

Koopman Operator Inference for the Chemical Master Equation

Aditya Dendukuri* and Linda R. Petzold†

Abstract. We infer kinetic rates of stochastic reaction networks from molecular-state pairs separated by a fixed time lag, without constructing the large state space of the Chemical Master Equation. Extended dynamic mode decomposition estimates the finite-time evolution of selected functions of the molecular state. A matrix logarithm then gives a projected generator whose species rows determine the rates through the known reaction stoichiometry. A finite set of functions introduces a systematic bias. We construct additional functions from the reaction channels and prove that each level of closure moves the first bias term to higher order. Spectral formulas identify the omitted modes responsible for this bias and show when the logarithm amplifies sampling error. Exact calculations and experiments on five reaction networks confirm the predicted error behavior and show when the added functions improve finite-sample rate recovery.

Key words. Koopman operator, extended dynamic mode decomposition, chemical master equation, stochastic reaction network, matrix logarithm, reaction rate inference

MSC codes. 65F60, 37M25, 92C42, 60J27

1. Introduction. The Chemical Master Equation (CME) governs the probability distribution of molecular copy numbers in a well-mixed stochastic reaction network [8, 18]. On a finite state space, the probability vector $p(t)$ satisfies

$$\frac{d}{dt}p(t) = \mathbf{A}(c)p(t),$$

where the reaction stoichiometry, propensities, and rate vector c determine the generator $\mathbf{A}(c)$ [9]. Our inverse problem is to recover c from pairs of molecular states observed at a fixed time separation.

The molecular state space grows rapidly with the number of species and their copy numbers. A likelihood based on the full CME requires a transition model whose dimension is the number of retained states. Finite-state projection reduces this cost by truncation [19]. Moment methods replace the distribution by a finite set of moments, but generally require a closure approximation [6, 23]. We instead pose the inverse problem on selected functions of the molecular state. The number of these observables, which we also call the dictionary, need not grow with the number of molecular states.

Let $\mathbf{X}(t)$ denote the vector of molecular copy numbers at time t , and let $\mathbf{x}$ denote a possible initial state. An *observable* is a function f of these copy numbers. Its conditional expected value after time t is

$$(\mathcal{K}_t f)(\mathbf{x}) = \mathbb{E}[f(\mathbf{X}(t)) \mid \mathbf{X}(0) = \mathbf{x}].$$

*Department of Computer Science, University of California Santa Barbara (adendukuri@ucsb.edu).

†Departments of Mechanical Engineering and Computer Science, University of California Santa Barbara (petzold@engineering.ucsb.edu).

The operators $\{\mathcal{K}_t\}_{t \geq 0}$ form the Koopman semigroup. The instantaneous rate of change of an observable is given by

$$(\mathcal{A}f)(\mathbf{x}) = \lim_{t \downarrow 0} \frac{(\mathcal{K}_t f)(\mathbf{x}) - f(\mathbf{x})}{t}.$$

This operator $\mathcal{A}$ is the Koopman generator. For the CME, $\mathcal{A} = \mathbf{A}(c)^\top$. From observations separated by a time h , extended dynamic mode decomposition (EDMD) approximates $\mathcal{K}_h$ on d observables [25]. Its principal matrix logarithm approximates $\mathcal{A}$, and the rows associated with the species determine the reaction rates. The fitted matrix has dimension d , independent of the number of molecular states.

The estimator has three stages:

$$(1.1) \quad \begin{aligned} \hat{\mathbf{K}}_h &= \Psi_Y \Psi_X^\dagger && \text{(finite-time Koopman matrix),} \\ \hat{\mathbf{A}}_h &= h^{-1} \log \hat{\mathbf{K}}_h && \text{(Koopman generator matrix),} \\ \hat{c} &= \mathcal{R}(\hat{\mathbf{A}}_h) && \text{(reaction rates).} \end{aligned}$$

Here Ψ_X and Ψ_Y contain the selected observables evaluated at the two states in each pair, and $\mathcal{R}$ extracts rates from the species rows. The fitted square matrices have dimension d , the number of selected observables.

Prior work uses Koopman lifting and the logarithm of an EDMD matrix to recover governing equations from snapshot pairs [17, 5]. Other work uses generator EDMD to approximate the infinitesimal operator directly and identify drift and diffusion coefficients [16]. Stochastic-model parameters can also be fitted to an estimated operator [22], while estimates at several observation intervals can be combined to recover a Markov-process generator [24].

For Markov jump processes, Niemann, Klus, and Schütte recover transition-rate combinations from a Koopman-generator model of the large-population diffusion limit [20]. Their construction uses short-time Kramers–Moyal estimates of drift and diffusion. Classical generator-inference methods fit a state-to-state rate matrix from a discretely observed transition matrix [3]. We restrict the exact jump-process generator to selected mass-action observables, so the matrix dimension equals the dictionary dimension rather than the number of molecular states.

Gupta and Khammash approximate the continuous-time Koopman operator of a specified reaction network through its dominant decay modes [10]. They use this representation to predict moments across initial states and compute parameter sensitivities and spectral densities. In our inverse problem, we determine unknown reaction rates from state pairs at one fixed separation through (1.1). The finite-dimensional projection therefore occurs before logarithmic inversion and can bias the rates even when the projected finite-time operator is known exactly.

Mass-action kinetics provides a natural observable set. If ϕ_{ν_k} is the propensity monomial for reaction k , then

$$(1.2) \quad \mathcal{A}x_s = \sum_{k=1}^M \zeta_{s,k} c_k \phi_{\nu_k}.$$

The coefficient of each propensity is therefore a stoichiometric change multiplied by a reaction rate. Once the observable set contains these monomials, identifiable rates can be read from the species rows of the exact Koopman generator.

The same structure gives an observable hierarchy. The generator is affine in the rates,

$$(1.3) \quad \mathcal{A}(c) = \sum_{k=1}^M c_k \mathcal{A}_k, \quad (\mathcal{A}_k f)(\mathbf{x}) = \phi_{\nu_k}(\mathbf{x}) [f(\mathbf{x} + \zeta_k) - f(\mathbf{x})],$$

and every unit-rate channel operator $\mathcal{A}_k$ is known once the reaction set is specified. Repeated generator images therefore have the expansion

$$\mathcal{A}(c)^j x_s = \sum_{k_1, \dots, k_j=1}^M c_{k_j} \cdots c_{k_1} \mathcal{A}_{k_j} \cdots \mathcal{A}_{k_1} x_s.$$

At any fixed depth, all functions on the right can be generated without the unknown rates. They form a reaction-defined closure hierarchy for the species rows that contain the physical parameters. The falling-factorial basis makes the growth of this hierarchy explicit. If $\mathcal{P}_q$ denotes polynomials of degree at most q and reaction k has order $o_k = |\nu_k|$, then

$$(1.4) \quad \mathcal{A}_k \mathcal{P}_q \subseteq \mathcal{P}_{q+o_k-1}.$$

The jump difference lowers degree before multiplication by the propensity. Polynomial spaces are therefore invariant for networks with only zero- and first-order reactions, and the projected logarithm is exact for sufficiently small admissible h . For nonlinear reactions, (1.4) bounds the new degrees introduced at each closure depth. This gives exact closure for linear reaction networks and a controlled, rate-independent enlargement for nonlinear networks.

Finite-time data determine a projected operator. If P projects onto the selected observable space $\mathcal{F}$, then exact finite-time expectations give the generator

$$(1.5) \quad h^{-1} \log \left(P e^{h\mathcal{A}}|_{\mathcal{F}} \right), \quad h^{-1} \log \left(P e^{h\mathcal{A}}|_{\mathcal{F}} \right) \neq P\mathcal{A}|_{\mathcal{F}}$$

in general. We therefore obtain biased species rows even before sampling error enters.

We reduce this error by adding successive generator images of the species counts used for rate extraction. An observable space has *closure depth* r when it contains $\mathcal{A}x_s, \dots, \mathcal{A}^r x_s$ for every selected species x_s . This definition links the reaction-network identity directly to the finite-time error in the physical rates.

The analysis gives a species-row identifiability condition, a modal-residue description of the omitted Koopman modes, and a rate-bias theorem at every closure depth. We also propagate sampling error through the logarithm. Exact finite-state matrices recover the predicted bias orders, and independent trajectory data measure the effect of the added observables on four further networks. Throughout, the reaction set is known, observations occur in pairs, and the selected species coefficients identify the rates. The unknowns are the rates within this fixed reaction set.

2. Reaction Networks and the Chemical Master Equation. We first describe the reaction network and its two generators. The forward generator evolves probabilities, while the backward generator evolves the observables used to recover the rates.

2.1. Reaction network model. Consider a network of N molecular species and M reaction channels.

Definition 2.1 (Stochastic reaction network). A stochastic reaction network on N species $\mathcal{S}_1, \dots, \mathcal{S}_N$ consists of M reaction channels indexed by $k = 1, \dots, M$, each specified by:

- a stoichiometric vector $\zeta_k = \beta_k - \nu_k \in \mathbb{Z}^N$, where $\nu_{ik}, \beta_{ik} \in \mathbb{N}$ are the reactant and product stoichiometric coefficients for species $\mathcal{S}_i$; and
- a propensity function $\alpha_k : \mathbb{N}_0^N \rightarrow [0, \infty)$, where, conditional on state $\mathbf{x}$ at time t , the probability that reaction k fires in $[t, t + dt)$ is $\alpha_k(\mathbf{x}) dt + o(dt)$.

The state $\mathbf{x} = (x_1, \dots, x_N)^\top \in \mathbb{N}_0^N$ records the molecular copy numbers, and $\mathbf{c} = (c_1, \dots, c_M)^\top$ denotes the vector of reaction rates.

Under mass-action kinetics, reaction k has propensity

$$(2.1) \quad \alpha_k(\mathbf{x}) = c_k \phi_{\nu_k}(\mathbf{x}), \quad \phi_{\nu_k}(\mathbf{x}) := \prod_{i=1}^N \binom{x_i}{\nu_{ik}},$$

where $c_k > 0$ is the rate constant and the binomial factor counts the possible reactant groups. We call ϕ_{ν_k} the *propensity monomial* of reaction k . The reactant stoichiometry fixes this monomial, while c_k is unknown. A source reaction has monomial 1, and a unimolecular reaction in species i has monomial x_i . A bimolecular reaction has either $x_i x_j$ or $x_i(x_i - 1)/2$; trimolecular reactions give the analogous degree-three factors.

2.2. Chemical master equation. The reaction channels define a continuous-time Markov chain and its CME generator.

Definition 2.2 (CME generator). The probability mass function $p(\mathbf{x}, t) = \mathbb{P}[\mathbf{X}(t) = \mathbf{x}]$ satisfies the Chemical Master Equation [7, 9]

$$(2.2) \quad \frac{d}{dt} p(\mathbf{x}, t) = \sum_{k=1}^M [\alpha_k(\mathbf{x} - \zeta_k) p(\mathbf{x} - \zeta_k, t) - \alpha_k(\mathbf{x}) p(\mathbf{x}, t)].$$

On a finite closed state set $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, let $\mathbf{p}(t) = (p(\mathbf{x}_1, t), \dots, p(\mathbf{x}_n, t))^\top$. Then $\dot{\mathbf{p}}(t) = \mathbf{A}(\mathbf{c}) \mathbf{p}(t)$, where $\mathbf{A}(\mathbf{c}) \in \mathbb{R}^{n \times n}$ has entries

$$\mathbf{A}(\mathbf{c})_{\mathbf{x}_i, \mathbf{x}_j} = \sum_{k: \mathbf{x}_i = \mathbf{x}_j + \zeta_k} \alpha_k(\mathbf{x}_j) \quad (i \neq j), \quad \mathbf{A}(\mathbf{c})_{\mathbf{x}_j, \mathbf{x}_j} = - \sum_{k: \zeta_k \neq 0} \alpha_k(\mathbf{x}_j),$$

and all columns sum to zero. The state set is closed if every enabled reaction remains in $\mathcal{X}$. An FSP truncation adds a sink state for transitions that leave its retained set.

An inverse problem posed directly in terms of $\mathbf{A}(\mathbf{c})$ uses probability vectors of dimension $n = |\mathcal{X}|$, which can grow combinatorially with the copy numbers. The Koopman formulation follows selected conditional expectations.

3. From Koopman Observables to Reaction Rates. The Koopman generator acts on functions of the molecular state rather than on state probabilities. For mass-action networks, its action on a species count is a known linear combination of the reaction rates. This identity places the unknown rates in selected rows of a small generator matrix.

3.1. Koopman semigroup and generator.

Definition 3.1 (Koopman semigroup). An observable is a measurable function $f : \mathcal{X} \rightarrow \mathbb{R}$. For bounded observables f , let the Koopman semigroup $\{\mathcal{K}_t\}_{t \geq 0}$ act by

$$(3.1) \quad (\mathcal{K}_t f)(\mathbf{x}) = \mathbb{E}[f(\mathbf{X}(t)) \mid \mathbf{X}(0) = \mathbf{x}].$$

The semigroup satisfies $\mathcal{K}_0 = I$ and $\mathcal{K}_{t+s} = \mathcal{K}_t \mathcal{K}_s$ for all $t, s \geq 0$. It is the evolution of observable functions dual to the probability evolution generated by $\mathbf{A}(c)$.

Definition 3.2 (Koopman generator). Let the Koopman generator denote the instantaneous rate of change of an observable under the Koopman semigroup:

$$(3.2) \quad (\mathcal{A}f)(\mathbf{x}) := \lim_{t \downarrow 0} \frac{(\mathcal{K}_t f)(\mathbf{x}) - f(\mathbf{x})}{t},$$

whenever this limit exists. For a stochastic reaction network, $\mathcal{A} = \mathbf{A}(c)^\top$, and we evaluate the limit to obtain

$$(3.3) \quad (\mathcal{A}f)(\mathbf{x}) = \sum_{k=1}^M \alpha_k(\mathbf{x}) [f(\mathbf{x} + \boldsymbol{\zeta}_k) - f(\mathbf{x})]$$

for any bounded observable f . It follows that $\frac{d}{dt} \mathbb{E}[f(\mathbf{X}(t))] = \mathbb{E}[(\mathcal{A}f)(\mathbf{X}(t))]$, which describes the evolution of expected observables. For unbounded polynomial observables, these identities require the corresponding moment and generator-domain conditions.

Remark 3.3 (Observable-space restriction). On a finite state space, the forward CME generator and backward Koopman generator are dual:

$$(3.4) \quad \begin{aligned} \langle f, \mathbf{A}(c)p \rangle &= \langle \mathbf{A}(c)^\top f, p \rangle, \\ \mathcal{A} &= \mathbf{A}(c)^\top, \\ \mathbb{E}_{p_0}[f(\mathbf{X}(t))] &= f^\top e^{t\mathbf{A}(c)} p_0 = p_0^\top e^{t\mathcal{A}} f. \end{aligned}$$

On a countable state space, the same dual relation holds when the generators and pairing have compatible domains, and both spaces may be infinite dimensional. The finite model used here is the restriction of $\mathcal{K}_t$ to the observables needed for rate recovery. Its dimension is the number of retained functions; its approximation error is analyzed in Sections 4–5.

3.2. Mass-action basis.

Definition 3.4 (Mass-action basis). For a multi-index $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_N) \in \mathbb{N}_0^N$, the corresponding mass-action monomial is

$$(3.5) \quad \phi_{\boldsymbol{\gamma}}(\mathbf{x}) = \prod_{i=1}^N \binom{x_i}{\gamma_i}.$$

For an order q at least as large as the largest reaction order, the mass-action dictionary

$$(3.6) \quad \boldsymbol{\psi}(\mathbf{x}) = (\psi_1(\mathbf{x}), \dots, \psi_d(\mathbf{x}))^\top$$

contains the monomials ϕ_γ with $|\gamma| \leq q$, ordered so that the constant and species coordinates appear first. Before dependencies are removed, it contains $d = \binom{N+q}{q}$ functions. We retain an independent subset on the state space under study while preserving the constant, the species, and a representation of every propensity monomial.

Before conservation reduction, each reaction propensity has the coordinate representation

$$(3.7) \quad \alpha_k(\mathbf{x}) = c_k \psi_{\ell(k)}(\mathbf{x}), \quad k = 1, \dots, M,$$

where $\ell(k)$ is the dictionary index for $\phi_{\nu_k} = \psi_{\ell(k)}$. After conservation reduction, the same monomial may instead be represented by a linear combination of dictionary functions. The corresponding extraction formula follows the basic single-coordinate case.

3.3. Generator applied to species counts. Let $\mathcal{F} = \text{span}\{\psi_1, \dots, \psi_d\}$ be the subspace spanned by the mass-action basis in Definition 3.4. For each reaction channel, let the unit-rate reaction operator be

$$(3.8) \quad (\mathcal{A}_k f)(\mathbf{x}) = \phi_{\nu_k}(\mathbf{x}) [f(\mathbf{x} + \zeta_k) - f(\mathbf{x})].$$

The full Koopman generator is therefore $\mathcal{A} = \sum_{k=1}^M c_k \mathcal{A}_k$. The specified reactions determine every $\mathcal{A}_k$ without knowledge of the rates.

The classical mass-action first-moment identity [9] now gives the coefficient rows used for rate identification and for the analysis of the projected matrix logarithm.

Lemma 3.5 (Generator applied to a species count). *Let ψ be the mass-action basis of Definition 3.4. Then, for each species s ,*

$$(3.9) \quad (\mathcal{A}x_s)(\mathbf{x}) = \sum_{k=1}^M \zeta_{s,k} c_k \phi_{\nu_k}(\mathbf{x}).$$

Here $\zeta_{s,k}$ is the s th component of ζ_k . In particular, $\mathcal{A}x_s \in \mathcal{F}$, and the expected species evolution satisfies

$$(3.10) \quad \frac{d}{dt} \mathbb{E}[x_s(\mathbf{X}(t))] = \sum_{k=1}^M \zeta_{s,k} c_k \mathbb{E}[\phi_{\nu_k}(\mathbf{X}(t))]$$

exactly at the species level.

Proof. A firing of channel k changes x_s by $\zeta_{s,k}$. We multiply this increment by its propensity and sum over channels to obtain both identities. ■

Matrix convention. Let P denote projection onto $\mathcal{F}$, and let $\mathbf{A}$ be the $d \times d$ coordinate matrix of $P\mathcal{A}|_{\mathcal{F}}$ in the ordered physical dictionary ψ , with coefficients stored by row. Thus $(\mathbf{A})_{i,j}$ is the coefficient of ψ_j in $P\mathcal{A}\psi_i$. The species coordinates are retained explicitly and, by Lemma 3.5, $P\mathcal{A}x_s = \mathcal{A}x_s$. Hence, the corresponding rows of $\mathbf{A}$ contain exact linear combinations of the rates. Later applications of the generator can leave $\mathcal{F}$, and the resulting projection error is the subject of Section 5.

Species-row structure. With rows labeled by the observable acted upon and columns by the resulting dictionary coordinate, the matrix has the form

$$(3.11) \quad \mathbf{A} = \begin{matrix} 1 \\ x_1 \\ \vdots \\ x_s \\ \vdots \\ x_N \\ \text{higher observables} \end{matrix} \begin{pmatrix} \psi_1 & \cdots & \psi_j & \cdots & \psi_d \\ 0 & \cdots & 0 & \cdots & 0 \\ (\mathbf{A})_{1,1} & \cdots & (\mathbf{A})_{1,j} & \cdots & (\mathbf{A})_{1,d} \\ \vdots & \ddots & \vdots & & \vdots \\ (\mathbf{A})_{s,1} & \cdots & (\mathbf{A})_{s,j} & \cdots & (\mathbf{A})_{s,d} \\ \vdots & & \vdots & \ddots & \vdots \\ (\mathbf{A})_{N,1} & \cdots & (\mathbf{A})_{N,j} & \cdots & (\mathbf{A})_{N,d} \\ * & \cdots & * & \cdots & * \end{pmatrix}.$$

The constant row is zero because $\mathcal{A}1 = 0$. For a species row, the generator action can be collected by dictionary column:

$$(3.12) \quad \begin{aligned} P\mathcal{A}x_s &= \mathcal{A}x_s = \sum_{k=1}^M \zeta_{s,k} c_k \psi_{\ell(k)} \\ &= \sum_{j=1}^d \left(\sum_{k: \ell(k)=j} \zeta_{s,k} c_k \right) \psi_j. \end{aligned}$$

Consequently,

$$(3.13) \quad (\mathbf{A})_{s,j} = \sum_{k: \ell(k)=j} \zeta_{s,k} c_k, \quad s = 1, \dots, N, \quad j = 1, \dots, d.$$

Equivalently, each propensity column has the species block

$$(3.14) \quad \begin{bmatrix} (\mathbf{A})_{1,j} \\ \vdots \\ (\mathbf{A})_{N,j} \end{bmatrix} = \sum_{k: \ell(k)=j} c_k \boldsymbol{\zeta}_k.$$

Reaction k therefore contributes the scaled stoichiometric vector $c_k \boldsymbol{\zeta}_k$ to its propensity column $\ell(k)$. The starred rows describe the action on higher observables and need not close in the retained dictionary.

When several reactions share a propensity function, one coefficient no longer determines one rate. The species rows still separate the rates when their stoichiometric changes are linearly independent. The following corollary states this condition and defines the extraction map.

Corollary 3.6 (Rate extraction). *Suppose reactions $k_1, \dots, k_g$ share the propensity observable ψ_j . Their rates and the corresponding species-column coefficients satisfy*

$$(3.15) \quad \begin{bmatrix} (\mathbf{A})_{1,j} \\ \vdots \\ (\mathbf{A})_{N,j} \end{bmatrix} = \begin{bmatrix} \zeta_{1,k_1} & \cdots & \zeta_{1,k_g} \\ \vdots & & \vdots \\ \zeta_{N,k_1} & \cdots & \zeta_{N,k_g} \end{bmatrix} \begin{bmatrix} c_{k_1} \\ \vdots \\ c_{k_g} \end{bmatrix}.$$

The rates are identifiable from this column if and only if the displayed stoichiometric matrix has full column rank. In that case, we recover the rates by multiplying the species-column vector by the pseudoinverse.

Proof. Collect the coefficients of the shared propensity in (3.9). The resulting linear system has a unique rate vector exactly when its stoichiometric matrix has full column rank. ■

Let $\mathcal{R}$ denote the fixed linear map obtained by collecting all full-rank instances of (3.15) from the species rows to the rate vector: $\mathcal{R}(\mathbf{A}) = \mathbf{c}$. This notation also covers conservation-reduced dictionaries, where a propensity is represented by a known linear combination of columns. The rank condition is verified for every numerical experiment; we may use additional observable rows to identify rates that the species rows alone cannot separate.

Species-row closure is weaker than invariance of $\mathcal{F}$ under $\mathcal{A}$. If channel k has reaction order o_k , the jump difference lowers the observable degree by at least one, so

$$\mathcal{A}_k \mathcal{P}_q \subseteq \mathcal{P}_{q+o_k-1}, \quad \mathcal{A}_k \mathbf{1} = 0,$$

where $\mathcal{P}_q$ contains polynomials of degree at most q . Thus, repeated application of the reaction operators for bimolecular and trimolecular channels can produce functions outside a fixed dictionary. The local expansions are therefore proved on a finite closed state space. Their extension to an unbounded network would require additional domain and moment bounds; the unbounded examples in Section 6 lie outside the stated scope of the theorem.

4. Estimating the Finite-Time Koopman Operator. Each observation is a pair $(\mathbf{X}(t_i), \mathbf{X}(t_i + h))$ separated by the same interval h . EDMD fits the finite-time Koopman operator from these pairs. We take its principal matrix logarithm, extract the rates, and then propagate the sampling error through these calculations.

For m paired observations, arrange the evaluated observables as columns:

$$(4.1) \quad \begin{aligned} \Psi_X &= [\psi(\mathbf{X}(t_1)) \quad \cdots \quad \psi(\mathbf{X}(t_m))], \\ \Psi_Y &= [\psi(\mathbf{X}(t_1 + h)) \quad \cdots \quad \psi(\mathbf{X}(t_m + h))]. \end{aligned}$$

When Ψ_X has full row rank, the empirical moment matrices and their induced operator are

$$(4.2) \quad \begin{aligned} \hat{\mathbf{G}}_0 &= \frac{1}{m} \Psi_X \Psi_X^\top, \\ \hat{\mathbf{G}}_1 &= \frac{1}{m} \Psi_Y \Psi_X^\top, \\ \hat{\mathbf{K}}_h &= \hat{\mathbf{G}}_1 \hat{\mathbf{G}}_0^{-1} = \Psi_Y \Psi_X^\dagger, \\ \hat{\mathbf{A}}_h &= h^{-1} \log \hat{\mathbf{K}}_h, \\ \hat{\mathbf{c}} &= \mathcal{R}(\hat{\mathbf{A}}_h). \end{aligned}$$

The dictionary contains the constant, selected species, and every mass-action propensity required for extraction. Fixed feature scaling is undone before the linear map $\mathcal{R}$ is applied. Rate estimates require full dictionary rank and a real principal logarithm.

4.1. Infinite-sample limit. Let μ be the law of the current state, possibly a mixture of transient observation laws. The population moment matrices have entries

$$(4.3) \quad \begin{aligned} (\mathbf{G}_0)_{ij} &= \mathbb{E}_\mu[\psi_i(\mathbf{X}(t))\psi_j(\mathbf{X}(t))], \\ (\mathbf{G}_1)_{ij} &= \mathbb{E}_\mu[\psi_i(\mathbf{X}(t+h))\psi_j(\mathbf{X}(t))]. \end{aligned}$$

The infinite-sample, or population, EDMD matrix is $\mathbf{K}_h = \mathbf{G}_1\mathbf{G}_0^{-1}$. It represents the $L^2(\mu)$ projection of $\mathcal{K}_h$ onto the dictionary span. For independent trajectories, or for one stationary ergodic trajectory, we apply entrywise laws of large numbers to obtain $\hat{\mathbf{G}}_i \rightarrow \mathbf{G}_i$. If $\mathbf{G}_0$ is invertible and $\mathbf{K}_h$ is admissible, we use continuity of inversion and of the principal logarithm to obtain

$$(4.4) \quad \hat{c} = \mathcal{R}(h^{-1} \log \hat{\mathbf{K}}_h) \longrightarrow \mathcal{R}(h^{-1} \log \mathbf{K}_h) = c_h.$$

Paired data therefore recover $\mathbf{K}_h$ and c_h in the infinite-sample limit. Increasing the sample size at fixed h removes sampling error but leaves the projection bias $c_h - c$. This convergence requires finite expectations for the Gram-matrix entries and applies to independent trajectories or to a stationary ergodic trajectory.

Sampling perturbations. Let $\Delta_i = \hat{\mathbf{G}}_i - \mathbf{G}_i$. At fixed invertible $\mathbf{G}_0$, we use first-order matrix perturbation to obtain

$$(4.5) \quad \hat{\mathbf{G}}_0^{-1} = \mathbf{G}_0^{-1} - \mathbf{G}_0^{-1}\Delta_0\mathbf{G}_0^{-1} + o_p(m^{-1/2}),$$

$$(4.6) \quad \begin{aligned} \hat{\mathbf{K}}_h - \mathbf{K}_h &= \Delta_1\mathbf{G}_0^{-1} - \mathbf{G}_1\mathbf{G}_0^{-1}\Delta_0\mathbf{G}_0^{-1} + o_p(m^{-1/2}) \\ &= (\Delta_1 - \mathbf{K}_h\Delta_0)\mathbf{G}_0^{-1} + o_p(m^{-1/2}). \end{aligned}$$

Applying the Fréchet derivative of the matrix logarithm and then the rate-extraction map, we obtain

$$(4.7) \quad \hat{c} - c_h = h^{-1}\mathcal{R}(D \log(\mathbf{K}_h) [(\Delta_1 - \mathbf{K}_h\Delta_0)\mathbf{G}_0^{-1}]) + o_p(m^{-1/2}).$$

The covariance of the two Gram matrices is therefore propagated jointly through inversion, the matrix logarithm, and stoichiometric extraction. For m independent trajectory contributions with finite second moments, the multivariate central limit theorem and delta method make the left side asymptotically normal, with covariance obtained from the displayed linear map. For independently sampled pairs, its short-interval rate variance is $\nu/h + O(1)$; hence the standard deviation is of order $(mh)^{-1/2}$. Whole-trajectory resampling is used below when pairs overlap.

5. Reducing Finite-Time Bias with Reaction-Generated Observables. Projection introduces bias because an observable can leave the retained space and return before the next observation. We describe these excursions in two ways. A block expansion gives their order in the observation interval, while a resolvent formula identifies the omitted Koopman modes that carry them. Reaction-generated observables remove the shortest excursions from the species rows used for rate extraction.

5.1. Projection bias in the species rows. Consider a full-support measure on a finite closed state space and let P be the orthogonal projection onto the independent dictionary span $\mathcal{F}$. Let

$$(5.1) \quad \begin{aligned} \mathbf{K}_h &:= P e^{h\mathbf{A}}|_{\mathcal{F}}, \\ \mathbf{A}_h &:= h^{-1} \log \mathbf{K}_h, \\ c_h &:= \mathcal{R}(\mathbf{A}_h). \end{aligned}$$

Here $\mathbf{K}_h$ is the projected finite-time action, $\mathbf{A}_h$ is its logarithmic generator, and c_h is the corresponding infinite-sample rate estimate. A hat marks an empirical quantity. We call a real projected matrix *admissible* if it has no eigenvalues on $(-\infty, 0]$, so that its real principal logarithm exists. An observation interval is admissible when its projected matrix is admissible.

Choose bases for $\mathcal{F}$ and $\mathcal{F}^\perp$ that are orthonormal in this measure. Together, these bases are adapted to $\mathcal{F} \oplus \mathcal{F}^\perp$. In the row-coefficient convention of EDMD, the full generator has the block form

$$(5.2) \quad \mathbf{A}_{\text{full}} = \begin{bmatrix} \mathbf{A}_{JJ} & \mathbf{A}_{JJ'} \\ \mathbf{A}_{J'J} & \mathbf{A}_{J'J'} \end{bmatrix}, \quad \mathbf{K}_h = (e^{h\mathbf{A}_{\text{full}}})_{JJ}.$$

Let J index the retained observables and J' their complement. The block $\mathbf{A}_{JJ'}$ contains the coefficients outside $\mathcal{F}$ that appear when the generator acts on a retained observable. The matrix $\mathbf{K}_h$ is the row-convention representation of $P e^{h\mathbf{A}}|_{\mathcal{F}}$, and the subscript “sp” selects the species rows used by $\mathcal{R}$. A change of retained basis is absorbed into $\mathcal{R}$, so it does not change the extracted rates.

Lemma 5.1 (Projected Koopman operator). *For fixed projection and $h \rightarrow 0$,*

$$(5.3) \quad \begin{aligned} \mathbf{K}_h &= I + h\mathbf{A}_{JJ} + \frac{h^2}{2}(\mathbf{A}_{JJ}^2 + \mathbf{A}_{JJ'}\mathbf{A}_{J'J}) \\ &+ \frac{h^3}{6}(\mathbf{A}_{JJ}^3 + \mathbf{A}_{JJ}\mathbf{A}_{JJ'}\mathbf{A}_{J'J} + \mathbf{A}_{JJ'}\mathbf{A}_{J'J}\mathbf{A}_{JJ} + \mathbf{A}_{JJ'}\mathbf{A}_{J'J'}\mathbf{A}_{J'J}) + O(h^4). \end{aligned}$$

Proof. We use block multiplication to obtain the required blocks of the second power:

$$\begin{aligned} (\mathbf{A}_{\text{full}}^2)_{JJ} &= \mathbf{A}_{JJ}^2 + \mathbf{A}_{JJ'}\mathbf{A}_{J'J}, \\ (\mathbf{A}_{\text{full}}^2)_{J'J} &= \mathbf{A}_{J'J}\mathbf{A}_{JJ} + \mathbf{A}_{J'J'}\mathbf{A}_{J'J}. \end{aligned}$$

We multiply once more to obtain

$$\begin{aligned} (\mathbf{A}_{\text{full}}^3)_{JJ} &= \mathbf{A}_{JJ}(\mathbf{A}_{\text{full}}^2)_{JJ} + \mathbf{A}_{JJ'}(\mathbf{A}_{\text{full}}^2)_{J'J} \\ &= \mathbf{A}_{JJ}^3 + \mathbf{A}_{JJ}\mathbf{A}_{JJ'}\mathbf{A}_{J'J} + \mathbf{A}_{JJ'}\mathbf{A}_{J'J}\mathbf{A}_{JJ} + \mathbf{A}_{JJ'}\mathbf{A}_{J'J'}\mathbf{A}_{J'J}. \end{aligned}$$

Using $e^{h\mathbf{A}_{\text{full}}} = I + h\mathbf{A}_{\text{full}} + h^2\mathbf{A}_{\text{full}}^2/2 + h^3\mathbf{A}_{\text{full}}^3/6 + O(h^4)$, its JJ block is

$$(e^{h\mathbf{A}_{\text{full}}})_{JJ} = I + h\mathbf{A}_{JJ} + \frac{h^2}{2}(\mathbf{A}_{JJ}^2 + \mathbf{A}_{JJ'}\mathbf{A}_{J'J}) + \frac{h^3}{6}(\mathbf{A}_{\text{full}}^3)_{JJ} + O(h^4).$$

We replace $(\mathbf{A}_{\text{full}}^3)_{JJ}$ by the expression above to obtain (5.3). ■

5.2. How projection changes the Koopman spectrum. We interpret the projection in (5.2) spectrally. If $\mathcal{F}$ is invariant under $\mathcal{A}$, then $\mathbf{A}_{JJ'} = 0$ and

$$(5.4) \quad \mathbf{K}_h = e^{h\mathbf{A}_{JJ}}, \quad \sigma(\mathbf{K}_h) = e^{h\sigma(\mathbf{A}_{JJ})}.$$

Thus the fitted Koopman eigenvalues are exponentials of generator eigenvalues. If $\sigma(h\mathbf{A}_{JJ})$ lies in the open strip $\{z : -\pi < \text{Im } z < \pi\}$, we use the principal logarithm to reverse this mapping and obtain $h^{-1} \log \mathbf{K}_h = \mathbf{A}_{JJ}$. Outside that strip, generator frequencies separated by $2\pi/h$ cannot be distinguished. These are the matrix-function form of the sampling and branch conditions for Koopman spectral inference [11, 2].

When the retained observable space is not invariant, paths through J' alter the spectrum. Let Γ be a contour in the common resolvent set that encloses the spectra of $\mathbf{A}_{\text{full}}$ and $\mathbf{A}_{J'J}$. We use block inversion and the Cauchy functional calculus to obtain

$$(5.5) \quad [(zI - \mathbf{A}_{\text{full}})^{-1}]_{JJ} = [zI - \mathbf{A}_{JJ} - \mathbf{A}_{JJ'}(zI - \mathbf{A}_{J'J'})^{-1}\mathbf{A}_{J'J}]^{-1},$$

$$(5.6) \quad \mathbf{K}_h = \frac{1}{2\pi i} \int_{\Gamma} e^{hz} [zI - \mathbf{A}_{JJ} - \mathbf{A}_{JJ'}(zI - \mathbf{A}_{J'J'})^{-1}\mathbf{A}_{J'J}]^{-1} dz.$$

Proposition 5.2 (Modal residues of the omitted space). Suppose $\mathbf{A}_{J'J'}$ is diagonalizable, with right and left eigenvectors $\mathbf{v}_\ell$ and $\mathbf{w}_\ell$ normalized by $\mathbf{w}_\ell^* \mathbf{v}_k = \delta_{\ell k}$, and eigenvalues μ_ℓ , where $*$ denotes conjugate transpose. Then

$$(5.7) \quad \mathbf{A}_{JJ'}(zI - \mathbf{A}_{J'J'})^{-1}\mathbf{A}_{J'J} = \sum_{\ell} \frac{\mathbf{C}_\ell}{z - \mu_\ell}, \quad \mathbf{C}_\ell = \mathbf{A}_{JJ'}\mathbf{v}_\ell\mathbf{w}_\ell^*\mathbf{A}_{J'J}.$$

Consequently, an omitted mode enters the projected Koopman operator through its decay rate μ_ℓ and coupling residue $\mathbf{C}_\ell$. If $\mathbf{C}_\ell = 0$, that mode is invisible to the retained finite-time operator.

Proof. We substitute the spectral resolution $(zI - \mathbf{A}_{J'J'})^{-1} = \sum_{\ell} \mathbf{v}_\ell \mathbf{w}_\ell^* / (z - \mu_\ell)$ into (5.5) to obtain (5.7). We obtain the final statement from (5.6). ■

The middle term in (5.5) represents a frequency-dependent return through the omitted observables. It shows how unresolved modes perturb the projected Koopman spectrum. When either coupling block vanishes, (5.4) is recovered. This Schur-complement resolvent is standard in invariant-subspace and matrix-function analysis [12, 11].

Equation (5.7) also connects the analysis with the stochastic Koopman approximation of Gupta and Khammash [10]. They estimate iterated resolvents to recover dominant decay modes of a specified reaction network. Here each pole μ_ℓ is paired with a residue $\mathbf{C}_\ell$ that measures how the mode leaves and re-enters the retained observable space. Their method uses dominant modes for prediction. We use the mode-residue pairs to explain bias in an inferred generator. Eigenvalues alone are insufficient because a slowly decaying mode with a small residue can have little effect on the recovered rates.

The same calculus separates the size of the projection defect from the conditioning of the logarithm. Suppose a contour Ξ surrounds the spectra of $\mathbf{K}_h$ and $e^{h\mathbf{A}_{JJ}}$ without crossing $(-\infty, 0]$, and the strip condition above holds. Then

$$(5.8) \quad \log \mathbf{K}_h - h\mathbf{A}_{JJ} = \frac{1}{2\pi i} \int_{\Xi} \log z (zI - \mathbf{K}_h)^{-1} (\mathbf{K}_h - e^{h\mathbf{A}_{JJ}})(zI - e^{h\mathbf{A}_{JJ}})^{-1} dz.$$

We take norms in (5.8) to bound the rate error by the projection defect, amplified by the two resolvents and by $1/h$ after applying $\mathcal{R}$. The contour can cease to exist when the projected spectrum reaches the principal-logarithm cut. The expansion below complements this global spectral statement by resolving the first small- h contribution that survives in the selected species rows.

5.3. Added observables and bias order. To remove these paths through a chosen order, we retain $\mathcal{A}x_s, \dots, \mathcal{A}^r x_s$ for each species used in rate extraction. Only these species rows must close; the full dictionary need not be invariant. The first surviving rate error is determined by one excursion from the retained space into its complement and back.

Definition 5.3 (Species closure depth). *The dictionary space $\mathcal{F}$ has closure depth r on a collection of species observables if*

$$(5.9) \quad \mathcal{A}^j x_s \in \mathcal{F}, \quad j = 1, \dots, r,$$

for every species used by the rate-extraction map.

Theorem 5.4 (Rate bias from finite-depth closure). *Let the state space be finite and closed, and let a full-support measure independent of h induce P . Suppose that $r \geq 1$ and*

$$\mathcal{A}^j x_s \in \mathcal{F}, \quad j = 1, \dots, r,$$

for every species used by the extraction map. If $h\|\mathbf{A}_{\text{full}}\| < \log 2$, then

$$(5.10) \quad c_h - c = -\frac{B_{r+1}}{(r+1)!} h^{r+1} \mathcal{R}(\mathbf{A}_{JJ}^r \mathbf{A}_{JJ'} \mathbf{A}_{J'J}) + O(h^{r+2}).$$

where we let B_n denote the Bernoulli numbers satisfying $z/(e^z - 1) = \sum_{n=0}^{\infty} B_n z^n / n!$. Consequently,

$$(5.11) \quad c_h - c = \begin{cases} O(h^{r+1}), & r \text{ odd}, \\ O(h^{r+2}), & r \text{ even}. \end{cases}$$

For odd r , the order is exactly $r+1$ if and only if $\mathcal{R}(\mathbf{A}_{JJ}^r \mathbf{A}_{JJ'} \mathbf{A}_{J'J}) \neq 0$.

Proof. The closure conditions imply

$$(5.12) \quad [\mathbf{A}_{JJ}^{j-1} \mathbf{A}_{JJ'}]_{\text{sp}} = 0, \quad j = 1, \dots, r.$$

Indeed, after $j-1$ generator applications the selected species rows have no complementary component; after one further application, we obtain the expression in (5.12). Any term that leaves the retained space must later return through $\mathbf{A}_{J'J}$. The first such word that can survive in a selected species row is therefore

$$(5.13) \quad [\mathbf{A}_{JJ}^r \mathbf{A}_{JJ'} \mathbf{A}_{J'J}]_{\text{sp}}.$$

Insert the exponential expansion of Lemma 5.1 into the ordinary power series for $\log \mathbf{K}_h$ and collect this word. At its first possible degree, terms containing $\mathbf{A}_{J'J}$, a retained factor

to the right of $\mathbf{A}_{J'J}$, or a second excursion cannot contribute. The remaining coefficients are generated by

$$(5.14) \quad \frac{1}{z} - \frac{1}{e^z - 1} = \frac{1}{2} - \frac{z}{12} + \frac{z^3}{720} - \dots$$

Hence the coefficient of the word in (5.13) is $-B_{r+1}/(r+1)!$. Dividing by h and applying $\mathcal{R}$ gives (5.10). Under the condition $h\|\mathbf{A}_{\text{full}}\| < \log 2$, the logarithm series converges and the remainder is uniform. Finally, $B_{r+1} = 0$ for positive even r , which proves (5.11). ■

Remark 5.5 (Closure as spectral decoupling). Under condition (5.12), every short block walk vanishes if it starts in an extracted species row and enters the omitted space during the first r generator applications. The full Koopman spectrum is unchanged, but the retained–omitted decomposition prevents the modal residues in Proposition 5.2 from reaching the extracted rows along these short paths. Theorem 5.4 gives the resulting delay in the rate bias.

Corollary 5.6 (Depth-one projection bias). *If $\mathcal{A}x_s \in \mathcal{F}$ for every species used by the extraction map, then*

$$(5.15) \quad c_h - c = -\frac{h^2}{12}\mathcal{R}(\mathbf{A}_{JJ}\mathbf{A}_{JJ'}\mathbf{A}_{J'J}) + O(h^3).$$

Proof. Let $r = 1$ in Theorem 5.4 and use $B_2 = 1/6$. ■

The specified reaction channels generate a space with the closure required by Definition 5.3. Let $\mathcal{F}_0$ be a base dictionary that contains the constant, the species, and the propensity functions needed for extraction, and let

$$(5.16) \quad \mathcal{F}_r = \mathcal{F}_0 + \text{span}\{\mathcal{A}_{k_j} \cdots \mathcal{A}_{k_1}x_s : 1 \leq j \leq r, 1 \leq k_1, \dots, k_j \leq M, s \text{ is used by } \mathcal{R}\}.$$

Zero functions and linear combinations already present are discarded when the dictionary is assembled.

Corollary 5.7 (Closure under reaction operators). *For every rate vector c , the space $\mathcal{F}_r$ has closure depth r on the selected species. Consequently, its rate bias has the order in (5.11).*

Proof. We use $\mathcal{A} = \sum_{k=1}^M c_k \mathcal{A}_k$ repeatedly. For each $j = 1, \dots, r$ and each selected species x_s , we expand

$$(5.17) \quad \mathcal{A}^j x_s = \sum_{k_1, \dots, k_j=1}^M c_{k_j} \cdots c_{k_1} \mathcal{A}_{k_j} \cdots \mathcal{A}_{k_1} x_s.$$

Every function on the right belongs to the span in (5.16). Hence $\mathcal{A}^j x_s \in \mathcal{F}_r$ for every $j \leq r$, and the space has closure depth r . We now apply Theorem 5.4 to obtain the stated bias order. ■

Remark 5.8 (Partial closure). The order theorem requires the complete space in (5.16). The Lotka–Volterra and Brusselator comparisons use this complete depth-two space. The Michaelis–Menten and MAPK comparisons retain selected directions from it. Under depth-one

closure, the complementary coefficients of $\mathcal{A}^2 x_s$ are $[\mathbf{A}_{JJ}\mathbf{A}_{JJ'}]_{\text{sp}}$. The leading bias therefore satisfies

$$\|\mathcal{R}(\mathbf{A}_{JJ}\mathbf{A}_{JJ'}\mathbf{A}_{J'J})\| \leq \|\mathcal{R}\| \|\mathbf{A}_{JJ}\mathbf{A}_{JJ'}\|_{\text{sp}} \|\mathbf{A}_{J'J}\|.$$

Adding directions from the residuals of $\mathcal{A}^2 x_s$ reduces the middle factor. A rate-independent partial-closure rule orders the empirical residual directions by singular value and retains the largest observed L^2 components first. The final factor and the extraction map also enter the bound, so partial closure need not reduce every rate error monotonically.

5.4. Constructing the added observables. We construct the spaces in (5.16) one depth at a time. At each depth, we apply every unit-rate reaction operator to the functions added at the preceding depth and discard zero functions and linear combinations already represented by the dictionary. Before this reduction, there are M^j ordered sequences of j reaction operators. Different sequences may produce the same polynomial or a function already represented by the dictionary, so several sequences can collapse to one coordinate.

For the reversible dimerization model used below, let channel 1 be $2X \rightarrow Y$ and channel 2 be $Y \rightarrow 2X$. Starting from $\{1, Y, \binom{X}{2}\}$, the functions $\mathcal{A}_1 Y$ and $\mathcal{A}_2 Y$ already lie in the base dictionary. At length two,

$$\mathcal{A}_1 \mathcal{A}_1 Y = \binom{X}{2}(3 - 2X), \quad \mathcal{A}_2 \mathcal{A}_1 Y = Y(2X + 1), \quad \mathcal{A}_1 \mathcal{A}_2 Y = -\binom{X}{2}, \quad \mathcal{A}_2 \mathcal{A}_2 Y = Y.$$

Using the conservation law $X + 2Y = 24$, the last three functions lie in the base dictionary. We add the component of $\binom{X}{2}(3 - 2X)$ that is independent of the existing functions. One new coordinate is enough to enforce depth-two closure in this example.

From Theorem 5.4, we obtain orders of at least 2, 4, 4, and 6 at depths one through four. The repeated fourth order has a different explanation at the two middle depths: the depth-two Bernoulli coefficient is zero, while the depth-three coefficient is $1/720$. The depth-three function can therefore reduce the fourth-order coefficient without raising the order. For polynomial libraries, each additional closure level raises the required degree by at most $\max_k \deg \phi_{\nu_k} - 1$. The Brusselator has a trimolecular channel, so a full polynomial dictionary that enforces depth two must include terms through degree five.

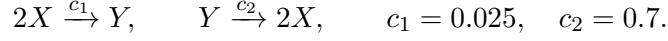
6. Numerical Validation. The numerical experiments separate the three errors that govern rate recovery. Exact finite-state matrices isolate projection bias and verify its predicted order. Simulated trajectories then show how sampling error, projection bias, and the logarithm branch condition interact in finite data.

6.1. Experimental setup. The dimerization model resolves finite-time bias by closure depth and omitted Koopman mode. A sweep over observation intervals for the Brusselator follows rate recovery to the principal-logarithm branch cut. The final comparison tests complete or selected reaction-generated closure on oscillatory, conserved, higher-order, and multi-tier networks.

The dimerization state space is small enough to enumerate. Lotka–Volterra supplies an oscillatory network; Michaelis–Menten tests conservation reduction and shared propensities; the Brusselator contains a third-order propensity; and a reduced MAPK cascade supplies a signaling network with a large conservation class.

For each network, we fix the species coefficients used to extract its rates and retain the constant, the physical scaling, and the full-rank EDMD fit as we change dictionaries. The Michaelis–Menten conservation class is finite. Lotka–Volterra and the Brusselator have unbounded natural state spaces and therefore test the method beyond the setting of the finite-state bias theorem.

6.2. Exact finite-state validation. Consider the reversible dimerization network



Using the conservation law $X + 2Y = 24$, we reduce the model to 13 states and enumerate them to form the exact Gram matrices from the full matrix exponential. We denote the initial probability of state i by ω_i and choose these probabilities proportional to $1, \dots, 13$ in increasing Y order. The dictionary $\psi = (1, Y, \binom{X}{2})^\top$ contains $\mathcal{A}Y = c_1 \binom{X}{2} - c_2 Y$. Repeated generator applications leave this dictionary. No simulated data enter the resulting bias calculation.

Figure 1 resolves the projection error by omitted Koopman mode. We diagonalize the complementary block in the weighted orthogonal decomposition used by the population EDMD calculation. At $h = 0.1$, the largest residue belongs to $\mu = -8.38 \pm 0.30i$. Decoupling this conjugate pair changes the inferred relative-rate vector by $4.84e-4$. The slower pair $-4.75 \pm 1.09i$ has a smaller residue and changes the vector by $3.07e-4$, while the modes near -10.8 and -11.8 have effects below 10^{-6} . Decay rate alone therefore does not rank the modes that matter for inference. Their coupling residues are also needed. The leave-one-mode-out changes are diagnostic rather than additive because the matrix logarithm couples all modes.

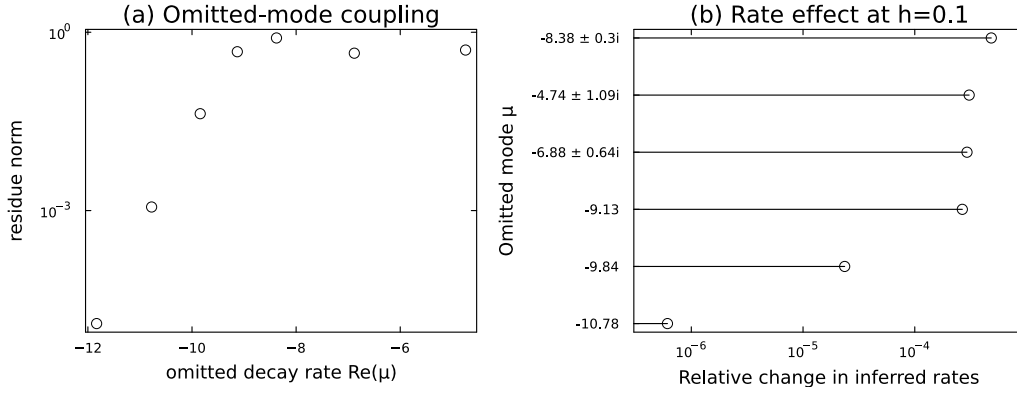


Figure 1. Omitted Koopman modes in the exact dimerization calculation. (a) Eigenvalues of the complementary generator block plotted against the Frobenius norm of their coupling residues $\mathbf{C}_\ell$. Conjugate pairs are grouped. (b) Change in the inferred relative-rate vector after decoupling one mode or conjugate pair at $h = 0.1$. This leave-one-mode-out quantity measures influence through the nonlinear logarithm and is not an additive decomposition of the total bias.

6.2.1. Finite-depth generator closure. We test the order theorem without truncation or Monte Carlo error in the finite dimerization model by adding $\mathcal{A}^j x_s$ to the dictionary for successive values of j . Starting with $\{1, Y, \binom{X}{2}\}$, we add the weighted-orthogonal residuals of $\mathcal{A}^2 Y$, $\mathcal{A}^3 Y$, and $\mathcal{A}^4 Y$ one at a time. We call the largest retained power the closure depth.

For this network, the same nested spaces are obtained by successively applying the reaction operators in Corollary 5.7 and removing linearly dependent functions.

Order verification. In Figure 2(a), we plot the infinite-sample bias across all four depths, for which the fitted slopes are 1.996, 3.994, 4.084, and 5.994. These values agree with the orders 2, 4, 4, and 6 in Theorem 5.4, so we conclude that its lower bounds are attained in this model. At depth three, we use the theorem to obtain the fourth-order coefficient $(1/720)\mathcal{R}(\mathbf{A}_{JJ}^3\mathbf{A}_{JJ'}\mathbf{A}_{JJ'})$. For the two rates, its values are $-3.531\text{e-}6$ and $1.053\text{e-}4$; by directly evaluating $(c_h - c)/h^4$ at $h = 5.00\text{e-}4$, we obtain $-3.544\text{e-}6$ and $1.056\text{e-}4$. The depth-three function therefore changes the fourth-order coefficient, and the depth-four Bernoulli coefficient vanishes.

Bias and sampling error. An enlarged dictionary has lower first-order mean squared error when its reduction in squared projection bias exceeds its increase in sampling variance. We compare closure depths using a common measure of physical rate error. For M rates, let $v_k(h)$ be the single-unit asymptotic variance coefficient for rate k obtained from the linearized estimator. With m independent sampling units, the first-order relative mean squared error is

$$(6.1) \quad \text{MSE}(h, m) \approx \frac{1}{M} \sum_{k=1}^M \left(\frac{c_{h,k} - c_k}{c_k} \right)^2 + \frac{1}{mM} \sum_{k=1}^M \frac{v_k(h)}{c_k^2}.$$

The first sum is the squared projection bias. The second is the sampling variance, with $v_k(h)$ predicted by the Gram-matrix perturbation calculation in (4.7) for one sampling unit and then divided by m . The sampling unit is an independent pair, an independent triple, or a whole trajectory, according to the estimator and the way observations are collected.

The short-interval variance following (4.7) explains why closure should be compared after choosing an interval for each dictionary. The following corollary separates this effect from the fixed-interval comparison below.

Corollary 6.1 (Optimized error rate). *Suppose that the aggregate relative bias is $\beta h^p + o(h^p)$ and the single-pair relative variance coefficient is $\nu/h + O(1)$, where $\beta, \nu > 0$. Then the leading mean squared error is minimized at*

$$(6.2) \quad h_*(m) = \left(\frac{\nu}{2p\beta^2 m} \right)^{1/(2p+1)},$$

and its minimum equals

$$(6.3) \quad (2p+1)\beta^2 \left(\frac{\nu}{2p\beta^2 m} \right)^{2p/(2p+1)}.$$

Thus the ordinary depth-one logarithm has the rate $m^{-4/5}$, while depth-two closure has the rate $m^{-8/9}$ when its fourth-order coefficient is nonzero.

Proof. The leading terms are $\beta^2 h^{2p} + \nu/(mh)$. We differentiate to obtain $2p\beta^2 h^{2p+1} = \nu/m$ and hence (6.2). At this value, the variance term is $2p$ times the squared-bias term; their sum is (6.3). ■

Optimized observation interval. In Figure 2(b), we minimize the exact first-order mean squared error over $2.00\text{e-}4 \leq h \leq 0.5$ separately for each depth and sample size. Depth two first improves on depth one at approximately $1.00\text{e}6$ pairs, because its smaller bias permits an interval near 0.3 while the depth-one optimum decreases with the sample size. At $1.00\text{e}10$ pairs, the optimized depth-two mean squared error is 26.0% of the optimized depth-one value. Depths three and four have larger sampling coefficients and do not improve on depth two over the plotted range.

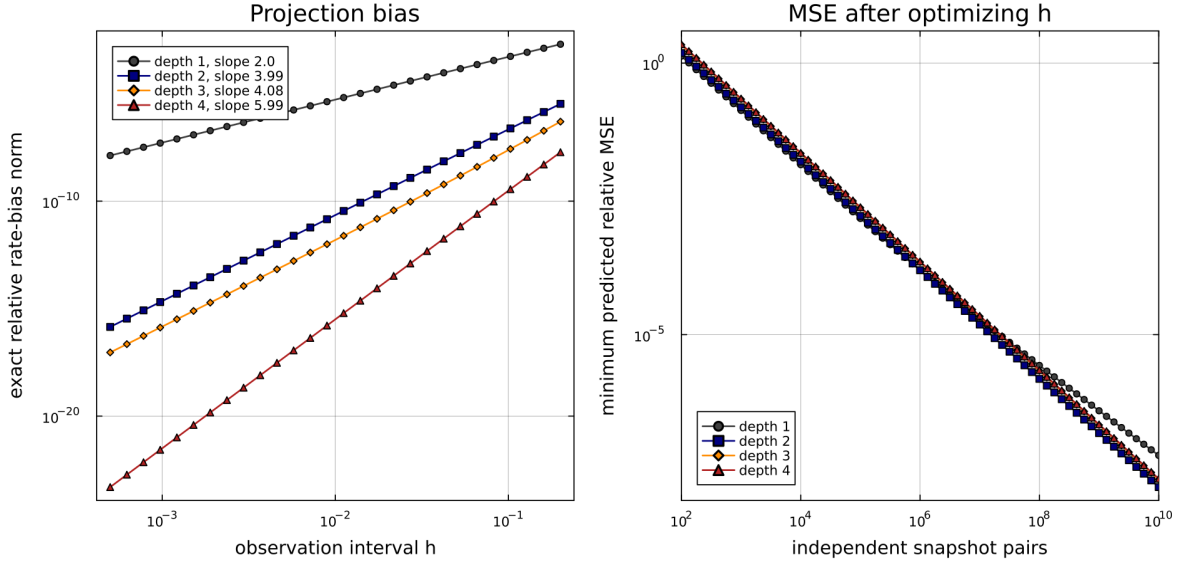
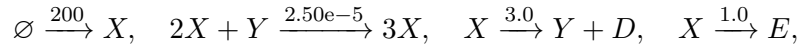


Figure 2. Finite-depth generator closure in the exactly enumerable dimerization model. (a) Exact relative rate bias, computed with 256-bit arithmetic. The dictionary at depth r contains $\mathcal{A}Y, \dots, \mathcal{A}^r Y$. (b) First-order relative mean squared error after minimizing over the observation interval for each depth and sample size; its bias contribution is evaluated from the exact projected operators.

6.3. Brusselator over a wide range of observation intervals. Exact enumeration separated projection bias from sampling error in the dimerization study. We now use simulated Brusselator trajectories to examine the practical choice of observation interval. The experiment begins in the accurate short-interval regime, passes through projection-dominated error, and ends where the fitted logarithm is no longer admissible.

The Brusselator [21] combines a constant source, autocatalysis, conversion, and degradation. It tests source-rate recovery and the separation of two reactions with the same linear X propensity. The volume-scaled stochastic model with $\Omega = 200$ is



which starts from $\mathbf{x}_0 = (200, 600)$ and runs to $T = 50$ with $\Delta t = 0.1$, while only (X, Y) are retained as dynamical species. The ten-dimensional mass-action basis retains the constant and contains all factorial monomials through degree three. Let $\phi(x, y) = x(x-1)y/2$ denote

the trimolecular propensity feature. The effective stochastic rate constants are $c_1 = 200$, $c_2 = 2.50\text{e-}5$, $c_3 = 3.0$, and $c_4 = 1.0$; any chemostatted factor is absorbed into c_3 . Applying Lemma 3.5, we obtain the species-row entries

$$\begin{aligned} (\mathbf{A})_{X,x} &= -(c_3 + c_4) = -4.0, & (\mathbf{A})_{X,\phi} &= c_2 = 2.50\text{e-}5, \\ (\mathbf{A})_{Y,x} &= c_3 = 3.0, & (\mathbf{A})_{Y,\phi} &= -c_2. \end{aligned}$$

We add the species actions to cancel both autocatalysis and conversion:

$$(6.4) \quad \mathcal{A}(X + Y) = c_1 - c_4 X.$$

Hence $c_4 = -(\mathbf{A})_{X,x} - (\mathbf{A})_{Y,x}$ is recovered from the coefficient of X . The controlled studies below use this relation together with the individual c_3 and autocatalysis coefficients.

The controlled calculation uses 1000 independent direct-SSA trajectories (seed 714) with the stated propensities and initial state. We retain the constant and all ten factorial features through degree three. At $h = 0.1$, the bootstrap standard deviations for c_3, c_4 are 0.0696, 0.0188, while the linearized predictions are 0.0697, 0.0187. The estimate $\hat{c}_3 = 4.524$ differs substantially from the true value 3. We vary h and use an independent reference ensemble to separate sampling variation from this systematic error.

The extended calculation covers 14 intervals from $5.00\text{e-}4$ to 6.4, with every fit using the same factorial dictionary and rate-extraction formulas. We retain the intercept and all nonzero singular values and require a real principal logarithm. All intervals share current times $0 : 0.1 : 49.6$ and 100 paired whole-trajectory bootstrap index sets. Simulation extends to time 56 to supply the longest intervals. The source and autocatalytic rates are now included, with autocatalysis extracted by averaging the two opposite species coefficients.

An independent ensemble of 10,000 trajectories (seed 2718), divided into twenty groups, provides the empirical reference. Variation among the groups gives linearized Monte Carlo standard errors. The reference therefore retains quantified Monte Carlo uncertainty, while every interval in the study uses the same trajectories and sampling mixture. Figure 3 reports all four rates together with the bootstrap and reference uncertainties.

In the figure, we distinguish three regimes. At short intervals, the bootstrap standard deviations for c_3 and c_4 fall from 0.614 and 0.278 at $h = 5.00\text{e-}4$ to 0.112 and 0.0558 at $h = 0.01$, while the order- h^2 projection bias remains below the sampling resolution. At intermediate intervals, bias becomes rate dependent: increasing h from 0.02 to 0.1 leaves c_4 near one but moves the conversion estimate from 3.177 to 4.443. From the independent reference, we obtain 4.413 ± 0.024 at $h = 0.1$, confirming that this discrepancy is systematic. At $h = 0.8, 1.6$, and 3.2, every bootstrap fit and the reference fit fail the real principal-logarithm check. Thus sampling error dominates at short intervals, projection bias dominates selected rates at intermediate intervals, and the fitted logarithm becomes inadmissible at long intervals.

Figure 4 gives the corresponding spectral view. Define the branch margin

$$(6.5) \quad \delta_h = \min_{x \leq 0} \sigma_{\min}(\mathbf{K}_h - xI),$$

This quantity measures distance to the negative real axis in a singular-value sense and therefore accounts for nonnormality. For the central fit, δ_h falls from $1.12\text{e-}2$ at $h = 0.1$ to $2.39\text{e-}4$ at

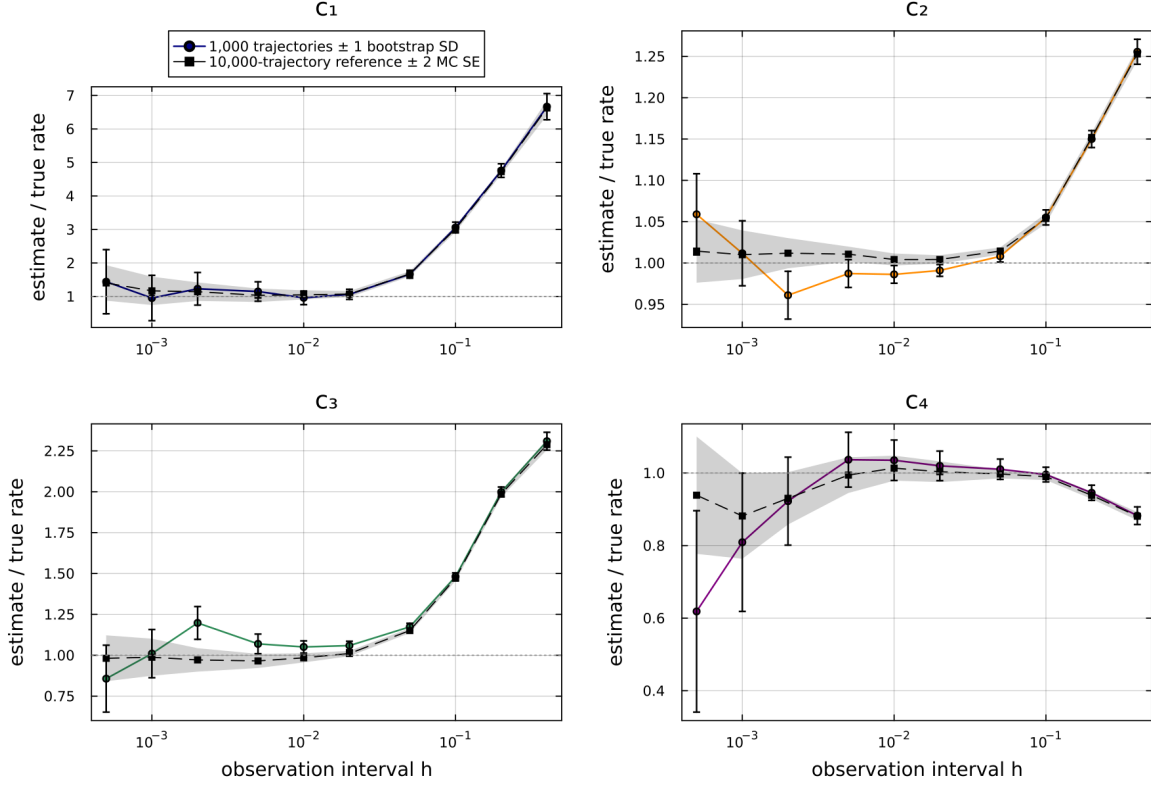


Figure 3. Rate recovery across observation intervals, normalized by each true rate. We plot one bootstrap standard deviation for 1000 trajectories as error bars and two Monte Carlo standard errors for the independent 10,000-trajectory reference as gray bands. The plotted range ends at $h = 0.4$, the largest tested interval for which the central fitted matrix has a real principal logarithm.

$h = 0.4$. It vanishes at $h = 0.8$ when an eigenvalue reaches the branch cut. The independent reference gives nearly identical values, so the loss of a real logarithm is reproducible across the two ensembles.

With the current sample sizes, the asymptotic bias slope remains unresolved at the shortest intervals.

6.4. Comparison across network classes. The dimerization and Brusselator studies isolate bias order, sampling cost, and interval selection. We next test a small number of added observables on four reaction networks with different structures. The additional functions do not depend on the unknown rates. For every channel pair (ℓ, k) , we evaluate $\phi_{\nu_\ell}(\mathbf{x})[\phi_{\nu_k}(\mathbf{x} + \boldsymbol{\zeta}_\ell) - \phi_{\nu_k}(\mathbf{x})]$. A function is discarded if it vanishes on all current observations. We subtract the empirical projection of each remaining function onto the base dictionary and compute an SVD of the residuals. Independent residuals are added in decreasing order of singular value, while the base coordinates remain unchanged. The resulting dictionaries are nested, beginning with the ordinary estimator and ending with full depth-two closure.

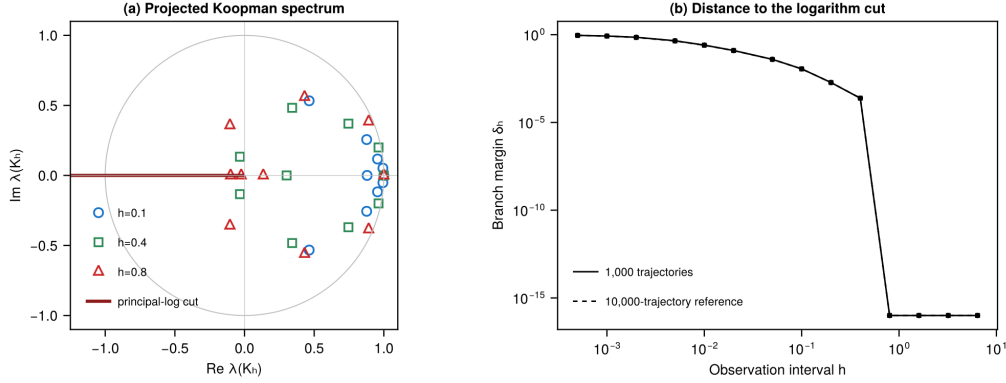


Figure 4. Spectral diagnostics for the projected Brusselator Koopman matrices. (a) Eigenvalues of the central 1000-trajectory fit at three observation intervals; the red segment is the principal-logarithm branch cut and the gray circle is the unit circle. (b) The branch margin (6.5) for the central and independent reference ensembles. Values plotted at machine scale are zero to working precision because an eigenvalue lies on the cut.

Separate exploratory ensembles choose the observation interval and the number of added functions. Among the admissible settings considered, we select the largest paired reduction in mean absolute relative rate error. This criterion uses one common interval for the base and closure estimators. We then fix the interval and number of functions before generating twenty new evaluation datasets for each network. The base and closure fits use identical observations within each dataset.

The largest model in this comparison is a reduced Huang–Ferrell MAPK cascade [13], with one phosphorylation at the upper tier and two at each of the lower tiers. After omitting explicit enzyme–substrate complexes, the model has ten mass-action channels: five phosphorylations catalyzed by the stimulus or the active species in the tier above, and five first-order dephosphorylations. The conserved tier totals are 40, 60, and 60, and twelve initial conditions vary the activity of each tier separately while including several intermediate activation states. After the reduction, the model has six independent dynamical coordinates and a 12-function base dictionary. At a fixed stimulus level, its conservation class still contains $41 \binom{62}{2}^2 = 146,611,121$ states.

Table 1 compares the two dictionaries on the same observations. Within each dataset, we first average the absolute relative errors over all rates. The error columns then report the median and interquartile range across twenty datasets. The condition numbers are the corresponding medians for the scaled empirical Gram matrices. Every projected matrix has a real principal logarithm. The Lotka–Volterra and Brusselator rows use overlapping intervals within each trajectory. Each Michaelis–Menten and MAPK pair comes from an independent short trajectory. We draw an initial condition and a current time from their fixed joint distribution, simulate to that time, and then continue once over the observation interval. Exploratory seeds 8101 and 9101–9501 select the intervals and added functions. The reported seeds 10101–12001, in increments of 100, are reserved for evaluation.

Figure 5 reports the paired comparison at the selected observation intervals and sample

Table 1

Independent comparison of observable sets using the same observations. Errors are percentages. The dimension column lists the base, selected, and complete depth-two dimensions. Added coordinates are orthogonalized against the scaled base dictionary. Their Gram eigenvalues lie within the extremal eigenvalues of the base Gram matrix, so the displayed condition numbers are unchanged. Intervals and selected dimensions were fixed on exploratory data before the twenty evaluation datasets were generated.

Network	h	Pairs	$d_0/d_{\text{sel}}/d_2$	Base error	Closure error	Condition numbers
Lotka–Volterra	0.2	100,000	6/7/7	0.936 (0.653, 1.05)	0.767 (0.378, 1.02)	$2.84\text{e}3 \rightarrow 2.84\text{e}3$
Michaelis–Menten	6.0	64,000	4/5/6	5.56 (5.22, 6.13)	2.79 (1.50, 4.19)	$8.48\text{e}2 \rightarrow 8.48\text{e}2$
Brusselator	0.04	248,500	10/11/11	15.9 (13.7, 17.9)	4.88 (2.61, 6.04)	$3.58\text{e}6 \rightarrow 3.58\text{e}6$
MAPK cascade	1.6	96,000	12/15/21	4.90 (3.92, 5.60)	2.90 (2.59, 3.67)	$2.46\text{e}4 \rightarrow 2.46\text{e}4$

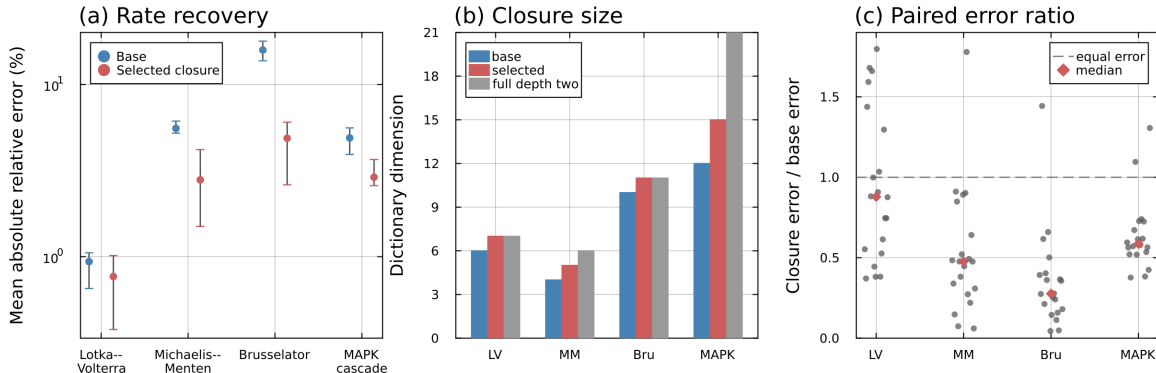


Figure 5. Independent comparison after adding selected reaction-generated observables. (a) Median mean absolute relative rate error across twenty datasets; bars span the interquartile range. (b) Base, selected, and complete depth-two dictionary dimensions. (c) Paired ratio of the closure error to the base error for each dataset; diamonds mark medians and the dashed line marks equal error. Each base and closure fit uses the same observations. All four networks use evaluation seeds that were excluded from interval and dimension selection.

sizes. The median errors decrease from 0.936% to 0.767% for Lotka–Volterra, 5.56% to 2.79% for Michaelis–Menten, 15.9% to 4.88% for the Brusselator, and 4.90% to 2.90% for MAPK. With closure, we lower the paired error in 13, 19, 19, and 17 of the twenty datasets, respectively. Exact two-sided paired Wilcoxon tests yield $p = 0.202, 1.68\text{e}^{-4}, 3.81\text{e}^{-6},$ and 1.34e^{-4} in the same order. Thus the Lotka–Volterra reduction is modest and unresolved at this sample size, while we observe clear paired reductions in the other three comparisons. We estimate one and three added functions for the Michaelis–Menten and MAPK calculations, respectively.

The finite Michaelis–Menten conservation class also permits a comparison without sampling error. At $h = 6$, the mean absolute relative error is 5.23% with the base dictionary and 0.484% with full depth-two closure. The infinite-sample calculation therefore agrees with the predicted bias reduction. The new trajectory data show how much of this reduction remains when one additional function must be estimated from a finite sample.

The MAPK conservation class contains 146,611,121 molecular states. We fit a 15×15 Koopman matrix whose observables retain the species and propensity coefficients needed to

identify all ten rates.

7. Discussion and Conclusions. The method recovers reaction rates from paired molecular states without constructing the full CME transition matrix. Mass-action structure places the rates in selected rows of the Koopman generator. EDMD and a matrix logarithm estimate these rows from finite-time data.

Generator paths that leave the retained observable space and return to the species rows describe the finite-time projection error. Closure through depth r represents these paths through the order stated in Theorem 5.4. The first remaining coefficient is $-B_{r+1}/(r+1)!$, which gives order h^{r+1} at odd depth and at least order h^{r+2} at even depth. The block-resolvent formula describes the same error spectrally. Each omitted mode contributes a pole μ_ℓ and a coupling residue C_ℓ . In the dimerization calculation, the residues, rather than the decay rates alone, rank the modes by their effect on the inferred rates. Closure suppresses their shortest paths into the extracted species rows, while the resolvents in (5.8) quantify the remaining amplification near the logarithm branch cut.

The finite dimerization calculation verifies these orders without state-space truncation or Monte Carlo error. It also shows why the observation interval and closure depth must be selected together. After optimizing the interval, depth-two closure first improves on the ordinary logarithm at about 1.00e6 pairs. At 1.00e10 pairs, its predicted mean squared error is 26.0% as large.

The Brusselator sweep reaches the complementary spectral boundary. The branch margin collapses at $h = 0.8$, where both the central fit and the independent reference lose a real principal logarithm. At shorter intervals, we use the same calculation to separate sampling-dominated and projection-dominated rate error.

Independent evaluation datasets test the added observables when they must be estimated from trajectories. Median errors decrease in all four networks, with clear paired reductions for Michaelis–Menten, the Brusselator, and MAPK. For Lotka–Volterra, the median gap is smaller than the variation among datasets, and the paired test is unresolved. This result agrees with the finite-sample error budget. Reducing projection bias gives only a modest gain when sampling variation has comparable size.

The observable formulation exchanges statistical efficiency for a calculation whose dimension equals the number of selected functions. In the 13-state model, the ordinary estimator has a standard deviation 6.05 times the Cramér–Rao limit attained asymptotically by the correctly specified full likelihood. In the MAPK calculation, a 15×15 observable matrix replaces a state-space matrix with about 1.47e8 rows and columns. These examples show the intended trade-off. The method avoids enumeration of the molecular state space while retaining a quantitative account of its bias and sampling error.

Future work. Extending the finite-state remainder bound to unbounded networks will require generator-domain conditions and moment estimates. Finite-data applications would also benefit from automatic selection or regularization of the generated candidate functions in place of the present exploratory selection. Such extensions should account explicitly for the covariance of overlapping intervals. Long observation intervals require a separate treatment of logarithm branches and admissibility, closely related to the classical embedding problem [15, 14, 4, 1]. We leave measurement error, discovery of missing reactions, and matrix-free

logarithm calculations to future work.

Code and data availability. The code and data used to reproduce every table and figure are available at https://github.com/AdityaDendukuri/cme_koopman.

REFERENCES

- [1] M. CASANELLAS, J. FERNÁNDEZ-SÁNCHEZ, AND J. ROCA-LACOSTENA, *The embedding problem for Markov matrices*, Publicacions Matemàtiques, 67 (2023), pp. 411–445, <https://doi.org/10.5565/PUBLMAT6712308>.
- [2] S. CHANDRASEKARAN, *Core matrix analysis*. Lecture notes, 2016. Version dated September 18, 2016.
- [3] D. T. CROMMELIN AND E. VANDEN-EIJNDEN, *Fitting timeseries by continuous-time Markov chains: A quadratic programming approach*, Journal of Computational Physics, 217 (2006), pp. 782–805, <https://doi.org/10.1016/j.jcp.2006.01.045>.
- [4] E. B. DAVIES, *Embeddable Markov matrices*, Electronic Journal of Probability, 15 (2010), pp. 1474–1486, <https://doi.org/10.1214/EJP.v15-733>.
- [5] Z. DRMAČ, I. MEZIĆ, AND R. MOHR, *Identification of nonlinear systems using the infinitesimal generator of the Koopman semigroup: A numerical implementation of the Mauroy–Goncalves method*, Mathematics, 9 (2021), p. 2075, <https://doi.org/10.3390/math9172075>.
- [6] S. ENGBLOM, *Computing the moments of high dimensional solutions of the master equation*, Applied Mathematics and Computation, 180 (2006), pp. 498–515, <https://doi.org/10.1016/j.amc.2005.12.032>.
- [7] S. N. ETHIER AND T. G. KURTZ, *Markov Processes: Characterization and Convergence*, Wiley, New York, 1986, <https://doi.org/10.1002/9780470316658>.
- [8] D. T. GILLESPIE, *Exact stochastic simulation of coupled chemical reactions*, The Journal of Physical Chemistry, 81 (1977), pp. 2340–2361, <https://doi.org/10.1021/j100540a008>.
- [9] D. T. GILLESPIE, *A rigorous derivation of the chemical master equation*, Physica A: Statistical Mechanics and its Applications, 188 (1992), pp. 404–425, [https://doi.org/10.1016/0378-4371\(92\)90283-V](https://doi.org/10.1016/0378-4371(92)90283-V).
- [10] A. GUPTA AND M. KHAMMASH, *A spectral Koopman approximation framework for stochastic reaction networks*, 2025, <https://arxiv.org/abs/2511.23114>.
- [11] N. J. HIGHAM, *Functions of Matrices: Theory and Computation*, SIAM, Philadelphia, 2008, <https://doi.org/10.1137/1.9780898717778>.
- [12] R. A. HORN AND C. R. JOHNSON, *Matrix Analysis*, Cambridge University Press, second ed., 2013, <https://doi.org/10.1017/CBO9781139020411>.
- [13] C.-Y. F. HUANG AND J. FERRELL, JAMES E., *Ultrasensitivity in the mitogen-activated protein kinase cascade*, Proceedings of the National Academy of Sciences of the United States of America, 93 (1996), pp. 10078–10083, <https://doi.org/10.1073/pnas.93.19.10078>.
- [14] R. B. ISRAEL, J. S. ROSENTHAL, AND J. Z. WEI, *Finding generators for Markov chains via empirical transition matrices, with applications to credit ratings*, Mathematical Finance, 11 (2001), pp. 245–265, <https://doi.org/10.1111/1467-9965.00114>.
- [15] J. F. C. KINGMAN, *The imbedding problem for finite Markov chains*, Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete, 1 (1962), pp. 14–24, <https://doi.org/10.1007/BF00531768>.
- [16] S. KLUS, F. NÜSKE, S. PEITZ, J.-H. NIEMANN, C. CLEMENTI, AND C. SCHÜTTE, *Data-driven approximation of the Koopman generator: Model reduction, system identification, and control*, Physica D: Nonlinear Phenomena, 406 (2020), p. 132416, <https://doi.org/10.1016/j.physd.2020.132416>.
- [17] A. MAUROY AND J. GONÇALVES, *Koopman-based lifting techniques for nonlinear systems identification*, IEEE Transactions on Automatic Control, 65 (2020), pp. 2550–2565, <https://doi.org/10.1109/TAC.2019.2941433>.
- [18] D. A. MCQUARRIE, *Stochastic approach to chemical kinetics*, Journal of Applied Probability, 4 (1967), pp. 413–478, <https://doi.org/10.2307/3212214>.
- [19] B. MUNSKY AND M. KHAMMASH, *The finite state projection algorithm for the solution of the chemical master equation*, The Journal of Chemical Physics, 124 (2006), p. 044104, <https://doi.org/10.1063/1.2145882>.
- [20] J.-H. NIEMANN, S. KLUS, AND C. SCHÜTTE, *Data-driven model reduction of agent-based systems using*

- the Koopman generator*, PLOS ONE, 16 (2021), p. e0250970, <https://doi.org/10.1371/journal.pone.0250970>.
- [21] I. PRIGOGINE AND R. LEFEVER, *Symmetry breaking instabilities in dissipative systems. II*, The Journal of Chemical Physics, 48 (1968), pp. 1695–1700, <https://doi.org/10.1063/1.1668896>.
 - [22] A. N. RISETH AND J. P. TAYLOR-KING, *Operator fitting for parameter estimation of stochastic differential equations*, 2018, <https://doi.org/10.48550/arXiv.1709.05153>, <https://arxiv.org/abs/1709.05153>.
 - [23] D. SCHNOERR, G. SANGUINETTI, AND R. GRIMA, *Approximation and inference methods for stochastic biochemical kinetics—a tutorial review*, Journal of Physics A: Mathematical and Theoretical, 50 (2017), p. 093001, <https://doi.org/10.1088/1751-8121/aa54d9>.
 - [24] R. SECHI, A. SIKORSKI, AND M. WEBER, *Estimation of the Koopman generator by Newton’s extrapolation*, Multiscale Modeling & Simulation, 19 (2021), pp. 758–774, <https://doi.org/10.1137/20M1333006>.
 - [25] M. O. WILLIAMS, I. G. KEVREKIDIS, AND C. W. ROWLEY, *A data-driven approximation of the Koopman operator: Extending dynamic mode decomposition*, Journal of Nonlinear Science, 25 (2015), pp. 1307–1346, <https://doi.org/10.1007/s00332-015-9258-5>.