Accelerating Community-Wide Synthesis in Molecular and Cellular Biology: Recommendations for an NSF-MCB Synthesis Center

Edward O'Brien
Penn State University
University Park, PA

Dave Thirumalai
University of Texas Austin
Austin, TX

Liana Lareau
University of California Berkeley
Berkeley, CA

Howard Salis
Penn State University
University Park, PA

Premal Shah
Rutgers University
Piscataway, NJ

Susan Marqusee
University of California Berkeley
Berkeley, CA

Table of Contents

1. Background and Introduction

1.1 Molecular and cellular biology have seen an explosive growth of data

The fields of molecular and cellular biology have seen rapid growth of publicly available data over the past three decades. Three experimental methodologies have driven this growth. Next-generation sequencing (NGS), in which short DNA fragments are sequenced in a high-throughput manner. Mass Spectrometry (MS) through which the composition and conformational dispositions of proteins and metabolites can be ascertained. And X-ray crystallography that allows the three-dimensional structure of biomolecules to be determined. There are 5 million samples available from NGS from 500 different species in the GEO database. There are some five billion protein mass spectra available from the ProteomeXchange consortium databases and two billion metabolite spectra from GNPS and MetaboLights databases. And 140,000 biomolecular structures available in the Protein Data Bank (PDB), the vast majority of which were determined by X-ray crystallography. Imaging and Cryo-EM techniques, which are seeing greater use, will fuel even faster data growth over the coming decade.

Several factors led to the explosion of diverse, publicly available MCB data. First, while it was common for X-ray crystal structures to be made publicly available since the 1990's, it wasn't until the 2010's that there was the widespread adoption of rules by journals and funding body's requiring NGS and MS data be deposited in centralized repositories. NGS and MS data are now regularly deposited in the GEO[1,2] and ProteomeXchange[3,4] databases upon publication. Second, experimental techniques to generate such datasets have become common in biochemistry, molecular biology, and synthetic biology labs due to falling instrument costs and support from university core facilities, making the creation of large datasets routine[5]. Third, through clever sample preparation, NGS and MS assays have been adapted to measure a wide variety of cellular components and their interactions[6], different aspects of subcellular processes[7], and the relative timing of these processes[8] in a high-throughput manner. These assays have diversified into hundreds of variants, each providing a snapshot of subcellular components, interactions, processes, and rates in some cases.

The mosaic of molecular information now available is astonishing. Most aspects of the central dogma have been characterized by these assays. The location of each and every nucleosome along the *S. cerevisiae* genome[9] has been measured from ChIP-seq[10] (a nucleosome needs to move for a transcription factor to bind). The location of each and every transcription factor bound to a regulatory element along the genome[11] is reported via ChIP-seq (a transcription factor must bind to control initiation by RNA polymerase). The location of each and every RNA-polymerase II molecule along the genome[12] is determined by NET-Seq[13] (RNA-polymerase synthesizes mRNA). For each transcribed mRNA, the fraction that is unspliced and immature (*i.e.*, still containing introns) is measured by RNA-seq[14]. The location of each and every 'scanning' ribosome, 'initiating' ribosome, and 'elongating' ribosome, respectively, during translation is characterized by RCP-Seq[15], TI-Seq[16], and Ribo-seq[17], respectively. The location of ribosomes that have collided with one another during translation is obtained from Disome-Seq[18] (ribosome collisions promote mRNA degradation). The point during protein synthesis at which different protein biogenesis factors engage with the ribosome nascent chain complex – such as co-translationally acting chaperones and targeting factors – is measured via Sel-Ribo-seq[19]. And finally, the half-lives of all the mRNAs and proteins are available from RNA-Seq and MS and pulse-chase labeling[20].

NGS and MS assays can map the influence of differential molecular interactions between components taking part in subcellular processes, characterize connections between genotype and phenotype, as well as measure the rates of subcellular processes when coupled with high-throughput mutagenesis or pulse-chase strategies. For example, such assays have revealed the influences that the 5' UTR sequence features and translation initiation regions have on translation efficiency[21,22]. In the case of transcription this approach was able to dissect the interactions most relevant to transcription factor binding[23]. When cellular phenotype is simultaneously measured, deep scanning mutagenesis has identified new mechanisms connecting genotype to phenotype[24-26]. While pulsing a cell with a small-molecule that inhibits translation initiation, followed by a chase of NGS measurements at subsequent time points has allowed the rate of protein synthesis to be measured across highly expressed transcripts[27]. Thus, these datasets provide multiple levels of information at high resolution, in many cases at the level of individual nucleotides (in the case of NGS), individual amino acids (in the case of MS), or macromolecular interactions and cellular location (in the case of 'spatial' MS).

1.2 Machine learning is accelerating discovery in molecular and cellular biology

This growth of data has allowed computational, data-centric approaches, such as machine learning, to solve some long-standing problems in molecular and cellular biology[28]. A prime example is AlphaFold2[29]. In this methodology, a novel machine learning approach that when combined with large protein sequence data sets and structures from the PDB makes accurate predictions of protein three-dimensional structures that rival the resolution of X-ray crystallographic structures[29]. Thus, AlphaFold2 has solved a grand challenge in the field of structural biology by predicting structure from sequence. Other examples include the automated identification of peptides and proteins in MS[30] - doubling the number of annotated results in the ProteomeXchange, and the automated identification of cell types from high-throughput imaging.

The continuing advances in machine learning hold the promise of accelerating insight and discovery across the fields of molecular and cellular biology through a combination of AI-human collaboration. Within this context there are challenges that must be addressed to maximally exploit these advances, including: (*i*) inferring cause from abstract machine learning knowledge representations; (*ii*) competency or fluency in ML concepts by MCB scientists; (*iii*) the transferability of ML models between different datasets and systems; and (*iv*) maximizing the ease with which ML-human collaboration can occur.

1.3 Theory and modeling's essential role in advancing molecular and cellular biology

The promise of accelerating knowledge synthesis in MCB by machine learning is most welcome. However, to substantially enhance our understanding of major unsolved problems in MCB the results from machine learning have to be integrated with theory and biophysical models. The NSF MCB's division codifies this in its mission statement to "support quantitative, mechanistic, predictive, and theory-driven fundamental research designed to promote understanding of complex living systems at the molecular, subcellular, and cellular levels". Why this is valued comes from the physical sciences, where scientific understanding is best demonstrated if simple equations or 'toy' models can be created that capture the essential physics and chemistry underlying a phenomenon in complex biological systems. In this framework, explanatory power of a model is valued as much as its predictive power. The utility of theory and modeling in advancing the goals of MCB spans the field's lifetime. In the 1940's, mathematical reasoning allowed biophysicists to demonstrate that mutations occur spontaneously, and not as a response to selection pressures[31]. A theoretical physicist modeled and solved the structure of DNA[32] and viral capsids[33] in the 1950's. In the 1960's Darwin's theory of evolution by natural selection was codified mathematically, and the fields of theoretical ecology and mathematical epidemiology started[34,35]. Since then, theorists have advanced our understanding of how DNA, RNA, and proteins assemble and carry out their function across multiple spatial and time scales[36], allowing us to understand the origins of various diseases[37], and how information is relayed within and between cells[38]. This demonstrated value of theory and modeling in synthesizing diverse data to gain new insights means that theory and modeling will continue to play a central role in advancing discovery in MCB.

This report summarizes the outcome of a workshop that brought together thirty scientists with diverse expertise including theoreticians, computationalists, and experimentalists. Participating were machine learning experts, biophysicists, molecular biologists, bioinformaticians, synthetic biologists, physicists, biophysical chemists, bioengineers, and genomic experts. This group was asked to identify grand challenges that are primed to be addressed through coordinated community-wide synthesis efforts, and how a future Synthesis Center could be organized to support the community in their synthesis efforts.

2. Classification Scheme for Types of Synthesis Questions

Community-wide synthesis problems can be broken into three classes. **Class 1 problems** can be addressed through readily available data and established computational or theoretical tools, and constitute the 'low-hanging fruit' that are likely to have substantial impact with the least amount of effort. At the community level, Class 1 problems require scientific domain experts and computational experts to come together to collaborate in organizing, analyzing, and synthesizing the data, but the up-front cost of data organization and analyses are low. **Class 2 problems** are challenges that either have (*i*) data readily available for analysis, or (*ii*) computational tools readily available to apply, but not both. In Class 2 problems the data may be a challenge, for example, because it is dispersed across different databases or published datasets, or has inconsistent formatting across sources. Even if the data is readily available, in Class 2 problems the computational analysis may be a challenge because existing computer codes or theoretical models need to be tailored to the problem at hand. In Class 2 problems, domain experts and computational experts will need to collaborate more extensively than in Class 1 problems to either gather or standardize the data, or handle modifications to analysis methodologies. In **Class 3 problems**, relevant data is not readily available and new computational and theoretical models must be created from scratch. "Not readily available" does not mean the data does not exist, but substantial effort must be invested to gather the data into a useful format for analysis. In addition, in Class 3 problems new computational methodologies and theoretical models must be implemented or created from scratch. Class 3 problems therefore require extensive collaboration between scientific domain and computational experts with progress timelines that may be difficult to estimate.

3. Scientific Challenges that can be Accelerated by Community-wide Synthesis

3.1 Understanding gene regulation through biophysical machine learning

Machine learning is becoming central to computational biology, creating an opportunity for a new synthesis between machine learning and biophysically-motivated modeling approaches. These methods are not in conflict; rather, they have huge potential as complementary approaches, each informing the other. A synthesis center should bring together knowledge from the biophysics modeling community and the machine learning community to enable major scientific goals such as a complete model of genome function and how sequences lead to function. The outcome of such synthesis will enable solutions to outstanding problems in synthetic biology, interpretation of genetic variation, and cellular programming. It will lead to interpretable, causal models that can generate hypotheses, and incorporate the dynamic understanding of time-dependent processes that one gets from biophysical modeling but are often restricted to a few classes of problems.

In the broadest sense, one challenge the proposed synthesis can address is, given the input information about a system, how can it be converted into the most precise and complete model of the system, without restricting this to any single modeling approach. A barrier to this integration is the lack of physical insight from machine learning models, which leads to reluctance to use ML methods in biophysical modeling. Improvements such as "visible" machine learning methods and models of uncertainty can decrease the barrier to communication between the two communities. In addition, training across disciplines can improve the compatibility of approaches and break through culture clashes.

To merge the cultures of disparate fields, a synthesis center must bring together domain experts and encourage the theoreticians to engage with machine-learning experts. It must create structures around modeling methods and build understanding of the methods used to generate biological datasets so that computational researchers can work with data from many different experiments. This major effort in building collaborations would be enabled by a synthesis center that brings people together physically into one space and builds community around shared goals and complementary methods. We envision a synthesis center with a focus on bringing people together through workshops and semester-long programs that facilitate scientists at all levels from different communities to engage deeply with how others address the same questions from entirely different perspectives. This could extend to pairs or trios of postdocs who work as teams at the synthesis center addressing specific challenges.

Work at a synthesis center would attempt to address big questions in combining modeling approaches. What is the right model for each type of question? What are the newest best ideas? There are few or no textbooks on these topics; a synthesis center could compile a constantly evolving, updated set of best practices and best understanding. Creating, curating, and maintaining this as a central resource is an academic endeavor in its own right that could be incentivized with sabbaticals for senior faculty who want to contribute to a 'textbook' or integrated into the educational component of NSF CAREER proposals from affiliated junior faculty.

A synthesis center could provide concrete resources as well as interaction. Genomics data are plentiful but it is essential to winnow through these to identify the best datasets. Cloud infrastructure can enable vast computational projects, but requires funding and infrastructure to interface with it. For instance, data sets that are currently downloaded by each research group could be stored in a shared, public location in the cloud, accessible without further data transfer costs. Access to GPUs, such as through Google's Colab project, would let a wider range of groups reinterpret and model these data. Further, workshops could address challenges such as data sharing standards.

Training is integral to the success of these goals, including both training in advanced methods and earlier training to ensure a diverse and qualified scientific workforce. Improving early training in computation is inherently equalizing, lowering barriers to those coming into research with existing computational skills or confidence. Supporting early university training in computational methods for people from different stem fields can improve the balance in who participates at later stages. The synthesis center could develop curriculum suggestions and training modules for undergraduate biology students to move into computation, as well as focused summer courses in machine learning for computational biology aimed at graduate and postdoctoral researchers. It will be critical to avoid self-selecting who participates in synthesis center activities, by being careful to recruit broadly, and by providing resources like childcare funding to broaden participation. Overall, integration of modeling approaches from machine learning and biophysics will rely on building the human resources that enable cross-disciplinary synthesis.

3.2 Predicting protein function from sequence and environment

The last several decades have produced an explosion of NGS and MS data that catalogs and characterizes the parts list of proteins in a given system. We now have the sequences and behavior of proteins from different species, organisms, and distinct individuals within a species. Over fifty years ago, the central dogma of molecular biology outlined a predictive, mechanistic understanding of information flow from genetics (DNA) to proteins (the amino acid sequence)[39]. With all of this data, a new question can be addressed: how does variation in sequence and environment affect protein function? Today, thanks to advances in machine learning, genomic databases, and the large number of structures in the PDB, we have the ability to accurately predict a structural model for the folded structure of a protein given only its amino-acid sequence[40]. While incredibly important, structure alone does not define the function of a protein and phenotype of an organism or cell[41]. Small variations in sequence and environment - which feeds evolution, phenotypic diversity, and the mechanistic details of function - rarely changes the overall structure of a protein. Proteins, the molecular machines of life, cannot be thought of as a single static structure, but an ensemble of structures that interconvert to create function and contribute to phenotype[42]. Addressing this grand challenge will require a synthesis of data at many different scales, from biophysical studies to genetic and population studies.

Diverse scientists representing all areas of MCB research can contribute to this challenge, spanning from detailed studies of individual proteins, their fluctuations and function, to high-throughput genetic studies. For example, in addition to the wealth of data in the BRENDA database[43] (the comprehensive enzyme information system) that catalogs enzyme function and the biophysical parameters that define an enzyme, recent developments in microfluidics have allowed the characterization of enzymological parameters for every single site variant of a single enzyme. Detailed biophysical studies on individual proteins determining their stability have been cataloged in the Protherm database[44], and while many site-specific mutations have been studied, high-throughput effects on protein stability are still lacking. The function and behavior of many proteins are now assayed via a DNA readout that allows for high-throughput screens and selections, such as those generated by deep mutational scanning[45] (DMS) or massively parallel reporter assays[5]. The dynamics, energetics, folding and function of many proteins and their variants are analyzed using data from MS – via amide hydrogen exchange, protease sensitivity, and sensitivity to modifications such as hydroxy-radical footprinting[46]. Genetic screens identify the effects of sequence environmental variation on phenotype[47]. Finally, simulation and theory has investigated the effect of sequence variation on the energy landscape and protein folding[48]. Independently, all of these experiments have provided us with important insight into the functioning of individual proteins and specific sequences. The time is ripe to synthesize these databases and develop predictive models.

A synthesis center is needed to harmonize these diverse data sets, to reconcile issues in existing data sets, such as BRENDA, which were not designed for data mining and thus are currently not used for synthesis purposes, and to serve as a hub, bringing different communities together. We believe that creation of such a synthesis center will stimulate the focus of researchers in their independent laboratories on some agreed upon systems and variants, both in experimental and computational studies, which, compared to the current uncoordinated approaches, would accelerate the overall mission. This could also result in the sharing of reagents such as DNA synthesis libraries, which would lower the overall cost of doing science and would better enable the science in individual labs. Thinking forward, this approach will create an iterative cycle where the computational studies will drive new experiments and vice versa. The ML training models will help define the needed or missing data, which will lead to a focus on new experimental developments in these needed areas. Thus, a synthesis center is both required to solve this grand challenge and will catalyze new and exciting science and methods development.

The time is ripe for addressing this challenge in a sustainable and scalable fashion. In fact, starting with small steps, such as coordination of many existing databases and archives would help the community better utilize data for machine-learning purposes. The synthesis center should provide the infrastructure and resources to link the databases in a coordinated way. The center should serve as both the integration hub and a community to set the standards for databases, to provide the definition of what types of data are needed, and to define how they should be integrated and linked. The ultimate goal would be a Federation of coordinating databases, which would serve as analytical tools for scientists worldwide. We believe that the individual communities have the expertise in the needs for their specific databases - but a two-way interaction with the synthesis center is needed to define what other users may want to best utilize/reuse the data. For instance, for some raw data might be very useful, while for others more processed outputs and results are what are needed. The center should figure out when and how intermediate steps in processing should be available. The center should lower the existing barriers in harnessing the massive amount of data currently available in Supplementary Information files that are not data minable. Additionally, there ought to be an obvious trajectory for such datasets to mature into databases when appropriate. In sum, an effective synthesis center would serve as a ‘clearinghouse’ to provide resources for both the ‘buyers’ and ‘sellers’, enhancing the individual components to be more than the sum of the parts and to accelerate science.

By its very nature, a synthesis center that allows access to different types of data, data that is usually expensive and obtained at R1 Universities, is inherently democratizing. Just like was seen with the PDB, access to the hub enables science and scientists anywhere in the country to synthesize and learn from the data. The resulting science will be enhanced by having such a diversity of participants. We believe that such a synthesis center will play a unique role in its ability to highlight the value of science and its global impact to the broader community. The larger communal discoveries enabled by synthesis – such as the recent developments in protein structure prediction – highlight the importance of the incremental steps in science that come together to make big discoveries. In addition to accelerating science, the synthesis center can function as a hub for diversity activities by utilizing and highlighting scientists from diverse backgrounds at the hub level and reducing the burden and minority tax that so many suffer from having to do it all on their own at individual institutions.

On the education front, to help train a future generation of biological scientists and encourage those with skills and interest in computation to work in the biological sciences, the synthesis center could develop online tutorials for machine learning that directly use examples drawn from the biomolecular sciences. We believe that even at the high school level, computer science classes could have examples drawn from biology.

3.3 Genotype to Phenotype in the Environmental Context

Both phenotype and genotype are environmentally determined. Evolutionary selection pressures select against phenotype, shaping genotype, which encodes phenotype[49-52]. A challenge then is understanding how, at the molecular and system's level, genotype gives rise to phenotype in different environmental contexts[53,54]. For example, organisms undergo varying environmental conditions (salinity, pH, temperature, oxygen stress, starvation, etc.) - how do the "functional modules" that carry out various subcellular processes respond to these changes? Addressing this challenge would advance our understanding of how cells maintain cellular homeostasis, and which functional modules respond to different environmental conditions.

To address this challenge, practical questions need to be answered. (i) Can we use existing data to conceptualize how to identify functional modules? (ii) How do functional modules communicate with each other in a given environmental setting? And (iii) how modular are the functional modules and how might they change in a new environment? In addition, can we predict the changes in the functional modules, predicted in one environment, to another?

Indeed, it is likely easier to identify functional modules by comparing across environments, rather than identifying a functional module by studying a single environment. Environmental changes represent a perturbation to the organism and subcellular processes that can be used to classify groups of genes, proteins, and biochemical pathways that act in a coordinated fashion. Such approaches have been used to identify new heat shock proteins, for example. But the potential for much widespread identification of functional modules is large if existing data is harnessed and analyzed.

Addressing this challenge is difficult, as there is only one database, STRING, run by a consortium in Europe which may be useful as a starting point. However, the database that is needed for this challenge is to modify the database using feedback so that testable hypotheses can be generated. In addition, the database that is needed should include environmental information, which is important if one wants to understand mechanisms for adaptation. A database that integrates mutagenesis data, including information from different species, environments, is badly needed. Such a database must include biophysical/biochemical parameters, along with high-throughput mutagenesis (phenotype/fitness/growth rate) data. As part of this effort, statistical metrics to assess accuracy of data should be applied to all public databases so that the broader community can use them with confidence.

The nature of challenges posed above clearly requires a broad spectrum of scientific expertise. To fruitfully attack outstanding problems requires expertise in data science, computer programs, and experts with biological understanding. Because at the core this requires synthesis, we also need theorists with broad experience as well. Because the expertise in a number of fields, starting with formulating answerable questions using the existing databases, is needed, creation of a Synthesis Center is the optimal choice.

The Synthesis Center will address the following issues: (1) Curating the existing data sets and analyzing their appropriateness, especially the quality of the datasets, is the first task that would allow one to abstract general questions. (2) In order to accomplish this task, ideas in ML will be fruitful with the challenge that one needs to devise new neural network models. (3) The generation of models from data should have predictive power, which means one ought to come up with extrapolatable models. To accomplish some of these goals, it is necessary to bring scientists, programmers, and statisticians to work together.

Although certain goals can be accomplished, eventually the community will demand curated databases tying together this information on all scales (molecular, to gene expression, to phenotype). If the right database is created, it will lead to community standards for reporting quality of data, e.g., high-throughput mutagenesis, beyond 'coverage', PHRED scores. To accomplish this task, one should convene community standards working groups in the context of a synthesis center. Such a working group must have members with expertise who may not nominally be associated with the Synthesis Center.

3.4 Extrapolating across species

Given the constraints of time and resources, the scientific community is restricted to studying a handful of model organisms[55]. Therefore, in order to obtain knowledge about the large biodiversity of species around the world, we will need to synthesize knowledge from these model organisms[56] to then apply them to non-model organisms[57,58]. More specifically, a grand-challenge for synthesis is (1) to identify what genotype-phenotype relationships can be transferred from model organisms to non-model organisms[56]; (2) to identify the minimal set of experiments we would need to perform in non-model organisms for specific questions; (3) to disentangle the contributions of environmental selection from those of molecular constraints across a phylogenetic tree[59]. Synthesis of data/knowledge in this regard also has biomedical implications. For instance, knowledge synthesis across species can help identify organisms that might be more suitable models for specific human pathologies[59-62].

The scientific challenge in achieving this synthesis is in mapping what is 'similar' in model systems to what is cognate in the new one and identifying the novel elements of the new one. In other words, how do we determine which new elements will affect predictions of function/behavior made from the old system? This requires ontological mapping of elements from one system to another (biomolecular 'ids' via analyses such as homology; activities such as predicted reactions or growth rates, and measurements), and mapping of experimental conditions and designs effectively.

3.5 Experimental error, combining models, and automated meta-analysis

Quantification[63-66], coupled cellular processes[67-69], and repeated meta-analyses[70] are an essential part of research in systems and synthetic biology. There is a community-wide challenge associated with each of these research aspects.

First, for each type of experimental measurement, data can have random or systematic biases that are poorly characterized and not often considered during model training and testing[71,72]. This is a common challenge as almost all measurements in molecular and systems biology provide indirect proxies of the quantity-of-interest. Read counts are a numerical quantity that measure nucleic acid (DNA or RNA) concentrations[73]. Fluorescence levels are a numerical quantity that measure the concentration of a fluorescent dye or protein[74]. Cycle thresholds from RT-qPCR measurements are a numerical quantity that measures nucleic acid (DNA or RNA) concentrations[75]. While we often assume that these measurements yield proportional proxies for the quantities-of-interest, there are common scenarios where the relationship is not proportional (*e.g.,* non-linear)[76], including sub-sampling and detection limits.

It's possible to develop a model of the measurement process itself in order to understand and compensate for random and/or systematic biases. Such models can be expressed using probability theory and stochastic processes to quantify the flow of material and information through the multi-step experimental workflow used to generate measurements. Numerical techniques, e.g. Monte Carlo, can be used to simulate outcomes of the experimental measurements in order to determine their intrinsic variability. The development of these models would create a "calibration curve" that transforms measured numerical values into the actual quantities-of-interest (e.g. expressed as a probability distribution function) to take into account the biases of the experimental workflow.

Second, it remains a challenge to combine multiple model predictions together into a self-consistent, seamless model, particularly when the individual models are formulated differently[77-79]. It also remains a challenge to explicitly take into account the uncertainty of each model prediction when feeding it into another model.

One way to combine multiple model predictions together is to reformat all model outputs as a probability distribution function (PDF), which explicitly includes the error and uncertainty in the model's predictions. Accordingly, one could then feed one model prediction as an input into another model prediction in terms of PDFs, which allows error and uncertainty to be processed and transformed as any other function.

Third, computational modeling and analysis techniques are constantly evolving and improving[29,80]. New approaches are needed to ensure 100% reproducibility when applying older techniques on older datasets. Additionally, when new computational models and analysis techniques are developed, new approaches are needed to apply these new techniques on all prior datasets.

New web-based platforms can be developed to enable: (i) upload of raw data; (ii) automated analysis of data with selectable workflows and options tailored for each type of data; (iii) generation of well-labeled and analyzed dataset; (iv) visualizability of analyzed data; and (v) links to the complete record of data and analysis. The platform can be readily extensible and 100% reproducible using new modeling and analysis pipelines by taking advantage of virtual environments, source code revision repositories, and containerization, for example, combining Docker and GitHub.

Overall, any web-based platform must be readily accessible to researchers who do not have experience with programming. There must be a clear incentive to upload data onto the platform, for example, by enabling improved data analysis. Convincing researchers to change their practices remains challenging. Rather than composing persuasive arguments, it would be better to develop "Show and Tell" educational examples that illustrate tangible and beneficial outcomes. This is a "smooth" way of convincing people to spend a bit of extra effort (automated) instead of defaulting to their "old way". If an approach solves a community-wide problem and if it's easy to carry out the approach, then it will be used.

3.6 Other outstanding challenges

At the workshop, 80 different challenges were proposed for MCB Synthesis, and the participants then voted on what they thought were the most exciting challenges. Thus, while the aforementioned represent the top five challenges, others are worthy of mention. They are:

1. How can we integrate the 'epigenetics view' and the 'proteostasis view' of cellular aging? These two molecular perspectives on aging mechanisms are siloed - there are no bridges being built between these two views. With large amounts of data associated with each view, this question is primed for synthesis.
2. Can we detect deviations from steady state and understand their impact on molecular and cellular biology? In the face of sampling statistics and other forms of noise, are there generalizable analysis methods that can help us identify from big data deviations of a system, or components of a system, away from steady state? Answering this question would allow for interesting phenomena to be automatically or rapidly identified in big data.
3. Can we predict which proteins behave similarly in vivo and in vitro? MCB researchers are actively experimenting to answer this question. Can it be answered with a synthetic approach?
4. How is the network of interactions (protein-protein; protein-DNA; protein-mRNA) wired under different cellular conditions? Cell environments and growth conditions are constantly changing. Can we take a synthetic approach to understand how these networks change under these different conditions?

4. Recommendations

4.1 A strategy for rapid impact

Two high-level strategies are needed for a synthesis center to have rapid impact.

4.1.1 *Focus on Class 1 and 2 problems.* Class 1 problems represent synthesis efforts that would yield the most impact for the least input of center resources. Therefore, it is advised that for the first several years the center prioritizes support for Class 1 problems and some Class 2 problems. As the center delivers on these efforts, approaches and procedures to support synthesis will be refined and iterated upon, making it possible to efficiently deliver on more challenging Class 3 problems. As the center matures, the portfolio of synthesis challenges should shift to a larger proportion of Class 2 and 3 problems.

Class 1 problems (which have data and modeling tools 'readily' available) supported by the Center must (*i*) represent a threshold of effort and expertise that is beyond the capabilities of a super-majority of MCB labs, and (*ii*) not represent a logical, tractable extension of ongoing research projects within a working group member's lab. These requirements will ensure that center resources are used for community-wide synthesis, and not for the benefit of one particular lab.

4.1.2 *Take an 'Agile' approach to Community-wide Synthesis.* In the 1970's it was common for software firms, such as IBM, to spend years and large amounts of resources developing software for users, only to find it didn't meet user needs and therefore wasn't widely used. This top-down software development approach has been largely replaced by an 'Agile' workflow, which involves an 'iterative approach to project management and software development that allows teams to deliver value by responding to their customers' needs faster. Instead of betting everything on a "big bang" launch, an agile team delivers work to users in small, but consumable, increments.' (From Atlassian.com)

Software development offers lessons for effectively supporting synthesis research efforts. There are innumerable data sets, databases, computational and theoretical tools that can be utilized for synthesis. It is a natural tendency to think that all these diverse data and tools should be brought together under one umbrella - either through a unified database, or software that ties all these analysis tools together. The history of software development suggests this costly, upfront approach is ill-fated. Instead, technical support of synthesis should meet only the goals of the communities of scientists who are addressing specific, focused questions in an iterative and timely process with technical support staff. The development of computational methodologies or processed datasets for synthesis should be limited to those needed to address specific synthesis questions. This approach will avoid a synthesis center squandering resources on a top-down approach that might not end up widely used.

As diverse synthesis efforts are supported over several years, it will become evident which data and tools have wide-spread use. Only at this point should a synthesis center invest strategically in creating general databases and software tools.

4.2 Structure of a Synthesis Center

4.2.1 *Leadership Team.* A leadership structure involving a director, co-director, assistant directors and their staff teams is a minimum requirement for a Synthesis Center. The director, with the aid of the Scientific Advisory Board, sets the vision for the center. The co-director handles the day-to-day operations, and oversees assistant directors who lead teams of (*i*) technical staff; (*ii*) education and outreach, and (*iii*) event coordination.

4.2.2 *Scientific Advisory Board.* A scientific advisory board should be created consisting of a diverse team of scientists who serve three-year terms and provide input on strategic planning, review requests for support, and make recommendations for programmatic focus to the Leadership Team. These board members represent the diverse scientific and computational drivers relevant to community-wide synthesis efforts.

4.2.3 *Technical Staff.* Technical staff are a pool of PhD trained scientists, data scientists, and software/database engineers that are matrixed out to support working groups, catalyst meetings, and Center postdoctoral Fellows in their synthesis efforts. 'Matrixing out' refers to matching technical staff with the needs of particular synthesis projects. A technical staff member will typically support up to four synthesis projects simultaneously, with project commitments lasting between a few weeks to a few months. Technical staff expertise of value to the MCB community are machine learning, statistics, data science, bioinformatics, computational biology, and molecular and theoretical modeling.

4.2.4 *Academic Consultants.* Matching the right data and tools to an impactful question is often a barrier to synthesis. When working groups, technical staff, and the scientific advisory board are unable to make this match, a working group will be able to reach out to academic consultants that can serve as either 'Data guides' or 'Tool guides' and are financially compensated for their time. A Data Guide is an academic (professor, postdoc, or graduate student) who has expert knowledge in a particular class of data who can help a working group find the best data to use, provide working knowledge about that data, or match them with another expert who can do this. A Tool Guide, similarly, is an academic who has expertise in computational or theoretical tools that technical staff or working groups do not possess, but a working group requires to make progress. Critically, these consultants are not members of a working group, nor collaborators on the project. They are consultants, who help make these matches to data and tools for a short, predefined time (typically less than 10 hours of consulting time). For this reason, they are not identified as authors on resulting publications. Conflicts of interest arise if a person takes on the dual role of consultant and collaborator on a project - as this can be viewed as "double dipping", getting extra pay for scientific research that the consultant is also academically benefitting from in terms of publications. To avoid this potential conflict of interest, regular auditing should be carried out of publication outputs by consultants.

We recognize that paid academic consultants are a foreign concept to academic research, as many scientists are motivated by a love for knowledge and discovery. However, we must also recognize that academics have many demands on their time, and those academics with the most useful data or tool knowledge will likely get multiple requests for consultation from working groups. Therefore, to motivate over-stretched academics, we believe financially remunerating academic consultants will accelerate synthesis research and discovery.

4.2.5 *Virtual components.* Many aspects of collaboration, education and outreach can be done remotely, as has been demonstrated by scientists during the Covid pandemic. Virtual components should be carried out where it is clear that such remote work will accelerate synthesis efforts. For example, virtual meetings can be held before and after an in-person catalyst or working group meetings to maximize the efficiency and impact of the in-person meetings. Additionally, outreach and education efforts can maximize their impact through hybrid meetings, archived video recordings, and materials that are made publicly available.

4.2.6 *Multiple meeting sites.* As a national and international resource, a Synthesis center should support domestic off-site working group and catalyst meetings when such off-site meetings have a clear benefit that cannot be provided by hosting a meeting at the site of the Synthesis center. This is a more expensive option, including supporting travel of center staff to the site. Therefore, budgeting will factor into the number of such off-site meetings in a given year.

4.2.7 *Education and Outreach Assistant Director.* Education and outreach are an essential part of a synthesis center. This Assistant Director will coordinate and plan education and outreach efforts of the center that are described in Section 4.3.

4.2.8 *Ad-hoc reviewers.* When a request for center resources involves expertise outside the knowledge of the scientific advisory board, ad hoc reviewers should be utilized to fill these knowledge gaps.

4.2.9 *Scientific Communication.* Novel means of supporting communication of Center efforts and results should be considered. A synthesis center could leverage journalism majors interested in scientific communication to write news pieces and center newsletters. This would provide an educational opportunity for students as well as communicate the activities and results of the center

4.2.10 *Center Postdoctoral Fellows.* The center should support a new cohort of around five postdocs each year whose projects focus on the synthesis of existing MCB data. These postdocs would receive three years of support from the Center, and be independent. Meaning that they are free to choose the research projects they work on, and are not under the employ of a professor. They will select three mentors from across the world who will advise them scientifically and on their career. Postdoc applications to the center will require proposal of a synthesis project. At steady state, around 15 postdocs will be active in the center, providing a community within which they can work and synergize. Postdocs are free to choose whether they take part in working group and catalyst meetings.

4.2.11 *Graduate Fellowship lines to support Community Synthesis.* A synthesis center should leverage the resources of the host university to support synthesis. An effective way to do this is to support the equivalent of a ‘Teaching Assistantship’ in which graduate students in Statistics, Machine Learning, Bioinformatics, Genomics, Computational Biology, and Biophysics are paid to be assistants mentored under the Technical Staff to support synthesis efforts of working groups, catalyst meetings, and postdoctoral fellows. As in conventional TA lines, students supported by this mechanism would be expected to provide, on average, 20 hours of work each week. This would have the added benefit of training the next generation of scientists.

4.2.12 *Post-bacs to support Community Synthesis.* Another mechanism to leverage and train young scientists is to provide two-year post-baccalaureate positions in which scientists with relevant bachelor’s degrees would be mentored by Technical staff to work on supporting synthesis efforts of working groups, catalyst meetings, and postdoctoral fellows. This would provide these post-bacs with real-world research experience and make them more competitive for graduate school, medical school, or industry.

4.3 Activities of a Synthesis Center

4.3.1 *Working Groups.* A highly effective mechanism for Synthesis are working groups, and have been used in other NSF supported centers. This is a diverse group of scientists (typically less than 15) that come together two-to-three times a year to collaborate on a big question in MCB that can be addressed through synthesis. These working groups are investigator led, and apply for support from the center for their activities. Importantly, working groups can be proposed and led by postdocs. Before obtaining support, applicants must demonstrate that there is sufficient data available to address the synthesis question, and that there are computational methods available or that can be developed during the time period the working group exists.

4.3.2 *Catalyst Meetings.* These are one-time meetings that bring together a diverse group of scientists to identify grand-challenges that can be addressed by synthesis, identify data needed to address that challenge as well as analysis tools, and build networks of scientists to address the challenge. Catalyst meetings result in potential future efforts for the synthesis center through working groups, postdoctoral fellows, and other activities at the center. Proposal’s for catalyst meetings can come from graduate students, postdocs, or faculty.

4.3.3 *Center Postdoctoral Fellows.* As described in Section 4.2, a cohort of 15 postdocs would be supported by the center. These fellows should be selected through a national competition in which applicants submit a research project centered on MCB synthesis. The advisory board would select the most meritorious applicants. Postdoctoral fellows would thus form one of the drivers for synthesis.

4.3.4 *Sabbaticals*. Faculty sabbaticals should be supported to allow faculty to spend time at the center to give them time to focus on starting a synthesis project, or continue an ongoing synthesis project already supported by the center. Sabbatical applications should provide a plan on how the faculty member will use the time and center resources to move their synthesis efforts forward.

4.3.5 *Short-term visitors.* Scholars from around the country and world should be invited to visit the center for short periods (2 weeks to 3 months) to carry out synthetic research that leverages center resources. These visitors can be graduate students through full professors. The diversity of expertise and research efforts of short-term visitors will contribute to a vibrant intellectual environment at the center.

4.3.6 *Community Competitions.* There are likely to be times when Catalyst and Working Groups put together data and synthesis questions that may be amenable to simple, rapid innovative solutions. In such situations, the center should support “community competitions”, which are virtual (or hybrid) events with a well-defined goal where individuals or teams can enter to win money or an award if they demonstrate they can achieve the goal. Similar to “hack-a-thons”, these community competitions would involve individuals or teams competing against each other.

4.3.7 *Journalist in residence.* Science and society intersect at a synthesis center. As a place of scholarship studying cutting-edge questions in MCB that can have relevance to society, a journalist in residence program should be arranged where journalists can spend up to 3 weeks at the center learning the latest science that they would be expected to report on through various media: print, radio, TV, or internet.

4.3.8 *Philosopher in residence.* Philosophers of Science have a unique opportunity to study MCB research at a unique time in its history, where the scholars doing the science can gather diverse data from multiple labs to get greater insight. Therefore, it is suggested a Philosopher in residence program be instituted where they are able to observe and take part in the various center activities to understand how MCB Synthesis research is done.

4.3.9 *Open Science.* As a community resource, a synthesis center should follow best practices for early stage, open science. In this case, this means the results of catalyst and working group meetings should be posted as soon as possible. Datasets and databases should be posted as they are created (before publication) to allow other scientists to start to analyze them.

4.3.10 *Reproducible Science.* The synthesis center should support reproducible science standards by requiring best practices in the research it supports. This includes the use of Jupyter Notebooks to run code, GitHub to provide source code and examples, Docker containers to allow anyone to run code remotely, data provenance standards - such as providing doi’s for new compiled datasets, and other such approaches.

4.3.11 *Classes.* Short courses on different topics should be offered, including Introductory Python programming; Introduction to Jupyter Notebooks and GitHub; Introductory and Intermediate Machine Learning using Scikit-Learn; Experimental methods for modelers; Computation for Experimentalists; Molecular Modeling; A responsive approach to community needs will result in other courses.

4.3.12 *Pipeline from undergraduate to graduate school*. Summer internships for undergraduates from primarily undergraduate institutions and Minority Serving Institutions could help increase the number of under-represented minorities entering STEM graduate studies. These internships could see these undergraduates be introduced to computation and synthesis and carry out research.

4.3.13 *Scientific Communication.* The center should support the training of its members in effective scientific communication. This includes across fields, within working groups and catalyst meetings. And in communicating with the public. To achieve this, when a new catalyst meeting or working group is approved by the center, the professional facilitator/management trainer will meet with the leaders of each to introduce them to scientific communication techniques when dealing with a diverse team. The trainer will also help them come up with plans for effective management and timelines.

4.3.14 *Compendium of Best Practices.* As a national resource, the center can promote and make standard best practices in data, computation, and modeling for synthesis. Therefore, an effort should be made up-front to build a compendium of best practices for these various aspects of MCB synthesis. This can be made part of the various activities of the center.

4.3.15 *Outreach.* By following the recommendations above, a number of outreach goals will be achieved, including the training of undergraduates, graduates, postdocs, as well as classes for cross-disciplinary training and facilitator led educational efforts on scientific communication. The center should involve minority serving institutions and primarily undergraduate institutions to engage them in synthesis research. Activities can include both virtual and in-person summer camps on big data and how computation is solving long standing challenges in MCB. Summer internships for undergraduates from these institutions would allow them to get hands on experience with synthetic research. Outreach to the public is important too. Popular science lectures, coding camps, and illustrating how biological data and computation are coming together to create advances that benefit society are outreach activities worth pursuing.

4.3.16 *Federated Database Creation.* We anticipate that a natural evolution will occur within a center regarding data. Initial synthesis efforts will involve creation of specific datasets for answering specific questions. As time progresses it will become apparent that some datasets or resources are in much higher demand than others. In those cases, it may accelerate synthesis to create a ‘Federated’ database, which pulls information from existing, publicly available databases to make access to relevant, diverse data more dynamic and easy. A federated database uses an API to pull information from other databases, and presents the information to the end user in an easy to use format.

5. Summary

Many areas of science are now making progress through the use of large data sets and computation. The Molecular and Cellular Biology community is primed for rapid and sustained scientific discovery over the next decade if the National Science Foundation provides the infrastructure to enable diverse groups of scientists to efficiently bring together existing large-scale data sets with domain experts in computation and theory. The community can only go after the grand challenges in this report through community-wide efforts as no individual lab in the United States possesses the diverse expertise and resources to pursue them. A synthesis center, well designed and implemented, will foster, coordinate, and accelerate these efforts through the strategic activities we have recommended. This will benefit scientific research, scientific training and education, and make the United States more economically competitive. Questions that were not hitherto able to be addressed have the potential to be solved. Engaging scientists at diverse career stages, disciplines, backgrounds and outreach and education to the general public will create a better trained STEM workforce and scientifically literate society. Some of the basic research discoveries made through an MCB synthesis center will undoubtedly have impact on economically important sectors such as the life sciences, medicine, biotechnology, and artificial intelligence. Thus, now is the time for the National Science

Foundation’s Molecular and Cellular Biology Division to support a synthesis center that will catalyze US scientific research with results that will reverberate around the world.

6. References

1. Edgar, R.; Domrachev, M.; Lash, A. E., Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 2002, *30* (1), 207-210, doi: https://doi.org/10.1093/nar/30.1.207.

2. Barrett, T.; Wilhite, S. E.; Ledoux, P.; Evangelista, C.; Kim, I. F.; Tomashevsky, M.; Marshall, K. A.; Phillippy, K. H.; Sherman, P. M.; Holko, M., NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* 2012, *41* (D1), D991-D995, doi: https://doi.org/10.1093/nar/gks1193.

3. Vizcaino, J. A.; Deutsch, E. W.; Wang, R.; Csordas, A.; Reisinger, F.; Rios, D.; Dianes, J. A.; Sun, Z.; Farrah, T.; Bandeira, N.*, et al.*, ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* 2014, *32* (3), 223-226, doi: https://doi.org/10.1038/nbt.2839.

4. Deutsch, E. W.; Csordas, A.; Sun, Z.; Jarnuczak, A.; Perez-Riverol, Y.; Ternent, T.; Campbell, D. S.; Bernal-Llinares, M.; Okuda, S.; Kawano, S.*, et al.*, The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition. *Nucleic Acids Res.* 2017, *45* (D1), D1100-D1106, doi: https://doi.org/10.1093/nar/gkw936.

5. Kinney, J. B.; McCandlish, D. M., Massively parallel assays and quantitative sequence–function relationships. *Annual review of genomics and human genetics* 2019, *20*, 99-127, doi: https://doi.org/10.1146/annurev-genom-083118-014845.

6. Inoue, F.; Kreimer, A.; Ashuach, T.; Ahituv, N.; Yosef, N., Identification and massively parallel characterization of regulatory elements driving neural induction. *Cell Stem Cell* 2019, *25* (5), 713-727. e10, doi: https://doi.org/10.1016/j.stem.2019.09.010.

7. Trauernicht, M.; Martinez-Ara, M.; van Steensel, B., Deciphering Gene Regulation Using Massively Parallel Reporter Assays. *Trends Biochem. Sci* 2020, *45* (1), 90-91, doi: https://doi.org/https://doi.org/10.1016/j.tibs.2019.10.006.

8. Qiu, Q.; Hu, P.; Qiu, X.; Govek, K. W.; Cámara, P. G.; Wu, H., Massively parallel and time-resolved RNA sequencing in single cells with scNT-seq. *Nat. Methods* 2020, *17* (10), 991-1001, doi: https://doi.org/10.1038/s41592-020-0935-4.

9. Johnson, D. S.; Mortazavi, A.; Myers, R. M.; Wold, B., Genome-wide mapping of in vivo protein-DNA interactions. *Science* 2007, *316* (5830), 1497-1502, doi: https://doi.org/10.1126/science.1141319.

10. Barski, A.; Cuddapah, S.; Cui, K.; Roh, T.-Y.; Schones, D. E.; Wang, Z.; Wei, G.; Chepelev, I.; Zhao, K., High-resolution profiling of histone methylations in the human genome. *Cell* 2007, *129* (4), 823-837, doi: https://doi.org/10.1016/j.cell.2007.05.009.

11. Schmid, C. D.; Bucher, P., ChIP-Seq data reveal nucleosome architecture of human promoters. *Cell* 2007, *131* (5), 831-832, doi: https://doi.org/10.1016/j.cell.2007.11.017.

12. Churchman, L. S.; Weissman, J. S., Native elongating transcript sequencing (NET-seq). *Curr. Protoc. Mol. Biol.* 2012, *98* (1), 14.4. 1-14.4. 17, doi: https://doi.org/10.1002/0471142727.mb0414s98.

13. Zhu, J.; Liu, M.; Liu, X.; Dong, Z., RNA polymerase II activity revealed by GRO-seq and pNET-seq in Arabidopsis. *Nature plants* 2018, *4* (12), 1112-1123, doi: https://doi.org/10.1038/s41477-018-0280-0.

14. Tilgner, H.; Knowles, D. G.; Johnson, R.; Davis, C. A.; Chakrabortty, S.; Djebali, S.; Curado, J.; Snyder, M.; Gingeras, T. R.; Guigó, R., Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res.* 2012, *22* (9), 1616-1625, doi: https://doi.org/10.1101/gr.134445.111.

15. Giess, A.; Cleuren, Y. N. T.; Tjeldnes, H.; Krause, M.; Bizuayehu, T. T.; Hiensch, S.; Okon, A.; Wagner, C. R.; Valen, E., Profiling of small ribosomal subunits reveals modes and regulation of translation initiation. *Cell reports* 2020, *31* (3), 107534, doi: https://doi.org/10.1016/j.celrep.2020.107534.

16. Zhang, P.; He, D.; Xu, Y.; Hou, J.; Pan, B. F.; Wang, Y.; Liu, T.; Davis, C. M.; Ehli, E. A.; Tan, L.*, et al.*, Genome-wide identification and differential analysis of translational initiation. *Nat. Commun.* 2017, *8* (1), 1-14, doi: https://doi.org/10.1038/s41467-017-01981-8.

17. Ingolia, N. T., Ribosome profiling: new views of translation, from single codons to genome scale. *Nature reviews genetics* 2014, *15* (3), 205-213, doi: https://doi.org/10.1038/nrg3645.

18. Zhao, T.; Chen, Y.-M.; Li, Y.; Wang, J.; Chen, S.; Gao, N.; Qian, W., Disome-seq reveals widespread ribosome collisions that promote cotranslational protein folding. *Genome Biol.* 2021, *22* (1), 1-35, doi: https://doi.org/10.1186/s13059-020-02256-0.

19. Oh, E.; Becker, A. H.; Sandikci, A.; Huber, D.; Chaba, R.; Gloge, F.; Nichols, R. J.; Typas, A.; Gross, C. A.; Kramer, G.*, et al.*, Selective ribosome profiling reveals the cotranslational chaperone action of trigger factor in vivo. *Cell* 2011, *147* (6), 1295-1308, doi: https://doi.org/10.1016/j.cell.2011.10.044.

20. Lugowski, A.; Nicholson, B.; Rissland, O. S., Determining mRNA half-lives on a transcriptome-wide scale. *Methods* 2018, *137*, 90-98, doi: https://doi.org/10.1016/j.ymeth.2017.12.006.

21. Sample, P. J.; Wang, B.; Reid, D. W.; Presnyak, V.; McFadyen, I. J.; Morris, D. R.; Seelig, G., Human 5′ UTR design and variant effect prediction from a massively parallel translation assay. *Nat. Biotechnol.* 2019, *37* (7), 803-809, doi: https://doi.org/10.1038/s41587-019-0164-5.

22. Gao, R.; Yu, K.; Nie, J.; Lian, T.; Jin, J.; Liljas, A.; Su, X.-D., Deep sequencing reveals global patterns of mRNA recruitment during translation initiation. *Sci. Rep.* 2016, *6* (1), 1-11, doi: https://doi.org/10.1038/srep30170.

23. Kinney, J. B.; Murugan, A.; Callan, C. G.; Cox, E. C., Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc. Natl. Acad. Sci.* 2010, *107* (20), 9158-9163, doi: https://doi.org/10.1073/pnas.1004290107.

24. Mighell, T. L.; Evans-Dutson, S.; O'Roak, B. J., A saturation mutagenesis approach to understanding PTEN lipid phosphatase activity and genotype-phenotype relationships. *The American Journal of Human Genetics* 2018, *102* (5), 943-955, doi: https://doi.org/10.1016/j.ajhg.2018.03.018.

25. Matreyek, K. A.; Starita, L. M.; Stephany, J. J.; Martin, B.; Chiasson, M. A.; Gray, V. E.; Kircher, M.; Khechaduri, A.; Dines, J. N.; Hause, R. J.*, et al.*, Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat. Genet.* 2018, *50* (6), 874-882, doi: https://doi.org/10.1038/s41588-018-0122-z.

26. Mighell, T. L.; Thacker, S.; Fombonne, E.; Eng, C.; O'Roak, B. J., An integrated deep-mutational-scanning approach provides clinical insights on PTEN genotype-phenotype relationships. *The*

American Journal of Human Genetics 2020, *106* (6), 818-829, doi: https://doi.org/10.1016/j.ajhg.2020.04.014.

27. Biswas, J.; Liu, Y.; Singer, R. H.; Wu, B., Fluorescence imaging methods to investigate translation in single cells. *Cold Spring Harb. Perspect. Biol.* 2019, *11* (4), a032722, doi: https://doi.org/10.1101/cshperspect.a032722.

28. Greener, J. G.; Kandathil, S. M.; Moffat, L.; Jones, D. T., A guide to machine learning for biologists. *Nature Reviews Molecular Cell Biology* 2022, *23* (1), 40-55, doi: https://doi.org/10.1038/s41580-021-00407-0.

29. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.*, et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* 2021, *596* (7873), 583-589, doi: https://doi.org/10.1038/s41586-021-03819-2.

30. Meyer, J. G., Deep learning neural network tools for proteomics. *Cell Reports Methods* 2021, *1* (2), 100003, doi: https://doi.org/10.1016/j.crmeth.2021.100003.

31. Luria, S. E.; Delbrück, M., Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* 1943, *28* (6), 491, doi: https://doi.org/10.1093/genetics/28.6.491.

32. Watson, J. D.; Crick, F. H., Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature* 1953, *171* (4356), 737-738, doi: https://doi.org/10.1038/171737a0.

33. Crick, F. H.; Watson, J. D., Structure of small viruses. *Nature* 1956, *177* (4506), 473-5, doi: https://doi.org/10.1038/177473a0.

34. Brauer, F., Mathematical epidemiology: Past, present, and future. *Infectious Disease Modelling* 2017, *2* (2), 113-127, doi: https://doi.org/10.1016/j.idm.2017.02.001.

35. McIntosh, R. P., The background and some current problems of theoretical ecology. *Synthese* 1980, *43* (2), 195-255, doi: https://doi.org/10.1007/BF00413926.

36. Mugnai, M. L.; Hyeon, C.; Hinczewski, M.; Thirumalai, D., Theoretical perspectives on biological machines. *Rev. Mod. Phys.* 2020, *92* (2), 025001, doi: https://doi.org/10.1103/RevModPhys.92.025001.

37. Biswas, P., Theoretical and computational advances in protein misfolding. *Adv. Protein Chem. Struct. Biol.* 2019, *118*, 1-31, doi: https://doi.org/10.1016/bs.apcsb.2019.08.010.

38. Thurley, K.; Wu, L. F.; Altschuler, S. J., Modeling Cell-to-Cell Communication Networks Using Response-Time Distributions. *Cell Systems* 2018, *6* (3), 355-367.e5, doi: https://doi.org/https://doi.org/10.1016/j.cels.2018.01.016.

39. Crick, F., Central dogma of molecular biology. *Nature* 1970, *227* (5258), 561-563, doi: https://doi.org/10.1038/227561a0.

40. Pearce, R.; Zhang, Y., Toward the solution of the protein structure prediction problem. *J. Biol. Chem.* 2021, *297* (1), 100870, doi: https://doi.org/10.1016/j.jbc.2021.100870.

41. Hunter, D. J., Gene–environment interactions in human diseases. *Nature Reviews Genetics* 2005, *6* (4), 287-298, doi: https://doi.org/10.1038/nrg1578.

42. Nussinov, R.; Tsai, C.-J.; Jang, H., Protein ensembles link genotype to phenotype. *PLoS Comp. Biol.* 2019, *15* (6), e1006648, doi: https://doi.org/10.1371/journal.pcbi.1006648.

43. Chang, A.; Jeske, L.; Ulbrich, S.; Hofmann, J.; Koblitz, J.; Schomburg, I.; Neumann-Schaal, M.; Jahn, D.; Schomburg, D., BRENDA, the ELIXIR core data resource in 2021: new developments and updates. *Nucleic Acids Res.* 2021, *49* (D1), D498-D508, doi: https://doi.org/10.1093/nar/gkaa1025.

44. Nikam, R.; Kulandaisamy, A.; Harini, K.; Sharma, D.; Gromiha, M. M., ProThermDB: thermodynamic database for proteins and mutants revisited after 15 years. *Nucleic Acids Res.* 2021, *49* (D1), D420-D424, doi: https://doi.org/10.1093/nar/gkaa1035.

45. Fowler, D. M.; Fields, S., Deep mutational scanning: a new style of protein science. *Nat. Methods* 2014, *11* (8), 801-807, doi: https://doi.org/10.1038/nmeth.3027.

46. Petrotchenko, E. V.; Borchers, C. H., Protein Chemistry Combined with Mass Spectrometry for Protein Structure Determination. *Chem. Rev.* 2021, doi: https://doi.org/10.1021/acs.chemrev.1c00302.

47. Burgess, D. J., Shining a light on genetic screen strategies. *Nature Reviews Genetics* 2018, *19* (1), 6-7, doi: https://doi.org/10.1038/nrg.2017.99.

48. Röder, K.; Wales, D. J., The Energy Landscape Perspective: Encoding Structure and Function for Biomolecules. *Frontiers in Molecular Biosciences* 2022, *9*, doi: https://doi.org/10.3389/fmolb.2022.820792.

49. Mousavizadeh, L.; Ghasemi, S., Genotype and phenotype of COVID-19: Their roles in pathogenesis. *J. Microbiol. Immunol. Infect.* 2020, *54*, 159-163, doi: https://doi.org/10.1016/j.jmii.2020.03.022.

50. Heslot, N.; Akdemir, D.; Sorrells, M. E.; Jannink, J. L., Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. *Theor. Appl. Genet.* 2014, *127*, 463-480, doi: https://doi.org/10.1007/s00122-013-2231-5.

51. Lee, S. H.; Ripke, S.; Neale, B. M.; Faraone, S. V.; Purcell, S. M.; Perlis, R. H.; Mowry, B. J.; Thapar, A.; Goddard, M. E.; Witte, J. S.*, et al.*, Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat. Genet.* 2013, *45*, 984-994, doi: https://doi.org/10.1038/ng.2711.

52. Vidal, M.; Cusick, M. E.; Barabási, A. L., Interactome networks and human disease. *Cell* 2011, *144*, 986-998, doi: https://doi.org/10.1016/j.cell.2011.02.016.

53. Ritchie, M. D.; Holzinger, E. R.; Li, R.; Pendergrass, S. A.; Kim, D., Methods of integrating data to uncover genotype-phenotype interactions. *Nature Reviews Genetics* 2015, *16*, 85-97, doi: https://doi.org/10.1038/nrg3868.

54. Cobb, J. N.; DeClerck, G.; Greenberg, A.; Clark, R.; McCouch, S., Next-generation phenotyping: Requirements and strategies for enhancing our understanding of genotype-phenotype relationships and its relevance to crop improvement. *Theor. Appl. Genet.* 2013, *126*, 867-887, doi: https://doi.org/10.1007/s00122-013-2066-0.

55. Fields, S.; Johnston, M., Whither model organism research? *Science* 2005, *307*, 1885-1886, doi: https://doi.org/10.1126/science.1108872.

56. Wang, H.; Robinson, J. L.; Kocabas, P.; Gustafsson, J.; Anton, M.; Cholley, P. E.; Huang, S.; Gobom, J.; Svensson, T.; Uhlen, M.*, et al.*, Genome-scale metabolic network reconstruction of model animals as a platform for translational research. *Proc. Natl. Acad. Sci. U. S. A.* 2021, *118*, doi: https://doi.org/10.1073/pnas.2102344118.

57. Xiao, M.; Zhang, Y.; Chen, X.; Lee, E. J.; Barber, C. J. S.; Chakrabarty, R.; Desgagné-Penix, I.; Haslam, T. M.; Kim, Y. B.; Liu, E.*, et al.*, Transcriptome analysis based on next-generation sequencing of non-model plants producing specialized metabolites of biotechnological interest. *J. Biotechnol.* 2013, *166*, 122-134, doi: https://doi.org/10.1016/j.jbiotec.2013.04.004.

58. Council, N. R., Application of Toxicogenomics to Cross-Species Extrapolation: A Report of a Workshop. 2006, 58, doi: https://doi.org/doi:10.17226/11488.

59. Wang, L.; Tan, Y.; Yang, X.; Kuang, L.; Ping, P., Review on predicting pairwise relationships between human microbes, drugs and diseases: from biological data to computational models. *Briefings in Bioinformatics* 2022, 1-25, doi: https://doi.org/10.1093/bib/bbac080.

60. Sung, J. H.; Srinivasan, B.; Esch, M. B.; McLamb, W. T.; Bernabini, C.; Shuler, M. L.; Hickman, J. J., Using physiologically-based pharmacokinetic-guided “body-on-a-chip” systems to predict mammalian response to drug and chemical exposure. *Exp. Biol. Med.* 2014, *239* (9), 1225-1239, doi: https://doi.org/10.1177/1535370214529397.

61. Raies, A. B.; Bajic, V. B., In silico toxicology: computational methods for the prediction of chemical toxicity. *WIREs Computational Molecular Science* 2016, *6* (2), 147-172, doi: https://doi.org/https://doi.org/10.1002/wcms.1240.

62. Barton, H. A.; Chiu, W. A.; Woodrow Setzer, R.; Andersen, M. E.; Bailer, A. J.; Bois, F. Y.; Dewoskin, R. S.; Hays, S.; Johanson, G.; Jones, N.*, et al.*, Characterizing uncertainty and variability in physiologically based pharmacokinetic models: State of the science and needs for research and implementation. *Toxicol. Sci.* 2007, *99*, 395-402, doi: https://doi.org/10.1093/toxsci/kfm100.

63. Picotti, P.; Aebersold, R., Selected reaction monitoring-based proteomics: Workflows, potential, pitfalls and future directions. *Nat. Methods* 2012, *9*, 555-566, doi: https://doi.org/10.1038/nmeth.2015.

64. Singh, H.; Tiwari, K.; Tiwari, R.; Pramanik, S. K.; Das, A., Small Molecule as Fluorescent Probes for Monitoring Intracellular Enzymatic Transformations. *Chem. Rev.* 2019, *119*, 11718-11760, doi: https://doi.org/10.1021/acs.chemrev.9b00379.

65. Zhang, J.; Campbell, R. E.; Ting, A. Y.; Tsien, R. Y., Creating new fluorescent probes for cell biology. *Nature Reviews Molecular Cell Biology* 2002, *3* (12), 906-918, doi: https://doi.org/10.1038/nrm976.

66. Caicedo, J. C.; Cooper, S.; Heigwer, F.; Warchal, S.; Qiu, P.; Molnar, C.; Vasilevich, A. S.; Barry, J. D.; Bansal, H. S.; Kraus, O.*, et al.*, Data-analysis strategies for image-based cell profiling. *Nat. Methods* 2017, *14*, 849-863, doi: https://doi.org/10.1038/nmeth.4397.

67. Wang, D.; Lippard, S. J., Cellular processing of platinum anticancer drugs. *Nature Reviews Drug Discovery* 2005, *4*, 307-320, doi: https://doi.org/10.1038/nrd1691.

68. Stein, K. C.; Morales-Polanco, F.; van der Lienden, J.; Rainbolt, T. K.; Frydman, J., Ageing exacerbates ribosome pausing to disrupt cotranslational proteostasis. *Nature* 2022, *601*, 637-642, doi: https://doi.org/10.1038/s41586-021-04295-4.

69. Mayer, M. P.; Bukau, B., Hsp70 chaperones: Cellular functions and molecular mechanism. *Cell. Mol. Life Sci.* 2005, *62*, 670-684, doi: https://doi.org/10.1007/s00018-004-4464-6.

70. Otwombe, K. N.; Ogutu, B., Improving the quality of reports of randomised controlled trials: The consort statement. *East Afr. Med. J.* 2002, *79*, 394-396, doi: https://doi.org/10.1016/s0002-9394(14)70422-2.

71. Taub, M. A.; Corrada Bravo, H.; Irizarry, R. A., Overcoming bias and systematic errors in next generation sequencing data. *Genome Med.* 2010, *2* (12), 87, doi: https://doi.org/10.1186/gm208.

72. Yaffe, E.; Tanay, A., Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat. Genet.* 2011, *43* (11), 1059-1065, doi: https://doi.org/10.1038/ng.947.

73. Magi, A.; Tattini, L.; Pippucci, T.; Torricelli, F.; Benelli, M., Read count approach for DNA copy number variants detection. *Bioinformatics* 2012, *28* (4), 470-478, doi: https://doi.org/10.1093/bioinformatics/btr707.

74. Royer, C. A., Probing Protein Folding and Conformational Transitions with Fluorescence. *Chem. Rev.* 2006, *106* (5), 1769-1784, doi: https://doi.org/10.1021/cr0404390.

75. Nolan, T.; Hands, R. E.; Bustin, S. A., Quantification of mRNA using real-time RT-PCR. *Nat. Protoc.* 2006, *1* (3), 1559-1582, doi: https://doi.org/10.1038/nprot.2006.236.

76. Serrano, A. L.; Waegele, M. M.; Gai, F., Spectroscopic studies of protein folding: Linear and nonlinear methods. *Protein Sci.* 2012, *21* (2), 157-170, doi: https://doi.org/https://doi.org/10.1002/pro.2006.

77. Jacobson, L. D.; Bochevarov, A. D.; Watson, M. A.; Hughes, T. F.; Rinaldo, D.; Ehrlich, S.; Steinbrecher, T. B.; Vaitheeswaran, S.; Philipp, D. M.; Halls, M. D.*, et al.*, Automated Transition State Search and Its Application to Diverse Types of Organic Reactions. *J. Chem. Theory Comput.* 2017, *13* (11), 5780-5797, doi: https://doi.org/10.1021/acs.jctc.7b00764.

78. Nebgen, B.; Lubbers, N.; Smith, J. S.; Sifain, A. E.; Lokhov, A.; Isayev, O.; Roitberg, A. E.; Barros, K.; Tretiak, S., Transferable Dynamic Molecular Charge Assignment Using Deep Neural Networks. *J. Chem. Theory Comput.* 2018, *14* (9), 4687-4698, doi: https://doi.org/10.1021/acs.jctc.8b00524.

79. Wu, Z.; Milano, G.; Müller-Plathe, F., Combination of Hybrid Particle-Field Molecular Dynamics and Slip-Springs for the Efficient Simulation of Coarse-Grained Polymer Models: Static and Dynamic Properties of Polystyrene Melts. *J. Chem. Theory Comput.* 2021, *17* (1), 474-487, doi: https://doi.org/10.1021/acs.jctc.0c00954.

80. Alford, R. F.; Leaver-Fay, A.; Jeliazkov, J. R.; O'Meara, M. J.; DiMaio, F. P.; Park, H.; Shapovalov, M. V.; Renfrew, P. D.; Mulligan, V. K.; Kappel, K.*, et al.*, The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput.* 2017, *13* (6), 3031-3048, doi: https://doi.org/10.1021/acs.jctc.7b00125.

7. Acknowledgements

We thank the NSF for financial support (Grant number 2133405 “MoCeIS-DCL: Planning Workshops for Synthesis of Massively Parallel Assays and Molecular Physiology”), the professional facilitators from KnowInnovation and Dr. Sarah Leininger, the participants who took part in the workshop listed in Appendix 1, as well as the NSF Program Officers who took part as observers.

8. Appendix 1: Participants at Synthesis Workshop

Adam Arkin, Professor, University of California Berkeley

Adam Arkin is an expert in computational biology who created the US DOE systems biology Knowbase, KBase, that aims to enable researchers to predict and ultimately design biological function through a community shared platform.

Recent highlights:

- A method for achieving complete microbial genomes and improving bins from metagenomics data. PLoS Computational Biology, 2021.
- KBase: The United States Department of Energy Systems Biology Knowledgebase. Nature Biotechnology, 2018.

Minkyung Baek, Postdoctoral Scholar, University of Washington

Minkyung Baek has created a machine learning model that accurately predicts the folded structures of proteins while working in David Baker’s lab.

Recent highlights:

- Deep learning and protein structure modeling. Nature Methods, 2022.
- Accurate prediction of protein structures and interactions using a three-tract neural network. Science, 2021.

Wout Bittremieux, Postdoctoral Scholar, University of California San Diego

Wout Bittremieux develops algorithmic solutions and machine learning methods to analyze MS-based proteomics and metabolomics data to attempt to solve fundamental biological questions.

Recent highlights:

- A learned embedding for efficient joint analysis of millions of mass spectra. Nature Methods, 2022.
- Fast open modification spectral library searching through approximate nearest neighbor indexing. Journal of Proteome Research, 2018.

Atreya Dey, Graduate Student, University of Texas Austin

Atreya Dey uses computational biophysics to explore the packing structures and defects of flexible polymers, including chromosomes.

Recent highlights:

- Predicting the organization of mitotic chromosomes using the generalized Rouse model. *Biophysical Journal*, 2020.
- Toroidal condensates by semiflexible polymer chains: Insights into nucleation, growth and packing defects. *The Journal of Physical Chemistry B*, 2017.

Ken Dill, Professor, Stony Brook University

Ken Dill specializes in theoretical and computational studies of protein homeostasis.

Recent highlights:

- Nanoscale catalyst chemotaxis can drive the assembly of functional pathways. *The Journal of Physical Chemistry B*, 2021.

- Accelerating protein folding molecular dynamics using inter-residue distances from machine learning servers. *Journal of Chemical Theory and Computation*, 2022.

Meghan Driscoll, Postdoctoral Scholar, University of Texas Southwestern
Meghan Driscoll works in cellular imaging and the application of data science to cellular imaging to determine how cellular shape affects migration and cellular dynamics.
Recent highlights:

- Robust and automated detection of subcellular morphological motifs in 3D microscopy images. *Nature Methods*, 2019.
- Spatiotemporal relationships between the cell shape and the actomyosin cortex of periodically protruding cells. *Cytoskeleton*, 2015.

Judith Frydman, Professor, Stanford University
Judith Frydman advances various omics techniques and applies them to protein biogenesis.
Recent highlights:

- Ageing exacerbates ribosome pausing to disrupt cotranslational proteostasis. *Nature*, 2022.
- Differentiation drives widespread rewiring of the neural stem cell chaperone network. *Molecular cell*, 2020.

Arjun Krishnan, Assistant Professor, Michigan State University
Arjun Krishnan is an expert in combining large-scale datasets with machine learning to gain biological insights into a variety of living systems.
Recent highlights:

- A computational framework for genome-wide characterization of the human disease landscape. *Cell Systems*, 2019.
- Robust normalization and transformation techniques for constructing gene coexpression networks from RNA-seq data. *Genome Biology*, 2022.

Liana Lareau, Assistant Professor, University of California Berkeley
Liana Lareau combines next-generation sequencing and machine learning to discover biological insights.
Recent highlights:

- Coverage-dependent bias creates the appearance of binary splicing in single cells. *eLife,* 2020.
- Accurate design of translational output by a neural network model of ribosome distribution. *Nature Structural & Molecular Biology*, 2018.

Travis La Fleur, Graduate Student, Penn State University
Travis La Fleur models gene expression using biophysical modeling and bioengineering data.
Recent highlights:

- Automated model-predictive design of synthetic promoters to control transcriptional profiles in bacteria. *bioRxiv*, 2021.

Steffen Lindert, Associate Professor, Ohio State University
Steffen Lindert predicts protein structures by using various datasets, including MS and protein footprinting data, and simulation.
Recent highlights:

- Accurate protein structure prediction with hydroxyl radical protein footprinting data. *Nature Communications*, 2021.
- Utility of covalent labeling MS data in protein structure prediction with Rosetta. *Journal of Chemical Theory and Computation*, 2019.

Zan Luthey-Schulten, Professor, University of Illinois Urbana-Champaign
Zan Luthey-Schulten incorporates various data sets into the computation of whole-cell modeling.
Recent highlights:

- Generating chromosome geometries in a minimal cell from cryo-electron tomograms and chromosome conformation capture maps. *Frontiers in Molecular Biosciences*, 2021.
- Generalized correlation-based dynamical network analysis: a new high-performance approach for identifying allosteric communications in molecular dynamics trajectories. *Journal of Chemical Physics*, 2020.

Shaun Mahoney, Associate Professor, Penn State University
Shaun Mahoney uses neural networks to make data-driven discoveries in transcription factor binding and gene regulation.
Recent highlights:

- Domain adaptive neural networks improve cross-species prediction of transcription factor binding. *Genome Research*, 2022.
- An interpretable bimodal neural network characterizes the sequence and preexisting chromatin predictors of induced transcription factor binding. *Genome Biology*, 2021.

Serghei Mangul, Assistant Professor, University of Southern California
Serghei Mangul studies gene expression and has analyzed public omics data for reusability in a variety of forms.
Recent highlights:

- Integrating big data computational skills in education to facilitate reproducibility and transparency in pharmaceutical sciences. *Journal of the American College of Clinical Pharmacy*, 2021.
- Improving the usability and archival stability of bioinformatics software. *Genome biology*, 2019.

Susan Marqusee, Professor, University of California Berkeley
Susan Marqusee examines protein folding through biophysical and MS experimentation.
Recent highlights:

- Exploring the evolutionary history of kinetic stability in the α-lytic protease family. *Biochemistry*, 2021.
- Site-specific ubiquitination affects protein energetics and proteasomal degradation. *Nature Chemical Biology*, 2020.

Wallace Marshall, Professor, University of California San Francisco
Wallace Marshall employ at integrated combination of techniques, to understand how cells solve geometric engineering problems.
Recent highlights:

- Towards computer aided design of cellular structure. *Physical Biology*, 2020.
- Multi-scale spatial heterogeneity enhances particle clearance in airway ciliary arrays. *Nature Physics*, 2020.

Kenneth Matreyek, Assistant Professor, Case Western Reserve University
Kenneth Matreyek develops and deploys massively parallel assays to discover protein function and adaptation to viruses.
Recent highlights:

- Integrating thousands of PTEN variant activity and abundance measurements reveals variant subgroups and new dominant negatives in cancers. *Genome Medicine*, 2021.

- An improved platform for functional assessment of large protein libraries in mammalian cells. *Nucleic Acids Research*, 2019.

Matteo Mori, Postdoc, University of California San Diego
Matteo Mori uses multi-level models and the analysis of omics data to connect gene expression and cellular phenotypes.
Recent highlights:

- Disruption of transcription-translation coordination in Escherichia coli leads to premature transcriptional termination. *Nature Microbiology*, 2019.
- On the optimality of the enzyme-substrate relationship in bacteria. *PLOS Biology*, 2021.

Ed O’Brien, Associate Professor, Penn State University
Ed O’Brien integrates bioinformatics, chemistry, and simulations to better understand the translation process.
Recent highlights:

- Combinations of slow-translating codon clusters can increase mRNA half-life in Saccharomyces cerevisiae. *PNAS*, 2021.
- Ribosome elongation kinetics of consecutively charged residues are coupled to electrostatic force. *Biochemistry*, 2021.

Kim Reynolds, Associate Professor, University of Texas Southwestern
Kim Reynolds creates statistical models that can explain, predict, and design cellular behaviors in the lab.
Recent highlights:

- Structurally distributed surface sites tune allosteric regulation. *Elife*, 2021.
- A simplified strategy for titrating gene expression reveals new relationships between genotype, environment, and bacterial growth. *Nucleic Acids Research*, 2020.

Gabriel Rocklin, Assistant Professor, Northwestern University
Gabriel Rocklin uses molecular modeling and experimental techniques to study protein biophysics.
Recent highlights:

- Dissecting the stability determinants of a challenging de novo protein fold using massively parallel design and experimentation. *BioRxiv*, 2021.
- Global analysis of protein folding using massively parallel design, synthesis and testing. *Science*, 2017.

Andrej Sali, Professor, University of California San Francisco
Andrej Sali develops computational tools that integrate diverse experimental data to study the structure and dynamics of assemblies.
Recent highlights:

- Integration of software tools for integrative modeling of biomolecular systems. *Journal of Structural Biology*, 2022.
- Bayesian metamodeling of complex biological systems across varying representations. *PNAS*, 2021.

Howard Salis, Associate Professor, Penn State University
Howard Salis combines biophysical and kinetic models with next-generation sequencing data of gene expression.
Recent highlights:

- Systematic quantification of sequence and structural determinants controlling mRNA stability in bacterial operons. *ACS Synthetic Biology*, 2021.
- The synthesis success calculator: predicting the rapid synthesis of DNA fragments with machine learning. *ACS Synthetic Biology*, 2020.

Premal Shah, Assistant Professor, Rutgers University

Premal Shah combines bioinformatics and chemical kinetic modeling to study transcription and translation.

Recent highlights:

- Promoter-sequence determinants and structural basis of primer-dependent transcription initiation in Escherichia coli. *PNAS*, 2021.
- XACT-seq comprehensively defines the promoter-position and promoter-sequence determinants for initial-transcription pausing. *Molecular Cell*, 2020.

Eugene Shakhnovich, Professor, Harvard University

Eugene Shakhnovich combines molecular biophysics theory and data with cellular properties on a broad range of topics.

Recent highlights:

- Effect of RNA on morphology and dynamics of membraneless organelles. *The Journal of Physical Chemistry B*, 2021.
- Validation of DBFOLD: an efficient algorithm for computing folding pathways of complex proteins. *PLOS Computational Biology*, 2020.

Ian Sitarik, Graduate Student, Penn State University

Ian Sitarik uses a combination of computational protein models and experiments to understand the fundamental principles driving protein folding and function.

Recent highlights:

- Subpopulations of soluble, misfolded proteins commonly bypass chaperones: how it happens at the molecular level. *BioRxiv*, 2021.
- Universal protein misfolding intermediates can bypass the proteostasis network and remain soluble and non-functional. *BioRxiv*, 2021.

Dave Thirumalai, Professor, University of Texas Austin

Dave Thirumalai performs theoretical and computational synthesis of diverse molecular biology processes and integrates Hi-C data.

Recent highlights:

- Multiscale coarse-grained model for the stepping of molecular motors with application to kinesin. *Journal of Chemical Theory and Computation*, 2021.
- Iterative annealing mechanism explains the functions of the GroEL and RNA chaperones. *Protein Science*, 2020.

Bin Zhang, Associate Professor, Massachusetts Institute of Technology

Bin Zhang uses biophysical modeling techniques to study chromatin structure in atomistic detail.

Recent highlights:

- Multiscale modeling of genome organization with maximum entropy optimization. *Journal of Chemical Physics*, 2021.
- Data-driven polymer model for mechanistic exploration of diploid genome organization. *Biophysical Journal*, 2020.

9. Appendix 2: Workshop schedule

Saturday, March 12, 2022
11 am to 4 pm EST

- **11 am:** Welcome and Overview
- **11:30 am:** Breakout 1–Identifying Grand Challenges
- **12:25 pm:** Breakout 2–Identifying Grand Challenges
- **1:15 pm:** Voting on Most Important Challenges
- **1:30 pm:** Lunch Break
- **2:00 pm:** Breakout 3–Diving Deeper - Developing Grand Challenges
- **3:20 pm:** Next Steps and Closing

Sunday, March 13, 2022
11 am to 4 pm EST

- **11 am:** Welcome and Overview
- **11:20 am:** Breakout 1–Identifying Barriers to Synthesis
- **12:15 pm:** Breakout 2–Identifying Solutions to Enable Synthesis
- **1:05 pm:** Lunch
- **1:35 pm:** Breakout 3–Diversity, Training, and Broader Impacts
- **2:10 pm:** Full Report Backs with Feedback
- **3:05 pm:** Breakout 4–Incorporating Feedback
- **3:40 pm:** Next Steps and Closing