Learning from an Informatics Approach Applied to Materials Design

Dr. Michael Krein*, Jason Poleski**, Dr. Richard Barto***

* Lockheed Martin ATL, 3 Executive Campus, Cherry Hill, NJ, USA, michael.krein@lmco.com
** Lockheed Martin MST, 199 Borton Landing Road, Moorestown, NJ, USA, jason.poleski@lmco.com
*** Lockheed Martin ATL, 3 Executive Campus, Cherry Hill, NJ, USA, rick.barto@lmco.com

ABSTRACT

Lockheed Martin has designed an informatics system, the Nanotechnology Materials Data Mining, Modeling & Management (NMD-M3), to capture and guide nanotechnology-focused experimentation. The system is built upon open software standards, is modular, and supports multiple experimental configurations. Data is captured via a series of macros tailored to experimental needs; input fields are subjected to data provenance filters, and missing data is flagged. Current challenges in data provenance, visualization, and exploration are highlighted. NMD-M3 system capabilities, practical uses, limitations are explored, and directions forward towards pervasive tight integration of informatics in experimental design are discussed.

Keywords: materials, informatics, modeling, visualization, nanotechnology

1 INTRODUCTION

Nanomaterials will significantly impact commercial and military applications when context and understanding of their vast design and property space is achieved. Identification of important materials design parameters is a first step in a materials design workflow that can optimize system-specific properties of interest. This workflow must be flexible enough to handle disparate problems, but still allow data to be visualized in familiar context. As NMD-M3 was being developed, it was quickly discovered that a machine learning framework alone would provide limited value. Without a basic understanding of the underlying data and informatics models, the information content within the data, the domain of applicability of the models [1], and the conclusions drawn from both would be arbitrary. Initial work focused on data provenance definition: how to validate, represent, store, and retrieve experimental data.

Defining dataset and informatics best practices was necessary to obtain actionable modeling outcomes, and to set user expectations of model accuracy. It was discovered that there was no single optimum set of practices [2]. NMD-M3 users need a diverse set of best practices. The focus of the tool is to present information about dataset and model quality to promote exploration of the data and continual application and refinement of informatics models.

At the same time, user feedback is critical to the development of the user interface and overall capabilities of the tool; as an exemplar of agile software development, over 60 development releases in a three year timeframe rapidly matured the modeling workflow, shown in Figure 1.

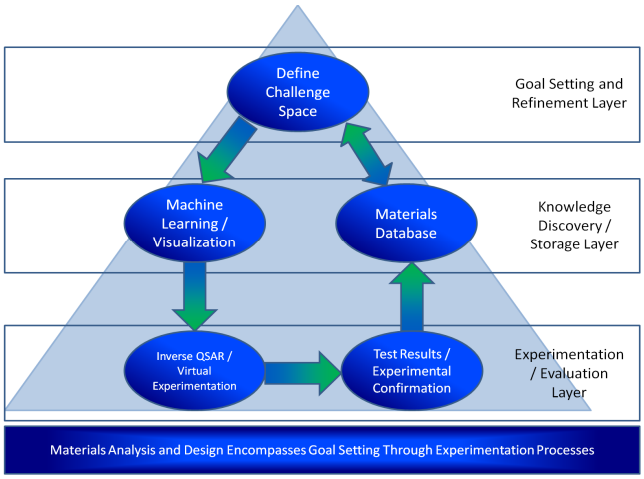

Figure 1: High-level user workflow of the NMD-M3 materials informatics tool, where distinct user concepts are represented as layers.

2 DATA AND REPRESENTATION

The path towards the success or failure of data analysis begins with appropriate understanding and handling of materials data.

2.1 Data Tagging and Import

Experimental test data is often acquired from multiple instruments with unique conditions; when possible, environmental (location, temperature, relative humidity, date and time, etc.) and handling information is stored. Input data canonicalization is a challenge solved locally and manually by frequent discussions with domain experts. This information can be used to partition data and understand systematic error. Estimates of error are captured for modeling context, and can be used as inputs to independent modeling efforts. Data arrives in numerous formats with the requirements that experiments be tagged by unique identifiers and that experimental features be also uniquely

identified. Microsoft Excel, comma-, and tab-delimited files are checked for conformity via a series of macros, and are stored in a relational database for ease of retrieval.

2.2 Representing Non-Numeric Data

Beyond the challenges of acquiring numerical data, there is a wealth of information collected that can not be used in standard machine learning approaches. Ordinal representation of categorical data may not be appropriate, and a sparse representation of mixed alphanumeric data as binary features can lead to very large datasets with limited information content. Automated image processing routines that return numeric feature vectors for SEM or TEM imagery and simulation-based features complicate the data landscape.

Such representations require expert knowledge of the systems, and dataset-by-dataset inspection. The effectiveness of such representations was best judged by the incremental improvement in statistical model performance over baseline models.

2.3 Preprocessing Considerations

When incorporating experimental data, features of that data are often disparate, and must be preprocessed before statistical models can be built and evaluated. Within NMD-M3, the user can choose to leave data unscaled, standardized, or median absolute centered. Collinear features beyond a user-defined threshold may be removed automatically to reduce model complexity and improve model interpretability. Features whose variance is greater than a user-specified threshold may be identified as "noisy" and can be ignored, removed, or have their maximal variance components capped to the defined cutoff. For the majority of internal data, data standardization, conservative collinear feature removal (removal decisions at > 0.9 R^2 correlation), and noisy feature capping at 6 sigma from the mean provided a good initial preprocessing strategy [2].

2.4 Visualization Strategies

The visualization of and interaction with high-dimensional data is a field unto itself, and there are standard projection and data clustering techniques readily available. NMD-M3 uses these methods side-by-side to present multiple views of the same preprocessed data.

Principal Component Analysis (an example of which is shown in Figure 2) [3], Independent Component Analysis, and Multidimensional Scaling [4] of datasets can be displayed in 2D and interactive 3D modes, and k-means clustering of these projections can be used to automatically select subsets of data on which to perform additional analyses.

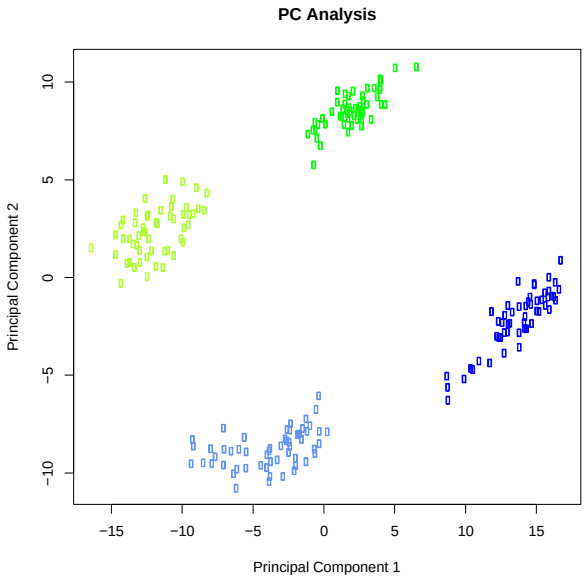

Figure 2: An exemplar of Principal Component Analysis. High dimensional data is represented as a projection in a lower dimensional space. Four clusters of data are clearly seen in this projection.

3 MODELING AND INTERPRETATION

3.1 Methods and Metrics

Similar to data visualization, machine learning is a large field with a variety of well-established, off-the-shelf learning algorithms. From a regression view, Multiple Linear Regression, Partial Least Squares Regression, Support Vector Machines, and Random Forests are widely used[2] and represent common linear and non-linear learning paradigms. Regression model performance characteristics, such as the coefficient of determination and root mean squared error of the model predictions are reported side-by-side with graphical representations of models to compare and contrast modeling methods. The relevance and consistency of features can be determined by analysis of automatically generated starplots [5], where a feature importance is plotted over subsets of data, as shown in Figure 3.

Global measures of feature importance can be characterized by sensitivity analysis, where changes in features lead to deviations in predictions.

Assessments of accuracy of the models are estimated through ten-fold internal crossvalidation [6] and external test set handling. The range of predictions generated through ten rounds of tenfold crossvalidation can be used to assess both accuracy and prediction and can be represented as error bars on standard *predicted vs. actual* plots (Figure 4) where predictions are compared to known experimental

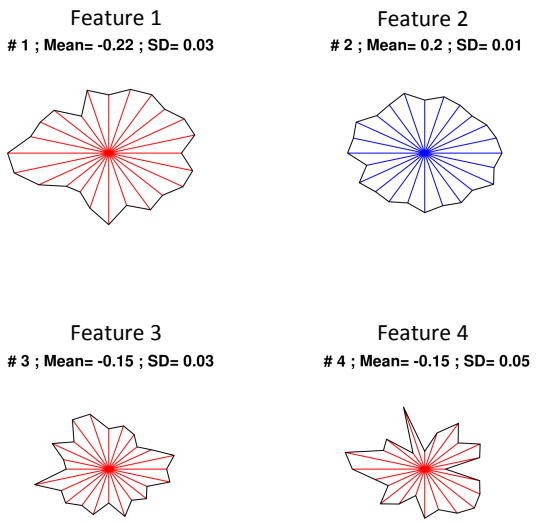

Figure 3: Example starplots – the uniformity of the importance of individual features is captured by the relative differences in spoke length, where each spoke is a unique subset of data. Blue and red colors represent positive and negative importance of the features, respectively.

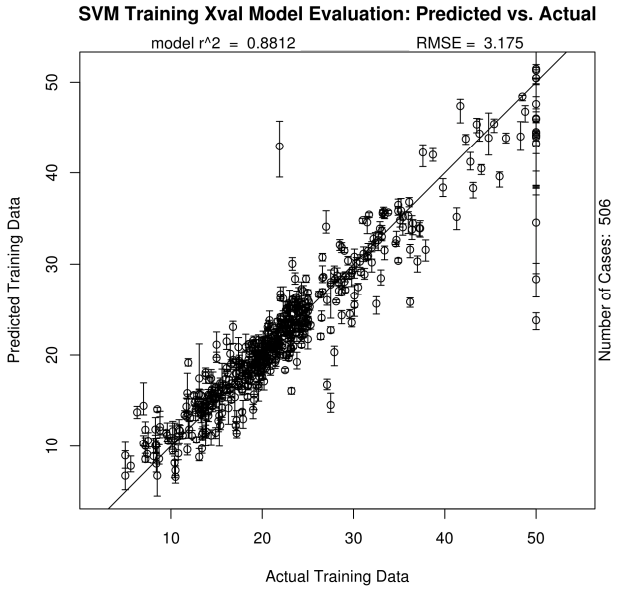

Figure 4: Predicted-vs.-Actual plot of a fictitious dataset. Title indicates SVM modeling, and error bars on each data point represent the range of predictions from crossvalidation. Summary model quality measures can be seen at the top of the figure.

values. Crossvalidation statistics, starplots, sensitivity analysis, and predicted vs. actual plots are model agnostic and can be used synergistically to compare one method to another, and to optimize a set of models. Crossvalidation statistics automatically tune modeling methods via a grid search of possible model parameters, and sensitivity analysis results can automatically refine models by deselecting features that are regarded as unimportant to the models. Once optimized and validated, models can be used for prospective predictions. A virtual design of experiments can be carried out and further refined to focus experimental efforts. This approximation to solve the inverse problem of materials design is a first step in high-dimensional tradespace analysis.

4 DISCUSSION

Continued usage and refinement of the NMD-M3 modeling system has provided key insights to materials systems of interest, and has also demonstrated current system limitations and need for future developmental work.

4.1 Current Limitations

Practical limitations to modeling using the current version of NMD-M3 include the dataset size in experiment and feature space, and the granularity of the hyperparameter grid used. For active modeling, all preprocessed data and models are stored within memory, and multithreaded execution forks the entire variable space. Thus, memory will typically be the limiting factor in working with large data (beyond hundreds of thousands of experiments).

A current limitation in model optimization is that model parameters are chosen via crossvalidated grid search: a very finely-sampled grid does lead to prohibitive execution time, without guaranteed convergence. Thus, alternative model optimization methods are currently being investigated.

4.2 Directions Forward

Customer interaction has revealed primary interests in feature development, consensus- and multiobjective-modeling that motivate future work. There is a significant need for intelligent data canonicalization and feature representation. Subsequent to representation, there is a desire for the fusing of simulations and experimental characterization using a variety of methods, *including appropriate treatment and propagation of errors*. At the same time, implementation of consensus modeling and voting schemes, where different materials models are compared and contrasted intelligently, is necessary to better inform future experimental decisions. Further, true multiobjective modeling is necessary to accurately navigate complex materials tradespaces. A holistic representation and treatment of materials systems from the atoms up is a grand vision of materials informatics and integrated computational materials engineering (ICME). We believe data mining tools like NMD-M3 will be an inportant part of ICME frameworks going forward.

REFERENCES

[1] S. Weaver and M. P. Gleeson, “The importance of the domain of applicability in QSAR modeling,”

Journal of Molecular Graphics and Modelling, vol. 26, pp. 1315-1326, 2008.

[2] M. Krein, T.-w. Huang, L. Morkowchuck, D. K. Agrafiotis, and C. M. Breneman, "Developing Best Practices for Descriptor-Based Property Prediction: Appropriate Matching of Datasets, Descriptors, Methods, and Expectations," in Statistical Modelling of Molecular Descriptors in QSAR/QSPR, M. Dehmer, K. Varmuza, and D. Bonchev, Eds., ed: Wiley-VCH, 2012.

[3] L. Eriksson, P. Andersson, E. Johansson, and M. Tysklind, "Megavariate analysis of environmental QSAR data. Part I – A basic framework founded on principal component analysis (PCA), partial least squares (PLS), and statistical molecular design (SMD)," Molecular Diversity, vol. 10, pp. 169-186, 2006.

[4] D. K. Agrafiotis, D. N. Rassokhin, and V. S. Lobanov, "Multidimensional scaling and visualization of large molecular similarity tables," Journal of Computational Chemistry, vol. 22, pp. 488-500, 2001.

[5] C. M. Breneman, K. P. Bennett, M. J. Embrechts, S. Cramer, M. Song, J. Bi, and N. Sukumar, "Descriptor Generation, Selection and Model Building in Quantitative Structure-Property Analysis," in Experimental Design for Combinatorial and High Throughput Materials Development, J. N. Cawse, Ed., ed New York: John Wiley, 2002, pp. 203-238.

[6] R. D. Cramer, J. D. Bunce, D. E. Patterson, and I. E. Frank, "Crossvalidation, Bootstrapping, and Partial Least Squares Compared with Multiple Regression in Conventional QSAR Studies," Quantitative Structure-Activity Relationships, vol. 7, pp. 18-25, 1988.