STOCHASTIC CHEMICAL KINETICS

Dan Gillespie

Dan T Gillespie Consulting
GillespieDT@mailaps.org

Rethinking Chemical Kinetics

A Chemically Reacting System consists of …

- Molecules of N **chemical species** $S_1,\ldots,S_N$.
  - Inside a volume Ω, at some temperature T.

- M "elemental" **reaction channels** $R_1,\ldots,R_M$.
  - R_j describes a *single instantaneous physical event*, which changes the population of at least one species.
  - "Elemental" means that R_j is one of two types:

$$S_i \rightarrow \text{ something else } \quad (\textit{unimolecular}),$$
$$\text{or}$$
$$S_i + S_{i'} \rightarrow \text{ something else } \quad (\textit{bimolecular}).$$

  - All other types (trimolecular, reversible, etc.) are made up of a series of two or more elemental reactions.

How does a chemically reacting system evolve in time?

The *traditional* answer, for *spatially homogeneous* systems:

"According to the ***reaction rate equation*** (RRE)."

- A set of coupled, first-order ODEs.
- Derived using ad hoc, phenomenological reasoning.
  - Is *more* than the "mass action equations" of thermodynamics, which apply only to systems in *equilibrium*.
- Implies the system evolves *continuously* and *deterministically*, even though molecules come in integer numbers and react stochastically.
- Is empirically accurate for large (test tube size) systems.
- But is often not adequate for very small (cell-size) systems.

* * *

Let's take a fresh look at this question.

Doing it "right": Molecular Dynamics

- The most exact way of describing the system's evolution.
- The "motion picture" approach: Tracks the position and velocity of every molecule in the system.
- Simulates *every* collision, *non-reactive* as well as *reactive*.
- Shows changes in species populations and their spatial distributions.
- ***But*** . . . it's *unfeasibly slow* for nearly all realistic systems.

A great simplification occurs *if* successive *reactive* collisions tend to be separated in time by *very many non-reactive* collisions.

- The overall effect of the non-reactive collisions is to *randomize*
  - the *velocities* of the molecules (Maxwell-Boltzmann distribution).
  - the *positions* of the molecules (spatially uniform or **well-stirred**),
- Then, instead of having to describe the system's state as the *position, velocity and species of each molecule*, we need only give

$$\mathbf{X}(t) \triangleq \big(X_1(t), \dots, X_N(t) \big),$$

$$X_i(t) \triangleq \text{ the } \textit{number} \text{ of } S_i \text{ molecules at time } t.$$

But this ***well-stirred simplification***, which . . .

- ignores the non-reactive collisions,
- truncates the definition of the system's state,

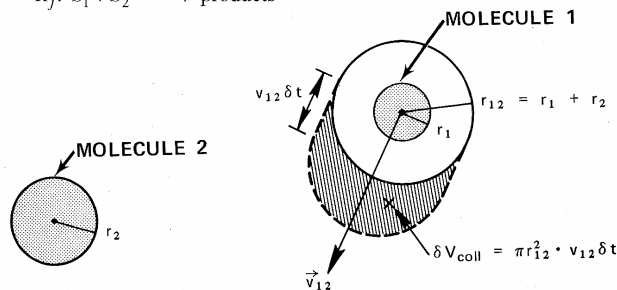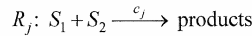   . . . comes at a price:

$\mathbf{X}(t)$ ***must now be viewed as a stochastic process.***

➢ But in fact, ***the system was never deterministic to begin with.***
  Even if molecules moved according to classical mechanics . . .
  - Unimolecular reactions always involve randomness (QM).
  - Bimolecular reactions usually do too.
  - A system of many colliding molecules is so *sensitive to initial conditions* that, for all practical purposes, it evolves "randomly".
  - The system is not isolated. It's in a *heat bath*, which keeps it "at temperature T " – via essentially random interactions.

For well-stirred systems, each R_j is completely characterized by …

- a *propensity function* $a_j(\mathbf{x})$: Given the system is in state $\mathbf{x}$,

  $a_j(\mathbf{x})\,dt \triangleq$ *probability* that one R_j event will occur in the next dt .

  - The *existence* and *form* of $a_j(\mathbf{x})$ follow from *molecular physics*.

- a *state change vector* $\boldsymbol{\nu}_j \equiv \left(\nu_{1j},\ldots,\nu_{Nj}\right)$:

$$\nu_{ij} \triangleq \text{the } change \text{ in } X_i \text{ caused by one } R_j \text{ event.}$$

  - R_j induces $\mathbf{x} \to \mathbf{x}+\boldsymbol{\nu}_j$. $\left\{\nu_{ij}\right\} \equiv$ the "stoichiometric matrix."

Examples:

$$S_1 + S_2 \underset{c_2}{\overset{c_1}{\rightleftharpoons}} 2S_1 \ : \ \begin{cases} a_1(\mathbf{x}) = c_1 x_1 x_2, & \boldsymbol{\nu}_1 = (+1,-1,0,\ldots,0) \\ a_2(\mathbf{x}) = c_2 \dfrac{x_1(x_1-1)}{2}, & \boldsymbol{\nu}_2 = (-1,+1,0,\ldots,0) \end{cases}$$

$$R_j:\ S_1 + S_2 \xrightarrow{\ c_j\ } \text{products}$$
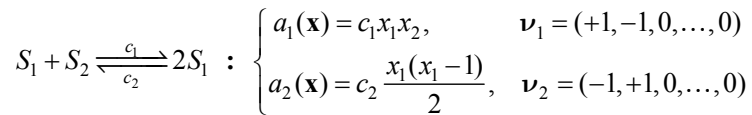
$$\text{Prob}\{v_{12}\text{-collision in } dt\} = \frac{(\pi r_{12}^2)(v_{12}dt)}{\Omega}. \quad \text{Prob}\{R_j\,|\,v_{12}\text{-collision}\} \triangleq p_j(v_{12}).$$

$$\underbrace{\left\langle \left(\frac{(\pi r_{12}^2)(v_{12}dt)}{\Omega}\right) \times p_j(v_{12}) \right\rangle_{v_{12}}}_{\substack{\text{Prob that a randomly chosen} \\ S_1\text{-}S_2 \text{ pair does an } R_j \text{ in next } dt}} \times x_1 x_2 = \underbrace{\left(\frac{\pi r_{12}^2}{\Omega}\left\langle v_{12}p_j(v_{12})\right\rangle_{v_{12}}\right)}_{c_j} x_1 x_2\,dt = \underbrace{c_j x_1 x_2\,dt}_{a_j(\mathbf{x})}$$

$$R_j \text{ iff } \textit{"collisional K.E."} > E_j \ \Rightarrow \ \left\langle v_{12}p_j(v_{12})\right\rangle_{v_{12}} = \underbrace{\sqrt{\frac{8k_B T}{\pi m_{12}}}}_{\langle v_{12}\rangle}\underbrace{\exp\left(-\frac{E_j}{k_B T}\right)}_{\text{Arrhenius}}$$

Two exact, rigorously derivable consequences . . .

> **1.** The *chemical master equation* (CME):

$$\frac{\partial P(\mathbf{x},t\,|\,\mathbf{x}_0,t_0)}{\partial t} = \sum_{j=1}^{M}\left[a_j(\mathbf{x}-\boldsymbol{\nu}_j)P(\mathbf{x}-\boldsymbol{\nu}_j,t\,|\,\mathbf{x}_0,t_0) - a_j(\mathbf{x})P(\mathbf{x},t\,|\,\mathbf{x}_0,t_0)\right].$$

- $P(\mathbf{x},t\,|\,\mathbf{x}_0,t_0) \triangleq \mathrm{Prob}\{\mathbf{X}(t)=\mathbf{x},\text{ given }\mathbf{X}(t_0)=\mathbf{x}_0\}$ for $t\ge t_0$.
- Follows from the *probability* statement

$$P(\mathbf{x},t+dt\,|\,\mathbf{x}_0,t_0) = P(\mathbf{x},t\,|\,\mathbf{x}_0,t_0)\times\left[1-\sum_{j=1}^{M}\left(a_j(\mathbf{x})dt\right)\right]$$
$$+\sum_{j=1}^{M}P(\mathbf{x}-\boldsymbol{\nu}_j,t\,|\,\mathbf{x}_0,t_0)\times\left(a_j(\mathbf{x}-\boldsymbol{\nu}_j)dt\right).$$

- But the CME is usually too hard to solve.

- Averages: $\qquad \left\langle f\left(\mathbf{X}(t)\right)\right\rangle \triangleq \sum_{\mathbf{x}} f(\mathbf{x})P(\mathbf{x},t\,|\,\mathbf{x}_0,t_0)$.

  If we multiply the CME through by $\mathbf{x}$ and then sum over $\mathbf{x}$, we find

$$\frac{d\left\langle\mathbf{X}(t)\right\rangle}{dt} = \sum_{j=1}^{M}\boldsymbol{\nu}_j\left\langle a_j\left(\mathbf{X}(t)\right)\right\rangle.$$

- *If* there were **no fluctuations**,

$$\left\langle a_j\left(\mathbf{X}(t)\right)\right\rangle = a_j\left(\left\langle\mathbf{X}(t)\right\rangle\right) = a_j\left(\mathbf{X}(t)\right),$$

  and the above would reduce to:

$$\frac{d\mathbf{X}(t)}{dt} = \sum_{j=1}^{M}\boldsymbol{\nu}_j a_j\left(\mathbf{X}(t)\right).$$

  - **This is the reaction-rate equation (RRE).**
  - It's usually written in terms of the *concentration* $\mathbf{Z}(t)\triangleq\mathbf{X}(t)/\Omega$.

  ▪ **But as yet, we have *no justification* for ignoring fluctuations.**

➢ **2.** The ***stochastic simulation algorithm*** (SSA):

A procedure for constructing *sample paths* or *realizations* of $\mathbf{X}(t)$.

> *Idea*: Generate *properly distributed random numbers* for
> - the time τ to the *next* reaction,
> - the index j of that reaction.

- $p(\tau, j | \mathbf{x}, t)\, d\tau \triangleq$ probability, given $\mathbf{X}(t) = \mathbf{x}$, that the *next* reaction

$$\text{will occur in } [t+\tau, t+\tau+d\tau), \text{ and will be } R_j.$$

$$= P_0(\tau) \times a_j(\mathbf{x})\, d\tau, \quad P_0(\tau) \triangleq \Pr(no \text{ reactions in time } \tau).$$

$$P_0(\tau + d\tau) = P_0(\tau) \times \left(1 - a_0(\mathbf{x})\, d\tau\right), \text{ where } a_0(\mathbf{x}) \triangleq \sum_1^M a_{j'}(\mathbf{x}).$$

$$\text{Implies } \frac{dP_0(\tau)}{d\tau} = -a_0(\mathbf{x}) P_0(\tau). \quad \text{Solution: } P_0(\tau) = e^{-a_0(\mathbf{x})\tau}.$$

$$\therefore \ p(\tau, j | \mathbf{x}, t) = e^{-a_0(\mathbf{x})\tau}\, a_j(\mathbf{x}) = a_0(\mathbf{x}) e^{-a_0(\mathbf{x})\tau} \times \frac{a_j(\mathbf{x})}{a_0(\mathbf{x})}.$$

Thus,
- τ is an *exponential random variable* with mean $1/a_0(\mathbf{x})$,
- j is an *integer random variable* with probabilities $a_j(\mathbf{x})/a_0(\mathbf{x})$.

The "Direct" Version of the SSA

1. In state $\mathbf{x}$ at time t, evaluate $a_1(\mathbf{x}), \ldots, a_M(\mathbf{x})$, and $a_0(\mathbf{x}) \equiv \sum_{j'=1}^{M} a_{j'}(\mathbf{x})$.

2. Draw two unit-interval uniform random numbers r_1 and r_2, and compute τ and j according to

- $\tau = \dfrac{1}{a_0(\mathbf{x})} \ln\left(\dfrac{1}{r_1}\right),$

- $j =$ the *smallest integer* satisfying $\displaystyle\sum_{k=1}^{j} a_k(\mathbf{x}) > r_2\, a_0(\mathbf{x})$.

3. Replace $t \leftarrow t + \tau$ and $\mathbf{x} \leftarrow \mathbf{x} + \boldsymbol{\nu}_j$.

4. Record $(\mathbf{x}, t)$. Return to Step **1**, or else end the simulation.

<div align="center">

A Simple Example: $S_1 \xrightarrow{\;c_1\;} 0$.

</div>

$$a_1(x_1) = c_1 x_1, \quad \nu_1 = -1 . \;\; \text{Take } X_1(0) = x_1^0 .$$

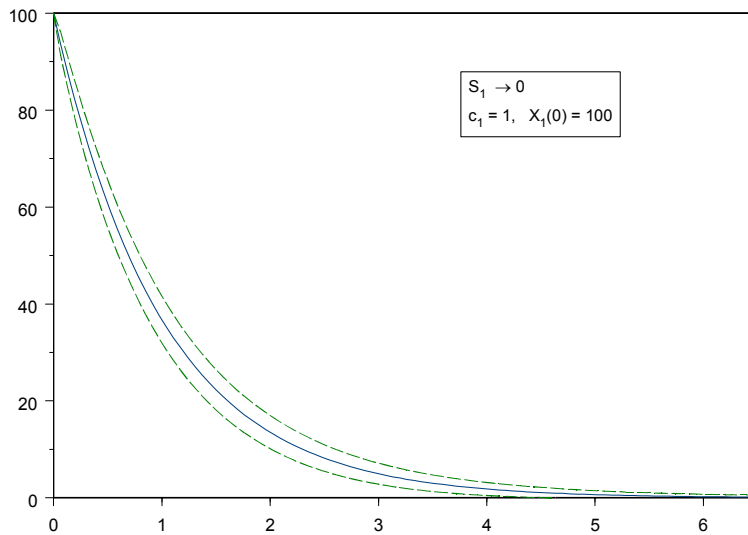**RRE**: $\dfrac{dX_1(t)}{dt} = -c_1 X_1(t)$. Solution is $X_1(t) = x_1^0 \, e^{-c_1 t}$.

CME: $\dfrac{\partial P(x_1,t\,|\,x_1^0,0)}{\partial t} = c_1 \Big[(x_1+1) P(x_1+1,t\,|\,x_1^0,0) - x_1 P(x_1,t\,|\,x_1^0,0) \Big]$.

Solution: $P(x_1,t\,|\,x_1^0,0) = \dfrac{x_1^0 !}{x_1!\,(x_1^0 - x_1)!} \, e^{-c_1 x_1 t} \left(1 - e^{-c_1 t}\right)^{x_1^0 - x_1} \; (x_1 = 0,1,\ldots,x_1^0)$

which implies $\langle X_1(t) \rangle = x_1^0 \, e^{-c_1 t}$, $\text{sdev}\{X_1(t)\} = \sqrt{x_1^0 \, e^{-c_1 t} \left(1 - e^{-c_1 t}\right)}$.

SSA: Given $X_1(t) = x_1$, generate $\tau = \dfrac{1}{c_1 x_1} \ln\left(\dfrac{1}{r}\right)$, then update:

$$t \leftarrow t + \tau, \quad x_1 \leftarrow x_1 - 1 .$$

$S_1 \to 0$
$c_1 = 1, \;\; X_1(0) = 100$

S_1 → 0
$c_1 = 1$, $X_1(0) = 100$
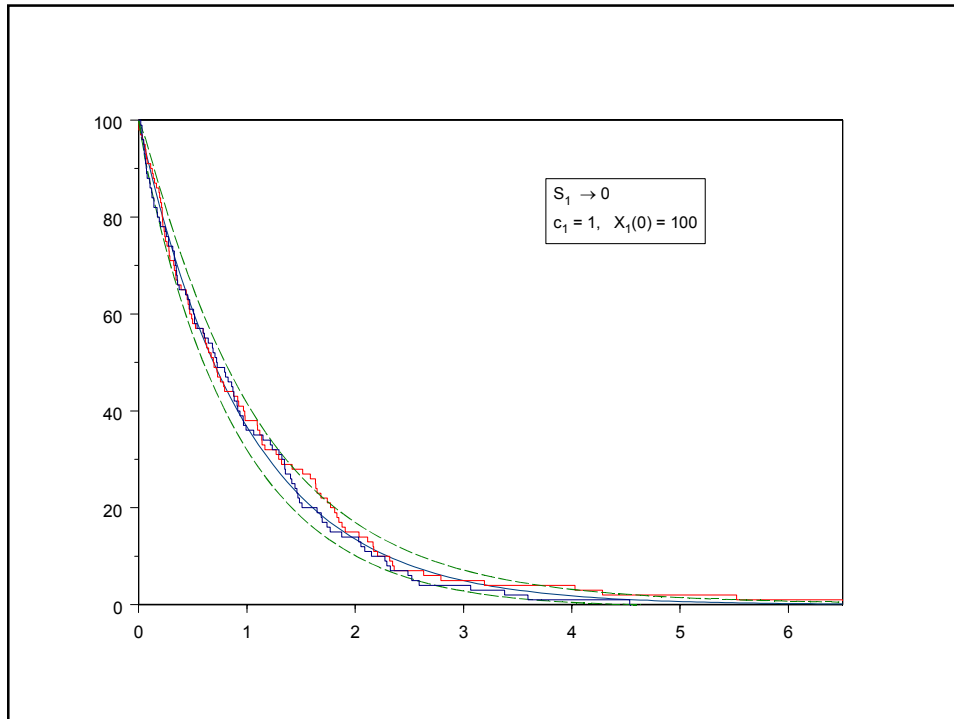
The SSA . . .

- Is *exact*.
- Does *not* entail approximating "*dt*" by "*Δt*".
- Is logically on par with the CME (but is *not* a method for numerically solving the CME).
- Is *procedurally simple*, even when the CME is intractable.
- Comes in a variety of implementations …
  - Direct Method (Gillespie, 1976)
  - First Reaction Method (Gillespie, 1976)
  - Next Reaction Method (Gibson & Bruck, 2000)
  - First Family Method (Lok, 2003)
  - Modified Direct Method (Cao, Li & Petzold, 2004)
  - Sorting Direct Method (McCollum, et al. 2006)
- ***Remains too slow for most practical problems***: Simulating *every* reaction event *one* at a time just takes too much time if any reactants are present in very large numbers.

We would be willing to sacrifice a little exactness . . .
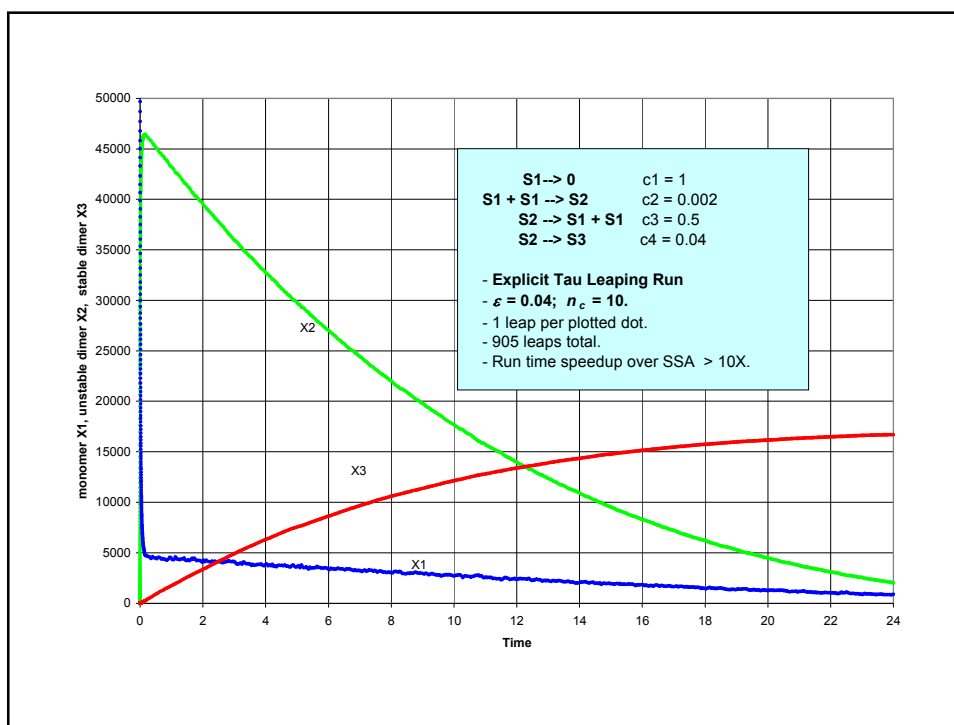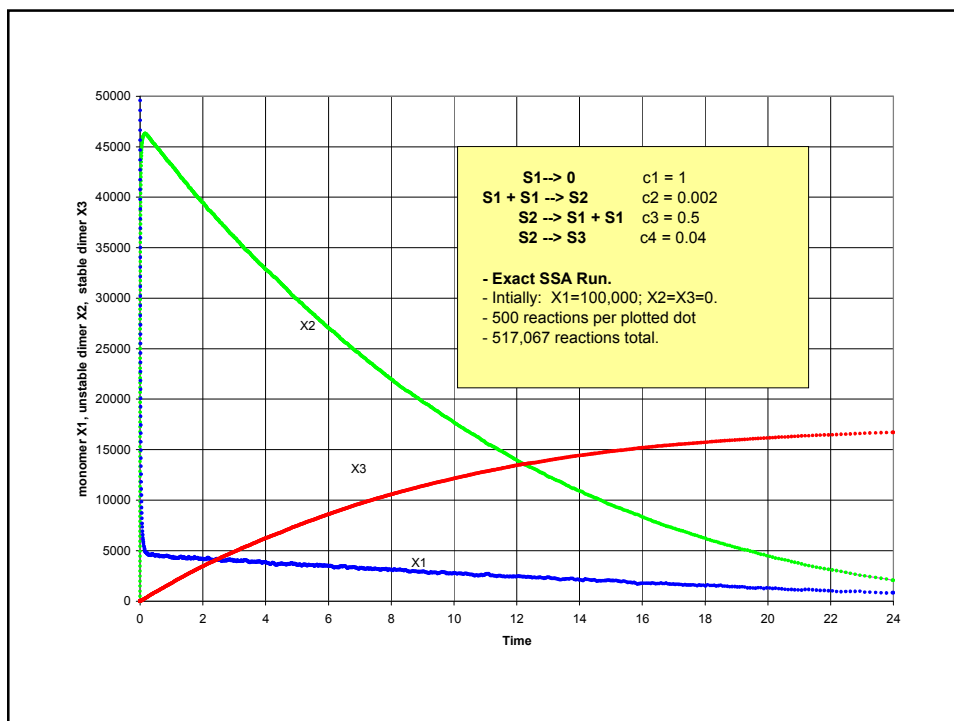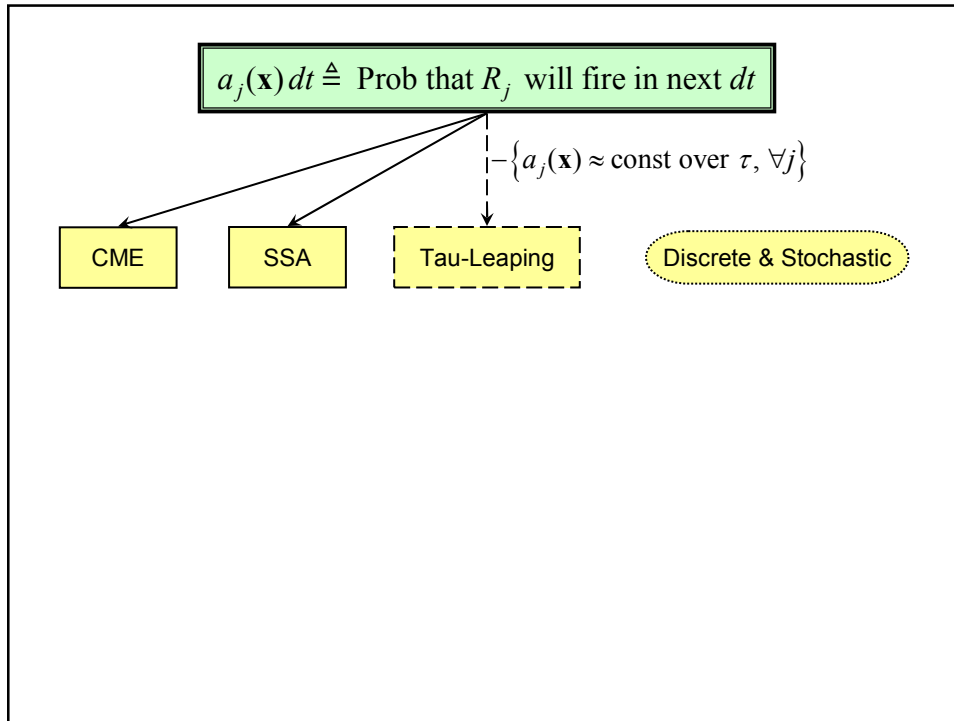**. . . if that would buy us a faster simulation.**

Tau-Leaping

- *Approximately* advances the process by a *pre-selected* time τ, which may encompass *more than one* reaction event.

- *Key*: The definition of "the Poisson random variable with mean $a\tau$":

    $\mathcal{P}(a\tau) \triangleq$ the *number of events* that will occur in a time τ,
    given that the probability of an event in any dt is $a\,dt$
    where a can be any positive *constant*.

- With $\mathbf{X}(t) = \mathbf{x}$, let us choose τ *small enough* to satisfy the

    Leap Condition: Each $a_j(\mathbf{x}) \approx$ constant in $[t, t+\tau]$.

- Then: The number of R_j firings in $[t, t+\tau] \approx \mathcal{P}\big(a_j(\mathbf{x})\tau\big)$.

$$\mathbf{X}(t+\tau) \doteq \mathbf{x} + \sum_{j=1}^{M} \mathcal{P}_j\big(a_j(\mathbf{x})\tau\big)\,\boldsymbol{\nu}_j$$

- Practical Implementation of Tau-Leaping -

- We have *two control parameters*, ε and n_c:
    - To satisfy the Leap Condition, restrict τ so that $\big|\Delta_\tau a_j / a_j\big| \leq \varepsilon, \forall j$.
    - To avoid populations <0, allow only *one* firing of all *critical* reactions ($\triangleq$ reactions that are within n_c firings of exhausting any reactant).

- We take $\tau = \min(\tau', \tau'')$, where:
    - τ' maximally satisfies the Leap Condition for firings of the *non-critical* reactions. (We have a fairly efficient way to estimate τ'.)
    - τ'' is the time to the *next critical* reaction . (Generate τ'' by applying the SSA to the *critical* reactions.)

- For each *non-critical* R_j, generate k_j as a sample of $\mathcal{P}\big(a_j(\mathbf{x})\tau\big)$.

- If $\tau' < \tau''$: Set the k_j's for all the *critical* R_j's to 0.

    If $\tau'' \leq \tau'$: Use the SSA to determine which critical reaction fires, set its k_j to 1, and set all other critical k_j's to 0.

- *Leap*: $t \leftarrow t + \tau$ and $\mathbf{x} \leftarrow \mathbf{x} + \sum_{j=1}^{M} k_j \boldsymbol{\nu}_j$.

➢ Becomes the SSA if <u>all</u> reactions are <u>critical</u> ($n_c \to \infty$).

monomer X1, unstable dimer X2, stable dimer X3

50000
45000
40000
35000
30000
25000
20000
15000
10000
5000
0

X2

X3

X1

Time
0 2 4 6 8 10 12 14 16 18 20 22 24

S1--> 0   c1 = 1
S1 + S1 --> S2   c2 = 0.002
     S2 --> S1 + S1   c3 = 0.5
     S2 --> S3   c4 = 0.04

- **Exact SSA Run.**
- Intially: X1=100,000; X2=X3=0.
- 500 reactions per plotted dot
- 517,067 reactions total.

monomer X1, unstable dimer X2, stable dimer X3

50000
45000
40000
35000
30000
25000
20000
15000
10000
5000
0

X2

X3

X1

Time
0 2 4 6 8 10 12 14 16 18 20 22 24

S1--> 0   c1 = 1
S1 + S1 --> S2   c2 = 0.002
     S2 --> S1 + S1   c3 = 0.5
     S2 --> S3   c4 = 0.04

- **Explicit Tau Leaping Run**
- $\varepsilon = 0.04$; $n_c = 10$.
- 1 leap per plotted dot.
- 905 leaps total.
- Run time speedup over SSA > 10X.

10

$$a_j(\mathbf{x})\,dt \triangleq \text{ Prob that } R_j \text{ will fire in next } dt$$

$-\{a_j(\mathbf{x}) \approx \text{const over } \tau, \ \forall j\}$

CME SSA Tau-Leaping Discrete & Stochastic

Speeding up Tau-Leaping: The Langevin Equation

- Two math facts:
  - If $m \gg 1$, then $\mathcal{P}(m) \approx \mathcal{N}(m,m)$.
  - $\mathcal{N}(m,\sigma^2) = m + \sigma\mathcal{N}(0,1)$.

- So, with $\mathbf{X}(t) = \mathbf{x}$, suppose we can choose τ *small enough* to satisfy the Leap Condition, *yet also large enough that* $a_j(\mathbf{x})\tau \gg 1, \forall j$.

Then . . .

$$\mathbf{X}(t+\tau) \doteq \mathbf{x} + \sum_{j=1}^{M} \mathcal{P}_j\big(a_j(\mathbf{x})\tau\big)\boldsymbol{\nu}_j$$

$$\doteq \mathbf{x} + \sum_{j=1}^{M} \mathcal{N}_j\big(a_j(\mathbf{x})\tau, a_j(\mathbf{x})\tau\big)\boldsymbol{\nu}_j$$

$$\doteq \mathbf{x} + \sum_{j=1}^{M} \Big[a_j(\mathbf{x})\tau + \sqrt{a_j(\mathbf{x})\tau}\,\mathcal{N}_j(0,1)\Big]\boldsymbol{\nu}_j$$

❖ $$\mathbf{X}(t+\tau) \doteq \mathbf{x} + \sum_{j=1}^{M} \boldsymbol{\nu}_j\, a_j(\mathbf{x})\tau + \sum_{j=1}^{M} \boldsymbol{\nu}_j\, \sqrt{a_j(\mathbf{x})}\,\mathcal{N}_j(0,1)\,\sqrt{\tau}\ .$$

$$\mathbf{X}(t+\tau) \doteq \mathbf{x} + \sum_{j=1}^{M} \boldsymbol{\nu}_j \, a_j(\mathbf{x})\tau + \sum_{j=1}^{M} \boldsymbol{\nu}_j \sqrt{a_j(\mathbf{x})} \, \mathcal{N}_j(0,1)\sqrt{\tau}$$

- This is the ***Langevin leaping formula***.
- It's faster than the ordinary tau-leaping formula, because
    - $a_j(\mathbf{x})\tau \gg 1$ means *lots* of reaction events get leapt over in τ ;
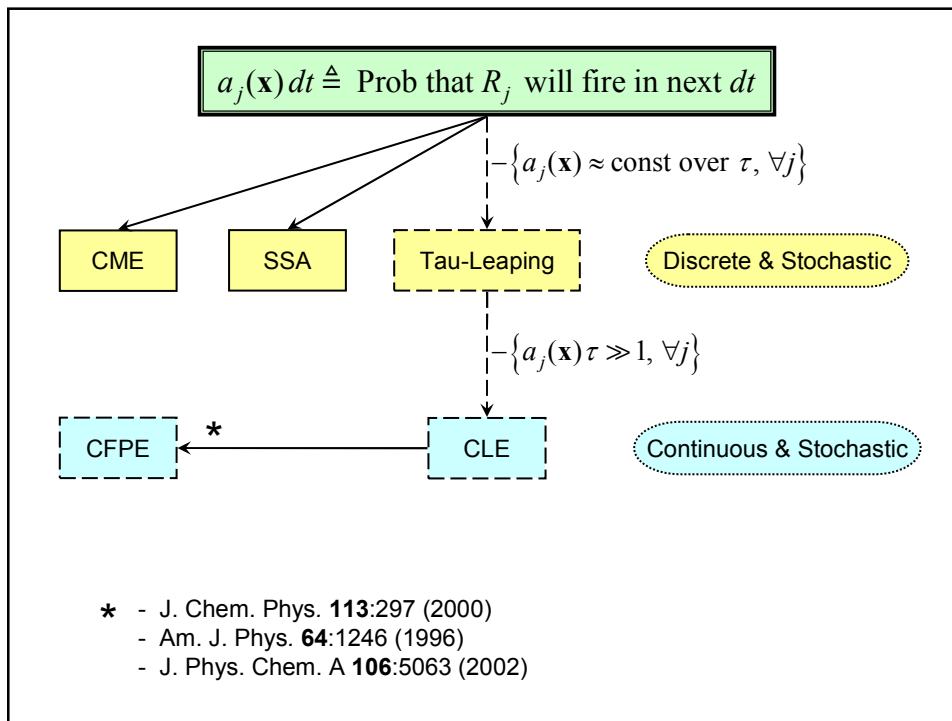   - *normal* random numbers can be generated faster than Poissons.
- It directly implies, and is entirely equivalent to, a SDE called
    the ***chemical Langevin equation*** (CLE):

$$\frac{d\mathbf{X}(t)}{dt} \doteq \sum_{j=1}^{M} \boldsymbol{\nu}_j \, a_j(\mathbf{X}(t)) + \sum_{j=1}^{M} \boldsymbol{\nu}_j \sqrt{a_j(\mathbf{X}(t))} \, \Gamma_j(t) \; .$$

   - *Gaussian white noise*: $\Gamma(t) \triangleq \lim_{dt \to 0^+} \dfrac{\mathcal{N}(0,1)}{\sqrt{dt}} \equiv \lim_{dt \to 0^+} \mathcal{N}\!\left(0, \dfrac{1}{dt}\right)$.

   - Satisfies $\left\langle \Gamma_j(t)\,\Gamma_{j'}(t')\right\rangle = \delta_{jj'}\,\delta(t-t')$.

- Our *discrete stochastic* process $\mathbf{X}(t)$ has now been *approximated* as a
    <u>*continuous stochastic*</u> process.

$$\boxed{a_j(\mathbf{x})\,dt \triangleq \text{ Prob that } R_j \text{ will fire in next } dt}$$

$\vdash \left\{ a_j(\mathbf{x}) \approx \text{const over } \tau, \; \forall j \right\}$

| CME | SSA | Tau-Leaping | Discrete & Stochastic |

$\vdash \left\{ a_j(\mathbf{x})\tau \gg 1, \; \forall j \right\}$

| CFPE | $\ast$ ← | CLE | Continuous & Stochastic |

$\ast$ - J. Chem. Phys. **113**:297 (2000)
   - Am. J. Phys. **64**:1246 (1996)
   - J. Phys. Chem. A **106**:5063 (2002)

<div style="border:1px solid black; padding:1em;">

The Thermodynamic Limit

Def: All $X_i \to \infty$, and $\Omega \to \infty$, with X_i/Ω constants.

- $\left. \begin{array}{l} a_j = c_j x_1 \sim x_1 \\ a_j = c_j x_1 x_2 \sim \Omega^{-1} x_1 x_2 \sim x_2 \end{array} \right\} \Rightarrow \left\{ \begin{array}{l} \text{In the thermodynamic limit,} \\ \textbf{all } a_j\text{'s grow like (system size).} \end{array} \right.$

- So in the thermodynamic limit, we see that in the CLE

$$\frac{d\mathbf{X}(t)}{dt} \doteq \sum_{j=1}^{M} \boldsymbol{\nu}_j \, a_j\big(\mathbf{X}(t)\big) + \sum_{j=1}^{M} \boldsymbol{\nu}_j \sqrt{a_j\big(\mathbf{X}(t)\big)} \, \Gamma_j(t) \ ,$$
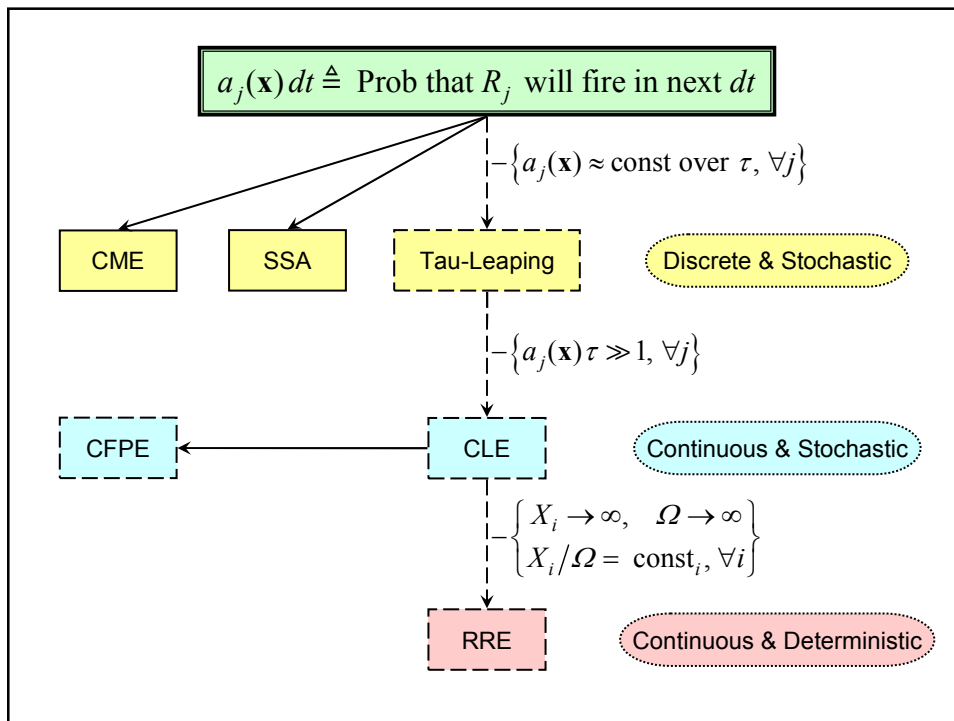
  - the *deterministic* term grows like (system size),
  - the *stochastic* term grows like $(\text{system size})^{1/2}$.

- $\Rightarrow$ Rule of Thumb: ***Relative fluctuations die off as*** $(\textbf{\textit{system size}})^{-1/2}$.

- **At the thermodynamic limit** the stochastic term disappears, leaving

$$\frac{d\mathbf{X}(t)}{dt} \doteq \sum_{j=1}^{M} \boldsymbol{\nu}_j \, a_j\big(\mathbf{X}(t)\big) \ \ \dots \text{ the } \textbf{RRE } \textit{\dots derived!}$$

  $\mathbf{X}(t)$ has now become a ***continuous deterministic*** process.

</div>

Complications from "Stiffness"

- Some R_j may be very fast, others very slow.

- Some X_i may be very fast, others very slow.

- "Fast" and "slow" are **interconnected** – not easy to separate.

- Often manifests as *dynamical stiffness*, a known ODE problem.

- SSA still works, and is exact. But it's agonizingly slow.

- Tau-leaping remains accurate, but the Leap Condition restricts τ to the shortest (fastest) time scale of the system. Still very slow.

- ***One approach:*** *Implicit Tau-Leaping*
  - A stochastic adaptation of the implicit Euler method for ODEs.

- ***Another approach:*** The *Slow-Scale Stochastic Simulation Algorithm*
  - Skips over the fast reactions and *simulates only the slow reactions*, using specially modified propensity functions. An adaptation of the partial equilibrium / quasi steady-state method for RREs.

Collaborators and Associates

Linda Petzold (UCSB)
John Doyle (Caltech)
Michael Hucka (Caltech & Beckman BNMC)
Muruhan Rathinam (UMBC)
Yang Cao (Virginia Tech)
Sotiria Lampoudi (UCSB)
Hong Li (UCSB)