Machine Learning Magnetism

Team: Ahmed Fahmy, Murod Mirzhalilov, Brandon Abrego, Sayok Chakravarty

GitHub: https://github.com/AhmedFahmy177/MachineLearningMagneticOrderClassification/tree/main

Project Overview: As the demand for scalable, accurate, and cost-effective materials discovery grows, data-driven approaches have become essential to accelerating scientific breakthroughs. Magnetism plays a fundamental role in a wide range of advanced technologies, including spintronic devices, magnetic memory storage, and emerging quantum information systems. However, traditional methods for identifying magnetic properties—such as neutron scattering experiments or density functional theory (DFT) calculations—are either resource-intensive or computationally demanding, making them unsuitable for large-scale screening.

This project aims to develop a machine learning framework capable of predicting the magnetic ordering of materials using structural, chemical, electrical, and thermodynamic coarse-grained descriptors that are already available for a wide range of inorganic materials, with two-fold strategy. The first is to build a ML model that can efficiently predict the type magnetic ordering based on training on the *Materials Project* database, the primary database for inorganic compounds, to facilitate future large-scale magnetic classifications instead of performing time-consuming DFT calculations. However, it is well known that the magnetic labels on this database suffer from bias towards the Ferromagnetic (FM) class, an issue which roots in the used DFT methods that fail with the absence of 'semi-'empirical inputs. That being said, it's still very efficient in identifying magnetic vs. nonmagnetic materials. To partially address this issue, we use another database, *MAGNDATA,* the most comprehensive database of experimental neutron scattering-based magnetic structures. By training another ML classifier on this database, we can efficiently predict the magnetic propagation vector of a given magnetic material, a quantity that carries some information about the underlying magnetic structure. Finally, this is adapted to the *Materials Project* database to correct some of the wrongly labelled FMs by showing that the *MAGNDATA* classifier predicts them to have nonzero propagation vector, contradicting the DFT-assigned labels.

Stakeholders: Materials scientists, Spintronics and quantum technology developers, Experimental condensed matter physicists, Computational materials researchers.

KPIs: - Accuracy: Improvement in predicted classification of magnetic orders compared to a baseline classifier (i.e., dummy classifier with a stratified strategy) and compared to a recent research paper (Helena A. Merker, et al., "Machine learning magnetism classifiers from atomic coordinates", iScience 2022).

- Throughput: Efficient reduction of computational time compared to time-consuming DFT calculations.

- Bias correction in *Materials Project*: How many FM compounds on the *Materials Project* can we predict though the *MAGNDATA* classifier to be non-FMs.

Approach: The first part focused on supervised classification models to predict the magnetic ordering of materials on the *Materials Project* which contains ~150,000 materials with magnetic labels, and the second part focused on building an efficient propagation vector classifier on *MAGNDATA,* containing ~ 2,100 commensurate magnetic materials, and then applied to compounds on *Materials Project*.

- **Feature Engineering:** The features of our data set include numerical and categorical data spanning structural, compositional, electronic, and thermodynamic descriptors that are physically relevant to magnetism. Based on random forest feature importance, PCS, and features correlation matrix, we find that all our numerical data are needed to explain a majority of the variance of our data. Much of our categorical data is of high cardinality relative to the size of our dataset. Some of these features were removed since they served as unique identifiers and only biased our model. Other features such as chemical composition were broken down into its composition and one-hot encoded based on each part of the composition rather than the unique combination of parts, which we found to boost the accuracy of our classifiers. We removed rows which did not contain our final key features, resulting in just over 100,000 compounds. For the case of materials on *MAGNDATA*, we used the same set of descriptors after importing them from *the Materials Project*.
- **Training & Validation & Testing:** 60% of the data on the Materials Project were used to train different classifiers: Random Forest, XGBoost, LightGBM, Support Vector Machine, Logistic Regression, k-Nearest Neighbors, Decision Tree and Naive Bayes. Performance of the classifiers was assessed using both the accuracy and the macro average F_1 score. The former is chosen to check the overall classifier performance over all classes, and the latter is chosen to check the performance over the minority classes. Four-fold cross validation scores, along with performance over a separate 20% validation set, were used to compare the performance of the different classifiers. For all classifiers, hyperparameter tuning was performed using GridSearchCV.

  For the MAGNDATA classifier, 67.5% of the data was used to train two classifiers: Random Forest and XGBoost, with a 6-fold cross validation strategy. Performance was assessed using a separate 22.5% validation set. For both databases, test sets were left as a final sanity check to the best classifier performance.

- **Merging different subsets**: we also check the cases where we merge different classes together, a strategy commonly used in literature to address e.g. the issue of the underrepresented Ferrimagnetic (FiM) class by merging it with the FM class since in both cases, the material has a nonzero net magnetization. Additionally, in order to compare our classifiers with those of the recent research study (mentioned above), we consider the subset of materials which have at least one element that belongs to transitions metals, lanthanides, or actinides. This subset is further modified by merging the FM and the FiM classes together, a strategy used in the aforementioned research article.

Results:

- **XGBoost classifier with 10 numerical features and 2 categorical features** achieved **an impressive accuracy of 89%** for magnetic class prediction, representing a huge improvement over the 45% accurate baseline dummy classifier with stratified strategy.
- **LightGBM achieves the highest macro average F_1 score of 67% and 86% accuracy (third highest).**
- SMOTE was used to slightly enhance the performance over the minority classes for XGBoost.

- In all models, feature importance analysis was performed. Additionally, SHAP was applied to XGBoost (for the case of all data) to study the impact of each descriptor over the model's performance.
- Across the merging of different subsets of data, both XGBoost, LightGBM, and Random Forest stay as the best three classifiers, however changing their rankings in some cases.
- In the case of materials with at least one magnetic element, **our LightGBM classifier outperforms the mentioned research study over both the NM and FM/FiM classes**, *as detailed in the powerpoint presentation*.
- Magnetic class prediction within a few seconds compared to the hours/days DFT calculation can take.
- **94% accurate Random Forest classifier for propagation vector on *MAGNDATA*.**
- Applying our *MAGNDATA* propagation vector classifier on the Materials Project database, we **Identified 12,609 materials on Materials Project that initially DFT-labelled as FMs to have nonzero propagation vector, suggesting that these materials are probably either AFMs or FiMs,** *therefore* ***partially correcting the FM bias in the Materials Project*.**

<u>Future Directions:</u>

- Including more detailed features to predict details of noncollinear magnetic structures.
- Implementation of Neural Networks to better capture the complexity of the connection between the structure and electronic features to the emergence of magnetic orders.