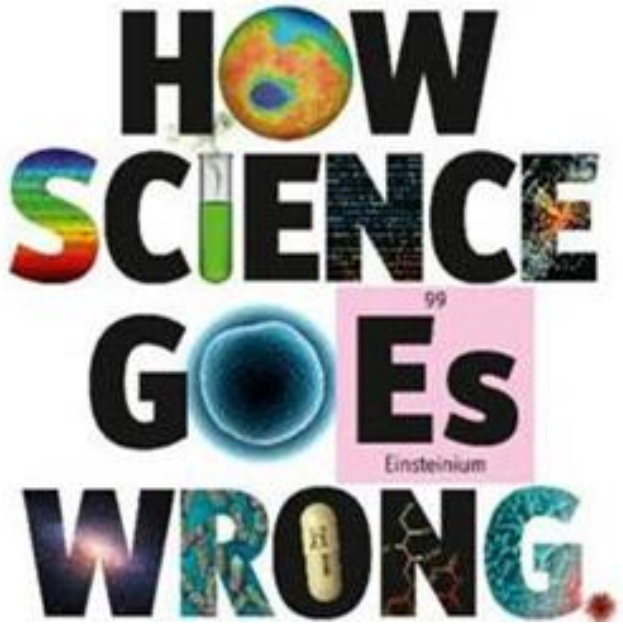
Approaches to Reproducibility in Systems and Physiological Modeling

University of Washington
Herbert Sauro

UNIVERSITY of WASHINGTON

What do we mean by reproducibility?



The results of a scientific experiment are **reproducible** if an **independent** investigator accessing published work can replicate them.

- **Computational repeatability:** a result can be replicated with the same data and software
- **Algorithmic reproducibility:** a result can be replicated with the same data and different software implementing the same algorithm
- **Scientific reproducibility:** a result can be replicated with the same data and a different algorithm
- **Empirical reproducibility:** a result can be replicated with independent data and algorithms

Stronger Claim

The Economist, 2013[2]

Some Definitions:

Repeatability:

From NIST:

In metrology, the component of measurement precision that is the variability in the short term, and that occurs under highly controlled situations (e.g. same metrology instrument, same operator, same setup, same ambient environment, etc.)

From SIX SIGMA:

In **Measurement Systems Analysis**, repeatability is the variation between measurements that occurs when one person measures the same item several times, using the same measuring equipment.

NIST Technical Note 1297
1994 Edition (Supersedes 1993 Edition)
http://physics.nist.gov/Pubs/guidelines/appd.1.html

Some Definitions:

Reproducibility:

From NIST:

In metrology, the total measurement precision, especially including the components of variability that occur in the long term, and occurring from one measurement instrument to another, one laboratory to another, etc.

From SIX SIGMA

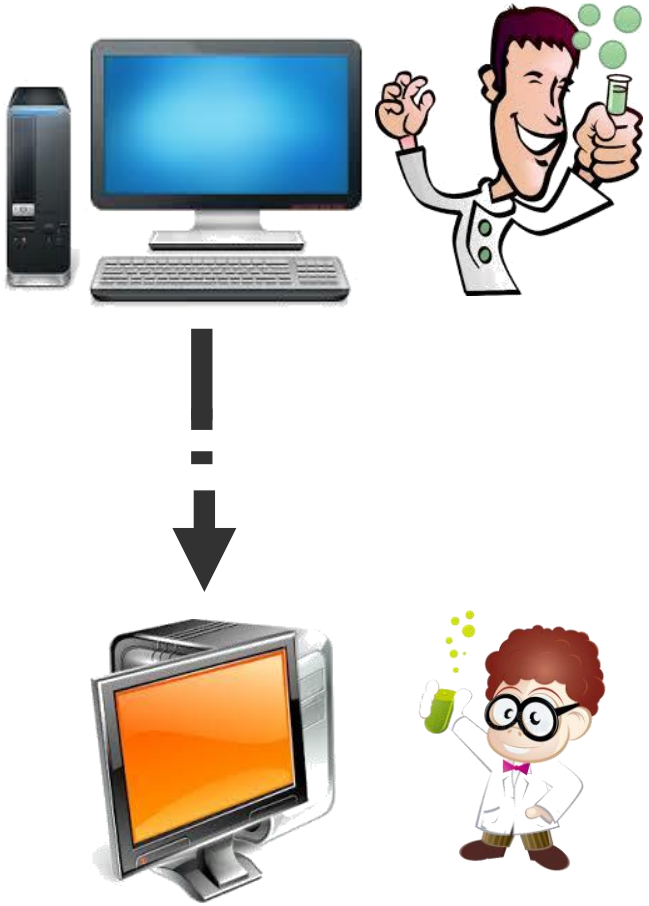
The amount of variation in a measurement system assigned to differences in employees, measurement tools and equipment, techniques, setup or other physical factors. Any factor can be used for reproducibility, but typically employees or measurement tools are the most commonly used variables.

NIST Technical Note 1297
1994 Edition (Supersedes 1993 Edition)
http://physics.nist.gov/Pubs/guidelines/appd.1.html

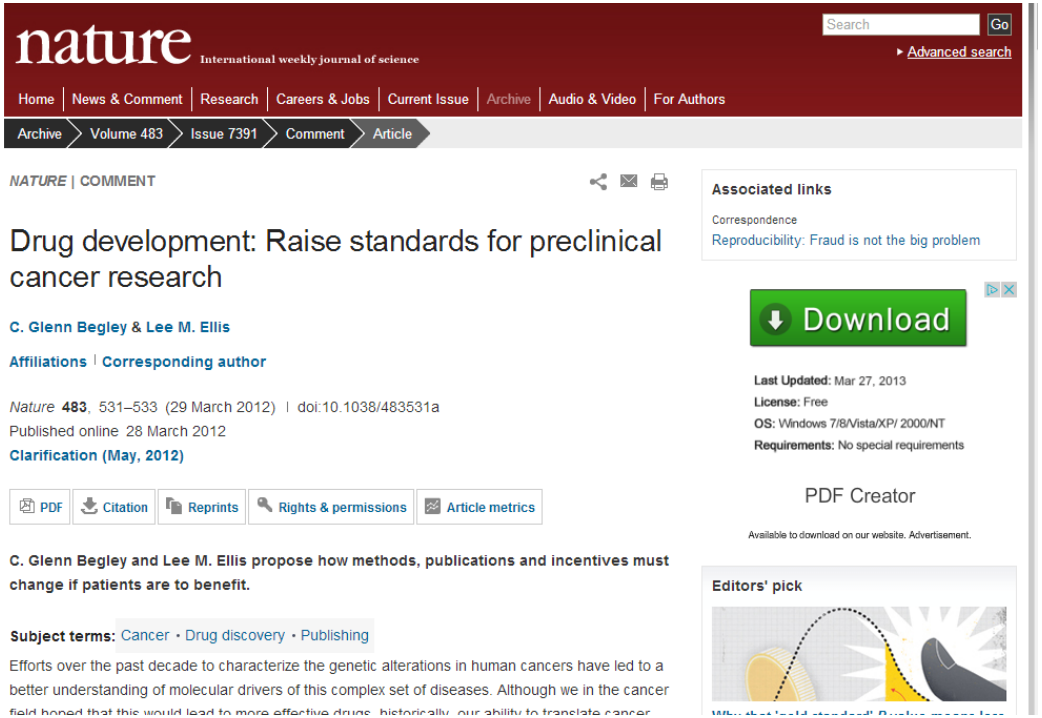
Some Definitions:

Repeatability

Reproducibility



11% of Preclinical Studies Could not be Reproduced

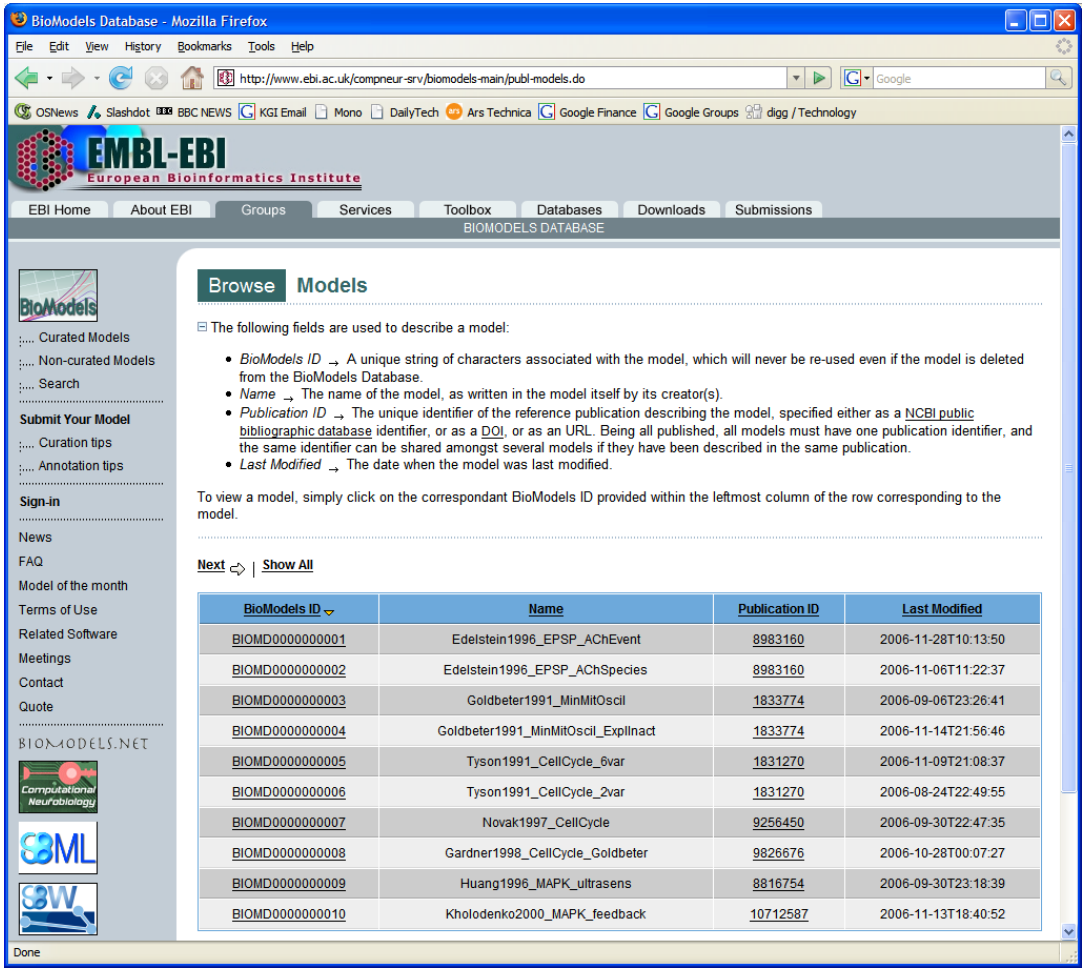


Model repositories

BioModels.net

640 curated models as of June 2017

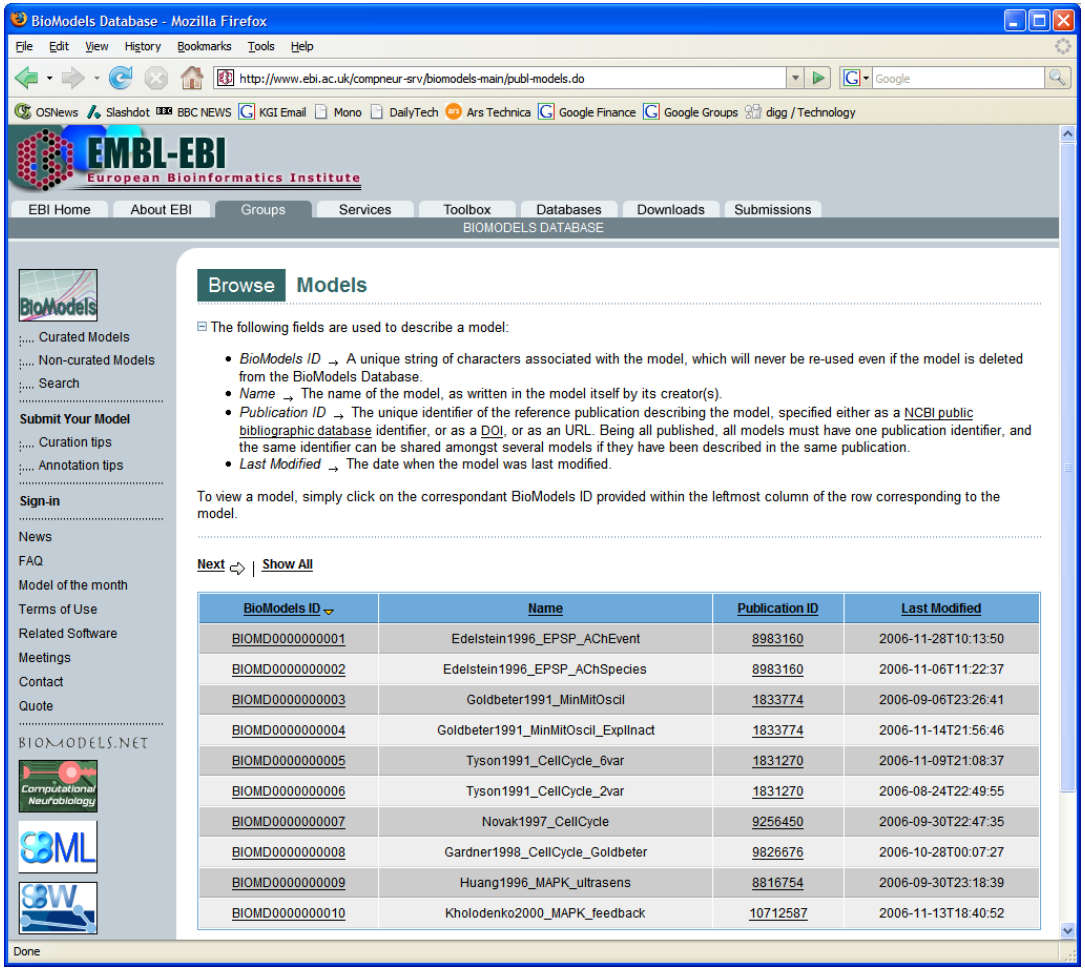
995 Models in total (included non-curated)



Model repositories

BioModels.net

Over 90% of curated models could not be reproduced.



Needle in a Haystack

2006; Goldberg et al. 2007), we examined the affects of variations in the level of expression of $Na_v1.7$ in DRG neurons in which $Na_v1.8$ is present.

MATERIALS AND METHODS

Computer simulations. The electrical properties of small sensory neurons were simulated using the NEURON program (version 7.1) (Hines and Carnevale 1997). Sodium and potassium conductances in the present study incorporated parameters described in previous reports (Herzog et al. 2001; Sheets et al. 2007). Conductances were modeled using Hodgkin and Huxley-type (HH) descriptions (Hodgkin and Huxley 1952) of the various voltage-dependent currents. For analysis of theoretical electric charge movements passing though model sodium channels, we calculated area under the sodium current using OriginPro 8.1 software (Microcal, Northampton, MA).

Passive membrane properties of the model neuron. Action potential firing was studied using a single compartment cylindrical model of length 30 μm and radius 23 μm, simulating a small sensory neuron with a 2,168 μm^2 surface area and 20.2 pF/cm^2 capacitance, based on electrically and microscopically measured values (Choi et al. 2007). The specific resistance of the cytosol was set to 123 Ω/cm. Simulations were performed assuming a temperature of 22°C, the temperature at which the experimental data were recorded (Choi et al. 2007). The integration method was Backward Euler at an

K_{DR} potassium current. The K_{DR} current was defined as: $I_{KDR} = g_{KDR} * n * (V - E_k)$, where g_{KDR} is the delayed rectifier potassium conductance and n is a dimensionless activation variable that varies between 0 and 1. The kinetic characterization of the channel described by Schild et al. (1994) has been used with alpha$_n$ = 0.001265 * (V + 14.273)/{1 − exp[(V + 14.273)/−10]}; beta$_n$ = 0.125 * exp(V + 55/−2.5); and n_{inf} = 1/{1 + exp[(V + 14.62)/−18.38]}. The peak conductance for K_{DR} (g_{KDR}) was set to 0.0035 S/cm^2, which corresponds to 6 nA potassium current at 0 mV.

K_A potassium current. The K_A current was defined as: $I_{KA} = g_{KA} * n * h * (V - E_k)$, where g_{KA} is the A-type potassium conductance and n and h are dimensionless activation and inactivation variables, respectively, that vary between 0 and 1. The kinetic characterization of the channel described by Gold et al. (1996b) has been used with $dn/dt = (n_{inf} - n)/n_{tau}$; $dh/dt = (h_{inf} - h)/h_{tau}$; n_{inf} = (1/{1 + exp[−(v + 5.4)/16.4]}) ^ 4; n_{tau}= (0.25 + 10.04 * exp(−[(v + 24.67) ^ 2]/(2 * 34.8 ^ 2))); h_{inf} = 1/{1 + exp[(v + 49.9)/4.6]}; h_{tau} = (20 + 50 * exp{− [(v + 40) ^ 2]/(2 * 40 ^ 2)}); if $h_{tau} < 5$ then $h_{tau} = 5$. The peak conductance for K_A (g_{KA}) was set to 0.0055 S/cm^2, which corresponds to 1 nA potassium current at 0 mV.

$Na_v1.7$ sodium current. The TTX-S sodium conductance from small DRG neurons was fitted to the conventional HH model for sodium conductance: $I_{Nav1.7} = g_{Nav1.7} * m * m * m * h * s * (V - E_{Na})$, where $g_{Nav1.7}$ is the WT fast inactivating $Na_v1.7$ sodium conductance and m, h, and s are dimensionless activation, fast inac-

respectively. The size of the current was adjusted so that its amplitude corresponded to an input resistance of 579 MΩ: g_{Leak} = 0.0000575 S cm^{-2}, similar to that reported for small DRG neurons (Choi et al. 2007).

Voltage-dependent currents. The DRG neuron model included a leak conductance, two potassium conductances (A-type and delayed rectifier), and, since we wanted to study the interactions of $Na_v1.7$ and $Na_v1.8$, two voltage-sensitive sodium conductances: $Na_v1.7$ conductance and $Na_v1.8$ conductance. We included a delayed rectifier conductance and a transient potassium conductance to reflect the predominant potassium conductances in small DRG neurons, a sustained (delayed rectifier type, K_{DR}) conductance and a transient (A-type, K_A) conductance (Gold et al. 1996b). The majority of small (<25 μm diameter) DRG neurons exhibit both TTX-S and -R currents (Cummins and Waxman 1997). Although many small DRG neurons express more than one TTX-S sodium channel isoform (Black et al. 1996), the major TTX-S current in the majority of small DRG neurons appears to be produced by $Na_v1.7$ (Cummins et al. 1998). For this reason, and because we wanted to study $Na_v1.7$ and $Na_v1.8$ in isolation, the only TTX-S conductance included in our model simulates $Na_v1.7$. Two TTX-R channels, slowly inactivating $Na_v1.8$ and persistent $Na_v1.9$ are present in small DRG neurons (Cummins et al. 1999). The persistent TTX-R current is largely inactivated by ultra-slow inactivation and tends to be negligible when the cells are held at approximately −60 mV (Cummins et al. 1999). Because of this, and because we wanted to study $Na_v1.7$ and $Na_v1.8$ in isolation, we only included $Na_v1.8$ as the sole TTX-R current in our model neuron. Specific details of the current models are given below. Throughout the text, we refer to the level of

best fit with a HH model that employed only one activation gate: $I_{Nav1.8} = g_{Nav1.8} * m * h * (V - E_{Na})$, where $g_{Nav1.8}$ is the $Na_v1.8$ sodium conductance and m and h are dimensionless activation and inactivation variables that vary between 0 and 1. Based on previous reports (Herzog et al. 2001; Sheets et al. 2007), we defined the following equations for $Na_v1.8$ m and h: $m = m + [1 - \exp(-dt/tau_m)] * (m_{inf} - m)$, $h = h + [1 - \exp(-dt/tau_h)] * (h_{inf} - h)$, alpha$_m$ = 2.85 − (2.839)/{1 + exp [(v − 1.159)/13.95]}; beta$_m$ = (7.6205)/{1 + exp[(v + 46.463)/8.8289]}; tau$_m$ = 1/(alpha$_m$ + beta$_m$); m_{inf} = alpha$_m$/(alpha$_m$ + beta$_m$); tau$_h$ = (1.218 + 42.043 * exp{−[(v + 38.1) ^ 2]/(2 * 15.19 ^ 2)}); h_{inf} = 1/{1 + exp[(v + 32.2)/4]}. The peak current of 25 nA was modeled by setting the peak value $g_{Nav1.8}$ to 0.026 S/cm^2, which was chosen to match experimental values (Choi et al. 2007; Cummins and Waxman 1997).

RESULTS

To investigate how the levels of expression of $Na_v1.7$ and $Na_v1.8$ affect the excitability of DRG neurons, we simulated sensory neuron membrane conductances and analyzed firing properties in the single compartment model using the NEURON modeling program (Hines and Carnevale 1997). We used four HH-type ion channel conductances ($Na_v1.7$, $Na_v1.8$, K_{DR}, and K_A) that have been previously described in the literature (Gold et al. 1996b; Herzog et al. 2001; Schild et al. 1994; Sheets et al. 2007) and that are shown in Fig. 1. The current amplitudes were adjusted to match maximum conductances to the results of previous reports (Choi et al. 2007; Yang et al. 2004). The resultant peak currents for Na 1.7 and Na 1.8 were 15.2 nA at

Physiological interactions between $Na_v1.7$ and $Na_v1.8$ sodium channels: a computer simulation study

Can you spot the error?

Published 2011

Needle in a Haystack

MATERIALS AND METHODS

Computer simulations. The electrical properties of small sensory neurons were simulated using the NEURON program (version 7.1) (Hines and Carnevale 1997). Sodium and potassium conductances in the present study incorporated parameters described in previous reports (Herzog et al. 2001; Sheets et al. 2007). Conductances were modeled using Hodgkin and Huxley-type (HH) descriptions (Hodgkin and Huxley 1952) of the various voltage-dependent currents. For analysis of theoretical electric charge movements passing though model sodium channels, we calculated area under the sodium current using OriginPro 8.1 software (Microcal, Northampton, MA).

Passive membrane properties of the model neuron. Action potential firing was studied using a single compartment cylindrical model of length 30 μm and radius 23 μm, simulating a small sensory neuron with a 2,168 μm^2 surface area and 20.2 pF/cm^2 capacitance, based on electrically and microscopically measured values (Choi et al. 2007). The specific resistance of the cytosol was set to 123 Ω/cm. Simulations were performed assuming a temperature of 22°C, the temperature at which the experimental data were recorded (Choi et al. 2007). The integration method was Backward Euler at an integration time step dt of 0.025 ms. Simulations were performed assuming free ionic concentrations of sodium ($[Na^+]_o$ = 140 mM; $[Na^+]_i$ = 10 mM) and potassium ($[K^+]_o$ = 5 mM; $[K^+]_i$ = 140 mM), which were used to calculate Nernst reversal potentials of +67.1 mV (E_{Na}) and −84.7 mV (E_K), respectively. By analogy to the HH model of action potential electrogenesis, the linear leakage current was defined as $I_{Leak} = g_{Leak} * (V - E_{Leak})$, where g_{Leak} is the leak conductance, V is the membrane potential, and E is the reversal

Can you spot the error?

Needle in a Haystack

MATERIALS AND METHODS

Computer simulations. The electrical properties of small sensory neurons were simulated using the NEURON program (version 7.1) (Hines and Carnevale 1997). Sodium and potassium conductances in the present study incorporated parameters described in previous reports (Herzog et al. 2001; Sheets et al. 2007). Conductances were modeled using Hodgkin and Huxley-type (HH) descriptions (Hodgkin and Huxley 1952) of the various voltage-dependent currents. For analysis of theoretical electric charge movements passing though model sodium channels, we calculated area under the sodium current using OriginPro 8.1 software (Microcal, Northampton, MA).

Passive membrane properties of the model neuron. Action potential firing was studied using a single compartment cylindrical model of length 30 μm and radius 23 μm, simulating a small sensory neuron with a 2,168 μm^2 surface area and 20.2 pF/cm^2 capacitance, based on electrically and microscopically measured values (Choi et al. 2007). The specific resistance of the cytosol was set to 123 Ω/cm. Simulations were performed assuming a temperature of 22°C, the temperature at which the experimental data were recorded (Choi et al. 2007). The integration method was Backward Euler at an integration time step dt of 0.025 ms. Simulations were performed assuming free ionic concentrations of sodium ($[Na^+]_o$ = 140 mM; $[Na^+]_i$ = 10 mM) and potassium ($[K^+]_o$ = 5 mM; $[K^+]_i$ = 140 mM), which were used to calculate Nernst reversal potentials of +67.1 mV (E_{Na}) and −84.7 mV (E_K), respectively. By analogy to the HH model of action potential electrogenesis, the linear leakage current was defined as $I_{Leak} = g_{Leak} * (V - E_{Leak})$, where g_{Leak} is the leak conductance, V is the membrane potential, and $E_{\quad}$ is the reversal

Can you spot the error?

Needle in a Haystack

MATERIALS AND METHODS

Computer simulations. The electrical properties of small sensory neurons were simulated using the NEURON program (version 7.1) (Hines and Carnevale 1997). Sodium and potassium conductances in the present study incorporated parameters described in previous reports (Herzog et al. 2001; Sheets et al. 2007). Conductances were modeled using Hodgkin and Huxley-type (HH) descriptions (Hodgkin and Huxley 1952) of the various voltage-dependent currents. For analysis of theoretical electric charge movements passing though model sodium channels, we calculated area under the sodium current using OriginPro 8.1 software (Microcal, Northampton, MA).

Passive membrane properties of the model neuron. Action potential firing was studied using a single compartment cylindrical model of length 30 μm and radius 23 μm, simulating a small sensory neuron with a 2,168 μm^2 surface area and 20.2 pF/cm^2 capacitance, based on electrically and microscopically measured values (Choi et al. 2007). The specific resistance of the cytosol was set to 123 Ω/cm. Simulations were performed assuming a temperature of 22°C, the temperature at which the experimental data were recorded (Choi et al. 2007). The integration method was Backward Euler at an integration time step dt of 0.025 ms. Simulations were performed assuming free ionic concentrations of sodium ($[Na^+]_o$ = 140 mM; $[Na^+]_i$ = 10 mM) and potassium ($[K^+]_o$ = 5 mM; $[K^+]_i$ = 140 mM), which were used to calculate Nernst reversal potentials of +67.1 mV (E_{Na}) and −84.7 mV (E_K), respectively. By analogy to the HH model of action potential electrogenesis, the linear leakage current was defined as $I_{Leak} = g_{Leak} * (V - E_{Leak})$, where g_{Leak} is the leak conductance, V is the membrane potential, and $E_{\quad}$ is the reversal

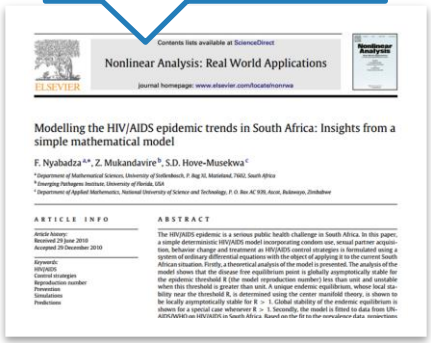
Can you spot the error?

23 um refers to the diameter

The area 2,168 um2 does not include the ends.

Common Problems in Reproducible Models

Incomplete parameters


Contents lists available at ScienceDirect

Nonlinear Analysis: Real World Applications

journal homepage: www.elsevier.com/locate/nonrwa

Modelling the HIV/AIDS epidemic trends in South Africa: Insights from a simple mathematical model

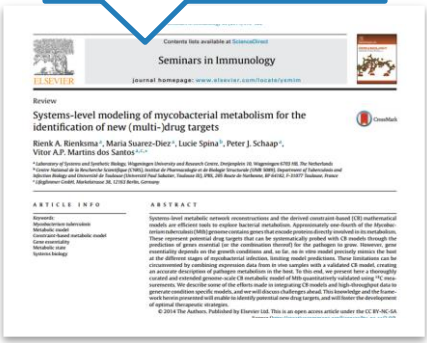
F. Nyabadza[a,*], Z. Mukandavire[b], S.D. Hove-Musekwa[c]

ARTICLE INFO

ABSTRACT

The HIV/AIDS epidemic is a serious public health challenge in South Africa. In this paper, a simple deterministic HIV/AIDS model incorporating condom use, sexual partner acquisition, behavior change and treatment as HIV/AIDS control strategies is formulated using a system of ordinary differential equations with the object of applying it to the current South African situation. Firstly, a theoretical analysis of the model is presented. The analysis of the model shows that the disease free equilibrium point is globally asymptotically stable for the epidemic threshold R (the model reproduction number) less than unit and unstable when this threshold is greater than unit. A unique endemic equilibrium, whose local stability near the threshold R, is determined using the center manifold theory, is shown to be locally asymptotically stable for R > 1. Global stability of the endemic equilibrium is shown for a special case whenever R > 1. Secondly, the model is fitted to data from UNAIDS/WHO on HIV/AIDS in South Africa. Based on the fit to the prevalence data, numerical

Incomplete model definition


Contents lists available at ScienceDirect

Seminars in Immunology

journal homepage: www.elsevier.com/locate/ysmim

Review

Systems-level modeling of mycobacterial metabolism for the identification of new (multi-)drug targets

Rienk A. Rienksma[a], Maria Suarez-Diez[a], Lucie Spina[b], Peter J. Schaap[a,*], Vitor A.P. Martins dos Santos[a,c,d,e]

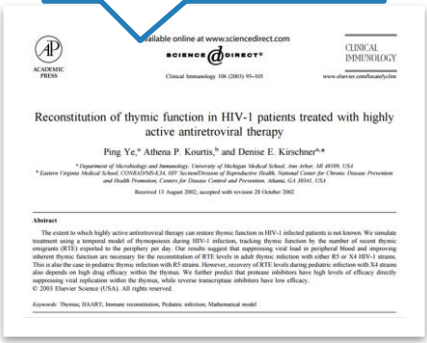
ARTICLE INFO

ABSTRACT

Systems-level metabolic network reconstructions and the derived constraint-based (CB) mathematical models are efficient tools to explore bacterial metabolism. Approximately one-fourth of the Mycobacterium tuberculosis (Mtb) genome contains genes that encode proteins directly involved in its metabolism. These represent potential drug targets that can be systematically probed with CB models through the prediction of genes essential for the Mtb metabolism that may also be essential for pathogen survival. However, the predictive accuracy of such models can be limited by the conditions set for, in vitro simulations. Some of these limitations can be circumvented by combining expression data from in vivo samples with a validated CB model, creating an accurate description of pathogen metabolism in the host. To this end, we present here a thoroughly curated and extended genome-scale CB metabolic model of Mtb quantitatively validated using ^{13}C measurements. We describe some of the efforts made in integrating CB models and high-throughput data to generate conditions specific models, and we will discuss challenges ahead. This knowledge and the framework herein presented will enable to identify potential new drug targets, and will foster the development of optimal therapeutic strategies.

© 2014 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-SA

No parameter values


available online at www.sciencedirect.com

SCIENCE DIRECT

CLINICAL IMMUNOLOGY

Clinical Immunology 106 (2003) 95–105

Reconstitution of thymic function in HIV-1 patients treated with highly active antiretroviral therapy

Ping Ye,[a] Athena P. Kourtis,[b] and Denise E. Kirschner[a,*]

Received 13 August 2002; accepted with revision 28 October 2002

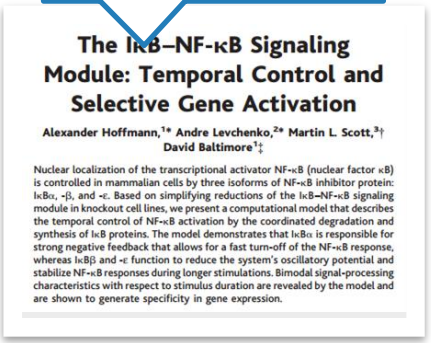
Abstract

The extent to which highly active antiretroviral therapy can restore thymic function in HIV-1 infected patients is not known. We simulate treatment using a temporal model of thymopoiesis during HIV-1 infection, tracking thymic function by the number of recent thymic emigrants (RTE) exported to the periphery per day. Our results suggest that suppressing viral load in peripheral blood and improving inherent thymic function are necessary for the reconstitution of RTE levels in adult thymic infection with either R5 or X4 HIV-1 strains. This is also the case in pediatric thymic infection with R5 strains. However, recovery of RTE levels during pediatric infection with X4 strains also depends on high drug efficacy within the thymus. We further predict that protease inhibitors have high levels of efficacy directly suppressing viral replication within the thymus, while reverse transcriptase inhibitors have low efficacy.

© 2003 Elsevier Science (USA). All rights reserved.

Keywords: Thymus; HAART; Immune reconstitution; Pediatric infection; Mathematical model

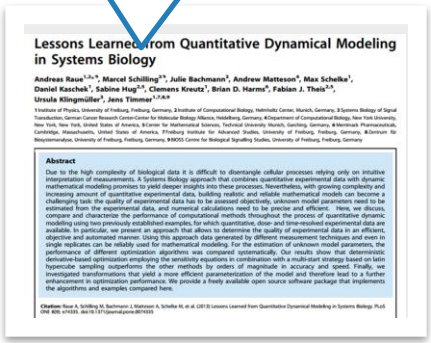
Irreproducible


The IκB–NF-κB Signaling Module: Temporal Control and Selective Gene Activation

Alexander Hoffmann,[1*] Andre Levchenko,[2*] Martin L. Scott,[3†] David Baltimore[1‡]

Nuclear localization of the transcriptional activator NF-κB (nuclear factor κB) is controlled in mammalian cells by three isoforms of NF-κB inhibitor protein: IκBα, -β, and -ε. Based on simplifying reductions of the IκB–NF-κB signaling module in knockout cell lines, we present a computational model that describes the temporal control of NF-κB activation by the coordinated degradation and synthesis of IκB proteins. The model demonstrates that IκBα is responsible for strong negative feedback that allows for a fast turn-off of the NF-κB response, whereas IκBβ and -ε function to reduce the system's oscillatory potential and stabilize NF-κB responses during longer stimulations. Bimodal signal-processing characteristics with respect to stimulus duration are revealed by the model and are shown to generate specificity in gene expression.

Undescribed analysis procedure


Lessons Learned from Quantitative Dynamical Modeling in Systems Biology

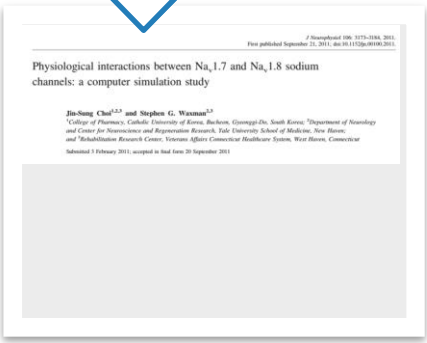
Andreas Raue[1,2,3,*], Marcel Schilling[4,5], Julie Bachmann[1], Andrew Matteson[6], Max Schelke[7], Daniel Kaschek[1], Sabine Hug[2,8], Clemens Kreutz[2], Brian D. Harms[6], Fabian J. Theis[2,9], Ursula Klingmüller[3], Jens Timmer[1,7,8,9]

Abstract

Due to the high complexity of biological data it is difficult to disentangle cellular processes relying only on intuitive interpretation of measurements. A Systems Biology approach that combines quantitative experimental data with dynamic mathematical modeling promises to yield deeper insights into these processes. Nevertheless, with growing complexity and increasing amount of quantitative experimental data, building realistic and reliable mathematical models can become a challenging task: the quality of experimental data has to be assessed objectively, unknown model parameters need to be estimated from the experimental data, and numerical calculations need to be precise and efficient. Here, we discuss, compare and characterize the performance of computational methods throughout the process of quantitative dynamical modeling using two previously established examples, for which quantitative, dose- and time-resolved experimental data are available. In particular, we present an approach that allows to determine the quality of experimental data in an efficient, objective and automated manner. Using this approach data generated by different measurement techniques and even in single replicates can be reliably used for mathematical modeling. For the estimation of unknown model parameters, the performance of different optimization algorithms was compared systematically. Our results show that deterministic derivative-based optimization employing the sensitivity equations in combination with a multi-start strategy based on latin hypercube sampling outperforms the other methods by orders of magnitude in accuracy and speed. Finally, we investigated transformations that yield a more efficient parameterization of the model and therefore lead to a further enhancement in optimization performance. We provide a freely available open source software package that implements the algorithms and examples compared here.

Citation: Raue A, Schilling M, Bachmann J, Matteson A, Schelke M, et al. (2013) Lessons Learned from Quantitative Dynamical Modeling in Systems Biology. PLoS ONE 8(9): e74335. doi:10.1371/journal.pone.0074335

Parameter incorrectly annotated

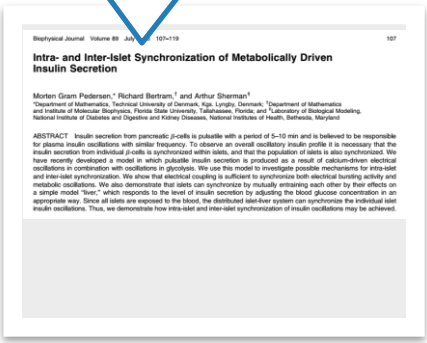

J Neurophysiol 106: 3173–3184, 2011.
First published September 21, 2011; doi:10.1152/jn.00160.2011.

Physiological interactions between $Na_v1.7$ and $Na_v1.8$ sodium channels: a computer simulation study

Jin-Sung Choi[1,2,3] and Stephen G. Waxman[2,3]

Submitted 3 February 2011; accepted in final form 20 September 2011

Irreproducible

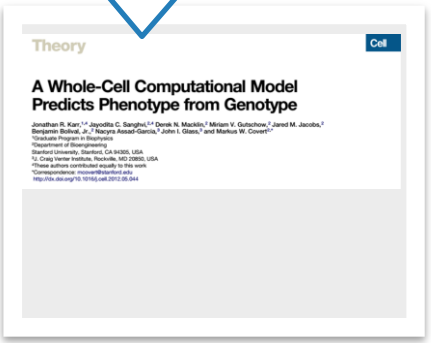

Biophysical Journal Volume 89 July 2005 107–119

Intra- and Inter-Islet Synchronization of Metabolically Driven Insulin Secretion

Morten Gram Pedersen,* Richard Bertram,† and Arthur Sherman‡

ABSTRACT Insulin secretion from pancreatic β-cells is pulsatile with a period of 5–10 min and is believed to be responsible for plasma insulin oscillations with similar frequency. To observe an overall oscillatory insulin profile it is necessary that the insulin secretion from individual β-cells is synchronized within islets, and that the population of islets is also synchronized. We have recently developed a model in which pulsatile insulin secretion is produced as a result of calcium-driven electrical oscillations in combination with oscillations in glycolysis. We use this model to investigate possible mechanisms for intra-islet and inter-islet synchronization. We show that electrical coupling is sufficient to synchronize both electrical bursting activity and metabolic oscillations. We also demonstrate that islets can synchronize by mutually entraining each other by their effects on a simple model "liver," which responds to the level of insulin secretion by adjusting the blood glucose concentration in an appropriate way. Since all islets are exposed to the blood, the distributed islet-liver system can synchronize the individual islet insulin oscillations. Thus, we demonstrate how intra-islet and inter-islet synchronization of insulin oscillations may be achieved.

No language for describing large models


Theory

Cell

A Whole-Cell Computational Model Predicts Phenotype from Genotype

Jonathan R. Karr,[1,4] Jayodita C. Sanghvi,[2,4] Derek N. Macklin,[2] Miriam V. Gutschow,[2] Jared M. Jacobs,[2] Benjamin Bolival, Jr.,[2] Nacyra Assad-Garcia,[3] John I. Glass,[3] and Markus W. Covert[2,*]

[1]Graduate Program in Biophysics
[2]Department of Bioengineering
Stanford University, Stanford, CA 94305, USA
[3]J. Craig Venter Institute, Rockville, MD 20850, USA
[4]These authors contributed equally to this work
[*]Correspondence: mcovert@stanford.edu
http://dx.doi.org/10.1016/j.cell.2012.05.044

Why are computational models not reproducible?

1. Missing data
2. Incorrect data (units wrong, values wrong)
3. Undefined terms/graph axes
4. Mismatch between text and model
5. Wrong model supplied with paper
6. Only one model supplied but multiple simulations described
7. Simulation environment no longer available
8. Model no longer available (url points to null)
9. Model only supplied as a binary

Replicating Computational Experiments

Laboratory of Genomics, Evolution and Development
Michigan State University

Title: A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data

C. Titus Brown, Adina Howe, Qingpeng Zhang, Alexis B. Pyrkosz, and Timothy H. Brom

arXiv preprint

Deep shotgun sequencing and analysis of genomes, transcriptomes, amplified single-cell genomes, and metagenomes has enabled investigation of a wide range of organisms and ecosystems. However, sampling variation in short-read data sets and high sequencing error rates of modern sequencers present many new computational challenges in data interpretation. These challenges have led to the development of new classes of mapping tools and *de novo* assemblers. These algorithms are challenged by the continued improvement in sequencing throughput. We here describe digital normalization, a single-pass computational algorithm that systematizes coverage in shotgun sequencing data sets, thereby decreasing sampling variation, discarding redundant data, and removing the majority of errors. Digital normalization substantially reduces the size of shotgun data sets and decreases the memory and time requirements for *de novo* sequence assembly, all without significantly impacting content of the generated contigs. We apply digital normalization to the assembly of microbial genomic data, amplified single-cell genomic data, and transcriptomic data. Our implementation is freely available for use and modification.

Online resources and data:

Replicating Computational Experiments

Online resources and data:

- A tutorial for running khmer on microbial genomes and eukaryotic transcriptomes.
- Git repository for khmer: github.com/ged-lab/khmer/tree/2012-paper-diginorm
- Git repository for paper & data analysis pipeline: github.com/ged-lab/2012-paper-diginorm
- Instructions on running the paper analysis pipeline & reproducing the paper
- HTML view of the ipython notebook containing code and scripts to reproduce the figures in the paper. (See the pipeline notes for a runnable version.)
- Data required to run the pipeline (.tar.gz, 7.9gb)
- Assembled microbial genomes and eukaryotic transcriptomes (.tar.gz, 110 mb)

The entire analysis is replicated on a virtual machine using the same data and software

Why computational biology isn't reproducible


Keeping Critical Information Confidential


Poor training of new scientists


Researchers lack the technological capabilities


Researchers lack the knowledge or time

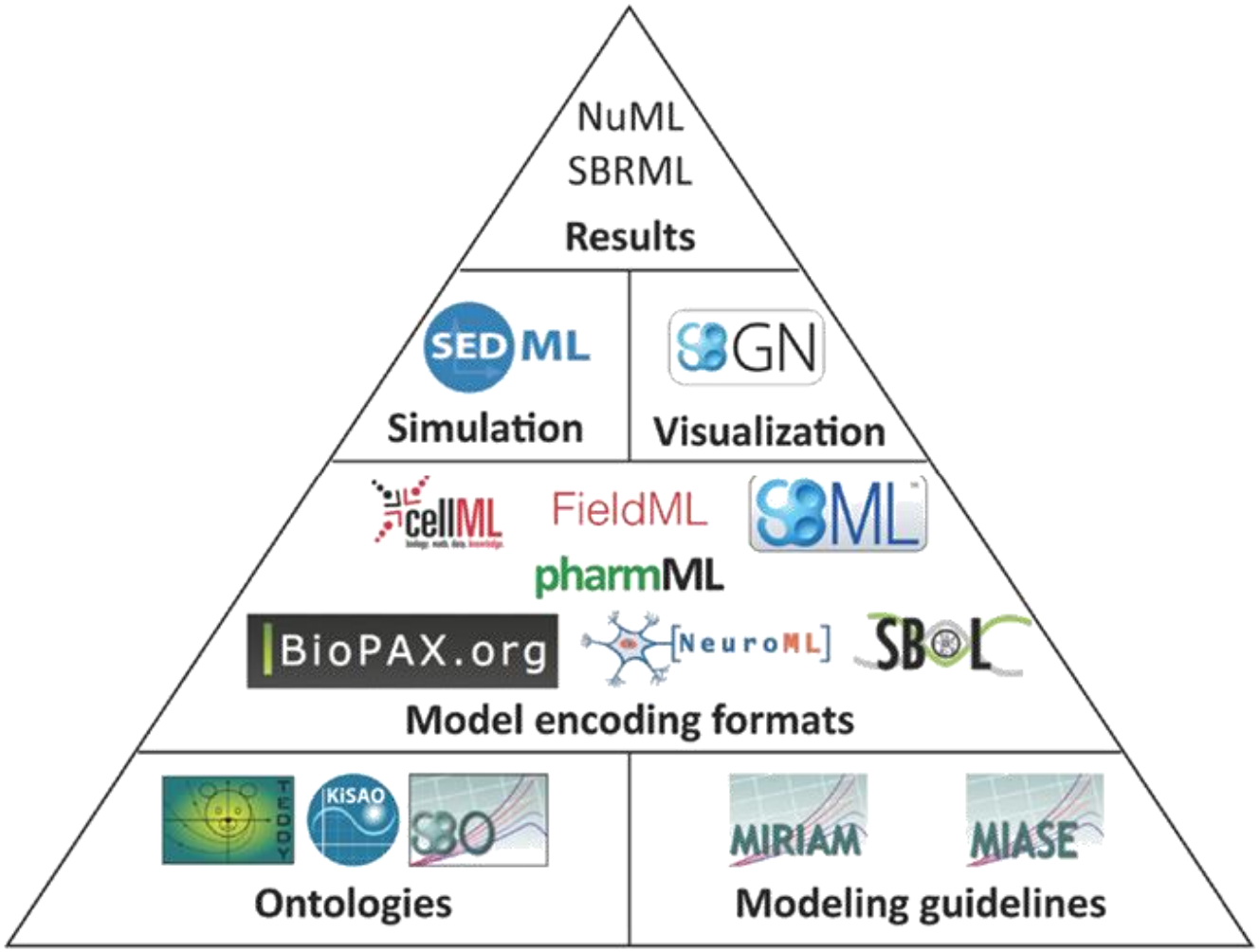

Researchers lack incentives and journals lack rewards

Why not just use a Programming Language?

Why not use an executable language such as Matlab, Python, Java etc to exchange and reproduce models?

1. To reproduce a model in a different programming language it would need to be manually translated to another language. This can be difficult and error prone.
2. There is no means to share such models because other groups might use different programming languages
3. Combining such models is extremely difficult.
4. If is difficult to annotate models that use an executable language.

COMBINE Standards



SBML in a Nutshell
"Systems Biology Markup Language"

- A **machine-readable** format for representing computational models in systems biology
- Domain: systems of **biochemical reactions**
- Specified using **XML**
- Components in SBML reflect the **natural conceptual constructs** of the domain
- Over 200 tools use SBML

SED-ML: Simulation Experiment Description ML

```
<sbml xmlns="http://www.sbml.org/sbml/level2/version4" xmlns:rdf="http://www.w3.org/
biomodels.net/biology-qualifiers/" xmlns:bqmodel="http://biomodels.net/model-qualifi
  <model metaid="metaid_decroly82" id="decroly82" name="Decroly1982_Enzymatic_Oscill
  <listOfUnitDefinitions>
    ...
  </listOfUnitDefinitions>
  <listOfCompartments>
    <compartment metaid="meta_cell" id="cell" name="cell" size="1">
    </compartment>
  </listOfCompartments>
  <listOfSpecies>
    <species metaid="_462445" id="alpha" name="alpha" compartment="cell" initialConc
    <species metaid="_462448" id="beta" name="beta" compartment="cell" initialConcen
    <species metaid="_462451" id="gamma" name="gamma" compartment="cell" initialConc
  </listOfSpecies>
  <listOfReactions>
    <reaction metaid="_462452" id="r1" reversible="false" sboTerm="SBO:0000176">
      <listOfProducts>
        <speciesReference species="alpha"/>
      </listOfProducts>
      <kineticLaw>
        <math xmlns="http://www.w3.org/1998/Math/MathML">
          <ci> v_Km1 </ci>
        </math>
        <listOfParameters>
          <parameter metaid="_462462" id="v_Km1" value="0.45" units="per_sec" sboTer
        </listOfParameters>
      </kineticLaw>
    </reaction>
    <reaction metaid="_462455" id="r2" reversible="false" sboTerm="SBO:0000176">
      <listOfReactants>
        <speciesReference species="alpha"/>
      </listOfReactants>
      <listOfProducts>
        <speciesReference species="beta" stoichiometry="50"/>
      </listOfProducts>
      <kineticLaw>
        <math xmlns="http://www.w3.org/1998/Math/MathML">
          <apply>
```

BIOMD0000000319 in BioModels Database

Decroly & Goldbeter, PNAS, 1982

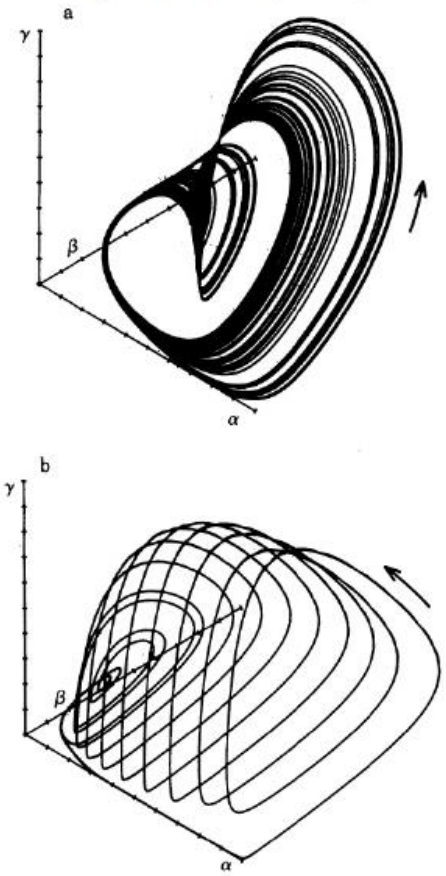


FIG. 4. Trajectories in the phase space (α, β, γ) associated with chaos (a) and with complex periodic behavior (b). The curves correspond to the substrate evolution depicted in Fig. 2 c and d, respectively, and have been obtained by integration of the kinetic equations from t = 0–5,000 sec. The ranges of variation of α, β, and γ in a are α = 28.44–50.6, β = 50.05–351.1, and γ = 0.05–2.28 and in b are α = 28.18–190.5, β = 0.14–604.0, and γ = 0.00014–8.8.

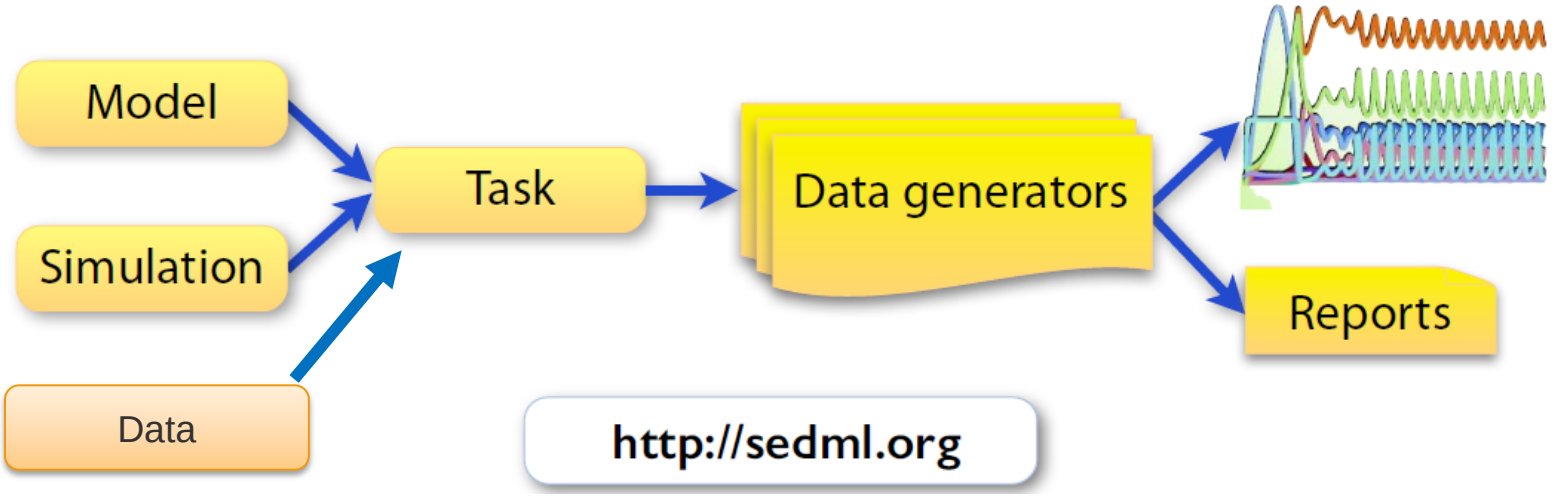
SED-ML: Simulation Experiment Description ML

Application-independent format

- Captures procedures, algorithms, parameter values

Can be used for

- Simulation experiments encoding parametrizations & perturbations
- Simulations using more than one model and/or method
- Data manipulations to produce plot(s)



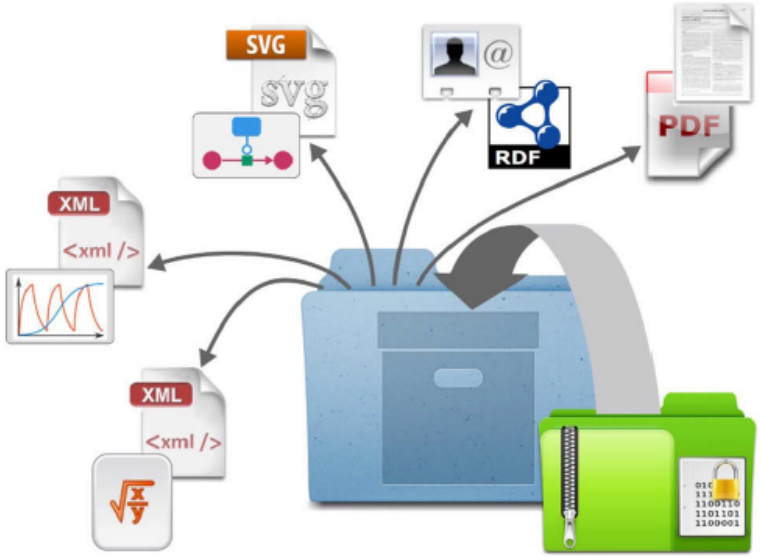
Multiple Files Make up a Model

A Complete Modeling Story is made of multiple files:

1. Model (s)
2. Simulation setup (s)
3. Parameter sets (virtual patients)
4. Diagrams
5. Raw Data
6. PDF Documents
7. etc

Exchange Format (OMEX)

COMBINE Archive format = single file that supports exchange of all information necessary for any modeling and simulation experiment



- Not SBML-specific at all
- Not programming-language specific
- Not domain specific

OMEX = file format for COMBINE Archive

- ZIP file containing manifest file (in XML form) + other files
- Use of ZIP leverages many existing programming libraries

http://co.mbine.org/documents/archive

All the Pieces Exist

Fig.: DOI: 10.1038/35002125

Model code

Network of reactions, entities, compartments

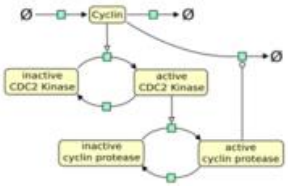


Fig.: BioModels Database

Model meta-data

Publication: Goldbeter pubmed:1833774

Organism: Human Taxonomy:9606
Compartment: Cell GO:0005623

M = inactive CDC2 Kinase UniProt:CDK1A_XENLA

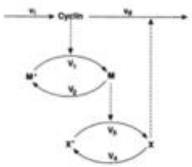
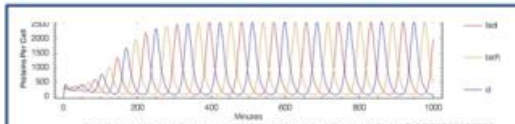


Fig.: DOI: 10.1038/35002125

Simulation / Data



Behavior: Oscillation TEDDY_0000006
Simulation algorithm: Gillespie KiSAO:000029

All the Pieces Exist

combine the computational modeling in biology network

Model code

Network of reactions, entities, compartments

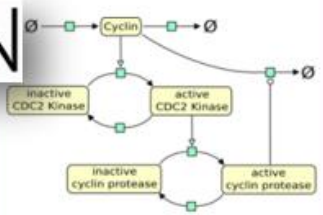


Fig.: BioModels Database

Model meta-data

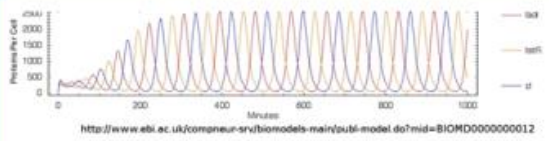
Publication: Goldbeter pubmed:1833774

Organism: Human Taxonomy:9606
Compartment: Cell GO:0005623

M = inactive CDC2 Kinase UniProt:CDK1A_XENLA

Fig.: DOI: 10.1038/3500

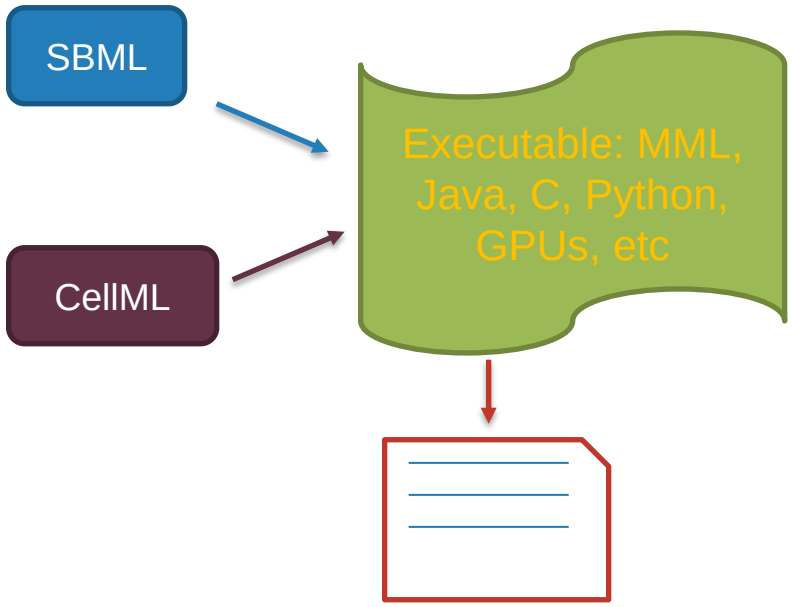
Simulation / Data



Behavior: Oscillation TEDDY_0000006
Simulation algorithm: Gillespie KiSAO:000029

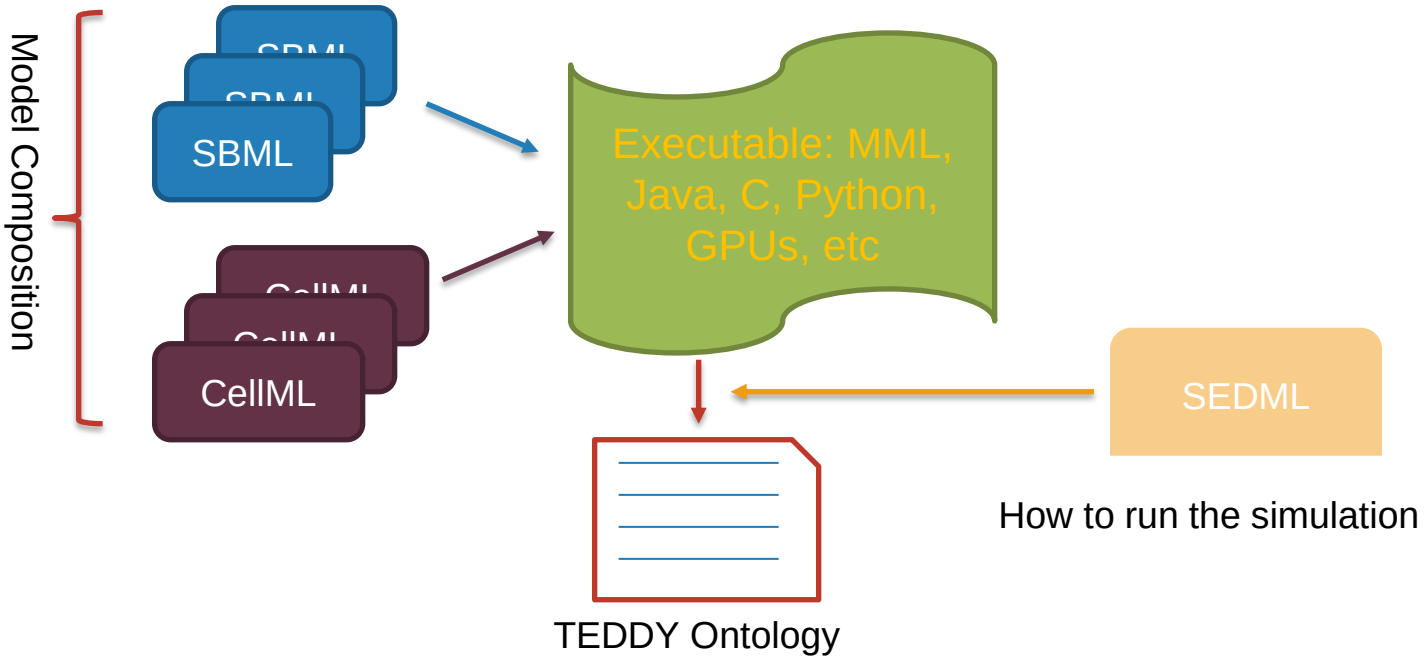
Possible Solution

Describe a model and how the model is run using a non-executable language that is independent of the application on which the model will be run. And then get the community to agree up upon it.



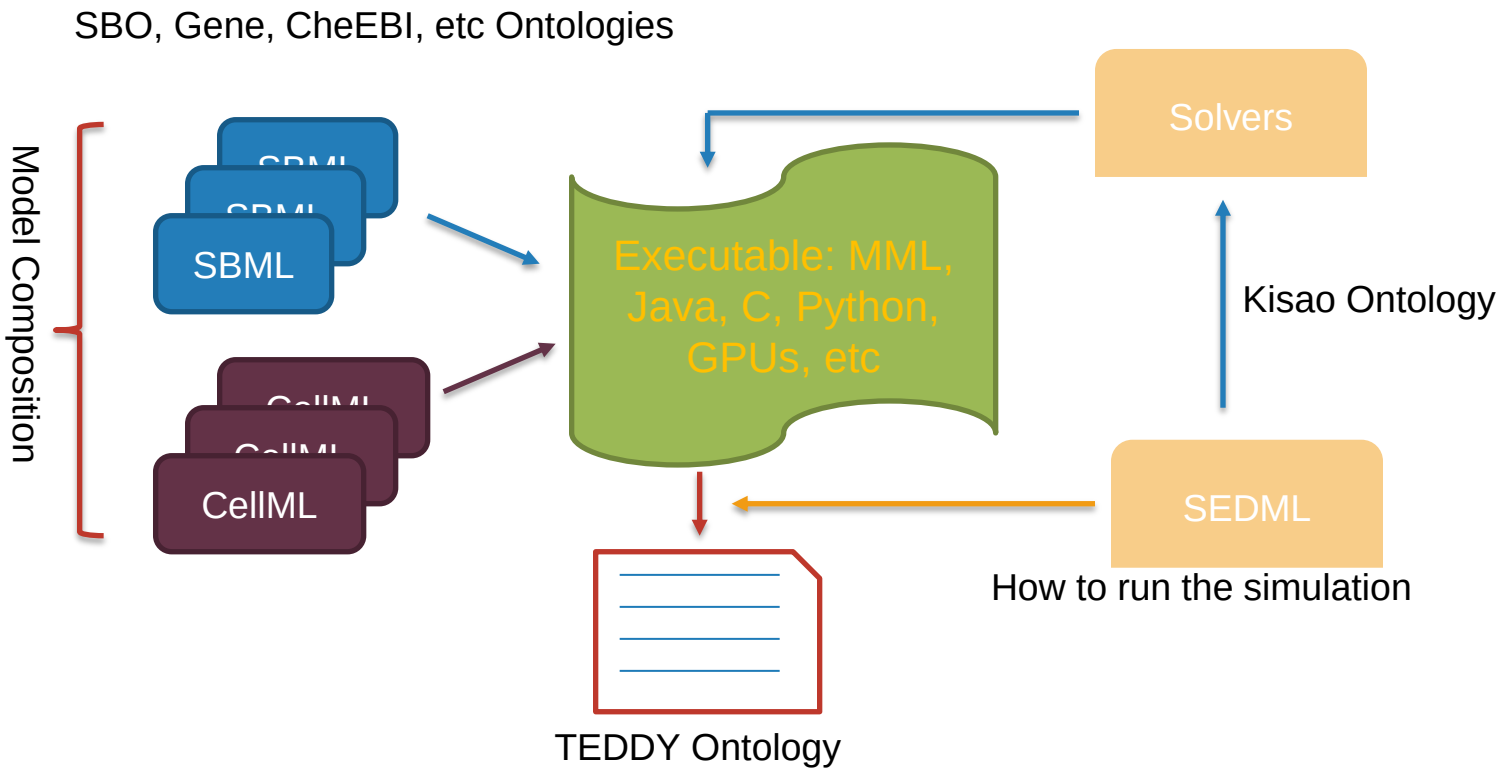
Possible Solution

Describe a model and how the model is run using a non-executable language that is independent of the application on which the model will be run. And then get the community to agree up upon it.



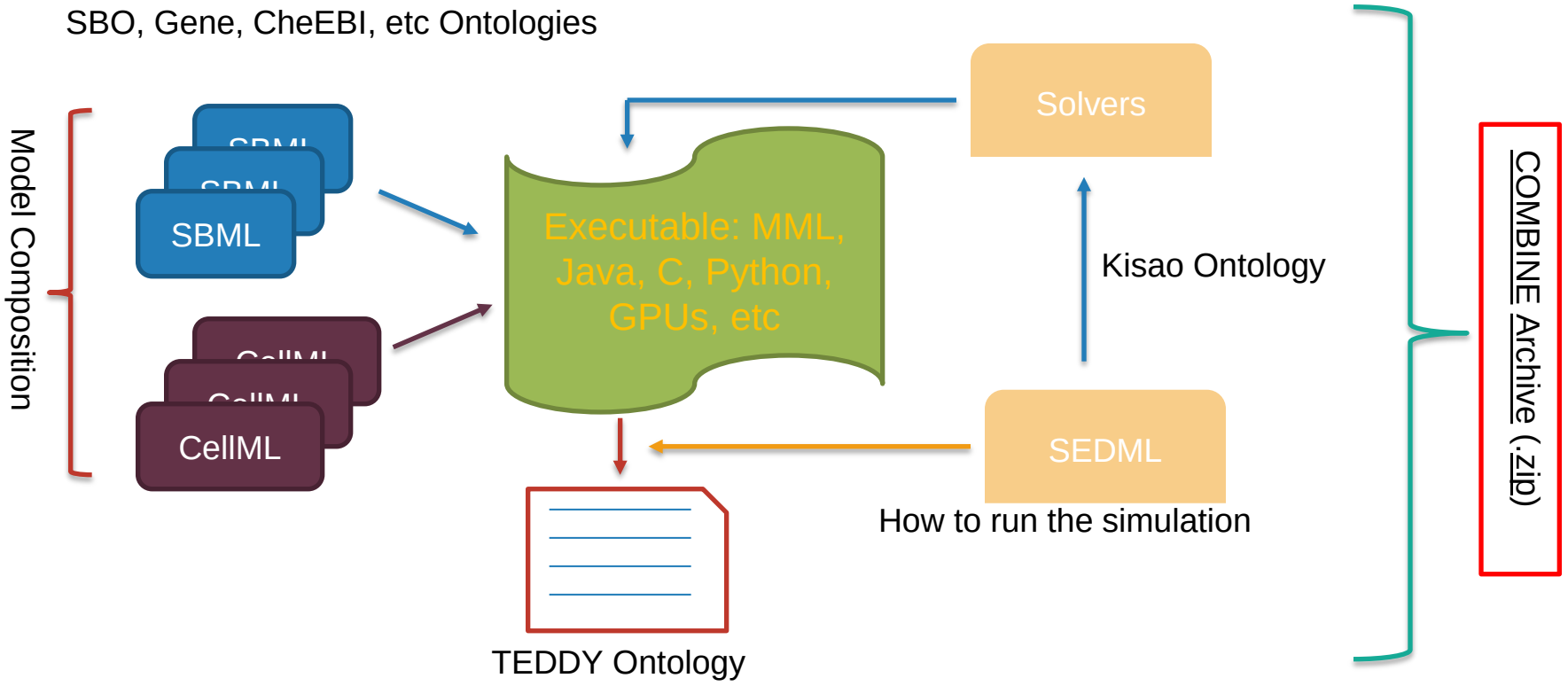
Possible Solution

Describe a model and how the model is run using a non-executable language that is independent of the application on which the model will be run. And then get the community to agree up upon it.



Possible Solution

Describe a model and how the model is run using a non-executable language that is independent of the application on which the model will be run. And then get the community to agree up upon it.

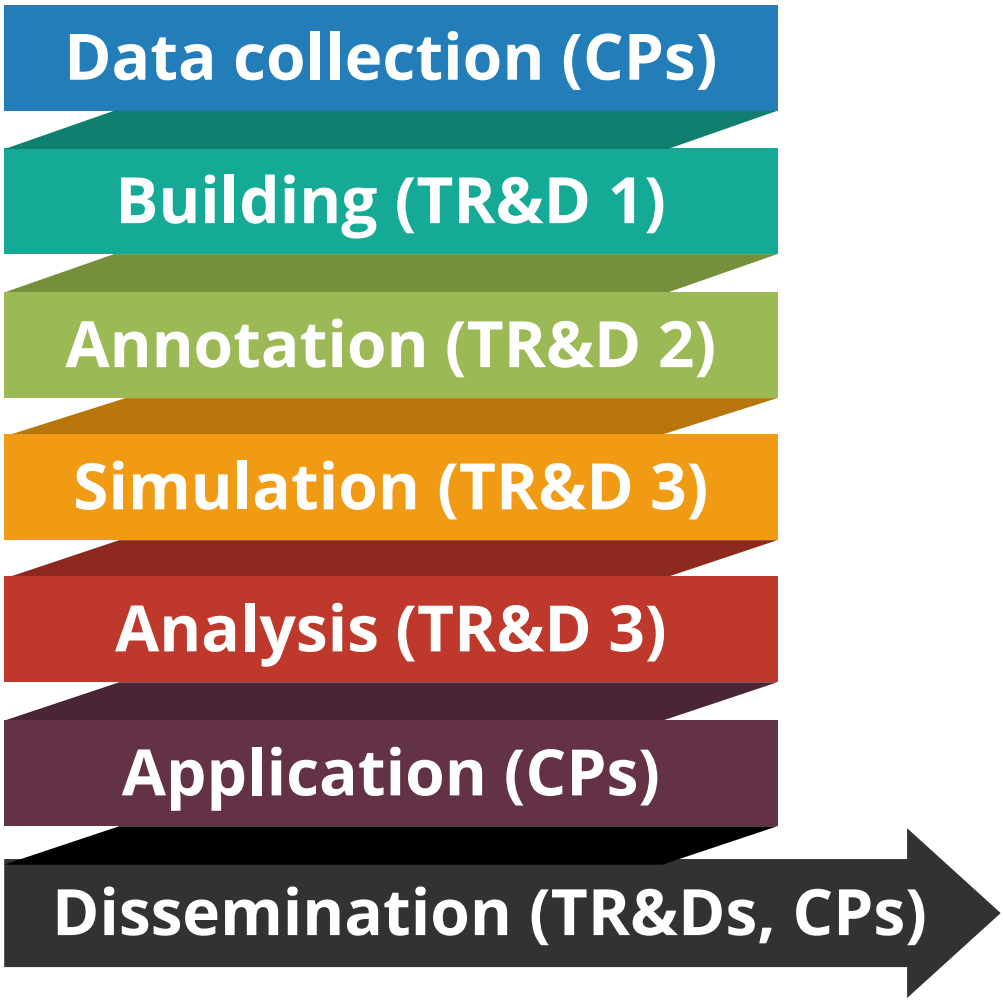


The Future

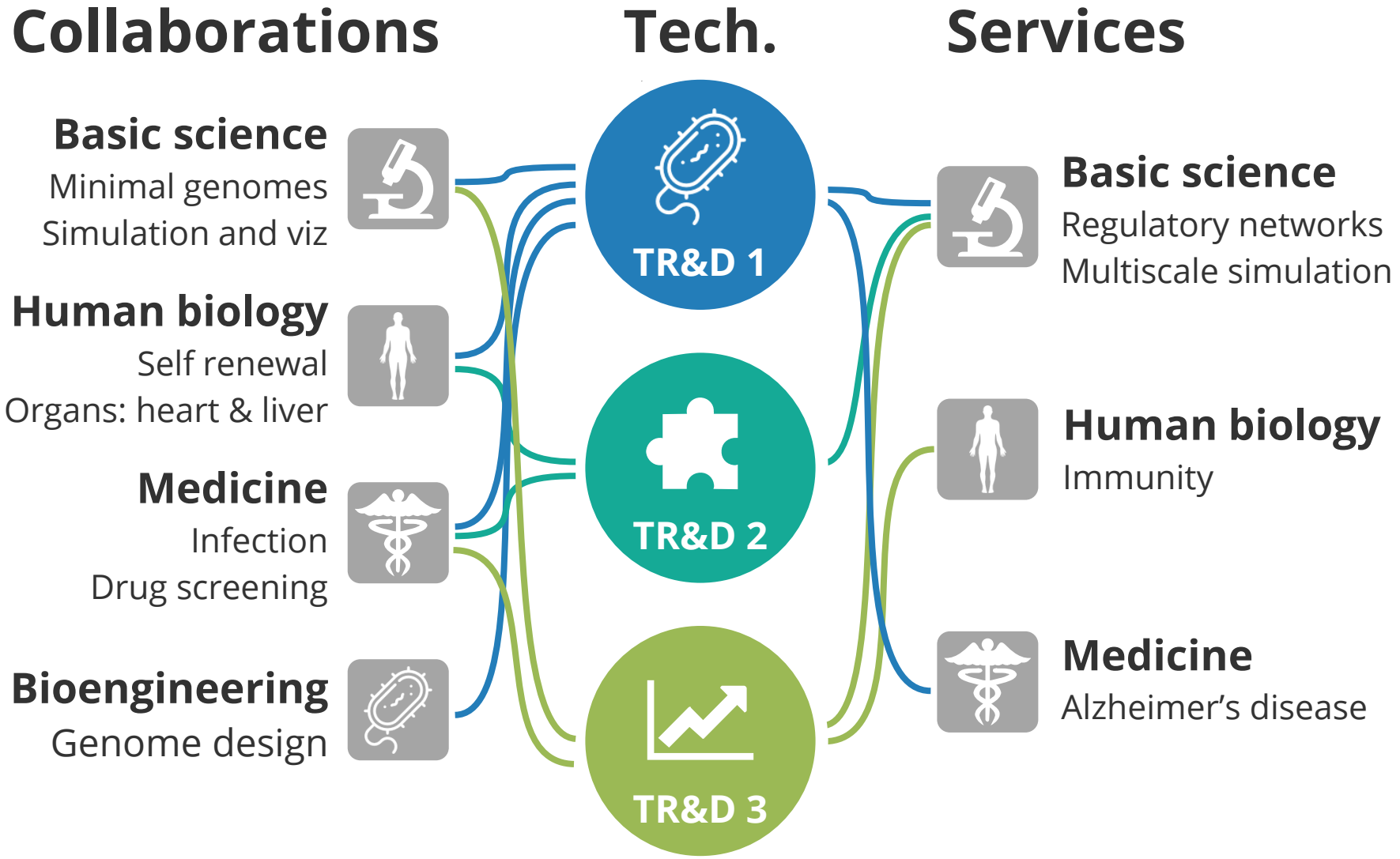
P41 NIH Center for Reproducibility of Systems and Physiological Models

CENTER FOR REPRODUCIBLE BIOMEDICAL MODELING

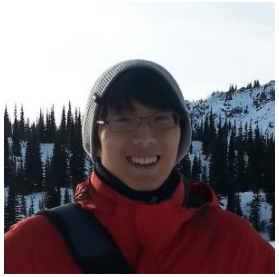
Goal: Enhance reproducibility of biomodeling



CPs and SPs span numerous applications

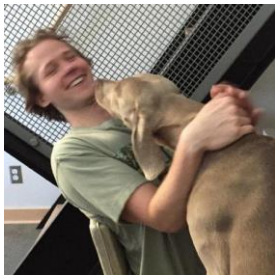


Acknowledgements



Kiri Choi
SED-ML, COMBINE,
Tellurium, phrasedML

Frank Bergmann
SED-ML



Kyle Medley
SBML, libRoadRunner,
Jupyter Interface

Lucian Smith
SBML, SED-ML,
phrasedml

NATIONAL INSTITUTES OF HEALTH

Acknowledgements

P41 Collaborators:

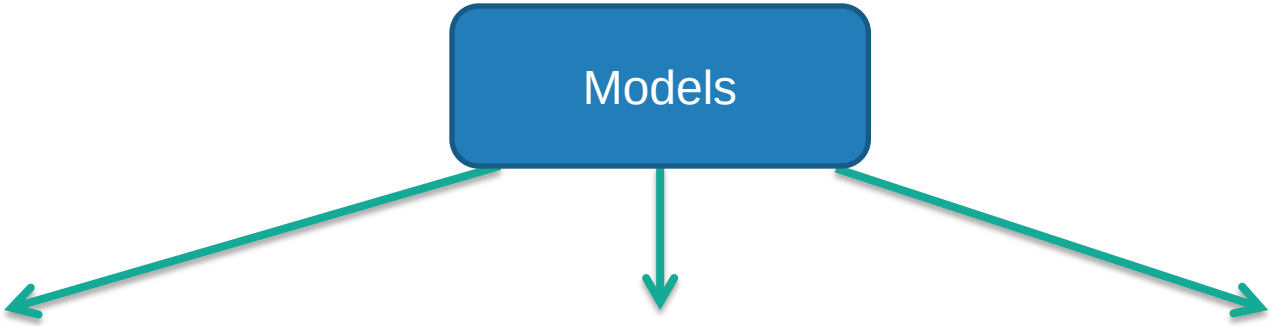
Jonathan Karr
Whole-Cell

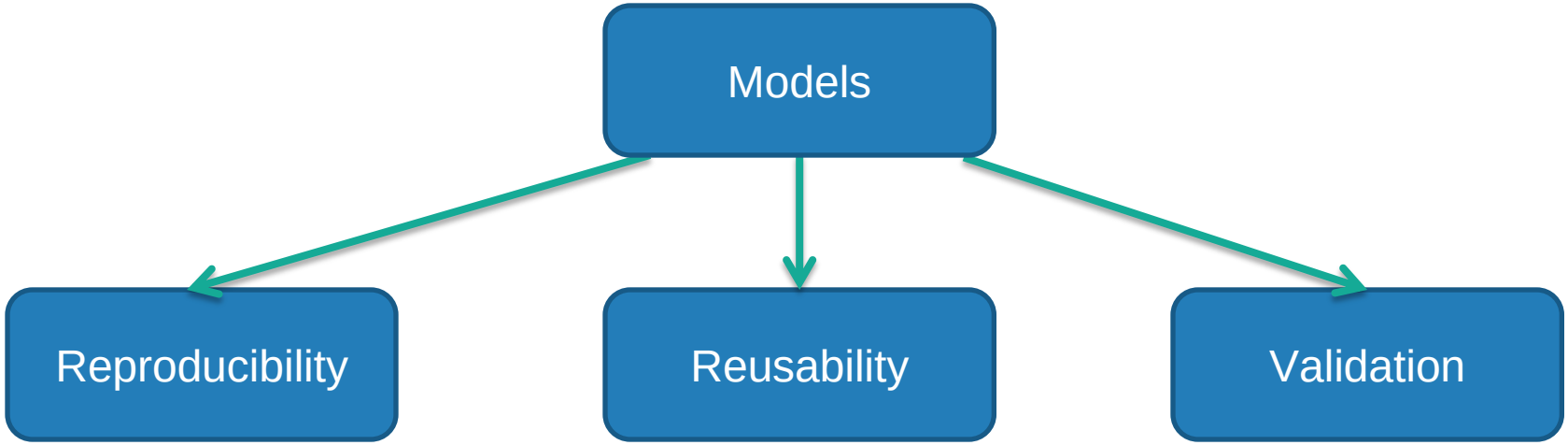
Ion Moraru
VCell

John Gennari
Annotation

David Nickerson
CellML, SEDML

NATIONAL INSTITUTES OF HEALTH

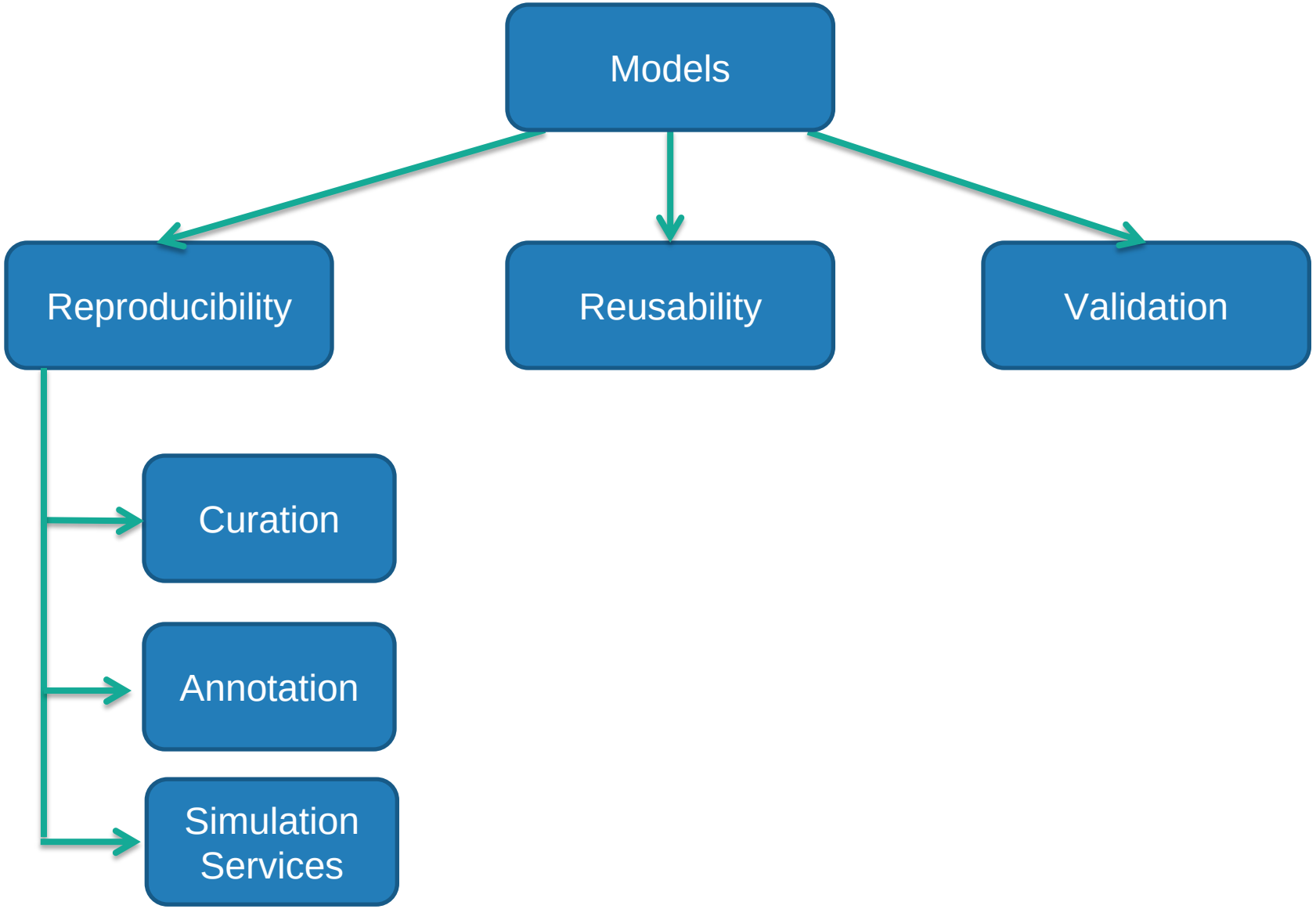

Models

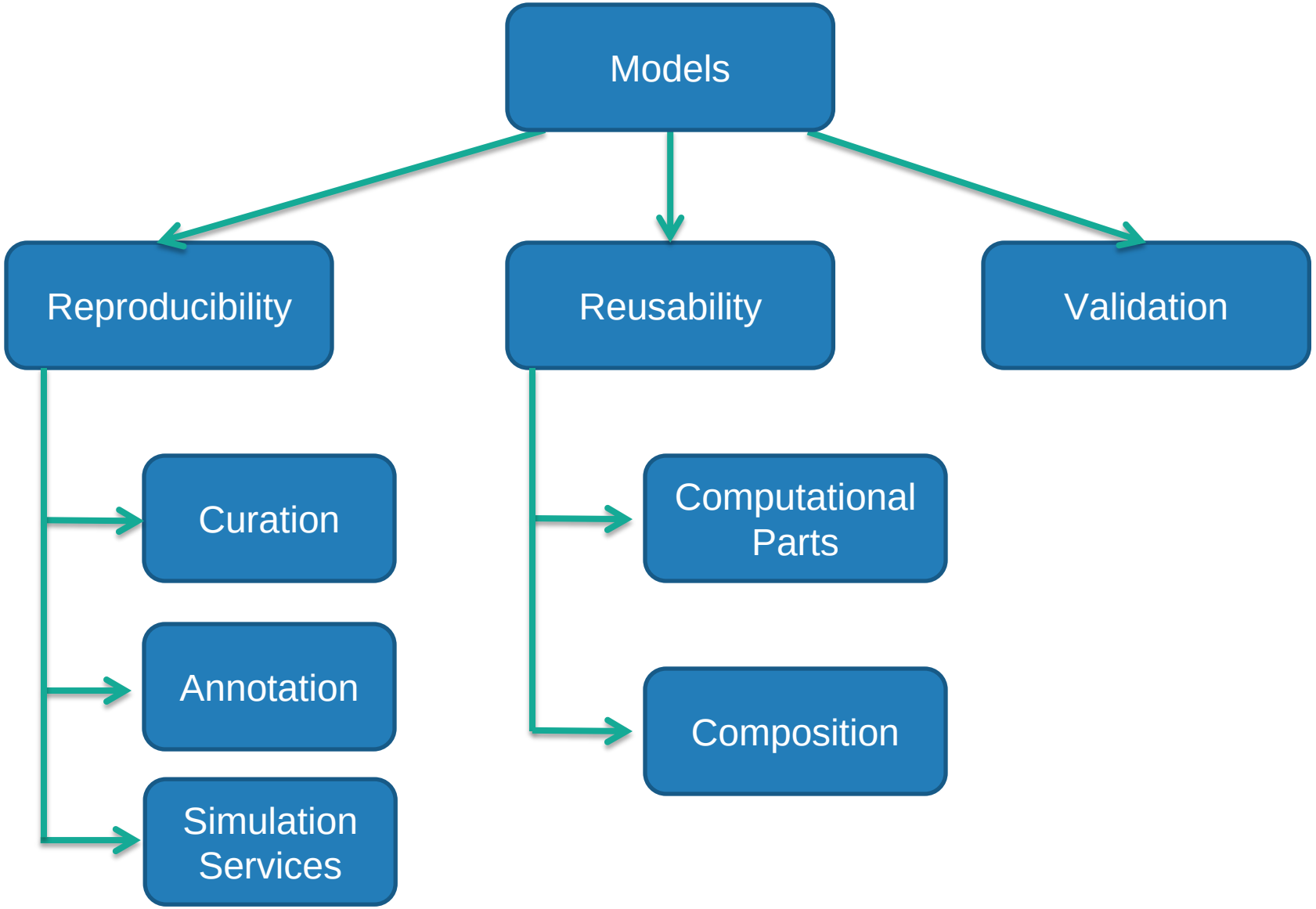


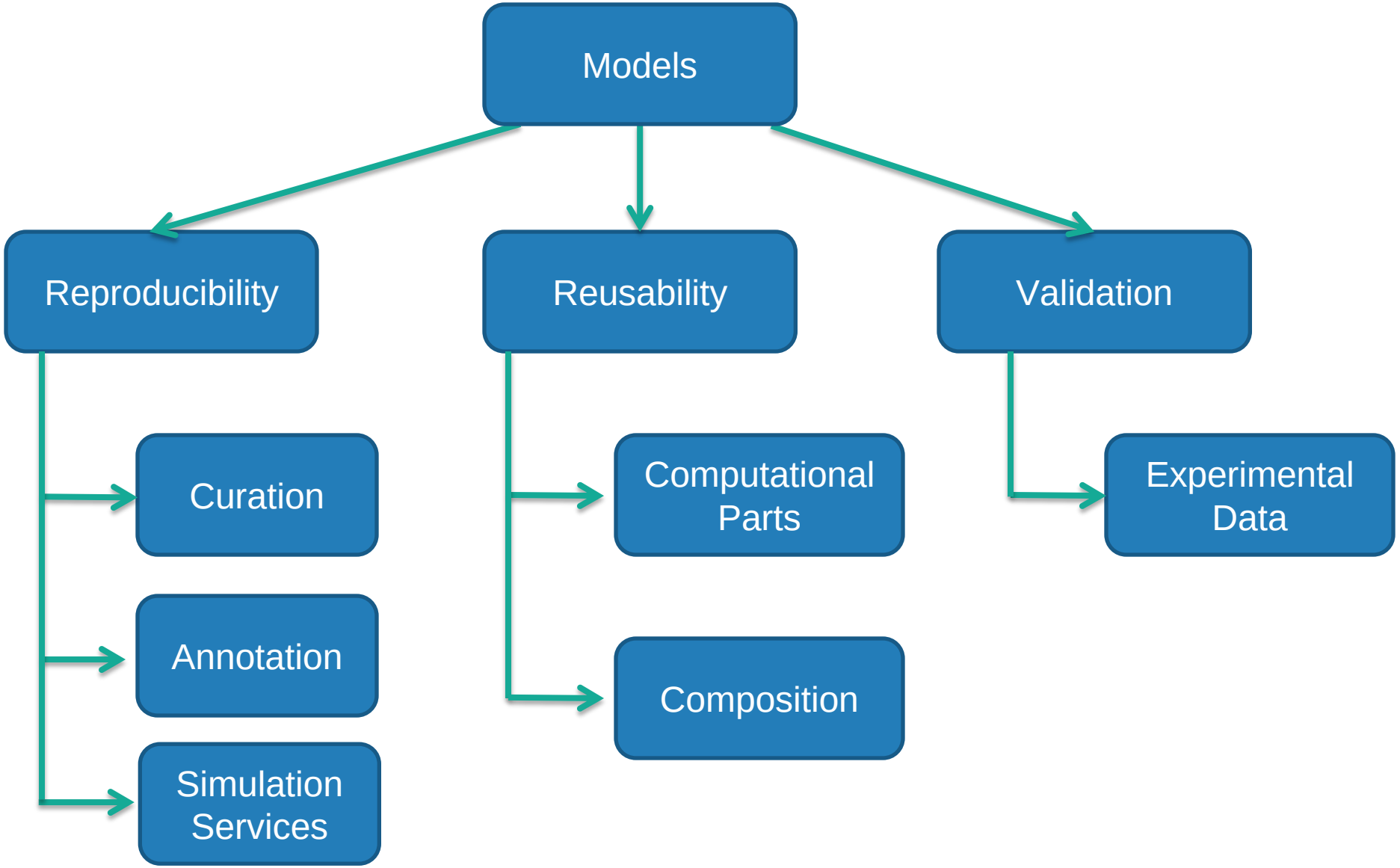
Being able to recreate published simulations

Being able to reuse published models in new applications.

Being able to show that the model has credibility.


Models
Reproducibility
Reusability
Validation
Curation
Annotation
Simulation Services


Models
Reproducibility
Reusability
Validation
Curation
Annotation
Simulation Services
Computational Parts
Composition


Models
Reproducibility
Reusability
Validation
Curation
Annotation
Simulation Services
Computational Parts
Composition
Experimental Data