Accelerated Materials Discovery

Turab Lookman
Los Alamos National Laboratory
Collaborators

A. Gopakumar, A. Emery (Northwestern), R. Dehghannasiri (Texas A&M), A. Mannodi-kannakkithodi, D. Trujillo (UConn), Doewon Seo (Illinois), D. Karpov (NMSU), B. Kowalski (Case Western), N. Lubbers (BU), B. Leduc (Cambridge), Z. Liu (Harbin), R. Yuan, Deqing Xue (Xian Jiatong)

P. Balachandran, D. Xue, G. Pilania, A. Kumar

Students/Postdocs

J. Gubernatis, E. Ben-Naim, J. Hogden, J. Theiler, K. Barros
R. Prasankumer, Q. Jia (Buffalo), F. J. Alexander (BNL)
E. Fohtung

LANL

NMSU/LANL

E. Dougherty, X. Qian, *Center for Genomic Signaling*
A. Sehirlioglu
A. Zunger
C. Wolverton, J. Randonelli
W. Cao
C. Humphrey, E. Salje
L. Varshney
X. Ding, X. Ren

Texas A & M
Case Western
NREL, Colorado
Northwestern
PSU/Harbin
Cambridge
IBM/ Illinois
Xian Jiatong

Postdocs at LANL
Q: How do we accelerate the process of selecting chemistry, composition, processing conditions to find materials with targeted properties?
Q: How do we accelerate the process of selecting chemistry, composition, processing conditions to find materials with targeted properties?
Q: How do we **accelerate** the process of selecting chemistry, composition, processing conditions to find materials with **targeted** properties?
Q: How do we accelerate the process of selecting chemistry, composition, processing conditions to find materials with targeted properties?
Q: How do we **accelerate** the process of selecting chemistry, composition, processing conditions to find materials with **targeted** properties?
Q: How do we accelerate the process of selecting chemistry, composition, processing conditions to find materials with targeted properties?
Materials discovery involves small amounts of data:

- Uncertainties, Multicomponent, vast search space
- Search for better Pb-free piezoelectrics

**Chemical space of** $63^2 = 3969$

**Screen for perovskite compounds**

- Nonmetallic $\Delta_{KS} > 0.25 \text{eV}$
- Small $\Delta E$ for distorted phases across MPB $\Delta E < 0.5 \text{eV}$

**Computational Design**

$ABO_3$

**Experimental Design**

$\text{Ba(Ti}_{0.8}\text{Zr}_{0.2})\text{O}_3$

$\text{(Ba}_{0.7}\text{Ca}_{0.3})\text{TiO}_3$

$\text{Ba(Ti}_{1-m}\text{X}_m)\text{O}_3-x(Ba}_{1-n}\text{Y}_n)\text{TiO}_3$

$n=0.12$ ; $m=0.30$

$n=0.18$ ; $m=0.40$

$20-30$ samples, search space $\sim 10^3$

Armiento et al. PRB (2011) 49

Xue et al. PNAS (2016)
Adaptive Learning for materials discovery

- iteratively improve predictions

1. Domain knowledge
   Physics models

2. Surrogate model
   $y(x) = f(x) \pm e(x)$

3. Optimal experimental design
   balance trade-offs

4. Computations
   Materials synthesis and characterization

5. Data

Success
Adaptive Learning for materials discovery

- iteratively improve predictions

1. Domain knowledge
   Physics models

2. Surrogate model

3. Optimal experimental design

4. Computations
   Materials synthesis and characterization

5. Data

- Current state of art single steps (5 to 2, 2 to 4 or 4 to 5);
  No inner feedback loop

\[ y(x) = f(x) \pm e(x) \]

Global search in high-dimensional space: ‘Exploit vs Explore’
Bayesian Global Optimization for large scale computational tasks OR experiments

Preliminary Experiments
Variables
Initial Designs

Response surfaces

Surrogate model
Bayesian

\[ \hat{y} \]

New designs
Infill points

\[ y \]

\[ x \]

High fidelity, Expensive Computer Calculation

Value of Information  (Howard’ 65)
Utility functions  (Lindley’ 55, ’72)

Bayesian Optimization

P[I] Kushner’64
E[I] Mockus’78
E[I] Jones’88
Knowledge gradient (Powell)

Measurement Policies
Online: bandit problems
Offline : ranking and selection
: stochastic search

Probability Collectives
(Wolpert’98)

Constraints,
Multiobjective,
Multifidelity
Parallel function evaluations

Probability Collectives
(Wolpert’98)

Mean Objective Cost of Uncertainty (MOCU)
(Dougherty’2013)

Design of Experiments (DOE)

Guide choice of experiments in an efficient manner

- **Factors** $X_1, \ldots, X_k \rightarrow Y$

- **Designs**
  - Randomized Complete Block
  - Factorial: Full, Fractional, Composite
    - Box-Behnken
  - Random
  - Latin HyperCube

**Optimal Design**: Find sample to optimize $Y$

$$y_i = f_\theta(x_i) \pm e_i$$

$$e \sim \mathcal{N}(0, \sigma)$$

$$y_i = \theta^T x_i \pm e_i$$

$$\text{Var}(\theta) = \sigma^2 (X^T X)^{-1}$$

$$\text{Var}(y_i) = \sigma^2 (X^T X)^{-1} x_i$$

<table>
<thead>
<tr>
<th>Optimal design</th>
<th>Objective</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$-optimal</td>
<td>minimize trace ${M_X^{-1}}$</td>
</tr>
<tr>
<td>$D$-optimal</td>
<td>minimize ${\det M_X}^{-\frac{1}{p}}$</td>
</tr>
<tr>
<td>$E$-optimal</td>
<td>minimize max eigenvalue ${M_X^{-1}}$</td>
</tr>
<tr>
<td>$G$-optimal</td>
<td>minimize max var ${f(x)}$, $x \in R$</td>
</tr>
<tr>
<td>$I$-optimal</td>
<td>minimize trace ${MM_X^{-1}}$</td>
</tr>
</tbody>
</table>

Santner et al.’ 2003
Experimental design via Bayesian Optimization

\[
\max_{x \in \mathbb{R}^d} f(x)
\]

Step 1: Estimate objective

\[ y_i = f(x_i) + \varepsilon_i \]

Data

\[ \mathcal{D}_{1:t} = \{x_{1:t}, f(x_{1:t})\} \]

Gaussian process

\[ f(x) \sim \mathcal{GP}(m(x), k(x, x')) \]

Posterior

\[ P(f_{t+1} | \mathcal{D}_{1:t}, x_{t+1}) = \mathcal{N}(\mu_t(x_{t+1}), \sigma_t^2(x_{t+1})) \]

O'Hagan, 70
Step 2: Sample with utility function

Choose next point to **maximize uncertainty**:

Choose next point to **minimize uncertainty**:

$$x_t = \text{argmax}_x u(x | D_{1:t-1})$$

(150 data points, 3 in training set.)
Utility functions: encoding decision criteria

- Expected Improvement $\rightarrow$ EGO (Kushner, Mockus, Jones)
- Expected Quantile Improvement (Picheny)
- Lower or Upp. Confidence Bounds (Cox and Johnson, Srinivas)
- Sequential Kriging Optimization (Huang)
- Knowledge Gradient (Powell, Frazier)
- Mean Objective Cost of Uncertainty (Yoon, Xian, Dougherty)
- Maximum Variance
Results

\[ N_{\text{train}} = \text{End 3} \]
Results

\( N_{\text{train}} = \text{sample 3} \)

\[
MAE = \frac{\sum_{i=1}^{n}|y_i - \mu|}{n}
\]
**Expected improvement**

Mockus’78, Jones’88

**Improvement:** \( I(x) = \max(Y - f_{\text{max}}, 0) \)

\[
E[I(x)] = \int_{f_{\text{max}}}^{\infty} (Y - f_{\text{max}}) P(Y \mid x') dY
\]

\[
= (\hat{y} - f_{\text{max}}) \Phi\left( \frac{\hat{y} - f_{\text{max}}}{\sigma} \right) + \sigma \phi\left( \frac{\hat{y} - f_{\text{max}}}{\sigma} \right)
\]

Choose sample with largest “Expected Improvement”: \( x_i = \arg\max_x [E(I(x))] \)

**Limiting cases:**

- Small \( \sigma \) \( E \rightarrow \hat{y} - f_{\text{max}} \)
- Large \( \sigma \) \( E \rightarrow \sigma \)

Choose \( \hat{y} \) better than best-so-far \( f_{\text{max}} \)

**Exploitation** (local, utilize model)

Choose \( \hat{y} \) with largest uncertainty \( \sigma \)

**Exploration** (global, improve model)
Mean Objective Cost of Uncertainty

Identify experiment expected to maximally reduce MOCU one step ahead

Assume cost function, \( g(\theta; c) \)

\( \theta \) Unknown parameters
\( c \) characterizes an experiment

If \( \theta \) known, then the optimal material would be

\[
c'_\theta = \arg \min_{c \in C} g(\theta; c)
\]

If \( \theta \) unknown, then the robust material is

\[
c'' = \arg \min_{c \in C} E_{\theta}[g(\theta; c)]
\]

Expected Loss in using robust instead of optimal

\[
MOCU = E_{\theta}[g(\theta; c'_\theta) - g(\theta; c'')]\]

Property of optimal material at state \( \theta \)
Property of IBR material across all \( \theta \)

MOCU: Selection of next experiment

- Experiments $C1, C2, C3, \textit{etc}$
- Outcomes: $C_1, C_2, C_3$

Remaining after outcome $c_i$

$$MOCU = E_{\theta | C_i = c_i} \left[ g(\theta; c'') - g(\theta; c'_\theta) \right]$$

Conditional distribution $f(\theta | C_i = c_i)$

Using Bayes, obtain $f(c | \theta)$

Expected remaining MOCU given that experiment $c_i$ undertaken:

$$MOCU_i = E_{c_i} E_{\theta | C_i = c_i} \left[ g(\theta; c'') - g(\theta; c'_\theta) \right]$$

$c_i^* = \arg \min_{i \in 1, \ldots, n} MOCU_i$
Problem: Minimize energy dissipation by selecting optimal dopants in as few experiments as possible

Use physics based model as surrogate for objective:

\[ g(h, \sigma, c; b) = \max \left( \frac{1}{(b_1 h + b_2 \sigma + b_3 h \sigma + b_4) c + b_5 h + b_6 \sigma + b_7 h \sigma + b_8}, 0 \right), \]
Active learning loop

Experimental design → Chosen measurement $i^*, c^*$ → Conducting measurement → Robust material $i(\theta|X_{i^*,c^*} = \hat{x}), c(\theta|X_{i^*,c^*} = \hat{x})$ → Posterior $f_{\mathcal{M}}(h, \sigma|X_{i^*,c^*} = \hat{x})$ → Prior $f_{\mathcal{M}}(h, \sigma)$

Training data → Data fitting → fitted model → posterior distribution → Update prior distribution → Experimental design → Making measurement

material with low energy dissipation
Comparison with several strategies

- Proposed
- Pure exploitation
- Random selection
- Optimal material

Average energy dissipation vs. Number of measurements
Optimize codes: LED structures for highest efficiencies at high currents

Quantum Efficiencies

APSYS code

Features:
- barrier height
- 5 quantum wells

Gaussian Process + EGO

Leduc et al., 2016
**Examples**

**Data-driven → Experiments**

**Strategy:** Ti$_{50}$ Ni$_{50-x-y-z}$ Cu$_x$ Pd$_y$ Fe$_z$

**Total possibilities:** 800,000 (22 known)

**Target:** Minimize Thermal Hysteresis

---

**Experiments**

- **Largest electrostrain**

\[
(Ba_{100-x-y}Ca_xSr_y)(Ti_{100-u-v}Zr_uSn_v)O_3
\]

**Yuan et al., Advanced Materials, 1702884 2018**

**Xue et al., Nature Comm., 2016**

---

**Use theory with data → Experiments**

**Strategy:** Ti$_{50}$ Ni$_{50-x-y-z}$ Cu$_x$ Pd$_y$ Fe$_z$

**Total possibilities:** 800,000 (22 known)

**Target:** Minimize Thermal Hysteresis

- 36 experiments
- 14 with superior properties

- **Xue et al., PNAS, 2016**
 Computations

**Predict new materials**

- Ruddlesden-Popper Phases
  - Balachandran et al., Nature Comm., 2017
  - Ruddlesden-Popper Phases

**Multi-fidelity data**

- Contours: Number of HSE06 Bandgap Training Data Points (n_t)
  - Number of PBE Bandgap Training Data Points (n_p)

**Accelerate codes**

- Quantum Efficiencies
  - APSYS code

- Target: Multi-fidelity data
  - lower accuracy cheaper with fewer
  - higher accuracy expensive
  - accuracy comparable to expensive

- Balachandran et al., Nature Comm., 2017
  - Pilania et al., Computational Materials Sci, 2017
**Example:** Find NiTi alloy with lowest hysteresis

**Strategy:** \( Ti_{50} Ni_{50-x-y-z} Cu_{x} Pd_{y} Fe_{z} \)

Search space of multicomponent alloy ~800K
22 training samples (.003%) with measured \( \Delta T \)

\( x \leq 20 \%, y \leq 5 \%, z \leq 20 \% \)

50 – \( x – y – z \geq 30 \% \)

Composition control 0.1%

**Knowledge:**
- Transition temperatures influenced by valence electron concentration
- Hysteresis influenced by atomic size
- Relative stability influenced by changes in electron number

**Features:**
- Valence electron number
- Radii: metallic, Clementi, Zunger
- Electronegativity
- Pettifor Chemical scale

Each \( Ti_{50} Ni_{50-x-y-z} Cu_{x} Pd_{y} Fe_{z} \) weighted fraction of features
Adaptive Design for Alloy Discovery

Feedback from experiments: augmented dataset with 4 new alloys

(i) Alloy Dataset

Ti50Ni50-x-y-zCu_xFe_yPd_z

- 22 initial alloys
- \( \Delta T \) measured

(ii) Valence electron number
- Pauling electronegativity
- Pseudopotential radius
- Metallic radius
- Atomic radius
- Pettifor chemical scale

(iii) Train model for predicting \( \Delta T \) from features with associated uncertainties

Ti50Ni50-x-y-zCu_xFe_yPd_z

- 797,482 alloys with features
- \( \Delta T \) predicted from model with uncertainties
- \( x \leq 20\% \)
- \( y \leq 5\% \)
- \( z \leq 20\% \)
- 50-x-y-z \( \geq 30\% \)

(iv) Choice of candidate alloys for experiments

Xue et al., Nature Comm., 2016
Experiments

Melting

Rolling

Heat treatment

Measurement: DSC

Thermal hysteresis
Material Performance for Synthesized Alloy

### Figure B

- Heat flow (w g⁻¹)
- Temperature (K)

### Table I

<table>
<thead>
<tr>
<th>Iterations</th>
<th>Composition</th>
<th>ΔT (K)</th>
<th>Transformation temperature (K)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>Ti₅₀Ni₄₆.₈Cu₀.₈Fe₂.₀Pd₀.₃</td>
<td>2.64</td>
<td>289.95</td>
</tr>
<tr>
<td>6</td>
<td>Ti₅₀Ni₄₄.₂Cu₁.₉Fe₃.₈Pd₀.₁</td>
<td>2.53</td>
<td>243.43</td>
</tr>
<tr>
<td>6</td>
<td>Ti₅₀Ni₄₆.₇Cu₀₈Fe₂.₃Pd₀.₂</td>
<td>1.₈₄</td>
<td>2₈₁.₇₇</td>
</tr>
<tr>
<td>7</td>
<td>Ti₅₀Ni₄₈.₁Cu₀₂Fe₁₅Pd₀₂</td>
<td>2.₀₉</td>
<td>3₀₁.₈₆</td>
</tr>
<tr>
<td>7</td>
<td>Ti₅₀Ni₄₆.₅Cu₁₂Fe₂₂Pd₀₂</td>
<td>2.₃₂</td>
<td>2₈₃.₇₉</td>
</tr>
</tbody>
</table>

Five new best alloys, which were found in iterations 6 and 7, amongst 14 with the lowest ΔT. From a total of 9 iterations, which resulted in 36 new alloys, 14 had a ΔT smaller than 3.15 K, the lowest in the original training set of 22. Transformation temperature is given by the endothermic peak in the DSC curve.
How good does the model have to be?

Blue: in-sample
Green: out-of-sample

next selection

\[ \mu^* \]

\[ \mu \]
Piezoelectrics

Electrical energy

Mechanical energy

Applications

**Piezoelectric Applications**

**Automotive**
- Spark Ignition
- Fuel Atomization
- Knock Sensors
- Tire Pressure
- Safety/Security Alarms
- Sound Systems
- Motion sensors
- Flow level sensors

**Consumer Goods**
- Telephone Ringers
- Speakers
- Humidifier Atomizers
- Smoke Detectors
- Automatic Lighting
- Security Systems
- Jewelry Cleaners

**Commercial**
- Hydrophones
- Fans
- Production sensors
- Counters
- Security
- Robotics/Toys
- Game machines
- Massaging

**Medical**
- Fetal Heart Monitors
- Blood Flow Diagnosis
- Flow Meters/Controllers
- Insulin Pumps
- Vaporizers/Nebulizers
- Ultrasonic Surgery
- Ultrasonic Imaging

**Industrial**
- Ultrasonic Cleaning
- Ultrasonic Welding
- Ultrasonic Machining
- Flow Detection
- Pneumatic Valves

**Computers**
- Ink Jet Printers
- Disk Drive
- Keyboards

**Aerospace/Defense**
- Sonar
- Hydrophones
- Ring Laser Gyros

Los Alamos National Laboratory
LDRD
Search for BaTiO₃ based solid solutions with relatively large electrostrains

\[(Ba_{100-x-y}Ca_xSr_y)(Ti_{100-u-v}Zr_uSn_v)O_3\]  

\[100 - x - y \geq 60\% \quad x \leq 40\% \quad y \leq 30\%\]

\[100 - u - v \geq 60\% \quad u \leq 30\% \quad v \leq 30\%\]

Compositional control 1%
Figure of merit and target

- BT-based ceramics
- BCT-BZT(BST)
- BNT-based ceramics
- KNN-based ceramics
- Lead-based ceramics
- This work

- BCT-0.5BZT under 20kV/cm
- BCT-0.5BZT under 5kV/cm
Experiments - fabrication

Conventional solid-state reaction method

Standard, but one composition in the phase diagram needs the whole procedure.

- Not high throughput
- Two days for four samples
- High dopant concentration samples need to tune the sintering conditions.
Experimental Comparison of design strategies: Search for BaTiO3-based large electrostrains

1. **Input dataset**
   - Material features, \( x \)
   - Targeted property, \( y \)
     - Electronegativity
     - Ionic radius
     - Volume
     - Ionic displacements
     - Polarization
     - Dopant effects on transitions

2. **Statistical Inference and machine learning**
   \[ y = f(x) + \text{uncertainties} \]

3. ~605000 possible compounds

4. **Design and optimization**: choice of candidates for next experiment

5. **Synthesis and characterization**
   Four recommended samples for synthesis and characterization

6. **Feedback**: Augmented dataset

7. **Success?**

**Categories**:
- Exploitation
- Exploration
- Trade-off between exploitation & exploration
- Random Selection

---

MaRIE
Relative Figure of Merit

\((\text{Ba}_{0.84}\text{Ca}_{0.16})(\text{Ti}_{0.90}\text{Zr}_{0.07}\text{Sn}_{0.03})\text{O}_3)\)
Comparing outcome to predictions

(a) Measured strain (%) vs. # Iterations
- Training data
- Trade-off
- Exploitation
- Exploration
- Random

(b) Predicted strain (%) vs. # Iterations
- Initial training data
- Group 1
- Group 2
- Group 3
- Group 4
- Group 5

(c) Measured strain (%) vs. Predicted strain (%)
- Initial training data
- Group 1
- Group 2
- Group 3
- Group 4
- Group 5
Comparison of BCT-BZT based piezoelectrics

(Ba$_{0.84}$Ca$_{0.16}$)(Ti$_{0.90}$Zr$_{0.07}$Sn$_{0.03}$)O$_3$

(Ba$_{0.84}$Ca$_{0.16}$)(Ti$_{0.90}$Zr$_{0.10}$)O$_3$

(Ba$_{0.84}$Ca$_{0.15}$)(Ti$_{0.90}$Zr$_{0.10}$)O$_3$
**Q:** What combination of \( m, n \) and chemistries (Al, Li, ...) will optimize phase boundaries, response?

- **Design criterion: Vertical**

- \( n = 0.12 ; m = 0.30 \)

- \( n = 0.18 ; m = 0.40 \)

- **Example: Importance of knowledge**
(111) \[ f_R = \frac{1}{2}a_2(\tau - \tau_C)p_R^2 + \frac{1}{6}(a_6 - \frac{1}{9}a_6' - \frac{1}{27}a_6'')p_R^6 \]

(100) \[ f_T = \frac{1}{2}a_2(\tau - \tau_C)p_T^2 + \frac{1}{6}(a_6 - a_6')p_T^6 \]

At MPB \[ f_R = f_T \]

\[ \tau = f(\tau_C, a_2, a_6, .., p_R, p_T) \]
Learning from theory + data

$x(Ba^{1-m}Ca^m)TiO_3 - Ba(Zr^nTi^{1-n})O_3$

$18% < m < 50%; 15% < n < 30%$

(1200 phase diagrams)

**Features:**

- Order parameters: Polarization, Strain

\[ \Delta V = V_T - V_R \quad u_T, u_R \]

\[ t_f = \frac{R_A + R_o}{R^R + R^o} \quad r_{\text{eff-nuc}} = \frac{A_{\text{enc}}}{B_{\text{enc}}} \quad r_{\text{elec-neg}} = \frac{A_{\text{en}}}{B_{\text{en}}} \]

**Training data:** 19 phase diagrams

\[ \tau = f(\tau_C, a_2, a_6, \ldots, p_R, p_T) \]

- Prior distribution subject to constraints

- Take samples from posterior
Fig. S 2. 20 phase diagrams used in the present study. 1 - 19 serve as the training data whereas 20 is the new system predicted by our Bayesian approach and adaptive design. All phase diagrams are characterized by a morphotropic phase boundary as shown in red symbols and curves. (both show composition dependence)
Predictions/ synthesis from model + data

Best in the training data:
$$\text{Ba}(\text{Zr}_{0.2}\text{Ti}_{0.8})\text{O}_3 - (\text{Ba}_{0.7}\text{Ca}_{0.3})\text{TiO}_3$$

New Prediction:
$$\text{Ba}(\text{Zr}_{0.3}\text{Ti}_{0.7})\text{O}_3 - (\text{Ba}_{0.5}\text{Ca}_{0.5})\text{TiO}_3$$

Xue et al., PNAS, 2016
Challenges – Path forward

- **No free lunch theorem!**  
  (Wolpert, 98)

- **Integrate** physics models

- **Guiding principles** for different classes of problems (materials, ..)
BioManufacturing
with Intelligent Adaptive Control

LASL’s MANIAC (Mathematical Analyzer, Numerical Integrator, and Computer or Mathematical Analyzer, Numerator, Integrator, and Computer) was an early computer built under the direction of Nicholas Metropolis, 1952.
What Problem are we Trying to Solve?

• Plastics–synthetic polymers made from petrochemicals–have revolutionized our society.

• However, plastics are over-engineered for durability, and plastic pollution is now a scourge on our planet.

• We propose to help solve this problem with an innovative process to discover, design, and develop new biopolymers with improved functionalities, balancing durability with faster degradability in the environment.

• Use biology as a template for new chemistries.

The Plastic Lifecycle. Top: Past: Cradle to Grave. Plastic from petroleum is used for most packaging, but ends up in landfills or in the environment. Bottom: Future: Cradle to Cradle. Algae can be used to produce bio-based plastics. Along with improved recycling methods (e.g. P&G’s Head & Shoulders shampoo bottle made from pelletized beach plastic) biodegradable compostable bioplastic can be broken down into basic molecules that can be captured and re-used as nutrients in agriculture.
A model or recipe that will predict the synthesis conditions/chemistries to make polymers with targeted performance from the biology or chemistry routes with given confidence levels.

A general approach that allows for efficient exploration of vast chemical spaces and synthesis conditions using ML to guide chemistry and/or biology based synthesis routes

Novel, degradable biopolymers (with a favorable combination of other properties, e.g., breathability, durability, mechanical strength) synthesized via the proposed biosynthesis route