

# Modeling signal transduction networks: A comparison of two stochastic kinetic simulation algorithms

Michel F. Pettigrew<sup>a)</sup>

*Cell Systems Initiative, Department of Bioengineering, University of Washington, P.O. Box 358070, 960 Republican Street, Seattle, Washington 98195-8070*

Haluk Resat<sup>b)</sup>

*Computational Biology and Bioinformatics Group, Pacific Northwest National Laboratory, P.O. Box 999, MS: K7-90, Richland, Washington 99352*

(Received 21 October 2004; accepted 13 July 2005; published online 20 September 2005)

Computational efficiency of stochastic kinetic algorithms depend on factors such as the overall species population, the total number of reactions, and the average number of nodal interactions or connectivity in a network. These size measures of the network model can have a significant impact on computational efficiency. In this study, two scalable biological networks are used to compare the size scaling efficiencies of two popular and conceptually distinct stochastic kinetic simulation algorithms—the random substrate method of Firth and Bray (FB), and the Gillespie algorithm as implemented using the Gibson-Bruck method (GGB). The arithmetic computational efficiencies of these two algorithms, respectively, scale with the square of the total species population and the logarithm of the total number of active reactions. The two scalable models considered are the size scalable model (SSM), a four compartment reaction model for a signal transduction network involving receptors with single phosphorylation binding sites, and the variable connectivity model (VCM), a single compartment model where receptors possess multiple phosphorylation binding sites. The SSM has fixed species connectivity while the connectivity between species in VCM increases with the number of phosphorylation sites. For SSM, we find that, as the total species population is increased over four orders of magnitude, the GGB algorithm performs significantly better than FB for all three SSM compartment models considered. In contrast, for VCM, we find that as the overall species population decreases while the number of phosphorylation sites increases (implying an increase in network linkage) there exists a crossover point where the computational demands of the GGB method exceed that of the FB. © 2005 American Institute of Physics. [DOI: [10.1063/1.2018641](https://doi.org/10.1063/1.2018641)]

## I. INTRODUCTION

Conventional formulations of reaction kinetic problems have traditionally been defined using macroscopic quantities that are based on species concentrations. This viewpoint is founded on the assumption of a sufficiently large number of molecules in a finite volume, where concentration as a well-defined quantity varies continuously over time. In such cases, the evolution in time of concentrations in a well-stirred reacting mixture can usually be described by a set of kinetic rate equations. While the macroscopic formulation has been successfully employed in a wide variety of important problems, the assumptions on which it is based break down for many dynamical systems in cell biology. Therefore, it has been advocated that the mesoscopic view,<sup>1</sup> in which chemical species appear in small copy numbers, is a more appropriate formulation for dynamical systems in biology in that changes in species populations are discrete and occur as a consequence of stochastic single reaction events.<sup>2</sup> Spatially averaged mesoscopic approaches, too, assume that the

chemically reacting mixture is well stirred and a thermodynamic equilibrium is maintained. It is with such stochastic kinetic systems that we are concerned in this study.

Dynamic biological systems are typically characterized by a large set of species types, some of which can exist at very small copy numbers. Another common occurrence for receptor signaling systems is the existence of many different forms of the receptor, such as the different phosphorylation states of the receptor or the different complexes that receptors form with their ligands or with other receptors or adaptor proteins.<sup>3</sup> As the number of species types increases, the number of possible interactions in the system increases proportionally too. This results in a large set of multistate interactions between complex types and makes tracking the distribution of the species a difficult task. Although it has been suggested that, for dynamical systems with these characteristics, the random substrate method of Firth and Bray<sup>4</sup> (FB) could be more efficient than the more conventional direct Gillespie algorithm,<sup>5,6</sup> how the increase in size of the kinetic models affects the numerical performance of the stochastic simulation algorithms has yet to be addressed. In this study, we have devised two scalable signal transduction reaction networks to compare and assess the computational perfor-

<sup>a)</sup>Electronic mail: [mpettigr@u.washington.edu](mailto:mpettigr@u.washington.edu)

<sup>b)</sup>Author to whom correspondence should be addressed. FAX: 509-372-4720. Electronic mail: [haluk.resat@pnl.gov](mailto:haluk.resat@pnl.gov)

mance of the FB algorithm with a well-known Gillespie variant—the Gibson-Bruck method<sup>7</sup> (GGB) through numerical experiments. Since large data sets obtained in high-throughput systems biology experiments are now making the construction of larger and larger biological networks feasible, knowing how the kinetic algorithms scale with network size will be important in choosing the right algorithm.

We base both of our scalable models on the epidermal growth factor receptor (EGFR) system which regulates cell proliferation and differentiation. After ligand activation, the EGFR is rapidly internalized by endocytosis.<sup>8,9</sup> Following endocytosis, receptor ligand and other receptor-bound complexes are sorted into different cellular compartments with distinct properties.<sup>2,10</sup> Receptor endocytosis and the resulting receptor trafficking within cells is a way for the receptors to be exchanged between cellular compartments that eventually leads to receptor deactivation and degradation. A similar receptor trafficking also occurs for the *G*-protein-coupled receptor as well as other receptor systems too.<sup>2</sup> More details on the network models for EGFR signaling may be found in the recent papers.<sup>11–15</sup>

## II. KINETIC MODELS AND COMPUTATIONAL METHODS

We have constructed two scalable models for benchmarking purposes in comparison studies. The first model, the size scalable model (SSM), is a scalable four compartment model that is typical for a receptor tyrosine kinase signal transduction network. In the SSM, the state of the receptor is described by phosphorylation at a single amino acid site, and the receptor becomes active in signal transduction upon phosphorylation. A subcompartmentalization process allows for the creation of refined SSM models where the size of the model increases in proportion to the number of created subcompartments while maintaining the connectivity between species almost unchanged. The SSM model is discussed at length in Sec. II A and in Appendix B. The second model, the variable connectivity model (VCM), is a scalable single compartment signal transduction network in which receptors have multiple phosphorylation sites so that the linkage of the network model can be altered by changing the number of phosphorylation sites. The VCM model is discussed in Sec. II B.

### A. Formulation of the SSM

#### 1. Domain and complex characterization

To facilitate an evaluation of how stochastic algorithms scale with network size, we employ a subcompartmentalization process to multiply the number of reactions included in the kinetic models. The system volume  $V$  in all simulations of the SSM model described in this paper is first divided into four major domains (i.e., distinct parent compartments), where particle exchange between the compartments is allowed. To simplify the numerical problem, we lower the dimensionality and convert the three-dimensional (3D) problem into a two-dimensional (2D) lattice problem. Figure 1 shows a representation of the abstract construct that we use in the reported simulations. We note that although the geom-

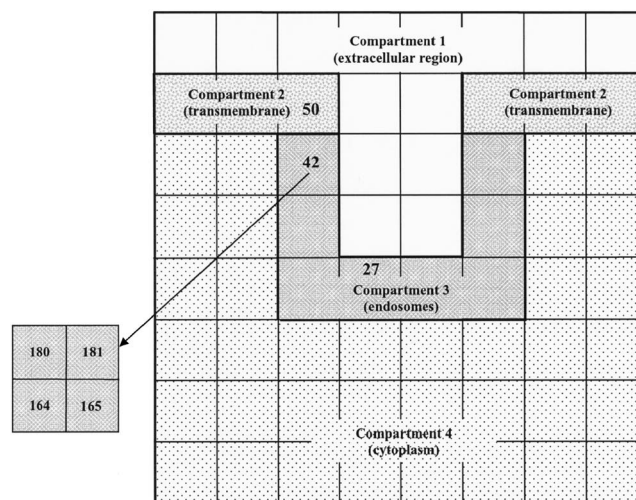


FIG. 1. The four compartment signal transduction model: SSM 1 (64 subcompartments). Expansion on the left shows the subcompartments of the SSM 2 that correspond to unit 42 of the first model.

etry of the fictitious construct changes, this reduction in the spatial dimension and reflecting the biological problem onto a lattice preserves the underlying true nature of the investigated biological system and has no effect on the fundamentals of the algorithm efficiency issue that we are addressing in this study. To keep the connection with the biological signal transduction networks, we label the major compartments of the model as (1) extracellular region, (2) cell plasma membrane, (3) intravesicular endosomes, and (4) cytoplasm of a mammalian cell.

#### 2. Subcompartmentalization

We create different versions of our four compartment model through the process of further partitioning as a way to facilitate the evaluation of numerical efficiency on network size. Our approach effectively is equivalent to creating a locally homogeneous model to include the spatial inhomogeneities where the coarseness of the model depends on the number of used subcompartments. Figure 1 illustrates the 64 subcompartment model, which is created by dividing the simulated system into 64 equal size units. This model will be referred to as SSM model 1 in the remainder of this report. For the employed square 2D setup, SSM model 1 is constructed by simply dividing the  $x$  and  $y$  edges into eight equally spaced sections, i.e.,  $8 \times 8 = 64$  subcompartments, while preserving the four major compartments intact. We again note that cells are obviously three dimensional but since our main aim is to compare the computational efficiency of stochastic algorithms, we opted to characterize the cell as a fictitious object that can be characterized as a two-dimensional grid. Generalization of our approach to three dimensions is straightforward; however, three dimensional representation would increase the complexity of the problem with no obvious benefits to our purpose. Therefore, a two-dimensional model was preferred in this study.

In addition to the 64 subcompartment model, we have created two more models with different sizes by partitioning the cell into even smaller units. These additional models con-

TABLE I. Size description of the SSM model.

Size model	Number of subcompartments	Number of complex types	Number of reactions
1	64	142	518
2	256	568	2268
3	1024	2272	9464

tain 256 and 1024 subcompartments, and will be labeled SSM model 2 and SSM model 3, respectively. Table I compares the sizes of the created models. As it is clear from Table I, the number of reactions included in the network models increases considerably in going from a 64 to 1024 subcompartment representation. It should also be noted that further partitioning the system using equal volume subcompartment units preserve the volume ratio of the different major cellular compartments, and hence, makes the different-sized models compatible with the parent model.

### 3. Indexing and particle tracking

For the system subdivided into  $N^2$  subcompartments, starting at the lower left-hand unit, the subcompartment units are labeled from 0 to  $N^2 - 1$  for unique identification. All molecular complexes are labeled with reference to the subcompartment that they reside in, and a molecular species of type  $S$  in subcompartment  $K$  of major compartment  $m$  is denoted by  $S_{m_K}$ . The population count of each species in each subcompartment is updated when internal or mass transfer reactions occur in that subcompartment. It is important to note that there is only one independent volume parameter in our multicompartment SSM, which is the volume  $V$  of the container. Volumes of subcompartments are simply fixed fractions of the total volume. Therefore, when the particles of a certain species are distributed among the compartments of the system in proportion to compartment volumes, the molar concentration of the species stays uniform between the subcompartments.

### 4. Models with different number of molecules

To investigate how the efficiency of the studied algorithms depends on the number of molecules (molecular copy numbers) in the system, we have created four comparable population models (A–D) for every subcompartment model by adjusting the system's total volume. When the volume is decreased, it is assumed that the number of molecules decreases by the same factor on average so the molar concentrations do not change. Although changing the cellular volume is not biologically sensible, this is a numerical trick that

makes it possible to modify the copy numbers of molecules in the computations without changing the molecular concentrations. Since the mean results for the concentrations of the molecules will be the same, this trick allows us to compute the computational requirements of the investigated algorithms as a function of the number of molecules in the system.

Table II reports the initial number of molecules for all species types in population model A. Ligand and adaptor molecules are initially placed in the extracellular (first major compartment) and cytoplasm (fourth) compartments, respectively. Receptors initially exist in their free and unphosphorylated ( $R$ ) form and are distributed between plasma membrane (90%) and endosomes (10%), Table II. Using a typical volume for the eukaryotic cells, in the population model A, we assume that a concentration of 100 pM equals to 180 molecules in the total cellular volume. For the population model B, the volume is reduced with a factor of 10 and 100 pM concentration corresponding to 18 molecules. Similarly, in population models C and D the volume is further reduced by factors of 10 and 100, respectively. With four different populations for each of the three different size networks, the studied SSM reaction system can thus be simulated in twelve different ways at different scales. In the remainder of the paper, the models will be referred to as SSM model  $XY$ , where  $X(=1-3)$  and  $Y(=A-D)$ , respectively, represent the network size (i.e., the number of subcompartments) and population (i.e., size in terms of number of molecules) of the referred model.

### 5. Boundary conditions

In order to keep simulations as simple as possible, a no-flux boundary condition on all complexes is assumed for subcompartments adjacent to the outer cell boundary.

### 6. Signal transduction aspects of the size scalable model

Major processes included in the SSM network model are unimolecular and bimolecular reactions and particle diffusion or exchange between the subcompartments. Reactions defined within a major compartment are defined per subcompartment basis while diffusion is considered as a first-order mass transfer between subcompartments.

The multicompartment signal transduction network used in this report contains a set of reactions between eight molecular species, which are listed in Table III. Most of the species are receptor complexes and we note that the molecular species that involve the receptor (i.e.,  $R$ ,  $RL$ ,  $RP$ ,  $RLP$ ,  $RPA$ , and  $RLPA$ ) will collectively be referred to as receptor

TABLE II. Number of molecules in the SSM population A model. These molecule copy numbers were used as the initial conditions at the start of the simulation runs.

Compartment	$L$	$R$	$A$	$RP$	$RL$	$RLP$	$RPA$	$RLPA$
1	36 000	0	0	0	0	0	0	0
2	0	16 200	0	0	0	0	0	0
3	0	1800	0	0	0	0	0	0
4	0	0	45 000	0	0	0	0	0



TABLE III. Molecular species included in the SSM signal transduction network.

Symbol	Description
$L$	Ligand (for example, growth factor or major histocompatibility complex)
$R$	Free receptor (for example, receptor tyrosine kinases or $T$ -cell receptors)
$A$	Adaptor protein
$RL$	Ligand-bound receptor, i.e., ligand:receptor complex
$RP$	Phosphorylated free receptor
$RLP$	Phosphorylated receptor in $RL$ complex
$RPA$	Phosphorylated receptor $RP$ complexed with adaptor protein $A$
$RLPA$	$RLP$ complexed with adaptor protein $A$

complexes in the remainder of this paper. The signal transduction network of SSM consists of two unimolecular and five bimolecular reversible reactions with mass transfer between compartments. The reactions constituting the model are tabulated in Table IV. Reversible reactions 1, 3, and 5 represent ligand ( $L$ )-receptor complex association. Reversible reactions 2 and 4 represent phosphorylation and dephosphorylation of receptors. Reactions 6 and 7 represent the recruitment and dissociation of adaptor proteins ( $A$ ) to phosphorylated receptors. By associating with the activated receptors, adaptor proteins [for example, the Grb2, Shc, Ras, Raf, and mitogen-activated protein kinase (MAPK) cascade] are instrumental in transmitting the cell signal initiated by external stimuli to the interior of the cells. This is the rationale for taking the number of receptor-adaptor protein complexes as the measure of the amplitude of the cell signal within the SSM model.

Further details of the SSM model and the list of used reaction rates can be found in Appendix B.

## B. Formulation of the VCM

Our second model, the variable connectivity model (VCM), is a scalable single compartment signal transduction network involving two unimolecular and two bimolecular reversible reactions for 3 species—ligand, receptor, and ligand-receptor complexes. The species and the reactions with rates for the VCM model are given in Tables V and VI, respectively. The receptors in this model may have multiple phosphorylation binding sites where ligand-bound receptors

TABLE IV. Reactions defining the SSM model.

No.	Reaction	Number	Reaction
1	$L + R \xrightleftharpoons[k_{-1}]{k_1} RL$	5	$RPA + L \xrightleftharpoons[k_{-5}]{k_5} RLPA$
2	$RL \xrightleftharpoons[k_{-2}]{k_2} RLP$	6	$RP + A \xrightleftharpoons[k_{-6}]{k_6} RPA$
3	$RP + L \xrightleftharpoons[k_{-3}]{k_3} RLP$	7	$RLP + A \xrightleftharpoons[k_{-7}]{k_7} RLPA$
4	$R \xrightleftharpoons[k_{-4}]{k_4} RP$		

TABLE V. Molecular species included in the simplified signal transduction VCM network.

Symbol	Description
$L$	Ligand
$R$	Free receptor with multiple phosphorylation binding sites
$RL$	Ligand-bound receptor, i.e., ligand:receptor complex
$RP$	Phosphorylated free receptor (phosphorylated at one or more sites)
$RLP$	Phosphorylated receptor in $RL$ complex

are phosphorylated at a much higher rate than the ligand-free receptors. We choose the simplest realistic assumptions concerning phosphorylation in our VCM with  $N$  phosphorylation binding sites: (a) any one of the  $N$  phosphorylation sites may be phosphorylated or dephosphorylated with equal probability, i.e., the rates of phosphorylation reactions are the same for every site; and (b) phosphorylation state of a site does not affect the (de)phosphorylation rates of other sites, i.e., possible cooperativity effects are neglected in our VCM.

To trace the phosphorylation status of the receptors we use the multistate notation  $R_{\{F_1 \cdots F_N\}}$  where the flags  $F_j$  shows the state (on/off) of the  $j$ th phosphorylation site. For example,  $R_{00001000}$  in a VCM would represent that the receptors have eight phosphorylation sites and this particular species has the fifth site phosphorylated while the other sites are unphosphorylated. In our VCM model each multistate form is treated as a distinct species. Therefore, according to the reaction scheme for a VCM with  $N$  phosphorylation sites (Table VI), each multistate is connected to  $N$  other multistates as well as to ligand  $L$  and to the complex  $LR$ . Clearly the linkage between the species increases in proportion to the number of phosphorylation sites.

## C. Stochastic algorithms for simulation of model problems

In this paper, we report our results for the dependence of the numerical efficiency of two stochastic kinetic simulation algorithms on the size and connectivity of the investigated model network. We particularly investigate the size dependence on the number of reactions included in the simulations and on the total number of molecules in the system. We chose a variant of the highly popular direct Gillespie algorithm,<sup>5,6</sup> the Gibson-Bruck method,<sup>7</sup> as one of the two stochastic algorithms that we utilize. This will be referred as

TABLE VI. Reactions defining the VCM model.

No.	Reaction	Rate constants
1	$R \xrightleftharpoons[k_{-1}]{k_1} RP$	$k_1 = 10^{-4}$ , $k_{-1} = 10^{-3}$
2	$RL \xrightleftharpoons[k_{-2}]{k_2} RLP$	$k_2 = 4 \times 10^{-3}$ , $k_{-2} = 10^{-3}$
3	$R + L \xrightleftharpoons[k_{-3}]{k_3} RL$	$k_3 = 5 \times 10^1$ , $k_{-3} = 10^{-3}$
4	$RP + L \xrightleftharpoons[k_{-4}]{k_4} RLP$	$k_4 = 5 \times 10^1$ , $k_{-4} = 10^{-3}$

the GGB algorithm. It has been suggested that, for dynamical systems where a large set of multistate interactions exist between molecular complex types, the inexact Firth and Bray (FB) algorithm<sup>4</sup> can be a very efficient method especially when the total complex population is small. The reactions defining the signal transduction VCM model studied here are dominated by the interactions of various receptor complex types, which correspond to the interactions of various receptor multistate forms. For this reason, we chose the FB algorithm as the second method in this study.

Major aspects of the FB and GGB algorithms are summarized in Appendix A. In conducting our numerical experiments, both the FB algorithm (as documented and coded in the C++ StochSim simulation software developed by Morton-Firth in Ref. 4) and the GGB algorithm were efficiently implemented in FORTRAN 95 in SIGTRAN by Cell Systems Initiative ([www.csi.washington.edu](http://www.csi.washington.edu)). SIGTRAN is a publicly available deterministic and stochastic simulation software package for kinetic simulations and it supports a number of algorithms including the stochastic FB and GGB algorithms. It is clear that the process of specifying the reaction networks for SSM models with 64, 256, and 1024 compartments require automation. To achieve this, a FORTRAN 95 program CMPTPARSER was created. The input to this program are various simulation parameters such as the simulation duration and the number of trajectories, the eight basic complex types, and the initial populations for each species type as well as the basic reaction and mass transfer set for each of the major compartments of the model. Using the input information together with the total number  $N^2$  of subcompartments, CMPTPARSER program creates the simulation, complex, and reaction input files for the SIGTRAN program. For each major compartment of the SSM model, CMPTPARSER uniformly distributed the species populations amongst the subcompartments. A FORTRAN 95 program SIMPLEPARSER, similar to CMPTPARSER, was created to generate the input files for the VCM model to uniformly distribute each multistate species population amongst the species states. All computations were carried out on a 2.0-GHz Dell Pentium(R) 4 with 512-Mbytes random access memory (RAM) running Lahey/Fujitsu FORTRAN LF 95 v.5.6 Pro.

We note that the FB and GGB algorithms were chosen for this study only because they are two *distinctly* different approaches to the stochastic simulation of biochemical reaction systems. Our aim was not to provide computational support for the most efficient algorithm—the answer to this question clearly depends on the network size and connectivity as well as on the implementation of these algorithms for a given computer architecture. Our main purpose was to investigate the scaling characteristics of these two distinct classes of algorithms. We also would like to point out that GGB algorithm may not be the most efficient variant of Gillespie-type approaches but it was chosen only because it is widely used in stochastic simulation studies.

### III. RESULTS AND DISCUSSION

#### A. Size scalable model

Table VII summarizes the simulations for different subcompartment and population size models that were per-

TABLE VII. Summary of the simulations performed with the Firth-Bray (FB) and Gillespie-Gibson-Bruck (GGB) algorithms for the SSM model.

SSM model	Population model A	Population model B	Population model C	Population model D
1 (64)	GGB	GGB	FB, GGB	FB, GGB
2 (256)	GGB	GGB	FB, <sup>a</sup> GGB	FB, GGB
3 (1024)	GGB	GGB	FB, <sup>a</sup> GGB	FB, GGB

<sup>a</sup>FB—only five trajectories completed.

formed to investigate the scaling properties of the FB and GGB algorithms. As will be discussed in detail below, simulations employing the FB algorithm are quite time consuming so the FB algorithm was used in fewer models than the GGB algorithm. For each model, each simulation consisted of an ensemble of 50 distinct trajectories run for 8000 s. Results reported in the figures below are the averages of the 50 stochastic simulation trajectories.

To verify our scalable model, we first analyzed and compared the distribution of complex types in selected subcompartments between models. We note that, as mass transfer is allowed to occur only between next neighbor subcompartments, constituency of subcompartments of a major compartment may show differences depending on their location relative to the boundary of the major compartment. Although we have cross verified our results in many subcompartments, for our 64 subcompartment model (SSM model 1, Sec. II A), we discuss our results by reporting the molecule distribution in three representative subcompartment units (Fig. 1): subcompartments 27 and 42 of the second major compartment (transmembrane region) and subcompartment 50 of the third major compartment (endosome region). Subcompartment unit 42 is at the border between the second and third major compartments, while unit 27 is located away from the boundary (Fig. 1). For the SSM model 2 (256 subcompartment model), discussed units will be subcompartments 180, 181, 164, and 165. These four units, respectively, are the upper left, upper right, lower left, and lower right subpartitions of SSM model 1 subcompartment unit 42 (Fig. 1). Discussion for the results for other subcompartments parallel the discussion for these chosen units.

We first investigated whether having small or large copy numbers in the simulations have an effect on the model predictions by comparing the results for the different population models (see Sec. II A). Figure 2 reports the mean concentration of RLPA, ligand-bound, and phosphorylated receptors that is in complex with the adaptor protein A (Table III), for SSM model 1 as calculated with the GGB algorithm. The results for the mean RLPA concentration agree very well among different population models. The most significant difference in the results is in the fluctuation levels. As the copy numbers of the molecules in the compartments decrease, there is a noticeable increase in the fluctuations about the mean values. This is a common occurrence in stochastic kinetic systems and an expected result because, at the low copy number limit, a change of one molecule in the compartment can lead to a large change in the species concentration. The same trend is also evident in the results for other molecular types (results not shown). We note the difference in the re-

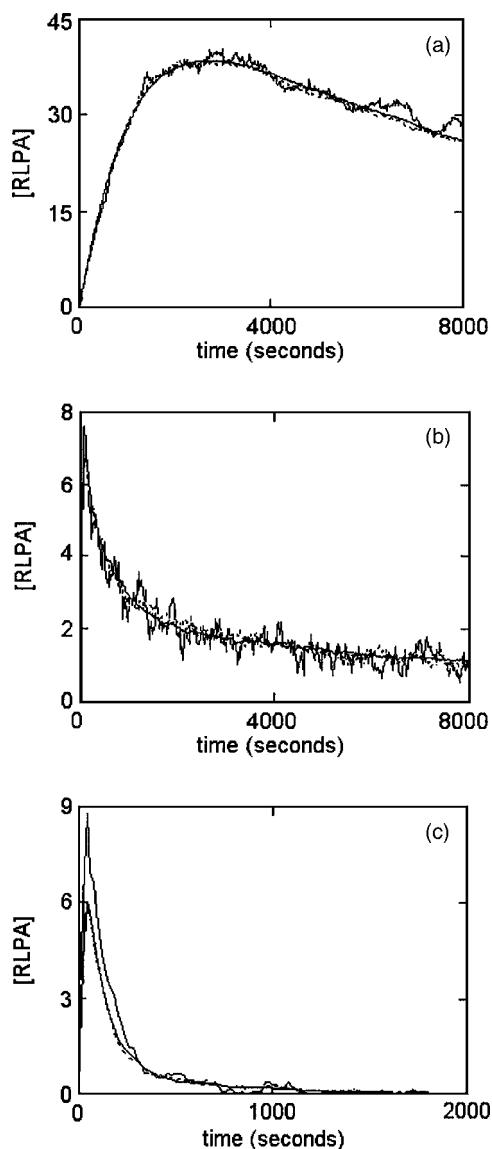


FIG. 2. Concentration of complex type RLPA in subcompartments: (a) 50, (b) 42, and (c) 27 of the SSM model 1 as obtained by averaging the results of 50 stochastic simulations employing the Gillespie-Gibson-Bruck algorithm. Results are shown for different population models: curves with the thick solid [in (c) higher solid curve], dashed, and solid lines show the results for the population models A, B, and C, respectively.

sults between subcompartments 27 and 42 that belong to the same major compartment. As mentioned above, because of the mass transfer events, subcompartment locations would affect the molecular distributions within a major compartment and the spectrum of the differences also depend on the molecular species (results not shown).

In Fig. 3, the averaged solution with the GGB algorithm for complex-type RLPA in unit 42 of the SSM model is contrasted for different size and population models. As is evident from the figures, RLPA levels for the coarser SSM model 1 show a sharp initial transient for all three population models, a feature not found with the finer compartment models. This is likely a result of the faster buildup due to easier mass transfer in the layout with smaller number of compartments. We note that all three size models appear to converge

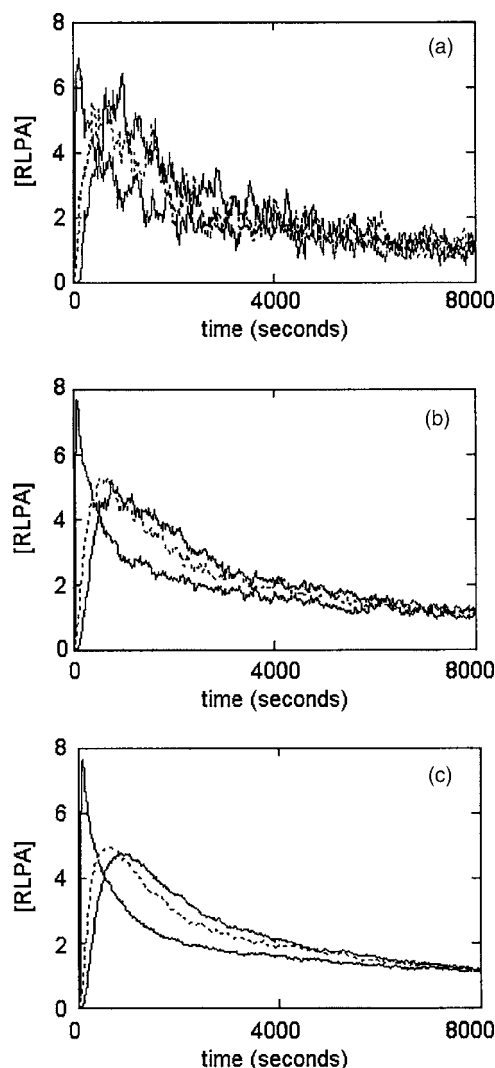


FIG. 3. Concentration of complex type RLPA in unit 42. Shown results are the average of the 50 stochastic trajectories obtained with the Gillespie-Gibson-Bruck algorithm. Population models are (a) C, (b) B, and (c) A. Results are shown for different SSM Models: curves with the thick solid (curve nearest to the y axis), dashed, and solid lines show the results for the SSM models 1, 2, and 3, respectively.

to a steady value of approximately 1.2 nM at 8000 s and that the results for different population models show very good agreement between themselves.

We have also investigated the agreement between the FB and GGB algorithms when both algorithms were employed for the same model. We demonstrate the comparison of the results of the two algorithms with a typical result. Figure 4 reports the RLPA levels for the SSM model 1C in various subcompartments. As they should, results for the FB and GGB algorithms agree with each other very well, not only having the same means but also similar fluctuations.

To compare the computational efficiency of the FB and GGB algorithms, we list the computational expenses of the performed simulations in Tables VIII and IX. As one can easily see from these tables, compared to the GGB algorithm, the computational cost of running the FB algorithm is exorbitant. It should however, be kept in mind that computational costs can strongly depend on how the algorithms are implemented and optimized in the simulation software and

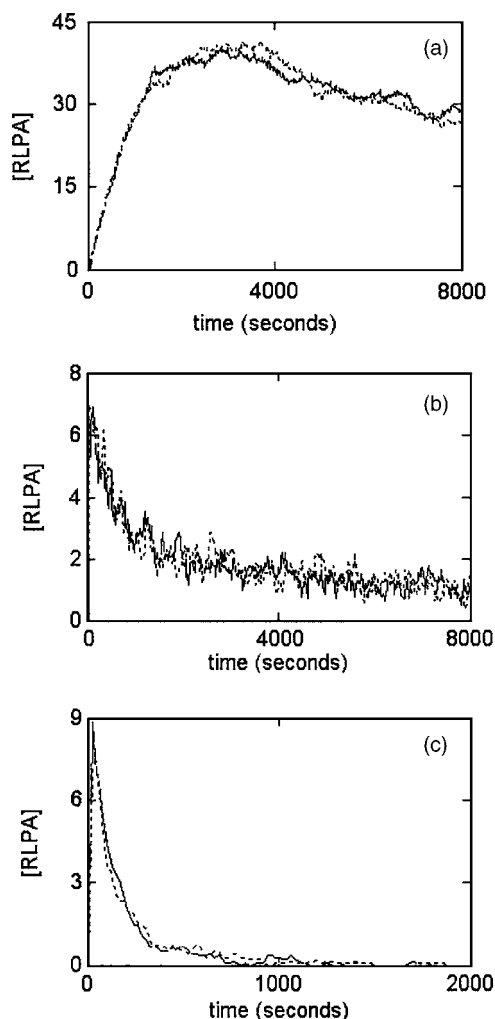


FIG. 4. Concentration of complex type RLPA in units: (a) 50, (b) 42, and (c) 27 of the SSM model 1 with population C, i.e., SSM 1C, as obtained by averaging the results of 50 stochastic simulations. Figure compares the results for the Gillespie-Gibson-Bruck (solid line) and Firth-Bray (dashed line) algorithms.

on the suitability (integer versus floating point operation efficiency) of computer architectures on which the simulations were run. It should also be noted that the efficiency of the algorithms would also depend on the types of the investigated models.

TABLE VIII. Comparison of CPU times and number of reaction events for simulation runs with the Firth-Bray algorithm for the SSM model.

Population model		SSM model 1	SSM model 2	SSM model 3
C	$T$ ( $10^8 \mu s$ )	8400	45 400	252 000
	$T_{event}$ ( $\mu s$ )	0.525	0.568	0.63
	$N_{events}$ ( $10^8$ )	16 000	80 000	400 000
	$N_{reactions}$	518	2268	9464
	$N_{molecules}$ ( $10^3$ )	0.99	0.99	0.99
D	$T$ ( $10^8 \mu s$ )	750	4220	16 200
	$T_{event}$ ( $\mu s$ )	0.469	0.528	0.608
	$N_{events}$ ( $10^8$ )	1600	8000	26 650
	$N_{reactions}$	518	2268	9464
	$N_{molecules}$ ( $10^3$ )	0.099	0.099	0.099

TABLE IX. Comparison of CPU times and number of reaction events for the simulation runs with the Gillespie-Gibson-Bruck algorithm for the SSM model.

Population model		SSM model 1	SSM model 2	SSM model 3
A	$T$ ( $10^8 \mu s$ )	972	2000	9630
	$T_{event}$ ( $\mu s$ )	6.80	7.55	9.44
	$N_{events}$ ( $10^8$ )	143	265	1020
	$N_{reactions}$	518	2268	9464
	$N_{molecules}$ ( $10^3$ )	99	99	99
B	$T$ ( $10^8 \mu s$ )	96.6	202	984
	$T_{event}$ ( $\mu s$ )	6.76	7.59	9.74
	$N_{events}$ ( $10^8$ )	14.3	26.6	101
	$N_{reactions}$	518	2268	9464
	$N_{molecules}$ ( $10^3$ )	9.9	9.9	9.9
C	$T$ ( $10^8 \mu s$ )	10.4	21.4	112
	$T_{event}$ ( $\mu s$ )	7.32	8.08	11.1
	$N_{events}$ ( $10^8$ )	1.42	2.65	10.1
	$N_{reactions}$	518	2268	9464
	$N_{molecules}$ ( $10^3$ )	0.99	0.99	0.99
D	$T$ ( $10^8 \mu s$ )	1.2	2.65	12.3
	$T_{event}$ ( $\mu s$ )	8.57	9.81	11.9
	$N_{events}$ ( $10^8$ )	0.140	0.270	1.03
	$N_{reactions}$	518	2268	9464
	$N_{molecules}$ ( $10^3$ )	0.099	0.099	0.099

As biological data obtained in high-throughput experiments starting to make it possible to construct larger and more complete network models, knowing how the efficiency of the popular kinetic algorithms scales with the network size will be important in choosing the right algorithm to use. For this reason, one of the main aims of this study was to quantify how FB and GGB algorithms scale with the problem size. It is clear from Tables VIII and IX that the total running times for the models studied in this report are dependent on the molecular population as well as on the number of reactions included in the models.

For the FB runs, in all cases reported in Table VIII, nearly optimal values were used for the time steps  $\Delta t \propto V_{reaction}/[n(n+n_G)]$ , where  $V_{reaction}$  is the reaction volume,  $n$  is the total species population, and  $n_G$  is the ghost molecule population (see Appendix A for details). We notice that the time complexity of the FB algorithm does indeed scale directly with the square of the total species population and inversely with the reaction volume with a weak dependency on the number of reactions. For example, moving across rows C and D in Table VIII the total CPU time increases from four- to fivefold corresponding to a fourfold subcompartment volume reduction between constant population models. Additionally, moving upward from row D to C the total population increases tenfold while the reaction volume decreases tenfold thus resulting in an approximately tenfold increase in total CPU time.

It took approximately 25 h to run one simulation with the FB for model 1C which implies a total running time of over 52 days to complete the 50 simulations for comparison with the other data points. Consequently, we limited the us-



age of the FB algorithm to only a few cases as indicated in Table VII. Being able to run the FB algorithm for only a few cases (Table VII) did not allow us to analyze its size scaling analytically. In contrast, manageable total running times of the GBB algorithm made it possible to obtain CPU statistics for all 12 models that we have used (Table IX). Using all 12 data points in Table IX, where  $N_{\text{events}}$ ,  $N_{\text{molecules}}$ ,  $N_{\text{reactions}}$ , and  $T$  are the number of reaction events (in units of  $10^8$ ), total complex population (in units of  $10^5$ ), total number of reactions in the model, and CPU time (in  $10^8 \mu\text{s}$ ), respectively, and a time complexity estimate of  $O(N_{\text{reactions}} + N_{\text{events}} \log N_{\text{reactions}})$ ,<sup>16</sup> a least-squares fit for the regression curve  $T = aN_{\text{molecules}}N_{\text{reactions}} + bN_{\text{molecules}}N_{\text{events}} \ln N_{\text{reactions}}$  yielded values for the fitted coefficients of  $a=1.02$  and  $b=2.24 \times 10^{-3}$ . Our regression analysis suggests that GGB algorithm scales weakly with the logarithm of the number of reactions in the model.

For SSM, we find that, as the total species population is increased over four orders of magnitude, the GGB algorithm performs significantly better than FB for all three SSM compartment models considered. However, extrapolating from the data for population model D (99 molecules), the FB algorithm might become competitive with the GGB algorithm for simulations of SSM model assuming a 1000-fold reduction in species molarity. Although it corresponds to an extreme case, this limit reflects the design characteristics of the FB algorithm. As discussed above and in Appendix A, the Firth-Bray algorithm was not designed as a competitive stochastic algorithm for reaction networks with thousands of species but was rather intended to show its utility for reaction networks characterized by a moderate number of complex types, each with a very small population, but with a high degree of connectivity amongst the complex types. So

the FB algorithm was expected to perform worse than the GBB algorithm for the SSM and our results confirm this expectation. Our results for the SSM also nicely show that the FB method might be the algorithm of choice at the limit that the FB is structured to perform better.

## B. Variable connectivity model (VCM)

Our analysis of the SSM showed that GGB is in general a more efficient algorithm but for certain cases, the use of FB might be more advantageous. Although in designing SSM we have tried to keep a good balance between mass transfer and biochemical reactions, between the time scales of the included events, and between molecular species types, the structure of the SSM could be favoring the GBB algorithm. For this reason, extrapolating the CPU requirements to predict that under certain limiting circumstances the FB might perform better than the GBB is a rather indirect conclusion. To directly show that this is actually the case, we have devised a second scalable network model, the variable connectivity model (VCM) that has the features favorable for the FB algorithm. Our VCM has been described in detail in Sec. II B. Although Gillespie's direct and next reaction methods are closely related, because of the numerical requirements of the methods, one may be more efficient than the other. For this reason, to investigate if our conclusions depend on the form of the Gillespie type algorithms, we have repeated the simulations with the direct method ( $G$ ) of Gillespie.<sup>5,6</sup>

Simulations for the VCM were run for 8000 s and mean values were obtained by averaging the trajectories of 50 runs. The CPU times necessary to initialize the simulations are not included in the reported computation times. The runs were started from an initial distribution where ligand:receptor ratio

TABLE X. Average running times for Firth-Bray (FB), Gillespie-Gibson-Bruck (GGB), and direct Gillespie ( $G$ ) algorithms for the VCM model.

Number of binding sites ( $N$ )	Number of reactions <sup>a</sup>	Reactions with $2^N + 2N$ adjacent edges <sup>b</sup>	Reactions with $2N + 1$ adjacent edges <sup>b</sup>	Initial population (192)			Initial population (1920)		
				FB (s)	GGB (s)	$G$ (s)	FB (s)	GGB (s)	$G$ (s)
1	8	50.0, 4	50.0, 3	0.69	0.39	0.31	56.01	3.84	2.80
2	24	33.3, 8	66.7, 5	1.20	0.63	0.48	100.81	7.22	4.52
3	64	25.0, 14	75.0, 7	1.91	1.17	0.69	148.45	12.11	6.94
4	160	20.0, 24	80.0, 9	2.42	1.75	1.02	196.11	17.81	10.11
5	384	16.7, 42	83.3, 11	3.14	2.69	1.87	246.03	25.73	14.94
6	896	14.3, 76	85.7, 13	3.98	4.17	3.67	298.83	37.23	23.30
7	2048	12.5, 142	87.5, 15	5.09	6.34	6.25	356.77	56.70	40.91
8	4608	11.1, 272	88.9, 17	6.38	10.22	43.25	420.14	84.69	187.21
9	10 240	10.0, 530	90.0, 19	8.31	16.81	235.03	484.03	140.69	645.00
10	22 528	9.1, 1044	90.9, 21	10.09	35.01	570.86	548.17	321.68	2085.68
11	49 152	8.3, 2070	91.7, 23				622.48	798.14	6239.66
12	106 496	7.7, 4120	92.3, 25				725.97	1464.10	16 316.40

<sup>a</sup>In a VCM with  $N$  phosphorylation sites, the total number of molecular species (i.e., complexes) is  $2^{N+1} + 1$ . For such a system with interaction rules given in Sec. II B, the number of reactions included in the model is  $2^{N+1}(N+1)$ .

<sup>b</sup>In the GGB algorithm, firing of a reaction requires the update of a set of related interactions. In the VCM, there are two classes of reactions: Reactions involving ligand binding are connected to a series of reactions (i.e., the adjacent edges in the reaction network graph) that require updating the propensities of a large number ( $2^N + 2N$ ) of reactions. In contrast, when they occur, phosphorylation reactions only require updating the propensities of  $2N + 1$  reactions. In these columns, the first and the second numbers, respectively, are the percentage of reactions belonging to the category and the number of reactions whose propensities need updating.



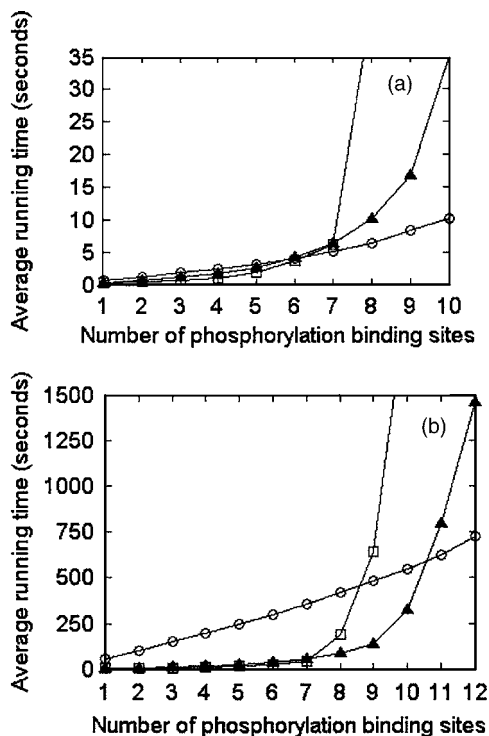


FIG. 5. Average running time as a function of the number of phosphorylation sites for the VCM model. Points show the average running times as reported in Table X. Line curves present the interpolations for the CPU times showing the scaling behavior for the Firth-Bray ( $\circ$ ), the Gillespie-Gibson-Bruck ( $\blacktriangle$ ), and the direct Gillespie ( $\square$ ) algorithms. Results are for the population case of (a) 192 complexes and (b) 1920 complexes.

was 2:1 with no existing ligand-receptor complexes. The volume of the reaction chamber was  $1.6 \times 10^{-6}$  pl. Table X and Fig. 5 report the CPU times for the Firth-Bray ( $\circ$ ), Gillespie-Gibson-Bruck ( $\blacktriangle$ ), and the direct Gillespie ( $\square$ ) algorithms as a function of the number of phosphorylation binding sites for two distinct sets of simulations of the VCM. In the first low population set, Fig. 5(a), the total complex population was initially set at 192 complexes (128 ligands and 64 receptors) while for the second set, Fig. 5(b), the total complex population was initially set at 1920 complexes with the 2:1 ratio of ligands to receptors unchanged (i.e., 1280 ligands and 640 receptors). In carrying out the simulations for the two population cases with the FB algorithm, the maximum number of complexes (cf. Appendix A) was set at 192 and 1920 complexes, respectively. It is clear from Fig. 5 and Table X that for the smaller population set, the FB algorithm becomes more efficient than GGB when the number of phosphorylation sites  $N$  reaches 6 whereas in the larger population set the GGB is clearly superior for small  $N$ 's with a crossover occurring between  $N=10$  and  $N=11$ . In turn the GGB is superior to the  $G$  method for larger values of  $N$  for both population cases while the  $G$  method performs slightly better than the GGB for smaller  $N$ . So for the VCM case, our results clearly establish that depending on the linkage level of the network either GBB or FB can be more efficient; the use of FB-type algorithms should be considered for reaction networks with high connectivity among the involved species. We note that the conclusion supported by our timing data is expected from the following arguments.

Part of the computational increase in the running times of the simulations is due to the fact that the sum  $\Sigma$  of the probability rates for all reactions in the network increases with the number of phosphorylation sites  $N$ , and this reduces the average waiting time  $1/\Sigma$  to the next reaction event (see Appendix A). The arithmetic computational cost per step of the FB method is  $c_0^{\text{FB}} + 3c_{\text{md}} + c_1^{\text{FB}}k_{\text{FB}}$  where  $c_{\text{md}}$  is the cost of computing a random number from the uniform distribution over  $[0,1]$ , and  $k_{\text{FB}}$  is the number of reactions involving both objects that are selected by the two grab process which is a  $O(N)$  process for the VCM model (Appendix A). For the GGB method the arithmetic computational cost per step<sup>7</sup> is  $c_0^{\text{GGB}} + c_{\text{md}} + c_1^{\text{GGB}}k_{\text{GGB}} + c_2^{\text{GGB}}k_{\text{GGB}}\log_2 N_r$  where  $N_r$  is the number of reactions,  $k_{\text{GGB}}$  is the number of edges in the reaction dependency graph (i.e., the number of reactions whose propensity needs to be updated), and the logarithmic term is due to the updating of the indexed priority queue.

For the VCM model it is clear from Fig. 5 that the dependence of  $k_{\text{GGB}}$ , and hence the CPU time, on the number of phosphorylation sites  $N$  is nonlinear. Occurrence of a reaction in the GGB method requires updating the propensities of the reactions that are adjacent in the interaction network graph. In terms of the adjacency graph properties, it can be shown that there are two types of reactions in the VCM model. A large fraction,  $N/(N+1)$ , of the reactions has  $2N+1$  adjacent edges while the remaining  $1/(N+1)$  fraction exhibit an exponential number  $2^N+2N$  of edges (see Table X). Most of the reactions with exponential number of adjacent edges correspond to ligand receptor binding. These reactions have a relatively high rate constant and therefore, can occur frequently and lead to exponential increase in CPU requirement for both the GGB and the direct Gillespie methods. Steep increases seen in Fig. 5 for the computation requirements for the VCM models with large number of phosphorylation sites are due to the dominance of the reactions with an exponential number of adjacency edges.

Even without specifying the values of the constants  $c^{\text{FB}}$  and  $c^{\text{GGB}}$  in these expressions the cost per step for the FB can clearly be less than the GGB when the number of reactions is large. It can be shown that (Appendix A) the optimal time step for the FB method varies inversely with both the number of phosphorylation binding sites and the square of the total species population. As for the GGB, Cao *et al.*<sup>17</sup> argue that the cost of accessing and maintaining the heap data structure for the indexed priority queue can be significantly higher than the associated arithmetic computational costs given above when the connectivity of the reaction network is large. This cost can have a major impact on the average running time of the GGB method. Additionally Cao *et al.*<sup>17</sup> also present an optimized version of the direct Gillespie method which is claimed superior to the GGB when network connectivity is high. Consequently, for the VCM model, if the overall species population is held fixed at sufficiently small levels while increasing the number of phosphorylation sites (implying an increase in network linkage) there can exist a crossover point where the simulation time step for the FB is relatively large compared to the average time progression step in the GGB method so that the computational demands of the GGB method exceed that of the FB method.

Thus, it can be stated that the algorithm of choice for the VCM model is the FB when the connectivity is sufficiently high and the total complex population sufficiently low.

#### IV. SUMMARY

In this study, we have developed and presented two scalable biological networks, the SSM and VCM, to investigate the size scaling efficiencies of two popular and distinct stochastic kinetic simulation algorithms. Studied network models were constructed for use in future benchmarking studies for comparing the size scaling behavior of stochastic kinetic simulation algorithms. Our results have shown that, at least with our implementation of the simulation algorithms and for the utilized scalable models, the GGB algorithm performs significantly better than the FB algorithm under realistic biological conditions. However, we have presented evidence that the FB algorithm can be competitive with the GGB algorithm when the total number of molecules in the system is very low and the molecular species are highly interconnected, i.e., can form many different types of complexes with the other species. We would like to point out that methods which speed up the computations while insignificantly sacrificing the exactness can be combined with the GGB or other Gillespie methods to further improve the performance. Examples of such approaches are the tau-leap method,<sup>18</sup> the implicit tau method,<sup>20</sup> the probability weighted dynamic Monte Carlo method,<sup>13</sup> and others.<sup>19,21,22</sup> In the future, we will include these algorithms in our studies and quantitatively investigate their efficiency performance.

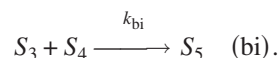
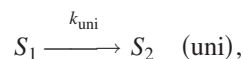
We also note that the used signal transduction models, 2D representation of the cell and its intracellular compartments, and other simplifications that we use in the studied scalable model may distort the underlying biology of the model. However, as the aim of this study is to develop a mechanism for a scalable model as a mean to investigate the size scaling properties of the kinetic simulation algorithms, and as we do not attempt to extract any biological information from the reported simulation results, the reality of the studied models is irrelevant for the purpose of this study.

#### ACKNOWLEDGMENTS

This work was supported in part by the University of Washington/Cell Systems Initiative and PNNL Joint Program in Cell Signaling, by the Laboratory Directed Research and Development program at the Pacific Northwest National Laboratory (PNNL), and by research grants and gifts to the UW CSI from numerous private individuals, from the Washington Research Foundation and from the G. Harold and Leila Y. Mathers Charitable Foundation. We thank Dr. B. Robert Franza, Director of CSI, for his encouragement and interest in the work. We also would like to acknowledge P. Divalentin and Dr. A. Sarkar for stimulating discussions on the geometry of the scalable model and furthermore, thank Dr. Sarkar for carrying out a number of simulations. PNNL is a multiprogram national laboratory operated by Battelle Memorial Institute for the U.S. Department of Energy under Contract No. DE-AC05-76RL01830.

#### APPENDIX A

Consider a dynamical system, in reaction volume  $V_{\text{reaction}}$ , of  $N$  elementary reactions of  $M$  chemical species  $S_i$  with population vector  $\mathbf{p}$  at time  $t$ . The elementary reactions that we shall consider are unimolecular and bimolecular reactions such as



Now the inexact uniform time-stepping method of Firth and Bray,<sup>4</sup> also known as the random substrate method by Schwehm,<sup>23</sup> introduces ghost molecules as an algorithmic artifact allowing both types of elementary reactions to be treated with a single (simultaneous and independent) two-step selection or grab process. For example, consider the unimolecular reaction (uni). According to the law of mass action, the rate of change in the number of molecules  $n_1$  over a time step  $\Delta t$  is, to  $O(\Delta t^2)$ ,

$$\Delta n_1 = -k_{\text{uni}} n_1 \Delta t. \quad (\text{A1})$$

Assuming a ghost molecule population of  $n_G$  and a total species population of  $n = \sum p_i$  (here  $n$  is taken as the maximum number of complexes at any time over the simulation), Eq. (A1) may be rewritten as

$$\begin{aligned} \Delta n_1 &= - \left[ \left( \frac{n_1}{n} \right) \left( \frac{n_G}{n + n_G} \right) \right] \left[ \frac{k_{\text{uni}} n (n + n_G)}{n_G} \Delta t \right] \\ &= [\text{Pr}(G_{\text{uni}}^2)] [\text{Pr}_t(R|G_{\text{uni}}^2) \Delta t]. \end{aligned} \quad (\text{A2})$$

In step 1 of the Firth-Bray method an object is selected from the set of all molecules while in step 2 an object is selected from the set of molecules and ghost molecules. Thus the first bracketed expression in Eq. (A2) is the probability of selecting the correct substrate for unimolecular reaction (uni) while the second bracketed expression is the product of a probability rate  $\text{Pr}_t$  and  $\Delta t$ , and gives the conditional probability of the unimolecular reaction, occurring over time interval  $\Delta t$ , given an appropriate substrate. Similarly, for bimolecular reaction (bi), the rate of change in population  $n_3$  [to  $O(\Delta t^2)$ ],

$$\Delta n_3 = - \frac{k_{\text{bi}}}{N_{\text{Avogadro}} V_{\text{reaction}}} n_3 n_4 \Delta t, \quad (\text{A3})$$

may be rewritten as

$$\begin{aligned} \Delta n_3 &= \left[ 2 \left( \frac{n_3}{n} \right) \left( \frac{n_4}{n + n_G} \right) \right] \left[ \frac{k_{\text{bi}} n (n + n_G)}{2 N_{\text{Avogadro}} V_{\text{reaction}}} \Delta t \right] \\ &= [\text{Pr}(G_{\text{bi}}^2)] [\text{Pr}_t(R|G_{\text{bi}}^2) \Delta t]. \end{aligned} \quad (\text{A4})$$

Again, the first bracketed expression in Eq. (A4) is the probability of selecting a pair of substrates  $S_3$  and  $S_4$  for bimolecular reaction (bi) while the second bracketed expression is the conditional probability of the bimolecular reaction occurring, over time interval  $\Delta t$ , given these substrates.

Morton-Firth<sup>4</sup> estimates the number of ghost molecules  $n_G$ ,

$$n_G = \left\lceil 2N_{\text{Avogadro}} V_{\text{reaction}} \frac{k_{\text{uni,max}}}{k_{\text{bi,max}}} \right\rceil$$

$$\text{where } (\lfloor x \rfloor \equiv \min\{n, n \geq x, \text{integer } n\}), \quad (\text{A5})$$

by equating  $\text{Pr}_i$  for unimolecular and bimolecular reactions, which coincidentally helps to mitigate stiffness issues between the first- and second-order reactions. The uniform time  $\Delta t$  is computed from  $\sum_{i=1}^N \text{Pr}_i(R_i | G^2) \Delta t \leq 1$ . Algorithmically, the two-step grab is accomplished by selecting two uniform random numbers from the intervals  $[1, n]$  and  $[1, n + n_G]$ . If the chosen substrates participate in a nonzero set  $M$  of reactions then a third uniform random number  $s$  from the unit interval  $[0, 1)$  selects the reaction number as the first index  $i_p$  satisfying  $\sum_{k=i_p}^{k=i_p} \text{Pr}_i(R_k \in M | G^2) \Delta t > s$ .

Other than updating population counts of complexes affected by the occurrence of a reaction, no other quantities, including the reaction probabilities, need be modified since the time step is fixed and uniform. Morton-Firth, in his thesis,<sup>4</sup> showed that the number of steps in a simulation  $N_{\text{steps}} \propto (\Delta t)^{-1} \propto n^2 / V_{\text{reaction}}$  so that the algorithm depends strongly on the total complex population  $n$  if the reaction volume is held fixed. It is clear that increasing the population of each species increases the number of nonproductive steps since the probability decreases that any two randomly chosen complexes actually interact.

An important parameter in the FB algorithm is the total population  $n$  of complexes. As an estimate for  $n$ , the FB algorithm uses the maximum number of complexes over the duration of the simulation. However, if the total population varies widely over time or is not easily estimated, it is possible that significant error may be introduced.

In 1976 Gillespie<sup>5,6</sup> developed an exact algorithm which avoided the generally intractable problem of solving for the probabilities of all possible state trajectories described by the master equation.<sup>1</sup> His approach was based on selecting (a) the next reaction and (b) the time of the next reaction according to the probability distributions that underlie the master equation for the particular reaction system in question. Given a dynamical system, in reaction volume  $V_{\text{reaction}}$ , of  $N$  elementary reactions of  $M$  chemical species  $S_i$  with population vector  $\mathbf{p}$  at time  $t$ , define the joint probability density function  $\wp(\kappa, \tau | \mathbf{p}, t)$  or more briefly  $\wp(\kappa, \tau)$ . The quantity  $\wp(\kappa, \tau) d\tau$  gives the probability that  $\kappa$  is the index of the next reaction and that  $\kappa$  occurs over the time interval  $[t, t + \tau + d\tau)$ . Gillespie proved that  $\wp(\kappa, \tau)$  is given by

$$\wp(\kappa, \tau) = \left( \frac{\sigma_\kappa}{\Sigma} \right) (\Sigma e^{-\Sigma \tau}), \quad \Sigma = \sum_{i=1}^N \sigma_i, \quad (\text{A6})$$

where  $\sigma_i$  is the reaction probability rate or propensity<sup>16</sup> of occurrence of the  $i$ th reaction. Summing  $\wp(\kappa, \tau) d\tau$  over all reactions reveals that the waiting time  $\tau$  for the next reaction event at time  $t$  is an exponential random variable with decay constant  $\Sigma$  and mean  $1/\Sigma$ ,

$$\wp_1(\tau) = \Sigma \exp(-\tau \Sigma), \quad (\text{A7})$$

while the integration of  $\wp(\kappa, \tau)$  over the time interval  $[0, \infty)$  reveals that the reaction index of the next event at time  $t$  is the integer random variable  $\kappa$

$$\wp_2(\kappa) = \frac{\sigma_\kappa}{\Sigma}. \quad (\text{A8})$$

Gillespie<sup>5,6</sup> gave the following direct method for sampling from the distributions characterized by the master equation: given two random numbers  $s_1$  and  $s_2$  from the uniform distribution over the unit interval, an exact pair of random variables  $(\kappa, \tau)$  from the joint probability density function  $\wp$  are generated from

$$\tau = -\frac{1}{\Sigma} \ln(s_1) \quad \text{and} \quad \sum_{i=1}^{\kappa} \sigma_i > s_2 \Sigma. \quad (\text{A9})$$

Gillespie<sup>5</sup> introduced a second equivalent algorithm, the first reaction method, in which for each reaction  $\kappa$ , a waiting time  $\tau_\kappa$  is determined from an exponential distribution based on the corresponding propensity  $\sigma_\kappa$ . Since it was thought that a new set of waiting times required  $N$  uniform deviates at each time step, the first reaction method was overlooked for a number of years, until recently, when Gibson and Bruck<sup>7</sup> introduced a variant, the next reaction method (referred to as Gillespie-Gibson-Bruck in this manuscript), which overcomes this issue. This, together, with the use of a reaction dependency graph to minimize reaction probability rate updates and the organization of waiting times into an indexed priority queue, yields an algorithm which is especially efficient for sparse networks. Gibson and Bruck<sup>16</sup> have shown that following the initial time step, the algorithm requires a single random deviate per event and has an overall time complexity of order  $O[N_r + N_E \log(N_r)]$  where  $N_r$  is the number of active reactions and  $N_E$  is the number of events. This does not include the possibly significant overhead incurred to maintain the indexed priority queue and updating based on the reaction dependency graph.

TABLE XI. Conditions imposed on the molecular species in terms of their participation in the reactions for the SSM model.

Condition	Description
1	Only ligand complexes $L$ , originally restricted to compartment 1, may be exchanged between compartments 1 and 2 or 3 by mass transfer.
2	Ligand complexes $L$ , originally restricted to compartment 1, exist unbound only in compartment 1.
3	Receptors $R$ , phosphorylated receptors $\text{RP}$ , and phosphorylated ligand-receptor complexes $\text{RLP}$ are restricted to compartments 2 and 3. Similarly, complexes $\text{RPA}$ and $\text{RLPA}$ exist only in compartments 2 and 3.
4	Adaptor complexes $A$ , originally restricted to compartment 4, exist unbound only in compartment 4.
5	Receptor complexes with attached adaptor protein $A$ may be exchanged between compartments 4 and 2 or 3.
6	Receptor complex exchange is a reversible process combining reaction and mass transfer.
7	Compartments 2 and 3 may exchange all permissible complexes.
8	No-flux boundary conditions exist between model and the environment.

TABLE XII. Detailed description of the reactions constituting the SSM signal transduction model. Given rates are for the SSM model 1 and, as explained in the text, the rates of the mass transport reactions in other size models are adjusted according to the system size.

Cmpt No.	Process type <sup>a</sup>	Process <sup>b</sup>
1	B	None
1	M.1	$L_{-1}_K \xrightarrow{8.6 \times 10^{-2}} L_{-1}_K$
1	B,M.2	$L_{-1}_K + R_{-2}_K \xrightarrow{10^{-3}} RL_{-2}_K$ $L_{-1}_K + RP_{-2}_K \xrightarrow{3 \times 10^{-3}} RLP_{-2}_K$ $L_{-1}_K + RPA_{-2}_K \xrightarrow{6 \times 10^{-2}} RLPA_{-2}_K$
1	B,M.3	$L_{-1}_K + R_{-3}_K \xrightarrow{1.4 \times 10^{-4}} RL_{-3}_K$ $L_{-1}_K + RP_{-3}_K \xrightarrow{3 \times 10^{-3}} RLP_{-3}_K$ $L_{-1}_K + RPA_{-3}_K \xrightarrow{6 \times 10^{-2}} RLPA_{-3}_K$
2	B	$RL_{-2}_K \xrightleftharpoons[9]{1.0} RLP_{-2}_K$ $R_{-2}_K \xrightleftharpoons[9]{10^{-4}} RP_{-2}_K$
2	M.1	$R_{-2}_K \xrightarrow{8.6 \times 10^{-4}} R_{-2}_K$ $RL_{-2}_K \xrightarrow{6.4 \times 10^{-4}} RL_{-2}_K$ $RP_{-2}_K \xrightarrow{8.6 \times 10^{-4}} RP_{-2}_K$ $RLP_{-2}_K \xrightarrow{6.4 \times 10^{-4}} RLP_{-2}_K$ $RPA_{-2}_K \xrightarrow{6.4 \times 10^{-4}} RPA_{-2}_K$ $RLPA_{-2}_K \xrightarrow{4.3 \times 10^{-4}} RLPA_{-2}_K$
2	M.2	$R_{-2}_K \xrightarrow{8.6 \times 10^{-4}} R_{-3}_K$ $RL_{-2}_K \xrightarrow{6.4 \times 10^{-4}} RL_{-3}_K$ $RP_{-2}_K \xrightarrow{8.6 \times 10^{-4}} RP_{-3}_K$ $RLP_{-2}_K \xrightarrow{6.4 \times 10^{-4}} RLP_{-3}_K$ $RPA_{-2}_K \xrightarrow{6.4 \times 10^{-4}} RPA_{-3}_K$ $RLPA_{-2}_K \xrightarrow{4.3 \times 10^{-4}} RLPA_{-3}_K$
2	B,M.3	$RL_{-2}_K \xrightarrow{2.7 \times 10^{-3}} L_{-1}_K + R_{-2}_K$ $RLP_{-2}_K \xrightarrow{6 \times 10^{-5}} L_{-1}_K + RP_{-2}_K$ $RLPA_{-2}_K \xrightarrow{1.0} L_{-1}_K + RPA_{-2}_K$
2	B,M.4	$RPA_{-2}_K \xrightarrow{6.4 \times 10^{-4}} A_{-4}_K + RP_{-2}_K$ $RLPA_{-2}_K \xrightarrow{4.3 \times 10^{-4}} A_{-4}_K + RLP_{-2}_K$
3	B.1	$RL_{-3}_K \xrightleftharpoons[9]{1.0} RLP_{-3}_K$ $R_{-3}_K \xrightleftharpoons[9]{10^{-4}} RP_{-3}_K$
3	B.2	$RL_{-3}_K \xrightarrow{9.1 \times 10^{-4}} \emptyset$ $RLP_{-3}_K \xrightarrow{6 \times 10^{-6}} \emptyset$ $RLPA_{-3}_K \xrightarrow{0.01} \emptyset$

TABLE XII. (Continued.)

Cmpt No.	Process type <sup>a</sup>	Process <sup>b</sup>
3	M.1	$R_{-3}_K \xrightarrow{8.6 \times 10^{-4}} R_{-3}_K$ $RL_{-3}_K \xrightarrow{6.4 \times 10^{-4}} RL_{-3}_K$ $RP_{-3}_K \xrightarrow{8.6 \times 10^{-4}} RP_{-3}_K$ $RLP_{-3}_K \xrightarrow{6.4 \times 10^{-4}} RLP_{-3}_K$ $RPA_{-3}_K \xrightarrow{6.4 \times 10^{-4}} RPA_{-3}_K$ $RLPA_{-3}_K \xrightarrow{4.3 \times 10^{-4}} RLPA_{-3}_K$
3	M.2	$R_{-3}_K \xrightarrow{8.6 \times 10^{-4}} R_{-2}_K$ $RL_{-3}_K \xrightarrow{6.4 \times 10^{-4}} RL_{-2}_K$ $RP_{-3}_K \xrightarrow{8.6 \times 10^{-4}} RP_{-2}_K$ $RLP_{-3}_K \xrightarrow{6.4 \times 10^{-4}} RLP_{-2}_K$ $RPA_{-3}_K \xrightarrow{6.4 \times 10^{-4}} RPA_{-2}_K$ $RLPA_{-3}_K \xrightarrow{4.3 \times 10^{-4}} RLPA_{-2}_K$
3	B,M.3	$RL_{-3}_K \xrightarrow{1.1 \times 10^{-2}} L_{-1}_K + R_{-3}_K$ $RLP_{-3}_K \xrightarrow{6 \times 10^{-5}} L_{-1}_K + RP_{-3}_K$ $RLPA_{-3}_K \xrightarrow{1.0} L_{-1}_K + RPA_{-3}_K$
3	B,M.4	$RPA_{-3}_K \xrightarrow{6.4 \times 10^{-4}} A_{-4}_K + RP_{-3}_K$ $RLPA_{-3}_K \xrightarrow{4.3 \times 10^{-4}} A_{-4}_K + RLP_{-3}_K$
4	B	None
4	M.1	$A_{-4}_K \xrightarrow{8.6 \times 10^{-2}} A_{-4}_K$
4	B,M.2	$A_{-4}_K + RP_{-2}_K \xrightarrow{10^{-4}} RPA_{-2}_K$ $A_{-4}_K + RLP_{-2}_K \xrightarrow{10^{-1}} RLPA_{-2}_K$
4	B,M.3	$A_{-4}_K + RP_{-3}_K \xrightarrow{10^{-4}} RPA_{-3}_K$ $A_{-4}_K + RLP_{-3}_K \xrightarrow{10^{-1}} RLPA_{-3}_K$

<sup>a</sup>B=biochemical; M=mass transfer.

<sup>b</sup> $K_s$  ( $K_t$ )=source (target) subcompartment index (see text for compartment index notation).

## APPENDIX B

Table XI reports the conditions imposed on the molecular species in the SSM model in terms of their involvement in the reactions as a function of their compartments. As stated in Table XI, unbound adaptor proteins reside in the cytoplasm only. However, upon forming a complex with receptors, adaptor proteins become part of the compartment where the interacting receptor resides in. Similarly, upon dissociation from the receptor complex, the adaptor protein returns to the cytoplasm. Therefore, reactions 1, 3, 5, 6, and 7 actually occur between two molecules in different compartments. Also, reactions 6 and 7 involve the mass transfer of an adaptor protein from a source subcompartment to a target subcompartment. Further details of mass transfer processes and reactions 6 and 7 are reported below. It is not difficult to check that species connectivity remains very nearly constant for SSM models 1, 2, and 3.



Localization and spatial distribution of molecules in the SSM system is ruled by particle exchange (i.e., mass transfer) between the subcompartments. A subcompartment may only exchange material with four adjacent subcompartments with which it has a common edge. Material exchange with diagonal neighbor subcompartments is not permitted. Material exchange between two subcompartments that belong to the same main compartment results in material redistribution within that major cellular compartment. In contrast, material exchange between two subcompartments that belong to different major compartments results in material transport between cellular compartments. Rather than using the diffusion equation, we model the mass transfer reactions as first-order reactions. Involved rate constants are guessed using the relationship between the average squared displacement  $\langle r^2 \rangle$  of a particle and its diffusion coefficient  $D$  for random two-dimensional motion  $\langle r^2 \rangle \sim 4Dt$  at large times. Given a subcompartment, the number of complexes with diffusion coefficient  $D$  leaving the circle of radius  $r$  spanning the subcompartment per unit time can be estimated as  $\sim 4D/\langle r^2 \rangle$ . As subcompartments have four edges, the mass transfer rate through any one of the four sides would be  $D/\langle r^2 \rangle$ . If there are  $N^2$  subcompartments in a model,  $\langle r^2 \rangle$  is proportional to  $A/N^2$  where  $A$  is the area of the system. Thus, it follows that the mass transport rate is proportional to  $N^2$ , and this factor was used to equalize the rates among different size models. Mass transfer rate constants that were used in our simulations, as well as the rate constants of the included biochemical reactions occurring in each of the four major compartments are tabulated in Table XII.

We note that compartments may have different characteristics and that the reaction rate of a particular reaction may not be the same in all compartments. For example, the  $pH$  (acidity) can have a strong effect on the association rates and cell compartments are known to have different  $pH$

values.<sup>14,15</sup> In this study, we include such effects by using different reaction rates in different compartments for the same type of reaction (Table XII).

- <sup>1</sup>D. T. Gillespie, *Physica A* **188**, 404 (1992).
- <sup>2</sup>H. H. McAdams and A. Arkin, *Trends Genet.* **15**, 65 (1999); A. Arkin, J. Ross, and H. H. McAdams, *Genetics* **149**, 1633 (1998).
- <sup>3</sup>C. Stanford and R. Horton, *Receptors: Structure and Function*, 2nd ed. (Oxford University Press, New York, 2001).
- <sup>4</sup>C. J. Morton-Firth, Ph.D. thesis, University of Cambridge, Cambridge, UK, 1998.
- <sup>5</sup>D. T. Gillespie, *J. Comput. Phys.* **22**, 403 (1976).
- <sup>6</sup>D. T. Gillespie, *J. Phys. Chem.* **81**, 2340 (1977).
- <sup>7</sup>M. A. Gibson and J. Bruck, *J. Phys. Chem. A* **104**, 1876 (2000).
- <sup>8</sup>S. S. Schmid, *Annu. Rev. Biochem.* **66**, 511 (1997).
- <sup>9</sup>A. Sorkin, *Biochem. Soc. Trans.* **29**, 480 (2001).
- <sup>10</sup>P. P. DiFiore and G. N. Gill, *Curr. Opin. Cell Biol.* **11**, 483 (1999).
- <sup>11</sup>B. N. Kholodenko, O. V. Demin, G. Moehren, and J. B. Hoek, *J. Biol. Chem.* **274**, 30169 (1999).
- <sup>12</sup>A. R. Asthagiri and D. A. Lauffenburger, *Biotechnol. Prog.* **17**, 227 (2001).
- <sup>13</sup>H. Resat, H. S. Wiley, and D. A. Dixon, *J. Phys. Chem. B* **105**, 11026 (2001).
- <sup>14</sup>H. Resat, J. A. Ewald, D. A. Dixon, and H. S. Wiley, *Biophys. J.* **85**, 730 (2003).
- <sup>15</sup>A. R. French, D. K. Tadaki, S. K. Niyogi, and D. A. Lauffenburger, *J. Biol. Chem.* **270**, 4334 (1995).
- <sup>16</sup>M. Gibson and J. Bruck, California Institute of Technology Report No. ETRO26, October 1998.
- <sup>17</sup>Y. Cao, H. Li, and L. Petzold, *J. Chem. Phys.* **121**, 4059 (2004).
- <sup>18</sup>D. Gillespie, *J. Chem. Phys.* **115**, 1716 (2001).
- <sup>19</sup>E. Haseltine and J. Rawlings, *J. Chem. Phys.* **117**, 6959 (2002).
- <sup>20</sup>M. Rathinam, L. R. Petzold, Y. Cao, and D. T. Gillespie, *J. Chem. Phys.* **119**, 12784 (2003); D. T. Gillespie and L. R. Petzold, *J. Chem. Phys.* **119**, 8229 (2003).
- <sup>21</sup>C. Rao and A. Arkin, *J. Chem. Phys.* **118**, 499 (2003).
- <sup>22</sup>T. Kiehl, R. Matheyses, and M. Simmons, *Bioinformatics* **20**, 316 (2004).
- <sup>23</sup>M. Schwehm, *Parallel Stochastic Simulation of Whole-Cell Models*, Proceedings of the Second International Conference on Systems Biology (ICSB 2001), Los Angeles, CA, 4–7 November 2001 (Omni Press, Madison, 2001), pp. 333–341.