Charting a new frontier of science by integrating mathematical modeling to understand and predict complex biological systems

Katharine White (University of Notre Dame, Department of Chemistry and Biochemistry, kwhite6@nd.edu), **Kira McEntire** (Trinity University, Department of Biology, kmcentir@trinity.edu), **Nicole Buan** (University of Nebraska-Lincoln, Department of Biochemistry, nbuan@unl.edu), **Kingshuk Ghosh** (University of Denver, Dept of Physics and Astronomy, Molecular and Cellular Biophysics, kingshuk.ghosh@du.edu), **Lecia Robinson (**Tuskegee University, Department of Biology, lrobinson@tuskegee.edu), **Elisar Barbar** (Oregon State University, Department of Biochemistry and Biophysics, barbare@oregonstate.edu).

*Authorship determined by random order.

Critical time for biological modeling

Biological systems are staggeringly complex. To untangle this complexity and make predictions about biological systems is a continuous goal of biological research. One approach to achieve these goals is to emphasize the use of quantitative measures of biological processes. Advances in quantitative biology data collection and analysis across scales (molecular, cellular, organismal, ecological) has transformed how we understand, categorize, and predict complex biological systems. Simultaneously, thanks to increased computational power, mathematicians, engineers and physical scientists -- collectively termed theoreticians -- have developed sophisticated models of biological systems at different scales. But there is still a disconnect between the two fields. This surge of quantitative data creates an opportunity to apply, develop, and evaluate mathematical models of biological systems and explore novel methods of analysis. The novel modeling schemes can also offer deeper understanding of principles in biology. *In the context of this paper, we use "models" to refer to mathematical representations of biological systems.*

This data revolution puts scientists in a unique position to leverage information-rich datasets to improve descriptive modeling. Moreover, advances in technology allow inclusion of heterogeneity and variability within these datasets and mathematical models. This inclusion may lead to identifying previously undetermined variables driving or maintaining heterogeneity and diversity. Improved inclusion of variation may even improve biologically meaningful predictions about how systems will respond to perturbations. Although some of these practices are mainstream in specific sub-fields of biology, such practices are not widespread across all fields of biological sciences. With resources dedicated to better integrating biology and mathematical modeling, we envision a transformational improvement in the ability to describe and predict complex biological systems.

We envision a future where an integrated scientific community of biologists and mathematicians work toward common goals, including but not limited to:

i) *Fully harnessing information-rich biological datasets to improve models.* Technological advances have resulted in an unprecedented access to volumes of biological data. Utilizing models that can handle and incorporate multivariable and complex datasets is critical for improving descriptive and predictive models.

ii) *Using model-directed collection of biological data.* Not all biological data is collected in such a way that enables use in model development or validation. Collecting data with a

mind on where and how it will be used in modeling is necessary to better integrate biology and mathematical modeling.

iii) *Determining how to better transfer and validate models developed at one biological system (or scale) to other systems or scales.* Mathematical models that can translate and scale to other biological systems would be transformative in creating common language and nodes of understanding between fields. Identifying models and biological systems to develop in depth as "anchor" models/systems is a critical and complex goal.

iv) *Developing and applying better predictive models to streamline experimental hypothesis testing.* Descriptive models are useful in explaining biological phenomena, revealing control points, and identifying regulators. Turning our focus towards building robust predictive models would enable us to design better biological validation experiments and improve scientific outcomes with limited time, money, and people.

v) *Applying predictive models to generate testable "dataset" space in fields where dataset collection is prohibitive (i.e. costly, low sample size, ethical constraints, latent processes, etc.).* Developing predictive models in representative biological systems would enable us to create virtual datasets. This would reduce the need to collect large datasets for all similar systems and would enhance our ability to predict outcomes of successively more complex biological systems in a rapidly changing world.

Achieving just a subset of these goals would be transformative for our ability to understand and predict complex biological systems. In order to reach these goals, we need to expand the use of existing mathematical models in biology, develop new models, overcome technological challenges, and nurture a cultural shift towards interdisciplinary and cross-field interactions. We will explore each of these approaches below.

Expanding the use of existing models in biology

A range of models can be applied to biological problems. While many of these models are well known in sub-fields, they are often unknown beyond their immediate sphere of application. For example, a biologist's familiarity with the following models: agent based models (ABM), statistical physics models (including all-atom simulations), bioinformatics-based prediction algorithms, geometric and graph based algorithms for network analysis, image processing algorithms, software testing algorithms, and global climate models, is likely to be entirely dependent on their subfield or particular expertise. *As a result, the vast majority of existing modeling tools remain underutilized in biology.* For example, simple statistical physics and agent based models (ABMs) have been proven to be highly valuable to derive insights in widely varying fields: from aquatic plant ecosystems, to forest fires, to bacterial biofilms, to disease propagation. Thus, applying this modeling method to biology at different levels of organization has profoundly improved our understanding of disparate biological processes at varying scales. Better communication across different sub-disciplines will increase awareness and promote wide-spread usage of existing technology. Furthermore, broader dissemination and cross-talk between diverse communities may even generate new modeling platforms by creatively combining existing complementary methods.

Developing new models to better describe and predict complex biology

At present, most modeling schemes fall within two extremes: reductionist and systems-level. Reductionist models are built by incorporating the minimum number of variables to accurately describe or predict biology. While these models produce results that sometimes don't scale to more complex systems, the results (and key variables) are usually readily identifiable. Systems-level models are often combined with machine learning, and can give accurate classifications when trained on correct data, but rarely produce human-interpretable models and make hypothesis generation and testing more difficult. Models represent our understanding of the

world we observe, and hence should be representative of the system in the simplest way possible to gain deeper understanding: “everything should be made as simple as possible but not simpler.”

Existing mathematical models struggle with the challenge of balancing necessary simplification with accurately capturing biological complexity. Issues of applicability also invariably arise when considering how to adapt models to new systems. A fundamental conundrum facing biologists is whether to choose a model from the existing range of models or to develop a new model that better fits or explains the data. This dilemma can obscure the interpretation of biological results and appropriate resolution requires input and feedback from theoreticians, which can be difficult to obtain.

Current models often do not fully harness available data, and have a particularly difficult time dealing with biological variances, or distributions, resulting from biological phenomena. From ecology to single-cell measurements to single-molecule imaging, we observe fluctuations in data that are intrinsic to the system, can hold key information, and may be biologically meaningful. Models that allow us to better decouple different sources of noise, experimental error, and intrinsic variability would be transformative for uncovering biological roles for stochasticity and heterogeneity.

To address these challenges, we need to transform existing bio-math to better incorporate probabilistic tools for model building and borrow tools from data science. The probabilistic framework allows us to explore the importance of intrinsic noise while still allowing us to decouple “signal” from “noise”. When low-dimension models -- derived from physical laws and insights generated from reductionist approach -- are developed and benchmarked with data we can move from descriptive modeling to predictive modeling. Development and dissemination of emerging models that combine physical reasoning and machine learning would be paradigm shifting.

Overcoming technical and cultural barriers

To better integrate mathematical modeling in complex biological systems, we must *overcome technical barriers.* These technical barriers range from data collection to model validation. For data collection, we need to make sure to collect data that can be used in model development or validation. This may require collecting data in multiple ways for the same question. At the same time, we need to ensure the data generated is high quality and determine better ways to evaluate measurement error and meaningful biological variation. Technological advances such as probabilistic models make it possible to grapple with variation in data sets and the appropriate analysis of this variation is critical for understanding biology.

Models often tend to become black-boxes and lose the power to generate insights/principles because of unresolved complexity. A challenge many biologists have faced how to select the best model when competing models fit the same data. Guided generation of new data combined with data science tools can transform model validation. In parallel, we should also advance theoretical models that try to unravel the black box of deep learning. A handful of such efforts are underway that need further studies and support. A deeper understanding of principles behind deep learning may enable us to provide models with more insights not only in biology but well outside biology. The final technical barrier are challenges to large dataset storage and curation. To help with data accumulation and dissemination, we propose the development of databases for storage of experimental results with standardized metadata to make it easily accessible for theoreticians.

Some of these technical barriers are being addressed in specific sub-fields. However, theorists and experimental biologists do not often cross paths, do not speak the same language, and do not always understand each-others' tools. To break these *cultural barriers,* we suggest creating user-friendly and publicly-available modeling systems that are accessible to biologists not well versed in theory. One fine example of this is PhysiCell, which is computationally powerful yet user-friendly for novices (with easy-to-follow tutorials). In another approach, the Department of Energy KBase is an evolving bioinformatics and modeling platform where users can upload or use open-source tools with the help of exemplary narratives. Funds could support training workshops and hack-a-thons to use, disseminate, and develop novel user-friendly applications. In addition, we propose incorporation of funding mechanisms that support and accelerate interdisciplinary research at the interface of biology and physical sciences. These would include dedicated funding opportunities for pilot projects between multiple researchers as well as funds to organize regional conferences that bring together scientists from diverse biological backgrounds with mathematicians, physicists, and computational chemists. These meetings would facilitate crosstalk between experimental researchers and the theoreticians who create computational resources.

The current culture of science encourages researchers who work in discipline-specific silos, often to the detriment of research advances. Mathematical model frameworks can serve as a unifying device connecting the fields of biology and mathematics. The most effective route to *overcome cultural barriers* includes training a new generation of scientists to work at the interdisciplinary interface between mathematics and biology. We propose a threefold systematic approach that incorporates training, curriculum, and outreach.

Training the next generation of scientists to work at the interface between math and biology requires dedicated funding opportunities for pre-doctoral and post-doctoral fellowships to support trainees wanting to work at this interface. Curriculum expansion within math and biology degree programs at the undergraduate and graduate levels will serve to integrate these fields early on in a student's post-secondary training. Creating a database of training modules that include datasets, code, and model tutorials would help faculty and teachers create integrated lessons for undergraduate and graduate education. Finally, outreach at middle and high schools is critical for building up a strong undergraduate cohort interested in the interface of mathematics and biology. The training modules described above could be easily simplified and adapted for use in middle and high schools to show students how math has improved our understanding of biological questions relevant to society (i.e. climate change, human health, etc.).

Transdisciplinary research can be uncomfortable, difficult, and humbling. Critical to overcoming cultural issues is emphasizing the creation of a warm and welcoming environment of like-minded researchers who are motivated to learn collectively from diverse perspectives. Special attention needs to be placed on setting a stage that lowers barriers to building a community of "**learning teachers**" rather than experts. At the same time, creating such an atmosphere would undoubtedly further encourage people from all backgrounds to continue at the interface between biology and math and enhance the creative potential of the field.

Scientific Outcomes

All biological sub-disciplines could benefit tremendously from better integrating theoretical modeling approaches as proposed here. The theoretical modeling approaches developed in physics and chemistry disciplines are powerful in their reductionist simplicity to elucidate

principles and can be highly predictive under defined or constrained conditions. Conversely, the complexity of biological systems necessitates new ideas on how to express higher-order model behavior and how to scale reductionist models to higher levels of complexity. Unifying biology with physics, chemistry and mathematics/statistics through the use of common model methodologies has the potential to revolutionize our fundamental understanding of the rules of life.

Furthermore, accurate and predictive models have the power to transform society by serving as a foundation for technological innovation. If biological models could approach the predictive accuracy of physical models, we would be able to predict and design biology with the ease that we can design a computer. Predictive biological models may eventually forecast the evolutionary trajectory of organisms similar to how we can predict the weather, design organisms to counteract the effects of climate or habitat change, reduce the time to harvest food crops to feed a growing world population, cure disease and prevent the spread of emerging threats, and design biological technologies to generate clean renewable energy. These outcomes cannot be achieved unless we foster transdisciplinary collaboration and train the next generation of scientists to explore a New Science frontier.