Inference of analytical thermodynamic models for biological networks

Inference of analytical thermodynamic models for biological networks

Eliodoro Chiavazzo,[1,*] Matteo Fasano,[1] and Pietro Asinari[1]

[1] *Energy Department, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy*

(Dated: November 13, 2012)

We present an automated algorithm for inferring analytical models of closed reactive biochemical mixtures, on the basis of standard approaches borrowed from thermodynamics and kinetic theory of gases. As an input, the method requires a number of steady states (i.e. an equilibria cloud in the phase-space), and at least one time series of measurements for each species. Validations are discussed for both the Michaelis-Menten mechanism (four species, two conservation laws) and the mitogen-activated protein kinase - MAPK - mechanism (eleven species, three conservation laws).

PACS numbers: 05.20.Dd, 82.39.-k, 82.20.Wt

INTRODUCTION

The reverse engineering of biological networks from experimental observations has recently gained an increasing attention, owing to remarkable advancements in modern high-throughput techniques for the generation of time series data on metabolites, genes and other components of biological relevance [1]. However, due to high dimensionality, the latter still remains a demanding task that often requires *a priori* knowledge on the system structure. Predictive mathematical models are highly desirable, for instance, for the external control of cellular functions, and this has motivated an intense effort in such a direction [2, 3]. In this work, we intend to investigate the ability of some classical thermodynamic approaches for the automatic prediction of equilibria, dynamical behavior and system structure. The main advantage of such an approach is that it does not require prior knowledge on the underlying biochemical mechanism, and it is solely based on measurements of species concentrations in closed systems.

We focus on biological systems formed by several species interacting according to a web of (bio)chemical reactions in closed systems under fixed temperature T and volume V. We further assume that dissipation is ensured by the existence of a global Lyapunov function G, which is typically linked to a thermodynamic potential, and a unique steady state (equilibrium) is reached after a sufficiently long time. Let the concentration of n species evolve in time according to an autonomous system of ordinary differential equations (ODEs):

$$\dot{x} = \frac{dx}{dt} = f\left(x\right), \tag{1}$$

with $x = [x_1, ..., x_n]^T$ defining the system state (e.g. in terms of molar concentrations x_i). Let x^{eq} and $G(x)$ be the unique equilibrium state of the ODEs (1) and its global Lyapunov function, respectively. Hence, at all instant t, the time derivative of G is non-positive, $\dot{G} = \nabla G f \leq 0$, and it vanishes at steady state: $\dot{G}\left(T, V, x^{eq}\right) = 0$. Time dynamics (1) is often characterized by linear constraints (e.g. due to conservation of the mole number of elements forming the chemical species). Thus, assuming the presence of r conserved quantities, there exists a fixed $(r \times n)$ matrix M such that, at all time instants t:

$$Mx\left(t\right) = C, \tag{2}$$

with C being an $r-$component column of fixed quantities (conserved moieties).

SEARCHING FOR CONSERVATION LAWS

Neither the number nor the expressions of the conservation laws (2) are typically known when investigating on a new biological phenomenon, unless pre-existing knowledge on the reaction stoichiometry is available. For addressing the above issues, the suggested approach is based upon the analysis of a collection of scattered steady states (experimental equilibrium cloud), and at least one time series of species concentrations evolving from an arbitrary initial state. In this work, we perform inspection of the equilibrium cloud by means of *principal component analysis* - PCA - [4] in order to estimate the cloud dimension which, as discussed below, indicates the number of conservation laws. For the sake of completeness, it is worth stressing that more recent non-linear techniques, such as *diffusion maps* [5], may be also adopted for estimating the dimension r.

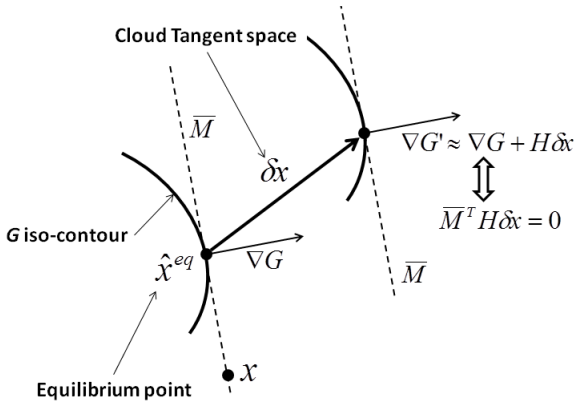

FIG. 1: Geometry underpinning the constrained minimization problems (8) and (14): Solutions are located where the affine hyperplane $x + \overline{M}$ is tangent to the G function iso-lines. Vectors δx spanning the local tangent space to the equilibrium cloud are thus linked to both the null space of the matrix M ($\overline{M} = \ker M$) and the second derivative matrix H of the Lyapunov function G. As a result, orthogonality between the columns of $\overline{M}$ and the gradient of G (∇G) implies: $\overline{M}^T H\delta x = 0$.

We notice that, in a perfectly closed system, thermodynamics and conservation laws rule the geometry of the manifold collecting all the equilibrium states. As a result, the matrix M is fully determined upon computation of the local tangent space to such a manifold. According to the pictorial representation in Fig. 1, we notice that the following relationship holds: $\overline{M}^T H\Delta X = 0$. Here, the columns of the $n \times (n-r)$ matrix $\overline{M} = \ker(M)$ span the null space of M, H is the second derivative matrix of the Lyapunov thermodynamic function G while the columns of the $(n \times r)$ matrix ΔX form a basis of the local tangent space of the equilibrium cloud. As a result, the following equation holds:

$$M' = \left[\ker\left[\left(\ker \Delta X^T\right)^T H^{-1}\right]\right]^T, \tag{3}$$

where the superscript ^{T} and the prime symbol $'$ denote transposition and the orthonormal basis respectively. For numerical purposes, a generic column δx of the matrix ΔX can be conveniently approximated by local interpolation or finite differences. Nevertheless, we stress that adoption of (3) may lead to inaccurate results due to a poor estimate of the vectors δx, or even to a lack of knowledge on the function G such as the activity coefficients for non-ideal mixtures (see below). Therefore, below we describe a stochastic method, based on the Metropolis algorithm [6], enabling an accurate computation of M by processing the time series of species concentrations. First, we initialize the k-th row, M_k, of the conservation law matrix: This can be made either stochastically or even on the basis of the estimate (3). We assume that all species concentrations are recorded at a discrete set of time instants (t_j) between t_1 and t_m: The latter data is thus available in the form of a time series, stored in the $(n \times m)$ data array $\hat{X} = \{\hat{x}_{ij}\}$, with the generic element $\hat{x}_{ij}$ denoting the concentration of the i-th species at the time instant t_j. Next, we compute the following deviation quantity:

$$d_k = \sum_{j=1}^{m} \left|\hat{C}_{kj} - \bar{C}_k\right|, \tag{4}$$

where, for the k-th conservation law, the j-th term of the time series $\{\hat{C}_{kj}\}$ and its time-averaged value $\bar{C}_k$ are defined as:

$$\hat{C}_{kj} = \sum_{i=1}^{n} \hat{x}_{ij} M_k(i),\ \bar{C}_k = \left(\sum_{j=1}^{m} \hat{C}_{kj} dt_j\right) \Big/ (t_m - t_1),$$

with dt_j the duration of the time interval corresponding to the term $\hat{C}_{kj}$.

Eq. (4) provides with a non negative term, that vanishes if M_k describes a conservation law of the dynamical system (1). Next, a refinement process is set, where at each step one randomly chosen element of M_k undergoes a mutation. If mutation occurs at position i^*, we impose: $M_k(i^*) \pm g$, with g being a positive quantity. Mutation is accepted with probability one, if it induces a decrease in the deviation quantity d_k. Conversely, when an increase of d_k happens, acceptance only occurs with probability: $e^{-\Delta d_k/\tilde{T}}$, with Δd_k and $\tilde{T}$ the incremental deviation quantity and a tuning parameter that stipulates the frequency acceptance for $\Delta d_k > 0$, respectively. Upon convergence (i.e. d_k is stationary), if M_k is linearly independent with respect to the vector basis $\{M_1, ..., M_{k-1}\}$, we retain the solution and k is updated to $k+1$ (until $k < r$). Otherwise, the procedure is repeated starting from a new random vector M_k.

EQUILIBRIA OF IDEAL MIXTURES

Let μ_j be the chemical potential of the $j-$th species. For ideal systems, μ_j reads:

$$\mu_j = \mu_j^0 + RT \ln x_j, \tag{5}$$

where R, V and $x_j = n_j/V$ are the universal gas constant, the system volume and the molar concentration of the j-th species, respectively, with n_j denoting the corresponding number of moles.

For closed systems with fixed volume V and temperature T, the thermodynamic Lyapunov function G is provided by Helmholtz potential as follows:

$$G = V \sum_{j=1}^{n} x_j \mu_j - pV = V \sum_{j=1}^{n} \left(x_j \mu_j^0 + RT x_j \ln x_j \right) - RTV \sum_{j=1}^{n} x_j, \tag{6}$$

where pressure is dictated by the ideal gas law of state: $p = RT \sum_{j=1}^{n} x_j$. The equilibrium state of (1) can be computed by global minimization of the function $\tilde{G}$:

$$\tilde{G} = G + \sum_{i=1}^{r} \lambda_i \left(M_i x - C_i \right), \tag{7}$$

where λ_i, M_i and C_i are the $i-$th Lagrange multiplier, the $i-$th row of the matrix M and the $i-$th component of C, respectively. In other words, x^{eq} fulfills the following algebraic system:

$$\begin{aligned} \frac{1}{V} \frac{\partial \tilde{G}}{\partial x_j} = \mu_j^0 + RT \ln x_j + \sum\nolimits_{i=1}^{r} \lambda_i M(i,j) = 0, \quad \forall j = 1, ..., n \\ \frac{\partial \tilde{G}}{\partial \lambda_k} = M_k x - C_k = 0, \ \forall k = 1, ..., r. \end{aligned} \tag{8}$$

Let $\hat{x}_0^{eq}$ be a reference equilibrium point of (1) (e.g. measured by experiments), it follows:

$$\mu_j^0 = -RT \ln \hat{x}_{0j}^{eq} - \sum_{i=1}^{r} \lambda_{0i} M(i,j), \ \forall j = 1, ..., n. \tag{9}$$

Any other equilibrium state (corresponding to an arbitrary initial condition x^{in}) can be computed by means of (8) recast in the more explicit form:

$$\begin{aligned} &-RT \ln \hat{x}_{0j}^{eq} + RT \ln x_j + \sum\nolimits_{i=1}^{r} (\lambda_i - \lambda_{0i}) M(i,j) = 0, \ j = 1, ..., n \\ &M_k x - C_k = 0, \ k = 1, ..., r \end{aligned}$$

hence,

$$\begin{aligned} &\ln \left(x_j / \hat{x}_{0j}^{eq} \right) + \sum\nolimits_{i=1}^{r} \bar{\lambda}_i M(i,j) = 0, \ j = 1, ..., n \\ &M_k x - M_k x^{in} = 0, \ k = 1, ..., r. \end{aligned} \tag{10}$$

with $\bar{\lambda}_i = (\lambda_i - \lambda_{0i})/RT$. The nonlinear algebraic system (10) enables to compute the equilibrium state of (1) x^{eq} corresponding to the initial condition x^{in} (when the ideal system assumption is valid). More specifically, according to (10), equilibrium states only depend on the r conserved quantities (under fixed temperature and volume) computed by x^{in}:

$$x^{eq} = x^{eq}(C_1, ..., C_r) = x^{eq}\left(M x^{in} \right), \tag{11}$$

and can be readily computed once an arbitrary equilibrium state $\hat{x}_0^{eq}$ and the matrix M are set.

EQUILIBRIA OF NON-IDEAL MIXTURES

For non-ideal systems, the chemical potential (5) takes the more general form [7, 8]:

$$\mu_j = \mu_j^0 + RT\ln\gamma_j x_j, \tag{12}$$

where the $j-$th *activity coefficient* γ_j is typically a non-trivial function of the system state, while the thermodynamic Lyapunov function G (for systems under fixed volume V and temperature T) takes the form:

$$G = V\sum\nolimits_{j=1}^{n}\left(x_j\mu_j^0 + RTx_j\ln\gamma_j x_j\right) - pV. \tag{13}$$

The steady state condition ($\nabla\tilde{G} = 0$) thus yields:

$$\begin{array}{l}\dfrac{1}{V}\dfrac{\partial\tilde{G}}{\partial x_j} = \mu_j^0 + RT\left(\ln\gamma_j x_j{+}1\right) + RT\sum\nolimits_{i=1}^{n}\dfrac{x_i}{\gamma_i}\dfrac{\partial\gamma_i}{\partial x_j} - \dfrac{\partial p}{\partial x_j} \\ \qquad\qquad + \sum\nolimits_{i=1}^{r}\lambda_i M\left(i,j\right) = 0,\ \forall j = 1,...,n \\ \frac{\partial\tilde{G}}{\partial\lambda_k} = M_k x - C_k = 0,\ \forall k = 1,...,r.\end{array} \tag{14}$$

Since the manifold collecting all steady states x_j^{eq} can be conveniently parameterized by conserved quantities C_i, in a neighborhood of an arbitrary point x_j^{eq}, we assume that activities depend on the quantities C_i: $\gamma_j = \gamma_j\left(C_i\right)$. Moreover, due to the fundamental relationship in thermodynamics stipulating that (for any system) the j-th chemical potential (12) is the partial derivative of Helmholtz potential with respect to the moles of the j-th species under fixed volume, temperature and remaining species concentrations, namely

$$\mu_j = \frac{1}{V}\left.\frac{\partial G}{\partial x_j}\right|_{V,T,x_{i\neq j}}, \tag{15}$$

the system (14) can be written as follows:

$$\begin{array}{l}\dfrac{1}{V}\dfrac{\partial\tilde{G}}{\partial x_j} = \mu_j^0 + RT\ln\gamma_j x_j + \sum\nolimits_{i=1}^{r}\lambda_i M\left(i,j\right) = 0,\ \forall j = 1,...,n \\ \frac{\partial\tilde{G}}{\partial\lambda_k} = M_k x - C_k = 0,\ \forall k = 1,...,r.\end{array} \tag{16}$$

Let us consider a reference solution whose equilibrium condition is $\hat{x}_0^{eq}$ (e.g. measured by experiments). The quantities μ_j^0 can be computed as follows:

$$\mu_j^0 = -RT\ln\gamma_{0j}\hat{x}_{0j}^{eq} - \sum\nolimits_{i=1}^{r}\lambda_{0i}M\left(i,j\right) \tag{17}$$

Upon substitution of (17) in (16), we obtain the following condition for equilibrium states of non-ideal mixtures:

$$\begin{array}{l}-RT\ln\gamma_{0j}\hat{x}_{0j}^{eq} - \sum\nolimits_{i=1}^{r}\lambda_{0i}M\left(i,j\right) + RT\ln\gamma_j x_j + \sum\nolimits_{i=1}^{r}\lambda_i M\left(i,j\right) = 0,\ \forall j = 1,...,n \\ M_j x - C_j = 0,\ k = 1,...,r.\end{array} \tag{18}$$

Around the reference point $\hat{x}_0^{eq}$, activities can be approximated by a polynomial series up to a certain order, hence the system (18) takes the final form:

$$\begin{array}{l}\ln\dfrac{\gamma_j x_j}{\gamma_{0j}\hat{x}_{0j}^{eq}} + \sum\limits_{i=1}^{r}\bar{\lambda}_i M\left(i,j\right) = 0,\ \forall j = 1,...,n \\ M_k x - M_k x^{in} = 0,\ k = 1,...,r \\ \gamma_j \approx \gamma_{0j} + \sum\nolimits_{i=1}^{r}\left.\frac{\partial\gamma_j}{\partial C_i}\right|_0\left(C_i - C_{0i}\right) + ...\ ,\ \forall j = 1,...,n.\end{array} \tag{19}$$

In the following, the expansion of activity coefficients is considered up to first order. Owing to the latter expansion, the system (19) can be adopted for predicting the steady state of non-ideal mixtures in a vicinity of a reference point $\hat{x}_0^{eq}$, once the functions $\gamma_j = \gamma_j(C_i)$ and the matrix M are properly estimated. Let x^{in} be an arbitrary initial state of (1). A possible criterion for choosing $\hat{x}_0^{eq}$ from the available cloud in the vicinity of the target equilibrium

corresponding to x^{in} may be based on the minimal Euclidean distance between the points Mx^{in} and $M\hat{x}_0^{eq}$, in the space of conserved quantities.

Unknown constants in (19) (i.e. $\gamma_{0j}, \quad \partial\gamma_j/\partial C_i|_0$) can be estimated through optimization. To this end, here we adopted a random procedure based on the Metropolis algorithm [6]. All the unknowns are first initialized: $\gamma_{0j} = 1$, $\partial\gamma_j/\partial C_i|_0 = 0, \ \forall i, j$. Next, a refinement process starts, where at each step one randomly chosen quantity undergoes a random mutation: $\pm\varepsilon$, with ε being a small positive number. Mutations are accepted with probability one, if they induce a decrease of the following error measure:

$$E = \sum_{j=1}^{s} \sum_{i=1}^{n} \left| \frac{x_j^{eq}(i) - \hat{x}_j^{eq}(i)}{\hat{x}_j^{eq}(i)} \right|, \tag{20}$$

where $\hat{x}_j^{eq}$ is an equilibrium point from the cloud, x_j^{eq} denotes the solution of (19) at $M\hat{x}_j^{eq}$, while s is the number of cloud points within a fixed neighborhood of interest around the reference point $\hat{x}_0^{eq}$. Conversely, mutations that induce an increase of the quantity (20) are accepted with probability: $e^{-\Delta E/\tilde{T}}$, with ΔE and $\tilde{T}$ the incremental deviation quantity and a tuning parameter that stipulates the acceptance frequency for $\Delta E > 0$, respectively. Finally, the process is terminated when E becomes stationary. In the previous relative error measure (as in the following one), a small threshold in computing zero concentrations avoids numerical singularities.

Remark-In this work, we make use of the above simple instance of a *genetic algorithm* for minimizing the functions (4) and (20). However, to this end, more advanced strategies may be also adopted (e.g. alternative evolving strategies of the unknowns) such as the ones reported in the classical work by Goldberg [9]. Moreover, we notice that inclusion of higher order terms in the expansion of activity coefficients, may lead to a more general expression of the system (19), although at the cost of a more demanding minimization of the error measure (20).

INFERENCE OF A THERMODYNAMIC MODEL

Modeling of biological systems often reveals processes with significant disparity of time scales [10]. As a result, the dynamics of the latter *multiscale* systems is characterized by short bursts towards a low-dimensional manifold in the phase space (also known as *slow invariant manifold* - SIM), where the subsequent dynamics is slower and it proceeds along the manifold itself. This phenomenon may occur several times until a steady state is reached.

With this picture in mind, here we assume that, starting from arbitrary initial states, time evolution of the biological systems under study can be regarded as a sequence of relaxation processes towards lower and lower dimensional manifolds in the phase-space. More specifically, we assume that the ordinary differential equations (ODEs) governing a closed biochemical system can be expressed in terms of several Bhatnagar-Gross-Krook (BGK) operators [11] as follows:

$$\begin{aligned} \tfrac{dx}{dt} = &-\tfrac{1}{\tau_1}\left(x - x^1\right) - \sum_{i=2}^{d} \tfrac{1}{\tau_i}\left(x^{i-1} - x^i\right) - \\ &\tfrac{1}{\tau_\infty}\left[x^d - x^{eq}(C_1, ..., C_r)\right], \end{aligned} \tag{21}$$

with fixed relaxation times, $\tau_1 < ... < \tau_d < \tau_\infty$, and x^i belonging to a $(d - i + 1)$-dimensional SIM. The equation system (21) is fully determined upon the definition of a procedure for computing all the manifold points x^i. To this end, here we follow one of the simplest way for approximating the SIM for dissipative systems, by constructing the so called *Quasi Equilibrium Manifold* - *QEM* [12, 13] (also referred to as *Constrained Equilibrium Manifold*). Hence, by a definition, a generic point x^i on a QEM is computed by solving the following constrained minimization problem:

$$G(x) \to \min, \ Mx = C, \ Nx = \Xi, \tag{22}$$

where N is a $(d - i + 1) \times n$ matrix, while the elements of Ξ represent $(d - i + 1)$ additional conserved quantities with respect to the problems (8) and (14). The proposed equation system (21) has been inspired by the works of Levermore [14] and Gorban and Karlin (see, e.g., [15–17] and the kinetic model equations in the section 2.4 of [18]). We also notice that a simple instance (i.e. two-step relaxation model) of (21) has been successfully employed for the kinetic modeling of multicomponent gas mixtures [19, 20].

Relying upon the time series of species concentrations (reference data), an optimization procedure aiming at minimizing deviations between the solution of (21) and the reference data can be set up for estimating the unknown

quantities (i.e. the relaxation times τ_i and the matrix N). To this end, an optimization procedure may be adopted for minimizing the following objective function:

$$E = \sum_{j=1}^{m} \sum_{i=1}^{n} \left| \frac{x_j\,(i) - \hat{x}_j(i)}{\hat{x}_j(i)} \right|, \tag{23}$$

where x_j is a sample point along the solution trajectory of the dynamical system (21) (at the time instant t_j, with $t_1 < t_j < t_m$), whereas $\hat{x}_j(i)$ represents the concentration of the i-th species at the same time instant t_j (e.g. recorded by experiments). In this work, for validation purposes, $\hat{x}_j(i)$ are provided by sample points along solution trajectories of detailed kinetic models from the literature. To this end, the latter models are adopted as a black-box to produce input data for the proposed approach, namely an equilibrium cloud and one time series for each species concentration. No knowledge on conserved moieties (see the above section on their stochastic search) and on the web of species interactions is requested. Conversely, some insights on species interactions may be extracted as an output of the method, as described below.

Jacobian with linear conservation laws

Once the dynamical model (21) is determined, inspection of the corresponding Jacobian matrix J enables to draw a network of interactions between the biochemical species (in a vicinity of a given state), where a large absolute value of the element $J(i,j) = \partial \dot{x}_i / \partial x_j |_{C_1,...,C_r}$ denotes a strong direct influence of the species j on the dynamics of the species i (hence a link from j to i). The adopted notation implies that the latter derivative is to be computed under fixed quantities $(C_1, ..., C_r)$, hence its numerical approximation is not straightforward, and the following procedure is adopted. A Jacobian matrix J fulfills the relation: $f\,(c + dc) \cong f\,(c) + J\,(c)\,dc$. Let us assume that the dynamics f obeys r conservation laws, namely there exist a $(r \times n)$ matrix M such that:

$$Mf\,(c + dc) = Mf\,(c) + MJdc \Rightarrow MJ = 0. \tag{24}$$

The Jacobian J can be expressed as a linear combination of vectors spanning the null space of M. Let $\overline{M}$ be a $(n - r) \times n$ matrix whose rows span the null space of M, the Jacobian J takes the general form $J = \Lambda \overline{M}$, with Λ being an unknown $n \times (n - r)$ matrix. Let the $(n - r)$ directional derivatives

$$\frac{\partial f}{\partial k_i} \simeq \frac{f\,(c + \varepsilon k_i) - f\,(c)}{\varepsilon}, \tag{25}$$

be collected in a $n \times (n - r)$ matrix D, with k_i the $i-$th row of $\overline{M}$ and the small parameter ε equals to the square root of machine precision. We notice that the derivatives (25) are computed under the fixed quantities $(C_1, ..., C_r)$ by a construction. Matrix J satisfies the condition: $J\overline{M}^T = \Lambda \overline{M}\overline{M}^T = D$, namely the Jacobian takes the following explicit form:

$$J = D\left(\overline{M}\overline{M}^T\right)^{-1} \overline{M}. \tag{26}$$

Finally, interactions among species may be inferred by inspection of a *cumulative Jacobian* matrix, obtained by summing up the absolute value of elements in the Jacobian matrix (26), at a discrete set of points along a given solution trajectory of (21).

EXAMPLE: THE MICHAELIS-MENTEN MECHANISM

Let us consider the four chemical species (A_1, A_2, A_3 and A_4) involved in the two reversible reactions [12]: 1) $A_1 + A_2 \leftrightarrow A_3$, 2) $A_3 \leftrightarrow A_2 + A_4$. A kinetic model of the above mechanism can be expressed on the basis of the *mass action law* as follows (see [10] for details):

$$\dot{x} = \begin{bmatrix} k_1^- x_3 - k_1^+ x_1 x_2 \\ k_1^- x_3 - k_1^+ x_1 x_2 + k_2^+ x_3 - k_2^- x_2 x_4 \\ k_1^+ x_1 x_2 - k_1^- x_3 + k_2^- x_2 x_4 - k_2^+ x_3 \\ k_2^+ x_3 - k_2^- x_2 x_4 \end{bmatrix}, \tag{27}$$

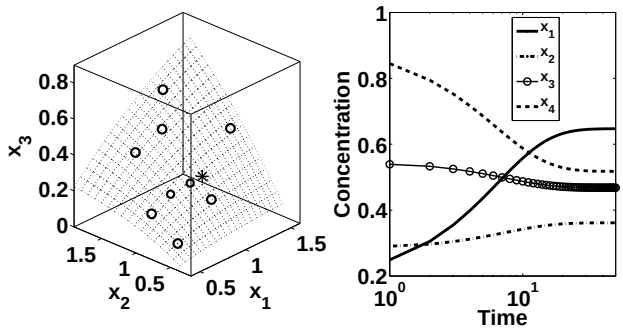

FIG. 2: Michaelis-Menten mechanism: A cloud of ten (randomly chosen) steady states are represented in the subspace: $A_1 - A_2 - A_3$. The reference point x_0^{eq} is reported by a star, while circles are adopted for the remaining steady states. Dashed lines represent the equilibrium manifold as predicted by (10) (left-hand side). A time series for each species is reported starting from an arbitrary initial condition (right-hand side).

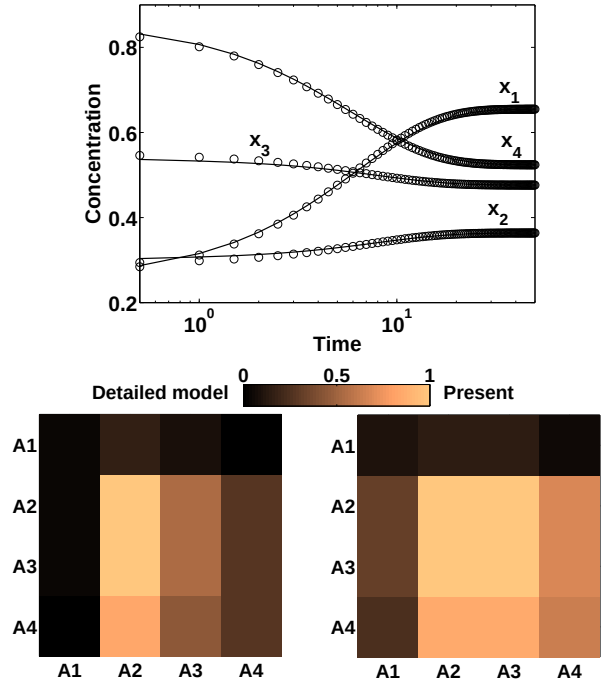

FIG. 3: (Color online). Top: Solution trajectories as predicted by the detailed kinetic model (27) (circles) and the present model (21) with $d = 1$ (lines). In the latter case, optimal parameters were found to be: $\tau_1 = 0.21$, $\tau_\infty = 6.2$, $N = [-0.675, -0.273, 0.211, 0.652]$. Down: Interaction among species as predicted by inspection of a *cumulative Jacobian* of the detailed model (27) (left-hand side) and of the present model (right-hand side).

with $k_i^{\pm}$ being the direct and reverse reaction rate constants of the reaction i, respectively. Here, the dynamical system (27) is used as a black-box (with $k_1^+ = 0.3$, $k_1^- = 0.15$, $k_2^+ = 0.8$ and $k_2^- = 2.0$), whose behavior is to be emulated by means of the proposed approach. Similarly, the two above reversible reactions are assumed unknown. Conversely, as shown in Fig. 2, a discrete set of steady states and one time series for each species concentration are the only accessible data. Here, principal component analysis [4] of the ten steady states in Fig. 2, reveals that they lay on a two-dimensional surface with less than 1% variance. Eq. (3) and the stochastic search of conservation laws (with $g = 1$ and $\tilde{T} = 10^{-4}$) deliver the following expressions for M:

$$M' = \begin{bmatrix} 0.258 & 0.516 & 0.775 & 0.258 \\ 0.577 & -0.578 & 0 & 0.577 \end{bmatrix},$$
$$M = \begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix},$$

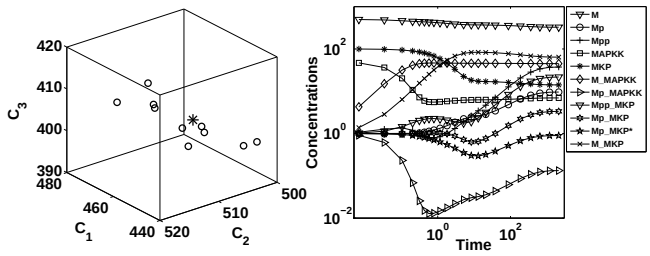

FIG. 4: Input data generated by the MAPK mechanism at the elementary step level in [21]. Left-hand side: Ten random equilibria (circles), around a reference state (star), are reported in the conserved quantity space: $C_1 - C_2 - C_3$. Right-hand side: A time series for each species conservation with: $C_{01} = 456$, $C_{02} = 508$, $C_{03} = 404$.

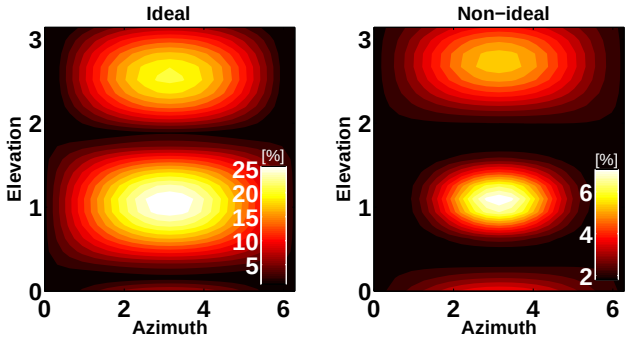

FIG. 5: (Color online). MAPK cascade: Equilibrium prediction in a neighborhood of a reference point $C_{01} = 456$, $C_{02} = 508$, $C_{03} = 404$ where the matrix M is given by (28). The relative error, $(100/n)\sum_{i=1}^{n} |x_i^{eq} - \hat{x}_i^{eq}| / \hat{x}_i^{eq}$, with $\hat{x}_{eq}$ being consistent with the detailed mechanism in [21], while x_i^{eq} is predicted according to the ideal mixture approximation (10) (left-hand side) and non-ideal mixture approximation (19) (right-hand side). Results refer to a spherical surface in the conserved quantity space, with center $C_0 = [C_{01}, C_{02}, C_{03}]$ and radius 7.

respectively, with the latter being the exact value of M (consistently with the detailed kinetic model (27)), and the former only an approximation. As shown in Fig. 3 (top), an excellent agreement between the solution trajectories of the detailed kinetic model (27) and the present system (21) is found, where $d = 1$ and initial conditions are the ones reported on the right-hand side of Fig. 2. Here, the Lyapunov function G takes the form (6) while, by means of the genetic algorithm [9] readily available in MATLAB [22], the optimal parameters in (21) and (22), are found to be: $\tau_1 = 0.21$, $\tau_\infty = 6.2$, $N = [-0.675, -0.273, 0.211, 0.652]$. Finally, the cumulative Jacobian matrix is computed for both the detailed model (27) and the present model (21). In Fig. 3 (down) we report a graphic representation, where in both cases the Jacobian is normalized such that: $0 < J(i,j) < 1$. The comparison shows that the present approach is indeed capable of predicting the main features of the interaction network underlying the Michaelis-Menten mechanism. For instance, consistently with the above reversible reactions, a weak (direct) interaction between the first and the fourth species is observed.

EXAMPLE: MAPK MECHANISM

Let us consider the mitogen-activated protein kinase (MAPK) mechanism at the elementary step level [21], where eleven species (ordered as: M, Mp, Mpp, MAPKK, MKP, M_MAPKK, Mp_MAPKK, Mpp_MKP, Mp_MKP, Mp_MKP$*$, M_MKP) interact according to a ten reaction mechanism. PCA [4] of the steady states in Fig. 4, reveals that they lay on a three-dimensional surface with less than 1% variance, while Eq. (3) and implementation of the

above stochastic search for conservation laws (with fixed $g = 1$ and $\tilde{T} = 10^{-4}$) yield:

$$M' = \begin{bmatrix} 3.21 & 3.49 & 3.78 & -0.263 & -0.516 \\ 5.33 & 1.85 & -1.64 & -5.37 & -1.85 \\ 0.31 & -2.44 & -5.17 & 2.17 & 4.92 \\ 2.98 & 4.13 & 3.26 & 3.15 & 2.98 & 2.72 \\ 0 & -2.54 & -3.49 & -2.08 & -0.003 & 3.19 \\ 2.49 & -0.02 & -0.25 & 0.851 & 2.49 & 5.01 \end{bmatrix}, \quad (28)$$

$$M = \begin{bmatrix} 1 & 1 & 1 & -1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & -1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix},$$

respectively. The latter matrix is in perfect agreement with the three conserved moieties described in [21] (i.e. linearly dependent), while the former is a rough approximation. In the following computations, we always make use of the latter exact matrix M. As shown in Fig. 5, the ideal mixture assumption introduces significant deviations between predictions by (10) and the detailed model in [21], even in the vicinity of a reference state. However, a remarkable improvement can be achieved by means of the system (19) for non-ideal mixtures. Here, around a reference point ($C_0 = [456, 508, 404]$), the parameters γ_0 and $\partial\gamma_i/\partial C_k|_0$ were estimated trough the above optimization procedure (with $\epsilon = 10^{-4}$, $\tilde{T} = 10^{-5}$) applied to the ten equilibrium points in Fig. 4 (circles on the left-hand side). As shown in Fig. 5, adoption of the system (19) with optimal parameters significantly improves the accuracy of equilibrium predictions (in the neighborhood of a reference point) compared to solutions of the system for ideal mixtures (10). Moreover, a dynamical system (21) has been inferred with $d = 1$, $d = 2$ and $d = 3$, upon minimization of the function (23) by means of the genetic algorithm [9] readily available in the MATLAB package [22]. The adopted input data $\hat{x}_j(i)$ are shown on the right-hand side of Fig. 4 while, for simplicity, the function G takes the ideal mixture expression (6). In Fig. (6), a comparison between the solutions of the detailed kinetic model in [21] and of the dynamical system (21) is reported for $C_1 = 415$, $C_2 = 354$, $C_3 = 467$. In this case, we observe that a three relaxation process ($d = 2$) is sufficient for achieving a good agreement. In the latter case, optimal parameters were found to be:

$$\tau_1 = 0.2,\ \tau_2 = 1.2,\ \tau_\infty = 250,$$
$$N = \begin{bmatrix} -0.145 & 0.83 & 1.01 & -4.76 & 2.24 \\ -0.147 & -3.38 & 1.53 & -0.256 & -1.52 \\ -4.91 & -3.76 & 3.24 & 3.08 & 2.78 & 2.17 \\ -0.604 & 0.973 & -0.518 & -6.48 & -3.40 & 1.35 \end{bmatrix}. \quad (29)$$

Finally, as shown in Fig. 7, the cumulative Jacobian matrix of (21) (normalized such that: $0 < J(i, j) < 1$) with $d = 3$ and the parameters (29) shares some common patterns with the cumulative Jacobian of the detailed model provided in [21]. As an example, it arises the central role of the fifth species (MKP) in the biological network. Conversely, from our simulations, we also notice that the present approach shows a tendency to over-predict the number of species interactions (see, e.g., interactions of the first and eleventh species in Fig. 7). Moreover, for the sake of comparison, for the case of Fig. 7, we computed that the $2-$norm of the matrix $(J - J_d)$ is 3.08, with J and J_d being the (normalized) cumulative Jacobian matrix of (21) and detailed system, respectively. Conversely, on average, the $2-$norm of the matrix $(J_{rand} - J_d)$ is 5.16 (+68% of prediction error), where J_{rand} is a uniformly distributed pseudorandom guess of the (normalized) cumulative Jacobian.

CONCLUSIONS

In this work, we have presented an automatic procedure to infer analytical models for closed reactive biochemical systems. The present method is solely based on a set of random steady states and one time series of species concentrations, with the main advantage that no prior knowledge on the underlying biochemical mechanism and system structure is requested as an input. Firstly, the proposed method involves a search for conservation laws, typically due to conservation of elements involved in the biochemical reactions. Secondly, equilibrium states are predicted by minimizing a thermodynamic potential under the above linear conservation laws, and two formulations are suggested for both ideal and non-ideal (bio)chemical mixtures. Moreover, inspired by other applications from the kinetic theory of gases [15–17, 19, 20], the species dynamics is imagined to have the general form of a sequence of hierarchical collapses onto low dimensional manifolds (slow invariant manifolds - SIMs) in the phase-space at fixed rates. Hence, optimization tools (i.e. genetic algorithms) have been adopted to both approximate points on the SIMs and evaluate relaxation

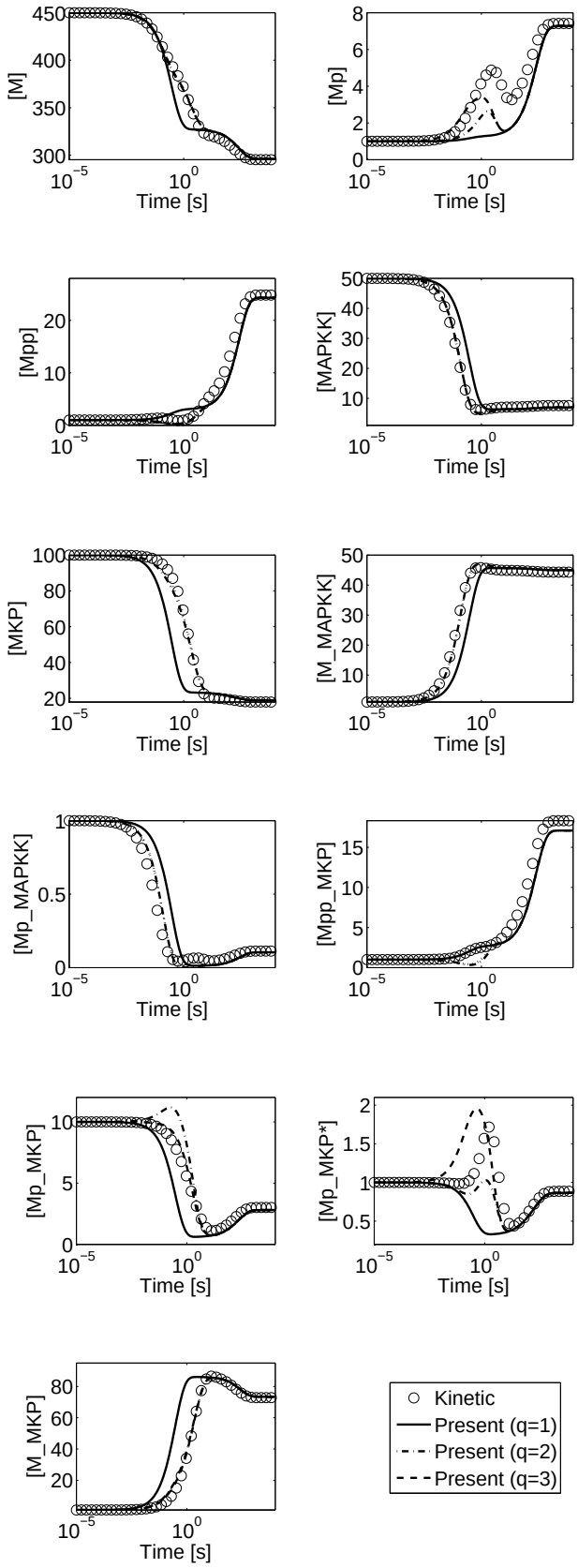

FIG. 6: MAPK mechanism: Solution trajectories as predicted by the detailed kinetic model in [21] and the model (21) with $d = 1$, $d = 2$ and $d = 3$. Conserved quantities are: $C_1 = 415$, $C_2 = 354$, $C_3 = 467$.

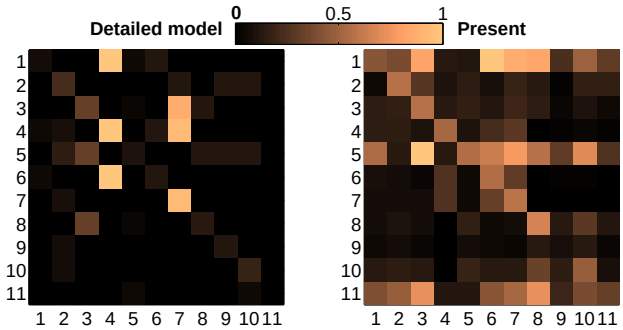

FIG. 7: (Color online). Interaction among species as predicted by inspection of a *cumulative Jacobian* J of the detailed kinetic model in [21] (left-hand side) and of the present model (21) with $d = 3$ and the optimal parameters (29) (right-hand side). The 2−norm of the matrices $(J - J_d)$ and $(J_{rand} - J_d)$ is 3.08 and 5.16 (+68% of prediction error), respectively. Here, J_d and J_{rand} denote the Jacobian of the detailed kinetic model and a uniformly distributed pseudorandom guess of the (normalized) cumulative Jacobian, respectively.

times. Finally, based on the Jacobian matrix of the latter dynamical system, some knowledge on the structure of the interaction network may be revealed. Validation is carried out for two proven examples: The Michaelis-Menten mechanism [12], and the MAPK cascade [21]. In both cases, we experience a good agreement of steady state and time dynamics predictions, at least in a neighborhood of a given reference state (i.e. from experiments) and if enough degrees of freedom are considered. Moreover, we notice that inspection of the cumulative Jacobian matrix can indeed provide some preliminary insights on the biological network (i.e. species interactions) underlying the detailed kinetic mechanism of the biological phenomenon under study. In this respect, we also stress that, from current computations, a drawback of the proposed method was noticed, where a typically larger number of species interactions is predicted by the method (false positives), compared to the true network. Here, for the sake of simplicity, smooth data from detailed kinetic models were adopted as a reference. However, in the near future, we plan to investigate the performances of the suggested method in real experiments, where time series of biochemical species are certainly affected by noise.

ACKNOWLEDGMENTS

Authors wish to thank Paolo Provero from the Molecular Biotechnology Center (MBC) at the University of Turin for discussions and suggestions. Referees are gratefully acknowledged for the help in improving the quality of the manuscript.

* Electronic address: `eliodoro.chiavazzo@polito.it`

[1] I. Chou and E. Voit, Math. Biosci. **219**, 57 (2009).
[2] C. A. Penfold and D. L. Wild, Interface focus **1**, 857 (2011).
[3] B. A. McKinney, J. E. Crowe, H. U. Voss, P. S. Crooke, N. Barney, and J. H. Moore, Phys. Rev. E **73**, 021912 (2006).
[4] I. T. Jolliffe, *Principal Component Analysis* (Springer, 2nd Edition, 2002).
[5] R. Coifman and S. Lafon, App. Comp. Harmonic Analysis **21**, 5 (2006).
[6] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, Jour. Chem. Phys. **21**, 1087 (1953).
[7] K. Denbigh, *The Principles of Chemical Equilibrium* (Cambridge University Press, 1981).
[8] Y. Damirel, *Nonequilibrium Thermodynamics. Transport and Rate Processes in Physical, Chemical and Biological Systems.* (Elsevier, 2007).
[9] D. E. Goldberg, *Genetic Algorithms in Search, Optimization & Machine Learning* (Addison-Wesley, 1989).
[10] O. Demin and I. Goryanin, *Kinetic Modelling in Systems Biology* (CRC Press, 2009).
[11] P. L. Bhatnagar, E. P. Gross, and M. Krook, Phys. Rev. **94(3)**, 511 (1954).
[12] A. N. Gorban and I. V. Karlin, *Invariant Manifolds for Physical and Chemical Kinetics* (Springer, Berlin, 2005).
[13] E. Chiavazzo and I. Karlin, Phys. Rev. E **83**, 036706 (2011).
[14] C. D. Levermore, Journal of Statistical Physics **83**, 1021 (1996).

[15] A. Gorban, I. Karlin, V. Zmievskii, and T. Nonnenmacher, Physica A **231**, 648 (1996).
[16] A. Gorban, I. Karlin, and V. Zmievskii, Transport Theor. Statist. Phys. **28(3)**, 271 (1999).
[17] A. Gorban and I. Karlin, Physica A **206**, 401 (1994).
[18] A. Gorban, I. Karlin, and A. Zinovyev, Physica A **333**, 106 (2004).
[19] S. Arcidiacono, J. Mantzaras, S. Ansumali, I. V. Karlin, C. Frouzakis, and K. B. Boulouchos, Phys. Rev. E **74** (2006).
[20] S. Arcidiacono, I. V. Karlin, J. Mantzaras, and C. E. Frouzakis, Phys. Rev. E **76** (2007).
[21] N. Markevich, J. Hoek, and B. Kholodenko, The Journal of Cell Biology **164**, 353 (2004).
[22] MATLAB, *The Language Of Technical Computing*, http://www.mathworks.com/products/matlab/.