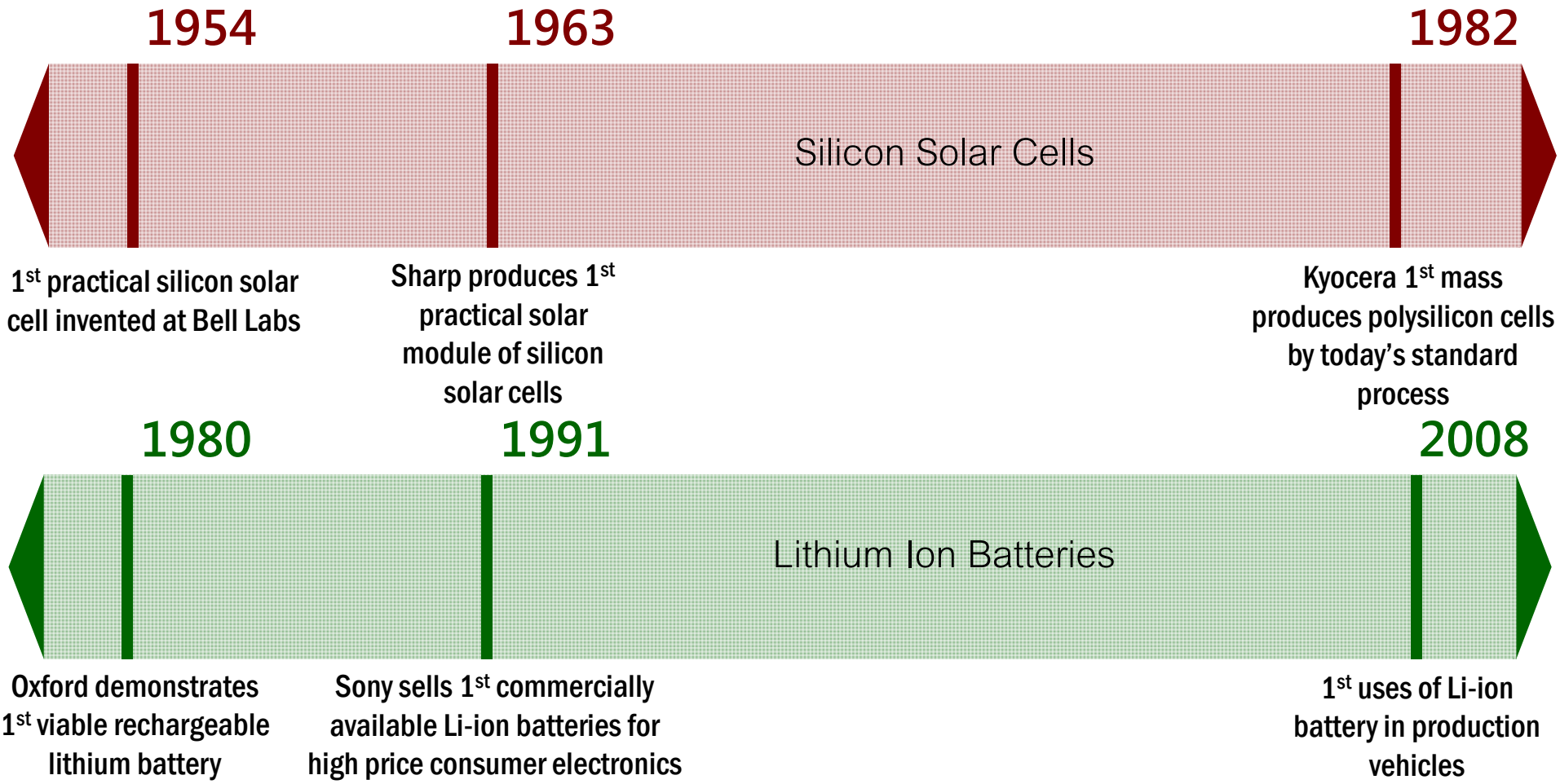
Text and Data Mining for Material Synthesis

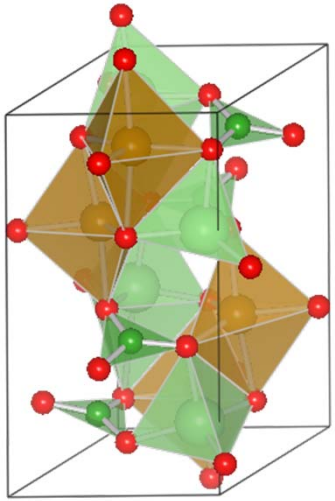
Elsa Olivetti, MIT Gerbrand Ceder, UC Berkeley
Departments of Materials Science & Engineering

Andrew McCallum, UMass Amherst
Department of Computer Science & Engineering

Challenges for technology development: Timeline for development is long



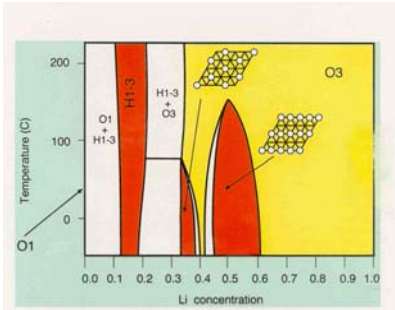
Modern data-driven and first-principles materials design accelerates pace of what to make…



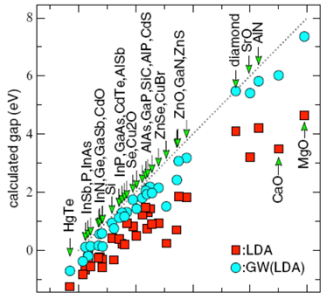
+

+ $$H = \sum_{i=1}^{N_e} \nabla_i^2 + \sum_{i=1}^{N_e} V_{nuclear}(r_i) + \frac{1}{2}\sum_{i}^{N_e}\sum_{j \neq i}^{N_e} \frac{1}{|\boldsymbol{r}_j - \boldsymbol{r}_i|}$$

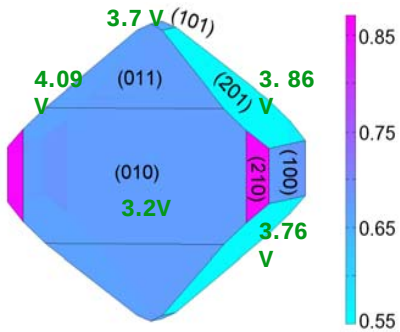
Phase diagrams

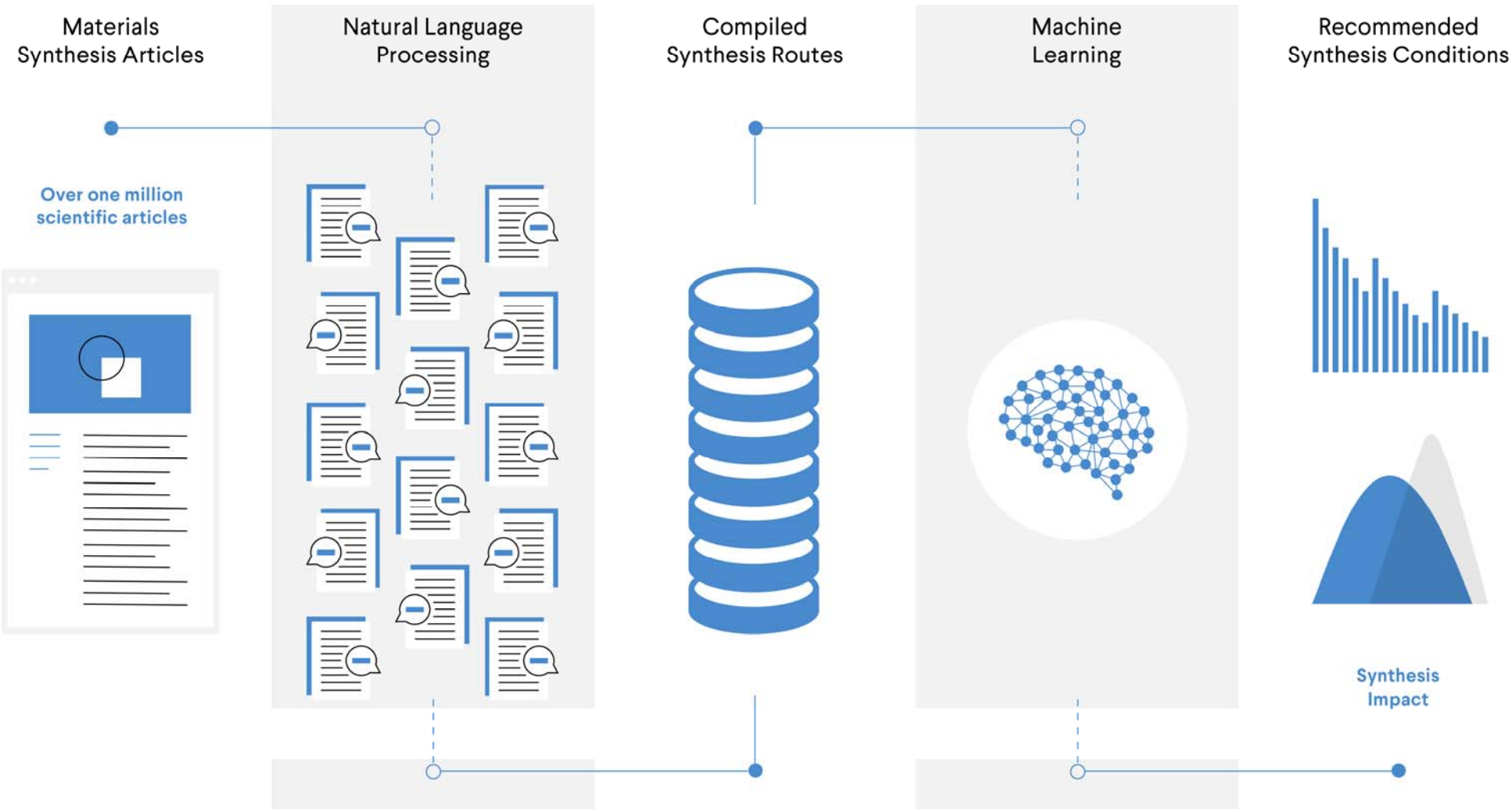


Bandgaps

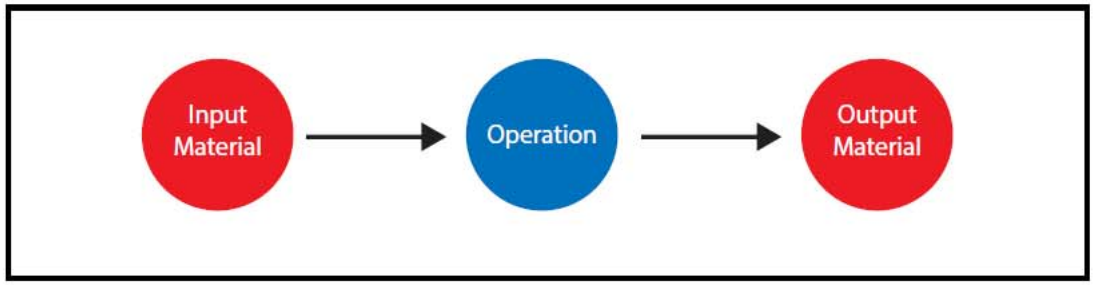


Surfaces

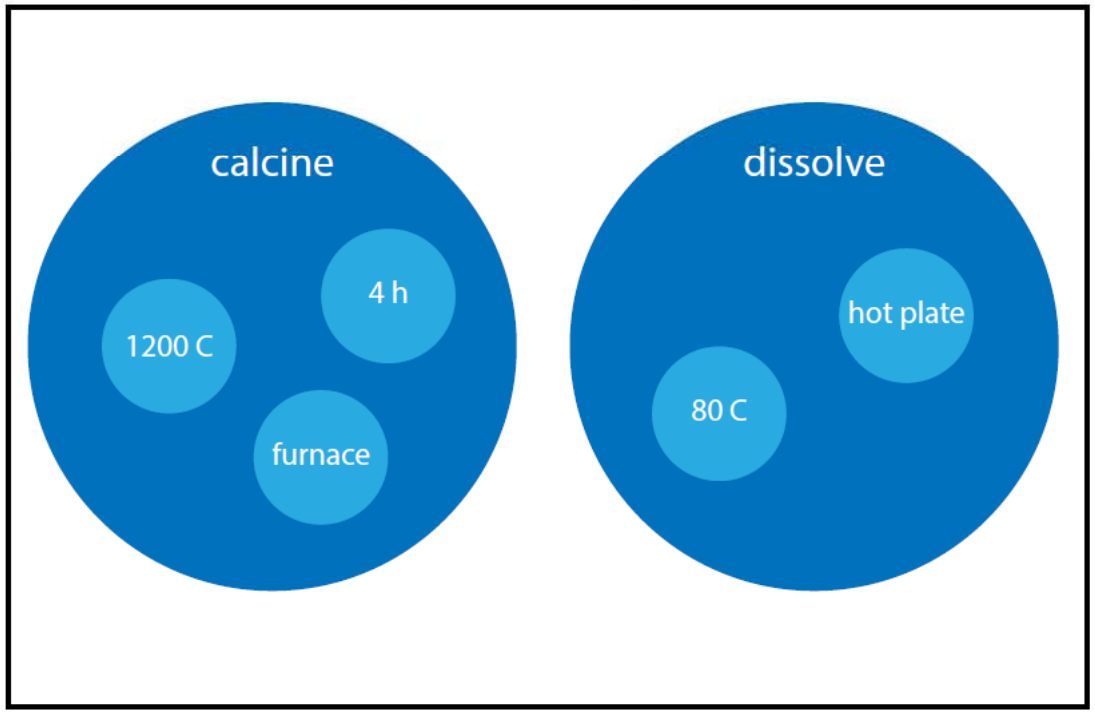



Materials
Synthesis Articles
Over one million
scientific articles
Natural Language
Processing
Compiled
Synthesis Routes
Machine
Learning
Recommended
Synthesis Conditions
Synthesis
Impact

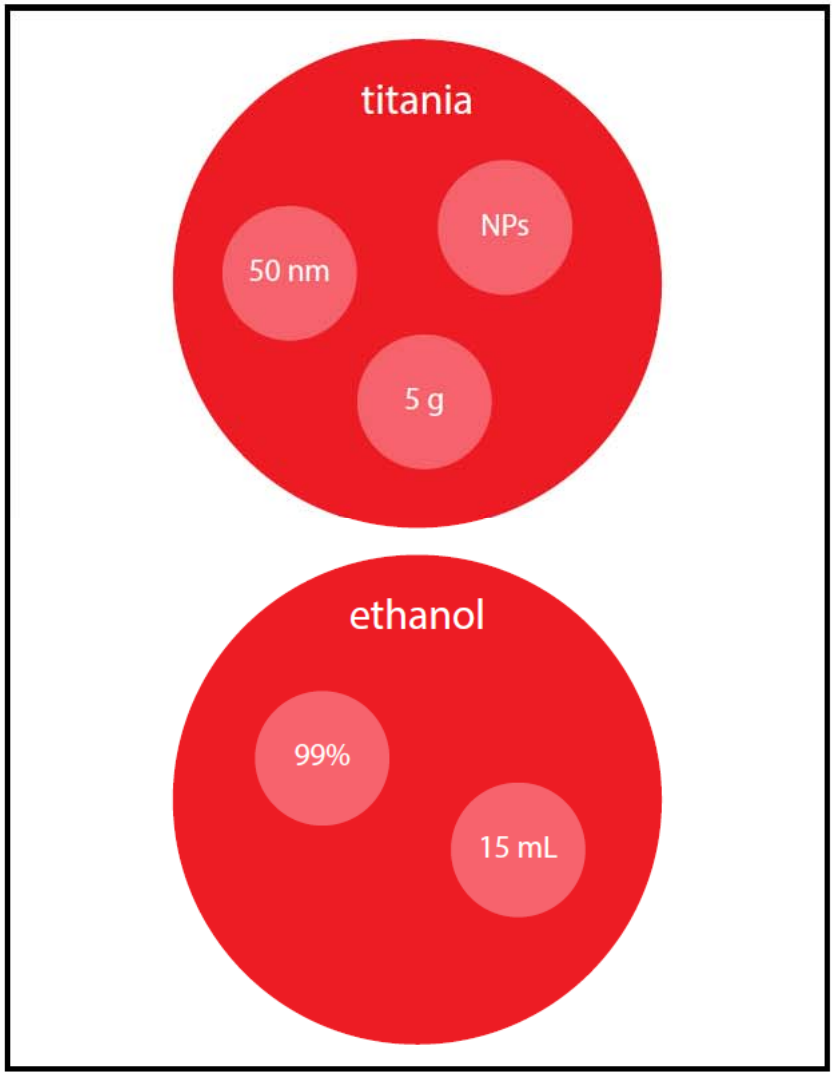
Extracted Synthesis Recipe Schema



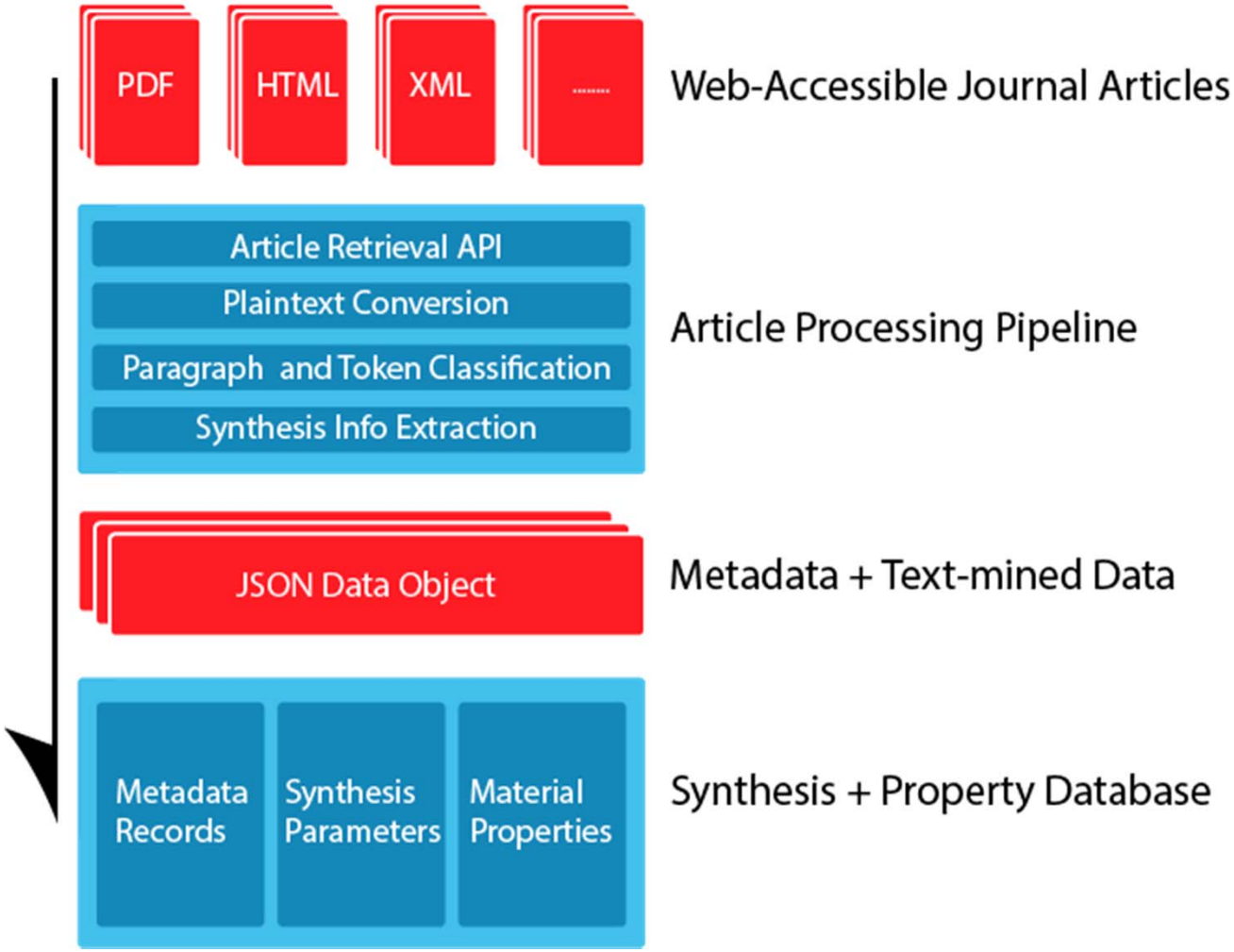
Example Extracted Operations



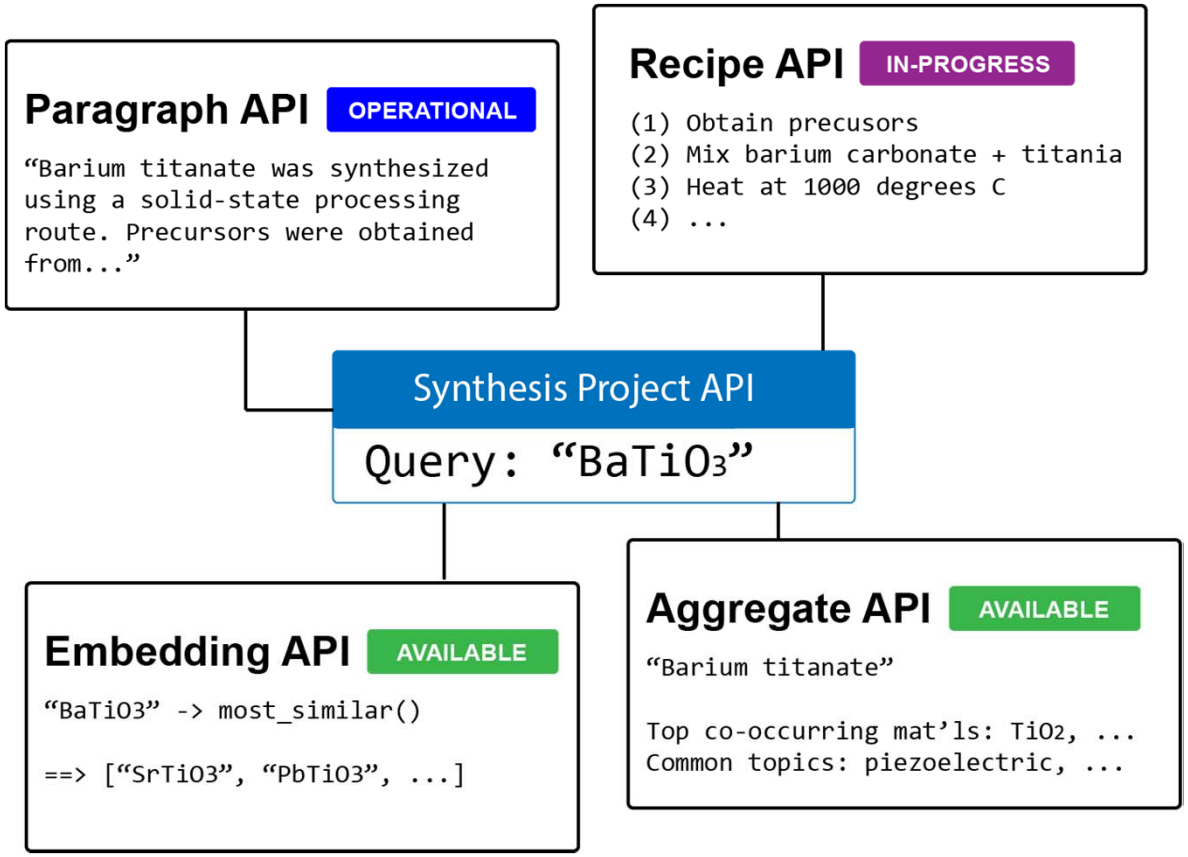
Example Extracted Materials

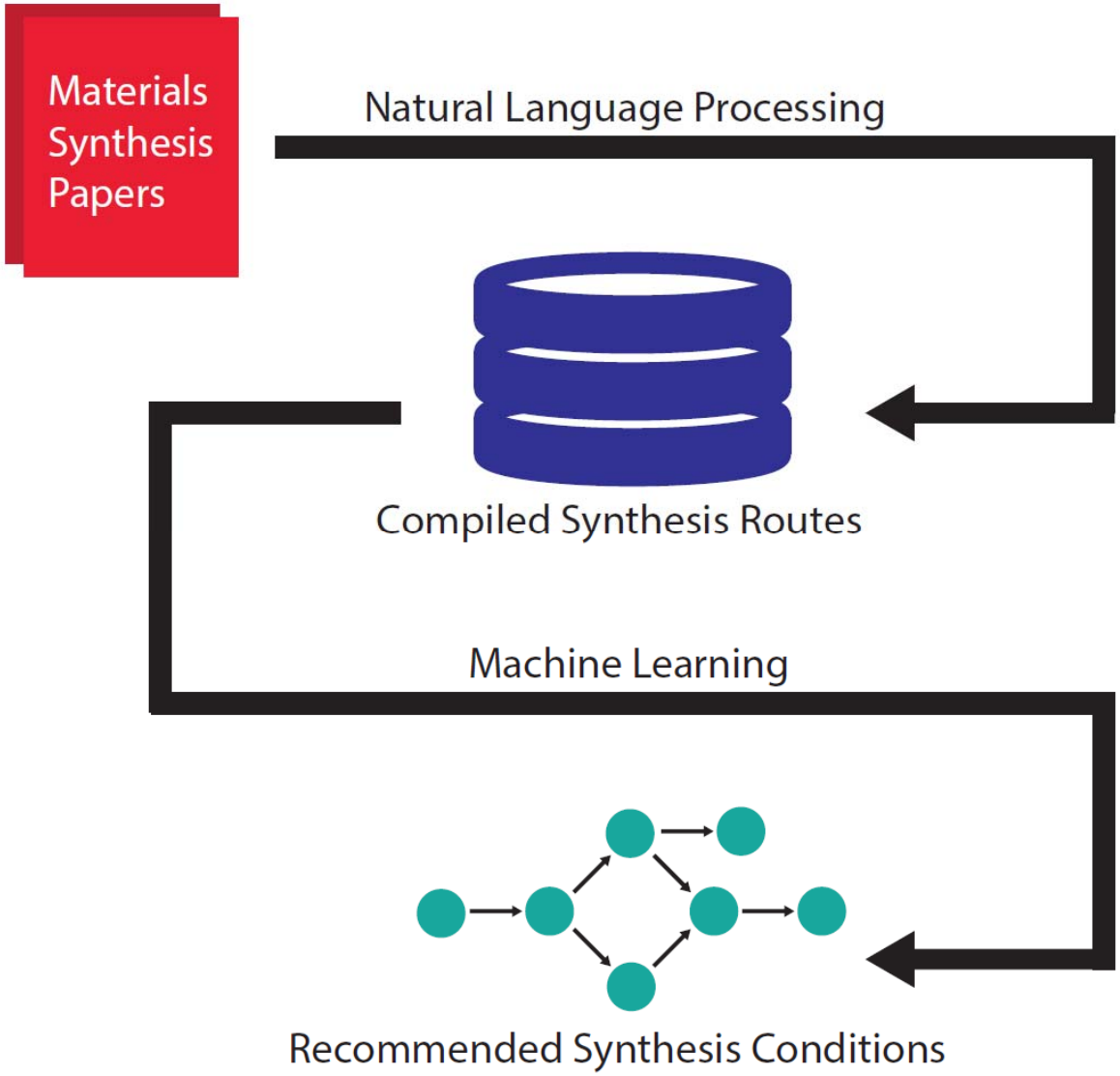


Text extraction workflow

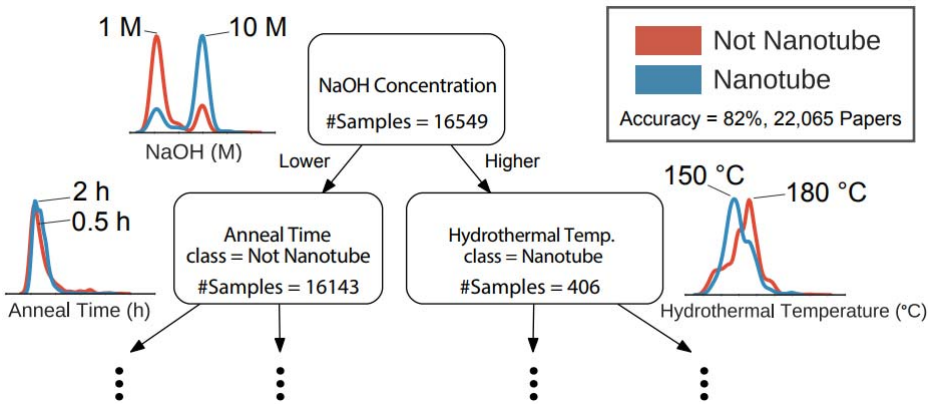


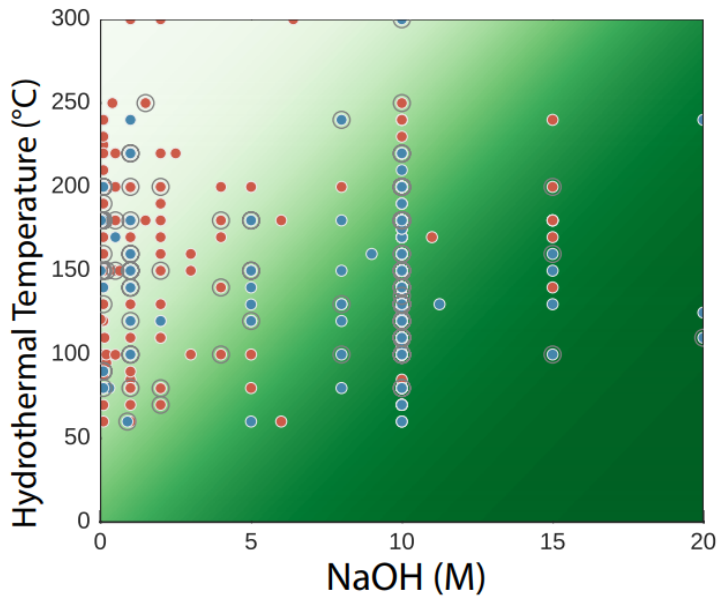
Status of data dissemination

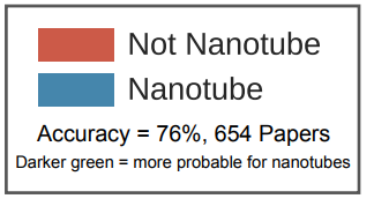



Materials Synthesis Papers
Natural Language Processing
Compiled Synthesis Routes
Machine Learning
Recommended Synthesis Conditions

Example: suggesting synthesis conditions for a specific morphology in titania







*Grey circled points = accuracy test points
All other points = training data points

Experimentally-accessible (and reported) variables to facilitate practical synthesis route planning.

Edward Kim et al., *Chemistry of Materials 2017*

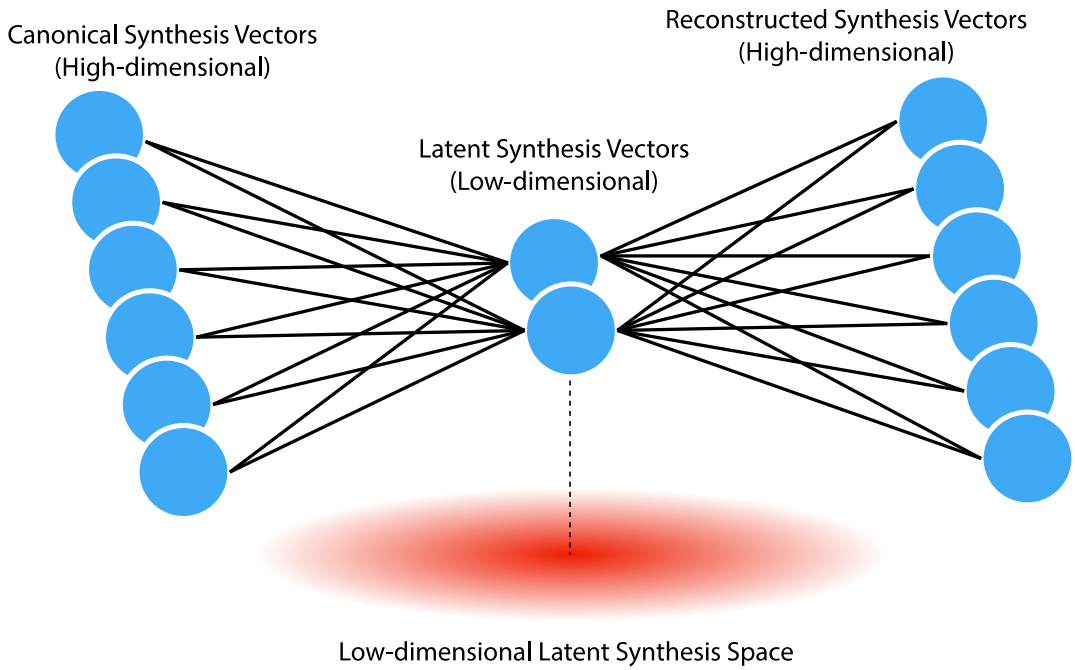
Virtual *synthesis* screening is hard: data is sparse & scarce

"Sparse" = high-dimensional vector of synthesis actions

"Scarce" = materials of interest → not many papers published to train on

Can deep learning / generative models be useful for synthesis screening?

Schematic of Variational Autoencoder



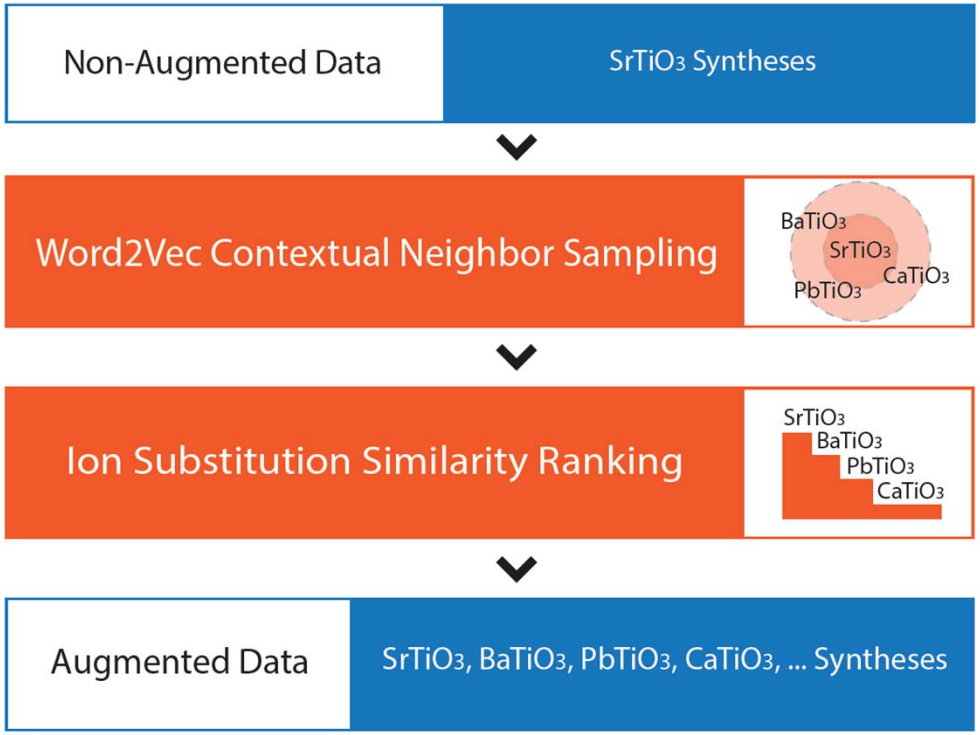
Variational autoencoder:

- Loss = reconstruction + f(Gaussian)
- Also a generative model

Collaborator, Stefanie Jegelka, CSAIL, MIT

Edward Kim et al., *npj Computational Materials 2017*

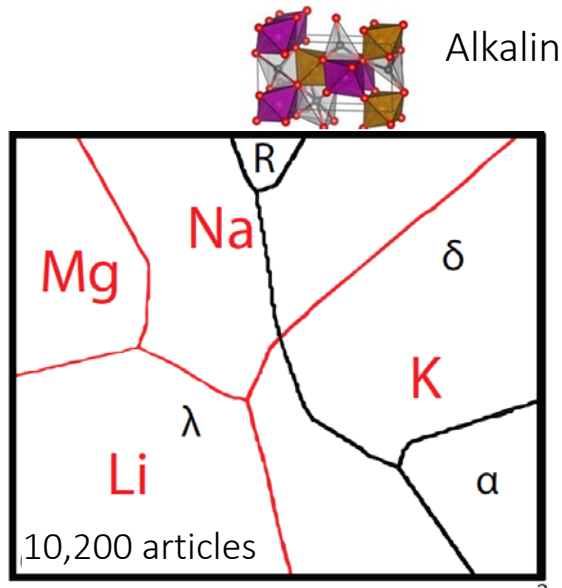
Data augmentation with text & data mining



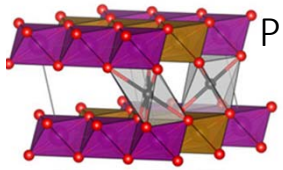
Edward Kim et al., *npj Computational Materials 2017*

Example: suggesting synthesis conditions for stabilizing desired materials

Polymorphs for MnO_2 overlaid with most probable alkali-ion use in synthesis (intercalation-based phase stability)

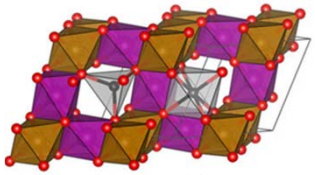


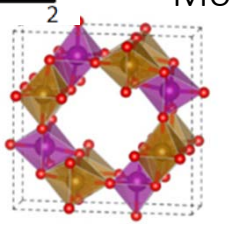
Alkaline batteries



Photocatalysts

Lithium-ion batteries

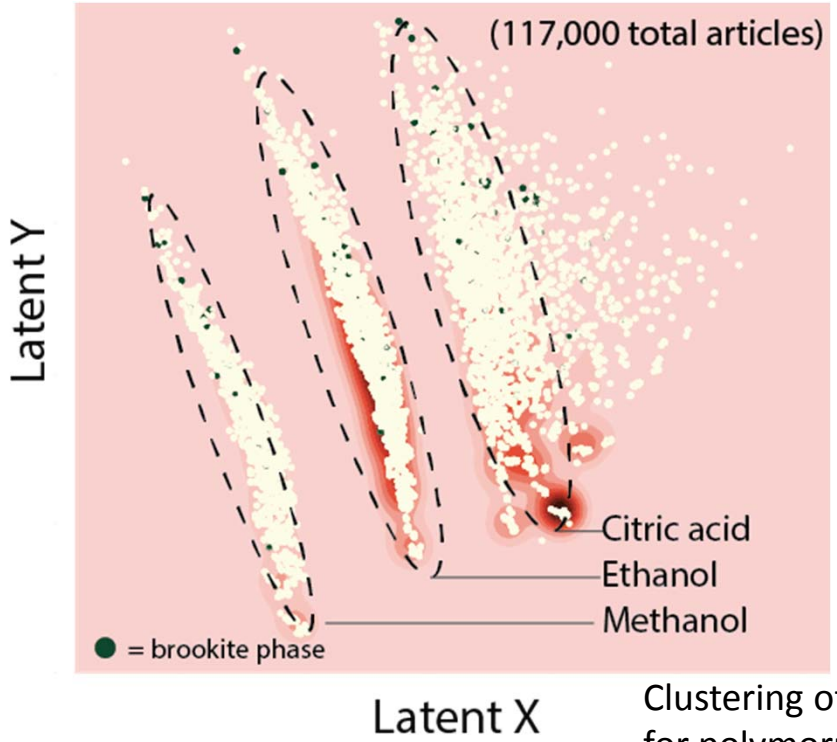


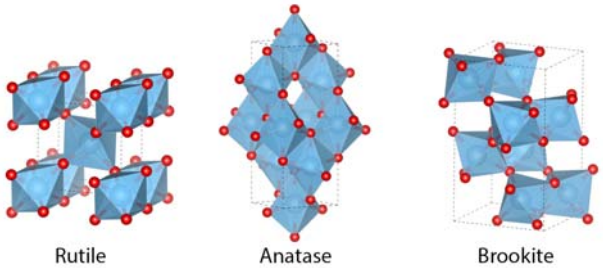


Molecular sieves

Edward Kim et al., *npj Computational Materials 2017*

Exploratory: Rare phase in common material





Clustering of latent space shows driving conditions for polymorph of TiO_2 for photocatalysis

Edward Kim et al., *npj Computational Materials 2017*

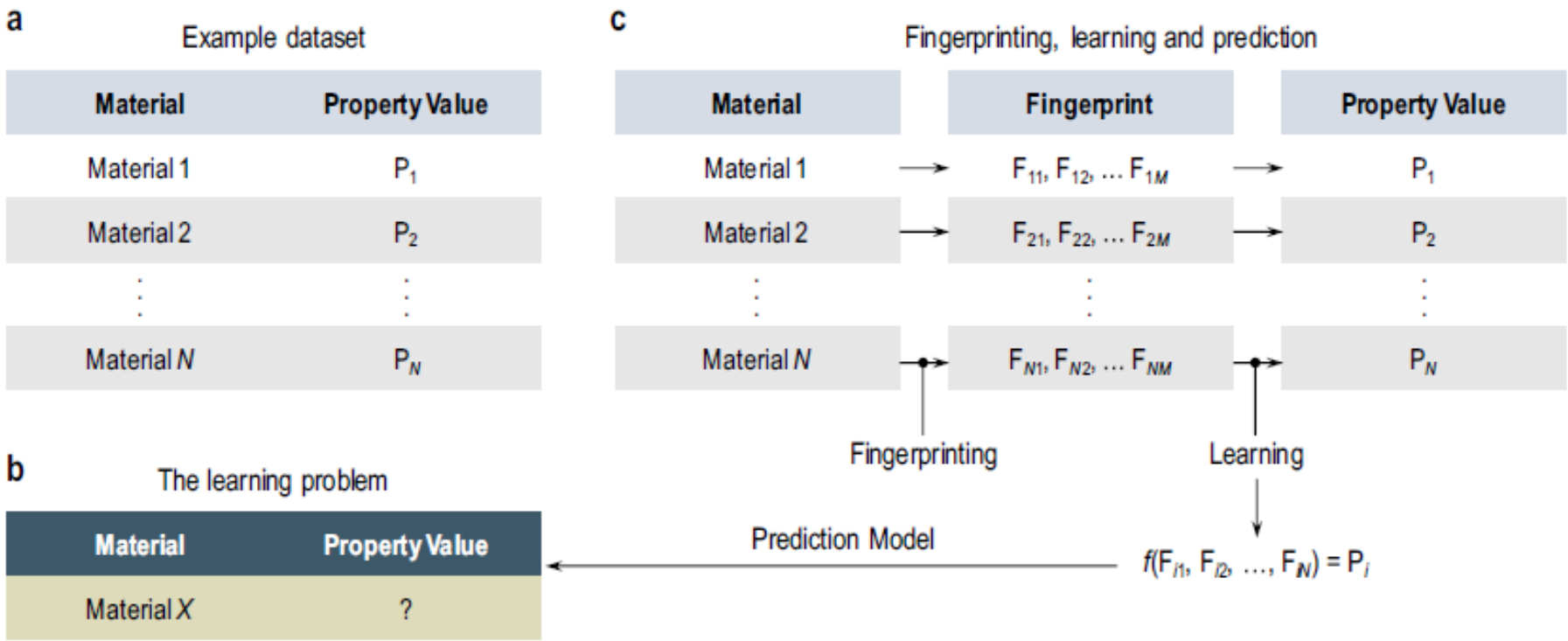
Comparison of literature / virtual samples for $SrTiO_3$ synthesis

Calcination	Sintering	Annealing	NaOH (M)	Reference
800C, 2h	-	-	1	Ye et al, 2016
800C, 2h	1250C, 2h	-	-	Zhao et al, 2004
1000C, 12h	-	500C, 2h	-	Zhao et al, 2015
600-750C, 4h	-	-	-	Puangpetch et al, 2008
721C, 1.8h	-	468C, 0.4h	-	N/A
-	-	450C, 0.9h	1	N/A
955C, 6h	1182C, 7.5h	-	-	N/A

One cannot train a model exclusively on literature data and classify something as successful or not, since there are no negative examples in the literature

Edward Kim et al., *npj Computational Materials 2017*

Key elements in machine learning for energy technology



Ramprasad et al npj computational materials, 2017

Applicability and Next steps

- Continue to improve pipeline and disseminate information to the community
- Inform structure for data going forward
- Use cases in:
  - Solid state synthesis, hydrothermal and sol gel methods
  - Alloy design
  - Electrolyte performance

Thank you
olivetti.mit.edu
synthesisproject.org