

Turab Lookman
Francis Alexander
Krishna Rajan *Editors*

Information Science for Materials Discovery and Design

Springer Series in Materials Science

Volume 225

Series editors

Robert Hull, Charlottesville, USA

Chennupati Jagadish, Canberra, Australia

Richard M. Osgood, New York, USA

Jürgen Parisi, Oldenburg, Germany

Tae-Yeon Seong, Seoul, Korea, Republic of (South Korea)

Shin-ichi Uchida, Tokyo, Japan

Zhiming M. Wang, Chengdu, China

The Springer Series in Materials Science covers the complete spectrum of materials physics, including fundamental principles, physical properties, materials theory and design. Recognizing the increasing importance of materials science in future device technologies, the book titles in this series reflect the state-of-the-art in understanding and controlling the structure and properties of all important classes of materials.

More information about this series at <http://www.springer.com/series/856>

Turab Lookman · Francis J. Alexander
Krishna Rajan
Editors

Information Science for Materials Discovery and Design

Editors

Turab Lookman
Theoretical Division
Los Alamos National Laboratory
Los Alamos, NM
USA

Francis J. Alexander
Computer and Computational Sciences
Division
Los Alamos National Laboratory
Los Alamos, NM
USA

Krishna Rajan
Department of Materials Design
and Innovation
University at Buffalo—The State University
of New York
Buffalo, NY
USA

ISSN 0933-033X ISSN 2196-2812 (electronic)
Springer Series in Materials Science
ISBN 978-3-319-23870-8 ISBN 978-3-319-23871-5 (eBook)
DOI 10.1007/978-3-319-23871-5

Library of Congress Control Number: 2015952059

Springer Cham Heidelberg New York Dordrecht London
© Springer International Publishing Switzerland 2016

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media
(www.springer.com)

Preface

Accelerating materials discovery has been the theme of a number reports from the Department of Energy's Office (DOE) of Basic Energy Science (BES), the National Science Foundation (NSF), the National Academies and other government agencies and professional societies. As a driver for accelerating materials discovery, the Materials Genome Initiative, announced by the President, is part of a bold plan to boost US manufacturing over the next few decades by halving the time it takes to discover and design new materials. In this plan, accelerating discovery relies on using in the material sciences large databases, computation, mathematics, and information science in a manner similar to the way that they were used to make the Human Genome Initiative a success for the biological sciences. Novel approaches are therefore being called for that can explore the enormous phase space presented by complex materials and processes. If we are to achieve the desired performance gains, then we must have a predictive capability that can guide experiments and computations in the most fruitful directions by reducing the possibilities that need to be tried.

Despite advances in computational and experimental techniques to generate large volumes of data to screen the vast search space, it is clear that the outstanding challenge remains to integrate information-theoretic tools and materials knowledge, in the form of constraints imposed by theory, to develop, robust, predictive tools for materials design and discovery. The rapidly emerging field of materials informatics provides the critical methodology that enables the discovery, identification, and harnessing of the materials "genes" for accelerated materials discovery and design.

We provide in this book a collection of articles in this nascent field which integrates contributions from the information sciences and materials communities. The collection is partly derived from a workshop held at Santa Fe, New Mexico, February 4–7, 2014 that was organized by the editors and sponsored with support from the Centers for Nonlinear Studies and Information Science and Technology at Los Alamos National Laboratory and National Science Foundation (Grant #: 13-07811). It outlines challenges and opportunities in the use of information-theoretic tools and

evaluates the state of the art on a number of materials-motivated problems. Presented are contrasting but complementary approaches, such as those based on high-throughput calculations or experiments, as well as data-driven discovery, together with the merits and challenges of machine-learning and statistical inference methods to accommodate searches within a high dimensional feature space.

The book is organized into *three* parts. In the *first* part, following a perspective of the state of the art in materials design and discovery, Chaps. 2–6 focus largely on information-theoretic tools and how they apply to specific materials problems. Chaps. 2 and 3 discuss how aspects of decision theory within a Bayesian framework can be used for optimal experimental design. In particular, Chap. 2 discusses how to decide on the best pair of experiments for inferring the parameters of a given model, as well as how to choose an experiment to distinguish between competing models. Chapter 3 discusses strategies based on methods for global optimization for choosing the next experiment to find a material with a desired property. Proceeding from problems involving regression to those requiring classification, Chap. 4 focuses on Bayesian methods for classifying objects, especially in the limit of small samples where classifier design procedures, which work well with large samples, can have problems when data is limited. The first part of this monograph is concluded with Chaps. 5 and 6 which deal with different aspects of clustering. Chapter 5 considers the effectiveness of data visualization algorithms that look for groupings of features and materials. Chapter 6 discusses how community detection, studied in statistical physics, can be used to partition a complex system into decoupled subsets at different spatial and temporal scales.

The focus of the *second* part of the book, Chaps. 7–12, is the application of informatics tools to materials science problems. Chapter 7 discusses how parameters in the additive manufacturing process may be constrained by combining simulations and experiments using feature selection and data-driven models. Learning from high-throughput data generated from electronic structure calculations is the emphasis of Chaps. 8–11. Techniques such as principal component analysis (PCA), support vector regression (SVR), partial least squares, and Kriging using Gaussian process modeling suggest new features and materials with specified properties. Chapter 8 shows how suitable dopants in an oxide may be identified for increasing water-splitting processes. Applications in Chap. 9 include the discovery of cathode materials for lithium-ion batteries and thermoelectrics. Chapter 10 focuses on the layered compounds known as MAX phases, and Chap. 11 discusses *ab initio* methods and applied crystallography tools for descriptor development to establish structure–property relationships. Chapter 12 describes hybrid methods that integrate statistical learning techniques, to extract features from the density of states for predicting elastic properties, such as bulk modulus and yet unexplored chemistries.

The *third* and final part, Chaps. 13 and 14, discusses high-throughput experiments, which generate large amounts of data. With appropriate characterization tools, the idea is to quickly identify the subspace of the large parameter space where a new compound with desired properties may be found. Such experiments, together with informatics tools, provide opportunities for “combinatorial materials science.”

Chap. 13 provides a review in the context of multifunctional materials, and Chap. 14 incorporates aspects of informatics with a focus on solar fuel applications and multicomponent oxide catalysts.

The book is aimed at an interdisciplinary audience as the subject spans aspects of statistics, computer science, and materials science and will be of timely appeal to those interested in learning about this emerging field. We are grateful to all the authors for their articles as well as their support of the editorial process.

Los Alamos, USA

Los Alamos, USA

Buffalo, NY

Turab Lookman

Francis J. Alexander

Krishna Rajan

Contents

Part I Data Analytics and Optimal Learning

1 A Perspective on Materials Informatics: State-of-the-Art and Challenges	3
T. Lookman, P.V. Balachandran, D. Xue, G. Pilania, T. Shearman, J. Theiler, J.E. Gubernatis, J. Hogden, K. Barros, E. BenNaim and F.J. Alexander	
1.1 Introduction	4
1.2 Statistical Inference and Design: Towards Accelerated Materials Discovery	5
1.3 Progress and Concluding Remarks	9
References	11
2 Information-Driven Experimental Design in Materials Science	13
R. Aggarwal, M.J. Demkowicz and Y.M. Marzouk	
2.1 Introduction	13
2.2 The Tools of Optimal Experimental Design	15
2.2.1 Bayesian Inference	15
2.2.2 Information Theoretic Objectives	16
2.2.3 Computational Considerations	18
2.3 Examples of Optimal Experimental Design	20
2.3.1 Film-Substrate Systems: Design for Parameter Inference	21
2.3.2 Heterophase Interfaces: Design for Model Discrimination	29
2.4 Outlook	37
References	39

3	Bayesian Optimization for Materials Design	45
	Peter I. Frazier and Jialei Wang	
3.1	Introduction	45
3.2	Bayesian Optimization	47
3.3	Gaussian Process Regression.	47
3.3.1	Choice of Covariance Function.	49
3.3.2	Choice of Mean Function.	51
3.3.3	Inference	51
3.3.4	Inference with Just One Observation	53
3.3.5	Inference with Noisy Observations	54
3.3.6	Parameter Estimation.	57
3.3.7	Diagnostics	58
3.3.8	Predicting at More Than One Point.	60
3.3.9	Avoiding Matrix Inversion.	61
3.4	Choosing Where to Sample	61
3.4.1	Expected Improvement	62
3.4.2	Knowledge Gradient	65
3.4.3	Going Beyond One-Step Analyses, and Other Methods	68
3.5	Software.	69
3.6	Conclusion	69
	References	73
4	Small-Sample Classification.	77
	Lori A. Dalton and Edward R. Dougherty	
4.1	Introduction	77
4.2	Classification	78
4.3	Error Estimation	82
4.4	Validity	84
4.5	MMSE Error Estimation.	87
4.6	Optimal Bayesian Classification	90
4.7	The Gaussian Model	91
4.8	Optimal Bayesian Classifier in the Gaussian Model	94
4.9	Concluding Remarks	97
	References	99
5	Data Visualization and Structure Identification	103
	J.E. Gubernatis	
5.1	Introduction	103
5.2	Theory.	104
5.3	Results.	106
5.3.1	The Piezo Data.	107
5.3.2	The Pls Data	108
5.3.3	The Tree Data	108

5.4	Concluding Remarks	109
	References	113
6	Inference of Hidden Structures in Complex Physical Systems by Multi-scale Clustering	115
	Z. Nussinov, P. Ronhovde, Dandan Hu, S. Chakrabarty, Bo Sun, Nicholas A. Mauro and Kisor K. Sahu	
6.1	The General Problem	116
6.2	Ensemble Minimization	117
6.3	Community Detection and Data Mining	118
6.4	Multi-scale Community Detection	121
6.5	Image Segmentation	123
6.6	Community Detection Phase Diagram	126
6.7	Casting Complex Materials and Physical Systems as Networks	128
6.8	Summary	133
	References	135
Part II	Materials Prediction with Data, Simulations and High-throughput Calculations	
7	On the Use of Data Mining Techniques to Build High-Density, Additively-Manufactured Parts	141
	Chandrika Kamath	
7.1	Introduction	141
7.1.1	Additive Manufacturing Using Laser Powder-Bed Fusion	142
7.2	Optimizing AM Parts for Density: The Current Approach	142
7.3	A Data Mining Approach Combining Experiments and Simulations	144
7.3.1	Using Simple Simulations to Identify Viable Parameters	145
7.3.2	Using Simple Experiments to Evaluate Simulation Results	150
7.3.3	Determining Density by Building Small Pillars	152
7.4	Experimental Results	154
7.5	Summary	154
	References	154
8	Optimal Dopant Selection for Water Splitting with Cerium Oxides: Mining and Screening First Principles Data.	157
	V. Botu, A.B. Mhadeshwar, S.L. Suib and R. Ramprasad	
8.1	Introduction	158
8.2	Screening Framework	160

8.3	First Principles Studies	160
8.3.1	Methods and Models	160
8.3.2	Enforcing the 3-Step Criteria	161
8.4	Data Analysis	164
8.4.1	Principal Component Analysis	165
8.4.2	Random Forest	166
8.5	Summary and Outlook	168
	References	168
9	Toward Materials Discovery with First-Principles Datasets and Learning Methods	173
	Isao Tanaka and Atsuto Seko	
9.1	Introduction	173
9.2	High Throughput Screening of DFT Data—Cathode Materials of Lithium ion Batteries	175
9.3	Combination of DFT Data and Machine Learning I—Melting Temperatures	177
9.4	Combination of DFT Data and Machine Learning II—Lithium ion Conducting Oxides	182
9.5	Combination of DFT Data and Machine Learning III—Thermoelectric Materials	185
	References	186
10	Materials Informatics Using <i>Ab initio</i> Data: Application to MAX Phases	187
	Wai-Yim Ching	
10.1	Introduction	187
10.2	MAX Phases: A Unique Class of Material	189
10.3	Applications of Materials Informatics to MAX Phases	191
10.3.1	Initial Screening and Construction of the MAX Database	191
10.3.2	Representative Results on Mechanical Properties and Electronic Structure of MAX	192
10.3.3	Classification of Descriptors from the Database and Correlation Among Them	197
10.3.4	Verification of the Efficacy of the Materials Informatics Tools	198
10.4	Further Applications of MAX Data	201
10.4.1	Lattice Thermal Conductivity at High Temperature	201
10.4.2	Universal Elastic Anisotropy in MAX Phases	203
10.5	Extension to Other Materials Systems	205
10.5.1	MAX-Related Systems, MXenes, MAX Solid Solutions, and Similar Layered Structures	205
10.5.2	CSH-Cement Crystals	206

10.5.3	Extension to Other Materials Systems: Bulk Metallic Glasses and High Entropy Alloys	209
10.6	Conclusions	210
	References	211
11	Symmetry-Adapted Distortion Modes as Descriptors for Materials Informatics	213
	Prasanna V. Balachandran, Nicole A. Benedek and James M. Rondinelli	
11.1	Introduction	213
11.2	Distortion Modes as Descriptors	214
11.3	Perovskite Nickelates	216
11.3.1	Statistical Correlation Analysis	217
11.3.2	Principal Component Analysis (PCA)	218
11.4	Summary	220
	References	221
12	Discovering Electronic Signatures for Phase Stability of Intermetallics via Machine Learning	223
	Scott R. Broderick and Krishna Rajan	
12.1	Introduction	223
12.2	Informatics Background and Data Processing	224
12.3	Informatics-Based Parameterization of the DOS Spectra	228
12.4	Identifying the Bulk Modulus Fingerprint	233
12.5	Summary	237
	References	237
Part III	Combinatorial Materials Science with High-throughput Measurements and Analysis	
13	Combinatorial Materials Science, and a Perspective on Challenges in Data Acquisition, Analysis and Presentation	241
	Robert C. Pullar	
13.1	Combinatorial Materials Science—20 Years of Progress?	242
13.2	Combinatorial Materials Synthesis	247
13.3	High-Throughput Measurement and Analysis	254
13.4	Data Analysis and Presentation	261
	References	267
14	High Throughput Combinatorial Experimentation + Informatics = Combinatorial Science	271
	Santosh K. Suram, Meyer Z. Pesenson and John M. Gregoire	
14.1	Tailoring Material Function Through Material Complexity: The Utility of High Throughput and Combinatorial Methods . . .	272
14.2	Materials Datasets as an Instance of Big Data	272

14.3	High Throughput Experimental Pipelines: The Example of Solar Fuels Materials Discovery	275
14.4	An Illustrative Dataset: Ni-Fe-Co-Ce Oxide Electrocatalysts for the Oxygen Evolution Reaction	276
14.5	Automating Sample Down-Selection for Maximal Information Retention: Clustering by Composition-Property Relationships	277
14.5.1	Down-Selection for Maximal Information Content	279
14.5.2	Information-Theoretic Approach	280
14.5.3	Genetic Programming Based Clustering	282
14.5.4	Calculating Membership	283
14.5.5	Application to a Synthetic Library.	284
14.5.6	Experimental Dataset.	285
14.6	The Simplex Sample Space and Statistical Analysis of Compositional Data	286
14.6.1	The Closure Effects—Induced Correlation	288
14.6.2	Illustrative Example	289
14.6.3	Sub-Compositional Coherence	290
14.6.4	Principled Analysis of Compositional Data.	290
14.6.5	Composition Spread and Distances	292
14.6.6	Interpolation of Compositional Data: Composition Profiles from Sputtering.	294
14.7	Summary and Conclusions	296
	References	297
	Index	301

Contributors

Raghav Aggarwal Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

F.J. Alexander CCS Division, Los Alamos National Laboratory, Los Alamos, USA

Prasanna V. Balachandran Theoretical Division, Los Alamos National Laboratory, Los Alamos, USA

K. Barros Theoretical Division, T-1, Los Alamos National Laboratory, Los Alamos, USA

E. BenNaim Theoretical Division, Los Alamos National Laboratory, Los Alamos, USA

Nicole A. Benedek Department of Materials Science and Engineering, Cornell University, Ithaca, USA

V. Botu Department of Chemical and Biomolecular Engineering, University of Connecticut, Storrs, CT, USA

Scott R. Broderick Department of Materials Design and Innovation, University at Buffalo—The State University of New York, Buffalo, NY, USA

S. Chakrabarty Department of Physics, Indian Institute of Science, Bangalore, India

Wai-Yim Ching Curators Professor of Physics, Kansas City, MO, USA

Lori A. Dalton The Ohio State University, Columbus, OH, USA

M.J. Demkowicz Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

Edward R. Dougherty Texas A&M University, College Station, TX, USA

Peter I. Frazier School of Operations Research & Information Engineering, Cornell University, Ithaca, NY, USA

John M. Gregoire Joint Center for Artificial Photosynthesis, California Institute of Technology, Pasadena, CA, USA

J.E. Gubernatis Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM, USA

J. Hogden CCS Division, CCS-3, Los Alamos National Laboratory, Los Alamos, USA

Dandan Hu Washington University in St. Louis, St. Louis, MO, USA

Chandrika Kamath Lawrence Livermore National Laboratory, Livermore, CA, USA

T. Lookman Theoretical Division, Los Alamos National Laboratory, Los Alamos, USA

Y.M. Marzouk Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge, MA, USA

Nicholas A. Mauro North Central College, Naperville, IL, USA

A.B. Mhadeshwar Center for Clean Energy and Engineering, University of Connecticut, Storrs, CT, USA; Present Address: ExxonMobil Research and Engineering, Annandale, NJ, USA

Z. Nussinov Washington University in St. Louis, St. Louis, MO, USA; Department of Condensed Matter Physics, Weizmann Institute of Science, Rehovot, Israel

Meyer Z. Pesenson Joint Center for Artificial Photosynthesis, California Institute of Technology, Pasadena, CA, USA

G. Pilania Materials Science Division, Los Alamos National Laboratory, Los Alamos, USA

Robert C. Pullar Departamento de Engenharia de Materiais e Cerâmica/CICECO—Aveiro Institute of Materials, Universidade de Aveiro, Campus Universitário de Santiago, Aveiro, Portugal

Krishna Rajan Department of Materials Design and Innovation, University at Buffalo—The State University of New York, Buffalo, NY, USA

R. Ramprasad Institute of Materials Science, University of Connecticut, Storrs, CT, USA; Department of Materials Science and Engineering, University of Connecticut, Storrs, CT, USA

James M. Rondinelli Department of Materials Science and Engineering, Northwestern University, Evanston, USA

P. Ronhovde Findlay University, Findlay, OH, USA

Kisor K. Sahu School of Minerals, Metallurgical and Materials Engineering, Indian Institute of Technology, Bhubaneswar, India

Atsuto Seko Department of Materials Science and Engineering, Kyoto University, Kyoto, Japan

T. Shearman Program in Applied Mathematics, University of Arizona, Tucson, USA

S.L. Suib Department of Chemistry, University of Connecticut, Storrs, CT, USA; Institute of Materials Science, University of Connecticut, Storrs, CT, USA

Bo Sun Washington University in St. Louis, St. Louis, MO, USA

Santosh K. Suram Joint Center for Artificial Photosynthesis, California Institute of Technology, Pasadena, CA, USA

Isao Tanaka Department of Materials Science and Engineering, Kyoto University, Kyoto, Japan

J. Theiler ISR Division, Los Alamos National Laboratory, Los Alamos, USA

Jialei Wang School of Operations Research and Information Engineering, Cornell University, Ithaca, NY, USA

D. Xue Theoretical Division, Los Alamos National Laboratory, Los Alamos, USA

Part I
Data Analytics and Optimal Learning

Chapter 1

A Perspective on Materials Informatics: State-of-the-Art and Challenges

**T. Lookman, P.V. Balachandran, D. Xue, G. Pilania, T. Shearman,
J. Theiler, J.E. Gubernatis, J. Hogden, K. Barros, E. BenNaim
and F.J. Alexander**

Abstract We review how classification and regression methods have been used on materials problems and outline a design loop that serves as a basis for adaptively finding materials with targeted properties.

T. Lookman (✉) · P.V. Balachandran · D. Xue · J.E. Gubernatis · E. BenNaim
Theoretical Division, T-4, Los Alamos National Laboratory, Los Alamos 87545, USA
e-mail: txl@lanl.gov

P.V. Balachandran
e-mail: pbalachandran@lanl.gov

D. Xue
e-mail: xdz@lanl.gov

J.E. Gubernatis
e-mail: jg@lanl.gov

E. BenNaim
e-mail: ebn@lanl.gov

G. Pilania
Materials Science Division, MST-8, Los Alamos National Laboratory,
Los Alamos 87545, USA
e-mail: gpilania@lanl.gov

T. Shearman
Program in Applied Mathematics, University of Arizona, Tucson 85721, USA
e-mail: toby.shearman@gmail.com

J. Theiler
ISR Division, Los Alamos National Laboratory, Los Alamos 87545, USA
e-mail: jt@lanl.gov

J. Hogden
CCS Division, CCS-3, Los Alamos National Laboratory, Los Alamos 87545, USA
e-mail: hogden@lanl.gov

K. Barros
Theoretical Division, T-1, Los Alamos National Laboratory, Los Alamos 87545, USA
e-mail: kbarros@lanl.gov

F.J. Alexander
CCS Division, Los Alamos National Laboratory, Los Alamos 87545, USA
fja@lanl.gov

© Springer International Publishing Switzerland 2016
T. Lookman et al. (eds.), *Information Science for Materials
Discovery and Design*, Springer Series in Materials Science 225,
DOI 10.1007/978-3-319-23871-5_1

1.1 Introduction

There has been considerable interest over the last few years in accelerating the process of materials design and discovery. The Materials Genome Initiative (MGI) [1], Integrated Computational Materials Engineering (ICME) [2] and Advanced Manufacturing [3] initiatives have spurred considerable activity and brought new researchers into the nascent field of materials informatics which includes the accelerated design and discovery of new materials. The activity has also highlighted some of the open questions in this emerging area and our objective here is to provide a perspective of the field in terms of general problems and information science methods that have been used to study classes of materials, and point to some of the outstanding challenges that need to be addressed. We are guided here by our own recent work at the Los Alamos National Laboratory (LANL).

One of the earliest-studied problems in modern materials informatics relates to the classification of AB solids into their stable crystal structures, based on key attributes of the chemistry and properties of the individual A and B constituents. The emphasis was on finding features that can give rise to easily visualized two-dimensional structural maps by “drawing” boundaries between classes. The problem was first studied in the 1960s [4] but Chelikowski and Phillips [5], studying the same problem in 1978, recognized the connections to information science. Realizing that energy differences between structures were rather small, they observed that “from the point of view of information theory, ...the available structural data already contain a great deal of information: about 120 bits, in the case of the AB octet compounds. Thus one can reverse the problem, and attempt to extract from the available data quantitative rules for chemical bonding in solids.” They realized that suitable combinations of orbital radii of the individual A and B atoms were appropriate features for predicting the crystal structure of the AB solids. Over the last few years, this problem has been revisited with a variety of machine learning methods (decision trees, support vector machines, gradient boosting, etc.) [6–8] and there have been a number of studies that have classified different materials classes, such as perovskites [9]. Feature selection from data remains a fundamental exercise and here principal component analysis and correlation maps have been widely employed. Recently, high-throughput approaches have been utilized to form combinations of features from a given set and then certain key combinations are down-selected [6].

The problem of materials design is about predicting the composition and processing of materials with a more desired targeted property and therefore involves regression that leads to an inference model from training data. For example, for ferroelectrics one may wish to discover lead-based or lead-free piezoelectrics with a high transition temperature or high piezoelectric coefficient. For shape memory alloys, one may seek compounds with reduced dissipation or low hysteresis. Typically, such materials are usually found in an Edisonian fashion using intuition and time-consuming trial and error. In recent years, theory has become powerful enough to predict very accurately some material characteristics, for example, *ab initio* calculations predict elastic constants, inter-atomic distances, crystal structure, polarization, etc. However, the parameter space is just too large and there are too many

possibilities, and even if nature rules out many of the possible combinations, the numbers are still staggering. Moreover, physical and chemical constraints make the realization of many theoretically possible materials impossible. Thus, one needs to successively improve or learn from available data candidate materials for further experiments and calculations. Recently, a number of studies have utilized regression methods to predict materials with given properties. However, most research in materials design has been based on high throughput approaches using electronic structure calculations. Typically, a large database is assembled with calculated properties and this is successively screened for materials with desired properties. High-throughput experiments have also been undertaken more recently to screen for candidate materials for further experiments [10, 11]. When it comes to multicomponent alloys or solid solutions, these methods have limitations. Moreover, very few studies have combined statistical inference with the high-throughput approach.

1.2 Statistical Inference and Design: Towards Accelerated Materials Discovery

Figure 1.1 illustrates our vision for the overall materials informatics/design problem. This shows a feedback loop that starts with the available assembled data (box 5), which may be obtained from multiple sources, including experiments or calculations. Materials knowledge (box 1) is then key in selecting the features and prescribing the constraints amongst them. Our aim is to train a statistical inference model that estimates the property (regression) or classification label with associated uncertainties (box 2). Classification models answer categorical questions: Is a compound stable?

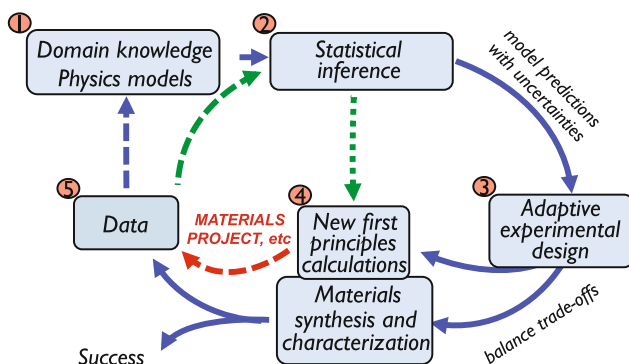


Fig. 1.1 Statistical Inference and design: A feedback loop to find a material with a desired targeted property. Prior or domain knowledge, including features, provide input to an inference model that predicts a label or a property with uncertainty. An experimental design or decision making module balances trade off between exploiting information or further exploring the high dimensional search space where the desired material may be found. A material is suggested for experimentation or calculation and the process repeats itself incorporating updated information

Is it a piezoelectric? What is its crystal symmetry? Regression models produce numerical estimates: What is the material's piezoelectric coefficient? What is its transition temperature? Because there usually is a limited quantity of training data, and because the space of possibilities is so high-dimensional, incorporation of domain knowledge is of potentially great value. Here explicitly Bayesian approaches, in which this knowledge is coded into prior probability distributions, and more traditional machine learning algorithms (such as support vector machines) in which case the domain knowledge could be incorporated as constraints or folded into the kernel design, become important [12].

Much existing work is essentially based on going from box 1 to box 4 in Fig. 1.1. A case in point are projects such as the Materials Project [13] and AFLOWLIB [14] focused on establishing databases using electronic structure calculations to make predictions. However, there are a few studies that use inference to make predictions. Examples include predictions of melting temperature [7, 8, 15] or piezoelectrics with high transition temperatures [16]. The search for piezoelectrics serves as a good example to contrast the two approaches. Extensive *ab initio* calculations were performed on a chemical space represented by $63^2 = 3969$ possible perovskite ABO_3 (up to Bi but excluding a few such as H and inert gases) end structures [17]. The number of possibilities were filtered down to 49 by discarding compounds that are nonmetallic or whose structures have small energy barriers to distortions across the morphotropic phase boundary (MPB) according to preset values. Almost no optimization or learning tools are used other than what may be involved in seeking an optimal minimum energy solution at zero temperature. All the physics is contained in this first-principles calculation, and we are not aware if any of this group's predictions of piezoelectricity have been verified experimentally. On the other hand, the approach of Balachandran et al. [16] on the same type of problem was to focus on a given subclass of piezoelectrics (e.g. Bi based) with known crystallographic and experimental data and use off-the-shelf inference tools to obtain candidates with high transition temperatures and that were formable. The tools included principal component analysis (PCA) for dimensionality reduction, partial least squares (PLS) regression for predicting transition temperatures and recursive partitioning (or decision trees) with a metric such as Shannon entropy for classification. The training data sets for PCA or regression studies were rather small (about 20 data points, 30 features) but data sets with 350 data points were also used to identify stable/formable perovskite compounds. Two new compounds were predicted, of which one has been synthesized [18], with the predicted transition temperature differing by 30–40 %. However, a key element lacking is the issue of uncertainties in predictions.

In Fig. 1.2, we demonstrate using an example, where we have used bootstrap methods (i.e. sampling with replacement) to estimate prediction uncertainties. Here, we took the same Bi-based piezoelectrics data set as that utilized in the work of Balachandran et al. [16] We generated a large number of bootstrapped samples (as opposed to using just one in the earlier work of Balachandran et al.) and utilized support vector regression (SVR) for predicting the Curie temperature (T_C). Our results with uncertainties are shown in Fig. 1.2. On average, we obtained a standard deviation of 37°C from the mean value of predicted T_C . More importantly, we also

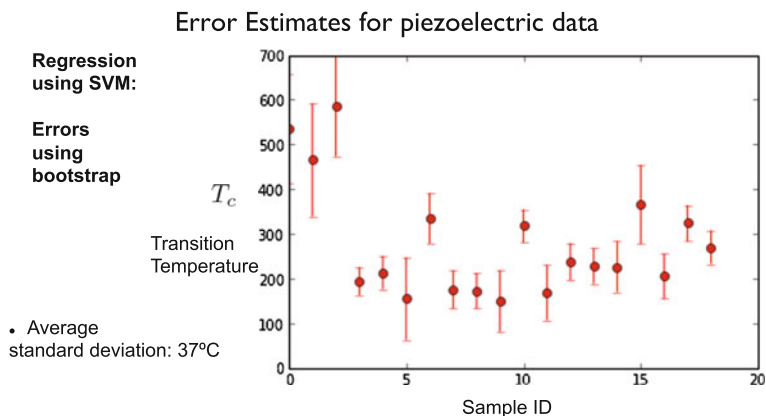


Fig. 1.2 Predictions using support vector regression (SVR) with uncertainties from bootstrap method. The piezoelectric data set of Bi-based PbTiO_3 solid solutions was used for machine learning. T_c (in $^{\circ}\text{C}$, y-axis) is the predicted ferroelectric Curie temperature at the morphotropic phase boundary (MPB). We use the SVR model and predicted T_c for two new compounds, $\text{BiLuO}_3\text{-PbTiO}_3$ and $\text{BiTmO}_3\text{-PbTiO}_3$, to be 552.5 ± 79 and 564.2 ± 97 $^{\circ}\text{C}$, respectively. Experimentally, T_c for $\text{BiLuO}_3\text{-PbTiO}_3$ was measured as 565 $^{\circ}\text{C}$ [18]

predicted the T_c for two new compounds, $\text{BiLuO}_3\text{-PbTiO}_3$ and $\text{BiTmO}_3\text{-PbTiO}_3$, to be 552.5 ± 79 and 564.2 ± 97 $^{\circ}\text{C}$, respectively with 95 % confidence. Experimentally, T_c for $\text{BiLuO}_3\text{-PbTiO}_3$ was measured as 565 $^{\circ}\text{C}$ [18], in close agreement with the current results from SVR. On the other hand, PLS predicted the T_c for $\text{BiLuO}_3\text{-PbTiO}_3$ to be 705 $^{\circ}\text{C}$. The merit of this example is that it shows in a rather modest manner that the informatics approach, even if manual and piecemeal, is potentially capable of predicting new materials.

A key aspect of our design loop is the uncertainty associated with the properties predicted from inference (box 2). These play a role in the adaptive experimental design (box 3) which suggests the next material to be chosen for further experiments or calculation (box 4) by balancing the tradeoffs between “exploration and exploitation”. That is, at any given stage a number of samples may be predicted to have given properties with uncertainties. The tradeoff is between exploiting the results by choosing to perform the next experiment on the material predicted to have the largest property or further improving the model by performing the experiment or calculation on a material where the predictions have the largest uncertainties. By choosing the latter, the uncertainty in the property is expected to (given the model, statistics) decrease, the model will probably improve and this will influence the results of the next iteration in the loop. While there is a considerable literature on error estimation methodologies, accurate and reliable error estimation with limited data is harder than simple prediction, and there is an even a stronger case for incorporating domain knowledge. [8, 19, 20]

Extracting measures of confidence, while at the same time encoding prior knowledge, is not an easy task but recent research in cancer genomics has demonstrated that increasing confidence in classification analysis built on small databases benefits

significantly from using prior knowledge [21, 22]. Prior domain knowledge constrains statistical outcomes by producing classifiers that are superior to those designed from data alone. How to use prior knowledge in classification and regression is a problem not only for materials and cancer genomics but for machine learning generally. Developing ways of constructing and using prior domain knowledge will distinguish the materials machine learning approach to classification and regression. The lesson learned from high-throughput genomics concerning classification is that, in high dimensional, small-sample settings, model-free classification is virtually impossible. The reason is that the salient property of any classifier is its error rate because the error rate quantifies its predictive capacity, which is the essential issue pertaining to scientific validity. Since the error rate must be estimated, there must be an estimation procedure and, with small samples, this procedure must be applied to the same data as that used for designing the classifier. In cancer genomics, Dalton and Dougherty [19, 20] addressed the problem by formulating error estimation as an optimization problem in a model-based framework and leads to a minimum-mean-square-error (MMSE) estimate of the classifier error. They formulate a prior probability distribution over a class of possible distributional models governing the features to be measured and the possible decisions to be made, each such model being known as a feature-label distribution. They then design a classifier from the data and an optimal MMSE error estimate is derived from the data. How well this approach will work for materials problems remains an open question.

In Figs. 1.3 and 1.4 we provide more details of our loop. Figure 1.3 shows how the loop would actually work in practice, and some of the algorithms that may be

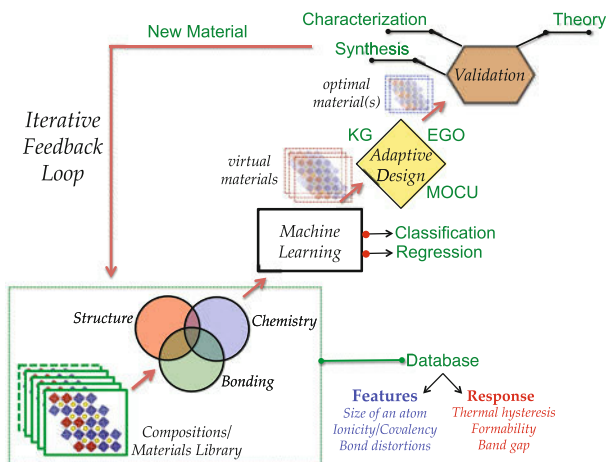


Fig. 1.3 The design loop in practice showing different stages of machine learning and adaptive design strategies with an iterative feedback loop. For completeness, we have also included experiments (synthesis and characterization), which are vital for validation and feedback. KG, EGO and MOCU stand for knowledge gradient, effective global optimization and mean objective cost of uncertainty, respectively

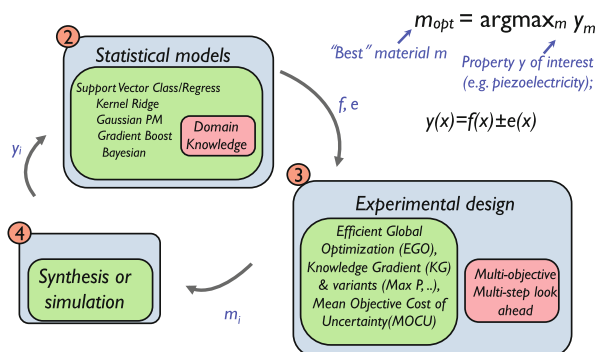


Fig. 1.4 A sub-component of our adaptive design loop showing the synergy between statistical models (box 2), experimental design (box 3) and validation (typically via experimental synthesis or simulation as shown in box 4). Statistical models use the available data to fit a regression model (f) along with an uncertainty measure (e). The experimental design component then evaluates the tradeoff between exploitation and exploration and suggests the “best” material (y_i) for validation. Here the term “best” need not correspond to a material with the optimal response. Alternatively, it refers to the choice of a material that would reduce the overall uncertainty in our model. Different statistical learning (including Bayesian learning) and adaptive design methods are given

used as part of the statistical inference and design tools, are shown in greater detail in Fig. 1.4. The green emphasize algorithms that can be utilized today and the red represent areas requiring further study and development. Design algorithms include well known exploitation-exploration strategies such as efficient global optimization (EGO) [23], and the closely related knowledge gradient(KG) [24] based on single-step look ahead.

1.3 Progress and Concluding Remarks

Our work at LANL has involved studying a number of materials problems along the lines of the approach described. These include problems involving classification learning and regression, which essentially involve an inner loop of Fig. 1.1 with boxes 2, 4 and 5. We have examined the role of features in classifying AB octet solids [8] and perovskites [9], as well as predicting new ductile RM intermetallics, where R and M are rare earth and transition metal elements, respectively [25]. These studies have suggested new features that led to better classification as well as new materials. In the case of RM intermetallics, we have shown that machine learning methods naturally uncover the functional forms that mimic most frequently used features in the literature, thereby providing a mathematical basis for feature set construction without *a priori* assumptions [25]. Our classification models (Fig. 1.5) that use orbital radii as features predicted that ScCo, ScIr, and YCd should be ductile, whereas each was previously proposed to be brittle. These results show it is possible to design

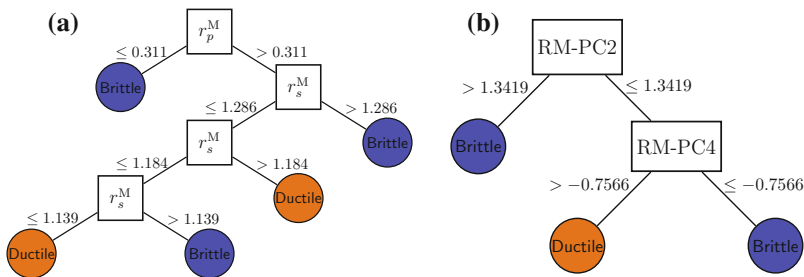


Fig. 1.5 Classification learning using decision trees to predict whether a given RM intermetallic, where R and M are rare earth and transition metal elements, respectively, is brittle and ductile. **a** Decision tree that uses the orbital radii as features and (b) Decision tree that uses the principal components (RM-PC2 and RM-PC4) that automatically extracts features in the form of the linear combinations of orbital radii. For example, RM-PC2 is defined as $-0.70r_p^M + 0.08r_s^M - 0.71r_d^M$. Features r_p^M , r_s^M and r_d^M are the p -, s - and d -orbital radii of atom-M, respectively

targeted mechanical properties in intermetallic compounds, which has significant implications for next-generation multi-component alloy discovery.

Our on-going work on multi-objective regression includes predicting functional polymers with large band gaps, as well as large dielectric constants for energy storage applications. Similarly, we are also performing high-throughput density functional theory (DFT) calculations to generate large data sets, which are subsequently mined using machine learning methods to identify new and previously unexplored candidate water splitting compounds for catalysis.

In the area of adaptive design, our focus has been on demonstrating the feedback loop of Figs. 1.1 or 1.3 with tight coupling to an “oracle”, which can be experiments (synthesis and characterization) or calculations. Specific materials studies include discovering new low thermal dissipation shape memory alloys, as well as Pb-free piezoelectric solid-solutions starting from experimental data on specific multicomponent systems. The search spaces can be well defined, for example, they can be a factor of 10^5 greater than the size of the training data. In addition, extensive databases from *ab initio* calculations become invaluable in benchmarking the various algorithms. For example, elastic moduli data for the hexagonal layered M_2AX phases consist of a library of 240 compounds. The *ab initio* data of the elastic constants and moduli were taken from the literature [26] with results well calibrated to experiments. In the M_2AX phases, X-atoms reside in the edge-connected M octahedral cages and the A atoms reside in slightly larger right prisms [27]. These M_2AX phases represent a unique family of materials with layered crystal structure and both metallic- and ceramic-like properties. We used orbital radii of M, A, and X atoms from the Waber-Cromer scale [28] as features, which include the s -, p -, and d -orbital radii for M, while the s - and p -orbital radii were used for A and X atoms. With the M_2AX data, we benchmarked our adaptive design strategy, i.e. explored different training set sizes, regressors, regressor/optimization combinations, etc., and uncovered invaluable guidelines that were eventually useful for real materials design problems.

Implementing the loop using simulation codes allows us to optimize the use of these codes in seeking a well defined set of parameters or constraints for given targeted outcomes. For example, an industry standard code for simulating semiconducting materials is APSYS (Advanced Physical Models of Semiconductor Devices). It is based on 2D/3D finite element analysis of electrical, optical and thermal properties of compound semiconductor devices, with silicon as a special case with an emphasis on band structure engineering and quantum mechanical effects. Inclusion of various optical modules allows one to configure applications involving photosensitive or light emitting diodes (LEDs). We have been recently using APSYS to investigate how to optimize the LED structure (number of quantum wells, indium concentration) of GaAs based systems for highest internal quantum efficiencies at high currents.

In summary, the use of classification and regression methods, in combination with optimization strategies, has the potential to impact discovery and design in materials science. What is needed is to establish how these tools perform on an array of materials classes with differing physics in order to distill some guiding principles for use by the materials community at large.

Acknowledgments We acknowledge funding support from a Laboratory Directed Research and Development (LDRD) DR (#20140013DR) at the Los Alamos National Laboratory (LANL).

References

1. *Materials Genome Initiative for Global Competitiveness* (2011)
2. S.R. Kalidindi, M. De Graef, Materials data science: current status and future outlook. *Ann. Rev. Mater. Res.* **45**(1), 171–193 (2015)
3. T.D. Wall, J.M. Corbett, C.W. Clegg, P.R. Jackson, R. Martin, Advanced manufacturing technology and work design: towards a theoretical framework. *J. Organ. Behav.* **11**(3), 201–219 (1990)
4. E. Mooser, W.B. Pearson, On the crystal chemistry of normal valence compounds. *Acta Crystallogr.* **12**, 1015–1022 (1959)
5. J.R. Chelikowsky, J.C. Phillips, Quantum-defect theory of heats of formation and structural transition energies of liquid and solid simple metal alloys and compounds. *Phys. Rev. B* **17**, 2453–2477 (1978)
6. L.M. Ghiringhelli, J. Vybiral, S.V. Levchenko, C. Draxl, M. Scheffler, Big data of materials science: critical role of the descriptor. *Phys. Rev. Lett.* **114**, 105503 (2015)
7. Y. Saad, D. Gao, T. Ngo, S. Bobbitt, J.R. Chelikowsky, W. Andreoni, Data mining for materials: computational experiments with AB compounds. *Phys. Rev. B* **85**, 104104 (2012)
8. G. Pilania, J.E. Gubernatis, T. Lookman, Structure classification and melting temperature prediction of octet AB solids via machine learning. *Phys. Rev. B* **91**, 124301 (2015)
9. G. Pilania, P.V. Balachandran, J.E. Gubernatis, T. Lookman, Predicting the formability of ABO₃ perovskite solids: a machine learning study. *Acta Crystallogr. B* **71**, 507–513 (2015)
10. S.M. Senkan, High-throughput screening of solid-state catalyst libraries. *Nature* **394** (6691), 350–353, 07 (1998)
11. H. Koinuma, I. Takeuchi, Combinatorial solid-state chemistry of inorganic materials. *Nat. Mater.* **3**, 429–438 (2004)
12. T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning* (Springer, New York, 2008)

13. A. Jain, S.P. Ong, G. Hautier, W. Chen, W.D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K.A. Persson, Commentary: the materials project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**(1) (2013)
14. S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R.H. Taylor, L.J. Nelson, G.L. Hart, S. Sanvito, M. Buongiorno-Nardelli, N. Mingo, O. Levy, AFLOWLIB.ORG: a distributed materials property repository from high-throughput ab initio calculations. *Comput. Mater. Sci.* **58**, 227–235 (2012)
15. A. Seko, T. Maekawa, K. Tsuda, I. Tanaka, Machine learning with systematic density-functional theory calculations: application to melting temperatures of single- and binary-component solids. *Phys. Rev. B* **89**, 054303 (2014)
16. P.V. Balachandran, S.R. Broderick, K. Rajan, Identifying the inorganic gene for high-temperature piezoelectric perovskites through statistical learning. *Proc. R. Soc. A: Math. Phys. Eng. Sci.* **467**(2132), 2271–2290 (2011)
17. R. Armiento, B. Kozinsky, M. Fornari, G. Ceder, Screening for high-performance piezoelectrics using high-throughput density functional theory. *Phys. Rev. B* **84**, 014103 (2011)
18. W. Hu, Experimental search for high Curie temperature piezoelectric ceramics with combinatorial approaches. Ph.D. dissertation, Iowa State University (2011)
19. L.A. Dalton, E.R. Dougherty, Optimal classifiers with minimum expected error within a Bayesian framework—Part I: discrete and Gaussian models. *Pattern Recognit.* **46**(5), 1301–1314 (2013)
20. L.A. Dalton, E.R. Dougherty, Optimal classifiers with minimum expected error within a Bayesian framework—Part II: properties and performance analysis. *Pattern Recognit.* **46**(5), 1288–1300 (2013)
21. K.E. Lee, N. Sha, E.R. Dougherty, M. Vannucci, B.K. Mallick, Gene selection: a Bayesian variable selection approach. *Bioinformatics* **19**(1), 90–97 (2003)
22. E.R. Dougherty, A. Zolnari, U.M. Braga-Neto, The illusion of distribution-free small-sample classification in genomics. *Curr. genomics* **12**(5), 333–341 (2011)
23. D.R. Jones, M. Schonlau, W.J. Welch, Efficient global optimization of expensive black-box functions. *J. Glob. Optim.* **13**(4), 455–492 (1998)
24. W. Powell, I. Ryzhov, *Optimal Learning*, Wiley Series in Probability and Statistics (Wiley, Hoboken, 2013)
25. P.V. Balachandran, J. Theiler, J. M. Rondinelli, T. Lookman, Materials Prediction via Classification Learning. *Sci. Rep.* **5**, 13285 (2015)
26. M.F. Cover, O. Warschkow, M.M.M. Bilek, D.R. McKenzie, A comprehensive survey of M_2AX phase elastic properties. *J. Phys.: Condens. Matter* **21**(30), 305403 (2009)
27. M.W. Barsoum, M. Radovic, Elastic and mechanical properties of the MAX phases. *Ann. Rev. Mater. Res.* **41**, 195–227 (2011)
28. J.T. Waber, D.T. Cromer, Orbital radii of atoms and ions. *J. Chem. Phys.* **42**(12), 4116–4123 (1965)

Chapter 2

Information-Driven Experimental Design in Materials Science

R. Aggarwal, M.J. Demkowicz and Y.M. Marzouk

Abstract Optimal experimental design (OED) aims to maximize the value of experiments and the data they produce. OED ensures efficient allocation of limited resources, especially when numerous repeated experiments cannot be performed. This chapter presents a fully Bayesian and decision theoretic approach to OED—accounting for uncertainties in models, model parameters, and experimental outcomes, and allowing optimality to be defined according to a range of possible experimental goals. We demonstrate this approach on two illustrative problems in materials research. The first example is a parameter inference problem. Its goal is to determine a substrate property from the behavior of a film deposited thereon. We design experiments to yield maximal information about the substrate property using only two measurements. The second example is a model selection problem. We design an experiment that optimally distinguishes between two models for helium trapping at interfaces. In both instances, we provide model-based justifications for why the selected experiments are optimal. Moreover, both examples illustrate the utility of reduced-order or surrogate models in optimal experimental design.

2.1 Introduction

Experiments are essential prerequisites of all scientific research. They are the basis for developing and refining mathematical models of physical reality. Experimental data are used to infer model parameters, to improve the accuracy of model-based predictions, to discriminate among competing models, to assess model validity, and

R. Aggarwal · M.J. Demkowicz (✉)

Department of Materials Science and Engineering, Massachusetts Institute of Technology,
77 Massachusetts Avenue, Cambridge, MA 02139, USA

e-mail: demkowicz@mit.edu

Y.M. Marzouk

Department of Aeronautics and Astronautics, Room 37-451, Massachusetts Institute
of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA

e-mail: ymarz@mit.edu

© Springer International Publishing Switzerland 2016

T. Lookman et al. (eds.), *Information Science for Materials*

Discovery and Design, Springer Series in Materials Science 225,

DOI 10.1007/978-3-319-23871-5_2

to improve design and decision-making under uncertainty. Yet experimental observations can be difficult, time-consuming, and expensive to acquire. Maximizing the value of experimental observations—i.e., designing experiments to be *optimal* by some appropriate measure—is therefore a critical task. Experimental design encompasses questions of where and when to measure, which variables to interrogate, and what experimental conditions to employ.

Conventional experimental design methods, such as factorial and composite designs, are largely used as heuristics for exploring the relationship between input factors and response variables. By contrast, *optimal* experimental design uses a concrete hypothesis—expressed as a quantitative model—to guide the choice of experiments for a particular purpose, such as parameter inference, prediction, or model discrimination. Optimal design has seen extensive development for linear models (where the measured quantities depend linearly on the model parameters) endowed with Gaussian distributions [5]. Extensions to nonlinear models are often based on linearization and Gaussian approximations [15, 21, 36], as analytical results are otherwise impractical or impossible to obtain. With advances in computational power, however, optimal experimental design for nonlinear systems can now be tackled directly using numerical simulation [48, 49, 64, 65, 84, 89, 93, 96].

This chapter will present an overview of model-based optimal experimental design, connecting this approach to illustrative applications in materials science—a field replete with potential applications for optimal experimentation. We will take a fully Bayesian and decision-theoretic approach. In this formulation, one first defines the *utility* of an experiment and then, taking into account uncertainties in both the parameter values and the observations, chooses experiments by maximizing an *expected utility*. We will define these utilities according to information theoretic considerations, reflecting the particular experimental goals at hand.

The evaluation and optimization of information theoretic design criteria, in particular those that invoke complex physics-based models, requires the synthesis of several computational tools. These include: (1) statistical estimators of expected information gain; (2) efficient optimization methods for stochastic or noisy objectives (since expected utilities are typically evaluated with Monte Carlo methods); and (3) reduced-order or surrogate models that can accelerate the estimation of information gain. For a simple film-substrate system, we will present an example of such a reduced-order model, derived from physical scaling principles and an “offline” set of detailed/full model simulations. This is but one example; reduced-order models constructed through a variety of techniques have practical use in a wide range of optimal experimental design applications [48].

The rest of this chapter is organized as follows. Section 2.2 will present the foundational tools of optimal experimental design, beginning with Bayesian inference and proceeding to discuss several information theoretic design criteria. It will also discuss the computational challenges presented by this formulation. Section 2.3 will illustrate the information-driven approach with two examples: optimal design for *parameter inference*, in the context of a film-substrate system; and optimal design for *model selection*, in the context of heterophase interfaces in layered metal composites. Section 2.4 will discuss open questions and topics of ongoing research.