W3: KAGGLE-PREDICTING TRANSPARENT CONDUCTORS

Heigo Ers, Taido Purason

https://github.com/taidopurason/nomad2018

INTRODUCTION

Transparent, conductive materials are compounds, which have electrical conductivity and low absorbance in the visible range. In spite of appealing properties, a small number of compounds possess such qualities. Therefore, computational screening is necessary to select compounds for experimental verification to lower the development costs.

Using machine learning methods we have estimated the band gap and formation energy of materials. The models were trained using attributes, which describe the symmetry of crystal combined with qualities that describe crystal's chemical environment.

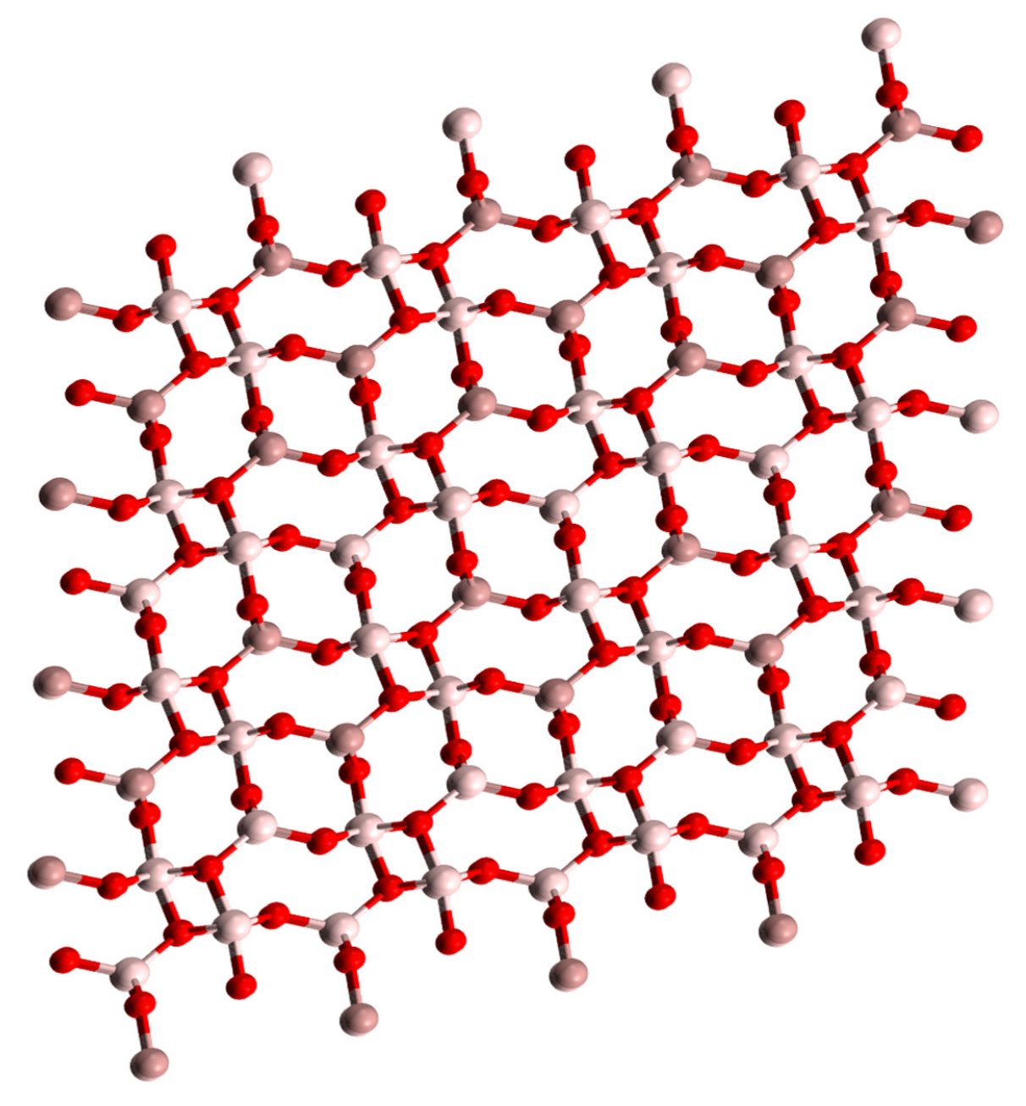


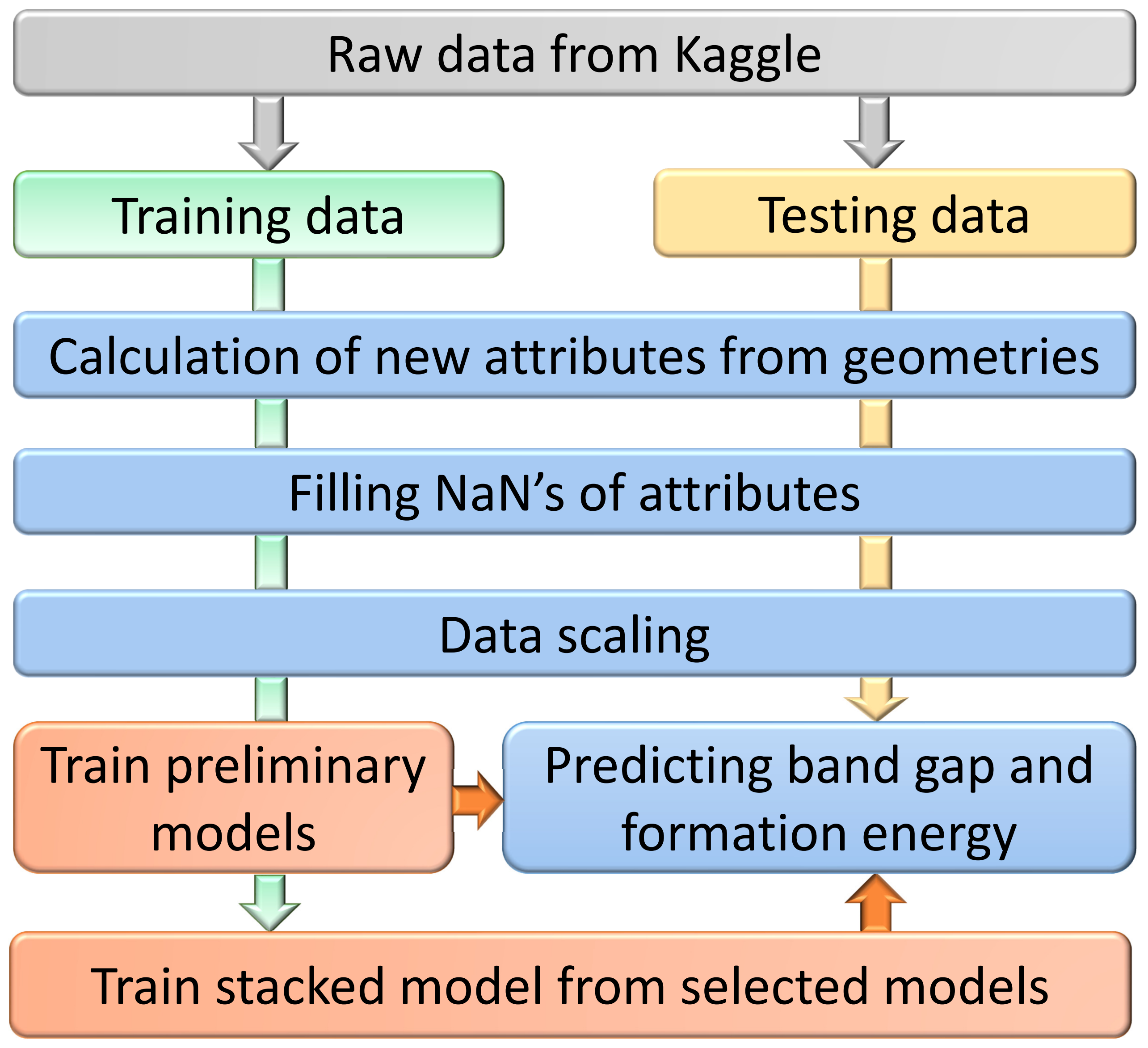
Figure 1. *Crystal's unit cell*

DATA AND METHOD

The initial data from Kaggle contained attributes about symmetry and composition of 3000 materials, containing Al, Ga, In and O atoms.

As attributes from Kaggle weren't comprehensive for the task, we constructed additional attributes for each material:

- Bond lengths of each atom pair
- Electrostatic interaction
- Atoms' partial charges
- Coordination numbers of each atom pair
- Equilibrated electronegativity
- Average volume of atom



Raw data from Kaggle

Training data

Testing data

Calculation of new attributes from geometries

Filling NaN's of attributes

Data scaling

Train preliminary models

Predicting band gap and formation energy

Train stacked model from selected models

Figure 2. *Overview of project's workflow*

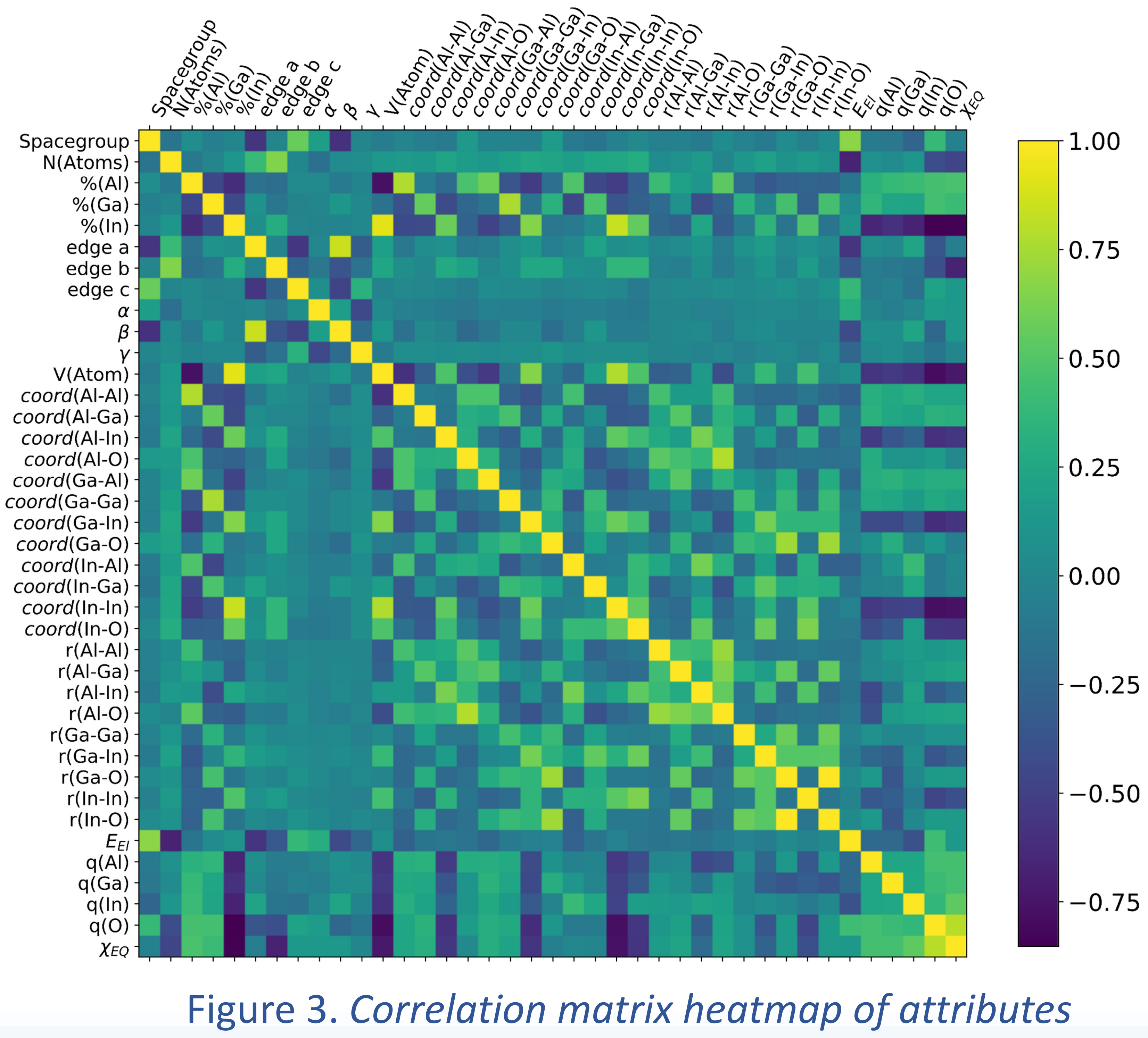


Figure 3. *Correlation matrix heatmap of attributes*

RESULTS

For this project root mean square logarithmic error (RMSLE) score was used : $\sqrt{\frac{1}{n}\sum\left(\ln(x_{real}+1)-\ln\left(x_{pred}+1\right)\right)^2}$

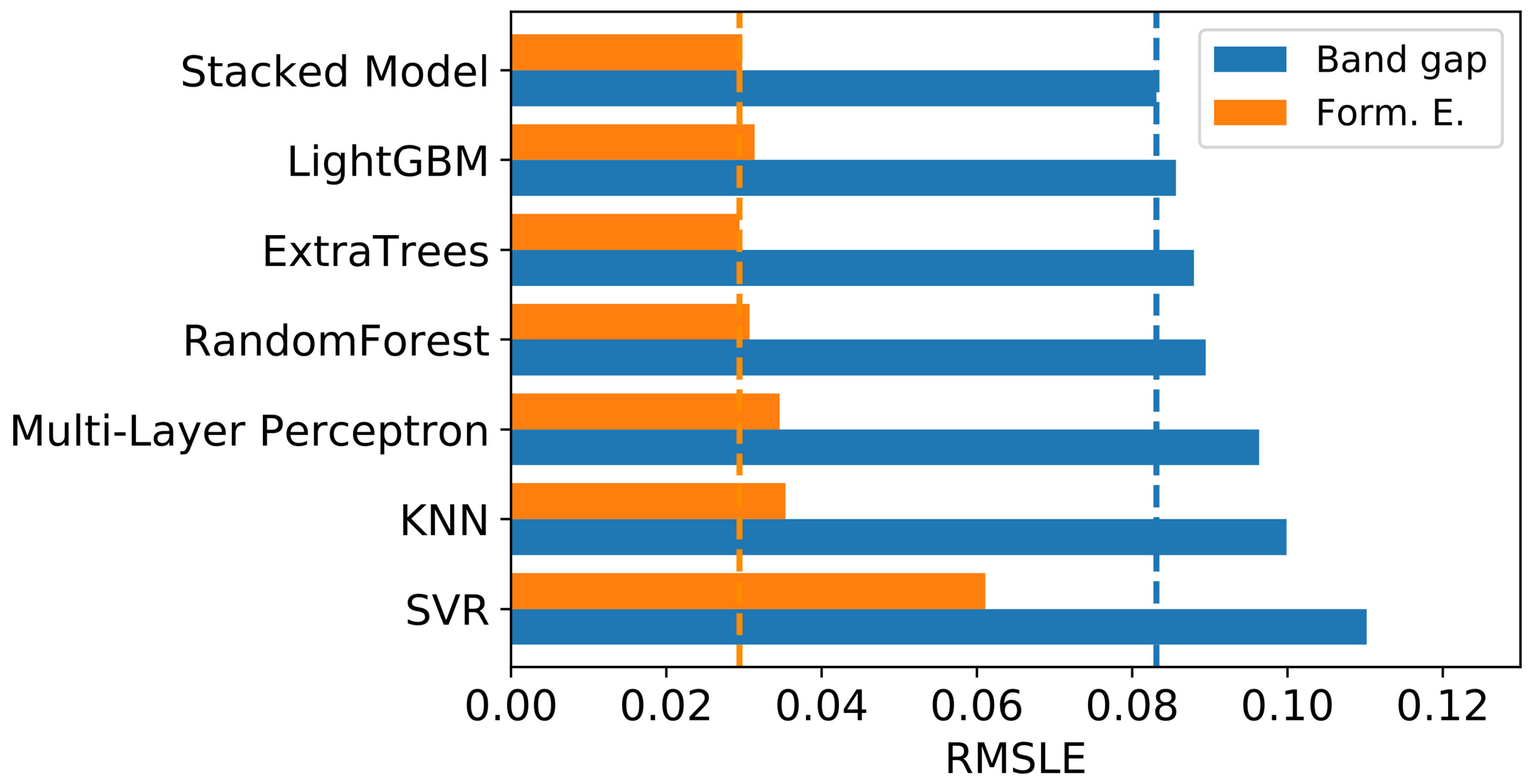


Figure 4. *Best trained models with cross-validated RMSLE score on the training data. For both target values, separate models were developed.*

THE BEST MODELS

- **Band gap energy - Stacked model**

  It's 1st level regressors were SVR, MLP, RF, ExtraTrees and LightGBM. Ridge regression was used as a meta-regressor.

- **Formation energy - ExtraTrees model**

The combination of these two models scored 0.06376 on the private Kaggle dataset (top 7%).

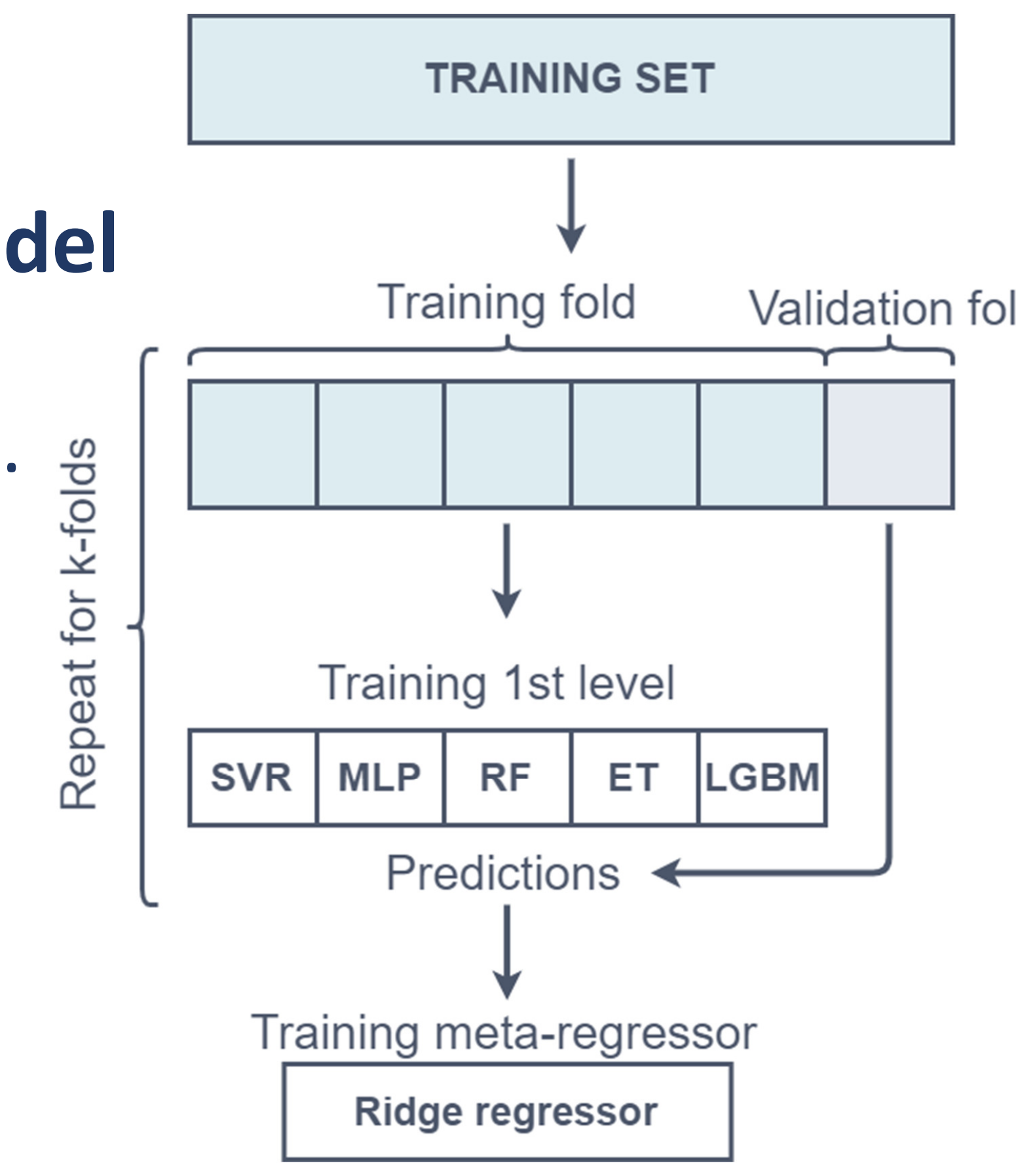


Figure 5. *The stacked model used for BE prediction.*

CONCLUSIONS

- Band gap and formation energy of transparent conductors can be successfully predicted using machine learning
- Computation of additional features was crucial for lowering the error rate and hence obtaining a accurate models.
- Even without elaborate models the error was relatively low.
- The 3 most important features for each target value:
  - Band gap – *V*(atom) (engineered feature), %(In), %(Al);
  - Formation energy – edge c, spacegroup, *r*(Al-In) (engineered feature).
- Further engineering of the models is necessary to obtain a competitive result.