IEEE Xplore® Digital Library
http://ieeexplore.ieee.org/Xplore

# Toward community standards and software for whole-cell modeling

Dagmar Waltemath*,†, Jonathan R. Karr*,†, Frank T. Bergmann, Vijayalakshmi Chelliah, Michael Hucka, Marcus Krantz, Wolfram Liebermeister, Pedro Mendes, Chris J. Myers, *Fellow, IEEE,* Pinar Pir, Begum Alaybeyoglu, Naveen K Aranganathan, Kambiz Baghalian, Arne T. Bittig, Paulo E. Pinto Burke, Matteo Cantarelli, Yin Hoon Chew, Rafael S. Costa, Joseph Cursons, Tobias Czauderna, Arthur P. Goldberg, Harold F. Gómez, Jens Hahn, Tuure Hameri, Denis Kazakiewicz, Ilya Kiselev, Vincent Knight-Schrijver, Christian Knüpfer, Matthias König, Daewon Lee, Audald Lloret-Villas, Nikita Mandrik, J. Kyle Medley, Bertrand Moreau, Hojjat Naderi-Meshkin, Sucheendra K. Palaniappan, Daniel Priego-Espinosa, Martin Scharm, Mahesh Sharma, Kieran Smallbone, Natalie J. Stanford, Je-Hoon Song, Tom Theile, Milenko Tokic, Namrata Tomar, Vasundra Touré, Jannis Uhlendorf, Thawfeek M Varusai, Leandro H. Watanabe, Florian Wendland, Markus Wolfien, James T. Yurkovich, Yan Zhu, Argyris Zardilis, Anna Zhukova, and Falk Schreiber†

Manuscript received August XX, 2015; revised March XX, 2016; accepted April XX, 2016. Date of publication XXX XX, 201X; date of current version XXX XX, 201X. The Rostock and Utah meetings were supported by the Volkswagen Foundation (Grant 88495 to D. W. and F. S.). J. .R. K was supported the James S. McDonnell Foundation Postdoctoral Fellowship Award in Studying Complex Systems and National Science Foundation grant 1548123. J. C. was supported by the Australian Research Council Centre of Excellence in Convergent Bio-Nano Science and Technology (project CE140100036). *Asterisk indicates corresponding authors: Dagmar Waltemath dagmar.waltemath@uni-rostock.de, Jonathan R. Karr karr@mssm.edu, Falk Schreiber falk.schreiber@uni-konstanz.de. Dagger indicates these authors contributed equally to this work.*

*D. Waltemath, A. T. Bittig, M. Scharm, T. Theile, V. Touré, F. Wendland and M. Wolfien are with the Institute of Computer Science, University of Rostock, 18051 Rostock, Germany.

J. R. Karr, Y. H. Chew and A. P. Goldberg are with the Department of Genetics & Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA.

F. T. Bergmann is with BioQuant, University of Heidelberg, 69120 Heidelberg, Germany.

V. Chelliah, N. K. Aranganathan and A. Lloret-Villas are with the European Bioinformatics Institute (EMBL-EBI), European Molecular Biology Laboratory, Cambridge CB10 1SD, UK.

M. Hucka is with the Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA 91125, USA.

M. Krantz, J. Hahn and J. Uhlendorf are with the Department of Biology, Humboldt University of Berlin, 10115 Berlin, Germany.

W. Liebermeister and M. König are with the Institute of Biochemistry, University Medicine Charité Berlin, 10117 Berlin, Germany. M. König is also with the Institute for Theoretical Biology, Humboldt-University Berlin, Invalidenstrae 43, 10115 Berlin, Germany.

P. Mendes is with the Manchester Institute of Biotechnology and the School of Computer Science, University of Manchester, Manchester M1 7DN, UK and also with the Center for Quantitative Medicine and the Department of Cell Biology, University of Connecticut Health Center, Farmington, CT 06030, USA.

C. J. Myers and L. H. Watanabe are with the Department of Electrical and Computer Engineering, University of Utah, Salt Lake City, Utah 84112, USA.

P. Pir is with the Gebze Technical University, Kocaeli 41400, Turkey.

B. Alaybeyoglu is with the Department of Chemical Engineering, Boğaziçi University, Bebek 34342, Turkey.

K. Baghalian is with the Department of Plant Sciences, University of Oxford, South Parks Road, Oxford, UK.

P. E. Pinto Burke is with the Institute of Science and Technology, Federal University of São Paulo, Brazil.

Y. H. Chew is with the Centre for Synthetic and Systems Biology, University of Edinburgh, Edinburgh EH9 3BF, UK.

M. Cantarelli is with OpenWorm.

R. S. Costa is with the Centre of Intelligent Systems-IDMEC, Instituto Superior Técnico, University of Lisbon, 1049-001 Lisboa, Portugal.

J. Cursons is with the Systems Biology Laboratory, Melbourne School of Engineering, University of Melbourne, Parkville, VIC 3010, Australia, and also with the ARC Centre of Excellence in Convergent Bio-Nano Science and Technology, Melbourne School of Engineering, University of Melbourne, Parkville, VIC 3010.

Digital Object Identifier 10.1109/TBME.2016.2560762

T. Czauderna is with the Faculty of Information Technology, Monash University, Clayton, VIC 3800, Australia.

H. F. Gómez is with the Department of Biosystems Science and Engineering, ETH Zürich, Basel, Switzerland.

T. Hameri and M. Tokic are with the Laboratory of Computational Systems Biotechnology (LCSB), Swiss Federal Institute of Technology (EPFL), CH-1015 Lausanne, Switzerland. M. Tokic is also with the Swiss Institute of Bioinformatics (SIB), CH-1015 Switzerland.

D. Kazakiewicz is with the Center for Statistics, Universiteit Hasselt, Hasselt BE3500, Belgium, and also with the Center for Innovative Research, Medical University of Białystok, Białystok 15-089, Poland.

I. Kiselev is with the Design Technological Institute of Digital Techniques, Siberian Branch of the Russian Academy of Sciences, Novosibirsk 630090, Russia.

V. Knight-Schrijver is with the Babraham Institute, Cambridge CB22 3AT, UK .

C. Knüpfer is with the Institut für Informatik, University of Jena, 07743 Jena, Germany.

D. Lee and J.-H. Song are with the Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology, Daejeon 305-701, Republic of Korea.

N. Mandrik is with the Sobolev Institute of Mathematics, Siberian Branch of the Russian Academy of Sciences, Novosibirsk 630090, Russia.

J. K. Medley is with the Department of Bioengineering, University of Washington, Seattle, WA 98195, USA.

H. Naderi-Meshkin is with the Stem Cell and Regenerative Medicine Research Department, Iranian Academic Center for Education, Culture Research (ACECR), Khorasan Razavi Branch, Mashhad, Iran.

B. Moreau is with the CoSMo Company, Lyon, France.

S. K. Palaniappan is with the Rennes - Bretagne Atlantique Research Centre, Institute for Research in Computer Science and Automation, 35042 Rennes Cedex, France.

D. Priego-Espinosa is with the Instituto de Ciencias Físicas, Universidad Nacional Autónoma de México, México.

M. Sharma is with the Department of Pharmacoinformatics, National Institute of Pharmaceutical Education and Research, Punjab 160062, India.

K. Smallbone and N. J. Stanford are with the Manchester Centre for Integrative Systems Biology, University of Manchester, Manchester M1 7DN, UK.

N. Tomar is with the Department of Dermatology, University Medicine, Friedrich-Alexander University of Erlangen-Nürnberg, Erlangen, Germany.

T. M. Varusai is with the Department of Systems Biology Ireland, University College Dublin, Belfield, Dublin 4, Ireland.

J. T. Yurkovich is with the Department of Bioengineering, University of California, San Diego, La Jolla, CA 92093, USA.

Y. Zhu is with Monash Institute of Pharmaceutical Sciences, Monash University, Parkville, VIC 3052, Australia.

A. Zardilis is with the Centre for Synthetic and Systems Biology, University of Edinburgh, UK.

A. Zhukova is with the Institut de Biochimie et Génétique Cellulaires, National Center for Scientific Research, and also with the University of Bordeaux, France, 33077 Bordeaux Cedex, France.

F. Schreiber is with the Faculty of Information Technology, Monash University, Clayton, VIC 3800, Australia and also with the Department of Computer and Information Science, University of Konstanz, 78457 Konstanz, Germany.

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/

***Objective:*** **Whole-cell (WC) modeling is a promising tool for biological research, bioengineering, and medicine. However, substantial work remains to create accurate, comprehensive models of complex cells.** ***Methods:*** **We organized the 2015 Whole-Cell Modeling Summer School to teach WC modeling and evaluate the need for new WC modeling standards and software by recoding a recently published WC model in SBML.** ***Results:*** **Our analysis revealed several challenges to representing WC models using the current standards.** ***Conclusion:*** **We, therefore, propose several new WC modeling standards, software, and databases.** ***Significance:*** **We anticipate that these new standards and software will enable more comprehensive models.**

***Index Terms*****—Whole-cell modeling, Systems biology, Computational biology, Simulation, Standards, Education**

## I. Introduction

COMPUTATIONAL modeling is a powerful tool for biological research, bioengineering, and medicine to understand complex systems. It has been used to identify gene functions [1], engineer metabolic pathways [2], and identify drug targets [3]. Computational models also have the potential to help bioengineers design new microorganisms that can synthesize high value chemicals, sense toxins, and decontaminate waste, as well as help clinicians interpret individual 'omics profiles and personalize medical therapy [4]. Realizing this potential requires more comprehensive models that can predict phenotype from genotype. In turn, this requires improved modeling and simulation standards and software [5–10].

Recently, Karr et al. developed the first *whole-cell* (WC) model which represents every individual gene function [11]. The model represents the life cycle of a single *Mycoplasma genitalium* bacterial cell and predicts the dynamics of every molecular species. The model is composed of 28 pathway submodels that are represented using multiple mathematical formalisms including *stochastic simulation*, *ordinary differential equations* (ODEs), *flux balance analysis* (FBA), and *Boolean rules* (BRs). The model was implemented in MATLAB.

The *M. genitalium* model has been used to gain novel insights into non-genetic cell cycle regulation mechanisms [11], learn unknown kinetic rate parameters from phenotypic data [12], calculate the metabolic costs of synthetic circuits [13], and repurpose antibiotics [14].

Karr et al. extensively documented the model; developed the WholeCellKB [15], WholeCellSimDB [16], and WholeCellViz [17] software tools to provide user-friendly interfaces to the model; and published the model open-source. This has enabled other researchers to reuse the model [12–14].

However, significant domain expertise is still needed to reuse the model or to develop new WC models. The multi-algorithm modeling methodology is complex. The model is difficult to understand, reuse, and extend because it is described directly in terms of its numerical simulation rather than in a software-independent format. The model code is difficult to learn and reuse because it is large, complex, and intertwined with the details of the *M. genitalium* model. The simulation code is also slow. Furthermore, the simulation code requires the proprietary MATLAB software package.

New standards and software tools are needed to help researchers build and simulate WC models. They would help researchers reuse, reproduce, and compare models, as well as share models through repositories such as BioModels [18].

Several systems biology standards have been developed by the *COmputational Modeling in BIology NEtwork* (COMBINE) [8], including the *Systems Biology Markup Language* (SBML) [19], CellML [20], the *Simulation Experiment Description Markup Language* (SED-ML) [21], and the *Systems Biology Graphical Notation* (SBGN) [22] (Table I). SBML and CellML are formats for representing mathematical models. CellML describes the mathematics whereas SBML describes biological processes. Both support several modeling formalisms including ODEs and FBA. SED-ML describes and enables researchers to reproduce computational experiments. SBGN is a visual notation for describing biological processes. However, none of these standards have been used for WC modeling.

We organized the 2015 Whole-Cell Modeling Summer School to train students in WC modeling and to evaluate the need for new WC modeling standards and software. The school focused on creating a reusable WC model by recoding the *M. genitalium* model in SBML. We focused on SBML because SBML is the most widely used systems biology standard and there was insufficient time to evaluate multiple standards. The school also aimed to improve numerous details of the model, visualize the model with SBGN, and describe model simulations with SED-ML. The latest versions of our SBML-encoded submodels and SBGN diagrams are available at https://github.com/whole-cell-tutors/wholecell/releases/tag/meeting-report.

Most importantly, the school generated extensive community discussion on how to best build and simulate WC models. This report describes the outcome of these discussions, including our recommendations for new standards and software to accelerate WC modeling. We also describe our progress toward recoding the *M. genitalium* model in SBML and the lessons that we learned about organizing research-based schools.

## II. The 2015 Whole-Cell Modeling Summer School

The school was held March 9-13, 2015, at the University of Rostock, Germany. It was organized by D. Waltemath and F. Schreiber and funded by the Volkswagen Foundation. 43 students and nine instructors participated in the school. A follow up meeting involving 15 of the original and six additional participants was held October 10-11, 2015, at the University of Utah, USA. All of the materials for the school are available at http://sites.google.com/site/vwwholecellsummerschool.

We advertised the school through community mailing lists, conference calendars, and websites. Applicants were asked to describe their experience and interest in WC modeling. We chose 43 participants from 118 applicants based on three criteria. (1) We identified the most qualified and enthusiastic applicants. (2) We gave preference to students, female applicants, and applicants from developing countries. (3) We

Table I
SYSTEMS BIOLOGY STANDARDS AND STANDARDIZATION EFFORTS.

| Acronym | Name | Type | Description | Ref. |
|---|---|---|---|---|
| CellML | CellML | Standard | Describes models in terms of mathematical relationships | 20 |
| COMBINE | COmputational Modeling in BIology NEtwork | Community | Develops computational biology standards and software | 8 |
| SBGN | Systems Biology Graphical Notation | Standard | Describes biochemical pathway diagrams | 23 |
| SBML | Systems Biology Markup Language | Standard | Describes models in terms of biochemical processes | 24 |
| SBML Arrays | SBML Package: Arrays | Standard | Describes arrays | 25 |
| SBML Comp | SBML Package: Hierarchical Model Composition | Standard | Describes how model are composed from other models | 26 |
| SBML Distrib | SBML Package: Distributions | Standard | Describes random distributions | 27 |
| SBML FBC | SBML Package: Flux Balance Constraints | Standard | Describes constraint-based models | 28 |
| SBML Multi | SBML Package: Multistate and Multicomponent Species | Standard | Supports rule-based modeling | 25 |
| SBML Spatial | SBML Package: Spatial Processes | Standard | Describes spatially-resolved models | 29 |
| SED-ML | Simulation Experiment Description Markup Language | Standard | Describes computational experiments | 21 |

selected participants to represent a broad range of scientific disciplines. We used the same criteria to select instructors.

The school began with introductory lectures on WC modeling and the existing systems biology standards by J. Karr and M. Hucka and introductory discussions on model composition, state representation, and stochastic modeling. Most of the school was devoted to active learning sessions in which the students and instructors were divided into 11 groups and challenged to use SBML to recode the *M. genitalium* model, use SBGN to visualize the model, and use SED-ML to simulate the model. Groups 1-8 encoded submodels. Group 9 developed a submodel integration scheme. Group 10 annotated and visualized the model. Group 11 helped all of the other groups understand, encode, and improve the model. Table SII lists the groups and participants of both meetings. Each day concluded with community discussions. In addition, the school included a poster session and networking activities.

The students learned about state-of-the-art WC modeling; the open challenges to building more complex models; open-source modeling software; the importance of reproducibility; and the SBML, SED-ML, and SBGN standards. The students also expanded their professional networks. Several of the students reported that the skills and knowledge they gained from the school would enhance their research.

We learned several lessons about organizing research-based schools. (1) Students enjoy working on research problems more than solving prescribed exercises. This engages students in the field, challenges them, and helps them build practical skills. (2) Research-based schools should have clear background knowledge expectations, learning objectives, and research goals. This helps students decide whether to participate, prepare, and learn efficiently. (3) Research-based schools should have a flexible schedule, multidisciplinary participants, and a high teacher-to-student ratio. This allows students to engage in impromptu discussions, draw on multiple perspectives, and get feedback and iterate quickly.

## III. TOWARD AN IMPROVED SBML-ENCODED WC MODEL

In addition to teaching students about WC modeling and the systems biology standards, the school aimed to improve the *M. genitalium* model and to encode the model in SBML.

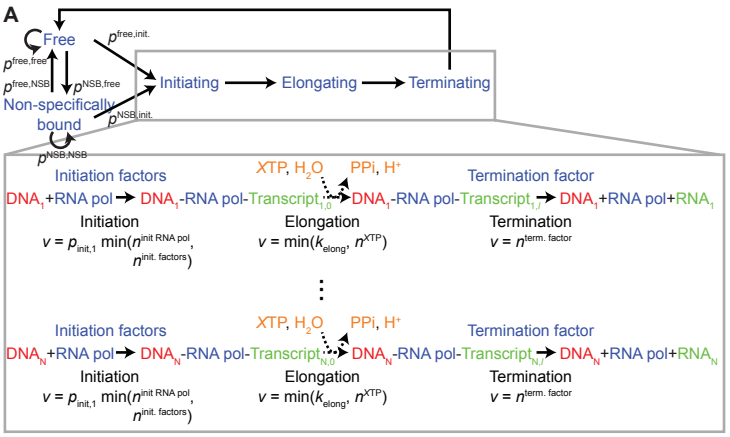
### A. Submodel encoding

We pursued several strategies to encode submodels in SBML. Several groups encoded submodels by (1) reading the original documentation of the model, (2) drawing pathway diagrams using software tools such as CellDesigner [30] and VANTED [31], and (3) writing scripts to generate SBML from the diagrams. Other groups used model design tools such as Antimony [32], BioUML [33], COBRApy [34], COPASI [35], iBioSim [36], and libRoadRunner [37] to recode submodels based on the original documentation. A few of the groups encoded submodels by converting the MATLAB code to SBML. As an example, Fig. 1 and File S1 illustrate how we recoded the transcription submodel.

We encountered several challenges to encoding the submodels in SBML. First, understanding the submodels was time-consuming because many students were not familiar with the modeled biology, many of the submodel details are described only in the MATLAB code, and the model documentation only summarizes the model. For these reasons, J. Karr, one of the authors of the original model, helped all of the groups understand the modeled biology and mathematics. Dr. Karr also helped several groups simplify their encoding tasks by recommending that they recode only the most important model components. For example, Dr. Karr suggested that the transcription group represent the transcription of each RNA species as a single lumped reaction rather than hundreds of thousands of individual base elongation reactions. It would have been challenging to recode the model without Dr. Karr. The essentiality of Dr. Karr's guidance underscores the need for improved WC modeling methods and standards.

Second, it was difficult to encode the original serial and randomized algorithms into SBML because SBML does not explicitly represent sequential operations and plain SBML does not support random number generation. We overcame these problems by formalizing submodels as Gillespie algorithm stochastic simulations [38].

Third, in many cases, we had to either enumerate the particle-based state representations used by the original model or approximate the original model. For example, the translation group approximated the original model by lumping all of the elongation reactions for each protein into a single reaction. The replication group used indicator variables to enumerate the particle-based chromosome representation from the origi-




| Model component | Original | SBML |
|---|---|---|
| Species | 542,808 | 346 |
| DNA-RNA pol | 335 | |
| DNA-RNA pol-Transcript | 542,126 | |
| Enzymes | 5 | 4 |
| Metabolites | 7 | 7 |
| RNA | 335 | 335 |
| Reactions | 542,461 | 335 |
| Elongation | 541,791 | |
| Initiation | 335 | 335 |
| Termination | 335 | |
| Parameters | 340 | 335 |
| RNA pol state transitions | 4 | |
| RNA polymerization rate | 1 | |
| Transcription init. rates | 335 | 335 |

Figure 1. Comparison of the original and SBML transcription submodels. (A) The original transcription submodel included two sub-submodels: (1) a Markov model that describes how RNA polymerase switches among freely diffusing, non-specifically bound, and initiating states and (2) an ad hoc stochastic model that describes how RNA polymerase initiates transcription, elongates individual bases by walking along DNA, and terminates transcripts. (B) We created the SBML transcription submodel by simplifying the original submodel. The SBML submodel only represents transcription initiation, elongation, and termination; lumps the initiation, elongation, and termination of each RNA species into a single reaction; and does not explicitly represent DNA-protein binding. (C) An equivalent population-based ad hoc stochastic simulation algorithm for the original submodel. The original submodel was implemented using a more efficient particle-based algorithm. To facilitate comparison with the population-based SBML version, we have described an equivalent population-based algorithm. (D) We also improved the SBML submodel by replacing the ad hoc stochastic simulation algorithm with the Gillespie algorithm. (E) Statistics of the original and improved transcription submodels in population-based representations.

nal model. However, this enumerated representation requires millions of variables, which is prohibitively expensive, and makes it difficult to represent the exclusion of multiple proteins from binding the same base. Furthermore, it is impractical to edit this verbose enumerated representation.

Fourth, we had to enumerate all of the arrays used by the original model because few SBML simulators support arrays. This created verbose SBML files that are difficult to interpret and maintain and slow to simulate.

In summary, we concluded that it is currently difficult to encode WC models in SBML. WC modeling would be accelerated by expanded software support for model composition, rule-based modeling, arrays, and random number generation.

### B. Submodel improvement

We also improved several aspects of the original model. As described above, we replaced the ad hoc stochastic simulation algorithms and rate laws used by the original submodels with the Gillespie algorithm and mass action kinetics. As an example, Fig. 1 and File S1 compare the original and SBML versions of the transcription submodel. We anticipate that these changes will improve the biological accuracy of WC models. The original model used these ad hoc algorithms and rate laws to achieve sufficient performance. Going forward, a high-performance parallel simulator is needed to achieve adequate performance of the Gillespie algorithm.

### C. Model integration

The integration group created a scheme for combining the submodels. First, they defined the global species as the union of all submodel species. Second, they standardized the species names to create consistent submodel-global species interfaces.

Third, the group designed a new multi-algorithm simulation strategy to overcome the limitations of the original simulation algorithm. In particular, the group sought to correctly implement the arrow of time by integrating submodels within the same time step based on the same input state. The integration group also sought to develop an algorithm that has a variable time step that can be optimized to balance accuracy and performance. (1) The group considered sequentially integrating the submodels within each time step and setting the time step small enough that only one submodel would advance

Table II
NEW STANDARDS AND SOFTWARE NEEDED TO ACCELERATE WC MODELING.

| Type | Description |
|---|---|
| Database | Expanded molecular biological databases such as ChEBI [39] |
| Software | Data curation tools for aggregating the data to build models |
| Software | Pathway/genome database to organize model training data |
| Standard | Sequence- and rule-based multi-algorithmic modeling language |
| Software | Model design tools that generate models from pathway/genome databases |
| Software | Distributed parameter estimation tools |
| Software | Frameworks for systematically verifying model |
| Software | High-performance, parallel, rule-based multi-algorithm simulator |
| Standard | Extended SBGN standard for hybrid maps containing Process Description, Entity Relationship, and Activity Flow nodes |
| Software | Visualization software that supports contextual zooming |

the cell state within each time step. However, this strategy is prohibitively expensive. (2) The group considered generalizing the original algorithm by dividing each of the global species pools into multiple, independent sub-species pools for each submodel; integrating the submodels in parallel; and merging the sub-species to update the global species. However, it is difficult to apply this strategy to coupled variables such as those that represent the protein occupancy of the chromosome. (3) The group decided to interpret the species changes predicted by each submodel as requests and implement a central controller that accepts or rejects these changes at the end of each time step to update the global species. This strategy is computationally efficient and generalizable.

Lastly, the group explored implementing this algorithm using both the SBML hierarchical model composition package [26] and SED-ML shared variables. The group concluded that both implementations are feasible. The group used iBioSim to test these strategies because iBioSim is one of the only SBML-compatible simulators that supports model composition.

### D. Annotation, documentation, and visualization

The documentation group was responsible for annotating the model. The group aimed to define every model element independently from external databases and to provide cross references to databases where possible to help users interpret the model. For example, they used InChI [40] to define small molecule species in terms of structures. They defined DNA, RNA, and protein species as polymers of small molecules. The group wrote scripts to identify cross references for each model entity. However, many entities are not represented by any database. The group contributed the missing metabolite structures to ChEBI [39] and concluded that the biological databases must be expanded to help aggregate data for models.

The group also helped the other groups visualize submodels by providing advice on SBGN and diagramming tools such as SBGN-ED [41], a VANTED add-on for creating, editing and validating SBGN diagrams. The main visualization problem encountered by the group was that WC models require large, intuitive diagrams that are difficult to lay out automatically.

### E. Progress and future work

We produced draft SBML and SBGN versions of the submodels. However, significant work remains to combine, identify, and verify the submodels. Using the lessons learned, a subgroup of the participants are continuing to recode the submodels and integrate the submodels into a single model. We expect that the final model will be more scalable, extensible, and easy to use than the original model. We also plan to build an SBML-compatible multi-algorithm simulator by expanding analysis tools, such as iBioSim and BioUML.

After recoding the model, we plan to identify and validate the new model. We will validate the model in two steps. (1) We will use the experimental data that was used to validate the original model. (2) To more rigorously validate the new model, we will compare the model to newly published single-gene deletion strain growth rates [12] that were not available when the original model was developed.

We aim to publish the SBML-encoded model to BioModels, along with SED-ML tests, SBGN diagrams, and textual documentation. Publication in BioModels will make the model searchable, retrievable, and reusable. We believe this valuable community resource will demonstrate how to describe WC models in standard formats, and it will help other researchers build upon the model.

## IV. TOWARD SBML-, SED-ML-, AND SBGN-BASED STANDARDS FOR WC MODELING

The school was the first attempt to encode a WC model using standards. Thus, we were not surprised to learn that the current standards and community software do not easily support WC modeling. Importantly, the school generated ideas for new WC modeling standards and software that will enable researchers to build vastly more comprehensive models.

### A. New standards

Two new standards are needed to facilitate WC modeling. A new SBML package should be created to support DNA, RNA, and protein sequence-based reaction patterns. This would enable researchers to easily model sequence-dependent reactions such as the methylation or protein binding of specific DNA motifs. This package would also help integrate genomics and bioinformatics with systems modeling.

SBGN must also be expanded to support (1) hybrid diagrams that contain Process Description, Entity Relationship, and Activity Flow elements and (2) visualizations at multiple levels of granularity.

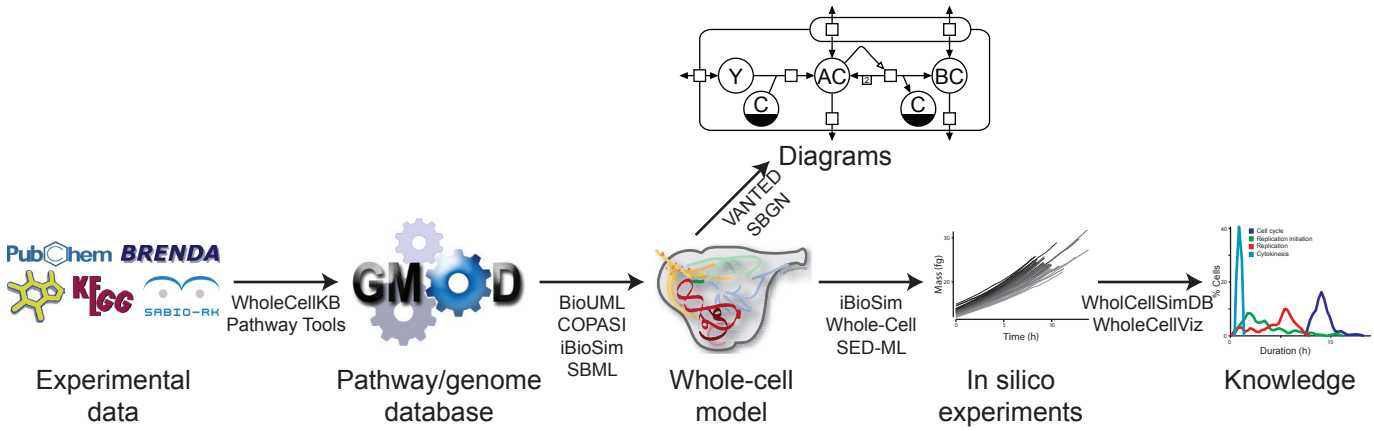



Figure 2. WC modeling workflow. Researchers will (1) assemble data into pathway/genome databases, (2) use these databases to construct models, (3) identify and verify models, (4) use multi-algorithm simulators to conduct *in silico* experiments, and (5) analyze these experiments to discover biology.

### B. New software tools and databases

Several new software tools and databases are needed to accelerate WC modeling (Table II). A high-performance simulator must be developed. This simulator should be parallelized to enable the simulation of vastly larger models that require more computing and memory than are available on a single machine. This requires research to determine how to concurrently integrate mathematically heterogeneous submodels that share state. The simulator should leverage recent advances in parallel discrete event simulation [42].

The simulator must also implement the SBML Multistate and Multicomponent Species package [43] to support rule-based modeling. This will enable more succinct model descriptions, making models easier to understand and edit. For example, translation could be described using a single reaction pattern parameterized by mRNA-specific translation initiation rates rather than by enumerating each individual reaction. By separating mathematical descriptions from parameter values, reaction patterns will also clarify the connection between dynamical models and their underlying data. Implementing this package would also enable modelers to efficiently simulate models with combinatorial state spaces, which, in turn, will enable the encoding of more complex models.

Ultimately, to accurately predict phenotypes, WC models must also represent spatially-dependent processes. Currently, researchers are independently pursuing WC and spatial modeling. For example, the *M. genitalium* model only represents three compartments, and the most advanced spatial models only represent individual pathways. WC and spatial modeling should be combined by adding support for the SBML Spatial Processes package [29] to the new WC simulator.

New model design software must be developed to help researchers quickly build WC models. This software should help researchers systematically build WC models from experimental data organized into pathway/genome database. In turn, this software will help researchers build bigger models.

New data curation tools are needed to aggregate data to build more comprehensive models. The software should automatically aggregate data from public databases, as well as accelerate manual curation from individual publications. This software will also make WC models more reproducible by automatically recording each data source. Natural language processing [44], crowdsourcing [45], and machine learning should also be explored to accelerate data curation.

New pathway/genome database software is needed to organize the data required to build WC models. To clarify the connection between computational models and their underlying experimental data, this software should use semantic annotations to describe how experimental data is used to build computational models.

New model parameter estimation and model verification tools are also needed to identify and verify computationally expensive WC models. To better estimate WC models, we must generalize our model reduction methods and adopt distributed numerical optimization techniques [46]. To more systematically verify WC models, we should adopt formal probabilistic verification techniques from electrical engineering [47].

New algorithms are needed to automatically create intuitive visualizations of large networks and the SBGN viewers should utilize contextual zooming to display diagrams at multiple levels of granularity.

In addition, biological databases, such as ChEBI, must be expanded to help researchers annotate WC models in terms of external entities.

### C. Systematic WC modeling pipeline

The new standards and software tools will enable a five step approach to WC model-driven discovery (Fig. 2). (1) Researchers will use data curation tools to aggregate heterogeneous data into pathway/genome databases. These databases will use semantic annotations to describe the connection between models and their underlying data. (2) Researchers will use design tools to build WC models from pathway/genome databases. These tools will export models to software-independent formats such as SBML. (3) Model identification and verification tools will be used to estimate parameters and test models. (4) A multi-algorithm simulator will be used to conduct *in silico* experiments. (5) Simulation databases and visualization software such as WholeCellSimDB

and WholeCellViz will be used to discover new biology by visualizing and analyzing *in silico* experiments.

Together, this pipeline will enable more researchers to more easily build, manage, simulate, and reproduce WC models. These new tools will also enable researchers to build more comprehensive models of more complex eukaryotic cells. Ultimately, this will enable WC modeling to support synthetic biology and personalized medicine.

## V. Conclusion

The 2015 Whole-Cell Modeling Summer School trained young scientists in WC modeling and standards by challenging them to recode a WC model in SBML. Additional courses are needed to provide theoretical training in multi-algorithm modeling, model reduction, and parameter estimation, as well as practical training in WC model building.

We made significant strides toward recoding the model in SBML. We also improved the model by replacing the ad hoc algorithms and rate laws used by the original model with the Gillespie algorithm and mass action kinetics. We designed an improved multi-algorithm simulation meta-algorithm. Through validating the model by comparison to quantitative growth rate measurements, we anticipate that we will also discover and add several unknown parallel pathways to the model. We have produced preliminary SBML versions of all of the submodels of the *M. genitalium* model and we are working to develop a software program to simulate the combined model. We plan to publish the new SBML-encoded model to BioModels.

Most importantly, our community discussions generated clear goals for new WC modeling software and standards. We recommend that researchers develop a new SBML-compatible simulator that supports both model composition and sequence- and rule-based modeling, as well as develop new model design, parameter estimation, model testing, and visualization tools. We also recommend expanding the biological databases to facilitate model building and annotation. Furthermore, we believe that SBGN should be extended to support hybrid diagrams, advanced graph layout, and contextual zooming. Lastly, we recommend evaluating CellML as another potential WC modeling standard.

In summary, we believe that WC modeling will be an important tool for biological science, bioengineering, and medicine. Achieving this potential requires new WC modeling software and standards. In turn, this requires expanding the WC modeling field, including training young researchers.

## References


[1] J. L. Reed, T. R. Patel, K. H. Chen *et al.*, "Systems approach to refining genome annotation," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 103, no. 46, pp. 17 480–17 484, 2006.

[2] J. W. Lee, D. Na, J. M. Park *et al.*, "Systems metabolic engineering of microorganisms for natural and non-natural chemicals," *Nat. Chem. Biol.*, vol. 8, no. 6, pp. 536–546, 2012.

[3] D. S. Lee, H. Burd, J. Liu *et al.*, "Comparative genome-scale metabolic reconstruction and flux balance analysis of multiple *Staphylococcus aureus* genomes identify novel antimicrobial drug targets," *J. Bacteriol.*, vol. 191, no. 12, pp. 4015–4024, 2009.

[4] J. Carrera and M. W. Covert, "Why build whole-cell models?" *Trends Cell Biol*, vol. 25, no. 12, pp. 719–722, 2015.

[5] D. N. Macklin, N. A. Ruggero, and M. W. Covert, "The future of whole-cell modeling," *Curr. Opin. Biotechnol.*, vol. 28, pp. 111–115, 2014.

[6] J. R. Karr, K. Takahashi, and A. Funahashi, "The principles of whole-cell modeling," *Curr. Opin. Microbiol.*, vol. 27, pp. 18–24, 2015.

[7] J. R. Karr, A. H. Williams, J. D. Zucker *et al.*, "Summary of the DREAM8 Parameter Estimation Challenge: Toward parameter identification for whole-cell models," *PLoS Comput. Biol.*, vol. 11, no. 5, p. e1004096, 2015.

[8] M. Hucka, D. P. Nickerson, G. D. Bader *et al.*, "Promoting coordinated development of community-based information standards for modeling in biology: the COMBINE initiative," *Frontiers in Bioengineering and Biotechnology*, vol. 3, no. 19, 2015.

[9] E. Klipp, W. Liebermeister, A. Helbig *et al.*, "Systems biology standards—the community speaks," *Nat. Biotechnol.*, vol. 25, no. 4, pp. 390–391, 2007.

[10] F. Büchel, N. Rodriguez, N. Swainston *et al.*, "Path2Models: large-scale generation of computational models from biochemical pathway maps," *BMC Syst. Biol.*, vol. 7, no. 1, p. 116, 2013.

[11] J. R. Karr, J. C. Sanghvi, D. N. Macklin *et al.*, "A whole-cell computational model predicts phenotype from genotype," *Cell*, vol. 150, no. 2, pp. 389–401, 2012.

[12] J. C. Sanghvi, S. Regot, S. Carrasco *et al.*, "Accelerated discovery via a whole-cell model," *Nat. Methods*, vol. 10, no. 12, pp. 1192–1195, 2013.

[13] O. Purcell, B. Jain, J. R. Karr *et al.*, "Towards a whole-cell modeling approach for synthetic biology," *Chaos*, vol. 23, no. 2, p. 025112, 2013.

[14] D. Kazakiewicz, J. R. Karr, K. Langner, and D. Plewczynski, "A combined systems and structural modeling approach repositions antibiotics for mycoplasma genitalium," *Comput. Biol. Chem.*, vol. 59, pp. 91–97, 2015.

[15] J. R. Karr, J. C. Sanghvi, D. N. Macklin *et al.*, "WholeCellKB: model organism databases for comprehensive whole-cell models," *Nucleic Acids Res.*, vol. 41, no. Database issue, pp. D787–D792, 2013.

[16] J. R. Karr, N. C. Phillips, and M. W. Covert, "WholeCellSimDB: a hybrid relational/hdf database for whole-cell model predictions," *Database*, vol. 2014, no. pii, p. bau095, 2014.

[17] R. Lee, J. R. Karr, and M. W. Covert, "WholeCellViz: data visualization for whole-cell models," *BMC Bioinformatics*, vol. 14, p. 253, 2013.

[18] V. Chelliah, N. Juty, I. Ajmera *et al.*, "BioModels: ten-year anniversary," *Nucleic Acids Res.*, vol. 43, no. D1, pp. D542–D548, 2015.

[19] M. Hucka, A. Finney, H. M. Sauro *et al.*, "The Systems Biology Markup Language (SBML): A medium for representation and exchange of biochemical network

models," *Bioinformatics*, vol. 19, no. 4, pp. 524–531, 2003.

[20] W. J. Hedley, M. R. Nelson, D. P. Bullivant, and P. F. Nielson, "A short introduction to CellML," *Philos. Trans. R. Soc. Lond. A*, vol. 359, pp. 1073–1089, 2001.

[21] D. Waltemath, R. Adams, F. Bergmann *et al.*, "Reproducible computational biology experiments with SED-ML—the Simulation Experiment Description Markup Language," *BMC Syst. Biol.*, vol. 5, no. 1, p. 198, 2011.

[22] N. Le Novère, M. Hucka, H. Mi *et al.*, "The Systems Biology Graphical Notation," *Nat. Biotechnol.*, vol. 27, pp. 735–741, 2009.

[23] N. Le Novère, M. Hucka, H. Mi *et al.*, "The Systems Biology Graphical Notation," *Nat. Biotechnol.*, vol. 27, no. 8, pp. 735–741, 2009.

[24] M. Hucka, F. T. Bergmann, S. Hoops *et al.*, "The Systems Biology Markup Language (SBML): Language specification for Level 3 Version 1 core." *Journal of Integrative Bioinformatics*, vol. 12, no. 2, p. 266, 2015.

[25] D. Waltemath, F. T. Bergmann, C. Chaouiya *et al.*, "Meeting report from the fourth meeting of the Computational Modeling in Biology Network (COMBINE)," *Stand. Genomic Sci.*, vol. 9, no. 3, 2014.

[26] L. P. Smith, M. Hucka, S. Hoops *et al.*, "SBML level 3 package: Hierarchical model composition, version 1 release 3," *Journal of Integrative Bioinformatics*, vol. 12, no. 2, p. 268, 2015.

[27] S. L. Moodie, L. P. Smith, N. Le Novère *et al.*, "The distributions package for SBML level 3," 2015, accessed: 2016-02-26. [Online]. Available: http://sourceforge.net/p/sbml/code/HEAD/tree/trunk/specifications/sbml-level-3/version-1/distrib/sbml-level-3-distrib-package-proposal.pdf?format=raw

[28] B. G. Olivier and F. T. Bergmann, "The Systems Biology Markup Language (SBML) level 3 package: Flux balance constraints," *Journal of Integrative Bioinformatics*, vol. 12, no. 2, p. 269, 2015.

[29] J. C. Schaff, A. Lakshminarayana, L. P. Smith *et al.*, "SBML level 3 package: Spatial processes," 2015.

[30] A. Funahashi, Y. Matsuoka, A. Jouraku *et al.*, "CellDesigner 3.5: a versatile modeling tool for biochemical networks," *Proc. IEEE*, vol. 96, no. 8, pp. 1254–1265, 2008.

[31] H. Rohn, A. Junker, A. Hartmann *et al.*, "VANTED v2: a framework for systems biology applications," *BMC Syst. Biol.*, vol. 6, p. 139, 2012.

[32] L. P. Smith, F. T. Bergmann, D. Chandran, and H. M. Sauro, "Antimony: a modular model definition language," *Bioinformatics*, vol. 25, no. 18, pp. 2452–2454, 2009.

[33] F. Kolpakov, "BioUML: visual modeling, automated code generation and simulation of biological systems," *Proc. BGRS*, vol. 3, pp. 281–285, 2006.

[34] A. Ebrahim, J. A. Lerman, B. O. Palsson, and D. R. Hyduke, "COBRApy: constraints-based reconstruction and analysis for python," *BMC Syst. Biol.*, vol. 7, no. 1, p. 74, 2013.

[35] S. Hoops, S. Sahle, C. Lee *et al.*, "COPASI — a COmplex PAthway SImulator," *Bioinformatics*, vol. 22, pp. 3067–3074, 2006.

[36] C. Madsen, C. J. Myers, T. Patterson *et al.*, "Design and test of genetic circuits using iBioSim," *IEEE Des. Test Comput.*, vol. 29, no. 3, 2012.

[37] E. T. Somogyi, J.-M. Bouteiller, J. A. Glazier *et al.*, "libRoadRunner: a high performance SBML simulation and analysis library," *Bioinformatics*, 2015.

[38] D. T. Gillespie, "Exact stochastic simulation of coupled chemical reactions," *J. Phys. Chem.*, vol. 81, no. 25, pp. 2340–2361, 1977.

[39] J. Hastings, G. Owen, A. Dekker *et al.*, "Chebi in 2016: Improved services and an expanding collection of metabolites," *Nucleic Acids Res*, vol. 44, no. D1, pp. D1214–D1219, 2016.

[40] S. R. Heller, A. McNaught, S. Stein *et al.*, "Inchi-the worldwide chemical structure identifier standard," *J Cheminform*, vol. 5, no. 1, p. 7, 2013.

[41] T. Czauderna, C. Klukas, and F. Schreiber, "Editing, validating and translating of sbgn maps," *Bioinformatics*, vol. 26, no. 18, pp. 2340–2341, 2010.

[42] A. P. Goldberg, Y. H. Chew, and J. R. Karr, "Toward scalable whole-cell modeling of human cells," in *Prin Adv Discret Simul*. ACM SIGSIM, 2016.

[43] F. Zhang and M. Meier-Schellersheim, "SBML Level 3 Package Specification: Multistate/Multicomponent Species (Version 1, Release 0.1 Draft 369)," 2015, accessed: 2015-05-25. [Online]. Available: http://sbml.org/Documents/Specifications/SBML_Level_3/Packages/multi

[44] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, "Natural language processing: an introduction," *J Am Med Inform Assoc*, vol. 18, no. 5, pp. 544–551, 2011.

[45] B. M. Good and A. I. Su, "Crowdsourcing for bioinformatics," *Bioinformatics*, p. btt333, 2013.

[46] A. F. Villaverde, J. A. Egea, and J. R. Banga, "A cooperative strategy for parameter estimation in large scale systems biology models," *BMC Syst Biol*, vol. 6, no. 1, p. 1, 2012.

[47] M. Kwiatkowska, G. Norman, and D. Parker, "PRISM 4.0: Verification of probabilistic real-time systems," in *Comput Aided Verification*. Springer, 2011, pp. 585–591.

**The 2015 Whole-Cell Modeling Summer School in Rostock** included the 54 participants listed in Table SI. Photo: University of Rostock IT and Media Center.