

ON A POSITIVE-PRESERVING, ENERGY-STABLE NUMERICAL SCHEME TO MASS-ACTION KINETICS WITH DETAILED BALANCE*

CHUN LIU[†], CHENG WANG[‡], AND YIWEI WANG[§]

Abstract. In this paper, we provide a detailed theoretical analysis of the numerical scheme introduced in [C. Liu, C. Wang, and Y. Wang, J. Comput. Phys., 436:110253, 2021] for the reaction kinetics of a class of chemical reaction networks that satisfies detailed balance condition. In contrast to conventional numerical approximations, which are typically constructed based on ordinary differential equations (ODEs) for the concentrations of all involved species, the scheme is developed using the equations of reaction trajectories, which can be viewed as a generalized gradient flow of a physically relevant free energy. The unique solvability, positivity-preserving, and energy-stable properties are proved for the general case involving multiple reactions, under a mild condition on the stoichiometric matrix.

Keywords. structure-preserving scheme; reaction kinetics; energy stable; positivity-preserving; energetic variational approach.

AMS subject classifications. 37D35; 65L04; 65L20; 65K10; 80M30; 92C45.

1. Introduction

Chemical reactions are fundamental to many physical, chemical, and biological processes [1, 6, 8, 16]. Mathematically, the reaction kinetics are often described by a system of nonlinear ODEs in terms of concentrations of all involved species [8].

Consider a chemical reaction network (CRN) consisting of N species $X_1, X_2, \dots, X_N$ and M reversible chemical reactions:

$$\alpha_1^l X_1 + \alpha_2^l X_2 + \dots + \alpha_N^l X_N \rightleftharpoons \beta_1^l X_1 + \beta_2^l X_2 + \dots + \beta_N^l X_N, \quad l = 1, \dots, M, \quad (1.1)$$

where $\alpha_i^l, \beta_i^l \geq 0$ are stoichiometric coefficients for the l -th reaction. The reaction kinetics is often formulated as [7]

$$\frac{d\mathbf{c}(t)}{dt} = \mathbf{S}\mathbf{r}(\mathbf{c}). \quad (1.2)$$

Here, $\mathbf{c}(t) = (c_1, c_2, \dots, c_N)^T \in \mathbb{R}_+^N$ represents the concentrations of all involved species, $\mathbf{r}(\mathbf{c}) = (r_1(\mathbf{c}), r_2(\mathbf{c}), \dots, r_M(\mathbf{c})) \in \mathbb{R}^M$ denotes the reaction rates of the M reactions, and $\mathbf{S} \in \mathbb{R}^{N \times M}$ is the stoichiometric matrix, where each element S_{il} is defined as $\beta_i^l - \alpha_i^l$. It is often assumed that $N \geq M$ and $\text{rank}(\mathbf{S}) = M$ [7]. The latter assumption indicates that the M reactions are linearly independent in this reaction network. Under this assumption, we have

$$\frac{d}{dt}(\boldsymbol{\gamma} \cdot \mathbf{c}) = \boldsymbol{\gamma} \cdot (\mathbf{S}\mathbf{r}) = (\mathbf{S}^T \boldsymbol{\gamma}) \cdot \mathbf{c} = 0, \quad \boldsymbol{\gamma} \in \text{Ker } \mathbf{S}^T, \quad (1.3)$$

which indicates that the reaction kinetics (1.2) employs $N - \text{rank}(\mathbf{S})$ conserved quantities.

*Received: February 05, 2024; Accepted (in revised form): December 26, 2024. Communicated by Jie Shen.

[†]Department of Applied Mathematics, Illinois Institute of Technology, Chicago, IL 60616, USA (cliu124@iit.edu).

[‡]Mathematics Department, University of Massachusetts–Dartmouth, North Dartmouth, MA 02747, USA (cwang1@umassd.edu).

[§]Corresponding author. Department of Mathematics, University of California–Riverside, Riverside, CA 92521, USA (yiweiw@ucr.edu).

The reaction rate for the l -th reaction, $r_l(\mathbf{c})$, is often expressed as the difference between the forward and backward reaction rates, denoted as $r_l^+(\mathbf{c})$ and $r_l^-(\mathbf{c})$, i.e., $r_l(\mathbf{c}) = r_l^+(\mathbf{c}) - r_l^-(\mathbf{c})$. These rates $r_l^\pm(\mathbf{c})$ are commonly specified by the law of mass action (LMA) [8]. The empirical law states that the reaction rate is directly proportional to the product of the concentrations of the reactants, i.e.,

$$r_l^+(\mathbf{c}) = k_l^+ \mathbf{c}^{\boldsymbol{\alpha}^l}, \quad r_l^-(\mathbf{c}) = k_l^- \mathbf{c}^{\boldsymbol{\beta}^l}, \quad (1.4)$$

where $\mathbf{c}^{\boldsymbol{\alpha}^l} = \prod_{i=1}^N c_i^{\alpha_i^l}$, $\mathbf{c}^{\boldsymbol{\beta}^l} = \prod_{i=1}^N c_i^{\beta_i^l}$. Consequently, the reaction kinetic equation (1.2) is generally a highly nonlinear ODE system.

At a numerical level, solving the reaction kinetic equation (1.2) is often a challenge, mainly due to the stiffness and nonlinearity [7]. Moreover, many standard ODE solvers may fail to preserve the basic physical properties of the original system, such as the positivity of $\mathbf{c}$ and the intrinsic conservation laws. Although there has been a long history of developing robust numerical methods for reaction kinetics [2, 3, 7, 18] to preserve the positivity, as well as the conservation property, a significantly small step size is often needed for most existing methods.

It has been well-known that if the reaction kinetics (1.2) with LMA (1.4) satisfies the **detailed balance** condition, i.e., there exists a positive equilibrium point $\mathbf{c}^\infty \in \mathbb{R}_+^N$, such that

$$k_l^+(\mathbf{c}^\infty)^{\boldsymbol{\alpha}^l} = k_l^-(\mathbf{c}^\infty)^{\boldsymbol{\beta}^l}, \quad (1.5)$$

the reaction kinetics (1.2) admits a Lyapunov function or free energy [1, 14, 20], given by

$$\mathcal{F}[\mathbf{c}] = \sum_{i=1}^N c_i (\ln(c_i/c_i^\infty) - 1). \quad (1.6)$$

Under the detailed balance condition (1.5), it was shown in [19] that the system can be viewed as a generalized gradient flow of the reaction trajectory $\mathbf{R} \in \mathbb{R}^M$ [15, 19], which accounts for the “number” of forward chemical reactions that have occurred by time t , with respect to the free energy (1.6). More precisely, for the general reaction network (1.1), one can introduce a reaction trajectory $\mathbf{R}(t) \in \mathbb{R}^M$, and $\mathbf{c}(t)$ will be determined by the kinematics

$$\mathbf{c}(t) = \mathbf{c}(\mathbf{R}(t)) = \mathbf{c}_0 + \mathbf{S}\mathbf{R}(t), \quad (1.7)$$

where $\mathbf{S} = (S_{il}) \in \mathbb{R}^{N \times M}$ is the stoichiometric matrix and $\mathbf{c}_0$ is the initial concentration. Subsequently, the reaction kinetics with LMA (1.4) can be viewed as a generalized gradient flow of $\mathbf{R}$, satisfying the energy-dissipation law

$$\frac{d}{dt} \mathcal{F}[\mathbf{c}(\mathbf{R})] = -D(\mathbf{R}, \dot{\mathbf{R}}), \quad D(\mathbf{R}, \dot{\mathbf{R}}) = \sum_{l=1}^M \dot{R}_l \ln \left(\frac{\dot{R}_l}{k_l^-(\mathbf{c}(\mathbf{R}))^{\boldsymbol{\beta}^l}} + 1 \right), \quad (1.8)$$

where $D(\mathbf{R}, \dot{\mathbf{R}})$ is the rate of energy dissipation.

Indeed, by a standard variational procedure, one can show that $R_l(t)$ satisfies a nonlinear ODE

$$\ln \left(\frac{\dot{R}_l}{k_l^-(\mathbf{c}(\mathbf{R}))^{\boldsymbol{\beta}^l}} + 1 \right) = -\frac{\delta \mathcal{F}}{\delta R_l}, \quad \frac{\delta \mathcal{F}}{\delta R_l} = \sum_{i=1}^N S_{li} \mu_i, \quad l = 1, 2, \dots, M, \quad (1.9)$$

where $\mu_i = \frac{\delta \mathcal{F}}{\delta c_i}$ is the chemical potential of i -th species, $\frac{\delta \mathcal{F}}{\delta R_l}$ is known as the chemical affinity of l -th chemical reaction [9]. Using (1.5), one can rewrite (1.9) as

$$\dot{R}_l = k_l^+(\mathbf{c}(\mathbf{R}))\alpha^l - k_l^-(\mathbf{c}(\mathbf{R}))\beta^l, \quad (1.10)$$

which is the LMA. Reaction kinetics beyond the law of mass action can be obtained by choosing the dissipation in (1.8) differently. We refer the interested readers to [19] for more detailed discussions. It is worth mentioning that, unlike mechanical systems, $D(\mathbf{R}, \dot{\mathbf{R}})$ is no longer quadratic in terms of $\dot{\mathbf{R}}$ [19]. However, near chemical equilibrium, i.e., $\dot{R}_l \approx 0$, $\forall l$, we have $D(\mathbf{R}, \dot{\mathbf{R}}) \approx \sum_{l=1}^M |\dot{R}_l|^2 / (k_l^- \mathbf{c}^{\beta^l})$. Hence, the linear response assumption is still valid at the last stage of chemical reactions [5].

The variational formulation (1.8) indicates that the reaction kinetics with the detailed balance condition can be viewed as a generalized gradient flow of the reaction trajectory. As a consequence, most numerical techniques for an L^2 -gradient flow can be effectively applied to the reaction kinetics systems of this type. In [10], the authors proposed a numerical scheme that discretizes the reaction trajectory Equation (1.9) directly (see Section 2 for more details). The unique solvability, unconditional energy stability, and the positivity-preserving property are established for the case with $M=1$. The convergence analysis has been provided in [12], and an extension to the second-order numerical algorithm has been reported in [11].

Although the numerical tests in [10, 11] show that the proposed numerical schemes work for cases with $M > 1$, the theoretical analysis in [10, 11] is limited to the case of $M=1$. The purpose of this short note is to provide a theoretical justification for the proposed numerical scheme, in particular in terms of the positivity-preserving property, unique solvability, and unconditional energy stability for the multiple reaction case, with $M > 1$. To clarify the idea, we only write down the details for the case with $M=2$ and $N=4$, but the proof strategy works for the general case where $N \geq M$ and $\text{rank}(\mathbf{S}) = M$.

The remainder of this paper is organized as follows. The structure-preserving numerical scheme is recalled in Section 2. The theoretical justification of the positivity-preserving analysis and unique solvability is provided in Section 3.

2. The structure-preserving numerical discretization

In this section, we briefly review the numerical scheme for the reaction kinetics, proposed in [10]. Instead of solving the reaction kinetics equation for the concentrations of all involved species (1.2), the numerical discretization is constructed on the reaction trajectory Equation (1.9), which can be viewed as a generalized gradient flow of $\mathbf{R}$. Similar to an L^2 -gradient flow, a first-order semi-implicit discretization to (1.9) can be written as

$$\ln \left(\frac{R_l^{n+1} - R_l^n}{k_l^-(\mathbf{c}^n)\beta^l \Delta t} + 1 \right) = -\frac{\delta \mathcal{F}}{\delta R_l}(\mathbf{R}^{n+1}), \quad 1 \leq l \leq M, \quad (2.1)$$

where $\mathbf{c}^n = \mathbf{c}_0 + \mathbf{S}\mathbf{R}^n$ and Δt is the temporal step-size. Although this equation is non-linear with respect to R_l^{n+1} , its variational structure allows us to reformulate it as an optimization problem:

$$\mathbf{R}^{n+1} = \operatorname{argmin}_{\mathbf{R} \in \mathcal{V}^n} J^n(\mathbf{R}), \quad J^n(\mathbf{R}) = d_R^2(\mathbf{R}, \mathbf{R}^n) + \mathcal{F}[\mathbf{c}(\mathbf{R})]. \quad (2.2)$$

Here, $\mathbf{c}(\mathbf{R}) = \mathbf{c}_0 + \mathbf{S}\mathbf{R}$, $d_R^2(\mathbf{R}, \mathbf{R}^n)$ is a function measuring the difference between $\mathbf{R}$ and $\mathbf{R}^n$, defined as

$$d_R^2(\mathbf{R}, \mathbf{R}^n) = \sum_{l=1}^M \left((R_l - R_l^n + k_l^-(\mathbf{c}^n)\beta^l \Delta t) \ln \left(\frac{R_l - R_l^n}{k_l^-(\mathbf{c}^n)\beta^l \Delta t} + 1 \right) - (R_l - R_l^n) \right), \quad (2.3)$$

and the admissible set is given by

$$\mathcal{V}^n = \{\mathbf{R} \in \mathbb{R}^M \mid \mathbf{c}_0 + \mathbf{S}\mathbf{R} \in \mathbb{R}_+^N, \quad R_l - R_l^n + k_l^-(\mathbf{c}^n)^{\beta^l} \Delta t > 0, \quad J^n(\mathbf{R}^{n+1}) \leq J^n(\mathbf{R}^n)\}. \quad (2.4)$$

Of course, $\mathcal{V}^n$ is a non-empty set, since $\mathbf{R}^n \in \mathcal{V}^n$. Moreover, noticing that $d_R^2(\mathbf{R}, \mathbf{R}^n) \rightarrow \infty$ if $\|\mathbf{R}\| \rightarrow \infty$ and $\mathcal{F}[\mathbf{c}(\mathbf{R})]$ is bounded from below, we conclude that $\mathcal{V}^n$ is a bounded subset of $\mathbb{R}^M$. The set $\{\mathbf{R} \in \mathbb{R}^M \mid \mathbf{c}_0 + \mathbf{S}\mathbf{R} \in \mathbb{R}_+^N\}$ is called a stoichiometric compatibility class for the initial condition $\mathbf{c}_0$ [1]. It is straightforward to verify that

$$\frac{\delta J^n(\mathbf{R})}{\delta R_l} = \ln \left(\frac{R_l - R_l^n}{k_l^-(\mathbf{c}^n)^{\beta^l} \Delta t} + 1 \right) + \frac{\delta \mathcal{F}}{\delta R_l}, \quad \forall l. \quad (2.5)$$

Hence, a critical point of $J^n(\mathbf{R})$ in $\mathcal{V}^n$ gives a solution of the nonlinear Equation (2.1).

REMARK 2.1. It is worth mentioning that an explicit treatment of $\mathbf{R}$ in the term $k_l^-(\mathbf{c}(\mathbf{R}))^{\beta^l}$ turns out to be crucial, and it enables the definition of $d_R^2(\mathbf{R}, \mathbf{R}^n)$. Moreover, if $\frac{R_l - R_l^n}{k_l^-(\mathbf{c}^n)^{\beta^l} \Delta t}$ is small for any l , we observe the following Taylor expansion:

$$d_R^2(\mathbf{R}, \mathbf{R}^n) \approx \sum_{l=1}^m \frac{1}{k_l^-(\mathbf{c}^n)^{\beta^l} \Delta t} (R_l - R_l^n)^2 + \text{higher order terms}. \quad (2.6)$$

Therefore, the numerical scheme is a natural generalization for the minimizing movement scheme for an L^2 -gradient flow.

It is straightforward to prove the following unconditional energy stability result by using the property of $d_R^2(\mathbf{R}, \mathbf{R}^n)$.

PROPOSITION 2.1. *If $\mathbf{R}^{n+1}$ is a global minimizer of $J^n(\mathbf{R})$ in $\mathcal{V}^n$, then the numerical scheme is unconditionally energy stable.*

Proof. Define $f(x) = (x+a)\ln(x/a+1) - x$, where $a > 0$ is a given constant. It is clear that $f(x)$ is a monotonic increasing function of x for $x \geq 0$ and $f(0) = 0$. Consequently, $d_R^2(\mathbf{R}, \mathbf{R}^n) \geq 0$ in $\mathcal{V}^n$ and $d_R^2(\mathbf{R}, \mathbf{R}^n) = 0$ if and only if $\mathbf{R} = \mathbf{R}^n$. Hence, if $\mathbf{R}^{n+1}$ is a global minimizer of $J^n(\mathbf{R})$ in $\mathcal{V}^n$, we have

$$\mathcal{F}(\mathbf{R}^{n+1}) \leq J^n(\mathbf{R}^{n+1}) \leq J^n(\mathbf{R}^n) = \mathcal{F}(\mathbf{R}^n), \quad (2.7)$$

which gives the unconditional energy stability. $\square$

3. The positivity-preserving analysis and unique solvability

The main theoretical question associated with the numerical scheme (2.2) is the existence and uniqueness of the global minimizer of $J^n(\mathbf{R})$ in $\mathcal{V}^n$. This property has been proved in [10] for the case with $M = 1$. In this section, we demonstrate that the result can be generalized to the general case of $M \leq N$ and $\text{rank}(\mathbf{S}) = M$. More precisely, we have the following theorem.

THEOREM 3.1. *If $M \leq N$ and $\text{rank}(\mathbf{S}) = M$, then given $\mathbf{R}^n \in \mathbb{R}^M$, with $\mathbf{c}^n = \mathbf{c}_0 + \mathbf{S}\mathbf{R}^n \in \mathbb{R}_+^N$, there exists a unique solution $\mathbf{R}^{n+1} \in \mathcal{V}^n$ for the numerical scheme (2.1).*

To prove this result, we first observe the following lemma:

LEMMA 3.1. *If $M \leq N$ and $\text{rank}(\boldsymbol{\sigma}) = M$, $0 < c_i \leq A^*$, then $J^n(\mathbf{R})$ is a convex function of $\mathbf{R}$ in $\mathcal{V}^n$.*

Proof. Denote $g(\mathbf{R}) = d_R^2(\mathbf{R}, \mathbf{R}^n)$. A direct calculation implies that

$$\frac{\partial^2 g}{\partial R_l^2} = \frac{1}{R_l - R_l^- + k_l^-(c^n)^{\beta_l} \Delta t} > 0, \quad \frac{\partial^2 g}{\partial R_l \partial R_k} = 0 \text{ if } l \neq k, \quad \forall \mathbf{R} \in \mathcal{V}^n. \quad (3.1)$$

Hence, $g(\mathbf{R})$ is a convex function of $\mathbf{R}$ over $\mathcal{V}^n$. For $\mathcal{F}[c(\mathbf{R})]$, we recall that $\mathbf{c}(\mathbf{R}) = \mathbf{c}_0 + \mathbf{S}\mathbf{R}$, and a direct calculation gives

$$\nabla_{\mathbf{R}}^2 F(\mathbf{R}) = \mathbf{S}^T (\nabla_{\mathbf{c}}^2 F(\mathbf{c})) \mathbf{S}, \quad (3.2)$$

where $\nabla_{\mathbf{c}}^2 F = \text{diag}(\frac{1}{c_1}, \frac{1}{c_2}, \dots, \frac{1}{c_N})$. Because of the definition of $\mathcal{V}^n$, we have a uniform bound of c_i , i.e., $0 < c_i \leq A^*$, which results in

$$\lambda_{\min}(\nabla_{\mathbf{R}}^2 F) \geq \frac{1}{A^*} \lambda_{\min}(\mathbf{S}^T \mathbf{S}) > 0.$$

Henceforth, $F(\mathbf{R})$ is a convex function of $\mathbf{R}$ over $\mathcal{V}^n$. $\square$

Since $\mathcal{V}^n$ is a bounded set of $\mathbb{R}^M$, and $J^n(\mathbf{R})$ is a convex function of $\mathbf{R}$ in $\mathcal{V}^n$, then there exists a unique minimizer of $J^n(\mathbf{R})$ in $\mathcal{V}^n$. The key point of the proof is to show that the minimizer of $J^n(\mathbf{R})$ over $\mathcal{V}^n$ cannot occur on the boundary of $\mathcal{V}^n$, so that the global minimizer of $J^n(\mathbf{R})$ is a critical point of $J^n(\mathbf{R})$, which turns out to be a solution of (2.1).

To illustrate this idea, we present the case with $M=2$ and $N=4$. The analysis can be extended to different values of M and N following the same strategy. First, we define a linear transformation of R_i

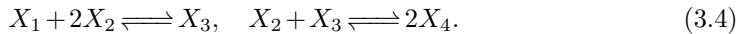
$$\tilde{R}_1 = c_1^0 + \sum_{j=1}^2 S_{1j} R_j, \quad \tilde{R}_2 = c_2^0 + \sum_{j=1}^2 S_{2j} R_j. \quad (3.3)$$

The positive stoichiometric compatibility class can be written in terms of $\tilde{R}_1$ and $\tilde{R}_2$, given by

$$\{(\tilde{R}_1, \tilde{R}_2) | \tilde{R}_i > 0, c_3^0 + \sum_{j=1}^2 \tilde{S}_{3j} \tilde{R}_j > 0, \quad c_4^0 + \sum_{j=1}^2 \tilde{S}_{4j} \tilde{R}_j > 0\},$$

where $\tilde{S}_{3j}$ and $\tilde{S}_{4j}$ are transformed stoichiometric coefficients in terms of $\tilde{R}_1$ and $\tilde{R}_2$.

EXAMPLE 3.1. We consider a concrete example of a reaction network



In turn, the stoichiometric matrix is given by

$$\mathbf{S} = \begin{pmatrix} -1 & 0 \\ -2 & 1 \\ 1 & -1 \\ 0 & 2 \end{pmatrix}. \quad (3.5)$$

Assume that $\mathbf{c}_0 = (1, 1, 1, 1)^T$, then the positive stoichiometric compatibility class corresponds to the set in the reaction space

$$\{(R_1, R_2) | 1 - R_1 > 0, 1 - 2R_1 + R_2 > 0, 1 + R_1 - R_2 \geq 0, 1 + 2R_2 > 0\}.$$

In this case, the linear transformation of R_i is defined as

$$\tilde{R}_1 = 1 - R_1, \quad \tilde{R}_2 = 1 - 2R_1 + R_2,$$

and the stoichiometric compatibility class becomes

$$\{(\tilde{R}_1, \tilde{R}_2) | \tilde{R}_i > 0, 3 - 3\tilde{R}_1 + \tilde{R}_2 > 0, \quad -1 + 4\tilde{R}_1 - 2\tilde{R}_2 > 0.\}$$

It is important to notice that the boundary of the stoichiometric compatibility class turns out to be $c_3 = 0$ and (or) $c_4 = 0$.

Without ambiguity, we omit the tilde notation in the following description. With a linear transformation, the kinematics can be rewritten as

$$c_1 = R_1, \quad c_2 = R_2, \quad c_3 = c_0^3 + S_{31}R_1 + S_{32}R_2, \quad c_4 = c_0^4 + S_{41}R_1 + S_{42}R_2, \quad (3.6)$$

and the free energy becomes

$$\mathcal{F}[R_1, R_2] = R_1 \ln \left(\frac{R_1}{c_1^\infty} - 1 \right) + R_2 \ln \left(\frac{R_2}{c_2^\infty} - 1 \right) + c_3 \ln \left(\frac{c_3}{c_3^\infty} - 1 \right) + c_4 \ln \left(\frac{c_4}{c_4^\infty} - 1 \right). \quad (3.7)$$

Denote $\gamma_l^n = R_l^n - k_l^-(c^n)\Delta t$. Since $R_l^n > 0$, it is clear that $\gamma_l^n \geq 0$ for Δt significantly small. Without loss of generality, we assume that $\gamma_l^n = 0$. In the case where $\gamma_l^n > 0$, we can adopt our approach to work on $R_l - \gamma_l^n$ instead. Moreover, to simplify the presentation, we take $\bar{c}_0^3 = \bar{c}_0^4 = 1$. Then the admissible set is given by

$$\mathcal{V}^n = \mathcal{V}_0^n \cap \{R | J^n(\mathbf{R}) \leq J^n(\mathbf{R}^n)\}, \quad (3.8)$$

where $\mathcal{V}_0^n = \{(R_1, R_2) \mid R_1 > 0, R_2 > 0, 1 + S_{31}R_1 + S_{32}R_2 > 0, c_1 + S_{41}R_1 + S_{42}R_2 > 0\}$. Figure 3.1(a)-(i) displays the possible geometry of the set $\mathcal{V}_0^n$.

It is important to note that the set $\mathcal{V}_0^n$ may not necessarily be bounded. Hence, it is crucial to consider $\mathcal{V}_0^n \cap \{R | J^n(\mathbf{R}) \leq J^n(\mathbf{R}^n)\}$. The boundedness of $\mathbf{R}$ comes from the condition $J^n(\mathbf{R}) \leq J^n(\mathbf{R}^n)$. Due to this bound, we have $0 < c_i(\mathbf{R}) < A^*$, $\forall \mathbf{R} \in \mathcal{V}^n$ for some constant A^* .

To show that the global minimizer of $J^n(R_1, R_2)$ over $\mathcal{V}^n$ cannot be obtained on the boundary, we only need to consider the following possible boundaries

$$\begin{aligned} \Gamma_1 &= \{(R_1, R_2) | R_1 = 0\}, & \Gamma_2 &= \{(R_1, R_2) | R_2 = 0\}, \\ \Gamma_3 &= \{(R_1, R_2) | c_3(R_1, R_2) = 0\}, & \Gamma_4 &= \{(R_1, R_2) | c_4(R_1, R_2) = 0\}. \end{aligned} \quad (3.9)$$

To this end, the following subset of $\mathcal{V}^n$ is taken into consideration:

$$\mathcal{V}_\delta^n = \{(R_1, R_2) \in \mathcal{V}^n \mid R_1, R_2 \geq g(\delta), c_3, c_4 \geq \delta\} \subset \mathcal{V}^n. \quad (3.10)$$

Let

$$\begin{aligned} \Gamma_1^\delta &= \{(R_1, R_2) | R_1 = g(\delta)\}, & \Gamma_2^\delta &= \{(R_1, R_2) | R_2 = g(\delta)\}, \\ \Gamma_3^\delta &= \{(R_1, R_2) | c_3(R_1, R_2) = \delta\}, & \Gamma_4^\delta &= \{(R_1, R_2) | c_4(R_1, R_2) = \delta\}, \end{aligned} \quad (3.11)$$

where $g(\delta)$ is a certain function that will be specified later. We only need to prove that the minimizer of J^n over $\mathcal{V}_\delta^n$ could not occur on $\Gamma_i^\delta \cap \mathcal{V}^n$ ($i = 1, \dots, 4$), if δ is taken significantly small. The strategy is to first assume that the minimizer of $J^n(R_1, R_2)$ over $\mathcal{V}_\delta^n$ occurs at a boundary point $(R_1^*, R_2^*) \in \Gamma_i^\delta$ for some i . In turn, if one can find

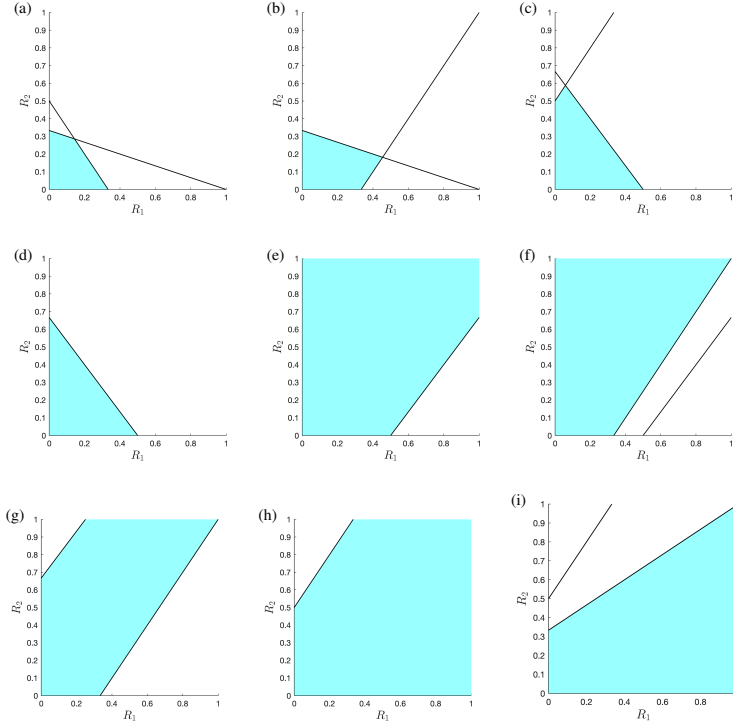


FIG. 3.1. Illustration of the transformed stoichiometric compatibility set $\mathcal{V}_0^n = \{(R_1, R_2) \mid R_1 > 0, R_2 > 0, c_3(R_1, R_2) > 0, c_4(R_1, R_2) > 0\}$ according to the signs of $(S_{31}, S_{32}, S_{41}, S_{42})$, (a) $(-, -, -, -)$; (b) $(-, +, -, -)$ or $(-, -, -, +)$; (c) $(+, -, -, -)$ or $(-, -, +, -)$; (d) $(+, +, -, -)$ or $(-, -, +, +)$; (e) $(+, +, -, +)$ or $(-, +, +, +)$; (f) $(-, +, -, +)$; (g) $(-, +, +, -)$ or $(+, -, -, +)$; (h) $(+, -, +, +)$ or $(+, +, +, -)$; (i) $(+, -, +, -)$. The case $(+, +, +, +)$ is not shown.

$(R'_1, R'_2) \in (\mathcal{V}_\delta^n)^\circ$ that $J^n(R'_1, R'_2) < J^n(R_1^*, R_2^*)$, then it leads to a contradiction. Such a strategy follows similar ideas as the positivity-preserving analysis reported in [4, 10]. At the beginning, we calculate the partial derivatives of $J^n(R_1, R_2)$ with respect to R_1 and R_2 . The derivatives are given by

$$\begin{aligned} \frac{\partial J^n}{\partial R_1} &= \ln \left(\frac{R_1 - R_1^n}{k_1^- (\mathbf{c}^n)^{\beta_1} \Delta t} + 1 \right) + \ln \left(\frac{R_1}{c_1^\infty} \right) + S_{31} \ln \left(\frac{c_3(\mathbf{R})}{c_3^\infty} \right) + S_{41} \ln \left(\frac{c_4(\mathbf{R})}{c_4^\infty} \right), \\ \frac{\partial J^n}{\partial R_2} &= \ln \left(\frac{R_2 - R_2^n}{k_2^- (\mathbf{c}^n)^{\beta_2} \Delta t} + 1 \right) + \ln \left(\frac{R_2}{c_2^\infty} \right) + S_{32} \ln \left(\frac{c_3(\mathbf{R})}{c_3^\infty} \right) + S_{42} \ln \left(\frac{c_4(\mathbf{R})}{c_4^\infty} \right). \end{aligned} \quad (3.12)$$

We will use these derivatives extensively in the subsequent analysis.

It is noticed that $\Gamma_1^\delta \cap \mathcal{V}^n$ and $\Gamma_2^\delta \cap \mathcal{V}^n$ are always two boundary sections of $\mathcal{V}_\delta^n$. We first consider the boundaries $\Gamma_1^\delta \cap \mathcal{V}^n$ and $\Gamma_2^\delta \cap \mathcal{V}^n$, by assuming the minimizer occurs at $R_1^* = g(\delta)$ or $R_2^* = g(\delta)$. Because of the symmetry, we only need to consider the case that $R_1^* = g(\delta)$, which in turn indicates that

$$\frac{\partial J^n}{\partial R_1} \Big|_{(R_1^*, R_2^*)} = \ln(g(\delta)) + \ln(g(\delta)) + S_{31} \ln(c_3^*) + S_{41} \ln(c_4^*) + Q_1, \quad (3.13)$$

where $Q_1 = -\ln(k_1^- (\mathbf{c}^n)^{\beta_1} \Delta t) - \ln c_1^\infty - S_{31} \ln c_3^\infty - S_{41} \ln c_4^\infty$ is a constant. Recall that

$\delta \leq c_3^* \leq A^*$ and $\delta \leq c_4^* \leq A^*$, and we always have

$$S_{31} \ln(c_3^*) + S_{41} \ln(c_4^*) \leq -|S_{31}| \ln \delta - |S_{41}| \ln \delta,$$

for some significantly small δ . One can always choose $g(\delta) = \delta^\alpha$ for some positive α such that $\frac{\partial J^n}{\partial R_1}|_{(R_1^*, R_2^*)} < 0$ with δ being significantly small. Then we can find $R'_1 > R_1^* = g(\delta)$ such that $J^n(R'_1, R_2^*) < J^n(R_1^*, R_2^*)$. Because of the fact that $(R'_1, R_2^*) \in \mathcal{V}_\delta$, this contradicts the assumption that (R_1^*, R_2^*) is a minimizer.

Next, we look at the possible boundary sections $\Gamma_3^\delta \cap \mathcal{V}^n$ and $\Gamma_4^\delta \cap \mathcal{V}^n$. The following different cases have to be discussed separately.

Case 1. $S_{31} < 0$, $S_{32} < 0$, $S_{41} < 0$, $S_{42} < 0$: In this case, the admissible set is sketched in Figure 3.1(a), and V_δ^n is the closed bounded set. We first assume that the minimizer occurs on $\Gamma_3^\delta \cap \mathcal{V}^n$. Since $S_{31} < 0$, $S_{32} < 0$, we see that either $|S_{31}|R_1^* \geq \frac{1}{3}$ or $|S_{32}|R_2^* \geq \frac{1}{3}$, if δ is significantly small. Without loss of generality, it is assumed that $|S_{31}|R_1^* \geq \frac{1}{3}$, so that $R_1^* \geq \frac{1}{-3S_{31}} := B_1^*$. Also notice that

$$\begin{aligned} \frac{\partial J^n}{\partial R_1}|_{(R_1^*, R_2^*)} &= \ln \left(\frac{R_1 - R_1^n}{k_1^-(c^n)^{\beta_1} \Delta t} + 1 \right) + \ln \left(\frac{R_1}{c_1^\infty} \right) + S_{31} \ln \left(\frac{\delta}{c_3^\infty} \right) + S_{41} \ln \left(\frac{c_4^*}{c_4^\infty} \right) \\ &= \ln R_1 + \ln R_1 + S_{31} \ln(\delta) + S_{41} \ln(c_4^*) + Q_1 \\ &\geq \ln B_1^* + \ln B_1^* + S_{31} \ln(\delta) + S_{41} \ln A^* + Q_1, \end{aligned} \quad (3.14)$$

where $Q_1 = -\ln(k_1^-(c^n)^{\beta_1} \Delta t) - \ln c_1^\infty - S_{31} \ln c_3^\infty - S_{41} \ln c_4^\infty$ is a constant, and A^* is the upper bound of c_4 in $\mathcal{V}^n \cap \{\mathbf{R} \mid J^n(\mathbf{R}) \leq J^n(\mathbf{R}^n)\}$. Since B_1^* , A^* and Q_1 are constants that are independent of Δt and δ , we are able to choose δ significantly small such that $\frac{\partial J^n}{\partial R_1}|_{(R_1^*, R_2^*)} > 0$. In other words, one can find $\delta < R'_1 < R_1$ such that $J(R'_1, R_2^*) \leq J(R_1^*, R_2^*)$. The fact that $c_3(R'_1, R_2^*) > c_3^* = \delta \in \mathcal{V}_\delta^n$ leads to a contradiction that (R_1^*, R_2^*) is a minimizer in $\mathcal{V}_\delta$. Using a similar argument, we are able to prove that the minimizer cannot occur at $c_4^* = \delta$, either.

Case 2. $S_{31} < 0$, $S_{32} < 0$, $S_{41} < 0$, $S_{42} > 0$, which corresponds to Figure 3.1(b): We first consider the boundary $\Gamma_3^\delta \cap \mathcal{V}^n$. On this boundary section, we see that either $R_1^* \geq \frac{1}{-3S_{31}} = B_1^*$ or $R_2^* \geq \frac{1}{-3S_{32}} = B_2^*$. In addition, denote $B_3^* = \min(B_1^*, -1/S_{41})$. If $R_1^* \geq B_3^*$, using similar arguments in the previous case, we have

$$\begin{aligned} \frac{\partial J^n}{\partial R_1}|_{(R_1^*, R_2^*)} &= \ln R_1 + \ln R_1 + S_{31} \ln(\delta) + S_{41} \ln(c_4^*) + Q_1 \\ &\geq \ln B_3^* + \ln B_3^* + S_{31} \ln(\delta) + S_{41} \ln A^* + Q_1. \end{aligned} \quad (3.15)$$

In turn, δ can be chosen significantly small, so that $\frac{\partial J^n}{\partial R_1}|_{(R_1^*, R_2^*)} > 0$. This leads to a contradiction. If $R_1^* \leq B_3^* \leq B_1^*$, we get $R_2^* \geq B_2^*$, and notice that

$$c_4^* = 1 + S_{41} R_1^* + S_{42} R_2^* \geq 1 + S_{41} B_3^* + S_{42} B_2^* \geq S_{42} B_2^*, \quad (3.16)$$

and

$$\begin{aligned} \frac{\partial J^n}{\partial R_2}|_{(R_1^*, R_2^*)} &= \ln(R_2^*) + \ln(R_2^*) + S_{41} \ln(\delta) + S_{42} \ln(c_4^*) + Q_2 \\ &\geq \ln B_2^* + \ln B_2^* + S_{31} \ln \delta + S_{42} \ln(S_{42} B_2^*) + Q_2. \end{aligned} \quad (3.17)$$

Again, since other terms are constants, we can choose δ significantly small, such that $\frac{\partial J^n}{\partial R_2}|_{(R_1^*, R_2^*)} > 0$. Therefore, one can find $R'_2 < R_2^*$, such that $J^n(R_1^*, R'_2) < J^n(R_1^*, R_2^*)$, which leads to a contradiction as $c_3(R_1^*, R'_2) \in \mathcal{V}_\delta$.

Next, we consider the case of $c_4^* = \delta$ (and $R_2^* > \delta$). Notice that, by choosing δ significantly small, we have

$$S_{41}R_1^* = \delta - 1 - S_{42}R_2^* \leq \delta - 1, \Rightarrow R_1^* \geq \frac{-1 + \delta}{S_{41}}. \quad (3.18)$$

By choosing δ significantly small, we get $R_1^* \geq -\frac{1}{2S_{41}} = B_4^*$. Therefore, the following inequality is valid:

$$\frac{\partial J^n}{\partial R_1} \big|_{(R_1^*, R_2^*)} \geq \ln B_4^* + \ln B_4^* + S_{31} \ln A^* + S_{32} \ln \delta + Q_1, \quad (3.19)$$

so that δ could be chosen significantly small satisfying $\frac{\partial J^n}{\partial R_1} \big|_{(R_1^*, R_2^*)} > 0$. Combining all these arguments, we conclude that a minimization point cannot occur at either $c_3^* = \delta$ or $c_4^* = \delta$, provided that δ is sufficiently small, in the case of $S_{31} < 0$, $S_{32} < 0$, $S_{41} < 0$, $S_{42} > 0$.

Due to the symmetry, the following cases (shown in Figure 3.1(c)) could be analyzed in a similar manner:

- $S_{31} < 0, S_{32} > 0, S_{41} < 0, S_{42} < 0$
- $S_{31} > 0, S_{32} < 0, S_{41} < 0, S_{42} < 0$
- $S_{31} < 0, S_{32} < 0, S_{41} > 0, S_{42} < 0$

Case 3. $S_{31} < 0$, $S_{32} > 0$, $S_{41} < 0$, $S_{42} > 0$, which corresponds to Figure 3.1(f): If a minimization point occurs at (R_1^*, R_2^*) with $c_4^* = (1 + S_{41}R_1^* + S_{42}R_2^*) = \delta$, we see that $R_1^* \geq \frac{-1}{S_{41}} := B_4^*$ (since $S_{42} > 0$). In turn, the following estimate could be derived:

$$\frac{\partial J^n}{\partial R_1} \geq \ln B_4^* + \ln B_4^* + S_{31} \ln A^* + S_{32} \ln \delta + Q_1. \quad (3.20)$$

Again, the value of $\ln B_4^* + \ln B_4^* + S_{31} \ln A^* + S_{32} \ln \delta$ becomes a fixed constant with a fixed Δt , and we could always choose δ significantly small such that $\frac{\partial J^n}{\partial R_1} \big|_{(R_1^*, R_2^*)} > 0$, which makes a contradiction to the assumption that $J(R_1, R_2)$ reaches a minimization point at (R_1^*, R_2^*) over V_δ . Using similar arguments, a minimization point cannot occur at (R_1^*, R_2^*) with $c_3^* = 1 + S_{31}R_1^* + S_{32}R_2^* = \delta$, either, in the case of $S_{31} < 0$, $S_{32} > 0$, $S_{41} < 0$, $S_{42} > 0$, if δ is sufficiently small. Because of the symmetry, the case of $S_{31} > 0$, $S_{32} < 0$, $S_{41} > 0$, $S_{42} < 0$, as shown in Figure 3.1(i), could be analyzed in a similar style (by switching R_1 and R_2).

Case 4. $S_{31} < 0$, $S_{32} > 0$, $S_{41} > 0$, $S_{42} < 0$, which corresponds to Figure 3.1(g): If a minimization point occurs at (R_1^*, R_2^*) with $c_3^* = (1 + S_{31}R_1^* + S_{32}R_2^*) = \delta$, we see that $R_1^* \geq \frac{-1}{S_{31}} := B_5^*$ (since $S_{31} > 0$, $S_{32} < 0$). This in turn indicates that

$$\frac{\partial J^n}{\partial R_1} \big|_{(R_1^*, R_2^*)} \geq \ln B_5^* + \ln B_5^* + S_{31} \ln A^* + S_{32} \ln \delta + Q_1. \quad (3.21)$$

Again, the uniform bound $c_4^* \leq A^*$ has been applied in the derivation. We could always choose δ significantly small so that $\frac{\partial J^n}{\partial R_1} \big|_{(R_1^*, R_2^*)} > 0$, which makes a contradiction to the assumption that $J(R_1, R_2)$ reaches a minimization point at (R_1^*, R_2^*) over V_δ . Using similar arguments, a minimization point cannot occur at (R_1^*, R_2^*) with $c_4^* = 1 + S_{41}R_1^* + S_{42}R_2^* = \delta$, either, due to the fact that R_2^* is bounded from below. Due to the symmetry, the case of $S_{31} > 0$, $S_{32} < 0$, $S_{41} < 0$, $S_{42} > 0$, could be analyzed in a similar fashion.

Case 5. $S_{31} > 0$, $S_{32} > 0$: In this case, the boundary section $c_3^* = \alpha_3(1 + S_{31}R_1^* + S_{32}R_2^*) = \delta$ will never be reached, because of the fact that $R_1^* > 0$, $R_2^* > 0$. In turn, the

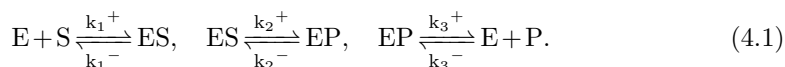
four boundary section constraints will be reduced to the three-boundary-section version, and the analysis in the previous cases could be recalled.

Case 6. $S_{41} > 0, S_{42} > 0$: Similarly, the boundary section $c_4^* = 1 + S_{41}R_1^* + S_{42}R_2^* = \delta$ will never be reached in this case, since $R_1^* > 0, R_2^* > 0$. Similarly, the four boundary section constraints will be reduced to the three-boundary-section version, and the analysis in the previous cases could be recalled.

Therefore, a combination of all these cases has demonstrated that the minimizer of $J(R_1, R_2)$ could not occur at a boundary point of V_δ where either $c_3 = \delta$ or $c_4 = \delta$, which completes the proof.

4. Numerical experiments

In this section, we provide numerical evidence to validate the proposed numerical scheme. We consider a generalized Michaelis-Menten equation that is widely used to model enzyme kinetics [8, 13, 17]. The corresponding reaction network is given by



Here, E is the enzyme that catalyzes the reaction $S \rightleftharpoons P$, SE and SP are two intermediates. It is often assumed that $k_2^- \ll k_2^+$ and $k_3^- \ll k_3^+$, so that most of S will be converted to E. This is a reaction network with 5 species and 3 reactions. Let c_i represent the concentration of species E, S, ES, EP, and P respectively, the generalized Michaelis-Menten equation can be written as

$$\begin{cases} \frac{dc_1}{dt} = -k_1^+ c_1 c_2 + k_1^- c_3 + k_3^+ c_4 - k_3^- c_1 c_5 \\ \frac{dc_2}{dt} = -k_1^+ c_1 c_2 + k_1^- c_3 \\ \frac{dc_3}{dt} = k_1^+ c_1 c_2 - k_1^- c_3 - k_2^+ c_3 + k_2^- c_4 \\ \frac{dc_4}{dt} = k_2^+ c_3 - k_2^- c_4 - k_3^+ c_4 + k_3^- c_1 c_5 \\ \frac{dc_5}{dt} = k_3^+ c_4 - k_3^- c_1 c_5 \end{cases}$$

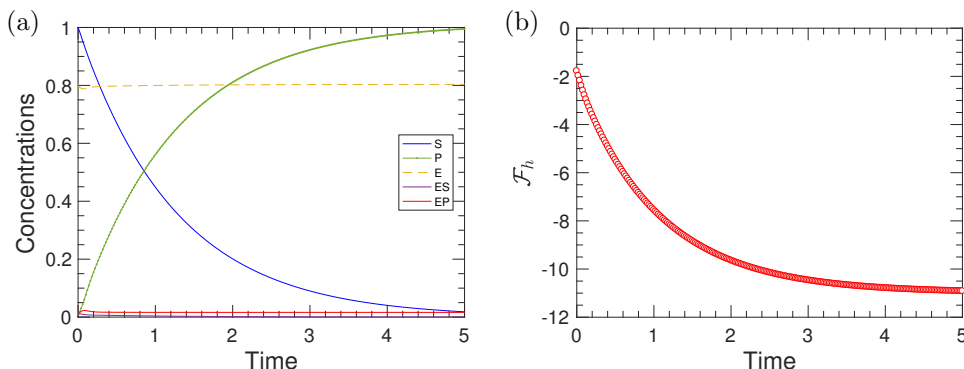


FIG. 4.1. Numerical results for the Generalized Michaelis-Menten kinetics with two intermediate states ($\Delta t = 1/50$): (a) The concentrations of different species with respect to time, (b) the numerical free energy with respect to time.

Let $\mathbf{R} = (R_1, R_2, R_3)^T$ denote three reaction trajectories, the energy-dissipation law of the system can be formulated as

$$\frac{d}{dt} \left(\sum_{i=1}^5 c_i (\ln c_i - 1 + U_i) \right) = - \int \sum_{l=1}^3 \partial_t R_l \ln \left(\frac{\partial_t R_l}{\eta_l(\mathbf{c}(\mathbf{R}))} - 1 \right). \quad (4.2)$$

where $U_1 = -\ln(k_1^- k_2^- k_3^-)$, $U_2 = \ln(k_1^+ k_3^-)$, $U_3 = -\ln k_2^-$, $U_4 = -\ln k_2^+$, $U_5 = -\ln(k_1^+ k_2^+ k_3^+)$, $\eta_1(\mathbf{c}) = k_1^- c_3$, $\eta_2(\mathbf{c}) = k_2^- c_4$, $\eta_3(\mathbf{c}) = k_3^- c_1 c_5$. We take $k_1^+ = 1$, $k_1^- = 0.5$, $k_2^+ = 100$, $k_2^- = 1$, $k_3^+ = 100$ and $k_3^- = 1$ in the numerical simulation.

It is difficult to preserve the positivity of all species, as the concentration of ES and EP will be around 0. Figure 4.1 shows the numerical result with initial condition $c_1 = 1, c_2 = 0.8, c_3 = c_4 = c_5 = 0.01$. The time step size used in the simulation is $\Delta t = 1/50$. The numerical result clearly shows the positivity-preserving and energy-stable properties of the proposed numerical scheme.

Acknowledgement. This work is partially supported by the National Science Foundation (USA) grants NSF DMS-2410742 (C. Liu), DMS-2012669 (C. Wang), DMS-2309548 (C. Wang), and DMS-2410740 (Y. Wang).

REFERENCES

- [1] D.F. Anderson, G. Craciun, M. Gopalkrishnan, and C. Wiuf, *Lyapunov functions, stationary distributions, and non-equilibrium potential for reaction networks*, Bull. Math. Biol., **77**(9):1744–1767, 2015. [1](#), [1](#), [2](#)
- [2] J. Bruggeman, H. Burchard, B.W. Kooi, and B. Sommeijer, *A second-order, unconditionally positive, mass-conserving integration scheme for biochemical systems*, Appl. Numer. Math., **57**(1):36–58, 2007. [1](#)
- [3] H. Burchard, E. Deleersnijder, and A. Meister, *A high-order conservative Patankar-type discretisation for stiff systems of production–destruction equations*, Appl. Numer. Math., **47**(1):1–30, 2003. [1](#)
- [4] W. Chen, C. Wang, X. Wang, and S.M. Wise, *Positivity-preserving, energy stable numerical schemes for the Cahn-Hilliard equation with logarithmic potential*, J. Comput. Phys. X, **3**:100031, 2019. [3](#)
- [5] S.R. De Groot and P. Mazur, *Non-equilibrium Thermodynamics*, Courier Corporation, 2013. [1](#)
- [6] T.A.J. Duke, *Molecular model of muscle contraction*, Proc. Natl. Acad. Sci., **96**(6):2770–2775, 1999. [1](#)
- [7] L. Formaggia and A. Scotti, *Positivity and conservation properties of some integration schemes for mass action kinetics*, SIAM J. Numer. Anal., **49**(3):1267–1288, 2011. [1](#), [1](#), [1](#)
- [8] J.P. Keener and J. Sneyd, *Mathematical Physiology*, Springer, **1**, 1998. [1](#), [1](#), [4](#)
- [9] D. Kondepudi and I. Prigogine, *Modern Thermodynamics: From Heat Engines to Dissipative Structures*, John Wiley & Sons, 2014. [1](#)
- [10] C. Liu, C. Wang, and Y. Wang, *A structure-preserving, operator splitting scheme for reaction-diffusion equations with detailed balance*, J. Comput. Phys., **436**:110253, 2021. [1](#), [2](#), [3](#), [3](#)
- [11] C. Liu, C. Wang, and Y. Wang, *A second-order accurate, operator splitting scheme for reaction-diffusion systems in an energetic variational formulation*, SIAM J. Sci. Comput., **44**(4):A2276–A2301, 2022. [1](#)
- [12] C. Liu, C. Wang, Y. Wang, and S.M. Wise, *Convergence analysis of the variational operator splitting scheme for a reaction-diffusion system with detailed balance*, SIAM J. Numer. Anal., **60**(2):781–803, 2022. [1](#)
- [13] L. Michaelis and M.L. Menten, *The kinetics of the inversion effect*, Biochem. Z., **49**:333–369, 1913. [4](#)
- [14] A. Mielke, *A gradient structure for reaction–diffusion systems and for energy-drift-diffusion systems*, Nonlinearity, **24**(4):1329, 2011. [1](#)
- [15] G.F. Oster and A.S. Perelson, *Chemical reaction dynamics. Part I: Geometrical structure*, Arch. Ration. Mech. Anal., **55**(3):230–274, 1974. [1](#)
- [16] J.E. Pearson, *Complex patterns in a simple system*, Science, **261**(5118):189–192, 1993. [1](#)
- [17] L. Peller and R.A. Alberty, *Multiple intermediates in steady state enzyme kinetics. I. The mechanism involving a single substrate and product*, J. Amer. Chem. Soc., **81**(22):5907–5914, 1959. [4](#)

- [18] A. Sandu, *Positive numerical integration methods for chemical kinetic systems*, J. Comput. Phys., **170**(2):589–602, 2001. [1](#)
- [19] Y. Wang, C. Liu, P. Liu, and B. Eisenberg, *Field theory of reaction-diffusion: Law of mass action with an energetic variational approach*, Phys. Rev. E, **102**(6):062147, 2020. [1](#), [1](#)
- [20] J. Wei, *Axiomatic treatment of chemical reaction systems*, J. Chem. Phys., **36**(6):1578–1584, 1962. [1](#)