CRUCIBLE: The AI Materials Classification Agent

Proposal by Nickolas Joyner

Metal oxides (MO) have wide applications in key chemical processes in energy production, catalysis, and infrastructure. These fields are growing rapidly, thus the examination of the thermochemical and electronic properties of known metal oxides is paramount in discovery of new materials to fulfill the increasing global demand. Investigators have often turned towards theoretical calculations such correlated molecular orbital theory and density functional theory (DFT) to characterize the underlying chemistry and thermodynamics of these systems. However, the innate electronic complexity of select metal oxide species, arising from strong electronic correlation,[1,2] poses a massive challenge in accurately modeling the thermodynamics of these systems. Recent research has deployed machine learning to address these challenges in materials discovery, as new ML algorithms excel in building relationships from complex data to predict the properties of materials.[3,4,5,6,7]

While recent research has successfully established high-accuracy property prediction for inorganic systems this proposal addresses the distinct challenge of material identification. I hypothesize that while individual thermodynamic values may be non-unique across species, the intersection of spectroscopic signatures and thermochemical stability creates a distinct, high-dimensional fingerprint for metal oxides. Machine learning algorithms can be trained to recognize these multi-modal patterns, allowing CRUCIBLE to accurately classify unknown species *in situ*. If valid, this protocol would allow researchers to characterize feedstock compositions in real-time, providing a distinct advantage in validating whether synthetic pathways are effectively producing target materials.

Chemical Foundations: ML in Metal Oxide Thermochemistry

The concept behind CRUCIBLE was developed during my doctoral research entitled: "Machine Learning for Accelerated Metal Oxide Thermochemical Predictions. "Within this study,[8] a ML algorithm was developed to predict the normalized clustering energies (NCE) of DFT-optimized M(II)Os containing 4 alkaline earth and selected predominantly 3d transition metals (M = Mg, Ca, Sr, Ba, Sc, Ti, V, Cr, Mn, Co, Cu, Zn, Cd). The NCE is a critical thermochemical property, describing the overall energy of the nanocluster relative to one base unit. From the NCE one can calculate cohesive energy, and the heat of formation of the material. Several machine learning algorithms were tested, including a kernel ridge regression (KRR), and a feed-forward artificial neural

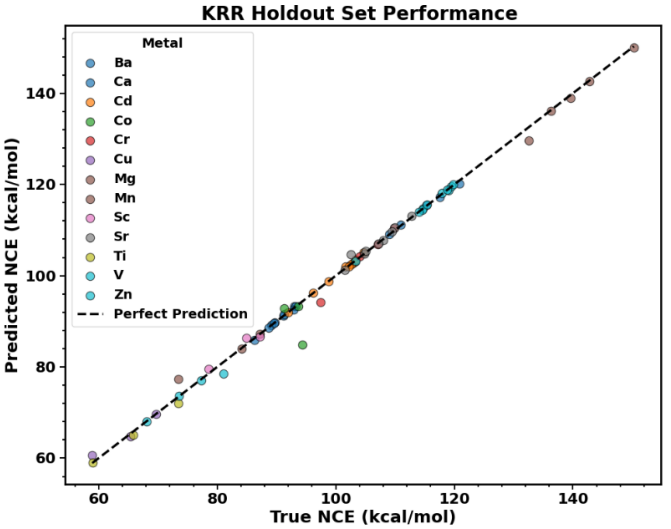

Figure 1. Comparison of KRR-predicted and DFT-calculated NCEs for M(II)O high-spin clusters. Each point represents a cluster; colors indicate the metal identity. R2 = 0.994, MAE = 0.619 kcal/mol.

network (ANN). On a held-out test set, KRR achieves the lowest error shown by Figure 1 (MAE = 0.619 kcal/mol, R^2 = 0.994), followed by ANN (MAE = 1.250 kcal/mol, R^2 = 0.986). By fitting the NCEs, bulk cohesive energies for closed-shell oxides (MgO, CaO, SrO, BaO, ZnO, CdO) are predicted to within ~1.5 kcal/mol of the corresponding DFT values. These results demonstrate that ML trained on modest cluster data can deliver accurate predictions of NCEs and bulk thermochemistry at a fraction of the cost of electronic-structure calculations. However, utilizing thermodynamic values alone is often insufficient for unique identification due to property overlap between species. Therefore, CRUCIBLE extends this foundational work by integrating these validated thermochemical signatures with spectroscopic data in order to provide a distinct "fingerprint" for the classification of complex materials.

While preliminary demonstrations validate the architecture, the primary obstacle to deploying this at scale is the acquisition of formatted, multimodal reference training data. National Laboratories are in a prime position to mitigate this challenge by leveraging decades of verified materials research, supplemented by online repositories such as the NIST Chemical Webbook. Crucially, the initial training dataset will act as a bridge between experiment and theory: it will correlate observable spectroscopic profiles (e.g., Raman shifts, IR frequencies) with fundamental thermochemical properties (e.g., NCE, heat of formation). Initially uniform metal oxides can be analyzed, followed by mixed metal oxides. This slow introduction of increasing compositionally complex materials allows for the characterization of uncertainty metrics within the model.

Model Architecture and AI Infrastructure

The analytical core of CRUCIBLE relies on a suite of robust machine learning classification algorithms implemented via Scikit-Learn and TensorFlow. To ensure comprehensive model benchmarking across varying data structures, the system employs a diverse ensemble of classifiers, including Random Forest, Support Vector Machines (SVM), Neural Networks, and K-Nearest Neighbors (KNN). The model input combines vibrational peak positions with thermodynamic stability data to distinguish between distinct material phases. These models are trained on formatted thermochemical and spectroscopic datasets and validated using a k-fold cross-validation technique (k = 5). This rigorous validation protocol allows for the precise characterization of uncertainty metrics, ensuring the model avoids overfitting and maintains predictive accuracy when analyzing novel metal oxide compositions.

To ensure access to these computational tools, user interaction is managed by a locally hosted Large Language Model (LLM), specifically Llama-3.1-8B-Instruct (Quantized), operating within a Retrieval-Augmented Generation (RAG) framework built on LlamaIndex. This local deployment strategy ensures strict compliance with national laboratory security requirements, as no sensitive feedstock data is transmitted to external cloud repositories. Crucially, the LLM serves not as a source of scientific fact but as a semantic orchestrator; it utilizes Python-based "FunctionTools" to query the database or execute classification scripts, ensuring that all material identifications are derived from the validated ML algorithms rather than generative text.

Figure 2 visualizes this integrated workflow, mapping the information lifecycle from raw data acquisition to final user insight. The diagram depicts the initial formatting of experimental and theoretical thermochemical data into the unified database, which feeds the continuous training of the classification algorithms. It further illustrates the inference pathway, where the LLM acts as the central interface, accepting natural language queries from the user, retrieving context from the vector store, and triggering the classification engine to output a validated material identity. This architectural flow demonstrates how CRUCIBLE bridges the gap between complex computational chemistry and accessible, interactive analysis.

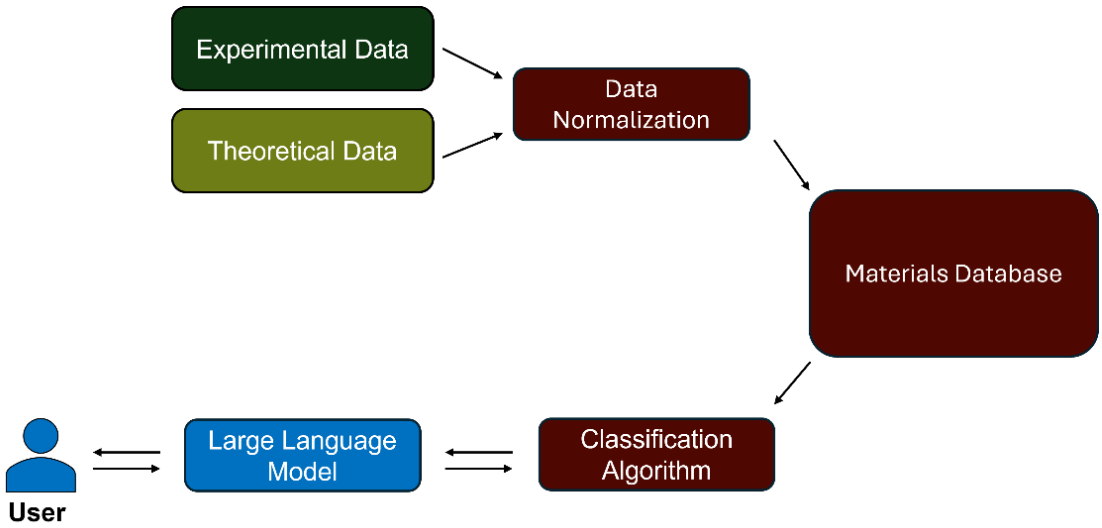

Figure 2: CRUCIBLE model architecture and workflow.

In conclusion, CRUCIBLE represents a pivotal advancement in materials characterization by rigorously validating that complex metal oxides can be accurately identified *in situ* solely through their thermodynamic and spectroscopic signatures. By coupling this high-fidelity machine learning backend with a secure, locally hosted Large Language Model, the platform transforms specialized computational analysis into an accessible, interactive partnership. Ultimately, this framework accelerates the validation of critical synthetic pathways, ensuring that data-driven discovery is both scientifically robust and intuitively available to the broader research community.

A demonstration of CRUCIBLE and example ML python scripts are located on GitHub (https://github.com/NAJoyner/Nickolas_Joyner_Crucible). This demonstration uses a random forest classifier trained on simple electronic and spectroscopic data of common materials. Furthermore, all example machine learning scripts are also included. This is an example model and is for demonstration purposes only.

---

[1]Romeu, J. G. F.; Joyner, N. A.; Dixon, D. A.; The Electronic Structure and Properties of First Row Transition Metal Oxides. *J. Chem. Theo. Comp.* Submitted Sept. 2025. In peer review

[2] Joyner, N. A.; Romeu, J. G. F.; Kent, B.; Dixon, D. A. The Electronic Structure of Diatomic Nickel Oxide. *PhysChemChemPhys*. **2024**, *26*, 19646-19657.

[3] Lee, J.; Seko, A.; Shitara, K.; Nakayama, K.; Tanka, I. Prediction Model of Band Gap for Inorganic Compounds by Combination of Density Functional Theory Calculations and Machine Learning Techniques. *Phys. Rev. B.* **2016**, *93,* 115104.

[4] Gao, Z.: Zhang, H.; Mao, G.; Ren, J.; Chen, Z.; Wu, C.; Gates, I.; Yang, W.; Ding, X.; Yao, J. Screening for Lead-Free Inorganic Double Perovskites with Suitable Band Gaps and High Stability Using Combined Machine Learning and DFT Calculation. *Ap. Surf. Sci.* **2021***, 568*, 150916.

[5] Thomas, J. C.; Bechtel, J. S.; Natarajan, A. R.; Van Der Ven, A. Machine Learning The Density Functional Theory Potential Energy Surface for the Inorganic Halide Perovskite $CsPbBr_3$. *Phys. Rev. B.* **2019,** *100,* 134101.

[6] Zhuo, Y.; Tehrni, A. M.; Brgoch, J. Predicting the Band Gaps of Inorganic Solids by Machine Learning. *J. Phys. Chem. Lett*. **2018**, *9,* 1668-1673.

[7] Janet, J. P.; Liu, F.; Nandy, A.; Duan, C.; Yang, T.; Lin, S.; Kulik, H. Designing in the Face of Uncertainty: Exploiting Electronic Structure and Machine Learning Models for Discovery in Inorganic Chemistry. *Inorg. Chem.*, **2019**, *58,* 10592-10606.

[8] Joyner, N. A.; Sprouse, S.; Herndon, H.; Butler, K. E.; Kaye, E.; Makoś, M. Z.; Dixon, D. A. Machine Learning for Accelerated Metal Oxide Thermochemical Predictions. *J. Chem. Theo. Comp.* Submitted Oct. 2025