Perspective

Data Science: Accelerating Innovation and Discovery in Chemical Engineering

David A. C. Beck
Department of Chemical Engineering, University of Washington, Seattle, WA

eScience Institute, University of Washington, Seattle, WA

James M. Carothers, Venkat R. Subramanian, and Jim Pfaendtner
Department of Chemical Engineering, University of Washington, Seattle, WA

DOI 10.1002/aic.15192
Published online February 28, 2016 in Wiley Online Library (wileyonlinelibrary.com)

Keywords: data science, machine learning, molecular science, biotechnology, energy systems

Introduction: What Is Data Science and Why Should Chemical Engineers Care About It?

All of science and engineering, including chemical engineering, is being transformed by new sources of data from high-throughput experiments, observational studies, and simulation. In this new era of data-enabled science and engineering, discovery is no longer limited by the collection and processing of data but by data management, knowledge extraction, and the visualization of information. The term *data science* has become increasingly popular across industry, and academic disciplines to refer to the combination of strategies and tools for addressing the oncoming deluge of data. The term *data scientist* is a common descriptor of an engineer or scientist from any disciplinary background who is equipped to seamlessly process, analyze, and communicate in this data-intensive context. The core areas of data science are often identified as data management, statistical and machine learning, and visualization. In this Perspective, we present an overview of these core areas, discuss application areas from within chemical engineering research, and conclude with perspectives on how data science principles can be included in our training.

As has been noted for several years,[1] chemical engineers of all varieties, from the practicing process engineer to the academic researcher, are being asked more and more often to manipulate, transform, and analyze complex data sets. The complexity often stems from the size of the data set itself, but this is not the only factor. For example, the stream of information available to an engineer in a modern plant is tremendous because of the proliferation of inexpensive instrumentation and the nearly ubiquitous high bandwidth and low-latency connectivity. In the area of research and discovery, a student or researcher conducting data-intensive experiments, such as high-resolution particle tracking, might generate more data in an afternoon than a student from a previous decade in the entire time spent earning his or her Ph.D. For those conducting mathematical modeling and computer simulations, advanced algorithms and hardware now give simulators unprecedented resolution but at the cost of massive increases in the data set. Underlying all of these examples is the cheap (near free) cost of data storage and the ubiquitous availability of our data from cloud-based services.

The aforementioned examples may appear to be vastly different from the outset. However, common themes in the limitation of our current approaches quickly emerge. Because our training of new chemical engineers (at all levels) has not kept pace with the explosion of data, each chemical engineer in the previous examples will likely approach her or his work in the same manner: manual searching for relationships in the data, classical visualization of monovariate or bivariate correlations in features, and a hypothesis-driven approach to science reminiscent of a data-poor era when the researcher or engineer could essentially manipulate relevant data in their mind. Simply put, without knowledge about and training on how to handle data skillfully, most of the information from our plants and refineries, our data-intensive experiments, and our computer simulations is thrown away, simply because we do not know how to extract knowledge from it. Fortunately, there is a potential solution on the horizon. Through the lens of the nascent field of data science, we can see an emergent (and limited) set of tasks needed by all of the previous chemical engineers: (1) to manage a huge data set consisting of ensembles of spatiotemporal data, (2) to sensibly read the data in a computationally scalable manner, and (3) to extract knowledge from this pile of information with robust techniques whose statistical reliability can be quantified. It also goes without saying that data science itself is not a panacea. Chemical engineering fundamentals are of the utmost importance, and

Correspondence concerning this article should be addressed to D. A. C. Beck at dacb@uw.edu and J. Pfaendtner at jpfaendt@uw.edu.

no amount of processing will allow someone to extract meaning from data that was collected incorrectly or has no useful informational content.

Chemical engineers have always been quick to adopt new methods and techniques for their toolbox. Indeed, because of the excellent math and computer skills that many chemical engineers possess, some of the methods and tools we discuss have been in use for some time in the chemical engineering community, or at minimum, for many chemical engineers, the adoption of these methods should prove to be straightforward. Additionally, however, the aforementioned challenges now extend to people who traditionally do not consider their work to be dependent on their computing skills and capacity. It is thus our hope that this article will provide a convenient framework and provide a common language to help guide not only those seeking to use these methods for the first time but also seasoned veterans who are expanding their capabilities.

Statistical, Machine Learning, and Visualization Tools Available to Chemical Engineers

In this section, we provide an overview of the keystone concepts of data science. A commonly used paradigm, which we adopt here, is the division of data science into three broad categories: data management, machine learning, and visualization. We stress that this section is not an exhaustive description of all possible methodologies in these topic areas. Instead, it is meant as a survey of many current, commonly used methods with short, informative but not overly technical descriptions and suggestions for further reading. Our goal is to help readers identify potential tools for addressing data-related challenges in their research and understanding the keywords (in boldface) in their application.

Data management

Data science begins with data. In **synthetic biology,** examples of data include genomics, transcriptomics, proteomics, and metabolomics, where datasets are often comparatively abundant. For **molecular and nanoscale phenomena,** data commonly take the form of trajectories or large ensembles of information. These data could be generated computer simulations or collected from data-intensive experiments, such as those from high-resolution/high-speed microscopy. At the **process and systems** level, data are often complex, interwoven time series, for example, complementary sensory and diagnostic information from a power device or chemical reactor process.

All of these are typified by the requirement to store, manage, integrate, and access data collected from one or multiple experiments or by streaming in directly from instruments. How we choose to organize, store, and manage our data significantly impacts the performance of downstream analyses, ease of sharing, and visualization. Since the early 1980s, spreadsheets have been an entry-level tool for handling two-dimensional data, that is, rows and columns. Their utility begins to break down, however, when data begin to get into the tens of thousands of rows or are of a higher dimensionality or when we require complex transformations of data or the integration of data across a large number of data sets. Moreover, spreadsheets do not have the flexibility of subsetting data or rapid programmatic access from external tools for machine learning or visualization.

Thus, the next step for most teams, as they begin to work with more complex data with increased analytical demands, is **relational databases,** which were first described in 1970.[2] Relational database management systems (RDBMs) implement variations of the Structured Query Language (**SQL;** commonly pronounced like *sequel*), which enables users to describe data in terms of two-dimensional **tables** and relationships between them, for example, one-to-one, one-to-many, and many-to-many. A molecule may have many atoms, a colloid may have many molecules, and so on. SQL enables us to ask complex **queries** across the relations, for example, to find all of the molecules with a specific moiety or all of the molecules above a given molecular weight. Tables can be indexed for speed of access at the expense of the storage footprint. As such, RDBMS and SQL provide powerful languages for data manipulation and extraction in complex datasets with high-dimensional data sets. Free and open-source RDBMs include Postgress SQL and MySQL.

The way we store data has increased steadily in sophistication and usability; this, in turn, has led to new challenges with the way we perform calculations on large-data sets. Computing with these data sets has begun to exceed the capabilities of a single desktop machines or single high-performance computers with tens of cores and lots of memory. Instead, new strategies for computing have been developed that leverage tightly coupled computers that share high-bandwidth, low-latency networks and very fast networked file systems, that is, shared storage space. Although these clusters have proven useful, particularly for tasks such as molecular simulation, where cluster nodes frequently exchange large amounts of data, they remain quite expensive. There has been a realization that a class of computational problems do not require tight coupling between computers working on the same problem and can make use a collection of loosely coupled, relatively inexpensive computers with standard network connectivity. Perhaps the quintessential example of this has been the MapReduce[3] algorithm developed at Google for processing large amounts of data, including Web pages. In the MapReduce model, an input data set is partitioned between commodity computers that work independently on a map step, which involves some transformation of the input data, and a reduce step, which aggregates the data from across all worker nodes. Multiple MapReduce iterations can be combined to perform arbitrarily complex tasks. The bedrock, open-source implementation of MapReduce is the Apache Hadoop,[4] and it has a rich software ecosystem to support it. The strengths of Hadoop lie in its ability to coordinate the execution, in a fault-tolerant manner, of lots of independent computers.

Environments such as Hadoop have been particularly useful when they have been used on the loosely coupled, pay-as-you-go computing platforms of cloud computing. The most popular examples of this are Amazon Web Services, which includes their Elastic Compute Cloud for computing and Simple Storage Service for storing data, and Microsoft's Azure. Both could be considered infrastructure as a service, where you pay for using the computational infrastructure, but what you run on top of that infrastructure is your responsibility. Data-intensive management and processing is not the only use

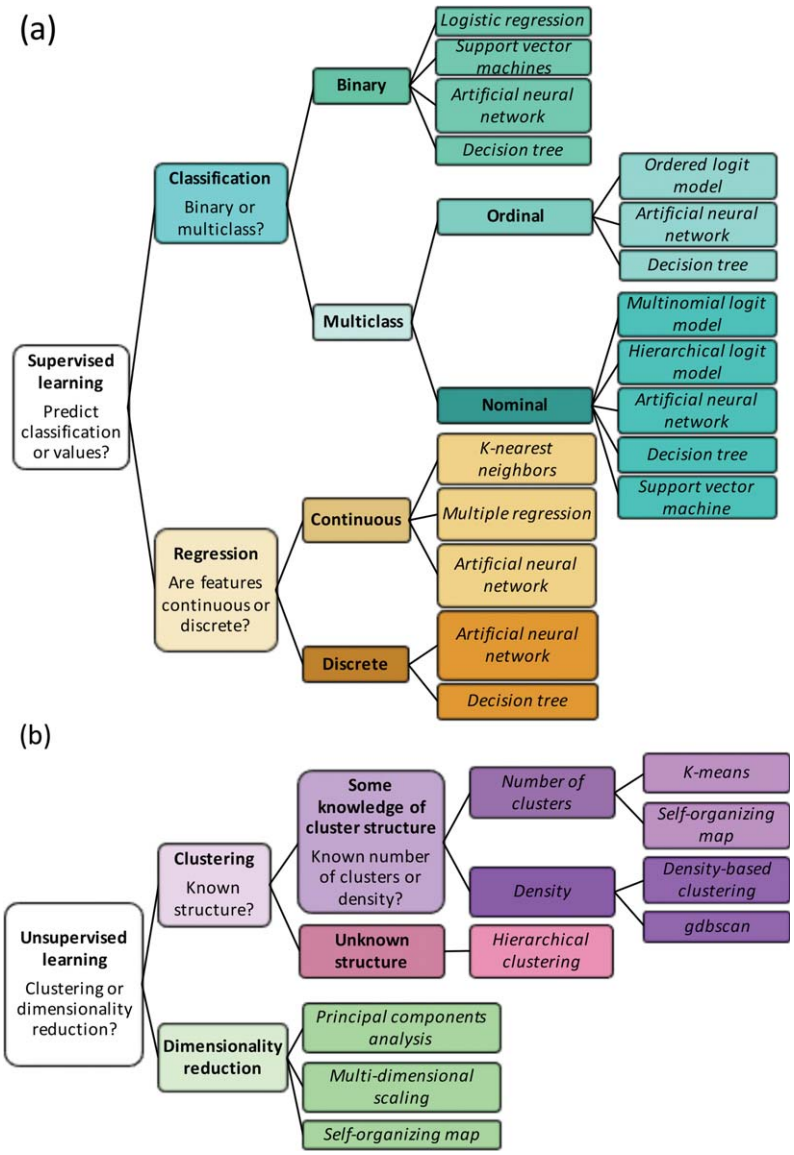

Figure 1. Simplified guide to statistical and machine learning choices. The two broad types of statistical and machine learning, (a) supervised and (b) unsupervised, are broken down into a simplified decision tree. An overview of the broad types (i.e., the first two levels of classification) and definitions of the key terms are in the main text.

for cloud computing for scientists and engineers. For loosely coupled computing tasks, where user demand is not constant, cloud computing can offer significant savings over the purchase and maintenance of computer clusters, particularly as the cost of electricity, cooling, and space grows. Thus, cloud computing is emerging as a complementary tool for traditional high-performance computing for certain classes of problems.

Statistical and machine learning

Roughly, statistical and machine learning is broken down into two types: supervised and unsupervised. In *supervised learning* (Figure 1a), the task is to define a model that can be used to accurately predict an output or outputs from a set of inputs. The inputs are described in terms of features or predictor variables, and the outputs are described as labels or response variables. For example, in a simplified quantitative structure-activity relationship[5] model for predicting boiling points, a single feature/predictor for each input molecule might be the number of pairs of atomic bonds that do not share an atom, and the response variable is the boiling point.[6] Most learning tasks have multiple features or dimensions describing the feature space with the values for a given sample described by a point in this space, which is referred to as the *feature vector*. Feature spaces can be continuous or discrete (e.g., yes or no; low, medium, or high).

In the previous example, the output was a real valued quantity, and this maps well to many problems involving the prediction of continuous outputs, such as binding affinities, reaction rates, and probabilities. This process is generally referred to as *regression* and is not confined to simple linear or nonlinear regression, with which every chemical engineer is familiar. However, we are not limited to continuous outputs. Often, the task is one of classification, where the intent is to place samples into an appropriate bin or to assign a correct

label according to the samples' features. The simplest form of this is binary classification, where there are only two possible classes, such as in the case where want to predict whether a material will self-assemble or not (i.e., yes or no). A related case is the multiclass problem, where an array of possible classifications is available and the goal is to identify the best classification on the basis of a sample's features, such as whether a given molecular configuration is in the reactant, transition, or product state of a free-energy landscape.

A key component of supervised learning is the presence of a training set or corpus from which to build the predictive model. Typically, during training, after each example is passed through the model, a cost or loss function is computed; this function evaluates the model prediction against the known output of the training example. This loss function then defines the error or quality of the prediction. The objective of a learning task is then to minimize the loss function. When training a model on a training set, it is important to reserve some of the training set for the evaluation of the model; this is often referred to as the *test set*. That is, a portion of the training set should not be used to train the model but should instead be kept back and used only after the model is fixed to estimate the accuracy of the model.

Three pitfalls potentially limit both the success of machine learning implementation and the final accuracy. Noise can be a problem in the training data and can lead to poor classification accuracy or response predictions. Generally, one can view noise in training data as arising from two sources: the features/predictors or the labels/responses. In the case of underfitting, we chose a model of insufficient complexity. In a trivial example of this, take our self-assembly binary classification problem again. Conversely, if a model is overfitting, components of noise are included, and it will not generalize well for new samples. In this case, predictive accuracy can be very high for the training set but poor for the evaluation examples or new samples.

In contrast to supervised learning, where the goal is to predict a label or response variable from an input set of features or predictor variables, a separate class of problems exists around the identification of the hidden structure in data on the basis of some feature set. This is generally known as *unsupervised learning* (Figure 1b), where no labels are attributed to the training set and the goal is to directly infer the relationships between samples. This challenge substantially differs from supervised learning, as there is no cost or loss function to indicate the quality of a model. Instead, each method has specialized metrics that need to be considered carefully.

One of the most familiar unsupervised learning approaches is clustering. In this paradigm, unlabeled data are grouped by some measure of similarity into clusters, which link related samples such that some sort of labeling can be inferred. Two factors describe a clustering strategy: the nature of the similarity or distance metric and the clustering algorithm. The most well-known distance metric is the Euclidean distance, but there are a variety of others, including the Manhattan distance, Pearson correlation, and Spearman correlation. The choice of a distance or similarity metric is predicated on the nature of the data, its underlying distribution, and the potential for noise and error. Common clustering algorithms include centroid methods such as k means,[7] density-based methods,[8,9] self-organizing maps,[10] and hierarchical clustering.[11] As with the choice of a distance or similarity metric, each of these methods has strengths that depend on the nature of the data and *a priori* knowledge of the behavior of the system to be modeled.

Suggested reading

The area of statistical and machine learning is a rapidly evolving discipline, with new algorithms being developed continuously. We can, however, recommend two texts that survey the most common methods in both supervised and unsupervised learning. The first, *An Introduction to Statistical Learning,*[12] is readily approachable, and includes some thoughts on deciding between algorithm choices, and provides example code. The second, in some sense an extension of the first, *Elements of Statistical Learning,*[13] is more comprehensive and positioned for the mathematical and statistical oriented researcher; it has superb detail about the range of methods.

Visualization

Once the data have been cleaned, structured, and integrated in a data management system, users need the ability to explore them. Thus, the third and final keystone component of data science is the visualization and interactive exploration of the data and quantities computed from it. It is important to note that the role of visualization is not to replace the statistical algorithms just described but rather to use them to support decision making and knowledge by users.

Vision is our most powerful communication vehicle. Anyone who has tried to find the pattern in a table of numerical values versus a heat map of the same data is readily familiar with this. However, the capacity of our vision is limited to two or three dimensions at a time. High-dimensional feature spaces make visualization difficult, if not impossible. To combat this, we can turn to dimensionality-reduction techniques. However, if the goal is to simply reduce the dimensionality in continuous spaces without the need for classification, simpler forms of dimensionality reduction exist; these include principal component analysis[14] (PCA) and multidimensional scaling.[15] Both methods approximate the distance of data in a high-dimensional feature space in a lower dimensional space, for example, two dimensions. When one uses these methods for dimensionality reduction, it is important to examine the measure of fit. For PCA, this is the cumulative amount of variance in the feature data explained by the number of dimensions chosen for the reduction, and in multidimensional scaling, it is the strain of the fit from the high-dimensional feature space into the lower dimensional visualization space.

The work of Edward Tufte[16–19] has been influential in the visualization and presentation of data. Visualization tools, such as Data-Driven Documents,[20] have simplified the creation of stunningly beautiful visuals from traditional scatterplots and bar charts to nontraditional chart types, such as chord diagrams (e.g., Circos[21]); graphs; network diagrams (e.g., Cytoscape[22]); tree maps that are used to describe hierarchical partitioning; Sankey diagrams, which can describe the flow through a system (see Sankey's 1898 model of a steam engine); and interactive sunburst charts,[23] which are hierarchical pie charts.

Areas of strength and opportunities in chemical engineering research

In this section, we highlight a few selected research areas within the chemical engineering field where there are growing uses of data science as illustrative examples of how these methods may be applied to our field. Of course, this list is not exhaustive nor could an article such as this even begin to fully do justice to (1) the potential impacts data science will make in the future or (2) areas of excellence in chemical engineering research where these tools are already being used. As for the former, we believe complex data intensive work is here to stay, and we hope many readers will be inspired and motivated by these examples. As for the latter, interested readers are referred to the rich literature on complex systems[24,25] and process systems engineering[26] for representative examples.

Computational Molecular Science and Engineering. Examples of the effective use of data science methods in the area of molecular science and engineering can already be found in the literature. Although many current applications in this area are predominantly found in the domain of computation and theory, we expect that data intensive wet-lab experimentation in molecular science will soon feel the impacts of these useful methods. In this section, we briefly review four subareas of (computational) molecular science and engineering with a broad survey of research progress enabled by various data science methods. The literature is rich with many additional examples (with the numbers growing weekly), and our goal in this subsection is, in particular, to provide a broad overview of different types of data science approaches that can be applied to molecular science.

Making* ab initio *calculations faster with neural-network potentials. There are countless research groups making routine use of quantum chemical (*ab initio*) methods to study the electronic and energetic properties of a huge range of systems. As a general rule, first-principles methods offer the promise of quantitative predictive power at a high level of accuracy but at the penalty of great computational cost. Therefore, the idea has emerged to use supervised learning techniques [viz., artificial neural networks (ANN)] to create efficient potentials that represent the complex quantum mechanical potential energy surface (PES). Proof of concept with an empirical potential was provided 20 years ago by Blank et al.,[27] who demonstrated that a simple ANN for a PES could be faithfully used to carry out more complex analyses (e.g., transition-state theory calculations). Around a decade ago, Lorenz et al.[28] demonstrated how a real quantum chemical potential for a heterogeneous system could be obtained. A short while later, Behler and Parrinello[29] offered a solution to address the difficulties in applying ANNs to high-dimensional quantum chemical potentials for systems such as bulk silicon. The structure of the ANN was reformulated on an atom-centered basis, inspired by the topology of molecules and materials themselves; this ultimately conferred much more transferability and generalizability to the approach. As an illustrative example, an ANN model was developed for silicon; this led to a speedup of five orders of magnitude in energy calculations with only minimal accuracy losses. Many further examples, including ANN potentials for challenging systems such as sodium[30] and graphite,[31] have been created with this approach. A general software tool for creating ANN potentials from *ab initio* calculations has been released,[32] and this can significantly increase the usability of the approach. Finally, the application of data science tools to *ab initio* calculations is not limited to ANN representations of the PES. This is illustrated in recent examples demonstrating on-the-fly machine learning of quantum mechanical forces to facilitate molecular dynamics (MD),[33] the prediction of atomization energies,[34] and even applications of the discovery of the underlying density functional with machine learning.[35]

Discovering the properties of molecules and macromolecules. Moving up in scale from the electronic/atomistic domain, it is possible to find many applications of data science methods to the study of molecular, biomolecular, and macromolecular systems. A common use of supervised learning algorithms is the so-called quantitative structure-property relations, which relate calculated or measured properties (usually performance metrics) to underlying features in the molecular structure. Arguably, these relations are best known for their applications in protein/ligand screening.[36] However, recent examples in the area of predicting the properties of ionic liquids[37] and natural products[38] have demonstrated potential uses far beyond the pharma and drug industries. Machine learning has found great use in the design of organic electronics.[39,40] A common motif is to use a large training set of *ab initio* calculations to predict many properties and then to use a supersized learning technique, such as ridge kernel regression, to optimize a target molecular feature to optimize a particular desired set of features. This is a general approach in the spirit of the recent Materials Genome Initiative (MGI) to accelerate the speed of discovery of new useful materials and to facilitate more efficient and effective use of resources devoted to the synthesis and characterization of new molecules and materials. As the number and extent of these examples grow, we expect that these methods will become commonplace, and it will become even easier for other researchers to adopt these approaches.

The area of molecular/biomolecular science also contains many examples of the effective use of unsupervised learning approaches. The most common example is the use of dimensionality-reduction techniques, such as PCA or related methods. The diffusion map approach[41] is a recent example that has been applied to the discovery of collective descriptors of chain dynamics. In the area of nonequilibrium MD simulations, a self-learning algorithm, reconnaissance metadynamics,[42] was introduced to facilitate discovery of slow-coarse degrees of freedom in a general way. These approaches are important for the general class of problems in which high-dimensional systems (e.g., an all-atom MD calculation) need to be represented by a few (nonobvious) slow degrees of freedom. Finally, the area of data management and visualization is also one that has seen significant progress by researchers in this field. The Dynameomics project,[43] among other advances, has led to the creation of a massive infrastructure for data management and processing in molecular simulations.[44,45] The SketchMap framework is another example of a complex data visualization method that can be applied to advanced MD calcuations.[46] The illustration of a complex visualization of a molecular simulation via SketchMap[47] is shown in Figure 2.

Applications of data science in materials science. The aforementioned MGI (https://www.mgi.gov), launched by the

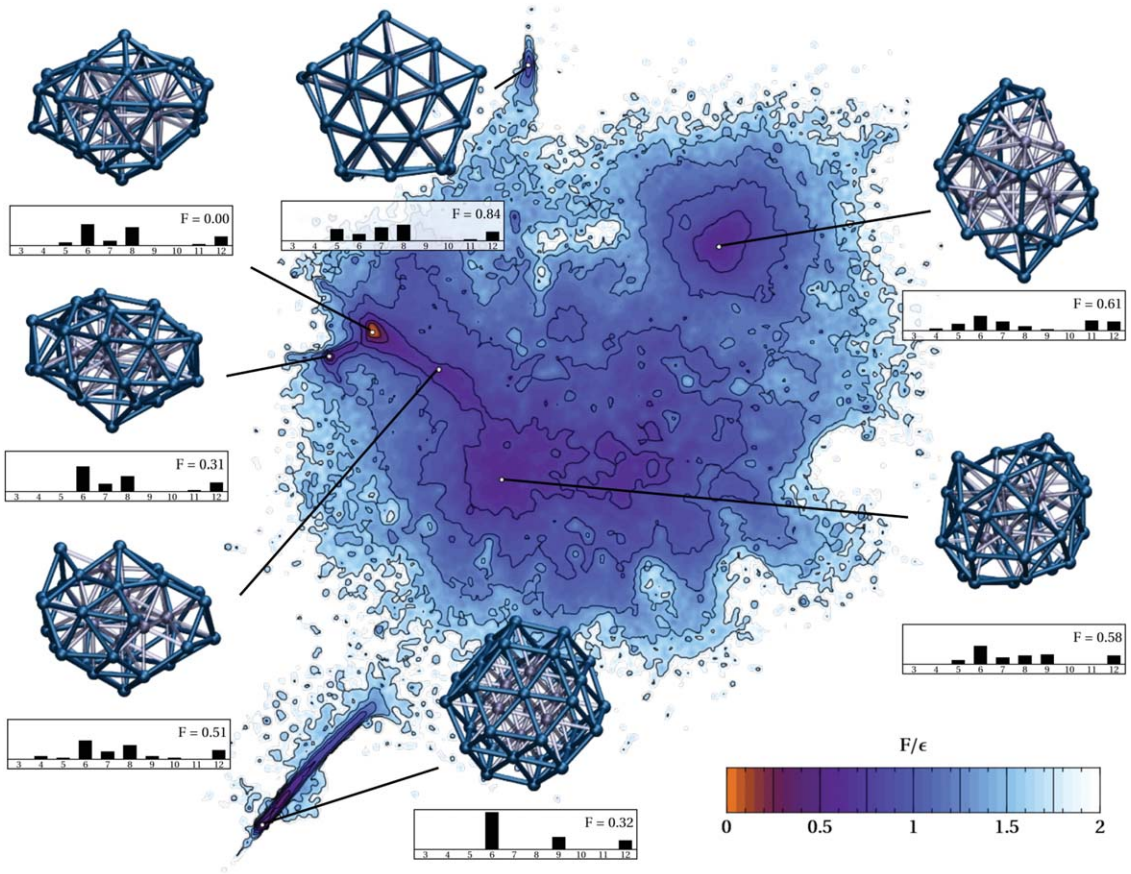

Figure 2. Visual analysis of the free-energy landscape of clusters of 38 Lennard-Jones spheres at the system's melting temperature. The free-energy is directly projected onto the SketchMap coordinates. The histogram insets relate relative to populations of different configurations. Reprinted with permissions from ref. 47. Copyright American Chemical Society.

U.S. government in 2011[48] as a multiagency initiative to accelerate the speed of discovery of new materials, has led to significant new applications of data science in all areas of the materials field.[49] Applications of supervised learning have been deployed to make unprecedented use of massive data sets coming from high-performance computing. For example, the Open Quantum Materials Database (https://oqmd.org) from the Wolverton research group[50] allows a wide range of tasks from searching existing materials to hypothesizing phase diagrams for new materials. In reflection of modern computer usage, the database can be directly queried through Twitter. The Materials Project[51] is a similarly inspired materials genome approach, which is devoted to data mining from large-scale quantum chemical calculations. It contains, at this writing, over 65,000 materials in a searchable database form. The Harvard Clean Energy Project,[52] which uses distributed worldwide computing, is another example and is specifically devoted to the discovery of next-generation photovoltaics. These examples, in particular, demonstrate the power of the effective use of the large data sets resulting from massive parallel computing that are rapidly becoming commonplace. In the area of the discovery of porous materials, significant discovery has been achieved through automated screen approaches[53] inspired by many of the previous methods, in particular, in the application of the discovery and design of zeolites and metal-organic frameworks.[54–57]

Predicting the chemical reactions of molecules and materials. In the area of complex reaction networks,[58] there have been fewer applications of advanced data science approaches. This is in contrast to the growing number of examples (several highlighted previously) of the use of data science methods to predict the reactivity of specific reactions. The connection between a series of chemical reactions and a neural network has long been noted,[59] albeit from the perspective of the design of new methods to carry out calculations. Examples of the prediction of the time-dependent behavior of combustion systems with ANNs have been demonstrated.[60,61] A general approach for discovering chemical reaction mechanisms with ANNs has also been proposed.[62] Finally, in the area of physics-based (*ab initio* MD) simulations, a generic, nonequilibrium approach based on the metadynamics framework and inspired through spectral graph theory was proposed to predict chemical reaction networks (CRNs);[63] it was further demonstrated by the discovery of reaction mechanisms applied to methanol combustion.[64]

Synthetic Biology. Model-driven synthetic biology is beginning to reduce the time and resources needed to engineer biological systems for applications in materials, chemicals,

energy, the environment, and health.[65–69] Computational approaches for synthetic biology are becoming much more effective with the advent of high-performance clusters and data science methods for rapidly simulating and analyzing very large numbers of potential system designs. Design rules for the generation of new components and system specifications are being identified through large-scale experimentation, data mining, and machine learning. Finally, advancements in visualization and cloud-based data management are facilitating the distribution of functional designs and streamlining the engineering process. In this subsection, we draw on literature examples and the work of one of the authors (J.M.C.) to show how data-enabled methods are propelling a shift toward the creation of models and simulation tools for predictable biological components and system engineering.

Large-scale design space mapping for complex system engineering. Traditionally, it has been difficult to create effective models for engineering cellular systems because the underlying design spaces are vast, there are a large number of components, and the relevant parameter values are largely unknown and frequently changing.[68,70] However, the challenge of engineering complex components and systems has become more tractable through large-scale simulation analysis. For instance, coarse-grained kinetic model construction combined with large sampling-based methods have enabled the identification of metabolic control circuit designs to significantly improve production titers[71] and has allowed automated metabolic network reconstruction.[72] Genome-wide simulation analysis has even been used successfully to predict growth phenotypes in model organisms.[73] Large-scale simulation and advanced statistical analysis can be used to search design spaces and identify component specifications for meeting target performance criteria. By reducing the need for trial-and-error experimentation, we can gain access to biological device and system designs with complexities not otherwise achievable.[68]

We previously developed model-driven approaches for engineering RNA-based genetic control devices to quantitatively program pathway and circuit gene expression. Coarse-grained mechanistic models[74] were first formulated as CRNs[75] for simulating genetic outputs in the system. To map the high-dimensional design space, sets of Ordinary Differential Equation(s) (ODEs) corresponding to the CRNs were solved with parameter values drawn from biochemically plausible ranges. Global sensitivity analysis[76] provided statistical tools for building up an understanding of the fitness landscape, even though these were necessarily sparse samplings of the design space. Monte Carlo filtering[77] was used to cluster the simulations as behavioral or nonbehavioral according to the output values; this generated the component specification ranges expected to give desired activities. The 94% correlation between the predicted and experimentally measured gene expression levels for 25 different engineered RNA devices validated the models and simulation analysis-based approach.

Engineering efforts aimed at harnessing cellular metabolism for the production of chemical and materials introduce stresses that the host cell may not have evolved to easily accommodate.[78] Engineered metabolic pathways have been successfully constructed with dynamic regulatory controllers that increase production titers by minimizing the buildup of toxic pathway intermediates[71] and enzymes and by balancing the supply and demand for cellular resources.[79] Large-scale computational simulation can inform the design and testing of engineered control circuitry. Sampling-based approaches were used to map the space of potential designs for dynamic sensor-regulator systems to produce fatty-acid-based chemicals and biodiesel molecules in *Escherichia coli*.[71] Global sensitivity analysis indicated that biodiesel production titers could be improved with dynamic pathway control across a broad range of design parameters compared to systems comprised only of static controllers. Consistent with model-derived predictions, the strains with engineered dynamic control system gave three times more fatty acid ethyl ester biodiesel equivalents and reached 28% of the theoretical maximum.

In principle, almost any biological control problem can be solved by dynamic systems that convert cues about the internal and external environments into programmable outputs that facilitate resource load balancing in changing conditions.[72,77,80] Recent work has shown that computational simulation and global sensitivity analysis can be used to uncover successful control architectures even within enormous biochemical design spaces.[81] Through the automation of the process of coarse-grained mechanistic model generation for 728 unique control architectures, the production of an industrial aromatic from a 15-gene engineered pathway was simulated for 3×10^6 distinct biochemical implementations. With Latin-hypercube sampling, clustering algorithms, and methods for statistical analysis under conditions of parameter uncertainty (e.g., bootstrapping), experimentally tractable design specifications were identified to solve pathway control problems and enable greater than ninefold increases in production (20% of the practical maximum, see Figure 3).

As efforts to create more full-fledged computer-aided design platforms[70] continue, large-scale design space mapping will play an increasingly important role. The development of next-generation approaches that simulate behaviors across multiple time scales and levels of complexity[82] (i.e., pathways, networks, and cells) will be important for further advancements. Ultimately, data-enabled design strategies could help realize the construction of complex, multilayered information processing and control systems to program the cell state across the levels of components, pathways, and networks for a wide range of applications.

Discovering biological design rules through data mining. Pre-existing biological components and systems have been subjected to unknown evolutionary trajectories and contain embedded functions that are difficult to discern. For genetic expression components, context-dependent differences in the need to maintain flexibility in the type, time scale, or sensitivity of the response may confound efforts to derive meaningful design rules that could be used to engineer new components and systems. Further complicating the analyses of pre-existing metabolic networks is the fact that the topologies, or connections between the sensed molecules and the actuation functions, are not completely known.[83] Dramatic reductions in the cost of massively scaled DNA synthesis and sequencing are resulting in new opportunities for fabricating libraries of components and systems and then applying statistical analysis to multiplexed experimental data to uncover engineering design rules.[84] In this way, forward engineering approaches are

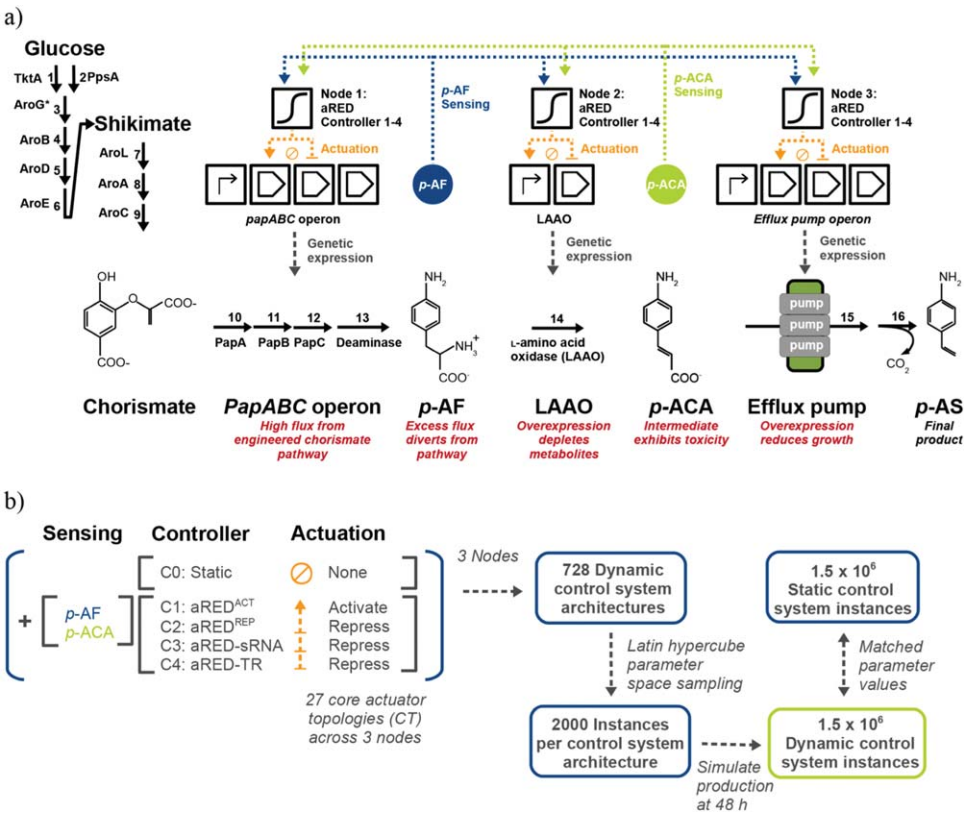

Figure 3. Engineered industrial aromatic [p-aminostyrene (p-AS)] production pathway and control system design space. (A) In principle, pathway control problems (red) can be solved by with the implementation of a dynamic genetic control system; gene product names (in bold) are indicated next to numbered pathway steps, the blue and green dashed lines indicate sensing functions, carried out by aptazyme-Regulated Expression Devices (aREDs) engineered to be responsive to metabolites, and orange dashed lines indicate actuation functions. (B) Massively scaled kinetic model simulation from 3 × 106 implementations and global sensitivity analysis identified experimentally tractable design specifications expected to solve pathway control problems and improve production in the system by more than ninefold. Here, four types of dynamic genetic control mechanisms were investigated: aREDACT (aRED-activating) and aREDREP (aRED-repressing) for control through transcript turnover, aRED-small RNA (aRED-sRNA) for translational control, and aRED-Transcriptional Repression (TR) for direct transcriptional control. Reprinted with permissions from ref. 95. Copyright American Chemical Society.

beginning to render otherwise intractable questions about the relationships between sequences and functions much more accessible to experiments.

Design rule discovery can be achieved through the iterative addition of variables to models, which are then evaluated for accuracy and fit to sets of data obtained through functional library screening. For example, rules for describing important aspects of genetic expression in bacteria (transcriptional termination) have been uncovered by the integration of biophysical modeling and cross-validation, where model accuracy was scored by the calculation of an average coefficient of correlation.[85] Progress toward the automated design of control elements for protein translation has been made through the integration of thermodynamic models with sequence library screening.[86] Through the combination of diversity-oriented sequence design with experimental filtering (*sort-seq*), mechanistic understanding of how naturally occurring gene sequences have evolved to modulate protein expression have even been obtained.[84]

Looking ahead, there is great potential for unsupervised learning approaches, such as PCA and dimensionality reduction, to further improve the process of design rule and model discovery. In this respect, recent work that measured splicing patterns from more than 2 million synthetic minigenes and then used machine learning to train models was particularly promising. In this case, additive effects from nearby sequence elements could be identified, and models trained only on synthetic data could, nonetheless, be applied to naturally occurring biological systems as well.[87]

Data management and visualization in synthetic biology. The formal separation of functional design from physical implementation has been a key advancement for building complex systems in a number of engineering disciplines, including electronic circuit design.[86] For biological components and systems, this will be essential if we are to realize the goal of assembling large, complex, fully functioning systems from separately generated and characterized components.[6] With increasing systems complexity comes a need for data structures and visualization techniques to support computational design and meaningful human interaction with those data.[87–89]

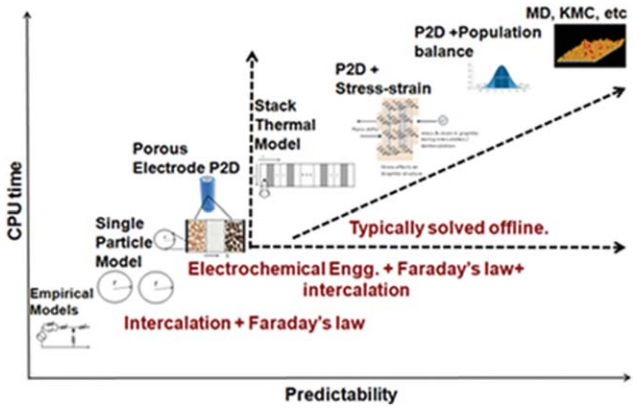

Figure 4. Wide range of physical phenomena that dictate different computation demands. Abbreviations: P2D = Porous 2 Dimensional , KMC = Kinetic Monte Carlo.

Tools for visualization and integration with rigorously defined data models could allow for cycles of design-build-test systems engineering aimed at complex system construction in teams distributed across multiple sites. Web databases of standard biological parts containing thousands of components have been commonly used by students of synthetic biology for more than a decade.[86] More recently, community-driven efforts have led to a proposed Synthetic Biology Open Language, which formalizes data standards for exchanging designs and for visual representations to make it easier for engineers to create and communicate them to others.[90] Synthetic Biology Open Language developers have already provided a practical demonstration of distributed design and engineering with the new data-exchange standards. When used alongside data-enabled tools for simulation and design, sophisticated data exchange could seamlessly integrate expertise across multiple industrial and academic sites to dramatically increase the sizes and complexities of the biological systems that can be engineered.

Energy systems and management. There is widespread popular support for the use of renewable energy, particularly solar and wind, which provide electricity without giving rise to any carbon emissions.[91] However, the energy generation from these resources is intermittent. The variability of these sources increases the need for power system storage and backup generation capacity to maintain the power balance and to meet the load demands in different operating scenarios.[92] Electricity providers must have enough installed power capacity to match peak demands and must continuously operate at enough capacity to meet real-time demands. This requires the use of a large number of distributed large-scale energy storage devices within the grid systems. Electrochemical energy storage devices offer the flexibility in capacity, sitting, and rapid response required to meet the application demands over a much wider range of functions than many other types of storage.

Within the electrochemical energy storage research area, data science will play a critical role in the discovery and invention of new electrode materials, electrolytes, membranes, and so on. In addition to the aforementioned MGI, an electrolyte genome initiative[93] has been initiated in the search for cheaper energy storage devices, which are expected to meet the cost barrier of $100/kWh.

A wide range of models have been developed and used to understand the performance of lithium-ion batteries (Figure 4).[94] Similar models are used for flow batteries and account for the changes in design, chemistry, and dynamics. To this date, despite significant advances in modeling, in particular at the microscale and nanoscale, the open-circuit voltage of ternary alloys used in lithium-ion batteries cannot be accurately predicted (within a typical experimental measurement error of 5 mV). It is expected that results from the aforementioned MGI will bring new multiscale models capable of bridging this gap. This effort will be highly interdisciplinary and will require contributions from physicists, chemists, material scientists, computer/data scientists, and chemical/mechanical engineers contributing to forward and inverse simulation in Density Functional Theory (DFT) calculations.

A potential area of opportunity for data science approaches to make an immediate impact in the quest to improve energy systems is in the area of nonlinear model predictive control (NMPC). Many of the aforementioned methods will play a significant role in the NMPC of batteries and grids for improving the efficiency (reduce cost), stability (prevent black outs), and safety (prevent battery explosions) and for enabling a higher level of penetration (increasing the number of off-grid installations based on renewable sources). The NMPC approach has been successfully demonstrated on several challenging problems, including batch nonisothermal reactors,[95] batch crystallization processes,[96] Tennessee Eastman plant control,[97] and distillation units.[98] There are also several excellent reviews and perspectives on NMPC[99,100] and economic model predictive control.[101]

We use a classic example from reaction engineering[95] to illustrate the advantages of addressing rich nonlinear dynamic problems through NMPC and some of the current challenges that could be addressed through data science. The model for an exothermic reaction $A \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} R$ in an unsteady state CSTR is given as follows:

$$\frac{dC_A}{dt} = \frac{C_{Ai} - C_A}{\tau} - k_1 C_A + k_{-1} C_R$$

$$\frac{dC_R}{dt} = \frac{C_{Ri} - R}{\tau} + k_1 C_A - k_{-1} C_R$$

$$\frac{dT_r}{dt} = \frac{-H}{\rho C_p}[k_1 C_A - k_{-1} C_R] + \frac{T_i - T_r}{\tau}$$

where T_i is the feed temperature, which is the manipulated variable, and C_R is the controlled variable. For a particular set of parameters (see ref. 95), this model exhibits sign change in the process gain. It can be theoretically shown that the traditional control and linear or linearized control for this model will be unstable, whereas NMPC offers stable control, even in the presence of uncertainties in the states or parameters. As described later, the expanding uses of NMPC will require that researchers go beyond traditional solution methods, which are typically limited to small-core laptop or desktop machines. Energy systems and devices, in particular energy-storage devices (i.e., batteries), are energy intensive and typically operate over significantly different scales (capacitors operate only for few seconds, batteries operate from seconds to hours,

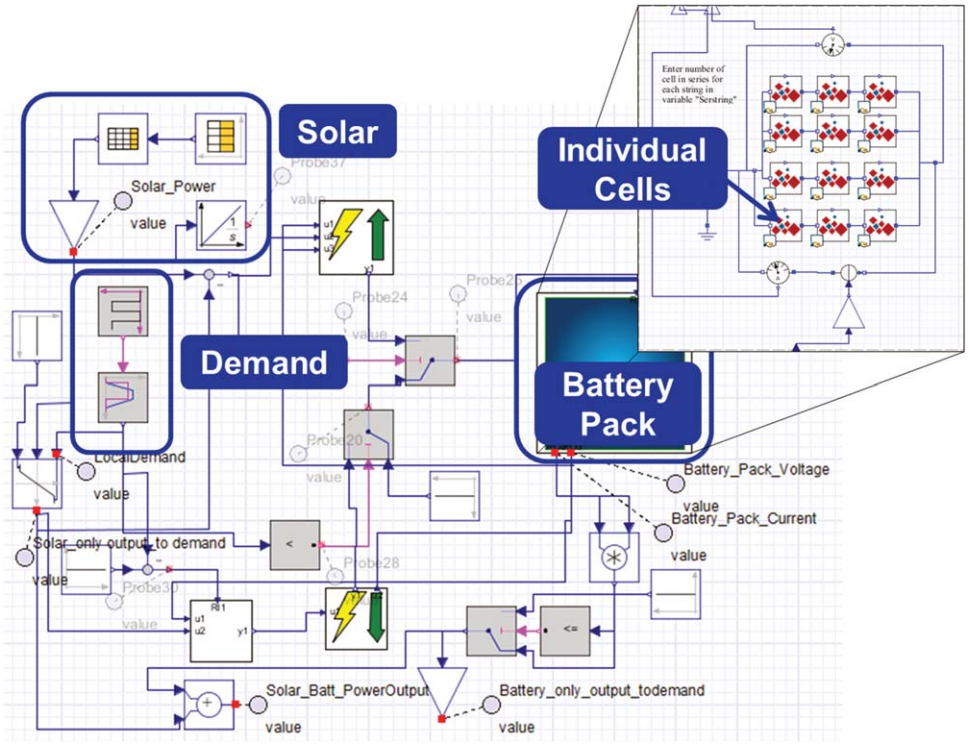

Figure 5. Simple microgrid connecting solar generation, lithium-ion battery storage, and demand forecast.

and flow batteries operate from minutes to hours). Most of these devices include electrochemical reactions, operation far from equilibrium, undesirable side reactions, thermal effects and runaway (batteries exploding), significant resistance, and delays caused by transport in multiple phases across multiple-length scales. Fortunately, continuum-level models (which require experimental open-circuit data) have been validated for the prediction of performance curves (charge-discharge curves) for lithium-ion batteries over a wide range of operating conditions and for different design parameters or configurations.[102] Battery models are stiff, and the model parameters have significant uncertainties with unstable dynamics; these are caused by the exothermic reaction, which can lead to thermal runaway. NMPC is the only option to guarantee optimal performance and stable operation as the battery degrades with life and use.[95]

Status quo in microgrid design. Today, an energy microgrid is designed and operated in a manner similar to how traditional large-scale energy grids are controlled (e.g., a dispatchable energy-generating unit might be replaced by intermittent weather-driven sources). A representative energy microgrid is shown in Figure 5. The typical approach for the design of a micogrid control system is encapsulated in four steps:

1. The assumption of a discrete model-in-time domain for each component.

2. The assumption of the validity of simple/linear dynamic models (i.e., a few differential equations for each component).

3. The use of a power conversion unit separately connected to both solar generation units and battery units (this reduces the efficiency of each of these systems by at least 8%).

4. The use of steady-state models or simple (linearized) approximations for dynamic models (e.g., with transfer functions) for each component.

The typical approaches that include batteries for microgrid control are based on highly simplified models of the batteries, where these are represented as black boxes (i.e., either steady-state or empirical fits).

This approach has served the community well, and there are several advantages to building our design and control strategies with the approach described previously. First, as the models are inherently linear or trivially linearizable, optimization is easily performed, for example, with the IBM CPLEX Optimization Studio. Second, the microgrid model is modular, and additional components can be added or removed (e.g., imaging changing between solar, wind, or other renewable energy inputs). Finally, linear models are computationally efficient and robust and consume very little RAM.

However, as they pertain to advanced predictive control, the weaknesses of the current approach clearly suggest that data science methods may be very useful in the future. The major disadvantage is the fact that the current status quo uses lithium-ion or other storage devices as simple black-box or empirical models. One consequence of the use of such models is that batteries cannot be used for the entire swing of depth of battery discharge (40% instead of 100% discharge). For example, a recent publication summarizes the tradeoffs in control of microgrids.[103] Unfortunately, simple empirical models were used to determine the performance of the batteries. The batteries were assumed to operate only in a narrow window (80% of the quoted capacity, which is roughly 65% of real capacity for most batteries, as manufacturers typically overstock the amount of materials to guarantee performance). In addition, charging rates were

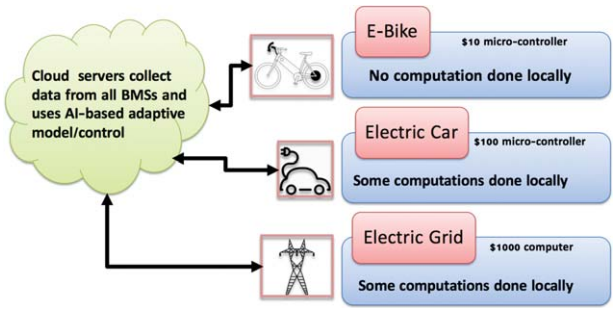

Figure 6. Conceptual framework for a cloud-based battery management system.

assumed to be low in these models, as higher rates would damage the batteries because of the increase in temperature.

Furthermore, because the battery lifetime modeling (i.e., long-term performance degradation) could not be included in the aforementioned commonly used schemes, the cost of the microgrid could not be optimized with respect to the battery life and replacement costs. That is, only constraints for the energy grid were achieved, an approach that precludes the maximization of the useable lifetime of the battery and may preclude maximization of the safety of the battery. Finally, because a black-box approach was used, it was impossible to impose dynamic path constraints on internal nonmeasurable variables within the micogrid. For example, if the internal temperature of the battery is not modeled, the battery must be operated at very low rates to ensure that the internal temperature does not rise high enough to reduce battery life and/or create unsafe operating conditions.

Potential of NMPC in energy microsystems. The potential benefits of a multiscale approach with the previously described NMPC methodology should lead to improved models and design, better control, and significantly improved battery operation. An overall framework for implementing NMPC in an energy microsystem could be as follows:

1. The development of detailed physics-based models for lithium-ion batteries that predict lifetime, safety, and temperature and enable the aggressive operation and control of batteries.

2. The development of detailed models capable of predicting the stochastic behavior of renewable-based power generation and the dynamic behavior of the controllable generation around their preferred operating points and solving the relevant power grid equations.

3. The simultaneous simulation and optimization of an integrated system (similar to Figure 5) to achieve the optimal level of control at the grid level (e.g., to meet power grid constraints) and at the same time optimize batteries (i.e., each cell/stack) individually at the node level.

4. The development of a conversion of the performance of each controller/action to the cost based on the life of the batteries and the use of this information to provide monetary units for a leveled cost of energy in real time for both batteries and rerouting of power within the grid.

5. Batteries that are very similar will have different life and performance characteristics depending on the ambient conditions, customer demand, and locations. Data will be available from various distributed installations, and data science is expected to help not only analyze the performance of these devices but also control these devices for improved life, safety, and economic benefit.

The development of such an approach would have immediate impacts on the efficiency and economics of using renewable energy systems. There are many inherent advantages to moving to a scheme such as NMPC that enables on-the-fly optimization and control. For example, the lifetime and performance of the battery are natively included within the action of the controller; this enables 100% drainage per cycle and extends the lifetime of the battery. The entire cost of the microsystem installation could be reduced (up to 40% by some estimates[104,105]) because of the concomitant reductions in battery cost. Real-time information with respect to energy routing/rerouting can be given on demand to customers, who can then decide to store, reroute, or use energy on the basis of fluctuating cost, needs, and demand. Finally, there is the advantage of additional potential reductions in the operating costs of the microgrid by the harmonious operation of the batteries over their entire range of operation in conjunction with available power-grid resources.

However, there are reasons why this approach has not yet gained widespread adoption. The NMPC framework, implemented in a way that maximizes predictive control, leads to severe computational challenges that must be addressed. First of all, multiple cells in stacks lead to 10^4 or more differential algebraic equations (DAEs), which arise from the discretization of the partial differential equations governing the behavior of the batteries. These equations must be solved and optimized simultaneously.[106] Second, computational demands are increased by the sheer scale of the model combined with the nonlinearities of the equations and uncertainties in the parameters and mechanisms. Although the numerical solution of DAEs and large-scale models have reasonably matured,[107,108] significant RAM and CPU requirements still persist for higher index DAEs (and even index 1 DAEs with a nonlinear nature) during the integration of thousands of equations connecting multiple devices within a grid, in particular for optimization, state, and parameter estimations. Finally, some individual components of the entire microgrid control model might include algorithms for particular components within the grid that are in the form of black-box models (e.g., because of licensing or intellectual property issues). This also reduces the efficiency of system-level simulation and control.

Future impacts of data science in energy microsystems management. With the aforementioned computational challenges, there is an opportunity for the field of data science to help make gains in the implementation of the NMPC framework. Specific examples are listed as follows:

- The improved data management and visualization of streaming data sets related to sensors and diagnostics that are ubiquitous within energy microgrids.
- The use of cloud computing technologies to provide fast and robust processing of large data sets within the NMPC framework.
- Systematic improvements in the integration of different level of algorithms, control architecture, protocols within a microgrid for power, and information flow.

- The use of distributed computing to support communication and filtering of solar and demand forecast data within the grid and across multiple connected/disconnected grids.
- Improved parameter fitting and model operation for on-demand estimation of the microgrid state (e.g., the development of a microgrid model that can handle multiple batteries in real-time within a grid and multiple batteries across multiple grids owned by utilities).

Figure 6 provides an envisioned role of cloud computing in the management of batteries for different applications (e-bikes, vehicles, and grids). The level of computing resources varies according to the application, and self-learning algorithms can play a critical role in optimizing and managing these systems. Although it is not possible to move all of the predictive models to online control for all of the applications, cloud computing with infinite resources can provide updated models, parameters, and control policies to individual Battery Management Systems (BMS) units. The time delay in data transfer between individual units and the cloud is, of course, best handled by NMPC algorithms compared to standard control schemes.

The optimization of large systems of models for grids operating at various locations with data collected at distinct locations presents a grand challenge at the intersection of data science and systems engineering. Meeting this challenge would provide significant benefits to individual residential customers by providing them options for their daily routines/energy use through mobile computing and would thereby make tangible changes in the way society consumes energy. Although some progress has been made in the scaling of linearized/linear predictive control algorithms to arbitrarily large systems, further research is needed to scale NMPC algorithms to arbitrarily large systems.[108] As a note to interested readers, a brief history on process control and future needs is provided elsewhere,[109] wherein the need to integrate process control with data science is specifically addressed.

Conclusions: Ideas for Making Data Science Mainstream in Chemical Engineering

One obvious challenging in raising our competency (as a profession) in data science is how to (or if we should) include it in the university curriculum or in a professional development context. We end this Perspective with a few closing thoughts on how this might be achieved in a sensible yet effective manner.

We begin by pointing out that the in the area of usability of data science methods, there has been an explosion of excellent free software tools that support data management, statistical and machine learning, and visualization. The growth of DIY learning online also means there is an accompanying amount of help readily available. A great starting point is the programming environment Python. An interactive version of Python is available on every computing platform in a convenient Web site framework (i.e., IPython Notebooks). There are countless YouTube videos available today to teach all aspects of this, from installing the software to using complex machine learning or visualization algorithms. This is just one example (of many), and the ubiquity of data science across many disciplines of science and engineering means that there is widespread support.

In the area of university education (both graduate and undergraduate), we point out that there may be room for small tweaks in the curriculum to accommodate data science. Most undergraduate curricula in chemical engineering offer a numerical methods or applied computing course. These courses were developed to help students implement methods to solve complex math problems related core engineering coursework (e.g., solving systems of 10 or fewer nonlinear equations or the numerical integration of a handful of ODEs). However, the growth in computing power and the ubiquity of multicore processors combined with the many available off-the-shelf codes (or even free Web sites) to solve these problems means that we could streamline the introduction and instruction on the use of these methods while still maintaining a focus on essential knowledge that chemical engineers need. An obvious alternate choice would be the addition of elective coursework or professional development workshops. The aforementioned growth of free online self-guided tutorials is another way to supplement the traditional chemical engineering curriculum. We expect that any of these approaches would be tractable without the sacrifice of any of the rigor and fundamental knowledge at the heart of our disciplinary training.

In closing, we believe the future of data science within the chemical engineering field is very bright. The methods and tools offered by the data science field have much to add to many aspects of our work across a wide range of subfields in our discipline, and chemical engineers are well-poised to dive into a data-rich future.

Acknowledgments

One of the authors (J.P.) acknowledges Blake Hough and Wesley Beckner for their careful reading of a portion of this article. Another author (V.S.) acknowledges S.Y. Lee for assistance in the preparation of Figure 5.

Literature Cited

1. Venkatasubramanian V. Drowning in data: informatics and modeling challenges in a data-rich networked world. *AIChE J*. 2009;55:2–8.
2. Codd EF. A relational model of data for large shared data banks. *Commun ACM*. 1970;13:377–387.
3. Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. *Commun ACM*. 2008;51:107–113.
4. White T. *Hadoop: The Definitive Guide*. 2nd ed. Farnham, United Kingdom: O'Reilly; 2010.
5. Dehmer M, Varmuza K, Bonchev D, Emmert-Streib F. *Statistical Modelling of Molecular Descriptors in QSAR/QSPR*. 2nd ed. Hoboken, NJ: Wiley; 2012.
6. Hosoya H. Topological index—newly proposed quantity characterizing topological nature of structural isomers of saturated hydrocarbons. *Bull Chem Soc Jpn*. 1971;44:2332–2339.
7. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. *J R Stat Soc Series B Stat Methodol*. 2001;63:411–423.
8. Sander J, Ester M, Kriegel HP, Xu XW. Density-based clustering in spatial databases: the algorithm GDBSCAN and its applications. *Data Min Knowl Disc*. 1998;2:169–194.
9. Kriegel HP, Kroger P, Sander J, Zimek A. Density-based clustering. *Wires Data Min Knowl Discov*. 2011;1:231–240.

10. Kohonen T. The self-organizing map. *Proc IEEE*. 1990; 78:1464–1480.
11. Manning CD, Raghavan P, Schütze H. *Introduction to Information Retrieval*. New York, NY: Cambridge University Press; 2008.
12. James G. *An Introduction to Statistical Learning With Applications in R*. New York, NY: Springer; 2013.
13. Hastie T, Tibshirani R, Friedman JH. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York, NY: Springer; 2009.
14. Leskovec J, Rajaraman A, Ullman JD. *Mining of Massive Datasets*. 2nd ed. Cambridge, United Kingdom: Cambridge University Press; 2014.
15. Borg I, Groenen PJF. *Modern Multidimensional Scaling Theory and Applications*. 2nd ed. New York, NY: Springer; 2005.
16. Tufte ER. *Envisioning Information*. Cheshire, CT: Graphics Press; 1990.
17. Tufte ER. *Visual Explanations: Images and Quantities, Evidence and Narrative*. Cheshire, CT: Graphics Press; 1997.
18. Tufte ER. *The Visual Display of Quantitative Information*. 2nd ed. Cheshire, CT: Graphics Press; 2001.
19. Tufte ER. *Beautiful Evidence*. Cheshire, CT: Graphics Press; 2006.
20. Bostock M, Ogievetsky V, Heer J. D-3: data-driven documents. *IEEE Trans Vis Comput Graph*. 2011;17:2301–2309.
21. Krzywinski M, Schein J, Birol I, et al. Circos: an information aesthetic for comparative genomics. *Genome Res*. 2009;19:1639–1645.
22. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;13:2498–2504.
23. Stasko J, Catrambone R, Guzdial M, McDonald K. An evaluation of space-filling information visualizations for depicting hierarchical structures. *Int J Hum Comput Stud*. 2000;53:663–694.
24. Amaral LAN, Ottino JM. Complex networks. *Eur Phys J B*. 2004;38:147–162.
25. Amaral LAN, Scala A, Barthélémy M, Stanley HE. Classes of small-world networks. *Proc Natl Acad Sci*. 2000;97:11149–11152.
26. Qin SJ. Process data analytics in the era of big data. *AIChE J*. 2014;60:3092–3100.
27. Blank TB, Brown SD, Calhoun AW, Doren DJ. Neural network models of potential-energy surfaces. *J Chem Phys*. 1995;103:4129–4137.
28. Lorenz S, Groß A, Scheffler M. Representing high-dimensional potential-energy surfaces for reactions at surfaces by neural networks. *Chem Phys Lett*. 2004;395: 210–215.
29. Behler J, Parrinello M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys Rev Lett*. 2007;98:146401.
30. Eshet H, Khaliullin RZ, Kühne TD, Behler J, Parrinello M. *Ab initio* quality neural-network potential for sodium. *Phys Rev B*. 2010;81:184107.
31. Khaliullin RZ, Eshet H, Kühne TD, Behler J, Parrinello M. Nucleation mechanism for the direct graphite-to-diamond phase transition. *Nat Mater*. 2011;10:693–697.
32. Khorshidi A, Peterson A. *Amp: Machine-Learning for Atomistics v0.3alpha*. Zenodo; 2015. DOI: 10.5281/zenodo.12665.
33. Li Z, Kermode JR, De Vita A. Molecular dynamics with on-the-fly machine learning of quantum-mechanical forces. *Phys Rev Lett*. 2015;114:096405.
34. Rupp M, Tkatchenko A, Müller K-R, von Lilienfeld OA. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys Rev Lett*. 2012; 108:058301.
35. Snyder JC, Rupp M, Hansen K, Müller KR, Burke K. Finding density functionals with machine learning. *Phys Rev Lett*. 2012;108:253002.
36. Ballester PJ, Mitchell JBO. A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. *Bioinformatics*. 2010;26:1169–1175.
37. Billard I, Marcou G, Ouadi A, Varnek A. *In silico* design of new ionic liquids based on quantitative structure-property relationship models of ionic liquid viscosity. *J Phys Chem B*. 2011;115:93–98.
38. Rupp M, Bauer MR, Wilcken R, et al. Machine learning estimates of natural product conformational energies. *PLoS Comput Biol*. 2014;10:e1003400.
39. Sharma V, Wang C, Lorenzini RG, et al. Rational design of all organic polymer dielectrics. *Nat Commun*. 2014;5:4845.
40. Kanal IY, Owens SG, Bechtel JS, Hutchison GR. Efficient computational screening of organic polymer photovoltaics. *J Phys Chem Lett*. 2013;4:1613–1623.
41. Ferguson AL, Panagiotopoulos AZ, Debenedetti PG, Kevrekidis IG. Systematic determination of order parameters for chain dynamics using diffusion maps. *Proc Natl Acad Sci*. 2010;107:13597–13602.
42. Tribello GA, Ceriotti M, Parrinello M. A self-learning algorithm for biased molecular dynamics. *Proc Natl Acad Sci*. 2010;107:17509–17514.
43. Beck DAC, Jonsson AL, Schaeffer RD, et al. Dynameomics: mass annotation of protein dynamics and unfolding in water by high-throughput atomistic molecular dynamics simulations. *Protein Eng Des Sel*. 2008;21:353–368.
44. Bromley D, Rysavy SJ, Su R, Toofanny RD, Schmidlin T, Daggett V. DIVE: a data intensive visualization engine. *Bioinformatics*. 2014;30:593–595.
45. Rysavy SJ, Bromley D, Daggett V. DIVE: a graph-based visual-analytics framework for big data. *IEEE Comput Graph Appl*. 2014;34:26–37.
46. Ceriotti M, Tribello GA, Parrinello M. Simplifying the representation of complex free-energy landscapes using sketch-map. *Proc Natl Acad Sci*. 2011;108:13023–13028.
47. Ceriotti M, Tribello GA, Parrinello M. Demonstrating the transferability and the descriptive power of Sketch-Map. *J Chem Theory Comput*. 2013;9:1521–1532.
48. National Science and Technology Council. Materials genome initiative for global competitiveness. https://www.whitehouse.gov/sites/default/files/microsites/ostp/materials_genome_initiative-final.pdf. Published June 2011. Accessed November 15, 2015.
49. Curtarolo S, Hart GLW, Nardelli MB, Mingo N, Sanvito S, Levy O. The high-throughput highway to computational materials design. *Nat Mater*. 2013;12:191–201.

50. Saal J, Kirklin S, Aykol M, Meredig B, Wolverton C. Materials design and discovery with high-throughput density functional theory: the Open Quantum Materials Database (OQMD). *JOM*. 2013;65:1501–1509.
51. Jain A, Ong SP, Hautier G, et al. Commentary: the Materials Project: a materials genome approach to accelerating materials innovation. *APL Mater*. 2013;1:011002.
52. Hachmann J, Olivares-Amaya R, Atahan-Evrenk S, et al. The Harvard Clean Energy Project: large-scale computational screening and design of organic photovoltaics on the world community grid. *J Phys Chem Lett*. 2011;2: 2241–2251.
53. Li Y, Li X, Liu J, Duan F, Yu J. *In silico* prediction and screening of modular crystal structures via a high-throughput genomic approach. *Nat Commun*. 2015;6: 8328.
54. Keskin S, Sholl DS. Screening metal-organic framework materials for membrane-based methane/carbon dioxide separations. *J Phys Chem C*. 2007;111:14055–14059.
55. Haldoupis E, Nair S, Sholl DS. Finding MOFs for highly selective CO_2/N_2 adsorption using materials screening based on efficient assignment of atomic point charges. *J Am Chem Soc*. 2012;134:4313–4323.
56. Kim J, Abouelnasr M, Lin LC, Smit B. Large-scale screening of zeolite structures for CO_2 membrane separations. *J Am Chem Soc*. 2013;135:7545–7552.
57. Colon YJ, Snurr RQ. High-throughput computational screening of metal-organic frameworks. *Chem Soc Rev*. 2014;43:5735–5749.
58. Broadbelt LJ, Pfaendtner J. Lexicography of kinetic modeling of complex reaction networks. *AIChE J*. 2005; 51:2112–2121.
59. Hjelmfelt A, Weinberger ED, Ross J. Chemical implementation of neural networks and Turing machines. *Proc Natl Acad Sci*. 1991;88:10983–10987.
60. Shenvi N, Geremia JM, Rabitz H. Efficient chemical kinetic modeling through neural network maps. *J Chem Phys*. 2004;120:9942–9951.
61. Zhou Z, Lü Y, Wang Z, Xu Y, Zhou J, Cen K. Systematic method of applying ANN for chemical kinetics reduction in turbulent premixed combustion modeling. *Chin Sci Bull*. 2013;58:486–492.
62. Kayala MA, Azencott CA, Chen JH, Baldi P. Learning to predict chemical reactions. *J Chem Inf Model*. 2011; 51:2209–2222.
63. Pietrucci F, Andreoni W. Graph theory meets *ab initio* molecular dynamics: atomic structures and transformations at the nanoscale. *Phys Rev Lett*. 2011;107:085504.
64. Zheng S, Pfaendtner J. Car-Parrinello molecular dynamics + metadynamics study of high-temperature methanol oxidation reactions using generic collective variables. *J Phys Chem C*. 2014;118:10764–10770.
65. Medema MH, van Raaphorst R, Takano E, Breitling R. Computational tools for the synthetic design of biochemical pathways. *Nat Rev Microbiol*. 2012;10:191–202.
66. Lee JW, Na D, Park JM, Lee J, Choi S, Lee SY. Systems metabolic engineering of microorganisms for natural and non-natural chemicals. *Nat Chem Biol*. 2012;8:536–546.
67. Keasling JD. Manufacturing molecules through metabolic engineering. *Science*. 2010;330:1355–1358.
68. Carothers JM. Design-driven, multi-use research agendas to enable applied synthetic biology for global health. *Syst Synth Biol*. 2013;7:79–86.
69. Carothers JM, Goler JA, Keasling JD. Chemical synthesis using synthetic biology. *Curr Opin Biotechnol*. 2009; 20:498–503.
70. Carothers JM, Goler JA, Juminaga D, Keasling JD. Model-driven engineering of RNA devices to quantitatively program gene expression. *Science*. 2011;334: 1716–1719.
71. Zhang FZ, Carothers JM, Keasling JD. Design of a dynamic sensor-regulator system for production of chemicals and fuels derived from fatty acids. *Nat Biotechnol*. 2012;30:354–359.
72. Monk J, Nogales J, Palsson BO. Optimizing genome-scale network reconstructions. *Nat Biotechnol*. 2014;32: 447–452.
73. O'Brien EJ, Palsson BO. Computing the functional proteome: recent progress and future prospects for genome-scale models. *Curr Opin Biotechnol*. 2015;34:125–134.
74. Chau PC. *Process Control: A First Course with MATLAB*. Cambridge, United Kingdom: Cambridge University Press; 2002.
75. Feinberg M, Horn FJM. Dynamics of open chemical systems and the algebraic structure of the underlying reaction network. *Chem Eng Sci*. 1974;29:775–787.
76. Saltelli A, Ratto M, Andres T, et al. *Global Sensitivity Analysis: The Primer*. Hoboken, NJ: Wiley; 2008.
77. Dunlop M, Keasling J, Mukhopadhyay A. A model for improving microbial biofuel production using a synthetic feedback loop. *Syst Synth Biol*. 2010;4:95–104.
78. Hoffman F, Rinas U. Stress induced by recombinant protein production in *Escherichia coli*. In: Enfors SO, ed. *Physiological Stress Responses in Bioprocesses*. New York, NY: Springer; 2004.
79. Brockman IM, Prather KLJ. Dynamic metabolic engineering: new strategies for developing responsive cell factories. *Biotechnol J*. 2015;10:1360–1369.
80. Holtz WJ, Keasling JD. Engineering static and dynamic control of synthetic pathways. *Cell*. 2010;140:19–23.
81. Stevens JT, Carothers JM. Designing RNA-based genetic control systems for efficient production from engineered metabolic pathways. *ACS Synth Biol*. 2015;4:107–115.
82. Guimera R, Amaral LAN. Functional cartography of complex metabolic networks. *Nature*. 2005;433:895–900.
83. Breitling R, Achcar F, Takano E. Modeling challenges in the synthetic biology of secondary metabolism. *ACS Synth Biol*. 2013;2:373–378.
84. Kosuri S, Goodman DB, Cambray G, et al. Composability of regulatory sequences controlling transcription and translation in *Escherichia coli*. *Proc Natl Acad Sci U S A*. 2013;110:14024–14029.
85. Cambray G, Guimaraes JC, Mutalik VK, et al. Measurement and modeling of intrinsic transcription terminators. *Nucleic Acids Res*. 2013;41:5139–5148.
86. Way JC, Collins JJ, Keasling JD, Silver PA. Integrating biological redesign: where synthetic biology came from and where it needs to go. *Cell*. 2014;157:151–161.
87. Rosenberg AB, Patwardhan RP, Shendure J, Seelig G. Learning the sequence determinants of alternative

splicing from millions of random sequences. *Cell*. 2015; 163:698–711.
88. Hillson NJ, Rosengarten RD, Keasling JD. j5 DNA assembly design automation software. *ACS Synth Biol*. 2012;1:14–21.
89. Chen J, Densmore D, Ham TS, Keasling JD, Hillson NJ. DeviceEditor visual biological CAD canvas. *J Biol Eng*. 2012;6:1–12.
90. Galdzicki M, Clancy KP, Oberortner E, et al. The Synthetic Biology Open Language (SBOL) provides a community standard for communicating designs in synthetic biology. *Nat Biotechnol*. 2014;32:545–550.
91. Kassakian JG, Hogan WW, Schmalensee R, Jacoby HD. *The Future of the Electric Grid*. Boston, MA: MIT Press; 2011.
92. Meliopoulos AS, Cokkinides GJ, Huang R, et al. Smart grid technologies for autonomous operation and control. *IEEE Trans Smart Grid*. 2011;2:1–10.
93. Qu X, Jain A, Rajput NN, et al. The Electrolyte Genome project: a big data approach in battery materials discovery. *Comput Mater Sci*. 2015;103:56–67.
94. Ramadesigan V, Northrop PW, De S, Santhanagopalan S, Braatz RD, Subramanian VR. Modeling and simulation of lithium-ion batteries from a systems engineering perspective. *J Electrochem Soc*. 2012;159:R31–R45.
95. De Oliveira Kothare SL, Morari M. Contractive model predictive control for constrained nonlinear systems. *IEEE Trans Automat Contr*. 2000;45:1053–1071.
96. Nagy ZK, Braatz RD. Robust nonlinear model predictive control of batch processes. *AIChE J*. 2003;49:1776–1786.
97. Ricker N, Lee J. Nonlinear model predictive control of the Tennessee Eastman challenge process. *Comput Chem Eng*. 1995;19:961–981.
98. Huang R, Zavala VM, Biegler LT. Advanced step nonlinear model predictive control for air separation units. *J Process Control*. 2009;19:678–685.
99. Rawlings JB. Tutorial overview of model predictive control. *IEEE Control Syst Mag*. 2000;20:38–52.
100. Bequette BW. Non-linear model predictive control: a personal retrospective. *Can J Chem Eng*. 2007;85:408–415.
101. Ellis M, Durand H, Christofides PD. A tutorial review of economic model predictive control methods. *J Process Control*. 2014;24:1156–1178.
102. Doyle M, Fuller TF, Newman J. Modeling of galvanostatic charge and discharge of the lithium/polymer/insertion cell. *J Electrochem Soc*. 1993;140:1526–1533.
103. Zachar M, Daoutidis P. Understanding and predicting the impact of location and load on microgrid design. *Energy*. 2015;90:1005–1023.
104. Beetz B. Li-ion battery costs to fall 50% in next 5 years, driven by renewables. http://www.pv-magazine.com/news/details/beitrag/li-ion-battery-costs-to-fall-50-in-next-5-years–driven-by-renewables_100022051/. Accessed November 30, 2015.
105. Ivanova N. Lithium-ion costs to fall by up to 50% within five years. http://analysis.energystorageupdate.com/lithium-ion-costs-fall-50-within-five-years.
106. Brown PN, Hindmarsh AC, Petzold LR. Using Krylov methods in the solution of large-scale differential-algebraic systems. *SIAM J Sci Comput*. 1994;15:1467–1488.
107. Wanner G, Hairer E. *Solving Ordinary Differential Equations II*. Vol 1. Berlin, Germany: Springer-Verlag; 1991.
108. Matni N, Leong YP, Wang YS, You S, Horowitz MB, Doyle JC. Resilience in large scale distributed systems. *Proc Comput Sci*. 2014;28:285–293.
109. Misbah A. Technology paving the way for the next generation of process control. http://blog.ifac-control.org.dedi161.flk1.host-h.net/2015/11/04/technology-paving-the-way-for-the-next-generation-of-process-control/. Published November 4, 2015.