A large-scale assessment of exact model reduction in the BioModels repository

Isabel Cristina Perez-Verona[1], Mirco Tribastone[1], and Andrea Vandin[2]

[1] IMT School for Advanced Studies Lucca, Italy
[2] DTU Lyngby, Denmark

Abstract. Chemical reaction networks are a popular formalism for modeling biological processes which supports both a deterministic and a stochastic interpretation based on ordinary differential equations and continuous-time Markov chains, respectively. In most cases, these models do not enjoy analytical solution, thus typically requiring expensive computational methods based on numerical solvers or stochastic simulations. Exact model reduction techniques can be used as an aid to lower the analysis cost by providing reduced networks that preserve the dynamics of interest to the modeler. We hereby consider a family of techniques for both deterministic and stochastic networks which are based on equivalence relations over the species in the network, leading to a coarse graining which provides the exact aggregate time-course evolution for each equivalence class. We present a large-scale empirical assessment on the BioModels repository by measuring their compression capability over 667 models. Through a number of selected case studies, we also show their ability in yielding physically interpretable reductions that can reveal dynamical patterns of the bio-molecular processes under consideration.

Keywords: Model reduction · Biological systems · Equivalence relations

1 Introduction

Computational models in systems biology combine biochemical and physiological knowledge to inform highly detailed mechanistic models of biological networks such as signaling pathways, protein-protein interaction networks, and genetic cascades. Mathematical models which equip such interaction networks with kinetic information generally lead to a dynamical-system representation in terms of a formal chemical reaction network (CRN), with two main interpretations based on ordinary differential equations (ODEs) and continuous-time Markov chains (CTMCs), respectively. In either case the model tracks the time-course evolution of all biochemical species in the network. In the ODE interpretation each species is associated with a variable of a system of (typically nonlinear) ODEs, which are analyzed from an initial condition that represents the initial concentration of each species [51]. In the CTMC interpretation [27], species are tracked discretely and each state is a vector of molecular counts, one component for each species.

It is well known that these two representations can be formally related to each other under appropriate conditions, with the ODEs being the thermodynamical limit when the number of molecules in the CRN is large enough [33].

Often it is useful to consider both interpretations—one would take the CTMC semantics as the ground-truth representation and the ODE as an approximation that estimates the first-order moments. Unfortunately, in both cases the analysis can be expensive due to the lack of analytical solutions in general. Indeed, the modeler is typically left with computational approaches such as the numerical integration of ODEs (e.g., [1]) or stochastic simulation [27]. This is a major motivating issue for several lines of research aiming at easing the computational cost of the analysis, including efficient simulation methods (e.g., [26]), approximation methods for stochastic chemical kinetics (e.g., [44]), and simplification techniques for multi-scale biochemical CRNs (e.g., [43]) and rule-based models [24,23,25].

A complementary approach that can be seen as a generic pre-analysis step consists in the use of an *exact model reduction* algorithm which, given an input CRN, produces a smaller CRN (i.e., consisting of fewer species and reactions) that preserves the output dynamics of interest to the modeler (e.g., [38,49]). This would lead to a coarse-grained CRN which still allows the full observation of the time evolution of some original species (e.g., the phosphorylated forms of downstream molecular complexes in a signaling pathway) while collapsing the behavior of other species into macro-variables. Such an approach may bring about two main advantages. First, being a CRN-to-CRN transformation, the coarse-grained CRN can still be subjected to other techniques to reduce the complexity of the analysis, including *approximate* model reduction methods. Second, the very collapse of several species into one may carry a physical interpretation that increases our understanding of the biology. The latter point appears to be of scientific relevance regardless of the CRN reduction ratio. Therefore, two suitable indicators of the relevance of exact model reduction techniques in practice are the effectiveness and the intelligibility of the reductions.

This paper presents a large-scale assessment on biological models in the literature for recent reduction techniques for CRNs, supporting the ODE and the CTMC semantics [8,7,10,11,13]. The techniques share two main unifying ideas:

i) Identifying criteria on the species and reactions of a CRN inducing a suitable *species equivalence*, i.e., a partition of the species such that an exactly reduced CRN can be written having a macro-species per partition block.
ii) Developing an algorithm for computing the largest species equivalence using partition refinement [41], based on iterative refinements of a given initial partition of species (with which, for instance, one can isolate the observable species to be preserved in the reduction).

The definitions of the species equivalences differ according to the underlying semantics to which they are applicable, the assumptions made on the input CRN, and the kind of reduction that they yield. Specifically, forward equivalence (FE) and backward equivalence (BE) apply to CRNs with ODE semantics based on mass-action kinetics and identify reduced models where each macro-species preserves the sums of original variables belonging to a block [10]; while with FE the

time-course of one species cannot be recovered, BE aggregates species that have the same solutions at all time points. Forward differential equivalence (FDE) and backward differential equivalence (BDE) are generalizations that can be applied to CRNs where the underlying ODEs have nonlinearities beyond polynomials such as rational expressions in Hill kinetics [7]. Finally syntactic Markovian bisimulation (SMB) is the species equivalence for stochastic CRNs [11]. It identifies a partition of species which induces a coarse graining of the underlying CTMC in terms of ordinary lumpability [5], aggregating CTMC states that have equal sums of molecular counts across each partition block of species.

Assisted by ERODE [9], a publicly available software tool that implements the aforementioned species equivalences, we carry out an assessment of the BioModels database [37], a well-known repository of quantitative models of biochemical systems.[3] Our goal is to answer the following three evaluation questions:

Q1. *How restrictive are the assumptions required by the species equivalences?* We answer this question by detailing how we translated the BioModels descriptions, available in the SBML format, into the input format of ERODE.

Q2. *What is the effectiveness of exact model reduction by species equivalence?* We measure effectiveness as the percentage of models that can be aggregated, as well as the compression ratio provided by the largest species equivalence that preserves the observables specified in the original model.

Q3. *What is the physical interpretation of the reductions?* For this question, we present a more detailed discussion of a selected number of models.

2 Background

In order to make the paper self-contained, in this section we briefly overview the main results regarding the species equivalences used in our assessment. We refer to the original papers for the details and further examples, while unifying tutorial-like presentations are given in [48,50].

Chemical reaction networks. First, we fix the notation and terminology for reaction networks. A CRN is a pair $(\mathcal{S}, \mathcal{R})$ consisting of a finite set of species $\mathcal{S}$ and a finite set of reactions $\mathcal{R}$, where each reaction is in the form $\rho \xrightarrow{f} \pi$ consisting of: a multiset of species ρ, with the multiplicity of species S denoted by ρ_S, that represents the *reactants*; a multiset of species π (the *products*); and the *propensity function* $f : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}_{\geq 0}$. Roughly speaking, it gives the rate at which the reaction fires based on the current system state; the *net stoichiometry* $\pi - \rho$ gives the state update upon the reaction firing.[4]

Example 1. Let us use a CRN $(\mathcal{S}_E, \mathcal{R}_E)$ with species S_1, S_2, S_3, S_4, S_5, and reactions $S_1 \xrightarrow{2} S_5$, $S_1 \xrightarrow{1} 2S_3$, $S_3 + S_5 \xrightarrow{3} S_3$, $S_2 \xrightarrow{2} S_3$, $S_2 \xrightarrow{1} 2S_5$, $S_4 + S_5 \xrightarrow{3} S_3$.

[3] The models are available at `https://sysma.imtlucca.it/tools/erode/cmsb2019/`

[4] As usual, the `+` and `-` operators denote multiset union and difference, respectively, while the multiplicity of a species denotes its stoichiometric coefficient.

According to the deterministic semantics of CRNs [51], a CRN is associated with an ODE system which tracks the time course of the vector of concentrations of the species at time t, $X(t) = (X_S(t))_{S\in\mathcal{S}}$, as follows:

$$\frac{dX_S(t)}{dt} = \sum_{(\rho \xrightarrow{f} \pi)\in\mathcal{R}} (\pi_S - \rho_S)\cdot f(X(t)).$$

In a deterministic *mass-action CRN*, each reaction is associated with a kinetic parameter $\lambda > 0$, and the propensity function, denoted by f_λ, is given by $f_\lambda(x) = \lambda \cdot \prod_{S\in\mathcal{S}} x_S^{\rho_S}$, where ρ is the multiset of reactants. The CRN $(\mathcal{S}_E, \mathcal{R}_E)$ is a mass-action CRN. For example, the ODEs for S_1 and S_2 are:

$$\frac{dX_1(t)}{dt} = -3\cdot X_1(t)) \qquad \frac{dX_2(t)}{dt} = -3\cdot X_2(t)$$

According to the stochastic semantics of CRNs [27], a CRN is represented as a Markov population process, a CTMC where each state is a vector $n = (n_S)_{S\in\mathcal{S}}$ of nonnegative integers that tracks the molecular counts of each species. The *initial state* is a vector representing the initial (integer) populations of each species. A transition between any two states n and $n+\pi-\rho$ occurs according to an exponential distribution with parameter $f(n)$ for each reaction $\rho \xrightarrow{f} \pi$. The CTMC underlying a CRN for an initial state consists of all states and transitions generated by applying exhaustively the reactions on all generated states, starting from the initial one. An *elementary* mass-action CRN has reactions in the form $\rho \xrightarrow{f_\lambda} \pi$ where $|\rho| \leq 2$ (i.e., at most two molecules can interact), $\lambda > 0$ is the kinetic parameter, and $f_\lambda(n) = \lambda \cdot \prod_{S\in\mathcal{S}} \binom{n_S}{\rho_S}$, where n is the source state. The CRN in Example 1 is elementary.

Forward and backward equivalence (FE and BE). FE and BE are two reduction techniques for deterministic mass-action CRNs given as equivalence relations on species which can be efficiently checked by using only structural conditions on the reactions [10]. For $\chi \in \{\mathrm{FE}, \mathrm{BE}\}$, both notions can be expressed as:

> Given a CRN $(\mathcal{S}, \mathcal{R})$, a partition $\mathcal{H}$ of species is χ if and only if for any two blocks $H, H' \in \mathcal{H}$ and any two species $S_i, S_j \in H$ it holds
> $$\mathbf{c}_\chi(S_i,\eta,H') = \mathbf{c}_\chi(S_j,\eta,H') \quad \forall \eta.\ \exists (S_k + \eta \xrightarrow{\lambda} \pi)\in\mathcal{R} \text{ for } S_k \in \{S_i, S_j\}$$

where $\mathbf{c}_\chi$ maps a species (S_i,S_j), a multiset of reagent partners (η) and a block (H') into a real number computed by inspecting once the reactions [10].

Fig. 1 shows FE partition $\mathcal{H}_f$ and BE partition $\mathcal{H}_b$, as well as their respective reduced CRNs, for the running example (We observe that $\mathcal{H}_b$ is a refinement of $\mathcal{H}_f$, but in general, FE and BE are not comparable [8,6]). FE relates species such that it is possible to rewrite the ODEs underlying the CRN in terms of sums of the variables in each block. Each macro-species in the FE-reduced CRN represents the sum of the corresponding species in the original CRN. For example, in Fig. 1(a) species $S_{1,2}$ and $S_{3,4}$ can be used to study the concentration of the sums of original variables $S_1 + S_2$ and $S_3 + S_4$, respectively.

$\mathcal{H}_f = \{\{S_1, S_2\}, \{S_3, S_4\}, \{S_5\}\}$
$\mathcal{S}_f = \{S_{1,2}, S_{3,4}, S_5\}$
$\mathcal{R}_f = \{S_{1,2} \xrightarrow{1} S_5,\ S_{1,2} \xrightarrow{0.5} 2S_{3,4}, S_{3,4} + S_5 \xrightarrow{3} S_{3,4},\ S_{1,2} \xrightarrow{1} S_{3,4},\ S_{1,2} \xrightarrow{0.5} 2S_5\}$

(a) FE-reduction

$\mathcal{H}_b = \{\{S_1, S_2\}, \{S_3\}, \{S_4\}, \{S_5\}\}$
$\mathcal{S}_b = \{S_{1,2}, S_3, S_4, S_5\}$
$\mathcal{R}_b = \{S_{1,2} \xrightarrow{1} S_5,\ S_{1,2} \xrightarrow{0.5} 2S_3,\ S_3 + S_5 \xrightarrow{3} S_3,\ S_{1,2} \xrightarrow{1} S_3,\ S_{1,2} \xrightarrow{0.5} 2S_5,\ S_4 + S_5 \xrightarrow{3} S_3\}$

(b) BE-reduction

Fig. 1: Coarsest FE/BE, and FE/BE-reductions of $(\mathcal{S}_E, \mathcal{R}_E)$ from Example 1.

BE relates species that have same ODE solution at any point in time (which therefore must have same initial condition). In the BE-reduced CRN in Fig. 1(b), $S_{1,2}$ represents the sum of original species $S_1 + S_2$. However, BE ensures that S_1 and S_2 have same ODE solution at all times. Therefore, we can recover each individual solution of by halving that of $S_{1,2}$.

Forward and backward differential equivalence (FDE and BDE). FDE and BDE are generalizations of FE and BE, respectively, for deterministic CRNs beyond mass-action [7,13]. FDE and BDE capture the same dynamical properties of FE and BE, and collapse to them for mass-action deterministic CRNs. The greater generality of FDE/BDE comes at the cost of a more computationally expensive implementation based on encodings in satisfiability modulo theory (SMT) formulas. For instance, the following formula $\psi^{\mathcal{H}_b}$ encodes the check whether partition $\mathcal{H}_b$ is a BDE:

$$\psi^{\mathcal{H}_b} := \ (X_1 = X_2) \implies (-3 \cdot X_1 = -3 \cdot X_2)$$

which checks that if all variables in same block are equal (the premise) then they must evolve in the same way, i.e. their derivative should evaluate to the same value (the conclusion). The formula has two free real variables, X_1 and X_2, corresponding to S_1 and S_2. By using an SMT solver, e.g., Z3 [19], we can check if $\mathcal{H}_b$ is a BDE by checking for the satisfiability of $\neg\psi^{\mathcal{H}_b}$. If there exists an assignment for X_1 and X_2 that makes $\neg\psi^{\mathcal{H}_b}$ true, then $\mathcal{H}_b$ is not a BDE. This is not the case, and hence it is a BDE (as expected from it being a BE).

Syntactic Markovian bisimulation (SMB). SMB is a reduction technique for stochastic mass-action elementary CRNs [11]. It is given as an equivalence on species, in the same spirit of FE and BE. Indeed, SMB can be seen as an instantiation of FE to the stochastic semantics of CRNs. We discuss this through our running example. The partition $\mathcal{H}_s = \{\{S_1\}, \{S_2\}, \{S_3, S_4\}, \{S_5\}\}$ is an SMB for the CRN $(\mathcal{S}_E, \mathcal{R}_E)$ from Example 1. The very same notion of FE/BE-reduced CRN applies to SMB as well. The $\mathcal{H}_s$-reduction of $(\mathcal{S}_E, \mathcal{R}_E)$ has species $\mathcal{S}_S = \{S_1, S_2, S_{3,4}, S_5\}$ and reactions $\mathcal{R}_S = \{S_1 \xrightarrow{2} S_5,\ S_1 \xrightarrow{1} 2S_{3,4}, S_{3,4} + S_5 \xrightarrow{3} S_{3,4}, S_2 \xrightarrow{2} S_{3,4},\ S_2 \xrightarrow{1} 2S_5\}$. A state of a CTMC of $(\mathcal{S}_E, \mathcal{R}_E)$ is a vector of size $|\mathcal{S}_E|$ counting the population of each original species, while a state of a CTMC of $(\mathcal{S}_S, \mathcal{R}_S)$ is a vector of size $|\mathcal{S}_S|$ counting the cumulative population

of each block of $\mathcal{H}_s$. The CTMCs of $(\mathcal{S}_S, \mathcal{R}_S)$ are reductions in terms of CTMC ordinary lumpability [5] of the ones obtained from $(\mathcal{S}_E, \mathcal{R}_E)$. All states of the original CTMC containing same number of $\mathcal{H}_s$-equivalent species get collapsed in the same macro-state in the reduced CTMC. Therefore, similarly to FE, SMB allows to obtain a coarse-grained version of the original CRN which allows to reason in terms of sums of variables. For example, the states $S_1 + 2S_3$, $S_1 + S_3 + S_4$, and $S_1 + 2S_4$ form an ordinary lumpable partition of a CTMC of the original CRN, and therefore get collapsed in the state $S_1 + 2S_{3,4}$ for the reduced CRN.

We note that $\mathcal{H}_s$ is a refinement of $\mathcal{H}_f$. Indeed, it has been shown that SMB implies FE, but not vice versa [11]. This will be confirmed in Section 3.4.

Partition refinement. Each equivalence is supported by a partition refinement algorithm which refines an initial partition of species (splitting its blocks) until a fixed point. The initial partition can be chosen, e.g., to isolate species that must not be aggregated because they are observables of interest to the modeler. The examples shown in this section are largest refinement of the singleton partition where all species are in a block. Other initial partitions will be used in Section 3.

3 Experimental set-up

3.1 Overview of the BioModels repository

The BioModels Database is a repository of computational models of biological processes [37]. It hosts dynamical quantitative models described in peer-reviewed scientific literature as well as models generated automatically from pathway resources such as KEGG [32], BioCarta [40], MetaCyc [14], PID [42] and SABIO-RK [52]. BioModels covers a wide range of models from several biological categories such as biochemical reaction systems, kinetic models, metabolic networks, steady-state models and signaling pathways. Models are available in the Systems Biology Markup Language (SBML) [30], a well-known machine-readable format based on XML for representing quantitative models of biological systems.

The BioModels repository is divided into two sections: the *curated branch* and the *non-curated branch*. The former contains models that have been manually checked and their components annotated using unambiguous identifiers [31] that refer to external biological databases [22,46,17] or ontologies (such as Gene Ontology [2], SBO [18] or ChEBI [20]). Models are curated following the Minimum Information Required in the Annotation of Models guidelines (MIRIAM) [35]. Models that are not MIRIAM-compliant are stored in the non-curated branch, which also contains non-kinetic models such as flux balance analysis models. A more detailed description of BioModels is available at [16].

3.2 Model conversion

We developed a prototype for translating SBML models into ERODE's format, using the workflow in Fig. 2. SBML files are read using the *jsbml* library (version 1.2) [36,21]. Here we briefly explain the main phases of the conversion process.

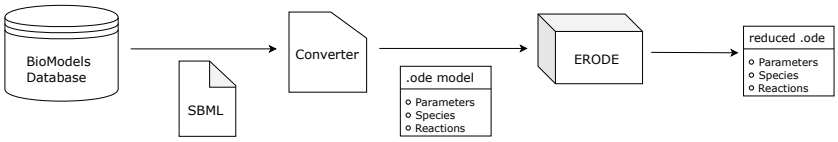


Fig. 2: Workflow overview. Models were downloaded from the BioModels repository in the SBML format. We implemented a tool to translate the SBML description into the CRN-like input (`.ode` format) of ERODE. The output of ERODE is a reduced CRN with reactions involving *macro-species*, each representing the sum within an equivalence class of original species. We manually inspected the ERODE output to provide a physical interpretation of the obtained equivalences.

The CRN input format of ERODE contains lists of parameters (to be used in kinetic rates), of species (with corresponding initial conditions), and of reactions. This is followed by a list of commands for analysis, reduction, and export.

The following SBML snippet, from *BIOMD0000000030*, specifies a parameter

```
<parameter id="k1" metaid="metaid_0000019" name="k1" value="0.02"/>
```

This is translated into `k1 = 0.02` within the parameters list (delimited by `begin parameters/end parameters`) of the ERODE description.

The next SBML snippet, adapted from the same model by removing the `annotation` tag containing links to external databases, defines the species M:

```
<species compartment="cell" id="M" initialConcentration="800"
        metaid="metaid_0000005" name="MAPK"/>
```

It describes the compartment in which the species is located, the initial concentration and an identifier. We translate this into `M = 800` within ERODE's species declaration section (delimited by `begin init/end init`).

Instead, the conversion of the reactions is less straightforward, particularly to recognize mass-action models to which the specialized FE, BE, and SMB can be applied. Indeed, SBML allows the direct specification of mass-action reactions by means of appropriate SBO labels in the `kineticLaw` tag (other labels identify different kinetics such as Michaelis-Menten and Hill). However, we encountered cases of reactions that, although not tagged with mass-action labels, were clearly so upon inspection of the reactions. One such example is given in Fig. 3. It shows the specification of a reaction containing a list of reactants, products (as well as modifiers, not used in this reaction, to model, e.g., catalysts or intermediates in the reaction). The `reaction` has an optional attribute `reversible`, by default set to `true`, indicating if the reaction is reversible. We inferred the forward and reverse rate functions as the left and right operand, respectively, of the topmost `minus` MathML tag (Line 16). This leads to the two following ERODE irreversible reactions (as ERODE does not support reversible reactions):

```
M + MAPKK -> M_MAPKK_Y, arbitrary cell *  k1 * M * MAPKK
M_MAPKK_Y -> M + MAPKK, arbitrary cell * k_1 * M_MAPKK_Y
```

```
<reaction id="reaction_0000001" metaid="metaid_0000046"
        name="binding MAPKK on Tyr site of MAPK">
  <listOfReactants>
    <speciesReference metaid="_063184" species="M"/>
    <speciesReference metaid="_063196" species="MAPKK"/>
   </listOfReactants>
   <listOfProducts>
     <speciesReference metaid="_063208" species="M_MAPKK_Y"/>
   </listOfProducts>
   <kineticLaw metaid="_063220">
     <math xmlns="http://www.w3.org/1998/Math/MathML">
      <apply>
        <times/>
          <ci>cell</ci>
            <apply>
              <minus/>
                <apply>
                  <times/><ci>k1</ci><ci>M</ci><ci>MAPKK</ci>
                </apply>
                <apply>
                  <times/><ci>k_1</ci><ci>M_MAPKK_Y</ci>
                </apply>
            </apply>
       </apply>
      </math>
    </kineticLaw>
</reaction>
```

Fig. 3: Sample SBML reaction adapted from *BIOMD0000000030*

Here, the left- and right-hand sides of the reactions are taken from the SMBL lists (and modifiers are added in both sides if present), whereas the `arbitrary` keyword denotes a reaction with a generic non-mass-action propensity function. However, one can notice that these two reactions are actually equivalent to mass-action reactions with kinetic parameters `cell * k1` and `cell * k_1`, respectively. We manually detected such occurrences of non-tagged mass-action reactions and translated into ERODE mass-action ones. In this example we get:

```
M + MAPKK -> M_MAPKK_Y,  cell *  k1
M_MAPKK_Y -> M + MAPKK,  cell * k_1
```

ERODEcan export the ODEs underlying a model as a Matlab function. Likewise, in BioModels all models come with an encoding as Matlab functions. We tested our converter over a large random selection of BioModels files by checking that their Matlab functions and those exported by ERODE corresponded.

3.3 Repository Preprocessing

In our experiments we used the BioModels repository snapshot 26 July 2017. It consists of 640 models in the curated branch (from id *BIOMD0000000001* to *BIOMD0000000640*) and 1000 models in the non-curated branch (with ids ranging from *MODEL0072364382* to *MODEL9811206584*).

We performed a preprocessing step to filter out models that could not be used for the analysis (cf. evaluation question **Q1** in Section 1). In the non-curated

branch only 491 models are kinetic models described as ODE systems, while the others are described in formalisms, such as logical or flux balance analysis models, that are outside the scope of applicability of species equivalences.

Overall, we could process 448 models from the curated branch and 219 from the non-curated one, for an overall sanitized dataset of 667 models. Of these, 43 were recognized as mass-action CRNs (as detailed in Section 3.2); all of them were found to be elementary mass-action CRNs, hence analyzable by SMB. The most frequent reasons for discarding a model were (within parenthesis we give the frequency in the curated branch, which we assume to be more stable):

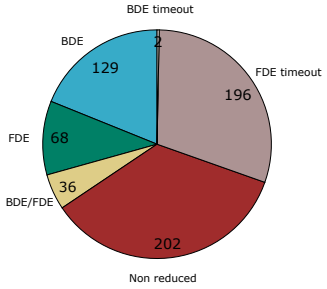
- syntactic limitations in our converter prototype, including the lack of support for models without explicit reactions where the dynamics is given by rate rules over a set of parameters, e.g., as in *BIOMD0000000020* (114);
- models with unsupported propensity functions such as tanh and exp (31);
- models with species with *Assignment Rules*, used to model features such as delayed equations and hybrid systems, not supported by ERODE (47).

3.4 Reduction results

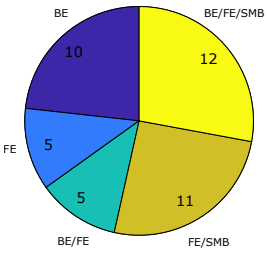
Here we report the summary of the reduction results. Non mass-action models were analyzed using FDE and BDE, while for mass-action ones we used FE, BE, and SMB. In a preliminary analysis we considered the maximal equivalences for all cases, computed by starting the partition-refinement algorithms with the initial singleton partition with a single block containing all species in the CRN. However, in 32 cases we found that the maximal FDE/FE collapsed *all* species and reactions. This is because these CRNs are *closed* and *mass-preserving*, meaning that the concentrations (represented by the ODE solutions for each species) just flow among the species, but the *total* cumulative concentration is constant. Therefore these systems can be self-consistently written as a single-equation ODE with zero derivative (and initial concentration equal to that total cumulative concentration). We dismissed such partitions as degenerate/uninteresting. Instead, for these cases we built more meaningful (ad-hoc) initial partitions to be used in the partition-refinement algorithm: we isolated variables of interest to the modeler, as evinced from the related scientific publication.

For each equivalence we computed the reduced CRN, recording the resulting number of species and reactions as a measure of the effectiveness of the exact reduction techniques (cf. **Q2** in Section 1). Figure 4 counts the models that could be reduced by at least one technique, regardless of the reduction ratio. For the non mass-action models (Fig. 4a), 233 models (37%) could be reduced. In particular, only 36 models could be reduced by both FDE and BDE, proving that they are not comparable. Several models (196, 31%) could not be analyzed due to the excessive computational cost of FDE, while only 2 due to BDE (we used a time-out of 8 hours). This is consistent with the more (and more complex) SMT checks required by FDE with respect to BDE [7].

All the mass-action models (Fig. 4b) could be reduced by at least one equivalence relation. Ten models (23%) could be reduced with BE and 5 (12%) with



(a) Non mass-action



(b) Mass-action

Fig. 4: Reduction results.

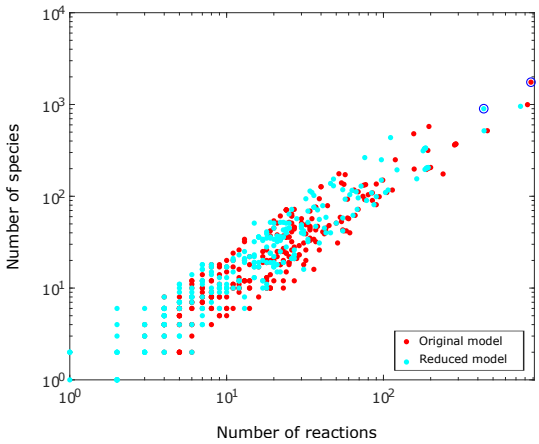
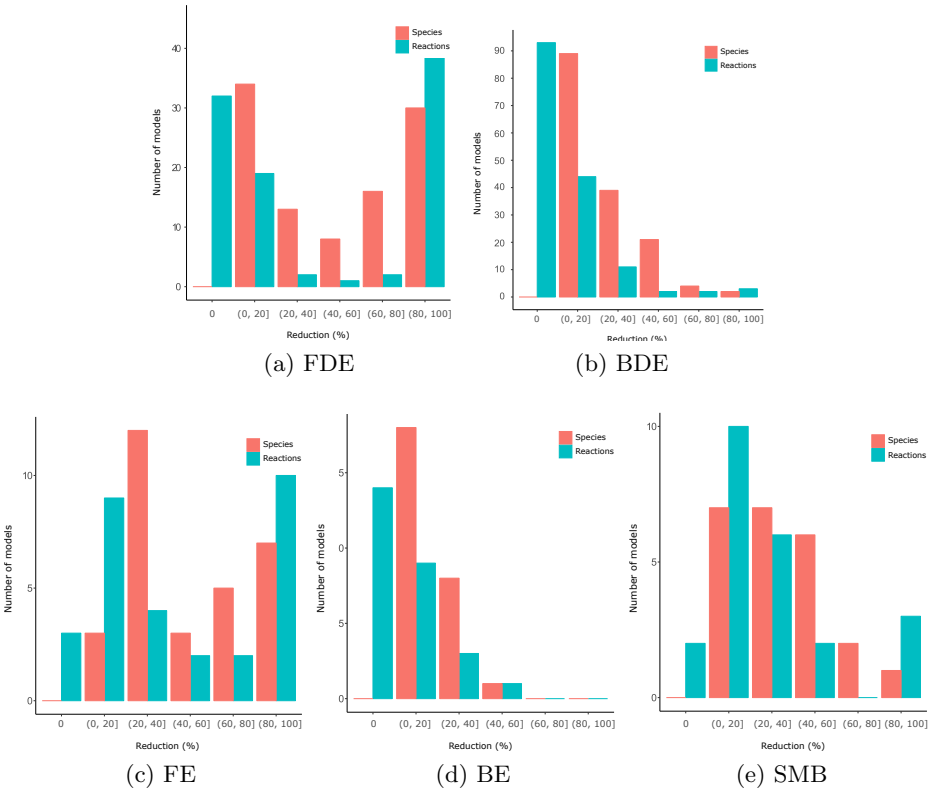


Fig. 5: Comparison among original and reduced species and reactions (log scale).

FE only. Twelve models (28%) could be reduced with the three methods, while 11 (25%) could be reduced with SMB and FE, and 5 (12%) with FE and BE. The presence of models that could be reduced only by FE and not SMB shows that FE does not imply SMB, while the converse is true, as discussed.

Figure 5 shows a scatter plot to summarize the reduction ratio for each model using the species equivalence that yielded the best reduction. *MODEL3632127506*, the largest model processed, denoted with blue circled dots in the figure, was reduced from 872 species and 1750 reactions to 436 species and 900 reactions, with a reduction of about 50% in the number of species and reactions. Overall, the average compression ratio is 36% for the species and 26% for the reactions.

The average reduction ratio in the number of species and reactions varies with each method: BDE (23% for species, 8% for reactions), FDE (50%, 48%), BE (19%, 8%), FE (51%, 47%), SMB (35%, 29%). Figure 6 illustrates the reductions obtained. For each species equivalence, we group the models in 5 histogram bins (0%-20%, . . . , 80%-100%) in two series showing the reduction ratio of the species (red) and the reactions (blue). It is possible to observe cases with models showing



(a) FDE (b) BDE

(c) FE (d) BE (e) SMB

Fig. 6: Reduction ratios for the species and reactions for each species equivalence.

no reductions in the number of reactions. This can be due to an equivalence among species with no dynamical role in the network, as they can be interpreted as distinct auxiliary species that are used to model zero-order reactions, such as I in reaction $I \rightarrow I + A$, a purely catalytic species C in a reaction like $A + C \rightarrow B + C$, or *SINK* in a degradation reaction such as $A \rightarrow SINK$. In the first two cases, these species are associated with zero-derivative variable, while in the last case the variable for *SINK* does not appear in any ODE in the system.

4 Case Studies

We hereby report selected case studies to highlight the physical interpretability of the reductions (cf. **Q3** in Section 1).

BE example: MAPK double phosphorylation. Multisite phosphorylation is a well-studied model in computational systems biology [29,47]. The double (de)phospho-rylation model depicted in Fig. 7 reflects the changes in the phosphorylated state of MAPK in *BIOMOD0000000030*. MAPK cascades are evolutionary conserved and consist of several (usually 3) levels, where the activated

(a) (b)

Fig. 7: (a) Mechanisms for the initial interaction of M and Mpp with MAPKK and MKP from [39]. Phosphorylation of M starts with the binding of MAPKK in either of terminus (T or Y) or M. Dephosphorylation occurs when MKP binds to an active molecule of M, in this case Mpp. (b) Reduced mechanism. BE equates the molecular complexes up to their their phosphorylated residue.

(a) (b)

Fig. 8: (a) Adaptation of the SPOC dynamical model from [15]. The SPB compartment is depicted in the yellow-circle background. Reactions crossing the compartment boundary represent the intrinsic Tem1 (blue rectangle) GTPase-cycle and reversible SPB association in terminal T. (b) Reduced mechanism where both FE and SMB equate all Tem1 molecules up to their GTP (green)- or GDP (red)-bound state (indicated by the green/red ellipsis).

kinase at each level phosphorylates the kinase at the next level down the cascade. MAPK (M) is a molecule with two residues: tyrosine (Y) and threonine (T), thus requires double phosphorylation from a MAPK Kinase to become active (Mpp), and double dephosphorylation from a MAPK phosphathase to return to its original inactive state. This dynamics is represented in a model with 18 species and 32 reactions. BE equates the MAPK complexes regardless of their binding with MAPK or MKP, yielding a reduced CRN with 16 species and 28 reactions.

FE example: SPOC. Model *BIOMOD0000000705* is a CRN of the Spindle Position Checkpoint (SPOC) [15]. SPOC intervenes in the process of cell division by verifying all requirements to pass to the next phase in the cell cycle. In particular, it prevents the separation of the duplicated chromosomes until each chromosome is properly attached to the spindle apparatus. The most upstream event of the pathway involves GTPase Tem1. Tem1 binds to the yeast centrosomes (called spindle pole bodies, SPBs) via GAP-dependent and GAP-independent sites (Fig. 8a). The intrinsic GTPase switching cycle of Tem1 is modeled as a reversible first-order reaction that converts $\mathrm{Tem}_1^{\mathrm{GTP}}$ into $\mathrm{Tem}_1^{\mathrm{GDP}}$ and vice versa. The model consists of 24 species and 71 reactions. FE equates the two forms of the GPTase Tem1 (Fig. 8), moreover this equivalence extends

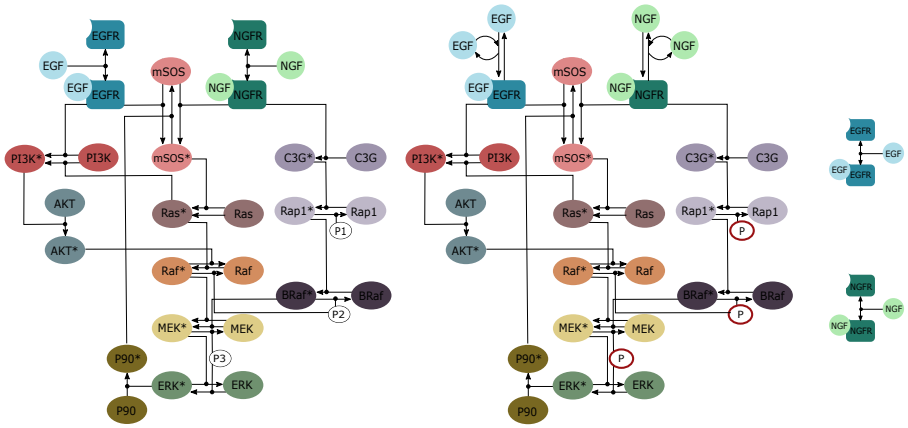
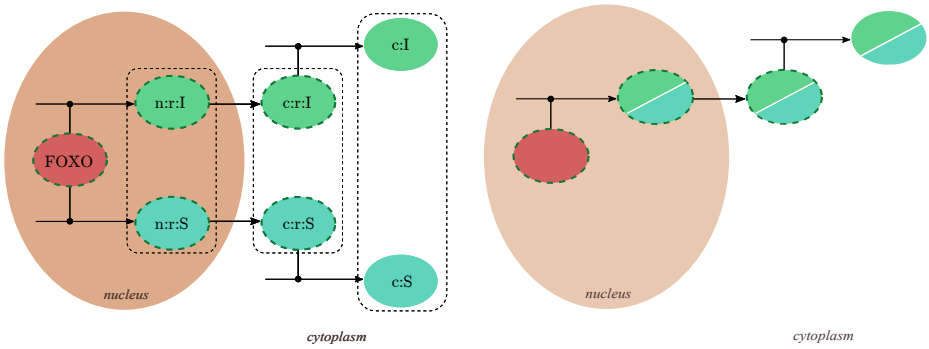


Fig. 9: (a) Adaptation of the signaling network in [4]. The activation of the molecular SOS by either of the receptors triggers the Ras cascade, concluding in ERK activation. EGF can also use the left branch involving PI3K to modulate Erk activity through Raf1 downregulation, and NGF can upregulate Mek using the right branch containing Rap1. P1,P2 and P3 represent unregulated phosphatases. Molecular components in (a) with the same color are grouped together in the same FDE equivalence class. (b) BDE reduction. (c) FDE reduction.

to all Tem1 molecular complexes, yielding a reduced model with 16 species and 36 reactions . In this example, the largest SMB yields the same reduction.

BDE/FDE example: Signaling cascade. Model *BIOMOD0000000033* is a signaling pathway concluding in ERK activation [4]. Its most upstream event (Fig. 9) starts with the binding of EGF and NGF to their respective receptors (EGFR, NGFR). Once bound, both receptors can activate molecular SOS and trigger the Ras cascade. Here, molecular components are modeled representing the species active and inactive state, i.e mSOS* and mSOS, yielding a model with 32 species and 26 reactions. For BDE, the free EGF and the free receptor EGFR are aggregated, simplifying the process of EGF binding to EGFR. Similarly, this occurs for NFG and NFGR. Finally, phosphatases P1, P2, and P3, whose role is purely catalytic, are aggregated (in a macro-species denoted by P). The BDE reduction has 27 species and 26 reactions (Fig. 9b). Instead, FDE collapses the active and inactive form of those species. Moreover, the dynamics of the active and inactive species sum up to zero if aggregated. As above, the phosphatases P1,P2, and P3 are aggregated in the same class. This results in the FDE reduction in Fig. 9c, with 18 species and 4 reactions.

SMB example: Proteins with same synthesis mechanism. As observed, our methods can help detecting symmetries among molecular components. We show this in *BIOMOD0000000705*, a FOXO-dependent synthesis mechanism in-



(a) (b)

Fig. 10: (a) Adaptation of the FOXO-dependent IsnR and Sod2 synthesis mechanism in [45]. Species labels are x:y:z , where x is the species compartment, y indicates binding with molecular RNA, and z is the first letter of the name of the protein, e.g., n:r:I encodes the nuclear RNA-bound IsnR, c:S encodes cytoplasmic Sod2. RNA-bound molecules are rounded by a dotted circle. SMB equivalences are represented by a dotted rectangle. (b) SMB/FE reduced mechanism.

volving IsnR and Sod2. Forkhead Box-type O (FOXO) is a family of transcription factors responsible for various biological processes including apoptosis, cell metabolism, differentiation, and drug resistance [34]. The model has 56 species and 135 reactions describing processes such as FOXO-dependent and basal transcription, export, translation, and degradation of RNA and proteins. The kinetic parameters for FOXO-dependent protein synthesis (Fig. 10a) for both IsnR and Sod2 are assumed to be equal. This gives an SMB reduction with 36 species and 110 reactions where IsnR and Sod2 molecules are aggregated in each step of the protein synthesis mechanism (Fig. 10b). FE leads to the same reduction.

5 Concluding Remarks

The empirical assessment of exact model reduction on the BioModels repository has provided a number of findings along the main evaluation questions **Q1**–**Q3** introduced in Section 1, which can be summarized as follows.

Q1. Assumptions for applicability of model reductions. In the preprocessing phase (Section 3.3), we found 300 models not supported by ERODE. Among the reasons for incompatibility it is worth commenting on the models which included exponential expressions in rate functions. This is not accepted by FDE/BDE because the underlying theory is not decidable. A workaround has been sketched in [10,13] and builds on a systematic technique which transforms an initial value problem for an ODE system with derivatives containing rational and exponential expressions into an equivalent problem with polynomial derivatives [28], to which BE and FE can be applied. In future work we plan to

implement such a transformation in order to extend the range of applicability of species equivalences. Instead, the limitation of SMB to elementary CRNs did not turn out to be practically impeding for the analysis of the BioModels repository, since all the CRNs were in this form; it is however theoretically interesting to extend the theory to non-elementary mass-action kinetics.

Q2. Effectiveness of the reductions. Overall, we found exact model reductions effective in terms of both the number of cases in which a CRN could be reduced by at least one technique (40%) and the overall compression ratio achieved on average (36% for number of species and 26% for the number of reactions). Unfortunately, the analysis of FDE on a rather appreciable number of models (196) was not conclusive due to timeouts, because of the relative complexity of the SMT checks that are required. This challenges the practical applicability of FDE to realistic case studies (BDE, on the other hand, timed out only twice in our tests whereas BE, FE, and SMB are supported by minimization algorithms that enjoy polynomial time and space complexity), prompting alternative approaches to computing FDE, for example by parallelizing the computations.

Q3. Physical interpretability. In the selected case studies herein presented, the exact model reductions have revealed that symmetries in certain signalling pathways carry over to equivalences at the level of the underlying quantitative semantics. Given their moderate size, the considered models would be computationally tractable even without reduction. However, the equivalences can be used as an aid in developing more complex models where such symmetries are present in some components. In addition, we remark that exact model reduction can still be useful when the complexity is due to the many repetitions that are required (e.g., for sensitivity analysis or for simulation with tight confidence intervals) or for particularly difficult analyses such as parametric inference [44].

Future work. This empirical study suggests potential benefits in the application of exact model reduction techniques in biological models from the literature. This motivates the development of our ERODE translator into a more mature tool to be further integrated with BioModels/SBML. The availability of ready-to-use model conversions in a simple CRN format such as ERODE's might stimulate similar assessments with other model reduction techniques (e.g., [3,12]).

In this paper we focused on reducing models with parameterizations given as in the respective original publications. If we wish to draw more general conclusions about the relevance of the reductions and the presence of certain symmetrical patterns in signaling pathways, it becomes important to test their *robustness* with respect to the model parameters. Theoretically, this does not seem to be particularly difficult, at least for CRNs with deterministic semantics. For example, model parameters could be interpreted as further variables in the SMT formulas used for checking FDE and BDE. Such an extension is currently not implemented in ERODE and is subject to the aforementioned caveats about the scalability of SMT-based reduction techniques, hence left for future work.

Acknowledgement. The authors are grateful to Andreas Dräguer (Institut für Informatik Zentrum für Bioinformatik Tübingen) for his support with JSBML.

References

1. Ascher, U.M., Petzold, L.R.: Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations. SIAM (1988)
2. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al.: Gene ontology: tool for the unification of biology. Nature genetics **25**(1), 25 (2000)
3. Boreale, M.: Algebra, coalgebra, and minimization in polynomial differential equations. In: 20th International Conference on Foundations of Software Science and Computation Structures (FOSSACS). pp. 71–87 (2017)
4. Brown, K.S., Hill, C.C., Calero, G.A., Myers, C.R., Lee, K.H., Sethna, J.P., Cerione, R.A.: The statistical mechanics of complex signaling networks: nerve growth factor signaling. Physical biology **1**(3), 184 (2004)
5. Buchholz, P.: Exact and Ordinary Lumpability in Finite Markov Chains. Journal of Applied Probability **31**(1), 59–75 (1994)
6. Cardelli, L., Tribastone, M., Tschaikowski, M., Vandin, A.: Efficient syntax-driven lumping of differential equations. In: TACAS. pp. 93–111 (2016)
7. Cardelli, L., Tribastone, M., Tschaikowski, M., Vandin, A.: Symbolic computation of differential equivalences. In: POPL. pp. 137–150 (2016). https://doi.org/10.1145/2837614.2837649
8. Cardelli, L., Tribastone, M., Tschaikowski, M., Vandin, A.: Forward and backward bisimulations for chemical reaction networks. In: 26th International Conference on Concurrency Theory, CONCUR. pp. 226–239 (2015). https://doi.org/10.4230/LIPIcs.CONCUR.2015.226
9. Cardelli, L., Tribastone, M., Tschaikowski, M., Vandin, A.: Erode: A tool for the evaluation and reduction of ordinary differential equations. In: International Conference on Tools and Algorithms for the Construction and Analysis of Systems. pp. 310–328. Springer (2017)
10. Cardelli, L., Tribastone, M., Tschaikowski, M., Vandin, A.: Maximal aggregation of polynomial dynamical systems. Proceedings of the National Academy of Sciences **114**(38), 10029–10034 (2017)
11. Cardelli, L., Tribastone, M., Tschaikowski, M., Vandin, A.: Syntactic Markovian bisimulation for chemical reaction networks. In: Models, Algorithms, Logics and Tools, pp. 466–483. Springer (2017)
12. Cardelli, L., Tribastone, M., Tschaikowski, M., Vandin, A.: Guaranteed error bounds on approximate model abstractions through reachability analysis. In: 15th International Conference on Quantitative Evaluation of Systems (QEST) (2018)
13. Cardelli, L., Tribastone, M., Tschaikowski, M., Vandin, A.: Symbolic computation of differential equivalences. Theoretical Computer Science (2019)
14. Caspi, R., Altman, T., Billington, R., Dreher, K., Foerster, H., Fulcher, C.A., Holland, T.A., Keseler, I.M., Kothari, A., Kubo, A., et al.: The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. Nucleic acids research **42**(D1), D459–D471 (2013)
15. Caydasi, A.K., Lohel, M., Grünert, G., Dittrich, P., Pereira, G., Ibrahim, B.: A dynamical model of the spindle position checkpoint. Molecular systems biology **8**(1), 582 (2012)
16. Chelliah, V., Laibe, C., Novère, N.L.: Biomodels database: a repository of mathematical models of biological processes. Encyclopedia of Systems Biology pp. 134–138 (2013)

17. Consortium, U.: Uniprot: a hub for protein information. Nucleic acids research **43**(D1), D204–D212 (2014)
18. Courtot, M., Juty, N., Knüpfer, C., Waltemath, D., Zhukova, A., Dräger, A., Dumontier, M., Finney, A., Golebiewski, M., Hastings, J., et al.: Controlled vocabularies and semantics in systems biology. Molecular systems biology **7**(1), 543 (2011)
19. De Moura, L., Bjørner, N.: Z3: An efficient SMT solver. In: TACAS. pp. 337–340 (2008)
20. Degtyarenko, K., De Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcántara, R., Darsow, M., Guedj, M., Ashburner, M.: Chebi: a database and ontology for chemical entities of biological interest. Nucleic acids research **36**(suppl_1), D344–D350 (2007)
21. Dräger, A., Rodriguez, N., Dumousseau, M., Dörr, A., Wrzodek, C., Le Novère, N., Zell, A., Hucka, M.: JSBML: a flexible Java library for working with SBML. Bioinformatics **27**(15), 2167–2168 (06 2011). https://doi.org/10.1093/bioinformatics/btr361
22. Federhen, S.: The ncbi taxonomy database. Nucleic acids research **40**(D1), D136–D143 (2011)
23. Feret, J., Henzinger, T., Koeppl, H., Petrov, T.: Lumpability abstractions of rule-based systems. Theoretical Computer Science **431**, 137–164 (2012)
24. Feret, J., Danos, V., Krivine, J., Harmer, R., Fontana, W.: Internal coarse-graining of molecular systems. Proceedings of the National Academy of Sciences **106**(16), 6453–6458 (2009). https://doi.org/10.1073/pnas.0809908106
25. Ganguly, A., Petrov, T., Koeppl, H.: Markov chain aggregation and its applications to combinatorial reaction networks. Journal of Mathematical Biology **69**(3), 767–797 (Sep 2014). https://doi.org/10.1007/s00285-013-0738-7, `https://doi.org/10.1007/s00285-013-0738-7`
26. Gillespie, D.T.: Stochastic simulation of chemical kinetics. Annual Review of Physical Chemistry **58**(1), 35–55 (2007)
27. Gillespie, D.: Exact stochastic simulation of coupled chemical reactions. Journal of Physical Chemistry **81**(25), 2340–2361 (December 1977)
28. Gu, C.: QLMOR: A Projection-Based Nonlinear Model Order Reduction Approach Using Quadratic-Linear Representation of Nonlinear Systems. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems **30**(9), 1307–1320 (2011). https://doi.org/10.1109/TCAD.2011.2142184
29. Gunawardena, J.: Multisite protein phosphorylation makes a good threshold but can be a poor switch. Proceedings of the National Academy of Sciences of the United States of America **102**(41), 14617–14622 (2005). https://doi.org/10.1073/pnas.0507322102
30. Hucka, M., Finney, A., Sauro, H.M., Bolouri, H., Doyle, J.C., Kitano, H., Arkin, A.P., Bornstein, B.J., Bray, D., Cornish-Bowden, A., et al.: The systems biology markup language (sbml): a medium for representation and exchange of biochemical network models. Bioinformatics **19**(4), 524–531 (2003)
31. Juty, N., Le Novere, N., Laibe, C.: Identifiers. org and miriam registry: community resources to provide persistent identification. Nucleic acids research **40**(D1), D580–D586 (2011)
32. Kanehisa, M., Goto, S.: Kegg: kyoto encyclopedia of genes and genomes. Nucleic acids research **28**(1), 27–30 (2000)
33. Kurtz, T.G.: The Relationship between Stochastic and Deterministic Models for Chemical Reactions. Journal of Chemical Physics **57**(7) (1972)

34. Lam, E.W.F., Brosens, J.J., Gomes, A.R., Koo, C.Y.: Forkhead box proteins: tuning forks for transcriptional harmony. Nature Reviews Cancer **13**, 482 EP – (06 2013)
35. Le Novère, N., Finney, A., Hucka, M., Bhalla, U.S., Campagne, F., Collado-Vides, J., Crampin, E.J., Halstead, M., Klipp, E., Mendes, P., et al.: Minimum information requested in the annotation of biochemical models (miriam). Nature biotechnology **23**(12), 1509 (2005)
36. Le Novère, N., Rodriguez, N., Wrzodek, F., Mittag, F., Fröhlich, S., Hucka, M., Thomas, A., Palsson, B.Ø., Lewis, N.E., Dräger, A., Myers, C.J., Watanabe, L., Vazirabad, I.Y., Kofia, V., Gómez, H.F., Diamantikos, A., Netz, E., Matthes, J., Eichner, J., Keller, R., Rudolph, J., Wrzodek, C.: JSBML 1.0: providing a smorgasbord of options to encode systems biology models. Bioinformatics **31**(20), 3383–3386 (06 2015). https://doi.org/10.1093/bioinformatics/btv341
37. Li, C., Donizelli, M., Rodriguez, N., Dharuri, H., Endler, L., Chelliah, V., Li, L., He, E., Henry, A., Stefan, M.I., Snoep, J.L., Hucka, M., Le Novère, N., Laibe, C.: BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models. BMC Systems Biology **4**, 92 (Jun 2010)
38. Li, G., Rabitz, H.: A general analysis of exact lumping in chemical kinetics. Chemical Engineering Science **44**(6), 1413 – 1430 (1989). https://doi.org/http://dx.doi.org/10.1016/0009-2509(89)85014-6
39. Markevich, N.I., Hoek, J.B., Kholodenko, B.N.: Signaling switches and bistability arising from multisite phosphorylation in protein kinase cascades. The Journal of cell biology **164**(3), 353–359 (2004)
40. Nishimura, D.: Biocarta. Biotech Software & Internet Report: The Computer Software Journal for Scient **2**(3), 117–120 (2001)
41. Paige, R., Tarjan, R.: Three partition refinement algorithms. SIAM Journal on Computing **16**(6), 973–989 (1987). https://doi.org/10.1137/0216062
42. Pratt, D., Chen, J., Welker, D., Rivas, R., Pillich, R., Rynkov, V., Ono, K., Miello, C., Hicks, L., Szalma, S., et al.: Ndex, the network data exchange. Cell systems **1**(4), 302–305 (2015)
43. Radulescu, O., Gorban, A.N., Zinovyev, A., Noel, V.: Reduction of dynamical biochemical reactions networks in computational biology. Frontiers in Genetics **3**(131) (2012). https://doi.org/10.3389/fgene.2012.00131
44. Schnoerr, D., Sanguinetti, G., Grima, R.: Approximation and inference methods for stochastic biochemical kinetics—a tutorial review. Journal of Physics A: Mathematical and Theoretical **50**(9), 093001 (2017)
45. Smith, G.R., Shanley, D.P.: Modelling the response of FOXO transcription factors to multiple post-translational modifications made by ageing-related signalling pathways. PLOS One **5**(6), e11092 (2010)
46. Stoesser, G., Baker, W., van den Broek, A., Camon, E., Garcia-Pastor, M., Kanz, C., Kulikova, T., Leinonen, R., Lin, Q., Lombard, V., et al.: The embl nucleotide sequence database. Nucleic acids research **30**(1), 21–26 (2002)
47. Thomson, M., Gunawardena, J.: Unlimited multistability in multisite phosphorylation systems. Nature **460**(7252), 274–277 (07 2009), `http://dx.doi.org/10.1038/nature08102`
48. Tribastone, M., Vandin, A.: Speeding up stochastic and deterministic simulation by aggregation: an advanced tutorial. In: 2018 Winter Simulation Conference, WSC 2018, Gothenburg, Sweden, December 9-12, 2018. pp. 336–350 (2018). https://doi.org/10.1109/WSC.2018.8632364
49. Turanyi, T., Tomlin, A.S.: Analysis of Kinetic Reaction Mechanisms. Springer (2014)

50. Vandin, A., Tribastone, M.: Quantitative abstractions for collective adaptive systems. In: SFM 2016, Bertinoro Summer School. pp. 202–232 (2016). https://doi.org/10.1007/978-3-319-34096-8_7
51. Voit, E.O.: Biochemical systems theory: A review. ISRN Biomathematics **2013**, 53 (2013), http://dx.doi.org/10.1155/2013/897658%]897658
52. Wittig, U., Kania, R., Golebiewski, M., Rey, M., Shi, L., Jong, L., Algaa, E., Weidemann, A., Sauer-Danzwith, H., Mir, S., et al.: Sabio-rk—database for biochemical reaction kinetics. Nucleic acids research **40**(D1), D790–D796 (2011)