

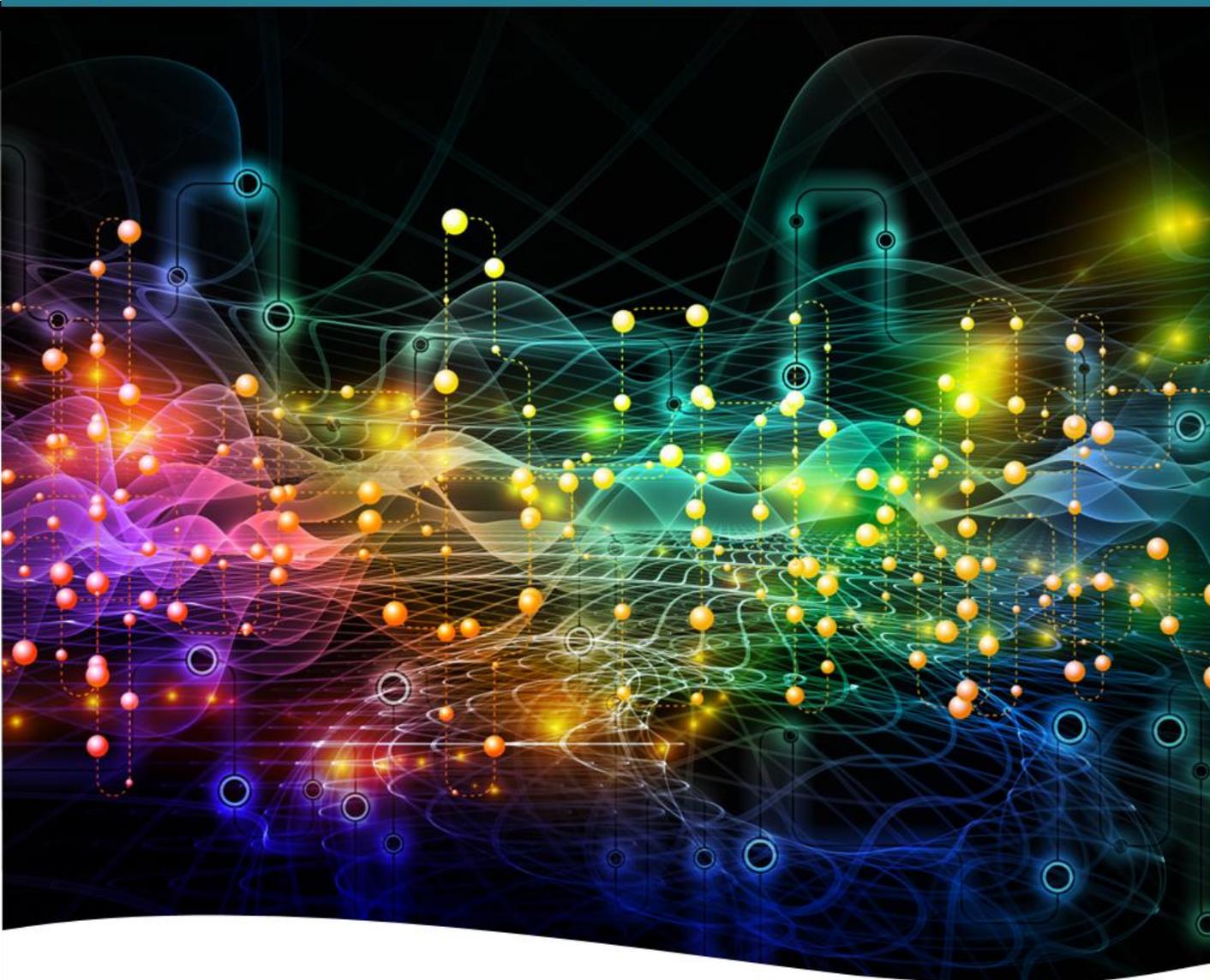
CHE WORKSHOP:

Framing the Role of Big Data and Modern Data Science in Chemistry

April 18-19, 2017 Arlington, VA



Final Report



Workshop Organizers:

Dr. Johannes Hachmann (University at Buffalo, The State University of New York)

Dr. Theresa Windus (Iowa State University)

Dr. John McLean (Vanderbilt University)

Additional Workshop Reporters:

Vanessa Allwardt (Vanderbilt University)

Dr. Alexandra Schrimpe-Rutledge (Vanderbilt University)

Mohammad Atif Faiz Afzal (University at Buffalo, The State University of New York)

Mojtaba Haghighatlari (University at Buffalo, The State University of New York)

Final Report of the National Science Foundation's Division of Chemistry Workshop on

Framing the Role of Big Data and Modern Data Science in Chemistry

Funded Under Award CHE-1733626

Disclaimer: This material is based upon work supported by the National Science Foundation under grant CHE-1733626. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the workshop participants and do not necessarily reflect the views of the National Science Foundation.

EXECUTIVE SUMMARY

The 2-day workshop “**Framing the Role of Big Data and Modern Data Science in Chemistry**” was conducted in order to spearhead a broad discussion about the role of big data research and modern data science in chemistry. The workshop set out to articulate the tremendous potential of this emerging field, to address the needs that have to be met – both now and in the long term – in order to fully develop this potential, and to offer suggestions on how this development could be supported beyond existing funding mechanisms. While there is now broad agreement on the value of data-driven approaches and the closely related ideas of rational design, there is still a significant disconnect between its possibilities and the realities of every-day research in the chemical domain. Data science and the use of advanced data mining tools are not part of the regular training of chemists, and the community is thus oftentimes reluctant to engage them. Conversely, chemical applications are generally well beyond the scope of most data and computer scientists, who are the actual experts with respect to these powerful methods. This workshop attempts to chart a path that will allow us to bridge this disconnect, to support and guide the activities of researchers, to provide consensus community directions, and to ultimately advance and shape this emerging field as a focus area. Our long-term objective is to help **pioneer a fundamental transformation of the discovery process in chemistry.**

TABLE OF CONTENTS

EXECUTIVE SUMMARY	2
I. BACKGROUND AND MOTIVATION	4
II. GRAND CHALLENGES	7
II.1. Identifying the main scientific challenges, drivers, and opportunities for big data research in chemistry.....	7
II.2. Aiding experimental and computational efforts for big data acquisition, storage, and dissemination (including advances in database technology; ontologies and semantics; hardware)	8
II.3. Adopting data science techniques for the chemical domain	13
II.4. Facilitating the use of data science for the creation of predictive models, innovative method developments, and decision making in chemical research.....	15
II.5. Coordinating the development of comprehensive, integrated, general-purpose, user-friendly tools	18
II.6. Building a community for data-driven chemistry, fostering collaborations between stakeholders, and engaging the data and computer science field.....	20
II.7. Promoting education and workforce development in modern data science for chemists	23
III. BROADER IMPACT OF DATA SCIENCE IN CHEMISTRY	26
IV. CONCLUSION.....	27
REFERENCES.....	28
APPENDICES	35
APPENDIX A: Workshop Participants & Aids.....	36
APPENDIX B: Workshop Program Schedule	38

I. BACKGROUND AND MOTIVATION

Principal Challenges. Two of the main challenges in creating new chemistry are that the behavior of chemical systems is governed by complicated structure-property and structure-activity relationships,¹⁻³ and that chemical space is practically infinite.⁴⁻⁶ Traditional trial-and-error research approaches that focus on individual compounds, materials, and chemical transformations and that are driven by experimental work are increasingly ill-equipped to meet these challenges, in particular since advanced chemical applications require more and more intricate property profiles.⁷⁻⁹ While there is obvious value in studying particular systems of interest, the insights gained in these small-scale studies cannot easily be transferred or generalized.

Opportunities. Experimentally-driven trial-and-error research is typically motivated by experience, intuition, conceptual insights, and guiding hypotheses, but it still often comes with distinct inefficiencies, shortcomings, and limitations due to its time-, labor-, and cost-intensive nature. The shift towards a *data-driven research paradigm* and the use of *modern data science* promises to mitigate many of the prevalent issues and there is now a growing recognition of the tremendous opportunities that are arising with this development. High-throughput methods can facilitate the large-scale exploration of chemical space, and its uncharted domains are expected to hold new classes of compounds and chemical transformations with game-changing characteristics. Machine learning and informatics are ideally suited to mine the large-scale data sets that result from such investigations in order to develop an understanding of the hidden mechanisms that govern chemical behavior. These insights are a prerequisite for rational design and inverse engineering capabilities.¹⁰⁻¹⁹ Data-driven research thus promises to advance our capacity to tackle complex discovery and design challenges, facilitate an increased rate and quality of innovation, and improve our understanding of the associated molecular and condensed matter systems. It will dramatically accelerate, streamline, and ultimately transform the chemical development process. The benefits of moving away from trial-and-error searches towards a rational design process have become increasingly evident. The *Materials Genome Initiative*²⁰ and other high-profile funding programs (including those from industrial sponsors) reflect this visionary development. A multitude of investments have already been made to advance big data science in chemistry and other disciplines. Past U.S. federal investments include for example the *Big Data Research and Development (R&D)* initiative started in 2010 and designed “to transform our ability to use Big Data for scientific discovery, environmental and biomedical research, education, and national security”.²¹ Three years later a *National Strategic Computing Initiative (NSCI)*, which also included “increasing coherence between the technology base used for modeling and simulation and that used for data analytic computing” as one of its five objectives.²² In addition, several other initiatives have been launched such as *NSF Earthcube* and *CyVerse* programs, focused at developing cyberinfrastructure collaboratives in geoscience and plant science respectively; the *NSF TRIPODS (Transdisciplinary Research in Principles of Data Science)* program and *DARPA’s Big Mechanism* program; or the *NIH Big Data to Knowledge (BD2K)* program, the *NSF*

Cyberinfrastructure Framework for 21st Century (CIF21) program, and the *NASA/NOAA/EPA Remote Sensing Information Gateway (RSIG)*, whose goal it is to enhance the interoperability of data.²³⁻³⁰ The *NSF Division of Chemistry (CHE)* is investing in promoting not only data-driven discovery research for an advanced understanding of chemical systems through initiatives related to NSF Big Idea “*Harnessing the Data Revolution*”, but is also providing infrastructure and offering training opportunities for workforce expansion as an active participant in *NSF Computational and Data-Enabled Science and Engineering (CDS&E)*, *Software Infrastructure for Sustained Innovation (SI2)*, *Data Infrastructure Building Blocks (DIBBs)* and the *BD Hubs/Spokes* programs.³¹⁻³⁵

At the same time, similar investments have been made globally and the European Union’s *BIGCHEM* program for example was started to enable collaborations of academia, the pharmaceutical industry, large academic societies, as well as small to medium-sized businesses in order to “develop computational methods specially for Big Data analysis”.³⁶

Finally, as of this writing, the NSF CSSI program has already released a new solicitation focused on research and tool development for an advanced data and software cyberinfrastructure.³⁷

Key Obstacles. Despite the apparent value of adopting data science for chemistry, there is still a significant disconnect between its possibilities and the realities of every-day research in the chemical disciplines. The three key obstacles that need to be addressed are: **(i)** data-driven research is beyond the scope and reach of most chemists due to a lack of available and accessible tools; **(ii)** many fundamental and practical questions on how to make data science work for chemical research remain unresolved; **(iii)** data science and the use of advanced data mining tools are not part of the formal training of chemists, and the community thus oftentimes lacks the necessary experience and expertise to utilize them (see **Fig.1**). Conversely, chemical applications are generally well beyond the comfort zone of most data and computer scientists, who are the experts on these powerful tools.

The Goal. The notion of utilizing modern data science in the chemical context is so recent that much of the basic infrastructure has not yet been developed, or is still in its infancy.³⁸⁻⁴⁰ The existing tools and expertise tend to be in-house, specialized, or otherwise unavailable to the community at large, so that **data science is practically beyond the scope and reach of most researchers in the field**. The goal is to overcome this situation, to fill the prevalent infrastructure gap, to enable and advance this emerging field by building the foundations that will make data-driven research a viable and widely accessible proposition in our community and thus an integral part of the chemical enterprise.

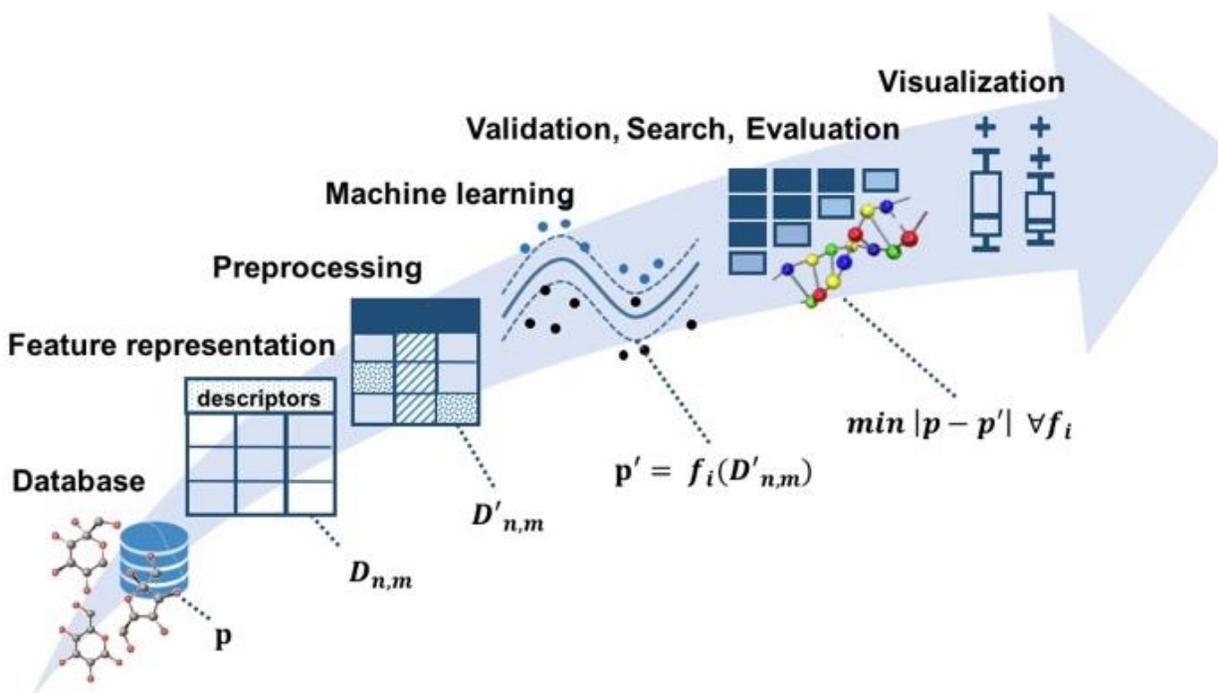


Fig. 1. A typical workflow and mathematical setup of a machine learning application in chemistry (example from the *ChemML* program package).

The NSF Division of Chemistry already recognizes and supports this paradigm shift as is evident from the recent *Dear Colleague Letter on Data-Driven Discovery Science in Chemistry (D3SC)*⁴¹, and it has signaled an interest in making it a priority. Concrete challenges that need to be addressed in order to deliver a transformative impact include:

- I.** Identifying the main scientific challenges, drivers, and opportunities for big data research in chemistry.
- II.** Aiding experimental and computational efforts for big data acquisition, storage, and dissemination (including advances in database technology; ontologies and semantics; hardware).
- III.** Adopting data science techniques for the chemical domain.⁴²⁻⁴⁶
- IV.** Facilitating the use of data science for the creation of predictive models, innovative method developments, and decision making in chemical research.
- V.** Coordinating the development of comprehensive, integrated, general-purpose, user-friendly tools.
- VI.** Building a data-driven research community, fostering collaborations between key stakeholders and engaging the data and computer science community.
- VII.** Promoting education at all levels and workforce development in modern data science for chemists.

This workshop explored the above aspects of big data and modern data science in chemistry by bringing together a diverse group of research leaders in the chemical sciences with specific interest and expertise in the development of this field, and to leverage the experience from their pioneering efforts (see *Appendix A* for a list of workshop participants).

II. GRAND CHALLENGES

II.1. Identifying the main scientific challenges, drivers, and opportunities for big data research in chemistry

In the following paragraphs, key aspects of drivers and challenges are outlined as they were discussed during the workshop:

Outreach Opportunities. The use of modern data science offers an opportunity to extend the scope of chemical research from specific scientific questions to a broader conceptual scope, thus enabling the work of the wider chemistry community. A prerequisite for these opportunities to materialize, however, is that the chemical community has to become equipped with knowledge of the capabilities of data science. Opportunities exist for gathering, analyzing, and merging vast amounts of experimental and computational data generated by labs of varying sizes, from single principal investigators to large multi-institutional centers. Further impact can be achieved by using data science approaches to dramatically lower the cost of computational research, and by integrating data-driven research into the evaluation or prediction properties of both chemical compounds and transformations. The true potential of employing modern data science is that it can yield insights beyond such individual studies, i.e., by facilitating the exploration of chemical space and by revealing underlying patterns and relationships. Chemical research is generally hampered by issues such as the complexity of processes, variable length and time scales, and incompatibility of modeling approaches. These challenges must be accounted for in the application of data science in order to build models that are capable of driving the research forward and reducing the cost and time associated with experimental research.

Scientific Challenges and Opportunities. Specific scientific challenges were discussed during the workshop, encompassing a breadth of opportunities for future data science endeavors. Representative examples include: mapping the covalent versus noncovalent chemisphere in order to apply multiscale methods connecting molecular mechanisms to cell-signaling, designing medicinal chemicals, identifying peaks in experimental spectra, determining functional descriptors leading to the development of novel catalysts for energy, and designing optic and photonic materials that have difficult to model non-linear optical properties. The ability of data science to advance analytical chemistry was

discussed during a breakout session. Chemistry's abundance of analytical data can be harnessed and organized, enabling the use of molecular features with the most predictive power to avoiding biases from human cognition. Therefore, a systematic exploration may begin with a screening that considers synthetic accessibility as well as broad regions of chemical space with high uncertainty (i.e., a high risk of synthetic inaccessibility, but high payoff if success is found). These efforts have a wide range of applications including monitoring environments, drugs, and food for safety, security and defense. The key findings are summarized as:

- ◆ Expose traditional researchers to big data and modern data science so that they may realize and further the potential applications of this developing field. Conversely, data scientists need to be versed in chemistry problems, for example by submitting data to Machine Learning competitions that chemists find important as well. Macro-exposure environments may include short-courses, conference presentations, publications, and symposia. Micro-exposure environments could include collaborations and direct integration and acceptance of data scientists into experimental research environments.

II.2. Aiding experimental and computational efforts for big data acquisition, storage, and dissemination (including advances in database technology; ontologies and semantics; hardware)

Background. Experimental and computational high-throughput screening are used to explore a variety of research areas including drug discovery (combinatorial biochemistry), bioassay screening, polymer science (e.g., organic semiconductors, photovoltaics, energy harvesting), organometallic catalysts, and mechanistic applications (catalysis). High-throughput screening research often requires a multi-disciplinary team (e.g., robotics, chemistry, biology, data science) to generate a broad, diverse set of data that is publicly accessible and manipulable. In addition, maintenance of this data for re-evaluation or novel assessment is a crucial component of data science. This should generate an appropriate amount of data necessary for downstream analyses (e.g., machine learning) as acquisition of orthogonal data is important for differentiating relevant from irrelevant information and for refining models. Though experimental datasets tend to be much smaller than computational datasets, the abundance of data, particularly for -omics measurements, is arduous to analyze given the speed of instrumentation for acquiring multi-dimensional data relative to the human time required to interrogate complex systems. Further, since analytical and biological variability is a concern for experimental screening efforts, metadata and sample variables are used to assess biases or batch effects that may be masked and thus maintaining knowledge of them is imperative.

Expectations to remain transparent and disseminate data have been realized, but a consensus of best practices for data acquisition, storage, and dissemination has yet to be achieved or defined. Data sets continue to become more complex and larger in size. New, improved, or faster computational methods are useful for high-throughput screens. However, there is often little incentive for authors to develop, maintain, and publish software for the community because of the time and effort involved.

The following paragraphs elaborate on key aspects of big data acquisition, storage, and dissemination as they were discussed during the workshop:

Data Access. One of the cardinal problems of data-driven research in chemistry today is access to suitable data sets. This mirrors, to a certain degree, the situation of the cheminformatics and quantitative structure-activity relationship (QSAR) field during its heyday in the 1990s.⁵ This field was in some sense well ahead of its time, but it often lacked in key aspects, including access to training data with the necessary volume and veracity.⁴⁷ (It also often had to rely on early, relatively simplistic data mining techniques.) These issues had a negative impact on the utility and reputation of the field.

Natural Language Processing and Machine Learning. In the wake of the booming field of bioinformatics and the *Materials Genome Initiative*, there have been concerted efforts at solving the data volume and veracity problem (both in chemical and materials research), e.g., by combining *first-principles* electronic structure theory with high-throughput computation and by combining robotics with chemical synthesis and characterization.^{48,49} A significant portion of the data of interest is only available in the literature. While there has also been early progress in automated text recognition applications for the extraction of structured data from the published literature,^{50,51} there is still much room for improvement in literature data mining. There is a critical need to use natural language processing and machine learning to derive more meaningful information beyond merely identifying chemicals in text, such as adding context (e.g., the chemical's role in synthesis – precursor, catalyst, coordinating solvent). Even so, the generation and collection of large-scale data sets has consequently never been easier than today.

Data Complexity and Retention. Workshop participants note that they desire access to comprehensive data collections including legacy data, however, logistical concerns exist. Computational data can in principle be reproduced on demand (albeit with some cost of effort) so the desire to retain inputs as well as primary results is often present. Many other properties are computed incidentally. Quantum chemistry methods are already getting more complex (combinations of approximations, multi-step local correlation models, etc.), so the ability to store the relevant cutoffs and tolerances to guarantee reproducibility will become even more important in the future. Access to legacy data is valuable – even in cases where the underlying methods may not be state-of-the-art any more – as it allows the

community to build on prior results in order to streamline the exploration of new areas of chemical space, re-mine old data for new applications, re-evaluate original findings with respect to their errors and predictive uncertainty, and compare data with future models. Another important aspect of the reusability of legacy data is its annotation with meta-data.

Data Storage Resource Needs. The availability of and access to legacy data sets are still difficult issues. Data collections oftentimes remain siloed in the groups that generate or compile them for a number of reasons, including ownership considerations, desire for competitive advantages, but also due to lack of a central repository, i.e., a physical infrastructure that would make data sharing practically viable. In addition, groups that generate large-scale data sets often face the problem of storing their data in the first place, as it is difficult to apply for such resources through regular research grants. The *Harvard Clean Energy Project*⁵²⁻⁵⁵ for instance generated about one petabyte of results from density functional theory calculations on organic semiconductor compounds for photovoltaic applications. The storage of this data was only financially viable through a generous donation by the hard drive company Hitachi/HGST and by constructing an in-house, low-budget storage array (see **Fig. 1**).

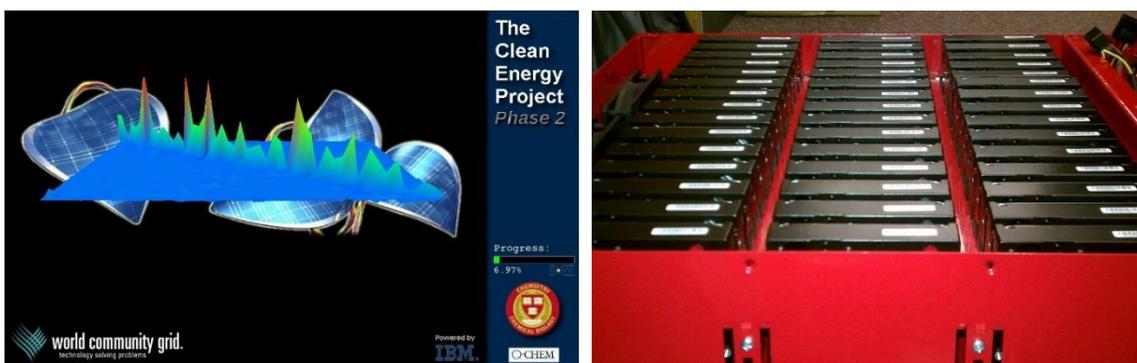


Fig. 1: The *Harvard Clean Energy Project* has harnessed distributed volunteer computing via a screensaver application of the IBM World Community Grid to generate quantum chemistry data on organic electronic compounds at a massive scale. The disc-based, home-built storage solution for this project called ‘Jabba’ is shown on the right.⁵²⁻⁵⁵

However, such a storage solution (not to mention the corresponding backup) is generally not accessible to most research groups, i.e., data sets generated as part of data-driven investigations may not be stored (at least not in their entirety), which represents a significant loss and missed opportunities for the field. When research teams do make their data available through website front ends, the richness and accessibility of the corresponding database backend is typically lost (see **Fig. 2** and the *NIST WebBook*⁵⁶ for examples).

Data Quality and Accuracy. Database content and inaccuracies are a concern during evaluation of many high-throughput findings, including –omics analyses. There are a

multitude of open databases and libraries that can be used for the purposes of screening and non-targeted analyses. However, many have issues in terms of data quality. Though they are recognized as imperfect, these databases are accepted and widely-used as they are free and widely accessible. A need exists for data-checking, manual curation to validate content, and to indicate uncertainty estimates as it is not uncommon for different answers to be generated for the same data. To the extent possible, automated validation tools should be developed to ensure integrity and internal consistency, such as with the *Protein Data Bank (PDB)* and the *Cambridge Crystallographic Data Centre (CCDC)*.^{57,58} In addition, data repositories should be archival (e.g., who entered what data, when, and how), provenance/audit logs should be retained and unusual or irreproducible results should be reported. Ideally, the community should push for research standards via a peer review mechanism. Indeed, even failed or negative results are of great use in the machine learning context⁵⁹, if properly annotated, and the gathering and curation of the failed data should be encouraged. One of the data sharing issues discussed is that constraints placed by the journals on the type/extent of data to be published may increase the barrier to entry for publication. For many types of data (e.g., in crystallography) it is an established norm that data is published in community data hubs as a prerequisite for publication. This publisher-accepted approach could likely be leveraged, though a challenge is to broaden it to new types of data. The latter will require the formation of multidisciplinary teams that can successfully implement such repositories in other fields, which will then allow researchers to easily share, publish, and extract open data in a centralized fashion.

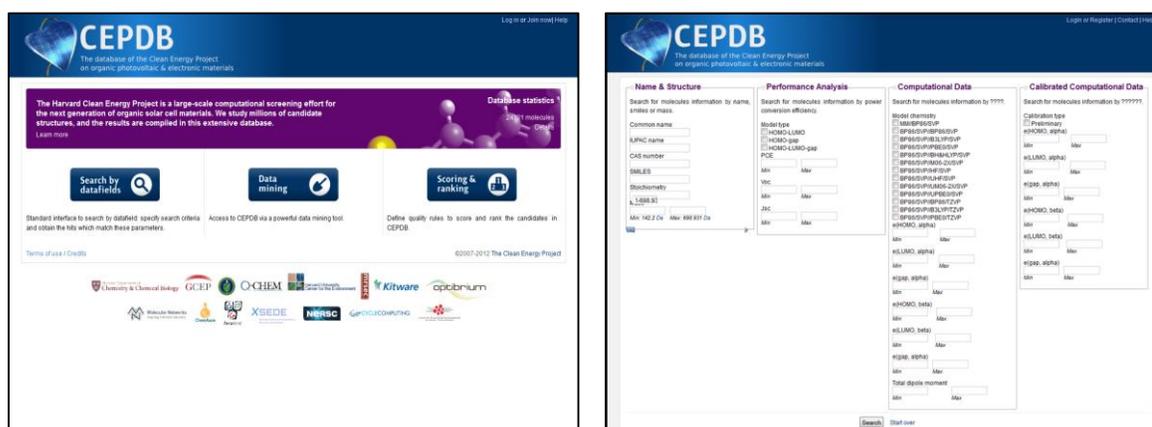


Fig. 2. Web frontend of the *Harvard Clean Energy Project Database*.⁵²⁻⁵⁵

Data Sharing Limitations. For data analysis, the backend is considerably more valuable. However, complete database dumps are rarely shared by the owners of such data sets. Complete databases or even raw data compilations may also be too large for electronic data transfer and may thus require physical shipping. For instance, the *Harvard Clean Energy Project* shared about 10TB of its data with the *Open Chemistry Project* for benchmarking and testing purposes, and the only viable option for the data transfer was shipping of hard drives by mail. This and similar situations could be avoided if the analysis and mining work

on the data were performed on site, where the corresponding tools have direct access to the underlying data architectures (see example in **Fig. 3**).

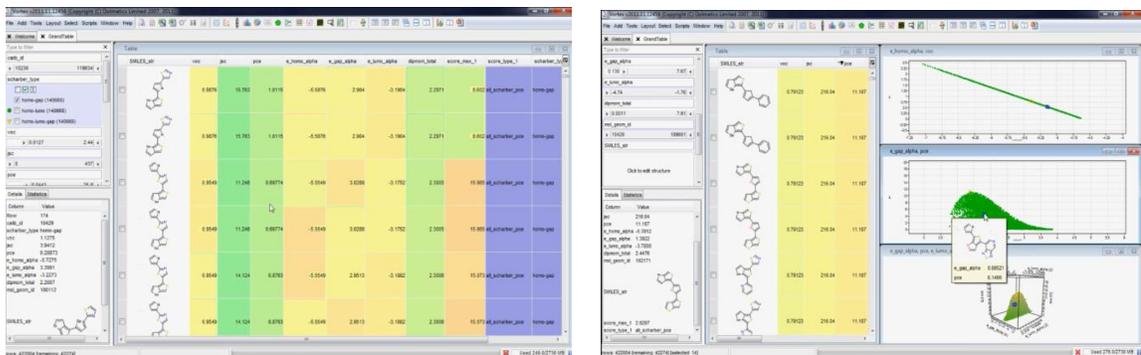


Fig. 3. Application of the Vortex drug discovery data mining tool on the Harvard Clean Energy Project Database backend.⁵²⁻⁵⁵

Differences in Data Formats. Finally, data sharing has been established for several disciplines, with funders setting expectations and publishers driving community norms. Within the materials community there are several repositories of data aimed at materials genome applications. Similarly, there are several different formats for sharing chemical structures and data. A standard format/database for data within the chemical data-driven discovery field is highly desirable. As the field of data-driven research is still relatively young and most researchers have not had the benefit of formal training in data science, there is a lack of established (or at least widely adopted) data standards, formats, architectures, etc. There is also very limited experience with domain specific issues with respect to hardware, database management systems and engines, etc.

Data in Supplementary Materials. Many publishers have traditionally been satisfied with the generation of PDFs of supplementary materials, which severely limited the accessibility and utility of the data it contained. However, there is a distinct shift towards requiring comprehensive compilations of supplementary data (including details of both physics- and data-derived models) in formats that are readily accessible and reproducible. As such, the community should continue advocating and pushing for improved accessibility, and enable data parsing by making available supplementary materials (e.g., *Jupyter* notebooks, *Python* scripts, *Docker* containers, databases, electronic notebooks⁶⁰). Data collections should be housed on publicly accessible sites and associated with digital object identifiers (DOIs), so that they can be adequately cited. To support this concept, perhaps grants could include researchers from different fields such as computer science. These supplementary material repositories may potentially be funded through public-private partnerships, i.e., by the funding agency that supported the research and by the publisher that disseminated the results.

In summary, it has become evident that the lack of a central, shared hardware infrastructure that hosts the important data sets of the chemistry community, that provides access and a storage solution for this data, and that offers an on-site platform for data mining and the exploration of the afore-mentioned issues, is a major roadblock on the path to progress in this emerging field. Recommendations to the community to resolve these roadblocks include:

- ◆ Implement a consensus of best practices for data acquisition, storage, and dissemination while the field is still young.
- ◆ Develop a community published data hub to ensure transparency. Users should be able to share and extract open data in a centralized fashion.
- ◆ Promote a push for research standards via a peer review mechanism for data quality; motivate publishers to drive community norms.

II.3. Adopting data science techniques for the chemical domain

Background. One of the greatest assets of data science is that it is (at least in principle) agnostic to the question it is applied to. There is thus no good reason why the successes of data science in other domains should not translate to chemistry. In fact, applications in the natural sciences are on much firmer footing than, e.g., in the social and behavioral sciences, since they are dealing with much more deterministic questions. There is also the option to incorporate fundamental physics into data science approaches. That being said, there are many challenges for how to adapt data science techniques for the chemical domain, in particular in the area of feature representations, dealing with the peculiarities of chemical data (size, bias, etc), model selection, and efficiency.

Data Set Variations. Even with data being accessible, curated, and manipulatable, the type and amount of data and the data representation and structure to be used in the ML algorithms are essential components to the success of applying the algorithms to the scientific problem to be explored. In chemistry, these components can be highly dependent on the type of problem to be solved. Some of the research areas such as reaction mechanisms have small data-sets with biased distributions that only include positive results, while others such as molecular dynamics simulations and structure data involve very large sets of data that are dependent on the parameters of the simulations and experiments. For the former case, the workshop participants suggest that one-shot methods that are more interpolative instead of extrapolative may be appropriate. For the latter case, there may be a need to reduce high-dimensional data in a principled way, while avoiding overfitting. In addition, integration of different data sources is challenging, especially with contradictory data in the sources. Consequently, there is an essential need for research to identify the types of data for the spectra of data and problems to be examined.

Descriptors. While there has been some progress in identifying appropriate descriptors for specific chemical contexts, more research is required to identify new descriptors – and perhaps identifying the distinction between structural (connectivity, spacial orientation), property (electronic, thermal, dynamic), and functional descriptors (catalytic activity and selectivity). However, as the wide-ranging viewpoints expressed by the workshop participants attest, it may be too early yet to identify which representations and feature spaces will yield results across the varied needs of the chemical ecosystem. Therefore, there is a need for a continuum of representations from raw data to higher representations that humans can appreciate. Even with the descriptors that have already been identified, there are issues of nontransferability of descriptors across types of matter. For example, descriptors that work well for organic molecules fail for inorganic solids and molecules. Only by creating a wide pool of descriptors and using them for various scientific purposes will the answers be found for what descriptors will perform best. Therefore, additional research will be required to make significant advancements in the use of machine learning techniques.

Data Bias. Related to the representations is the question of which data to include. For example, data with lower inherent error bars should not be used on the same footing as data with larger error bars. Not all data even has associated error bars, such as results from some computational methods. While this is manageable within the machine learning framework, it may cause unintended biases in the data. In addition, inclusion of negative results is useful in machine learning, but these are often not available in the data, which may result in biases that influence the algorithms. Resolving these issues requires close collaboration with data scientists and statisticians. Additionally, benchmarking studies that produce at least average error bars for data without error bars are required.

Hybrid Models. There is significant interest in exploring the benefits of merging physics-based and data-derived prediction models.⁶¹ This hybrid approach is expected to yield more robust, reliable, and accurate models with a greater range of applicability, as the underlying physics provides the correct framework for the overall approach. A related concept is that of delta models that bridge the gap between physics-based modeling results of idealized systems and observations in complex, non-ideal reality. There are also downsides of incorporating physics into data-derived models, i.e., cost and potential bias of the physics-based components (e.g., issues of DFT results would propagate into derived models). There is thus also a legitimate space for pure data-models.

Summary recommendations to the community for the transformation of data for machine learning purposes in the chemical domain thus are:

- ◆ Develop a wide pool of publicly accessible data and descriptors to be used in various chemical machine learning contexts to determine appropriate descriptors for particular chemical problems.
- ◆ Develop centers of excellence that combine chemists (experimental, computational and theoretical), data scientists and statisticians to resolve issues associated with data bias and uncertainty.
- ◆ Encourage benchmarking activities to produce better data with error bars to calibrate machine learning techniques and research.

II.4. Facilitating the use of data science for the creation of predictive models, innovative method developments, and decision making in chemical research

Machine Learning Applications. There are many different “scales” along which to invoke a machine learning approach such as solving the Schrödinger equation, predicting solvation energies, determining accurate molecular properties with a speed comparable to molecular mechanics, predicting reaction products, guiding experimental investigations, etc. Outcomes from each level can serve as inputs to a higher level that includes more complexity of the target system. For example, information from quantum mechanics-based benchmark sets (calculated in the traditional manner) can serve as data for machine learning in deriving density functional methods⁶²⁻⁶⁴ or predicting atomic interactions^{65,66} that do not have a traditional functional form. However, machine learning should not be narrowed down to applications in quantum mechanics, molecular dynamics, or material sciences but be broadly defined as an approach to investigate chemical systems that feature a high chemical complexity. Applications may also include environmental chemistry or chemical analyses of biological samples that are internally governed by many reaction pathways and/or interact with their chemical environment.

Hypothesis Extraction and Decision Making. Additional research is needed to support data-driven studies by extracting higher-level emergent models from data or to identify hypotheses of interest for transfer of knowledge from data scientists to experiments. In addition, high level models can be developed from the data using prior knowledge of systems in a hypothesis-driven approach. Important questions include:

- What is the mathematical structure of the data?
- Why is that math working?
- What is the physical interpretation?

Statistical and Machine Learning. Pure statistical learning can help discover interesting topics to explore in a discovery-driven approach. Machine learning tools can help in

exploration of chemical space in a more systematic and less-biased manner and, hopefully, increase creativity at the same time. For example, inductive logic programming can be used to generate new hypotheses automatically and in synthesizing contextual information.

Uncertainty Quantification. Related to hypothesis- and data-driven research is the need to decrease uncertainty in order to develop and improve decision-making processes. Active learning and domain learning can help in identifying outliers in data sets, determining the biggest factors of uncertainty in the models and evaluating systematic errors versus white noise. This uncertainty quantification helps evaluate the predictive capabilities of the developed models and provides a basis for decision making. The workshop suggests that examples of questions that could be answered in a decision-making process include:

- **What should be done next?** Data methods could help determine new experiments, improve models, find model problems that will generate significant insight at a low cost, evaluate the predictive capabilities of the models, plan synthesis, or determine when to stop an unprofitable line of research.
- **How should an experiment or computation be performed?** Machine learning could be used for synthetic accessibility and cost/difficulty metrics, or as a scientific digital assistant/advisor that advises on the best model and manner to perform a calculation in complex situations.
- **Was there value added?** For example, did a new technique provide a substantial advantage over existing alternatives? Machine learning could give provable quality of advice by providing confidence bounds, theoretical guarantees, etc.
- **Why is the prediction working?** Current work on attribution or interpretation of nonlinear models to understand important features needs to be continued to extract design rules latent in the data.

Data Interpretation Approaches. The latter question directly relates to the need to create models that are interpretable. Often machine learning is primarily heuristic knowledge of correlations, whereas physical interpretation is needed to extract causality from these correlations. While chemically relevant descriptors can help, they insufficiently capture physical interpretations from the machine learning models. Workshop participants are not in agreement on the appropriate approach to obtain interpretability. One tactic would be to combine machine learning approaches for predictions with Hamiltonian-based methods to understand the physical properties. Another approach could be to build physics into the models. This can be accomplished, for example, by using chemical knowledge and prior insights into the chemistry through constraints in the model or by ensuring that model components are chemically relevant. For example, in a neural network model, latent layer engineering might add an atomic latent layer that discovers insights that might look like oxidation number, coordination, etc. Then a diatomics-in-molecules latent layer could

discover bond length/strength correlations.⁶⁷ By building multiple layers, chemical insights could be obtained.

Collaborations for Continuous Machine Learning. Inherent in all of these approaches is the deep integration among experiment, computation and data modeling. As the data modeling makes new predictions, experiments and computations will be required to verify or reject the predictions, which will then need to be propagated back into the algorithms. This includes predictions of both positive and negative results if the machine learning algorithms are to be improved to the point of predictability for a large variety of situations. These types of collaborations could be supported through the development of centers of excellence in data science.

Machine Learning Standards. Standards for determining the correct machine learning approach for a given application will be necessary to enable broad applicability within chemistry. It is also useful to identify poor or weak machine learning techniques to facilitate reviews and development in the field. While there is still significant research required to determine these standards, collaborations with data scientists and knowledge gained from machine learning in other fields will facilitate these standard development activities. Ensuring that the train/validation/test sets are chosen responsibly is a well-developed and understood method in the data sciences. Identifying the number of layers and nodes in a network and understanding the expense of training in the data generation are all considerations that must be taken into account. Finally, version control methods will be required to ensure that results are reproducible, especially if on-the-fly generation of models are used.

Summary example requirements include:

- ◆ Research is needed to understand what makes for useful and reproducible data-driven studies.
- ◆ Uncertainty in predictions must decrease to enable decision-based processes. This will require collaborative efforts among experimental and computational chemists, data scientists and statisticians.
- ◆ Research is needed to enable interpretability of the machine learning models.
- ◆ Standards for determining the correct machine learning approach for a given application will be necessary to enable broad applicability within chemistry.

II.5. Coordinating the development of comprehensive, integrated, general-purpose, user-friendly tools

Background. The workshop calls for the creation of open, general-purpose software tools for big data analysis (i.e., the use of machine learning, informatics, and database technology for the validation, mining, and modeling of resulting datasets).⁶⁸ A key consideration is to make these tools as comprehensive and user-friendly as possible, so that they can readily be employed by interested researchers without the need for excessive expert knowledge. This implies the use of grey-box solutions that provide established workflows and default settings that encapsulate best practices, while simultaneously giving users the flexibility to fully customize their work, if desired, and thus explore the largely uncharted utility of data science in chemistry. In addition to delivering production-level capabilities to the community, these tools should also provide development evaluation facilities for innovation in the underlying methods, algorithms, and protocols. These would allow the community to gain insights into the performance of existing techniques as well as new ones that emerge as the field is evolving. Approaches to establish guidelines and best practices that will provide added value to these tools is another area of interest. Many of these developments will be driven by concrete molecular design problems, which will allow the community to assess the efficacy of these new research approaches. The bottom-up formulation of grand challenges will help to move the field forward.

In the following paragraphs, key aspects of open software tools and tool development are outlined as discussed during the workshop:

User- and Contributor-Friendliness. There are many examples of available codes (e.g., in genomics) that are limited to in-house use due to a lack of user-friendliness and intuitiveness. It is generally agreed that user-friendliness is a central concern, as tools are oftentimes not created to serve the community at large. There are many interesting methods, algorithms, and techniques that are often limited to the creator's own group. Modular software design and the use of well-articulated libraries are the complementary issues for contributor-friendliness, as the community seeks flexibility and extensibility of software packages and the broad buy-in by stakeholders. Other factors are documentation, tutorials, training resources, and the development of (virtual) support communities (e.g., software forums, email lists, blogs). Workshops, online courses, and other means of user- and developer-community building also tie in with the issue of user- and contributor-friendliness and are discussed in more detail in Sec. 2.6 and 2.7.

Extensibility and Sustainability. As previously stated, there is a need to encourage a broad buy-in and engagement in the software development process. The community should strive for a smaller number of comprehensive, high-performance program packages developed by larger teams rather than a large number of single-use scripts with low

performance and very limited capabilities developed by individuals. However, there is value in having some code diversity and the difficulty of reconciling the interests and preferences of many stakeholders, while acknowledging that constantly reinventing the proverbial wheel and starting projects from scratch is not a productive development model. A notable parallel to the computational chemistry field is the increasing number of quantum chemistry log file parsers, molecular dynamics post-processors, and viewer applications. To address this issue, the community should strive for the creation of infrastructure that is modular, library-centered, and that offers better application program interfaces. Patches and features that are already developed can then be easily integrated. Moreover, adding technical information to the code, guides on how to contribute, and teaching best practices can motivate users to extend existing tools rather than redevelop them. Easily extensible codes are likely to be more sustainable, too. The consistent use of file formats, well-documented manuals, and regular software workshops/tutorials are also pointed out as practical approaches to achieve sustainability. The commitment for the support is an important concern, since the extensibility and sustainability of the software development efforts depend on a longer project assurance, and ideally will include hiring staff programmers. It is crucial to have careful project management and a good point-person, e.g., professional software developer, who can take incoming contributions and standardize, test, document, clean them up, etc. For open source projects, it is unrealistic to ask developers to also provide the support. Access to experienced staff programmers would thus be very desirable, but may not be a realistic expectation from a funding perspective. Adequate training (including in software engineering best practices) for the graduate student researchers that often perform the hands-on implementation of new codes is another important issue in order to achieve extensibility and sustainability. Finally, the workshop discussed the cultural barriers that can potentially prevent a major contribution to existing development efforts. When it comes to evaluation, researchers usually prefer to see the original contribution (i.e., their own implementations), which potentially causes conflicts of interest (e.g., regarding the uniqueness, attribution, tenure, citations, or convenience). This workshop suggests that evaluation metrics can be changed for software developers (e.g., by number of downloads, contributions, or software citations) to encourage collaborations or at least to lower contribution barriers.

Validation. Proper software and method validation is another significant aspect in the successful creation of open software tools. The validation can be achieved by developing good test cases with corresponding data sets, creating blind competitions (similar to the crystal structure prediction field), and allowing for user comments such as Github issue tracking, forums, wikis, etc. Consistent benchmarking of codes and techniques will also help in delivering higher quality tools to the community. One significant challenge is the interplay of commercial, closed-source codes, as companies can be reluctant to provide benchmarking and validation.

Funding and Resources. Several participants are concerned about the existing resources and funding needed to realize the above points. There are no straightforward funding mechanisms, in particular for the long-term support and maintenance of open-source software. Many development efforts in the community have been supported in an *ad hoc* fashion or were performed without support. Developing and maintaining open source software takes time and energy, so it is understandable if groups are reluctant to take their tools mainstream and forfeit technical advantages. This workshop recognizes the importance of adequate incentives to develop and support open source code from funding agencies and other stakeholders.

Overall, there are many important lessons – both positive and negative – to be learned from the field of computational chemistry. The field of data-driven research can thus build on 50 years of experience from computational chemistry tool development, and can avoid mistakes that result from a completely organically evolving field that did not have a template to follow. The main concerns covered in this section can be addressed only if the community and funding organizations work together to tailor the field in the right direction:

- ◆ Create open, general-purpose software tools.
- ◆ Make these tools as user-friendly as possible to reach non-expert users.

II.6. Building a community for data-driven chemistry, fostering collaborations between stakeholders, and engaging the data and computer science field

Background. While data-driven research is important in all scientific areas, data-driven chemistry is not yet well established as a cohesive and distinct field. A particular challenge is that data-driven chemistry draws its stakeholders from many different areas that represent traditional chemical domains (which at times have co-existed in a siloed fashion). There is a need to develop a core community that will drive the new field of data-driven chemistry (e.g., by creating the scientific foundations, techniques, and tools that underpin it). At the same time, it should also advance the notion that data-driven research can play an important role in all branches of chemistry and that it should become a ubiquitous approach in the every-day chemical enterprise at large. To achieve such a widespread adoption in chemistry, the core community should strive to democratize the use of data science (similar to the approach the computational chemistry community has been pursuing) and thus maximize its reach and impact.

Opportunities for Interaction and Collaboration. One of the consequences of the absence of a well-established community is the lack of opportunities to interact, exchange ideas and experiences, and collaborate. This is, e.g., reflected in the deficit of dedicated topical conferences, meetings, and workshops that would bring together investigators from

different backgrounds with a common interest in data science. Forging alliances between academia, industry, and government (including international partners) is another important action item. Partnerships and joint ventures on all levels (ranging from specific studies by small research teams to large-scale initiatives by institutions and agencies) will allow the community to better identify and frame common problem settings for which common solutions can be developed. There are great opportunities for cross-fertilization and transfer of knowledge between different domains that can be facilitated by researchers of different backgrounds. The diversity from which the data-driven chemistry community will be able to draw is clearly a strength. The common issues of the underlying chemical data problem may thus help bridge gaps of traditional fields for which data science questions represent a unifying theme.

Collaborations with Computer and Data Scientists. A related concern is the engagement of the data and computer science community, which has obviously much expertise to contribute to the field of data-driven chemistry. A key prerequisite is that communication barriers have to be torn down or at least reduced. The challenge for the data-driven chemistry community is to encourage interactions with computer and data scientists, and to identify ways in which these interactions can yield more valuable outcomes. Given the differences in priorities, perspectives, and educational background, this is a non-trivial task and may require new incentives. Interdisciplinary workshops and funding programs that specifically target collaborations between chemists and data scientists are two ideas that were offered by the workshop. The workshop participants suggest including investigators from the computer and data science field in planning meetings to harness their critical input and offer them the opportunity for active participation in the implementation of the resulting action items.

Institutionalized Support Framework. In order to facilitate the goal of community building and the fostering of collaborations in a sustained and lasting fashion, an institutionalized support framework is required. Such a framework should include a center-level structure and corresponding funding to provide data hardware, data tool building, data education and outreach, and the data science support of chemistry grand challenge projects. The *Molecular Sciences Software Institute (MolSSI)* that was launched in 2017 for the computational chemistry community could serve as a template for the data-driven chemistry community. In fact, MolSSI and the data-driven chemistry community share many common goals – for instance in establishing data standards and accessibility. Other suitable avenues would be a topical *Big Data Innovation Spoke* or *Center for Chemical Innovation*.^{69,70} In the spirit of a true community effort, any center-level initiative should cast a wide net to engage as many leaders and pioneers of this emerging field as possible. They all have a vested interest in the success of the field and can contribute their unique expertise. Workshops and planning meetings can be used to gain initial input and subsequent feedback. Any interested stakeholder should have the opportunity to actively

contribute to its implementation and the corresponding creative process. Contributing partners should also be rewarded for their work by having access to the resources created as part of this initiative. In addition, other funding mechanisms that address the specific needs for data-driven research, e.g., for data storage and mining hardware as well as the corresponding software developments are required. Another cornerstone of the institutionalized framework of a new community could be a new topical subdivision of the *American Chemical Society*, ideally affiliated with all the relevant divisions, i.e., *PHYS*, *COMP*, and *CINF*. The workshop participants recognize that by building an institutionalized support framework the community will be in a stronger position to meet the needs – both now and in the long term – in order to fully develop the potential of this emerging field.

The MGI Template. One important question is how to define success for an emerging community, field of research, and potentially a corresponding center-level effort. There is much to learn from the successes and failures of the *Materials Genome Initiative (MGI)*.⁷¹ The MGI had the clear missions of “materials discovery” and delivering for the “marketplace”. Similarly, data-driven chemistry is interested in fundamental methodological advances, the application of these methods to real-world problem settings, and tangible successes (i.e., the creation of new chemistry) resulting from these projects. It is worth contemplating if the MGI actually achieved its marketplace aspect, although the ultimate judgement is probably still out. The MGI has suffered from across-the-board funding cuts due to the 2012 Sequester, but fluctuating funding levels are beyond the community’s control. Another problem of MGI was the lag in the integration of experimental work. The workshop suggests that funding mechanisms tailored to incentivizing more integrated efforts between experimental, computational, and data thrusts may help improve this situation.

Vital tactics to building a community for data-driven chemistry, fostering collaborations between stakeholders, and engaging the data and computer science field thus include:

- ◆ Build a core community to drive the field.
- ◆ Democratize the field to maximize its reach and impact.
- ◆ Harness the interdisciplinary nature of the field for cross-fertilization.
- ◆ Create an institutionalized support framework for sustainability and longevity of the field.

II.7. Promoting education and workforce development in modern data science for chemists

The final grand challenge considered in this workshop is that data science and the use of advanced data mining tools are not part of the regular training of chemists, and the wider community thus oftentimes lacks the necessary experience and expertise to utilize them. Correspondingly, the workshop set out to outline a strategy to bridge this disconnect by supporting and guiding the activities of educators.

Education and Workforce Development Need. The paradigm shift towards data-driven research is disruptive and changes the playing field for the chemistry community. The qualitative novelty and inherent interdisciplinarity of this approach give rise to numerous educational challenges and opportunities. Means to addressing these issues include cross-cutting curricular and course developments, the creation of interactive teaching materials, skill-building (e.g., in partnership with industry and other stakeholders), and outreach. The goal of any education and workforce development initiative has to be to help update and adapt education to this changing research landscape in order to adequately equip the next generation of scientists and engineers, to build a competent and skilled workforce for the cutting-edge R&D of the future, and to ensure the competitiveness of our students in the job market. If the community succeeds in creating such an innovative and timely educational framework (including the necessary education of educators), then it will be in a strong position to attract more bright young minds to the emerging field and secure its future. There are several funding programs (e.g., the *NSF NRT program*⁷²), which the community may be able to harness for these efforts.

Multi-Disciplinarity. The overarching theme of the community's educational efforts should be the harnessing of the cross-cutting nature of data-driven work, as well as of the benefits of transferring skills and techniques into new application domains. This includes the need for a broadened horizon, new skills, and the pooling and integration of specialized expertise from different backgrounds. There is a need to promote communication and problem awareness that will transcend the boundaries of traditional disciplines. This approach promises to trigger creative, out-of-the-box thinking and to instill new perspectives on problem solving.

Contemporary Chemistry Curricula. The inclusion of data science ideas into chemistry curricula (both at the undergraduate and graduate level) should be promoted. Traditional curricula do not consider them. Currently, most students are settled with chemical core courses and have very limited opportunities/freedom to complete coursework offered in other domains such as computer science, statistics, and applied mathematics. Most students thus learn about data science in an *ad hoc* manner without formal coursework, which can easily result in critical gaps in their data science knowledge. Curricular rigidity can in part

be traced back to a traditionalist understanding of what constitutes the essence of chemical knowledge, i.e., content that needs to be conveyed by any university-level program. Another reason may be the assumption that students (and potential employers) are not interested in data science. Given the ever-expanding chemical knowledge, it is undoubtedly a great challenge to find the right content balance for contemporary curricula, in particular since the rise of data-driven chemistry is not the only new direction. It is an important task to advocate for flexibility, openness, and for a redefining of what represents the core of chemical knowledge. As instant changes to the curriculum will be difficult to adapt, the workshop suggests focusing on creating awareness in the short-term and collaborating with the chemistry community at large to affect long-term change.

Data-Driven Chemistry Education Tracks. While some essential data science content must become part of every chemistry curriculum, other content will have to remain optional as part of electives that students can choose to specialize in. Some universities have already implemented elective tracks that are relevant for data-driven chemistry, and this may be a model for other schools and degree levels as well. Specialized degree tracks and certificate programs may increase the visibility and marketability of chemistry programs, in particular in areas that emphasize skills (such as data science) that can be transferred, are in high demand, and can help recruit top students. Students with a strong interest in data science or applied mathematics will generally join other departments if chemistry does not offer suitable and attractive options. This could result in chemistry departments missing out on good students and potentially creating a shortage of mathematical skills in our discipline.

Course Content. Data science content may either be taught by other departments, as part of new courses within chemistry, or as new modules/sections within existing chemical core courses. Topics for new foundational courses could include scientific programming and computing, data mining and machine learning in chemistry, uncertainty quantification, and cheminformatics. These courses should be augmented by modules covering basic concepts of probability, statistics, informatics, programming, and numerical methods, which will provide the vertical integration of data science in other classes. Introducing these modules early and revisiting data science topics on a regular basis will help build interest in the students and a sense of the importance for the subject matter. The workshop also suggests revisiting the design of lab courses to incorporate hands-on data science questions. Making graduate level courses open to undergraduates was another suggestion.

Further, there is a need to contribute to the development of corresponding courses and common core course contents that meet the needs of the data-driven research field (see details below). These should connect key aspects of traditional computational work with techniques that enable data-driven approaches. A comprehensive training will put students in a position to perform data-driven research without the need farm out the technical aspects to third party experts.

Educational Materials and Resources. Due to the nascent nature of the subject matter, the shape of new courses and modules in this area is at this point still in flux. Data-driven research requires specialized expertise from many directions across different departments, and the community will have to put particular emphasis on accommodating these demands. Aside from teaching basic concepts, the community should focus on translating success stories of data-driven chemistry to engage students in data science while keeping their interest in chemistry. Teaching the scientific/societal importance of data science can further motivate students to learn and understand. Deep expertise is not required for every student, but general foundations and an understanding of basic data science concepts will be vital for every student. In addition, students would benefit from a basic understanding of what cheminformatics, computational chemistry, and data science can do for them, and the kinds of questions these techniques can help answer.

Central Repository and Training Modules. In addition to framing and coordinating new course contents, the community should also work on efficient ways to deliver them. To this end, stakeholders should contribute to the development and dissemination of course materials and resources that should be made available to the community at large. The workshop participants suggest the creation of a central repository for lectures and other educational material.

In addition to creating dedicated teaching material, the community should also augment research codes with extensive tutorials and training modules. The notion to integrate educational functionality into research software offers an opportunity to reduce the perceived discrepancy between research and classroom education.

Workshops on Educational Challenges. Workshops (or even a workshop series) can be used to raise awareness, bring together stakeholders from different backgrounds, and thus advance educational and workforce development issues associated with data science in chemistry. This includes hands-on training for educators on new tools and techniques (e.g., on the utility of *Python* machine learning libraries in research and education or cloud-based platforms such as *Jupyter HUB*). These workshops can also tackle the creation of new courses, programs, curricula, course contents, teaching materials, guidelines, and recommendations as discussed above. Concrete pilot implementations of educational and outreach initiatives, materials, and resources are expected to result from these meetings. Workshops like *IPAM* and *PASC* are good examples to build on, as they have played a valuable role in the initial development of the field.

Industry-Education Partnerships. A particular point to workshop participants is that industry collaborations could play a considerable role in the rapid adaptation of data science in the chemistry community. With the industry sponsoring ACS symposia on data issues (in particular within the CINF division) and an increase in job postings requiring

both data science and computational chemistry skills, it is evident that industry has both a vested interest in and specific demands for this new field. Industry can offer internships that directly involve students in its data-centric work as well as feedback on specific data science needs. This feedback will allow the academic community to also tailor its educational mission to real-world industry requirements. One suggestion is to follow the *Semiconductor Research Corp (SRC)* model for industrial involvement, i.e., to create funding programs that support direct cross-disciplinary collaborations for both research and education. This model will also create opportunities for students to enter the job market.

Diversity. Notably, outreach and educational efforts have to promote diversity and participation that do not discard any talent and human capital, as neither society nor science can afford to pass on the contributions that stakeholders promise to make.

The education and workforce development in modern data science for chemists would thus benefit from the following:

- ◆ Pursue cross-cutting curricular and course developments to put data-driven chemistry on a solid educational foundation.
- ◆ Advance the creation of interactive teaching materials that address the needs of our educational mission.
- ◆ Seek the input of and collaboration with industry stakeholders to benefit from their unique perspective.
- ◆ Emphasize outreach to harness the potential of underrepresented racial, ethnic, gender, and socioeconomic groups in the development of the field.

III. BROADER IMPACT OF DATA SCIENCE IN CHEMISTRY

Practical solutions to many of the grand challenges of our time can be found in the discovery and development of novel compounds, materials, and processes. These drive innovation, which in turn drives economic development, prosperity, and a rising standard of living in both developed and developing countries.⁷³ They also offer answers to pressing questions in the areas of energy, sustainability, economic competitiveness, human well-being, and national security. Our capacity to address these questions will – perhaps more than ever before – shape the future of our society and planet. Prime examples for this notion are renewable energy capture, conversion, and storage; green catalysts and solvents that drive our industry, conserve resources, and protect the environment; as well as semi- and superconductors for the technologies, devices, and consumer products of the future. The application of big data and modern data science will boost our capacity

to address these grand challenges by transforming the process that creates innovation. A paradigm shift away from trial-and-error searches and towards data-driven discovery and rational design – as outlined by the NSF CHE *Dear Colleague Letter on Data-Driven Discovery Science in Chemistry*⁴¹ – has far-reaching technological implications. It also promises a significant increase in the return on public and private investments, both in terms of resources and time. Educational programs and initiatives will have to complement these ideas: there is an evident need **(i)** to adapt education to such a changing research landscape in order to adequately equip the next generation of scientists and engineers, **(ii)** to build a competent and skilled workforce for the cutting-edge R&D of the future, and **(iii)** to ensure the competitiveness of our students in the job market, and the global competitiveness of the United States economy.

IV. CONCLUSION

There is great excitement about the promise and prospects for the emerging field of machine learning and data-driven research in chemistry. The rise of modern data science has been transformative in many application domains, and it is now primed to make an impact in the chemical sciences and engineering disciplines. An interesting aspect of this development is that the chemistry community does not have to start this development from scratch, but can build on extensive experience and strong foundations from the both the data science and application domain communities (including the materials community). Considering that this field is really only a few years old, there are already numerous impressive and inspiring success stories from pioneering efforts that underscore the potential of this new research paradigm. Nonetheless, there is still much work to be done to make this field a success, ranging from fundamental, basic science questions, to the development of robust techniques, methods, and tools, to the challenge of engaging and taking along the chemistry community at large. Opinions differ on the current state of the field, i.e., on how far along we have already come and how much more progress is necessary for machine learning and data-driven research to become a viable approach for the community and ultimately a cornerstone of chemistry. However, there is little doubt about the tremendous impact this field can have on the future of chemistry.

REFERENCES

1. Selassie, C. D. “**History of Quantitative Structure-Activity Relationships**” *Burger’s Medicinal Chemistry and Drug Discovery* 1–48 (2003).
2. Müller, K.-R., Rätsch, G., Sonnenburg, S., Mika, S., Grimm, M., and Heinrich, N. “**Classifying ‘Drug-Likeness’ with Kernel-Based Learning Methods**” *J. Chem. Inf. Model.* 45, no. 2, 249–253 (2005).
3. Bailly de Tilleghe, C. Le and Govaerts, B. “**A Review of Quantitative Structure-Activity Relationship (QSAR) Models**” *Technical Report 07027* no. Universite catholique de Louvain (2007).
4. Lipinski, C. and Hopkins, A. “**Navigating Chemical Space for Biology and Medicine**” *Nature* 432, no. 7019, 855–861 (2004).
5. Kirkpatrick, P. and Ellis, C. “**Chemical Space**” *Nature* 432, no. 7019, 823 (2004).
6. Dobson, C. M. “**Chemical Space and Biology**” *Nature* 432, no. 7019, 824–828 (2004).
7. Schneider, G. “**Virtual Screening: An Endless Staircase?**” *Nat. Rev. Drug Discov.* 9, no. 4, 273–6 (2010).
8. Scior, T., Medina-Franco, J. L., Do, Q.-T., Martinez-Mayorga, K., Yunes Rojas, J. A., and Bernard, P. “**How to Recognize and Workaround Pitfalls in QSAR Studies: A Critical Review**” *Curr. Med. Chem.* 16, no. 32, 4297–4313 (2009).
9. Zvinavashe, E., Murk, A. J., and Rietjens, I. M. C. M. “**Promises and Pitfalls of Quantitative Structure-Activity Relationship Approaches for Predicting Metabolism and Toxicity**” *Chem. Res. Toxicol.* 21, no. 12, 2229–2236 (2008).
10. Korth, M. “**Large-Scale Virtual High-Throughput Screening for the Identification of New Battery Electrolyte Solvents: Evaluation of Electronic Structure Theory Methods**” *Phys. Chem. Chem. Phys.* 16, no. 17, 7919–7926 (2014).
11. Hummelshøj, J. S., Landis, D. D., Voss, J., Jiang, T., Tekin, A., Bork, N., Dułak, M., Mortensen, J. J., Adamska, L., Andersin, J., Baran, J. D., Barmparis, G. D., Bell, F., Bezanilla, A. L., Bjork, J., Björketun, M. E., Bleken, F., Buchter, F., Bürkle, M., Burton, P. D., Buus, B. B., Calborean, A., Calle-Vallejo, F., Casolo, S., Chandler, B. D., Chi, D. H., Czekaj, I., Datta, S., Datye, A., DeLaRiva, A., Despoja, V., Dobrin, S., Engelund, M., Ferrighi, L., Frondelius, P., Fu, Q., Fuentes, A., Fürst, J., García-Fuente, A., Gavnholt, J., Goeke, R., Gudmundsdottir, S., Hammond, K. D., Hansen, H. A., Hibbitts, D., Hobi, E., Howalt, J. G., Hruby, S. L., Huth, A., Isaeva, L., Jelic, J., Jensen, I. J. T., Kacprzak, K. A., Kelkkanen, A., Kelsey, D., Kesanakurthi, D. S., Kleis, J., Klüpfel, P. J., Konstantinov, I., Korytar, R., Koskinen, P., Krishna, C., Kunkes, E., Larsen, A. H., Lastra, J. M. G., Lin, H., Lopez-Acevedo, O., Mantega, M., Martínez, J. I., Mesa, I. N., Mowbray, D. J., Mýrdal, J. S. G., Natanzon, Y., Nistor, A., Olsen, T., Park, H., Pedroza, L. S., Petzold, V., Plaisance, C., Rasmussen, J. A., Ren, H., Rizzi, M., Ronco, A. S., Rostgaard, C., Saadi, S., Salguero, L. A., Santos, E. J. G., Schoenhalz, A. L., Shen, J., Smedemand, M., Stausholm-Møller, O. J., Stibius, M., Strange, M., Su, H. B., Temel, B., Toftelund, A., Tripkovic, V., Vanin, M.,

- Viswanathan, V., Vojvodic, A., Wang, S., Wellendorff, J., Thygesen, K. S., Rossmeisl, J., Bligaard, T., Jacobsen, K. W., Nørskov, J. K., and Vegge, T. **“Density Functional Theory Based Screening of Ternary Alkali-Transition Metal Borohydrides: A Computational Material Design Project”** *J. Chem. Phys.* 131, no. 1, 14101 (2009).
12. Greeley, J., Jaramillo, T. F., Bonde, J., Chorkendorff, I. B., and Nørskov, J. K. **“Computational High-Throughput Screening of Electrocatalytic Materials for Hydrogen Evolution”** *Nat. Mat.* 5, no. 11, 909–913 (2006).
 13. Dias, J. R. **“The Polyhex/Polypent Topological Paradigm: Regularities in the Isomer Numbers and Topological Properties of Select Subclasses of Benzenoid Hydrocarbons and Related Systems”** *Chem. Soc. Rev.* 39, no. 6, 1913–1924 (2010).
 14. Wang, S., Wang, Z., Setyawan, W., Mingo, N., and Curtarolo, S. **“Assessing the Thermoelectric Properties of Sintered Compounds via High-Throughput *Ab-Initio* Calculations”** *Phys. Rev. X* 1, no. 2, 21012 (2011).
 15. Wilmer, C. E., Farha, O. K., Bae, Y.-S., Hupp, J. T., and Snurr, R. Q. **“Structure–Property Relationships of Porous Materials for Carbon Dioxide Separation and Capture”** *Energy Environ. Sci.* 5, no. 12, 9849 (2012).
 16. Hautier, G., Jain, A., and Ong, S. P. **“From the Computer to the Laboratory: Materials Discovery and Design Using First-Principles Calculations”** *J. Mater. Sci.* 47, no. 21, 7317–7340 (2012).
 17. Mueller, T., Hautier, G., Jain, A., and Ceder, G. **“Evaluation of Tavorite-Structured Cathode Materials for Lithium-Ion Batteries Using High-Throughput Computing”** *Chem. Mater.* 23, no. 17, 3854–3862 (2011).
 18. Wilmer, C. E., Leaf, M., Lee, C. Y., Farha, O. K., Hauser, B. G., Hupp, J. T., and Snurr, R. Q. **“Large-Scale Screening of Hypothetical Metal-Organic Frameworks”** *Nat. Chem.* 4, no. 2, 83–89 (2012).
 19. Castelli, I. E., Olsen, T., Datta, S., Landis, D. D., Dahl, S., Thygesen, K. S., and Jacobsen, K. W. **“Computational Screening of Perovskite Metal Oxides for Optimal Solar Light Capture”** *Energy Environ. Sci.* 5, no. 2, 5814–5819 (2012).
 20. White House Office of Science and Technology Policy **“Materials Genome Initiative for Global Competitiveness.”** (2011) Retrieved from https://www.mgi.gov/sites/default/files/documents/materials_genome_initiative-final.pdf Mar 26, 2018.
 21. NITRD **“The Federal Big Data Research and Development Strategic Plan”** (2016) Retrieved from <https://www.nitrd.gov/PUBS/bigdatardstrategicplan.pdf> Mar 20, 2018.
 22. The White House **“Executive Order - Creating a National Strategic Computing Initiative”** (2015) Retrieved from <https://obamawhitehouse.archives.gov/the-press-office/2015/07/29/executive-order-creating-national-strategic-computing-initiative> Mar 20, 2018.
 23. National Science Foundation **“NSF EarthCube”** (2017). Retrieved from <https://www.earthcube.org/info/about> Mar 20, 2018.

24. **CyVerse**. Retrieved from <http://www.cyverse.org/about> Mar 20, 2018.
25. NITRD (2017) "**The Networking and Information Technology Research and Development Program - Supplement to The President's Budget**" Retrieved from <https://www.nitrd.gov/pubs/2018supplement/FY2018NITRDSupplement.pdf> Mar 20, 2018.
26. National Science Foundation "**Transdisciplinary Research in Principles of Data Science (TRIPODS) Program**" Retrieved from https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505347 Mar 20, 2018.
27. Defense Advanced Research Projects Agency "**Big Mechanism Program**" Retrieved from <https://www.darpa.mil/program/big-mechanism> Mar 20, 2018.
28. National Institutes of Health "**Big Data to Knowledge (BD2K) Initiative**" Retrieved from <https://datascience.nih.gov/bd2k/about> Mar. 20, 2018.
29. National Science Foundation "**Cyberinfrastructure Framework for 21st Century Science and Engineering (CIF21)**" Retrieved from https://nsf.gov/funding/pgm_summ.jsp?pims_id=504730&org=ACI&from=home Mar. 20, 2018.
30. United States Environmental Protection Agency "**Remote Sensing Information Gateway (RSIG) Initiative**" Retrieved from <https://www.epa.gov/hesc/remote-sensing-information-gateway> Mar 20, 2018.
31. "**Harnessing Data for 21st Century Science and Engineering**" in 10 Big Ideas for Future NSF Investments. Retrieved from https://www.nsf.gov/about/congress/reports/nsf_big_ideas.pdf Mar 20, 2018.
32. National Science Foundation "**Computational and Data-Enabled Science and Engineering (CDS&E) Program**" Retrieved from https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504813 Mar 20, 2018.
33. National Science Foundation "**Software Infrastructure for Sustained Innovation (SSE, SSI, S2I2): Software Elements, Frameworks and Institute Conceptualizations Program**" Retrieved from https://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf17526 Mar. 20, 2018.
34. National Science Foundation "**Data Infrastructure Building Blocks (DIBBs) Program**" Retrieved from https://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf17500 Mar 20, 2018.
35. National Science Foundation "**Big Data Regional Innovation Hubs: Establishing Spokes to Advance Big Data Applications (BD Spokes) Program**" Retrieved from https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505264 Mar 20, 2018.
36. Tetko, I.V., Engkvist, E., Koch, U., Reymond J.-L., Chen, H. "**BIGCHEM: Challenges and Opportunities for Big Data Analysis in Chemistry**" *Mol. Inf.* 35, no. 11-12, 1868-1751 (2016).
37. National Science Foundation (2018) "**Cyberinfrastructure for Sustained Scientific Innovation (CSSI) - Data and Software: Elements and Frameworks**" Retrieved from <https://nsf.gov/pubs/2018/nsf18531/nsf18531.htm> Mar 20, 2018.

38. Rajan, K. **“Informatics for Materials Science and Engineering: Data-Driven Discovery for Accelerated Experimentation and Application”** Butterworth Heinemann (2013). Print.
39. Snyder, J. C., Rupp, M., Hansen, K., Müller, K.-R., and Burke, K. **“Finding Density Functionals with Machine Learning”** *Phys. Rev. Lett.* 108, no. 25, 253002 (2012).
40. Brown, N., Zehender, H., Azzaoui, K., Schuffenhauer, A., Mayr, L. M., and Jacoby, E. **“A Chemoinformatics Analysis of Hit Lists Obtained from High-Throughput Affinity-Selection Screening”** *J. Biomol. Screen.* 11, no. 2, 123–30 (2006).
41. **“Dear Colleague Letter: Data-Driven Discovery Science in Chemistry (D3SC)”**
Retrieved from https://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf17112 Mar 20, 2018.
42. Hansen, K., Biegler, F., Fazli, S., Rupp, M., Sche, M., Lilienfeld, O. A. Von, Tkatchenko, A., and Mu, K. **“Assessment and Validation of Machine Learning Methods for Predicting Molecular Atomization Energies”** *J. Chem. Theory Comput.* 9, no. 8, 3404–3419 (2013).
43. Huan, T. D., Mannodi-Kanakithodi, A., and Ramprasad, R. **“Accelerated Materials Property Predictions and Design Using Motif-Based Fingerprints”** *Phys. Rev. B* 92, no. 1, 1–10 (2015).
44. Ramakrishnan, R., Dral, P. O., Rupp, M., and Lilienfeld, O. A. von. **“Big Data Meets Quantum Chemistry Approximations: The Δ -Machine Learning Approach”** *J. Chem. Theory Comput.* 11, no. 5, 150423114552001 (2015).
45. Balabin, R. M. and Lomakina, E. I. **“Support Vector Machine Regression (LS-SVM)-an Alternative to Artificial Neural Networks (ANNs) for the Analysis of Quantum Chemistry Data?”** *Phys. Chem. Chem. Phys.* 13, no. 24, 11710–11718 (2011).
46. Bartók, A. P., Gillan, M. J., Manby, F. R., and Csányi, G. **“Machine-Learning Approach for One- and Two-Body Corrections to Density Functional Theory: Applications to Molecular and Condensed Water”** *Phys. Rev. B* 88, no. 5, 54104 (2013).
47. Broderick, S. and Rajan, K. **“Informatics Derived Materials Databases for Multifunctional Properties”** *Sci. Technol. Adv. Mater.* 16, no. 1, 13501 (2015).
48. Pyzer-Knapp, E. O., Changwon, S., Gómez-Bombarell, R., Aguilera-Iparraguirre, J., and Aspuru-Guzik, A. **“What Is High-Throughput Virtual Screening? A Perspective from Organic Materials Discovery”** *Annu. Rev. Mater. Res.* no. 45 (1) (July), 195–216 (2015).
49. Dallinger, D., Gorobets, N. Y., and Kappe, C. O. **“High-Throughput Synthesis of N3-Acylated Dihydropyrimidines Combining Microwave-Assisted Synthesis and Scavenging Techniques”** *Org. Lett.* 5, no. 8, 1205–1208 (2003).
50. Ananiadou, S., Kell, D. B., and Tsujii, J. **“Text Mining and Its Potential Applications in Systems Biology”** *Trends in Biotechnology* 24, no. 12, 571–579 (2006).
51. Kim, E., Huang, K., Saunders, A., McCallum, A., Ceder, G., Olivetti, E. **“Materials Synthesis Insights from Scientific Literature via Text Extraction and Machine Learning”** *Chem. Mater.* 29, no. 21, 9436–9444 (2017).
52. Hachmann, J., Olivares-amaya, R., Atahan-Evrenk, S., Amador-Bedolla, C., Sánchez-Carrera, R. S., Gold-Parker, A., Vogt, L., Brockway, A. M., and Aspuru-Guzik, A. **“The**

- Harvard Clean Energy Project: Large-Scale Computational Screening and Design of Organic Photovoltaics on the World Community Grid** *J. Phys. Chem. Lett.* 2, no. 17, 2241–2251 (2011).
53. Hachmann, J., Olivares-Amaya, R., Jinich, A., Appleton, A. L., Blood-Forsythe, M. A., Seress, L. R., Román-Salgado, C., Trepte, K., Atahan-Evrenk, S., Er, S., Shrestha, S., Mondal, R., Sokolov, A., Bao, Z., and Aspuru-Guzik, A. **“Lead Candidates for High-Performance Organic Photovoltaics from High-Throughput Quantum Chemistry – the Harvard Clean Energy Project”** *Energy Environ. Sci.* 7, no. 2, 698 (2014).
 54. Olivares-Amaya, R., Amador-Bedolla, C., Hachmann, J., Atahan-Evrenk, S., Sánchez-Carrera, R. S., Vogt, L., and Aspuru-Guzik, A. **“Accelerated Computational Discovery of High-Performance Materials for Organic Photovoltaics by Means of Cheminformatics”** *Energy Environ. Sci.* 4, no. 12, 4849 (2011).
 55. Amador-Bedolla, C., Olivares-Amaya, R., Hachmann, J., and Aspuru-Guzik, A. **“Towards Materials Informatics for Organic Photovoltaics”** *Informatics for Materials Science and Engineering*. Rayan, K. Butterworth-Heinemann (2013). Print.
 56. National Institutes of Standards and Technology (US Department of Commerce). **NIST Chemistry WebBook**. Retrieved from <http://webbook.nist.gov/chemistry/> Mar 18, 2018.
 57. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E. **"The Protein Data Bank"** *Nucleic Acids Res.*, no. 28, 235-242, 2000. doi:10.1093/nar/28.1.235 Retrieved from <http://www.rcsb.org/>
 58. **The Cambridge Structural Database Centre (CCDC)**. Retrieved from <https://www.ccdc.cam.ac.uk/> Mar 18,2018.
 59. **Journal of Negative Results - Ecology & Evolutionary Biology**. Retrieved from <http://www.jnr-eeb.org/index.php/jnr>. 10 Jan 2018.
 60. **"The Smart Laboratory." Dial-a-Molecule EPSRC Grand Challenge Network**. Retrieved from <http://generic.wordpress.soton.ac.uk/dial-a-molecule/the-network/lab-of-the-future-synthetic-route-selection/the-smart-laboratory/>. 10 Jan 2018.
 61. Afzal, M.A.F, Cheng, C., Hachmann, J. **"Combining First-Principles and Data Modeling for the Accurate Prediction of the Refractive Index of Organic Polymers"** *J. Chem. Phys.* 148, 241712, 2018.
 62. Brockherde, F., Vogt, L., Li, L., Tuckerman, M.E., Burke, K. **“Bypassing the Kohn-Sham Equations with Machine Learning”** *Nature Comm.*, 9, 872, 2017.
 63. Jacobsen, T.L., Jorgensen, M.S., Hammer, B. **“On-the-Fly Machine Learning of Atomic Potential in Density Functional Theory Structure Optimization”** *Phys. Rev. Lett.*, 12, 026102, 2018.
 64. Smith, J.S., Isayev, O., Roitberg, A.E. **“ANI-1: An extensible neural network potential with DFT accuracy at force field computational cost”** *Chem. Science*, 8, 3192, 2017.
 65. Behler, J. **“Perspective: Machine learning potentials for atomistic simulations”** *Chem. Phys.*, 145, 170901, 2016.

66. Gomes, J., Ramsundar, B., Feinberg, E.N., Pande, V.S. “**Atomic Convolutional Networks for Predicting Protein-Ligand Binding Affinity**” *CoRR*, 1703.10703, 2017.
67. Yao, K., Herr, J.E., Brown, S.N., Parkhill, J. **Intrinsic Bond Energies from a Bonds-in-Molecules Neural Network.** *J. Phys. Chem. Lett.* 8, 12, 2689–2694, 2017.
68. Hachmann, J., Afzal, M.A.F., Haghightlari, M., Pal, Y. "Building and Deploying a Cyberinfrastructure for the Data-Driven Design of Chemical Systems and the Exploration of Chemical Space", *Mol. Simul.*, submitted (2018).
69. National Science Foundation "Big Data Regional Innovation Hubs: Establishing Spokes to Advance Big Data Applications (BD Spokes)" Retrieved from https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505264 18 Mar 2018.
70. National Science Foundation "Centers for Chemical Innovation (CCI) Program" Retrieved from https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=13635 18 Mar 2018.
71. Jain, A., Ong, S. P., Hautier, G., Chen, W., Richards, W. D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., and Persson, K. A. “**Commentary: The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation**” *APL Materials* 1, 1, 2013.
72. National Science Foundation "NRT program" Retrieved from https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505015 18 Mar 2018.
73. Thornton, K. and Asta, M. “**Current Status and Outlook of Computational Materials Science Education in the US**” *Model. Simul. Mater. Sci. Eng.* 13, 2, R53–R69, 2005.
74. Hanwell, M. D., Curtis, D. E., Lonie, D. C., Vandermeersch, T., Zurek, E., and Hutchison, G. R. “**Avogadro: An Advanced Semantic Chemical Editor, Visualization, and Analysis Platform**” *J. Cheminform.* 4, 1, 17 (2012).
75. Jong, W. A. de, Walker, A. M., and Hanwell, M. D. “**From Data to Analysis: Linking NWChem and Avogadro with the Syntax and Semantics of Chemical Markup Language**” *J. Cheminform.* 5, no. 1, 25 (2013).
76. Kanal, I. Y., Keith, J. A., and Hutchison, G. R. “**A Sobering Assessment of Classical Force Field Methods for Low Energy Conformer Predictions**” *arXiv: https://arxiv.org/abs/1705.04308* 1–19.
77. Kanal, I. Y., Owens, S. G., Sharapov, A. B., and Hutchison, G. R. “**Inverse Design of Conjugated Polymers from Computed Electronic Structure Properties: Model Chemistries of Polythiophenes**” *arXiv: https://arxiv.org/abs/1706.10243*.
78. Gagorik, A. G., Mohin, J. W., Kowalewski, T., and Hutchison, G. R. “**Effects of Delocalized Charge Carriers in Organic Solar Cells: Predicting Nanoscale Device Performance from Morphology**” *Adv. Funct. Mat.* 25, no. 13, 1996–2003 (2015).
79. Zhang, S., Bauer, N. E., Kanal, I. Y., You, W., Hutchison, G. R., and Meyer, T. Y. “**Sequence Effects in Donor-Acceptor Oligomeric Semiconductors Comprising Benzothiadiazole and Phenylenevinylene Monomers**” *Macromolecules* 50, no. 1, 151–161 (2017).

80. Kanal, I. Y. and Hutchison, G. R. “**Rapid Computational Optimization of Molecular Properties Using Genetic Algorithms: Searching Across Millions of Compounds for Organic Photovoltaic Materials**” *arXiv: <https://arxiv.org/abs/1707.02949>* 0–8.
81. Zhang, S., Hutchison, G. R., and Meyer, T. Y. “**Sequence Effects in Conjugated Donor-Acceptor Trimers and Polymers**” *Macromol. Rapid Commun.* 37, no. 11, 882–887 (2016).
82. Baghbanzadeh, M., Bowers, C. M., Rappoport, D., Zaba, T., Gonidec, M., Al-Sayah, M. H., Cyganik, P., Aspuru-Guzik, A., and Whitesides, G. M. “**Charge Tunneling along Short Oligoglycine Chains**” *Angew. Chem. - Int. Ed.* 54, no. 49, 14743–14747 (2015).
83. Losilla, S. A., Watson, M. A., Aspuru-Guzik, A., and Sundholm, D. “**Construction of the Fock Matrix on a Grid-Based Molecular Orbital Basis Using GPGPUs**” *J. Chem. Theory Comput.* 11, no. 5, 2053–2062 (2015).
84. Zubarev, D. Y., Rappoport, D., and Aspuru-Guzik, A. “**Uncertainty of Prebiotic Scenarios: The Case of the Non-Enzymatic Reverse Tricarboxylic Acid Cycle**” *Sci. Rep.* 5, no. 1, 8009 (2015).
85. Rappoport, D., Galvin, C. J., Zubarev, D. Y., and Aspuru-Guzik, A. “**Complex Chemical Reaction Networks from Heuristics-Aided Quantum Chemistry**” *J. Chem. Theory Comput.* 10, no. 3, 897–907 (2014).
86. Cabalo, J. B., Saikin, S. K., Emmons, E. D., Rappoport, D., and Aspuru-Guzik, A. “**State-by-State Investigation of Destructive Interference in Resonance Raman Spectra of Neutral Tyrosine and the Tyrosinate Anion with the Simplified Sum-over-States Approach**” *J. Phys. Chem. A* 118, no. 41, 9675–9686 (2014).
87. Jinich, A., Rappoport, D., Dunn, I., Sanchez-Lengeling, B., Olivares-Amaya, R., Noor, E., Even, A. B., and Aspuru-Guzik, A. “**Quantum Chemical Approach to Estimating the Thermodynamics of Metabolic Reactions**” *Sci. Rep.* 4, (2014).

APPENDICES

APPENDIX A: Workshop Participants & Aids

PARTICIPANT	INSTITUTION	Email
Allison, Thomas	NIST	thomas.allison@nist.gov
Baker, Erin	PNNL	erin.baker@pnnl.gov
Bligaard, Thomas	Stanford	bligaard@slac.stanford.edu
Cheatham, Thomas	U Utah	tec3@utah.edu
Chen, Jason	Scripps	jschen@scripps.edu
Chu, Feixia	U New Hampshire	feixia.chu@unh.edu
Clementi, Cecilia	Rice	cecilia@rice.edu
Corbeil, Jacques	Laval	jacques.corbeil@crchudequebec.ulaval.ca
Deskins, Aaron	Worcester Polytech	nadeskins@wpi.edu
Frenkel, Anatoly	Stony Brook	anatoly.frenkel@stonybrook.edu
Hachmann, Johannes	U Buffalo	hachmann@buffalo.edu
Hanna, Tamara	ACS	t_hanna@acs.org
Hanwell, Marcus	Kitware	marcus.hanwell@kitware.com
Hutchison, Geoff	U Pittsburgh	geoffh@pitt.edu
Isayev, Olexandr	U North Carolina	olexandr@olexandrisayev.com
Kearnes, Steven	Google	kearnes@google.com
Kulik, Heather	MIT	hjkulik@mit.edu
Lewinski, Nastassja	Virginia Commonwealth U.	nalewinski@vcu.edu
Lipson, Hod	Columbia	hod.lipson@columbia.edu
Marom, Noa	Carnegie Mellon	nmarom@andrew.cmu.edu
McLean, John	Vanderbilt	john.a.mclean@vanderbilt.edu
Moore, Jonathan	Dow Chemical	jmoore2@dow.com
Mueller, Tim	Johns Hopkins	tmueller@jhu.edu
Nicklaus, Marc	NIH	mn1@helix.nih.gov
Pamidighantam, Sudhakar	Indiana U	pamidigs@iu.edu
Rajan, Krishna	U Buffalo	krajan3@buffalo.edu
Rice, Julia	IBM	jrice@us.ibm.com
Rinderspacher, Christopher	US Army Research Laboratory	berend.c.rinderspacher.civ@mail.mil
Roitberg, Adrian	U Florida	roitberg@ufl.edu
Schrier, Joshua	Haverford	jschrier@haverford.edu
Sherril, David	Georgia Tech	sherrill@gatech.edu
Shukla, Diwakar	UIUC	diwakar.shukla@shuklagroup.org

Vogt, Frank	U Tennessee	fvogt@utk.edu
West, Richard	Northeastern	r.west@neu.edu
White, Andrew	U Rochester	andrew.white@rochester.edu
Williams, Antony	EPA	williams.antony@epa.gov
Windus, Theresa	Iowa State	twindus@iastate.edu
Yaron, David	Carnegie Mellon	yaron@cmu.edu
Zimmerman, Paul	U Michigan	paulzim@umich.edu

AID	INSTITUTION	Email
Faiz Afzal, Atif	U Buffalo	m27@buffalo.edu
Haghighatlari, Mojtaba	U Buffalo	mojtabah@buffalo.edu
Schrimpe-Rutledge, Alexandra	Vanderbilt	a.rutledge@vanderbilt.edu

APPENDIX B: Workshop Program Schedule

DAY 0: MONDAY, APRIL 17

RUSTICO RESTAURANT & BAR (4075 WILSON BLVD, ARLINGTON, VA 22203)

Workshop Opening: 7:00 – 10:00 pm

Registration, seated opening dinner, opening remarks by the organizers on the motivation and goals for the workshop; discussion of program, schedule, and report writing process.

DAY 1: TUESDAY, APRIL 18

HOLIDAY INN ARLINGTON AT BALLSTON (4610 FAIRFAX DR., ARLINGTON, VA 22203)

Meeting room: Arlington/Clarendon Room

Breakfast, breaks, and working lunch will be catered by the Holiday Inn. Dinner on your own.

Continental Breakfast: 7:30 – 8:20 am

Welcome Remarks by the NSF CHE Director Dr. Angela Wilson: 8:20-8:30 am

Presentation on the NSF Division of Chemistry's ongoing activities and interests.

Session A: 8:30 – 10:30 am (Moderator: Dr. John McLean; Protocol: Workshop aides)

The session will address lead issues 1-3:

Introduction 1 (Invited Speaker: Dr. Erin Baker): 8:30 – 8:45 am

What is the current state of experimental high-throughput screening techniques?

Discussion 1: 8:45 – 9:10 am

What is the future role of experimental high-throughput screening techniques?

Introduction 2 (Invited Speaker: Dr. Goeff Hutchison): 9:10 – 9:25 am

What is the current state of computational high-throughput screening techniques?

Discussion 2: 9:25 – 9:50 am

What is the future role of computational high-throughput screening techniques?

Introduction 3 (Invited Speaker: Dr. Steven Kearnes): 9:50 – 10:05 am

What is the current state of data science (including database, descriptor, data mining, and informatics) techniques?

Discussion 3: 10:05 – 10:30 am

What is the future role of data science (including database, descriptor, data mining, and informatics) techniques?

Coffee Break: 10:30 – 10:50 am

Panel Discussion A: 10:50 – 11:20 am (Moderator: Dr. John McLean; Protocol: Workshop aides)
Panelists: TBD

Session B: 11:20 – 12:00 noon (Moderator: Dr. Johannes Hachmann; Protocol: Workshop aides)
The session will address lead issue 4:

Introduction 4 (Invited Speaker: Dr. David Yaron): 11:20 – 11:35 am
What is the current state of data science for the creation of predictive models?

Discussion 4: 11:35 – 12:00 noon
What is the future role of data science for the creation of predictive models?

Working lunch: 12:00 noon – 1:00 pm
Writing of the draft report will commence during the working lunch.

Overview of Break-Out Sessions: 12:45 – 1:00 pm (Presenter: John McLean)

Break-Out Session 1: 1:00 – 2:40 pm
~10 participants for each of 4 break-out groups – each break-out group to be assigned a specific question/task based on lead issues 1-4 discussed so far. Each group is assigned a facilitator and scribe.

Reports from Break-Out Sessions by Facilitators: 2:15 – 2:40 pm

Coffee Break: 2:40 – 3:00 pm

Session C: 3:00 – 4:20 pm (Moderator: Dr. Johannes Hachmann; Protocol: Workshop aides)
The session will address lead issues 5-6.

Introduction 5 (Invited Speaker: Dr. Adrian Roitberg): 3:00 – 3:15 pm
What is the current state of data science for method development?

Discussion 5: 3:15 – 3:40 pm
What is the future role of data science for method development?

Introduction 6 (Invited Speaker: Dr. Joshua Schrier): 3:40 – 3:55 pm
What is the current state of data science to support decision making in chemical research?

Discussion 6: 3:55 – 4:20 pm
What is the future role of data science to support decision making in chemical research?

Panel Discussion B/C: 4:20 – 4:50 pm (Moderator: Dr. Johannes Hachmann; Protocol: Workshop aides)

Wrap-Up: 4:50 – 5:00 pm (Presenter: Dr. John McLean)
Summarizing the results of the day.

Break 5:00 – 6:00 pm

Working dinner 6:00 – 9:00 pm (as individual break-out groups)
Writing of the draft report will commence during the working dinner.

Day 2: Tuesday, April 19

Holiday Inn Arlington at Ballston (4610 Fairfax Dr., Arlington, VA 22203)

Meeting room: Arlington/Clarendon Room
Breakfast, breaks, and working lunch will be catered by the Holiday Inn.

Continental Breakfast 8:00 – 8:30 am

Session D: 8:30 – 10:30 am (Moderator: Dr. Johannes Hachmann; Protocol: Workshop aides)
The session will address lead issues 7-12.

Introduction 7 (Invited Speaker: Dr. Markus Hanwell): 8:30 – 8:45 am
What is the current state of comprehensive, integrated, general-purpose, user-friendly tools and their development?
What are the main science successes of data-driven research?

Discussion 7: 8:45 – 9:10 am
What is the future role of comprehensive, integrated, general-purpose, user-friendly tools and their development?
What are the main science challenges and opportunities for data-driven research?

Introduction 8 (Invited Speaker: Dr. Cecilia Clementi): 9:10 – 9:25 am
What is the current state of education in modern data science for chemists?
What is the current state of engagement of the data and computer science community?

Discussion 8: 9:25 – 9:50 am
What is the future of education in modern data science for chemists?
What is the future of engagement of the data and computer science community?

Introduction 9 (Invited Speaker: Dr. Krishna Rajan): 9:50 – 10:05 am
What are the lessons from the Materials Genome Initiative and other big data initiatives?
What are the funding mechanisms to support the specific needs for data-driven research?

Discussion 9: 10:05 – 10:30 am
What are the lessons from the Materials Genome Initiative and other big data initiatives?
What are the funding mechanisms to support the specific needs for data-driven research?

Coffee Break 10:30 – 10:50 am

Break-Out Session 2: 10:50 – 12:20 pm

~10 participants for each of 4 break-out groups – each breakout group to be assigned a specific question/task based on lead issues – each group assigned a facilitator and scribe.

Reports from Break-Out Sessions by Facilitators: 12:00 – 12:20 pm

Working lunch: 12:20 – 1:20 pm

Writing of the draft report will commence during the working lunch.

Break-Out Session 3, pt 1: 1:20 – 2:40 pm

Summary of the Break-Out reports and task force organization for elaborating key ideas for the initial version of the workshop report.

Coffee Break: 2:40pm – 3:00 pm

Break-Out Session 3, pt 2: 3:00 – 3:50 pm

Summary of the Break-Out reports and task force organization for elaborating key ideas for the initial version of the workshop report.

Wrap-Up: 3:50 – 4:00 pm

Organizers summarizing the day's results

Closing Remarks and Adjourn: 4:00 pm