$ subject to (18).

S2.3 Computational Singular Perturbation (CSP)

CSP [3] provides a systematic local decomposition of fast and slow modes by analysing the Jacobian $\mathbf{J}(\mathbf{y})$ along trajectories. The method constructs fast and slow subspaces via eigen- or Schur-based decompositions, then projects the dynamics to suppress rapidly decaying components. Concretely, at each point $\mathbf{y}$ on the trajectory, the Jacobian is decomposed as

$$\mathbf{J}(\mathbf{y}) = \mathbf{Q}\,\Lambda\,\mathbf{Q}^{-1}, \tag{19}$$

where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_N)$ with $|\mathrm{Re}(\lambda_1)| \geq \cdots \geq |\mathrm{Re}(\lambda_N)|$. The fast subspace is spanned by eigenvectors corresponding to the p most negative eigenvalues, and the reduced dynamics are obtained by projecting onto the complementary slow subspace. In practice, CSP requires repeated eigendecompositions along trajectories and relies on a threshold to separate fast from slow modes, making it computationally expensive and sensitive to parameter choices in complex networks.

S2.4 Limitations of Classical Methods

Despite their interpretability, all three methods above share common limitations. They require prior knowledge of which species or reactions are fast, which is often unavailable in large networks, and the fast–slow classification can change along a trajectory, requiring regime-dependent re-tuning. In tightly coupled networks the assumption of clear timescale separation may not hold, leading to inaccurate reductions, and none of these methods adapt automatically to new operating conditions or incomplete mechanistic knowledge. These limitations motivate data-driven approaches that infer effective dynamics directly from trajectories.

S3 Reaction Mechanism Discovery: Full Model Specification

This section provides the complete mathematical specification of the differentiable gated Hill-type ODEs used for reaction mechanism discovery on the MCF7 dataset.

S3.1 Model Formulation

Let $\mathbf{x}(t) = (x_1(t), \ldots, x_P(t))^\top \in \mathbb{R}^P$ denote the normalised state of P observed species. For each target node $i \in \{1, \ldots, P\}$,

$$\frac{dx_i}{dt} = -\gamma_i x_i + b_i + \sum_{j \neq i} g_{ji}\, \beta_{ji}\, H_{ji}(x_j), \tag{20}$$

where $\gamma_i > 0$ is a first-order deactivation rate, $b_i \in \mathbb{R}$ is a basal input, $\beta_{ji} \in \mathbb{R}$ is the additive effect of the directed edge $j \to i$, $g_{ji} \in (0, 1)$ is the edge confidence gate (self-edges disallowed: $g_{ii} = 0$), and $H_{ji}(x_j) \in [0, 1]$ is a bounded Hill-type response function.

S3.2 Activation/Inhibition via a Continuous Sign Gate

Each edge $j \to i$ interpolates between the two Hill-type forms $H^{\text{act}}(x) = x^n/(K^n + x^n)$ and $H^{\text{inh}}(x) = K^n/(K^n + x^n)$ through a learnable sign gate $s_{ji} \in (0, 1)$:

$$H_{ji}(x_j) = s_{ji}\, \frac{x_j^{\,n_{ji}}}{K_{ji}^{\,n_{ji}} + x_j^{\,n_{ji}}} + (1 - s_{ji})\, \frac{K_{ji}^{\,n_{ji}}}{K_{ji}^{\,n_{ji}} + x_j^{\,n_{ji}}}, \qquad K_{ji} > 0,\ n_{ji} > 0, \tag{21}$$

where K_{ji} is the half-saturation constant, n_{ji} the Hill coefficient, and s_{ji} a learnable sign gate: $s_{ji} \to 1$ recovers the activation form H^{act} used in Eq. (3), and $s_{ji} \to 0$ recovers the inhibition form H^{inh}. The edge sign reported in Eq. (3) and Table 2 corresponds to the discretized limit of s_{ji}.

S3.3 Differentiable Topology

To avoid combinatorial search over discrete adjacency matrices, edge indicators are parameterised as

$$g_{ji}(\tau) = \sigma\left(\frac{\theta_{ji}}{\tau}\right), \qquad \theta_{ji} \in \mathbb{R},\ \tau > 0,\ g_{ii} = 0, \tag{22}$$

where θ_{ji} is an unconstrained logit, $\sigma(\cdot)$ is the sigmoid function, and the temperature τ is annealed during training so that gates become progressively sharper (approaching $\{0, 1\}$). The annealing schedule follows $\tau(n) = \tau_0 \cdot \rho^n$ where n is the training epoch, $\tau_0 = 1.0$, and $\rho = 0.99$.

S3.4 Interpretation of the Gated Additive Term

In Eq. (20), each regulator j enters through the gated term $g_{ji}\,\beta_{ji}\,H_{ji}(x_j)$. The gate g_{ji} controls whether the edge is present: at $g_{ji} \to 0$ the term vanishes, and at $g_{ji} \to 1$ the full effect $\beta_{ji}\,H_{ji}(x_j)$ is recovered. Here β_{ji} sets the sign and strength of the interaction, and $H_{ji}(x_j) \in [0, 1]$ bounds its response. Since regulators are summed, gating each edge keeps the topology differentiable, so edge presence, sign, and saturation can be learned jointly.

S4 HMM Statistical Analysis and Derivations

S4.1 Probabilistic Formulation and Variational Bound

We model the teacher-generated trajectory $\{\mathbf{y}_t\}_{t=1}^{T}$ as a stochastic process governed by a K-state HMM. At each time step t, a discrete latent variable $z_t \in \{1, \ldots, K\}$ represents the unobserved kinetic phase, which emits the observed concentration vector $\mathbf{y}_t \in \mathbb{R}^N$ according to phase-specific statistics.

The joint probability distribution factorizes as

$$p(\mathbf{y}, \mathbf{z} \mid \theta) = p(z_1) \prod_{t=2}^{T} p(z_t \mid z_{t-1}) \prod_{t=1}^{T} p(\mathbf{y}_t \mid z_t), \tag{23}$$

with parameters $\theta = \{\boldsymbol{\pi}, \mathbf{A}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$: initial probabilities $\pi_k = p(z_1 = k)$, transition matrix $A_{ij} = p(z_t = j \mid z_{t-1} = i)$, and Gaussian emissions $p(\mathbf{y}_t \mid z_t = k) = \mathcal{N}(\mathbf{y}_t; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$.

Training maximizes the marginal log-likelihood $\log p(\mathbf{y} \mid \theta) = \log \sum_{\mathbf{z}} p(\mathbf{y}, \mathbf{z} \mid \theta)$. Since direct summation over K^T paths is intractable, we employ the EM framework, which effectively maximizes the variational lower bound

$$\log p(\mathbf{y} \mid \theta) \geq \mathcal{L}(q, \theta) = \mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{y}, \mathbf{z} \mid \theta)] + \mathcal{H}[q(\mathbf{z})], \tag{24}$$

where $\mathcal{H}[q]$ is the entropy of the variational posterior. Optimization alternates between updating $q(\mathbf{z})$ (E-step) and θ (M-step), guaranteeing monotonic convergence of the likelihood.

S4.2 Derivation of Characteristic Phase Lifetimes

The diagonal elements A_{kk} of the learned transition matrix encode the persistence of dynamic regimes. Here we derive the mapping from these discrete probabilities to the continuous physical timescales τ_k.

In the discrete HMM, the number of steps L the system dwells in state k before switching is geometrically distributed with success probability $(1 - A_{kk})$. The expected dwell duration is

$$\mathbb{E}[\text{Dwell time}] = \frac{\Delta t}{1 - A_{kk}}. \tag{25}$$

With respect to the continuous time, the phase decay is modeled as an exponential process e^{-t/τ_k}. Equating the discrete transition probability to the continuous decay

over one step, we have

$$A_{kk} \approx e^{-\Delta t/\tau_k} \implies \tau_k \approx -\frac{\Delta t}{\ln A_{kk}}, \tag{26}$$

where $\Delta t = \text{median}(t_{i+1} - t_i)$ is robustly estimated from the sampling grid to handle non-uniform experimental sampling.

Large A_{kk} (near 1) implies $\ln A_{kk} \to 0$ and $\tau_k \to \infty$, correctly assigning high weight to long-lived stable states; transient states with lower A_{kk} receive lower weights.

S5 Consistency Analysis of Phase-Aware Distillation

This section establishes a statistical consistency justification for the proposed PAKD pipeline. We show that (i) the HMM-derived phase-aware weights converge to a well-defined population target, and (ii) the resulting empirical distillation objective consistently approximates its population counterpart.

S5.1 Consistency of Phase Posteriors, Lifetimes, and Weights

Let $\hat{\theta}_T$ denote the HMM parameter estimator from teacher-generated trajectories of length T, and let θ^* be the population parameter. Define the population phase posterior $\gamma_t^*(k) := p_{\theta^*}(z_t = k \mid \mathbf{y}_{1:T})$, its estimator $\hat{\gamma}_t(k) := p_{\hat{\theta}_T}(z_t = k \mid \mathbf{y}_{1:T})$, and the population lifetime $\tau_k^* := -\Delta t / \ln A_{kk}^*$ with plug-in estimator $\hat{\tau}_k := -\widehat{\Delta t} / \ln \hat{A}_{kk}$.

The population phase-aware weight is

$$w_t^* := \sum_{k=1}^{K} \gamma_t^*(k)\, \tau_k^*, \tag{27}$$

and the estimated weight is

$$\hat{w}_t := \sum_{k=1}^{K} \hat{\gamma}_t(k)\, \hat{\tau}_k. \tag{28}$$

Proposition 1 (Consistency of HMM-based phase quantities) *Assume:*

1. *the teacher-generated trajectories are sampled from an ergodic, identifiable K-state HMM with true parameter θ^*;*
2. *the parameter space Θ is compact, θ^* is an interior point, and the emission family is regular;*
3. *$\hat{\theta}_T \xrightarrow{p} \theta^*$ as $T \to \infty$;*
4. *$A_{kk}^* \in (0,1)$ for all k, and $\widehat{\Delta t} \xrightarrow{p} \Delta t > 0$.*

Then, for each fixed t and k,

$$\hat{\gamma}_t(k) \xrightarrow{p} \gamma_t^*(k), \qquad \hat{\tau}_k \xrightarrow{p} \tau_k^*, \qquad \hat{w}_t \xrightarrow{p} w_t^*. \tag{29}$$

Proof Since $\hat{\theta}_T \xrightarrow{p} \theta^*$ and the forward–backward map is continuous in the model parameters under standard positivity and regularity conditions, the continuous mapping theorem gives $\hat{\gamma}_t(k) \xrightarrow{p} \gamma_t^*(k)$.

The lifetime map $(\Delta t, A_{kk}) \mapsto -\Delta t / \ln A_{kk}$ is continuous on $(0,\infty) \times (0,1)$. Since $\widehat{\Delta t} \xrightarrow{p} \Delta t$ and $\hat{A}_{kk} \xrightarrow{p} A_{kk}^*$, a further application of the continuous mapping theorem gives $\hat{\tau}_k \xrightarrow{p} \tau_k^*$.

Finally, $\hat{w}_t = \sum_k \hat{\gamma}_t(k)\hat{\tau}_k$ is a finite sum of products of consistent estimators. Since products and finite sums preserve convergence in probability, $\hat{w}_t \xrightarrow{p} w_t^*$. □

Remark 1 Proposition 1 implies that the phase-aware importance weights used in PAKD consistently estimate a well-defined population quantity: long-lived slow phases receive asymptotically stable higher weights, while short-lived transient phases receive asymptotically smaller weights.

S5.2 Consistency of Phase-Aware Empirical Risk Minimisation

Let $x = (\mathbf{y}_0, t)$ be the student's input. Define the single-sample weighted loss

$$\ell(x; \vartheta, w) := w\Big(w_{\text{out}} \|f_S(x;\vartheta) - f_T(x)\|_2^2 + w_{\text{hid}} \|W_P h_S(x;\vartheta) - h_T(x)\|_2^2\Big), \tag{30}$$

the population risk $\mathcal{R}(\vartheta) := \mathbb{E}[\ell(X; \vartheta, w^*(X))]$, and the empirical phase-aware risk

$$\hat{\mathcal{R}}_n(\vartheta) := \frac{1}{n} \sum_{i=1}^{n} \ell(X_i; \vartheta, \hat{w}_i). \tag{31}$$

We impose the following regularity conditions:

(A1) The student parameter space Θ_S is compact.
(A2) For a.e. x, the map $\vartheta \mapsto \ell(x; \vartheta, w)$ is continuous.
(A3) There exists an integrable envelope $M(X)$ such that $|\ell(X; \vartheta, w)| \le M(X)$ for all ϑ and admissible w.
(A4) $\hat{w}_i - w_i^* \xrightarrow{p} 0$ for each i.
(A5) The class $\{\ell(\cdot; \vartheta, w^*(\cdot)) : \vartheta \in \Theta_S\}$ is Glivenko–Cantelli [4].

Theorem 2 (Uniform consistency of the empirical phase-aware risk) *Under assumptions* (A1)–(A5),

$$\sup_{\vartheta \in \Theta_S} \big|\hat{\mathcal{R}}_n(\vartheta) - \mathcal{R}(\vartheta)\big| \xrightarrow{p} 0 \qquad \text{as } n \to \infty. \tag{32}$$

Proof Decompose:

$$\sup_{\vartheta \in \Theta_S} |\hat{\mathcal{R}}_n(\vartheta) - \mathcal{R}(\vartheta)| \le I_n + II_n, \tag{33}$$

where

$$I_n := \sup_{\vartheta \in \Theta_S} \left| \frac{1}{n} \sum_i \big(\ell(X_i; \vartheta, \hat{w}_i) - \ell(X_i; \vartheta, w_i^*)\big) \right|, \tag{34}$$

$$II_n := \sup_{\vartheta \in \Theta_S} \left| \frac{1}{n} \sum_i \ell(X_i; \vartheta, w_i^*) - \mathbb{E}[\ell(X; \vartheta, w^*(X))] \right|. \tag{35}$$

By (A5), the Glivenko–Cantelli property gives $II_n \xrightarrow{p} 0$. Since ℓ is linear in w,

$$\ell(X_i; \vartheta, \hat{w}_i) - \ell(X_i; \vartheta, w_i^*) = (\hat{w}_i - w_i^*)\, g(X_i, \vartheta), \tag{36}$$

where g is uniformly bounded by the envelope in (A3). Since $\hat{w}_i - w_i^* \xrightarrow{p} 0$ by (A4), standard dominated convergence arguments give $I_n \xrightarrow{p} 0$. Combining the above results yields the conclusion. □

Theorem 3 (Consistency of the PAKD estimator) *Under the conditions of Theorem 2 and assuming the population minimiser* $\vartheta^* = \arg\min_{\vartheta \in \Theta_S} \mathcal{R}(\vartheta)$ *is unique,*

$$\hat{\vartheta}_n := \arg\min_{\vartheta \in \Theta_S} \hat{\mathcal{R}}_n(\vartheta) \xrightarrow{p} \vartheta^* \qquad \text{as } n \to \infty. \tag{37}$$

Proof By Theorem 2, the empirical risk converges uniformly to the population risk. Since Θ_S is compact and $\mathcal{R}(\vartheta)$ has a unique minimiser, the standard argmin consistency theorem for M-estimators applies. □

Remark 2 Theorems 2 and 3 show that once the HMM-derived quantities are consistently estimated, minimising the empirical PAKD objective is asymptotically equivalent to minimising the population objective in which different dynamical regimes are weighted according to their latent persistence and posterior occupancy.

Remark 3 (Approximation error of teacher model) The consistency statement is relative to the teacher model induced target risk, not the full physical ground-truth generator. The total discrepancy has three parts, the approximation error of the teacher model, the weight estimation error, and the distillation error of the student model. We address the latter two here. Reducing the teacher model's approximation error through better training or ensemble methods is a separate direction.

References

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

[2] 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)

[3] 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)

[4] Vaart, A.W.: Asymptotic Statistics vol. 3. Cambridge university press, Cambridge (2000)

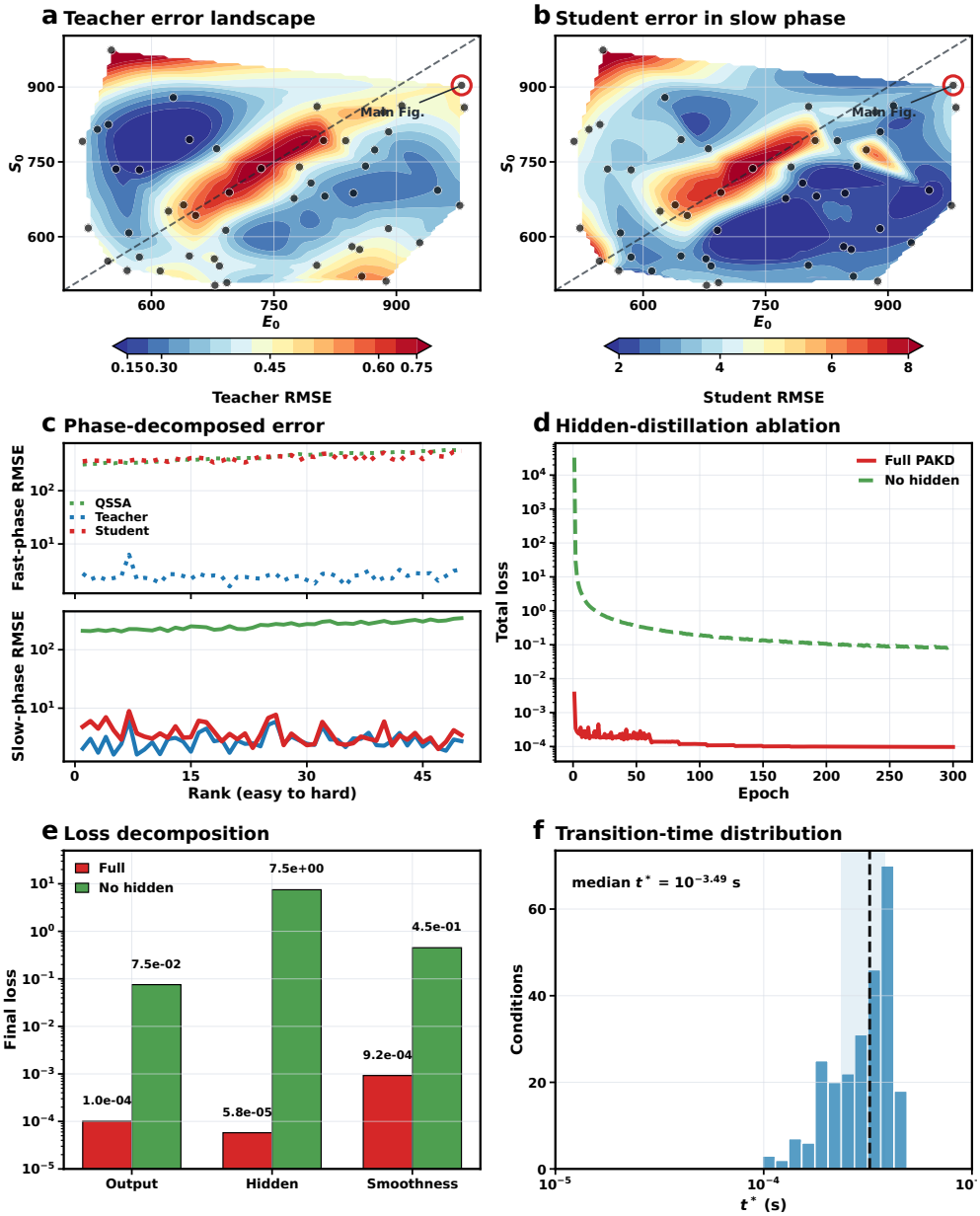

Fig. S1: **Training diagnostics and ablation analysis of the MM reaction surrogate.** (**a**) Teacher RMSE landscape over (E_0, S_0) for 50 test conditions. The red circle marks the main-figure condition. (**b**) Student RMSE in the slow phase only. (**c**) Phase-decomposed RMSE (fast, top; slow, bottom) by QSSA difficulty. (**d**) Hidden-distillation ablation. (**e**) Loss-component breakdown at epoch 300. Without hidden loss, the hidden-layer term reaches 7.5 (a.u.). Full PAKD suppresses it to 5.8×10^{-5}. Output and smoothness terms are also reduced. (**f**) Distribution of t^* across 50 test conditions (median $10^{-3.49}$ s), separating the fast and slow phases.

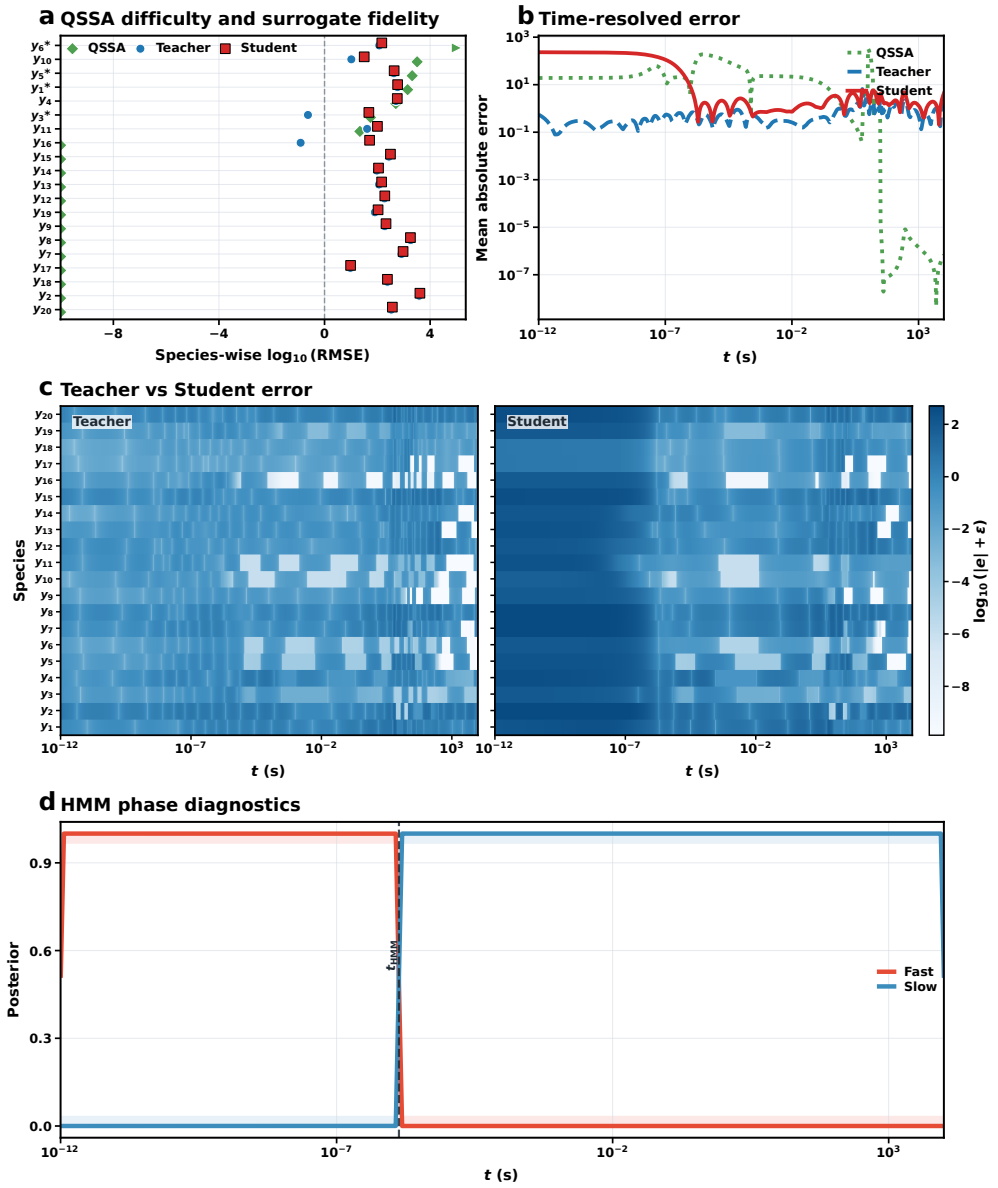

Fig. S2: **Diagnostics and error analysis of the POLLU surrogate.** (**a**) Species-wise RMSE ($\log_{10}$) ranked by QSSA difficulty. Asterisks mark the main-figure species. (**b**) Time-resolved mean absolute error for the base initial condition. (**c**) Error heatmaps (teacher, left; student, right) on a shared $\log_{10}$ scale. (**d**) HMM posterior probabilities for fast and slow phases. The dashed line marks the inferred transition time t_{HMM}.

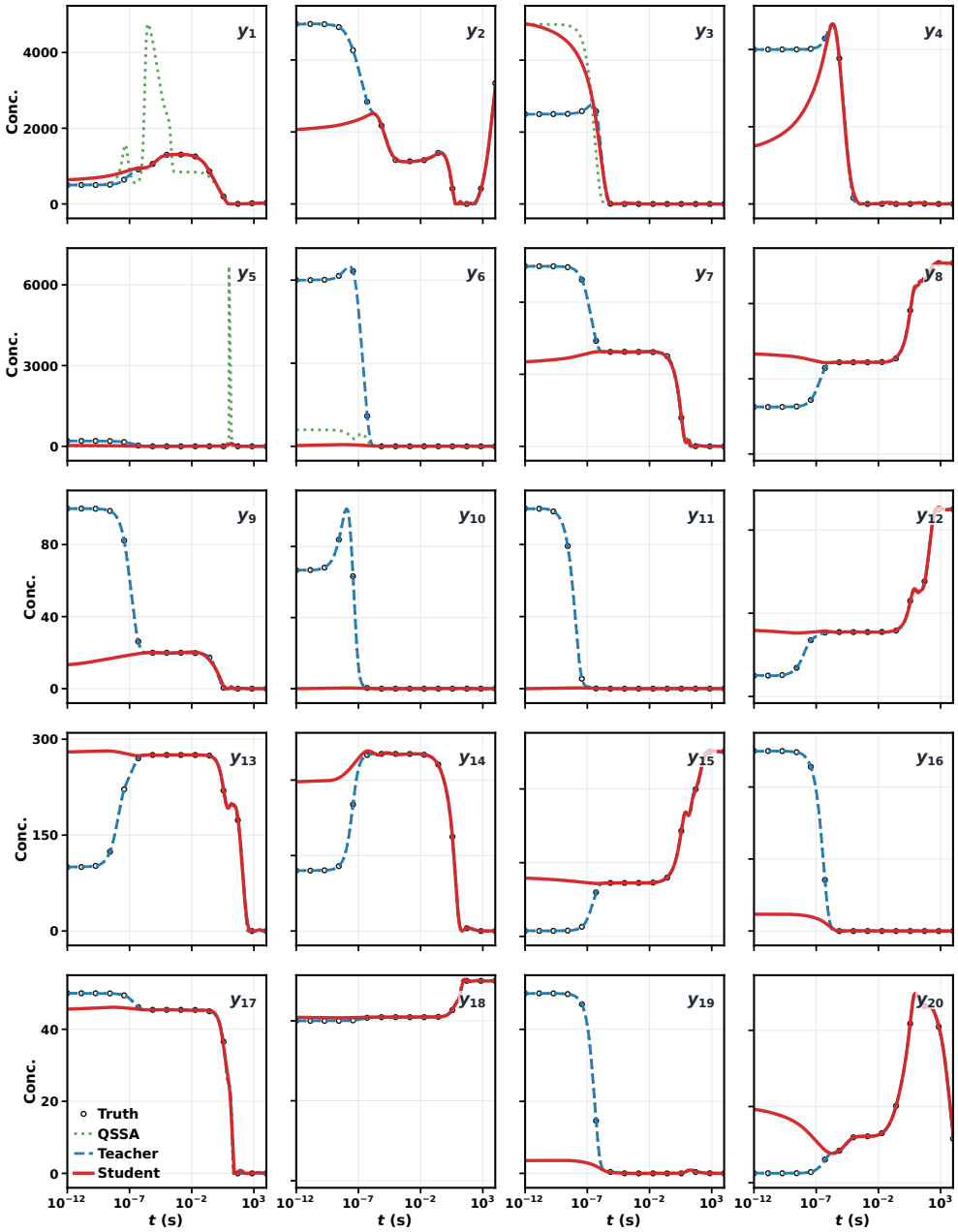

Fig. S3: **Trajectory overview for all 20 species in the POLLU problem.** Concentration profiles of species y_1 through y_{20} under the base initial condition, comparing the ground truth, QSSA approximation, teacher surrogate, and PAKD student.

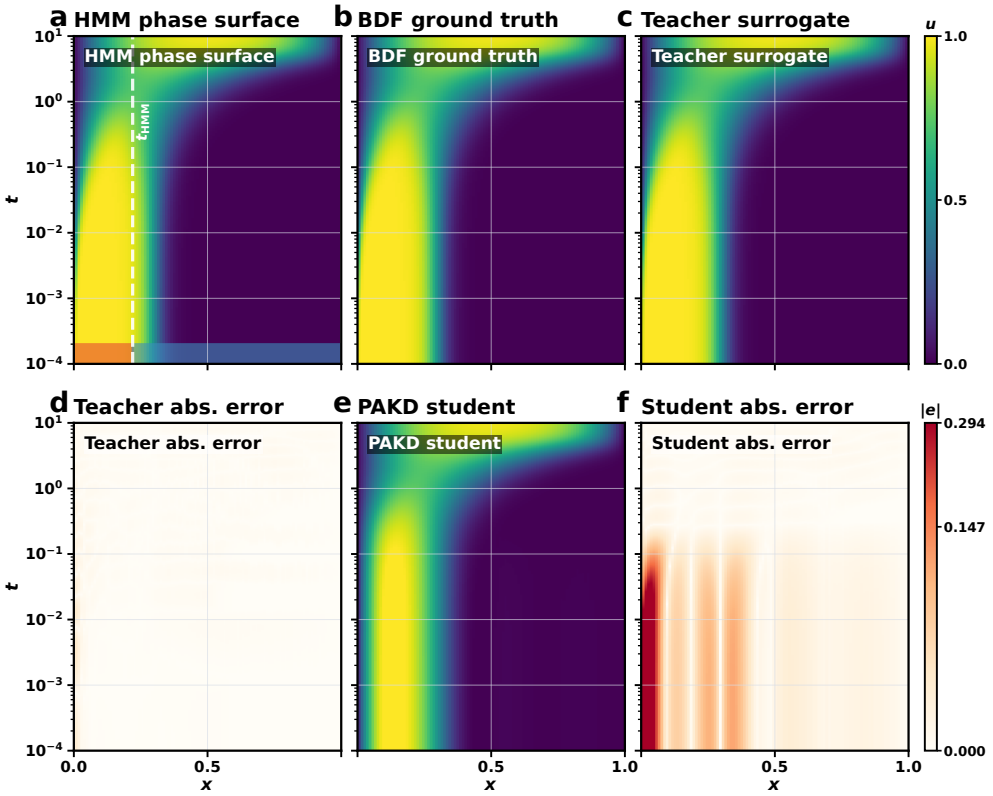

Fig. S4: **Space-time heatmaps of the Fisher-KPP surrogate and PAKD student.** (**a**) HMM-discovered phase boundary t^*_{HMM} overlaid on the BDF ground truth; the bottom color band marks the fast and slow regimes. (**b**–**c**) BDF reference and teacher surrogate. (**d**) Teacher absolute error, essentially zero across the domain. (**e**–**f**) PAKD student prediction and its absolute error, showing residual error concentrated in the early fast-phase transient ($t \lesssim t^*_{\mathrm{HMM}}$).

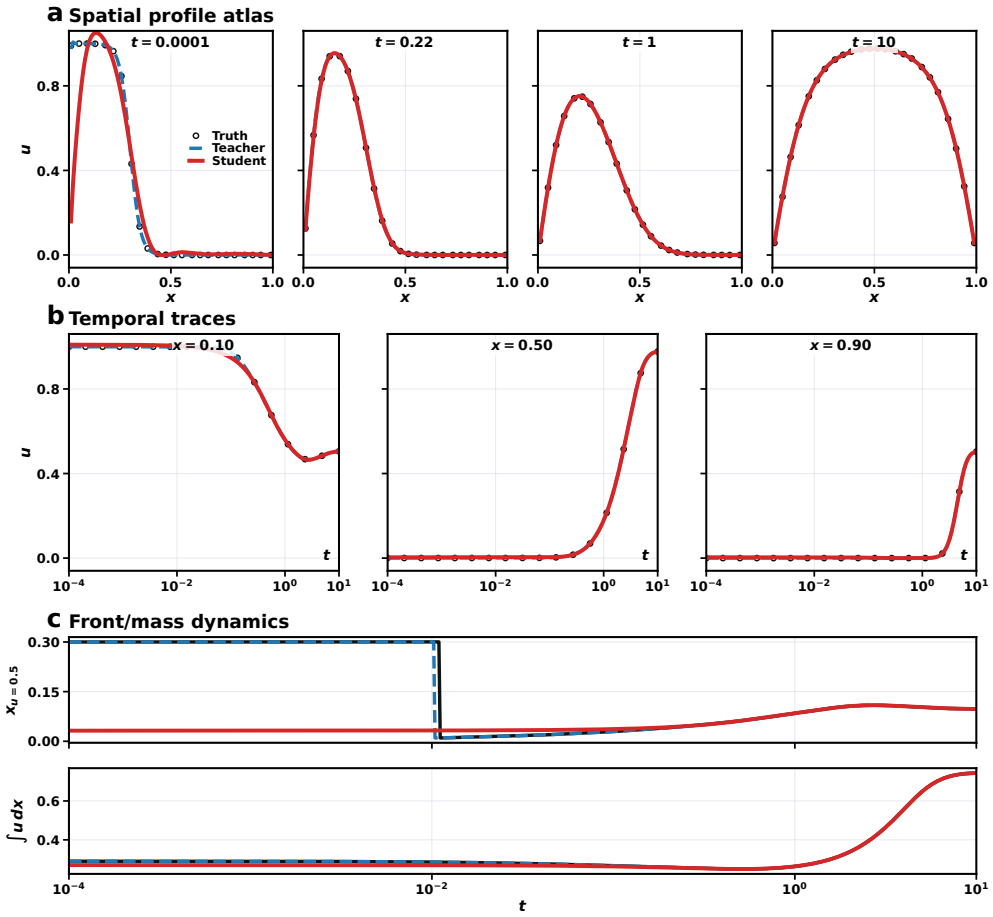

Fig. S5: **Extended trajectory verification of the Fisher–KPP surrogate. a,** Spatial profiles of $u(x,t)$ at four representative times ($t = 10^{-4}$, 0.22, 1, 10). **b,** Temporal traces at fixed spatial locations ($x = 0.10$, 0.50, 0.90), illustrating the early decay near the centre, the delayed rise at the domain edge, and the intermediate transition. **c,** Front position $x_{u=0.5}$ (top) and integrated mass $\int u\,\mathrm{d}x$ (bottom) versus time. Both macroscopic observables are closely reproduced by the teacher (dashed) and student (solid) compared with the BDF reference (markers), confirming accurate wave propagation speed and total mass dynamics.

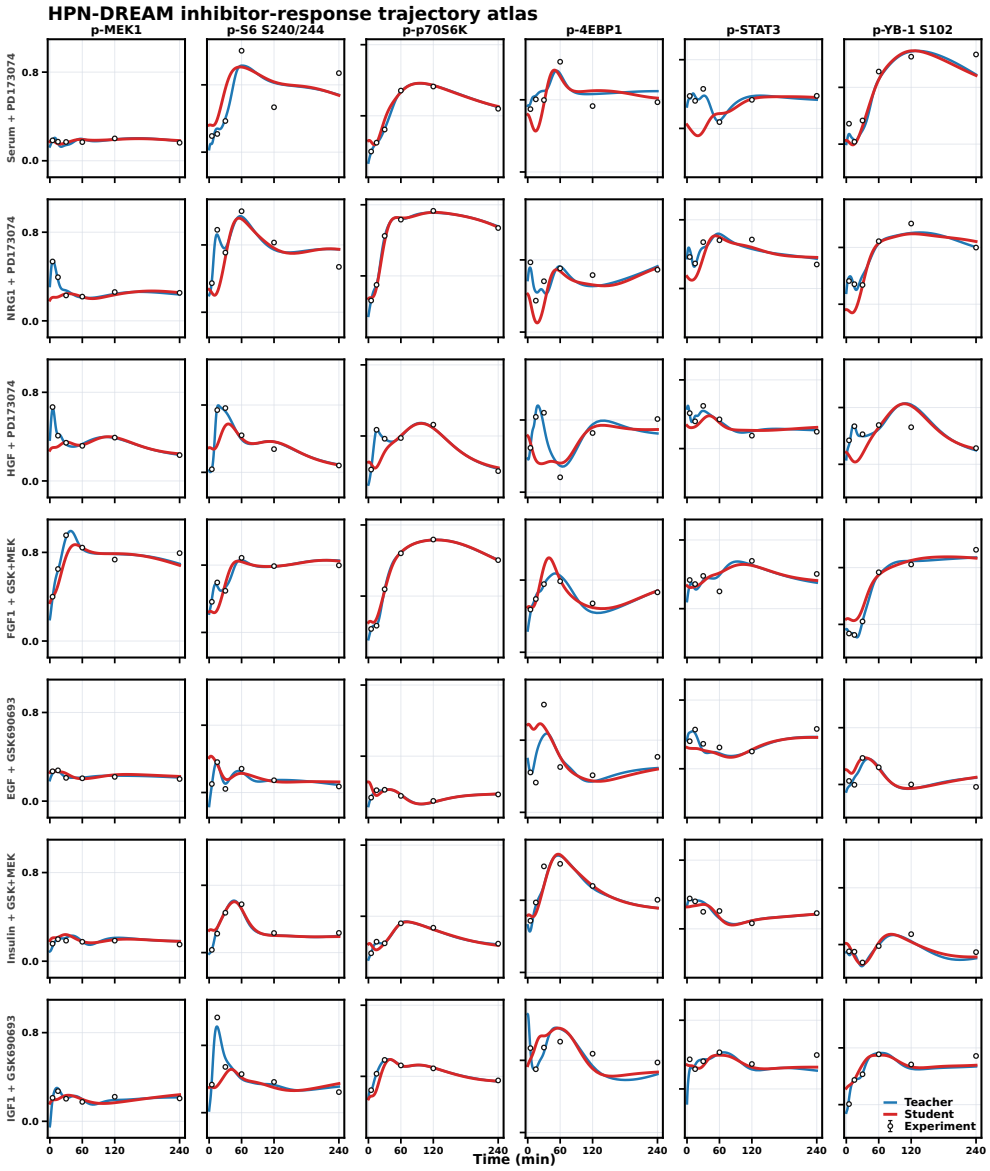

Fig. S6: **Supplementary inhibitor-response trajectory overview for the HPN-DREAM MCF7 signalling dataset.** Representative protein trajectories are shown across selected stimulation–inhibitor conditions, comparing the surrogate teacher, the phase-aware distilled student, and experimental measurements. Rows correspond to perturbation conditions involving pathway inhibitors, while columns correspond to selected phosphoprotein readouts, including p-MEK1, p-S6 S240/244, p-p70S6K, p-4EBP1, p-STAT3, and p-YB-1 S102.

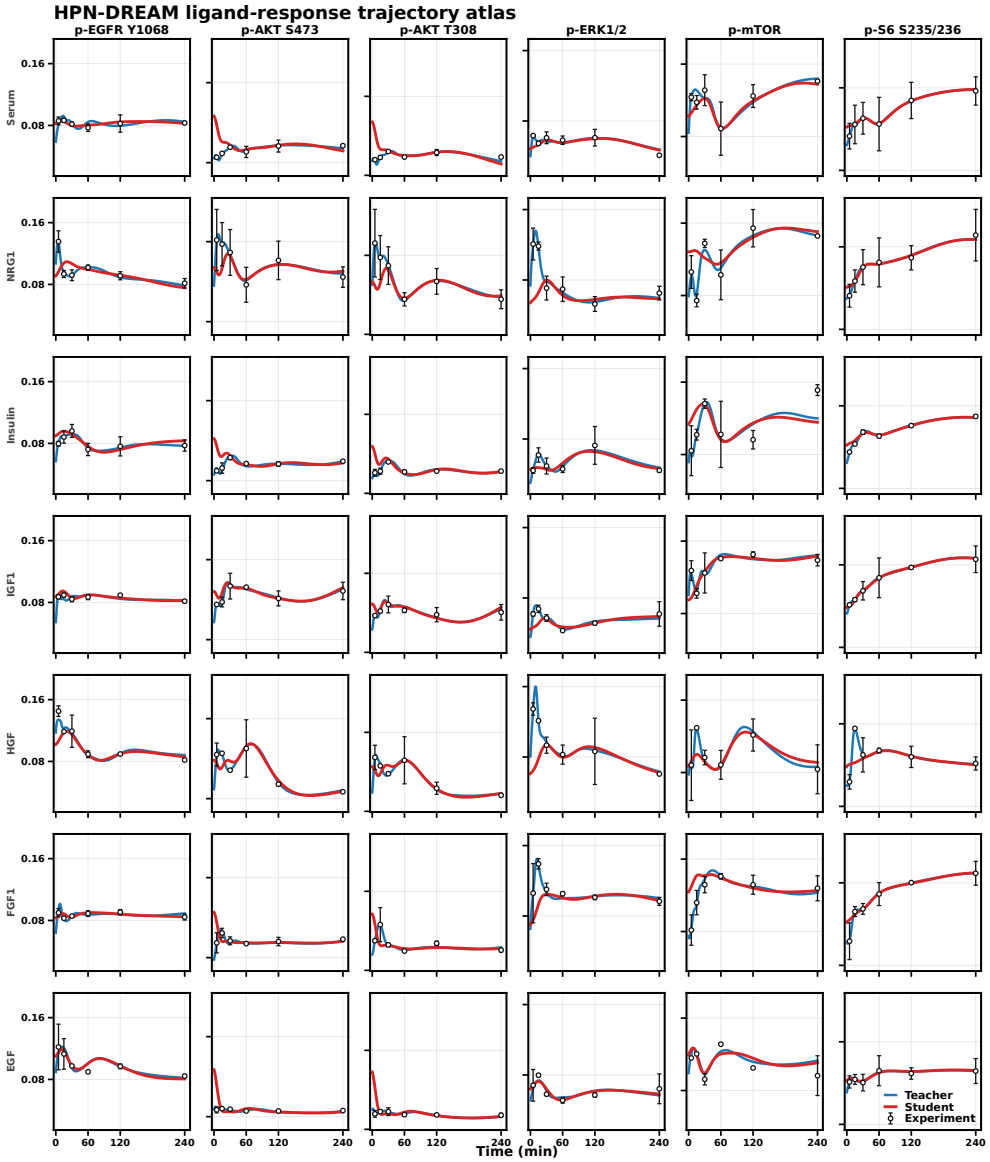

Fig. S7: **Supplementary ligand-response trajectory overview for the HPN-DREAM MCF7 signalling dataset.** Representative trajectories are shown across ligand stimulation conditions, comparing the surrogate teacher, the phase-aware distilled student, and experimental measurements with uncertainty bars. Rows correspond to extracellular ligands, including Serum, NRG1, Insulin, IGF1, HGF, FGF1, and EGF, while columns correspond to selected phosphoprotein readouts, including p-EGFR Y1068, p-AKT S473, p-AKT T308, p-ERK1/2, p-mTOR, and p-S6 S235/236.