Data-Mining-Driven Quantum Mechanics for the Prediction of Structure

G. Ceder, D. Morgan, C. Fischer, K. Tibbetts, and S. Curtarolo

Abstract

The prediction of crystal structure is a key outstanding problem in materials science and one that is fundamental to computational materials design. We argue that by combining the predictive accuracy of quantum mechanics with data mining tools to extract knowledge from a large body of historical experimental or computational results, this problem can be successfully addressed.

Keywords: ab initio, *crystal structure*, *simulation.*

Introduction

Over the last 40 years, *ab initio* methods have become ubiquitous tools in chemistry, physics, and materials science. *Ab initio* methods, which accurately solve the fundamental quantum mechanical equations (Schrödinger or Dirac) for the electrons of a system, hold the promise of virtual materials research, that is, learning the properties of materials completely by computation, before experimental synthesis and testing. In the last decade, significant advances in solid-state physics, fundamental materials science, and advanced computing have brought us closer to that objective, and accurate *ab initio* approaches now exist for many properties (e.g., diffusion, thermodynamic quantities, ferroelectricity, lattice parameters, elastic constants, etc.). The September 2006 issue of *MRS Bulletin* on density functional theory (guest-edited by J. Hafner, C. Wolverton, and G. Ceder) highlights some of the successes of *ab initio* methods in a variety of materials research areas.

Ab initio studies are still primarily used to further the understanding and rationalize the properties of well-known materials. Studies of this type bypass the problem of predicting the structure of a material, as it is usually known from experiment. If we peek into the future and imagine true virtual materials design, our efforts will need to extend beyond property prediction and address the problem of structure prediction. Most materials properties, from bandgaps to brittle fracture, melting temperature to magnetism, depend strongly on the structure of the materials involved, and without knowledge of the crystal structure, *ab initio* computations easily become irrelevant. Hence, the full power of *ab initio* calculations for materials design will only be unlocked if we address the problem of structure prediction. In this article, we focus on equilibrium crystal structure prediction, setting aside the even more difficult problem of predicting amorphous and metastable structures.

Data Mining Structure Prediction

Predicting the stable crystal structure of a material in essence requires one to find the atomic arrangement with lowest free energy (at non-zero temperature), or with lowest energy (at zero temperature and pressure). We only focus here on the zero-temperature, zero-pressure ground-state search, as it is also a key component of any finite-temperature, finite-pressure study. Models to construct the free energy of a given structure or class of structures (e.g., those on a fixed topology) are well developed and have led to a large number of successful phase diagram computations.[1–5]

Both an accurate description of the *energetics* of a material as well as a strategy to *search* through the almost infinite space of possible structures are needed to find the most stable structure. Decades of work with *ab initio* methods studying specific systems and/or properties, typically using the local density approximation (LDA) or generalized gradient approximation (GGA) to density functional theory (DFT), as well as a more recent large-scale comparison of computationally predicted ground states with experiments in 80 binary alloys,[6] indicates that the accuracy of DFT is not the limiting factor for predicting structure: Only rarely does DFT erroneously predict an equilibrium structure that is different from the experimentally observed one, though much effort has been spent on the few cases where DFT fails in this regard.[6]

The difficulty of predicting structure lies primarily in the searching strategy. One may think of this as an optimization problem in the space of $3N$ coordinates, where N is the number of atoms in the system. Due to the high dimensionality of the space, many local minima exist in this space, and the global minimum energy configuration usually cannot be found through intuitively appealing approaches, such as static or dynamic minimizations of the energy with respect to the positions of atoms. It is worth pointing out that promising genetic algorithms, a class of optimization strategies based on evolutionary approaches, have recently been explored.[7]

Instead, structure is often "predicted" by a *suggest and test* approach: candidate structures are selected by the researcher, and DFT is used to compare their relative stability. The process of determining good candidates from what is already known can be viewed as a problem in data mining. Currently, the candidates are usually suggested based on personal intuition. Such an approach is of course limited by human ability to include the correct structure in the guess.

We believe that by formalizing this data mining step, and by codifying and quantifying knowledge obtained from past experience—computational or experimental—one can make systematic informed guesses about the crystal structures that are likely to form in a new, unexplored system. It is important to understand that suggesting a reasonable, and short, candidate list of structures containing the true ground-state structure with a high probability essentially solves

the crystal structure prediction problem. This is because *ab initio* energy methods can easily be used to determine, with high accuracy, the lowest-energy structure from a short list of candidates. Hence, the problem of predicting crystal structure can be solved, for all practical purposes, by combining modern quantum mechanical methods with machine learning techniques[8] into a common framework. A machine learning method, trained on previously obtained results, inductively captures the underlying physical rules governing structural stability, and quantum mechanics provides the final accuracy. This is a deviation from past efforts that treat structure prediction either as a mathematical optimization problem on the DFT functional, with no embedded historical knowledge,[9,10] or as a heuristic problem trying to define simple rules such as ionic size and electronegativity to predict structure. Data mining structure prediction (DMSP), on the other hand, integrates the best of both of these approaches.

How can historical data on structure be embodied into rules or information that can intelligently steer *ab initio* energy methods toward the stable structure in a new system? Figure 1 shows the key ideas of the approach. "Knowledge" about structure can be extracted from large amounts of computational or experimental results. A variety of standard data mining methods such as principal component analysis, neural networks, clustering schemes, and so on, could in principle be used for this knowledge extraction, though at this point only principal component regression and Bayesian probability methods have been tested. This historical knowledge is then used to suggest candidate structures to be evaluated with DFT. Curtarolo et al. showed the success of this approach by extracting knowledge from a large set of computed data.[8]

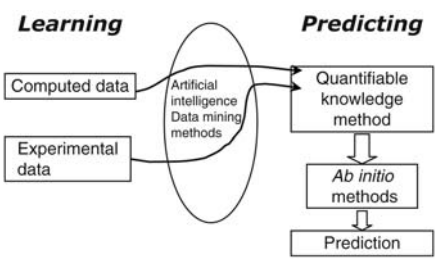

Figure 1. Knowledge about rules that govern crystal structure stability can be extracted using data mining tools on a large amount of experimental or computed structure information. This knowledge can then be used to inform quantum mechanical methods and drive them toward the most likely structures.

In a high-throughput mode, they computed the energy of 114 structure prototypes in each of 55 systems, and used principal component analysis to show that the DFT energy differences between crystal structures are strongly correlated across chemical systems.

The method proposed by Curtarolo exploits the fact that there are clear dependencies between the energies of different crystal structures. For example, if two structures are stabilized by ionic interactions, then their relative energy will tend to go up and down together as the system under study is more or less ionic. Rather than resort to physics-based models, data mining methods can capture this correlation in a purely mathematical and therefore less biased form. Using principal least-squares regression on the computed training data set, they were able to significantly accelerate finding the lowest-energy structure for a new system. Such a knowledge-driven search can be implemented in an iterative manner. Starting with some minimal information about the system (e.g., only element structure data), the correlations extracted from the training data set are used in DMSP to suggest low-energy structures for intermetallic compounds in the new system. Calculating the energy of these suggestions with *ab initio* methods provides more information about the system, which can then be used in a new iteration of the DMSP. Figure 2 shows the number of *ab initio* computations that were needed to predict the ground-state line of the 55 binary alloys tested with a given accuracy. With only 26 structure calculations per alloy

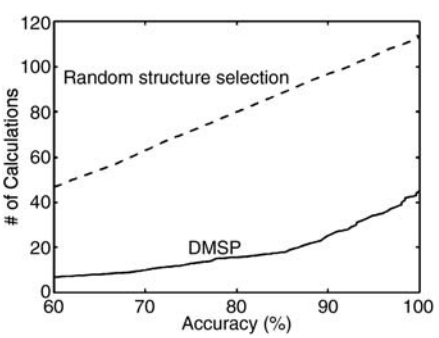

Figure 2. Knowledge based on a set of computed data in 55 binary alloys was used to inform this algorithm.[8] The DMSP curve shows how many calculations are needed on average to obtain the ground-state line of these alloys with a given accuracy. Comparison with random testing of structures shows clearly that knowledge is embedded in the data mining structure prediction (DMSP) method.

system—a very reasonable effort with modern computing—90% of the ground states are correct.

Experimental Versus Virtual Data

Even though high-throughput computations[11] make it possible to gather a large amount of information on which learning methods can be trained, experimentally obtained crystal structure data has the advantage of providing information on the true, most stable crystal structure over a large number of chemistries. Using experimental data to quantify knowledge poses different challenges from using computed data, as the former is non-numerical and discrete: A structure either occurs or does not, and information rarely exists as to how close in energy competing structures may be, limiting the use of energy-based data mining schemes. But as shown by Fischer et al.,[12] information can be extracted in other ways. For example, one may ask the question to what extent the occurrence of two crystal structures (e.g., at two different compositions in an A–B alloy) are correlated, expressed in mathematical terms as

$$f(x_i, x_j) = \frac{p(x_i, x_j)}{p(x_i)p(x_j)}, \quad (1)$$

where x_i and x_j are variables representing crystal structures appearing at compositions i and j in the alloy, $p(x_i, x_j)$ is the probability that both will occur in the same binary system picked at random, and $p(x_i)$ is the probability for the structure appearing at composition i. If nature were random, these correlations would converge to one for all pairs of structures when sampled over a large enough number of systems. Instead, it is expected that the underlying molecular-level interactions will drive some structures to frequently appear together, resulting in a correlation ratio greater than one, while others, independently stabilized by very different physical phenomena, will yield a correlation ratio approaching zero. Sampling the structure–structure and element–structure correlations present between 611 different structure types in 1335 binary metallic alloys in the Pauling file,[13] one of the largest databases for the structures of binary metals, shows strong correlation or anticorrelation between pairs of structures. For example, the Fe_3C-type structure at AB_3 composition and the $MgCu_2$-type structure at A_2B appear together in 52 of the 87 alloys in which Fe_3C is present, giving a correlation ratio f of 8.49. In other words, given that Fe_3C is present at AB_3, it is 8.49 times more likely that $MgCu_2$ will form at A_2B than if the structures were

uncorrelated. In this case, a physical origin for the correlation can be clearly identified, as both structures form in systems where the constituent elements A and B are of very different size. As such, these structures are very unstable when the roles of small and large atoms are interchanged, a fact which is also borne out by the data: the correlation ratio for Fe_3C forming at the same composition, AB_3, but $MgCu_2$ at AB_2 (instead of A_2B) reveals strong anticorrelation and $f \approx 0$. It is important to stress that while in some cases the physical effect driving (anti)correlation is clear, the strength of data mining approaches is that they can extract correlation information without first specifying an underlying physical mechanism.

It is possible to quantify the extent to which such correlations are useful for predicting structure. An often-encountered situation is that a compound is known to form at one composition in an alloy, but what appears at other compositions remains unknown. To what extent does knowledge of structures at one composition provide predictions for structures at other compositions? This concept can be represented mathematically by the mutual information[14] $I_{i,j}$ between compositions i and j, given by

$$I_{i,j} = \sum_{x_i,x_j} p(x_i, x_j) \ln\left[\frac{p(x_i, x_j)}{p(x_i)p(x_j)}\right], \qquad (2)$$

where the sum extends over all combinations of structures x_i and x_j that can appear at compositions i and j.

Figure 3 shows the mutual information between compositions, as well as between a composition and a constituent element (the brighter areas are the more mutually informative). Bright pixels for rows labeled "A" and "B" in Figure 3 reiterate the fact that crystal structure in a binary alloy is strongly influenced by the identity of its constituent elements, giving substantial credence to methods empirically relating structure to an elemental property.[14,15] Correlation between the structure of $A_{1-c}B_c$, where c is a composition variable, and element A in the two-row AB matrix in Figure 3 is understandably very strong when the system is rich in A, but decreases only slowly with increasing B content. The structure–structure correlations shown in Figure 3 reflect the fact that interactions present at one composition are indicative of interactions at other compositions, a result which should not be too surprising.

There are many ways to use such strong correlations in experimental data for more efficiently suggesting structure. A formal procedure was outlined and tested by

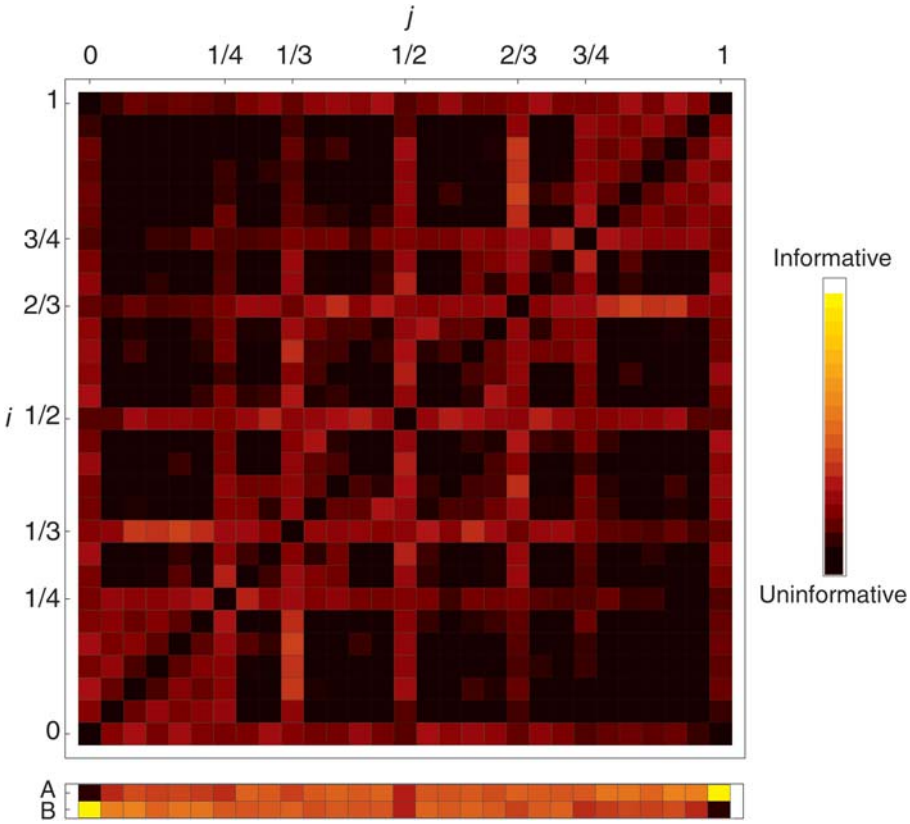

Figure 3. Correlation map for binary alloys, showing to what degree knowledge of crystal structure at one composition, i, *determines the structure at another composition,* j. *The two bars below the map show to what degree the crystal structure in an alloy is related to the knowledge of the element (A or B).*

Fischer et al.,[12] who constructed a probability function in the space of all possible structure combinations and chemistries, $P[\mathbf{X} = (A, B, x_i, \ldots, x_k)]$ where the variable x_i indicates the crystal structure present at composition c_i and A and B identify the constituent elements "A" and "B" present in the alloy. Note that the vector $\mathbf{X} = (A, B, x_i, \ldots, x_k)$ fully describes states of the alloy A–B. Such a probabilistic model provides a natural framework in which one can utilize the various degrees of partial knowledge about an alloy system in a consistent manner. For example, let **e** be the vector that contains what is already known about the system. Other than the chemical nature of the elements, this almost always includes the crystal structure of the elements, but can also include knowledge of structures at other compositions. Determining candidate structures at other compositions then consists of determining the conditional probability $P(\mathbf{X}|\mathbf{e})$, or translated, what is the probability for a particular set of ground states, **X**, based on what we already know about the system, **e**? Practical predictions require an explicit form of $P(\mathbf{X})$, which can, for example, be achieved with a cumulant expansion. In general, $P(\mathbf{X})$ should be constructed in a manner that is consistent with known information while remaining maximally noncommittal or spread out.[16]

Correlation between the occurrence of structures at different compositions—for example, sampled from a database of experimental structure information—can be used to inform a probability density $P(\mathbf{X})$ that in turn can be used to predict structures in new materials. Table I shows the results of such a prediction for $AgMg_3$, where experiments have indicated the presence of a compound but with a yet undetermined crystal structure. A probability density constructed from available data in the Pauling file and conditioned on the limited knowledge of structures occurring at other compositions in the Ag-Mg system predicts the Cu_3P structure as the most likely candidate for the ground state. Four other highly ranked structures are also shown in Table I. A DFT calculation of the energy of these five candidate structures as well as 26 other structures identified the Cu_3P structure as having the lowest energy, indicating that this data

Table I: Data Mining Prediction for the Crystal Structure of $AgMg_3$.

DMSP-Ranked List of Candidate Structure Types for $AgMg_3$	Rank Based on Frequency of Occurrence in Nature
1. Cu_3P	21
2. BiF_3	7
3. $IrAl_3$	15
4. $SrPb_3$	14
5. Mg_3Cd	4

Notes: The first column shows five structure prototypes suggested (and ranked) by the data mining algorithm based on their probability to be the ground state for $AgMg_3$, given knowledge of the elemental structures and the structures at Ag_3Mg and AgMg. The second column gives the frequency with which each of these structures occurs in nature at that stoichiometry. First-principles computations on these and 26 more structures revealed no lower-energy structures than Cu_3P.

mining algorithm is extremely efficient in suggesting low-energy structures.

Figure 4 shows on a larger scale how well data mining methods can suggest structure. By testing the data mining approach on 3975 compounds appearing at least twice in the Pauling file database of binary metallic alloys, Fischer et al. obtained statistics on the accuracy of this method. For each prediction in an A–B alloy, all information about that alloy was removed from the data, and an ordered list of candidate structures was generated for each composition by the data mining algorithm. Figure 4 shows how far one must descend down this list to have a certain probability of finding the true structure for three different methods. The data mining approach (DMSP, blue curve) is very effective in predicting the true ground state, requiring the investigation of only five structures for a 90% chance of finding the true structure. Comparing the data mining approach with two simpler data mining schemes shows that it embodies considerably more knowledge: Picking trial ground states based on the *frequency* with which they occur in nature (red curve in Figure 4) is an obvious approach but results in only 62% accuracy when investigating a five-structure suggestion list. As a point of reference, the result of guessing structures with a uniform probability (at random) is also shown (orange curve in Figure 4). The improved efficacy of DMSP is of practical value, as the length of the list is the number of *ab initio* computations one needs to perform to obtain the correct ground state with a certain probability. Clearly, once the DMSP method is trained by a body of historical results, this list is remarkably short.

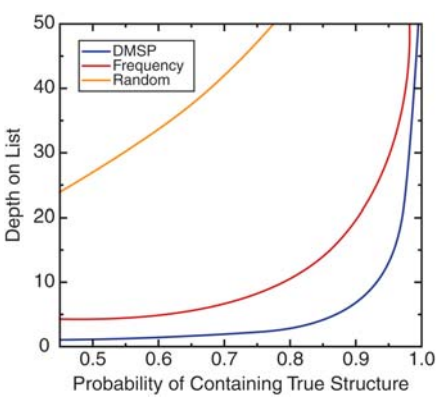

Figure 4. Knowledge based on the experimental information in the Pauling file is used to rank candidate structures in an alloy left out deliberately. The blue curve shows how long the candidate list produced by the data mining algorithm needs to be ("depth of list") to contain the true ground state with a given probability. For comparison, random guessing (orange curve) and picking structures based on the frequency with which they occur in nature (red curve) are also shown.

Conclusions

These examples indicate the power of data mining to capture historical knowledge in non-implicit rules, and to use them to rapidly drive quantum mechanics toward low-energy structures.

We believe that such a synergy between *knowledge methods* and *quantum mechanics* addresses the limitation that currently exists in structure prediction approaches. While computational quantum mechanics is highly *accurate*, it is not *suggestive*, and incorporates knowledge only through the experience of the user.

On the other hand, many heuristic methods have been developed to relate the structure of materials to simple properties of their constituent elements. Structure maps using electronegativity, atomic size, or simple position in the periodic table have had substantial success in organizing various structure types.[17]

While such heuristic methods create considerable insight into the physical mechanisms that control structure selection, their *accuracy* is limited by assumptions made regarding what constitutes relevant physical parameters, and often by their focus on only elemental properties.

An advantage of data mining methods is that the important parameters do not need to be identified *a priori*, and large, complex data sets, like the known intermetallic structures, can be automatically mined for their hidden information. Moreover, typical heuristic methods lack a mechanism to systematically improve upon prediction ability. Using data mining to suggest a meaningful candidate list of structures that can be sorted through quickly with quantum mechanics combines the suggestive character of knowledge methods with the accuracy of quantum mechanics and is a pragmatic and functional solution to the problem of predicting crystal structure.

Data mining approaches are a radical departure from the deductive logic approach in science and are typically found more often in consumer preference studies, business, and social sciences. Ideally, one would predict the macroscopic properties of a material by understanding and formalizing its relation to all relevant underlying microscopic phenomena. While such an approach is often intellectually more satisfying to the physical scientist, the complexity of many materials and the length and time scales that need to be bridged sometimes render the deductive approach virtually impossible.

For properties for which it is difficult to create or implement a formal coarse-graining theory to predict macroscopic properties from microscopic calculations, data mining, which allows the underlying relations to be established inductively, may be an alternative approach. Data mining techniques have been adapted for a number of materials properties, including boiling points, creep, mechanical weld properties, time–temperature–transformation diagrams, and catalytic activity (see review in Reference 18). Crystal structure is now joining a host of other materials properties that are being treated with data mining methods.

Acknowledgments

This work was supported by the National Science Foundation's ITR Program under contract DMR-0312537, and by the Department of Energy under grant DE-FG02–96ER45571. We thank John Rodgers for discussions on structure prototypes and experimental databases.

References

1. A. Van de Walle and G. Ceder, *J. Phase Equilib.* **23** (2002) p. 348.
2. D. de Fontaine, in *Solid State Physics*, edited by H. Ehrenreich and D. Turnbull (Academic Press, New York, 1994) p. 33.
3. A. Van der Ven, M.K. Aydinol, G. Ceder, G. Kresse, and J. Hafner, *Phys. Rev. B.* **58** (1998) p. 2975.
4. M. Asta, D. de Fontaine, M. Van Schilfgaarde, and M. Sluiter, *Phys. Rev. B.* **46** (1992) p. 5055.

5. V. Ozolins, C. Wolverton, and A. Zunger, *Phys. Rev. B.* **57** (1998) p. 6427.
6. S. Curtarolo, D. Morgan, and G. Ceder, *Calphad* **29** (2005) p. 163.
7. N.L. Abraham and M.I.J. Probert, *Phys. Rev. B.* **73** 224104 (2006).
8. S. Curtarolo, D. Morgan, K. Persson, and G. Ceder, *Phys. Rev. Lett.* **91** 135503 (2003).
9. G.H. Johannesson, T. Bligaard, A.V. Ruban, H.L. Skriver, K.W. Jacobsen, and J.K. Norskøv, *Phys. Rev. Lett.* **88** 255506 (2002).
10. M. Jansen, *Angew. Chem. Int. Ed.* **41** (2002) p. 3746.
11. D. Morgan, G. Ceder, and S. Curtarolo, *Meas. Sci. Technol.* **16** (2005) p. 296.
12. C. Fischer, K. Tibbetts, D. Morgan, and G. Ceder, *Nature Mater.* **5** (2006) p. 641.
13. P. Villars, M. Berndt, K. Brandenburg, K. Cenzual, J. Daams, F. Hulliger, T. Massalski, H. Okamoto, K. Osaki, A. Prince, H. Putz, and S. Iwata, *Pauling File: Binaries Edition*, Database on CD-ROM (ASM International, Materials Park, Ohio, 2002).
14. D. Morgan, J. Rodgers, and G. Ceder, *J. Phys.: Condens. Matter* **15** (2003) p. 4361.
15. D.G. Pettifor, *J. Phys. C: Solid State Physics* **19** (1986) p. 285.
16. E.T. Jaynes, *Probability Theory: The Logic of Science* (Cambridge University Press, Cambridge, UK, 2003).
17. P. Villars, in *Factors Governing Crystal Structures in Intermetallic Compounds: Principle and Practice*, edited by J.H. Westbrook and R.L. Fleischer (John Wiley & Sons, New York, 1994) p. 227.
18. D. Morgan and G. Ceder, in *The Handbook of Materials Modeling*, edited by S. Yip (Springer, New York, 2005) p. 395. □

MRS FUTURE MEETINGS

2007 SPRING MEETING
April 9-13
Exhibit: April 10-12
San Francisco, CA

Meeting Chairs:

Timothy J. Bunning
Air Force Research Laboratory
timothy.bunning@wpafb.af.mil

Harold Y. Hwang
University of Tokyo
hyhwang@k.u-tokyo.ac.jp

Debra Kaiser
National Institute of Standards and Technology
debra.kaiser@nist.gov

Jennifer Lewis
University of Illinois, Urbana-Champaign
jalewis@uiuc.edu

2007 FALL MEETING
November 26-30
Exhibit: November 27-29
Boston, MA

Meeting Chairs:

Duane Dimos
Sandia National Laboratories
dbdimos@sandia.gov

Mary Galvin
Air Products and Chemicals, Inc.
galvinme@airproducts.com

David Mooney
Harvard University
mooneyd@deas.harvard.edu

Konrad Samwer
Universitaet Goettingen I. Physikalisches Institut
ksamwer@gwdg.de