# Detection of Topological Materials using Machine Learning

Chaithanya Krishna Moorthy

Indian Institute of Technology
Madras

11-05-2021

# Outline



# Introduction

- Topological Quantum Chemistry (TQC) applied on Density Functional Theory (DFT) calculations has compiled a large catalogue of Topological Matrials.
- Machine Learning methods can offer faster topological classification of these materials and offer hints to identify which quantities decide the Topology of a material.

## Objective

To build an ML model to predict the DFT-computed topology of a given material with an accuracy of about 90%

# Dataset

- 70020 impurity free, crystalline materials were selected from the Inorganic Crystal Structure Database (ICSD) on which first-principle calculations as implemented in Vienna Ab-Initio Simulation Package - DFT and TQC were carried out.
- Topological Categories into which the above materials were grouped:
  - Trivial Topology (49.5% of total materials)
  - Topological Insulators (TI): NLC (6.5%) and SEBR (7%) [1]
  - Topological Semi-metals (TSM): ESM(10%) and ESFD(27%) [2]

[1] NLC: Not a Linear Combination of Electronic Band Representaions; SEBR: Split EBR

[1] ESM: Enforced Semi-metals; ESFD: Enforced Semi-metals with Fermi Degeneracy

# Dataset

- Each material is described by its stoichiometric formula, a space group (SG), a unit cell, and the positions of the atoms therein.
- Two materials exhibiting the same stoichiometry and SG also have the same topological classification and hence were put into the same equivalence classes, and a representative was taken
- Thus, the size of the dataset was reduced to 35009 effectively.
- The complexity of the materials ranged as:
  Number of atoms per primitive unit cell - 1 to 60
  215/230 different Space Groups
  92 different chemical elements

# Frequency of TI's

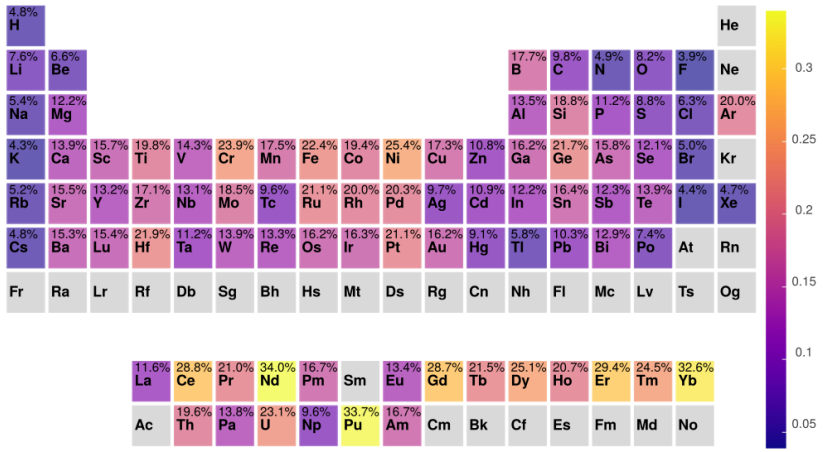


Figure 1: Conditional probability that a material is a Topological insulator, given that it contains the paricular element from the periodic table

- Elements forming Ionic compounds rarely form TI
- Alternating pattern in rare-earth elements due to even/odd f-electrons affecting the formation of fully filled f-electron bands

# Quality of Model

For binary classification:

- Accuracy: Fraction of data points classified correctly. Not a good in cases where the number of data points for each class is imbalanced. Compared against baseline accuracy (49.5% in this case).
- Precision: Fraction of positive predictions that are correct
- Recall: Fraction of positive class classified correctly.
- F1-score: Balance between precision and recall - harmonic mean of precision and probability.

Precision and Recall depend on the threshold chosen for probabilities. In this paper, F1-score for each class was chosen as a measure of the quality of the model as it is a single-valued score for each class and is fairly independent of the threshold.

# ML model - Gradient Boosted Trees (GBT)

- GBT uses an ensemble of weak learners (decision trees) to build a strong model (classifier here).
- Weak trees - with small number of nodes - are added iteratively to correct misclassifications of the previous trees.
- The performance of the model saturated at about 150 trees.
- The algorithm is trained to minimize the loss function.
- Hyperparameters - Maximal Tree Depth, Learning rate, Column subsampling by tree, Minimal Child Weight, $L^2$ regularization $\lambda$ and Column subsampling by node - are chosen by searching for the set that provides maximum efficiency on the cross-validation (k fold stratified Cross-validation method) data set.
- About 10% of the dataset was used as test set. Rest was used for training.
- Other ML models tested which gave lower efficiency are : random forests, k-nearest neighbour classifiers, SVM

# GBT

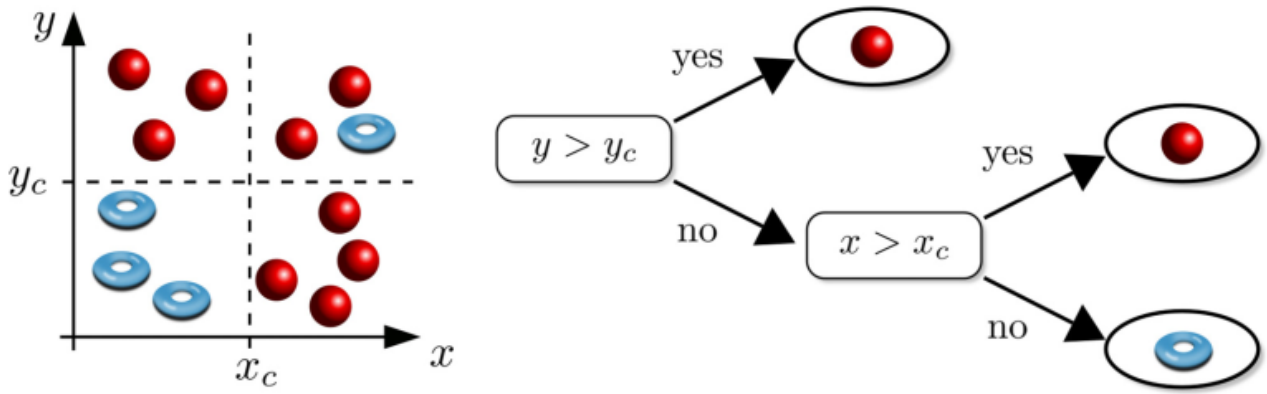


Figure 2: Example of GBT

Each node (containing a binary question) of a decision tree divides the space into to halves. Questions which lead to the maximum separation are chosen.

# Representation of Crystals for ML

- Single-atom statistics are used to map crystals to real vectors, usable in the ML model
- For each constituent atom chemical and physical single atom properties are concatenated in a single vector $\mathbf{X}_i$. For representing the entire crystal, the mean and variance of the vectors were carried out for the unit cell.
- The relevant single-atom statistics are mean number of s, p, d and f shell valence electrons and mean and standard deviations of the atom's row and column number in the periodic table.
- In addition to single-atom statistics, global properties of Space Group(SG) and number of electron per unit cell ($N_e$) were chosen.
- These quantities were deemed relevant after extensive testing and elimination from a list of quantities.
- Nearest-neighbour features such as positions of atoms in the unit cell with respect to each other were discovered to be irrelevant.

# Analysis of the GBT model

| Model | Descriptor | $d$ | Acc. (%) | $F_1$ Triv. (%) | $F_1$ TI (%) | $F_1$ TSM (%) |
|---|---|---|---|---|---|---|
| Full model (FM) | SG, $N_e$, $spdf+$, number of atoms from each periodic table row, and column | 49 | 89.7(5) | 94.0(3) | 70(1) | 92.0(5) |
| FM + Non-SOC | Features used by FM and topological classification of material obtained by DFT without SOC | 50 | 92.0(3) | 96.5(2) | 77(1) | 93.3(4) |
| Baseline model | SG, $N_e$, and baseline descriptor (number of atoms from each element in the stoichiometric formula) | 94 | 86.0(5) | 92.5(5) | 67(1) | 91.0(5) |
| $spdf+$ model | SG, $N_e$, and $spdf+$ features | 10 | 87.7(5) | 93.0(5) | 69(1) | 92.0(5) |
| FM + nearest neighbor | Features used by FM and nearest-neighbor difference features [42], defined in Sec. II C | 184 | 89.0(5) | 94.0(3) | 69(3) | 92.0(5) |
| FM without SG | $N_e$, $spdf+$, number of atoms from each periodic table row, and column | 48 | 84.0(5) | 91.5(3) | 57(2) | 86(1) |

Figure 3: Performance of GBT models measured by F1-score and accuracy

# Simple Decision Tree

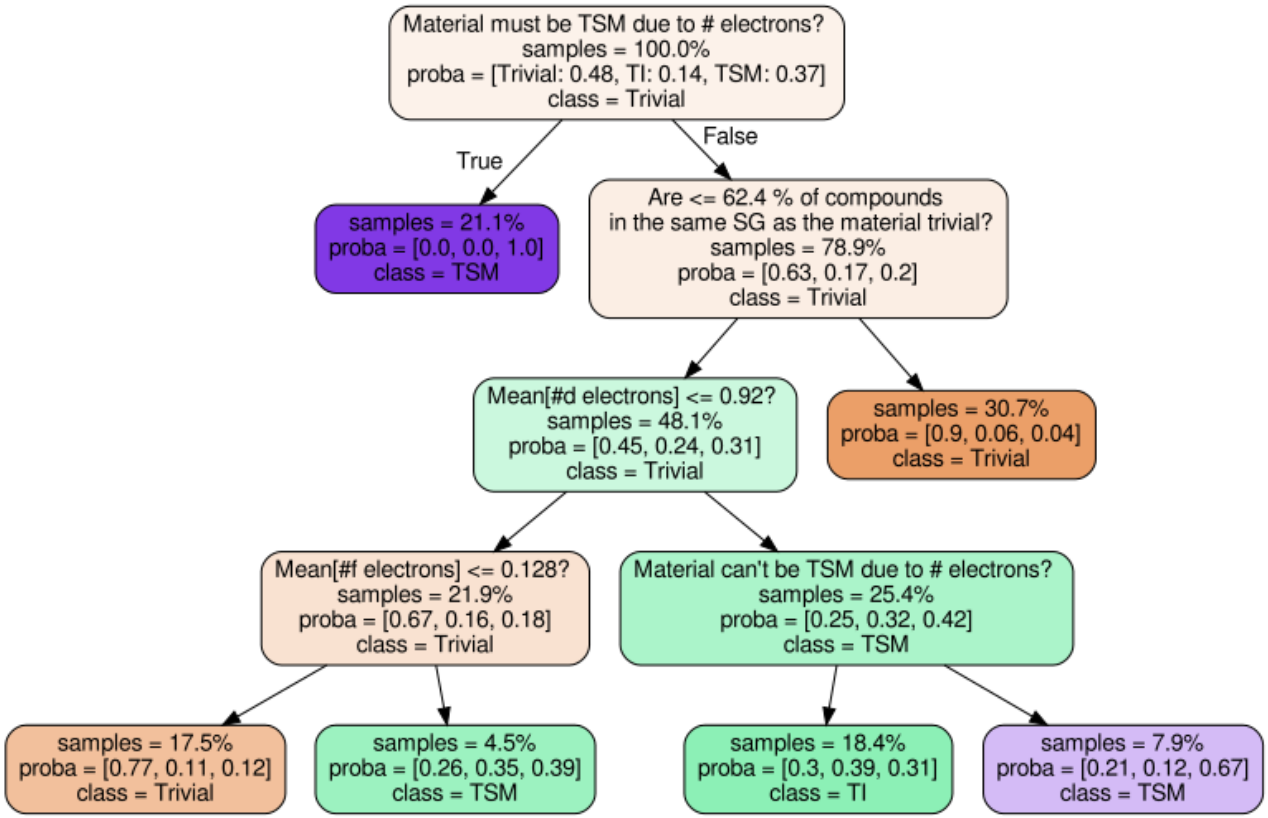


Figure 4: Simplified decision tree using spdf+ model in table 3.

# Observations

- SG's importance is seen by the drop in performance on its removal from the model.
- Inclusion of nearest-neighbour properties decreases the performance.
- TI's have a lower score which can be attributed to the less number of samples available in the dataset and difficulty in inferring the nature just using the features used in the model.
- A large number of d or f-electrons help turn a compound into a TI.
- TSM's are mostly determined by $N_e$ and SG.

# Limitations

1. Size of available data

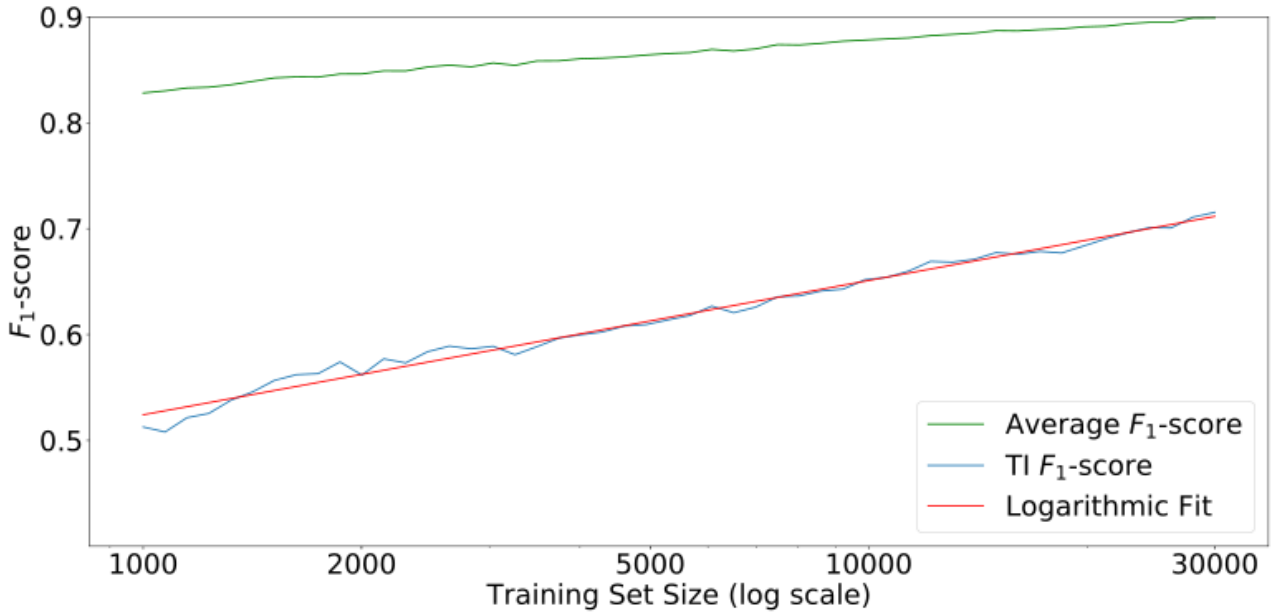


Figure 5: Variation of the average F1-score with the size of dataset used for training - increases logarithmically

The materials have a large variety of SG's and chemical elements - number of samples which are similar to a given material is very less, which makes it harder to infer patterns.

# Limitations

2. DFT calculations used to determine topological labels can be unreliable
   This is checked by building a model that uses *high-confidence materials* - with reliable DFT calculations.

| Total Accuracy (%) | F1 triv. | F1 NLC | F1 SEBR | F1 ES | F1 ESFD |
|---|---|---|---|---|---|
| 87.0 | 94.0 | 66 | 59 | 73 | 95.5 |

Table 1: Performance in classifying topological subclasses using the full data set for training

| Total Accuracy (%) | F1 triv. | F1 NLC | F1 SEBR | F1 ES | F1 ESFD |
|---|---|---|---|---|---|
| 92.0 | 96.0 | 72 | 68 | 79 | 99.0 |

Table 2: Performance in classifying topological subclasses using the high-confidence data set for training

# Limitations

3. Specific types of materials on which model performance is poor:
   error rate of the model depends on the symmetry properties of the material - the model performs poorly on cubic point groups and the hexagonal point group $D_{6h}$. These point groups contain the greatest number of distinct symmetry operations and are complicated.
4. Poor performance on materials containing halogens and alkali metals since they form less TI's.
5. Energetics can influence the topology ( of SEBR's for eg), but has not been included in the model.

# Discussions

- Using a simple GBT based algorithm, classification of listed Topological materials could be achieved without expensive calculations.
- The use of the ML model and analyzing the decision trees has determined the properties that are responsible for topological phenomenon - it is mostly determined by course-grained chemical composition and crystal symmetry and does not depend much on positions of atoms in the lattice. Materials with large number of d and f shell electrons or containing heavy metals with strong SOC are likely to be TI's

# Further research directions

- Including global properties like crystal symmetry.
- Using more sophesticated ML models such as crystal graph convolutional networks (CGNN's)
- Preprocess data using physical understanding for better performance of the model.
- Using *empty lattice approximation* which allows estimation of topological features from lattice constants.

# Database and Tools

- Database: B. Wieder, L. Elcoro, B. A. Bernevig, N. Regnault, and M. G. Vergniory, A Complete Catalogue of All Topological Materials (unpublished)
- `https://topologicalquantumchemistry.org/`
- Link to the online tool built : `https://www.topologicalquantumchemistry.com/mltqc/`