

# Toward standards for tomorrow's whole-cell models

Dagmar Waltemath<sup>1</sup>, Falk Schreiber<sup>2,3</sup>, Jonathan R. Karr<sup>4</sup>, Chris J. Myers<sup>6</sup>, Vijayalakshmi Chelliah<sup>9</sup>, Wolfram Liebermeister<sup>11</sup>, Begum Alaybeyoglu<sup>17</sup>, Yin Hoon Chew<sup>18</sup>, Rafael S. Costa<sup>19</sup>, Denis Kazakiewicz<sup>21,22</sup>, Ilya Kiselev<sup>23</sup>, Vincent Knight-Schrijver<sup>16</sup>, Christian Knüpfer<sup>24</sup>, Matthias König<sup>30</sup>, Nikita Mandrik<sup>25</sup>, J Kyle Medley<sup>31</sup>, Sucheendra K. Palaniappan<sup>26</sup>, Martin Scharm<sup>1</sup>, Kieran Smallbone<sup>27</sup>, Je-Hoon Song<sup>28</sup>, Tom Theile<sup>1</sup>, Namrata Tomar<sup>29</sup>, Jannis Uhlenndorf<sup>10</sup> and Markus Wolfien<sup>1</sup>

<sup>1</sup>Department of Systems Biology and Bioinformatics, University of Rostock

<sup>2</sup>Clayton School of Information Technology, Monash University

<sup>3</sup>Institute of Computer Science, Martin Luther University Halle-Wittenberg

<sup>4</sup>Department of Genetics & Genomic Sciences, Icahn School of Medicine at Mount Sinai

<sup>6</sup>Department of Electrical and Computer Engineering, University of Utah

<sup>9</sup>European Molecular Biology Laboratory, European Bioinformatics Institute

<sup>10</sup>Department of Biology, Humboldt University of Berlin

<sup>11</sup>Institute of Biochemistry, Charité Medical University of Berlin

<sup>16</sup>Babraham Institute

<sup>17</sup>Department of Chemical Engineering, Boğaziçi University

<sup>18</sup>Centre for Synthetic and Systems Biology, University of Edinburgh

<sup>19</sup>Instituto Superior Técnico, University of Lisbon

<sup>20</sup>Department of Bioinformatics, Mohammad Ali Jinnah University

<sup>21</sup>Center for Statistics, Universiteit Hasselt, Hasselt BE3500, Belgium

<sup>22</sup>Center for Innovative Research, Medical University of Białystok, Białystok 15-089, Poland

<sup>23</sup>Design Technological Institute of Digital Techniques of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk

<sup>24</sup>Institut für Informatik, University of Jena

<sup>25</sup>Sobolev Institute of Mathematics, Siberian Branch of the Russian Academy of Sciences, Novosibirsk

<sup>26</sup>Rennes - Bretagne Atlantique Research Centre, Institute for Research in Computer Science and Automation

<sup>27</sup>Manchester Centre for Integrative Systems Biology, University of Manchester

<sup>28</sup>Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology

<sup>29</sup>Department of Dermatology, University Medicine, University of Erlangen-Nuremberg

<sup>30</sup>Department of Computational Systems Biochemistry, University Medicine Charité Berlin, 10117 Berlin, Germany

<sup>31</sup>Department of Bioengineering, University of Washington, Seattle, WA 98195, USA

Whole-cell models are promising tools for biological research, bioengineering, and medicine. However, significant work remains to achieve complete and accurate whole-cell models, including developing a strong theoretical understanding of multi-algorithm modeling, a standardized whole-cell modeling language, and an efficient general-purpose simulator. We organized the 2015 Whole-Cell Modeling Summer School to teach whole-cell modeling, as well as to evaluate the need for new whole-cell modeling standards by attempting to encode a recently published whole-cell model into SBML. We propose three SBML extensions to support transparent, reproducible whole-cell modeling; support for multi-algorithm models, support for particle-based state representation, and support for template reactions. In addition, we describe several new software tools which are needed to enable researchers to encode and simulate whole-cell models including a user-friendly graphical model editor and a parallelized simulator. We also propose several new SGBN extensions. Together these new standards and software tools would accelerate whole-cell modeling.

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

## I. INTRODUCTION

OVER the past twenty years, computational modeling has become an essential and powerful tool for biological research, bioengineering, and medicine to analyze high-throughput molecular measurements and understand the molecular details of complex biological systems. Computational modeling has been used to identify new metabolic

genes [1], to add metabolic pathways to bacteria [2], to and identify potential new antimicrobial drug targets [3]. Computational models also have the potential to enable bioengineers to design new microorganisms for industrial applications such as chemical synthesis, biofuel production, and waste decontamination, as well as to enable clinicians to tailor therapy to individual patients. Realizing this potential requires more comprehensive and accurate computational models which are capable of predicting cellular behavior from genotype, as well

as standardized methods for exchanging models, simulation experiments, and model visualizations [4], [5], [6], [7].

Recently, researchers at Stanford University developed the first whole-cell model of the gram-positive bacterium *Mycoplasma genitalium* [8]. The model represents the life cycle of a single Mycoplasma cell including the copy number dynamics of each metabolite, RNA, and protein species and accounts for every known gene function. The model is composed of 28 sub-models, each of which is implemented using different mathematical representations including ordinary differential equations (ODEs), flux balance analysis (FBA), and Boolean rules (BRs), and trained using different experimental data.

The *M. genitalium* whole-cell model was implemented in MATLAB, is available open-source under the MIT license, and is extensively documented [9]. This has enabled other researchers to use the model for their own research.

However, the *M. genitalium* whole-cell model software is not transparent or reusable. The *M. genitalium* whole-cell model software is also not user-friendly, computationally efficient, or easily maintainable. Consequently, significant domain expertise is required to use the *M. genitalium* model or construct new whole-cell models. New whole-cell modeling standards and simulation tools are needed to enable more researchers to develop and simulate their own whole-cell models. Such standards and software tools would accelerate whole-cell modeling. They would enable researchers to develop models more quickly, to explore models more deeply, and to evaluate models more rigorously. Furthermore, standards would simplify the submission process to model repositories such as BioModels [10], [11]. In turn, this would make models more searchable, retrievable, reusable, and comparable.

Several systems biology standards have already been developed by the COmputational Modeling in BIology Network (COMBINE) [12] including the Systems Biology Markup Language (SBML) [13], the Cell Markup Language (CellML) [14], the Simulation Experiment Description Markup Language (SED-ML) [15], and the Systems Biology Graphical Notation (SBGN) [16]. SBML and CellML are languages for describing mathematical models including ODE, logical, and FBA models. Both have been used to build thousands of models of various intracellular pathways. SED-ML is a language for describing computational experiments. SED-ML enables scientists to reproduce a simulations by completely describing simulation setups, including the simulation algorithm and every parameter value. SBGN includes three languages for describing visual representations of models. None of these standards have been used to construct, simulate, or visualize models as complex as the *M. genitalium* whole-cell model.

We organized the 2015 Whole-Cell Modeling Summer School to train students in whole-cell modeling, as well as to evaluate the need for new standards for whole-cell modeling. The majority of the school was focused on trying to encode the *M. genitalium* whole-cell model using SBML to train students, as well as to evaluate the ability of SBML to encode whole-cell models. The ultimate scientific goal of the school was to develop an open-source whole-cell model encoded in SBML and simulated using SED-ML.

Here, we describe the summer school, outline our progress

toward encoding the *M. genitalium* model using SBML, and propose several SBML and SGBN extensions to support whole-cell modeling. First, we summarize the summer school. Second, we describe our progress toward encoding the *M. genitalium* whole-cell model using SBML. Lastly, we describe the SBML and SBGN expansions and software tools needed to support whole-cell modeling.

## II. THE 2015 WHOLE-CELL MODELING SUMMER SCHOOL

We organized the summer school to teach students how to build and encode models using COMBINE standards by attempting to encode the *M. genitalium* model using only standard representation formats.

### A. Organization

The 2016 Whole-Cell Modeling Summer School was held March 9-13, 2015 at the University of Rostock in Rostock, Germany. The school was organized by Dagmar Waltemath and Falk Schreiber and supported by the Volkswagen Foundation. 45 students, nine instructors, and two organizers participated in the five-day school.

The school began with two lectures which introduced whole-cell modeling and the existing systems biology standards. Jonathan Karr from the Icahn School of Medicine at Mount Sinai, USA presented an overview of whole-cell and multi-algorithm modeling. Michael Hucka from the California Institute of Technology, USA presented an overview of the SBML, SED-ML, and SBGN standards; several open-source software tools which support these standards; and the COMBINE initiative. We also organized three discussions on multi-algorithm model composition, particle-based state representation, and random number generation.

The majority of the school was dedicated to hands-on active learning sessions in which students learned about whole-cell modeling and the COMBINE standards by trying to encode parts of the *M. genitalium* whole-cell model using SBML. Students were divided into ten groups of four to five students, each of which was challenged to encode one or more sub-models using SBML. Each group was led by an experienced instructor.

Each day concluded with brief progress reports from each group. This facilitated discussion on common encoding challenges and model integration and provided an opportunity for groups to obtain feedback from each other.

We also organized a poster session, as well as several evening social activities to provide the students opportunities to network with each other and the instructors.

### B. Educational outcomes

We surveyed the students to assess the educational outcome of the school. Most students reported gaining deep knowledge of whole-cell modeling, increased appreciation for reproducible science, and increased understanding of the SBML, SED-ML, and SBGN standards. Many students also reported learning about open-source modeling software tools relevant to their own research.

In addition, many of the students reported that the school expanded their scientific network. Several students commented that the school introduced them to potential postdoctoral positions and next year's whole-cell modeling summer school (<http://www.wholecell.org/school-2016>).

### C. Lessons learned for organizing research-based schools

We learned several valuable lessons about how to best organize an open-ended, research-based school. First, we conclude that research-based schools should clearly outline the expected background knowledge and learning objectives and have well-organized learning activities. This helps students make informed decisions about whether to participate in the school, know how to prepare for the school, and learn efficiently. Second, we conclude that students greatly enjoy learning through open research problems rather than through prescribed training exercises. This makes students feel engaged, challenged, and connected to research. This also helps students build practical skills to complement their foundational undergraduate and graduate training. Third, we conclude that open-ended project-based schools require a high teacher-to-student ratio, a flexible schedule, and multidisciplinary project teams. A high teacher to student ratio allows students to get feedback and iterate through potential solutions quickly. A flexible schedule enables impromptu lectures and discussions. Multidisciplinary teams enable students to work through difficult problems by drawing on perspectives from multiple fields.

## III. TOWARD AN SBML-ENCODED WHOLE-CELL MODEL

In addition to training young computational systems biology researchers, the second goal of the school was to attempt to encode the *M. genitalium* whole-cell model (<https://github.com/CoverLab/WholeCell/releases/tag/v1.1>) into SBML. To achieve this goal, most of the course was devoted to active learning sessions in which students were challenged to encode sub-models of the *M. genitalium* into SBML, integrate sub-models into a single model, and simulate models using SED-ML. During these sessions, the students and instructors were divided into ten groups. Eight of the groups were tasked with encoding one or more sub-models. The ninth group was tasked with developing a standards-compliant scheme to integrate the sub-models into a single model. This group was responsible for defining the global state variables and sub-model interfaces and developing a SED-ML scheme to simulate the integrated model. The tenth group was responsible for developing an annotation scheme and helping the other groups document and visualize their sub-models. Table SI lists the ten groups and all of the students and instructors.

### A. Sub-model encoding

The eight sub-model encoding groups pursued various strategies to encode the sub-models using SBML. Several of the groups encoded sub-models by first reading the sub-model documentation, then drawing pathway maps using software tools such as Cell Designer [17] and VANTED [18], and

finally writing scripts to generate SBML models from their maps. Other groups used modeling software tools such as BioUML [19], COPASI [20], and iBioSim [21] to encode sub-models based on their documentation. A few of the groups encoded sub-models by converting the MATLAB code to SBML. These groups then generated SBGN maps from their SBML to better understand their sub-models.

The groups encountered several challenges to encoding the *M. genitalium* sub-models into SBML. First, most of the groups had to spend a significant amount of time reading the MATLAB code and documentation to understand the details of the *M. genitalium* sub-models because the connection between the sub-models and the associated pathway/genome database is not transparent, many of the sub-models details are implemented directly in MATLAB code rather than in a transparent language such as SBML, and the documentation only provides overviews of the sub-models. Fortunately, one of the authors of the *M. genitalium* model was available to answer questions about the model.

A second challenge to encoding the sub-models in SBML was encoding serially executed MATLAB sub-models into SBML which, because it is not a programming language, does not expose control over the order of simulation execution. This fundamental difference between programming languages and SBML makes quantitatively reproducing the *M. genitalium* model impossible. Most of the groups decided to tackle this problem by formalizing MATLAB sub-models as discrete stochastic models and simulating them using the Gillespie or other approximate algorithms. For several of the sub-models, this conversion imposed an explicit internal sub-model timescale which was not present in the original MATLAB sub-model due to the lack of kinetic data for the corresponding pathway.

The fact that SBML is not a programming language and does not expose methods for arbitrary random number generation also made it challenging for groups to encode the random algorithms used by the MATLAB sub-models into SBML. For example, the MATLAB translation sub-model includes a random algorithm which assigns amino acids to individual polypeptides. Because this algorithm is not equivalent to the Gillespie algorithm, the algorithm cannot easily be encoded into SBML. Most of the groups also solved this problem by formalizing sub-models as stochastic models. Even if it were possible to transcode the MATLAB sub-models directly into SBML, it would still be difficult to quantitatively reproduce the MATLAB simulations because SBML does not expose control over the random number generator algorithm or seed. Consequently, it would only be feasible to compare the first two moments of the MATLAB and transcoded model simulations.

To encode many of the sub-models into SBML, the groups also had to either enumerate the hybrid population/particle-based state representation used by the MATLAB sub-models or approximate the MATLAB sub-models. The groups responsible for the transcription and translation sub-models chose to approximate the MATLAB sub-models by eliminating the internal dynamics of the polymerization of each RNA and polypeptide. Consequently, these sub-models no longer track the progress of individual RNA polymerases and ribosomes,

account for base-specific transcription or translation rates, or predict RNA polymerase collisions. The groups responsible for the DNA sub-models including replication, replication initiation, and transcriptional regulation, chose to enumerate the sparse chromosome representation used by the MATLAB model by creating Boolean indicator variables to represent the existence and protein-binding status of each chromosome base. This enumerated representation requires millions of variables. Consequently, the corresponding SBML XML files are impractical to read and computationally expensive to simulate. Enumerating the rules which govern the joint values of the enumerated variables, such as the rules which represent the steric effects of DNA-bound proteins by preventing proteins from binding neighboring bases, is also impractical. Furthermore, the enumerated SBML files are impractical for humans to read, edit, or maintain.

The lack of universal SBML simulator support for arrays was another challenge to encoding the MATLAB sub-models into SBML. All of the groups overcame the lack of array support by enumerating individual array elements of the MATLAB and all matrix algebra computations. This creates verbose SBML files which are more difficult to interpret, maintain, and edit. Enumerating the matrix algebra computations also increases the computational cost of simulation.

Together, these five challenges made it very difficult for the groups to encode most of the MATLAB sub-models into SBML. Going forward, SBML and the SBML simulators should be expanded to provide support for random number generation, particle-based state representation, and arrays.

### B. Model integration

The integration group was responsible for assembling the individual sub-models into a single model including defining the global state variables, defining the interfaces exposed by the sub-models to the global state variables, and developing a scheme for managing simultaneous writing of shared state variables. The integration group defined the global state variables as the union of all state variables shared by at least two sub-models rather than by explicitly defining a set of global state variables as done by the MATLAB simulator. The advantages of this approach are that sub-model developers are not also required to develop global state variables and that it minimizes the number of global state variables. The disadvantages of this approach are that the total set of variables is less transparent and that it requires users to learn all of the sub-models and their naming conventions to analyze model simulations.

The integration group standardized the interfaces exposed by the individual sub-models by defining a variable naming convention. This naming convention ensures, for example, that the copy numbers of each protein species are represented by variables with the same names in each of the sub-models. This convention makes it clear how multiple local sub-model variables map onto the same global variable. Specifically, the integration group chose to use the same variable names as those used by the MATLAB implementation. Matrix and particle-based variables were enumerated by creating multiple

variables with names containing additional suffixes to indicate their identity.

The primary challenge faced by the integration group was how to handle concurrent editing of shared state variables by multiple sub-models. The integration group explored several potential strategies to manage concurrent writing. First, they explored sequentially simulating the sub-models and updating the global state variables. This avoids the need for more complex strategies to merge variable changes. However, under this approach sub-models are simulated with different variable values within each timestep. Consequently, simulation predictions are sensitive to the sub-model execution order.

The integration group also explored several more complex sub-model integration strategies which enable all of the sub-models to be simulated with the same variable values within each timestep. These strategies included reducing the sub-model integration timestep such that sub-models do not request conflicting variable changes; dividing each of the shared state variables into separate, independent sub-variables for each sub-model, simulating the sub-models in parallel, and merging the sub-variables to compute the update global values; and using semaphores to manage concurrent variable changes whereby at each timestep sub-models request sets of atomic state variables changes and a controller decides which change sets are processed. Each of these strategies has different advantages and disadvantages. The first strategy is the simplest to understand and implement, but is computationally expensive. The second strategy is simple to implement and computationally efficient for independent variables, but is difficult to implement for sets of non-independent variables such as those which represent the protein occupancy of the chromosome. The third strategy is the most complex to implement, but is more general than the second strategy and more computationally efficient than the first. The integration group tested their sub-model integration strategies using iBioSim because iBioSim is one of the only SBML simulators to support all of the existing needed SBML packages including hierarchical model composition (*comp*), arrays, and flux balance constraints (*fbc*).

The lack of an SBML simulator which supports multi-algorithm model composition was another challenge to integrating the sub-models. The integration group plans to overcome this limitation by adding support for multi-algorithm to iBioSim.

### C. Progress

The students produced preliminary SBML and SBGN-ML versions of many of the *M. genitalium* whole-cell model sub-models. However, significant work remains to finish encoding the sub-models, integrate the sub-models into a single model, and test the sub-models and the combined model. Complete drafts are only available for a few of the simplest sub-models. The current drafts of most of the more complex sub-models are also greatly simplified compared to their MATLAB versions. For example, the transcription and translation sub-models drafts do not represent the polymerization of individual bases and the DNA sub-models do not account for the protein occupancy of the chromosome. In addition, none of the



SBML-encoded sub-models have been thoroughly tested by replicating the unit tests from the MATLAB implementation. Furthermore, none of the SBML sub-models have been thoroughly documented and complete SBGN maps are not yet available for any of the sub-models.

#### D. Future steps

We hope to finish encoding the *M. genitalium* sub-models into SBML and integrate the SBML-encoded sub-models into a single model simulatable by open-source software tools such as COPASI, BioUML, and iBioSim. Several students and instructors have continued to keep working and meeting online. Several students and instructors also plan to participate in a second meeting in October 2015 which will be held at the University of Utah, USA immediately prior to the 2015 COMBINE Forum.

Going forward, we plan to publish SBML-encoded versions of each of the *M. genitalium* sub-models to BioModels, along with SED-ML tests, SBGN maps, and textual documentations. This would make the sub-models searchable, retrievable, and reusable by other scientists. We believe this would be a valuable community resource. It would demonstrate how to build a whole-cell model and enable other researchers to build upon the *M. genitalium* sub-models. Ultimately, SBML and the SBML simulation software tools also need to be extended to facilitate whole-cell modeling.

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

Prior to the school, SBML, SED-ML, and SBGN had never been used to represent whole-cell or similarly complex models. Consequently, not surprisingly, the summer school revealed several limitations of SBML, SBGN, and the existing simulation software for large models. Importantly, the summer school facilitated discussions among modelers, tool developers, and model curators about how to best expand the existing standards and software tools to overcome these limitations.

#### A. Current limitations

As discussed above, SBML does not support multi-algorithm modeling, particle-based state representation, or arbitrary random number generation and the current SBML simulators have limited support for arrays. Consequently, SBML cannot efficiently simulate models with large, combinatorial state spaces; represent arbitrary stochastic models; or efficiently simulate arbitrary mathematical models involving matrix algebra operations. State spaces must be enumerated, stochastic models must be described using a ratio time scale and simulated using the Gillespie or other approximate algorithm, and matrix algebra operations must be expanded. For the *M. genitalium* whole-cell model this would result in large, unmanageable XML files and millions of variables.

Furthermore, no general-purpose SBML simulation software is currently capable of efficiently simulating models involving millions of variables and no SBML simulation software supports multi-algorithm modeling. In addition, no

SBML editing tool provides researchers a user-friendly interface for editing SBML files containing millions of variables. Consequently, currently, whole-cell model SBML files must be generated by scripts and cannot be easily edited.

The summer school also revealed several limitations of the graphical representation scheme SBGN and the existing SBGN viewers for whole-cell models. First, whole-cell modeling requires hybrid SBGN maps which contain all three types of nodes: Process Descriptions, Entity Relationships, and Activity Flows. Currently, SBGN maps can only contain one type of node. Second, whole-cell modeling requires new automatic layout algorithms suitable for large maps. The existing layout algorithms are unable to construct intuitive maps of many of the pathways that need to be in whole-cell models. Third, in order to effectively visualize complex SBGN maps of whole-cell models, SBGN viewers must be able to display maps at different levels of granularity by automatically reducing maps.

The summer school did not reveal any limitations of SED-ML for describing whole-cell model simulations.

#### B. Standard extensions

Taken together, expanded standards and supporting software tools are needed to facilitate accurate, reproducible whole-cell modeling. We propose three SBML expansions to facilitate whole-cell modeling. First, the SBML comp package must be expanded to support models composed of sub-models implemented with different simulation algorithms. Currently, the comp package only supports models composed of sub-models each implemented using the same simulation algorithm. In addition, the existing SBML simulators must be expanded to support multi-algorithm simulations and/or new simulators must be developed which support the expanded package. This requires significantly more research to determine the best ways to integrate heterogeneous sub-models, including rigorously evaluating the advantages and disadvantages of each of the schemes proposed by the integration group. Significant effort will also be needed to develop a parallelized simulator which is capable of quickly simulating complex whole-cell models.

Second, a new SBML package must be created to support hybrid population/particle-based state representations such as those used by BioNetGen [22], [23] and NFSim [24]. In parallel, the existing SBML simulators must be expanded to support this new package and/or new SBML simulators must be developed. This would enable modelers to compactly describe and efficiently simulate models with large, combinatorial state spaces. The compact descriptions enabled by this package would also make models more transparent and easier to maintain and expand.

A new SBML package must also be created to support reaction templates so that, for example, translation could be described using a single reaction template and arrays of mRNA-specific translation initiation rates and codon-specific elongation rates. Such reaction templates would enable whole-cell and other large models to be compactly described, and consequently easily interpretable, maintainable, and editable. By separating the mathematical descriptions and quantitative parameter values, reaction templates would also make the

connection between dynamical models and the experimental data used to inform their parameter values more transparent. The new reaction templates could be expanded for backward compatibility with older SBML simulators.

New user-friendly graphical editors must also be developed to enable researchers to easily build SBML files which take advantage of these new features. These graphics editors must also allow researchers to transparently map model parameters onto experimental data organized in pathway/genome databases.

In addition, as discussed above, SBGN must be expanded to support hybrid diagrams and the SBGN viewers must be expanded to support more automatic layout algorithms, automatic model reduction, and contextual zooming. These features would enable researchers to use SBGN to map whole-cell and other large models.

Together, these SBML, SBGN, and software expansions would enable more researchers to more easily build, manage, simulate, and reproduce whole-cell models and simulations. These new standards and software tools would also enable researchers to build more comprehensive and more accurate models, including of human cells. Ultimately, these new standards and software would enable whole-cell modeling to support rational biological design and personalized medicine.

## V. CONCLUSION

The 2016 Whole-Cell Modeling Summer School provided 45 young scientists hands-on training in whole-cell and multi-algorithm modeling through attempting to encode the *M. genitalium* whole-cell model into SBML. Additional summer schools and courses are needed to provide students with deeper theoretical training in dynamical modeling, multi-algorithm modeling, model reduction, and parameter estimation, as well as practical training in model construction including data curation, model building, numerical optimization, model testing, and model analysis.

The summer school also made significant strides toward encoding the *M. genitalium* whole-cell model using SBML for simulation by open-source software. The students developed preliminary SBML versions of all of the sub-models of the *M. genitalium* model. Since the summer school, several students have continued to encode the *M. genitalium* model, and several of the students and instructors are participating in a second meeting prior to 2015 COMBINE Forum at the University of Utah in Salt Lake City, USA. Ultimately, we hope to publish an SBML-encoded version of the model to BioModels.

However, significant work remains to complete, test, integrate, simulate, and document the SBML-encoded version of the *M. genitalium* model. Currently, SBML cannot represent whole-cell models, there is no simulation software program which could efficiently simulate an SBML-encoded whole-cell model, and there is no graphical editor which is capable of constructing, editing, or visualizing an SBML-encoded whole-cell model. SBML must be expanded to support multi-algorithm modeling, template reactions, particle-based state representation, and arrays and an efficient simulation software program and a user-friendly model editor must be developed to enable modelers to easily construct and simulate whole-cell

models. New parameter estimation, model testing, and visual analysis tools must also be developed to enable researchers to effectively use SBML-encoded whole-cell models for scientific research. SBGN and the SBGN viewers must also be expanded to support hybrid diagrams, automatic graph layout, automatic graph reduction, and contextual zooming.

In summary, we believe that whole-cell modeling has the potential to be an important tool for biological discovery, bio-engineering, and medicine. Achieving this potential requires improved standards for describing whole-cell models, as well as new general-purpose simulation software for reproducibly simulating whole-cell models. In turn, this requires expanding the whole-cell modeling field including training additional young researchers.

## ACKNOWLEDGMENTS

The school was supported by a grant from the Volkswagen Foundation to DW and FS. JRK was supported by a James S. McDonnell Foundation Postdoctoral Fellowship Award in Studying Complex Systems.

## REFERENCES

- [1] J. L. Reed, T. R. Patel, K. H. Chen, A. R. Joyce, M. K. Applebee, C. D. Herring, O. T. Bui, E. M. Knight, S. S. Fong, and B. O. Palsson, "Systems approach to refining genome annotation," *PNAS*, vol. 103, no. 46, pp. 17 480–17 484, 2006.
- [2] J. W. Lee, D. Na, J. M. Park, J. Lee, S. Choi, and S. Y. Lee, "Systems metabolic engineering of microorganisms for natural and non-natural chemicals," *Nature Chemical Biology*, vol. 8, no. 6, pp. 536–546, 2012.
- [3] D. S. Lee, H. Burd, J. Liu, E. Almaas, O. Wiest, A. L. Barabási, Z. N. Oltvai, and V. Kapatral, "Comparative genome-scale metabolic reconstruction and flux balance analysis of multiple *Staphylococcus aureus* genomes identify novel antimicrobial drug targets," *Journal of Bacteriology*, vol. 191, no. 12, pp. 4015–4024, 2009.
- [4] D. N. Macklin, N. A. Ruggero, and M. W. Covert, "The future of whole-cell modeling," *Current Opinion in Biotechnology*, vol. 28, pp. 111–115, 2014.
- [5] J. R. Karr, K. Takahashi, and A. Funahashi, "The principles of whole-cell modeling," *Current Opinion in Microbiology*, vol. (in press), 2015.
- [6] M. Hucka, D. P. Nickerson, G. D. Bader, F. T. Bergmann, J. Cooper, E. Demir, A. Garny, M. Golebiewski, C. J. Myers, F. Schreiber *et al.*, "Promoting coordinated development of community-based information standards for modeling in biology: the COMBINE initiative," *Frontiers in Bioengineering and Biotechnology*, vol. 3, 2015.
- [7] E. Klipp, W. Liebermeister, A. Helbig, A. Kowald, and J. Schaber, "Systems biology standards - the community speaks," *Nature Biotechnology*, vol. 25, no. 4, pp. 390–391, 2007.
- [8] J. R. Karr, J. C. Sanghvi, D. N. Macklin, M. V. Gutschow, J. M. Jacobs, B. Bolival, Jr, N. Assad-Garcia, J. I. Glass, and M. W. Covert, "A whole-cell computational model predicts phenotype from genotype," *Cell*, vol. 150, no. 2, pp. 389–401, 2012.
- [9] Covert Lab. Whole Cell Model. [Online]. Available: <https://github.com/CovertLab/WholeCell/releases/tag/v1.1>
- [10] N. Juty, R. Ali, M. Glont, S. Keating, N. Rodriguez, M. Swat, S. Wimalaratne, H. Hermjakob, N. Le Novère, C. Laibe *et al.*, "Biomodels: Content, features, functionality, and use," *CPT: Pharmacometrics & Systems Pharmacology*, vol. 4, no. 2, pp. 1–14, 2015.
- [11] V. Chelliah, N. Juty, I. Ajmera, R. Ali, M. Dumousseau, M. Glont, M. Hucka, G. Jalowicki, S. Keating, V. Knight-Schrijver *et al.*, "Biomodels: ten-year anniversary," *Nucleic acids research*, vol. 43, no. D1, pp. D542–D548, 2015.
- [12] N. Le Novère, M. Hucka, N. Anwar, G. D. Bader, E. Demir, S. Moodie, and A. Sorokin, "Meeting report from the first meetings of the Computational Modeling in Biology Network (COMBINE)," *Standards in Genomic Sciences*, vol. 5, no. 2, p. 230, 2011.

- [13] M. Hucka, A. Finney, H. M. Sauro, H. Bolouri, J. C. Doyle, H. Kitano, A. P. Arkin, B. J. Bornstein, D. Bray, A. Cornish-Bowden, A. A. Cuellar, S. Dronov, E.-D. Gilles, M. Ginkel, V. Gor, I. I. Goryanin, W. J. Hedley, T. C. Hodgman, J.-H. Hofmeyr, P. J. Hunter, N. S. Juty, J. L. Kasberger, A. Kremling, U. Kummer, N. L. Novère, L. M. Loew, D. Lucio, P. Mendes, E. D. Mjolsness, Y. Nakayama, M. R. Nelson, P. F. Nielsen, T. Sakurada, J. C. Schaff, B. E. Shapiro, T. S. Shimizu, H. D. Spence, J. Stelling, K. Takahashi, M. Tomita, J. Wagner, and J. Wang, "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.
- [14] W. J. Hedley, M. R. Nelson, D. P. Bullivant, and P. F. Nielsen, "A short introduction to CellML," *Philosophical Transactions of the Royal Society of London A*, vol. 359, pp. 1073–1089, 2001.
- [15] D. Waltemath, R. Adams, F. Bergmann, M. Hucka, F. Kolpakov, A. Miller, I. Moraru, D. Nickerson, S. Sahle, J. Snoep, and N. Le Novère, "Reproducible computational biology experiments with SED-ML—the Simulation Experiment Description Markup Language," *BMC Systems Biology*, vol. 5, no. 1, p. 198, 2011.
- [16] N. Le Novère, M. Hucka, H. Mi, S. Moodie, F. Schreiber, A. Sorokin, E. Demir, K. Wegner, M. Aladjem, S. M. Wimalaratne, F. T. Bergman, R. Gauges, P. Ghazal, H. Kawaji, L. Li, Y. Matsuoka, A. Villéger, S. E. Boyd, L. Calzone, M. Courtot, U. Dogrusoz, T. Freeman, A. Funahashi, S. Ghosh, A. Jouraku, S. Kim, F. Kolpakov, A. Luna, S. Sahle, E. Schmidt, S. Watterson, G. Wu, I. Goryanin, D. B. Kell, C. Sander, H. Sauro, J. L. Snoep, K. Kohn, and H. Kitano, "The systems biology graphical notation," *Nature Biotechnology*, vol. 27, pp. 735–741, 2009.
- [17] A. Funahashi, Y. Matsuoka, A. Jouraku, M. Morohashi, N. Kikuchi, and H. Kitano, "CellDesigner 3.5: a versatile modeling tool for biochemical networks," *Proceedings of the IEEE*, vol. 96, no. 8, pp. 1254–1265, 2008.
- [18] H. Rohn, A. Junker, A. Hartmann, E. Grafahrend-Belau, H. Treutler, M. Klapperstück, T. Czauderna, C. Klukas, and F. Schreiber, "VANTED v2: a framework for systems biology applications," *BMC Systems Biology*, vol. 6, p. 139, 2012.
- [19] F. Kolpakov, "BioUML: visual modeling, automated code generation and simulation of biological systems," *Proceedings BGRS*, vol. 3, pp. 281–285, 2006.
- [20] P. Mendes, S. Hoops, S. Sahle, R. Gauges, J. Dada, and U. Kummer, "Computational modeling of biochemical networks using COPASI," *Methods in Molecular Biology*, vol. 500, pp. 17–59, 2009.
- [21] C. Madsen, C. J. Myers, T. Patterson, N. Roehner, J. T. Stevens, and C. Winstead, "Design and test of genetic circuits using iBioSim," *IEEE Des Test Comput*, vol. 29, no. 3, 2012.
- [22] W. S. Hlavacek, J. R. Faeder, M. L. Blinov, R. G. Posner, M. Hucka, and W. Fontana, "Rules for modeling signal-transduction systems," *Sci STKE*, vol. 2006, no. 344, p. re6, 2006.
- [23] J. S. Hogg, L. A. Harris, L. J. Stover, N. S. Nair, and J. R. Faeder, "Exact hybrid particle/population simulation of rule-based models of biochemical systems," *PLOS Comp Biol*, vol. 10, no. 4, p. e1003544, 2014.
- [24] M. W. Sneddon, J. R. Faeder, and T. Emonet, "Efficient modeling, simulation and coarse-graining of biological complexity with NFsim," *Nat Methods*, vol. 8, no. 2, pp. 177–183, 2011.



**2015 Whole-Cell Modeling Summer School** included the 56 researchers listed in Table SI.

Table SI  
2015 WHOLE-CELL MODELING SUMMER SCHOOL CONSORTIUM MEMBERS.

Group	Participant	Affiliation
Cytokinesis	Naveen Kumar Aranganathan Daniel Alejandro Priego Espinosa Ilya Kiselev Wolfram Liebermeister Yan Zhu	University Paris-Sud, France National Autonomous University of Mexico, Mexico Siberian Branch of the Russian Academy of Sciences Novosibirsk, Russia Charité Medical University of Berlin, Germany Monash University, Australia
DNA repair	Arne Bittig Vijayalakshmi Chelliah Audald Lloret-Vilas Mahesh Sharma Namrata Tomar	University of Rostock, Germany European Bioinformatics Institute, UK European Bioinformatics Institute, UK National Institute of Pharmaceutical Education and Research, India Friedrich-Alexander University of Erlangen-Nürnberg, Germany
Metabolism	Kambiz Baghalian Frank T. Bergmann Rafeal Sousa Costa Matthias König Kieran Smallbone Milenko Tokic	University of Oxford, UK California Institute of Technology, USA University of Lisbon, Portugal Charité Medical University of Berlin, Germany University of Manchester, UK Swiss Federal Institute of Technology in Lausanne, Switzerland
Protein	Begum Alaybeyoglu Matteo Cantarelli Yin Hoon Chew Marcus Krantz Daewon Lee	Boğaziçi University, Turkey OpenWorm, UK University of Edinburgh, UK Humboldt University of Berlin, Germany KAIST, South Korea
Replication	Vincent Knight-Schrijver Je-Hoon Song Jannis Uhlendorf Dagmar Waltemath James Yurkovich Anna Zhukova	Babraham Institute, UK KAIST, South Korea Humboldt University of Berlin, Germany University of Rostock, Germany University of California, San Diego, USA University of Bordeaux, France
Replication initiation	Harold Gomez Jens Hahn Michael Hucka Nikita Mandrik Martin Scharm Florian Wendland	Boston University, USA Humboldt University of Berlin, Germany California Institute of Technology, USA Siberian Branch of the Russian Academy of Sciences Novosibirsk, Russia University of Rostock, Germany University of Rostock, Germany
RNA	Tuure Hameri Jesse Kyle Medley Sucheendra Kumar Palaniappan Pinar Pir Natalie Stanford Markus Wolfien	Swiss Federal Institute of Technology in Lausanne, Switzerland University of Washington, USA Institute for Research in Computer Science and Automation, France Babraham Institute, UK University of Manchester, UK University of Rostock, Germany
Translation	Joseph Cursons Muhammad Haseeb Daniel Hernandez Denis Kazakiewicz Pedro Mendes Hojjat Naderi Meshkin	University of Melbourne, Australia Mohammad Ali Jinnah University, Pakistan Swiss Federal Institute of Technology in Lausanne, Switzerland University of Hasselt, Belgium University of Manchester, UK Academic Center for Education, Culture and Research, Iran
Integration	Paulo Eduardo Pinto Burke Tobias Czauderna Bertrand Moreau Chris J. Myers Thawfeek Mohamed Varusai Argyris Zardilis	Federal University of São Paulo, Brazil Monash University, Australia CoSMo Company, France University of Utah, USA University College Dublin, Ireland University of Edinburgh, UK
Visualization, annotation and documentation	Christian Knüpfer Falk Schreiber Tom Theile	University of Jena, Germany Monash University, Australia University of Rostock, Germany
Modeling instructor	Jonathan R. Karr	Icahn School of Medicine at Mount Sinai, USA