CSSI Elements: Software: **Autonomous, Robust, and Optimal In-Silico Experimental Design Platform for Accelerating Innovations in Materials Discovery**

PI: Byung-Jun Yoon[1], Co-PIs: Raymundo Arroyave[2], Xiaofeng Qian[2], Xiaoning Qian[1]
[1]Dept. Electrical & Computer Engineering, [2]Dept. Materials Science & Engineering, Texas A&M University

Award #: 1835690

Abstract

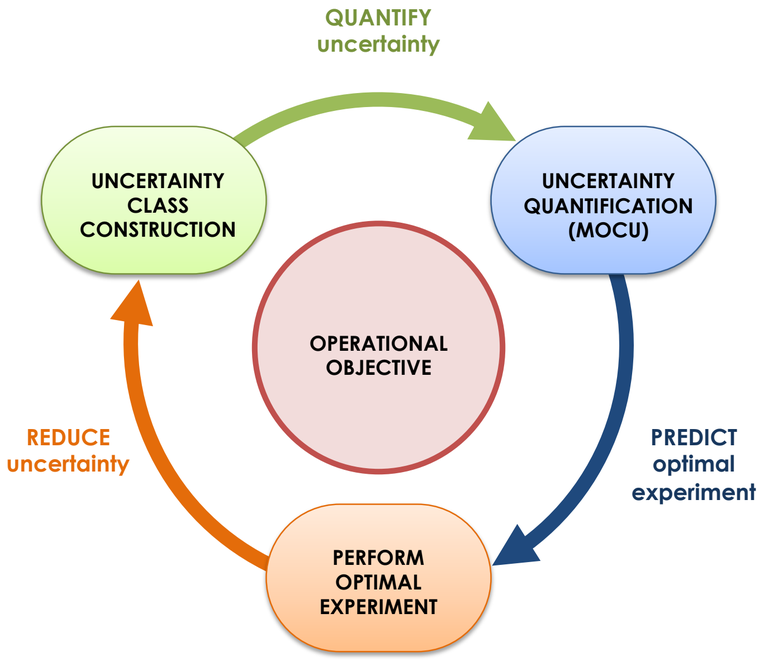
Accelerating the development of novel materials that have desirable properties is a critical challenge as it can facilitate advances in diverse fields across science, engineering, and medicine. However, the current prevailing practice in materials discovery relies on trial-and-error experimental campaigns and/or high-throughput screening approaches, which cannot efficiently explore the huge design space to develop materials with the targeted properties. Furthermore, measurements of material composition, structure, and properties often contain considerable errors due to technical limitations in materials synthesis and characterization, making this exploration even more challenging. This project aims to develop an effective in-silico experimental design platform to accelerate the discovery of novel materials. The platform is built on optimal Bayesian learning and experimental design methodologies that can translate scientific principles into predictive models, in a way that takes model and data uncertainty into account. The optimal Bayesian experimental design framework will enable the collection of smart data that can help exploring the material design space efficiently, without relying on slow and costly trial-and-error and/or high-throughput screening approaches.

Limitations of Current Existing Approaches Materials Discovery

- Most existing data-enabled approaches for materials discovery rely on costly data-collection based on (1) *trial-and-error* experimental campaigns or (2*) high-throughput screening* methods (both real/virtual)
- Such data requirements often become practical bottlenecks for accelerating materials discovery through computational approaches
- Existing experimental/computational approaches do not provide effective means for synergistic integration and optimal planning of real experiments and computational simulations
- Furthermore, existing strategies are often not built to improve the knowledge about a given system, hence the generated data from experiments/simulations do not necessarily enhance system knowledge

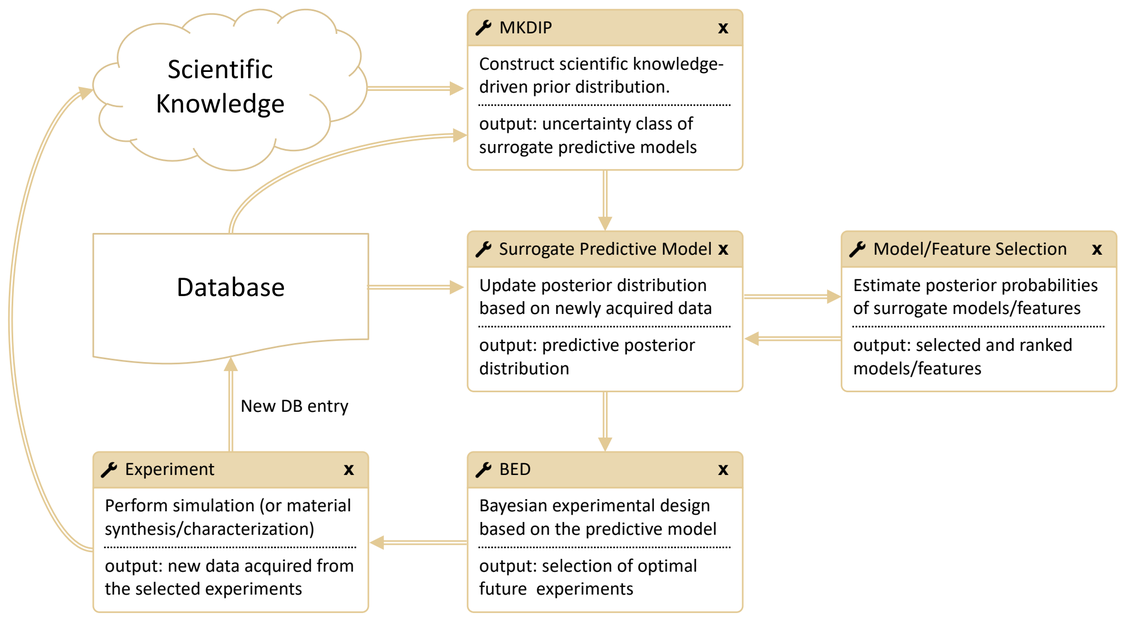
Addressing the Challenges Towards Autonomous Materials Discovery

- This CSSI project aims to design and implement a **robust and optimal *in-silico* experimental design platform for accelerated materials discovery**
- Furthermore, we aim to provide a preliminary demonstration of an effective "autonomous" simulation-based materials discovery platform
- Primary focus is on the development of **efficient Bayesian learning and experimental design framework** that features/enables
  1. Prior construction scheme that can incorporate scientific knowledge and first principles in materials science
  2. Integration of data and scientific knowledge
  3. Quantification of model/data uncertainty
  4. Optimal experimental design for uncertainty reduction
- The Bayesian framework in this project is built on the concept of **MOCU** (mean objective cost of uncertainty) – for **objective-based uncertainty quantification** and **objective-based optimal experimental design**



Open-Access Materials Discovery Platform

- The developed Bayesian learning/experimental design tools will benefit from modular design, and they will be included in **MSGalaxy Toolshed** thereby integrated within the MSGalaxy workflow management system
- Furthermore, the developed tools will enrich **TAMMAL** (Texas A&M Materials Modeling Automation Library)

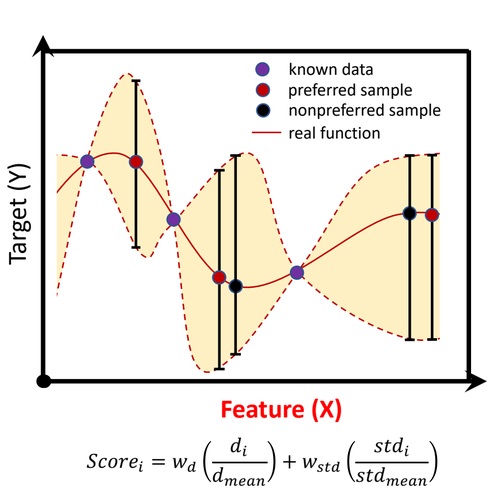


Research Progress Highlights

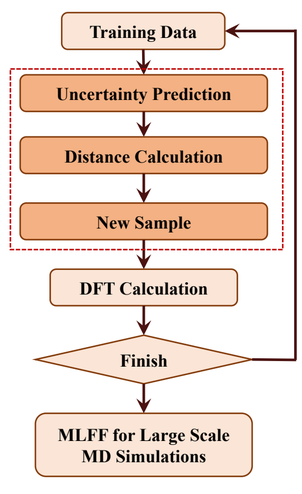
- **Automatic Feature Engineering**
  - We have developed RL-based scheme for automated feature engineering scheme that can find (1) physically meaningful features (2) that can accurately predict functional properties, to be used within our Bayesian learning & experimental design framework.
  - Tree-based semantic grammar has been adopted to guide feature engineering in both reinforcement learning and Bayesian optimization.
  - Promising preliminary results have been obtained when performance was compared to a state-of-the-art method called SISSO (Sure Independence Screening and Sparse Operation).
  - Results show "objective-based" or "goal-oriented" feature engineering is critically important

	selected descriptors	run time	pred. acc.	# of generated features
	Primary Features		82.5%	
DQL	$\sqrt[4]{IE_B} - IE_A + \frac{IE_A}{rcov_A}$	1 hour	97.5%	123,623
DQL	$\frac{IE_A IE_B (d_{AB} - r_{covA})}{exp(\chi_A)\sqrt{r_{covB}}}$	6.8 hour	100%	514,196
SISSO		8.6 hour	100%	830,877

- **Active Learning for Efficient Sampling of Configuration Space**
  - We have developed an active learning scheme to efficiently sample large configuration space for developing density-functional theory (DFT) based machine learning force field (MLFF) for large scale molecular dynamics (MD) simulations



$$Score_i = w_d\left(\frac{d_i}{d_{mean}}\right) + w_{std}\left(\frac{std_i}{std_{mean}}\right)$$



Benchmark: Atomic Force Prediction