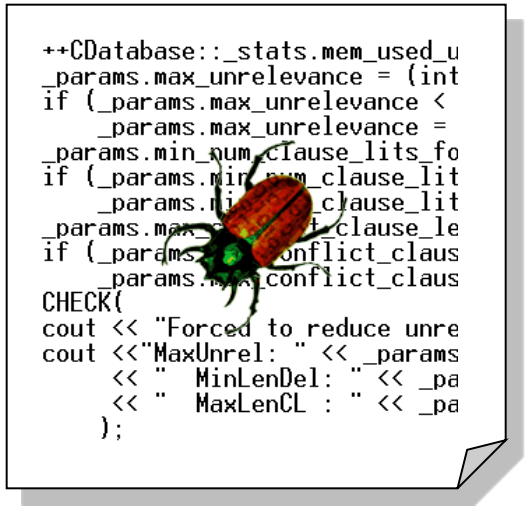
Statistical Model Checking with Applications to Systems Biology

Edmund M. Clarke
School of Computer Science
Carnegie Mellon University



Joint work with James Faeder, Sumit Jha, Chris Langmead, Andre Platzer, and Paolo Zuliani

Intel Pentium FDIV Bug

Try 4195835 – 4195835 / 3145727 * 3145727.

In 94’ Pentium, it doesn’t return 0, but 256.

Intel uses the SRT algorithm for floating point division. Five entries in the lookup table are missing.

Cost: $400 - $500 million

Xudong Zhao’s Thesis on Word Level Model Checking

P53-Mdm2 and DNA Repair Circuit

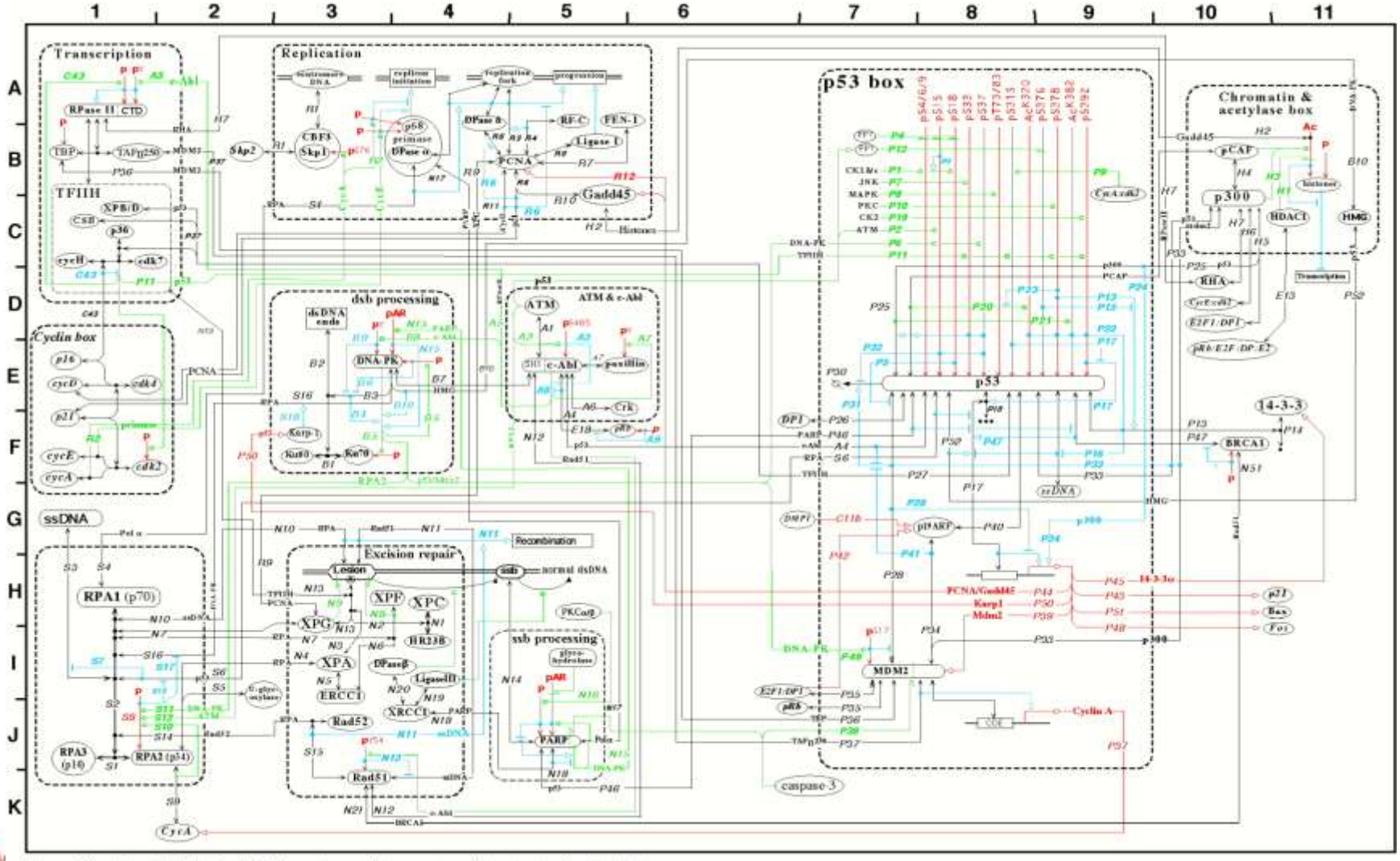


Figure 6B: The p53-Mdm2 and DNA repair regulatory network (version 2p - May 19, 1999)

Kurt W. Kohn, Molecular Biology of the Cell 1999

P53, DNA Repair, and Apoptosis

"The p53 pathway has been shown to mediate cellular stress responses; p53 can initiate DNA repair, cell-cycle arrest, senescence and, importantly, apoptosis. These responses have been implicated in an individual's ability to suppress tumor formation and to respond to many types of cancer therapy."

(A. Vazquez, E. Bond, A. Levine, G. Bond. The genetics of the p53 pathway, apoptosis and cancer therapy. Nat Rev Drug Discovery 2008 Dec;7(12):979-87.)

The protein **p53** has been described as the **guardian of the genome** referring to its role in preventing genome mutation.

In 1993, **p53** was voted ***molecule of the year*** by **Science Magazine**.

The State Explosion Problem

My 28 Year Quest:

- Symmetry Reduction
- Parametric Model Checking
- Partial Order Reduction
- Symbolic Model Checking
- Induction in Model Checking
- SAT based Bounded Model Checking
- Predicate Abstraction
- Counterexample Guided Abstraction Refinement
- Compositional Reasoning

. . .

The State Explosion Problem

My 28 Year Quest:

- Symmetry Reduction
- Parametric Model Checking
- Partial Order Reduction
- Symbolic Model Checking
- Induction in Model Checking
- SAT based Bounded Model Checking
- Predicate Abstraction
- Counterexample Guided Abstraction Refinement
- Compositional Reasoning

. . .

- ***Statistical Model Checking!***

Wait a minute!

Isn’t ***Statistical Model Checking*** an oxymoron?

Wait a minute!

Isn’t ***Statistical Model Checking*** an oxymoron?

I thought so for the first 27 years of my quest.

Wait a minute!

Isn’t ***Statistical Model Checking*** an oxymoron?

I thought so for the first 27 years of my quest.

Much easier to simulate a complex biological system than to build the transition relation for it.

Wait a minute!

Isn't ***Statistical Model Checking*** an oxymoron?

I thought so for the first 27 years of my quest.

Much easier to simulate a complex biological system than to build the transition relation for it.

Moreover, we can bound the probability of error.

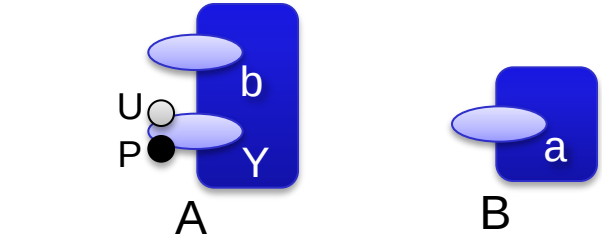
The BioNetGen Language

Jim Faeder, UPMC

begin molecule types

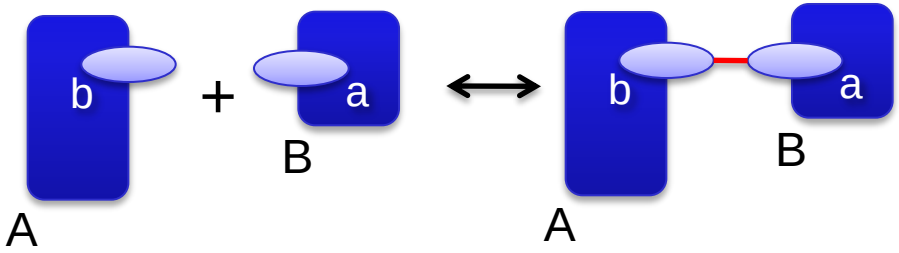
```
A(b,Y~U~P)
B(a)
```

end molecule types



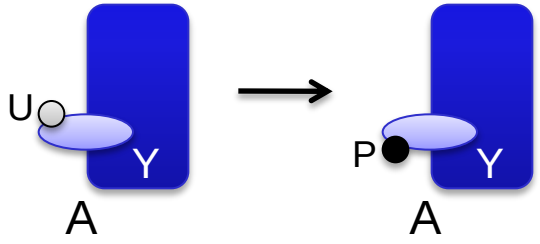
begin reaction rules

```
A(b)+ B(a)<-> A(b!1).B(a!1)
```



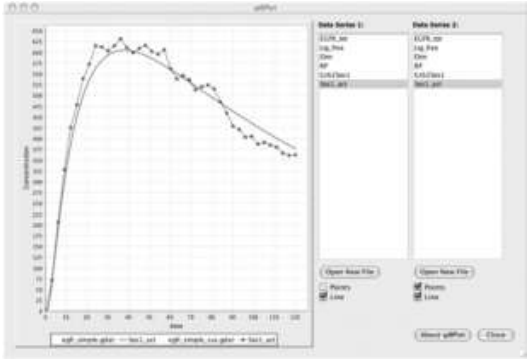

```
A(Y~U) -> A(Y~P)
```

end reaction rules



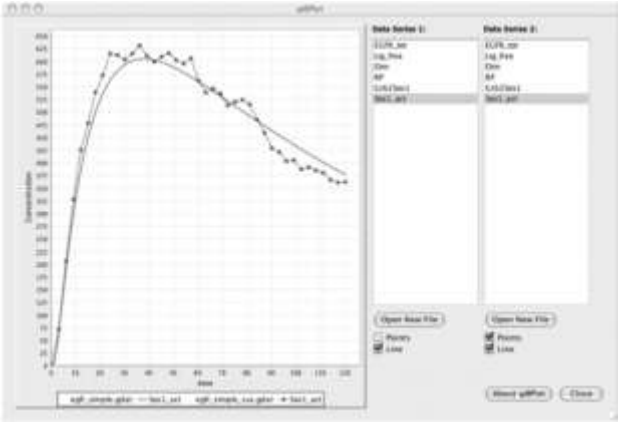
Faeder JR, Blinov ML, Hlavacek WS **Rule-Based Modeling of Biochemical Systems with BioNetGen.** In *Methods in Molecular Biology: Systems Biology*, (2009).

Existing Approach: Manual Analysis



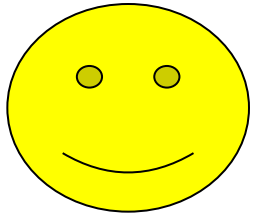
Many simulation traces need to be carefully analyzed !

Model Checking Approach



Automated Analysis !

BioLab 2.0



Bounded Linear Temporal Logic

- Bounded Linear Temporal Logic (BLTL): Extension of LTL with **time bounds** on temporal operators.
- Let $\sigma = (s_0, t_0), (s_1, t_1), \ldots$ be an execution of the model
  - along states $s_0, s_1, \ldots$
  - the system stays in state s_i *for time* t_i
- Example: Does the concentration of protein G stay above 6000 for 2 time units and fall below 6000 before 20 time units?
  - G^2 (*GProtein* > 6000) $\wedge$ F^{20} (*GProtein* < 6000)

Semantics of BLTL

The semantics of the **timed Until** operator:

- "within time t, Φ_2 will be true and Φ_1 will hold until then "
- σ^k: Execution trace starting at state k.
- $\sigma^k \models \Phi_1 \; \mathcal{U}^t \; \Phi_2$ iff there exists a number n such that
  1) $\sigma^{k+n} \models \Phi_2$
  2) $\Sigma_{i<n} \; t_{k+i} \leq t$
  3) for each $0 \leq j < n$, $\sigma^{k+j} \models \Phi_1$
- In particular: $F^t \; \Phi = true \; U^t \; \Phi$, $G^t \; \Phi = \neg F^t \; \neg \Phi$

Probabilistic Model Checking

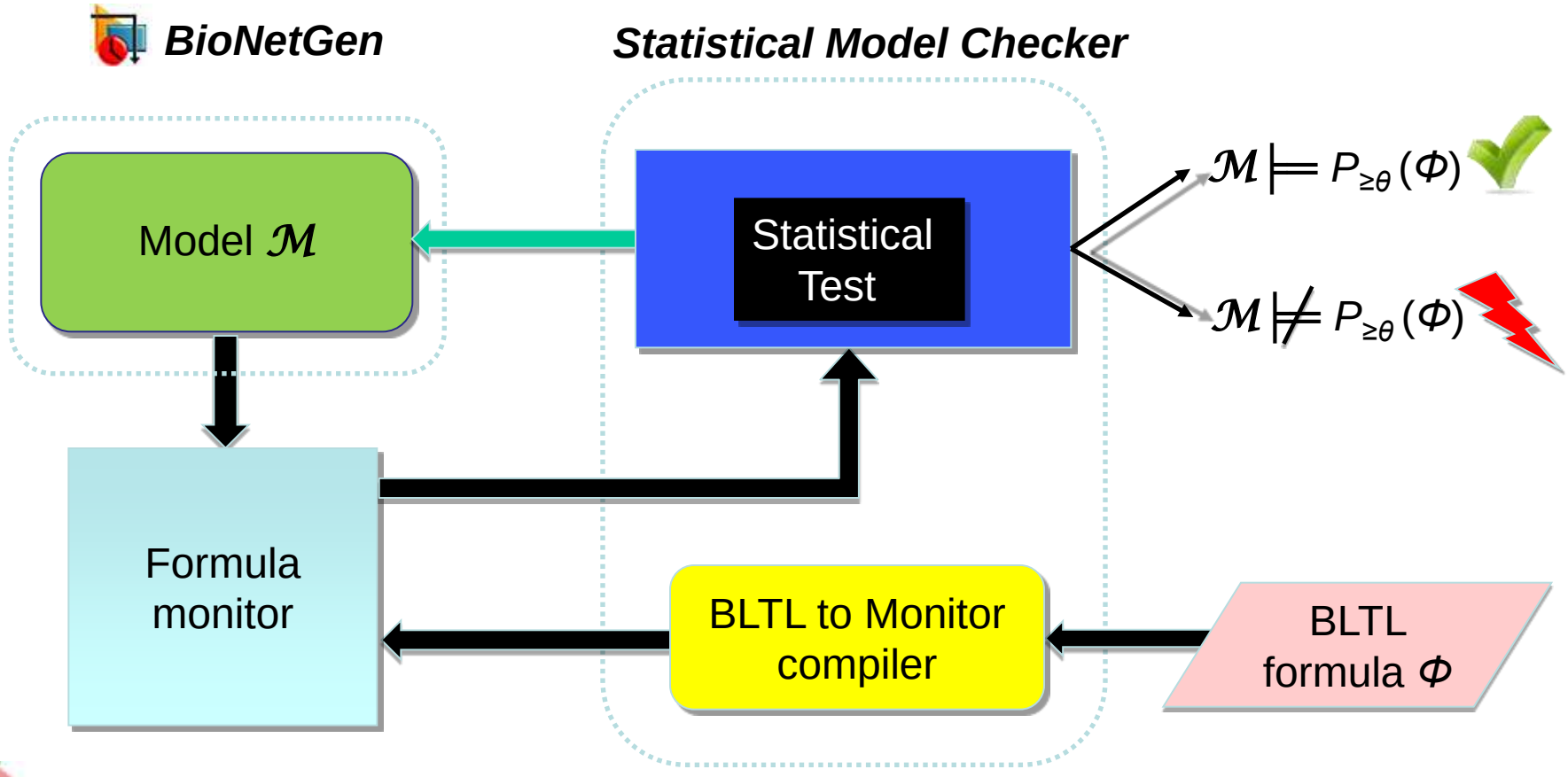
- Given a stochastic model $\mathcal{M}$ such as
  - a Discrete or Continuous Markov Chain, or
  - the solution to a stochastic differential equation
- a Bounded Linear Temporal Logic property ϕ and a probability threshold $\theta \in (0,1)$.
- Does $\mathcal{M}$ satisfy ϕ with probability at least θ?

$$\mathcal{M} \models P_{\geqslant \theta}(\phi)$$

- Numerical techniques compute the precise probability of $\mathcal{M}$ satisfying ϕ:
  - Does **NOT** scale to large systems.

BioLab 2.0

Model Checking Biochemical Stochastic models: $\mathcal{M} \models P_{\geq\theta}(\Phi)$?



Statistical Model Checking

- Decides between two mutually exclusive hypotheses:
  - Null Hypothesis $H_0 : \mathcal{M} \models P_{\geqslant \theta}(\phi)$
  - Alternate Hypothesis $H_1 : \mathcal{M} \models P_{< \theta}(\phi)$

- Statistical tests can determine the true hypothesis:
  - based on sampling the traces of system $\mathcal{M}$
  - answer may be wrong, but error probability is bounded.

- ***Statistical Hypothesis Testing*** ⟹ ***Model Checking!***

Motivation - Scalability

- State Space Exploration often infeasible for complex systems.
  - May be relatively easy to simulate a system
- Our Goal: Provide probabilistic guarantees using fewer simulations
  - How to generate each simulation run?
  - How many simulation runs to generate?
- Applications: BioNetGen, Stateflow / Simulink

BioLab: A Statistical Model Checker for BioNetGen Models.

E. Clarke, C. Langmead, J. Faeder, L. Harris, A. Legay and S. Jha. (*International Conference on Computational Methods in System Biology, 2008*)

Motivation – Parallel Model Checking

- Some success with explicit state Model Checking
- More difficult to distribute Symbolic MC using BDDs.
- Learned Clauses in SAT solving are not easy to distribute.
- Multiple simulations can be easily parallelized.
- Next Generation Model Checking should exploit
  - multiple cores
  - commodity clusters

Existing Work


- [Younes and Simmons 02-06] use Wald's SPRT
  - SPRT: Sequential Probability Ratio Test
- [Hérault et al. 04] use Chernoff bound:
  - Estimate the probability that $\mathcal{M} \models \Phi$
- [Sen et al. 04-05] use *p-value*:
  - "Approximates" the probability that $\mathcal{M} \models P_{\geq\theta}(\Phi)$ is true
- [Grosu and Smolka 05] randomized LTL model checking:
  - Finds counterexamples with high probability
- [Clarke et al. 09] Bayesian approach
  - Both hypothesis testing and estimation
  - Faster (fewer samples required)

Existing Work: SPRT

- [Younes and Simmons 06] use Wald's SPRT
  - SPRT: Sequential Probability Ratio Test
- The SPRT decides between
  - the simple null **hypothesis** $H_0' : \mathcal{M} \models P_{=\theta_0}(\phi)$

    vs
  - the simple alternate **hypothesis** $H_1' : \mathcal{M} \models P_{=\theta_1}(\phi)$
- SPRT is asymptotically optimal (when H_0' or H_1' is true)
  - Minimizes the expected number of samples
  - Among all tests with equal or smaller error probability.

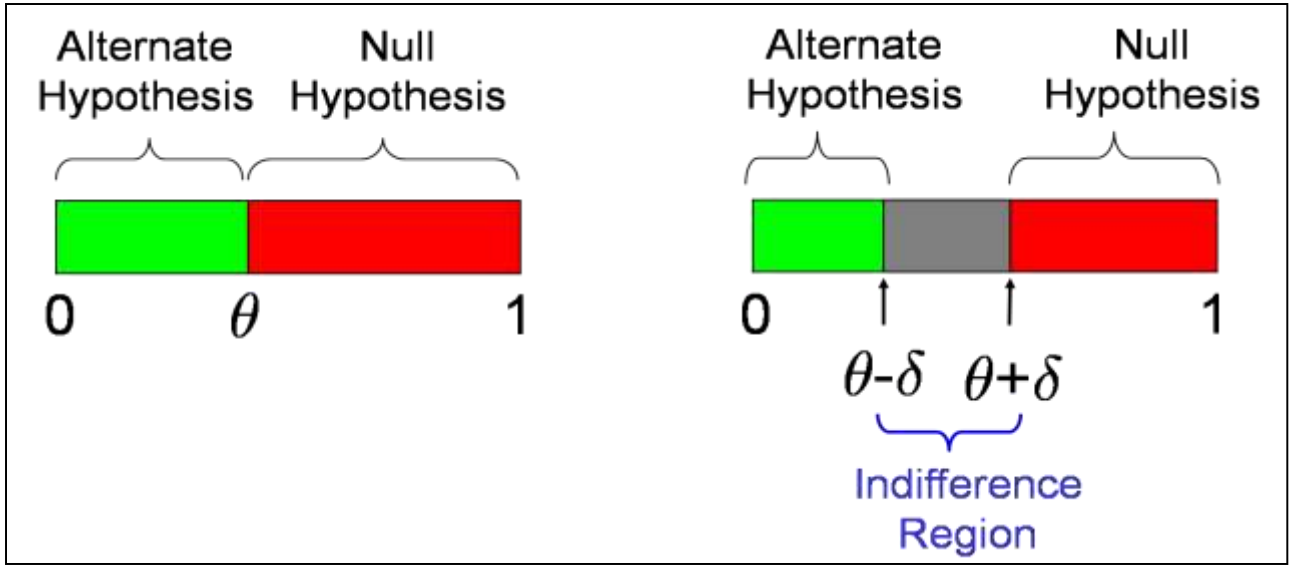
Existing Work: SPRT

- MC chooses between two composite hypotheses

$$H_1 : \mathcal{M} \models P_{<\theta}(\phi) \qquad H_0 : \mathcal{M} \models P_{\geqslant\theta}(\phi)$$

- Existing works use Wald's SPRT for hypothesis testing with an indifference region:

$$\mathcal{M} \models P_{=\theta-\delta}(\phi) \qquad \mathcal{M} \models P_{=\theta+\delta}(\phi)$$



Faster Statistical Model Checking!

- But MC chooses between two mutually exclusive composite hypotheses

  Null Hypothesis $H_0 : \mathcal{M} \models P_{\geqslant \theta}(\phi)$

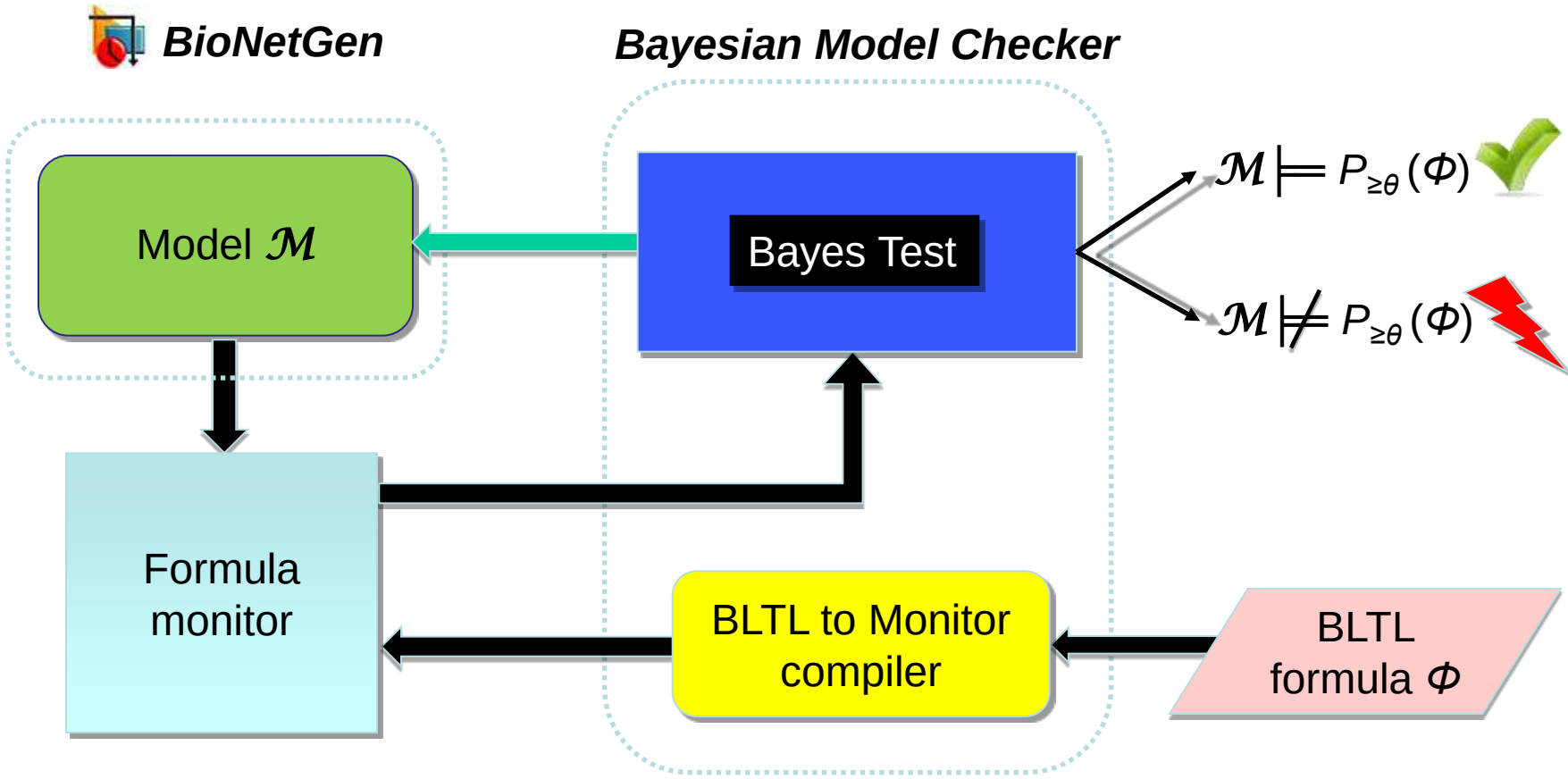
  vs

  Alternate Hypothesis $H_1 : \mathcal{M} \models P_{< \theta}(\phi)$

- We have developed a new statistical MC algorithm
  - Performs Composite Hypothesis Testing
  - Based on Bayes Theorem and the Bayes Factor.

BioLab 2.0

Model Checking Biochemical Stochastic models: $\mathcal{M} \models P_{\geq\theta}(\Phi)$?



Bayesian Statistical Model Checking

- Bayesian Approach to Statistical Model Checking
  - Faster than previous Statistical Model Checking.
- Uses prior knowledge about the model
- Revises prior knowledge in light of experimental data

$$P(H_0 \mid X) = \frac{P(X \mid H_0)P(H_0)}{P(X)}$$

Statistical Model Checking of Stochastic Systems
E. M. Clarke, S. K. Jha, A. Platzer, and P. Zuliani.
CMU CS Technical Report 09-162.

Bayesian Statistical Model Checking 1

- Model Checking $H_0 : \mathcal{M} \models P_{\geqslant \theta}(\phi)$
- Suppose $\mathcal{M}$ satisfies ϕ with (unknown) probability *u.*
  - *u* is given by a random variable *U* with density *g*.
  - *g* represents the prior belief that $\mathcal{M}$ satisfies ϕ.
- Generate independent and identically distributed (iid) sample traces.
- x_i: the ith sample trace σ satisfies ϕ .
  - $x_i = 1$ iff $\sigma_i \models \phi$
  - $x_i = 0$ iff $\sigma_i \not\models \phi$
- Then, x_i will be a Bernoulli trial with density

$$f(x_i|u) = u^{x_i}(1 - u)^{1-x_i}$$

Bayesian Statistical Model Checking 2

- $X = (x_1, \ldots, x_n)$ a sample of Bernoulli random variables.
- Bayes Theorem (Posterior Probability):

$$P(H_0 \mid X) = \frac{P(X \mid H_0)P(H_0)}{P(X)}$$

$$P(H_1 \mid X) = \frac{P(X \mid H_1)P(H_1)}{P(X)}$$

- Ratio of Posterior Probabilities:

$$\frac{P(H_0 \mid X)}{P(H_1 \mid X)} = \frac{P(X \mid H_0)}{P(X \mid H_1)} \frac{P(H_0)}{P(H_1)}$$

Bayes Factor

Bayesian Statistical Model Checking 3

- Bayes Factor: Measure of confidence in H_0 vs H_1
  - provided by the data $X = (x_1, \ldots, x_n)$
  - weighted by the prior *g*.
- Bayes Factor $>$ Threshold1: Accept Null Hypothesis H_0.
- Bayes Factor $<$ Threshold2: Reject Null Hypothesis H_0.

Definition: Bayes Factor $\mathcal{B}$ of sample *X* and hypotheses H_0, H_1

joint distribution of independent events

$$\mathcal{B} = \frac{P(X \mid H_0)}{P(X \mid H_1)} = \frac{\int_\theta^1 f(x_1 \mid u) \cdots f(x_n \mid u) \cdot g(u) du}{\int_0^\theta f(x_1 \mid u) \cdots f(x_n \mid u) \cdot g(u) du}$$

Bayesian Statistical Model Checking 4

Require: ***Property*** $P_{\geq\theta}(\Phi)$***, Threshold*** $T > 1$***, Prior density*** g

$n := 0$ {*number of traces drawn so far*}

$x := 0$ {*number of traces satisfying so far*}

repeat

σ := draw a sample trace of the system (iid)

$n := n + 1$

if $\sigma \models \Phi$ **then**

$x := x + 1$

end if

$\mathcal{B} := BayesFactor(n, x)$

until $(\mathcal{B} > T \ \vee \mathcal{B} < 1/T)$

if $(\mathcal{B} > T)$ **then**

return H_0 *accepted*

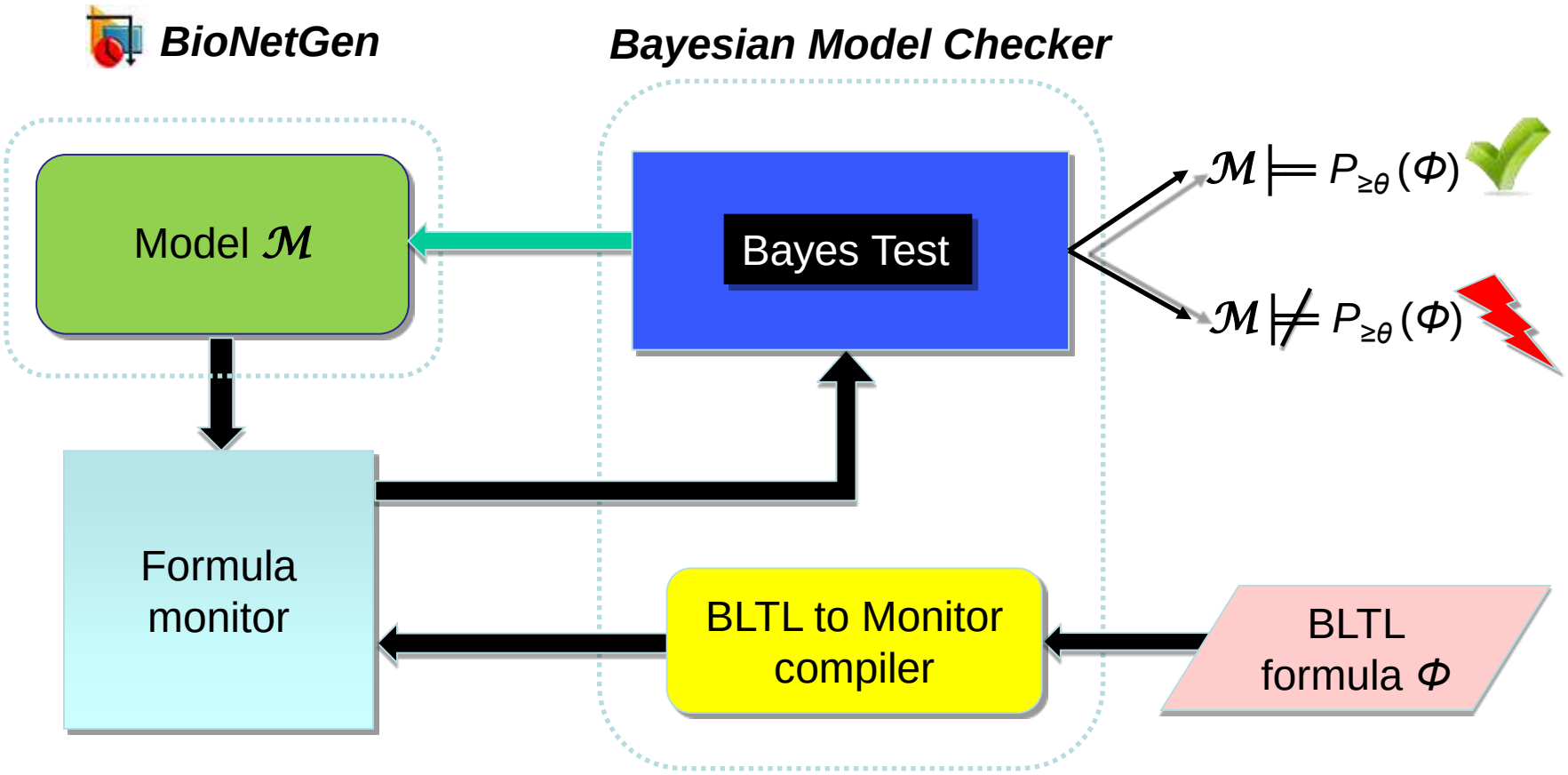
else

return H_1 *accepted*

end if

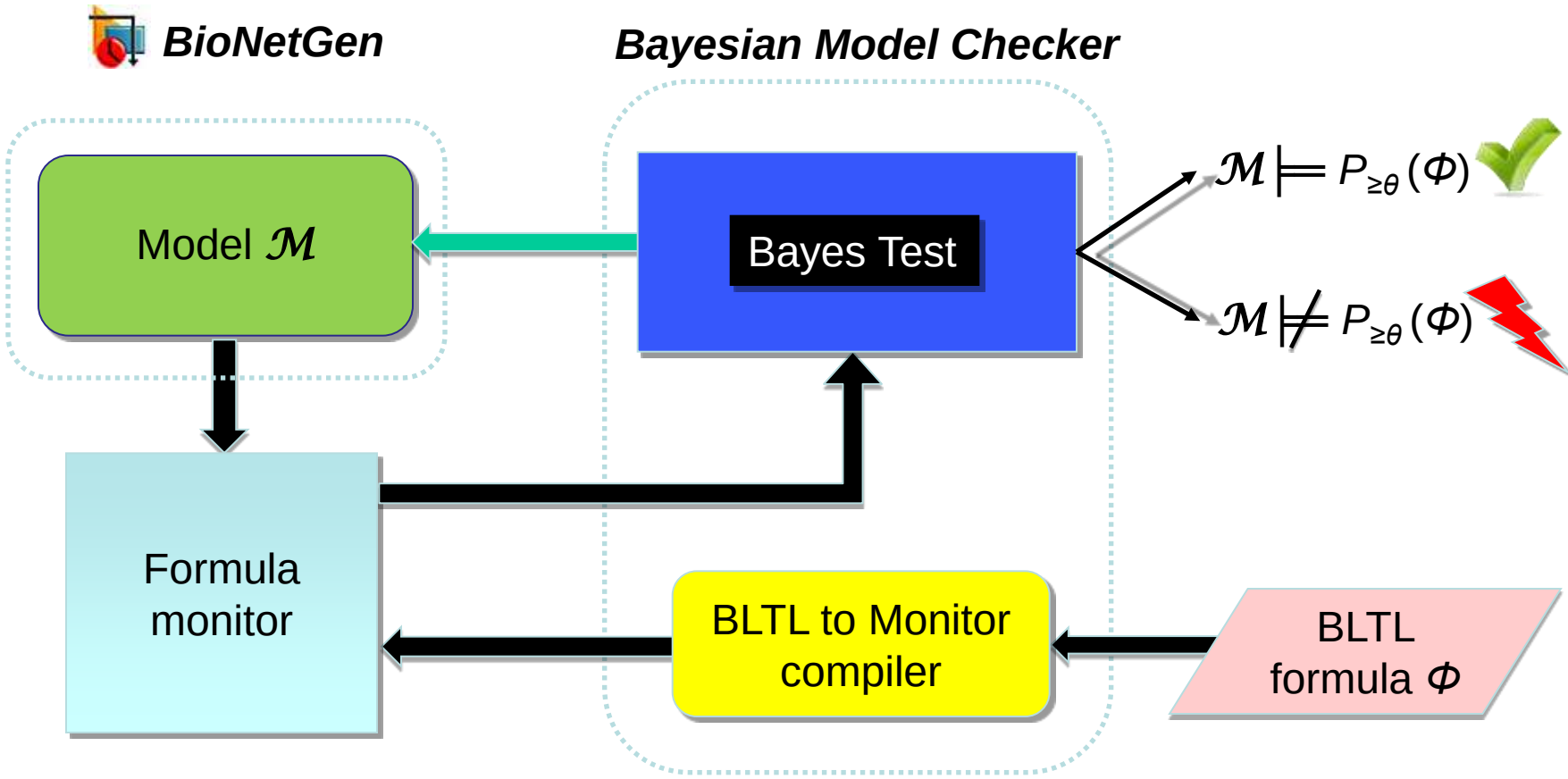
BioLab 2.0

Model Checking Biochemical Stochastic models: $\mathcal{M} \models P_{\geq\theta}(\Phi)$?



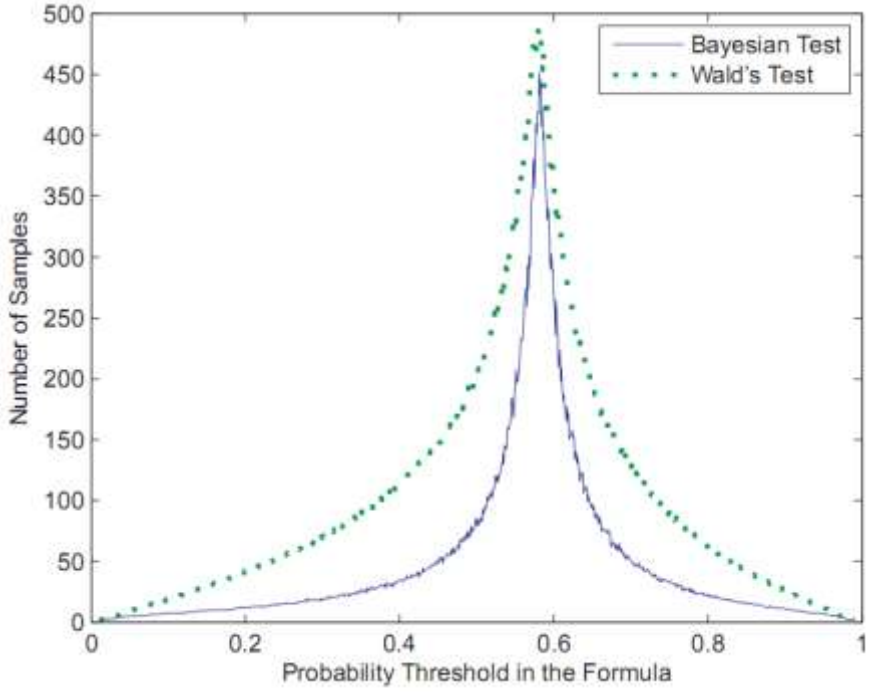
BioLab 2.0

Model Checking Biochemical Stochastic models: $\mathcal{M} \models P_{\geq\theta}(\Phi)$?

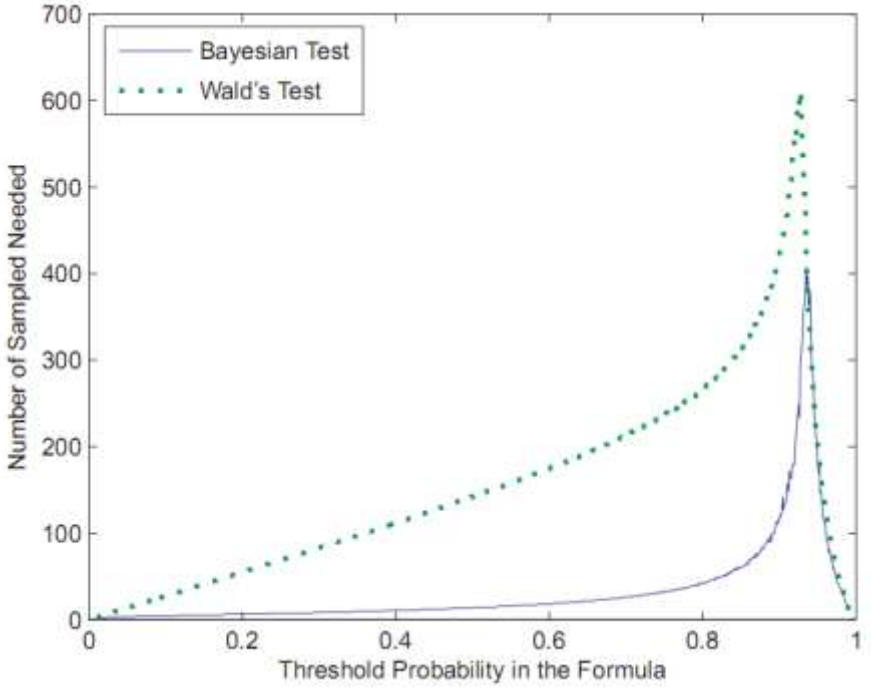


Bayesian Model Checking: Performance

Number of Samples Needed vs. Threshold θ in the Probability Formula



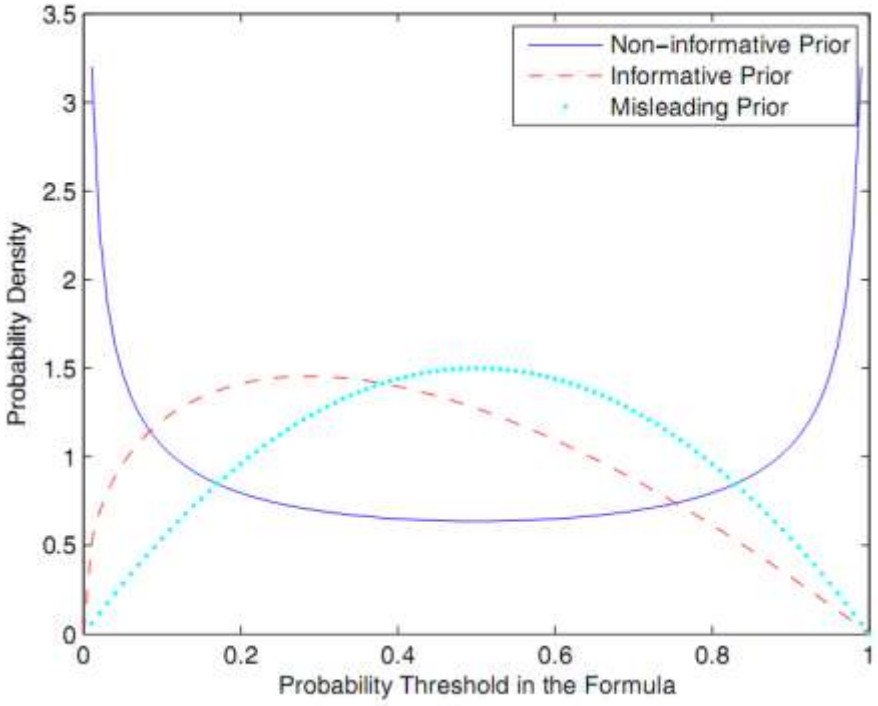
Actual Probability of the Formula being True = 0.58

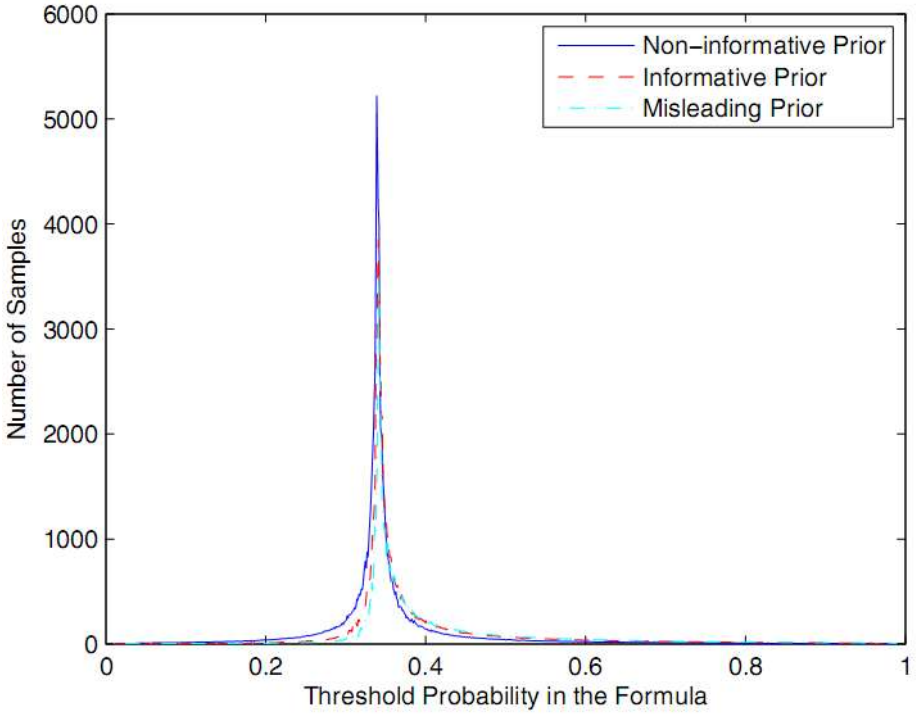


Actual Probability of the Formula being True = 0.93

Bayesian Model Checking: Priors

Number of Samples Needed vs. Different Choices of Prior Probability Distribution





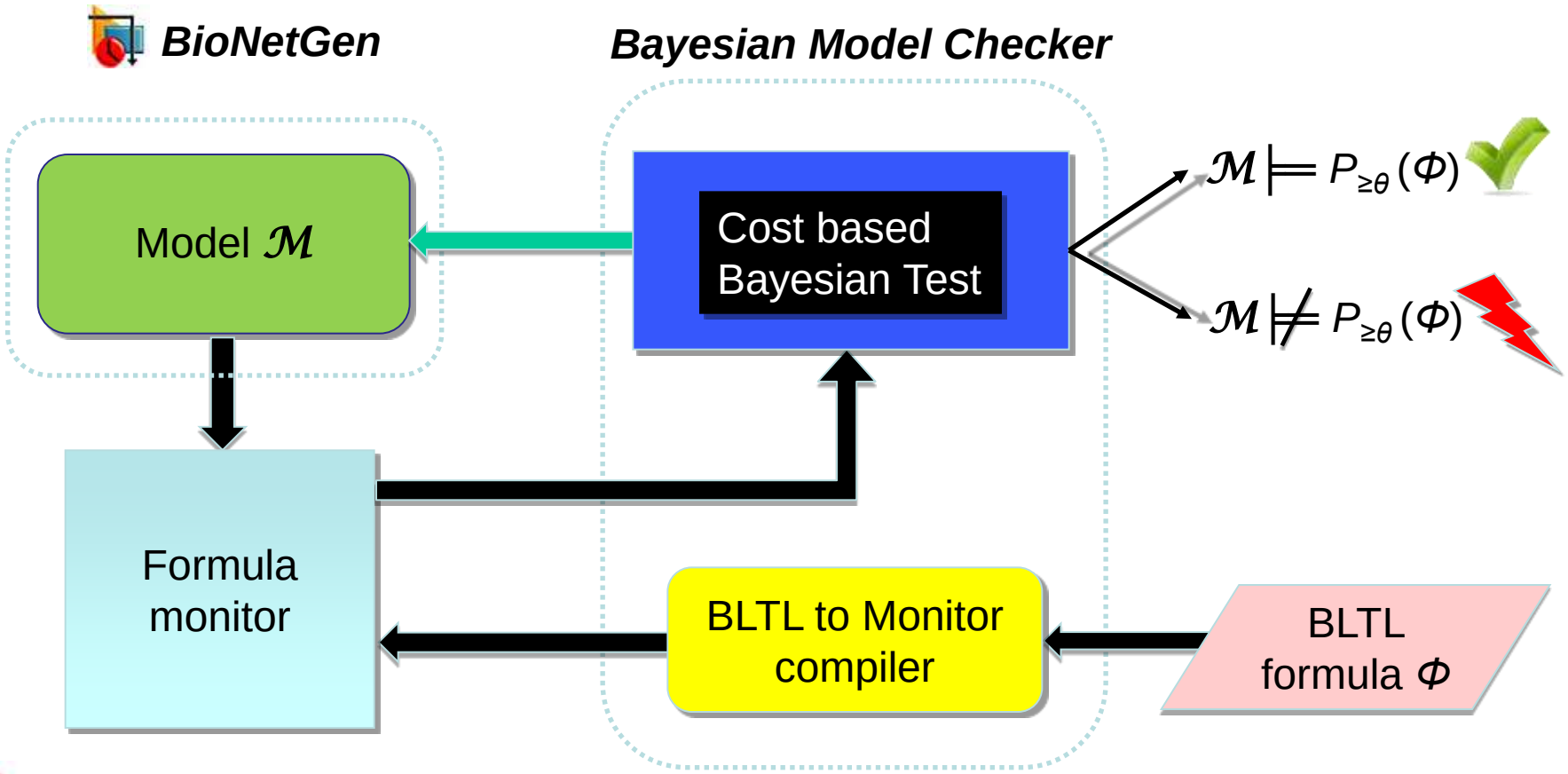
Future Work: Cost-Based Bayesian MC

- Model Checking query: $\mathcal{M} \models P_{\geq\theta}(\Phi)$, for $0 < \theta < 1$.
- $C(N)$: Cost of generating the N^{th} sample.
- $R(u,\theta)$: Cost of incorrectly deciding the MC query
  - u is the (unknown) probability that $\mathcal{M}$ satisfies Φ
  - θ is the probability threshold in the specification
- Then, the key problem is to compute $E[R(u,\theta) \mid X_N]$
  - expected cost of a wrong decision after observing N samples $X_N = (x_1, \ldots, x_N)$
- Stopping Criterion:
  - Stop when cost exceeds the reduction in the expected cost of making a wrong decision.

$$C(N+1) \geq E[R(u,\theta) \mid X_{N+1}] - E[R(u,\theta) \mid X_N]$$

BioLab (upcoming)

Model Checking Biochemical Stochastic models: $\mathcal{M} \models P_{\geq\theta}(\Phi)$?



Conclusions

- Some evidence that Statistical MC scales to large systems
  - BioNetGen Models
  - Matlab Simulink Models
- We have developed a Bayesian MC algorithm which
  - is faster than state-of-the-art approaches,
  - can use prior knowledge about the system.
- Initial experiments on BioNetGen / Matlab models are encouraging.
- Plan:
  - More complex BioNetGen and Stateflow / Simulink models
  - In particular, BioNetGen Models of **Pancreatic Cancer** from **TGen**

The End

Questions?