

## REVIEW

# Key challenges facing data-driven multicellular systems biology

Paul Macklin 

Department of Intelligent Systems Engineering, Indiana University, 700 N Woodlawn Ave, Bloomington, IN 47408, USA

\*Correspondence address. Paul Macklin, Department of Intelligent Systems Engineering, Indiana University, Bloomington, IN, USA.

macklinp@iu.edu  <http://orcid.org/0000-0002-9925-0151>

## Abstract

Increasingly sophisticated experiments, coupled with large-scale computational models, have the potential to systematically test biological hypotheses to drive our understanding of multicellular systems. In this short review, we explore key challenges that must be overcome to achieve robust, repeatable data-driven multicellular systems biology. If these challenges can be solved, we can grow beyond the current state of isolated tools and datasets to a community-driven ecosystem of interoperable data, software utilities, and computational modeling platforms. Progress is within our grasp, but it will take community (and financial) commitment.

**Keywords:** multicellular systems biology; data-driven; challenges; multidisciplinary; open source; open data; data standards; big data; simulations; machine learning

## Background

In the past decade, we have seen tremendous advances in measuring, annotating, analyzing, understanding, and even manipulating the systems biology of single cells. Not only can we perform single-cell multi-omics measurements in high throughput (e.g., [1–3]), but we can manipulate single cells (e.g., by CRISPR systems [4]), and we can track single-cell histories through novel techniques such as DNA barcoding [5].

As these techniques mature, new questions arise: How do single-cell characteristics affect multicellular systems? How do cells communicate and coordinate? How do systems of mixed cell types create specific spatiotemporal and functional patterns in tissues? How do multicellular organisms cope with single-cell mutations and other errors? Conversely, given a set of functional design goals, how do we manipulate single-cell behaviors to achieve our design objectives? Questions like these are at the heart of multicellular systems biology. As we move from understanding to designing multicellular behavior, we arrive at multicellular systems engineering.

High-throughput multiplex experiments are poised to create incredibly high-resolution datasets describing the molecular and behavioral state of many cells in three-dimensional tissue systems (e.g., [6]). Computational modeling—including dynamical simulation models and machine learning approaches—can help make sense of these data.

Modelers “translate” a biologist’s current set of hypotheses into simulation rules, then simulate the system forward in time. They compare these results to experimental data to evaluate the hypotheses, and refine them until simulations match experiments [7, 8]. Computational models allow us to ask “what if” questions [9]. What if we added a new cell type to the mix? What if we spliced in a new signaling pathway? How would our system change?

Machine learning and bioinformatics complement the dynamical modeling approach: analyses of large datasets—especially when annotated with expert-selected biological and clinical features—can be mined to discover new relationships between single-cell states and behaviors, multicellular organization, and emergent function. This, in turn, can drive new hy-

Received: 20 May 2018; Revised: 27 September 2019; Accepted: 30 September 2019,

© The Author(s) 2019. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

potheses in simulation models. Moreover, machine learning can provide novel analyses of simulation data, increasing what we learn from the efforts.

Examples of these approaches appear largely as isolated efforts. Most groups seek out their own data sources (previously published data and tailored experiments), build their own models, and perform their own analyses. Much of this work uses in-house tools created to work on datasets with *ad hoc*, non-interoperable data elements (see Fig. 1). Thus, any one group's work is by and large incompatible with any other group's, hindering or altogether preventing replication studies and modular reuse of valuable data and software.

It does not have to be this way. If we could solve key challenges, we could move beyond single-laboratory efforts to a community built around compatible data and software. Multiple experimental laboratories could pool their efforts to characterize common experimental model systems and record their data in centralized repositories. With a shared “data language,” labs could cooperatively build better simulation, analysis, and visualization tools. Multiple computational labs could build models off of these shared data and tools, find new biological insights, and feed them back into the community (see Fig. 2).

In this review, I explore some key challenges that we need to overcome before we can reach the full potential of an ecosystem of interoperable data and tools for multicellular systems biology.

While the challenges are not presented in any ranked order of importance or priority, they progress from the concrete challenges of standardized data representation and knowledge capture to community resources we could build with standardized data. We do not need to address these challenges sequentially. One of the great strengths of open research communities is that progress can occur by many groups in parallel, each contributing according to their individual skills, resources, and interests.

## Key Challenges

### Shared multicellular data standards

Data arising from high-throughput experiments need to be machine readable and stored in interoperable formats with biologically meaningful data elements. We need to move beyond shared drives of raw images and spreadsheets to extracted biological data elements that are useful for building models and machine learning. We need to store not only averaged cell data but also single-cell states for many cells at multiple time points. Measurements lose meaning without context: data must be stored with metadata including detailed cell line and (molecular) growth media details, biophysical culture conditions, who performed the measurements, what instruments were used, and what software tools were used for analysis.

#### Current progress

Great strides have been made towards this challenge. The Open Microscopy Environment (OME) has emerged as a biological image standard with extensive metadata [10], which has helped to make scientific instruments more interoperable. The ISA-Tab format [11] functions as a rich online file system: provenance and other metadata are bundled with raw data of any file type, allowing the contents to be indexed and searched without detailed knowledge of the data formatting. This has facilitated the creation of large databases of very heterogeneous data (such as GigaDB [12]), and it enables simple data exchange owing to its support for many data types.

While these formats facilitate file-level interoperability, they do not encode extracted biological data elements. Protocols.io was developed to share detailed experimental protocols [13], which can be cited in journal publications to help improve repeatability and reproducibility. However, the protocols are human-readable checklists; they do not use a machine-readable controlled vocabulary of growth factors and other culture conditions.

Ontologies such as the Medical Subject Headings (MeSH) [14, 15] and the Cell Behavior Ontology (CBO) can annotate many biological concepts [16], but they serve as controlled vocabularies rather than standardized data formats.

The Systems Biology Markup Language (SBML) is a well-established standard for single-cell systems biology [17], and efforts such as SBML-Dynamic are working to extend SBML to multicellular models. Domain experts in computational biology, experimental biology, and data science worked together to draft MultiCellDS, a standard for multicellular data [18]. MultiCellDS has a highly extensible representation of single-cell phenotype built from a variety of ontologies such as MeSH and CBO, which can be used to represent highly multiplex data (e.g., [1]) for many cells, along with metadata and micro-environmental context. The European Union-funded MULTIMOT project has been developing a community-driven standard for cell motility measurements (MIACME: Minimum Information about Cell Migration Experiments) [19], with a corresponding software ecosystem [20] that can interface with data in ISA-Tab and OME formats.

#### Future

None of these efforts has completely addressed this challenge. Ultimately, we should combine and extend them into a unified data format. ISA-Tab could bundle image data (using OME) and extracted biological features (e.g., with MultiCellDS and MULTIMOT), while storing experimental protocol details with a controlled vocabulary growing out of Protocols.io [13].

We must ensure that metadata not only annotate experimental protocols but also data extraction protocols: What algorithms were used to extract the biological data elements, and where is the source code permanently archived? Some popular data science software (e.g., Docker and Jupyter notebooks) allow users to export their computational pipelines to facilitate this reproducibility. Last, note that extracted biological data elements cannot *replace* raw data: end users must be free to reproduce (and improve!) the extraction of data elements, which necessitates access to the original data.

### Shared multicellular observational representations

Beyond quantitative measurements like cell division rates, we need a machine-readable encoding of qualitative observations and insights derived from raw biological data: when cells are in condition X, they do Y. When cells of type X and Y interact by contact, they tend to do Z. When cell line X looks like Y in an experiment, the cell culture medium lacked factor Z.

Laboratories and clinics are replete with such examples of hard-won knowledge, but until we can systematically record them, these insights will remain siloed, isolated, and destined to be relearned, lab by lab. If we could consistently record qualitative observations, we could progress from single-cell measurements to multicellular systems understanding, including annotation of critical cell-cell interactions.

Until we can specify “correct” model behavior with machine-readable annotations, our simulation studies will be rate-limited to how quickly humans can view simulations and assess them

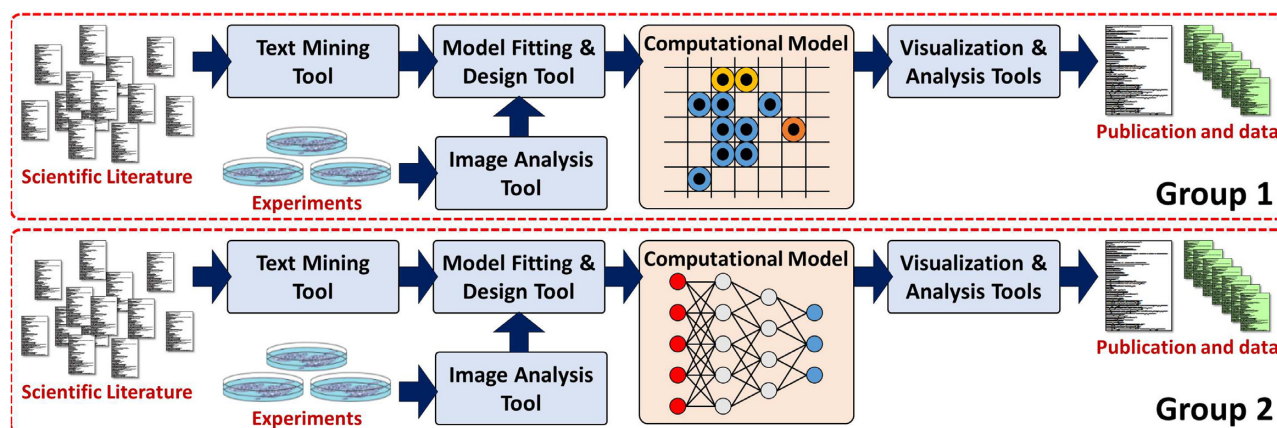


Figure 1: Currently, data-driven workflows are largely parallel, with custom-made, incompatible data and tools.

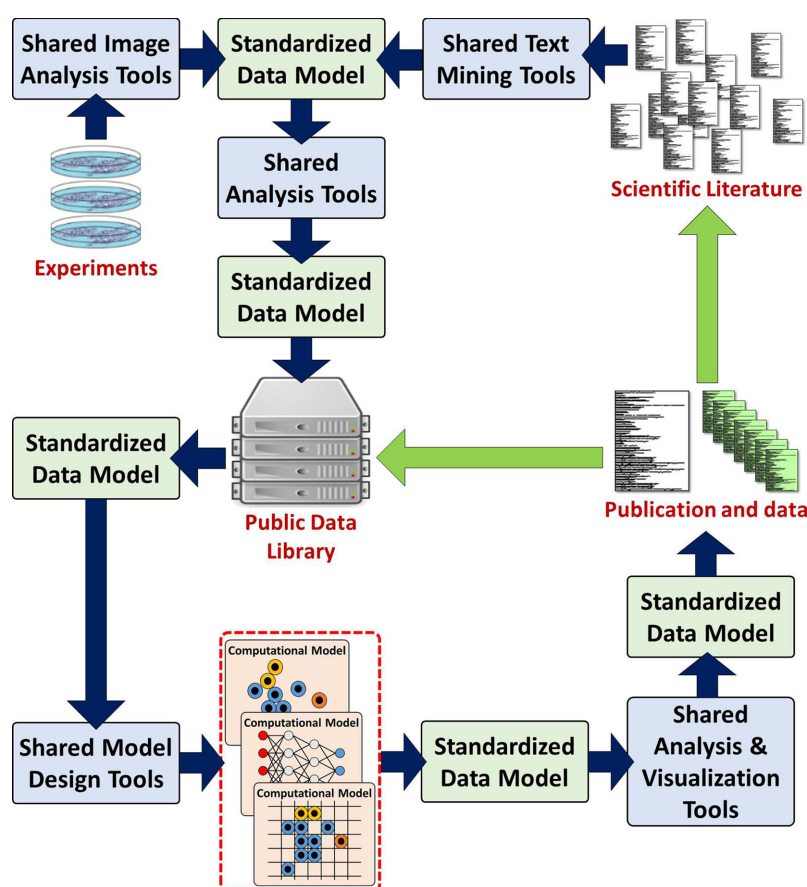


Figure 2: If the community can overcome key challenges, an ecosystem of interoperable computational modeling, analysis, configuration, visualization, and other tools could work on community-curated data and aggregate insights from many sources.

as more or less “realistic.” How do we say, in a generalized way, that a simulated tumor stays compact or becomes invasive? How do we know whether a simulated developmental process has the “right” amount of branching? What does it mean for simulated image  $X$  to “look like” experimental image  $Y$ , given that both the simulation and the experiment are single instances of stochastic processes? If we cannot record the qualitative behavior of simulations and experiments, we cannot automate processes to compare them.

#### Current progress

Progress on this challenge has been limited. The CBO [16] has developed a good starting vocabulary for observed cell behaviors. Extensions of SBML [17] could also potentially represent some of these multicellular and multiscale observations. Tailored image processing has been applied to individual investigations to extract (generally quantitative) representations, although to date we have seen few (if any) qualitative descriptors generated by systematic image analysis.

There has been greater progress in presenting phylogenetic relationships in multicellular populations with automatically extracted phylogenetic trees and other data visualizations, such as Muller plots (e.g., [21–23]). These techniques examine large multi-omics datasets (e.g., small conditional RNA-sequencing data [24]) to fit and represent lineage relationships between cell types (or classes) with directed graph data structures.

#### Future

This area is ripe for machine learning: given a set of qualitative descriptors such as “compact” versus “invasive,” “mixed” versus “separated,” “growing” versus “shrinking” or “steady,” a neural network could be trained to human classifications of experimental and simulation data. High-throughput multicellular simulators (e.g., [7]) could create large sets of training data in standardized formats with clear ground truths. Machine vision could also be used to analyze time series of multicellular data. These annotations could give rise to metrics that help us systematically compare the behavior of one simulation with another, or determine which simulation (in a set of hundreds or thousands of simulations) behaves most like an experiment.

Graph structures could also be applied to represent and visualize cell-cell interactions in multicellular populations [18], similarly to phylogenetic trees (e.g., [22, 23]), chemical reaction networks (e.g., [25, 26]), gene network diagrams (e.g., [27]), and emerging data formats for agent-based model rules (e.g., as in Morpheus [28]).

### Standards support in computational tools

For data standards to be truly useful, they must be broadly supported by a variety of interoperable tools.

#### Current progress

Single-cell systems biology has already shown the enabling role of stable data standards [29]: once SBML crystallized as a stable data language, a rich and growing ecosystem of data-compatible simulation and analysis software emerged. Multicellular systems biology has not yet reached this point: most computational models have custom configuration and output formats, sometimes with customized extensions of SBML to represent single-cell systems biology [18].

#### Future

If a multicellular data standard emerges, key open source projects [29] can implement read-and-write support in their software, either “natively” (i.e., at run-time) or as data converters. Hackathons or similar hosted workshops could facilitate this work. Ontologists need to provide user-friendly data bindings to simplify these development efforts. If standards are to be supported more broadly than just major open source packages, we must remember that most scientific software is created with little formal software engineering training; the data bindings must be well-documented, have simple syntax, and require minimal installation effort.

### Shared tools to configure models and explore data

It is not enough to simply read and write data into individual tools. We must reverse the current “lock-in” effect: because multicellular modeling software is difficult to learn, users (and often entire laboratories) focus their training on a single modeling approach. Because of this, replication studies are rare, even when a study’s source code and data are openly available.

To solve this, we need user-friendly tools to import and set biological and biophysical parameters, design the virtual geometry, and write standardized configuration files that initialize many modeling frameworks. Users could run models in multiple software packages, replicate the work of others, and avoid software-specific artifacts that can bias their conclusions.

Shared software to read, analyze, compare, and visualize outputs from multiple modeling packages could reduce the learning curve for new software. If the shared data exploration and analysis tools were written to work on a common format that includes segmented experimental data, they could also be used to explore experimental data, make and annotate new observations, and motivate new model hypotheses [30].

#### Current progress

Without a common format for multicellular simulation data, there has been little opportunity to develop shared tools for configuring, running, and visualizing multicellular simulations. Some individual simulation packages such as Morpheus [28] and CompuCell3D [31] have user-friendly graphical model editors, but they are currently limited to their individual user communities and not compatible with other simulation packages [29]. Commercially backed open source software such as Kitware’s ParaView [32] is commonly used to visualize multicellular simulation data, but only by writing customized, simulation-tailored data importers. ParaView is generally not used to visualize biological data.

Cloud-hosted tools have provided a means to share sophisticated tools with broad, multidisciplinary audiences without the need for downloading and compiling the tools. For example, the National Cancer Institute (NCI) has introduced NCI cloud resources as part of the NCI Cancer Research Data Commons [33]. Sophisticated simulation models can also be shared as web applications: the PhysiCell development team recently created xml2jupyter [34] to automatically create Jupyter-based GUIs for PhysiCell-based multicellular simulations, which can then be cloud-hosted on platforms like nanoHUB [35].

Other model and data-sharing paradigms that emerged to address related issues in reproducibility may also encourage reuse, such as bundling data and software with Binder [36] or GigaScience’s recent partnership with CodeOcean to pair papers with cloud-hosted executable platforms [37]. However, these typically are single-purpose workflows (specialized to a specific data analysis for a single paper) that are not designed for modular reuse in new research workflows. They tend to lack standardized data formats to facilitate connection with other tools, and latency issues will challenge their use in high-throughput workflows. Moreover, note that while cloud-hosted executable codes increase accessibility and availability, they must not substitute for (or circumvent) sharing source code for full reproducibility.

#### Future

It will be difficult to make progress on this challenge without stable standards for multicellular input and output data. However, progress could be made using current draft standards, such as MultiCellDS [18]. ParaView could use customized plugins to support emerging standards for multicellular data. If projects like Morpheus implemented standards, their graphical model editors could become valuable community resources.

Hackathons can help to rapidly prototype new tools (particularly if they are paired with benchmark datasets), but they must aim to create well-documented, engineered software that is maintained in the long term. We may need new funding paradigms to support small open source teams. The form of



these funding paradigms is not fully clear. Hackathons and similar forms of focused, small-team collaboration could possibly be sponsored through existing federal and philanthropic mechanisms for meetings and travel grants. Crowdsourcing could potentially fund some focused community tool development and maintenance. There is also room for creativity among funding organizations for smaller grants with faster review cycles for community tool-building efforts.

Last, shared code platforms such as the NCI Data Commons could provide an environment to connect data and tools in online, easy-to-use workflows that encourage scientists to “mix and match” data software components into unique research. However, it will be important to avoid “lock-in” effects that prevent moving data and tools from one platform to another. Moreover, as workflows come to incorporate more web services (in differing platforms), they could become vulnerable to technical failures, business failures, or malicious attacks. Open source software has largely solved these issues by mirroring software repositories. Web services may need similar mirroring, and open science norms will need to encourage source code sharing and data/tool portability for web platforms just as they have for offline code.

### High-quality, multiscale benchmarking datasets

Once we have standardized data formats and an ecosystem of compatible software to support them, we need high-quality datasets to drive the development of computational models. The ideal datasets would sufficiently resolve single-cell morphologies and multi-omic states in 3D tissues, along with microenvironmental context (e.g., spatial distribution of oxygen).

To capture the behavioral states of cells, we need standard immunohistochemical panels that capture multiple dimensions of cell phenotype: cycle status, metabolism, death, motility (including markers for the leading edge), adhesiveness, cell mechanics, polarization, and more. We will need to capture these details simultaneously in many cells at multiple time points, using massively multiplexed technologies.

These datasets would be used to formulate model hypotheses and assumptions (through data exploration using standardized tools), to train models, and to evaluate them. Moreover, as the community develops new computational models, they could be evaluated against benchmark datasets. Benchmark datasets are domain specific: separate datasets are needed for developmental biology, avascular and vascular tumor growth, autoimmune diseases, and other problems. It is important that these datasets be easily accessible with open data licenses to promote the broadest use possible. Adhering to FAIR (Findability, Accessibility, Interoperability, and Reusability) data principles would be ideal [38].

#### Current progress

Cancer biology has made perhaps the greatest progress on this challenge, where the NIH-funded Cancer Genome Atlas hosts many genomic, microscopy, and other large datasets [39]. Typically, these consist of many samples at a single time, rather than time course data. Highly multiplex multicellular data are generally not available. DREAM challenges have assembled high-quality datasets to drive model development (through competitions) [40], but these have not typically satisfied the multiplex, time-series ideals outlined above. Private foundations are using cutting-edge microscopy to create high-quality online datasets (e.g., the Allen Cell Explorer Project [41] and the Human BioMolecular Atlas Program [6]).

The technology for highly multiplexed measurements is steadily improving: CyTOF-based immunohistochemistry (e.g., as in [1]) can stain for panels of 30–50 immunomarkers on single slides at 1–2  $\mu\text{m}$  resolution or better. There are no standardized panels to capture the gamut of phenotypic behaviors outlined above. Social media discussions (e.g., [42]) have helped to drive community dialog on difficult phenotypic parameters, but no clear consensus has emerged for a “gold standard” panel of immunostains.

#### Future

Workshops of leading biologists should assemble the “dream panel” of molecular markers. Consortia of technologists will need to reliably implement these multiparameter panels in experimental workflows [1]. Workshops of bioinformaticians, data scientists, and modelers will be needed to “transform” these raw data into standardized datasets for use in models. All this will require federal or philanthropic funding, and contributions by multiple laboratories. Social media has great potential for public brainstorming, disseminating resources, and recruiting new contributors. Hackathons could help drive the “translation” of raw image data into standardized datasets, while developing tools that automate the process.

### Community-curated public data libraries

We need “public data libraries” to store and share high-quality, standardized data [43, 30]. Data should not be static: the community should continually update data to reflect scientific advances, with community curation to ensure data quality. Public libraries must store not only raw image data and extracted biological parameters but also qualitative observations and human insights. The public libraries should host data at multiple stages of publication: preliminary data (which may or may not be permanently archived), datasets under construction (i.e., the experiments are ongoing), data associated with a preprint or a paper in review, and data associated with a published work. Public data libraries should enable if not encourage versioned post-publication refinement, particularly for datasets arising from secondary analysis or curation of heterogeneously sourced primary raw data, such as digital cell lines [18]. Last, public data libraries need to be truly public by using licenses (e.g., Creative Commons CC0 or CC-BY) that encourage new derivative works, as well as aggregation into larger datasets.

#### Current progress

Numerous data portals exist, and more are emerging. Many are purpose-built for specific communities, such as The Cancer Genome Atlas [39]. The Image Data Resource [44] was recently launched to facilitate sharing bioimages using the OME data format [10], further demonstrating how standardized data can facilitate the creation of shared tools and resources. Others like GigaDB [12] and DRYAD [45] allow users to post self-standing datasets with unique DOIs to facilitate data reuse and attribution. These repositories are free for access, thus increasing the reach and impact of hosted data, but the data contributors must pay at the time of data publication. The fees often include editorial and technical assistance while ensuring long-term data availability.

Even within single data hosting repositories, individual datasets are largely disconnected and mutually non-interoperable beyond ISA-Tab compatibility. Thus, individual hosted datasets and studies are generally not bridged and recombined. Moreover, the datasets are usually static after pub-

lication, rather than actively curated and updated. BioNumbers has long served as a searchable resource of user-contributed biological parameters [46], but it lacks a unified data model. The MultiCellDS project proposed “digital cell lines,” which aggregate measurements from many sources for a single cell type [18]. Digital cell lines were intended to be continually updated and curated by the community, so that low-quality measurements could be replaced by better measurements as technology advances. However, this effort is currently manual, with no single, easily searchable repository for its pilot data.

An unfortunate consequence of the current data-hosting model is that all the burden rests on data donors: they generate the data, format it to standards, assemble it, document it, upload it, and then pay the hosting and scientific publication costs. This is a classic case of the “tragedy of the commons”: it is easy to benefit from shared resources, but the cost of contribution falls on contributors. Most repositories have fee waivers for scientists in low-income nations, but small and underfunded laboratories and citizen scientists are still at a disadvantage.

Nonprofit organizations like DRYAD have made great strides in creating sustainable resources to host data; currently (as of 2019), a one-time charge of US \$120 per dataset applies once the data are accepted by curators and publicly available [45]. This is a small fee compared to the data generation cost for experimental labs and within the means of well-funded labs. In cases where secondary analyses or simulations generate new datasets independent of grant funding, there may be greater hardship in these costs, particularly when coupled with open access publication fees.

#### Future

We need to develop more unified, scalable repositories that can bridge fields and collect our knowledge. The repositories should be indexed and community curated to encourage continuous refinement where possible. While there has been great progress to create financially sustainable, permanent data hosting, there is still room to explore alternative funding for data generated independently of specific grant funds. Moreover, these archive-oriented data stores still require curation and indexing if they are to grow from data storage to libraries.

Solutions to this challenge may well originate outside the bioinformatics community. Library scientists have longstanding domain expertise in collecting and curating knowledge across disciplines in unified physical libraries: this expertise would undoubtedly benefit any efforts to create public data libraries. The tremendous success of Wikipedia [47] in hosting its own image and video resources on Wikimedia Commons [48]—at no cost to contributors—could be a very good model. bioRxiv [49] has been similarly successful in hosting preprints at no cost to authors, although experimental data hosting costs are far higher than the cost of hosting manuscripts. Both of these have relied upon a combination of public donations, federal support, and philanthropy, channeled through appropriate nonprofit structures.

We note that public data libraries could become victims of their own success: as public repositories proliferate, finding information will become increasingly difficult, and the community of contributors could become fragmented. This, in turn, will make it difficult to recruit data curators to maintain the quality of the resources. Thus, the community may need to reach consensus on which libraries serve as the standard repositories for which types of data. Moreover, unified search engines and indexes may be needed to help unify knowledge in existing and new data libraries.

Last, to ensure robustness and sustainability, we need to encourage data mirroring with global searchability, and promote a culture that values and properly cites all contributions to shared knowledge: data generation, data analysis, and data curation. While badges can help [50, 51], we must ensure that data users can easily cite all these contributions in papers, that impact metrics reflect the breadth of contributions, and that tenure and other career processes truly value all contributions to community knowledge resources.

#### Quality and curation standards

Community-curated public libraries face new questions: how can we consistently decide which data are worth saving? How do we determine whether a new measurement is better than an old one? How do we monitor quality? Can we automatically trust one laboratory’s data contributions on the basis of prior contributions? And who gets to make these decisions?

#### Current progress

Little to none, aside from uncertainty quantification.

#### Future

This challenge is as much cultural as it is technical. We will need to hold workshops of leading biologists to identify community values and standards for assessing different measurement types. The community will need to determine whether gold standards can be devised for comparing measurements.

#### Linking data to models

We need to connect data to computational models. Data modelers should help design experiments, to determine what variables are needed to build useful models. We need to determine how to “map” biological measurements to model parameters.

#### Current progress

This challenge is currently being addressed on a study-by-study basis. Individual teams design experiments, devise their own model calibration methods, formulate model evaluation metrics, and create their own tools to analyze and compare experimental and simulation data.

#### Future

This challenge is both technical and cultural. Mathematicians, biologists, data scientists, and others will need to work together to determine what it means for an inherently stochastic simulation model to match an experiment. Any progress in creating standardized data elements and annotating multicellular systems behaviors will surely help in creating metrics to compare experimental and computational models. Once standardized biological parameters are extracted to create benchmark datasets, machine learning could help drive more systematic mappings from extracted biological parameters to computational model inputs.

#### Conclusions

The time is ripe for data-driven multicellular systems biology and engineering. Technological advances are making it possible to create high-resolution, highly multiplex multicellular datasets. Computational modeling platforms—including simulation and machine learning approaches—have advanced considerably, and they are increasingly available as open source

[29, 52]. Supercomputing resources are amplifying the power of these computational models [7, 8], while cloud resources are making them accessible to all [34, 35].

If we can solve these key challenges, we will connect big multicellular datasets with computational technologies to accelerate our understanding of biological systems. Steady, incremental progress towards any of the challenges benefits the community as we move towards this broader vision.

Some of the challenges are largely technical, such as creating data standards. Others are more cultural, such as shaping community values for data curation. All of the challenges share a need for community investment: developing and sharing compatible tools and data, hosting data, curating public data libraries, and ultimately funding these worthwhile efforts. Many groups are already contributing pieces of this puzzle, often with little financial support. In the future, we must reduce the individual burden in creating community goods. We may need newer, more rapid funding paradigms to help support and harden new software tools, scaling from small but simple proposals to the current large software grant mechanisms (which tend to have low funding rates). We may need to fund software labs rather than software projects, to encourage rapid response to emerging community needs.

We are on the cusp of accelerated, data-driven biological discovery of how cells work together, how they build things, and how this breaks to cause disease. If you are working towards solving any of these challenges (or if you have new ones to pose!), please consider sharing your advances here.

## Abbreviations

CBO: Cell Behavior Ontology; CC0: Creative Commons public domain license; CC-BY: Creative Commons attribution license; CRISPR: clustered regularly interspaced short palindromic repeats; CyTOF: cytometry by time of flight; DOI: Digital Object Identifier; DREAM: Dialogue for Reverse Engineering Assessments and Methods; FAIR: Findability, Accessibility, Interoperability, and Reusability; GUI: graphical user interface; ISA-Tab: Investigation-Study-Assay tabular format; MeSH: Medical Subject Headings; NCI: National Cancer Institute; NIH: National Institutes of Health; OME: Open Microscopy Environment; PI: principal investigator; SBML: Systems Biology Markup Language.

## Competing Interests

The author declares that he has no competing interests.

## Funding

P.M.'s work to develop computational tools and data standards for multicellular systems biology was funded by the Breast Cancer Research Foundation (PIs Agus, Ewald, Gilkes, and Macklin) and the Jayne Koskines Ted Giovanis Foundation for Health and Policy (PIs Ewald, Gilkes, and Macklin), the National Science Foundation (PI Fox, 1720625), and the National Cancer Institute (PIs Finley, Macklin, and Mumenthaler, U01-CA232137-01; PIs Agus, Atala, and Soker, 1R01CA180149; PI Hillis, 5U54CA143907).

## Author's information

PM has worked for >10 years in computational multicellular systems biology, with a focus on cancer biology and tissue engineering. He has written several open source tools for the field, in-

cluding BioFVM (a multisubstrate diffusion solver for biochemical cell-cell communication) [53], PhysiCell (a 3D agent-based modeling toolkit) [54], and MultiCellDS (a draft multicellular data standard) [18]. He is an Associate Professor of Intelligent Systems Engineering at Indiana University.

## Acknowledgments

P.M. thanks Nicole Nogoy and the staff of GigaScience for the opportunity to write this review, and for editorial support. P.M. thanks the reviewers and preprint readers for valuable comments and feedback.

## References

1. Levenson RM, Borowsky AD, Angelo M. Immunohistochemistry and mass spectrometry for highly multiplexed cellular molecular imaging. *Lab Invest* 2015;95(4):397–405.
2. Kim M, Rai N, Zorraquino V, et al. Multi-omics integration accurately predicts cellular state in unexplored conditions for *Escherichia coli*. *Nat Commun* 2016;7:13090.
3. Norris JL, Farrow MA, Gutierrez DB, et al. Integrated, high-throughput, multiomics platform enables data-driven construction of cellular responses and reveals global drug mechanisms of action. *J Proteome Res* 2017;16(3):1364–75.
4. Adli M. The CRISPR tool kit for genome editing and beyond. *Nat Commun* 2018;9(1):1–13.
5. Blundell JR, Levy SF. Beyond genome sequencing: Lineage tracking with barcodes to study the dynamics of evolution, infection, and cancer. *Genomics* 2014;104(6, Pt A):417–30.
6. (2019) The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. *Nature* 574, 187–192 10.1038/s41586-019-1629-x31597973
7. Ozik J, Collier N, Wozniak J, et al. High-throughput cancer hypothesis testing with an integrated PhysiCell-EMEWS workflow. *BMC Bioinformatics* 2018;19(Suppl 18):483.
8. Ozik J, Collier N, Heiland R, et al. Learning-accelerated discovery of immune-tumour interactions. *Mol Syst Des Eng* 2019;4:747–60.
9. Macklin P. When seeing isn't believing: how math can guide our interpretation of measurements and experiments. *Cell Sys* 2017;5(2):92–4.
10. Goldberg IG, Allan C, Burel JM, et al. The Open Microscopy Environment (OME) Data Model and XML file: open tools for informatics and quantitative analysis in biological imaging. *Genome Biol* 2005;6(5):R47.
11. Rocca-Serra P, Brandizi M, Maguire E, et al. ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. *Bioinformatics* 2010;26(18):2354–6.
12. GigaDB website. <http://gigadb.org>. Accessed 14 October 2019.
13. Teytelman L, Stoliartchouk A, Kindler L, et al. Protocols.io: virtual communities for protocol development and discussion. *PLoS Biol* 2016;14(8):e1002538.
14. Rogers FB. Communications to the editor. *Bull Med Libr Assoc* 1963;51(1):114–6.
15. Lipscomb CE. Medical Subject Headings (MeSH). *Bull Med Libr Assoc* 2000;88(3):265–6.
16. Sluka JP, Shirinifard A, Swat M, et al. The cell behavior ontology: describing the intrinsic biological behaviors of real and model cells seen as active agents. *Bioinformatics* 2014;30(16):2367–74.



17. Hucka M, Finney A, Sauro HM, et al. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* 2003;**19**(4):524–31.
18. Friedman SH, Anderson ARA, Bortz DM, et al. MultiCellDS: a community-developed standard for curating microenvironment-dependent multicellular data. *bioRxiv* 2016:090456. Accessed 14 October 2019.
19. Gonzalez-Beltran AN, Masuzzo P et al. Community standards for open cell migration data. *bioRxiv*. <https://doi.org/10.1101/803064>
20. Masuzzo P, Martens L. An open data ecosystem for cell migration research. *Trends Cell Biol* 2015;**25**(2):55–8.
21. Qiu P, Simonds EF, Bendall SC, et al. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat Biotech* 2011;**29**(10):886–91.
22. Hicks DG, Speed TP, Yassin M, et al. Maps of variability in cell lineage trees. *PLoS Comput Biol* 2019;**15**(2):e1006745.
23. Gatenbee CD, Schenck RO, Bravo R, et al. EvoFreq: visualization of the evolutionary frequencies of sequence and model data. *bioRxiv* 2019, doi:10.1101/743815.
24. Satija R, Farrell JA, Gennert D, et al. Spatial reconstruction of single-cell gene expression. *Nat Biotech* 2015;**33**(5):495–502.
25. Blinov ML, Yang J, Faeder JR, et al. Graph theory for rule-based modeling of biochemical networks. In: Priami C, Ingólfssdóttir A, Mishra B, et al., eds. *Transactions on Computational Systems Biology VII* Berlin. Berlin, Heidelberg: Springer; 2006:89–106.
26. Lambusch F, Waltemath D, Wolkenhauer O, et al. Identifying frequent patterns in biochemical reaction networks: a workflow. *Database (Oxford)* 2018 ;**2018**, doi:10.1093/database/bay051.
27. Zhang M, Li Q, Yu D, et al. GeNeCK: a web server for gene network construction and visualization. *BMC Bioinformatics* 2019;**20**(1):12.
28. Starruß J, de Back W, Brusch L, et al. Morpheus: a user-friendly modeling environment for multiscale and multicellular systems biology. *Bioinformatics* 2014;**30**(9):1331–2.
29. Macklin P, Friedman SH, MultiCell DS Project. Open source tools and standardized data in cancer systems biology. *bioRxiv* 2018, doi:10.1101/244319.
30. Zaritsky A. Sharing and reusing cell image data. *Mol Biol Cell* 2018;**29**(11):1274–80.
31. Swat MH, Thomas GL, Belmonte JM, et al. Multi-scale modeling of tissues using CompuCell3D. *Methods Cell Biol* 2012;**110**:325–66.
32. Kitware. ParaView website. <https://www.paraview.org/>. Accessed 14 October 2019.
33. NCI Cloud Resources. <https://datascience.cancer.gov/data-commons/cloud-resources>. Accessed 26 September 2019.
34. Heiland R, Mishler D, Zhang T, et al. xml2jupyter: mapping parameters between XML and Jupyter widgets. *J Open Source Softw* 2019;**4**(39):1408.
35. Madhavan K, Zentner L, Farnsworth V, et al. nanoHUB.org: cloud-based services for nanoscale modeling, simulation, and education. *Nanotech Rev* 2013;**2**(1):107–17.
36. Toward publishing reproducible computation with Binder. <https://elifesciences.org/labs/a7d53a88/toward-publishing-reproducible-computation-with-binder>. Accessed 26 September 2019.
37. Data Intensive Software Publishing & Sailing The Code Ocean. Q&A With Ruibang Luo. <http://gigasciencejournal.com/blog/data-intensive-software-publishing-sailing-the-code-ocean-qa-with-ruibang-luo/>. Accessed 26 September 2019.
38. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 2016;**3**:160018.
39. The Cancer Genome Atlas. <https://cancergenome.nih.gov/>. Accessed 14 October 2019.
40. DREAM Challenges Project Website. <http://dreamchallenges.org/>. Accessed 14 October 2019.
41. Allen Institute. Allen Cell Explorer Project. <http://www.allenell.org/cell-catalog.html>. Accessed 14 October 2019.
42. @MAG2ART, @MathCancer, @AlexisLomakin, @KidneyMyosin. Twitter discussion on cell motility immunostains. <https://twitter.com/MAG2ART/status/940942422074384384>. Accessed 14 October 2019.
43. Ellenberg J, Swedlow JR, Barlow M, et al. A call for public archives for biological image data. *Nat Methods* 2018;**15**(11):849–54.
44. Williams E, Moore J, Li SW, et al. Image Data Resource: a bioimage data integration and publication platform. *Nat Methods* 2017;**14**:775–81.
45. DRYAD Digital Repository. <https://datadryad.org/>. Accessed 14 October 2019.
46. Milo R, Jorgensen P, Moran U, et al. BioNumbers—the database of key numbers in molecular and cell biology. *Nucleic Acids Res* 2010;**38**(Suppl. 1):D750–3.
47. Wikipedia. <https://wikipedia.org>. Accessed 14 October 2019.
48. Wikimedia Commons. <https://www.wikimedia.org/>. Accessed 14 October 2019.
49. bioRxiv: The Preprint Server for Biology. <https://biorxiv.org>. Accessed 14 October 2019.
50. Kidwell MC, Lazarević LB, Baranski E, et al. Badges to acknowledge open practices: a simple, low-cost, effective method for increasing transparency. *PLoS Biol* 2016;**14**(5), doi:10.1371/journal.pbio.1002456.
51. Rowhani-Farid A, Allen M, Barnett AG. What incentives increase data sharing in health and medical research? A systematic review. *Res Integr Peer Rev* 2017;**2**(1):4.
52. , Macklin P, Frieboes HB, Sparks JL, Springer, 10.1007/978-3-319-42023-3\_12, Systems Biology of Tumor Microenvironment. *Advances in Experimental Medicine and Biology*, chap. 12 et alProgress towards computational 3-D multicellular systems biology. 2016;**936**:225–46, K Rejniak .
53. Ghaffarizadeh A, Friedman SH, Macklin P. BioFVM: an efficient, parallelized diffusive transport solver for 3-D biological simulations. *Bioinformatics* 2016;**32**(8):1256–8.
54. Ghaffarizadeh A, Heiland R, Friedman SH, et al. PhysiCell: an open source physics-based cell simulator for 3-D multicellular systems. *PLoS Comput Biol* 2018;**14**(2):e1005991.