Final Report

DoE Project #: DE-SC0004910

Project Title: MODELING AND SIMULATION OF HIGH DIMENSIONAL STOCHASTIC MULTISCALE PDE SYSTEMS AT THE EXASCALE

Principal Investigator: Nicholas Zabaras

Distribution: unlimited

Executive Summary: Predictive Modeling of multiscale and Multiphysics systems requires accurate data driven characterization of the input uncertainties, and understanding of how they propagate across scales and alter the final solution. This project develops a rigorous mathematical framework and scalable uncertainty quantification algorithms to efficiently construct realistic low dimensional input models, and surrogate low complexity systems for the analysis, design, and control of physical systems represented by multiscale stochastic PDEs. The work can be applied to many areas including physical and biological processes, from climate modeling to systems biology.

Accomplishments (compared to the original goals): The original goals of the project included: 1) developing data driven algorithms to construct stochastic input models of low dimensionality; 2) develop uncertainty quantification algorithms that solve high dimensional SPDEs; 3) develop model reduction algorithms with quantifiable error that solve high dimensional SPDEs; 4) create multiscale algorithms and integrate them with SPDE solvers and model reduction algorithms.

The Cornell portion of this project focused on goals 1 and 4, addressing the issues of optimization and thermodynamic characterization. We merged information theory with the cluster expansion to obtain a thermodynamic treatment in closer agreement with experiments. We addressed, in a rigorous way, the important issue of how much accuracy we sacrifice when replacing the expensive computer code with the surrogate. We develop an approach capable of quantifying the uncertainties in properties computing using the surrogate. The framework is general, but will be applied to the cluster expansion. We demonstrated a case of using surrogates for materials design. The cluster expansion surrogate model is employed to predict a material with a given specified property. We showed how the cluster expansion is useful for materials design and in this process we modify the traditional cluster expansion applicable to bulk alloy systems to now handle low-dimensional systems of arbitrary shapes.

Summary of Project Activities:

One of the biggest achievements of computational materials science would be to produce a full-fledged virtual materials design laboratory thus removing any need for costly real world experimental testing. We imagine inputing a set of desired materials properties into a computer

code, wait for a reasonable amount of time, and receive the output material optimal for the application at hand. Only a single material, the optimal one, needs to be produced in a real world laboratory and we know immediately that all other possible representations of this material are inferior.

We still have a long way to go to reach this goal successfully, but in order to do so, we must be skillful in at least the following two tasks: We must be able to optimize materials properties, such as the energy, the band gap, the thermal conductivity, etc., represented *in silico*, and second, we should be able to accurately characterize materials. An essential characterization is of thermodynamic nature. That is, we need to know, e.g., in which phases the material will exist at various external conditions and also how stable these phases are. In both of these tasks, we must have a clear understanding, and quantification, of the uncertainties associated in our work.

One of the main reasons why this virtual materials laboratory is not a reality at this time is the fact that computing the property of a material requires running temporally expensive computer codes. For example, if we want the accurate *ab initio* quantum mechanical energy of a single representation of a material, we can run the Vienna *ab initio* simulation package VASP, which currently takes hours even on a supercomputer. This introduces a serious problem since any optimization task needs to search typically many hundreds of millions different representations of the material in its way to the optimal answer. It seems that we are faced with an insurmountable computational effort.

To make matters worse, for thermodynamic modeling of a material, we rely on statistical thermodynamics which demands the evaluation of ensemble averages requiring the materials property evaluated for the system found in the most likely (in principle *all*) states at a set of external conditions such as temperature and pressure. Again, although not an optimization task per se, we find ourselves facing the same problem of an infeasible computational task due to the expensive computer code. By now it should be clear that, overcoming the computational cost associated with obtaining materials properties is of central importance in computational materials science and represents one of the most relevant research areas.

The above discussion motivates the development of so-called *surrogate models*. A surrogate model, or simply, a surrogate, is a replacement for the accurate, expensive, computer code. In a nutshell, the surrogate attempts to learn the output, also called the response, of the computer code for any given input, in orders of magnitude less time than it takes to run the expensive code. The outputs of the computer code for all possible inputs are collectively called the response surface.

When employing a surrogate to learn the response surface, we necessarily introduce uncertainty into computed materials properties. This generates essential questions such as, how do these uncertainties affect our final predictions about which material is best for a given application? We may tell the experimentalist to produce material A when in fact material B is better because we relied too much on a single surrogate, this is the danger of using non-

Bayesian methods. Bayesian methods on the other hand, as we will see in this thesis, accounts probabilistically for all possible surrogates consistent with a set of samples taken from the true response surface. Rigorously accounting for uncertainties with a probabilistic approach will tell us how much}we believe the best material is A over B, leaving us with a much more informed approach to materials design. Furthermore, in the context of thermodynamic characterization, how does the approximate surrogate affect our uncertainty about which phase a material is most stable in for various temperatures and pressures? Disregarding this uncertainty can be detrimental to future materials design.

Although the methods developed in this thesis are general, we have decided to focus our attention on a subset of materials, namely alloys, i.e., materials composed of a set of different chemical elements of which at least one element is metallic. Alloys are very commonplace in society. Indeed, mercury mixed with silver, tin, copper, and zinc forms dental fillings, iron mixed with aluminum, nickel, cobalt, and other elements, creates the magnets in loudspeakers, copper mixed with zinc produces door locks and bolts, iron mixed with carbon and silicon is used to build bridges and cookware, copper mixed with nickel and manganese is used to create coins, aluminum mixed with copper, magnesium, and manganese forms materials used in automobiles, for aircraft body parts, and for military equipment. The list goes on.

In the case of alloys represented in silico, a particular surrogate model called the cluster expansion has been employed for many decades to represent alloy configurational properties, i.e., properties that depend on exactly where the atoms in the alloy are placed, called a configuration, on the lattice (for example a face-centered cubic (fcc) lattice with a basis) defining the geometry of the alloy. The cluster expansion is useful because of its computational speed and can, in principle, be made arbitrarily accurate to the point where the true computer code is exactly represented. In practice, however, it is made approximate.

In this research, we addressed both the issue of optimization and that of thermodynamic characterization. We merged information theory with the cluster expansion to obtain a thermodynamic treatment in closer agreement with experiments. We addressed, in a rigorous way, the important issue of how much accuracy we sacrifice when replacing the expensive computer code with the surrogate. We develop an approach capable of quantifying the uncertainties in properties computing using the surrogate. The framework is general, but will be applied to the cluster expansion. We demonstrated a case of using surrogates for materials design. The cluster expansion surrogate model is employed to predict a material with a given specified property. We showed how the cluster expansion is useful for materials design and in this process we modify the traditional cluster expansion applicable to bulk alloy systems to now handle low-dimensional systems of arbitrary shapes.

Relative entropy

In this section, we seek to emphasize our successful approach in using information theory to improve computational alloy modeling.

Construction of binary alloy phase diagrams generally relies on a computationally tractable parametrized surrogate model for the quantum mechanical energy surface. The cluster expansion is a commonly used model which has been very successful in describing configurational properties of alloys. The model parameters are referred to as effective cluster interactions (ECI) and, for a typical system, range in number from 20 to 80. The ECI are often fitted to 50-100 observed energies, e.g., using least squares with cross validation, potentially coupled with genetic algorithms, or compressive sensing. These observations are made at a high computational cost, involving *ab initio* software, e.g., the Vienna *ab initio* simulation package VASP. Given the ECI, the energy of any configuration can be computed, and subsequently used for thermodynamic simulations.

In this work, we provided an intuitive argument as to why fitting the ECI to equally weighted observed energies does not necessarily lead to an optimal description of the thermodynamics of the system. Consider the case of a canonical ensemble. The Boltzmann factor, in equilibrium, dictates that more energetic states are exponentially less likely to be observed. So, at low temperatures, the partition function is mostly influenced by the few low energy states. The Boltzmann factor for all other states is essentially zero.
Therefore, the thermodynamic importance of a state is quantified by its Boltzmann factor. We conclude that, as the temperature is increased, so is the importance of any state. Furthermore, at infinite temperature all states are equally important.

Motivated by the previous discussion it is desirable to investigate techniques which obtain the ECI based on thermodynamic arguments. All the necessary information is encapsulated in the probability distribution over states (PDS). The idea is to bring the PDS induced by the cluster expansion (candidate PDS) as close as possible to the true one. We propose to measure this distance in terms of relative entropy (also known as the Kullback--Leibler divergence). Thus, we obtain a variational problem, namely, the minimization of the relative entropy functional with respect to the candidate PDS. Though theoretically sound this problem is computationally intractable. To cope with this, we show that the relative entropy functional can be approximated by the variance (with respect to the true PDS) of the difference between the true and the candidate cluster expansion energy. Restricting this approximation on the observed data leads to a weighted least squares problem making the proposed approach computationally attractive.

In this work, we tested the performance of our method in a study of canonical phase transformations in Si-Ge (two-phase coexistence to disorder) and Mg-Li (order to disorder) alloys at various compositions. We compare the relative entropy results to least squares with leave-one-out-cross-validation (least squares LOOCV) where we, for Mg-Li, observe noticeable differences in the transition temperatures. Our results are found to be in better agreement with guiding experimental data.

Bayesian global optimization in computational alloy modeling

This section emphasizes the exciting work on an approach to adaptively select simulations for the discovery of the ground state line of binary alloys with a limited computational budget.

First principles calculations are computationally expensive. This information acquisition cost, combined with exponentially high number of possible material configurations, constitutes an important roadblock towards the ultimate goal of materials by design. To overcome this barrier, one must devise schemes for the automatic and maximally informative selection of simulations. Such information acquisition decisions are task-dependent, in the sense that an optimal information acquisition policy for learning about a specific material property will not necessarily be optimal for learning about another. In this work, we develop an information acquisition policy for learning the ground state line (GSL) of binary alloys. Our approach is based on a Bayesian interpretation of the cluster expanded energy. This probabilistic surrogate of the energy enables us to quantify the epistemic uncertainty induced by the limited number of simulations which, in turn, is the key to defining a function of the configuration space that quantifies the expected improvement to the GSL resulting from a hypothetical simulation. We show that optimal information acquisition policies should balance the maximization of the expected improvement of the GSL and the minimization of the size of the simulated structure. We validate our approach by learning the GSLs of NiAl and TiAl binary alloys, where to establish the ground truth GSL we use the embedded-atom method (EAM) for the calculation of the energy of a given alloy configuration. Note that the proposed policies are directly applicable to the discovery of generic phase diagrams, if one can construct a probabilistic surrogate of the relevant thermodynamic potential.

Products: The primary products were publications. Here is a list:

Kristensen, J., Bilionis, I., and Zabaras, N. (2016, in print) “Adaptive Simulation Selection for the Discovery of the Ground State Line of Binary Alloys with a Limited Computational Budget”, in ”Recent Progress and Modern Challenges in Applied Mathematics, Modeling and Computational Science”, R. Melnik, R. Makarov, J. Belair (Eds.), Fields Institute Communications, 2016.

Aldegunde, M., Kristensen, J., and Zabaras N. (2016), “Quantifying uncertainties in first-principles alloy thermodynamics using cluster expansions”, Journal of Computational Physics (special issue)

Kristensen, J., and Zabaras, N. (2015), “Predicting low-thermal-conductivity Si-Ge nanowires with a modified cluster expansion method.” Physical Review B, 91(5)

Kristensen, J., and Zabaras, N. (2014), “Bayesian uncertainty quantification in the evaluation of alloy properties with the cluster expansion method.” Computer Physics Communications, 185(11)

Kristensen, J., Bilionis, I., and Zabaras, N. (2013), “Relative entropy as model selection tool in cluster expansions.” Physical Review B,87(17)