Automated Extraction and Database Creation for Ferroelectric Material Discovery

Ian Moog[1], Michael Wang[2], Brandon Schoener[1] (Advisor), Yuting Hu[3], Dr. Hao Zeng[1], Dr. Jinjun Xiong[3,4]

[1] Department of Physics, [2] Williamsville East HS, [3] CSE Department, [4] IAD

University at Buffalo
Institute for Artificial Intelligence and Data Science

Introduction

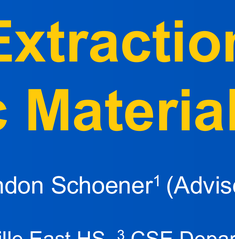
Ferroelectric materials are crystals that maintain an electric polarization even without an external field and can reverse polarity when a field is applied. Their coupled piezoelectric and pyroelectric properties enable applications in non-volatile memory, sensors, actuators, energy harvesting, tunable optics, or even neuromorphic computing, where their intrinsic non-volatility and polarization switching behavior can mimic synaptic activity for brain-like learning.

What Causes Ferroelectricity?

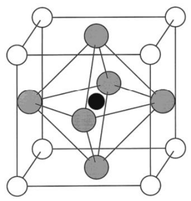
For ferroelectricity to arise in a material, it must have a non-centrosymmetric arrangement of ions and their corresponding electrons.

Many ferroelectric structures can be understood as derivatives of a *non-polar, centrosymmetric prototype phase*, the most widely studied being the perovskite structure.

Long-range Coulomb forces that favor polar distortion, compete with short-range repulsive forces, which favor symmetry. At lower temperatures, Coulomb forces tend to dominate, they can cause distortion in the center cation, pushing it off center.
This is why ferroelectricity will only be present below a certain Curie temperature.



ZnO shown here, has a non-centrosymmetric wurtzite structure, and has a spontaneous electric polarization, but does not switch in response to an applied electric field (not ferroelectric)



Cubic Perovskite structure
The small B cation (in black) is at the center of an octahedron of oxygen anions (in gray). The large A cations (white) occupy the unit cell corners

Methods

- Query OpenAlex + Unpaywall for URLs containing PDF links to open access papers.
- Direct download + Selenium download attempts.
- Conversion of PDFs to .md files via MinerU.

LLM Guided Property Extraction

- Utilize a long context model (Gemini-Flash-2.5) to analyze the .md files
- Return a structured response containing extracted properties

Example PDF Abstract + MinerU conversion

ABSTRACT

This study combines first-principles Density Functional Theory (DFT) calculations and experimental analysis to investigate the impact of Sr-doping on the structural, electronic, and piezoelectric properties of $Ba_{1-x}Sr_xTiO_3$(BST) perovskites at different Sr concentrations (x = 0, 0.125, 0.25 theoretically; x = 0, 0.1, 0.125, 0.15,

```
# ABSTRACT

This study combines first- principles Density Functional Theory (DFT) calculati
electronic, and piezoelectric properties of  $\mathrm{Ba_1}$ $\mathrm{xSr_xTiO_
theoretically;  $\mathbf{x} = 0$  0.1, 0.125, 0.15, 0.2, 0.25 experimentally).
```

Snippet from a .md file which can be passed to an LLM for extraction

Why use an ML Model?

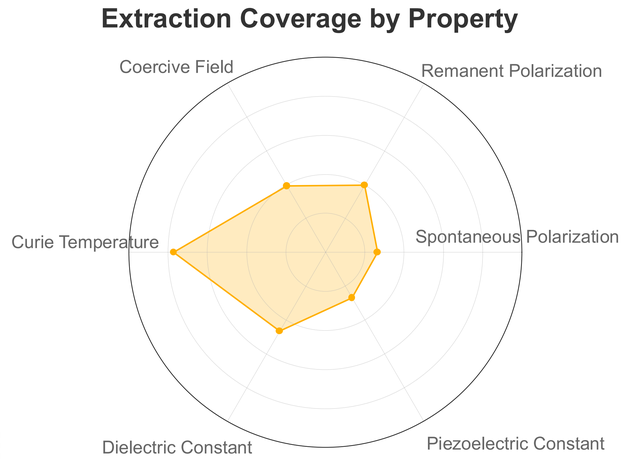
The space of possible crystal structures is essentially infinite, and first-principles calculations like DFT are too slow to explore it fully. A machine learning model allows property prediction of these candidate structures.

This enables a potential workflow:

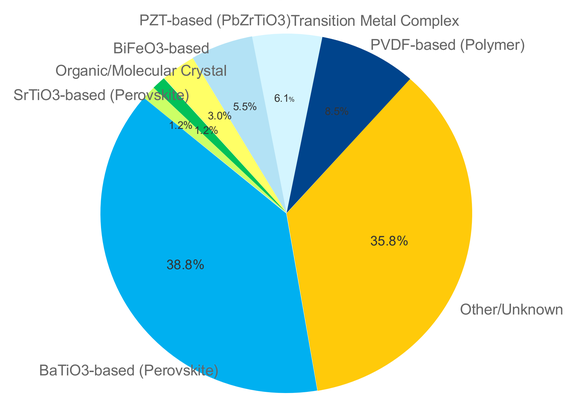
1. **Generate** new candidate structures (e.g., with a crystal GAN).
2. **Screen** them rapidly with the ML model for desirable properties.
3. **Validate** only the most promising ones through detailed simulations or experiments

Extraction Results

- From testing on 100 papers, we identified 165 unique ferroelectric materials
- Curie temperature was the most reported (77%)



This figure shows the extraction success rate by property, i.e., how often each property was extracted.



This figure shows the distribution of material types that we extracted. We can see that BaTiO3 based Perovskites were the most recorded type of ferroelectric, which is expected. This is beneficial because the physics is well understood, and future discoveries will likely be of this type.

Prompting Techniques

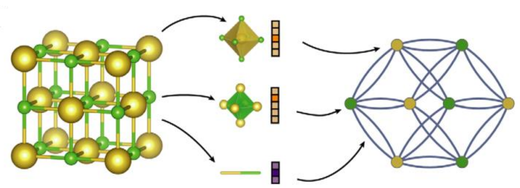
Due to the long context window (1 million tokens) of Gemini-Flash-2.5 we can safely pass our files. Prompting is straightforward, the API allows a schema to be defined, so it will only output one csv line. We simply describe the properties, their units, probable locations, etc. And the model can figure out the rest. Here is an example output:

```
BaTiO3,,14 μC/cm²,2.2 kV/cm,130 °C,,220 pC/N,DFT-based-and-experimental-study-on-Sr-doped-BaTiO
Ba0.9Sr0.1TiO3,,16 μC/cm²,2.4 kV/cm,100 °C,,180 pC/N,DFT-based-and-experimental-study-on-Sr-dop
Ba0.875Sr0.125TiO3,,17.5 μC/cm²,2.1 kV/cm,90 °C,,170 pC/N,DFT-based-and-experimental-study-on-S
```

As you can see, we successfully extracted properties for three materials from one paper, with 4/6 properties for each. We can flag extractions with low confidence, and send them to a higher cost model later, which will perform better.

Property Prediction

To predict properties, we can use a Crystal Graph Convolutional Neural Network (CGCNN). A graph is a data type made of nodes connected by edges. Crystal structures can be represented this way, where atoms are the nodes, and the distance between them can be represented by the edges.
So, our CGCNN can take a crystal structure as an input and give property values as output.



https://doi.org/10.1103/PhysRevLett.120.145301

Pipeline for Material Discovery

- Match crystal structures (CIF files) to our extracted properties
- Use this as input to train a CGCNN to predict properties from structure alone
- Once enough structure-property relationships are learned we can generate and feed it entirely new crystal structures.
- Use our model to rank these structures, and find potential candidates to test experimentally

References

1. T. Xie & J.C. Grossman, Phys. Rev. Lett. 120, 145301 (2018)
2. K. Rabe, Ch. H. Ahn & J.-M. Triscone, Springer (2007).
3. A. Jain et al., APL Mater. 1, 011002 (2013).
4. B. Schoener, UB, Comprehensive Exam

University at Buffalo