SISSO (SURE-INDEPENDENCE SCREENING AND SPARSIFYING OPERATOR)

OVERVIEW OF SISSO

- Regressor technique which yields predictive surrogate models expressed as analytical formula.
- Used in conjunction with compressed sensing techniques to filter features.(OMP)
- Overcomes limitation of smaller datasets.

MAIN PARAMETERS

- OMP(Orthogonal Matching Pursuit) – Selects features of data set to be used based off their coefficients. Number of features can be specified.

- Operation set – Specifies possible operations to be used. E.g "(+)(*)(^2)(^3)(^-1)(cos)(sin)“.

- Descriptor dimension/rung/complexity – Specifies number of features, composed operations and operations used.

- SISSO then selects model which yields lowest RMSE after specifying parameters.

OMP DEMO

- Applied to Mxenes dataset.
- Input:

```
OMP(descript_dim=3, #number of descriptors to filter out
           data=df_to_ML,
           data_fraction_for_test=0.2,
           split_random_seed=1,
           save_path=SAVE_PATH #set to 0 if don't want to save
    )
```

- Results:

```
  Descriptors     Coefficients
0      charge      373.1032887
1           c   -0.03389305751
2           a     0.6661875896
************************************
*              All done            *
************************************

Process finished with exit code 0
```

- 3 most correlated descriptors are found with their respective coefficients.

SISSO DEMO

- Applied to Mxenes dataset.

- Input:

```
SISSO(OMP_dim=0, #0 if no OMP use, else state dimension to chose via omp
      data=df_to_ML,
      CLEAN_RUN_DIR=False, #False to have SISSO eqn. Need to be false for OMP use too
      optree_depth=2, # rung of the feature space to be constructed
      op_set="(+)(*)(^2)(^3)(^-1)(cos)(sin)", # "(+)(*)(^2)(^3)(^-1)(cos)(sin)"
      descriptor_dim=3, # number of descriptors used, i.e number of columns composed
      maxi_complexity=3, # max feature complexity (number of operators in a feature)
      data_fraction_for_test=0.2,
      split_random_seed=4,
      save_path=SAVE_PATH #set to 0 if don't want to save
      )
```

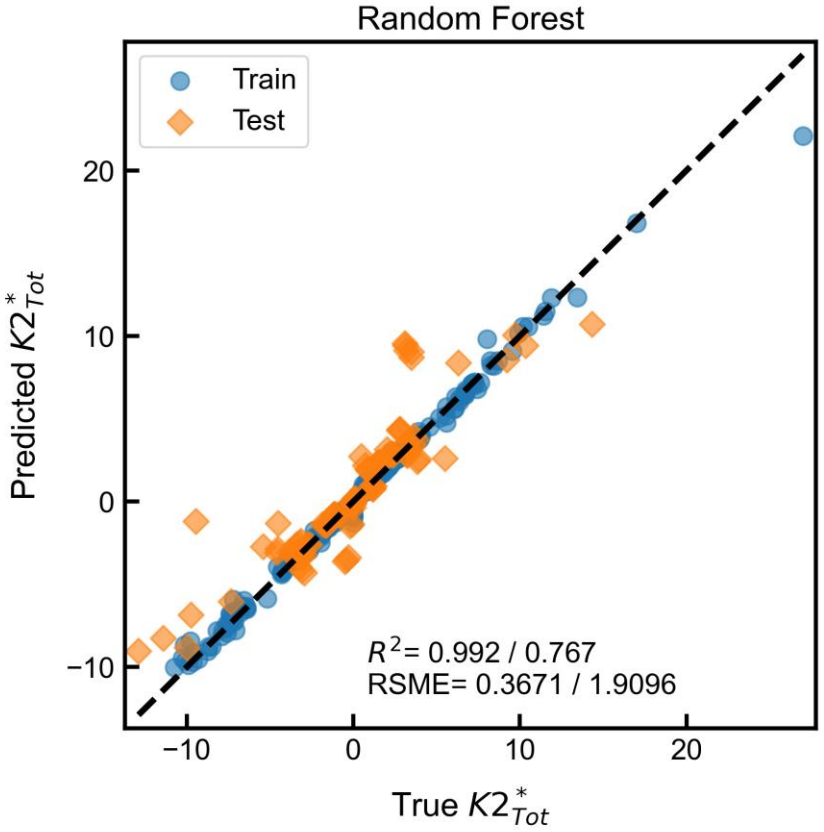
- Results:

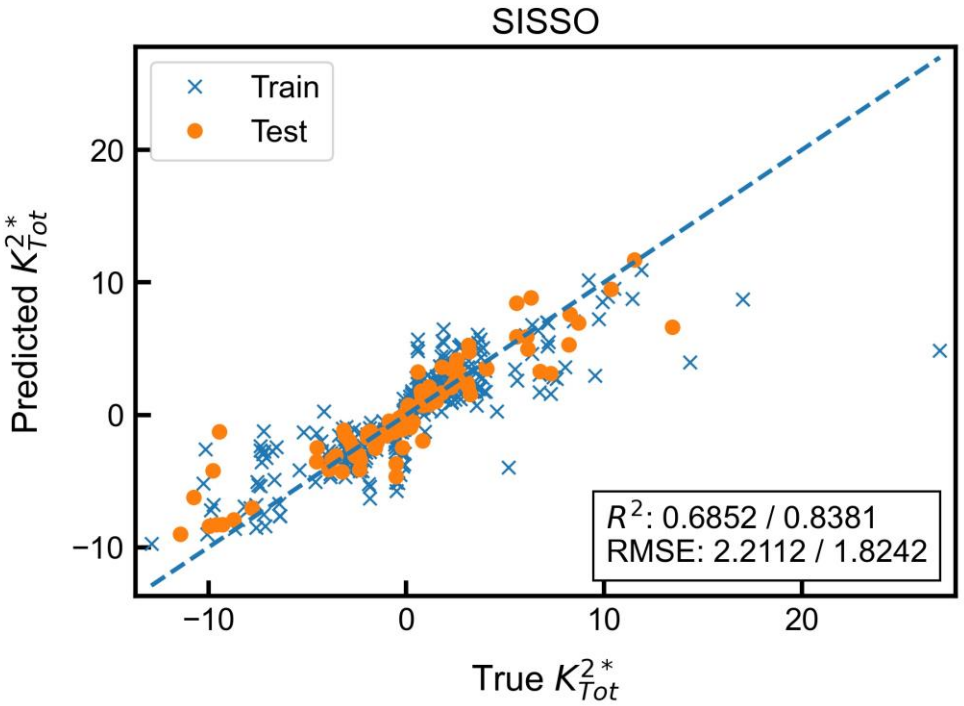
```
SISSO w/o OMP:
 +24689514(sin(charge)*(atomic_states*atomic_tot_elec))  -24688847(atomic_states*(charge*atomic_tot_elec))  +3839(sin(a)*(charge*atomic_states))
Dimension: 3
RMSE: 2.2112/1.8242
R^2: 0.6852/0.8381
*************************************
*               All done            *
*************************************
```

- Analytic equation is obtained with dimension, RMSE and R-squared values.

COMPARISON WITH GENERAL ML

- Applied to Mxenes dataset.
- Plots:



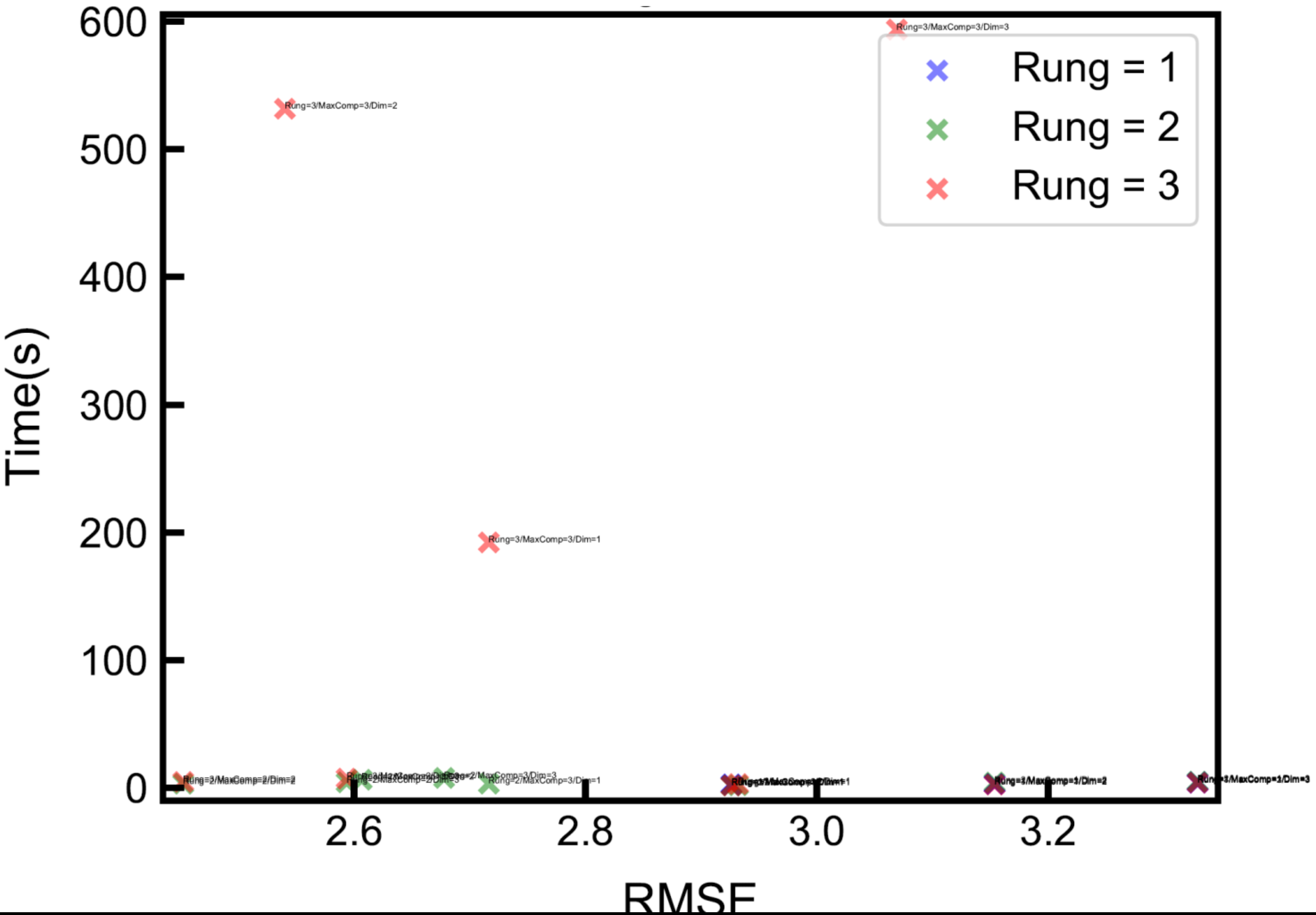


- Performance is comparable, while attaining an analytic equation.

SISSO PERFORMANCE AND RESULTS ANALYSIS

PERFORMANCE PLOT

- Applied to Mxenes dataset with target variable K_{Tot}^{2*}
- Points with varied parameters are plotted.(Rung, Maximum complexity, Dimension)
- Observed that rung = 3 (red) is not necessarily more optimal for accuracy or time as points are distributed evenly. (potentially overfitting)
- Observed that rung = 1 (blue) is less accurate as is distributed towards bottom right.

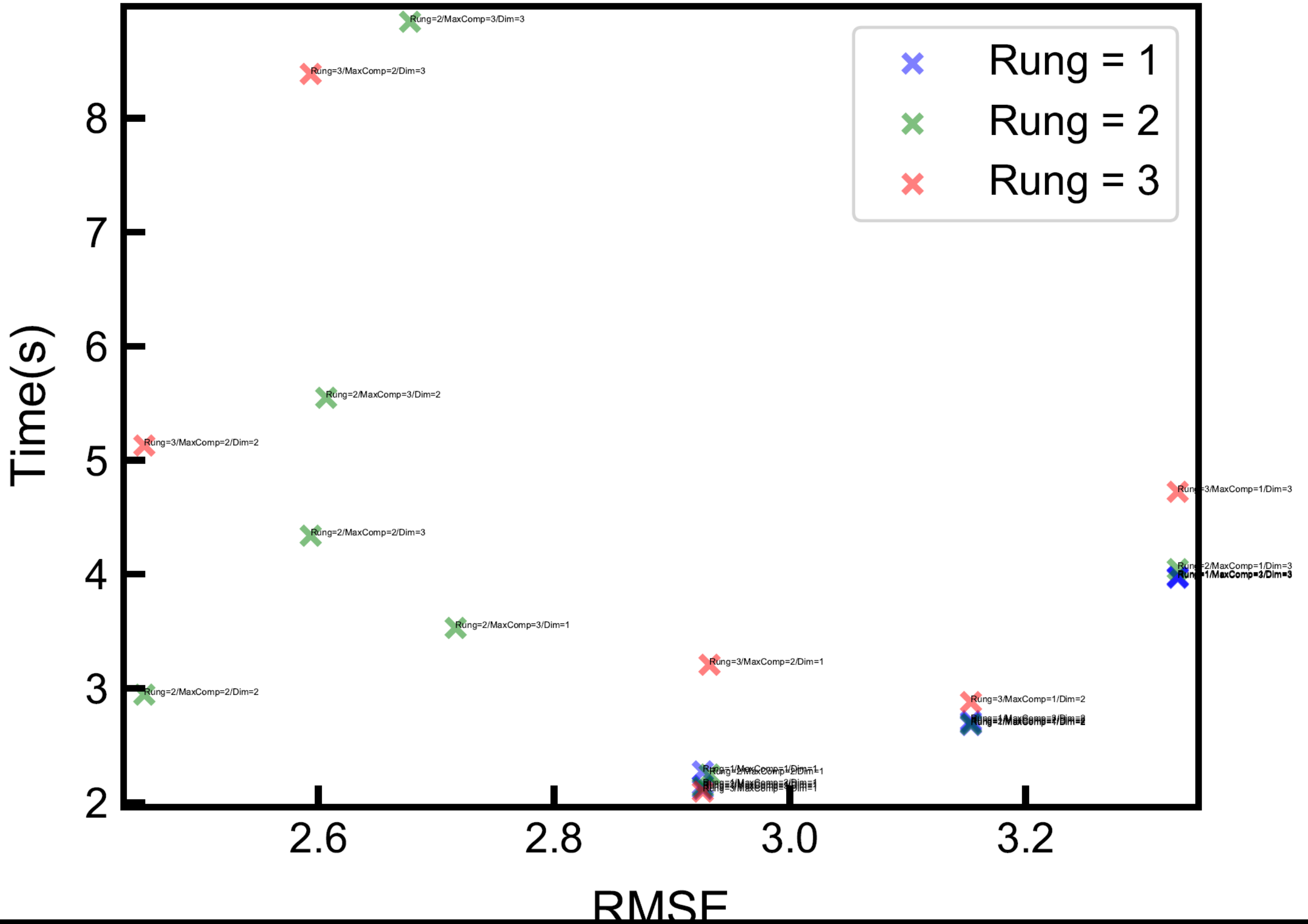


PERFORMANCE PLOT (ZOOMED)

- Zoomed into bottom portion of previous plot

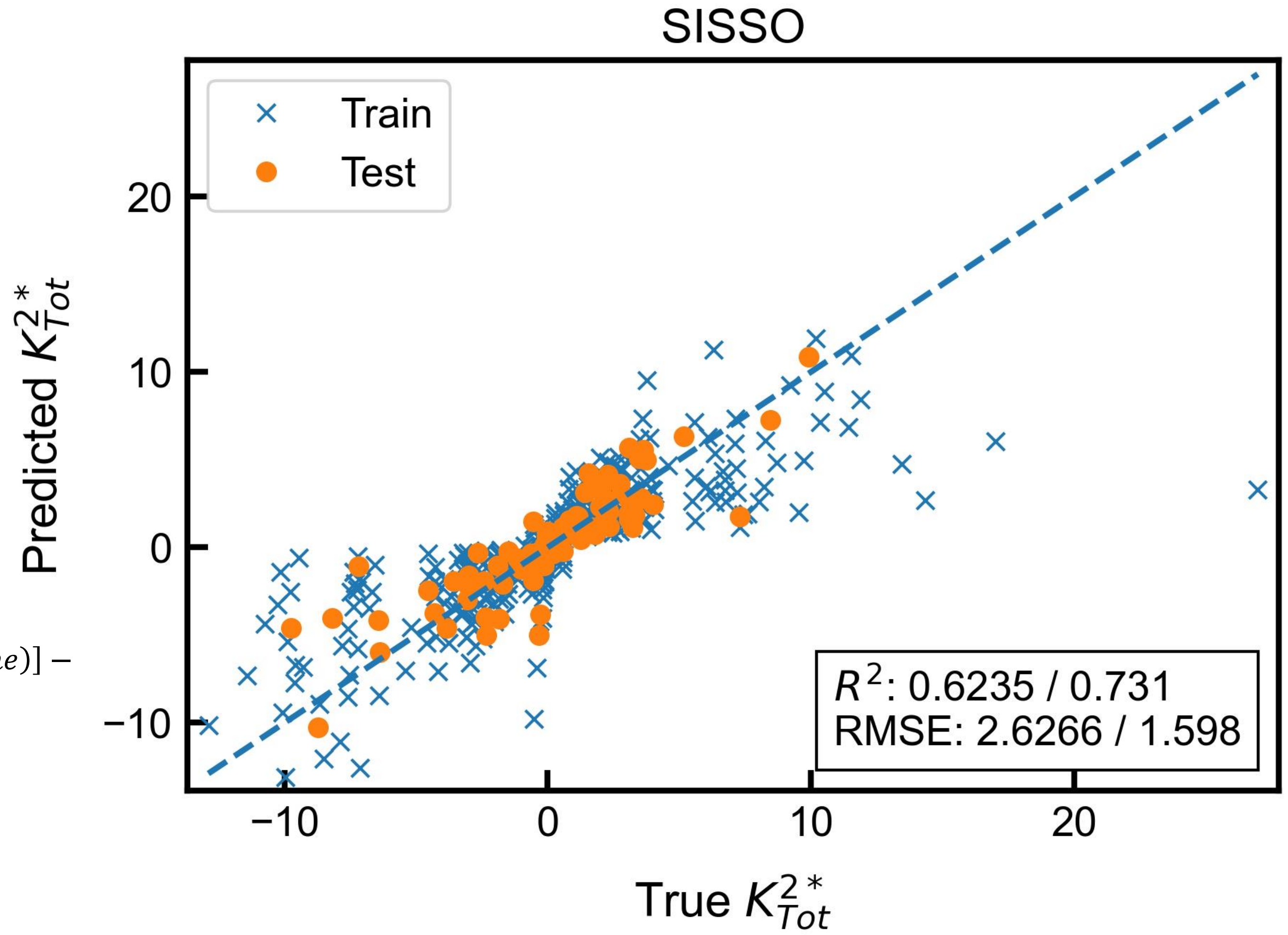
- Observed that rung = 2 (green) is optimal for both time and precision on mxenes dataset.

- Best performance is observed with parameters rung, maximum complexity and dimension = 2. (Bottom left)



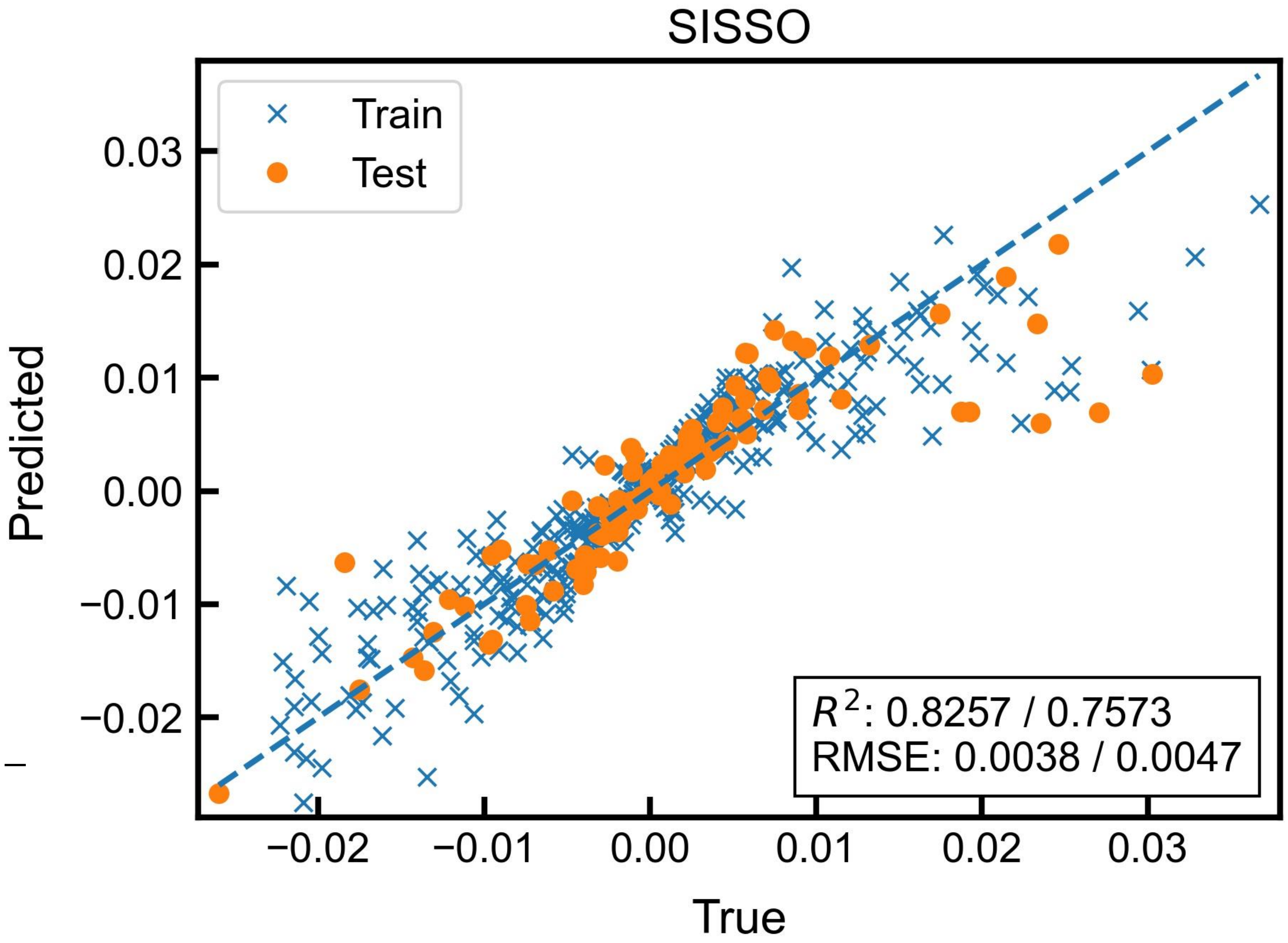
APPLICATION TO MXENES DATA(K^{2*}_{Tot})

- Applied best parameters from analysis (previous slide) to K^{2*}_{Tot}.
- Accuracy is high
- Analytic Equation:
  $6308.041021[atomicstates * (charge * totelecpervolume)] - 8.738623357[(c33 * charge) * atomicstates]$



APPLICATION TO MXENES DATA(EFFICIENCY)

- Applied same parameters to predict effitotsigned. (rung,maxcomp,dim = 2)
- Accuracy is high
- Analytic Equation: $0.008881534082[charge * abs(c11 - c33)] - 4.840361695e - 06[c11 * (mass * charge)]$

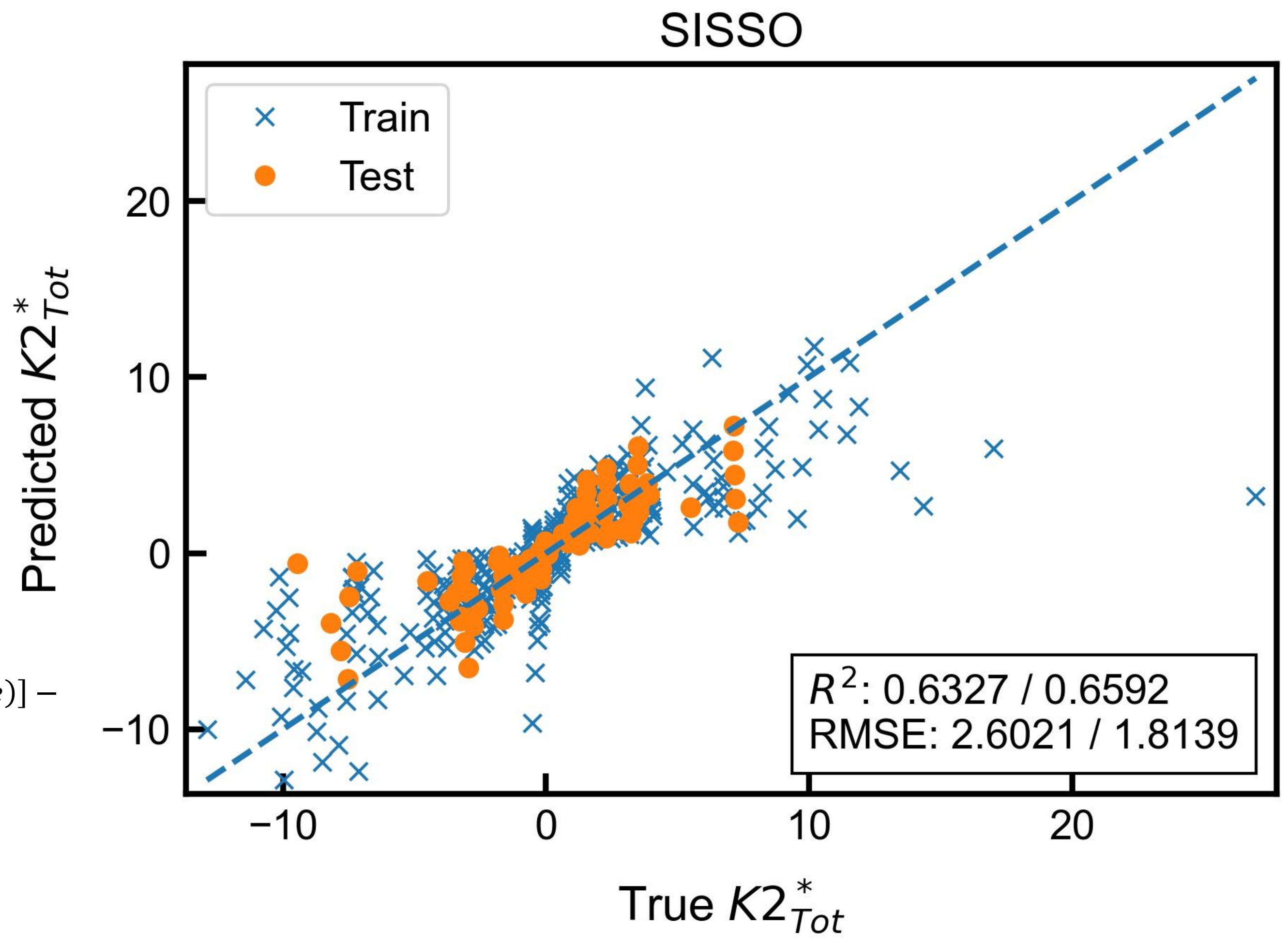


APPLICATION TO STRATIFIED MXENES DATA(K^{2*}_{Tot})

- Applied best parameters from analysis to K^{2*}_{Tot} with stratification by material group.

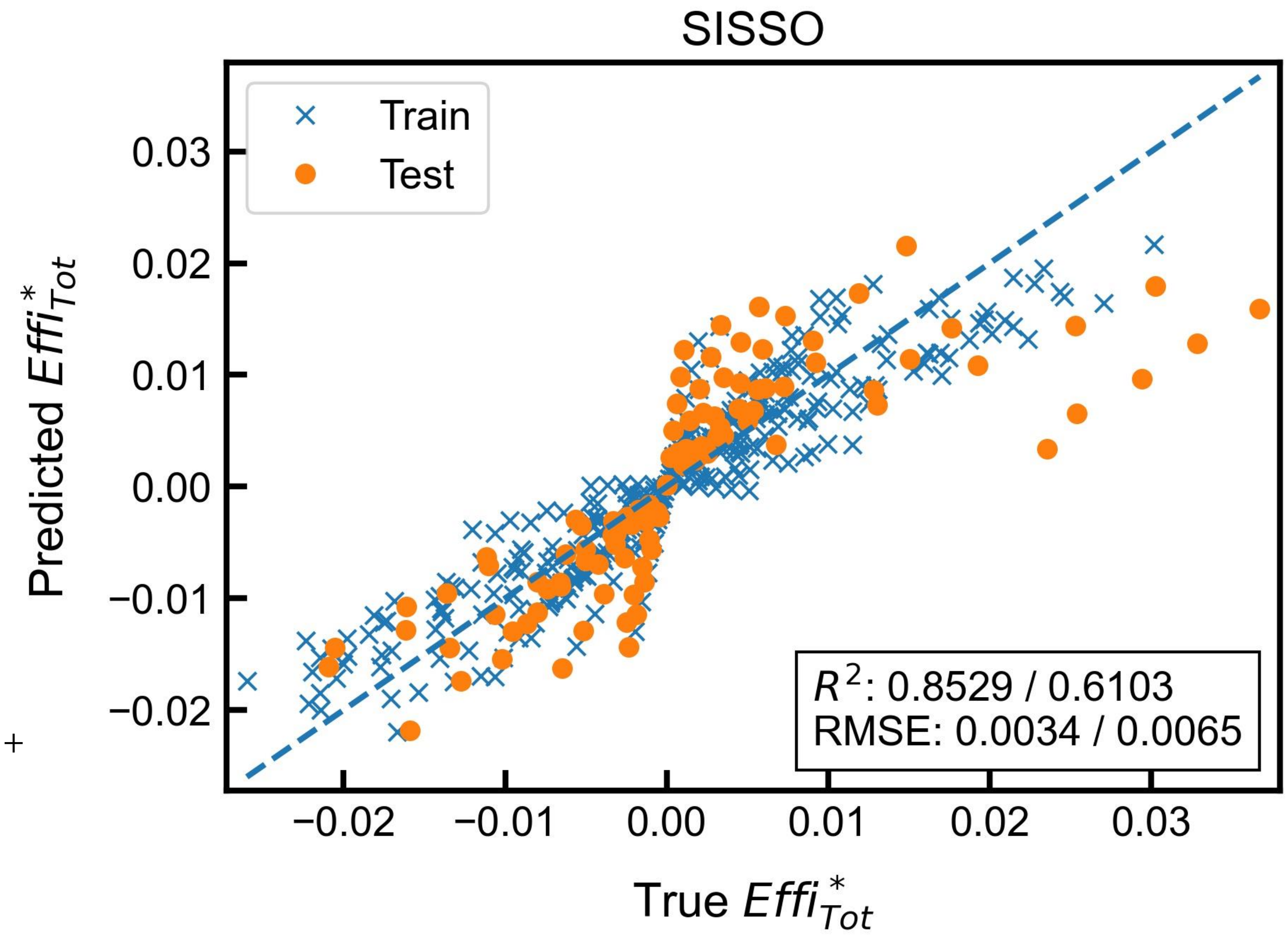
- Accuracy is high

- Analytic Equation:
  $6215.127894[atomicstates * (charge * totelecpervolume)] - 8.597285392[(c33 * charge) * atomicstates]$



APPLICATION TO STRATIFIED MXENES DATA(EFFICIENCY)

- Applied same parameters to predict effitotsigned with stratification by material group. (rung,maxcomp,dim = 2)
- Accuracy is high
- Analytic Equation: $0.3931274926[(charge * atomic\_tot\_elec) - charge] + 0.1873179569[charge * (atomic\_tot\_elec - density)]$

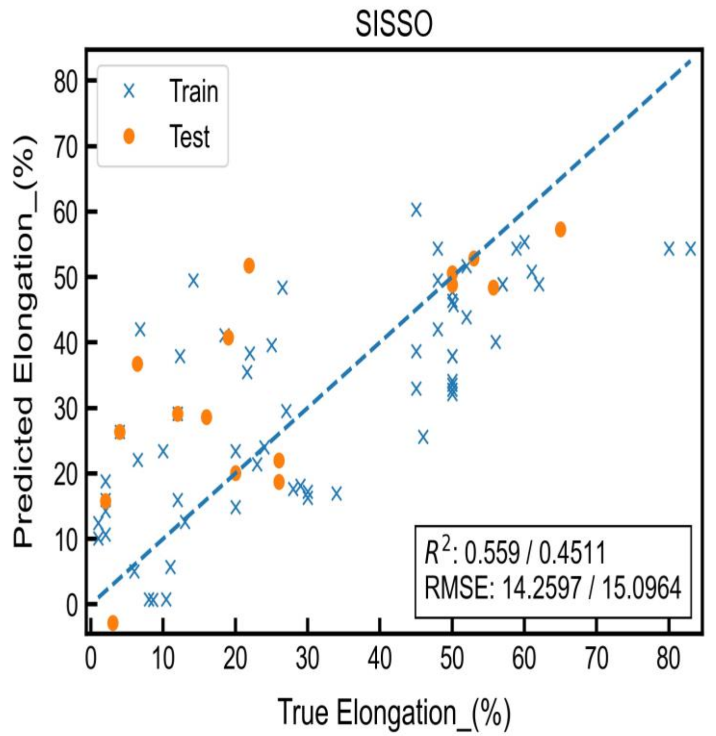
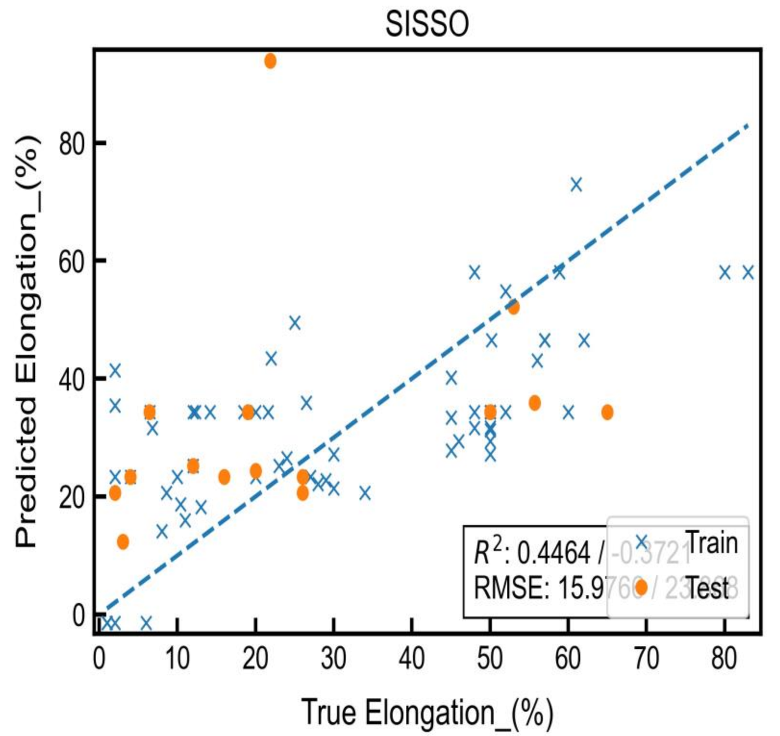
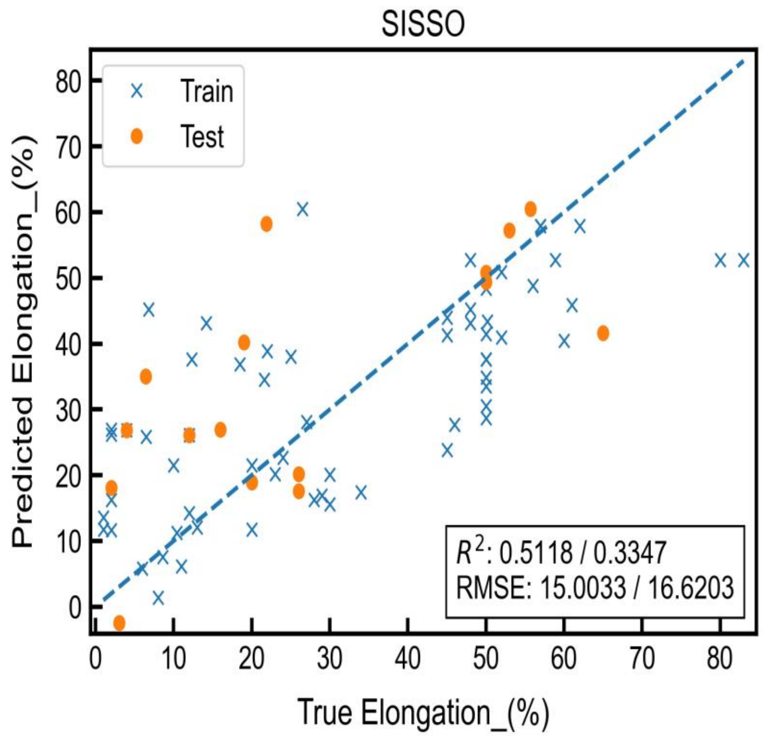


SISSO-HEAS RESULTS ANALYSIS

METHODOLOGY

- Selected target variable. (Elongation, YS, HV, UTS)
- Used OMP on HEAS data to select 5 most correlated features.
- Applied SISSO to the set of selected features, with hyperparameter tuning (range 1 to 3) and 5-fold cross-validation for rung, dimension and maximum complexity. (most frequent optimal combination of parameters: 1,2,1)
- Repeated for HEAS data with only composition features and with only physical features to compare results.

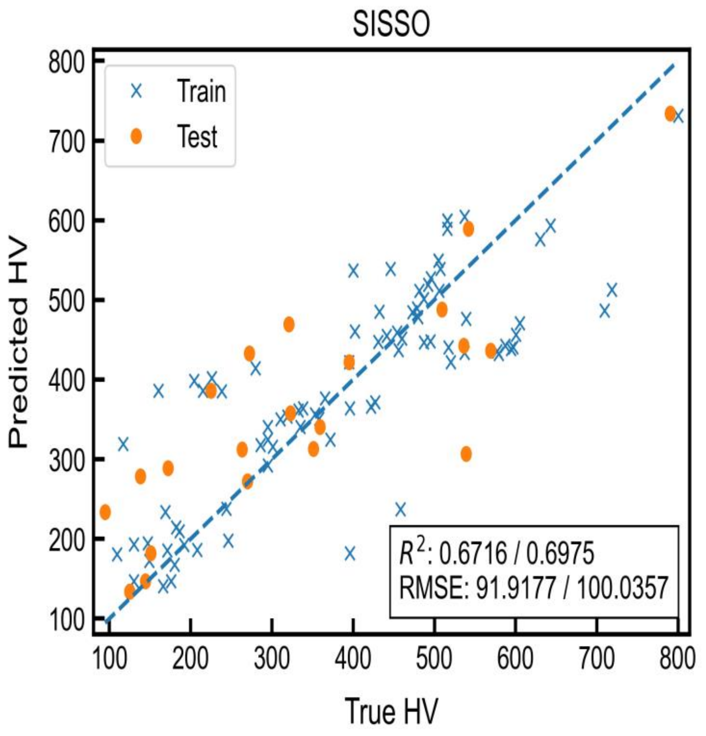
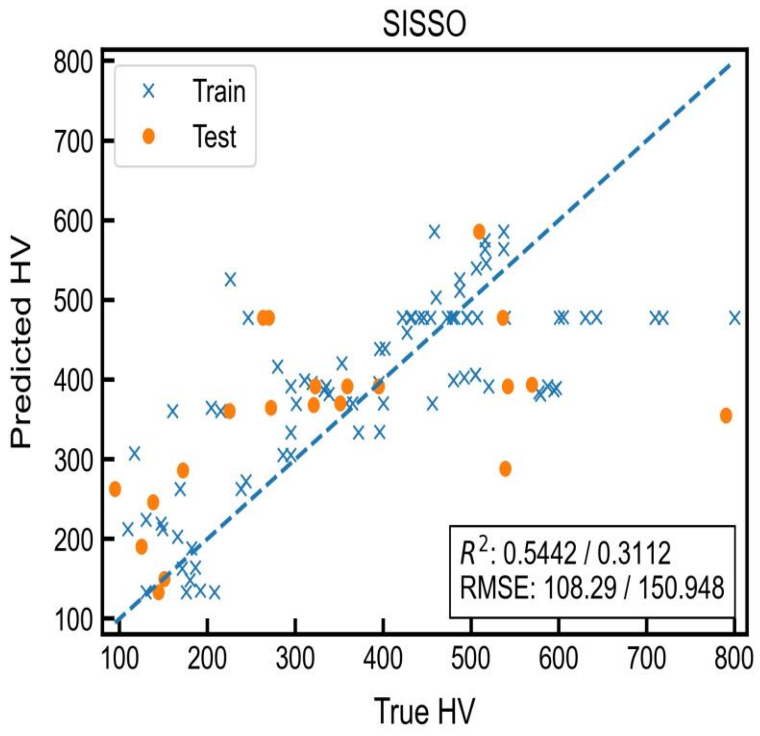
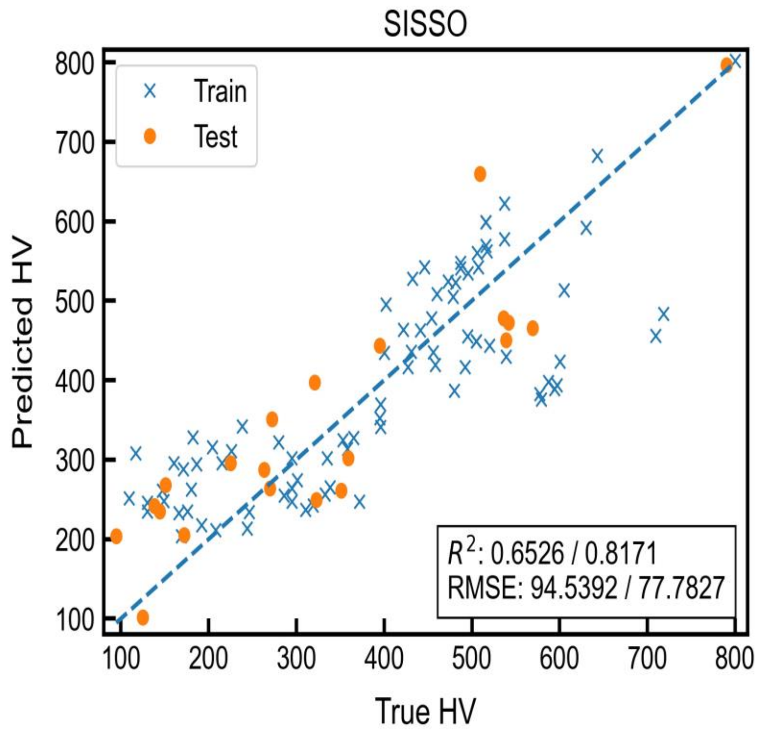
TARGET VARIABLE: ELONGATION(%)

<table>
<tr><th>Dataset</th><th>Whole data</th><th>Composition data</th><th>Physical data</th></tr>
<tr><td>Plot</td><td>
</td><td>
</td><td>
</td></tr>
<tr><td>Equation</td><td>$-52.97817093(Shear\_modulus\_delta - Ni) - 14.76128983(Mixing\_enthalpy * Electronegativity\_local\_mismatch)$</td><td>$-55.07104127(Al + Mo) + 1516.61397(Co)^3$</td><td>$-1.777507734(Mixing\_enthalpy - Interant\_electrons) - 4.300657093(Interant\_electrons * Shear\_modulus\_delta)$</td></tr>
</table>

Accuracy (worst to best): Composition -> Physical -> Whole

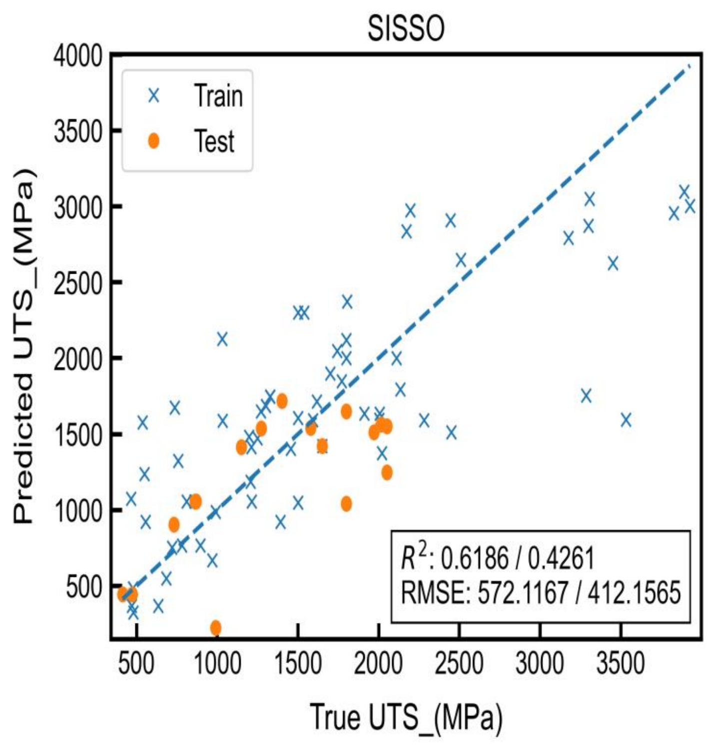
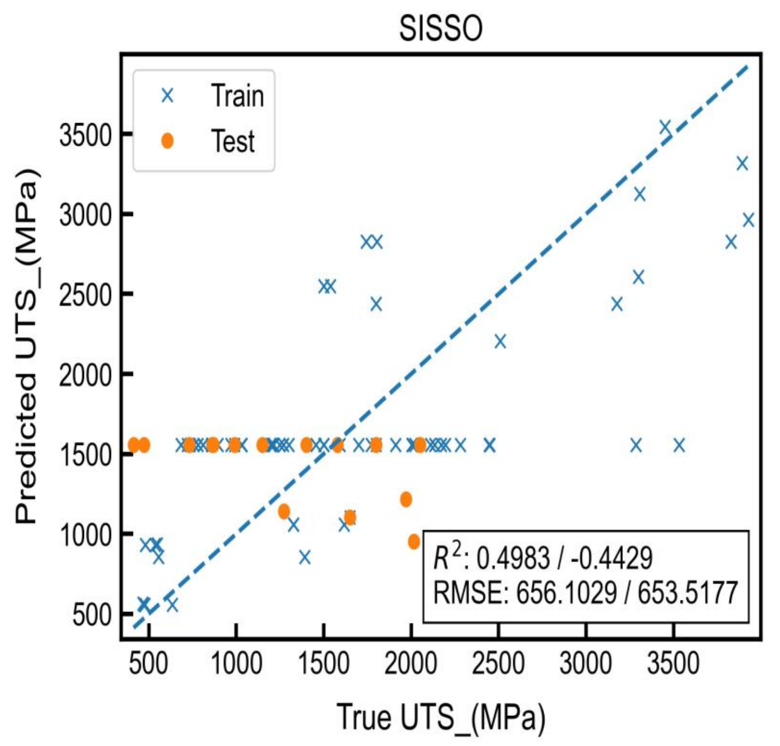
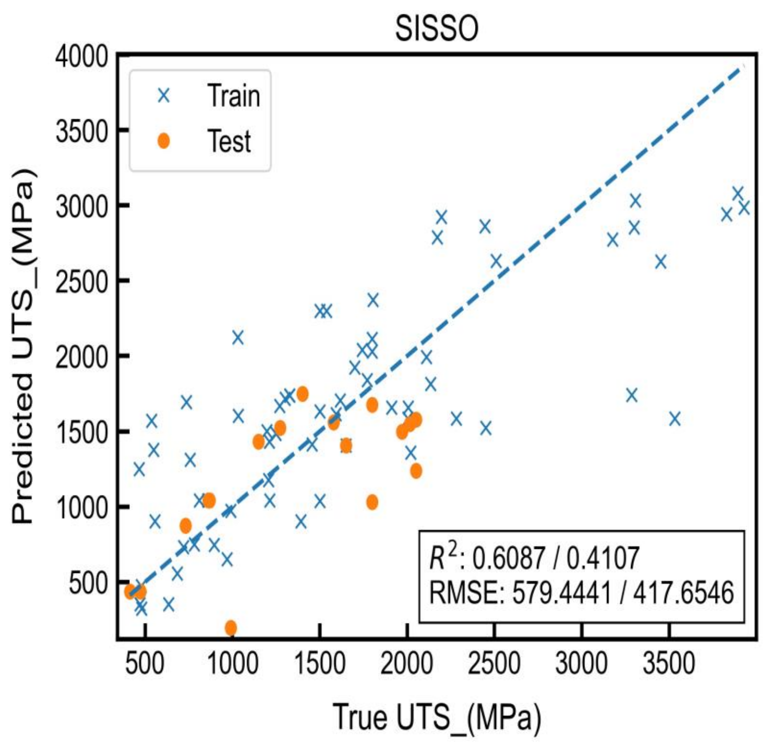
TARGET VARIABLE:

HV

<table>
<tr><th>Dataset</th><th>Whole data</th><th>Composition data</th><th>Physical data</th></tr>
<tr><td>Plot</td>
<td>
</td>
<td>
</td>
<td>
</td></tr>
<tr><td>Equation</td>
<td>$\boldsymbol{4194.979928 exp(Electronegativity\_delta) - 766.5006764(Mn + Shear\_modulus\_strength\_model)}$</td>
<td>$\boldsymbol{-861.5603106(Co + Mn) + 431.6943479(Al - Zr)}$</td>
<td>$\boldsymbol{1754.503458(Electronegativity\_delta - Shear\_modulus\_strength\_model) + 11.33349587(Mixing\_enthalpy - VEC\_mean)}$</td></tr>
</table>

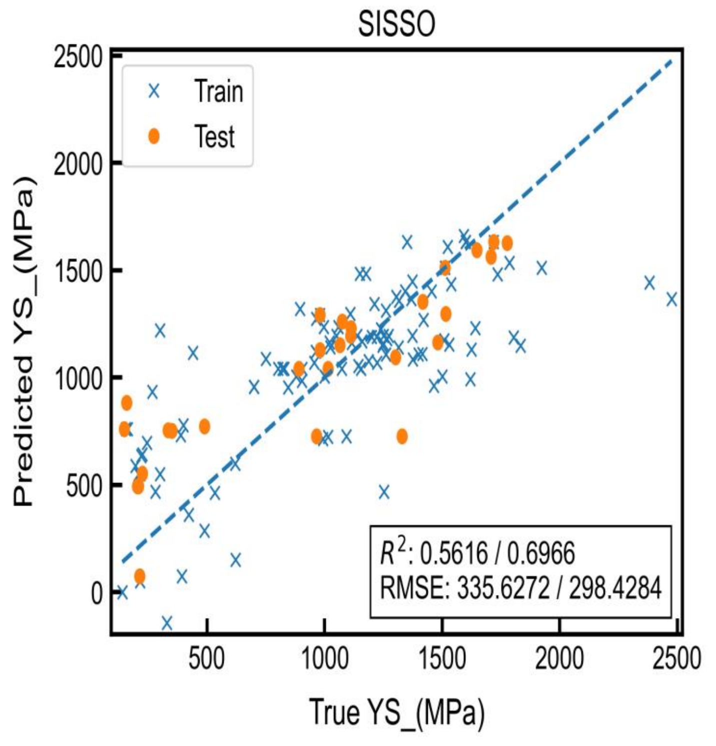
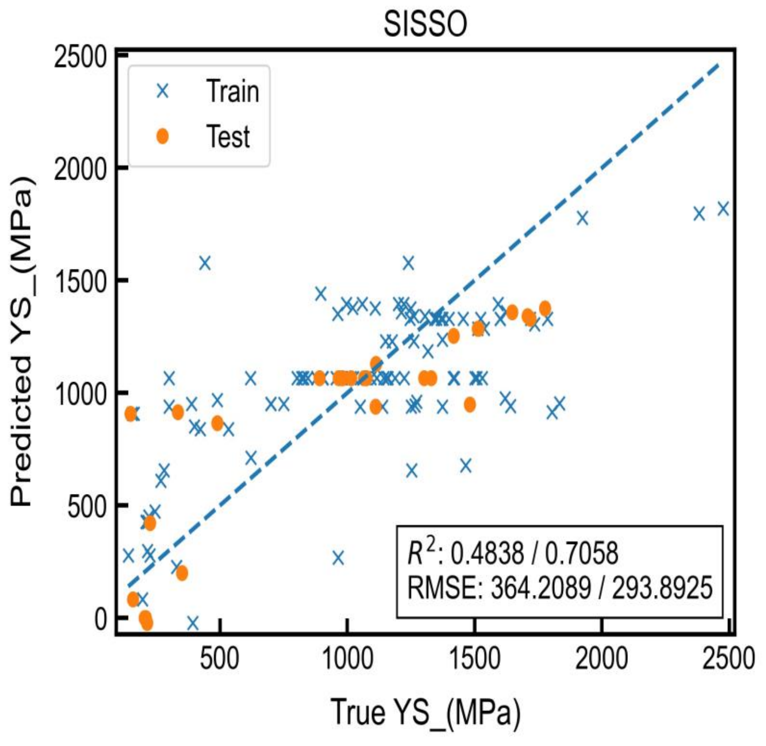
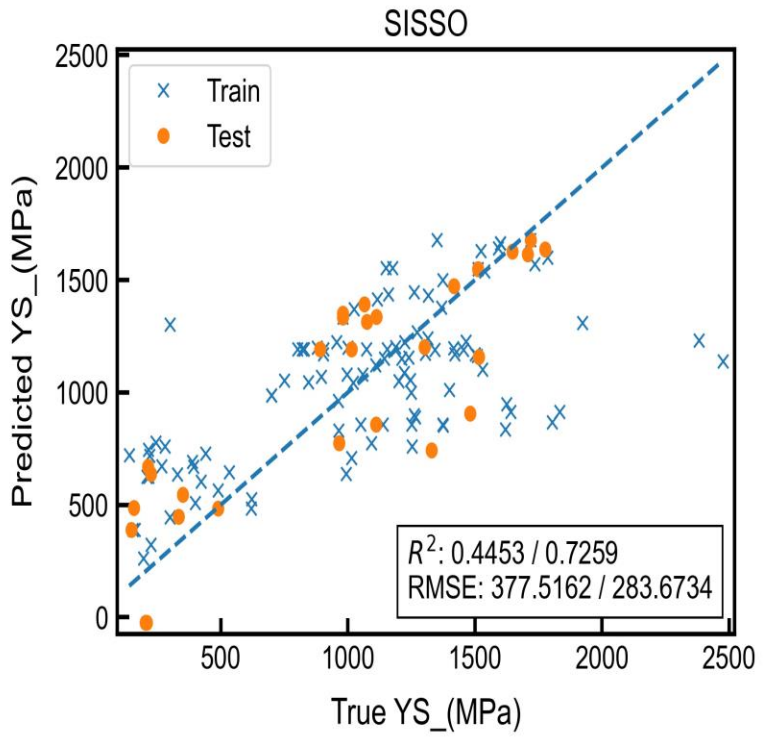
Accuracy (worst to best): Composition -> Whole -> Physical

TARGET VARIABLE: UTS(MPA)

Dataset	Whole data	Composition data	Physical data
Plot			
Equation	$36551.71097(Radii\_gamma * Shear\_modulus\_delta) - 35154.83364(Sn + Shear\_modulus\_delta)$	$31794.64019(Mo * Zr) - 2499.812945(Cu + Mn)$	$35335.70666(Radii\_gamma * Shear\_modulus\_delta) - 33784.28339(Shear\_modulus\_delta)$

Accuracy (worst to best): Composition -> Physical = Whole

TARGET VARIABLE: YS(MPA)

<table>
<tr><th>Dataset</th><th>Whole data</th><th>Composition data</th><th>Physical data</th></tr>
<tr><td>Plot</td><td>
</td><td>
</td><td>
</td></tr>
<tr><td>Equation</td><td>$13265.55335(Electronegativity\_delta - Mg) - 2067.858599(Cu + Mn)$</td><td>$-2561.36509(Mn - Cr) + 1314.036639abs(Co - Mo) - 4508.87323abs(Co - Mg)$</td><td>$9499.973507(Yang\_delta + Electronegativity\_delta) - 0.6182925161(VEC\_mean)^3$</td></tr>
</table>

Accuracy (worst to best): Whole -> Composition -> Physical

NOTABLE RESULTS

- For all targets, using only composition data was less accurate than only physical data.
- For targets HV, YS, whole data was less accurate than only physical. This implies that adding composition data as predictors reduced the overall accuracy.
- Only for target Elongation, whole data was more accurate than physical. This implies that composition data helped improve accuracy.
- For target YS, composition features made up a larger proportion of the features used in whole data compared to other targets. (3 comp. and 1 phy. feat. vs 1 comp. and 3 phy. feat.) However, using whole data performed worse than composition or physical data only in this case.

MULTI-TASK SISSO

OVERVIEW

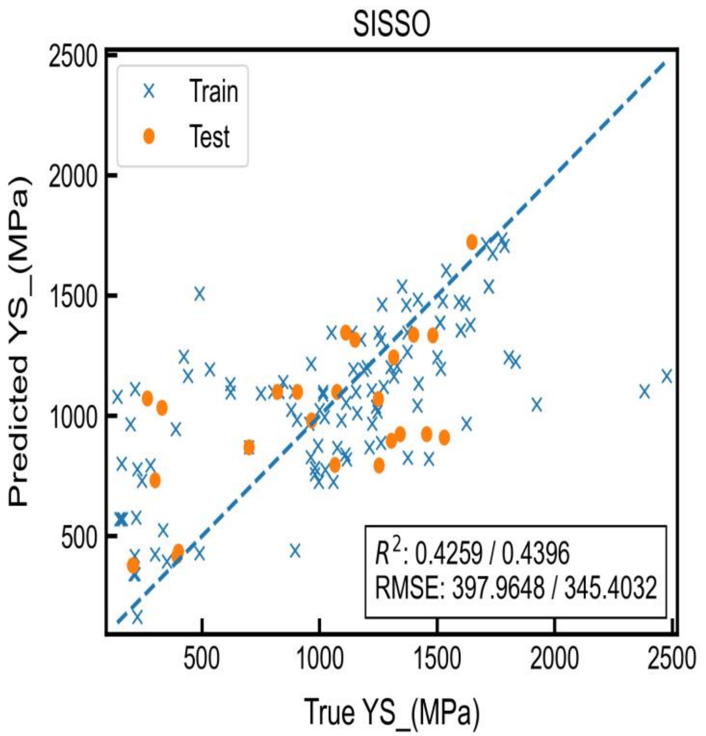
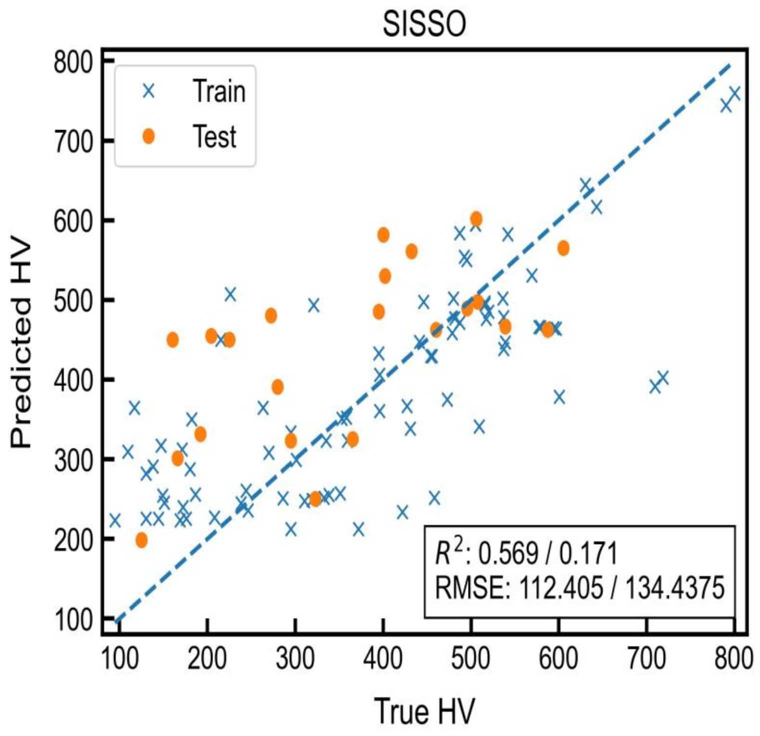
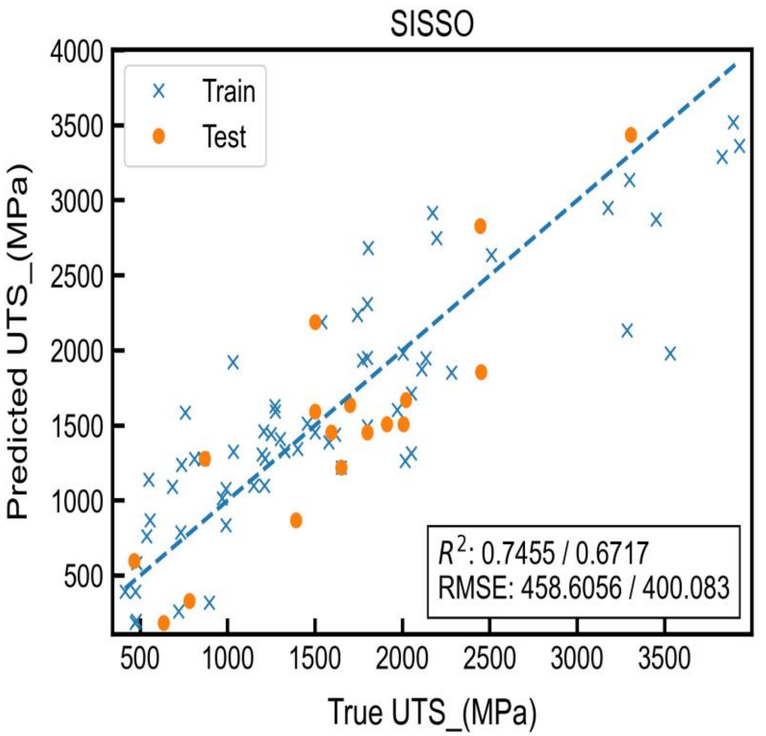
- Takes as inputs, “X” data columns and multiple “Y” target columns.
- For each feature (variable in the equation), single-task SISSO is run for each individual “Y” target column.
- RMSE for that feature is calculated as the average RMSE of each single task run, and the features which minimise RMSE are selected for usage in the equation.
- Generates analytic equation for each target, with a fixed set of features. i.e. All targets share the same features but with different coefficients in their equations.
- Advantages: Predictions are generated for NA values.

APPLICATION TO HEAS DATASET

- “Y” target variables are input as 'YS (MPa)', ’HV’ and 'UTS (MPa)’.
- Performed with only physical data.
- Run with configuration rung = 2, dimension = 2 and complexity 2.
- Run with configuration rung = 1, dimension = 2 and complexity 1.
- Compared with previous results with optimal config:1,2,1 for config:1,2,1.

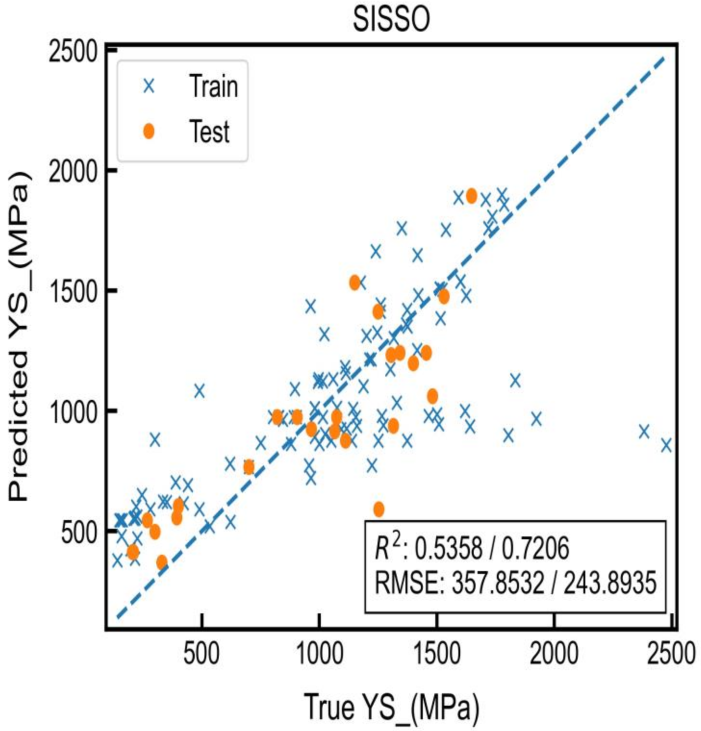
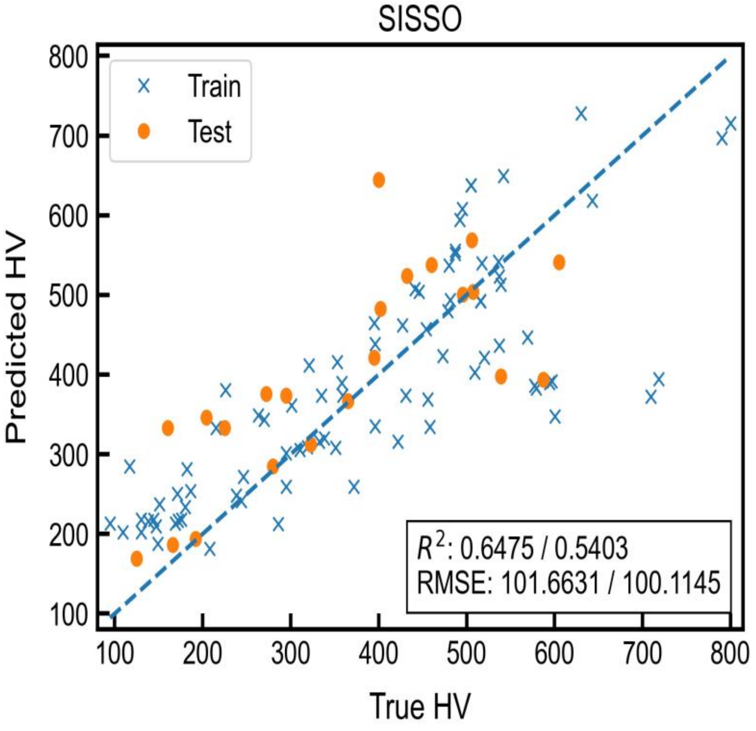
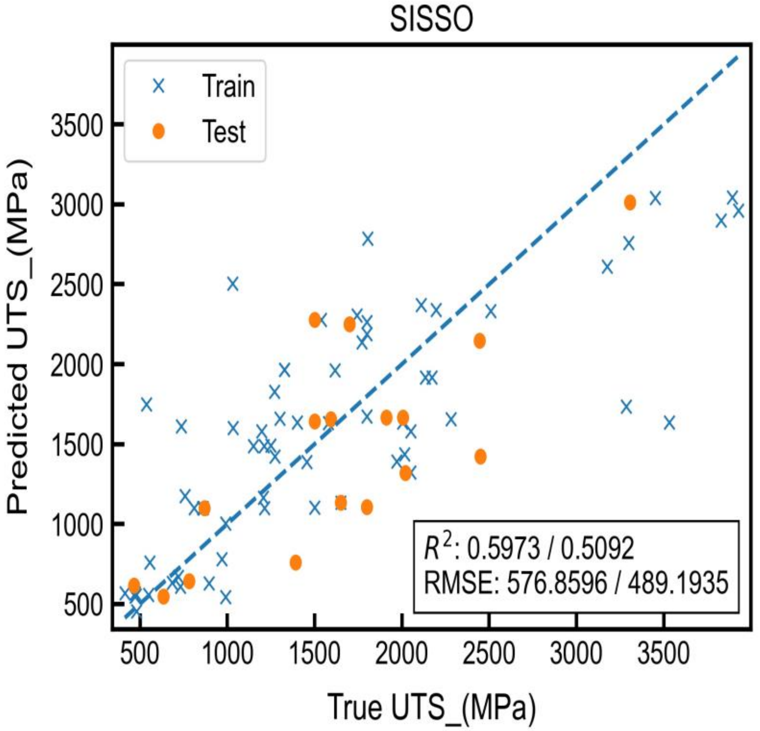
CONFIG:

RUNG = 2, DIM = 2, COMP = 2

<table>
<tr><th>Target</th><th>YS (MPa)</th><th>HV</th><th>UTS (MPa)</th></tr>
<tr><td>Plot</td><td>
</td><td>
</td><td>
</td></tr>
<tr><td>Equation</td><td>$-48827.35389(log(Radii\_gamma) * Shear\_modulus\_strength\_model) - 41.56337773abs(Interant\_d\_electrons - (Mixing\_enthalpy + Shear\_modulus\_local\_mismatch))$</td><td>$-51409.68208(log(Radii\_gamma) * Shear\_modulus\_strength\_model) - 0.6022698101abs(Interant\_d\_electrons - (Mixing\_enthalpy + Shear\_modulus\_local\_mismatch))$</td><td>$-188661.3072(log(Radii\_gamma) * Shear\_modulus\_strength\_model) - 51.10224241abs(Interant\_d\_electrons - (Mixing\_enthalpy + Shear\_modulus\_local\_mismatch))$</td></tr>
</table>

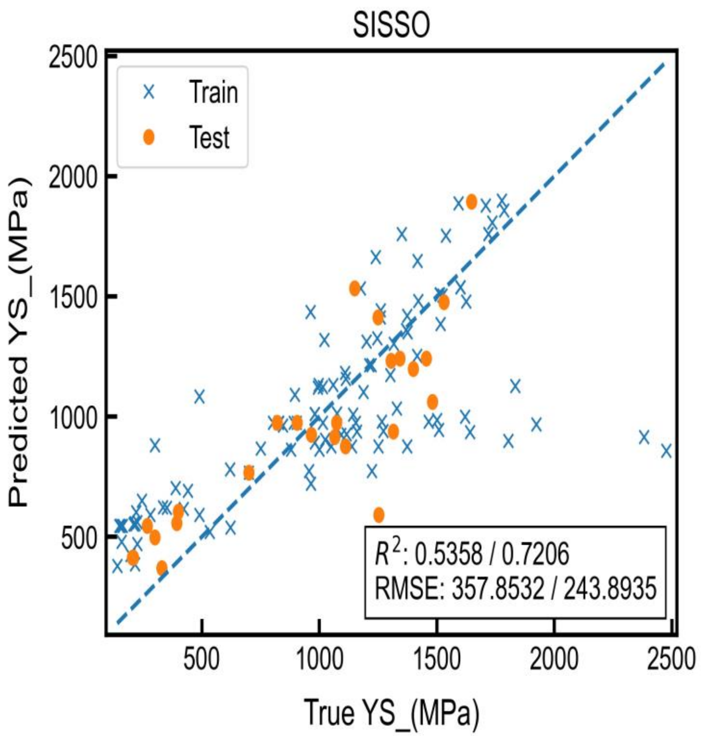
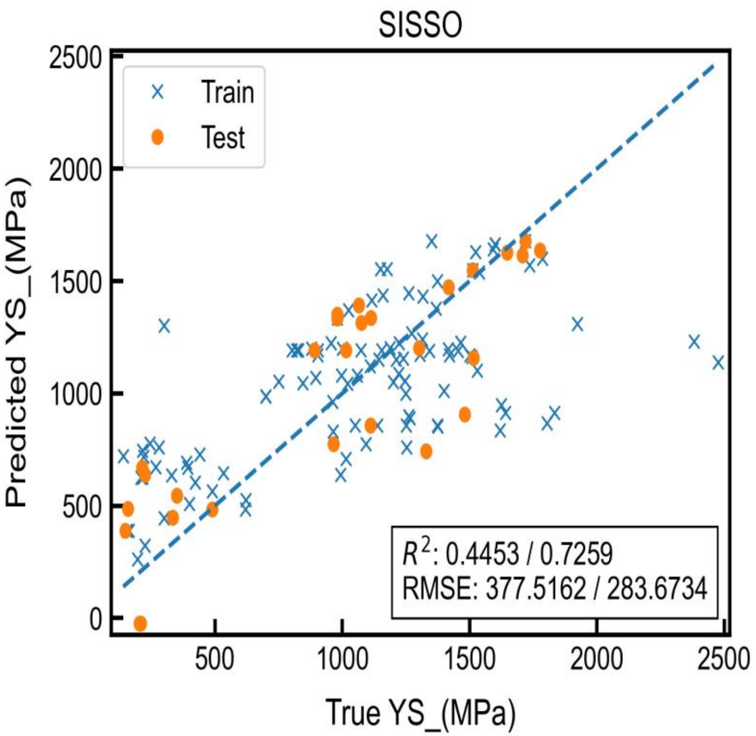
CONFIG:

RUNG=1,DIM=2,COMP=1

<table>
<tr><td>Target</td><td>YS (MPa)</td><td>HV</td><td>UTS (MPa)</td></tr>
<tr><td>Plot</td><td>
</td><td>
</td><td>
</td></tr>
<tr><td>Equation</td><td>$$-151253.3373(Yang\_delta * Shear\_modulus\_strength\_model) -127.484471 abs(VEC\_mean - Mean\_cohesive\_energy)$$</td><td>$$-77254.61692(Yang\_delta * Shear\_modulus\_strength\_model) -30.73113347 abs(VEC\_mean - Mean\_cohesive\_energy)$$</td><td>$$-348775.8453(Yang\_delta * Shear\_modulus\_strength\_model) -25.31638202 abs(VEC\_mean - Mean\_cohesive\_energy)$$</td></tr>
</table>

COMPARISON:

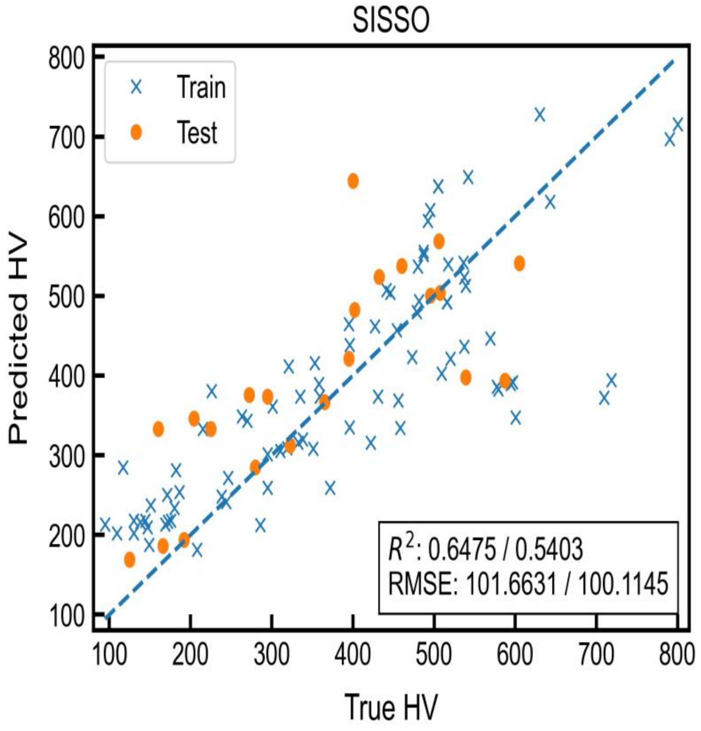
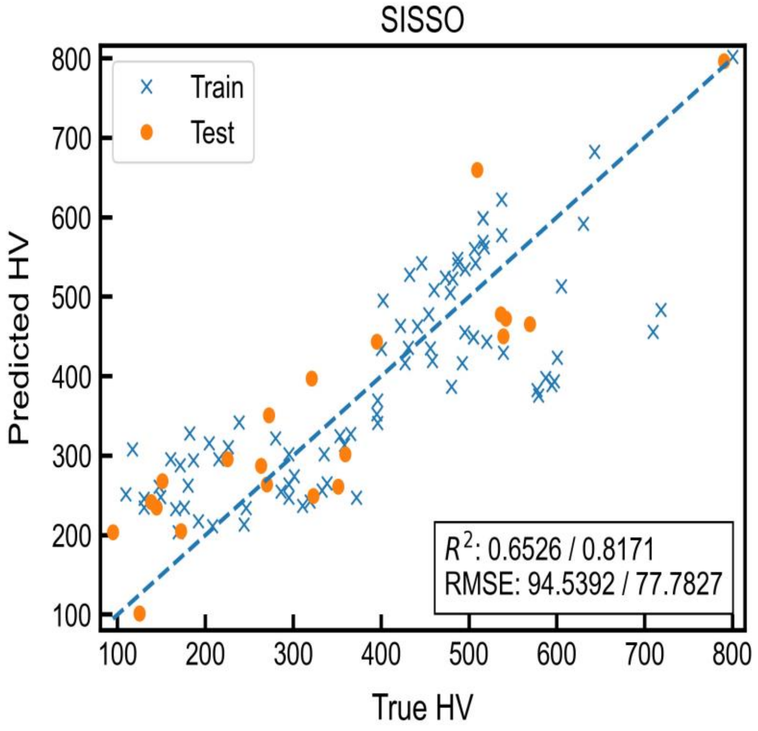
YS(MPA)

SISSO Type	Multi-task	Single-task
Plot		
Equation	$-151253.3373(Yang\_delta * Shear\_modulus\_strength\_model) - 127.484471 abs(VEC\_mean - Mean\_cohesive\_energy)$	$9499.973507(Yang\_delta + Electronegativity\_delta) - 0.6182925161(VEC\_mean)^3$

- Configuration: 1,2,1 is used for both.
- Multi-task performs better in comparison to single-task.
- YS data has 56 nan values out of 193. i.e. 29% nan values.

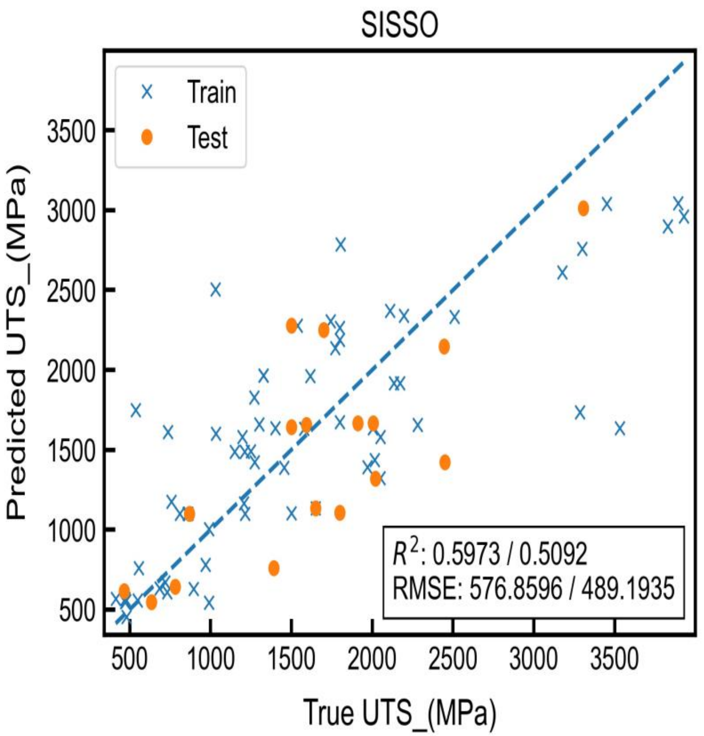
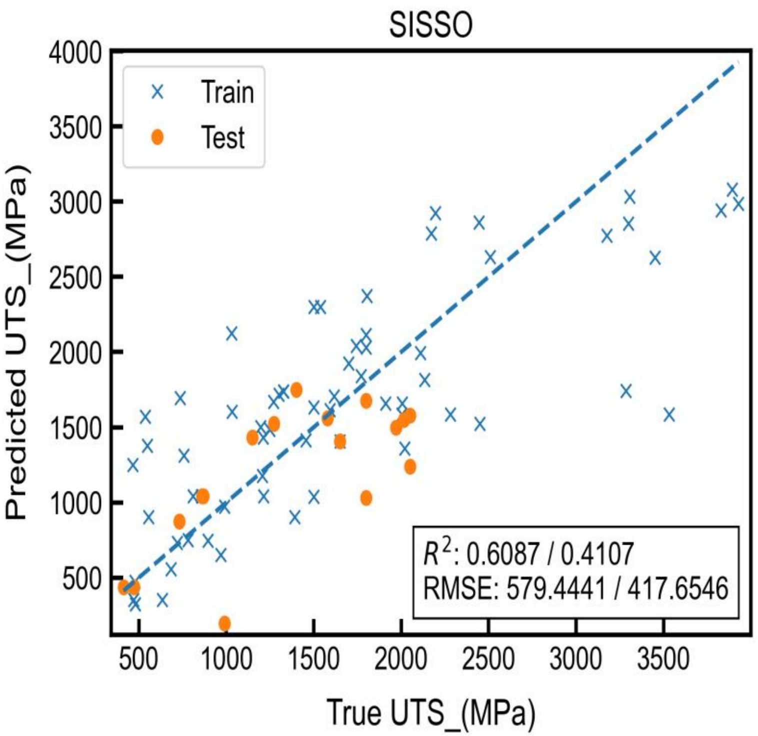
COMPARISON:

HV

SISSO Type	Multi-task	Single-task
Plot		
Equation	$-77254.61692(Yang\_delta * Shear\_modulus\_strength\_model) -30.73113347abs(VEC\_mean - Mean\_cohesive\_energy)$	$1754.503458(Electronegativity\_delta - Shear\_modulus\_strength\_model) + 11.33349587(Mixing\_enthalpy - VEC\_mean)$

- Configuration: 1,2,1 is used for both.
- Multi-task performs worse in comparison to single-task.
- HV data has 86 nan values out of 193 entries. i.e. 44.5% nan values.

COMPARISON: UTS(MPA)

SISSO Type	Multi-task	Single-task
Plot		
Equation	$-348775.8453(Yang\_delta * Shear\_modulus\_strength\_model) - 25.31638202 abs(VEC\_mean - Mean\_cohesive\_energy)$	$35335.70666(Radii\_gamma * Shear\_modulus\_delta) - 33784.28339(Shear\_modulus\_delta)$

- Configuration: 1,2,1 is used for both.
- Multi-task performs worse in comparison to single-task.
- UTS data has 111 nan values out of 193 entries. i.e. 57.5% nan values.

NOTABLE RESULTS

- It can be inferred that multi-task performance is close to single-task.
- It appears that columns with fewer missing entries benefit more from multi-task SISSO.
- Similarly, it appears that columns with more missing entries lose accuracy from using multi-task SISSO. It could be that the larger target datasets corrupt the accuracy by holding a larger influence on the selection of features.
- It is to be noted for the comparison that multi-task had been run without OMP (with multiple targets, it might not be possible to specify correlation in the given software) and without cross-validation (due to software limitations). Having both would further help validate the hypotheses above as single-task had been run with both.