PERSPECTIVE | NOVEMBER 08 2018

Perspective: Computational chemistry software and its advancement as illustrated through three grand challenge cases for molecular science

FREE

Anna Krylov ; Theresa L. Windus; Taylor Barnes; Eliseo Marin-Rimoldi; Jessica A. Nash ; Benjamin Pritchard; Daniel G. A. Smith ; Doaa Altarawy ; Paul Saxe; Cecilia Clementi ; T. Daniel Crawford ; Robert J. Harrison; Shantenu Jha; Vijay S. Pande; Teresa Head-Gordon

J. Chem. Phys. 149, 180901 (2018)
https://doi.org/10.1063/1.5052551
CHORUS

Perspective: Computational chemistry software and its advancement as illustrated through three grand challenge cases for molecular science

Anna Krylov,[1] Theresa L. Windus,[2] Taylor Barnes,[3] Eliseo Marin-Rimoldi,[3] Jessica A. Nash,[3] Benjamin Pritchard,[3] Daniel G. A. Smith,[3] Doaa Altarawy,[3] Paul Saxe,[3] Cecilia Clementi,[4,5] T. Daniel Crawford,[6] Robert J. Harrison,[7] Shantenu Jha,[8] Vijay S. Pande,[9] and Teresa Head-Gordon[10,a)]

[1]*Department of Chemistry, University of Southern California, Los Angeles, California 90089, USA*
[2]*Department of Chemistry, Iowa State University, Ames, Iowa 50011, USA*
[3]*Molecular Sciences Software Institute, Blacksburg, Virginia 24061, USA*
[4]*Department of Chemistry and Center for Theoretical Biological Physics, Rice University, 6100 Main Street, Houston, Texas 77005, USA*
[5]*Department of Mathematics and Computer Science, Freie Universitt Berlin, Arnimallee 6, 14195 Berlin, Germany*
[6]*Department of Chemistry, Virginia Tech, Blacksburg, Virginia 24061, USA*
[7]*Institute for Advanced Computational Science, Stony Brook University, Stony Brook, New York 11794, USA*
[8]*Electrical and Computer Engineering, Rutgers The State University of New Jersey, Piscataway, New Jersey 08854, USA*
[9]*Department of Bioengineering, Stanford University, Stanford, California 94305, USA*
[10]*Department of Chemistry, Department of Bioengineering, Department of Chemical and Biomolecular Engineering, Pitzer Center for Theoretical Chemistry, University of California, Berkeley, California 94720, USA*

(Received 19 August 2018; accepted 18 October 2018; published online 8 November 2018)

The field of computational molecular sciences (CMSs) has made innumerable contributions to the understanding of the molecular phenomena that underlie and control chemical processes, which is manifested in a large number of community software projects and codes. The CMS community is now poised to take the next transformative steps of better training in modern software design and engineering methods and tools, increasing interoperability through more systematic adoption of agreed upon standards and accepted best-practices, overcoming unnecessary redundancy in software effort along with greater reproducibility, and increasing the deployment of new software onto hardware platforms from in-house clusters to mid-range computing systems through to modern supercomputers. This in turn will have future impact on the software that will be created to address grand challenge science that we illustrate here: the formulation of diverse catalysts, descriptions of long-range charge and excitation transfer, and development of structural ensembles for intrinsically disordered proteins.
https://doi.org/10.1063/1.5052551

I. INTRODUCTION

Computational molecular science (CMS) is a core science area that underpins a broad spectrum of disciplines, including chemistry and biochemistry, catalysis, materials science, nanoscience, energy and environmental science, and geosciences. The CMS community has achieved fantastic success over its long history by creating computational models and algorithms that are now used by hundreds of thousands of scientists worldwide, via dozens of academic and industrial software packages stemming from decades of human effort. Their translation and deployment have resulted in innovative products coming from the chemical, pharmaceutical, information technologies, and advanced engineering industries that have and hopefully will continue to make lives better.

These scientific breakthroughs have been made possible by the evolution of dozens of CMS community codes—some with lifetimes reaching back to the earliest days of computing—which include both open-source and commercial packages. One of the strengths and at the same time one of the challenges of the CMS field is the multitude of different software packages used and the data that are generated from it. There are some key benefits to having such a robust software ecosystem. Multiple code bases ensure the agility of the developments and facilitate the testing of new ideas and paradigms, which are constantly emerging in this vibrant and rapidly developing field. Different packages exemplify different software design philosophies, and a healthy competition

a)Author to whom correspondence should be addressed: thg@berkeley.edu

between them allows the best of breed to emerge. Furthermore, there can be specific methodological and software niches served by different packages.

However, this also leads to the lack of algorithmic interoperability between codes, in which the current standard is to duplicate the most standard algorithms in each software platform, which can be inefficient, prone to translational errors, and suppresses innovation for new methodology. When the format of source data changes or is different from code to code, the warehouse, data repository, or software interface must be updated to read that source or it will not function properly. We are currently experiencing porting and scalability bottlenecks of community codes on traditional high performance computing (HPC) platforms, multicore clusters, and Graphics processing unit (GPUs). The bulk of the needed software modifications to address these issues involves low-level translation and integration tasks which typically require the full attention of domain experts. Together this has led to tremendous challenges regarding the sustainability, maintenance, adaptability, and extensibility of these early software investments.

In 2016, the National Science Foundation (NSF) selected the CMS community for the establishment of a Scientific Software Innovation Institute, which we have coined the Molecular Sciences Software Institute (MolSSI). The purpose of the MolSSI[1] is to serve as a long-term hub of excellence in software infrastructure and technologies to actively enable software development in the CMS community, by developing a culture of modern software engineering practices. The MolSSI aims to reach these goals by engaging the CMS community in multiple ways.

First, the MolSSI has formulated an interdisciplinary team of software scientists[2] who are developing software frameworks, interacting with community code developers, collaborating with partners in cyberinfrastructure, forming mutually productive coalitions with industry, government labs, and international efforts, and ultimately serving as future CMS experts and leaders. In addition, MolSSI supports and mentors a cohort of software fellows[3] of graduate students and postdoctoral scholars actively developing code infrastructure in CMS research groups across the U.S. MolSSI is guided by an internal Board of Directors[4] and an external Science and Software Advisory Board[5] comprising a diverse group of leaders in the field, who both work together with the software scientists and fellows to establish the key software priorities for MolSSI.

Furthermore, the MolSSI continues to sponsor multiple software workshops for the purpose of understanding the different needs of the diverse CMS community through capturing requirements and active development of use cases. In addition, MolSSI has encouraged the organization of a community-driven Molecular Sciences Consortium[6] to develop standards for code and data sharing. MolSSI is also actively developing and providing summer schools[7] and an on-line job search forum[8] for developing a diverse and broadly trained workforce for the future generation of CMS activities. In total, the MolSSI endeavors to fundamentally and dramatically improve molecular science software development to benefit the CMS community.

As a result, this new software infrastructure and support will create opportunities for new levels of scientific questions to be asked and answered. In this perspective, we discuss the software challenges for three illustrative grand challenge science areas: catalyst design, long-range charge transfer and excitation transfer, and intrinsically disorder proteins. We show that the theoretical, methodological, and algorithmic advances that provide the scientific approaches to these problems—i.e., what individual research groups excel at and where all of the true innovation will come from—will be the underlying engines for the software projects that MolSSI can address in these grand challenge scientific use cases.

II. CATALYST DESIGN

Effective catalysts decrease the energy consumption of reactions, improve the control and selectivity of undesirable by-products, and reduce the production of waste components.[9] Such species are at the heart of the worldwide chemical and biochemical product industry such as petroleum[10,11] and pharmaceuticals.[12] Existing and emerging technologies related to energy applications,[13–15] new synthetic routes for polymers[16] and drugs,[17] biomass conversion to light alkanes and alcohols,[18] and the development of designed enzymes[19,20] all hinge on a deeper understanding of catalytic mechanisms.

The goal of designing new catalysts is to ensure that they are highly active and selective with high turnover rates, while maintaining thermochemical stability such that the catalyst survives many reactions, thereby decreasing the industrial cost.[21] The modeling requirements to achieve this goal are to describe active sites with atomic precision and to accurately incorporate the multiscale, multiphase environments in which they operate.[22–24] This requirement of multi-physics complexity introduces significant scientific and computational challenges. For example, single site catalysts[25] are often affected by the electronic or physical structure in which they embedded—such as the enzyme scaffold,[26] the shape-selective effects on catalysis in porous zeolites,[27] or for driven catalysis at the electrocatalytic interface[28]—significantly changing the catalytic behavior due to the full system. Multiple catalytic sites[29] can have cooperative or destructive effects that can drastically change the product distributions and the overall activity of the reactive sites. In addition, solvent and non-equilibrium effects can also have huge effects on the stability and dynamics of the catalytic process.[30,31]

To illustrate, mesoporous silicon nanoparticles (MSNs) can be functionalized with many different active groups, have varying pore sizes, be utilized with different solvents, have variable catalytic active site concentrations, and therefore have designable ranges of chemical reactivity. It has been found that the relative rates of catalyzed aldol reactions in these MSNs can be inverted by 2 orders of magnitude by changing the solvent from hexane to water, showing that solvent cannot be ignored in these reactions.[32] The major differences suggested by the computational analysis found that solvent polarity and acidity were critically important features for understanding the changes in reaction rates.[32] Simulations on a portion of the MSN channel also showed that the curvature of the MSN

plays a significant role in the reaction mechanism that is not captured by simpler cluster models.[33] Spatially coarse-grained stochastic models and kinetic Monte Carlo simulations have recently been used to examine the diffusion processes coupled with the catalytic activity[34] and were able to explain the 20-fold enhancement found for a relative small increase in pore sizes.

An additional computational catalysis challenge is the first-principles modeling of a complete electrochemical device. The ultimate goal of whole device modeling in electrochemistry is to simulate the partial current densities for individual products as a function of catalyst composition, electrolyte composition (including pH), membrane composition, and applied voltage. Computational modeling of the various physical phenomena that must be considered is given in Fig. 1 for CO_2 reduction, in which Singh and co-workers proposed the integration of three levels of theory and computation that are needed to calculate the overall performance of the cell.[35] The first requirement is a continuum model for the species transport and reaction in a 1-D electrochemical cell; a microkinetic model is needed to describe the rates at which each product is formed to feed up to the continuum model; and finally Kohn-Sham (KS) Density Functional Theory (DFT)[36] is used to characterize intermediates and reaction barriers, which are supplied to the microkinetic model.

Although these studies show the power of theory and simulations to understand and quantify complex catalytic processes, there are still many scientific challenges to catalysis modeling that are limited by algorithmic and software bottlenecks. By definition, quantum mechanical (QM) methods are required to reproduce the reactive part of the system and are needed to provide the necessary accuracy to predict reaction barriers and thermochemistry. However, these methods are often too computationally expensive to describe the rest of the environment for the reaction (solvent, solid catalyst support, enzyme scaffolds, complexes, etc.). Furthermore, while KS-DFT[36] has become a standard bearer for QM modeling, especially using the dispersion corrected, generalized gradient approximation,[37] there is often a need for higher rung DFT functionals[38] as well as wavefunction methods for strongly correlated systems[39] that can make the computations even more difficult. The integration of high-end with lower-fidelity methods via quantum embedding[40] and quantum mechanics/molecular mechanics (QM/MM)

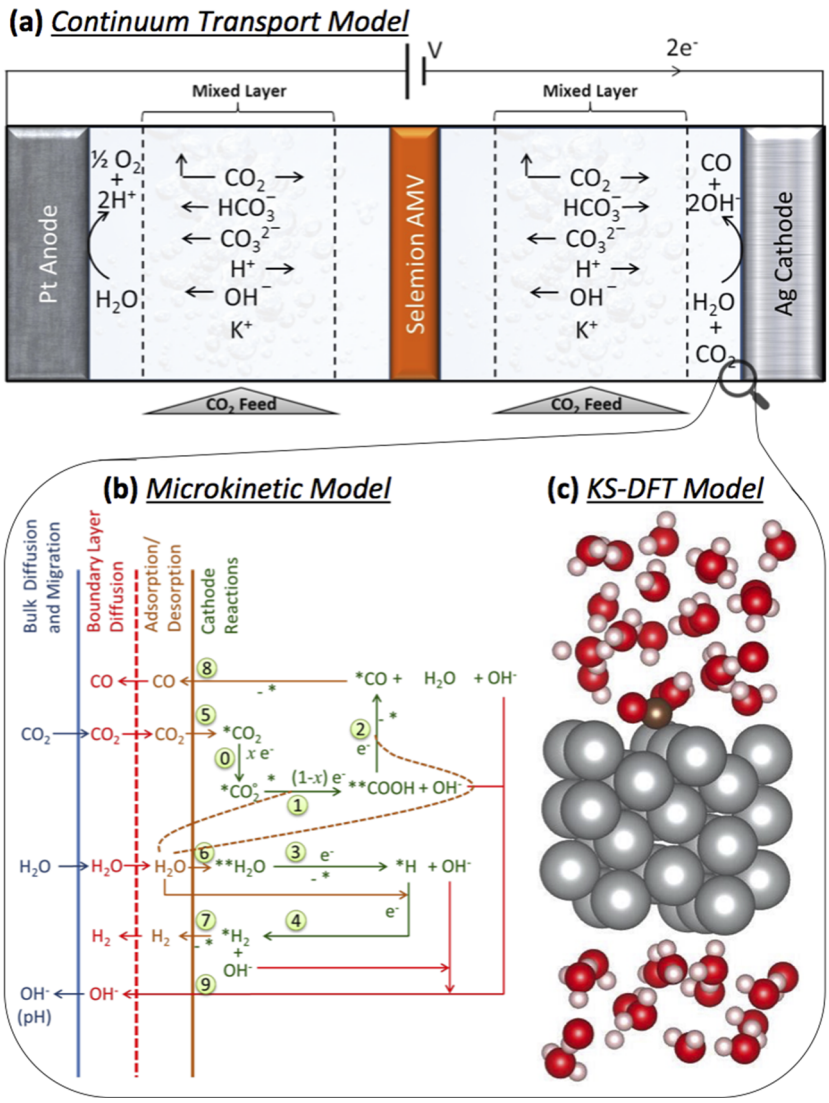

FIG. 1. Whole device model for electrocatalysis. (a) Continuum transport model for a 1-D electrochemical cell, illustrated for CO_2 reduction. (b) Microkinetic model showing elementary processes for CO_2RR and HER over Ag(110). (c) KS-DFT to supply the microkinetic model with relevant stationary points on the free energy surface such as intermediates and reaction barriers. Reprinted with permission from Singh *et al.*, Phys. Chem. Chem. Phys. **17**, 18924–18936 (2015). Copyright 2015 Royal Society of Chemistry.

methods,[41,42] and resolving the continuum to molecular resolution to describe statistical fluctuations are necessary. Efficient classical[43] and semiclassical[44] *ab initio* dynamics on one or more potential energy surfaces,[45] and methods for improving the sampling of complex high-dimensional energy landscapes for extended time scales,[46–48] and under non-equilibrium conditions[49] are particular theoretical needs for real catalytic systems.

Currently most computational chemistry software does not have this full suite of capabilities, emphasizing the need for interoperable software and frameworks to couple accurate electronic structure, statistical mechanics, and kinetics for progressively larger and more complex systems. While some of the components of a framework for coupling theories exist, a seamless integration of these components and making the components useful for many different applications is lacking. Enabling this software on massively parallel, heterogeneous hardware systems will also be required to enable computations on the scale needed to address the catalysis challenge.[50] With the advent of machine learning as an approach to recapitulating advanced potential energy surfaces,[51,52] their extension to the design of new catalysts with specific properties requires the accumulation of extensive amounts of data (both experimental and computational) and the ability to access these data through portals and/or databases. While there have been some attempts at developing standards for the chemistry community to make the data more accessible to a broader audience, there has been limited traction in this area.

III. LONG-RANGE CHARGE AND EXCITATION TRANSFER

A wide range of materials and biological systems make use of charge and/or excitation transport over long distances. These processes are the fundamental core of natural and artificial light harvesting,[53] cellular respiration, proton-coupled electron transfer in fuel cells and batteries, photonics, electronics, and spintronics. These inherently quantum phenomena span multiple time, energy, and length scales and involve numerous coupled degrees of freedom. Furthermore, they operate in a diverse range of ordered and disordered organic and inorganic materials.

As a fascinating example illustrating the complexity of these processes, consider electrochemically active bacteria,[54–56] the metabolic cycle of which involves shuttling electrons from the periplasm, via the outer-membrane, to solid external acceptors. At certain conditions, electron transport in these bacteria occurs via long (tens of μm) extensions called bacterial nanowires.[57] These species can be exploited in novel technologies including biophotovoltaics, microbial fuel cells, bioremediation of heavy metals, and more.

Experiments have measured nanowires' conductivity and their range of redox potentials exhibited at different conditions, identified the proteins responsible for electron transport, and determined their crystal structures, density, and location within the cellular membranes. And yet our mechanistic understanding of the electron transfer in these organisms is rather rudimentary, and the interpretation of these state-of-the-art experiments is hotly debated. Does electron transfer occur by tunneling-like ballistic transport or via hopping?[58] What is the exact role of different cofactors, such as flavin? Why there are redundant pathways and what controls the variation in expression of the multiple proteins associated with these pathways? And how can the high electrical currents[57] be supported by this fluctuating biological scaffold?

In one type of such organism, *Shewanella oneidensis MR-1*, electron transport proceeds via 3 distinct membrane proteins,[59] MtrF, MtrC, and OmcA, all of which are decaheme cytochromes, shown in Fig. 2. Theoretical modeling[60–63] has provided important insights into this system. For example, pioneering calculations by Blumberger and co-workers[60–62] determined that the redox landscape does not have a gradient. Therefore electron transport can proceed in either direction (10-to-5 or 5-to-10 as shown in Fig. 2), meaning that the bacteria can reverse the direction of the electron flow depending on their functional needs. This work also pointed out that the larger free-energy barriers between hemes are often compensated by larger electronic couplings, and vice versa. A more recent study[63] has quantified the effect of the overall redox state of the protein on the free energy landscape and explained the variations in measured redox potentials by different conductance regimes, i.e., hole hopping versus electron hopping. Even while these studies employed very advanced theoretical approaches for quantitative modeling of electron transfer,[64] together they revealed limitations of the currently available software and methodology, and thus left many intriguing questions unanswered.

To quantitatively model redox states of just one decaheme cytochrome (roughly 200 000 atoms, not counting the solvating water), one needs to compute free energies of each heme in a reduced and oxidized state for different combinations of redox states of the other hemes, which results in $2^9 = 512$

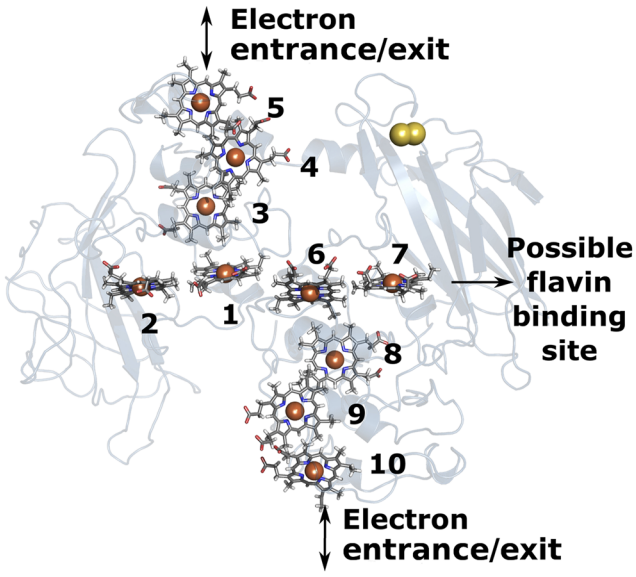

FIG. 2. A series of heme-c cofactors in the protein MtrF illustrating the current model of electron flow via hopping in the decaheme outer-membrane proteins in *Shewanella oneidensis MR-1*. The staggered-cross arrangement of the 10 hemes is counter-intuitive and is still not fully explained. Reprinted with permission from Barrozo *et al.*, Angew. Chem., Int. Ed. **57**, 6805–6809 (2018). Copyright 2018 John Wiley and Sons, Inc.

distinct redox states of the protein. Additional information needed to arrive at the redox energies should include the protonation states of various protein residues, as well as their possible pKa changes in response to redox states of the hemes. Furthermore, each redox calculation entails extensive sampling of configurations on the reduced and oxidized state.

Even if the sampling of structures is carried out with molecular dynamics, the evaluation of the electronic energy differences will require multiple QM/MM calculations. In the minimal setup needed to obtain quantitative agreement with the experiment,[63] when the QM region includes only one heme, the quantum system was comprised of 109 atoms. Such calculations[63] of the redox states of MtrF and MtrC in only 3 out of 512 distinct regimes (electron hopping, hole hopping, and electron hopping with heme 7 reduced) using KS-DFT, classical molecular dynamics, and a linear response approximation have burned well over 150 000 central processing unit (CPU)-hours at the Extreme Science and Engineering Discovery Environment (XSEDE) and University of Southern California High-Performance Computing Cluster (USC-HPCC) facilities, all while using the fastest and most efficient implementations of these methods.

To calculate electron-transfer rates and electron flow, one needs to go beyond this QM/MM approach and compute free energy changes for electron transfer between each pair of neighboring hemes (9 pairs) as well as their respective electronic couplings. To do so, the QM system should include 2 hemes and employ an electronic structure method capable of describing multiple electronic states (D/A and D^+/A^-) and their interactions. Obviously, such calculations are demanding, even when using the least sophisticated levels of theory (DFT, non-polarizable force-fields, and a linear response approximation). Modeling of excitation transfer and exciton dynamics is even more demanding. Given the ultrafast nature of photo-induced processes, a more correct calculation requires the departure from Marcus-type models toward theories appropriate for non-equilibrium processes, as shown in Fig. 3.

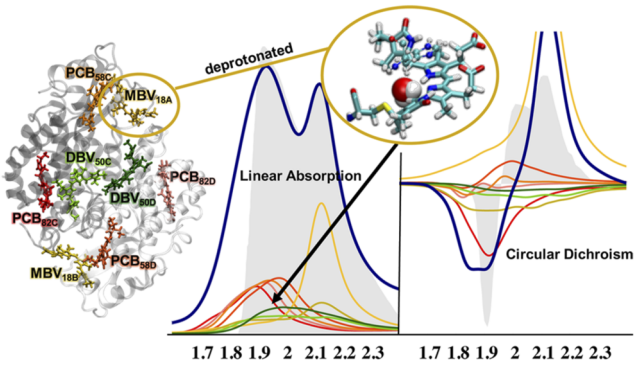

FIG. 3. Phycobiliprotein light harvesting complexes from cryptophyte algae. Excitation energy transfer among eight bilin chromophores is strongly modulated by their interactions with local environment and vibrational motions. A multiscale approach combining high-level electronic structure calculations and non-Markovian reduced density matrix description provided insight into inhomogeneous line broadening, excited-state lifetimes, and dissipative dynamics[66] in these systems. Reprinted with permission from Lee *et al.*, J. Am. Chem. Soc. **139**, 7803–7814 (2017). Copyright 2017 American Chemical Society.

These different types of calculations entail different workflows that affect the data exchange between the modules for this grand challenge science case. Some standard workflows, such as those used for redox potentials, are not automated and require diligent and expert human involvement at various stages of the calculation. Consequently, the software limitations affect productivity since too much precious research time is spent on fighting with idiosyncrasies of various packages, fixing broken interfaces, and trouble-shooting technical issues. A recent attempt to create a more automated workflow for high-throughput modeling of rhodopsins illustrates the progress that has been made, as well as the complexity and heterogeneity of the underlying theoretical models and the severe requirements for the software stack.[65]

And yet, this is still insufficient for a complete description of electron transport through *Shewanella's* nanowires! How does electron transfer occur between different proteins across the membrane protein complex? Do the properties of the proteins in immediate contact with the solid electron acceptor (i.e., an electrode) differ from those residing deep inside the membrane? Does bending of the nanowires affect their conductivity? What is the role of soluble electron shuttlers in the overall mechanism? Does the electron flow follow static one-dimensional pathways or dynamic three-dimensional networks? To answer these questions, the theoretical model must go beyond a single isolated protein, while preserving atomic-level resolution in describing the essential physics. This makes the software requirements even more challenging.

IV. INTRINSICALLY DISORDERED PROTEINS

While most of the effort in molecular biology over the last 30 years has focused on the characterization of the conformational changes and folding of structured proteins, it has long been known that regions of intrinsic disorder (see Fig. 4) are common in eukaryotic proteins.[67,68] Intrinsically disordered regions (IDRs) and proteins (IDPs) comprise approximately 25% of the human proteome, and their inherent disorder is required for function, such as in cellular regulation and signaling.[69] At the same time, numerous IDPs are associated with human diseases, including cancer, cardiovascular disease, amyloidoses, neurodegenerative diseases, and diabetes.[69]

One of the deep intellectual challenges of studying IDPs is how to build structural and dynamical models that allow researchers to gain insight and conceptualize their nature.[72] For folded proteins, crystal structures from X-ray crystallography provide concrete, predictive, and yet conceptually straightforward models, as represented by the often powerful connections between structure and protein function. IDPs require a broader framework to achieve comparable insights. As such, investigation of the properties of IDP structural ensembles will require significantly more than one dominant experimental tool and computational analysis approach than found when using X-ray crystallography beamlines. Although nuclear magnetic resonance (NMR)[73] and small angle X-ray spectroscopy (SAXS)[74] are typically used to characterize the solution structure and dynamical conformations of IDPs, the long time scale of these measurements limits the identification

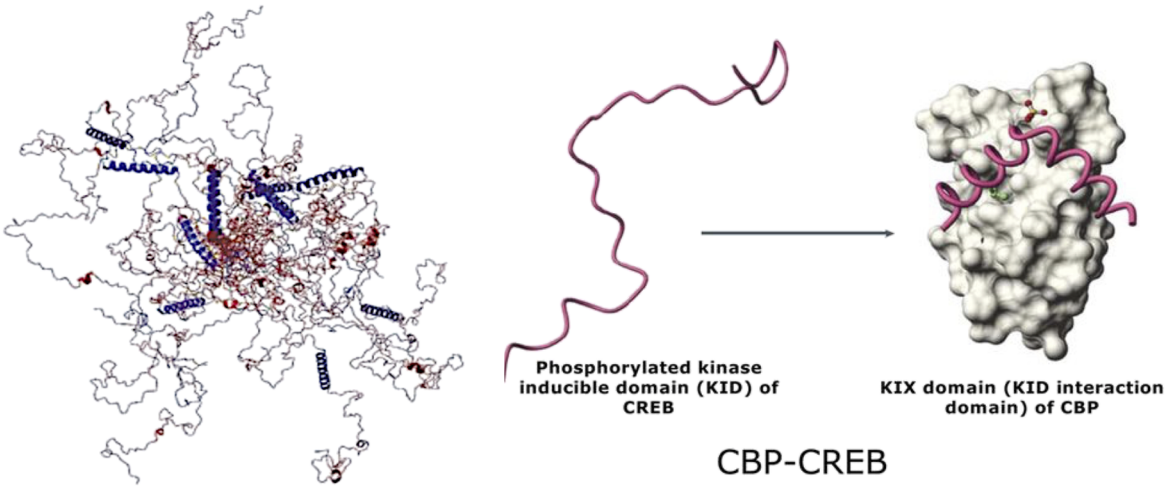

FIG. 4. (Left) Representative ensemble of conformers for CBP-ID4. Some parts of the CBP-ID4 protein take on a definite structure (red and blue), while other regions remain disordered (gray) in the unbound state. Reprinted with permission from Piai *et al.*, Biophys. J. **110**, 372–381 (2016). Copyright 2016, Biophysical Society.[70] (Right) Disordered protein regions can sometimes become structured upon binding with another protein as shown for the unstructured KID domain of CREB that folds upon binding with CBP. Reprinted with permission from Babu, Biochem. Soc. Trans. **44**, 1185 (2016). Copyright 2016 Biochemical Society.[71]

of the conformational substrates due to conformational averaging; these specific substrates are thought to have distinct functional roles, so even solution experimental approaches to IDPs are inherently underdetermined.

For these reasons, the demand for reliable computer simulations of IDPs has become increasingly intense in recent years.[72,75] However such computational tools have yet to realize their full potential due to serious software obstacles. These include the concurrency and scaling limits associated with some of the more aggressive sampling methods.[76,77] There is a need for improved force fields beyond that used for folded proteins,[78,79] which can entail additional computational expense when including many-body interactions such as polarization.[80] In turn, the trial ensembles that are generated need to be validated through back-calculation using more accurate property prediction (such as for NMR chemical shifts) derived from large-scale quantum chemical computations to compare to experiment since current heuristic property calculators[81] perform inadequately for IDPs. Finally, IDP structural ensembles must be evaluated with Bayesian and statistical tools to validate and interpret IDP ensembles,[82,83] due to the problem of underdetermination.

Therefore by analogy with X-ray crystallographic beamlines and their role in streamlining acquisition of structures, the IDP problem ultimately requires an integrative approach by combining diverse experimental data and simulation methodology into a single computational instrument, which we have previously referred to as a "computational beamline."[72] In terms of software and data, a computational beamline is currently a significant and unsolved software infrastructure challenge that requires a larger software framework and that would ideally be composed of the following elements.

Computational simulation codes such as Amber,[84] CHARMM,[85] NAMD,[86] or OpenMM[87] (to name a few) would be needed to create trial IDP structural ensembles using brute force MD, or taking up newer software modules that perform enhanced and adaptive sampling methods.[46,77,88,89] Additional software will be needed to allow a rich network of interactions to occur between the experimental data, such as NMR, SAXS, and Förster resonance energy transfer (FRET),[90] and their connection to structural or dynamical observables through back-calculations to validate the trial IDP ensemble. This would include agreement with chemical shifts and scalar couplings from more rigorous quantum mechanical codes such as CFOUR (www.cfour.de), GAMESS,[91] Gaussian,[92] NWChem,[93] Psi4,[94] or Q-Chem[95] (again, to name a few).

The co-location and integration of all the data and codes combined with the ability to run many concurrent ensembles will require sophisticated software frameworks. Fireworks,[96] Ensemble Toolkit,[97] and RepEx[98] leverage the computing power of supercomputers and mid-range clusters to accommodate the large number of runs needed to sample the large conformational space of IDPs and their complexes, including back-calculations of the experimental data for their validation. As part of the workflow, the output should be automatically curated, along with all the parameters used to create that output, for subsequent analysis.

A central data repository to integrate all available experimental and simulated data and which provides powerful and flexible search capabilities is needed. Data sharing would also occur to external IDP and NMR databases such as BioIsis,[99] pE-DB,[100] and BMRB[101] from the computational beamline repository. Making the data collected and generated by the computational beamline accessible via a central resource will significantly improve access, use, and reuse of data. Uploaded data can be processed further to build data objects and will be discoverable by relevant data attributes so that researchers will be able to find and retrieve experimental and computational data for validation, to determine constraints, and to perform advanced data analysis.

To make the data generated by the scalable workflows accessible to users for analysis, it is critical that we provide an end-station comprising robust statistical tools such

as regression, clustering, Bayesian inference, and Markov State Models, that can give insight into key structural aspects and to identify dynamical motifs, including repeated transient structure and more sophisticated correlative motions. The analysis end station could also be integrated with visualization software and utilize Jupyter Notebooks to enable a more integrated workspace where IDP scientists can collaborate.

In summary, IDPs require an unprecedented level of integration of multiple and complementary experimental data types, state-of-the art molecular simulation methodology, and a comprehensive set of statistical and data science analysis tools. Such software could connect observed structural or dynamical motifs to greater functional relevance for a wide range of IDP systems. The primary benefit from this ambitious software effort is to push IDP models closer to crystal structures in the goals of utility, understandability, and predictive power.

V. EARLY SOFTWARE EFFORTS AT THE MolSSI

MolSSI is working with the CMS community to address the software bottlenecks posed by these scientific use cases through the development of new software tools and improvements to software infrastructure. To illustrate, the eMap software developed by Bravaya and co-workers[102] is a community-led effort that has received partial support by MolSSI to address the theoretical modeling of electron transfer covered in Sec. III. The software issue stems from the scientific problem that if the crystal structures are not available or electron-transfer pathways are not immediately obvious, as in the case of photo-induced electron transfer in different mutants of the green fluorescent protein[103] shown in Fig. 5, eMap can narrow down the search of likely electron accepting residues by determining the shortest chain of aromatic residues connecting the chromophore with the surface. Currently such community led software projects are solicited through the software fellows program,[3] but this also entails a significant education program around software best practices for novices. Our expectation for the future is a process for more open and direct engagement with more senior software developers and end users to help lead MolSSI in new software project directions.

As we begin the 3rd year of MolSSI, there are currently three overarching directions in software projects that will ultimately address the larger software infrastructure needs for the grand challenge use cases described in Secs. II–IV. These include reproducibility and interoperability of different software packages, code curation efforts, as well as software development processes.

For example, manual conversion of the energy expression files between molecular simulation programs is error-prone, and consequently many automated tools have been developed to perform these conversions.[104,105] The Energy Expression Exchange (EEX)[106] developed at MolSSI is building on these efforts using an all-to-all Python package translator that converts the topology, force field and simulation parameters between two given simulation programs. The EEX uses a plug-in architecture that makes the application more modular, customizable, and extensible. The host application contains EEX's internal representation of the system and defines the interface of such representation with the external world (i.e., the various reader/writer plug-ins). Each of the MD or MC codes has an associated plug-in that interacts with EEX's host application to carry out the conversion. The EEX project will benefit the community by facilitating the interconversion of simulation inputs from one engine to another. This will allow researchers to reproduce the same calculation in different codes, and to leverage the capabilities of one software package that might not be available in other packages, or to compare two simulation codes. The agreement among complex computational tools is important for scientific reproducibility[107–109] and therefore impacts all scientific drivers reviewed here equally.

Similarly, MolSSI is involved in a rewrite of the popular Environmental Molecular Sciences Laboratory (EMSL) Basis Set Exchange,[110,111] which is a repository for basis sets used in QM calculations. As part of this project, basis sets are being curated and verified against the literature and other reputable sources. Different QM codes can have different internal data for a basis set, even though they will use the same name, resulting in computations that are not comparable between codes. The end product of the project will not only be a user-accessible repository of information related to basis sets, but a canonical source for verified data which programs can use to verify their own basis sets, and a place where users can download a specific basis set to use across many different codes, increasing the reproducibility and comparability of their computation.

The MolSSI Driver Interface[112] is a socket-based interface that enables an external driver to control the high-level program flow of QM and MM production codes. From a developer's perspective, the interface appears very similar to Message Passing Interface (MPI), with simple MDI-Send and MDI-Recv functions handling communication between the driver and production codes. The driver sends commands to the production codes, such as "receive a new set of nuclear coordinates from me" or "calculate and send the forces to me." When not performing a command, the production codes wait and listen for a new command. By sending particular commands to particular production codes in a particular order, a driver can orchestrate complex, multi-code calculations. Development of the interface was originally motivated by the challenges of QM/MM simulations, but it is sufficiently general that it can also support Path-Integral MD, *ab initio* MD, Metadynamics, and many other methods that can benefit from the cooperation of multiple codes.

Another example is the creation of a schema for quantum chemistry data. QCSchema,[113] will help define data interfaces that, if used by the community, can facilitate a more seamless environment for software components to work together. An initial draft of the schema is available at the MolSSI GitHub website and has been developed in partnership with many of the quantum chemistry code developers as well as developers of the consumers of that data—such as visualization, analysis software, and stand-alone geometry optimizers. An analogous schema for molecular mechanics and molecular dynamics data is in the planning stage.

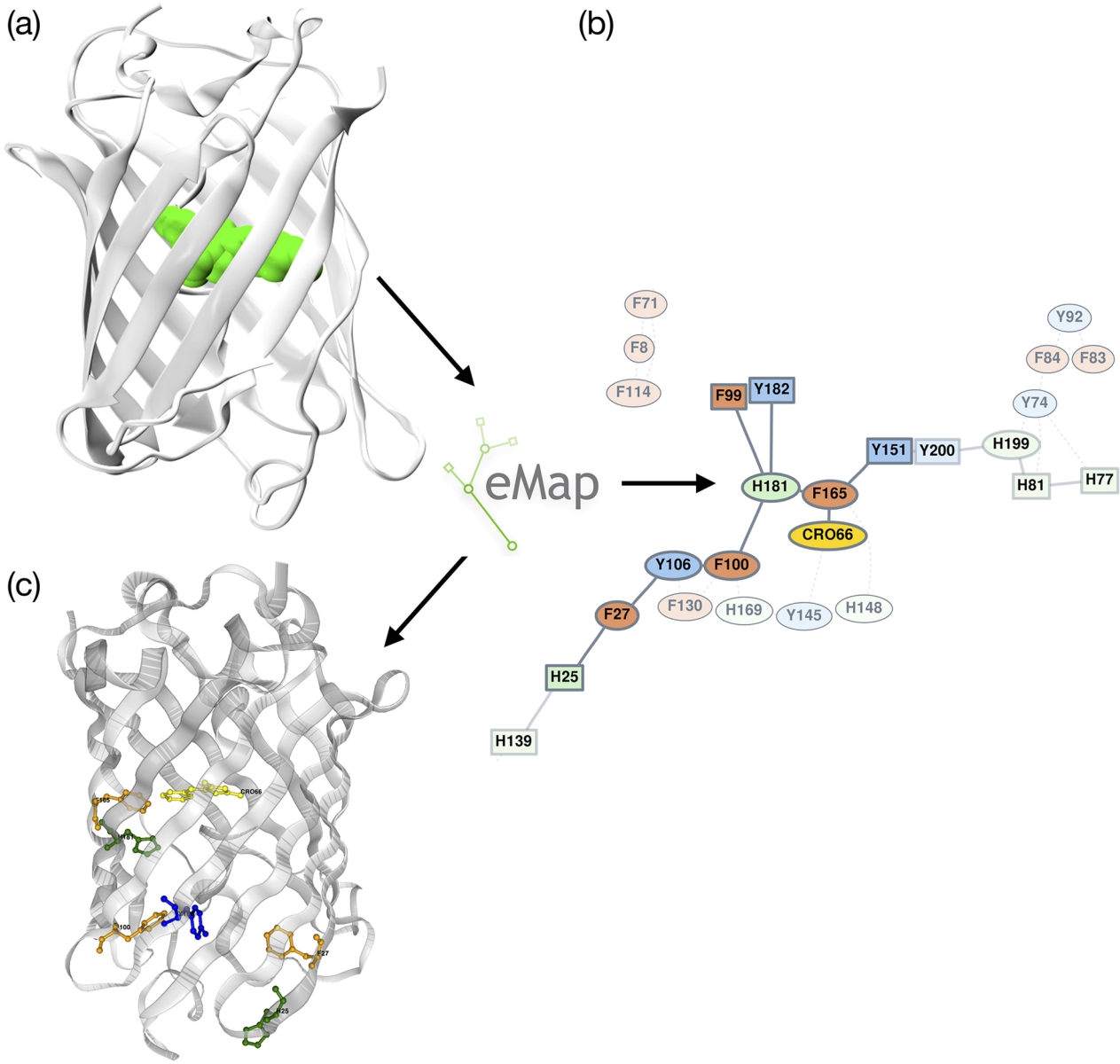

FIG. 5. Photoinduced electron transfer in green fluorescent protein. The first step in modeling the rate and yield of photoxidation requires finding a pathway of electron transfer from the chromophore (bright green) tucked inside a tight protein barrel (a) to an outside oxidant. The eMap software package determines such pathways by identifying amino acid residues that are the most likely intermediate electron acceptors and renders the resulting pathways in 3D, as shown in panels (b) and (c).

The MolSSI Quantum Chemistry Archive[114] sets out to answer a single fundamental question: How do we compile, aggregate, query, and share quantum chemistry data to accelerate the understanding of new method performance, fitting of novel force fields, and supporting the incredible data needs of machine learning for computational molecular science? The resulting project is a hybrid distributed computing and database program for quantum chemistry to make creation, curation, and distribution of large datasets more accessible to the entire CMS community. The QCSchema is used as the core transfer of quantum chemistry information to ensure that the project is not specific to any single quantum chemistry program. The project also has several distributed computing backends to choose from like Dask[115] and Fireworks[96] to ensure users can switch between flexibility and scalability as required for their projects.

Finally, the software created by the CMS community is broad and deep, and the MolSSI has developed the Community Code Database[116] for curation of community software metadata to move beyond basic Wikipedia lists of websites. The database is accessible through a web gateway as well as REST APIs, and it is searchable with many possible filters. The database includes more expanded information such as licensing, version release, requested citations(s), programming language(s), relevant compilers, graphical user interfaces, test suites and coverage, file formats, documentation, and much more. In addition, it has domain-specific information such as basis sets, element coverage, force field types, and sampling methods. Through a straightforward web interface, the CMS community can easily contribute to the database by submitting their own software products.

VI. CONCLUSION

There are many opportunities for developing the necessary models, software, and tools needed for simulating realistic catalytic systems,[50] long-range electron and charge transfer,[53] and intrinsically disordered proteins and their complexes.[72] MolSSI is at an early stage of systematically advancing software frameworks in all three of these scientific use cases. Looking ahead, MolSSI envisions its role in advancing CMS through active engagement with the community in the following software areas.

One is the more rapid deployment of the newest state-of-the-art methods and models from electronic structure, quantum dynamics, multiscale modeling, equilibrium and non-equilibrium dynamics, statistical mechanics, and coarse-graining. This would be formulated as a sophisticated set of modular software and containers for greater agility in uptake into CMS community codes.

Another is to further increase the robustness and interoperability of community codes deploying such models. At present, often the choice of computational protocols is dictated by the methodological availability in specific software packages rather than by the best theoretical considerations.

It would also be desirable for the CMS community to establish data and software standards. The timing for this endeavor is especially propitious given the emergence of data science and machine learning that will be employed in many grand challenge CMS problems, including the three discussed here.

Finally, the porting and scalability of CMS software must take advantage of multicore architectures, GPUs, and high performance computing. This is the best way to fully exploit high throughput computation and the inevitable advances toward exascale computing.[50] Of course new computing paradigms such as the emergence of quantum computing[117] will ensure that computer architectures will always be disruptive!

In summary, producing robust, scalable, and sustainable molecular simulation software requires a multi-disciplinary community of CMS domain scientists, computer science and software engineers, and applied mathematicians to advance new software initiatives. The MolSSI will provide the home and focal point for bridging and interfacing among different simulation communities to do a new level of science grand challenge problems not currently achievable within more specialized communities.

ACKNOWLEDGMENTS

The authors thank the National Science Foundation for support under Grant No. ACI-1547580. T.H.G. is thankful to Daniel Gunter and Julie Forman-Kay for discussions on experimental data and software needs for IDPs. The MolSSI team would also like to thank the many members of the CMS community for their participation in the Institute.

[1] See https://molssi.org/about/ for MolSSI.
[2] See https://molssi.org/molssi-software-scientists/ for Current MolSSI Software Scientists.
[3] See https://molssi.org/the-molssi-software-fellowship-program/ for MolSSI Software Fellowship Program.
[4] See https://molssi.org/people/ for MolSSI Board of Directors.
[5] See https://molssi.org/people/advisory-board/ for MolSSI Scientific Software Advisory Board.
[6] See https://molssi.org/the-molecular-sciences-consortium/ for Molecular Sciences Consortium.
[7] See https://molssi.org/education/summer-schools/ for MolSSI Summer Schools.
[8] See https://molssi.org/job-opportunities/ for CMS Job Postings.
[9] P. C. J. Kamer, D. Vogt, and J. W. Thybaut, *Contemporary Catalysis: Science, Technology, and Applications* (Royal Society of Chemistry, 2017).
[10] E. G. Derouane, "Catalysis in the 21st century, lessons from the past, challenges for the future," CATTECH **5**(4), 214–225 (2001).
[11] A. Primo and H. Garcia, "Zeolites as catalysts in oil refining," Chem. Soc. Rev. **43**, 7548–7561 (2014).
[12] C. A. Busacca, D. R. Fandrick, J. J. Song, and C. H. Senanayake, "The growing impact of catalysis in the pharmaceutical industry," Adv. Synth. Catal. **353**, 1825–1864 (2011).
[13] Y. Li and G. A. Somorjai, "Nanoscale advances in catalysis and energy applications," Nano Lett. **10**, 2289–2295 (2010).
[14] B. Parida, S. Iniyan, and R. Goic, "A review of solar photovoltaic technologies," Renewable Sustainable Energy Rev. **15**, 1625–1636 (2011).
[15] R. J. Lim *et al.*, "A review on the electrochemical reduction of CO_2 in fuel cells, metal electrodes and molecular catalysts," Catal. Today **233**, 169–180 (2014).
[16] J.-i. Kadokawa and S. Kobayashi, "Polymer synthesis by enzymatic catalysis," Curr. Opin. Chem. Biol. **14**, 145–153 (2010).
[17] K. C. Nicolaou, "Catalyst: Synthetic organic chemistry as a force for good," Chem **1**, 331–334 (2016).
[18] E. L. Kunkes *et al.*, "Catalytic conversion of biomass to monofunctional hydrocarbons and targeted liquid-fuel classes," Science **322**, 417 (2008).
[19] D. Baker, "An exciting but challenging road ahead for computational enzyme design," Protein Sci. **19**, 1817–1819 (2010).
[20] I. V. Korendovych and W. F. DeGrado, "Catalytic efficiency of designed catalytic proteins," Curr. Opin. Struct. Biol. **27**, 113–121 (2014).
[21] J. K. Norskov, T. Bligaard, J. Rossmeisl, and C. H. Christensen, "Towards the computational design of solid catalysts," Nat. Chem. **1**, 37 (2009).
[22] V. V. Welborn, L. R. Pestana, and T. Head-Gordon, "Computational optimization of electric fields for better catalysis design," Nat. Catal. **1**, 649 (2018).
[23] J. J. Bravo-Surez, R. V. Chaudhari, and B. Subramaniam, *Novel Materials for Catalysis and Fuels Processing* (American Chemical Society, 2013), Vol. 1132, Chap. 1, pp. 3–68.
[24] L. Falivene, S. M. Kozlov, and L. Cavallo, "Constructing bridges between computational tools in heterogeneous and homogeneous catalysis," ACS Catal. **8**, 5637–5656 (2018).
[25] J. D. A. Pelletier and J.-M. Basset, "Catalysis by design: Well-defined single-site heterogeneous catalysts," Acc. Chem. Res. **49**, 664–677 (2016).
[26] A. Bhowmick, S. C. Sharma, and T. Head-Gordon, "The importance of the scaffold for de Novo enzymes: A case study with Kemp Eliminase," J. Am. Chem. Soc. **139**, 5793–5800 (2017).
[27] E. Mansoor, J. Van der Mynsbrugge, M. Head-Gordon, and A. T. Bell, "Impact of long-range electrostatic and dispersive interactions on theoretical predictions of adsorption and catalysis in zeolites," Catal. Today **312**, 51 (2018).
[28] M. Dunwell, W. Luc, Y. Yan, F. Jiao, and B. Xu, "Understanding surface-mediated electrochemical reactions: CO_2 reduction and beyond," ACS Catal. **8**, 8121–8129 (2018).
[29] A. Zecchina, S. Bordiga, and E. Groppo, "The structure and reactivity of single and multiple sites on heterogeneous and homogeneous catalysts: Analogies, differences, and challenges for characterization methods," in *Selective Nanocatalysts and Nanoscience: Concepts for Heterogeneous and Homogeneous Catalysis* (Wiley, 2011).
[30] C. Sievers *et al.*, "Phenomena affecting catalytic reactions at solidliquid interfaces," ACS Catal. **6**, 8286–8307 (2016).
[31] S. Belsare, V. Pattni, M. Heyden, and T. Head-Gordon, "Solvent entropy contributions to catalytic activity in designed and optimized Kemp Eliminases," J. Phys. Chem. B **122**, 5300–5307 (2018).
[32] K. Kandel, S. M. Althaus, C. Peeraphatdit, T. Kobayashi, B. G. Trewyn, M. Pruski, and I. I. Slowing, "Solvent-induced reversal of activities between two closely related heterogeneous catalysts in the aldol reaction," ACS Catal. **3**, 265–271 (2013).
[33] A. P. de Lima Batista, F. Zahariev, I. I. Slowing, A. A. C. Braga, F. R. Ornellas, and M. S. Gordon, "Silanol-assisted carbinolamine formation in an

amine-functionalized mesoporous silica surface: Theoretical investigation by fragmentation methods," J. Phys. Chem. B **120**, 1660–1669 (2016).
[34]A. Garca, I. I. Slowing, and J. W. Evans, "Pore diameter dependence of catalytic activity: *p*-Nitrobenzaldehyde conversion to an aldol product in amine-functionalized mesoporous silica," J. Chem. Phys. **149**, 024101 (2018).
[35]M. R. Singh, E. L. Clark, and A. T. Bell, "Effects of electrolyte, catalyst, and membrane composition and operating conditions on the performance of solar-driven electrochemical reduction of carbon dioxide," Phys. Chem. Chem. Phys. **17**, 18924–18936 (2015).
[36]W. Kohn and L. J. Sham, "Self-consistent equations including exchange and correlation effects," Phys. Rev. **140**, A1133–A1138 (1965).
[37]J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized gradient approximation made simple," Phys. Rev. Lett. **77**, 3865–3868 (1996).
[38]N. Mardirossian and M. Head-Gordon, "Thirty years of density functional theory in computational chemistry: An overview and extensive assessment of 200 density functionals," Mol. Phys. **115**, 2315–2372 (2017).
[39]T. Helgaker, W. Klopper, and D. P. Tew, "Quantitative quantum chemistry," Mol. Phys. **106**, 2107–2143 (2008).
[40]F. Libisch, C. Huang, and E. A. Carter, "Embedded correlated wavefunction schemes: Theory and applications," Acc. Chem. Res. **47**, 2768–2775 (2014).
[41]M. W. van der Kamp and A. J. Mulholland, "Combined quantum mechanics/molecular mechanics (QM/MM) methods in computational enzymology," Biochemistry **52**, 2708–2728 (2013).
[42]X. Lu *et al.*, "QM/MM free energy simulations: Recent progress and challenges," Mol. Simul. **42**, 1056–1078 (2016).
[43]D. Marx and J. Hutter, *Ab Initio Molecular Dynamics: Basic Theory and Advanced Methods* (Cambridge University Press, New York, 2009).
[44]W. H. Miller, "The semiclassical initial value representation: A potentially practical way for adding quantum effects to classical molecular dynamics simulations," J. Phys. Chem. A **105**, 2942–2955 (2001).
[45]A. S. Petit and J. E. Subotnik, "Appraisal of surface hopping as a tool for modeling condensed phase linear absorption spectra," J. Chem. Theory Comput. **11**, 4328–4341 (2015).
[46]D. M. Zuckerman, "Equilibrium sampling in biomolecular simulation," Annu. Rev. Biophys. **40**, 41–62 (2011).
[47]M. A. Rohrdanz, W. Zheng, and C. Clementi, "Discovering mountain passes via torchlight: Methods for the definition of reaction coordinates and pathways in complex macromolecular reactions," Annu. Rev. Phys. Chem. **64**, 295–316 (2013).
[48]C. Dellago and P. Bolhuis, "Transition path sampling and other advanced simulation techniques for rare events," Adv. Polym. Sci. **221**, 167–233 (2008).
[49]U. Ray, G. K.-L. Chan, and D. T. Limmer, "Exact fluctuations of nonequilibrium steady states from approximate auxiliary dynamics," Phys. Rev. Lett. **120**, 210602 (2018).
[50]T. Windus, T. Devereaux, M. Banda, https://science.energy.gov/~/media/bes/pdf/reports/2017/BES-EXA_rpt.pdf.
[51]J. Behler, "Perspective: Machine learning potentials for atomistic simulations," J. Chem. Phys. **145**, 170901 (2016).
[52]K. Yao, J. E. Herr, and J. Parkhill, "The many-body expansion combined with neural networks," J. Chem. Phys. **146**, 014106 (2017).
[53]G. D. Scholes, "Introduction: Light harvesting," Chem. Rev. **117**, 247 (2018), Editorial for a special issue on light harvesting.
[54]K. H. Nealson, A. Belz, and B. McKee, "Breathing metals as a way of life: Geobiology in action," Antonie Van Leeuwenhoek Int. J. Gen. Mol. Microbiol. **81**, 215 (2002).
[55]J. K. Fredrickson, M. F. Romine, A. S. Beliaev, J. M. Auchtung, M. E. Driscoll, T. S. Gardner, K. H. Nealson, A. L. Osterman, G. Pinchuk, J. L. Reed, D. A. Rodionov, J. L. M. Rodrigues, D. A. Saffarini, M. H. Serres, A. M. Spormann, I. B. Zhulin, and J. M. Tiedje, "Towards environmental systems biology of Shewanella," Nat. Rev. Microbiol. **6**, 592 (2008).
[56]D. R. Lovley, J. D. Coates, E. L. Blunt-Harris, E. J. P. Phillips, and J. C. Woodward, "Humic substances as electron acceptors for microbial respiration," Nature **382**, 445 (1996).
[57]M. Y. El-Naggar, G. Wanger, K. M. Leung, T. D. Yuzvinsky, G. Southam, J. Yang, W. M. Lau, K. H. Nealson, and Y. A. Gorby, "Electrical transport along bacterial nanowires from Shewanella oneidensis MR-1," Proc. Natl. Acad. Sci. U. S. A. **107**, 18127 (2010).
[58]N. F. Polizzi, S. S. Skourtis, and D. N. Beratan, "Physical constraints on charge transport through bacterial nanowires," Faraday Discuss. **155**, 43 (2012).
[59]M. Breuer, K. M. Rosso, J. Blumberger, and J. N. Butt, "Multi-haem cytochromes in Shewanella oneidensis MR-1: Structures, functions and opportunities," J. R. Soc. Interface **12**, 20141117 (2014).
[60]J. Blumberger, "Free energies for biological electron transfer from QM/MM calculation: Method, application and critical assessment," Phys. Chem. Chem. Phys. **10**, 5651 (2008).
[61]M. Breuer, P. Zarzycki, L. Shi, T. A. Clarke, M. J. Edwards, J. N. Butt, D. J. Richardson, J. K. Fredrickson, J. M. Zachara, J. Blumberger, and K. M. Rosso, "Molecular structure and free energy landscape for electron transport in the decahaem cytochrome MtrF," Biochem. Soc. Trans. **40**, 1198 (2012).
[62]M. Breuer, K. M. Rosso, and J. Blumberger, "Electron flow in multiheme bacterial cytochromes is a balancing act between heme electronic interaction and redox potentials," Proc. Natl. Acad. Sci. U. S. A. **111**, 611 (2014).
[63]A. Barrozo, M. Y. El-Naggar, and A. I. Krylov, "Distinct electron conductance regimes in bacterial decaheme cytochromes," Angew. Chem., Int. Ed. **57**, 6805–6809 (2018).
[64]J. Blumberger, "Recent advances in the theory and molecular simulation of biological electron transfer reactions," Chem. Rev. **115**, 11191 (2015).
[65]F. Melaccio, M. C. Marín, A. Valentini, F. Montisci, S. Rinaldi, M. Cherubini, X. Yang, Y. Kato, M. Stenrup, Y. Orozco-Gonzalez, N. Ferré, H. L. Luk, H. Kandori, and M. Olivucci, "Toward automatic Rhodopsin modeling as a tool for high-throughput computational photobiology," J. Chem. Theory Comput. **12**, 6020 (2016).
[66]M. K. Lee, K. B. Bravaya, and D. F. Coker, "First-principles models for biological light-harvesting: Phycobiliprotein complexes from cryptophyte algae," J. Am. Chem. Soc. **139**, 7803–7814 (2017).
[67]P. Tompa, "Intrinsically disordered proteins: A 10-year recap," Trends Biochem. Sci. **37**, 509–516 (2012).
[68]H. J. Dyson and P. E. Wright, "Intrinsically unstructured proteins and their functions," Nat. Rev. Mol. Cell Biol. **6**, 197–208 (2005).
[69]A. K. Dunker, I. Silman, V. N. Uversky, and J. L. Sussman, "Function and structure of inherently disordered proteins," Curr. Opin. Struct. Bio. **18**, 756–764 (2008).
[70]A. Piai *et al.*, "Just a flexible linker? The structural and dynamic properties of CBP-ID4 revealed by NMR spectroscopy," Biophys. J. **110**, 372–381 (2016).
[71]M. M. Babu, "The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease," Biochem. Soc. Trans. **44**, 1185 (2016).
[72]A. Bhowmick, D. H. Brookes, S. R. Yost, H. J. Dyson, J. D. Forman-Kay, D. Gunter, M. Head-Gordon, G. L. Hura, V. S. Pande, D. E. Wemmer, P. E. Wright, and T. Head-Gordon, "Finding our way in the dark proteome," J. Am. Chem. Soc. **138**, 9730–9742 (2016).
[73]H. J. Dyson and P. E. Wright, "Elucidation of the protein folding landscape by NMR," Methods Enzymol. **394**, 299–321 (2005).
[74]P. Bernado and D. I. Svergun, "Structural analysis of intrinsically disordered proteins by small-angle X-ray scattering," Mol Biosyst. **8**, 151–167 (2012).
[75]N. L. Fawzi *et al.*, "Structure and dynamics of the $A\beta_{21-30}$ peptide from the interplay of NMR experiments and molecular simulations," J. Am. Chem. Soc. **130**, 6145–6158 (2008).
[76]A. Cavalli, C. Camilloni, and M. Vendruscolo, "Molecular dynamics simulations with replica-averaged structural restraints generate structural ensembles according to the maximum entropy principle," J. Chem. Phys. **138**, 094112 (2013).
[77]J. Lincoff, S. Sasmal, and T. Head-Gordon, "Comparing generalized ensemble methods for sampling of systems with many degrees of freedom," J. Chem. Phys. **145**, 174107 (2016).
[78]J. Huang *et al.*, "CHARMM36m: An improved force field for folded and intrinsically disordered proteins," Nat. Methods **14**, 71 (2016).
[79]P. S. Nerenberg and T. Head-Gordon, "New developments in force fields for biomolecular simulations," Curr. Opin. Struct. Bio. **49**, 129–138 (2018).
[80]A. Albaugh *et al.*, "Advanced potential energy surfaces for molecular simulation," J. Phys. Chem. B **120**, 9811–9832 (2016).
[81]B. Han, Y. Liu, S. W. Ginzinger, and D. S. Wishart, "SHIFTX2: Significantly improved protein chemical shift prediction," J. Biomol. NMR **50**, 43 (2011).
[82]C. K. Fisher, A. Huang, and C. M. Stultz, "Modeling intrinsically disordered proteins with Bayesian statistics," J. Am. Chem. Soc. **132**, 14919–14927 (2010).

[83]D. H. Brookes and T. Head-Gordon, "Experimental inferential structure determination of ensembles for intrinsically disordered proteins," J. Am. Chem. Soc. **138**, 4530–4538 (2016).
[84]D. A. Case *et al.*, "The Amber biomolecular simulation programs," J. Comput. Chem. **26**, 1668–1688 (2005).
[85]B. R. Brooks *et al.*, "CHARMM: The biomolecular simulation program," J. Comput. Chem. **30**, 1545–1614 (2009).
[86]J. C. Phillips *et al.*, "Scalable molecular dynamics with NAMD," J. Comput. Chem. **26**, 1781–1802 (2005).
[87]P. Eastman *et al.*, "OpenMM 4: A reusable, extensible, hardware independent library for high performance molecular simulation," J. Chem. Theory Comput. **9**, 461 (2013).
[88]C. Abrams and G. Bussi, "Enhanced sampling in molecular dynamics using metadynamics, replica-exchange, and temperature-acceleration," Entropy **16**, 163 (2014).
[89]R. C. Bernardi, M. C. R. Melo, and K. Schulten, "Enhanced sampling techniques in molecular dynamics simulations of biological systems," Biochim. Biophys. Acta, Gen. Subj. **1850**, 872–877 (2015).
[90]M. Aznauryan *et al.*, "Comprehensive structural and dynamical view of an unfolded protein from the combination of single-molecule FRET, NMR, and SAXS," Proc. Natl. Acad. Sci. U. S. A. **113**, E5389–E5398 (2016).
[91]M. W. Schmidt *et al.*, "General atomic and molecular electronic structure system," J. Comput. Chem. **14**, 1347–1363 (1993).
[92]M. J. Frisch *et al.*, GAUSSIAN 16, Revision B.01, Gaussian, Inc., Wallingford, CT, 2016.
[93]M. Valiev *et al.*, "NWChem: A comprehensive and scalable open-source solution for large scale molecular simulations," Comput. Phys. Commun. **181**, 1477–1489 (2010).
[94]R. M. Parrish *et al.*, "Psi4 1.1: An open-source electronic structure program emphasizing automation, advanced libraries, and interoperability," J. Chem. Theory Comput. **13**, 3185–3197 (2017).
[95]Y. Shao *et al.*, "Advances in molecular quantum chemistry contained in the Q-Chem 4 program package," Mol. Phys. **113**, 184–215 (2015).
[96]A. Jain *et al.*, "FireWorks: A dynamic workflow system designed for high-throughput applications," Concurr. Comput.: Pract. Exper. **27**, 5037–5059 (2015).
[97]V. Balasubramanian, A. Treikalis, O. Weidner, and S. Jha, "Ensemble toolkit: scalable and flexible execution of ensembles of tasks," in *Proceedings 45th International Conference on Parallel Processing, ICPP 2016* (IEEE, 2016), pp. 458–463.
[98]A. Treikalis, A. Merzky, H. Chen, T.-S. Lee, D. M. York, and S. Jha, "RepEx: A flexible framework for scalable replica exchange molecular dynamics simulations," in *2016 45th International Conference on Parallel Processing (ICPP)* (IEEE, 2016), pp. 628–637.
[99]See http://www.bioisis.net for BIOSIS, the resource for macromolecular SAXS.
[100]M. Varadi *et al.*, "pE-DB: A database of structural ensembles of intrinsically disordered and of unfolded proteins," Nucleic Acids Res. **42**, D326–D335 (2014).
[101]E. L. Ulrich *et al.*, "BioMagResBank," Nucleic Acids Res. **36**, D402–D408 (2008).
[102]R. N. Tazhigulov, J. G. Gayvert, M. Wei, and K. B. Bravaya, eMap: an Online Platform for Identifying and Visualizing Electron and Hole Transfer Pathways in Proteins (2018), https://emap.bu.edu/.
[103]A. Acharya, A. M. Bogdanov, K. B. Bravaya, B. L. Grigorenko, A. V. Nemukhin, K. A. Lukyanov, and A. I. Krylov, "Photoinduced chemistry in fluorescent proteins: Curse or blessing?," Chem. Rev. **117**, 758 (2017).
[104]S. Jo, T. Kim, V. G. Iyer, and W. Im, "CHARMM-GUI: A web-based graphical user interface for CHARMM," J. Comput. Chem. **29**, 1859–1865 (2008).
[105]J. Swails, ParmEd, https://github.com/ParmEd/ParmEd, 2018.
[106]J. Nash, E. Marin-Rimoldi, and D. G. A. Smith, The Energy Expression Exchange for Molecular Dynamics, https://github.com/MolSSI/EEX, 2018.
[107]R. D. Peng, "Reproducible research in computational science," Science **334**, 1226–1227 (2011).
[108]M. McNutt, "Journals unite for reproducibility," Science **346**, 679 (2014).
[109]J. P. Mesirov, "Accessible reproducible research," Science **327**, 415–416 (2010).
[110]K. L. Schuchardt, B. T. Didier, T. Elsethagen, L. Sun, V. Gurumoorthi, J. Chase, J. Li, and T. L. Windus, "Basis set exchange: A community database for computational sciences," J. Chem. Inf. Model. **47**, 1045–1052 (2007).
[111]B. Pritchard *et al.*, Basis Set Exchange, https://github.com/MolSSI-BSE/Basis_Set_Exchange, 2018.
[112]T. Barnes, MolSSI Driver Interface, https://github.com/MolSSI/molssi_driver_interface, 2018.
[113]D. G. A. Smith, QCSchema, https://github.com/MolSSI/QC_JSON_Schema, 2018.
[114]D. G. A. Smith *et al.*, QCFractal, https://github.com/MolSSI/QCFractal, 2018.
[115]Dask Development Team, Dask: Library for dynamic task scheduling, http://dask.pydata.org, 2016.
[116]D. Altarawy, Community Code Database, https://molssi.org/software-search/, 2018.
[117]I. Kassal, J. D. Whitfield, A. Perdomo-Ortiz, M.-H. Yung, and A. Aspuru-Guzik, "Simulating chemistry using quantum computers," Annu. Rev. Phys. Chem. **62**, 185–207 (2011).