CellPress

Cell Systems

Perspective

A forecast for large-scale, predictive biology: Lessons from meteorology

Markus W. Covert,[1,*] Taryn E. Gillies,[1] Takamasa Kudo,[2] and Eran Agmon[1]
[1]Department of Bioengineering, Stanford University, Stanford, CA 94305, USA
[2]Department of Chemical and Systems Biology, Stanford University, Stanford, CA 94305, USA
*Correspondence: mcovert@stanford.edu
https://doi.org/10.1016/j.cels.2021.05.014

SUMMARY

Quantitative systems biology, in which predictive mathematical models are constructed to guide the design of experiments and predict experimental outcomes, is at an exciting transition point, where the foundational scientific principles are becoming established, but the impact is not yet global. The next steps necessary for mathematical modeling to transform biological research and applications, in the same way it has already transformed other fields, is not completely clear. The purpose of this perspective is to forecast possible answers to this question—what needs to happen next—by drawing on the experience gained in another field, specifically meteorology. We review here a number of lessons learned in weather prediction that are directly relevant to biological systems modeling, and that we believe can enable the same kinds of global impact in our field as atmospheric modeling makes today.

INTRODUCTION

Numerical weather prediction is a comprehensive endeavor to integrate observations from around the world and in space, over multiple timescales, into a mathematical model which both holds an initial state of the global system and can produce forecasts of changes in the atmosphere several days in advance (Kerr, 2012). In terms of the benefits, prediction of storms such as Hurricane Sandy 10 days in advance of landfall—with the corresponding evacuation of hundreds of residents, saving both lives and property—could arguably be ranked among the great technical triumphs in human history (Alley et al., 2019). Considering further impacts, from agricultural planning to commerce to renewable energy, contributing to an overall roughly 10-fold return on global investment, the massive impact of quantitative large-scale modeling in this field is clear (Bauer et al., 2015).

In biology the stakes are even higher, especially with the technological game-changing advances in chromosomal modification and their potential applications (Baltimore et al., 2015). In particular, the use of the CRISPR system, for applications such as editing the human germline, or attempting to eradicate infectious disease by targeting massive insect populations with gene drives, demonstrates both incredible promise and the very real possibility of major catastrophes (Piergentili et al., 2021). More than ever before, biologists need to be able to accurately predict the possible outcomes and unintended consequences of their new designs.

In this context, our lab previously identified important areas required for large-scale modeling to realize its full potential—from wet-lab experimental interrogation to accelerated computation, as well as collaboration and community development (Figure 1) (Macklin et al., 2014). Re-examining the list for this perspective piece, we were struck by how similar it was to the areas that have already been required and developed for atmospheric modeling.

Thus, we describe here some of the major insights that have been developed over the past almost two centuries of numerical weather prediction and apply these lessons to the past and current state of quantitative systems biology. We then attempt to forecast the issues and opportunities that may arise and offer recommendations for moving forward in this exciting and rapidly developing field.

EXPERIMENTAL OBSERVATIONS HAVE EXPLODED IN THROUGHPUT AND VARIETY...

From the weather stations that log visibility, wind speed, temperature, barometric pressure, and the like on a regular basis, to the geostationary and low earth orbiting satellites sending images to scientists from outer space, weather forecasting depends on a massive accumulation of heterogeneous data. Making these measurements has required a significant investment in technology development, but they are the underpinnings of all weather models (Blum, 2019).

Similarly, astonishing progress has been made in biological measurement, particularly in sequencing-related measurement, to the point where some assert that extremely large data sets of any kind should now be called "genomical" rather than "astronomical," with data acquisition rates rapidly increasing and estimated to approach one zettabyte per year by 2025 (Stephens et al., 2015). This progress has led to exciting new "atlas" projects, for example, the *Tabula Muris*, a compendium of single-cell transcriptomic data from 20 mouse organs and tissues (Tabula Muris Consortium et al., 2018). Such projects stand to be further expanded by the incorporation of other single-cell data—including, for example, protein expression, lineage

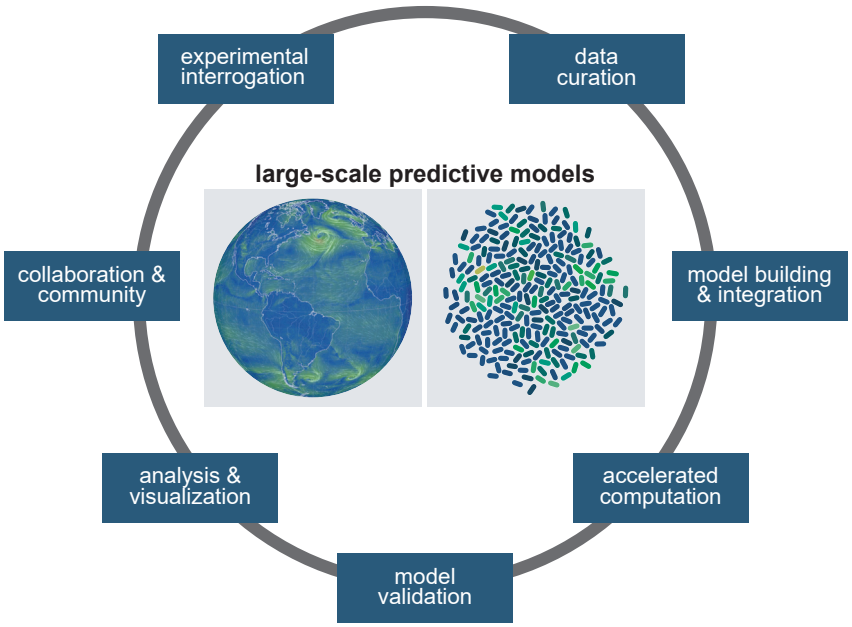

Figure 1. Numerical weather prediction and comprehensive mechanistic modeling of cells are complementary, interdisciplinary challenges

The seven areas connected in a circle were identified in "the future of whole-cell modeling" (Macklin et al., 2014) as necessary areas of expertise for modeling of cellular systems but are also highly relevant to weather forecasting. Two actual visualizations of large-scale simulations are shown in the middle. On the left is a screenshot of Earth Nullschool, an interactive visualization of simulations which predict global weather patterns (surface wind is shown in green) (https://earth.nullschool.net/, 2020). On the right is a visualization of a "whole-cell population" colony simulation, where every cell is an instantiation of our *E. coli* model (Skalnik et al., 2021). The heterogeneity of protein expression for a single gene across the colony is depicted in green.

information, and chromatin state (Stuart and Satija, 2019). New innovations in cellular and molecular imaging add further complementary observations to the mix, from cryoelectron tomography (Turk and Baumeister, 2020) to super-resolution imaging to light-sheet microscopy (Lemon and McDole, 2020; McDole et al., 2018). Live-cell imaging studies bring dynamic resolution to single-cell measurements as cellular behaviors emerge over time (Jeknić et al., 2019), including within more sophisticated cellular structures (Pokrass et al., 2020), and the palette of sensors and perturbation techniques continues to expand (Lee et al., 2020). All of the above is facilitated by new, cutting-edge analysis techniques (Moen et al., 2019; Stuart and Satija, 2019; Wang et al., 2020).

Taken together, these existing studies and technologies have put the biological sciences in a situation parallel to the atmospheric sciences: vast amounts of data, arriving from a variety of disparate sources, modalities, and groups, with increasing amounts generated every day.

...BUT FURTHER PROGRESS ALSO DEPENDS ON NOVEL ANALYTICAL AND MODELING TOOLS

In the 1895 first volume of the journal *Science*, meteorologist Cleveland Abbe called for a new path forward in predicting the weather: "It is not enough to know what the conditions have been and are, but we must know what they will be, and why so… further progress in meteorology demands a laboratory and the consecration of the physicist and the mathematician to this science" (Abbe, 1911).

Advances in measurement technology will also undoubtedly play a major role in future weather forecasting; however, Abbe's assertion remains true today. Successful predictions require not only a critical mass of data but also a mechanistic framework which assimilates and unites them. Importantly, such a framework enables both interpolation (e.g., weather estimates for remote areas which may not have state-of-the-art observation capacities) and extrapolation (prediction of future conditions) (Blum, 2019).

Abbe's assertion can also be applied directly to the biological sciences, which are experiencing an explosion in measurement technology, as described above. The bottleneck in developing a fundamental understanding of biological systems may not be the generation of more data, but instead the assimilation and integration of that data in ways that catalyze novel insight and understanding. This was underscored by a notable study in which a MOS 6507 microprocessor (perhaps best known as the processor in Atari's video game system) was used as a model neurological system (Jonas and Kording, 2017). The scientists attempted to understand three complex "behaviors"—the ability to run three well-known video games (Donkey Kong, Space Invaders, and Pitfall)—by performing experiments that were directly analogous to neuroscientific experiments. For example, microscopy was used to determine the processor's "connectome" (the wiring connections between transistors); comprehensive "lesioning experiments" were performed to infer whether or not a given transistor was involved in a particular behavior; "spike trains" (patterns of "off-to-on" transitions for each transistor) were analyzed during a particular behavior; and the like. Although, unlike in the brain, these studies could be performed with comprehensiveness and perfect precision, and even though the processor is already completely understood, the results of every experiment was found to be insufficient to understand the processor in a meaningful way. In fact, the results were often quite misleading, for example, by suggesting that certain parts of the microprocessor were "Space Invaders transistors" or "Donkey Kong transistors" when in fact none of the transistors had been specifically coded for such a purpose (Jonas and Kording, 2017). The authors concluded that "the problem is not that neuroscientists could not understand a microprocessor, the problem is that they would not understand it given the approaches they are currently taking" and suggested a need for better theory and data analysis, in addition to novel experimental approaches.

All of this is to say that the "consecration of the physicist and the mathematician" will be required in biology as well. In fact

such consecration has a storied history, from the early biochemistry work of Maud Leonora Menten and Leonor Michaelis in characterizing a natural invertase enzyme in 1913 (Gunawardena, 2012a; Michaelis and Menten, 2013); to Salvador Luria and Max Delbrück's landmark study of bacterial mutation rates and how they impacted susceptibility to bacteriophage infection in 1943 (Luria and Delbrück, 1943); to Alan Hodgkin and Andrew Huxley's pioneering work to understand ionic currents and action potentials in the axons of giant squid in 1952 (Hodgkin and Huxley, 1952); and many others. Modern systems biology arose as a field at around the turn of the millennium; one of its stated goals was to build more comprehensive models of cellular systems (Kitano, 2002). Network motifs or larger circuits, whether identified in nature (Brandman et al., 2005; Hoffmann et al., 2002; Milo et al., 2002) or designed *de novo* (Elowitz and Leibler, 2000; Gardner et al., 2000), were characterized and modeled to identify key properties of these systems. Since that time, quantitative modeling of bacterial and eukaryotic cellular systems has become ever more sophisticated and wide-reaching (Braniff and Ingalls, 2018; Hughey et al., 2010).

However, comparison of the modeling approaches used in meteorology and biology also reveals a significant difference between these fields. In the case of meteorology, the models are based on numerical solutions of the following well-established physical equations: Navier-Stokes, mass continuity, the first law of thermodynamics, and the ideal gas law. In contrast, the computational models used to describe cellular processes reflect many levels of abstraction (Ideker and Lauffenburger, 2003), from all-atom molecular dynamics simulations of cellular cytoplasm (Oliveira Bortot et al., 2020) and stochastic simulations (Roberts et al., 2011) on the more detailed side, to Boolean networks on the more abstracted side (Covert et al., 2004). The choice of modeling approach for a particular system is therefore often practical and subjective, balancing highly disparate factors against each other. Computational biologists may need to simultaneously consider the particular goals of a study together with the accessibility or quantitativeness of any relevant data, as well as the runtime required by different approaches, for example. This difference complicates efforts to coordinate research across labs.

STUDYING THE SYSTEM AS A WHOLE NOT ONLY LEADS TO INSIGHTS ON THE GLOBAL SCALE, BUT MAY ALSO BE THE BEST PATH TO UNDERSTANDING INDIVIDUAL COMPONENTS

A major transition in weather prediction occurred with the invention of the telegraph, which enabled news about the weather to travel faster than the weather itself. This enabled scientists (for example, at the Smithsonian Institution) to collect weather observations from many regions simultaneously, and for the first-time weather was correctly understood as a complex, global entity rather than simply a local phenomenon. Current weather models illustrate this point even more clearly, as local observations are incorporated to produce multiregional effects that can impact the entire globe.

Given the high level of complexity for both cellular and atmospheric systems, one might intuit that biological models must also grow more inclusive and holistic over time. This will require a dramatic increase in model scope and scale, as most computational models in biology are much smaller than their weather counterparts (Hughey et al., 2010). However, unfortunately there has been a stigma associated with larger-scale models, in part derived from an anecdote related by Enrico Fermi in which "my friend Johnny von Neumann used to say, with four parameters I can fit an elephant, and with five I can make him wiggle his trunk" (Dyson, 2004). While others have trusted this oft-cited assertion, to the point of actually plotting a cartoon elephant with the allowed parameters (Mayer et al., 2010), Fermi's statement is preposterous in the context of efforts such as weather prediction, in which the model involves "a system of nonlinear differential equations at about half a billion points per time step between the initial time and weeks to months ahead, and accounting for dynamic, thermodynamic, radiative and chemical processes working on scales from hundreds of meters to thousands of kilometers and from seconds to weeks" (Bauer et al., 2015). Similar situations exist in other fields, including microprocessor design and of course, complex biological systems—both of which also require thousands of parameters and more in order to be performed properly.

Accordingly, we and others have been working toward the "grand challenge" of developing whole-cell modeling technologies (Tomita, 2001). This effort spans nearly 40 years, beginning with the first physiology-based models (Domach et al., 1984), to kinetics-based descriptions (Jahan et al., 2016; Khodayari and Maranas, 2016; Tomita et al., 1999) to the development of large-scale models of cellular metabolism (Gu et al., 2019), and the integration of these models with other data or cellular processes (O'Brien et al., 2015; Øyås and Stelling, 2018).

Inspired by the extraordinary work that went before, our lab reported the first model of a bacterium that took all known gene functions explicitly into account (Karr et al., 2012). This model, of the simplest culturable bacterium *Mycoplasma genitalium*, was actually a highly integrated composite model with twenty-eight different submodels, representing the major biological functions in this organism. (There were modules for transcription, translation, replication, and metabolism, for example [Gunawardena, 2012b]). After this model was published, we changed our focus to *Escherichia coli* (Macklin et al., 2020), a much more complex organism with greater importance to science, medicine, and industry, and we continue to work (and invite all to join us) on the *E. coli* whole-cell modeling project (M.W.C., unpublished data).

Hundreds (*M. genitalium*) to tens of thousands (*E. coli*) of parameters have been obtained from the literature and incorporated into these models, enabling simulations of cellular behavior based on molecular and physiological measurements generated by hundreds of labs over many decades. The most exciting application of whole-cell modeling is the ability to predict global, whole-system behaviors, which can be explained mechanistically based on such measurements. Thus, we were encouraged to see that our simulations were able to predict larger-scale physiological behaviors, with insights related to chromosomal occupancy, cell-cycle duration, and cellular energetics in *M. genitalium* (Karr et al., 2012), and a critical role for subgenerational gene expression in *E. coli* (Macklin et al., 2020). These biological insights would have been extremely difficult to identify without a large-scale model, and the experimental validation of

these and other complex model predictions (currently in progress) promises to be a watershed moment for large-scale cellular modeling as a field.

Interestingly, although both large-scale weather and cellular models can rightly be thought of as holistic, they are also fundamentally reductionist, in that they explicitly account for smaller-scale and local phenomena. Such models may therefore be thought of as a mechanistic intertwining of the local and the global, the molecular and the cellular, the genotype and the phenotype—the system and its component parts. This is in contrast to most modeling based on machine learning approaches, in which the factors causing a particular behavior or observation are typically abstracted, making it difficult to determine causation (Ching et al., 2018; Yang et al., 2019).

Thus, it is not only that knowing the components and their interactions enables us to predict global behaviors but also that global observations enable predictions about the component parts. Returning to the impact of the telegraph, "the ability to know the weather in many places at one time was the first step toward knowing the weather in one place at many times" (Blum, 2019). Patterns and dynamics (for example, a "nor'easter") could now be identified, enabling residents of one region to guess at their future weather based on the present or past weather in another region. In other words, a global understanding of the overall system enabled local insights that would not be possible by considering the local system alone—even if studied in great detail.

This point was originally less intuitive to us, and so we were surprised by the capacity of our cellular models to predict behavior at the molecular scale, especially as we subsequently verified these predictions experimentally. In the case of *M. genitalium*, comparing model simulations to growth-rate measurements for a complete library of single-gene disruption strains led to the model's successful prediction of kinetic rate constants (k_{cat}) for three metabolic enzymes, which were characterized after and as a direct result of the simulation results (Sanghvi et al., 2013). For *E. coli*, comparing the model simulations to an independent proteomics dataset strongly suggested that certain of the model's protein-decay rates were incorrect; subsequent experiments verified this prediction as well. Further investigation of the CdsA protein in particular bridged the gap between the global and the local, as its decay rate was shown to be orders of magnitude lower than previously proposed – and this had a direct impact on the model's predicted growth rate (Macklin et al., 2020).

TACKLING THE CHALLENGES POSED BY COMPLEX, NONLINEAR SYSTEMS REQUIRES INNOVATIVE MATHEMATICAL APPROACHES

Both weather systems and biological systems are highly nonlinear, which challenges any efforts at model-based prediction. The recognition of atmospheric nonlinearity—and in particular, how a fully deterministic system could nonetheless produce sudden and surprising behaviors that seemed almost random—led to the development of new mathematical ideas, including chaos theory (Slingo and Palmer, 2011). Chaos theory has proven that nonlinearity can make a system's long-term behavior fundamentally unpredictable, since very small differences in initial conditions can lead to drastic differences in model output. Because such differences are not large enough to be distinguished by measurement, nonlinear models can produce equally plausible outcomes that are nonetheless highly divergent. This realization led to the development of new methods, including for example data assimilation, to better estimate initial states, and ensemble-based approaches, which can calculate the variability in predicted outcomes based on the uncertainty of those states. With regard to data assimilation, weather models run based on initial conditions that are imperfectly known, and physical processes that are imperfectly represented. As a result, errors will accumulate in the model as it runs, decreasing the value of its forecast. Data assimilation encompasses a suite of methods to maintain forecast utility over time by incorporating novel observational data into these models (Bannister, 2017). In ensemble-based approaches to weather prediction, a weather system is simulated many times with slightly varying initial conditions, parameters, and numerical representations. Ensemble-based and probabilistic approaches have enabled more robust estimations, not only of the future weather but also of the uncertainty or reliability of the predictions themselves (Gneiting and Raftery, 2005). Integrating both data assimilation and ensemble modeling into numerical weather prediction has dramatically improved the forecasting quality (Bauer et al., 2015).

While cells and the atmosphere are both multiscale, complex, nonlinear systems, certain aspects specific to biology complicate the challenge of modeling systems-level behavior. One complicating factor, described in detail above, is that our understanding of biological systems spans a wide range of levels of abstraction. Another fundamental challenge is the ubiquity of biological noise, which can manifest itself in the thermal fluctuations within a cell (Elowitz et al., 2002), bacterial persistence in the presence of antibiotics (El Meouche and Dunlop, 2018), or kernels on an ear of corn (McClintock, 1950). Biological systems also exhibit extensive nonlinearity, which suggests the possibility of chaotic behavior (Mackey and Glass, 1977). In this context, however, it is interesting to note that certain biological architectures also appear designed to support robustness, pushing an otherwise noisy system to an almost deterministic response (Barkai and Leibler, 1997; Brandman et al., 2005). Other circuits appear to exploit biological noise in order to produce differentiated phenotypes (Ahrends et al., 2014; Losick and Desplan, 2008) and in some of these cases even appear to preserve a certain amount of variability as a critical component of their behavior (Hughey et al., 2015).

Thus, whereas the weather models are deterministic with homogeneous representations across regions, but highly nonlinear, whole-cell simulations must be heterogeneous and stochastic, in addition to being nonlinear. To respond to these challenges, theoretical and systems biologists have grappled with stochasticity, heterogeneity, and nonlinearity. Using gene expression as an example (Hortsch and Kremling, 2018; McAdams and Arkin, 1997), stochastic simulations have been used to represent a variety of systems, including the repressilator (Elowitz and Leibler, 2000), the bacteriophage lambda lysis/lysogeny decision circuit (Arkin et al., 1998), NF-κB regulation (Tay et al., 2010), the *lac* operon (Roberts et al., 2011) and multiple antibiotic resistance networks (Garcia-Bernardo and Dunlop, 2013) in *E. coli*, and transcriptional bursting (Chong et al., 2014).

Whole-cell modeling also presents particular modeling and analytical challenges which must be addressed in order to understand, predict, and determine proper inputs and outputs (Babtie and Stumpf, 2017; Goldberg et al., 2018). Because whole-cell simulations are based on a single cell, and because some of the functional submodules have probabilistic components (including mRNA transcription, for example), each simulated cell has a different phenotype than the rest. As a result, any key model prediction also depends on an ensemble of simulations. Ensemble-based modeling also provides a framework for whole-cell models to systematically estimate parameter values at scale—for example, via massive combinatorial parameter perturbations and large-scale full factorial experimental designs for our simulations, as we demonstrated in the *E. coli* model (Macklin et al., 2020).

The state of an individual cell is also both highly variable and unspecified, which means that data-assimilation techniques will likely be paramount for improving prediction quality for initializing a whole-cell model. The *E. coli* model initialized the counts and properties of every molecular species using available omics data and a statistical model that, on average, fits experimental data. Although novel experimental measurements have been used to update models and improve predictions, in general, data-assimilation techniques are relatively underdeveloped in cellular modeling and represent an exciting area to learn from numerical weather prediction techniques.

Large-scale prediction is fundamentally an integrative, multidisciplinary effort

The wide range of expertise needed to build comprehensive models of highly complex processes (see Figure 1) requires interdisciplinary and collaborative efforts. Weather observation and numerical weather prediction has fostered such collaboration since the 1950s, with agreements and symposia to pool resources and information across the globe (Benjamin et al., 2019). Governments and international organizations play a massive role in these efforts, for example, by providing investment and funding support to catalyze scientific breakthroughs and business ventures, developing standards to ensure that data can be shared most efficiently and supporting low- or middle-income countries by making these data freely available and accessible (Thorpe and Rogers, 2018). Without such massive and coordinated efforts, numerical weather prediction would be nowhere near where it is today.

Biological modeling requires and deserves a similar effort. No individual lab can possibly hold either all of the areas of expertise in a state-of-the-art capacity simultaneously or the sheer number of people required to maintain such a large effort; thus, the construction, application, and dissemination of large-scale cellular and multicellular models is likely to depend on a large global community, comparable with what has already been achieved in atmospheric modeling and weather prediction. Interestingly, such a community was proposed several decades ago by none other than Sydney Brenner and Sir Francis Crick, who outlined a global, multidisciplinary effort called Project K (presumably referring to the K12 strain of *E. coli*) to obtain "a complete solution of *E. coli*" (Crick, 1973). Project K went so far as to call for a central laboratory to handle shared needs of the community, such as producing and distributing mutant strains, developing techniques and instrumentation to automate experiments, and the like. Unfortunately, no global collaboration or central laboratory was able to take hold, and from our own vantage point, the *E. coli* whole-cell modeling project remains very much incomplete, with immediate needs in all of the areas highlighted in Figure 1. So, although research on *E. coli* continues to progress, one can only speculate where we would be if Crick's advice—or the example of the meteorology community—had been followed.

Other bacterial- and cell-type specific efforts have arisen and must arise in the future. Brenner proposed not only Project K, but follow-on projects L (bacteriophage lambda), M (mouse), and of course N (nematode) (Brennen, 1975), the latter of which is beginning to be realized via the OpenWorm group (Sarma et al., 2018). Others are already working toward large-scale models in humans (Szigeti et al., 2018; Thiele et al., 2020). There have also been cross-cutting initiatives to facilitate collaboration for any organism or project; for example, the challenge posed by heterogeneity of mechanisms within biological systems has inspired the creation of standard modeling languages (Keating et al., 2020) and integrative frameworks that combine different mathematical models into unified representations (Agmon et al., 2021). While these efforts are in their relatively early stages, the weather community's experience with prediction strongly suggests that a larger and coordinated community dedicated to the overall practice of large-scale biological modeling would greatly accelerate any and all more specific modeling efforts.

MOVING FROM PREDICTIONS TO DECISIONS

Numerical weather prediction for the entire globe occurs every day, and forecasts are automatically compared and evaluated against the corresponding weather observations for that day. The metrics for comparison are many and varied, with more tests in active development (Bauer et al., 2015). Such frequent and rigorous comparison drives improvement, and in recent years (and thanks to many of the methods mentioned above), the forecasts have reached such a level of consistent accuracy that they can be relied upon. This transition has catalyzed a shift in focus from predictions to decisions: when the weather models forecast Hurricane Sandy, people began to evacuate.

Medicine and public health are also areas in which urgent, high-stakes decisions must be made on an increasingly regular basis. In some cases, such decisions have indeed involved the use of predictive models. For example, modeling the dynamics of HIV infection, both at the cellular scale (Perelson et al., 1996) and the human population level (Blower et al., 2000), have had a significant impact on the decisions on how to best combat AIDS. Similarly, recent models of the COVID-19 pandemic have impacted country guidelines to minimize the spread and impact of this disease (Edeling et al., 2021; Oyetunde and Obiaderi, 2020). Just as noted with weather forecasting, these models went beyond generating predictions to catalyzing actionable decisions. However, validation is on a much longer timescale and often less straightforward. For example, in the Blower HIV model, which predicted the emergence of resistance to antiretrovirals, new validation data emerged on the timescale of two years (Blower et al., 2003).

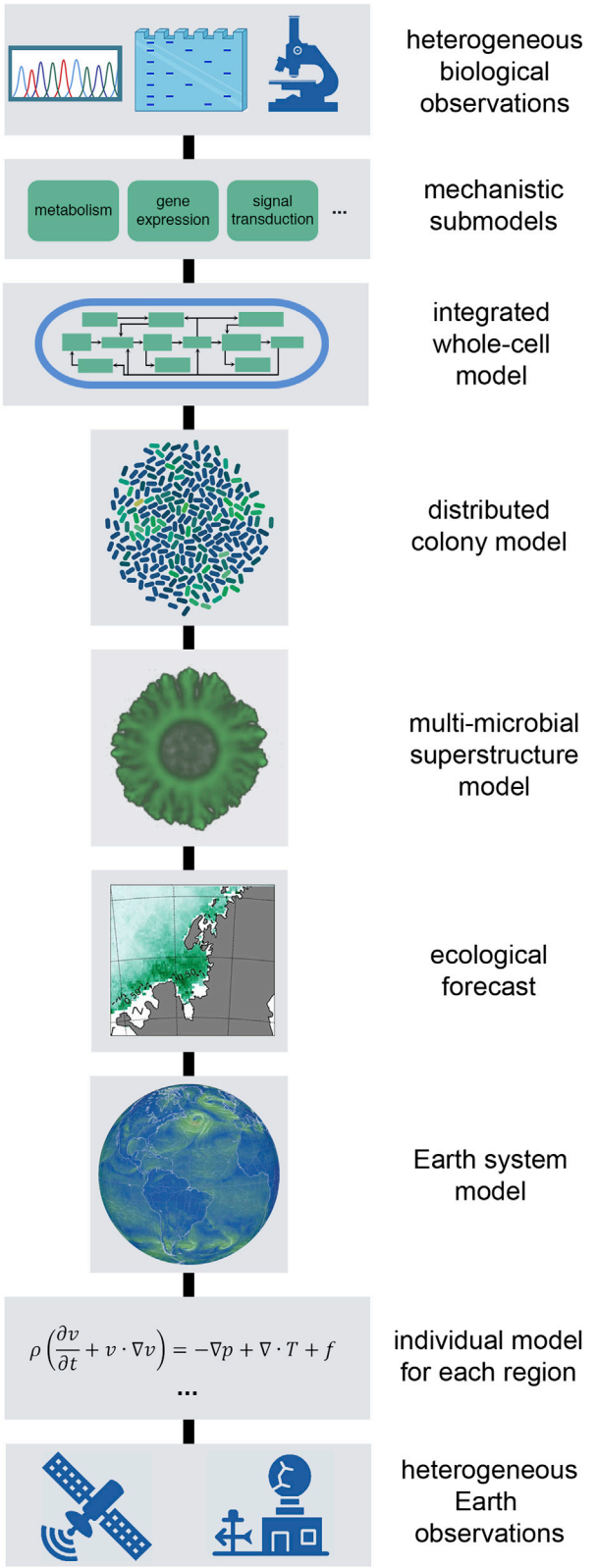

Figure 2. Current potential interactions between models of cellular systems and the Earth system

Whereas atmospheric models (toward the bottom), and biological models (toward the top), have been largely developed in isolation from one another, recent developments present the possibility for them to come together. Biological models can now span from the single-cell level, to the colony level with many heterogeneous cells interacting in a shared environment, and to larger models with multiple microbial species. Ecological forecasts already integrate geological and biological processes—for example, with models of algal blooms. Enabling these different kinds of models to interface with each other could facilitate our understanding of the role oceanic microbes play in shaping our climate. The superstructure and ecological forecast images are adapted from (Dukovski et al., 2020) and (Schultz et al., 2021), respectively.

Models of cellular systems present opportunities in driving decision making that are largely separate from models related to the atmosphere or public health, with the potential to accelerate both biological discovery and design. This is because of the ubiquity of perturbation in biological science research. In principle, any prediction made by a cellular model can be tested at the bench, and the model can be evaluated by direct comparison. Moreover, examination of model "failures"—i.e., discrepancies between model predictions and experimental observations—can often lead to new mechanistic insights about the system under study (Covert et al., 2004; Sanghvi et al., 2013). Large-scale cellular modeling can also be applied to design. We and others have spearheaded efforts intended to facilitate genome design for synthetic biology (Purcell et al., 2013; Rees-Garbutt et al., 2020). Others have shown the power of integrating quantitative biological modeling and analysis for the metabolic engineering of industrially relevant products (Jung et al., 2010), and at multiple steps of a drug development process (Schoeberl et al., 2017). Taken together, these studies crack open the door to a host of future possibilities—and yet they are exceptional. Large-scale cellular modeling has not yet completely crossed the threshold from prediction to decision.

TOWARD THE FUTURE: THE GAIA MODEL?

On a final, more speculative note: Could whole-cell and whole-Earth models eventually engage with each other? Developments over the last several years have made this possibility more plausible. For example, Earth system models are being developed which build on the physical equations used to create weather and climate models, but also include more explicit representations of chemical and biological processes (Bonan and Doney, 2018). Current Earth system models incorporate the biosphere, modeling the effects of insect outbreaks, crop growth, wildfires, and most notably for this perspective—plankton ecosystems, and their interactions with oceanic bacteria. The specific representations of these models are currently too coarse-grained to enable detailed local predictions, but complementary "ecological forecasting" (Dietze et al., 2018) and other mesoscale models are arising, which can predict phenomena such as algal blooms (Schultz et al., 2021), and thereby facilitate the parameterization required for more global models.

On the cellular-modeling side, although efforts to build comprehensive whole-cell models are already providing us with new insights about how and why cells behave the way they do, one critical limitation of such models is their isolation from other cells in the environment. Of course, in the real world—from biofilms to tissues to organs and beyond—cells interact in dynamic shared environments, and these interactions can be as important in driving cellular behavior as are the

molecular interactions within each cell. Several groups have focused on modeling these interactions and the resulting population behaviors (Dukovski et al., 2020; Koch et al., 2019; Norfleet et al., 2020; Segre, 2014; Talman et al., 2019). Our group's most recent work also led to our first "whole-colony" models, where a population of interacting cells in a shared environment was simulated, and each cell was an instance of the most current snapshot of the *E. coli* model (Skalnik et al., 2021).

Given that modeling efforts now exist at these various scales, and even occasionally overlap, one could therefore imagine an exciting future effort which attempts to combine or interface detailed molecular models of ocean bacterial populations together with representations of their context in the ocean and on the planet (Figure 2). Such efforts could lead to an increase in our understanding of the role of microbial metabolism in shaping our climate and also—given that protein evolution speed has now been shown to depend on temperature—how this role and these microbes may change over time as global temperatures rise.

LAST WORDS

In summary, our forecast for the field is that an accumulation of quantitative talent in the biological and medical sciences will precipitate a large-scale predictive capacity—across a spectrum that reaches from individual bacterial cells to human populations and ecological niches—that will in turn lead to a front of "blue-sky" applications. The benefits accrued by developing, supporting, organizing, and deploying this talent for such applications will likely, as we have seen in the atmospheric sciences, far outweigh the costs.

ACKNOWLEDGMENTS

Many thanks to M. Skylar-Scott and S. Doney, as well as our reviewers, including the first meteorologist (!) to ever review a Covert Lab manuscript; our work was greatly improved by their insights and feedback. This work was supported by the Paul G. Allen Frontiers Group through the Allen Discovery Center at Stanford, as well as by an R01 from the National Library of Medicine (LM013229), an F32 from the National Institute of General Medical Sciences (GM137528) to T.E.G., and (GM137464) to E.A.; and the Nakajima Foundation Scholarship to T.K.

REFERENCES

Abbe, C. (1911). The needs of meteorology. Science *34*, 602–603.

Agmon, E., Spangler, R.K., Skalnik, C.J., Poole, W., Peirce, S.M., Morrison, J.H., and Covert, M.W. (2021). Vivarium: an interface and engine for integrative multiscale modeling in computational biology. bioRxiv. https://doi.org/10.1101/2021.04.27.441657.

Ahrends, R., Ota, A., Kovary, K.M., Kudo, T., Park, B.O., and Teruel, M.N. (2014). Controlling low rates of cell differentiation through noise and ultrahigh feedback. Science *344*, 1384–1389.

Alley, R.B., Emanuel, K.A., and Zhang, F. (2019). Advances in weather prediction. Science *363*, 342–344.

Arkin, A., Ross, J., and McAdams, H.H. (1998). Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected Escherichia coli cells. Genetics *149*, 1633–1648.

Babtie, A.C., and Stumpf, M.P.H. (2017). How to deal with parameters for whole-cell modelling. J. R. Soc. Interface *14*, 20170237. https://doi.org/10.1098/rsif.2017.0237.

Baltimore, D., Berg, P., Botchan, M., Carroll, D., Charo, R.A., Church, G., Corn, J.E., Daley, G.Q., Doudna, J.A., Fenner, M., et al. (2015). Biotechnology. A prudent path forward for genomic engineering and germline gene modification. Science *348*, 36–38.

Bannister, R.N. (2017). A review of operational methods of variational and ensemble-variational data assimilation. QJR Meteorol. Soc. *143*, 607–633.

Barkai, N., and Leibler, S. (1997). Robustness in simple biochemical networks. Nature *387*, 913–917.

Bauer, P., Thorpe, A., and Brunet, G. (2015). The quiet revolution of numerical weather prediction. Nature *525*, 47–55.

Benjamin, S.G., Brown, J.M., Brunet, G., Lynch, P., Saito, K., and Schlatter, T.W. (2019). 100 years of progress in forecasting and NWP applications. Meteorol. Monogr. *59*, 13.1–13.67.

Blower, S.M., Aschenbach, A.N., and Kahn, J.O. (2003). Predicting the transmission of drug-resistant HIV: comparing theory with data. Lancet Infect. Dis. *3*, 10–11.

Blower, S.M., Gershengorn, H.B., and Grant, R.M. (2000). A tale of two futures: HIV and antiretroviral therapy in San Francisco. Science *287*, 650–654.

Blum, A. (2019). The Weather Machine: A Journey Inside the Forecast (Harper Collins).

Bonan, G.B., and Doney, S.C. (2018). Climate, ecosystems, and planetary futures: the challenge to predict life in earth system models. Science *359*, eaam8328.

Brandman, O., Ferrell, J.E., Jr., Li, R., and Meyer, T. (2005). Interlinked fast and slow positive feedback loops drive reliable cell decisions. Science *310*, 496–498.

Braniff, N., and Ingalls, B. (2018). New opportunities for optimal design of dynamic experiments in systems and synthetic biology. Curr. Opin. Syst. Biol. *9*, 42–48.

Brenner, S. (1975). Personal notes. *CSHL Archives Repository*, Reference SB/2/3/45. http://libgallery.cshl.edu/items/show/73963.

Ching, T., Himmelstein, D.S., Beaulieu-Jones, B.K., Kalinin, A.A., Do, B.T., Way, G.P., Ferrero, E., Agapow, P.M., Zietz, M., Hoffman, M.M., et al. (2018). Opportunities and obstacles for deep learning in biology and medicine. J. R. Soc. Interface *15*, 20170387.

Chong, S., Chen, C., Ge, H., and Xie, X.S. (2014). Mechanism of transcriptional bursting in bacteria. Cell *158*, 314–326.

Covert, M.W., Knight, E.M., Reed, J.L., Herrgard, M.J., and Palsson, B.O. (2004). Integrating high-throughput and computational data elucidates bacterial networks. Nature *429*, 92–96.

Crick, F.H.C. (1973). Project K: "the complete solution of E. coli." Perspect. Biol. Med. *17*, 67–70.

Dietze, M.C., Fox, A., Beck-Johnson, L.M., Betancourt, J.L., Hooten, M.B., Jarnevich, C.S., Keitt, T.H., Kenney, M.A., Laney, C.M., Larsen, L.G., et al. (2018). Iterative near-term ecological forecasting: needs, opportunities, and challenges. Proc. Natl. Acad. Sci. USA *115*, 1424–1432.

Domach, M.M., Leung, S.K., Cahn, R.E., Cocks, G.G., and Shuler, M.L. (1984). Computer model for glucose-limited growth of a single cell of Escherichia coli B/r-A. Biotechnol. Bioeng. *26*, 1140.

Dukovski, I., Bajić, D., Chacón, J.M., Quintin, M., Vila, J.C.C., Sulheim, S., Pacheco, A.R., Bernstein, D.B., Rieh, W.J., Korolev, K.S., et al. (2020). Computation Of Microbial Ecosystems in Time and Space (COMETS): an open source collaborative platform for modeling ecosystems metabolism. arXiv, arXiv:2009.01734.

Dyson, F. (2004). A meeting with Enrico Fermi. Nature *427*, 297.

Edeling, W., Arabnejad, H., Sinclair, R., Suleimenova, D., Gopalakrishnan, K., Bosak, B., Groen, D., Mahmood, I., Crommelin, D., and Coveney, P.V. (2021). The impact of uncertainty on predictions of the CovidSim epidemiological code. Nat. Comput. Sci. *1*, 128–135.

El Meouche, I., and Dunlop, M.J. (2018). Heterogeneity in efflux pump expression predisposes antibiotic-resistant cells to mutation. Science *362*, 686–690.

Elowitz, M.B., and Leibler, S. (2000). A synthetic oscillatory network of transcriptional regulators. Nature *403*, 335–338.

Elowitz, M.B., Levine, A.J., Siggia, E.D., and Swain, P.S. (2002). Stochastic gene expression in a single cell. Science *297*, 1183–1186.

Garcia-Bernardo, J., and Dunlop, M.J. (2013). Tunable stochastic pulsing in the Escherichia coli multiple antibiotic resistance network from interlinked positive and negative feedback loops. PLoS Comput. Biol. *9*, e1003229.

Gardner, T.S., Cantor, C.R., and Collins, J.J. (2000). Construction of a genetic toggle switch in Escherichia coli. Nature *403*, 339–342.

Gneiting, T., and Raftery, A.E. (2005). Atmospheric science. Weather forecasting with ensemble methods. Science *310*, 248–249.

Goldberg, A.P., Szigeti, B., Chew, Y.H., Sekar, J.A., Roth, Y.D., and Karr, J.R. (2018). Emerging whole-cell modeling principles and methods. Curr. Opin. Biotechnol. *51*, 97–102.

Gu, C., Kim, G.B., Kim, W.J., Kim, H.U., and Lee, S.Y. (2019). Current status and applications of genome-scale metabolic models. Genome Biol *20*, 121.

Gunawardena, J. (2012a). Some lessons about models from Michaelis and Menten. Mol. Biol. Cell *23*, 517–519.

Gunawardena, J. (2012b). Silicon dreams of cells into symbols. Nat. Biotechnol. *30*, 838–840.

Hodgkin, A.L., and Huxley, A.F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. J. Physiol. *117*, 500–544.

Hoffmann, A., Levchenko, A., Scott, M.L., and Baltimore, D. (2002). The IkappaB-NF-kappaB signaling module: temporal control and selective gene activation. Science *298*, 1241–1245.

Hortsch, S.K., and Kremling, A. (2018). Characterization of noise in multistable genetic circuits reveals ways to modulate heterogeneity. PLoS One *13*, e0194779.

Hughey, J.J., Gutschow, M.V., Bajar, B.T., and Covert, M.W. (2015). Single-cell variation leads to population invariance in NF-κB signaling dynamics. Mol. Biol. Cell *26*, 583–590.

Hughey, J.J., Lee, T.K., and Covert, M.W. (2010). Computational modeling of mammalian signaling networks. Wiley Interdiscip. Rev. Syst. Biol. Med. *2*, 194–209.

Ideker, T., and Lauffenburger, D. (2003). Building with a scaffold: emerging strategies for high- to low-level cellular modeling. Trends Biotechnol *21*, 255–262.

Jahan, N., Maeda, K., Matsuoka, Y., Sugimoto, Y., and Kurata, H. (2016). Development of an accurate kinetic model for the central carbon metabolism of Escherichia coli. Microb. Cell Fact. *15*, 112.

Jeknić, S., Kudo, T., and Covert, M.W. (2019). Techniques for studying decoding of single cell dynamics. Front. Immunol. *10*, 755.

Jonas, E., and Kording, K.P. (2017). Could a neuroscientist understand a microprocessor? PLoS Comput. Biol. *13*, e1005268.

Jung, Y.K., Kim, T.Y., Park, S.J., and Lee, S.Y. (2010). Metabolic engineering of Escherichia coli for the production of polylactic acid and its copolymers. Biotechnol. Bioeng. *105*, 161–171.

Karr, J.R., Sanghvi, J.C., Macklin, D.N., Gutschow, M.V., Jacobs, J.M., Bolival, B., Jr., Assad-Garcia, N., Glass, J.I., and Covert, M.W. (2012). A whole-cell computational model predicts phenotype from genotype. Cell *150*, 389–401.

Keating, S.M., Waltemath, D., König, M., Zhang, F., Dräger, A., Chaouiya, C., Bergmann, F.T., Finney, A., Gillespie, C.S., Helikar, T., et al. (2020). SBML Level 3: an extensible format for the exchange and reuse of biological models. Mol. Syst. Biol. *16*, e9110.

Kerr, R.A. (2012). Meteorology. Weather forecasts slowly clearing up. Science *338*, 734–737.

Khodayari, A., and Maranas, C.D. (2016). A genome-scale Escherichia coli kinetic metabolic model k-ecoli457 satisfying flux data for multiple mutant strains. Nat. Commun. *7*, 13806.

Kitano, H. (2002). Systems biology: a brief overview. Science *295*, 1662–1664.

Koch, S., Kohrs, F., Lahmann, P., Bissinger, T., Wendschuh, S., Benndorf, D., Reichl, U., and Klamt, S. (2019). RedCom: a strategy for reduced metabolic modeling of complex microbial communities and its application for analyzing experimental datasets from anaerobic digestion. PLoS Comput. Biol. *15*, e1006759.

Lee, H.N., Mehta, S., and Zhang, J. (2020). Recent advances in the use of genetically encodable optical tools to elicit and monitor signaling events. Curr. Opin. Cell Biol. *63*, 114–124.

Lemon, W.C., and McDole, K. (2020). Live-cell imaging in the era of too many microscopes. Curr. Opin. Cell Biol. *66*, 34–42.

Losick, R., and Desplan, C. (2008). Stochasticity and cell fate. Science *320*, 65–68.

Luria, S.E., and Delbrück, M. (1943). Mutations of bacteria from virus sensitivity to virus resistance. Genetics *28*, 491–511.

Mackey, M.C., and Glass, L. (1977). Oscillation and chaos in physiological control systems. Science *197*, 287–289.

Macklin, D.N., Ahn-Horst, T.A., Choi, H., Ruggero, N.A., Carrera, J., Mason, J.C., Sun, G., Agmon, E., DeFelice, M.M., Maayan, I., et al. (2020). Simultaneous cross-evaluation of heterogeneous E. coli datasets via mechanistic simulation. Science *369*.

Macklin, D.N., Ruggero, N.A., and Covert, M.W. (2014). The future of whole-cell modeling. Curr. Opin. Biotechnol. *28*, 111–115.

Mayer, J., Khairy, K., and Howard, J. (2010). Drawing an elephant with four complex parameters. Am. J. Phys. *78*, 648–649.

McAdams, H.H., and Arkin, A. (1997). Stochastic mechanisms in gene expression. Proc. Natl. Acad. Sci. USA *94*, 814–819.

McClintock, B. (1950). The origin and behavior of mutable loci in maize. Proc. Natl. Acad. Sci. USA *36*, 344–355.

McDole, K., Guignard, L., Amat, F., Berger, A., Malandain, G., Royer, L.A., Turaga, S.C., Branson, K., and Keller, P.J. (2018). In toto imaging and reconstruction of post-implantation mouse development at the single-cell level. Cell *175*, 859–876.e33.

Michaelis, L., and Menten, M.M.L. (2013). The kinetics of invertin action. 1913. FEBS Lett *587*, 2712–2720.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network motifs: simple building blocks of complex networks. Science *298*, 824–827.

Moen, E., Bannon, D., Kudo, T., Graf, W., Covert, M., and Van Valen, D. (2019). Deep learning for cellular image analysis. Nat. Methods *16*, 1233–1246.

Norfleet, D.A., Andre Norfleet, D., Park, E., and Kemp, M.L. (2020). Computational modeling of organoid development. Current Opinion in Biomed. Eng. *13*, 113–118.

O'Brien, E.J., Monk, J.M., and Palsson, B.O. (2015). Using genome-scale models to predict biological capabilities. Cell *161*, 971–987.

Oliveira Bortot, L., Bashardanesh, Z., and van der Spoel, D. (2020). Making soup: preparing and validating models of the bacterial cytoplasm for molecular simulation. J. Chem. Inf. Model. *60*, 322–331.

Øyås, O., and Stelling, J. (2018). Genome-scale metabolic networks in time and space. Curr. Opin. Syst. Biol. *8*, 51–58.

Oyetunde, A.A., and Obiaderi, U.N. (2020). Mathematical modeling analysis for COVID- 19 with contact tracing and quarantine control measures. Int. J. Math. Trends Technol. *68*, 311–315.

Perelson, A.S., Neumann, A.U., Markowitz, M., Leonard, J.M., and Ho, D.D. (1996). HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. Science *271*, 1582–1586.

Piergentili, R., Del Rio, A., Signore, F., Umani Ronchi, F., Marinelli, E., and Zaami, S. (2021). CRISPR-Cas and its wide-ranging applications: From human genome editing to environmental implications, technical limitations, hazards and bioethical issues. Cells *10*, 969.

Pokrass, M.J., Ryan, K.A., Xin, T., Pielstick, B., Timp, W., Greco, V., and Regot, S. (2020). Cell-cycle-dependent ERK signaling dynamics direct fate specification in the mammalian preimplantation embryo. Dev. Cell *55*, 328–340.e5.

Purcell, O., Jain, B., Karr, J.R., Covert, M.W., and Lu, T.K. (2013). Towards a whole-cell modeling approach for synthetic biology. Chaos *23*, 025112.

Rees-Garbutt, J., Chalkley, O., Landon, S., Purcell, O., Marucci, L., and Grierson, C. (2020). Designing minimal genomes using whole-cell models. Nat. Commun. *11*, 836.

Roberts, E., Magis, A., Ortiz, J.O., Baumeister, W., and Luthey-Schulten, Z. (2011). Noise contributions in an inducible genetic switch: a whole-cell simulation study. PLoS Comput Biol *7*, e1002010.

Sanghvi, J.C., Regot, S., Carrasco, S., Karr, J.R., Gutschow, M.V., Bolival, B., and Covert, M.W. (2013). Accelerated discovery via a whole-cell model. Nat. Methods *10*, 1192–1195.

Sarma, G.P., Lee, C.W., Portegys, T., Ghayoomie, V., Jacobs, T., Alicea, B., Cantarelli, M., Currie, M., Gerkin, R.C., Gingell, S., et al. (2018). OpenWorm: overview and recent advances in integrative biological simulation of Caenorhabditis elegans. Philos. Trans. R. Soc. Lond. B Biol. Sci. *373*, 20170382.

Schoeberl, B., Kudla, A., Masson, K., Kalra, A., Curley, M., Finn, G., Pace, E., Harms, B., Kim, J., Kearns, J., et al. (2017). Systems biology driving drug development: from design to the clinical testing of the anti-ErbB3 antibody seribantumab (MM-121). NPJ Syst. Biol. Appl. *3*, 16034.

Schultz, C., Doney, S.C., Hauck, J., Kavanaugh, M.T., and Schofield, O. (2021). Modeling phytoplankton blooms and inorganic carbon responses to sea-ice variability in the West Antarctic peninsula. J. Geophys. Res.: Biogeosci. *126*, e2020JG006227.

Segre, D. (2014). An open source platform for multi-scale spatially distributed simulations of microbial ecosystems. In USDOE Office of Science (SC), Biological and Environmental Research (BER). https://doi.org/10.2172/1228383.

Skalnik, C.J., Agmon, E., Spangler, R.K., Talman, L., Morrison, J.H., Peirce, S.M., and Covert, M.W. (2021). Whole-colony modeling of Escherichia coli. bioRxiv. https://doi.org/10.1101/2021.04.27.441666.

Slingo, J., and Palmer, T. (2011). Uncertainty in weather and climate prediction. Philos. Trans. A Math. Phys. Eng. Sci. *369*, 4751–4767.

Stephens, Z.D., Lee, S.Y., Faghri, F., Campbell, R.H., Zhai, C., Efron, M.J., Iyer, R., Schatz, M.C., Sinha, S., and Robinson, G.E. (2015). Big data: astronomical or Genomical? PLoS Biol *13*, e1002195.

Stuart, T., and Satija, R. (2019). Integrative single-cell analysis. Nat. Rev. Genet. *20*, 257–272.

Szigeti, B., Roth, Y.D., Sekar, J.A.P., Goldberg, A.P., Pochiraju, S.C., and Karr, J.R. (2018). A blueprint for human whole-cell modeling. Curr. Opin. Syst. Biol. *7*, 8–15.

Tabula Muris Consortium; Overall coordination; Logistical coordination; Organ collection and processing; Library preparation and sequencing; Computational data analysis; Cell type annotation; Writing group; Supplemental text writing group; Principal investigators (2018). Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. Nature *562*, 367–372.

Talman, L., Agmon, E., Peirce, S.M., and Covert, M.W. (2019). Multiscale models of infection. Current Opinion in Biomed. Eng. *11*, 102–108.

Tay, S., Hughey, J.J., Lee, T.K., Lipniacki, T., Quake, S.R., and Covert, M.W. (2010). Single-cell NF-kappaB dynamics reveal digital activation and analogue information processing. Nature *466*, 267–271.

Thiele, I., Sahoo, S., Heinken, A., Hertel, J., Heirendt, L., Aurich, M.K., and Fleming, R.M.T. (2020). Personalized whole-body models integrate metabolism, physiology, and the gut microbiome. Mol. Syst. Biol. *16*, e8982.

Thorpe, A., and Rogers, D. (2018). The future of the global weather enterprise: opportunities and risks. Bull. Am. Meteorol. Soc. *99*, 2003–2008.

Tomita, M. (2001). Whole-cell simulation: a grand challenge of the 21st century. Trends Biotechnol *19*, 205–210.

Tomita, M., Hashimoto, K., Takahashi, K., Shimizu, T.S., Matsuzaki, Y., Miyoshi, F., Saito, K., Tanida, S., Yugi, K., Venter, J.C., and Hutchison, C.A., 3rd (1999). E-CELL: software environment for whole-cell simulation. Bioinformatics *15*, 72–84.

Turk, M., and Baumeister, W. (2020). The promise and the challenges of cryo-electron tomography. FEBS Lett *594*, 3243–3261.

Wang, X., Rai, N., Merchel Piovesan Pereira, B., Eetemadi, A., and Tagkopoulos, I. (2020). Accelerated knowledge discovery from omics data by optimal experimental design. Nat. Commun. *11*, 5026.

Yang, J.H., Wright, S.N., Hamblin, M., McCloskey, D., Alcantar, M.A., Schrübbers, L., Lopatkin, A.J., Satish, S., Nili, A., Palsson, B.O., et al. (2019). A white-box machine learning approach for revealing antibiotic mechanisms of action. Cell *177*, 1649–1661.e9.