

14th International Conference on Nanomaterials and Nanotechnology, March 30-31, 2017 | Madrid, Spain -Materials Science in the AI age: high-throughput library generation, machine learning and a pathway from correlations to the underpinning physics**Mohammad Musa Yamali**

Anadolu University, Turkey

The use of advanced data analytics and applications of statistical and machine learning approaches ('AI') to materials science is experiencing explosive growth recently. In this prospective, we review recent work focusing on generation and application of libraries from both experiment and theoretical tools, across length scales. The available library data both enables classical correlative machine learning, and also opens the pathway for exploration of underlying causative physical behaviors. We highlight the key advances facilitated by this approach, and illustrate how modeling, macroscopic experiments and atomic-scale imaging can be combined to dramatically accelerate understanding and development of new material systems via a statistical physics framework. These developments point towards a data driven future wherein knowledge can be aggregated and used collectively, accelerating the advancement of materials science. The use of statistical and machine learning algorithms (broadly characterized as 'Artificial Intelligence' herein) within the materials science community has experienced a resurgence in recent years. However, AI applications to material science have ebbed and flowed through the past few decades. For instance, Volume 700 of the Materials Research Society's Symposium Proceedings was entitled "Combinatorial and Artificial Intelligence Methods in Materials Science," more than 15 years ago, and expounds on much of the same topics as those at present, with examples including high-throughput screening, application of neural networks to accelerate particle simulations, and use of genetic algo-

gorithms to find ground states. One may ask the question as to what makes this resurgence different, and whether the current trends can be sustainable. In some ways this mirrors the rises and falls of the field of AI, which has had several bursts of intense progress followed by 'AI winters'. The initial interest was sparked in 1956, where the term was first coined, and although interest and funding was available, computational power was simply too limited. A rekindling began in the late 1980s, as more algorithms (such as backpropagation for neural networks, or the kernel method for classification) were utilized. The recent spike has been driven in large part by the success of deep learning, with the parallel rise in GPU and general computational power. The question becomes whether the current, dramatic progress in AI can translate to the materials science community. In fact, the key enabling component of any AI application is the availability of large volumes of structured labeled data – which we term in this prospective "libraries." The available library data both enables classical correlative machine learning, and also opens a pathway for exploration of underlying causative physical behaviors. We argue in this prospective that, when done in the appropriate manner, AI can be transformative not only in that it can allow for acceleration of scientific discoveries, but also that it can change the way materials science is conducted.

The recent acceleration of adoption of AI/machine learning-based approaches in materials science can be traced back to a few key factors. Perhaps most

pertinent is the Materials Genome Initiative, which was launched in 2011 with an objective to transform manufacturing via accelerating materials discovery and deployment. This required the advancement of high-throughput approaches to both experiments and calculations, and the formation of online, accessible repositories to facilitate learning. Such databases have by now become largely mainstream with successful examples of databases including Automatic Flow for Materials Discovery (AFLOWLIB), Joint Automated Repository for Various Integrated Simulations (JARVIS-DFT), Polymer Genome, Citrination, Materials Innovation Network, etc. that host hundreds of thousands of datapoints from both calculations as well as experiments. The timing of the initiative coincided with a rapid increase in machine learning across commercial spaces, largely driven by the sudden and dramatic improvement in computer vision, courtesy of deep neural networks, and the availability of free packages in R or python (e.g., scikit-learn) to apply common machine learning methods on acquired datasets. This availability of tools, combined with access to computational resources (e.g., through cloud-based services, or internally at large institutions) was also involved. It can be argued that one of the main driving forces within the materials science community was an acknowledgement that many grand challenges, such as the materials design inverse problem, were not going to be solved with conventional approaches. Moreover, the quantities of data that were being acquired, particularly at user facilities such as synchrotrons or microscopy centers, was accelerating exponentially, rendering traditional analysis methods that relied heavily on human input unworkable. In the face of the data avalanche, it was perhaps inevitable that scientists would turn to the methods provided via data science and machine learning. Please note commercial software is identified to specify procedures. Such identification does not imply recommendation by the National Institute of Standards and Technology.

Thus, the question becomes, how can these newly found computational capabilities and 'big' data be leveraged to gain new insights and predictions for materials? There are already some answers. For example, the torrent of data from first principles simulations has been used for high throughput screening of candidate materials, with notable successes. Naturally, one asks the question as to what insights can be gained from similar databases based not on theory, but on experimental data, e.g. of atomically resolved structures, along with their functional properties. Of course, microstructures have long been optimized in alloy design. Having libraries (equivalently, databases) of these structures, with explicit mentioning of their processing history, can be extremely beneficial not just for alloys but for many other material systems, including soft matter. These databases can be used for e.g. utilizing known knowledge of similar systems to accelerate the synthesis optimization process, to train models to automatically classifying structures and defects, and to identify materials with similar behaviors that are exhibited, potentially allowing underlying causal relationships to be established.

In this prospective, we focus on the key areas of library generation of material structures and properties, through both simulations/theory, and imaging. High-throughput approaches enable both simulation and experimental databases to be compiled, with the data used to build models that enable property prediction, determine feature importance, and guide experimental design. In contrast, imaging provides the necessary view of microstates enabling the development of statistical mechanical models that incorporate both simulations and macroscopic characterization to improve predictions and determine underlying driving forces. Combining the available experimental and theoretical libraries in a physics-based framework can accelerate materials discoveries, and lead to lasting transformations of the way materials science research is approached worldwide.