CENTER FOR REPRODUCIBLE BIOMEDICAL MODELING

Annual Newsletter VOL.1 MARCH 2022

A Year of New Innovations

BY VERONICA PORUBSKY, JANIS SHIN, & HERBERT SAURO

CRBM Team of Directors. From left to right: Drs. Herbert Sauro, Jonathan Karr, John Gennari, and Ion Moraru

Welcome to the first newsletter from the Center for Reproducible Biomedical Modeling (CRBM) funded jointly by the NIBIB, NIGMS, and NSF. We hope this will be a biannual publication to broadcast the latest developments, commentary, and news in the world of biomedical modeling.

Our center started in 2018 as part of an effort to improve best practices in publishing and deploying biomedical models with a particular emphasis on subcellular and physiological models. These are the kinds of models one might construct using ordinary or partial differential equations, stochastic reaction-based models, etc. Such models describe mechanistic aspects of a disease or biological phenomena. In this first phase, our center is particularly concerned with the reproducibility of published biomedical models.

It is no secret that scientific reproducibility has been of some concern in the wider scientific community with numerous articles in the popular scientific press lamenting the fact that many published works are either difficult or impossible to replicate. Many lab experiments today are complex and perhaps it is no surprise that this can happen. However, one might expect computational experiments to be highly reproducible since we are dealing with very defined problems that essentially boil down to just ones and zeros. What could possibly go wrong? A lot it seems. Over the last decade or more we have had anecdotal evidence that a large fraction of published biomedical models were not reproducible. This was confirmed last year when a group from EBI, Cambridge (Tiwari et al., 2021) published a paper that analyzed the question more formally. They discovered that at least half of all published models were not reproducible.

Of course, this begs the question: Does it really matter? There are several aspects to this. The first is that the scientific method relies on

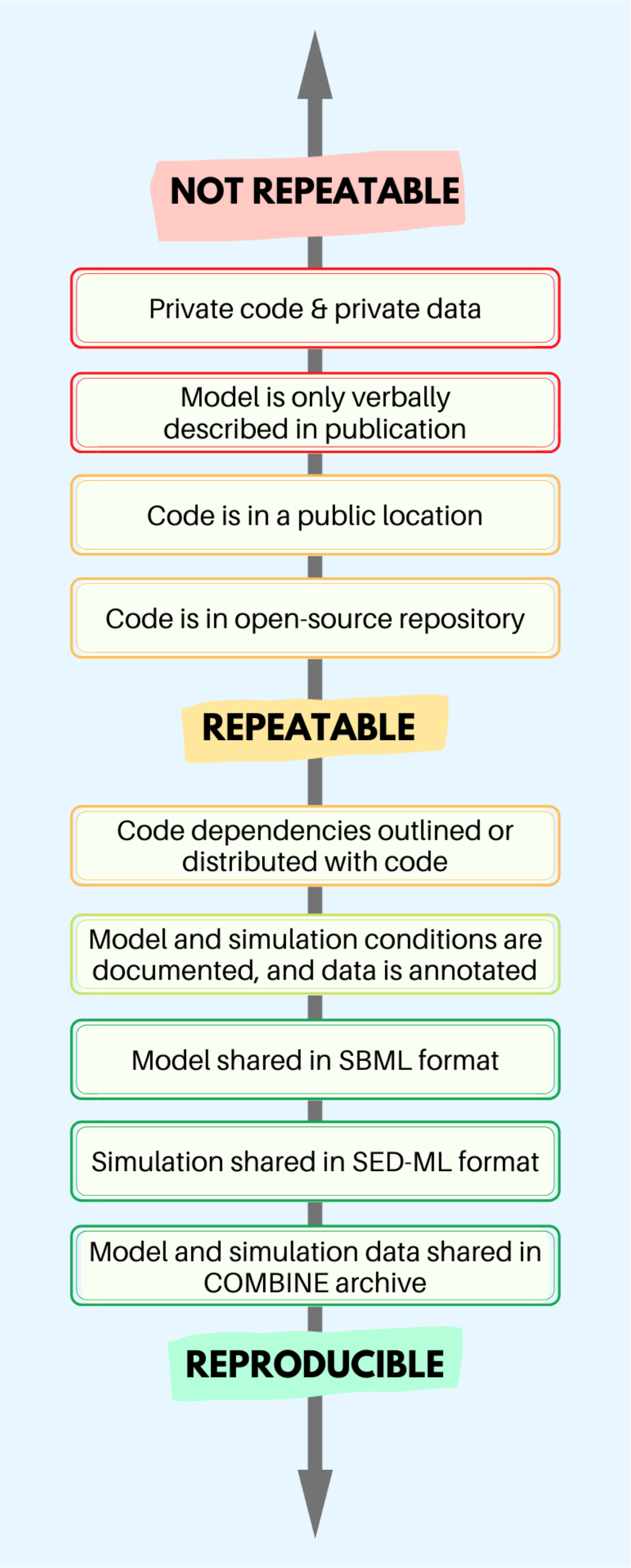
A Year of New Innovations (continued)

reproducibility to gain confidence in a given result. This is what distinguishes science from other efforts to understand the world. The second issue is simply one of economics. Our society expends resources on scientific research, and work that is not reproducible is a waste of those resources. In the process, we are losing intellectual capital while at the same time undermining the scientific method and - perhaps worst of all - we are undermining the public's trust in the scientific establishment.
One way to publish reproducible models is to encourage the use of standardized model and data formats, common naming conventions, and detailed model descriptions and annotations. The CRBM recommends modelers make these resources available and accessible by openly sharing software and data used to construct published models. In future newsletters, we will cover some of these aspects in more detail and describe the critical role that journals have to play.

This newsletter summarizes the progress that the CRBM has achieved in the past year to develop several technologies that support reproducible modeling. It features a spotlight on one of the CRBM developers and shares future plans to encourage reproducible biomodeling at a systemic level through collaboration with academic journals and extensive outreach to the greater modeling community.

The CRBM is generously supported by the NIBIB and NIGMS of the National Institutes of Health under award number P41-EB023912, as well as the National Science Foundation award 1933453.

(Right) Various levels of repeatability and reproducibility on a spectrum. Ideally, all data and simulation information would be shared in a common format and stored in a COMBINE archive.



NOT REPEATABLE

Private code & private data

Model is only verbally described in publication

Code is in a public location

Code is in open-source repository

REPEATABLE

Code dependencies outlined or distributed with code

Model and simulation conditions are documented, and data is annotated

Model shared in SBML format

Simulation shared in SED-ML format

Model and simulation data shared in COMBINE archive

REPRODUCIBLE

Technology progress

The Center for Reproducible Biomedical modeling is developing several projects in the domains of model building, model annotation, and online simulation and visualization. Here we describe some of the progress on tool development for each of these domains in collaboration with the greater modeling community.

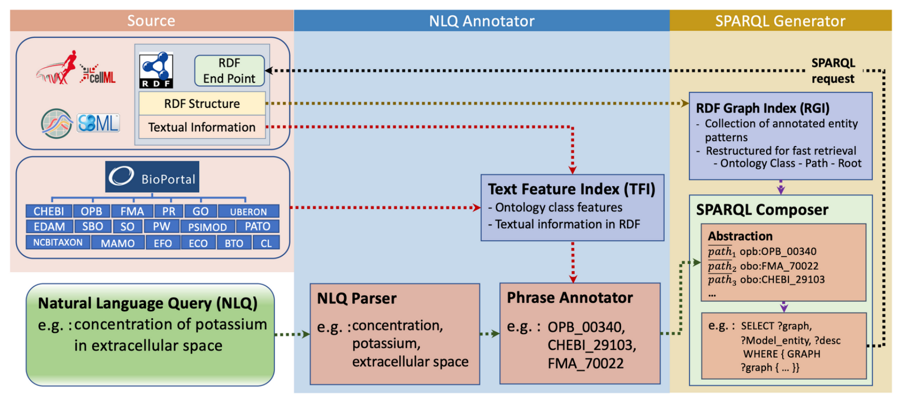
FORMALIZED OMEX METADATA SPECIFICATION

A common language to represent annotations in computational modeling of biological systems has been discussed in the COMBINE community, with the goal of encouraging reuse and reproducibility between researchers, enabling efficient search and retrieval of models and data using standard terms, and supporting semantic comparison across models. A specification was developed to encode annotations in the Open Modeling and EXchange (OMEX) archives. An updated specification (version 1.2) was recently published by Dr. John Gennari and his team, which adds model-level annotations atop the existing smaller scale annotations in the original specification. The new specification uses omex-library.org to provide a common root for all annotations and ultimately ensures that annotations can be distributed with the OMEX archive for a complete distribution of the model and all components that ensure it is explicitly annotated and queryable. See the Github page here.

LIBOMEXMETA C++ AND PYTHON LIBRARIES

With the growing complexity and quantity of biosimulation models, annotation tools which can accommodate a variety of modeling languages and simulation environments were needed to encourage models which could be easily understood and uniquely identified. The COMBINE community developed a consensus on the utilities required for this open-source, cross-platform software library for model annotation. libOmexMeta is a C++ library, developed by Dr. Ciaran Welsh, which has an associated Python front end, pyomexmeta. Together, libOmexMeta and pyomexmeta enable semantic annotation, which connects model elements to standardized vocabulary terms, and ideally encourages users to add biological ontology terms and standardized identifiers to describe the components of the model. libOmexMeta uses the Resource Description Framework (RDF) to represent the annotations in a structured format which can be stored independent from model descriptions. See the Github page here.

NLIMED


Source
NLQ Annotator
SPARQL Generator
cellML
SBML
RDF
RDF End Point
RDF Structure
Textual Information
SPARQL request
BioPortal
CHEBI
OPB
FMA
PR
GO
UBERON
EDAM
SBO
SO
PW
PSIMOD
PATO
NCBITAXON
MAMO
EFO
ECO
BTO
CL
RDF Graph Index (RGI)
- Collection of annotated entity patterns
- Restructured for fast retrieval
- Ontology Class - Path - Root
Text Feature Index (TFI)
- Ontology class features
- Textual information in RDF
SPARQL Composer
Abstraction
path1 opb:OPB_00340
path2 obo:FMA_70022
path3 obo:CHEBI_29103
...
Natural Language Query (NLQ)
e.g. :concentration of potassium in extracellular space
NLQ Parser
e.g. :concentration, potassium, extracellular space
Phrase Annotator
e.g. : OPB_00340, CHEBI_29103, FMA_70022
e.g. : SELECT ?graph, ?Model_entity, ?desc WHERE { GRAPH ?graph { ... }}

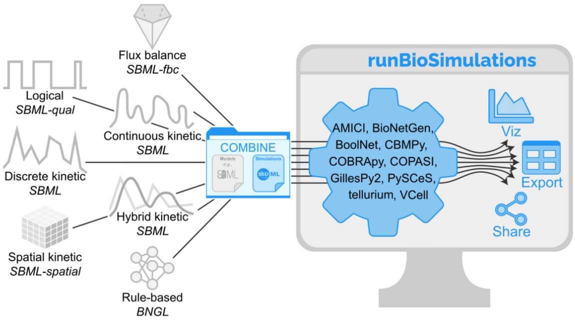
NLIMED workflow shows the conversion of a Natural Language Query to a SPARQL request. Figure from Munarko et al. (bioRxiv, 2021)

The Natural Language Interface for Model Entity Discovery in Biosimulation Model Repositories (NLIMED) was developed by Yuda Munarko to convert natural language queries into the SPARQL syntax which is typically used to search semantic annotations encoded in RDF. The SPARQL syntax is complex and rigid, making it difficult to understand and use for many modeling users interested in searching through RDF annotations. Thus far, NLIMED can be used to query the Physiome Model Repository (PMF) and the BioModels database, and is likely useful for other databases using RDF. See the Github page here.

Technology progress (continued)

BIOSIMULATIONS

BioSimulations, developed by Jonathan Karr, Bilal Shaikh and their team, is an online tool which makes it easy to share and reuse simulations and simulation tools for biological modeling. BioSimulations serves as a gateway for reproducible published work: it provides a central location to publish model simulations and to discover new simulation projects and tools, including resources for data visualization, and simulators that accommodate a variety of model formats and simulation algorithms. Users can use the search function to find papers and run published simulations on the site. The web-based tools enable users to run simulations and visualize and explore results in an interactive format. It even provides tools to modify simulation experiments to generate new simulations. The platform provides support for the COMBINE/OMEX archive format, KiSAO terms, and SED-ML; building off these standardized formats makes the simulation experiments stored on BioSimulations fully transparent. The runBioSimulations platform provides a containerized interface to ensure that authors can preview their simulations in an independent computing environment before publishing them, encouraging reproducible practices. BioSimulations is a community-driven effort, as numerous individuals and organizations have contributed to its development. Currently, over fifty simulators are supported in the BioSimulations tool registry, with ongoing curation efforts. See the Github page here.



runBioSimulations supports sharing and reuse of modes and simulations. Figure from Harmony Hackathon 2021.

ENHANCED SED-ML LIBRARIES AND SPECIFICATION

The Simulation Experiment Description Markup Language (SED-ML) is a standardized, community-based format which explicitly describes simulation processes for biochemical models, ensuring that researchers who are interested in a published model can efficiently reproduce and build upon that model. Its use is independent of specific operating systems or versions of software. The CRBM and COMBINE community actively collaborate to enhance the existing SED-ML libraries and specification. The members of the COMBINE community have worked to develop the most recent version of SED-ML, Level 1 Version 4, which features Kinetic Simulation Algorithm Ontology (KiSAO) integration. KiSAO terms expedite setting up new experiments—users can simply add a new KiSAO term to the SED-ML file. Currently, the CRBM is also developing a common API that can export simulation commands directly to SED-ML. See the Github page here.

SBMATE

SBMate is a Python package principally designed by Woosub Shin to evaluate the quality of annotations in systems biology models. The project proposes a framework which can be used to evaluate the model quality using three metrics - coverage, consistency, and specificity. This framework determines whether annotations for a given model element exist (coverage), whether an existing annotation is appropriate for the element (consistency), and reports on the level of the detail included in the annotation (specificity). SBMate was used to analyze 1,000 models from the BioModels repository, and has also been applied to models in the BiGG database. Common issues encountered in the models evaluated in this study include the absence of annotations, obsolete identifiers, inappropriate identifiers, or insufficient information required to explicitly identify the model element. The creation of SBMate provides an accessible Python package which can be modified and extended to effectively test model annotations. See the Github page here.

Spotlight on Dr. Lucian Smith

Dr. Lucian Smith, a research scientist in Dr. Herbert Sauro's lab, has worked in the systems biology domain for the past fourteen years. He serves as a key member of the COMBINE community, and has contributed to the development and maintenance of standards for years. He has previously served as an SBML and SED-ML editor and continues to work on various tools for systems biology. Here, we share Lucian's path into the field of systems biology, his ongoing work with the CRBM, and his goal for the field of reproducible computational modeling moving forward.

Smith received his Ph.D. in biochemistry and cell biology from Rice University, where he worked on apomyoglobin stability. Specifically, he worked to create a model system which could grow hemoglobin in bacteria to use as a blood substitute. After receiving his doctorate, Smith returned to his home, Seattle. There, he began working as a postdoctoral fellow for Dr. Mary Kuhner in the University of Washington Genome Sciences department. Kuhner first encountered Smith through a 1997 interactive fiction competition. Smith had entered a game that he had programmed himself, and won. Kuhner had played Smith's game, enjoyed it, and written a review about it. Given Smith's programming capabilities and the computational nature of her lab, Kuhner hired him to work on a program called LAMARC, which analyzed phylogenies.

Smith first became involved in systems biology when he was hired by Dr. Herbert Sauro as a staff scientist near the end of his postdoctoral fellowship. Smith's first project in the Sauro lab was Antimony. In 2000, Sauro had written a program called JARNAC, a text and simulation-based program for modeling biological systems. Smith's Antimony language built upon JARNAC by being modular and including a syntax for genetic circuits. It was written in C++, with bindings for a variety of languages including Python.

Dr. Lucian Smith enjoying the outdoors at the Black Hills in South Dakota.

Once the grant funding his work on the Antimony language ended, Smith left the Sauro Lab and started working with Dr. Mike Hucka at Caltech on the SBML team. Hucka assigned Smith with implementing hierarchy within SBML, so Smith read 11 years worth of different proposals for hierarchy and tried to collapse everything he had read into one system. This was his first project developing one of the core standards in computational systems biology. He consulted with the SBML community for his work and asked them questions like "Is this what you would use?" "Do you like this?" "What do you need?" He addressed people's concerns over several rounds of going back and forth between the community and writing up new things, adding new parts that people wanted or changing things that they did not want.

According to Smith, the process of creating new specifications is slow; it takes a long time to build community support around new tools and find someone who has the time to write out everything. Smith notes, "There were a number of short conferences about model standards like the SBML Forum and the SBML Hackathon. These

Spotlight on Dr. Lucian Smith (continued)

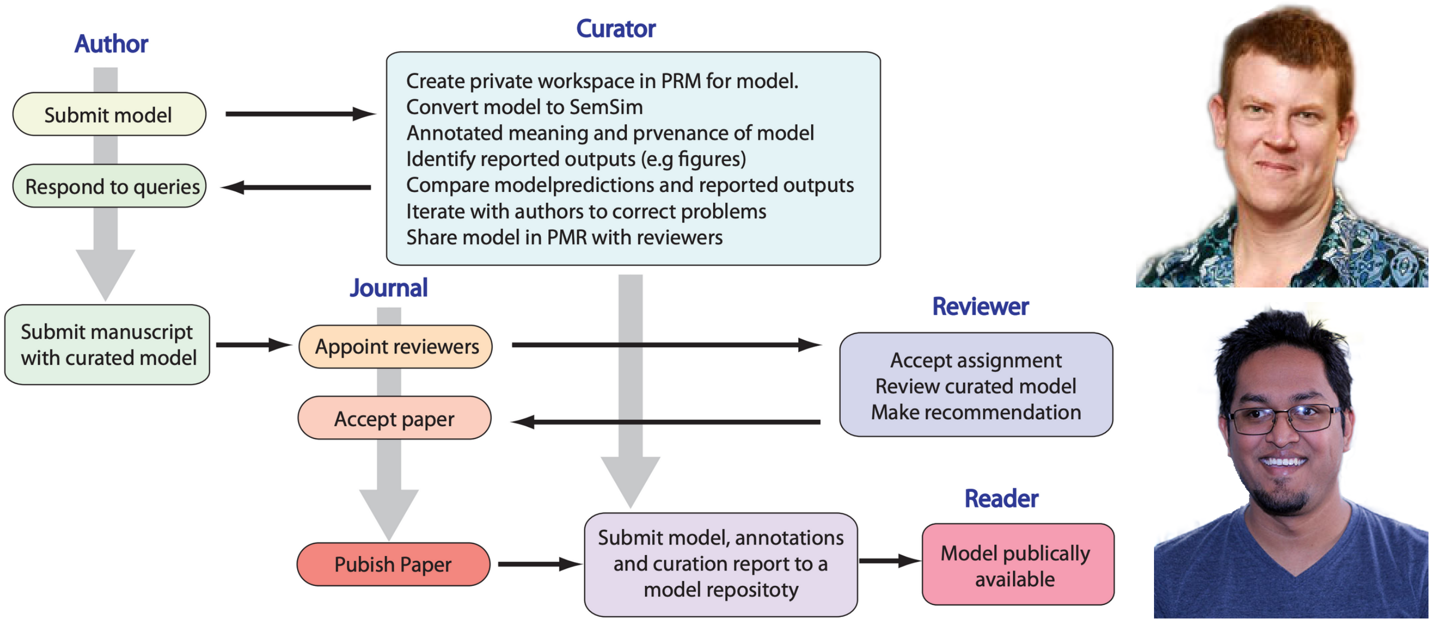
gatherings would attract the same people, and grew larger over time. Eventually, these people decided to form the COMBINE family of standards.” Smith gained significant experience writing specifications through his work on SBML. Though his main focus had been developing standards, after the first grant ended, standards development was pushed to the outer edges of new grants. In the coming years, Smith rotated between working with Sauro, Hucka, Kuhner, and Dr. James Bassingthwaighte on various modeling and software development projects. Since April of 2020, he has been working full time for Dr. Herbert Sauro, focused on improving roadrunner and developing the SED-ML standard and support, a project that had started through spontaneous discussions at a 'Super-hackathon' event in Okinawa. Dagmar Waltemath talked about the Minimum Information About a Simulation Experiment (MIASE), which inspired these discussions, and she headed the group that eventually produced SED-ML Level 1 Version 1. Eventually, the Sauro lab decided to create an Antimony-like text-based representation for SED-ML. Thus, phraSED-ML was created. Dr. Kiri Choi and Smith were the two main developers on the project; Choi did much of the design work and Smith did the programming.

Smith’s ongoing work with the CRBM involves expanding and enhancing the SED-ML specification and associated libraries, with the most recent version being SED-ML Level 1 Version 4. According to Smith, one of the most interesting features of the latest version of SED-ML is the ability to encode most descriptions of things, with references to the KiSAO ontology, which is under active development by Dr. Jonathan Karr. KiSAO and SED-ML were developed in parallel, but by slightly different groups. Smith notes, “Karr’s input has been valuable for the SED-ML project. Karr came with a fresh perspective and asked a lot of questions that had not previously been asked.” This outside perspective helped the SED-ML developers work on how to more clearly express what they wanted. Smith says “Anytime [Karr] had a question, the answer would seem obvious to its developers, so we would have to go to the spec and say ‘what made this unclear?’ and ‘how do we make what's obvious to me obvious to everyone else in writing?’” Smith is proud to have addressed many of Karr’s questions in the updated specification, and appreciates the value added to the specification by Karr’s input.

At present, these groups have merged. Karr’s group is working to add more terms to KISAO which has opened up a lot of possibilities for SED-ML. Smith says that the best part about working on SED-ML is figuring out specific design problems and working on solutions. Though Smith was not able to work on SED-ML as much as he had liked during his elected term as SED-ML editor, he continued to work extensively on SED-ML even after his term ended. Therefore, the most recent specification contains many of Smith’s own words.

In his free time before the pandemic, Smith enjoyed performing in various community plays, and singing in the Total Experience Gospel Choir with his daughter. He hopes those opportunities will arise again soon, but in the meantime has increased his time online with friends, playing various board-games-moved-online, such as Settlers of Catan or Ticket to Ride, as well as telling stories in different TTRPGs. With the conferences also moving online, his tradition of singing parody songs at the end has also moved online. And he still participates in the Interactive Fiction community that got him his first job all those years ago.

Partnership with academic journals to encourage reproducibility


Author
Submit model
Respond to queries
Submit manuscript with curated model
Curator
Create private workspace in PRM for model.
Convert model to SemSim
Annotated meaning and prvenance of model
Identify reported outputs (e.g figures)
Compare modelpredictions and reported outputs
Iterate with authors to correct problems
Share model in PMR with reviewers
Journal
Appoint reviewers
Accept paper
Pubish Paper
Reviewer
Accept assignment
Review curated model
Make recommendation
Submit model, annotations and curation report to a model repositoty
Reader
Model publically available

(Left) Curation service workflow shows how authors, journals, and curators from the CRBM come together during peer-review to check whether the submitted model is reproducible. (Right, from top to bottom) Model curators Drs. David Nickerson and Anand Rampadarath.

A major development to encourage policy change around reproducible modeling in academic publishing occurred through the partnership of the Center for Reproducible Biomedical Modeling and PLoS Computational Biology for a reproducibility pilot launched in July 2019. Authors that submitted articles containing relevant models had the option to select to participate in the reproducibility review. If they selected to participate, the authors received additional technical peer review and comments noting which elements of the model were reproducible, showing the results which could be generated. The model curation service offered by Dr. David Nickerson and Dr. Anand Rampadarath at the Auckland Bioengineering Institute evaluated a list of criteria established to assess the reproducibility of the associated models.

The authors of articles submitted for additional reproducibility peer review could select to have their peer review results submitted alongside the published article, further improving transparency. Not only does this additional check for reproducibility confirm that results can be independently reproduced, it necessitates the sharing of all software and data associated with the published results. Ultimately, the goal of the pilot, and of adding formal peer review for reproducibility of computational models, is to drive a cultural shift that encourages researchers to engage in Findable, Accessible, Interoperable, Reusable (FAIR) modeling practices in computational biology.

This pilot is now under review by the leaders at PLoS Computational Biology and a publication demonstrating the findings of this pilot is expected soon, which will answer many questions to determine whether this is a valuable resource that would be adopted by the greater modeling community. A collection of articles evaluated during this pilot study can be accessed here.

Outreach

Knowledgeable researchers are essential to improving the reproducibility of biomedical modeling. Therefore, the CRBM is actively involved in advocating for reproducible modeling, disseminating information about reproducible modeling, and training researchers to conduct modeling reproducibly through seminars, workshops, and meetings.

SEMINAR SERIES

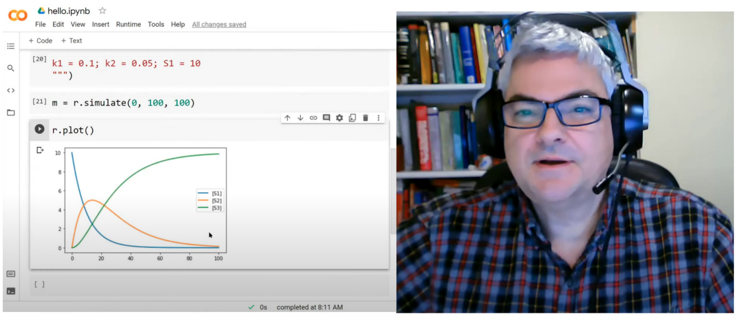
The seminar series this year featured six speakers who shared software tools and theoretical techniques they have used to address complex challenges in systems biology and computational modeling. To begin the series, Dr. Jay Bardhan discussed how electrical engineers address system complexity, and how the same techniques can be applied to biological systems.

The remaining speakers shared the tools they have been developing to support parameter estimation, annotation, and visualization challenges. Dr. Joe Hellerstein and Dr. Daniel Weindl both presented their approaches to parameter estimation: SBstoat and PEtab, respectively. SBstoat is a Python package which can be used for simple parameter estimation problems. PETab is a data format which is rapidly gaining popularity as a standardized input format for parameter estimation.

In the area of model annotation, Dr. David Nickerson shared his ongoing work on the library, libOmexMeta, which is used for annotating biosimulation models stored in the SBML format. He presented an interactive tutorial using pyomexmeta, the Python front-end for the library. Concluding the seminar series for 2021, Dr. Jin Xu and Dr. Matthias König presented tools with visualization capabilities. Xu shared the package, Coyote, an extendable reaction network visualization tool that allows users to add new plugins. König shared his work developing sbmlutils, a Python package to assist with writing, annotating, and documenting SBML models.

NETWORK MODELING SUMMER SCHOOL

In addition to this seminar series, the Network Modeling Summer School hosted by the CRBM was extremely successful at disseminating theoretical and practical knowledge of reproducible modeling to beginners in the field. The Center hosted a five-day virtual summer school, which included a symposium to showcase six invited speakers who shared tools being developed in their lab. More than sixty attendees completed the summer school, attending from around the world and bringing with them a diversity of experience in biological systems and computational approaches to study systems biology. These attendees were able to use the tools shared in the summer school in their own research projects after gaining fundamental knowledge. An archive of the summer school sessions was also published to YouTube to provide lasting educational resources, and these videos have generated ongoing interest from the greater modeling community.



```
[20] k1 = 0.1; k2 = 0.05; S1 = 10
     """)

[21] m = r.simulate(0, 100, 100)

     r.plot()
```

(Above) CRBM Director Dr. Herbert Sauro delivers a lecture for the Virtual Network Modeling Summer School held in July 2021.

Publications

Several publications have been published on the work the CRBM has accomplished, including publications to demonstrate the utility of developed tools and reviews and perspectives on the current challenges of reproducible research in computational systems biology and modeling.

Blinov ML, Gennari JH, Karr JR, Moraru II, Nickerson DP, Sauro HM. Practical resources for enhancing the reproducibility of mechanistic modeling in systems biology. (2021). Current Opinion in Systems Biology. DOI: 10.1016/j.coisb.2021.06.001

Chew YH & Karr JR. Centralizing data to unlock whole-cell models. (2021). Current Opinion in Systems Biology. DOI: 10.1016/j.coisb.2021.06.004

Gennari JH, König M, Misirli G, Neal ML, Nickerson DP, Waltemath D. OMEX metadata specification (version 1.2). (2021). Journal of Integrative Bioinformatics. DOI: 10.1515/jib-2021-0020

Mazein A, Rougny A, Karr JR, Saez-Rodriguez J, Ostaszewski M & Schneider R. Reusability and composability in process description maps: RAS-RAF-MEK-ERK signalling. (2021). Briefings in Bioinformatics. DOI: 10.1093/bib/bbab103.

Munarko Y, Sarwar DM, Rampadarath AK, Atalag K, Gennari JH, Neal ML, Nickerson DP. NLIMED: Natural Language Interface for Model Entity Discovery in Biosimulation Model Repositories. (2021). bioRxiv. DOI: 10.1101/756304

Shahidi N, Pan M, Safaei S, Tran K, Crampin EJ, Nickerson DP. Hierarchical semantic composition of biosimulation models using bond graphs. (2021). PLOS Computational Biology. DOI: 10.1371/journal.pcbi.1008859

Shaikh B, Smith LP, Blinov ML, Sauro HM, Moraru II, Karr, JR. Integrated models, model languages, model repositories, simulation experiments, simulation tools and data visualizations enable facile model reuse with biosimulations. (2022). Biophysical Journal. DOI: 10.1016/j.bpj.2021.11.2118

Shaikh B, Marupilla G, Wilson M, Blinov ML, Moraru II, Karr JR. runBioSimulations: an extensible web application that simulates a wide range of mathematical modeling frameworks, algorithms, and formats. (2021). Nucleic Acids Research. DOI: 10.1093/nar/gkab411

Shin W, Hellerstein JL, Munarko Y, Neal ML, Nickerson DP, Rampadarath AK, Sauro HM, Gennari JH. SBMate: A Framework for Evaluating Quality of Annotations in Systems Biology Models. (2021). bioRxiv. DOI: 10.1101/2021.10.09.463757

Smith L, Bergmann F, Garny A, Helikar T, Karr J, Nickerson D, Sauro H, Waltemath D, König M. The simulation experiment description markup language (SED-ML): language specification for level 1 version 4. (2021). Journal of Integrative Bioinformatics. DOI: 10.1515/jib-2021-0021

Welsh C, Nickerson DP, Rampadarath A, Neal ML, Sauro HM, Gennari JH. libOmexMeta: Enabling semantic annotation of models to support FAIR principles. (2021). Bioinformatics. DOI: 10.1093/bioinformatics/btab445